MODELOS DE AUTORIA NÃO-BOOLEANOS PARA BUSCA DE ESPECIALISTAS NA ACADEMIA

VÍTOR MANGARAVITE

MODELOS DE AUTORIA NÃO-BOOLEANOS PARA BUSCA DE ESPECIALISTAS NA ACADEMIA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

ORIENTADOR: RODRYGO LUIS TEODORO SANTOS

Belo Horizonte

Junho de 2016

Mangaravite, Vítor

M277m Modelos de autoria não-booleanos para busca de especialistas na academia / Vítor Mangaravite. — Belo Horizonte, 2016
xx, 95 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Rodrygo Luis Teodoro Santos

1. Computação — Teses. 2. Sistemas de recuperação da informação. 3. Mineração de dados (computação). 4. Banco de dados - Busca.
I. Orientador. II. Título.

CDU 519.6*72 (043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Modelos de autoria não-booleanos para busca de especialistas na academia.

VÍTOR MANGARAVITE

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. RODRYGO LUIS TEODORO SANTOS - Orientador
Departamento de Ciência da Computação - UFMG

PROF. ALBERTO HENRIQUE FRADE LAENDER
Departamento de Ciência da Computação - UFMG

PROF. LEANDRO BALBY MARINHO
Departamento de Sistemas e Computação - UFCG

PROF. MARCOS ANDRÉ GONÇALVES
Departamento de Ciência da Computação - UFMG

PROF. MARIO SÉRGIO FERREIRA ALVIM JÚNIOR
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 01 de junho de 2016.

Dedico este trabalho a Maria Luisa Santos Mangaravite, Maria Júlia de Medeiros Mangaravite, Cristiane dos Santos Silveira e Érica Mangaravite.

Agradecimentos

Primeiramente gostaria de agradecer aos meus pais: Maria Julia e José Carlos, por terem me motivado a começar esse investimento a longo prazo, que é a vida acadêmica. Eles foram primordiais para eu não desistir pelo caminho, e fundamentais para eu me estabilizar durante minha estadia em Belo Horizonte.

Não poderia deixar de agradecer à minha namorada Cristiane, por ter sido tantas vezes meu ombro amigo, conselheira e, por vezes, até consultora acadêmica. Pacificou meu coração de forma que nenhuma pessoa conseguiria fazer.

Agradeço, inclusive, a quem ainda nem entende o que seria mundo acadêmico: minha filha, Maria Luisa. Ela foi o fator motivador para insistir, quando a guerra parecia vencida. Foi daquele sorriso ingênuo que surgiu meu gosto real pela ciência e compreensão dos modelos apresentados nesta dissertação. Agradeço também aos meus irmãos: Érica e Igor, que tornaram esse momento da minha vida intenso, graças aos poucos momentos que pudemos rir juntos!

Agradeço à dedicação do meu orientador, Professor Rodrygo L.T. Santos, muitas vezes solícito e de invejável didática. Além dele, agradeço ao professor Alberto H. F. Laender, com o qual iniciei minha jornada no programa de mestrado na UFMG e que foi fundamental para escolha do tópico de pesquisa escolhido.

Não poderia deixar de agradecer também aos meus mentores e ídolos: Professores Anderson Almeida Ferreira, Guilherme T. Assis e Álvaro Pereira Jr. Foi graças a eles que tive espaço na academia e não poderiam ser esquecidos no passado, sendo tão responsáveis pelo meu presente. Agradeço, também a todos meus professores do programa de pós-graduação da UFMG, sem exceções, pois, por mais difíceis que suas disciplinas tenham sido, foram pilar para a vitória da guerra travada.

Tenho que fazer um agradecimento muito especial aos amigos no Laboratório de Banco de Dados (LBD) e, atual, CS+X pelas muitas conversas descontraídas. Meu carinho por esse *old* LBD é infinito e intensifico meus agradecimentos para os T[h]iago's, Amir e a galera do MLCT (*Machine Learning, Computer Thieves*): Sérgio, Clébson, Daniel e Hasan. Além disso, agradeço à galera que passou e formou ou partiu para o

CS+X: Harlley Lima, Isac, Michele Brito e Brandão, Rodrigo, Guilherme's, Natália e Lizardo.

Agradeço também a galera do LATIN. Tantas vezes tão desesperados com *deadline* quanto eu: Raul, Sabir e Alberto; obrigado a todos. Outros desesperados, foram os que moraram comigo. Eu agradeço a cada um de vocês, foram uma família durante a jornada: Carlos, Bráulio, *little* Jordan, Wellington e, mais recentemente, Juvenil (Gustavo).

Um agradecimento aos amigos que estão longe: Mayara, Japão, Dorley e Ramon's, que sempre acompanharam minhas lutas de perto e aos amigos da República Badalação e de Ouro Preto, com os quais fugia desse mundo acadêmico-caótico para descontração sem complexidades, sempre que surgia uma oportunidade.

Por último, mas não menos importante, agradeço ao Trello[1] e ao Francesco Cirillo, por criar, no final dos anos 1980, a técnica pomodoro.

[1] `http://trello.com/`

“ O cientista virou um mito. E todo mito é perigoso,
porque induz o comportamento e inibe o pensamento.
Esse é um dos resultados engraçados (e trágicos) da ciência.”
(Rubem Alves, Filosofia da Ciência)

Resumo

O problema de busca de especialistas visa recuperar *candidatos a especialistas* dada uma consulta em linguagem natural. As abordagens do estado-da-arte para busca de especialistas dependem de associações documento-candidato para inferir a expertise de uma pessoa para uma determinada consulta. Essas associações têm sido tradicionalmente modeladas como variáveis booleanas, indicando se um candidato é ou não autor de um determinado documento, sendo o peso dessa associação normalizado para penalizar candidatos prolíficos. Nesta dissertação, abordamos o problema de busca de especialistas em um ambiente acadêmico, onde a autoria de um documento pode ser determinada com razoável confiança. Assim, em contraste às abordagens tradicionais, propomos modelar associações como variáveis não-booleanas, refletindo a probabilidade de um documento ser informativo para a especialidade de um candidato. Além disso, introduzimos um esquema de normalização alternativo que mede o quão discriminativa uma associação documento-candidato é à luz de todas as associações que envolvem o documento ou o candidato. Através de um estudo de grande escala com acadêmicos especialistas de diversas áreas do conhecimento, demonstramos o desempenho das funções de associação e de normalização propostas para melhorar a eficácia de uma abordagem do estado-da-arte para busca de especialistas.

Palavras-chave: Busca de Especialistas, Recuperação de Informação.

Abstract

The goal of an expert search system is to retrieve candidate experts given a query expressed in natural language. State-of-the-art expert search approaches rely on document-person associations to infer the expertise of a candidate person for a given query. Such associations have traditionally been modeled as boolean variables, indicating whether or not a candidate authored a document, and further normalized to penalize prolific authorships. In this paper, we address expert search in academia, where the authorship of a document can be determined with reasonable certainty. In contrast to traditional approaches, we propose to model associations as non-boolean variables, reflecting the probability that a document is *informative* of the expertise of a candidate. Moreover, we introduce an alternative normalization scheme that measures how *discriminative* a particular document-person association is in light of all associations involving either the document or the person. Through a large-scale user study with academic experts from several areas of knowledge, we demonstrate the suitability of the proposed association and normalization schemes to improve the effectiveness of a state-of-the-art expert search approach.

Keywords: Expert Search, Information Retrieval.

Lista de Tabelas

2.1 Principais características das coleções de teste de busca de especialistas. . 25

4.1 Estatísticas das coleções de documentos e associações. 44
4.2 Exemplo de consultas e suas especificidades. 53
4.3 Ordem dos parâmetros e faixa de valores escolhidos para procedimento de treino dos modelos de L2R. 58

5.1 Resultados dos rankings dos *baselines* e das funções de associação propostas usando o modelo generativo e as normalizações tradicionais. 65
5.2 Taxa de *win-loss* dos resultados dos rankings do modelo generativo. 66
5.3 Resultados dos rankings das funções de normalizações propostas usando o modelo generativo. 70
5.4 Resultados dos modelos discriminativos baseados em agregação de ranking. 74
5.5 Resultados do *win-loss* das métricas para os modelos discriminativos baseados em agregação de ranking. 74
5.6 Resultados dos modelos discriminativos baseados em agregação de todos os rankings. 75
5.7 Resultados do *win-loss* das métricas para os modelos discriminativos baseados em agregação de todos os rankings. 76

A.1 Tabela da comparação das instanciações das funções de associação de dominância. 88

Sumário

Agradecimentos **ix**

Resumo **xiii**

Abstract **xv**

Lista de Tabelas **xvii**

1 Introdução **1**
1.1 Motivação 1
1.2 Argumento da Dissertação 5
1.3 Contribuições da Dissertação 5
1.4 Origens do Material 7
1.5 Organização da Dissertação 7

2 Trabalhos Relacionados **11**
2.1 Modelos de Ranking 11
2.1.1 Modelos de Votação 12
2.1.2 Modelos Probabilísticos 12
2.1.3 Modelos Baseados em Grafos 17
2.2 Modelos de Associação 19
2.2.1 Funções de Associação ρ 20
2.2.2 Funções de Normalização ψ 23
2.3 Avaliação de Rankings 24
2.3.1 Coleções de Teste 24
2.3.2 Métricas de Avaliação dos Resultados 26

3 Modelos de Associação Documento-Candidato **29**
3.1 Funções de Associação 30

3.1.1 Dominância de Conteúdo 31
3.1.2 Estabilidade da Dominância 33
3.1.3 Novidade no Domínio 35
3.1.4 Recência 35
3.2 Normalização Centrada na Auto-informação 40

4 Metodologia de Avaliação 43
4.1 Coleção de Teste 43
4.1.1 Construção dos Conjuntos de Documentos e Associações 43
4.1.2 Construção do Gabarito das Consultas 48
4.2 Configurações Iniciais e *Baselines* 54
4.3 Procedimentos de Treino e Teste 58

5 Avaliação Experimental 61
5.1 Questões de Pesquisa 61
5.2 Resultados Experimentais: Modelos Generativos 62
5.2.1 Funções de Associação 62
5.2.2 Funções de Normalização 67
5.3 Resultados Experimentais: Modelos Discriminativos 71
5.3.1 Análise de Correlação 72
5.3.2 Modelo Baseado em Atributos Agregados 73

6 Conclusões e Trabalhos Futuros 79
6.1 Conclusões 79
6.2 Trabalhos Futuros 80

Referências Bibliográficas 81

Apêndice A Comparação da Entropia Cruzada 87

Apêndice B Classificação de Especialidade 91

Apêndice C Correlação de Funções de Associação 93

Capítulo 1

Introdução

1.1 Motivação

Nos últimos anos, várias abordagens de busca de especialistas vêm sendo propostas na literatura. O problema de busca de especialistas visa recuperar *candidatos a especialistas* dada uma consulta em linguagem natural. Em geral, abordagens que propõem soluções para esse problema geram uma ordem para esses candidatos através da relevância dos documentos associados a cada um desses candidatos, onde um candidato é definido de acordo com o ambiente do problema. Por exemplo, um candidato e seus documentos podem ser representados como um pesquisador de uma instituição e suas respectivas publicações, um funcionário de uma empresa e as suas postagens em fóruns de discussão internos ou até mesmo o autor de um blog e suas postagens nesse blog.

Os usuários procuram um especialista sempre que precisam de conhecimento técnico específico em um determinado tópico [Balog et al., 2012]. Encontrar um especialista pode ser de grande ajuda, como quando é necessário um grau de confiança grande em uma consultoria, um problema em questão não pode ser respondido computacionalmente ou as informações necessárias para a solução não estão digitalizadas, não existem soluções conhecidas para um dado problema de uma área nova ou muito específica, ou é necessário construir um grupo de especialistas de uma área para uma dada companhia [Balog et al., 2012].

Segundo Balog et al. [2012], desde 1960, quando o problema de busca de especialistas começou efetivamente a ser estudado, muitos pesquisadores visavam representar o conhecimento do homem como modelos computáveis. Naquele momento, o problema foi descoberto quando o intuito era encontrar, em bibliotecas e coleções de periódicos, as informações que tentavam evidenciar qual a fonte do conhecimento que os pesquisadores e engenheiros usavam para seus estudos. Mais recentemente, com o crescente

volume de informações profissionais e acadêmicas publicadas diária e abertamente na Web, buscar e ranquear especialistas tornou-se um problema extremamente difícil de ser resolvido.

Desde o advento da Web, várias empresas vêm convertendo documentos armazenados fisicamente em versões virtuais. Tais versões virtuais ajudam na associação de documentos a pessoas e, com isso, no maior controle do desenvolvimento geral da empresa. Contudo, ao mesmo tempo em que tais coleções ajudam na definição de especialidades de pessoas, o crescimento de tais coleções, muitas vezes não-estruturadas, dificulta a sua manipulação. Isso acontece principalmente pelo tamanho que tais coleções podem chegar, tornando-se, muitas vezes, intratáveis ou sendo descartadas devido ao custo para manutenção [Balog et al., 2012].

Diante dessa situação, onde muitos documentos virtuais eram descartados desnecessariamente, a área de recuperação de especialistas se reergueu, absorvendo conceitos consolidados de recuperação de informação para a manipulação de tais bases [Balog et al., 2012]. Por essa razão, a academia e as empresas que detinham tais bases investiram no desenvolvimento de métodos para geração de grupos de especialistas [Hertzum, 2000; Davenport & Prusak, 1998].

O problema de busca de especialistas tem recebido considerável atenção da comunidade de recuperação de informação na última década, com um foco particular em encontrar especialistas dentro de uma organização empresarial [Bailey et al., 2007b; Balog et al., 2008a; Craswell et al., 2005; Soboroff et al., 2006]. Várias abordagens de busca de especialistas foram propostas tentando modelar os conhecimentos das pessoas e sua relevância dada uma consulta do usuário (por exemplo, Balog et al. [2006]; Fang et al. [2010a]; Macdonald & Ounis [2006]; Serdyukov & Hiemstra [2008]). Em comum, todas essas abordagens dependem de alguma forma de *associação* entre pessoas e documentos, a fim de modelar o perfil de especialidades de cada candidato [Bailey et al., 2007b].

Nesta dissertação, abordamos o problema de busca de especialistas na academia, onde encontrar especialistas pode auxiliar na tarefa de buscar candidatos a orientadores de projetos de pesquisa, avaliadores de artigos científicos em conferências/revistas e recomendar colaboradores no desenvolvimento de trabalhos acadêmicos.

Reconhecidamente como uma subárea de recuperação de informação, a maioria dos problemas que podem ser resolvidos com uma abordagem baseada em busca de especialistas visa encontrar pessoas que acumulam uma quantidade de conhecimento relacionada a uma dada área. Sendo assim, quando se trata de busca de especialistas, um dos meios de se estimar o grau de especialidade de um candidato é através dos documentos associados a esse candidato. Esses documentos cumprem o papel de evidenciar

a especialidade dos candidatos nos processos convencionais de busca de especialistas.

Paralelamente a isso, vemos novas plataformas de compartilhamento/armazenamento de trabalhos acadêmicos surgindo todos os dias, como, por exemplo, ResearchGate[1], DBLP[2], Google Scholar[3], Lattes[4], Semantic Scholar[5] e Microsoft Academic Search[6].

Segundo Balog et al. [2012], a maioria dos trabalhos na literatura para busca de especialistas tem o escopo específico de encontrar especialistas em ambientes empresariais. Esse tipo de ambiente se diferencia do ambiente acadêmico por alguns fatores, sendo eles: (1) No ambiente acadêmico, as evidências de autoria dos documentos são explícitas, ou seja, é possível identificar quais são os autores mencionados no conteúdo do documento; (2) Em geral, os documentos das bases acadêmicas são referências bibliográficas ou citações, ou seja, consistem de documentos com conteúdo textual menor; (3) As relações entre autores de um mesmo documento científico são evidências mais fortes de proximidade contextual entre os coautores.

Desse modo, podemos generalizar que a busca de especialistas em ambientes acadêmicos diverge da busca de especialistas em ambientes empresariais segundo o modo em que as abordagens representam a influência de um documento para os seus respectivos candidatos associados.

Em particular, nesta dissertação modelamos associações documento-especialista como uma combinação de duas funções: uma *função de associação*, responsável por quantificar a importância de um documento para os autores associados a ele, e uma *função de normalização*, responsável por ajustar o peso das associações ao contexto em que ela está inserida.

Nos cenários empresariais, as abordagens consideram o fato da menção de um dos identificadores do candidato no documento ser a evidência esperada para criar a associação. Esse tipo de abordagem é denominado ponderação booleana e possui um conjunto de fatores que deterioram o resultado do modelo de ranking. Dois dos principais fatores são: (1) desconsiderar uma ponderação para candidatos cujas associações são incertas [Balog & De Rijke, 2008], fato que acontece quando não há garantia durante a construção das relações documento-especialista, e (2) não representar o interesse central dos autores em relação ao conteúdo do documento [Macdonald et al., 2008]. Enquanto isso, no ambiente empresarial, as abordagens que não consideram esse

[1] `http://www.researchgate.net/`
[2] `http://dblp.uni-trier.de/`
[3] `http://scholar.google.com.br/`
[4] `http://lattes.cnpq.br/`
[5] `https://www.semanticscholar.org/`
[6] `http://academic.research.microsoft.com/`

tipo de associação, denominadas funções não-booleanas, consideram a ponderação da associação com um esquema que visa diminuir a incerteza inserida durante a construção das associações documento-candidato. Portanto, o objetivo central desse tipo de abordagem é determinar a chance do candidato mencionado no documento ser, de fato, o autor do documento.

Além disso, com exceção do trabalho de Macdonald et al. [2008] mencionado, poucos estudos buscam discriminar o grau de especialidade do candidato para os seus documentos através de funções que ponderam, diretamente, a importância do conteúdo do documento para o candidato. Os ambientes acadêmicos são mais propícios para esse tipo de abordagem, por denotarem associações explicitamente nos documentos, não havendo incerteza da autoria deles.

Assim, nesta dissertação, introduzimos funções de associação não-booleanas que atribuem diferentes interpretações para a relevância de cada documento para seus respectivos candidatos autores. O cerne da abordagem é ponderar as associações através de quatro instanciações dessas funções de forma que elas estimem a importância do documento para o candidato em função da *recência* do documento, do *domínio do conteúdo* do candidato para o documento, do quão *inovador* o candidato foi na proposta daquele conteúdo e do quanto o candidato tem facilidade de, recorrentemente, escrever sobre aquele conteúdo.

Adicionalmente, a função de normalização visa determinar um contexto para a função de associação, sendo esse contexto uma maneira de determinar em qual conjunto de associações a associação normalizada está inserida. Nos trabalhos tradicionais, as associações são normalizadas em relação ao documento ou ao candidato. Nesse caso, diz-se que a associação é normalizada em relação a todos os candidatos autores do documento ou, analogamente, a associação é normalizada em relação a todos os documentos do autor analisado. Essas funções de normalização têm semânticas diferentes, podendo, por exemplo, inferir qual candidato é o autor mais relevante para uma publicação ou qual documento é mais importante para um candidato.

Nas funções de normalização tradicionais, uma associação é ponderada de maneira inversamente proporcional ao número de associações paralelas a ela. Como resultado, documentos associados a muitos candidatos ou candidatos associados a muitos documentos acabam penalizados. Portanto, a normalização proposta tenta remover esse viés que é inserido por essas normalizações aplicando conceitos de teoria da informação em duas funções parametrizáveis. Assim, nesta dissertação, propomos funções de normalização que consideram o contexto da associação como sendo um fator de ponderação da associação.

1.2 Argumento da Dissertação

O principal argumento desta dissertação é que associações não-booleanas que reflitam o quão informativo um documento é para a especialidade de um candidato podem melhorar a busca de especialistas. Nesse aspecto, propomos uma solução para o problema de ponderação da associação documento-candidato através da apresentação de um processo em duas etapas, onde cada etapa é estudada separadamente. Essas etapas ocupam posições importantes no funcionamento devido do sistema de busca de especialistas, determinando meios de se criar um peso para a associação considerando conteúdo e idade do documento e, paralelamente, normalizando esse peso por uma função não-linear adaptada a partir de conceitos da teoria da informação.

Para tanto, apresentamos as perguntas de pesquisa a serem respondidas no decorrer desta dissertação:

Q1. Quão eficazes podem ser as **funções de associação** propostas na geração de rankings de especialistas?

Q2. As **funções de normalização** propostas geram rankings melhores em comparação com as funções de normalização propostas na literatura?

Q3. Quão **complementares** são as diferentes combinações de funções de associação e de normalização no processo ranking de especialistas?

Assim, com base nas perguntas de pesquisa, descrevemos as principais contribuições desta dissertação a seguir.

1.3 Contribuições da Dissertação

As principais contribuições desta dissertação são:

1. Apresentação do processo de ponderação de associação documento-candidato divido em duas etapas.

Nos trabalhos de busca de especialistas existentes, os autores descrevem os modelos de ponderação de associação como uma maneira monolítica para gerar peso para uma relação documento-candidato. Nesta dissertação, as associações são descritas como um processo dividido em duas etapas: (1) funções de associação, e (2) funções de normalização das associações. Atribuindo esse novo formato, é possível adaptar os modelos tradicionais de ponderação de associação como um processo mais formal,

inclusive permitindo a expansão para novos modelos e facilitando a identificação dos fatores principais para a eficácia da busca de especialistas.

2. Introdução de funções de ponderação de associação que consideram domínio de conteúdo e temporalidade do documento para o candidato.

Nas funções de ponderação de associação não-booleanas aplicados no ambiente corporativo, as relações documento-candidato são extraídas com um grau de incerteza. Nos ambientes acadêmicos, essa incerteza ocorre com menor frequência, dado que a menção aos candidatos autores do documento é explícita. Assim, nas abordagens de ponderações de busca de especialistas em ambientes acadêmicos, é possível trabalhar um conceito diferente da relação documento-candidato, onde as associações podem quantificar diretamente a especialidade do candidato perante o conteúdo do documento. Além disso, considerando que os documentos acadêmicos possuem uma data de publicação, é possível modelar a especialidade do candidato considerando fatores evolutivos da área de pesquisa da publicação. Nas funções de ponderação que consideram o tempo, é proposto um conjunto de fatores que denotam a qualidade do documento, por exemplo, documentos precursores em determinadas áreas podem ser primordiais para a determinação da especialidade do candidato.

3. Proposta de funções de normalização de associações documento-candidato como etapas baseadas na teoria da informação.

O segundo fator importante para a determinação do grau de especialidade do candidato para um documento é a forma como a ponderação dessa associação está relacionada a toda especialidade acumulada do candidato. Assim, é proposto uma função de normalização não linear que leva em consideração o peso de todas as associações do candidato ou documento para ponderar a relevância daquela associação. A função proposta mostrou-se eficaz para o ambiente acadêmico por não deteriorar o peso da associação proporcionalmente à quantidade de associações no documento/candidato.

4. Apresentação de uma coleção de teste para busca de especialistas validada por pesquisadores dos Institutos Nacionais de Ciência e Tecnologia (INCTs).

Outra contribuição para a área de busca de especialistas em ambiente acadêmico é a criação e a disponibilização de uma coleção de teste validada por especialistas. Como é descrito na Seção 4.1, uma gama de coleções de teste existe, mas um número restrito delas é ambientada na academia e, além disso, nenhuma possui níveis de relevância para os especialistas do gabarito das consultas. A coleção apresentada nesta

dissertação foi construída a partir da aplicação de um questionário, cujos candidatos respondentes são pesquisadores dos INCTs, um grupo seleto de pesquisadores que possuem um significativo reconhecimento na academia nacional e internacional em diversas áreas.

1.4 Origens do Material

Uma parte do material apresentado nesta dissertação foi aceito como dois artigos na conferência *ACM SIGIR Conference on Research and Development in Information Retrieval* 2016 (SIGIR 2016). Mais especificamente:

- No Capítulo 3 apresentamos a formalização do processo de geração de ponderações para associação, o qual é dividido em duas etapas. A primeira etapa, descreve a ponderação denominada dominância de conteúdo (Seção 3.1.1) e a segunda descreve as funções de normalização de associações, que são inspiradas no conceito de teoria da informação de auto-informação[7] [Mangaravite & Santos, 2016].

- Na Seção 4.1 apresentamos a coleção de teste utilizada para experimentação das abordagens propostas. Tal coleção é baseada nos currículos de um subconjunto de pesquisadores da plataforma Lattes, especificamente, dos currículos dos doutores da plataforma e as consultas de especialidade foram validadas através de questionários aplicados a pesquisadores dos Institutos Nacionais de Ciência e Tecnologia (INCTs). O trabalho desenvolvido por Mangaravite et al. [2016], descreve a coleção dos documentos e consultas usadas nesta dissertação.

1.5 Organização da Dissertação

Esta dissertação está organizada da seguinte forma:

- No Capítulo 2, apresentamos a formalização do processo de determinação das ponderações das associações documento-candidato. Inicialmente, apresentamos os trabalhos relacionados discutindo as principais abordagens de ranking de especialistas (Seção 2.1). Na Seção 2.2, fizemos uma discussão descrevendo as principais diferenças dos trabalhos tradicionais de ponderação de associação e das

[7] Do inglês, *self-information*.

funções de normalização e as etapas do processo proposto. Além disso, apresentamos uma descrição das abordagens de especialistas que usam funções de ponderação de associação (Seção 2.2.1) e as funções de normalizações (Seção 2.2.2), bem como o seus respectivos funcionamentos. Na Seção 2.3, apresentamos as principais coleções de testes que aplicam os modelos de ranking de especialistas tradicionais.

- No Capítulo 3, apresentamos as formas de determinar o peso da associação como uma das duas etapas do processo. Esse peso pode ser implementado como sendo a dominância do conteúdo do candidato em relação ao documento (Seção 3.1.1), a estabilidade temporal do candidato para o domínio de cada documento publicado por ele (Seção 3.1.2), a função que pondera a novidade do documento do candidato em relação aos documentos publicados anteriormente na coleção (Seção 3.1.3), e, por último, as funções puramente temporais que quantificam o quão recente o documento é para o candidato e para a coleção (Seção 3.1.4). Ao final do capítulo, na Seção 3.2, apresentamos a função de normalização não linear proposta para ambientes acadêmicos que remove o viés favorável a candidatos/documentos com poucas associações.

- No Capítulo 4, apresentamos as configurações experimentais usadas para comparação e validação das abordagens propostas. Na Seção 4.1 incluímos a descrição detalhada do procedimento de construção da coleção de teste usada nesta dissertação. Na Seção 4.2 apresentamos a descrição detalhada de como foi implementado cada *baseline* usado na comparação com o método proposto, bem como quais abordagens são comparáveis. Finalmente, na Seção 4.3 descrevemos o funcionamento do procedimento de treino-teste, usualmente aplicado para avaliação de problemas com soluções empíricas.

