RECOMENDAÇÃO DE ETIQUETAS PARA SUMARIZAÇÃO DE PERFIS ACADÊMICOS

ISAC SANDIN RIBEIRO

RECOMENDAÇÃO DE ETIQUETAS PARA SUMARIZAÇÃO DE PERFIS ACADÊMICOS

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: Alberto Henrique Frade Laender
Coorientador: Rodrygo Luis Teodoro Santos

Belo Horizonte
Julho de 2015

Ribeiro, Isac Sandin
R484r Recomendação de etiquetas para sumarização de perfis acadêmicos / Isac Sandin Ribeiro. — Belo Horizonte, 2015
xx, 45 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais. Departamento Ciência da Computação.

Orientador: Alberto Henrique Frade Laender.
Coorientador: Rodrygo Luis Teodoro Santos.

1. Computação - Teses. 2. Recuperação da informação. 3. Indexação. I. Orientador. II. Coorientador. III. Título.

CDU 519.6*73(043)

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

FOLHA DE APROVAÇÃO

Recomendação de etiquetas para sumarização de perfis acadêmicos

ISAC SANDIN RIBEIRO

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

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

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

PROF. LEANDRO BALBY MARINHO
Departamento de Informática - UFMA

PROF. LEONARDO CHAVES DUTRA DA ROCHA
Departamento de Ciência da Computação - UFSJ

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

Belo Horizonte, 30 de julho de 2015.

Agredecimentos

Agradeço primeiramente a Deus, à minha querida mãe Ilda, que já não está mais entre nós, mas que sempre me inspirou a ser uma pessoa de bom coração e de caráter. Agradeço ao meu pai pelo apoio e força, e por dedicar tanto tempo de sua vida trabalhando para ver seus filhos formados. Agradeço à minha esposa Denise pelo amor incondicional e apoio nos momentos felizes e nos momentos difíceis dessa jornada. Agradeço ao meu orientador professor Alberto Laender, pela sua paciência e dedicação. Ao professor Rodrygo Teodoro e ao professor Marcos Gonçalves pelas dicas valiosas e feedback. Quero agradecer também ao professor Leonardo Rocha, que desde o princípio buscou uma forma de me ajudar, quase teve muitos infartos por minha causa, mas que finalmente viu seu aluno formado e a caminho do mestrado, meu muito obrigado! Agradeço também aos colegas, Guilherme e Felipe Viegas pela amizade e companheirismo durante todo o curso, pelas brincadeiras descontraídas e pelas horas de estudo nas vésperas de provas, nossa república foi a minha segunda casa! Finalmente, quero agradecer ao pessoal do LBD, aos professores, pelos ensinamentos valiosos e aos colegas de laboratório, pelo companheirismo e momentos descontraídos, vocês são sensacionais!

Resumo

Construir perfis de especialidade é um ponto crucial para identificar especialistas em diferentes áreas do conhecimento. No entanto, sumarizar os tópicos de especialidade de um indivíduo é um grande desafio, principalmente pela natureza semiestruturada e heterogênea das evidências documentais disponíveis para esta tarefa. Nessa dissertação investigamos a aplicação de métodos de recomendação de etiquetas (do inglês *tags*) como mecanismo para construir perfis de especialidade. Em particular, realizamos um estudo em larga escala usando especialistas acadêmicos de diversas áreas do conhecimento para verificar a efetividade de vários recomendadores de etiquetas supervisionados e não-supervisionados, bem como para avaliar a efetividade de diversas fontes de evidência textual. Nossa análise revelou que os métodos tradicionais de recomendação baseados em conteúdo tiveram um bom desempenho em identificar etiquetas relacionadas às especialidades dos pesquisadores, com palavras-chave sendo a mais efetiva das fontes textuais para perfis de diferentes áreas do conhecimento e vários níveis de esparsidade de dados. Além disso, combinamos múltiplos recomendadores e fontes de evidência textual como sinais de aprendizado, demonstrando assim a efetividade das técnicas de recomendação de etiquetas para o problema da construção de perfis de especialidade.

Abstract

Building expertise profiles is a crucial step towards identifying experts in different knowledge areas. However, summarizing the topics of expertise of a given individual is a challenging task, primarily due to the semi-structured and heterogeneous nature of the documentary evidence available for this task. In this dissertation, we investigate the suitability of tag recommendation as a mechanism to produce effective expertise profiles. In particular, we perform a large-scale user study with academic experts from different knowledge areas to assess the effectiveness of multiple supervised and unsupervised tag recommendation approaches as well as multiple sources of textual evidence. Our analysis reveals that traditional content-based tag recommenders perform well for identifying expertise-oriented tags, with article keywords being a particularly effective source of evidence across profiles in different knowledge areas and with various levels of sparsity. Moreover, by combining multiple recommenders and sources of evidence as learning signals, we further demonstrate the effectiveness of tag recommendation for expertise profiling.

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Sumário

Capítulo 1

Introdução

1.1 Contextualização

Com o advento da web, as possibilidades de compartilhamento de experiências e conhecimentos entre pessoas tornaram-se sem precedentes. Um domínio que pode se beneficiar tremendamente dessas possibilidades de interação é a ciência. Em particular, pesquisadores de diversas áreas podem estabelecer novos contatos a partir de redes sociais e colaborar em problemas de pesquisa desafiadores. No entanto, o grande número de pesquisadores trabalhando *online* e procurando por contatos relevantes e parcerias requer novas ferramentas e serviços que permitam automaticamente encontrar especialistas em uma determinada área de pesquisa.

Entretanto, para que esses serviços funcionem de forma adequada dois desafios precisam ser resolvidos: (1) a construção de perfis que podem descrever a especialidade dos pesquisadores de forma significativa e completa e (2) a criação de algoritmos capazes de ordenar pesquisadores de acordo com seus perfis de especialidade. A segunda tarefa, chamada de busca de especialistas[1], tem recebido muita atenção da comunidade acadêmica atualmente [Balog et al., 2012]. A tarefa de sumarização de especialidades[2], por outro lado, tem recebido muito menos atenção [Rybak et al., 2014a; Serdyukov et al., 2011] e, portanto, é o nosso foco nesta dissertação.

1.2 Motivação

Existem muitas dificuldades relacionadas com a construção de perfis de especialidade significativos, tais como: (i) balanceamento entre concisão e representatividade, (ii)

[1]Do inglês *expert finding*.
[2]Do inglês *expertise profiling*.

identificação dos melhores tópicos para sumarizar carreiras acadêmicas muitas vezes longas, (iii) modelagem da evolução de tópicos de interesse no tempo e também (iv) extração e agregação de diferentes fontes de evidência, incluindo currículos *online*, páginas em diferentes bibliotecas digitais e redes sociais.

Para tratar alguns desses desafios, abordamos o problema de construção de perfis de especialidade como um problema de recomendação de etiquetas que representam especialidades de pessoas[3] (ou etiquetagem de pessoas[4]). Embora o problema de recomendação de etiquetas tem sido bastante estudado para diferentes tipos de mídia *online* [Canuto et al., 2013; Figueiredo et al., 2009; Garg & Weber, 2008; Heymann et al., 2008; Sigurbjörnsson & van Zwol, 2008a], a etiquetagem de pessoas é uma área que tem sido muito pouco investigada e que possui enormes oportunidades para melhorias.

1.3 Contribuições

Nesta dissertação, investigamos a aplicação de recomendadores de etiquetas para produzir perfis de especialidade efetivos. Para isso, realizamos um estudo em larga escala envolvendo 1.288 respondentes (de 5.355 contatados) dentre os mais proeminentes pesquisadores de diferentes áreas do conhecimento no Brasil, para avaliar a efetividade da técnica de etiquetagem de pessoas para a construção de perfis de especialidade. Como fonte de evidência da especialidade de cada pesquisador, usamos o seu currículo disponível na Plataforma Lattes[5], uma iniciativa reconhecida internacionalmente que mantém informações sobre ciência, tecnologia e inovação relacionadas a pesquisadores individuais e grupos de pesquisa no Brasil [Lane, 2010].

Nosso trabalho contrasta três recomendadores de etiquetas baseados em conteúdo representativos na literatura, explorando três diferentes fontes de evidência de especialidade baseadas nas publicações listadas no Currículo Lattes de cada pesquisador, a saber, seus títulos, resumos e palavras-chave, resultado em nove combinações. Nossos resultados experimentais demonstram a efetividade, completude e robustez dos perfis de especialidade construídos através da recomendação de etiquetas para pesquisadores em diferentes áreas do conhecimento e para vários níveis de esparsidade das fontes textuais utilizadas. Enquanto perfis de especialidade baseados em palavras-chave mostraram-se os mais efetivos, nós investigamos a aplicação de múltiplas estratégias de *Learning*

[3]Do inglês: *tag recommendation for people*.
[4]Do inglês: *people tagging*.
[5]`http://lattes.cnpq.br`

to Rank(L2R) para combinar os algoritmos de recomendação de etiquetas aplicados a diferentes fontes textuais como atributos de aprendizado.

Nossos resultados mostraram que a melhor das estratégias de *Learning to Rank*, que é baseada em um comitê de modelos de *ranking*, pode produzir ganhos em torno de 21.5% se comparado ao recomendador de etiquetas que obteve o melhor resultado (algoritmo TF aplicado a palavras-chave). Além disso, usando apenas duas características no modelo de *Learning to Rank*, mostramos que o resultado é equivalente a 80% do resultado obtido utilizando nove características, o que provê uma indicação sobre o potencial de aplicação dos recomendadores de etiquetas em ambientes reais. Até onde temos conhecimento, o volume deste trabalho em termos de cobertura (um país inteiro), diferentes áreas do conhecimento, diferentes algoritmos e fontes de evidência é único.

Em suma, as principais contribuições desta dissertação são:

1. Um estudo em larga escala para sumarização de perfis de especialidade envolvendo os mais proeminentes pesquisadores brasileiros de diferentes áreas do conhecimento e vários tamanhos de carreira.

2. Um minucioso estudo empírico da efetividade de múltiplos recomendadores de etiquetas supervisionados e não-supervisionados para construção de perfis de especialidade de forma automática.

Os resultados apresentados nesta dissertação foram consolidados em um artigo publicado na *Eighth ACM International Conference on Web Search and Data Mining* [Ribeiro et al., 2015].

1.4 Organização

O restante desta dissertação está organizado da seguinte forma. O Capítulo 2 cobre os trabalhos relacionados em diferentes tópicos como construção de perfis de especialidade, recomendação de etiquetas e avaliação de recomendadores de etiquetas.

O Capítulo 3 detalha a metodologia de nossa avaliação experimental, incluindo os procedimentos para extrair e ordenar as etiquetas candidatas, bem como a coleta dos julgamentos de relevância em nosso estudo.

O Capítulo 4 discute os resultados de nossa avaliação experimental. Nesse capítulo respondemos à nossas questões de pesquisa avaliando completude, robustez e complementariedade dos perfis de especialidade gerados pelas combinações de algoritmos de recomendação e fontes de dados.

Finalmente, o Capítulo 5 apresenta as nossas conclusões e discute direções para trabalhos futuros.

Capítulo 2

Trabalhos Relacionados

Neste capítulo, apresentamos uma revisão dos principais trabalhos encontrados na literatura relacionados ao tópico desta dissertação. Cada seção tratará de uma classe de problema. Inicialmente, abordamos a construção de perfis de especialidade e etiquetagem de pessoas, as técnicas envolvidas nesse processo e os desafios existentes. A seguir descrevemos algumas técnicas para tratamento de fontes textuais, como sumarização textual e extração de n-gramas. Em seguida abordamos técnicas de recomendação de etiquetas e discutimos a sua relação com o trabalho desenvolvido nesta dissertação, abordando ainda as principais estratégias para avaliação dos algoritmos de recomendação de etiquetas usados neste trabalho. Por último, apresentamos um pequeno resumo das diferentes técnicas de *Learning to Rank*.

2.1 Perfis de Especialidade e Etiquetagem de Pessoas

A construção de perfis de especialidade é uma etapa essencial para qualquer sistema de busca ou recuperação de especialistas [Balog et al., 2012]. O tema tem atraído a atenção de alguns pesquisadores como, por exemplo, Balog & de Rijke [2007] e de Rijke et al. [2010], mas o interesse persiste com trabalhos recentes dedicados a aspectos temporais da especialidade [Fang & Godavarthy, 2014a; Rybak et al., 2014a].

Nesta dissertação, abordamos o problema de construção de perfis de especialidade como sendo um problema de etiquetagem de pessoas [Serdyukov et al., 2011]. Essa solução tem várias vantagens, como: (i) representação concisa e uniforme do perfil, (ii) existência de métodos em outros domínios para sugerir um conjunto ordenado de etiquetas candidatas, (iii) capacidade de geração de diferentes "nuvens de etiquetas" para

diferentes períodos na carreira de um pesquisador, (iv) tratamento uniforme das fontes de informação textual como "conjuntos de etiquetas candidatas", etc. Essas vantagens podem ajudar os métodos de etiquetagem de pessoas a sobrepor algumas das dificuldades inerentes à construção de perfis de especialidade mencionadas no Capítulo 1. No entanto, a etiquetagem de pessoas é um problema muito menos investigado do que a recomendação de etiquetas para outros tipos de "objeto".