- No Capítulo 5, retomamos as questões de pesquisa a serem respondidas pela dissertação e a metodologia aplicada para responder essas questões. Em seguida, demonstramos duas aplicações práticas do processo de construção das ponderações de associação, aplicando os conceitos de peso de associação e normalização de associação em um modelo probabilístico generativo (Seção 5.2) e modelos discriminativos (Seção 5.3). Na descrição dos resultados, apresentamos uma avaliação dos potenciais fatores que levaram para os resultados positivos obtidos no ranking de especialistas na coleção acadêmica baseada no Lattes.

- Concluindo a dissertação apresentamos um sucinto resumo das contribuições no Capítulo 6. Ainda nesse capítulo descrevemos as conclusões obtidas pelo resul-

tado experimental e os possíveis caminhos a seguir como trabalhos futuros para as funções de associação e normalização propostas.

Capítulo 2

Trabalhos Relacionados

A busca de especialistas tem sido alvo de intensa atenção dos pesquisadores da área de recuperação de informação desde a apresentação da tarefa de busca proposta pela TREC (*Text REtrieval Conference*) em [Craswell et al., 2005]. De lá para cá, outras coleções de teste para busca de especialistas foram propostas [Soboroff et al., 2006; Bailey et al., 2007b; Berendsen et al., 2013b; Tang et al., 2008a], aumentando ainda mais o foco em busca de especialistas.

Como apresentado por Balog et al. [2012], um sistema de busca de especialistas possui dois componentes primordiais para seu funcionamento adequado: (1) o modelo de ranking dos candidatos a especialistas (Seção 2.1) e (2) o esquema de associação usado para representar as associações documento-candidato (Seção 2.2).

2.1 Modelos de Ranking

Nos últimos anos, vários modelos de ranking foram propostos para busca de especialistas. Os principais utilizam conceitos de (1) modelos de votação, (2) modelos probabilísticos generativos, (3) modelos probabilísticos discriminativos, e (4) modelos baseados em grafos.

Em geral, as abordagens de ranking de especialistas são divididas aplicando dois arcabouços diferentes: aquelas que agrupam todo o conteúdo dos documentos associados aos candidatos, para então, dada a consulta, avaliarem o grau de expertise diretamente nos candidatos; ou, aquelas que dividem a abordagem em duas etapas, a primeira que avalia a similaridade entre os documentos e a consulta, e a segunda que transfere a similaridade do documento para os candidatos associados.

2.1.1 Modelos de Votação

Macdonald & Ounis [2006] propuseram o problema de busca de especialistas de forma equivalente ao problema de votações ponderadas. Em seus trabalhos, os autores consideram que cada candidato a especialista é ponderado no processo de votação segundo a relevância dos documentos para a consulta e as combinações desses pesos são feitas por abordagens de fusão de ranking. Foram avaliadas diferentes formas de estimar a relevância do documento e doze propostas de abordagens de fusão de ranking.

Posteriormente, Macdonald et al. [2008] propuseram extensões aos modelos de votação incluindo evidências que qualificavam bons documentos ou candidatos a especialistas. Essas evidências foram combinadas às funções do processo de votação e melhoraram, significativamente, os modelos de votação. As evidências usavam conceitos de recuperação de informação para quantificar relações entre documentos e candidatos, usando, por exemplo, proximidade dos termos da consulta com os identificadores do candidato no documento, informações de *inlinks* dos documentos, posição do documento nos *clusters* de documentos do candidato (como é apresentado na Seção 2.2.1) e tamanho das URLs dos documentos. De acordo com os experimentos, o resultado do método que combina as evidências depende da coleção usada.

2.1.2 Modelos Probabilísticos

Nesta seção apresentaremos os principais modelos probabilísticos para ranking de especialistas. Esse tipo de modelo de ranking é divido em dois grupos de abordagens, denominadas, modelos probabilísticos generativos (Seção 2.1.2.1) e modelos probabilísticos discriminativos (Seção 2.1.2.2).

A principal diferença entre os dois tipos de abordagens é que nos modelos generativos a probabilidade é dada pela distribuição conjunta dos eventos, portanto, o ranking é gerado sem considerar possíveis gabaritos para a consulta. Enquanto isso, em modelos discriminativos, o modelo aprende a geração do ranking através da distribuição condicional dos eventos, treinando o modelo de ranking a partir do gabarito conhecido para as consultas [Jordan, 2002].

Nas seções a seguir, apresentaremos alguns dos modelos probabilísticos conhecidos na literatura para ranking de especialistas.

2.1.2.1 Modelos Generativos

Para os modelos generativos, a ideia central é responder à seguinte pergunta: Qual a probabilidade do candidato e ser um especialista dada a consulta q? Para isso, Balog

& de Rijke [2006] estimaram, pelo teorema de Bayes, que candidatos com os maiores valores da probabilidade $P(e|q)$ são candidatos mais propícios a serem especialistas, conforme a equação:

$$P(e|q) = \frac{P(q|e)P(e)}{P(q)}, \tag{2.1}$$

onde $P(e)$ é a probabilidade a priori do candidato e e $P(q)$ é probabilidade da consulta. Uma vez que $P(q)$ é uma constante em relação à consulta, pode-se ignorar esse fator por não alterar a ordem final do ranking de especialistas. Apesar de $P(e)$ poder ser estimado através de outros modelos, como é apresentado nos trabalhos de Petkova & Croft [2007] e Fang & Zhai [2007], no trabalho de Balog & de Rijke [2006] e nesta dissertação, estima-se que $P(e)$ tem distribuição uniforme sobre os candidatos retornados pela consulta.

Onde $P(e|q)$ é monotônico em $P(q|e)$, o problema de ranquear usando $P(e|q)$ se reduz ao problema de se ranquear usando $P(q|e)$ apenas. Portanto, o desafio principal dos modelos generativos, propostos por Balog & de Rijke [2006], é determinar uma maneira de estimar a probabilidade $P(q|e)$. Naquele trabalho, os autores propuseram duas maneira de representar essa estimação, usando *big document models* ou usando *small document models*. A principal diferença entre as duas abordagens é que a primeira (Modelo 1) infere um modelo linguístico para o candidato a partir do conteúdo de todos os seus documentos, enquanto a segunda utiliza apenas os documentos que melhor descrevem a especialidade dos candidatos (Modelo 2), em um processo de duas etapas, similar ao utilizado por Macdonald & Ounis [2006].

Considerando a premissa da independência entre os termos, a Equação 2.2 define como o Modelo 1, também conhecido na literatura como *candidate model*, estima a probabilidade $P(q|e)$[1] construindo um modelo linguístico para o candidato e calculando a verossimilhança da consulta[2] por

$$P_{\mathrm{M1}}(q|e) = \prod_{t \in q} P(t|\theta_e)^{n(t,q)}, \tag{2.2}$$

onde $n(t,q)$ é o número de ocorrências do termo t da consulta q, $P(t|\theta_e)$ é a probabilidade do termo t no modelo do candidato estimado como

$$P(t|\theta_e) = (1 - \lambda_e)P(t|e) + \lambda_e P(t), \tag{2.3}$$

[1]Note que a Equação 2.2 omite o coeficiente multinomial $K_q = {}^{\sum_t n(t,d)!}/_{\prod_t n(t,d)!}$, o que implica que $P(q|e)$ não representa uma distribuição de probabilidades. Essa simplificação é típica em modelos de ranking baseados em verossimilhança da consulta.

[2]Do inglês, *query likelihood*.

onde $P(t)$ é probabilidade do termo t na coleção de documentos e λ_e é o hiper-parâmetro da função de suavização, usada para evitar probabilidades zeradas durante o cálculo da probabilidade $P(t|\theta_e)$. Para estimar $P(t|e)$, os autores propõem um meio de inferir a probabilidade do termo t dado o candidato e relacionando os documentos associados ao candidato

$$P(t|e) = \sum_d P(t|d,e)P(d|e), \tag{2.4}$$

onde $P(t|d,e)$ é a probabilidade do termo t ocorrer no documento d e o candidato e e $P(d|e)$ é o peso da associação do candidato e e do documento d, sendo, nesse ponto, a probabilidade do candidato e ser um dos autores do documento d. Considerando que a ocorrência conjunta do documento e do candidato são independentes em relação ao termo t, pode-se dizer que $P(t|d,e) \approx P(t|\theta_d)$;

$$P(t|\theta_d) = (1-\lambda_d)P(t|d) + \lambda_d P(t), \tag{2.5}$$

onde λ_d é o hiper-parâmetro da suavização que, assim como na Equação 2.3, é proposta para evitar probabilidades zeradas, e sendo $P(t|d)$ a probabilidade do termo t ocorrer no documento d, tem-se

$$P(t|d) = \frac{n(t,d)}{\sum_{t'} n(t',d)} \tag{2.6}$$

onde $n(t,d)$ é o número de ocorrências do termo t no documento d. Assim, Balog & de Rijke [2006] obtiveram a seguinte formulação final para o Modelo 1:

$$P_{\mathrm{M1}}(q|e) = \prod_{t \in q} \Big\{ (1-\lambda_e)\Big(\sum_d P(t|\theta_d)P(d|e)\Big) + \lambda_e P(t) \Big\}^{n(t,q)}. \tag{2.7}$$

Para o Modelo 2, também conhecido como *document model*, os autores estimam a probabilidade $P(q|e)$ estimando, primeiramente, as probabilidades dos documentos associados ao candidato e. Assim,

$$P_{\mathrm{M2}}(q|e) = \sum_d P(q|d,e)P(d|e), \tag{2.8}$$

onde $P(q|d,e)$ é a probabilidade da consulta q ocorrer, conjuntamente, no documento d e candidato e e, assim como na Equação 2.4, $P(d|e)$ é o peso da associação do candidato

e e do documento d. Portanto,

$$P(q|d,e) = \prod_{t \in q} P(t|d,e)^{n(t,q)}, \tag{2.9}$$

onde, assim como no Modelo 1, os autores consideram que o documento e o candidato são independentes em relação ao termo t, aproximando $P(t|d,e)$ como $P(t|\theta_d)$, onde se obtém, ao final:

$$P_{\text{M2}}(q|e) = \sum_{d} \Big(\prod_{t \in q} P(t|\theta_d)^{n(t,q)} \Big) P(d|e). \tag{2.10}$$

Uma vantagem do uso do Modelo 2 em relação ao Modelo 1 é que, para o Modelo 1, é preciso construir um índice auxiliar para estimar a especialidade do candidato, enquanto que para o Modelo 2 ele é inferido a partir da probabilidade dos documentos individualmente.

Apesar de tanto $P_{\text{M1}}(q|e)$, quanto $P_{\text{M2}}(q|e)$ não resultarem uma distribuição de probabilidades, ambas funções são proporcionais às suas probabilidades nas distribuições. A justificativa dessas funções não serem distribuições de probabilidade é que ignoramos uma constante K_q que representa o coeficiente multinomial da consulta q, sendo que ignorar essa constante é usual em abordagens que usam a verossimilhança da consulta em modelos linguísticos de ranking.

2.1.2.2 Modelos Discriminativos

Diferentemente dos modelos generativos, que modelam o processo de geração da linguagem das consultas, os modelos discriminativos são treinados considerando os resultados já conhecidos das consultas. Apesar de menos estudados, recentemente a área tem olhado com mais atenção para esse tipo de abordagem, viabilizado pela disponibilidade de grandes coleções de treino [Fang et al., 2010b].

Fang et al. [2010b] introduziram dois modelos discriminativos para busca de especialistas em ambientes empresariais disponíveis. Em seus estudos, eles abordam o problema similarmente ao Modelo 2, proposto por Balog & de Rijke [2006], e a relevância de um candidato e para uma dada consulta q considera dois fatores: (1) se o documento d for relevante para a consulta q ($P(r_1|q,d)$); e (2) se o candidato e for relevante para o documento d ($P(r_2|e,d)$). Os autores partiram da premissa de que d

é independente de q e de e. Assim,

$$P_Z(r = 1|e, q) = \sum_d P(r_1 = 1|q, d)P(r_2 = 1|e, d)P(d), \tag{2.11}$$

onde $P(d)$ é a probabilidade a priori do documento d, Z é o modelo de otimização a ser treinado, e $P(r_1 = 1|q, d)$ e $P(r_2 = 1|e, d)$ são definidos como:

$$P(r_1 = 1|q, d) = \sigma\big(\sum_i^{N_g} \alpha_i g_i(q, d)\big) \tag{2.12}$$

$$P(r_2 = 1|e, d) = \sigma\big(\sum_j^{N_f} \beta_j f_j(e, d)\big) \tag{2.13}$$

onde $\sigma(x) = {}^1/_{(1+e^{-x})}$, $g_i(q, d)$ e $f_j(e, d)$ representam o i-ésimo e j-ésimo atributos do par consulta-documento e da associação documento-candidato, respectivamente. Algumas funções possíveis para $f_i(e, d)$ são apresentadas na Seção 2.2.1. Seja o modelo Z na Equação 2.11 definido pelos parâmetros de α e β das combinações lineares das Equações 2.12 e 2.13, Fang et al. [2010b] modelam o problema considerando a probabilidade do candidato ser ou não ser especialista treinando, com a função de otimização Quasi-Newton (Dennis Jr & Schnabel [1996]), o parâmetro Z maximizando a função de log verossimilhança

$$Z^* = \arg\max_Z \sum_{q \in Q} \sum_{e \in E} \Big(r_{qe} \log P_Z(r = 1|e, q) + (1 - r_{qe}) \log \big(1 - P_Z(r = 1|e, q)\big)\Big) \tag{2.14}$$

onde Q é o conjunto de todas as consultas de treino, E é o conjunto de todos os candidatos a especialistas, r_{qe} é a variável binária que determina se o candidato e é relevante para a consulta q ($r_{qe} = 1$) ou se não é relevante ($r_{qe} = 0$).

Além desse arcabouço, os autores apresentaram uma maneira de estimar $P_Z(r = 1|e, q)$ como sendo a média geométrica das evidências e experimentaram ambas as abordagens, demonstrando que os modelos discriminativos podem ser superiores aos modelos generativos com um grau de significância aceitável.

Um outro conceito de modelos discriminativos é aplicado em abordagens de *learning to rank*, com atributos de especialistas definidos a partir da agregação de atributos de documentos. Nesse tipo de abordagem, diferentes técnicas de ranking de especialistas são treinadas com algoritmos de *learning to rank* resultando em rankings, em geral, melhores. Uma abordagem que usa esse conceito é apresentada em Macdonald & Ounis

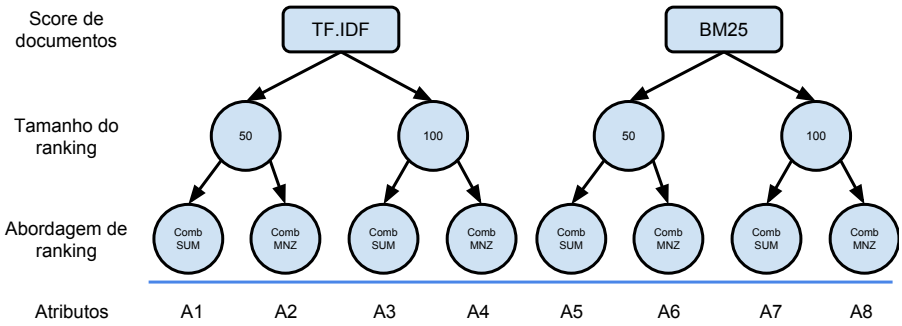

Figura 2.1: Exemplo de definição dos atributos dos candidatos a especialistas.

[2011]. Nesse trabalho, os autores aplicam técnicas de *ensemble learning* para o treino de diferentes algoritmos de aprendizagem fraca com intuito de criar um algoritmo de aprendizagem forte (Opitz & Maclin [1999]). Os atributos usados são os resultados de combinações de ranking de especialistas de duas diferentes instanciações de modelos de votos, variando as abordagens de peso de documentos e tamanho do ranking dos documentos.

A Figura 2.1 apresenta um exemplo de como gerar os atributos dos candidatos a especialistas através da técnica criada por Macdonald & Ounis [2011]. Nesse caso, são criados oito atributos para cada candidato, variando em todas as possibilidades a abordagem de geração do *score* dos documentos, a quantidade de documentos que serão recuperados por essa abordagem de *score* e as agregações dos rankings propostas por eles em [Macdonald & Ounis, 2006].

A principal diferença entre o modelo probabilístico discriminativo proposto por Fang & Zhai [2007], denominado aqui de arcabouço discriminativo de *modelos baseados em atributos simples*, e a abordagem *learning to ranking* proposta por Macdonald et al. [2008], denominada aqui de *modelos baseados em atributos agregados*, é na instanciação dos atributos e como são treinado os modelos finais. No primeiro caso, o treino está em um grão mais fino do processo de ranking de especialistas, treinando o peso das associações e documentos usando uma abordagem similar ao Modelo 2, proposto por Balog & de Rijke [2006]. Enquanto no segundo caso, o processo de treino é aplicado após a agregação de atributos.

2.1.3 Modelos Baseados em Grafos

Vindo com uma proposta diferente das abordagens apresentadas neste trabalho, os modelos que usam fatores sociais sobre o grafo de autoria têm o objetivo de avaliar a especialidade dos candidatos a especialistas usando as relações documento-documento

e candidato-candidato, propondo, por exemplo, uma forma de modelar essas associações em redes como de citações. Serdyukov et al. [2008], demonstraram que é possível melhorar o ranking dos especialistas a partir de uma ponderação dos candidatos considerando a região em que cada candidato se encontra no grafo de especialidade. Outra importante contribuição foi a definição de diferentes formas de representar o grafo de especialidade, onde, segundo eles, existem pelo menos três tipos de arestas que podem ser usadas na modelagem: (1) documento-candidato; (2) documento-documento; e (3) candidato-candidato. Dada a formalização dessas arestas, os autores propuseram quatro algoritmos baseados em caminhamento aleatório probabilístico[3].

A primeira proposta dos autores é denominada *Finite Random Walk*. Nessa abordagem, os autores consideram uma navegação sobre o grafo de especialidade bipartido, ou seja, usando apenas as associações documento-candidato. A intuição do funcionamento é que uma pessoa que está buscando especialistas sobre o grafo de especialidades navega de um documento altamente relevante para um candidato e vice-versa repetindo o processo iterativamente. Em seguida, os autores apresentam a abordagem denominada *Infinite Random Walk*, em que consideram o processo de navegação como um processo não-terminável e usam como critério de parada a convergência da execução. Diferente da abordagem a *Finite Random Walk*, essa abordagem considera a importância do candidato estar próximo de documentos relevantes e também assume a existência de uma distribuição estacionária, introduzindo a probabilidade de saltos nos vértices de candidatos no grafo de especialidade.

Seguindo o preceito de que alguns cenários de busca de especialistas podem incluir não apenas associações documento-candidato, mas também candidato-candidato e documento-documento, Serdyukov et al. [2008] apresentaram uma metodologia que não considera apenas grafos bipartidos. Nesse grafo de especialidade, os autores combinam duas ponderações de caminhamento aleatório probabilísticos, sendo uma baseada na probabilidade da navegação usando as associações documento-candidato e documento-documento e outra considerando a probabilidade da navegação usando as associações documento-candidato e candidato-candidato. Para representar a probabilidade da especialidade do candidato é usada, assim como na abordagem anterior, a probabilidade acumulada das iterações.

Na ultima abordagem apresentada, os autores representaram a busca de especialistas como um *absorbing random walk* em um grafo apenas com as associações documento-candidato. Segundo eles, essa abordagem possui várias vantagens teóricas sobre as demais porque, quando o grafo de especialistas possui tamanho fixo, esse mé-

[3]Do inglês, *Probabilistic Random Walk.*

todo não precisa de nenhum parâmetro. Nesse caso, o método pode ser visto como uma generalização do Modelo 2, apresentado por Balog & de Rijke [2006], onde $P(e|d)$ é representado não como propagação em um passo, mas com um número mínimo suficiente de passos. Além disso, não é necessária a execução iterativa do método, uma vez que existe uma representação matricial em que o método pode ser reescrito.

2.2 Modelos de Associação

Um segundo importante aspecto que deve ser levado em consideração no processo de criação de um sistema de busca de especialistas é a determinação da relevância de um documento para os candidatos associados a ele. Como mencionado, existem algumas diferenças fundamentais das abordagens de ranking de especialistas nos ambientes empresariais e acadêmicos no que se refere a construção das associações. Existem duas estratégias principais para identificação de uma associação documento-candidato: (1) onde os documentos possuem metadados informando explicitamente quais os candidatos estão associados a ele; e (2) onde os documentos não fazem menção explícita aos candidatos associados e é necessário extrair as associações do conteúdo não estruturado de cada documento.

No primeiro caso, usualmente encontrado em ambientes acadêmicos, uma associação documento-candidato possui uma semântica diferente, onde a pessoa associada ao documento denota diretamente algum conhecimento em relação aos tópicos cobertos pelo documento. Exemplos desses tipos de documento são encontrados em bases de publicações acadêmicas (Balog et al. [2007a]; Tang et al. [2008a]), orientações de trabalhos de conclusão de curso (Fang et al. [2009]; Liebregts & Bogers [2009]; Deng et al. [2008]), e de envio e recebimento de mensagens de e-mail (Balog & de Rijke [2006]; Fang et al. [2010b]; Petkova & Croft [2008]).

Enquanto isso, no segundo caso, são usadas técnicas de reconhecimento de entidades nomeadas[4] para tentar inferir quais as associações são encontradas no conteúdo dos documentos. Em geral, os documentos desse tipo de base são páginas HTML (*HyperText Markup Language*) e as associações são inferidas usando uma gama de representações dos identificadores dos candidatos, como, por exemplo, diferentes combinações do nome/sobrenome e endereço de e-mail (Balog et al. [2012]).

Mesmo tendo semânticas diferentes, pode-se dizer que, em ambos os aspectos de construção de uma associação documento-candidato, existe, para cada par documento-candidato, um valor quantitativo que mensura essa relação. Assim, podemos definir

[4]Do inglês *named entity recognition*.

uma *função de associação* ρ que pondera essa relação, onde $\rho(d, e)$ é o peso do documento d para um candidato e, ou vice versa.

Ademais, os modelos tradicionais de ponderação de associação apresentam uma variedade de formas de normalizar os pesos das relações documento-candidato. Assim como denotado para função ρ, podemos generalizar os esquemas de normalização como uma *função de normalização*. Nesse caso, dizemos que o peso final de uma associação documento-candidato é dado pela função:

$$f(d, e) = \psi(\rho(d, e)) \tag{2.15}$$

Essa taxonomia é introduzida nesta dissertação para auxiliar a compreensão dos principais fatores que contribuem para o funcionamento de um sistema de busca de especialistas que usa associações não-booleanas. Contudo, esse processo de construção não está limitado a esse arcabouço, podendo ser adaptado para abordagens baseadas em grafos, por exemplo.

Adiante são apresentadas algumas instanciações tradicionais para a função de associação ρ (Seção 2.2.1) e para função de normalização ψ (Seção 2.2.2) aplicadas a busca de especialistas.

2.2.1 Funções de Associação ρ

A primeira função de associação proposta foi apresentada em Balog & de Rijke [2006]. A ideia central desse tipo de peso é que as associações documento-candidato são independentes entre si, ou seja, todas as associações têm o mesmo peso. Nesse tipo de função de associação, dizemos que as associações são booleanas e a sua formalização é dada por

$$\rho(d, e) = \begin{cases} 1, & \text{se existe associação entre } e \text{ e } d \\ 0, & \text{caso contrário.} \end{cases} \tag{2.16}$$

Este tipo de função de associação é usual e, em alguns casos, prático [Balog et al., 2012]. Contudo, algumas funções mais genéricas aceitam o peso da associação documento-candidato como sendo um valor real, não-booleano. Uma abordagem popular dessa estimação foi proposta em Balog & De Rijke [2008], onde o peso da associação é determinado a partir da frequência das ocorrências dos identificadores dos candidatos no documento. Nessa função, os documentos são representados unicamente por esses

identificadores (chamada representação *lean*) e a função $f(d, e)$ é estimada como:

$$f(d,e) = (1-\lambda)\frac{n(e,d)}{\sum_{e'} n(e',d)} + \lambda\frac{\sum n(e)}{\sum_{e'} n(e')}, \tag{2.17}$$

onde λ é o hiper-parâmetro da função de suavização, $n(e, d)$ é o número de ocorrências dos identificadores do candidato e no documento d e $n(e)$ é o tamanho da representação *lean* do candidato e.

Ainda no trabalho de Balog & De Rijke [2008], os autores introduziram o conceito de *Semantic-Relatedness*, onde o número de ocorrências do candidato no documento é estimado pela importância desse documento para o candidato. Eles reformulam as representações *lean* para modelos linguísticos dos documentos (θ_d) e candidatos (θ_e), também usando os identificadores dos candidatos e substituem o número de ocorrências de e em d por,

$$n'(d,e) = \begin{cases} KL(\theta_e||\theta_d), & \text{se } n(e,d) > 0 \\ 0, & \text{caso contrário.} \end{cases} \tag{2.18}$$

onde $KL(\theta_e||\theta_d)$ é a distância de *Kullback-Leibler* e é representada por

$$\begin{aligned} KL(\theta_e||\theta_d) &= \sum_{i\in\theta} P(i|\theta_e)\log\left(\frac{P(i|\theta_e)}{P(i|\theta_d)}\right) \\ &= H(\theta_e||\theta_d) - H(\theta_e) \\ &= H(\theta_e||\theta_d) + const(e) \\ &\approx -\sum_{i\in\theta} P(i|\theta_e)\log P(i|\theta_d), \end{aligned} \tag{2.19}$$

onde $H(\bullet)$ é a entropia de Shannon (Cover & Thomas [2012]), $H(\theta_e||\theta_d)$ é a entropia cruzada entre o modelo do candidato θ_e e o modelo do documento θ_d, $-H(\theta_e) = const(e)$ é uma constante em relação ao candidato e. Assim, a modelagem proposta por Balog & De Rijke [2008] para a distância de *Kullback-Leibler* é aproximadamente equivalente a entropia cruzada, $H(\theta_e||\theta_d)$.

O peso das associações é usualmente estimado no nível de documentos, mas é possível estabelecer um peso em uma granularidade mais fina do processo. Essa intuição foi aplicada para criar funções de associações que consideram a ocorrência dos termos da consulta próximos a identificadores de candidatos no documento. Conhecido como abordagem baseada em janelas, esse tipo de ponderação determina o peso da associação durante o processo do calculo da similaridade da consulta com o candidato, alterando a probabilidade $P(q|d)$ condicionando o documento associado ao candidato ($P(q|d, e)$).

Petkova & Croft [2007] propõem uma maneira de capturar a dependência entre os termos e os candidatos a autores do documento usando a representação do documento baseada em kernels de proximidade. Assim, dado um termo t da consulta q, estima-se $P(t|d,e)$ para capturar a dependência dos termos e candidatos, substituindo a probabilidade $P(t|d)$ da Equação 2.5 pela formulação:

$$P(t|d,e) = \frac{1}{\sum_{i=1}^{N} k(t,e)} \sum_{i=1}^{N} \delta_d(i,t) k(t,e), \tag{2.20}$$

onde N é o tamanho do documento e

$$\delta_d(i,t) = \begin{cases} 1, & \text{se } i = t \\ 0, & \text{caso contrário.} \end{cases} \tag{2.21}$$

Qualquer função $k(t,e)$ não-booleana e não-crescente pode ser convertida para um kernel baseado em janela. Por exemplo, $k(t,e) = 1/N$ corresponde a uma representação *bag-of-words*, onde é determinada a mesma probabilidade para cada termo do documento. Três funções não-booleanas foram consideradas no trabalho de Petkova & Croft [2007]: kernel triangular, kernel Gaussiano e *step function*. Segundo os resultados empíricos, os três kernels têm performance similar e superior às funções constantes.

Analogamente ao que é proposto com as funções ρ apresentadas nesta dissertação, Macdonald et al. [2008] apresentaram uma maneira de determinar a proximidade do conteúdo dos documentos com o interesse central dos candidatos associados. Para isso, os autores agruparam os documentos de cada candidato através de um algoritmo single-pass de clusterização e determinaram o peso de cada documento para o candidato como sendo uma função monotonicamente decrescente em relação à posição no ranking dos clusters com mais documentos (Equação 2.22). A ideia central dessa proposta é que as áreas em que o candidato possui mais expertise tendem a corresponder aos clusters com mais documentos:

$$\rho(d,e) = \begin{cases} \frac{1}{cluster(d,e)}, & \text{se } cluster(d,e) \leq K \\ 0, & \text{caso contrário.} \end{cases} \tag{2.22}$$

onde $cluster(d,e)$ é a posição do ranking de clusters do candidato e que o documento d ocorreu e K é a posição máxima aceitável para o ranking dos clusters ser usado na abordagem, sendo a posição $cluster(d,e)$ do cluster relativa a ordem dos maiores clusters do candidato e.

2.2.2 Funções de Normalização ψ

Como mencionado, os trabalhos na área de busca de especialistas usam uma gama de funções de associação para representar diferentes semânticas de uma relação documento-candidato. Nos trabalhos estudados, boa parte das abordagens usam as funções de normalização apresentados em Balog et al. [2006]. O primeiro, denominado *document-centric*, estima a força de uma associação entre um documento d e um candidato e em termos da probabilidade $P(d|e)$. Assim eles definem essa função de normalização *document-centric* (DC) de acordo com:

$$\psi_{DC}(\bullet) \equiv \frac{\bullet}{\sum_{e' \in E_d} \rho(d, e')}, \tag{2.23}$$

onde $\bullet$ é a função de associação que deve ser normalizada e E_d é o conjunto de candidatos associados ao documento d, formalmente representado por $E_d = \{e' : e' \in E \wedge (e', d) \in A\}$, onde A é o conjunto de todas as associações da coleção de documentos, (e', d) é a associação do candidato e para o documento d e E é o conjunto de todos os candidatos a especialistas.

De forma análoga, os autores propuseram a função de normalização denominado *candidate-centric* (CC), onde a normalização estima a probabilidade $P(e|d)$ como sendo

$$\psi_{CC}(\bullet) \equiv \frac{\bullet}{\sum_{d' \in D_e} \rho(d', e)}, \tag{2.24}$$

onde D_e é o conjunto de documentos associados ao candidato e, formalmente representado por $D_e = \{d' : e \rightarrow d'\}$.

Cada função que Balog et al. [2006] propuseram tem características individuais. a função de normalização *document-centric* representa a associação como sendo uma proporção da importância de cada candidato mencionado no documento, tentando diferenciar a importância de cada autor para um mesmo documento e tem a vantagem de ser mais estável em diferentes bases. O *candidate-centric*, por sua vez, que representa a associação como sendo uma proporção da importância de cada documento com o qual o candidato tem associação, tem o intuito de diferenciar entre os documentos de um mesmo candidato e, em alguns casos, pode prejudicar candidatos com muitos documentos.

Como algumas bases de dados possuem muitos candidatos prolixos que não são necessariamente especialistas, Macdonald & Ounis [2011] propuseram uma função de normalização que visa normalizar a associação relacionando o peso da associação com o peso médio desse tipo de associação na coleção. Em seus estudos, os autores propuseram

duas medidas de associação para validar a função de normalização, onde as associações são ponderadas segundo o número de documentos associados ao candidato ou o número de termos encontrados na representação do modelo do candidato. Assim essa função de normalização é formalizado como

$$\psi_{Norm2}(\bullet) \equiv \log\Big(\frac{\overline{\rho(d,e)}}{\bullet} + 1\Big), \tag{2.25}$$

onde $\overline{\rho(d,e)}$ é a média aritmética dos pesos da associação, sendo representada por $\overline{\rho(d,e)} = \frac{\sum_{(d',e')}\rho(d',e')}{N_{(d,e)}}$, onde (d',e') é o conjunto de todas as associações da coleção, sendo que d' e e' tem associações entre si; e $N_{(d,e)}$ é o número de associações na coleção.

2.3 Avaliação de Rankings

2.3.1 Coleções de Teste

Um dos pontos principais para o progresso da área de pesquisa de busca de especialistas é desenvolver e disponibilizar coleções de testes. Além das abordagens de busca de especialistas apresentadas, uma gama de coleções de teste foram construídas nos últimos anos. Duas das primeiras coleções disponibilizadas foram apresentadas na *TREC Enterprise track* de 2005 a 2007, sendo essas coleções baseadas nas organizações *World Wide Web Consortium* (W3C) e *Commonwealth Scientific and Industrial Research Organisation* (CSIRO). A primeira delas, apresentada por Craswell et al. [2005] e estendida por Soboroff et al. [2006], é baseada nos membros de grupos de trabalho da W3C como candidatos a especialistas e está ambientada em uma área de conhecimento, padrões da *Web*. As avaliações de especialidade não foram construídas pelos especialistas, mas sim pelos grupos de trabalho da W3C e o conjunto de documentos consiste de domínios variados, podendo ser e-mail, códigos, páginas *Web* da corporação ou pessoais e páginas de *Wikis*, sendo que os e-mails representam a maioria dos documentos (59% da coleção).

A segunda coleção proposta pela TREC, denominada *CSIRO Enterprise Research Collection* (CERC), foi apresentada por Bailey et al. [2007a] e estendida por Bailey et al. [2007b] e Balog et al. [2008b]. Essa coleção consiste dos empregados da CSIRO como candidatos a especialistas, sendo que a avaliação dos especialistas foi feita pelos próprios pesquisadores da CSIRO. Os documentos foram coletados de 100 domínios pertencentes a CSIRO, somando mais de 370 mil documentos no total. Uma característica em destaque dessa coleção de teste é a variedade de áreas em que os especialistas estão distribuídos nas múltiplas divisões da companhia.

Duas coleções de teste de busca de especialistas em ambientes acadêmicos denominadas UvT e TU, foram propostas respectivamente por Balog et al. [2007b] e Berendsen et al. [2013a], com nome de TU. Nessas coleções, os candidatos a especialistas são pesquisadores associados à *Tilburg University*. Ambas tiveram suas consultas gabaritadas considerando a aplicação de um questionário onde os candidatos avaliam rótulos de sumarização de suas áreas de pesquisa, assim como o problema *expert profiling*. Os documentos das coleções têm características acadêmicas, podendo ser páginas pessoais dos professores/pesquisadores, artigos publicados, teses ou dissertações ou páginas de disciplinas ofertadas pelos candidatos, somando, no total, 36.699 e 32.567 documentos, para UvT e TU, respectivamente. Uma peculiaridade dessas coleções é que os documentos e consultas são divididos em dois idiomas: Holandês e Inglês.

Uma outra coleção de teste de busca de especialistas em ambiente acadêmico é a ArnetMiner [Tang et al., 2008b]. Essa coleção define, como candidatos a especialistas, pesquisadores com páginas na DBLP, um repositório de referências bibliográficas de publicações na área de ciência da computação. Com um limitado número de consultas (13 no total) e avaliação não validada por especialistas, essa coleção apresenta, como maior vantagem, o significativo número de documentos (mais de 1,6 milhão) e de candidatos (mais de 1 milhão), além de ser a única coleção de teste acadêmica estudada que deu início a uma máquina de busca de especialistas implementada e disponível, a AMiner.

	domínio	ano	multi-organização	multi-área	aval. por especialista	#candidatos	#documentos	#consultas	#qrels	tamanho
W3C	empresarial	2004				1.092	331.037	99	9.860	5,70
CERC	empresarial	2007		✔	✔	3.500	370.715	127	2.862	4,20
UvT	academia	2006		✔		1.168	36.699	1.491	4.318	0,31
TU	academia	2008		✔		977	32.567	1.266	2.112	1,11
ArnetMiner	academia	2008	✔		✔	1.033.050	1.632.440	13	1.781	0,85

Tabela 2.1: Principais características das coleções de teste de busca de especialistas.

Na Tabela 2.1, são apresentadas algumas das principais características de cada coleção de teste de busca de especialistas estudada. Os números de *qrel* referenciam a quantidade de respostas no gabarito das consultas e o tamanho das coleções está em GB, considerando as coleções descompactadas. Além disso, são marcadas as coleções que possuem mais de uma organização no conjunto de candidatos a especialistas (deno-

tada por "multi-organização"), as coleções que possuem mais de uma área no conjunto de consultas (denotada por "multi-área"), e se as respostas esperadas para as consultas foram avaliadas por especialistas ou por outros meios de validação do resultado (denotada por "aval. por especialista").

Na Seção 4.1, apresentaremos a coleção de teste proposta nesta dissertação, construída a partir dos currículos dos doutores da plataforma Lattes. As duas principais justificativas para construirmos uma coleção de teste nova são: (1) devido a características peculiares à plataforma Lattes, como informação estruturada e identificadores únicos para os autores; e (2) geração de um gabarito respondido exclusivamente por especialistas das consultas através da aplicação de dois questionários. Em particular, a aplicação de dois questionários evita a introdução de vieses associados à visão de um candidato quanto ao seu próprio nível de especialidade. O restante dos detalhes da construção da coleção serão apresentados na Seção 4.1.

2.3.2 Métricas de Avaliação dos Resultados

Nesta seção, apresentaremos as métricas usadas nesta dissertação que, na sua maioria, são tradicionais em avaliação de rankings. A primeira métrica usada para avaliação dos resultados computa a porcentagem de candidatos relevantes nos top-k candidatos retornados pelo modelo. Denominada $precision_k$ (P_k), essa métrica considera todos os candidatos com nível de relevância maior que zero como igualmente relevantes.

Além disso, foi usada a métrica *Mean Reciprocal Rank* (MRR) visando avaliar a qualidade do ranking em função da posição do primeiro candidato relevante. A Equação 2.26 apresenta a formalização da métrica:

$$MRR = \frac{1}{N_Q} \sum_{q} \frac{1}{first\_rel_q} \tag{2.26}$$

onde $first\_rel_q$ é a posição do primeiro relevante que ocorreu no ranking da consulta q e N_Q é a quantidade de consultas a serem avaliadas. Além disso, assim como a métrica P_k, os candidatos que ocorrerem no gabarito já são denotados como relevantes.

Por último, uma das mais importantes métricas para quantificação da qualidade do ranking é a $nDCG_k$ (*Normalized Discounted Cumulative Gain*), onde k é o tamanho do ranking gerado pelo modelo para uma determinada consulta que será avaliada. O resultado dessa métrica está no intervalo $[0, 1]$, e quanto maior esse valor, melhor o ranking. A Equação 2.27 apresenta a formalização dessa métrica de avaliação, dado

um ranking de tamanho k:

$$nDCG_k = \frac{DCG_k}{IDCG_k}, \tag{2.27}$$

onde $IDCG_k$ seria o DCG_k do ranking ideal, ou seja, o ranking determinado pelo gabarito da consulta e DCG_k é definido como:

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2 (i + 1)}, \tag{2.28}$$

onde rel_i é a relevância do i-ésimo candidato no ranking.

Todas as métricas mencionadas nesta seção foram baseadas na implementação disponível pela *Text REtrieval Conference* (TREC), denominada trec_eval[5].

[5] `http://trec.nist.gov/trec_eval/`

Capítulo 3

Modelos de Associação Documento-Candidato

Seguindo o mesmo processo de geração de peso de associação apresentado nos trabalhos relacionados, onde o peso de uma associação pode ser descrito como

$$f(d, e) = \psi(\rho(d, e)) \tag{3.1}$$

onde o modelo de associação $f(d, e)$ que pondera a associação do documento d para o candidato e é definido como uma combinação de uma função de associação (ρ) e uma função de normalização (ψ), propomos um conjunto de possíveis instanciações dessas funções descrevendo diferentes semânticas para cada peso. Em particular, neste capítulo são apresentadas quatro instanciações da função de associação (Seção 3.1) e duas da função de normalização, que visam remover o viés das normalizações propostas em Balog et al. [2006] para documentos ou candidatos com poucas associações (Seção 3.2).

A Figura 3.1 demonstra um exemplo de associações e entidades em um sistema de busca de especialistas. Nesse exemplo, os pesquisadores "1" e "2" são coautores do documento "D", além de cada um possuir sua própria lista de documentos publicados. Cada documento possui um ano de publicação, por exemplo, o documento "D" foi publicado em 2005, o documento "A" em 2002 e assim sucessivamente.

A função de associação é fundamental para a definição do peso de uma associação de forma individual. Nesse caso, seria como definir a importância do autor "1" para o documento "D", segundo o exemplo da Figura 3.1. Enquanto isso, as funções de normalização são importantes para ponderar a importância de cada associação em relação a todas as associações paralelas a ela. Por exemplo, dada uma função de

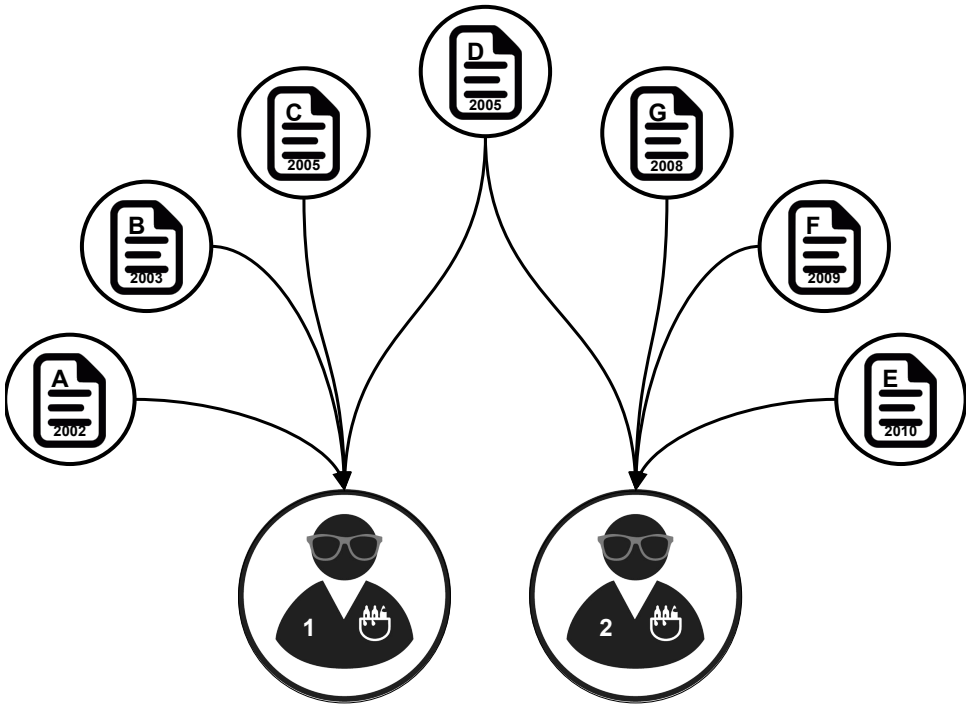

Figura 3.1: Exemplo de associações e entidades em um sistema de busca de especialistas.

associação, definir qual documento é mais importante para o pesquisador "1", qual autor é mais importante para o documento "D".

A seguir serão apresentadas as funções de associação e normalização propostas nesta dissertação, sempre exemplificando seguindo o mesmo caso da Figura 3.1.

3.1 Funções de Associação

Como apresentado nos trabalhos relacionados, raramente os estudos desenvolvidos em busca de especialistas tentam criar associações com semânticas mais intrínsecas da relação documento-candidato. Uma justificativa para esse número reduzido de trabalhos específicos sobre associações é que boa parte das coleções de busca de especialistas são fundamentadas em bases empresariais e a ponderação das associações visa inferir a probabilidade do candidato mencionado no documento realmente ser o especialista buscado.

Enquanto isso, nesta dissertação, introduzimos quatro formas de ponderar cada associação documento-candidato. Os conceitos usados tentam modelar a relação usando conteúdo textual, textual-temporal e puramente temporal. As funções de associação propostas serão descritas a seguir.

3.1.1 Dominância de Conteúdo

Enquanto Macdonald et al. [2008] modelam o peso da associação quantificando a distância que o documento está dos temas de maior interesse do candidato, a abordagem de dominância de conteúdo proposta nesta dissertação modela diretamente o interesse do candidato para o documento. A intuição por trás dessa função de associação é que documentos fortemente relacionados a todos os trabalhos do candidato tendem a ser documentos sobre os quais o candidato possui maior domínio.

Ao dizer que o candidato tem domínio sobre o documento, considera-se a capacidade do candidato de produzir, a partir de sua representação, o conteúdo textual do documento, sendo a representação do candidato criada a partir da união do conteúdo textual de todos os documentos que o candidato já publicou. Dessa forma, ao representar a dominância de conteúdo quantifica-se, seguindo o caso da Figura 3.1, o peso da associação entre o pesquisador candidato "1" e o documento "A", por exemplo, segundo o domínio que o pesquisador "1" tem sobre o conteúdo apresentado no documento "A".

Para representar essa função de associação, supõe-se que o conteúdo de todos os trabalhos publicados pelo candidato é uma consulta para o modelo θ_d da distribuição dos termos do documento d. No modelo linguístico convencional de recuperação de informação, podemos representar a verossimilhança da consulta como:

$$P(q|\theta_d) = \prod_{t \in q} P(t|\theta_d)^{n(t,q)}. \tag{3.2}$$

onde a ordem dos termos é independente, tal como os documentos são independente entre si.

É convencional, ainda nos modelos linguísticos, desconsiderar a constante do coeficiente multinomial K_q (Seção 2.1.2.1) e considerar a escala logarítmica da verossimilhança, por não alterar a ordem do ranking final e, por questões numéricas, ser uma escala mais fácil de ser computada. Assim, obtemos a log-verossimilhança da consulta como sendo:

$$\log P(q|\theta_d) = \sum_{t \in q} n(t,q) \log P(t|\theta_d). \tag{3.3}$$

Na suposição central da função de dominância de conteúdo, consideramos que a consulta é o modelo linguístico formado por todos os documentos do candidato, obtendo

assim:

$$\log P(\theta_e|\theta_d) = \sum_{i \in \theta_e} n(i, \theta_e) \log P(i|\theta_d). \tag{3.4}$$

Segundo Lavrenko & Croft [2003], uma versão mais genérica da log-verossimilhança da consulta é estimada considerando a entropia cruzada quando incluímos o fator de normalização em $n(i, \theta_e)$, dividindo esse valor pelo tamanho do modelo θ_e. Nesse caso, log-verossimilhança da consulta é um caso especial das abordagens de relevância considerando a entropia cruzada, como é apresentado na Equação 3.5.

$$\begin{aligned} \rho_d(d, e) &= -\sum_i P(i|\theta_e) \log P(i|\theta_d) \\ &= H(\theta_e||\theta_d) \end{aligned} \tag{3.5}$$

Usando essa modelagem ainda é possível interpretar a associação documento-candidato como sendo uma maneira de quantificar quão informativo é o documento em relação as áreas de interesse do candidato a especialista. Em teoria da informação, entropia cruzada mede o custo aproximado de transformar a distribuição θ_d, na distribuição esperada θ_e.

Balog & De Rijke [2008] introduziram a entropia cruzada em seu modelo denominado *Semantic-Relatedness* na etapa que pondera o número de ocorrências dos identificadores dos autores no modelo construído para o candidato. Assim, eles substituem a contagem dos identificadores por essa ponderação e atribuem uma nova interpretação para a associação: o peso de transformar o documento d no modelo do candidato associado e.

Ao contrário do que a intuição da entropia cruzada diz, onde valores mais altos da função representam associações mais fracas, essa modelagem de determinação do custo de transformar a distribuição dos termos do documento d na distribuição dos termos do candidato associado e demonstrou maior representatividade da dominância do conteúdo do candidato para o documento, como é demonstrado na tabela de resultados no Apêndice A. Portanto, deste ponto em diante, tornamos padrão o uso de entropia cruzada como medida de dominância de conteúdo, sendo essa usada nas próximas duas funções de associação, apresentadas na seção seguinte e na Seção 3.1.3.

3.1.2 Estabilidade da Dominância

Quando se trabalha com coleções ambientadas na academia, um fator que pode ser levado em consideração durante o processo de determinação da importância de uma associação é a idade do documento. Usando essa premissa, nesta dissertação é proposta uma função de ponderação de associação que estima a capacidade do autor de ser relevante em relação ao tema de um determinado documento durante um período de tempo.

Keikha et al. [2011], propuseram um modelo de estabilidade temporal aplicado ao problema de busca na blogsfera. Esse problema tem características similares à busca de especialistas, havendo um candidato a "especialista" (um blogger) e um conjunto de documentos associação a esse blog (as postagens do blogger). Uma das soluções para esse problema é similar às soluções generativas probabilísticas apresentadas nesta dissertação, onde o modelo estima a relevância de um blogger através das estimativas de relevância das postagens encontradas em seu blog. Além disso, os modelos de associação, assim como na busca de especialistas, são pouco estudados. Contudo, Keikha et al. [2011] apresentaram um modelo que estima, de certa forma, a probabilidade de uma postagem (um documento) para um blogger. O cerne do modelo apresentado em Keikha et al. [2011] é estimar a estabilidade da relevância do blogger no decorrer de suas postagens.

A partir da intuição de estabilidade de relevância apresentado por Keikha et al.

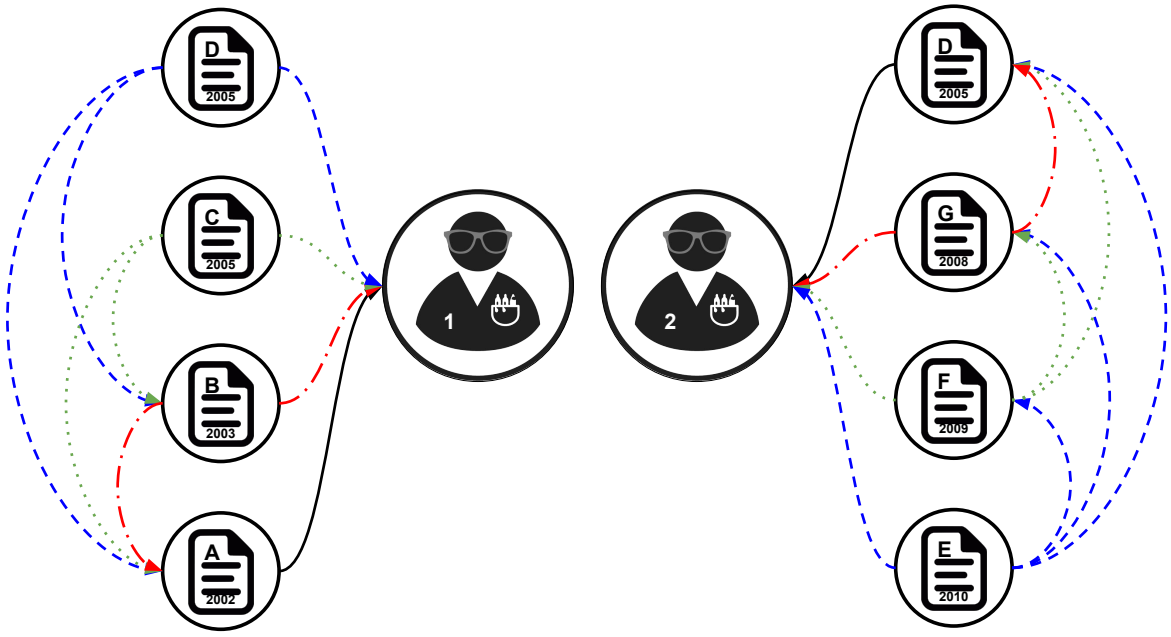

Figura 3.2: Exemplo do funcionamento da estabilidade da dominância.

[2011], adaptamos esse conceito para o problema de ponderação de associações da busca de especialistas. Nesse novo modelo, a estabilidade da relevância é estimada como a estabilidade do domínio do candidato a especialista diante dos documentos retornados pela consulta. Intuitivamente, isso poderia ser descrito como a capacidade do autor de se manter estável no tema da consulta.

Para fazer isso, consideramos que a estabilidade de um candidato para o tema de um documento é o valor médio do domínio de conteúdo que esse documento tem, em relação a todos os documentos publicados anteriormente pelo autor. Nesse caso, o candidato que publica constantemente documentos similares em diferentes anos é um candidato mais propício a ser um especialista para a consulta. Intuitivamente, isso demonstra que o candidato vem trabalhando no decorrer dos anos nesse tema, podendo ser um forte candidato a especialista dada a experiência em publicações daquele tema.

Exemplificando essa função no caso da Figura 3.1, demonstramos, na Figura 3.2, quais documentos seriam comparados para a construção de cada associação dos pesquisadores "1" e "2" com seus documentos. Por exemplo, para estimar o peso da associação entre o documento "D" e o candidato "1", estima-se a dominância de conteúdo (deste ponto em diante, denominado apenas como dominância) de "D" sobre os documentos "B" e "A", para então se estimar a estabilidade do conteúdo de "D" para o candidato "1". Enquanto isso, o documento "D" para o candidato "2" representa a sua primeira publicação, portanto, o peso de "D" para o candidato "2" é irrisório em comparação com o peso dos documentos "G", "F" e "E", desde que esses sejam documentos com conteúdo similar.

A formalização dessa função considera $t = time(d)$ como sendo o ano que o documento d foi publicado e o conjunto D_e^t como sendo o conjunto dos documentos que o candidato a especialista e publicou antes do tempo t, representado como $D_e^t = \{d' : time(d') < t \wedge d' \in D_e\}$. Assim, a função ρ que estima a estabilidade de dominância de conteúdo é dada por:

$$\rho_{ed}(d, e) = \frac{1}{|D_e^t|} \sum_{d' \in D_e^t} H(\theta_{d'} || \theta_d), \tag{3.6}$$

onde $H(\theta_{d'}||\theta_d)$ é a entropia cruzada entre o documento d a ser avaliado e o documento d' anteriormente publicado, $|D_e^t|$ é o tamanho do conjunto dos documentos publicados pelo candidato e antes do tempo do documento d.

3.1.3 Novidade no Domínio

Um terceiro aspecto que pode ser levado em consideração durante o processo de determinação da importância de um documento para um candidato é a novidade do conteúdo do documento para a coleção. A intuição por trás dessa ideia é que, para algumas áreas, o pesquisador precursor pode ser tão importante quanto um pesquisador que domina a área recentemente.