A abordagem mais próxima à nossa é apresentada por Serdyukov et al. [2011] que propuseram um método de etiquetagem de pessoas para construção de perfis de especialidade em um domínio empresarial. Enquanto nosso objetivo primário é a aplicação de técnicas de recomendação de etiquetas para a tarefa de construção de perfis de especialidade, há diferenças fundamentais entre o nosso trabalho e o trabalho proposto por esses autores. Primeiramente, nosso domínio é o meio acadêmico, significando que a evidência que exploramos é completamente diferente das evidências exploradas por eles. De fato, enquanto eles coletam características como documentos web, listas de discussão e *logs* de cliques em páginas de busca de uma empresa, exploramos características que são específicas de documentos científicos, como título, resumo e palavras-chave. Segundo, na avaliação que eles conduziram, foram contrastadas as etiquetas recomendadas com os perfis criados pelos empregados participantes do experimento realizado. Se uma etiqueta não estivesse no perfil criado pelos empregados, a etiqueta era automaticamente considerada irrelevante. Essa forma rigorosa de avaliação pode explicar os números baixos de desempenho relatados pelos autores. Em contraste, empregamos uma abordagem de *pooling* para reunir julgamentos de relevância para as etiquetas sugeridas por nove recomendadores diferentes (três recomendadores de etiquetas aplicados a três fontes de evidência textual distintas). Finalmente, Serdyukov et al. [2011] trataram a recomendação de etiquetas como um problema de classificação, usando um classificador de regressão logística para determinar se uma etiqueta era relevante ou não, com a confiança dessa classificação usada para induzir uma ordenação geral. Por outro lado, tratamos essa tarefa como uma tarefa de *Learning to Rank*, contrastando nove algoritmos do estado-da-arte usados para agregar os escores produzidos por nossos nove recomendadores de etiquetas considerados como características de aprendizado.

2.2 Sumarização Automática de Texto

Também podemos citar a sumarização automática de textos como um tópico relacionado à extração de etiquetas. O objetivo dessa técnica é criar um novo texto a partir de um ou mais textos existentes e que contenha a mesma informação do texto originail, ou

bem próximo disso. Em geral, um resumo não deve ultrapassar a metade do tamanho do texto original.

Hovy & Lin [1998] desenvolveram um algoritmo de sumarização automática de texto multi-lingual que chamaram de SUMMARIST. O algoritmo utiliza técnicas de processamento de linguagem natural e um *thesaurus* derivado da WordNet[1], aumentado por meio de dicionários e fontes similares. Os resultados mostraram que o algoritmo conseguiu resumos bem próximos daqueles elaborados por humanos.

Lin & Hovy [2000] propuseram a criação automática de "assinaturas de tópicos", ou seja, construções que poderiam ser utilizadas para identificar a presença de "conceitos complexos". Esses conceitos constituem vários componentes que se relacionam de maneira fixa e podem ser explorados para criação automática de resumos. Conroy & O'leary [2001] apresentaram uma abordagem para geração automática de resumos utilizando um modelo oculto de Markov[2]. O objetivo desse modelo é definir a probabilidade de uma dada sentença no texto pertencer ao resumo ou não.

No nosso caso, as técnicas de sumarização automática poderiam ser aplicadas em uma extensão do trabalho desenvolvido nesta dissertação, por exemplo, para extrair informações adicionais do corpo do texto. A sumarização ajudaria a manter o foco nas informações mais importantes do texto, diminuindo assim o tempo de processamento na extração de n-gramas e diminuindo também o ruído inerente a fontes de dados textuais muito verbosas.

Com relação a extração de n-gramas, que é uma técnica usual para representação de fontes textuais, podemos citar, por exemplo, o trabalho de Nascimento et al. [2011], que desenvolveram um sistema de recomendação de artigos científicos que utiliza fontes de dados disponíveis na Web para buscar por artigos similares a um artigo dado como entrada. Eles utilizaram n-gramas e sintagmas nominais[3] do título, resumo e corpo do texto do artigo de entrada, gerando assim consultas candidatas, que foram ordenados por relevância seguindo métricas definidas no artigo. Somente as consultas mais bem pontuadas eram submetidas às fontes de dados. Os artigos candidatos retornados por essas consultas eram ordenados por similaridade de acordo com o artigo de entrada do usuário. Nas avaliações feitas pelos autores, a consulta formada por n-gramas gerou resultados superiores à consulta obtida usando sintagmas nominais. Focando a análise nos tipos de campos do texto, foi feito um estudo utilizando cada um deles separadamente e suas combinações. Chegou-se à conclusão de que, embora o corpo do texto gere consultas que levam a uma revocação mais alta, a combinação de título e

[1] `http://wordnet.princeton.edu`
[2] Do inglês *hidden markov models*.
[3] Do inglês *noum phrases*.

resumo foi a que obteve a melhor relação custo-benefício pelo fato de gerar consultas com boa revocação e ser geralmente a informação mais disponível na Web, pois o corpo do artigo necessita do texto completo, que nem sempre está disponível.

A não-disponibilidade em 100% dos casos dos textos completos dos artigos foi um dos principais motivos pelos quais optamos por usar os títulos, resumos e palavras-chave em vez do texto completo dos artigos como fontes de evidência em nossos experimentos. Nascimento et al. [2011] corroboram nossas escolhas em relação ao uso de n-gramas e fontes textuais.

2.3 Recomendação de Etiquetas

Abordagens para recomendação de etiquetas têm sido propostas para uma grande variedade de tipos de mídia, principalmente na chamada Web 2.0. Como exemplos podemos citar trabalhos que exploram recomendação de etiquetas para redes sociais [Garg & Weber, 2008; Sigurbjörnsson & van Zwol, 2008a; Heymann et al., 2008], trabalhos que exploram novidade e diversidade em recomendação de etiquetas [Belém et al., 2011; Belém et al., 2013], trabalhos que exploram recomendação de etiquetas utilizando múltiplas características [Belém et al., 2014], que avaliam aplicação de *learning to rank* para recomendação de etiquetas [Canuto et al., 2013] e trabalhos que buscam evidências de qualidade em diversas características (*features*) textuais na Web [Figueiredo et al., 2009].

As técnicas do estado-da-arte para recomendação de etiquetas exploram padrões de coocorrência com etiquetas previamente selecionadas, expandindo assim um conjunto inicial de etiquetas I_O para um objeto O com outras etiquetas que coocorrem com as etiquetas de I_O, em diferentes objetos da coleção. As etiquetas mais relevantes podem então ser usadas para representar o objeto O. Canuto et al. [2013] compararam diversas abordagens de L2R aplicadas à tarefa de recomendação de etiquetas. O problema com essas abordagens é que elas precisam partir de um conjunto I_O que não esteja vazio. Quando não há etiquetas inicialmente disponíveis, como é o caso em muitos cenários de etiquetagem de pessoas, esses métodos não funcionam de forma adequada [Martins et al., 2016].

Outros trabalhos exploram conexões entre objetos que já possuem etiquetas e objetos ainda sem etiquetas [Lin et al., 2012; Siersdorfer et al., 2009; Song et al., 2011; Yin et al., 2013]. Isso pode ser visto como um tipo de abordagem baseada em filtragem colaborativa [Ekstrand et al., 2011]. Em contraste, neste trabalho adotamos uma abordagem puramente baseada em conteúdo, por explorar somente evidências

geralmente disponíveis nos conteúdos das publicações escritas pelos pesquisadores aos quais atribuiremos as etiquetas.

2.4 Avaliação de Nuvens de Etiquetas e de Recomendadores de Etiquetas

Um trabalho relevante para o nosso, ainda no contexto de atribuição de etiquetas a objetos, foi proposto por Venetis et al. [2011]. Em particular, eles tratam nuvens de etiquetas como uma lista ordenada de etiquetas e definem uma série de métricas que capturam propriedades estruturais dessas nuvens. Por exemplo, a cobertura de uma nuvem de etiquetas representa a fração de todos os objetos que podem ser recuperados pelas etiquetas pertencentes à nuvem.

Baseados nas métricas propostas, eles desenvolveram um modelo de satisfação para avaliar a qualidade de uma nuvem de etiquetas para uma tarefa específica de busca. Esse modelo considera a probabilidade de as etiquetas na nuvem falharem em satisfazer a necessidade de busca de um usuário hipotético por um documento da coleção. Usando esse modelo, eles realizaram uma análise quantitativa de diferentes algoritmos para seleção de etiquetas. Dentre os algoritmos analisados, o que apresentou o melhor desempenho foi usado como recomendador de etiquetas em nossa análise.

A avaliação de recomendadores de etiquetas é um campo de pesquisa por si só. Muitos dos trabalhos anteriores se baseiam em um processo de avaliação automática no qual uma parte (usualmente 50%) das etiquetas previamente assinaladas a um objeto do sistema é usada como treinamento enquanto as etiquetas remanescentes são usadas como gabarito do que deve ser predito pelo recomendador. Isso se deve em grande parte às dificuldades inerentes e ao custo associado com uma avaliação manual por parte dos usuários. Além disso, é geralmente difícil para um avaliador julgar se uma etiqueta assinalada por outro usuário (por exemplo, um usuário que não fez a carga do objeto ou não está familiarizado com ele) é relevante para um objeto cuja interpretação pode ser subjetiva (por exemplo, uma imagem ou um vídeo).

O problema com esse tipo de avaliação é que uma possível etiqueta relevante pode ser recomendada e não ser considerada relevante por não estar no conjunto usado como gabarito. Dadas as dificuldades mencionadas, apenas alguns trabalhos (por exemplo, Bi & Cho [2013]; Prokofyev et al. [2012]; Siersdorfer et al. [2009]; Sigurbjörnsson & van Zwol [2008b]; Wu et al. [2009]) avaliaram recomendadores de etiquetas usando avaliações manuais de usuários, geralmente feitas em pequena escala.

Em comparação, nesta dissertação apresentamos um estudo em larga escala en-

volvendo 1.288 respondentes (de 5.355 contatados) dentre os mais proeminentes pesquisadores em diferentes áreas do conhecimento atualmente trabalhando no Brasil. Além disso, esses respondentes podem ser considerados como os avaliadores ideais, pois eles avaliaram amostras de suas próprias produções científicas ao longo dos anos.

2.5 Evolução Temporal dos Perfis de Especialidade

Evolução de perfis de especialidade é um tópico relativamente novo e ainda pouco explorado, apesar de sua importância. Um exemplo disso é que se encontram poucos trabalhos na literatura tratando especificamente desse tema. Em nosso caso, usamos uma abordagem totalmente automática para geração dos perfis de especialidade e não usamos qualquer taxonomia para nos guiar no processo de descoberta de mudanças temporais no perfil de especialidade dos pesquisadores estudados.

O trabalho de Fang & Godavarthy [2014b], bem parecido com o cenário que estudamos, tem como objetivo modelar as dinâmicas da busca de especialistas. Eles aplicaram a metodologia no cenário acadêmico, em que os pesquisadores são os especialistas e suas publicações indicam sua especialidade. Os autores usaram as palavras-chaves dos artigos para definir os tópicos de especialidade dos pesquisadores. Todas as publicações associadas com uma devida palavra-chave definem a "área" de atuação desse pesquisador.

Para entender essa dinâmica, os autores analisaram como as áreas de atuação de um determinado pesquisador evoluem com o tempo. Eles assumem uma premissa markoviana de que as áreas de atuação no ano $t+1$, denotadas por a_{t_1}, dependem exclusivamente das áreas de atuação no ano t, denotadas por a_t. Eles definem então a probabilidade do especialista trabalhar na área a_{t+1} no ano $t+1$ como:

$$P(a_{t+1}|e) \;=\; \sum_{a_t} P(a_{t+1}|a_t, e)P(a_t|e), \tag{2.1}$$

em que $P(a_t|e)$ é a probabilidade de que a área a_t seja a área de especialidade do pesquisador no ano t e $P(a_{t+1}|a_t, e)$ é a probabilidade de que ele vá publicar na área a_{t+1}, dado que sua área corrente é a_t. A estimativa de $P(a_t|e)$ pode ser baseada na frequência relativa de a_t nas publicações do ano t do pesquisador. Especificamente, $P(a_t|e) = \frac{N_{a_t,e}}{N_e,t}$, onde (N_{a_t}, e) define o número de vezes que a_t ocorre nas publicações de e no ano t, e (N_e, t) o número total de vezes que qualquer uma das áreas ocorre nas publicações de e no ano t. A estimativa de $P(a_{t+1}|a_t, e)$ é o componente central discutido nesse trabalho, que caracteriza como o pesquisador e escolhe a próxima área

a_{t+1}, dado que sua área atual é a_t. Os autores consideram três fatores para estimar $P(a_{t+1}|a_t, e)$: (1) a personalidade do pesquisador para explorar uma nova área, ou seu conservadorismo em permanecer na mesma área; (2) a similaridade entre a nova área e as áreas atuais do pesquisador; e (3) a popularidade da nova área.