Da mesma maneira que a função de associação de estabilidade, essa função de associação, denominada novidade no domínio, modela a distribuição dos termos no tempo. Contudo, nesse caso, a distribuição está relacionada aos termos de todos os documentos na coleção publicados anteriormente ao ano do documento a ser comparado. Individualmente, essa forma de ponderar a associação não considera informação do candidato, sendo apenas uma representação da importância do documento.

Analogamente à modelagem da função de estabilidade de dominância, introduzimos a novidade do domínio como sendo uma função que representa a importância do documento avaliado com relação a um conjunto de documentos precursores. A diferença principal é que é usada uma representação *big document* da coleção temporalmente acumulada para determinar a dominância desse conjunto para o documento em questão. A modelagem dessa forma determinaria a probabilidade do documento ter sido publicado anteriormente e, sendo que queremos representar o contrário, invertemos esse valor, assim:

$$\rho_{nd}(d, e) = \frac{1}{H(\theta_\tau || \theta_d)}, \tag{3.7}$$

onde θ_τ é a representação do conteúdo de todos os documentos anteriores ao documento d. Para exemplificar essa situação, usando o caso apresentado na Figura 3.1, demonstramos, para os documentos "C" e "D", como seria calculada a função ρ de associação com o candidato "1" na Figura 3.3. Como é apresentado, ambos os documentos "C" e "D" têm a entropia cruzada calculada em relação a um modelo $\theta_{\tau_{2005}}$. Contudo, o resultado é dependente do conteúdo de cada documento, além de ser combinado com a relevância estimada de cada um deles em relação à consulta efetuada.

3.1.4 Recência

As ultimas duas funções de associação propostas são baseadas na idade do documento. Denominadas funções de recência, a intuição dessas funções é de que candidatos a especialistas com publicações mais recentes são candidatos cuja pesquisa é mais evoluída. A

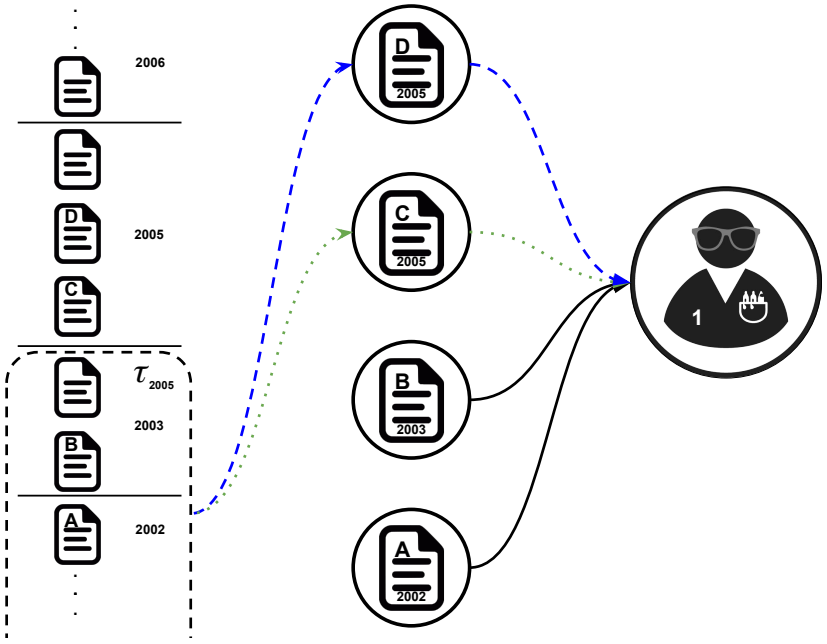

Figura 3.3: Exemplo do funcionamento da novidade do domínio.

ideia central delas é oposta à função de novidade no domínio, mas não usa o conteúdo textual do documento. Portanto, consistem de funções puramente temporais.

Para estimar a recência de um documento d com relação a um conjunto X, ajustamos o ano de publicação do documento com relação aos anos representados em X, de acordo com:

$$t(d, X) = \frac{time(d) - min\_time(X) + 1}{max\_time(X) - min\_time(X) + 1}, \tag{3.8}$$

onde X é o conjunto com relação ao qual o documento d deve ser normalizado; $time(d)$, como mencionado anteriormente, é o ano do documento d; e $min\_time(X)$ e $max\_time(X)$ é o menor e o maior ano da coleção X de documentos, respectivamente. Além disso, é usada a suavização laplaciana [Manning et al., 2008] para evitar resultados zerados durante a normalização.

Seguindo o mesmo caso da Figura 3.1, vamos supor que os candidatos "1" e "2" sejam um subconjunto de candidatos da coleção, cujo primeiro documento foi datado em 1961 e último em 2015 (tal como é a coleção de teste avaliada nesta dissertação). Através da função normalização descrita na Equação 3.8, pode-se ponderar o documento

"A" para o candidato "1" através da transformação

$$\begin{aligned} t(\text{“A”}, D_{e_1}) &= \frac{time(\text{“A”}) - min\_time(D_{e_1}) + 1}{max\_time(D_{e_1}) - min\_time(D_{e_1}) + 1} \\ &= \frac{2002 - 2002 + 1}{2005 - 2002 + 1} \\ &= 0,25 \end{aligned} \tag{3.9}$$

onde D_{e_1} é o conjunto dos documentos publicados pelo candidato "1", $max\_time(D_{e_1})$ é o ano do último documento publicado pelo candidato "1", que no caso, é o documento "D", de 2005 e $min\_time(D_{e_1})$ é o anjo do primeiro documento publicado pelo candidato "1", que no caso, é o próprio documento "A", de 2002, respectivamente. Assim, podemos dizer o quão recente o documento "A" é para o candidato "1", ponderando esse valor por 0,25. De forma análoga, podemos dizer o quão recente um documento "A" é em relação a todos os documentos da coleção, aplicando a Equação 3.8, como é apresentado abaixo

$$t(\text{“A”}, D) = \frac{2002 - 1961 + 1}{2015 - 1961 + 1} \approx 0,764. \tag{3.10}$$

Assim, dadas as representações temporais do documento em relação ao candidato e em relação a coleção, são propostas duas abordagens que combinam essas informações. Uma justificativa para combinarmos as informações é que ambos os aspectos temporais do documento devem ser levados em consideração quando se pretende mensurar o quão recente um documento é. Dessa forma, para uma relação ser considerada recente, o documento deve ser recente para a coleção e, ao mesmo tempo, para o candidato. Na situação exemplo, o documento "A" de 2002 é, relativamente, novo para a coleção, que tem seu primeiro documento datado em 1961. Enquanto isso, o documento "A" é o primeiro e mais antigo documento do candidato "1". Espera-se que esse tipo de documento seja menos relevante para o candidato do que para a coleção, dada a função de associação proposta.

Considerando essa premissa, a primeira função de associação puramente temporal proposta considera ambas as informações como igualmente importantes, contudo, dependentes entre si. Assim, obtemos a função de associação ρ_{rm} que denominamos de função de recência linear, definida como:

$$\rho_{rm}(d, e) = t(d, D_e) \times t(d, D) \tag{3.11}$$

A Figura 3.4 apresenta o gráfico de curvas de contorno onde as curvas demonstram

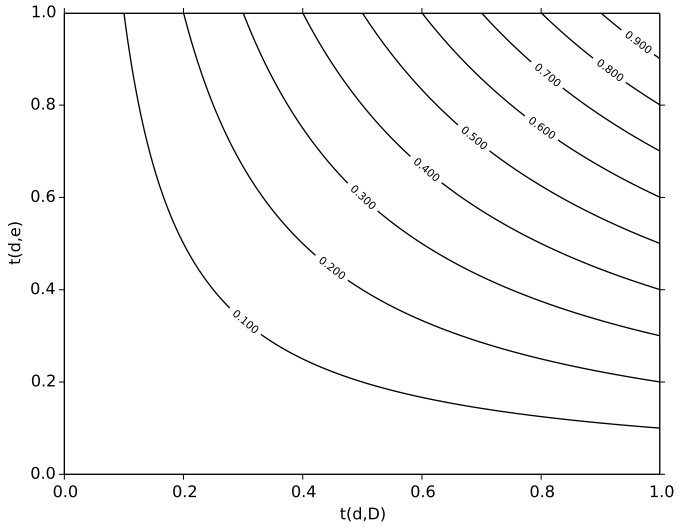

Figura 3.4: Exemplo do decaimento monotônico da função de associação ρ_{rm}.

os valores de decaimento resultantes da função ρ_{rm}, sendo os eixos X e Y iguais a $t(d, D)$ e $t(d, D_e)$, respectivamente. Quando fixamos qualquer um dos eixos, o valor final de $\rho_{rm}(d, e)$ obtido segue um decaimento linear. A intuição dessa característica foi derivada da quantificação da importância de associações entre coautores, apresentado por Xia et al. [2014], onde as associações de coautoria são ponderadas com um decaimento temporal em relação à sua ocorrência e aplicadas no processo de recomendação de coautorias.

Assim, se considerarmos o exemplo anterior, onde o objetivo é quantificar a associação entre o candidato "1" e o documento "A", obtemos através da função ρ_{rm} o valor para o peso da associação ("1","A") como:

$$\rho_{rm}(\text{“A”}, e_1) = t(\text{“A”}, D_{e_1}) \times t(\text{“A”}, D) \approx 0,191 \tag{3.12}$$

Enquanto isso, a segunda função de associação puramente temporal proposta, que se baseia no decaimento exponencial da importância do documento em relação ao tempo, é definida como:

$$\rho_{re}(d, e) = t(d, D)^{1-t(d,D_e)} \tag{3.13}$$

A intuição de nossa função foi adaptada da função proposta por Li & Croft [2003], onde é incorporada a probabilidade a priori do documento à função de ranking, definida como:

$$P(d|time(d)) = \lambda \times \exp(-\overline{t}(d)) \tag{3.14}$$

onde $\lambda \in [0, 1]$ é um hiper parâmetro da função; e

$$\bar{t}(d) = \frac{(max\_time(D) - time(d))^2}{2}. \tag{3.15}$$

No trabalho apresentado por Li & Croft [2003], os autores propuseram esse modelo exponencial de probabilidade a priori do documento sob a justificativa de que, para algumas consultas, a necessidade de informação da pessoa que está realizando a consulta é enviesada a favor de documentos mais recentes. Essa é uma das justificativas apresentadas para a avaliação das funções de associações puramente temporais que visam quantificar o quão recente uma associação é em relação ao candidato e à coleção.

Nessa formulação da probabilidade a priori, os autores queriam aproximar a Equação 3.15 utilizando uma outra função de probabilidade a priori do documento, também apresentado em Li & Croft [2003], onde a distribuição normal é utilizada para determinar a importância do documento no período em que foi datado. No caso da função de associação proposta, substituímos a função apresentada na Equação 3.15 pela função de normalização da Equação 3.8, usando como conjunto de comparação os documentos do candidato associado ao documento ($t(d, D_e)$).

Assim, a intuição da Equação 3.14 foi usada para a formulação da função de decaimento em relação ao candidato e à coleção, apresentada na Equação 3.13. Seguindo o mesmo exemplo da função de associação de recência linear, podemos ponderar o quão recente o documento "A" é para o candidato "1" usando a função de associação exponencial $\rho_{re}(\text{“A”}, e_1)$ de acordo com:

$$\rho_{re}(\text{“A”}, e_1) = t(\text{“A”}, D)^{1-t(\text{“A”}, D_{e_1})} = 0,806. \tag{3.16}$$

Numa comparação direta dos resultados das associações puramente temporais vemos que a função linear, que apresenta 0,191 como resultado, é mais rígida quanto à recência do documento para o candidato, enquanto a função exponencial é mais suave nesse aspecto, resultando um peso maior para a associação devido a recente idade do documento em relação à coleção.

A Figura 3.5 apresenta a curva de contorno da função $\rho_{re}(d, e)$, onde é demonstrado que, conforme a recência do documento para a coleção ($t(d, D)$) fica maior, é menos exigido que o documento seja recente para o candidato ($t(d, D_e)$).

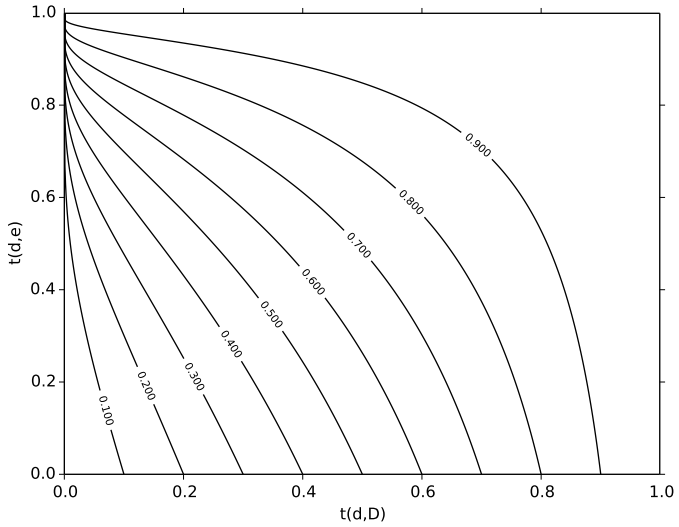

Figura 3.5: Exemplo do decaimento exponencial da função de associação ρ_{re}.

3.2 Normalização Centrada na Auto-informação

Como mencionado, as funções de normalização são o segundo importante componente de um modelo de associação documento-candidato no processo de busca de especialistas. No caso específico das funções de normalização, o objetivo principal é demonstrar qual a relação entre a associação comparada e as demais associações paralelas a ela.

No exemplo da Figura 3.1, os autores "1" e "2" têm um documento em comum nas suas respectivas listas de publicações: o documento "D". Segundo os trabalhos apresentados no Capítulo 2, existem três maneiras de normalizar o valor de uma associação, ou seja, três diferentes formas de dizer qual a importância daquela associação (ou o peso dela) em relação às outras associações paralelas. Nesta dissertação, apresentamos uma quarta maneira de normalizar essa associação em relação às associações paralelas, formulada como:

$$\psi_{SiC}(\bullet) \equiv \log\left(\frac{\left(\sum_{(d',e')\in X}\rho(d',e')\right)^{\alpha}+1}{\bullet+1}\right) \tag{3.17}$$

onde α é o hiper-parâmetro da normalização e o conjunto X é determinado pela centralidade usada na normalização, por exemplo, para X centrada no documento d, temos que $X = \{(d, e') : d \to e'\}$, ou, analogamente, para X centrada no candidato e, temos que $X = \{(d', e) : d' \to e\}$. Além disso, os valores 1 somados no divisor e no dividendo são respectivos a suavização Laplaciana, usada aqui para evitar inconsistências no cálculo da logarítmica. Portanto, em termos gerais, temos que a normalização proposta

pode ser definida como:

$$\psi_{SiC}(\bullet) \equiv \log\Big(\frac{\big(\sum_{(d',e')\in X}\rho(d',e')\big)^{\alpha}}{\bullet}\Big). \tag{3.18}$$

Denominada normalização centrada na auto-informação (SiC[1]), a função de normalização proposta é inspirada no conceito de auto-informação, em teoria da informação. Formulada como I(ω), onde I é a função que mensura o valor esperado do evento ω, a auto-informação é uma função crescente e positiva. Por definição, a quantidade de auto-informação contida no evento ω depende apenas da probabilidade do evento ocorrer: quanto menor for a sua probabilidade, maior é a auto-informação associada com a informação do evento. Até por isso, essa medida também tem sido chamada de *surprisal*, uma vez que representa a "surpresa" de ver o resultado (um resultado altamente improvável é muito surpreendente) [Tribus, 1961].

A Equação 3.19 apresenta a formulação tradicional da auto-informação. Nesse formato, dizemos que a entropia da informação de um evento aleatório ω é o valor esperado da auto-informação [Cover & Thomas, 2012]. Assim:

$$\mathrm{I}(\omega) = \log\Big(\frac{1}{P(\omega)}\Big), \tag{3.19}$$

onde $P(\omega)$ é a probabilidade do evento ω ocorrer.

Como mencionado, para o problema de modelagem de associação para busca de especialistas, dizemos que $P(\omega)$ pode deduzir duas interpretações: (1) a probabilidade da associação existir, que é equivalente à probabilidade do nome mencionado no documento ser do candidato associado; ou (2) a probabilidade da relação entre o documento-candidato em termos conteúdo-temporais. Em especial, ambas as alternativas possuem duas direções a se considerar: A probabilidade do documento gerar o candidato; ou a probabilidade do candidato gerar o documento. Apesar da aparente equivalência, podemos formalizar as direções como $P(e|d)$ e $P(d|e)$, respectivamente.

Além disso, essas duas direções já foram formalizadas nas abordagens generativas propostas por Balog et al. [2006]. Assim, podemos denotar $P(\omega)$ de duas maneiras, onde:

$$P(\omega) = \begin{cases} P(d|e), & \text{Se centrarmos a normalização no documento;} \\ P(e|d), & \text{Se centrarmos a normalização no candidato,} \end{cases} \tag{3.20}$$

sendo que $P(d|e)$ e $P(e|d)$ já foram introduzidas como funções de normalização na

[1]Do inglês, *Self-information Centric*

Seção 2.2.2, nas Equações 2.23 e 2.24, respectivamente.

A partir de $\alpha = 1$ é possível apresentar a conversão da função de auto-informação para a função de normalização considerando apenas a centralidade no documento, sendo que a centralidade no candidato consiste de um processo similar. Assim, digamos que $\psi_{SDC}(\bullet)$ consiste da função de normalização centrada na auto-informação do documento[2]. Portanto, formalização a função de normalização como:

$$\psi_{SDC}(\bullet) \equiv \log\Big(\frac{1}{P(d|e)}\Big). \tag{3.21}$$

Como mencionado anteriormente, no modelo de associação introduzido por Balog et al. [2006], $P(d|e)$ equivale a função de normalização $\psi_{DC}(\bullet)$ apresentada na Seção 2.2.2. Assim, sendo que essa função de normalização é apresentada na Equação 2.23, obtemos, pela formalização,

$$\begin{aligned}\psi_{SDC}(\bullet) &\equiv \log\Big(\frac{1}{\frac{\bullet}{\sum_{e' \in E_d} \rho(d,e')}}\Big) \\ &\equiv \log\Big(\frac{\sum_{e' \in E_d} \rho(d,e')}{\bullet}\Big).\end{aligned} \tag{3.22}$$

De forma similar, podemos construir a função $\psi_{SCC}(\bullet)$ de normalização centrada na auto-informação do candidato[3] como:

$$\psi_{SCC}(\bullet) \equiv \log\Big(\frac{\big(\sum_{d' \in D_e} \rho(d',e)\big)}{\bullet}\Big). \tag{3.23}$$

O parâmetro α foi inserido como um critério de estimação da auto-informação contida na normalização, sendo que, para $\alpha \geq 1$, a normalização trás um viés a favor das associações paralelas à associação avaliada, enquanto para $\alpha \leq 1$ o viés é a favor do peso da associação avaliada. Para o parâmetro $\alpha = 1$, a função de normalização equivale exatamente a auto-informação da associação, representada como $\mathrm{I}(\omega)$.

No próximo capítulo, descrevemos a configuração experimental que suporta a avaliação empírica das funções de associação e normalização propostas neste capítulo. Os resultados dessa avaliação são apresentados e discutidos no Capítulo 5.

[2] *SDC* vem do inglês, *Self-information Document-Centric.*
[3] *SCC* vem do inglês, *Self-information Candidate-Centric.*

Capítulo 4

Metodologia de Avaliação

Com intuito de esclarecer o processo de experimentação das funções de associação e normalização, neste capítulo são apresentados os principais pontos levantados para o funcionamento da experimentação. Tais pontos foram divididos em três aspectos principais: (1) a coleção de teste usada para a validação dos resultado, (2) os *baselines* utilizados e o processo de definição dos parâmetros e as suas configurações experimentais (Seção 4.2), e (3) o processo de treino e teste usado na validação das abordagens discriminativas de ranking de especialistas (Seção 4.3).

4.1 Coleção de Teste

Como mencionado no Capítulo 1, uma das contribuições desta dissertação é a apresentação de uma coleção de teste para busca de especialistas no ambiente acadêmico. De fato, poucas coleções desse tipo foram construídas e disponibilizadas, como mencionado na Seção 2.3, demonstrando a contribuição da apresentação desta coleção de teste.

Assim, o foco principal desta seção é apresentar os processos executados para a construção da coleção de teste de busca de especialistas. Tais processos são divididos em duas grandes etapas: (1) construção da coleção de documentos e associações, e (2) construção do gabarito das consultas do processo de busca. Essas etapas são apresentadas nas Seções 4.1.1 e 4.1.2, respectivamente.

4.1.1 Construção dos Conjuntos de Documentos e Associações

O conjunto de documentos da coleção de teste usada nesta dissertação foi baseado na plataforma Lattes. Tal plataforma representa a experiência do Conselho Nacional de

Desenvolvimento Científico e Tecnológico (CNPq)[1] na integração de bases de dados de currículos, de grupos de pesquisa e de instituições em um único sistema de informações. Sua dimensão atual se estende não só às ações de planejamento, gestão e operacionalização do fomento do CNPq, mas também de outras agências de fomento federais e estaduais, das fundações estaduais de apoio à ciência e tecnologia, das instituições de ensino superior e dos institutos de pesquisa.

Devido a sua grande adoção pela maioria das instituições de fomento, universidades e institutos de pesquisa do Brasil, a plataforma Lattes se tornou amplamente usada como coleção de teste para diferentes tarefas, como, por exemplo, caracterização da pesquisa brasileira [Balancieri et al., 2005; Barbosa et al., 2009; Digiampietri et al., 2012]; construção de ferramentas para extração de dados [Mena-Chalco, 2009; Alves et al., 2012]; teses de doutoramento em ontologia ou organização de informação [Castano, 2008; Silva, 2007]; tarefas de sumarização ou recomendação de especialidades de pesquisadores [Ribeiro et al., 2015]; e, como apresentado nesta dissertação, busca de especialistas no ambiente acadêmico [Mangaravite & Santos, 2016].

Segundo informações da plataforma, o Lattes tem hoje mais de 4 milhões de currículos de pesquisadores, alunos e funcionários associados a pesquisa em diferentes níveis técnicos. Para construção da coleção de teste, foram usados apenas os documentos extraídos dos currículos dos candidatos que informaram ser doutores na plataforma. Nas coleções coletadas em 2014 e 2015, compondo currículos de toda a plataforma Lattes ou apenas dos doutores, respectivamente, analisamos algumas estatísticas extraídas e as apresentamos na Tabela 4.1.

Tabela 4.1: Estatísticas das coleções de documentos e associações.

	Doutores	**% do total**	**Lattes**
#Documentos	11.942.014	72,26%	16.526.452
#Candidatos	223.853	6,54%	3.423.548
#Candidatos[2]	206.697	21,19%	975.470
#Associações	21.015.538	69,75%	30.128.338
Associações por Documento	1,76		1,82
Associações por Candidato[2]	101,67		30,89
Documentos por Candidato[2]	57,78		16,94

A Tabela 4.1 apresenta as seguintes estatísticas a respeito das duas coleções do Lattes coletadas: número de documentos na coleção (#Documentos), número de currículos, onde cada currículo pertence a apenas um pesquisador (#Candidatos), número

[1] `http://memoria.cnpq.br/web/portal-lattes/sobre-a-plataforma`.

[2] Candidatos com pelo menos uma associação.

de currículos que têm pelo menos uma publicação associada (#Candidatos (com associações)), número de associações entre os currículos e os documentos (#Associações), e, por último, três avaliações das médias de associações por documento (Associações por Documento), das associações por candidato que possui pelo menos um documento associado (Associações por Candidato) e número de documentos por candidato com pelo menos um documento associado (Documentos por Candidato).

A escolha de se usar apenas os documentos dos currículos de doutores se baseia em algumas justificativas: (1) Com 6,5% dos currículos do Lattes (quantidade de doutores), cobrem-se 72% de todos os documentos extraídos da coleção e 69% das associações encontradas (o processo de construção das associações será descrito posteriormente); (2) Apesar de a quantidade média de associações por documento aumentar para a coleção toda do Lattes, o número de associações por candidato cai de 101 para 30 e a quantidade de documentos por candidato cai de 57 para 16. Isso sugere que uma significativa parcela das publicações da comunidade acadêmica do Brasil tem, na lista dos autores, pelo menos um doutor identificado pelo processo de extração e construção de associação; e (3) Como será apresentado na Seção 4.1.2, todos os candidatos a especialistas no gabarito das consultas foram identificados como doutores na coleção.

Como se espera de uma plataforma de currículos, a plataforma Lattes permite armazenar uma quantidade muito variada de informações, como referências bibliográficas, experiências profissionais, instituições vinculadas, endereço profissional, etc. Considerando o escopo específico de construção de uma abordagem de busca de especialistas no ambiente acadêmico, consideramos apenas metadados bibliográficos referentes a documentos de cinco tipos específicos, sendo esses tipos e suas respectivas proporções da coleção: artigos completos (publicações em periódicos, 23,48%); trabalhos publicados em anais de eventos (50,58%); apresentação de artigos em eventos (19,90%); e livros completos ou capítulos de livros (6,04%).

Cada currículo armazenado na plataforma Lattes possui um identificador único, representado na coleção como sendo um número de onze dígitos. Além disso, nem todas as menções das autorias das referências bibliográficas dos currículos apontam os coautores pelos seus identificadores e, apesar de haver métodos sofisticados propostos como solução desse problema, reconciliar os coautores das publicações pela menção nominal seria de grande complexidade. Além disso, identificar erroneamente coautores de uma determinada publicação tornaria a solução de modelagem de associação proposta sensível a fatores externos às interpretações propostas para cada associação documento-candidato.

Apenas 52,5% das 58 milhões de menções aos autores nos documentos da plataforma Lattes têm o identificador único explícito dos currículos nas referências. Dessa

forma, não consideramos usar nenhum processo sofisticado de identificação de duplicatas das instâncias de documento, dada a complexidade que seria um processo de reconciliação de documentos duplicados. Para demonstrar um possível problema dessa abordagem, consideremos o mesmo exemplo usado na Seção 3, apresentado na Figura 3.1, onde os autores "1" e "2" são coautores do documento "D". Caso o autor "1" não coloque o documento "D" em sua própria lista de publicações no currículo na plataforma Lattes e o autor "2" não tenha referenciado explicitamente o identificador do autor "1" na lista dos autores do documento "D", o autor "1" não será reconhecido como um dos possíveis autores de "D" na coleção de teste apresentada.

Espera-se que esse tipo de situação seja um caso excepcional ou que não interfira significativamente no resultado final do ranking de especialistas devido a alguns motivos justificados pelas intuições: (1) se o autor "1" não tiver colocado o documento "D" em sua própria lista de publicações, isso pode demonstrar que o documento "D" não é significativo para o autor "1"; (2) dado o fato que nem todos os autores das publicações do Lattes são, exclusivamente, pesquisadores que têm currículo na plataforma, diminuímos os falso-positivos possíveis de um processo de reconciliação. Em outras palavras, as associações construídas pelo processo proposto produzem uma coleção de relações documento-candidato mais confiáveis do que processos de reconciliação conhecidos na literatura. Além disso, seria necessário uma coleção de teste para validação da qualidade das reconciliações de documento e, não sendo esse o objetivo principal da dissertação, não haveria justificativa suficiente para o desenvolvimento do processo de reconciliação para os documentos extraídos da plataforma Lattes.

Para identificação de instâncias de documentos repetidos em diferentes currículos, foi desenvolvido um processo que considera casar documentos do mesmo tipo de publicação, no mesmo ano e com títulos similares. A similaridade dos títulos se dá pela simplificação do conteúdo textual, por exemplo, removendo caracteres especiais, múltiplos espaços seguidos e convertendo todo título para minúsculo, para então identificarmos documentos equivalentes pelo casamento exato. Dado o processo, foram encontradas cerca de 21 milhões de associações documento-candidato únicas (69% do número total das menções explícitas com identificador do Lattes e 36% de todas as menções).

Em uma etapa posterior à coleta, filtragem e construção das associações documento-candidato das publicações da plataforma Lattes, foi aplicado um processo de enriquecimento dos conteúdos dos documentos por meio da agregação dos resumos (*abstracts*) das publicações. Esse processo foi divido em duas etapas: (1) coleta dos *abstracts* diretamente das páginas das publicações cuja *Uniform Resource Locator* (URL) se baseia na *Application programming interface* (API) do *Digital Object Iden-*

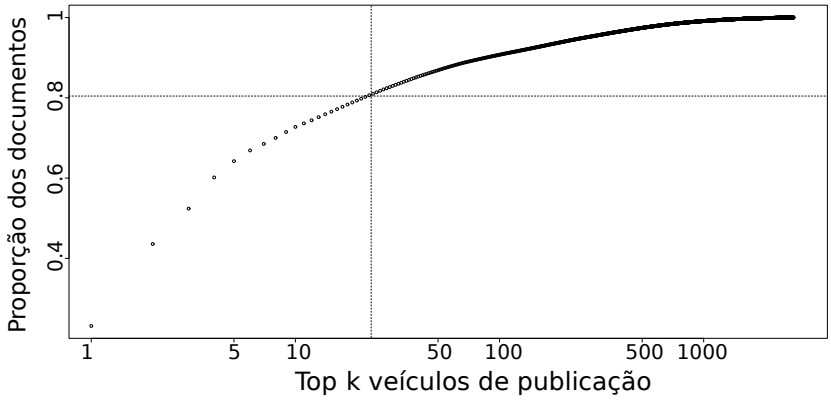

Figura 4.1: Número de documentos dos k veículos de publicação mais frequentes.

tifier (DOI) das publicações, e (2) coleta dos *abstracts* não recuperados pela primeira etapa através de consultas do DOI na API do Mendeley[2].

A primeira etapa do processo descrito anteriormente considera extrair os *abstracts* das páginas das publicações através do redirecionamento da URL do DOI. Dados os 672.893 DOIs extraídos dos documentos da plataforma Lattes, decidimos construir extratores para os veículos de publicação que cobrissem pelo menos 80% dos documentos com DOI. A Figura 4.1 apresenta a curva acumulada da quantidade de documentos dos veículos de publicação. Esse gráfico encontra-se em escala logarítmica e as retas tracejadas demonstram a posição dos top-22 veículos que cobrem 80% da coleção de documentos com DOI.

Assim, foram construídos coletores para cada veículo de publicação dos 22 domínios mais frequentes dos DOI dos documentos. Esse extratores conseguiram extrair 413.356 (61%) de toda coleção de documentos com DOI, sendo que isso equivale a, aproximadamente, 75% dos documentos dos 22 veículos mais frequentes. Dos 259.537 DOIs restantes, a segunda etapa do processo de enriquecimento dos documentos, que é baseada em consultas na API do Mendeley, conseguiu recuperar 69.866 *abstracts* (27% dos 259.537 DOIs), resultando em 483.222 (72%) dos documentos com DOI enriquecidos com *abstract*.

Entre as características dos currículos da plataforma Lattes, o fato de incluir múltiplos idiomas e múltiplas áreas é um diferencial entre a maioria das coleções de documentos acadêmicos conforme descritas na Seção 2.3. Para caracterizar a distribuição de idiomas na coleção, usamos um algoritmo de identificação de idioma em conteúdo textual[3] para reconhecer qual o idioma de cada documento extraído do Lattes consi-

[2] `http://dev.mendeley.com/getting_started/hello_mendeley.html`
[3] `https://pypi.python.org/pypi/langdetect`

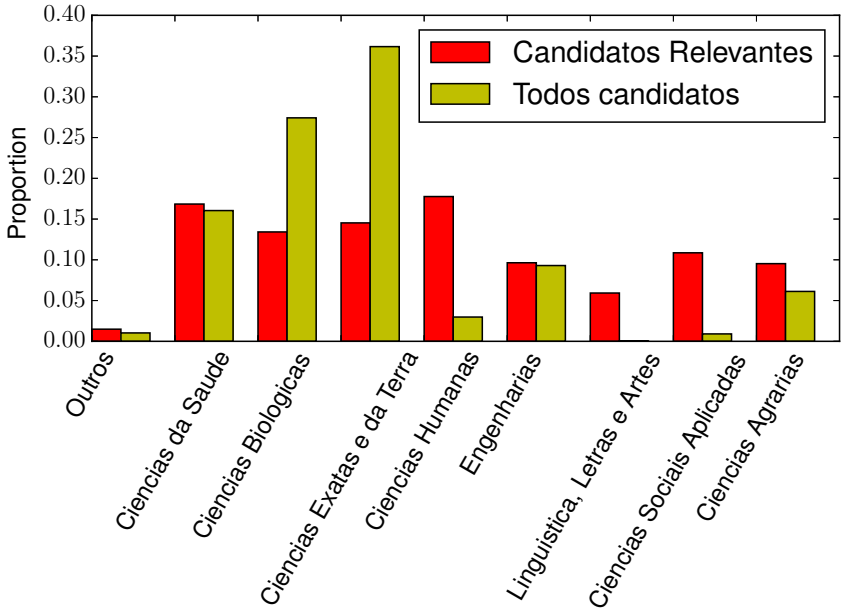

Figura 4.2: Candidatos especialistas por grande área de interesse.

derando o título das publicações. No resultado obtido, cerca de 61% dos documentos estão em português e dos 39% restantes, 56,5% estão em inglês, 22% em espanhol, 7,2% em italiano, 6,1% alemão e 7,9% em outros 26 idiomas.

Além disso, em relação às grandes áreas que abrangem os currículos do Lattes, a Figura 4.2 apresenta a distribuição das áreas dos currículos extraídos em relação a todos os doutores (barra verde) e as grandes áreas dos currículos dos candidatos a especialistas no gabarito (barra vermelha), sendo que o processo para construção do gabarito é apresentado na Seção 4.1.2.

Fixando o conjunto de documentos a usar na coleção de teste de busca de especialistas, analisamos também a distribuição dos tamanhos dos perfis dos candidatos, em número de *tokens* e número de documentos, conforme apresentado na Figura 4.3.

Como era esperado, os tamanhos dos perfis, em *tokens* e em documentos, seguem a distribuição de cauda longa, onde uma grande quantidade de perfis têm tamanhos menores e menos frequentes os perfis com muitos documentos ou *tokens*.

4.1.2 Construção do Gabarito das Consultas

Para uma coleção de teste prover uma avaliação realista, consultas devem ser selecionadas de forma que representem a necessidade real dos usuários [Sanderson, 2010]. Apesar de não ter acesso direto aos usuários realizando a busca de especialistas em

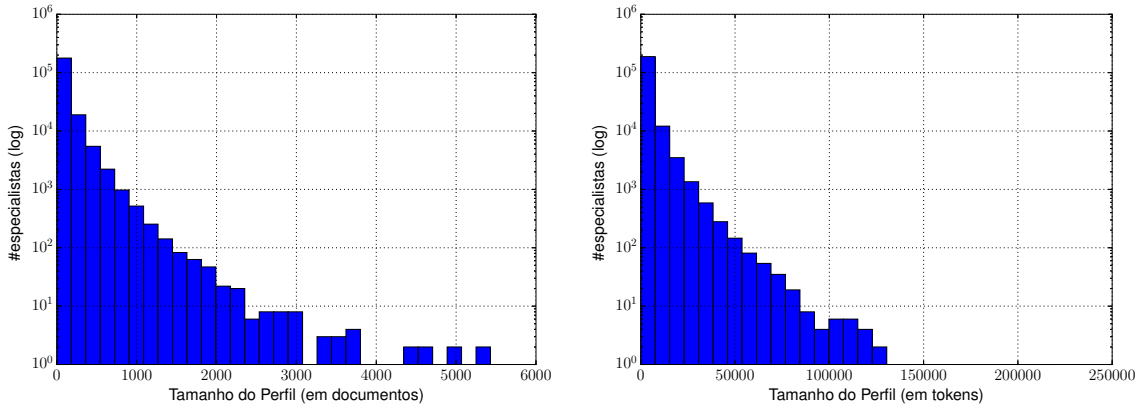

Figura 4.3: Distribuição do tamanho dos perfis.

nosso cenário, tivemos acesso ao perfil dos usuários especialistas, ou seja, os próprios especialistas. Com esse intuito, contactamos esses especialistas realização de julgamentos de relevância para as consultas através da aplicação de um questionário.

Consultas candidatas foram selecionadas do conjunto de rótulos recomendados pela abordagem de Ribeiro et al. [2015]. O problema de recomendação de rótulos, conhecido na literatura como *expert profiling*, busca sumarizar os conhecimentos relacionados a um determinado candidato a partir dos documentos associados a ele. A abordagem proposta por Ribeiro et al. [2015] recomendou rótulos de sumarização de especialidades dos mesmos candidatos usados na nossa abordagem de busca de especialistas. Dos 57.841 rótulos avaliados pelo questionário aplicado para construção do gabarito da solução proposta por eles (descrita também em [Mangaravite et al., 2016] e [Ribeiro et al., 2015]), selecionamos as consultas com pelo menos três candidatos que marcaram o rótulo como relevante ou fortemente relevante para sua expertise, resultando em 4.105 consultas.

O problema gerado pela conversão direta de rótulos de sumarização de expertise para consultas no processo de busca de especialistas pode ser ilustrada pela Figura 4.4. Idealmente, comparar julgamentos de relevância de graus diferentes no arcabouço de auto-avaliação pode resultar em julgamentos incoerentes. Nesse exemplo, o pesquisador “1” pode ser representado pelos rótulos R_1 e R_2, sendo suas relevâncias 3 e 1, isto é, altamente relevante e fracamente relevante, respectivamente. Enquanto isso, o pesquisador “2”, que pode ser um respondente mais modesto, é representado pelo rótulo R_1 apenas, respondendo com relevância 2, representando uma relevância moderada. Assim, na conversão, teremos o rótulo R_1 sendo uma consulta de especialidade para o candidato “1” e “2” e o rótulo R_2 apenas para o candidato “1”, resultado o ranking incoerente mencionado.

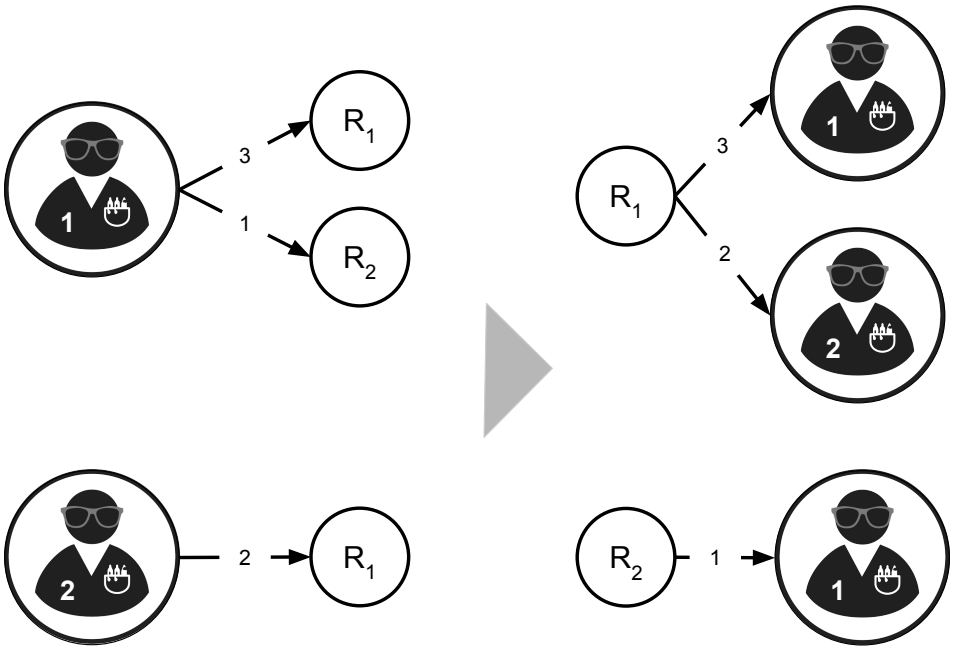

Figura 4.4: Exemplo de conversão de rótulos de sumarização de expertise para consulta de busca de especialistas.

Então, dado o conjunto das consultas a serem avaliadas, foi aplicado um novo questionário que envolveria 1.348 dos 5.356 pesquisadores associados a algum INCT. O objetivo desse questionário era avaliar a expertise dos candidatos a especialistas dada a opinião da comunidade sobre a especialidade dele. Cada pesquisador respondente recebeu para opinar, no seu questionário, para cada rótulo que marcou ser relevante ou fortemente relevante (valores 2 e 3, respectivamente), a lista dos candidatos a especialistas que também marcaram ser relevantes ou fortemente relevantes. As alternativas estavam entre 0 e 3, onde os valores significaram

0. **Indiscriminante**: O pesquisador que está respondendo o questionário disse não possuir elementos para mensurar o nível de especialidade do pesquisador avaliado, ou prefere não opinar sobre o grau de especialidade do candidato;

1. **Parcialmente relevante**: O pesquisador que está respondendo o questionário disse que o pesquisador avaliado possui **conhecimentos na área**;

2. **Relevante**: O pesquisador que está respondendo o questionário disse que o pesquisador avaliado **é um especialista na área**;

3. **Fortemente relevante**: O pesquisador que está respondendo o questionário disse que o pesquisador avaliado é a **principal referência na área**.

No Apêndice B apresentamos o conteúdo completo apresentado para cada respondente do questionário.

Dos 1.348 convidados a responder o questionário, 514 se prontificaram a participar (35% dos convidados), totalizando 7.210 respostas discriminantes (conjunto das respostas diferente de 0). Para cada consulta, foram escolhidos apenas os candidatos com pelo menos duas respostas discriminantes e apenas as consultas com pelo menos três candidatos a especialistas participantes desse conjunto, totalizando 235 consultas e 1635 relações consulta-especialista (qrel). Vale reforçar que as consultas estão em inglês e, que ao contrário da coleção inteira dos documentos do Lattes, onde 61% das publicações estão em português, as publicações dos pesquisadores do gabarito estão, majoritariamente em inglês (51%), seguido de português (37%), espanhol (5,4%) e outros 28 idiomas (6,1%).

A relevância atribuída para o candidato e que atender às exigências mencionadas para a consulta q, será a média das relevâncias que recebeu $\overline{R_{q,e}}$, sendo essa relevância final o arredondamento para o valor inteiro mais próximo:

$$\overline{R_{q,e}} = round\Big(\frac{1}{N}\sum_{i=1}^{N} R_{q,e,i}\Big) \tag{4.1}$$

onde $R_{q,e,i}$ é a relevância atribuída pelo i-ésimo respondente que avaliou o candidato e para a consulta q e N é a quantidade de respondentes que avaliaram o candidato e para a consulta q.

A distribuição dos níveis de especialidade dos candidatos nas 235 consultas é:

- 20% dos candidatos têm nível 1, ou fracamente relevante, sendo uma média de 1,9 candidatos por consulta;
- 34% têm nível 2, ou relevante, sendo uma média de 3,25 candidatos por consulta;
- 46% têm nível 3, ou fortemente relevante, sendo uma média de 4,41 candidatos por consulta.

Uma outra maneira de caracterizar a coleção das consultas é demonstrando o grau de "especificidade" que cada consulta tem em relação aos candidatos especialistas encontrados no gabarito. Definimos a especificidade de uma consulta como sendo a quantidade de grandes áreas associadas a todos os candidatos a especialistas no gabarito da consulta. Dessa maneira, uma consulta com muitas grandes áreas é uma consulta dita mais "genérica" por estar associada a temas ambíguos no seu sentido ou semântica, por exemplo, assim como demonstrado na Tabela 4.2, a consulta "amazon"

pode estar associada a diferentes contextos, podendo ser uma consulta associada a tratamento médico (Ciências da Saúde), tratamento de dados (Ciência Exatas e da Terra), tratamento em produção agrícola (Ciências Agrárias), etc. Alternativamente, consultas com poucas grandes áreas são consultas mais "específicas", demonstrando uma granularidade mais fina ou mais específica do contexto da consulta, por exemplo, "search engines" dificilmente estaria associada com outra área senão ciência da computação (Ciências Exatas e da Terra). Nesse contexto, a Figura 4.5 demonstra a distribuição das consultas em termos de especificidade. A título de ilustração da medida de especificidade das consultas, apresentamos algumas consultas e seus respectivos valores de especificidade na Tabela 4.2.

Para quantificar a concordância entre os respondentes para os candidatos a especialistas, avaliamos o erro médio absoluto[4] (Equação 4.2), estipulando como o valor esperado (ou valor correto) $\overline{R_{q,e}}$ (Equação 4.1) e cada relevância $R_{q,e,i}$ como a variável a ser avaliada. Nessa métrica, quanto menor o valor resultante, maior a concordância entre os respondentes. Assim, definimos o $MAE(q, e)$ como:

$$MAE(q, e) = \frac{1}{N} \sum_{i=1}^{N} |R_{q,e,i} - \overline{R_{q,e}}| \tag{4.2}$$

Valores de concordância por consultas foram obtidas calculando-se a média dos valores do MAE obtidos para todos os especialistas associados à consulta. A Figura 4.6 mostra a distribuição de concordância sobre todas as consultas, bem como sobre subconjuntos de consultas com diferentes níveis de especificidade.

[4]Do inglês, *Mean Absolute Error (MAE)*

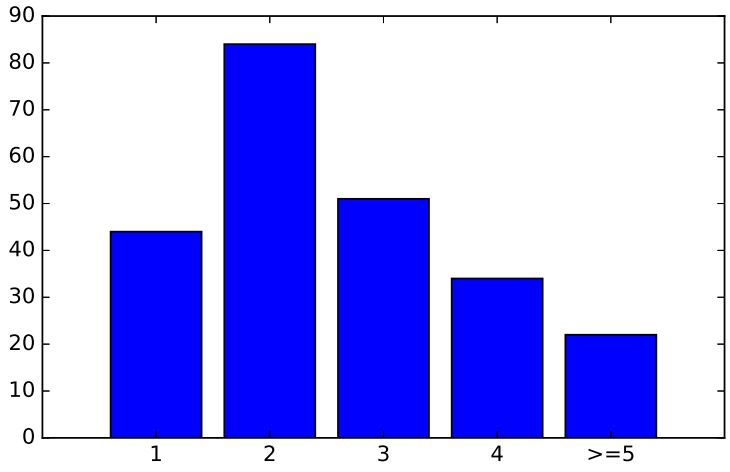

Figura 4.5: Distribuição da especificidade sobre as 235 consultas do gabarito.

Tabela 4.2: Exemplo de consultas e suas especificidades.

Especificidade	**Consultas**
>= 5	water quality brazil climate change amazon risk factors pregnancy treatment ...
...	
2	polymer chemical synthesis breast cancer rheumatic heart disease neurotoxicity somatic embryogenesis ...
1	development and validation quantum information and quantum mechanics condensed matter: structural, mechanical & thermal search engines circumstellar matter neurotoxin ...

Visualmente existe uma correlação positiva entre o MAE e a especificidade e, para demonstrar isso, analisamos a correlação entre os 50% dos dados mais próximos da mediana, sendo eles entre o primeiro quartil (Q1) e o terceiro quartil (Q3), de cada nível de especificidade e seus respectivos níveis de concordância. Usando Spearman como métrica, obtivemos 0,66 de correlação, sendo que esse valor é estatisticamente válido com p-value $< 0{,}01$. Usando apenas a média como entrada, a correlação torna-se ainda maior, chegando a 0,89 com p-value $< 0{,}05$. Com esses números, demonstramos que, quanto mais específica for a consulta, maior é a concordância entre os respondentes dela.

Assim, a coleção de teste construída nesta dissertação é a única coleção de teste de busca de especialistas, dentre as coleções estudadas, que é, ao mesmo tempo, multi-organização, multi-área e possui a avaliação feita pelos próprios especialistas. Além disso, é a segunda maior coleção em termos de número de documentos e candidatos, perdendo apenas para a ArnetMiner [Tang et al., 2008b] que possui apenas 13 consultas gabaritadas. Como resultado, acreditamos tratar-se de uma contribuição relevante para a comunidade de pesquisa em busca de especialistas.

4.2 Configurações Iniciais e *Baselines*

Dado o objetivo predominante da dissertação de demonstrar o funcionamento das abordagens de construção das associações propostas, não foi realizada nenhuma análise experimental da sensitividade dos parâmetros dos modelos de ranking de especialistas básicos descritos na Seção 2.1. Nesta seção, são apresentados e descritos, sucintamente, aspectos relevantes para a compreensão dos resultados em termos de parâmetros iniciais e seu procedimento de treino, os principais *baselines* para cada abordagem de ranking de especialistas estudada.

De fato, algumas das abordagens de ranking de especialistas são extremamente sensíveis aos seus hiper-parâmetros e funções de suavização escolhidas. Nessas condições, foi realizado, a priori, um estudo analisando algumas das melhores configurações experimentais para o funcionamento otimizado dos modelos de ranking de especialistas.

Toda análise da configuração inicial ideal foi mensurada considerando o modelo de ranking de especialistas tradicional onde as associações têm pesos uniformes, ou, em outras palavras, a função de associação constante e a normalização centrada em documento. Todo esse arcabouço tem por finalidade encontrar o *baseline* ideal onde as configurações não prejudiquem a comparação com as funções propostas.

As duas funções de suavização experimentadas, denominadas Jelinek-Mercer e Dirichlet, possuem os hiper-parâmetros α e β a serem treinados, respectivamente. A Equação 4.3 apresenta a formulação da suavização de Jelinek-Mercer aplicada na ve-

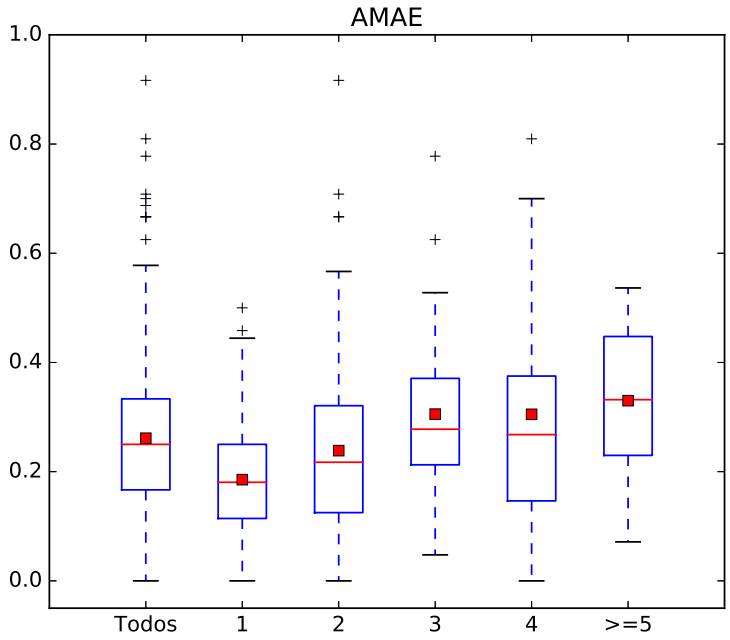

Figura 4.6: Mean Absolute Error (MAE) por grupos de especificidade.

rossimilhança de um termo da consulta para o modelo linguístico de um documento. Assim:

$$P(t|\theta_d) = (1 - \alpha)P(t|d) + \alpha P(t), \tag{4.3}$$

onde θ_d é o modelo linguístico formado para o documento d, $P(t|d)$ é a probabilidade do termo t ocorrer no documento d (apresentada na Equação 2.6) e $P(t)$ é probabilidade do termo t na coleção de documentos, como descrito na Seção 2.1.2.1.

A função de suavização Dirichlet é dada pela atribuição do hiper-parâmetro α como:

$$\alpha = \frac{\beta}{\beta + n(d)}, \tag{4.4}$$

onde $n(d)$ é a quantidade de termos encontrados no documento d e, como mencionado, β é o hiper-parâmetro da função de suavização de Dirichlet.

Assim, foram treinadas todas as possíveis combinações das seguintes configurações: (1) todas as combinações das quatro informações contidas nos documentos da coleção, sendo elas, título, áreas da publicação, palavra-chaves e resumo; (2) as funções de suavização de Dirichlet, variando o parâmetro β de 0 até 10000 de 1000 em 1000, ou a função de suavização de Jelinek-Mercer, variando o parâmetro λ de 0 até 1 de 0,1 em 0,1; e (3) seis diferentes quantidades de documentos a serem retornados pela função de ranking, sendo elas 10, 50, 100, 500, 1000 ou 2000.

Dessa maneira, foram avaliadas 1980 instâncias diferentes de configurações do modelo de ranking, para, então, fixar a melhor configuração considerando o Modelo 2 proposto por Balog et al. [2006] e usando a métrica nDCG@100 como critério de seleção. A configuração final selecionou a suavização de Dirichlet com parâmetro 2000, retornando 1000 documentos e a combinação dos atributos título, palavra-chaves e resumo como melhor configuração considerando essa base dados. Assim, deste ponto em diante da dissertação, não serão mais avaliadas nenhuma dessas condições do modelo de ranking estudado, sendo essa configuração usada, inclusive, na avaliação das funções de associação e normalização propostas.

Dada, então, a configuração inicial fixa, o arcabouço intuitivo para avaliação das funções de associação e normalização propostas é o Modelo 2 apresentado em Balog et al. [2006]. Em particular, usamos o Modelo 2 como base para testar as funções de associação e normalização propostas em comparação àquelas apresentadas na literatura e discutidas na Seção 2.2.

Assim, outro *baseline* usado na comparação com as abordagens de ponderação de

associação é a função de associação proposta por Macdonald et al. [2008]. Essa função determina a proximidade do conteúdo dos documentos com o interesse central dos candidatos associados através de uma função de agrupamento. Para replicar o resultado dos autores, foi usada a ferramenta de agrupamento denominada gmeans[5] que implementa o algoritmo *k-means*. Para determinação das distâncias entre os documentos, foi usada a função de cosseno na representação TF-IDF clássica [Baeza-Yates & Ribeiro-Neto, 2011] sobre o título das publicações dos candidatos. O parâmetro do número de grupos K foi replicado da abordagem de Macdonald et al. [2008], sendo fixado com valor 10. Além disso, no artigo os autores determinam que os candidatos só terão seus grupos construídos se, e somente se, forem associados com pelo menos 30 documentos. Na nossa replicação, esse número foi reduzido para 10 por serem os candidatos em nossa coleção menos prolíficos do que candidatos dos ambientes corporativos.