Em contraste, temos o trabalho de Rybak et al. [2014b] que focaram na tarefa de sumarização de especialistas com o objetivo de identificar e caracterizar mudanças na especialidade de indivíduos ao passar do tempo. De acordo com os autores, eles são os primeiros a propor essa tarefa. Para alcançar seu objetivo, eles primeiramente introduzem o conceito de perfil de especialidade hierárquica, onde áreas são organizadas em hierarquias e a especialidade é representada como sendo uma árvore ponderada. O perfil de especialidade temporal é então definido como uma série de perfis de especialidade hierárquicos divididos e organizados temporalmente. Em seguida, os autores desenvolveram métodos de detecção e caracterização de mudanças nos perfis dos especialistas. A ideia principal está na identificação dos chamados "nodos de foco", isto é, um nodo ou uma pequena quantidade de nodos que acumulam a grande maioria dos pesos, com relação a um nodo pai. Uma mudança ocorre se há uma diferença no conjunto de nodos de foco entre dois pontos no tempo. A mudança é então interpretada dependendo de qual nível da hierarquia de tópicos é afetado.

Diferente de Fang & Godavarthy [2014b], Rybak et al. [2014b] usam taxonomias pré-definidas pela ACM como entrada para classificação dos documentos em áreas nas quais o pesquisador publica e que portanto, assume-se que seja especialista. Os autores detectam então mudanças na estrutura dessa árvore de especialidades ao longo do tempo. O trabalho de Fang & Godavarthy [2014b] é bem mais parecido com o nosso, pois usa palavras-chave dos artigos científicos que, conforme demonstraremos no Capítulo 4, constituem uma fonte eficaz de informação para se extrair tópicos de interesse dos pesquisadores.

2.6 Learning to Rank

Nesta dissertação, abordamos o problema de sumarização de especialidades como um problema de *ranking*, uma nova tendência em recuperação de informação, particularmente em buscas na Web [Santos, 2013], e que visa aplicar técnicas de aprendizado de máquina para construir automaticamente um modelo de *ranking*. Isso é motivado por vários aspectos. Em buscas na Web há muitas evidências que podem representar relevância como, por exemplo, os textos âncora de um documento e o valor do seu *PageRank*[Brin & Page, 2012]. Incorporar essas informações ao modelo de *ranking* usando

aprendizado de máquina se tornou uma escolha natural [Hang, 2011]. Nas máquinas de busca, uma grande quantidade de *logs*, como dados de cliques em documentos, é acumulada. Isso torna possível derivar dados de treinamento e automaticamente criar um modelo de *ranking*. De fato, *Learning to Rank*(L2R) se tornou uma das tecnologias chave para as técnicas de busca modernas [Hang, 2011].

Como qualquer tarefa de aprendizado supervisionado ou semi-supervisionado, a técnica de *learning to rank* requer alguma forma de treinamento [Santos, 2013]. Como ilustrado na Figura 2.1, os dados de treinamento compõem uma amostra $\{(x_{ij}, y_{ij})\}_{j=1}^{n_{q_i}}$ para cada consulta q_i, incluindo um representação vetorial das características $x_{i,j}$ e uma etiqueta de saída $y_{i,j}$ para cada um dos n_{q_i} documentos retornados mais acima no *ranking*, dada a consulta q_i, usando algum dos algoritmos que discutiremos mais à frente nesta seção. As amostras de treinamento são usadas pelo módulo de aprendizado para produzir uma função de *ranking* h com efetividade ótima nas consultas de treinamento, como mensurado por uma função de perda Δ. Para reduzir a possibilidade de que a função aprendida esteja super ajustada aos dados de treinamento e não generalize bem para consultas ainda não vistas, exemplos de validação separados podem ser usados para guiar o algoritmo de aprendizado. Finalmente, dada uma consulta de teste q com uma amostra $\{(x_j, ?)\}_{j=1}^{n_q}$ compartilhando o mesmo espaço de características com as amostras de validação e treinamento, o módulo de *ranking* aplica a função h aprendida para produzir a permutação idealmente mais efetiva dos documentos do conjunto inicial.

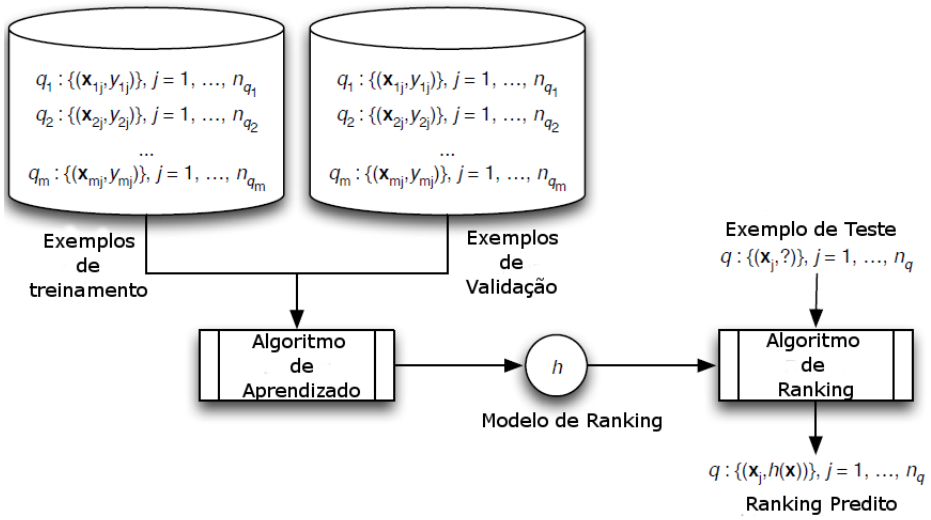


Figura 2.1: Arcabouço de aprendizado discriminativo.

Dependendo da escolha do espaço de entrada, do espaço de saída e do espaço de

hipóteses, bem como da função de perda, as abordagens de *learning to rank* podem ser classificadas em *pointwise*, *pairwise* ou *listwise* [Santos, 2013]. A abordagem *pointwise* considera o espaço de entrada compreendendo vetores de características construídos por documentos individuais e o espaço de saída compreendendo um único escore numérico para cada vetor de documento. Nesse caso, o processo é reduzido a uma tarefa de regressão básica, que é predizer o escore de relevância para cada par consulta-documento. Como resultado, vários algoritmos existentes podem ser aplicados à tarefa de *learning to rank*. Exemplos dessa abordagem são os algoritmos MART [Friedman, 2001] e Random Forests [Breiman, 2001].

Diferente da abordagem *pointwise*, a abordagem *pairwise* utiliza um espaço de entrada compreendendo pares de vetores de documentos e um espaço de saída compreendendo valores binários $\{-1, 1\}$, que denotam a preferência por um dos documentos no par sobre o outro. A abordagem *pairwise* minimiza o número médio de trocas no *ranking*. Exemplos dessa abordagem são os algoritmos RankNet [Burges et al., 2005], RankBoost [Freund et al., 2003] e LambdaRank [Quoc & Le, 2007].

Uma limitação das abordagens *pointwise* e *pairwise* é que elas ignoram o fato de que alguns (pares de) documentos são relacionados à mesma consulta. Para superar essa limitação, a abordagem *listwise* estende o espaço de entrada para incluir todos os exemplos para uma determinada consulta. Portanto, seu espaço de saída compreende também uma permutação completa dos exemplos de entrada, ou escores numéricos para todos eles. O espaço de saída também determina a função de perda. Em particular, se a saída é uma permutação, o erro de predição pode ser estimado pela diferença entre o gabarito e as permutações preditas. De outra forma, se existirem etiquetas no gabarito para todos os documentos, uma métrica padrão para avaliação pode ser usada para estimar o erro. Exemplos dessa abordagem são os algoritmos AdaRank [Xu & Li, 2007], LambdaMART [Wu et al., 2010], ListNet [Cao et al., 2007] e Coordinate Ascent [Metzler & Croft, 2007].

Capítulo 3

Metodologia Experimental

Abordagens de recomendação de etiquetas são tradicionalmente avaliadas particionando o conjunto existente de etiquetas em treino e teste. A avaliação da etiquetagem de pessoas impõe desafios adicionais, primeiramente porque as pessoas que recebem as etiquetas têm que concordar com a relevância das etiquetas que foram assinaladas para elas. A importância dessa aprovação é exacerbada quando as etiquetas têm um sentido de especialidade, que é o foco desta dissertação. Por outro lado, conduzir uma avaliação manual em larga escala é muitas vezes dispendioso em relação a tempo e recursos envolvidos no processo, o que de certa forma explica a escassez desse tipo de avaliação na literatura. Neste capítulo, descrevemos a metodologia experimental adotada em nosso trabalho. Em particular, discutimos a seleção de um conjunto representativo de especialistas, a aquisição das evidências de especialidade sobre eles, a geração dos perfis de especialidade candidatos a partir da recomendação de etiquetas e a avaliação desses perfis. A Figura 3.1 mostra uma visão geral de nossa metodologia.

3.1 Aquisição de Dados

O primeiro passo para produzir uma coleção de teste para perfis de especialidade é escolher um conjunto representativo de especialistas. Em 2008, o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) introduziu um programa de financiamento para fomentar redes de pesquisa colaborativa em várias áreas consideradas estratégicas para o país. Juntos, os 123 grupos de pesquisa contemplados, denominados de Institutos Nacionais de Ciência e Tecnologia (INCTs)[1], compreendem mais de 6000 dos mais proeminentes pesquisadores de todas as áreas do conhecimento trabalhando no Brasil. Nossa coleção de teste foi construída usando esses pesquisadores

[1] `http://goo.gl/FdjqRo`

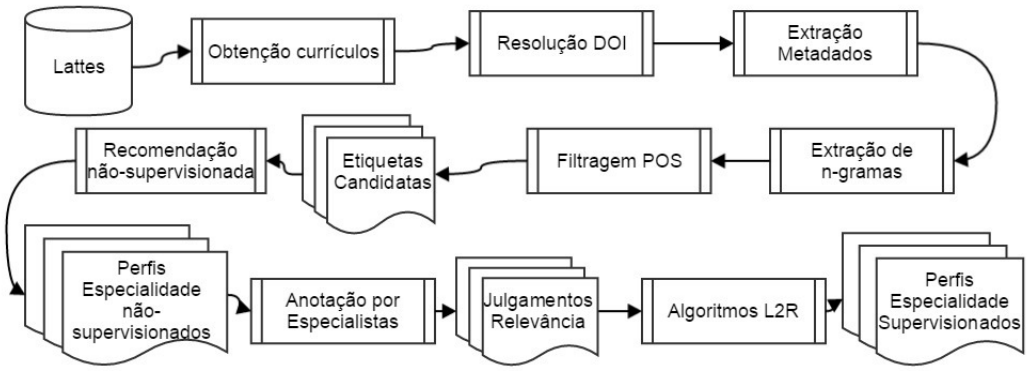


Figura 3.1: Visão geral de nossa metodologia experimental.

como um conjunto representativo de especialistas, para os quais construiremos os perfis de especialidade. Em particular, acreditamos que a amplitude de seus conhecimentos e suas carreiras heterogêneas apresentam desafios realistas para avaliação das nossas abordagens para geração de perfis de especialidade.

Tendo selecionado o conjunto de especialistas para a avaliação, precisamos coletar evidências de sua especialidade para serem exploradas e então produzir perfis de especialidade efetivos. Para isso, como discutido no Capítulo 1, recorremos à Plataforma Lattes, um repositório de dados acadêmicas de acesso público mantido pelo CNPq, que armazena currículos atualizados de pesquisadores que trabalham em instituições de pesquisa públicas e privadas. Em particular, dos mais de 6.000 pesquisadores em nosso grupo alvo, conseguimos coletar o currículo Lattes de 5.355 deles[2]. Para cada currículo coletado, extraímos dados sobre todas as publicações até abril de 2014. Para estabelecer uma base comum para os pesquisadores em diferentes áreas, nos concentramos em publicações de periódicos, que são geralmente vistos como o principal veículo para divulgação de pesquisa na maioria das áreas, e deixamos a exploração de outras fontes de evidência para trabalhos futuros.

Tendo coletado os currículos e extraído os títulos de todas as publicações em periódicos contidas neles, coletamos a seguir metadados adicionais correspondentes a cada publicação. Para as publicações sem o Identificador Digital do Objeto (DOI), realizamos uma busca com seus dados de citação utilizando a API do serviço CrossRef[3]. Uma vez descobertos os DOIs, o próximo passo foi coletar metadados utilizando os serviços providos por cada editora. Em particular, restringimo-nos às 20 editoras que concentram a maioria das publicações, correspondendo a mais de 80% do total de

[2]O currículo Lattes dos demais pesquisadores não pôde ser coletado devido a falhas persistentes ao serem baixados.

[3]`http://www.crossref.org`

artigos a serem coletados, de acordo com a Figura 3.2. Para cada publicação desses pesquisadores, extraímos o resumo e a lista de palavras-chave. Além do título extraído anteriormente dos currículos obtidos a partir da plataforma Lattes, esses metadados constituem três fontes de evidência textual para mineração de especialidades.

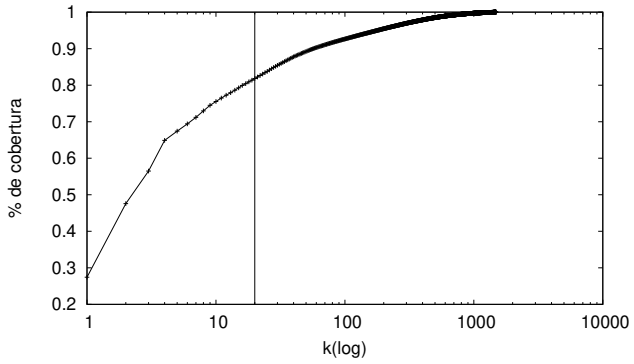


Figura 3.2: Número de publicações cobertas pelas k principais editoras.

3.2 Geração de Perfis Candidatos

Após adquirir metadados das publicações e a fim de construir diferentes fontes de evidência específicas, ou seja, título, resumo, e palavras-chave, o próximo passo foi então gerar um perfil de especialidade candidata para cada pesquisador. Para isso, extraímos n-gramas com tamanhos variados (de $n = 1$ a 3) a partir de cada uma das fonte de evidência textual. Os n-gramas extraídos representam as etiquetas do pesquisador.

A fim de descartar etiquetas improváveis de representar qualquer tópico de especialidade, realizamos um estudo piloto com voluntários, incluindo estudantes de pós-graduação e colaboradores de nosso grupo de pesquisa. Baseado em uma amostragem de 7.123 etiquetas, esse estudo piloto identificou 57 padrões gramaticais (POS[4]) [Manning & Schütze, 1999] comumente associados a etiquetas malformadas sintaticamente. Em particular, selecionamos padrões POS com um mínimo de 90% de confiança e um suporte mínimo de 15 etiquetas para maximizar precisão e revocação, foi estudada a variação desses parâmetros e os parâmetros que obtiveram os melhores resultados de precisão e revocação foram escolhidos. Em seguida, os padrões POS selecionados foram utilizados para filtrar etiquetas potencialmente malformadas em nosso conjunto de candidatas. Um exemplo de padrão POS seria NN-IN-VBG (i.e., substantivo, pre-

[4]Do inglês, *part-of-speach*.

posição, gerúndio), que removeu etiquetas como "proposal after executing" e "testing for validating".

Finalmente, a fim de evitar as limitações de um conjunto de avaliação de perfis criado por nós mesmos, como discutido na Seção 2, geramos um conjunto de diversas etiquetas para serem avaliadas por cada pesquisador. Para isso, escolhemos três algoritmos de recomendação de etiquetas baseados em conteúdo representativos, que são comumente utilizados na literatura [Venetis et al., 2011]. Nas seções 3.2.1 a 3.2.3 detalharemos cada um desses algoritmos.

Cada um dos algoritmos (TF,TFIDF e COV) foi aplicado a cada uma das três fontes de evidência consideradas (título, resumo e palavras-chave), produzindo assim nove recomendadores diferentes. Por último, reunimos o *ranking* produzido por cada um dos recomendadores considerando as 50 etiquetas mais bem pontuadas no intuito de produzir um conjunto final de etiquetas para serem avaliadas por cada pesquisador.

3.2.1 TF

Mostrar etiquetas populares nos perfis permite que usuários vejam o que o pesquisador mais produz em sua carreira. O algoritmo de recomendação de etiquetas TF [Venetis et al., 2011] retorna as etiquetas mais populares, dada uma métrica de popularidade e dada uma entrada que pode ser, por exemplo, a produção científica de um pesquisador. A métrica de popularidade adotada em nosso trabalho foi a frequência do termo (em inglês *term frequency*) que foi a mesma usada por Venetis et al. [2011] em seu algoritmo de extração de etiquetas por popularidade baseada em conteúdo. O Algoritmo 1 possui os principais detalhes dessa implementação.

Algoritmo 1: Pseudocódigo para o algoritmo TF.

```
/*E : Conjunto de etiquetas extraídas de todos os documentos do autor */
/*k : número de etiquetas no conjunto de saída */
Input: E, k
/*S: Conjunto de etiquetas escolhidas pelo algoritmo*/
Output: S ⊆ E
/*PQ é uma fila prioritária*/
PQ ← ∅, S ← ∅;
foreach e ∈ E do
    p ← tf(e);
    PQ.insert(e, p);
end
while (|S| ≥ k) ∧ (PQ.size() > 0) do
    S ← S ∪ PQ.top();
    PQ.top();
end
return S
```

3.2.2 TFIDF

Etiquetas muito frequentes, por outro lado, podem representar termos que não são discriminativos para a construção de um bom perfil de especialidade. Nesse caso, devemos usar um outro algoritmo que leva em consideração essa peculiaridade das coleções textuais. O TFIDF [Venetis et al., 2011], algoritmo tradicional de recomendação de etiquetas baseado em conteúdo, escolhe as k primeiras etiquetas de acordo com a pontuação TF×IDF, onde IDF é o inverso da frequência da etiqueta na coleção, que corresponde ao número de publicações onde a mesma ocorre. Esse algoritmo é similar ao TF, com a exceção de que o componente IDF auxilia a rebaixar a pontuação de etiquetas muito frequentes, favorecendo as mais discriminativas. O Algoritmo 2 possui os principais detalhes dessa implementação.

Algoritmo 2: Pseudocódigo para o algoritmo TFIDF.

```
/*E : Conjunto de etiquetas extraídas de todos os documentos do autor */
/*k : número de etiquetas no conjunto de saída */
Input: E, k
/*S: Conjunto de etiquetas escolhidas pelo algoritmo*/
Output: S ⊆ E
/*PQ é uma fila prioritária*/
PQ ← ∅, S ← ∅;
foreach e ∈ E do
    p ← tfidf(e);
    PQ.insert(e, p);
end
while (|S| ≥ k) ∧ (PQ.size() > 0) do
    S ← S ∪ PQ.top();
    PQ.top();
end
return S
```

3.2.3 COV

O algoritmo de maximização de cobertura (COV [Venetis et al., 2011]) tenta maximizar a cobertura do conjunto resultante de etiquetas com a restrição de que esse conjunto seja menor que um dado k. O algoritmo trata um problema clássico de cobertura máxima. Nosso universo de objetos consiste nos artigos do pesquisador que são representados pelas suas respectivas etiquetas. Nosso objetivo é encontrar um conjunto composto por no máximo k etiquetas que maximize o número de publicações que podem ser obtidas. Esse problema é conhecidamente NP-Completo [Venetis et al., 2011]. Entretanto, mesmo esse problema sendo NP-Completo, existem boas aproximações para ele [Venetis et al., 2011]. O algoritmo aqui implementado usa uma heurística gulosa para encontrar um conjunto com uma alta cobertura. A cada

iteração, o algoritmo adiciona uma etiqueta ao conjunto, que cobre o maior número de publicações não-cobertas até o momento. O algoritmo é finalizado depois de atingir k etiquetas ou não existirem mais etiquetas a serem adicionadas. O Algoritmo 3 possui os principais detalhes dessa implementação.

Algoritmo 3: Pseudocódigo para o algoritmo COV.

```
/*E : Conjunto de etiquetas extraídas de todos os documentos do autor */
/*k : número de etiquetas no conjunto de saída */
Input: E, k
/*S: Conjunto de etiquetas escolhidas pelo algoritmo*/
Output: S ⊆ E
/* V : conjunto de documentos cobertos pelas etiquetas de S */
V ← ∅, S ← ∅;
/*Maximizando cobertura de forma gulosa */
while (|S| ≥ k) ∧ (E ≠ ∅) do
    /* cov(e) : conjunto de documentos cobertos pela etiqueta e */
    ê ← arg max_{e∈E}{|V ∪ cov(e)|};
    S ← S ∪ ê;
    V ← V ∪ cov(e);
    E ← E − ê;
end
return S
```

3.3 Avaliação dos Perfis de Especialidade

O próximo passo em nossa metodologia de avaliação foi colher avaliações de relevantes para cada uma das etiquetas reunidas para cada pesquisador usando os nove recomendadores considerados. Para isso, convidamos os 5.355 pesquisadores para os quais produzimos um perfil candidato de especialidades para avaliar a relevância das etiquetas nesse perfil. Dos 5.355 pesquisadores contatados, 1.288 responderam ao nosso convite e participaram da avaliação em um período de duas semanas em Julho de 2014. A participação de aproximadamente 20% dos pesquisadores comprova a relevância do estudo para a própria comunidade acadêmica. A Figura 3.3 mostra a distribuição do número de participantes por área de conhecimento. A partir da figura, podemos notar que a taxa de resposta foi razoavelmente consistente entre todas as áreas, sendo a área de Ciências da Saúde — a maior comunidade em nosso estudo — a de maior participação.

A Figura 3.4 apresenta a tela da interface de avaliação de relevância. Durante a avaliação, a cada pesquisador foram apresentadas 60 etiquetas sem uma ordenação em particular, selecionadas por *round-robin* a partir do *ranking* de 50 etiquetas retornadas por cada um dos nove recomendadores reunidos. O pesquisador foi então solicitado a avaliar a relevância de cada etiqueta de acordo com a escala de quatro pontos a seguir:

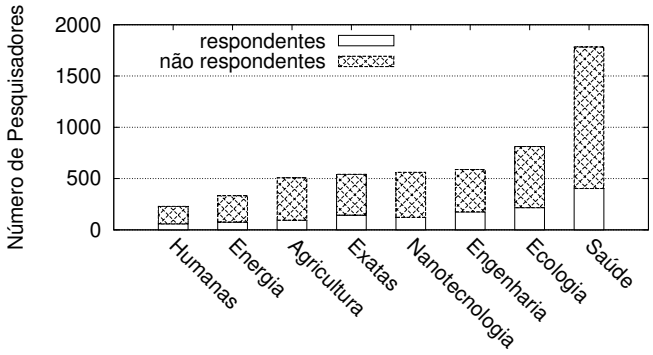


Figura 3.3: Número de pesquisadores por área de conhecimento

Dear Rodrygo Luis Teodoro Santos

As part of a research project of the National Institute of Science and Technology for the Web, we have developed new methods to generate a representative list of tags to describe the topics of expertise of researchers based on their scientific production. We would like to invite you to validate this list of topics for your particular case.

The tags listed below were automatically generated based on the publications available in our Lattes curriculum with the goal of describing the most representative topics of your research.

For each tag, please indicate one of the following options:

1. The tag **is malformed** (a spurious tag)
2. The tag is well-formed, but **is not relevant** to describe my work
3. The tag is well-formed, but **is only partially relevant** to describe my work
4. The tag is well-formed and **is highly relevant** to describe my work

Tag	Classification	Tag	Classification	Tag	Classification
learning-to-rank	○ 1 ○ 2 ○ 3 ○ 4	missing full text	○ 1 ○ 2 ○ 3 ○ 4	whens and hows	○ 1 ○ 2 ○ 3 ○ 4
web search engines	○ 1 ○ 2 ○ 3 ○ 4	query suggestions	○ 1 ○ 2 ○ 3 ○ 4	component-based software development	○ 1 ○ 2 ○ 3 ○ 4
expert search	○ 1 ○ 2 ○ 3 ○ 4	note that existing	○ 1 ○ 2 ○ 3 ○ 4	search	○ 1 ○ 2 ○ 3 ○ 4
search result diversification	○ 1 ○ 2 ○ 3 ○ 4	wizard tool	○ 1 ○ 2 ○ 3 ○ 4	query-dependent	○ 1 ○ 2 ○ 3 ○ 4
mimicking web search	○ 1 ○ 2 ○ 3 ○ 4	documents with respect	○ 1 ○ 2 ○ 3 ○ 4	web services	○ 1 ○ 2 ○ 3 ○ 4

Figura 3.4: Tela do sistema de avaliação de etiquetas.

1. “A etiqueta está malformada (uma etiqueta ilegítima)”

2. “A etiqueta está bem formada, mas não é relevante para descrever meu trabalho”

3. “A etiqueta está bem formada, mas é apenas parcialmente relevante para descrever meu trabalho”

4. “A etiqueta está bem formada e é altamente relevante para descrever meu trabalho”

A Figura 3.5 mostra a distribuição das etiquetas avaliadas pelos nove recomendadores de etiquetas utilizados. Pela figura, percebemos a alta incidência de etiquetas malformadas pelos recomendadores baseados nos resumos, mesmo com a filtragem POS realizada para remover tais etiquetas. Por outro lado, títulos produziram relativamente menos etiquetas malformadas, com palavras-chave possuindo a melhor performance nesse aspecto. Como iremos explorar no Capítulo 4, esse fato pode ser parcialmente

explicado pelas restrições de espaço inerentes a cada uma dessas fontes de evidência e o cuidado correspondente tomado pelo pesquisador ao criar cada uma delas.

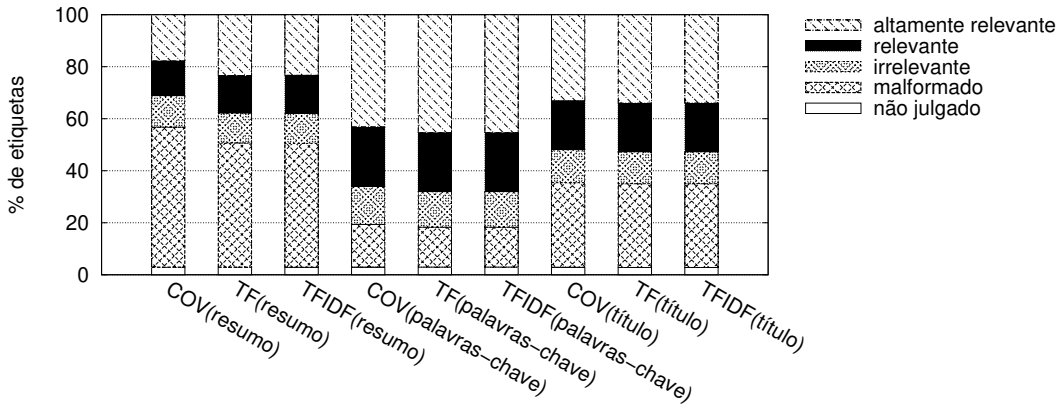


Figura 3.5: Distribuição de cada avaliação por recomendador.