Um terceiro *baseline* levado em consideração para validação das abordagens de função de associação propostas foi a função de associação denominada *Semantic-Relatedness*, apresentada em Balog & De Rijke [2008]. Para treinar essa abordagem, foram experimentadas as mesmas configurações usadas no treino da etapa de ranking de documentos. O melhor resultado foi usando a função de suavização Jelinek-Mercer, com parâmetro $\lambda = 0,1$.

Como *baseline* também, foi implementada e experimentada a função de normalização Norm2 proposta por Macdonald & Ounis [2011]. Para replicação, serão usadas as mesmas funções de associação que os autores propuseram no trabalho: quantidade de termos na representação do candidato e quantidade de documentos a que o candidato está associado. Assim, o único parâmetro a se treinar nessa função de normalização é quais informações serão consideradas na contagem do número de *tokens* para cada candidato. Dessa forma, foi experimentado e avaliado o mesmo conjunto de combinações das cinco informações contidas nos documentos usadas para o treino da função de ranking de documentos, contudo para a construção da representação dos candidatos.

A fim de investigar a complementaridade das funções de associação e normalização propostas, decidimos investigar seu uso conjunto como entrada para modelos discriminativos de busca de especialistas, conforme descrito na Seção 2.1.2.2.

Quando tratamos o problema de ranking de especialistas com soluções baseadas em modelos discriminativos, convertemos o problema de ranking de especialistas em duas diferentes abordagens discriminativas, denominadas, aqui, como: (1) Aprendizagem de agregação de rankings de especialistas (abordagem introduzida por Macdonald & Ounis [2011]), e (2) Aprendizagem de associação de candidatos-documento (aborda-

[5] `http://www.cs.utexas.edu/users/dml/Software/gmeans.html`

gem introduzida por Fang et al. [2010b]).

Quando é proposta uma solução para o problema de ranking de especialistas usando a primeira abordagem discriminativa citada, a abordagem de aprendizagem de agregação de rankings de especialistas, podemos dizer que convertemos o problema de ranking de especialistas usando soluções de engenharia de atributos para modelos de aprendizagem de ranking (L2R[6]). Nesse caso, então, existem aspectos que foram levados em consideração para determinação dos modelos de aprendizagem de rankings usados e do procedimento de escolha de cada um dos seus respectivos valores para os parâmetro.

Assim, primeiramente foram determinados quais seriam os modelos de aprendizagem de ranking usados. A princípio, escolhemos o AdaRank [Xu & Li, 2007] como um possível candidato por estar na lista dos métodos usados pelos autores do artigo Macdonald & Ounis [2011]. Então, em seguida, foram escolhidos os métodos LambdaMART [Burges, 2010], MART [Friedman, 2001] e RandomForest [Breiman, 2001]. A implementação usada dos métodos é baseada no conjunto ferramental de L2R denominado RankLib[7].

Cada modelo de L2R tem seu próprio conjunto de parâmetros a serem treinados. Assim, evitando experimentar todas as combinações das faixas de valores para cada parâmetro, escolhemos o melhor parâmetro de cada modelo de L2R iterativamente, fixando o melhor valor para os parâmetros já avaliados. Como esse procedimento é sensível à ordem escolhida para cada modelo de L2R, foram escolhidos os parâmetros que são, reconhecidamente, mais sensíveis para cada modelo.

Assim, determinamos a ordem e a faixa de valores de cada parâmetro a ser experimentado como apresentado na Tabela 4.3, lembrando que esse procedimento de seleção dos valores dos parâmetros é sensível a ordem. Como critério de seleção dos valores dos parâmetros, foi usado o nDCG_{100} e os melhores valores dos parâmetros de cada um dos modelos de L2R foram selecionados individualmente para cada *fold* de treino.

Na próxima Seção é apresentado o procedimento de particionamento dos dados usados e o procedimento de treino e teste das funções de associação e normalização para os modelos discriminativos e generativos.

[6]Do inglês, *Learning to rank.*

[7]`https://people.cs.umass.edu/~vdang/ranklib.html`.

4.3 Procedimentos de Treino e Teste

O procedimento de treino e teste utilizado conhecido na literatura como validação cruzada, busca aproximar o resultado obtido para as métricas de avaliação numa simulação do ambiente real do problema. Amplamente usada para avaliação de modelos de predição em diferentes problemas de Recuperação de Informação e Aprendizagem de Máquina, essa técnica tem como conceito central dividir o conjunto de dados em dois subconjuntos: dados de treino e dados para teste.

Existem diferentes métodos para a divisão dos subconjuntos de dados. Nesta dissertação, o método escolhido foi o *k-fold*. Esse método divide o conjunto de dados, no caso, conjunto de consultas, em k partes iguais, usando uma parte para teste dos resultados e as $k-1$ partes restantes para treino, estimação dos parâmetros e validação dos modelos. Esse processo é repetido k vezes, sempre mudando qual a parte será usada para o teste dos resultados.

Tabela 4.3: Ordem dos parâmetros e faixa de valores escolhidos para procedimento de treino dos modelos de L2R.

Modelo de L2R	**Parâmetro**	**Faixa de valores**
MART & LambdaMART	Taxa de aprendizagem	[0,0001; 0,001; 0,01; 0,1; 0,25; 0,5]
	Mínimo suporte por folha	[1, 5, 10, 20, 30, 40, 50]
	Número de folhas em cada árvore	[1, 2, 5, 10, 25, 50]
	Número de árvores	[1, 2, 5, 10, 25, 50]
RandomForest	Taxa de amostragem de atributos	[0,2; 0,3; 0,4; 0,5]
	Taxa de subamostragem	[0,5; 0,75; 1]
	Taxa de aprendizagem	[0,0001; 0,001; 0,01; 0,1; 0,25; 0,5]
	Mínimo suporte por folha	[1, 5, 10, 25]
	Número de bags	[1, 10, 50, 100, 150, 200, 250, 300, 350]
	Número de árvores	[1, 2, 5, 10]
	Número de folhas em cada árvore	[1, 5, 10, 25, 50, 100]
AdaRank	Tolerância entre dois rounds consecutivas de aprendizagem	[0,001; 0,002; 0,01; 0,05; 0,1; 0,25; 0,5]
	Número de rounds para treinar	[1, 5, 10, 50, 100, 200, 300, 400, 500, 600]
	O número máximo de vezes que um atributo pode ser consecutivamente selecionado sem alterar o resultado	[1, 2, 5, 7, 10, 20, 50]

Nesta dissertação, usamos o valor de $k = 10$ e, para o resultado final do processo inteiro das k execuções, usamos a média aritmética dos *folds* de teste. Além disso, foram usadas as mesmas divisões dos *folds* para a avaliação dos modelos generativos e discriminativos.

Capítulo 5

Avaliação Experimental

Neste capítulo, apresentamos os resultados experimentais das funções de associação e normalização propostas nesta dissertação. Como mencionado, o objetivo central do processo de avaliação é demonstrar a eficácia de um procedimento de ponderação de relações documento-candidato dividida em duas etapas: funções de associação e funções de normalização. Para isso, avaliamos, individualmente, a qualidade do resultado do ranking de especialistas em relação a cada etapa do processo, comparando os resultados obtidos pelo modelo de ranking generativo proposto por Balog et al. [2006] e os modelos de ranking discriminativos proposto por Macdonald et al. [2008] com as funções de associação e normalização conhecidas na literatura.

Para formalizar os principais aspectos a se considerar para a avaliação das funções propostas, segmentamos os objetivos em três questões de pesquisa a serem respondidas. Tais questões são apresentadas na próxima seção e são respondidas no decorrer deste capítulo.

5.1 Questões de Pesquisa

Assim, retomamos as perguntas de pesquisa apresentadas na Seção 1.2. Tais perguntas são fundamentais para uma avaliação estruturada dos resultados demonstrados neste capítulo.

Desse modo, as principais questões de pesquisa são:

Q1. Quão eficazes podem ser as **funções de associação** propostas na geração de rankings de especialistas?

Q2. As **funções de normalização** propostas geram rankings melhores em comparação com as funções de normalização propostas na literatura?

Q3. Quão **complementares** são as diferentes combinações de funções de associação e de normalização no processo ranking de especialistas?

As questões Q1 e Q2 estão relacionadas, diretamente, à avaliação das funções de associação e normalização aplicadas ao modelo generativo de ranking de especialistas proposto por Balog et al. [2006] (Seção 5.2). A questão Q3, avalia uma forma de combinação das funções de associação e normalização através de modelos discriminativos para ranking de especialistas (Seção 5.3).

5.2 Resultados Experimentais: Modelos Generativos

Como mencionado, as duas primeiras questões de pesquisa estão associadas diretamente com a avaliação das funções de associação e normalização nos modelos generativos de ranking de especialistas. Essas questões são abordadas no restante desta seção.

5.2.1 Funções de Associação

Para avaliarmos a eficácia das funções de associação propostas, consideramos um processo em duas etapas para responder a pergunta de pesquisa Q1, onde, primeiramente, estudamos a parametrização das funções de associação propostas, para, então, demonstrarmos a eficácia da abordagem considerando funções não-booleanas de associação.

Apresentamos o estudo e os resultados da primeira etapa na seção seguinte e os resultados experimentais relacionados as funções de associação na Seção 5.2.1.2.

5.2.1.1 Parametrização

Para analisar a melhor configuração para avaliação das funções de associação, descrevemos sucintamente cada etapa do processo de avaliação e experimentação do modelo generativo, começando pelo estudo dos parâmetros usados pela função de associação denominada *dominância de conteúdo* (Seção 3.1.1). A escolha dessa função de associação para o treino dos parâmetros está relacionada ao grau de dependência desses valores para a avaliação do restante das funções de associação, considerando que todas as funções de associação que usam o conteúdo formalizam seus pesos também através da entropia cruzada. Assim como os modelos generativos de ranking de especialistas, a função de associação *dominância de conteúdo* é sensível ao parâmetro de sua função de suavização, além de ser dependente da combinação dos campos do documento que serão usados na função de associação.

Portanto, como a maioria das funções de associação são dependentes dos mesmos parâmetros e configurações da função de associação *dominância de conteúdo*, foi realizado um estudo da sensitividade na escolha dessas configurações visando encontrar a combinação dos campos que obtém o melhor ranking dado um conjunto de possíveis valores para o parâmetro β da suavização de Dirichlet a serem experimentados.

O critério de comparação entre as combinações foi a média aritmética do nDCG_{100} da função de ranking generativo com a função de associação *dominância de conteúdo* aplicando duas diferentes funções de normalização: centrada em documento (Equação 2.23) e a função proposta também centrada em documento (Equação 3.22). Foram usadas todas as combinações do parâmetro β, variando nos valores do conjunto {0,1; 0,5; 1; 10; 100; 1000; 2000; 5000; 10000} e as combinações de campos para representação do documento, incluindo as informações contidas em nossa coleção de teste (Seção 4.1, sendo elas título, resumo e palavras-chave) e a representação *lean* do documento da mesma forma que proposta no trabalho de Balog & De Rijke [2008].

Dado o resultado que não tem diferença significativa entre os parâmetros, escolhemos os parâmetros com um critério empírico onde: (1) a maior número de campos dos documentos são selecionados; e (2) o parâmetro β com valor equivalente ao definido na função de similaridade da consulta com o conteúdo do documento. Assim, ao final usamos, para determinação da dominância do candidato perante o documento, todos os campos do documento e o parâmetro $\beta = 2000$ para o restante dos experimentos desta dissertação.

Dessa forma, apesar de não ser possível estimar qual a melhor combinação de campos para representação da dominância de um candidato perante um documento, é possível apresentar alguns fatores que denotariam quais campos unitários são menos representativos. Numa tentativa de estimar tais campos, ranqueamos as combinações de campos pelos menores resultados do nDCG_{100} e analisamos quais campos são repetidamente selecionados nos top-k rankings.

Assim, nove entre os dez rankings com menor nDCG_{100} usam exclusivamente o campo palavras-chave para modelar a dominância do conteúdo do candidato, sendo que essas nove configurações que usam exclusivamente o campo palavras-chave incluem todas as variações do parâmetro β. Ou seja, existe uma forte tendência a se acreditar que, considerando o nDCG_{100}, para o ambiente experimentado e qualquer β avaliado, usar apenas as palavras-chave para modelar a dominância dos candidatos perante o conteúdo dos documento é uma forma ineficaz.

Com os parâmetros da função de associação *dominância de conteúdo* fixados, atribuímos as mesmas configurações experimentais para o restante das funções de associação que usam o conceito de dominância introduzido aqui, sendo elas: *novidade no*

domínio e *estabilidade na dominância*.

5.2.1.2 Análise Comparativa

Partindo então para segunda etapa da experimentação das funções de associação nos modelos generativos, buscamos comprovar a significância entre a diferença dos rankings, usando o *t-Student test* pareado [Smucker et al., 2007]. Como mencionado, foram escolhidas três métricas de avaliação de ranking: *Mean Reciprocal Rank* (MRR), *Normalized Discounted Cumulative Gain* dos top-10 candidatos no ranking ($nDCG_{10}$) e precisão dos top-10 candidatos ranqueados.

Os resultados são apresentados na Tabela 5.1. Nesta fase da avaliação, comparamos as funções de associação propostas usando as normalizações centradas em documento (DC) e centradas em candidato (CC) propostas por Balog et al. [2006] com a função de associação também proposta por eles, denominada aqui, de booleana. Como *baselines* adicionais, as funções de associação *cluster-based* (CL), *semantic-relatedness* (SR) e Norm2 (de termo e de documento), descritas na Seção 2.2.1, são comparadas com a função booleana de associação usando a normalização DC. Os símbolos $\blacktriangle$ e $\blacktriangledown$ sobrescritos representam ganho/perda com relação à função de associação booleana com *p-value* $< 0,01$, respectivamente, enquanto os símbolos $\triangle$ e $\triangledown$ representam ganho/perda com *p-value* $< 0,05$, respectivamente. Além disso, os valores em negrito representam os maiores valores para cada métrica em relação ao modelo de ranking generativo, a função de associação e a função de normalização.

Assim, como é apresentado na Tabela 5.1, considerando a normalização DC, demonstramos que a função de associação *dominância de conteúdo* resulta em um ranking de especialistas significativamente superior nas três métricas de avaliação experimentadas, chegando a ganhos de 7,6% na métrica P_{10}. Enquanto isso, a função de associação *novidade no domínio* apresentou o pior resultado entre as abordagens propostas, sendo, inclusive, significativamente inferior em 9,7% com *p-value* $< 0,01$. Com relação às funções de associação CL, SR e Norm2, consideradas como *baselines* não-booleanos, observamos que a função de associação *dominância de conteúdo* proposta foi superior em relação a quase todos os *baselines*, obtendo apenas um empate estatístico com a abordagem SR no caso em que usamos a normalização DC.

Em contrapartida, considerando a função de normalização CC, a função de associação *dominância de conteúdo* apresentou um ranking significativamente inferior em relação à função booleana em termos de MRR. Porém, as funções de normalização CC apresentaram os piores rankings gerais em relação às métricas de avaliação. Uma justificativa para isso é o grau de dependência que essa função de normalização tem

Tabela 5.1: Resultados dos rankings dos *baselines* e das funções de associação propostas usando o modelo generativo e as normalizações tradicionais.

	$\mathbf{nDCG_{10}}$	$\mathbf{P_{10}}$	**MRR**	$\mathbf{nDCG_{10}}$	$\mathbf{P_{10}}$	**MRR**
CL	$0.022^{\blacktriangledown}$	$0.014^{\blacktriangledown}$	$0.06^{\blacktriangledown}$	-	-	-
SR	0.135	0.082	0.240	-	-	-
	Term			**Document**		
Norm2	$0.015^{\blacktriangledown}$	$0.01^{\blacktriangledown}$	$0.039^{\blacktriangledown}$	$0.009^{\blacktriangledown}$	$0.008^{\blacktriangledown}$	$0.028^{\blacktriangledown}$
	DC			**CC**		
Booleana	0.133	0.079	0.254	0.009	**0.008**	**0.028**
Dominância	$\mathbf{0.139}^{\blacktriangle}$	$\mathbf{0.085}^{\blacktriangle}$	$\mathbf{0.263}^{\blacktriangle}$	0.009	0.007	$0.025^{\blacktriangledown}$
Novidade	0.133	0.079	0.254	0.008	0.005	0.024
Estabilidade	$0.12^{\blacktriangledown}$	$0.071^{\blacktriangledown}$	$0.234^{\triangledown}$	0.008	0.007	0.027
Recência Lin.	0.133	0.082	0.245	0.009	0.006	0.026
Recência Exp.	0.132	0.079	0.254	**0.01**	**0.008**	**0.028**

em relação ao número de documentos associados a cada candidato, onde candidatos com muitos documentos tendem a ser prejudicados pelo número de publicações. Isso demonstra a fragilidade da função de normalização CC para a geração de rankings de especialistas no ambiente acadêmico, onde candidatos prolixos, ou seja, candidatos com muitos documentos, não necessariamente são candidatos menos propícios a ser especialistas.

Outra observação a se fazer está relacionada a função de associação *novidade no domínio* com a normalização DC. Como os documentos têm o mesmo peso de associação para os candidatos, essa função de associação na normalização DC produz um ranking exatamente igual à função de associação booleana. Isso não se repete em nenhuma outra função de associação em qualquer normalização e é uma peculiaridade da normalização DC onde associações uniformemente ponderadas, resultam nos mesmos rankings que a função booleana.

Com exceção das funções *novidade no domínio* e *estabilidade na dominância* com a normalização DC, o resultado apresentado na Tabela 5.1 demonstra uma proximidade muito grande entre os resultados das funções de associação propostas e os resultados da função booleana na maioria das normalizações experimentadas aqui. Assim, numa segunda análise efetuada para demonstrar o potencial das funções de associação propostas, apresentamos o critério *win-loss* na Tabela 5.2, que denota o número de consultas melhoradas/pioradas (do universo de 235 consultas) em relação à função de associação booleana, considerada como *baseline*. Os valores em negrito apresentam as funções de associação propostas que são superiores à função booleana segundo a métrica avaliada.

A linha que apresenta o total do *win-loss* representa o número de consultas que

Tabela 5.2: Taxa de *win-loss* dos resultados dos rankings do modelo generativo.

	Win/Loss					
	DC			**CC**		
	nDCG$_{10}$	**P$_{10}$**	**MRR**	**nDCG$_{10}$**	**P$_{10}$**	**MRR**
Dominância	**56/25**	**18/17**	**105/52**	3/10	2/3	68/144
Novidade	-			7/12	5/10	82/145
Estabilidade	31/57	11/29	58/112	9/9	5/7	102/122
Recência Lin.	**40/29**	**13/8**	**73/67**	6/8	2/5	80/131
Recência Exp.	**12/10**	2/3	29/35	**6/2**	**2/1**	**87/77**
Total	**88/79**	35/36	143/161	**21/15**	13/14	196/219

ocorreram no conjunto que ganha ou perde no resultado de alguma das funções de associação propostas. Por exemplo, na coluna do nDCG$_{10}$ da normalização DC, 88 diferentes consultas apareceram no conjunto *win* de alguma das funções de associação propostas, enquanto 79 consultas apareceram no conjunto *loss*.

Nesse caso, nota-se uma superioridade em número de consultas no conjunto *win* em algumas das funções de associação propostas, sendo elas, para normalização DC, *recência lin.* e *dominância de conteúdo* e, para normalização CC, *recência exp.* Ainda é possível observar uma superioridade do número de consultas *win* para *recência exp.* na normalização DC, quando é avaliada usando nDCG$_{10}$. Além disso, podemos observar que esse caso possui uma grande quantidade de consultas empatadas na métricas P$_{10}$, devido à baixa quantidade de consultas nos conjuntos *win* (duas consultas) e *loss* (três consultas). Como a função *recência exp.* não possui mais consultas superiores aos valores da função booleana na métrica MRR, podemos deduzir que ela não é superior em selecionar o primeiro especialista para as consultas. Contudo, ela ordena melhor os top-10 candidatos retornados, mesmo retornando quase a mesma quantidade de candidatos especialistas, como demonstrado pelo número de empates em P$_{10}$ e por ter mais consultas com nDCG$_{10}$ maiores do que o *baseline*.

Outro ponto a ser notado na Tabela 5.2 é a grande quantidade de empates no número de consultas com o mesmo valor de nDCG$_{10}$ na normalização CC. Na média, são 14 consultas com valores diferentes para as cinco funções de associação, numa distribuição de 8 a 19 consultas diferentes. Isso representa apenas 6,1% de todas as consultas experimentadas. Contudo, mais para frente demonstraremos que o conjunto das consultas que as cinco métricas demonstraram ser superiores possui uma interseção menor do que o conjunto das consultas que as métricas perdem.

Contudo, analisando a Tabela 5.2, é possível deduzir que algumas funções de associação propostas demonstram potencial de modelar melhor a importância de um dado documento para um candidato. De fato, considerando a normalização DC, as

funções *recência lin.* e *recência exp.* apresentam resultado positivo em relação à taxa de *win-loss* para $nDCG_{10}$, com *recência lin.* mostrando resultados positivos também para P_{10} e MRR. Considerando a normalização CC, a função *recência exp.* se destaca sendo superior nas três métricas avaliadas.

Isso demonstra que, para algumas consultas, o documento ser recente para o candidato e para a coleção é uma evidência forte de expertise para o candidato e pode, sim, beneficiar o modelo de ranking de especialistas generativo. Assim, respondendo a pergunta Q1 relacionada às funções de associação, os resultados das métricas avaliadas demonstraram que a função *dominância de conteúdo* é, significativamente, superior à função booleana para normalização DC. Quanto a normalização CC, não é possível afirmar qual das funções propostas podem ser significativamente superiores à função booleana. Além disso, embora inconsistentes, as funções de associação baseadas em recência demonstraram potencial para ganhos adicionais.

5.2.2 Funções de Normalização

Demonstrado o potencial e, até mesmo o significativo ganho das funções de associação propostas, entraremos na discussão das funções de normalização propostas. Retomamos, então, a pergunta de pesquisa Q2 como sendo o foco principal desta seção da dissertação:

Assim como a pergunta de pesquisa Q1, separamos a pergunta de pesquisa Q2 em duas etapas, onde a primeira, que será apresentada na próxima seção, representa a fase de parametrização da função e a segunda, que será apresentada na Seção 5.2.2.2, a análise dos resultados obtidos pela função proposta.

5.2.2.1 Parametrização

Dadas as etapas a serem estudadas, a primeira etapa é responsável pela análise da sensitividade do hiper-parâmetro α das funções de normalização propostas. Esse é um aspecto crucial para o funcionamento das normalizações por depender de características da coleção. Em nosso caso, trata-se de uma coleção de teste voltada para a área acadêmica, portanto, intuitivamente espera-se que o peso de todas as associações dos candidatos ou documentos sejam mais relevantes do que uma associação individual.

Como o interesse dessa normalização é não penalizar documentos ou candidatos com muitas associações, o valor de α, que é o parâmetro da função exponencial, determina o impacto da suavização que será aplicada ao peso da associação. Por exemplo, candidatos prolixos serão beneficiados se, e somente se, tiverem muitas associações

fortes com seus documentos, enquanto candidatos com poucas associações, mas relativamente fortes, serão mais beneficiados pela normalização nesse caso.

Portanto, analisamos os valores do parâmetro α nas funções de normalização propostas (Equações 3.22 e 3.23) variando de 0,01 até 2, em intervalos de 0,01 em 0,01, resultando em 40 experimentos para cada função de normalização e apresentamos o resultado na Figura 5.1. Como métrica de avaliação usamos o nDCG_{100} por demonstrar uma maior sensibilidade em relação aos valores de α.

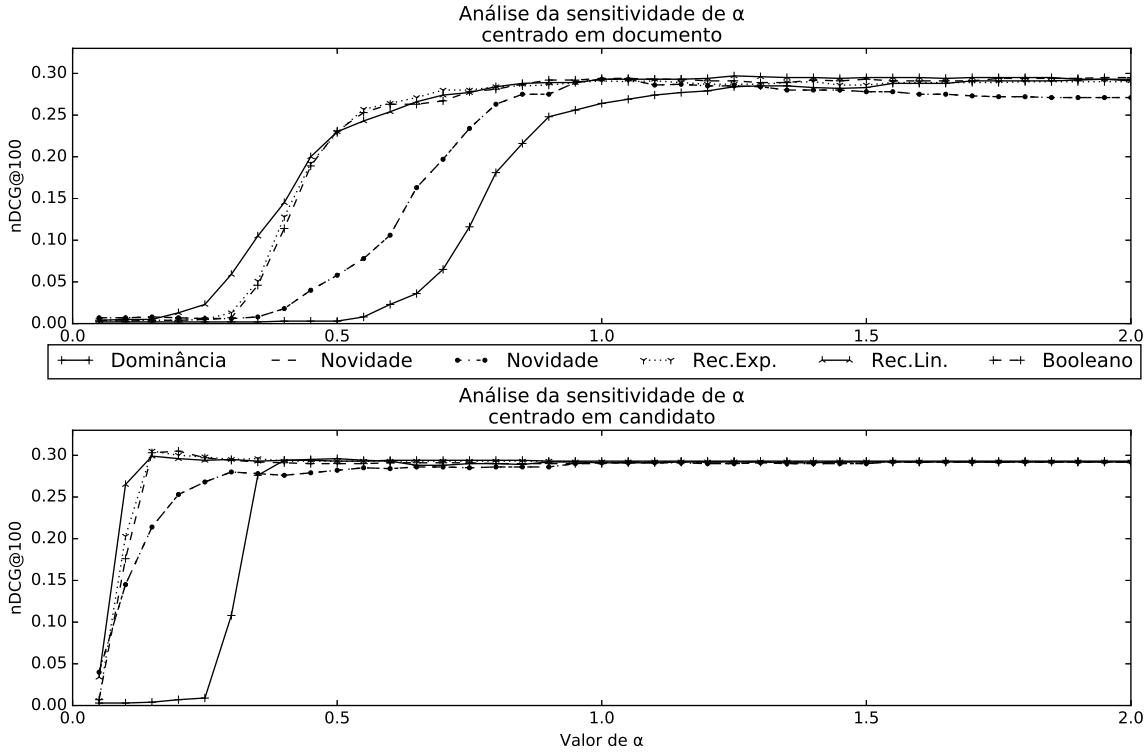

Figura 5.1: Análise da sensitividade do parâmetro α das funções de normalização propostas.