3.4 Coleção de Teste

Como resultado da metodologia de avaliação descrita neste capítulo, produzimos uma coleção de teste para avaliação de técnicas para sumarização de perfis de especialidade. Nesta seção descreveremos algumas estatísticas gerais sobre os documentos da coleção e o gabarito produzido para avaliação.

3.4.1 Documentos da Coleção

Explorando algumas estatísticas da coleção, obtivemos o tamanho do vocabulário para cada um dos campos textuais que foram utilizados em nossa avaliação. Observando os resultados mostrados pela Tabela 3.1, podemos perceber como o vocabulário de palavras-chave é mais restrito que os demais. Enquanto títulos e resumos possuem em torno de 80.000 n-gramas únicos, palavras-chave incluem em torno de 40.000. Isso enfatiza ainda mais o cuidado dos pesquisadores ao escolherem os termos pertencentes a esse campo em seus artigos.

Campo	Valor
Títulos	80.271
Resumos	83.656
Palavras-chave	39.333

Tabela 3.1: Tamanho do vocabulário dos diferentes campos coletados.

Avançando mais nas estatísticas inerentes à coleção de dados, temos a Tabela 3.2. Observando os dados da tabela podemos perceber um comportamento interessante: o grande número de artigos com um único autor existente na coleção (ppua), tanto para convidados quanto para respondentes. A porcentagem de artigos tem relação direta com o número de colaborações entre os participantes de nossa pesquisa. Um número baixo indica que há grande sobreposição entre as publicações coletadas em nosso experimento, ou seja, há muitos coautores entre os pesquisadores participantes. Mas o que vemos é o comportamento contrário, ou seja, ou (1) os pesquisadores de nossa coleção publicam muitos artigos sem coautores, ou (2) colaboram com pesquisadores que estão fora de nosso universo de coleta. Como mostraremos mais à frente, são raros os artigos que são publicados com apenas um autor. Portanto podemos concluir que a maioria das colaborações é feita com autores fora do universo de nossa coleta (possivelmente estudantes de pesquisadores não participantes dos INCTs).

	convidados			participantes		
área	npes	ppua	nmppp	npes	ppua	nmppp
Agrárias	507	59,91%	43,06	94	79,24%	50,06
Ecologia	812	74,49%	25,35	216	85,17%	27,00
Energia	333	76,77%	29,26	76	90,61%	32,22
Engenharia	588	76,04%	24,26	176	85,24%	21,68
Exatas	542	76,75%	49,95	144	91,18%	56,13
Humanas	229	78,10%	15,31	58	85,94%	15,33
Nanotecnologia	561	68,26%	50,32	122	86,91%	58,75
Saúde	1783	66,47%	48,13	403	84,62%	52,03
Total	5355	65,48%	39,41	1289	84,33%	41,82

Tabela 3.2: Estatísticas da coleção de teste gerada, incluindo o número de pesquisadores (npes), a porcentagem de publicações com um único autor no corpus (ppua) e o número médio de publicações por pesquisadores (nmppp), tanto para convidados quanto para pesquisadores participantes.

Para corroborar a afirmativa (2) feita no parágrafo anterior, geramos dois gráficos sobre a distribuição do número de autores por artigo de nossa coleção. A Figura 3.6 mostra essa estatística considerando todos os pesquisadores convidados e a Figura 3.7 mostra essa estatística considerando apenas os pesquisadores respondentes de nossa pesquisa.

Observando a Figura 3.7, podemos perceber padrões claros de publicação entre algumas áreas. A área de Saúde tem um pico de publicações com cinco autores. Percebemos também as áreas de Ciências Exatas e Ecologia com curvas de autoria bem parecidas com pico em três autores. As áreas de Agrárias e Nanotecnologia têm pico em quatro autores, Energia e Engenharia em três autores e Humanas com uma curva

completamente diferente das demais com um pico de publicações com apenas um autor. Quando olhamos a Figura 3.6, observamos que apenas a área de Agrárias altera um pouco o seu comportamento, passando a ter o pico em cinco autores.

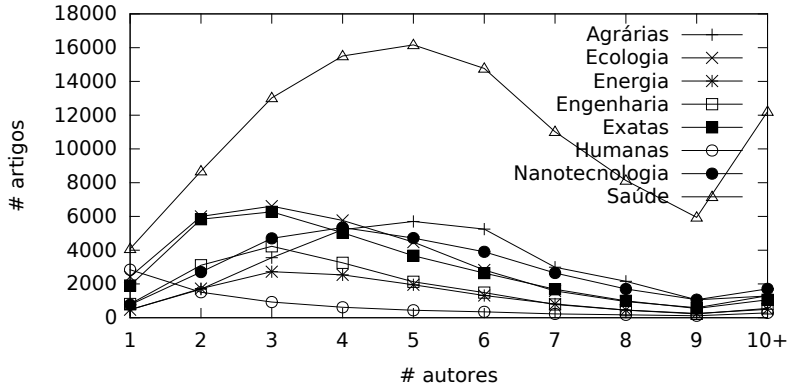


Figura 3.6: Distribuição do número de artigos por número de autores. Estatísticas coletadas entre todos os pesquisadores convidados para nossa pesquisa.

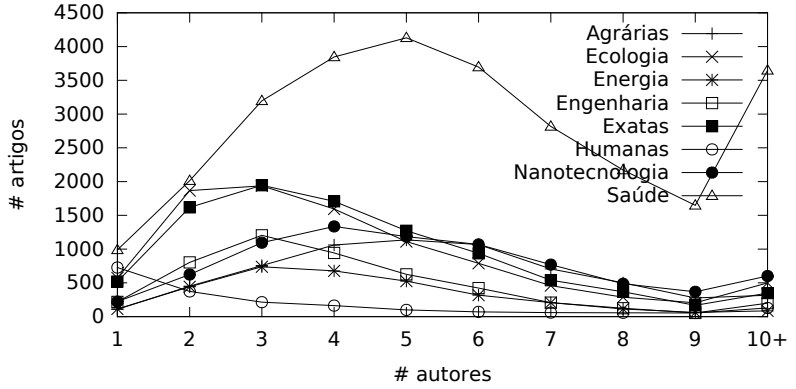


Figura 3.7: Distribuição do número de artigos por número de autores. Estatísticas coletadas entre os pesquisadores participantes de nossa pesquisa.

As Figuras 3.6 e 3.7 nos dizem que os pesquisadores efetivamente colaboraram na escrita da maioria dos artigos que coletamos. Resta-nos concluir se as colaborações são feitas com pesquisadores fora de nosso universo de coleta. Tal comportamento é de certa forma esperado, visto que uma publicação inclui autores com diversos níveis de especialidade e somente especialistas são contemplados em nossa coleção.

3.4.2 Gabarito

Em nossos experimentos consideramos cada pesquisador como uma "consulta". Os pesquisadores são representados pelas etiquetas extraídas de suas publicações. O gabarito,

ou conjunto de julgamentos de relevância, foi fornecido pelos próprios pesquisadores na interface construída para avaliação dos recomendadores de etiquetas. A cada pesquisador foi oferecido um total de 60 etiquetas recomendadas pelos algoritmos de *ranking* baseados em conteúdo. Cada pesquisador respondeu o questionário avaliando as etiquetas informadas de acordo com a escala de quatro níveis previamente mencionados (altamente relevante, relevante, irrelevante ou malformada).

A Figura 3.8 mostra a distribuição das etiquetas avaliadas pelos pesquisadores de cada uma das oito áreas do conhecimento consideradas em nosso estudo. A partir dessa figura, notamos primeiramente que a distribuição de etiquetas avaliadas de acordo com a escala de quatro níveis mencionada previamente é consistente dentre as diferentes áreas. Mais importante, notamos que, para todas as áreas do conhecimento, a proporção de etiquetas malformadas é mais baixa que nas demais classes obtidas de forma agrupada. Isso demonstra a adequação dos recomendadores escolhidos para identificar etiquetas potencialmente relevantes.

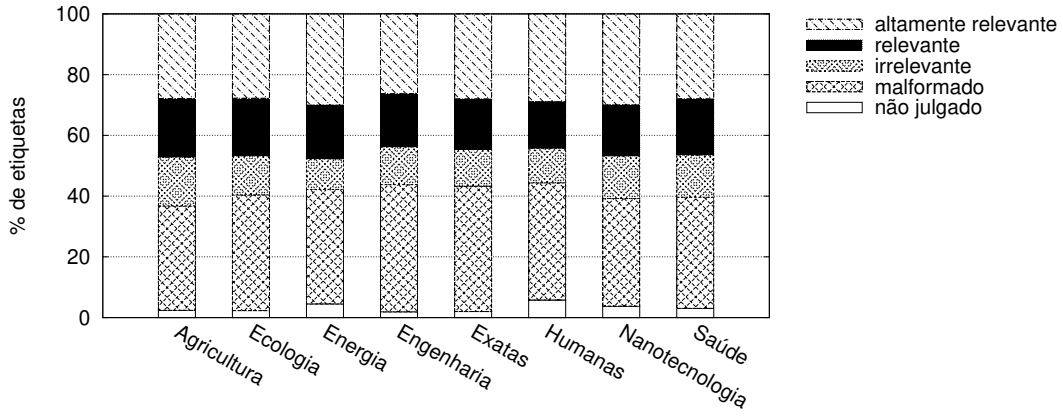


Figura 3.8: Distribuição das etiquetas avaliadas por área.

Avaliando também o tamanho do n-grama, podemos notar pela Figura 3.9 que a distribuição das respostas é desigual em relação aos vários tamanhos de n-gramas. O pior desempenho é obtido pelos n-gramas de tamanho 1, que têm o maior índice de etiquetas malformadas. Os n-gramas de tamanho 2 obtiveram o melhor resultado, com o maior número de etiquetas relevantes e altamente relevantes, classificadas pelos próprios pesquisadores. Os n-gramas de tamanho 3 obtiveram resultados intermediários, mas não menos expressivos, possuindo também um grande número de etiquetas relevantes e altamente relevantes. Os n-gramas de tamanho 4 e superior foram desconsiderados por serem inexpressivos comparados aos de tamanho 1, 2 e 3 respectivamente.

Estatísticas principais da coleção de teste estão resumidas na Tabela 3.3. A partir dessa tabela, vale a pena notar a representatividade dos dados da amostra obtida

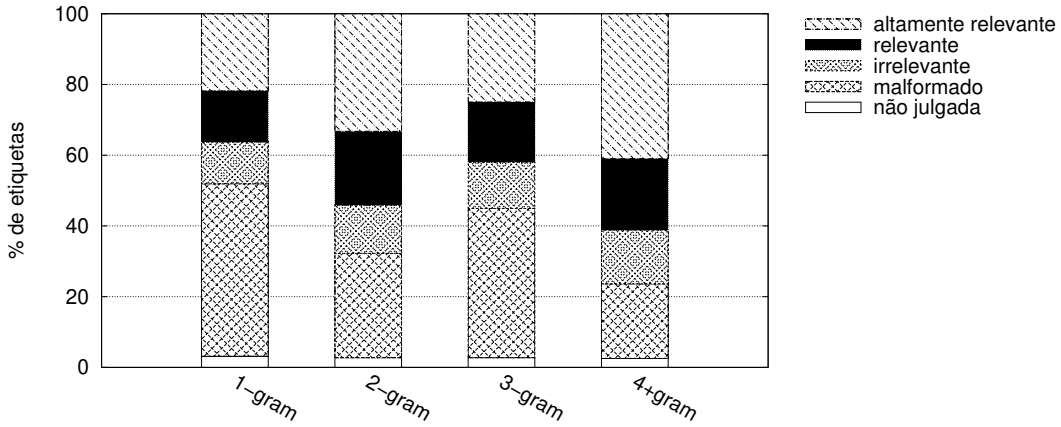


Figura 3.9: Distribuição de cada avaliação por tamanho do n-grama.

em relação ao conjunto total de pesquisadores convidados dos INCTs. Igualmente interessantes são as idiossincrasias de áreas específicas, particularmente em termos do número médio de publicações por pesquisador (nmppp), número médio de etiquetas consideradas parcialmente relevantes (ner) e número de etiquetas consideradas altamente relevantes (near) para representar a especialidade de cada pesquisador. Como discutiremos no Capítulo 4, esse conjunto de teste é usado para avaliar a relevância da recomendação de etiquetas como um mecanismo de geração automática de perfis de especialidades. Além disso, utilizamos as avaliações de relevância geradas como dados rotulados para aprender a classificar os perfis de especialidades.

	convidados		participantes			
área	npes	nmppp	npes	nmppp	ner	near
Agricultura	507	43,06	94	50,06	11,57	16,71
Ecologia	812	25,35	216	27,00	11,20	16,45
Energia	333	29,26	76	32,22	10,53	17,92
Engenharia	588	24,26	175	21,74	10,18	15,32
Exatas	542	49,95	144	56,13	9,85	16,65
Saúde	1.783	48,13	403	52,03	11,00	16,68
Humanas	229	15,31	58	15,33	8,69	16,36
Nanotecnologia	561	50,32	122	58,75	9,93	17,77
Total	5.355	39,41	1.288	41,85	10,60	16,62

Tabela 3.3: Estatísticas da coleção de teste gerada, incluindo o número de pesquisadores (npes), de publicações por pesquisador (nmppp), tanto para convidados quanto para pesquisadores participantes. Por último, também mostramos o número de etiquetas relevantes (ner) e altamente relevantes (near) por pesquisador.