Na Figura 5.1, podemos notar uma constância nos resultados de nDCG_{100} a partir de diferentes valores de α para cada normalização centrada em documento ou candidato. Por exemplo, na normalização SCC, a constância começa com valores menores de α, demonstrando uma necessidade de maior suavização da normalização em termos do contexto para modelar o peso de uma associação documento-candidato. Estima-se que isso acontece devido à maior quantidade de associações ligadas aos candidatos em comparação com a quantidade de associações ligadas a documentos. Ou seja, quanto mais associações paralelas usadas na normalização da associação avaliada, menor é necessidade de valores altos para α na suavização do contexto para uma modelagem mais eficaz. Isso é uma característica peculiar do ambiente acadêmico, onde uma proporção

do número de publicações ou do número de autores pode ser usado na modelagem da relevância dessa publicação para a expertise de cada candidato.

Portanto, para a coleção de teste em questão, o parâmetro α pode ser estipulado de uma forma quase visual, tendo em vista que ele estabiliza, para a maioria das funções de associação, conforme que se aproxima do valor 2. Assim, podemos dizer que o hiper parâmetro α pode ser definido, empiricamente, com valor maior do que 1 para normalização SCC e maior do que 1.5 para normalização SDC para a maioria das funções de associação. Visando padronizar os experimentos, estipulamos o valor $\alpha = 2$ para todas funções de normalização.

5.2.2.2 Análise Comparativa

Com o parâmetro α fixado em 2, partimos para a avaliação da eficácia das funções de normalização. O principal objetivo, nesta fase da análise experimental, é comparar os resultados obtidos pelas funções de normalização propostas com as funções de normalização conhecidas, respondendo a pergunta de pesquisa relacionada às funções de normalização, Q2.

Serão três funções de normalização comparadas as funções propostas, sendo elas: (1) As funções de normalizações DC e CC propostas por Balog et al. [2006]; (2) a normalização Norm2 proposta por Macdonald et al. [2008]; e (3) uma função inócua (ID) baseada na função identidade $\psi(\bullet) \equiv \bullet$. Uma observação a se fazer em relação ao *baseline* Norm2 é que, na sua formulação tradicional, a média usada pela função (representada pelo valor $\overline{\rho(d,e)}$ na Equação 2.25) seria a média geral de todas as associações. Por questões de escalabilidade, adaptamos essa média para a média de todas as associações dos candidatos que possuem pelo menos um documento na lista de documentos retornados pela consulta. Espera-se que esse valor não esteja muito distante da média geral da coleção, já que são retornadas aproximadamente 338 mil associações diferentes por consulta (cerca de 1,6% do número de associações na coleção). Portanto, supomos que esse valor adaptado de média não altere significativamente a ordem final dos candidatos a especialistas.

Os resultados das funções de normalização são apresentados na Tabela 5.3. Os símbolos ▲ e ▼; e △ e ▽, usados para as comparações da tabela, têm representação análoga aos apresentados para Tabela 5.1. Contudo, é incluído ainda o símbolo ∘ para representar a ausência de significância estatística ($p\text{-}value \geq 0,05$). Os resultados dos *baselines* Norm2 e ID foram comparados com os resultados da normalização DC, enquanto a normalização proposta foi comparada com os três *baselines* onde os símbolos de significância representam, respectivamente, comparação com os *baselines*

DC, Norm2 e ID. A escolha de não comparar a normalização SCC com a normalização CC foi devida ao baixo desempenho na eficácia desse *baseline* e para simplificação da apresentação dos resultados na Tabela 5.3.

Tabela 5.3: Resultados dos rankings das funções de normalizações propostas usando o modelo generativo.

	$\mathbf{nDCG_{10}}$	$\mathbf{P_{10}}$	**MRR**	$\mathbf{nDCG_{10}}$	$\mathbf{P_{10}}$	**MRR**
	Norm2			**ID**		
Booleana	$0{,}160^{\blacktriangle}$	$0{,}097^{\blacktriangle}$	$0{,}286^{\blacktriangle}$	$0{,}160^{\blacktriangle}$	$0{,}097^{\blacktriangle}$	$0{,}286^{\blacktriangle}$
Dominância	$0{,}132^{\circ}$	$0{,}079^{\circ}$	$0{,}246^{\circ}$	$0{,}169^{\blacktriangle}$	$0{,}102^{\blacktriangle}$	$0{,}302^{\blacktriangle}$
Novidade	$0{,}003^{\blacktriangledown}$	$0{,}001^{\blacktriangledown}$	$0{,}016^{\blacktriangledown}$	$0{,}098^{\blacktriangledown}$	$0{,}056^{\blacktriangledown}$	$0{,}205^{\blacktriangledown}$
Estabilidade	$0{,}022^{\blacktriangledown}$	$0{,}017^{\blacktriangledown}$	$0{,}063^{\blacktriangledown}$	$0{,}122^{\circ}$	$0{,}070^{\circ}$	$0{,}259^{\circ}$
Recência Lin.	$0{,}154^{\blacktriangle}$	$0{,}086^{\circ}$	$0{,}282^{\blacktriangle}$	$0{,}161^{\blacktriangle}$	$0{,}100^{\blacktriangle}$	$0{,}281^{\blacktriangle}$
Recência Exp.	$0{,}147^{\triangle}$	$0{,}092^{\blacktriangle}$	$0{,}282^{\triangle}$	$0{,}162^{\blacktriangle}$	$0{,}097^{\blacktriangle}$	$0{,}290^{\blacktriangle}$
	DC			**CC**		
Booleana	0,133	0,079	0,254	0,009	0,008	0,028
Dominância	0,139	0,085	0,263	0,009	0,007	0,025
Novidade	0,133	0,079	0,254	0,008	0,005	0,024
Estabilidade	0,120	0,071	0,234	0,008	0,007	0,027
Recência Lin.	0,133	0,082	0,245	0,009	0,006	0,026
Recência Exp.	0,132	0,079	0,254	0,010	0,008	0,028
	SDC			**SCC**		
Booleana	$0{,}166^{\blacktriangle\circ\circ}$	$0{,}099^{\blacktriangle\circ\circ}$	$0{,}295^{\triangle\circ\circ}$	$0{,}163^{\blacktriangle\circ\circ}$	$0{,}098^{\blacktriangle\circ\circ}$	$0{,}292^{\blacktriangle\circ\circ}$
Dominância	$0{,}163^{\blacktriangle\blacktriangle\circ}$	$0{,}097^{\blacktriangle\blacktriangle\circ}$	$0{,}293^{\triangle\blacktriangle\circ}$	$0{,}163^{\blacktriangle\blacktriangle\blacktriangledown}$	$0{,}097^{\blacktriangle\blacktriangle\triangledown}$	$0{,}292^{\blacktriangle\blacktriangle\triangledown}$
Novidade	$0{,}150^{\triangle\blacktriangle\blacktriangle}$	$0{,}089^{\triangle\blacktriangle\blacktriangle}$	$0{,}271^{\circ\blacktriangle\blacktriangle}$	$0{,}157^{\blacktriangle\blacktriangle\blacktriangle}$	$0{,}096^{\blacktriangle\blacktriangle\blacktriangle}$	$0{,}292^{\triangle\blacktriangle\blacktriangle}$
Estabilidade	$0{,}153^{\blacktriangle\blacktriangle\blacktriangle}$	$0{,}093^{\blacktriangle\blacktriangle\blacktriangle}$	$0{,}277^{\triangle\blacktriangle\circ}$	$0{,}162^{\blacktriangle\blacktriangle\blacktriangle}$	$0{,}099^{\blacktriangle\blacktriangle\blacktriangle}$	$0{,}295^{\blacktriangle\blacktriangle\triangle}$
Recência Lin.	$0{,}165^{\blacktriangle\triangle\circ}$	$0{,}098^{\blacktriangle\blacktriangle\circ}$	$0{,}291^{\triangle\circ\circ}$	$0{,}163^{\blacktriangle\blacktriangle\circ}$	$0{,}098^{\blacktriangle\blacktriangle\circ}$	$0{,}293^{\blacktriangle\circ\circ}$
Recência Exp.	$0{,}164^{\blacktriangle\circ\circ}$	$0{,}098^{\blacktriangle\triangle\circ}$	$0{,}290^{\triangle\circ\circ}$	$0{,}163^{\blacktriangle\triangle\circ}$	$0{,}098^{\blacktriangle\triangle\circ}$	$0{,}292^{\blacktriangle\circ\circ}$

Assim, dados os resultados apresentados na Tabela 5.3, é possível fazer algumas considerações a respeito das normalizações propostas. A primeira observação diz respeito ao desempenho da função de associação *dominância de conteúdo* sem o uso de nenhuma normalização (ID), o que, para todas as métricas experimentadas, resultou nos melhores números até o momento, com $\text{nDCG}_{10} = 0,169$, $\text{P}_{10} = 0,102$ e MRR$= 0,302$. Contudo, em termos de significância, a função de associação em questão com a normalização SDC ficou estatisticamente equivalente.

Outra observação pertinente é que, para todas as funções de associação estudadas, as funções de normalização propostas não são significativamente inferiores a normalização DC proposta por Balog et al. [2006]. Isso se repete para todos os casos, ocorrendo apenas um empate usando a métrica MRR da normalização SDC com a função de associação *novidade no domínio*.

Além disso, para as funções de associação *novidade no domínio* e *estabilidade na dominância*, as funções de normalização propostas são extremamente benéficas para o resultado do ranking de especialistas, alcançando ganhos expressivos de 27,5% e 12,8% no nDCG_{10} quando comparamos a normalização SDC com a normalização DC, respectivamente.

Portanto, podemos dizer que, considerando as funções de normalização propostas, 75% dos casos nossas normalizações foram superiores a função de normalização DC com *p-value* $\leq 0,05$, sendo 36 casos analisados (3 métricas de avaliação, 6 funções de associação e de 2 normalização). Enquanto isso, em relação às normalizações Norm2 e ID, os ganhos são menos expressivos, mas também acontecem, sendo 69% dos resultados de SDC ou SCC superiores aos resultados da normalização Norm2 e 91% dos casos são superiores ou estatisticamente equivalentes aos resultados da normalização ID. Dessa última porcentagem, observamos ganhos significativos em 30% dos casos e empates em 61% dos casos.

Portanto, ao final, observamos que as funções de normalização propostas constroem rankings superiores às normalizações tradicionais em 65% dos 108 casos avaliados (3 métricas de avaliação, 3 *baselines*, 6 funções de associação e 2 de normalização). Com isso, é possível formalizar uma resposta para a segunda pergunta de pesquisa relacionada às funções de normalização. Em particular, demonstramos que as funções de normalização propostas são superiores à maioria das normalizações tradicionais, com exceção de um caso da normalização ID que, usando a função de associação *dominância de conteúdo*, apresentou resultados superiores estatisticamente do que a normalização SCC usando a mesma função de associação.

Assim, podemos concluir que, para os modelos generativos de ranking de especialistas, o esquema proposto que separa a modelagem da associação documento-candidato como combinação de duas funções facilita a compreensão dos resultados e abre espaço para estudos mais aprofundados dessa área do problema de busca de especialistas.

5.3 Resultados Experimentais: Modelos Discriminativos

A fim de avaliar a complementaridade das funções de associação e normalização propostas, nesta seção analisamos a eficácia dessas funções como atributos para modelos discriminativos para busca de especialistas. Para isso, detalhamos o procedimento efetuado para responder nossa terceira e última pergunta de pesquisa e apresentamos a resposta construída a partir dos resultados no decorrer desta seção.

5.3.1 Análise de Correlação

Para justificar o uso de diferentes modelos de associação em arcabouços que buscam se beneficiar da complementaridade dos rankings, apresentamos, na Figura 5.2, a correlação entre os valores resultantes das funções de associação. Nessa fase da análise de complementaridade, estudamos a correlação entre o peso individual de cada associação usando a correlação de Spearman.

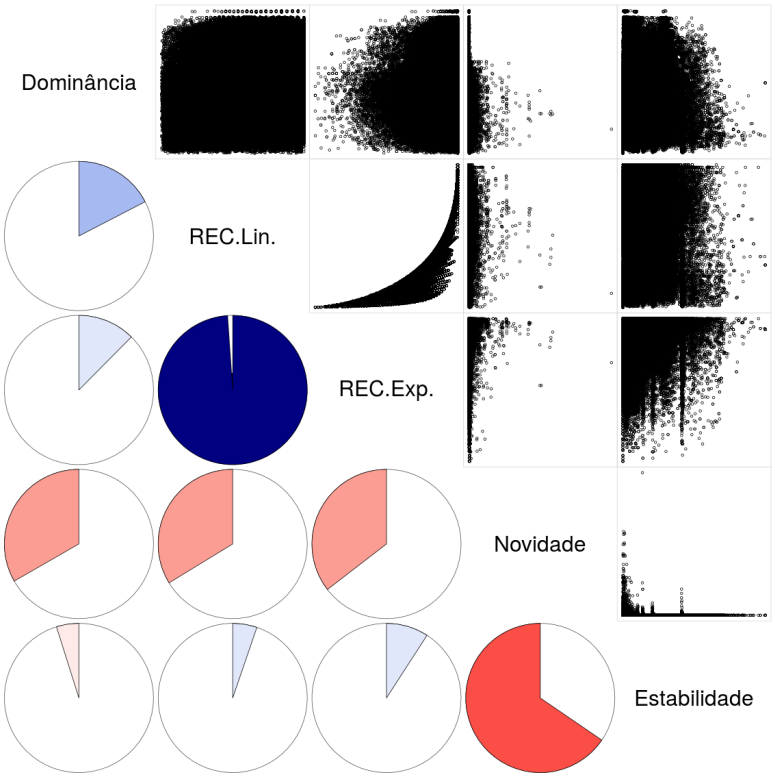

Figura 5.2: Correlação de Spearman entre os valores resultantes das funções de associação.

Na Figura 5.2 apresentamos, na diagonal inferior, o valor da correlação de Spearman entre os valores das funções de associação propostas, onde a cor azul significa correlação positiva (ou também pode ser lido onde o gráfico de pizza cresce em sentido horário) e a cor vermelha significa correlação negativa (ou também pode ser lido onde o gráfico de pizza cresce em sentido anti-horário) e, na diagonal superior apresentamos a distribuição das correlações. Como a função de associação booleana equivale a um valor constante, não foram estudadas correlações entre ela e as funções propostas.

Nessa figura podemos observar ainda que, como era esperado, as funções de associação baseadas na recência do documento são positivamente correlacionadas. Além disso, a estabilidade e a novidade são funções negativamente correlacionadas, o que

induz a interpretar que, conforme um trabalho é precursor em termos de conteúdo, os candidatos associados a ele tende a ser menos estáveis nesse conteúdo. Isso demonstra que nossa intuição corrobora com os valores esperados dos pesos das associações, onde um candidato não pode ser estável em um tema relativamente novo, tendo em vista que aquele conteúdo raramente foi mencionado anteriormente àquele trabalho.

Quanto à correlação entre as diferentes funções de associação para cada função de normalização, apresentamos, no Apêndice C, os gráficos que demonstram que os valores das funções de associação são sobrepostos pela normalização usada, tornando o peso final de cada associação correlacionadas aos pesos das associações formuladas através de outras funções de associação. Contudo, o objetivo principal dessa parte da dissertação é demonstrar que existe uma complementaridade entre os valores finais dos rankings de especialistas, mesmo que esses sejam correlacionados, para isso, analisamos os rankings gerados através do modelo discriminativo baseado em atributos agregados[1], conforme apresentado na Seção 2.1.2.2.

5.3.2 Modelo Baseado em Atributos Agregados

O *modelo discriminativo baseado em atributos agregados*, proposto por Macdonald et al. [2008], representa os diferentes rankings de especialistas gerando-os a partir de agregações de rankings generativos como atributos em algoritmos de *learning to rank* (L2R). Nessas abordagens, os autores propõem melhorar os rankings tentando usar a complementaridade de cada ranking gerado por modelos de associação diferentes. Assim, uma das técnicas aplicadas na fase de geração dos atributos foi considerar diferentes tamanhos do ranking de documentos gerado pela primeira fase do modelo de ranking de especialistas. Replicamos esse conceito e apresentamos, para cada técnica, diferentes modelos de rankings baseados em L2R.

Assim, apresentamos os resultados da agregação de ranking usando atributos agregados na Tabela 5.4. Comparamos os resultados com o melhor resultado (*) de cada normalização (demonstrado pelos símbolos sobrescritos). Além disso, os números em negrito representam os melhores valores da métrica para aquela normalização.

Analisando a tabela podemos notar que o *modelo baseado em atributos agregados* beneficia na maioria dos resultados, sendo inferior significativamente apenas na métrica P_{10} e $nDCG_{10}$ quando usamos o algoritmo de LambdaMART e atributos baseados na

[1]Durante o desenvolvimento desta dissertação, investigamos a complementaridade das funções de associação e normalização propostas como atributos do *modelo discriminativo baseados em atributos simples*, proposto, inicialmente, por Fang & Zhai [2007]. Contudo, durante a fase de experimentação, foi observado, a partir dos resultados, que esse arcabouço não explorou de maneira eficaz a complementaridade das funções propostas.

Tabela 5.4: Resultados dos modelos discriminativos baseados em agregação de ranking.

withcut	**nDCG$_{10}$**	**P$_{10}$**	**MRR**	**nDCG$_{10}$**	**P$_{10}$**	**MRR**
	DC			**CC**		
*	0,140	0,085	0,263	0,009	0,006	0,026
AdaRank	0,140	0,085	0,263	0,016$^{\triangle}$	0,013$^{\triangle}$	0,044$^{\blacktriangle}$
LambdaMART	0,142	0,084	0,260	0,106$^{\blacktriangle}$	0,065$^{\blacktriangle}$	0,197$^{\blacktriangle}$
MART	**0,151**	0,087	**0,284**	**0,115**$^{\blacktriangle}$	0,068$^{\blacktriangle}$	**0,204**$^{\blacktriangle}$
Random Forest	0,148	**0,088**	0,268	0,114$^{\blacktriangle}$	**0,071**$^{\blacktriangle}$	0,198$^{\blacktriangle}$
	SDC			**SCC**		
*	0,166	0,099	0,295	0,163	0,098	0,293
AdaRank	0,165	0,100	0,292	0,163	0,098	0,293
LambdaMART	0,147$^{\triangledown}$	0,085$^{\triangledown}$	0,280	0,161	0,094	0,312
MART	**0,173**	**0,102**	**0,311**	**0,173**	**0,101**	**0,316**
Random Forest	0,168	0,097	0,306	0,170	0,100	**0,316**$^{\triangle}$

normalização SDC. Além disso, o *modelo baseado em atributos agregados* é benéfico para todos os casos em que usamos rankings baseados na normalização CC, demonstrando a complementaridade dos rankings generativos baseados em diferentes funções de associações usados aqui como atributos para modelos de L2R.

Como era esperado, os modelos discriminativos que usam como atributos os rankings resultantes das normalizações propostas (SDC e SCC) são superiores aos modelos que usam os rankings das normalizações propostas por Balog et al. [2006] (DC e CC).

Assim como o resultado anterior, apresentamos, na Tabela 5.5, os resultados considerando o *win-loss* das métricas analisadas na Tabela 5.4. Os resultados que se destacam na Tabela 5.4 se repetem nesta tabela demonstrando que *modelos baseados em atributos agregados* são benéficos para os rankings finais dos especialistas.

Tabela 5.5: Resultados do *win-loss* das métricas para os modelos discriminativos baseados em agregação de ranking.

withcut	**nDCG$_{10}$**	**P$_{10}$**	**MRR**	**nDCG$_{10}$**	**P$_{10}$**	**MRR**
	DC			**CC**		
AdaRank	-	-	-	**21/8**	**96/2**	**173/59**
LambdaMART	58/63	**35/32**	90/104	**97/3**	**19/6**	**208/24**
MART	**98/2**	**38/29**	**108/88**	**98/2**	**94/3**	**208/23**
Random Forest	**71/49**	**33/24**	**97/88**	**98/2**	**97/2**	**212/21**
	SDC			**SCC**		
AdaRank	**35/33**	-	**65/51**	-	-	-
LambdaMART	42/86	29/39	69/120	61/67	29/39	**97/94**
MART	**69/56**	**31/24**	92/98	**73/56**	**31/24**	**94/88**
Random Forest	51/55	**21/17**	72/90	**70/46**	**21/17**	**91/73**

Tabela 5.6: Resultados dos modelos discriminativos baseados em agregação de todos os rankings.

	$\mathbf{nDCG_{10}}$	$\mathbf{P_{10}}$	**MRR**
*	0,166	0,099	0,295
AdaRank	0,163	0,098	0,293
LambdaMART	0,168	0,100	0,303
MART	**0,173**	**0,104**	0,311
Random Forest	**0,173**	0,102	**0,320**$^{\triangle}$

Assim, além dos resultados considerando a normalização CC, podemos notar que, para normalização SDC e SCC, os rankings gerados pelo MART cobrem uma quantidade de consultas maior, sendo, em 5 das 6 métricas avaliadas, superior. O segundo algoritmo que mais se destaca na cobertura das consultas é o RF, que é superior em 4 das 6 métricas avaliadas.

Quando consideramos as normalizações tradicionais (DC e CC), os algoritmos de RF e MART são superiores em todas as métricas avaliadas. Isso demonstra que os algoritmos de L2R são capazes de criar rankings de especialistas melhores do que os modelos generativos que usam apenas um ranking de documentos. Contudo, não é sempre que essa superioridade é estatisticamente significativa, como quando consideramos algoritmos como AdaRank. Ademais, rankings gerados por algoritmos como MART e RF apresentam, na maioria das vezes, melhores resultados em termos de MRR. Isso não se repete para todas as normalizações, mas para aquelas que são centradas em candidatos isso é intensificado.

Assim, demonstrado o potencial das agregações de ranking considerando subconjuntos de atributos divididos pela normalização usada, propomos executar um último experimento que utiliza todas as funções de associação e normalização propostas, incluindo também a normalização ID. Ao final, foram usados 120 rankings como atributos de entrada para o treino dos algoritmos de L2R (4 tamanhos de rankings de documentos, sendo eles $\{10, 100, 500, 1000\}$; 5 funções de normalização e 6 funções de associação).

Nesta bateria de experimentos, o objetivo é demonstrar que as diferentes normalizações se complementam na determinação de rankings de especialistas melhores. Assim, apresentamos o resultado dos rankings desta configuração experimental na Tabela 5.6. O *baseline* para essa configuração é o melhor resultado geral dos modelos generativos considerando o nDCG_{10}, definido como sendo a função de associação booleana e a normalização SDC.

Os resultados dos modelos discriminativos se comportaram de forma parecida com

o restante dos experimentos, tendo como destaque os algoritmos RF e MART. Apenas o resultado do MRR com o algoritmo RF foi significativamente superior ao *baseline*. Apesar disso, com exceção do algoritmo AdaRank, os resultados de agregação de rankings obtiveram valores superiores, chegando a ganhos de 8,5% para o MRR analisando o algoritmo RF. Além disso, esse valor possui significância estatística, resultando em um $p\text{-}value < 0{,}01$.

Em termos de *win-loss*, demonstramos, na Tabela 5.7, que os rankings gerados pela agregação usando todos os rankings como atributos foram superiores em quase todos os casos. Se analisarmos os dois casos em que os rankings não foram superiores, observamos que é uma abordagem *listwise* (LambdaMART) e uma abordagem *pairwise* (AdaRank). Esse resultado corrobora a observação de Balog et al. [2012] sobre a superioridade de algoritmos *pointwise* para L2R para busca de especialistas.

Assim, respondendo a pergunta de pesquisa proposta para os modelos discriminativos de agregação de ranking, dizemos que diferentes funções de associação e normalização podem beneficiar o resultado final dos rankings. Contudo, é preciso realizar experimentos mais finos para tentar identificar quais os principais fatores que levam aos resultados superiores de abordagens como RF e MART, sem que essas tenham diferença significativa com relação aos *baselines* estudados.

Numa análise posterior para tentar identificar os fatores que levaram aos resultados das métricas a serem relativamente baixos, analisamos a revocação das consultas nos rankings dos top-10 candidatos e o *Mean Average Precision* (MAP) dessas consultas em comparação com as mesmas abordagens em outras coleções de teste.

Dessa forma, o melhor resultado entre todas as abordagens generativas avaliadas marcou a revocação em 0,179, demonstrando que, na média, apenas 17,9% dos candidatos no gabarito são retornados nos top-10 candidatos ranqueados. Isso é uma evidência de porque os valores de nDCG_{10} e P_{10} podem estar baixos, sendo que a quantidade de candidatos relevantes devidamente retornados entre os 10 primeiros candidatos é relativamente pequena.

Ademais, considerando o MAP, medida que avalia a qualidade de um ranking con-

Tabela 5.7: Resultados do *win-loss* das métricas para os modelos discriminativos baseados em agregação de todos os rankings.

	nDCG_{10}	**P_{10}**	**MRR**
AdaRank	**61/59**	22/24	**104/69**
LambdaMART	**77/57**	34/36	**101/84**
MART	**72/61**	**33/24**	**104/80**
Random Forest	**68/56**	**29/23**	**104/65**

siderando a relevância binária dos candidatos para a consulta, nossa melhor abordagem generativa resultou em 0,122. As abordagens melhores colocadas nas três competições da TREC de busca de especialistas, que ocorreram em 2005 [Craswell et al., 2005], 2006 [Soboroff et al., 2006] e 2007 [Bailey et al., 2007a], resultaram 0,275, 0,643 e 0,463, respectivamente. Contudo, como mencionado na Seção 2.3, essas competições foram baseadas em coleções de teste no ambiente corporativo, portanto uma comparação direta dos resultados não beneficiaria a compreensão dos resultados.

Quando analisamos o MAP do melhor resultado encontrado no trabalho de Balog [2007] para a coleção UvT [Balog et al., 2007b], onde essa já é uma coleção de teste baseada no ambiente acadêmico, notamos que, na melhor configuração apresentada, o Modelo 2 resultou o MAP de 0,46. Considerando a proporção de candidatos avaliados na etapa de construção do gabarito (771 candidatos avaliados) e quantidade de candidatos na coleção (cerca de 206 mil candidatos), podemos dizer que nossa coleção existem menos candidatos avaliados (1635 associações candidato-consulta), portanto, não podemos inferir que os candidatos não encontrados no ranking resultante são, necessariamente, não-especialistas. Portanto, gerar um ranking de especialistas na coleção de teste experimentada demonstra ser mais difícil do que nas coleções de teste apresentadas na Seção 2.3.

Capítulo 6

Conclusões e Trabalhos Futuros

6.1 Conclusões

Nesta dissertação, apresentamos um modelo de construção de pesos de associações aplicado ao problema de busca de especialistas no ambiente acadêmico. O modelo de associação documento-candidato proposto combina uma função de associação, responsável por ponderar uma associação documento-candidato, e uma função de normalização, responsável por ajustar o peso das associações ao contexto que ela está inserida. Esse modelo de associação foi proposto visando quantificar a informação contida em cada publicação em relação aos seus respectivos autores, com o objetivo de aprimorar modelos de rankings de especialistas existentes. Com esse objetivo, realizamos um estudo em larga escala com uma coleção de teste construída com base em pesquisadores reconhecidamente especialistas, e demonstramos o potencial de múltiplas instâncias de funções de associação e normalização.