Capítulo 4

Avaliação Experimental

Neste capítulo, analisamos minunciosamente a aplicação da recomendação de etiquetas ao problema de construção de perfis de especialidade. Avaliamos os métodos propostos quanto à correção, completude e robustez dos perfis gerados. Avaliamos também a aplicabilidade desses métodos no cenário de *Learning to Rank*

4.1 Questões de Pesquisa

Nossa avaliação procura responder às seguintes questões de pesquisa:

Q1. Quão efetiva é a recomendação de etiquetas como um mecanismo de construção automática de perfis de especialidade?

Q2. Quão completa é a recomendação de etiquetas como um mecanismo de construção automática de perfis de especialidade?

Q3. Quão robusta é a recomendação de etiquetas como um mecanismo de construção automática de perfis de especialidade?

Q4. Podemos efetivamente aprender a recomendar etiquetas de especialidade?

Os resultados de nossa análise são discutidos nas seções subsequentes, que abordam cada uma das questões por vez.

4.2 Efetividade dos Perfis

Para responder à questão Q1, avaliamos três algoritmos de recomendação de etiquetas baseados em conteúdo descritos na literatura, especificamente TFIDF, TF e COV [Ve-

netis et al., 2011]. Como discutido no Capítulo 3, cada algoritmo recebe evidências textuais de especialidade acadêmica, derivadas dos títulos, resumos e palavras-chave dos artigos dos pesquisadores e, ao final, produz um perfil de especialidade para o pesquisador. A Figura 4.1 sumariza a efetividade dos perfis de especialidade gerados pelas diferentes combinações de algoritmos de recomendação de etiquetas aplicados a três evidências textuais. Para cada combinação, reportamos o *discounted cumulative gain* normalizado (nDCG) [Baeza-Yates & Ribeiro-Neto, 2011] para múltiplos valores de k. O nDCG avalia o quanto a ordenação de um *ranking* está correta, os valores possíveis vão de 0 a 1, com o valor 1 significando uma ordenação perfeita dos resultados. O valor de k é variado para avaliar o algoritmo considerando apenas uma parte dos resultados, ou seja, apenas os k primeiros itens do *ranking* são considerados na avaliação. Para computar o nDCG, assinalamos as etiquetas parcialmente relevantes e altamente relevantes aos níveis de relevância 1 e 2, respectivamente. Etiquetas malformadas e irrelevantes foram ambas assinaladas com nível de relevância 0. Para fins de legibilidade, as barras de erro indicando o intervalo de confiança de 95% foram omitidas pois eram menores que os símbolos nos pontos do gráfico.

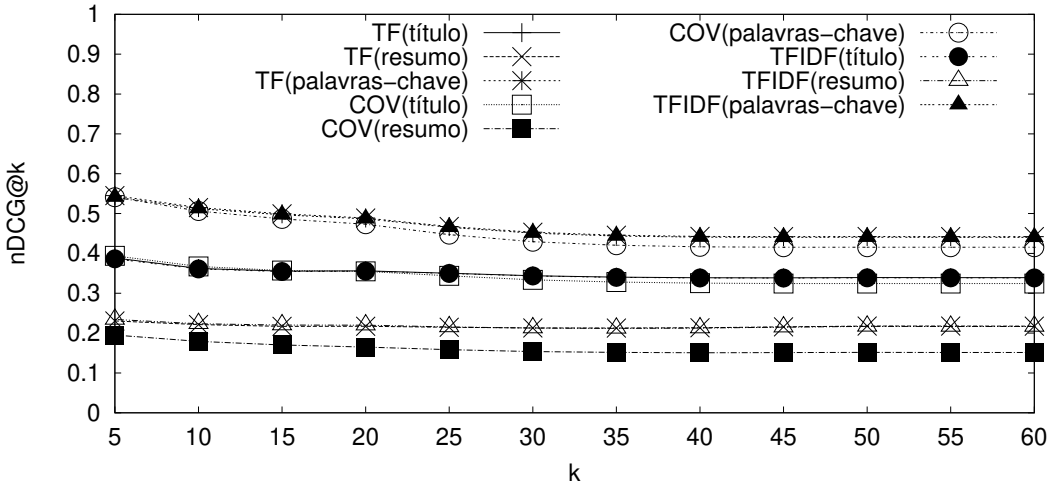


Figura 4.1: Efetividade dos perfis em termos do nDCG par várias posições k. Barras de erro omitidas por legibilidade.

Em relação aos diferentes algoritmos de recomendação de etiquetas considerados em nosso trabalho, pela Figura 4.1 notamos primeiramente que TFIDF se comporta similarmente ao algoritmo TF, o que de certa forma já era esperado. Em particular, como mostrado na Figura 4.3, perto de 80% de todas as etiquetas ocorrem somente uma vez e mais de 90% ocorrem pelo menos duas vezes em toda a coleção. Nesse cenário, com a maioria das etiquetas mostrando uma escarcidade similar, o componente IDF tem pouco impacto no *ranking* final produzido pelo algoritmo TFIDF. Como resultado,

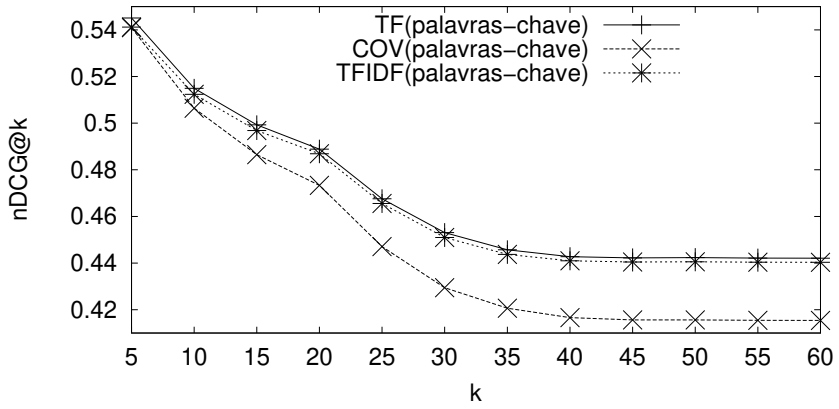


Figura 4.2: Mesmo gráfico da Figura 4.1 com foco nas combinações utilizando palavras-chave

o *ranking* produzido pelo algoritmo TFIDF se assemelha ao *ranking* produzido pelo algoritmo TF.

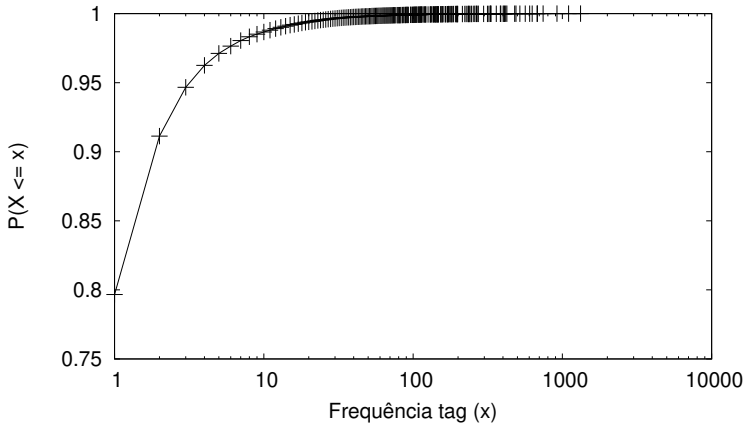


Figura 4.3: Distribuição cumulativa de frequência das etiquetas.

Pela Figura 4.1, notamos também que tanto TFIDF quanto TF têm resultados melhores que o algoritmo COV. Com o objetivo de explicar esse comportamento, a Tabela 4.1 mostra a sobreposição média par-a-par entre as 10 primeiras etiquetas obtidas pelo algoritmo TF, em termos do coeficiente de Jaccard, comparando as publicações onde essas etiquetas foram mencionadas. Pela tabela, observamos que as etiquetas do topo retornadas pelo algoritmo TF têm pouca sobreposição no que diz respeito às publicações que elas cobrem. De fato, a sobreposição média par-a-par das etiquetas do *ranking* gerado pelo algoritmo TF é em torno de 0,102805 no melhor caso, quando consideramos resumos como fonte de evidência. Como resultado, promover uma alta cobertura dessas publicações, como proposto pelo algoritmo COV, tem pouco benefício extra e pode inclusive piorar a efetividade dos recomendadores de etiquetas, como é o

caso quando se usam resumos. Além disso, o algoritmo COV é computacionalmente mais caro que os algoritmos TFIDF e TF, e portanto deve ser evitado em favor das outras abordagens mais rápidas, sempre que a eficácia for a principal preocupação ao se construir os perfis de especialidade.

evidência	Jaccard
resumo	0,102805 ± 0,009586
palavras-chave	0,097010 ± 0,007868
título	0,044094 ± 0,003939

Tabela 4.1: Coeficiente de Jaccard par-a-par médio para as 10 melhores etiquetas geradas pelo algoritmo TF usando diferentes evidências textuais.

Em relação a qual fonte textual é o indicador mais útil de especialidade, palavras-chave fornecem uma evidência de especialidade mais eficaz comparada com os títulos, que por sua vez são mais eficazes que os resumos. Essas observações são consistentes ao longo de toda a variação do nDCG na Figura 4.1 e sugere que, quanto mais restrita a fonte de evidência, mais útil ela é como indicador de especialidade. De fato, com espaço limitado para transmitir uma ideia, os pesquisadores precisam esforçar-se ao máximo para garantir que essa ideia seja cuidadosamente descrita. De modo recíproco, fontes de evidência menos restritas, como resumos, fornecem mais liberdade para geração de conteúdo, no entanto são mais suscetíveis a ruído. Como exemplo ilustrativo, a Tabela 4.2 mostra as cinco principais etiquetas geradas pelo algoritmo TF usando cada uma das fontes de evidência consideradas para produzir um perfil de especialidade para um pesquisador da área de recuperação de informação. Observando as etiquetas geradas, podemos notar que, enquanto etiquetas baseadas em palavras-chave são mais propensas a representar uma especialidade válida, como "web search" e "learning to rank", as outras fontes de evidência podem fornecer etiquetas razoavelmente boas e complementares às etiquetas de palavras-chave como, por exemplo, "search engines" e "weighting models".

Recordando a questão Q1, os resultados desta seção atestam a aplicabilidade de algoritmos tradicionais de recomendação de etiquetas baseadas em conteúdo para identificar tópicos relevantes de especialidade. Em particular, um recomendador baseado no algoritmo TF com etiquetas extraídas das palavras-chave das publicações gera os perfis de especialidade mais efetivos.

palavras-chave	título	resumo
web search	digital	weighting models
learning to rank	search result	learning to rank
relevance	learning to rank	combination
diversity	component-based	search result
digital libraries	search engines	ambiguous query

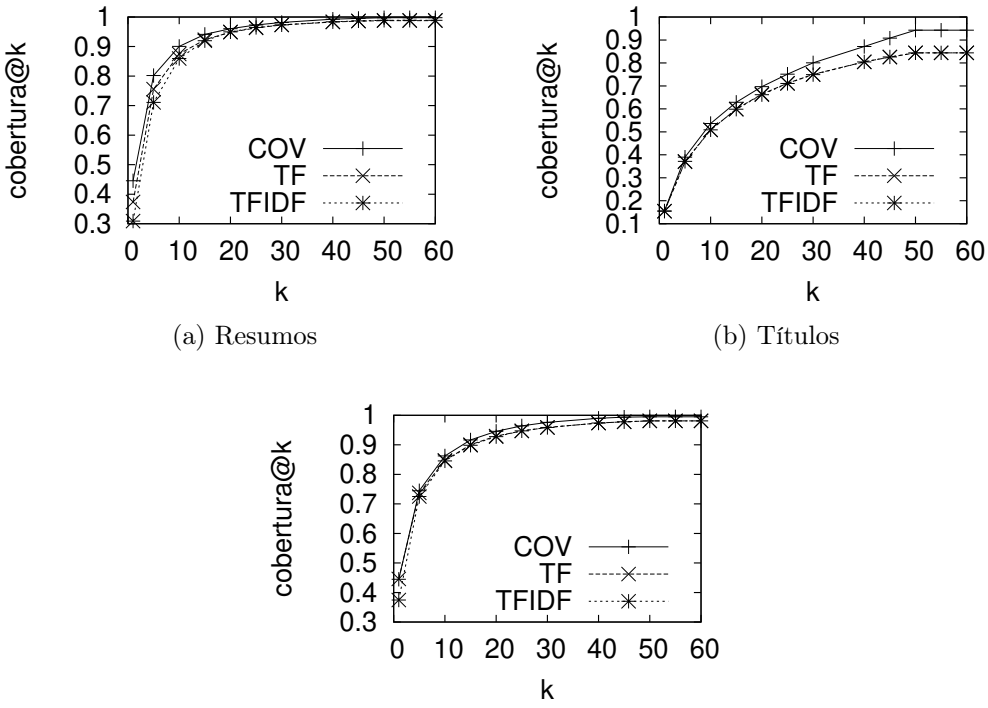
Tabela 4.2: Cinco etiquetas principais geradas pelo algoritmo TF usando, respectivamente, palavras-chave, título, e resumo para um pesquisador da área de recuperação de informação.

4.3 Completude dos Perfis

Os resultados da Seção 4.2 mostram que algoritmos tradicionais de recomendação de etiquetas baseados em conteúdo podem produzir perfis de especialidade eficazes. Uma questão natural que surge nesse cenário é quão completos são esses perfis. Em particular, para responder a questão Q2, analisamos a completude dos perfis de especialidade produzidos através da recomendação de etiquetas. Para avaliar a completude desses perfis, utilizamos a métrica de cobertura de etiquetas proposta por Venetis et al. [2011]. Em nosso caso, *cobertura@k* mede a fração de todas as publicações de um pesquisador que podem ser recuperadas pelas k principais etiquetas no seu perfil. As Figuras 4.4(a)-(c) mostram o valor de cobertura em diferentes posições k do *ranking* para todas as combinações consideradas de algoritmos de recomendação e evidências textuais. Novamente, barras de erro foram omitidas por legibilidade.

Pelas Figuras 4.4(a)-(c), observamos primeiramente que, com somente algumas das principais etiquetas, a maioria das combinações de recomendadores de etiquetas e evidências textuais é capaz de cobrir a maioria da produção científica dos pesquisadores. Em particular, perfis baseados nos resumos e nas palavras-chave têm cobertura@10 perto de 90%. Perfis baseados em títulos, por outro lado, precisam em torno de 50 etiquetas para conseguir o mesmo nível de cobertura. Em relação aos três algoritmos de recomendação considerados, COV, que diretamente procura aprimorar a cobertura, é particularmente o mais efetivo em posições mais altas do *ranking* quando aplicado aos resumos. Para títulos e palavras-chave, COV é mais eficaz em posições mais baixas no *ranking*, com ganhos estatisticamente significativos comparado tanto com TFIDF quanto com TF.

Voltando à questão Q2, os resultados nesta seção demonstram a completude dos perfis de especialidade construídos através da recomendação de etiquetas. De fato, todos os recomendadores considerados foram capazes de sumarizar de forma abrangente a especialidade dos pesquisadores com somente algumas das etiquetas do topo



(a) Resumos

(b) Títulos

(c) Palavras-Chave

Figura 4.4: Cobertura das publicações para múltiplos recomendadores de etiquetas e evidências textuais.

do *ranking*. O algoritmo COV, que procura promover etiquetas com um alto poder de cobertura das publicações de um pesquisador, mostrou-se particularmente vantajoso para essa finalidade.

4.4 Robustez dos Perfis

Os resultados apresentados até agora atestam a aplicabilidade dos algoritmos de recomendação de etiquetas para perfis de especialidade em termos da eficácia e completude das etiquetas recomendadas. Nesta seção, investigamos dois fatores que podem afetar esse desempenho. Em particular, para tratar a questão Q3, avaliamos a eficácia dos recomendadores de etiquetas para construir perfis de especialidade em diferentes áreas do conhecimento e em face de diferentes níveis de esparsidade das evidências textuais. Relativamente a essa última dimensão, a Figura 4.5 mostra a efetividade do perfil de todas as combinações consideradas de recomendadores e fontes de evidência em termos do nDCG@10 para pesquisadores com diferentes quantidades de publicações em periódicos.

Pela Figura 4.5, observamos que todas as abordagens têm um resultado razoa-

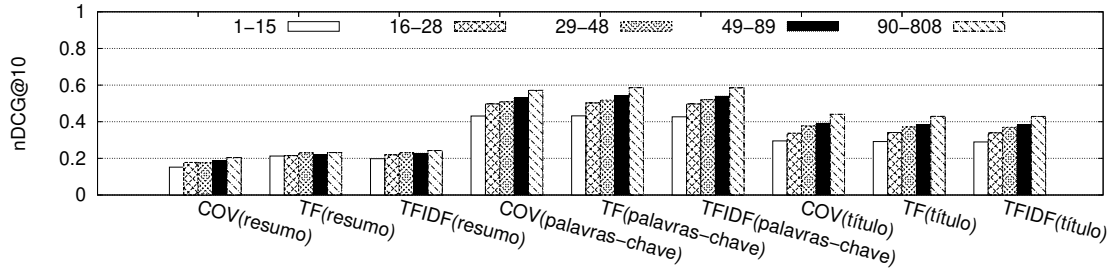


Figura 4.5: Eficácia do perfil para pesquisadores com diferentes volumes de publicações em periódicos.

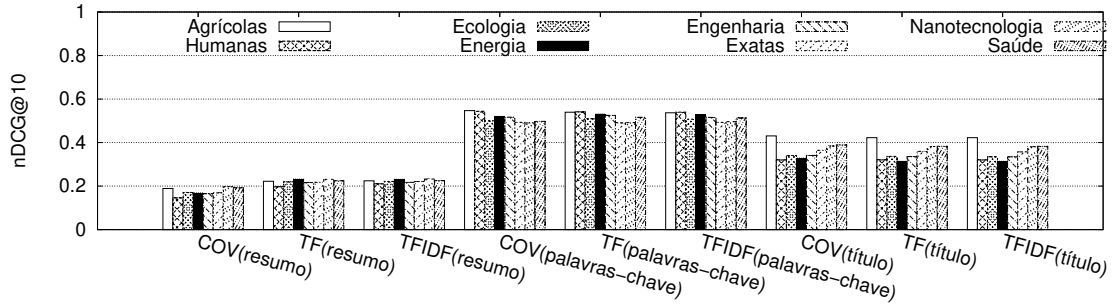


Figura 4.6: Eficácia do perfil para diferentes áreas do conhecimento.

velmente bom mesmo para pesquisadores com poucas publicações em seus currículos. Esse resultado revela a robustez de nossas abordagens quando aplicadas a evidências de especialidade esparsas. Não obstante, podemos notar que, quanto maior o currículo do pesquisador, mais eficazes são os perfis de especialidade produzidos. De fato, os currículos maiores tendem a oferecer espaço para cobrir temas de especialização distintos, bem como profundidade suficiente para enfatizar a importância de temas centrais. No entanto, currículos maiores podem também introduzir ruído no processo de recomendação de etiquetas. Esse é particularmente o caso para fontes de evidência mais verbosas, como resumos. Em particular, apesar de terem um desempenho inferior quando comparados a recomendadores baseados em títulos e palavras-chave, recomendadores baseados em resumos são menos influenciados por conjuntos de evidências mais esparsos.

Além de avaliar o impacto da quantidade de evidência de especialidade disponível, a Figura 4.6 mostra a efetividade do perfil gerado pelas já mencionadas combinações de algoritmo e evidências textuais para as oito áreas do conhecimento representadas em nossa coleção, como descritas no Capítulo 3. Pela Figura 4.6, podemos notar três perfis de eficácia claros entre as diferentes áreas do conhecimento, correspondendo às três diferentes fontes textuais de evidência empregadas. Em particular, resumos levam

a uma eficácia estável entre as diferentes áreas, enquanto palavras-chave e títulos são mais instáveis. Enquanto todas as áreas consideradas têm mais ou menos o mesmo número de etiquetas relevantes por pesquisador, como discutido no Capítulo 3, alguns resultados interessantes podem ser observados. Por exemplo, a área de Ciências Agrícolas geralmente mostra o melhor desempenho entre todas as áreas. A área de Saúde, que tem de longe o maior número de etiquetas relevantes bem como de pesquisadores em nossa coleção, tem um desempenho intermediário para todos os algoritmos e fontes de evidência textual considerados. Por último, a área de Humanas tem um dos piores desempenhos com resumos e títulos, mas um dos melhores com palavras-chave. Isso sugere que diferentes fontes de evidência podem prover benefícios complementares, como discutiremos na próxima seção.

Retomando à questão Q3, os resultados nesta seção demonstram a robustez dos recomendadores de etiquetas para construção de perfis de especialidade. Enquanto a maior disponibilidade de evidências de especialidade, como currículos maiores, aumenta a eficácia dos perfis produzidos, as combinações de algoritmos de recomendação e evidências textuais mantém-se robustas mesmo com dados esparsos. Diferentes áreas do conhecimento, por outro lado, têm impacto diferente na recomendação de etiquetas, no entanto sem comprometer visivelmente a eficácia atingida para nenhuma área em particular.

4.5 Aprendizado dos Perfis

Os resultados nas Seções 4.2 a 4.4 demonstram a eficácia, completude e robustez da recomendação de etiquetas para criação de perfis de especialidade, com as etiquetas baseadas em palavras-chave sendo particularmente as mais eficazes. A partir desses resultados, é possível argumentar que apenas palavras-chave são suficientes para gerar perfis de especialidade eficazes, dado que as palavras-chave são cuidadosamente fornecidas pelos próprios pesquisadores. Ainda assim, outras fontes de evidência podem prover etiquetas relevantes que não são mencionadas entre as palavras-chave das publicações dos pesquisadores. Nesta seção, avaliamos a complementariedade dessas fontes de evidência combinadas aos algoritmos de recomendação TFIDF, TF e COV. Para isso, abordamos a questão Q4 procurando aprender um modelo de *ranking* eficaz para perfis de especialidade.

Para abordar a questão Q4, avaliamos a aplicação dos recomendadores de etiquetas como características para o processo de *Learning to Rank*(L2R) ao criar os perfis de especialidade. Em particular, testamos nove algoritmos do estado-da-arte descritos

na literatura, como implementados pela biblioteca Ranklib[1]: MART [Friedman, 2001], RankNet [Burges et al., 2005], RankBoost [Freund et al., 2003], AdaRank [Xu & Li, 2007], Coordinate Ascent [Metzler & Croft, 2007], LambdaMART [Wu et al., 2010], LambdaRank [Quoc & Le, 2007], ListNet [Cao et al., 2007] e Random Forests [Breiman, 2001]. Para cada um desses algoritmos, conduzimos uma validação cruzada em cinco partições para otimizar o nDCG@10, com três partições usadas para treinamento, uma para validação e uma para teste. Na Figura 4.7, reportamos o nDCG@10 médio entre as partições de teste para cada um dos algoritmos L2R. Barras de erro indicam intervalo de confiança de 95% para as médias retornadas. Finalmente, uma linha horizontal indica a performance do recomendador de etiquetas mais eficaz usado isoladamente identificado na Figura 4.1, mais especificamente o recomendador TF(palavras-chave).

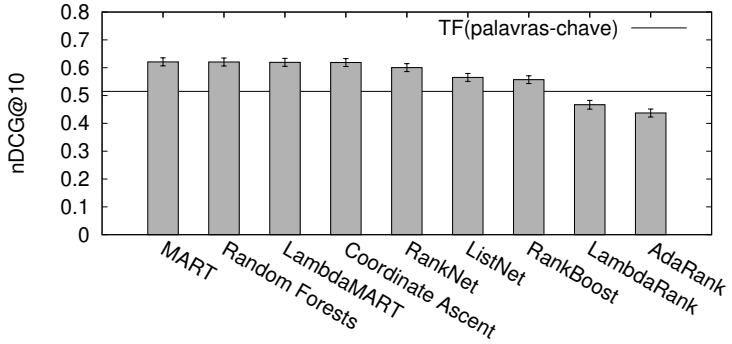


Figura 4.7: Eficácia dos perfis criados pelos algoritmos L2R.

Na Figura 4.7, observamos um empate estatístico entre MART, Random Forests, LambdaMART e Coordinate Ascent. Além disso, esses quatro algoritmos são estatisticamente superiores aos outros cinco (RankNet, Listnet, RankBoost, LambdaRank, and AdaRank). Também vale a pena notar que três dos quatro melhores algoritmos são baseados em combinação. Comparado com a nossa linha de base, o recomendador TF(palavras-chave), a maioria dos algoritmos L2R melhoram, muitas vezes de forma significativa, com ganhos de até 21,5% em termos de nDCG@10. Tal melhoria demonstra a complementariedade dos vários recomendadores considerados como atributos na tarefa de *learning to rank* conseguindo assim perfis de especialidade mais eficazes.

A fim de avaliar melhor a importância e o impacto de cada característica no desempenho dos modelos aprendidos, primeiramente ordenamos essas características pelo seu valor estimado de ganho de informação [Yang & Pedersen, 1997]. A Tabela 4.3 mostra os resultados dessa análise, listando todas as nove características consideradas em ordem crescente de ganho de informação, por exemplo, da menos informativa à

[1] `http://sourceforge.net/p/lemur/wiki/RankLib/`

mais informativa. Pela tabela, podemos ver que os resultados são consistentes com os resultados de eficácia avaliados anteriormente, com recomendadores baseados em palavras-chave tendo o maior valor de ganho de informação, que é substancialmente superior aos resultados das demais características.

Característica	Ganho de Informação
TFIDF(título)	0,007200
TF(título)	0,008010
COV(título)	0,008210
TFIDF(resumo)	0,008900
TF(resumo)	0,009730
COV(resumo)	0,023060
COV(palavras-chave)	0,045410
TFIDF(palavras-chave)	0,056220
TF(palavras-chave)	0,057830

Tabela 4.3: Ganho de informação dos características individuais.