Portanto, uma das conclusões que podemos obter desta dissertação é que a proposta do processo de ponderação de associação documento-candidato dividido em duas etapas facilita a compreensão dos resultados obtidos considerando o modelo generativo de ranking de especialistas. Isso é mais evidente observando os resultados apresentados no Capítulo 5, onde as funções de associação propostas melhoraram os rankings de especialistas para as normalizações tradicionais. Em contrapartida, as funções de normalização propostas mostram-se superiores às funções de normalizações tradicionais.

Vale ressaltar, ainda que a normalização *candidate-centric*, proposta por Balog et al. [2006], demonstrou ser ineficaz na geração de rankings de especialistas para a coleção de teste experimentada. Uma justificativa para os resultados inferiores está relacionada ao viés a favor de candidatos com poucas publicações no seu curriculum, uma característica que deteriora abordagens de ranking de especialistas no ambiente acadê-

mico, onde candidatos prolíficos não necessariamente são candidatos menos propícios a serem especialistas.

Contudo, modelos discriminativos que agregam resultados de rankings se comportaram melhores do que as melhores abordagens generativas para cada função de normalização. Como apresentado, isso ocorreu graças a capacidade que os modelos de L2R têm de modelar a qualidade de cada atributo dado como entrada. Isso foi demonstrado pelo *win-loss* melhor das consultas avaliadas, onde abordagens como *Random Forest* (RF) foram superiores em 60% mais consultas do que o melhor resultado generativo usando apenas um modelo de associação.

6.2 Trabalhos Futuros

Com o processo de construção das ponderações das associações dividido em duas etapas, é possível analisar os fatores discriminantes para a obtenção de bons rankings de especialistas em outros ambientes. Por exemplo, é possível analisar os fatores que beneficiam os resultados de rankings de especialistas no ambiente empresarial, identificando quais etapas (função de associação ou função de normalização) beneficiam mais esses rankings no processo de determinação dos especialistas, incluindo a experimentação das instâncias propostas e apresentação de novas configurações experimentais aplicadas a esse ambiente. Além disso, pretendemos avaliar os modelos propostos em outras coleções para busca de especialistas na academia, como TU e ArnetMiner.

Além disso, é possível analisar como se comportam as funções de associação e normalização propostas em abordagens baseadas em grafos, onde, iterativamente, os pesos dos candidatos e documentos são atualizados através de processos de caminhamento aleatório. Para esses processos, pode-se ir além, propondo ponderações para as associações documento-documento e candidato-candidato, mensurando a relação entre as entidades através de funções de associação similares às propostas nesta dissertação.

Por último, pode-se estimar a qualidade do resultado das consultas através da análise de *clicks* [Sanderson, 2010] aplicando o resultado das consultas em uma ferramenta real de validação do ranking. Esse tipo de sistema auxiliaria, inclusive, na apresentação do resultado dos rankings na tarefa de buscar candidatos a orientadores de projetos de pesquisa, avaliadores de artigos científicos e para recomendar colaboradores para o desenvolvimento de trabalhos acadêmicos.

Referências Bibliográficas

Alves, A.; Yanasse, H. & Soma, N. (2012). Lattesminer: uma linguagem de domínio específico para extração automática de informações da plataforma lattes. In *Workshop de Computação Aplicada*, volume 12.

Baeza-Yates, R. A. & Ribeiro-Neto, B. A. (2011). *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England. ISBN 978-0-321-41691-9.

Bailey, P.; Craswell, N.; Soboroff, I. & de Vries, A. P. (2007a). The CSIRO enterprise search test collection. In *ACM SIGIR Forum*, volume 41, pp. 42--45. ACM.

Bailey, P.; de Vries, A. P.; Craswell, N. & Soboroff, I. (2007b). Overview of the TREC 2007 Enterprise track. In *Proceedings of Text REtrieval Conference*.

Balancieri, R.; Bovo, A. B.; Kern, V. M.; Pacheco, R. d. & Barcia, R. M. (2005). A análise de redes de colaboração científica sob as novas tecnologias de informação e comunicação: um estudo na plataforma lattes. *Ciência da Informação*, 34(1):64--77.

Balog, K. (2007). People search in the enterprise. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 916--916. ACM.

Balog, K.; Azzopardi, L. & De Rijke, M. (2006). Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 43--50. ACM.

Balog, K.; Bogers, T.; Azzopardi, L.; De Rijke, M. & Van Den Bosch, A. (2007a). Broad expertise retrieval in sparse data environments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 551--558. ACM.

Balog, K.; Bogers, T.; Azzopardi, L.; de Rijke, M. & van den Bosch, A. (2007b). Broad expertise retrieval in sparse data environments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 551--558.

Balog, K. & de Rijke, M. (2006). Finding experts and their details in e-mail corpora. In *Proceedings of the 15th international conference on World Wide Web*, pp. 1035--1036.

Balog, K. & De Rijke, M. (2008). Associating people and documents. In *Proceedings of the 30th European Conference on IR Research*, pp. 296--308.

Balog, K.; Fang, Y.; de Rijke, M.; Serdyukov, P. & Si, L. (2012). Expertise retrieval. *Foundations and Trends in Information Retrieval*, 6(2–3):127--256.

Balog, K.; Soboroff, I.; Thomas, P.; Craswell, N.; de Vries, A. P. & Bailey, P. (2008a). Overview of the TREC 2008 Enterprise track. In *Proceedings of Text REtrieval Conference*.

Balog, K.; Thomas, P.; Craswell, N.; Soboroff, I.; Bailey, P. & De Vries, A. P. (2008b). Overview of the trec 2008 enterprise track. Relatório técnico, DTIC Document.

Barbosa, S. d. F. F.; Sasso, G. T. M. D. & Berns, I. (2009). Enfermagem e tecnologia: análise dos grupos de pesquisa cadastrados na plataforma lattes do cnpq. *Texto and Contexto Enfermagem*, 18(3):443.

Berendsen, R.; de Rijke, M.; Balog, K.; Bogers, T. & van den Bosch, A. (2013a). On the assessment of expertise profiles. *Journal of the American Society for Information Science and Technology*, 64(10):2024--2044.

Berendsen, R.; Rijke, M.; Balog, K.; Bogers, T. & Bosch, A. (2013b). On the assessment of expertise profiles. *Journal of the American Society for Information Science and Technology*, 64(10):2024--2044.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5--32.

Burges, C. J. (2010). From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11:23--581.

Castano, A. C. (2008). *Populando ontologias através de informaçoes em HTML-o caso do currıculo lattes*. Tese de doutorado, Universidade de Sao Paulo.

Cover, T. M. & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Craswell, N.; de Vries, A. P. & Soboroff, I. (2005). Overview of the TREC 2005 Enterprise track. In *Proceedings of Text REtrieval Conference*.

Davenport, T. H. & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Harvard Business Press.

Deng, H.; King, I. & Lyu, M. R. (2008). Formal models for expert finding on DBLP bibliography data. In *Proceedings of the Eighth IEEE International Conference on Data Mining*, pp. 163--172.

Dennis Jr, J. E. & Schnabel, R. B. (1996). *Numerical methods for unconstrained optimization and nonlinear equations*, volume 16. Siam.

Digiampietri, L.; Mena-Chalco, J.; de Jésus Pérez-Alcázar, J.; Tuesta, E. F.; Delgado, K. & Mugnaini, R. (2012). Minerando e caracterizando dados de currıculos lattes. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*.

Fang, H. & Zhai, C. (2007). *Probabilistic models for expert finding*. Springer.

Fang, Y.; Si, L. & Mathur, A. (2009). Ranking experts with discriminative probabilistic models. In *SIGIR Workshop on Learning to Rank for Information Retrieval,(LR4IR?09)*.

Fang, Y.; Si, L. & Mathur, A. P. (2010a). Discriminative models of integrating document evidence and document-candidate associations for expert search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 683--690.

Fang, Y.; Si, L. & Mathur, A. P. (2010b). Discriminative models of integrating document evidence and document-candidate associations for expert search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 683--690. ACM.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232. ISSN 00905364.

Hertzum, M. (2000). People as carriers of experience and sources of commitment: Information seeking in a software design project. *New Rev. Inf. Behav. Res.*, 1(January):135--149.

Jordan, A. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14:841.

Keikha, M.; Gerani, S. & Crestani, F. (2011). Relevance stability in blog retrieval. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, pp. 1119--1123. ACM.

Lavrenko, V. & Croft, W. B. (2003). *Relevance Models in Information Retrieval*, capítulo 2.

Li, X. & Croft, W. B. (2003). Time-based language models. In *Proceedings of the twelfth international conference on Information and knowledge management*, pp. 469--475. ACM.

Liebregts, R. & Bogers, T. (2009). Design and evaluation of a university-wide expert search engine. In *Proceedings of the 31th European Conference on IR Research*, pp. 587--594. Springer.

Macdonald, C.; Hannah, D. & Ounis, I. (2008). High quality expertise evidence for expert search. In *Proceedings of the 30th European Conference on IR Research*, pp. 283--295.

Macdonald, C. & Ounis, I. (2006). Voting for candidates: adapting data fusion techniques for an expert search task. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pp. 387--396.

Macdonald, C. & Ounis, I. (2011). Learning models for ranking aggregates. In *Proceedings of the 33rd European Conference on IR Research*, pp. 517--529.

Mangaravite, V. & Santos, R. L. T. (2016). On information-theoretic document-person associations for expert search in academia. In *Proceedings of the 39th annual international ACM SIGIR conference on Research and development in information retrieval*.

Mangaravite, V.; Santos, R. L. T.; Ribeiro, I. S.; Gonçalves, M. A. & Laender, A. H. F. (2016). The LExR collection for expertise retrieval in academia. In *Proceedings of the 39th annual international ACM SIGIR conference on Research and development in information retrieval*.

Manning, C. D.; Raghavan, P.; Schütze, H. et al. (2008). *Introduction to information retrieval*, volume 1. Cambridge University Press.

Mena-Chalco, J. (2009). Scriptlattes software: uma ferrramenta para extração e visualização de conhecimento a partir de currículos lattes. *São Paulo*.

Opitz, D. & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, pp. 169--198.

Petkova, D. & Croft, W. B. (2007). Proximity-based document representation for named entity retrieval. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 731--740. ACM.

Petkova, D. & Croft, W. B. (2008). Hierarchical language models for expert finding in enterprise corpora. *International Journal on Artificial Intelligence Tools*, 17(1):5--18.

Ribeiro, I. S.; Santos, R. L. T.; Gonçalves, M. A. & Laender, A. H. F. (2015). On tag recommendation for expertise profiling: a case study in the scientific domain. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pp. 189--198, Shanghai, China. ACM.

Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247--375.

Serdyukov, P. & Hiemstra, D. (2008). Modeling documents as mixtures of persons for expert finding. In *Proceedings of the 30th European Conference on IR Research*, pp. 309--320.

Serdyukov, P.; Rode, H. & Hiemstra, D. (2008). Modeling multi-step relevance propagation for expert finding. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pp. 1133--1142. ACM.

Silva, F. M. (2007). *Organização da Informação em sistemas eletrônicos abertos de Informação Científica & Tecnológica: Análise da Plataforma Lattes. 2007 163 f.* Tese de doutorado, Tese (Doutorado em Ciência da Informação)–Departamento de Biblioteconomia e Documentação, Universidade de São Paulo, São Paulo.

Smucker, M. D.; Allan, J. & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 623--632. ACM.

Soboroff, I.; de Vries, A. P. & Craswell, N. (2006). Overview of the TREC 2006 Enterprise track. In *Proceedings of Text REtrieval Conference*.

Tang, J.; Zhang, J.; Yao, L.; Li, J.; Zhang, L. & Su, Z. (2008a). Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM*

SIGKDD international conference on Knowledge discovery and data mining, pp. 990--998. ACM.

Tang, J.; Zhang, J.; Yao, L.; Li, J.; Zhang, L. & Su, Z. (2008b). ArnetMiner: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 990--998.

Tribus, M. (1961). *Thermostatics and thermodynamics*. Center for Advanced Engineering Study, Massachusetts Institute of Technology.

Xia, F.; Chen, Z.; Wang, W.; Li, J. & Yang, L. T. (2014). Mvcwalker: Random walk-based most valuable collaborators recommendation exploiting academic factors. *Emerging Topics in Computing, IEEE Transactions on*, 2(3):364--375.

Xu, J. & Li, H. (2007). Adarank: a boosting algorithm for information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 391--398. ACM.

Apêndice A

Comparação da Entropia Cruzada

Neste experimento, replicamos as melhores configurações experimentais encontradas pelo processo descrito na Seção 5.2, onde os símbolos △ e ▲ representam ganhos significativos, com *p-value* menores que 0,05 e 0,01, respectivamente e os símbolos ▽ e ▼ representam perdas significativas, com *p-value* menores que 0,05 e 0,01, respectivamente, sendo ambas as representações considerando o teste T-*student*.

Propusemos esses experimentos extras visando converter a função de entropia cruzada em versões que computam a similaridade entre as distribuições, demonstrando que, ao contrário da intuição apresenta, a dissimilaridade pode ser usada como métrica de dominância de conteúdo. Para isso, foram usadas duas estratégias: (1) Considerar o inverso da entropia cruzada (Equação A.1), e (2) a sigmoide da entropia cruzada negativa (Equação A.2). Ambas as adaptações foram usadas como função de associação nas diferentes instâncias de função de normalização e avaliadas com duas métricas de avaliação de ranking, nDCG_{10} e P_{10}.

Assim, definimos a primeira configuração, que considera o inverso da entropia cruzada, como

$$\rho_{inv\_d}(\theta_e||\theta_d) = {}^{1}/_{H(\theta_e||\theta_d)} \tag{A.1}$$

onde, $H(\theta_e||\theta_d)$ é a entropia cruzada entre o modelo θ_e do candidato e $\theta_e\theta_d$. E, a seguir, apresentamos a versão sigmoidal da entropia cruzada.

$$\rho_{sig\_d}(\theta_e||\theta_d) = \frac{1}{1+\exp(H(\theta_e||\theta_d))} \tag{A.2}$$

onde exp é a função exponencial.

Considerando que a entropia cruzada é uma medida assimétrica de dissimila-

ridade, ou seja, $H(\theta_e||\theta_d) \neq H(\theta_d||\theta_e)$, avaliamos também a função de associação de dominância de conteúdo considerando o custo de transformar a distribuição dos termos do candidato e na distribuição dos termos do documento d, onde a entropia cruzada é dada pela formulação $H(\theta_d||\theta_e)$.

Assim, nos lugares em que os símbolos de significância estiverem sobrescritos a comparação se dá entre as abordagens de mesma normalização, comparando as versões alternativas da entropia cruzada com a tradicional, formulada como $H(\theta_e||\theta_d)$. Enquanto isso, onde os símbolos de significância ocorrerem subscritos, a comparação se dá entre as mesmas normalizações, contudo comparando a versão da função de associação a que usa a entropia cruzada $H(\theta_d||\theta_e)$. Em relação a isso, determinamos que os *baselines* de comparação são aqueles que usam a função de entropia cruzada padrão, considerando $H(\theta_e||\theta_d)$.

Assim, com exceção da normalização $Norm2$, todas as funções de associação que usam outras versões da entropia cruzada demonstraram ser menos eficazes em comparação com a versão tradicional que usa, basicamente, $H(\theta_e||\theta_d)$. Enquanto isso,

Tabela A.1: Tabela da comparação das instanciações das funções de associação de dominância.

Entropia Cruzada	**nDCG$_{10}$**	**P$_{10}$**	**nDCG$_{10}$**	**P$_{10}$**
	ID		**Norm2**	
$\rho_d(e,d)$	0.169	0.102	0.132	0.079
$\rho_{inv\_d}(e,d)$	$0.098^{\blacktriangledown}$	$0.061^{\blacktriangledown}$	$0.169^{\blacktriangle}$	$0.101^{\blacktriangle}$
$\rho_{sig\_d}(e,d)$	$0.000^{\blacktriangledown}$	$0.000^{\blacktriangledown}$	$0.171^{\blacktriangle}$	$0.101^{\blacktriangle}$
$\rho_d(d,e)$	0.161	0.097	0.136	0.080
$\rho_{inv\_d}(d,e)$	$0.125^{\blacktriangledown}_{\blacktriangle}$	$0.073^{\blacktriangledown}_{\blacktriangle}$	$0.003^{\blacktriangledown}_{\blacktriangledown}$	$0.001^{\blacktriangledown}_{\blacktriangledown}$
$\rho_{sig\_d}(d,e)$	$0.147^{\triangledown}_{\blacktriangle}$	$0.089^{\triangledown}_{\blacktriangle}$	$0.164^{\blacktriangle}$	$0.097^{\blacktriangle}$
	DC		**CC**	
$\rho_d(e,d)$	0.140	0.085	0.009	0.007
$\rho_{inv\_d}(e,d)$	$0.114^{\blacktriangledown}$	$0.068^{\blacktriangledown}$	$0.012^{\vartriangle}$	0.010
$\rho_{sig\_d}(e,d)$	$0.050^{\blacktriangledown}$	$0.032^{\blacktriangledown}$	0.008	0.006
$\rho_d(d,e)$	$0.133_{\triangledown}$	$0.080_{\triangledown}$	0.008	0.007
$\rho_{inv\_d}(d,e)$	$0.130_{\blacktriangledown}$	$0.078_{\blacktriangledown}$	0.011	0.009
$\rho_{sig\_d}(d,e)$	$0.132_{\blacktriangledown}$	$0.079_{\blacktriangledown}$	0.010	0.007
	SDC		**SCC**	
$\rho_d(e,d)$	0.163	0.097	0.163	0.097
$\rho_{inv\_d}(e,d)$	0.161	0.093	0.161	0.097
$\rho_{sig\_d}(e,d)$	$0.108^{\blacktriangledown}$	$0.055^{\blacktriangledown}$	$0.012^{\blacktriangledown}$	$0.008^{\blacktriangledown}$
$\rho_d(d,e)$	0.166	0.101	0.165	0.099
$\rho_{inv\_d}(d,e)$	$0.154^{\triangledown}$	$0.093^{\triangledown}$	$0.149^{\blacktriangledown}_{\triangledown}$	0.092
$\rho_{sig\_d}(d,e)$	$0.163_{\blacktriangle}$	$0.095_{\blacktriangle}$	$0.163_{\blacktriangle}$	$0.098_{\blacktriangle}$

as versões da entropia cruzada que usam $H(\theta_d||\theta_e)$ como dominância de conteúdo foram, em situações pontuais, superiores às suas versões que consideram $H(\theta_e||\theta_d)$ como função de ponderação da dominância de conteúdo.

Um surpresa em relação aos resultados foi a função $\rho_{sig\_d}(d, e)$ usando a normalização $Norm2$, que obteve os melhores resultados gerais de nDCG_{10}, apesar de não significativamente diferente dos demais. Como observação final, podemos notar que, para normalização $Norm2$, que visa penalizar candidatos prolixos, abordagens de similaridade se comportam melhor, como demonstrado pelos resultados das abordagens que usam dominância de conteúdo a partir da entropia cruzada dada pela formulação $H(\theta_e||\theta_d)$, que são funções de associação que usam dissimilaridade.

Apêndice B

Classificação de Especialidade

Caro(a) ≪Nome do candidato≫

Como parte de um projeto do Instituto Nacional de Ciência e Tecnologia para a Web (InWeb), desenvolvemos um novo método para a identificação automática de especialistas em diferentes áreas do conhecimento e gostaríamos de contar com a sua colaboração para validá-lo.

Abaixo, para algumas de suas áreas de especialidade informadas, listamos outros pesquisadores também indicados como possíveis especialistas em cada área, juntamente com um link para seus respectivos currículos Lattes. Para cada área, por favor, indique o nível de especialidade de cada pesquisador listado, segundo a seguinte escala

0. **Indiscriminante**: O pesquisador que está respondendo o questionário disse não possuir elementos para mensurar o nível de especialidade do pesquisador avaliado, ou prefere não opinar sobre o grau de especialidade do candidato;

1. **Fracamente relevante**: O pesquisador que está respondendo o questionário disse que o pesquisador avaliado tem **sólidos conhecimentos na área**;

2. **Relevante**: O pesquisador que está respondendo o questionário disse que o pesquisador avaliado **é um especialista na área**;

3. **Fortemente relevante**: O pesquisador que está respondendo o questionário disse que o pesquisador avaliado é a **principal referência na área**.

Apêndice C

Correlação de Funções de Associação

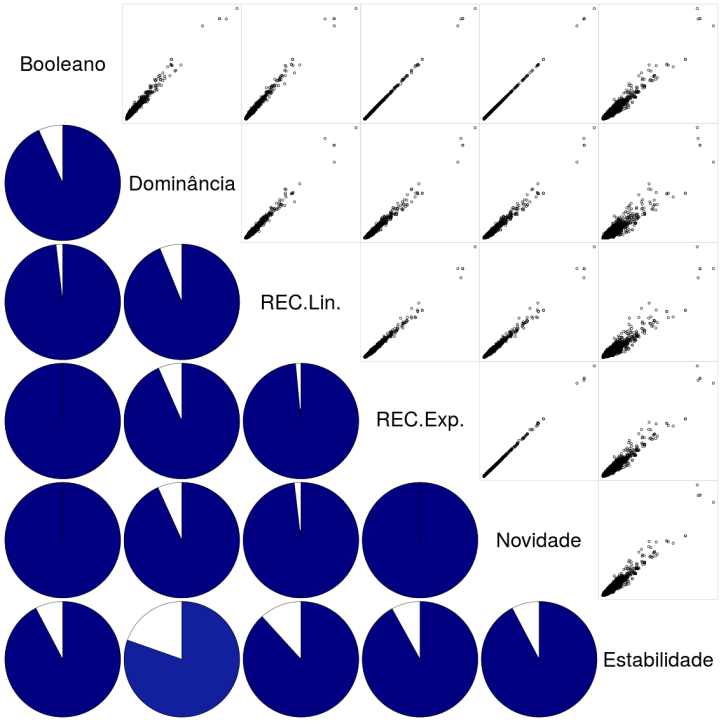

Figura C.1: DC

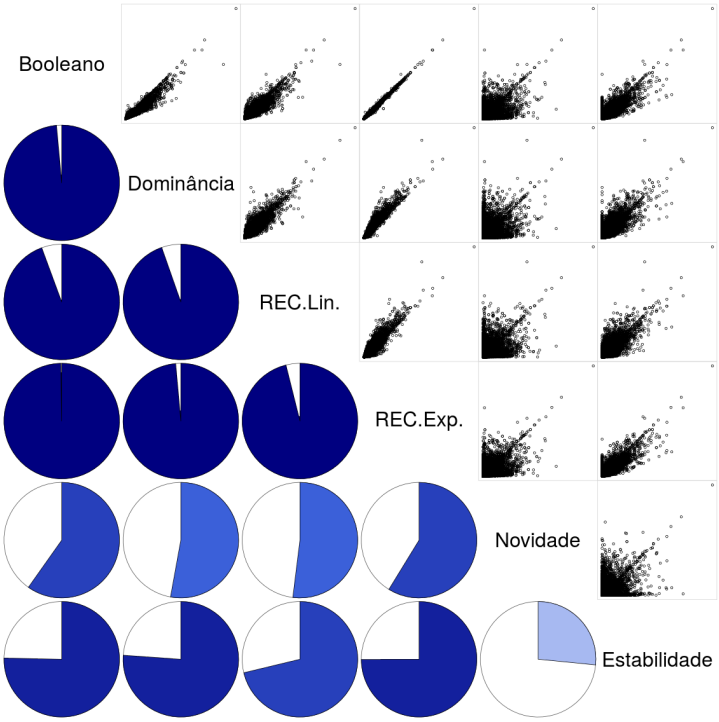

Figura C.2: CC

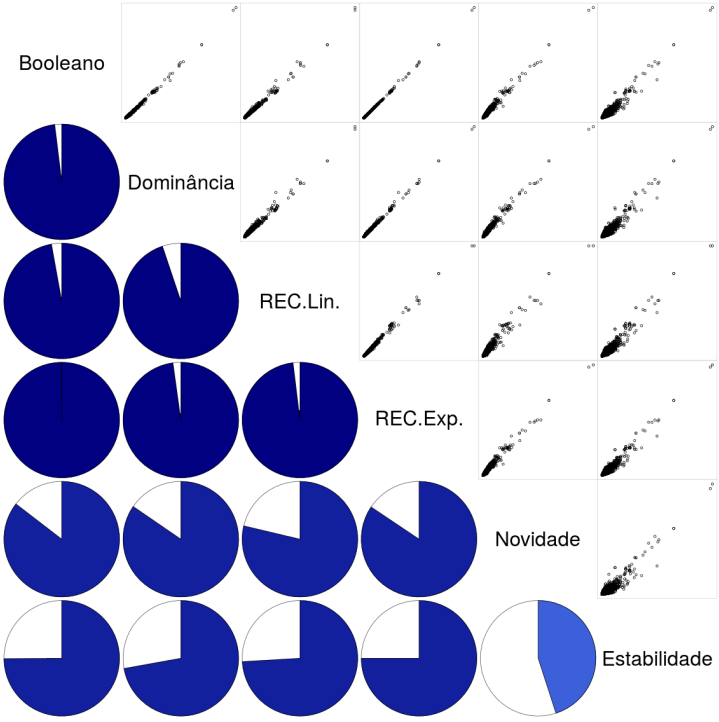

Figura C.3: SDC

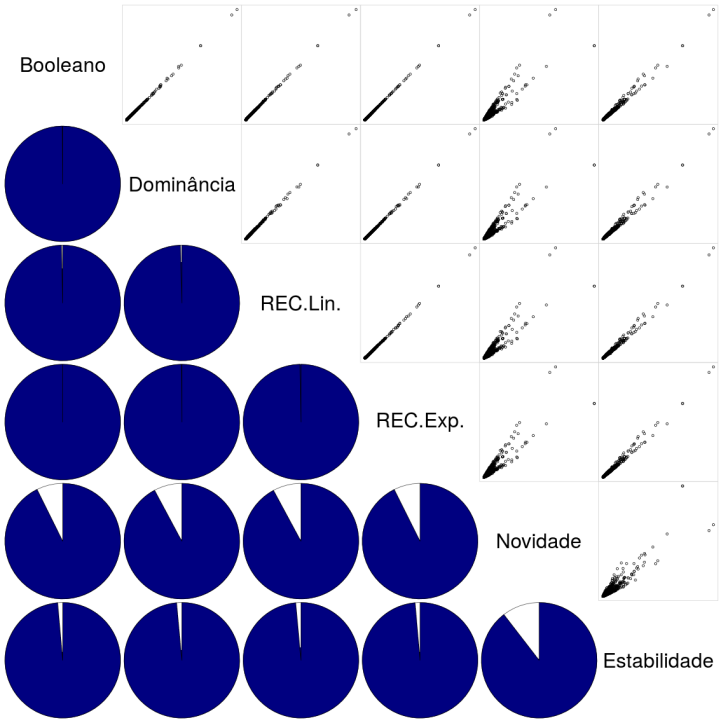

Figura C.4: SCC