Com a ordem imposta pelo ganho de informação computado em cada característica na Tabela 4.3, ainda avaliamos seu impacto no modelo de perfil de especialidade aprendido pelo MART, o algoritmo L2R que obteve o melhor desempenho na Figura 4.7. Em particular, começando com o modelo MART que usa todas as características, removemos uma característica por vez, da menos informativa à mais informativa. A cada passo de remoção, um novo modelo MART era aprendido baseado nas características restantes através de uma validação cruzada de cinco partições, até que tivéssemos uma única característica, especificamente, TF(palavras-chave). A Tabela 4.4 mostra os resultados desse experimento, em termos do nDCG@10 médio entre as partições de teste, com intervalo de confiança de 95%.

Característica	nDCG@10
todas as características	0,621862 ± 0,013161
-TFIDF(título)	0,621292 ± 0,013148
-TF(título)	0,619677 ± 0,013174
-COV(título)	0,612476 ± 0,013140
-TFIDF(resumo)	0,611220 ± 0,013126
-TF(resumo)	0,608421 ± 0,013274
-COV(resumo)	0,608024 ± 0,013301
-COV(palavras-chave)	0,605045 ± 0,013290
-TFIDF(palavras-chave)	0,511833 ± 0,012279

Tabela 4.4: Eficácia do perfil após remoção das características.

Como observado na Tabela 4.4, todas as características têm um impacto positivo no desempenho final do modelo aprendido, com as menos informativas causando perdas muito pequenas quando removidas. Isso é consistente com nossa análise anterior. No entanto, é interessante notar que se considerarmos apenas duas das melhores características para os algoritmos de L2R, TF(palavras-chave) e TFIDF(palavras-chave), os ganhos sobre a linha de base são ainda consideráveis, em torno de 18.2%, o que corresponde a aproximadamente 85% dos ganhos totais obtidos quando usamos as nove características juntas. Esse resultado tem impacto positivo em termos de "custo versus benefício" nas combinações L2R produzidas.

Os resultados desta seção demonstram a complementariedade dos recomendadores de etiquetas mencionados anteriormente para identificar tópicos de especialidade relevantes. Voltando à questão de pesquisa Q4, esses resultados mostram que uma eficácia ainda maior pode ser alcançada usando-se os recomendadores de etiquetas como características de um modelo L2R. Além disso, como demonstrado na análise de remoção de características, a combinação de um conjunto muito reduzido dessas features pode ainda levar a ganhos significativos em eficácia, enquanto potencialmente reduz o custo computacional gasto pela solução L2R.

Capítulo 5

Conclusões e Trabalhos Futuros

5.1 Conclusões

Nesta dissertação, investigamos a utilização de recomendação de etiquetas para a criação automática de perfis de especialidade, com um estudo de caso no domínio científico. Com esse objetivo, realizamos um estudo em larga escala com especialistas acadêmicos de diversas áreas do conhecimento trabalhando no Brasil, no intuito de produzir uma coleção de teste representativa para geração de perfis de especialidade. Com a coleção criada, avaliamos a efetividade, completude e robustez dos perfis de especialidade gerados por nove recomendadores diferentes, derivados de algoritmos representativos de recomendação de etiquetas baseados em conteúdo encontrados na literatura, aplicados a três diferentes fontes de evidência de especialidade: títulos das publicações, resumos e palavras-chave. Nossos experimentos demonstraram que a maioria das combinações produz resultados satisfatórios, sendo que os recomendadores baseados em TF e TFIDF se mostraram os melhores quando aplicados a palavras-chave. Diferentemente da maioria dos trabalhos anteriores [Fang & Godavarthy, 2014a; Rybak et al., 2014a; Serdyukov et al., 2011], nossas análises foram conduzidas com um grande número de especialistas reais que podem ser considerados como avaliadores ideais para a tarefa de criação automática de perfis de especialidade.

A baixa interseção dos resultados geradas pelos nove recomendadores também nos motivou a combiná-los utilizando estratégias de L2R. Para tanto, testamos nove algoritmos de L2R do estado da arte, sendo que os melhores produziram ganhos de mais de 20% em eficácia quando comparados ao melhor recomendador utilizado isoladamente. Finalmente, realizamos uma análise da evidência disponível para classificadores e descobrimos que a combinação de apenas algumas das melhores características produz a maior parte dos ganhos observáveis. Como resultado, as soluções de recomendação de

etiquetas investigadas aqui têm aplicação em potencial para cenários reais de sumarização de perfis de especialidade, como empresas, bem como serviços de redes sociais acadêmicos ou de negócios.

5.2 Trabalhos Futuros

Os resultados apresentados nesta dissertação abrem perspectivas para vários outros trabalhos. Uma extensão importante seria realizar uma análise dos perfis de especialidade dos pesquisadores ao longo do tempo. Nossa rica coleção poderia ser utilizada para entender melhor a evolução de tópicos de interesse em diferentes áreas do conhecimento. Outro trabalho interessante seria aprimorar a eficácia das recomendações de etiquetas geradas utilizando outras fontes de evidência de especialidade (por exemplo, redes de coautoria e de citações), bem como estratégias mais avançadas de filtragem de etiquetas malformadas.

Além disso, os resultados alcançados nesta dissertação podem ser utilizados para auxiliar a tarefa de busca de especialistas. O *feedback* sobre relevância das etiquetas fornecido pelos próprios pesquisadores é um recurso valioso para construir um sistema de busca de especialistas. Outra aplicação interessante de nossas técnicas seria um sistema de sugestão automática de etiquetas de especialidade em sistemas como ResearchGate[1] e Linkedin[2]. Essas etiquetas seriam utilizadas como sugestão para que usuários possam endossar outros usuários. Os documentos utilizados como base para a produção das etiquetas poderiam ser os artigos publicados pelos próprios usuários nas respectivas redes sociais.

[1] `https://www.researchgate.net/`
[2] `https://www.linkedin.com`

Referências Bibliográficas

Baeza-Yates, R. A. & Ribeiro-Neto, B. (2011). *Modern Information Retrieval*. Pearson Education Ltd., Harlow, UK, 2 edição.

Balog, K. & de Rijke, M. (2007). Determining expert profiles (with an application to expert finding). Em *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pp. 2657–2662.

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.

Belém, F.; Martins, E.; Pontes, T.; Almeida, J. & Gonçalves, M. (2011). Associative Tag Recommendation Exploiting Multiple Textual Features. Em *Proc. 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1033–1041.

Belém, F.; Santos, R. L. T.; Almeida, J. M. & Gonçalves, M. A. (2013). Topic diversity in tag recommendation. Em *Seventh ACM Conference on Recommender Systems, RecSys '13, Hong Kong, China, October 12-16, 2013*, pp. 141–148.

Belém, F. M.; Martins, E. F.; Almeida, J. M. & Gonçalves, M. A. (2014). Personalized and object-centered tag recommendation methods for web 2.0 applications. *Inf. Process. Manage.*, 50(4):524–553.

Bi, B. & Cho, J. (2013). Automatically generating descriptions for resources by tag modeling. Em *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, CIKM.

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

Brin, S. & Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks*, 56(18):3825–3833.

Burges, C.; Shaked, T.; Renshaw, E.; Lazier, A.; Deeds, M.; Hamilton, N. & Hullender, G. (2005). Learning to rank using gradient descent. Em *Proceedings of the 22nd International Conference on Machine Learning*, pp. 89–96. ACM.

Canuto, S.; Belém, F. M.; Almeida, J. M. & Gonçalves, M. A. (2013). A Comparative Study of Learning-to-Rank Techniques for Tag Recommendation. *Journal of Information and Data Management*, 4(3):453–468.

Cao, Z.; Qin, T.; Liu, T.-Y.; Tsai, M.-F. & Li, H. (2007). Learning to Rank: From Pairwise Approach to Listwise Approach. Em *Proceedings of the 24th International Conference on Machine Learning*, pp. 129–136. ACM.

Conroy, J. M. & O'leary, D. P. (2001). Text Summarization via Hidden Markov Models. Em *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 406–407.

de Rijke, M.; Balog, K.; Bogers, T. & van den Bosch, A. (2010). On the Evaluation of Entity Profiles. Em *Proceedings of the International Conference of the Cross-Language Evaluation Forum (CLEF 2010)*, pp. 94–99.

Ekstrand, M. D.; Riedl, J. T. & Konstan, J. A. (2011). Collaborative filtering recommender systems. *Found. Trends Hum.-Comput. Interact.*, 4(2):81–173.

Fang, Y. & Godavarthy, A. (2014a). Modeling the Dynamics of Personal Expertise. Em *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1107–1110.

Fang, Y. & Godavarthy, A. (2014b). Modeling the Dynamics of Personal Expertise. Em *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1107–1110.

Figueiredo, F.; Belém, F.; Pinto, H.; Almeida, J.; Gonçalves, M.; Fernandes, D.; Moura, E. & Cristo, M. (2009). Evidence of Quality of Textual Features on the Web 2.0. Em *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pp. 909–918.

Freund, Y.; Iyer, R.; Schapire, R. E. & Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *The Journal of machine learning research*, 4:933–969.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pp. 1189–1232.

Garg, N. & Weber, I. (2008). Personalized, Interactive Tag Recommendation for Flickr. Em *Proceedings of the 2008 ACM Conference on Recommender Systems*, pp. 67–74.

Hang, L. (2011). A short introduction to learning to rank. *IEICE Transactions on Information and Systems*, 94(10):1854–1862.

Heymann, P.; Ramage, D. & Garcia-Molina, H. (2008). Social Tag Prediction. Em *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 531–538.

Hovy, E. & Lin, C.-Y. (1998). Automated Text Summarization and the SUMMARIST System. Em *Proceedings of a Workshop on TIPSTER Text Program*, pp. 197–214.

Lane, J. (2010). Let's make science metrics more scientific. *Nature*, 464(7288):488–489.

Lin, C.-Y. & Hovy, E. (2000). The Automated Acquisition of Topic Signatures for Text Summarization. Em *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, pp. 495–501.

Lin, Z.; Ding, G.; Hu, M.; Wang, J. & Sun, J. (2012). Automatic Image Annotation Using Tag-Related Random Search Over Visual Neighbors. Em *Proc. 21st ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 1784–1788.

Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Martins, E. F.; Belém, F. M.; Almeida, J. M. A. & Gonçalves, M. A. (2016). On cold start for associative tag recommendation. *Journal of the Association for Information Science and Technology*, 67(1):83–105.

Metzler, D. & Croft, W. B. (2007). Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274.

Nascimento, C.; Laender, A. H.; da Silva, A. S. & Gonçalves, M. A. (2011). A Source Independent Framework for Research Paper Recommendation. Em *Proceedings of the 11th ACM/IEEE Joint Conference on Digital Libraries*, pp. 297–306.

Prokofyev, R.; Boyarsky, A.; Ruchayskiy, O.; Aberer, K.; Demartini, G. & Cudré-Mauroux, P. (2012). Tag Recommendation for Large-scale Ontology-based Information Systems. Em *Proceedings of the 11th International Conference on The Semantic Web - Volume Part II*, ISWC.

Quoc, C. & Le, V. (2007). Learning to rank with nonsmooth cost functions. *Neural Information Processing Systems*, 19:193.

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. Em *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pp. 189–198, Shanghai, China. ACM.

Rybak, J.; Balog, K. & Nørvåg, K. (2014a). Temporal expertise profiling. Em *Advances in Information Retrieval*, pp. 540–546. Springer.

Rybak, J.; Balog, K. & Nørvåg, K. (2014b). Temporal Expertise Profiling. Em *Proceedings of the 36th European Conference on Information Retrieval*, pp. 540–546.

Santos, R. L. T. (2013). *Explicit web search result diversification.* PhD thesis, School of Computing Science, University of Glasgow, Glasgow, UK.

Serdyukov, P.; Taylor, M.; Vinay, V.; Richardson, M. & White, R. W. (2011). Automatic People Tagging for Expertise Profiling in the Enterprise. Em *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pp. 399–410, Berlin, Heidelberg. Springer-Verlag.

Siersdorfer, S.; Pedro, J. S. & Sanderson, M. (2009). Automatic Video Tagging Using Content Redundancy. Em *Proc. 32nd International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 395–402.

Sigurbjörnsson, B. & van Zwol, R. (2008a). Flickr Tag Recommendation Based on Collective Knowledge. Em *Proceedings of the 17th International Conference on World Wide Web*, pp. 327–336.

Sigurbjörnsson, B. & van Zwol, R. (2008b). Flickr Tag Recommendation Based on Collective Knowledge. Em *Proc. 17th International Conference on World Wide Web (WWW)*, pp. 327–336.

Song, Y.; Zhang, L. & Giles, C. (2011). Automatic Tag Recommendation Algorithms for Social Recommender Systems. *ACM Transactions on the Web*, 5:1–31.

Venetis, P.; Koutrika, G. & Garcia-Molina, H. (2011). On the Selection of Tags for Tag Clouds. Em *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pp. 835–844.

Wu, L.; Yang, L.; Yu, N. & Hua, X.-S. (2009). Learning to Tag. Em *Proc. 18th International Conference on World Wide Web (WWW)*, pp. 361–370.

Wu, Q.; Burges, C. J.; Svore, K. M. & Gao, J. (2010). Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3):254–270.

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

Yang, Y. & Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization. Em *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pp. 412–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Yin, D.; Guo, S.; Chidlovskii, B.; Davison, B.; Archambeau, C. & Bouchard, G. (2013). Connecting Comments and Tags: Improved Modeling of Social Tagging Systems. Em *Proc. 6th ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 547–556.