O PROBLEMA DA ESCASSEZ DE MATCHING: QUANDO RECOMENDADORES NÃO CONECTAM EXTREMOS EM SERVIÇOS DE RECRUTAMENTO

ALAN DA SILVA CARDOSO

O PROBLEMA DA ESCASSEZ DE MATCHING: QUANDO RECOMENDADORES NÃO CONECTAM EXTREMOS EM SERVIÇOS DE RECRUTAMENTO

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

Orientador: Leonardo Chaves Dutra da Rocha
Coorientador: Fernando Henrique Jesus Mourão

São João del Rei
25 de dezembro de 2019

Cardoso, Alan da Silva

XXXX O Problema da Escassez de Matching: quando recomendadores não conectam extremos em serviços de recrutamento / Alan da Silva Cardoso. — São João del Rei, 2019
xxii, 69 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de São João del Rei
Orientador: Leonardo Chaves Dutra da Rocha

Sistemas de Recomendação, Recrutamento On-line, Modelagem de Usuários, Escassez de Matchings

CDU XXXXXXX

[Folha de Aprovação]

Quando a secretaria do Curso fornecer esta folha,
ela deve ser digitalizada e armazenada no disco em formato gráfico.

Se você estiver usando o `pdflatex`,
armazene o arquivo preferencialmente em formato PNG
(o formato JPEG é pior neste caso).

Se você estiver usando o `latex` (não o `pdflatex`),
terá que converter o arquivo gráfico para o formato EPS.

Em seguida, acrescente a opção `approval={`*nome do arquivo*`}`
ao comando `\ppgccufsj`.

Se a imagem da folha de aprovação precisar ser ajustada, use:
`approval=[`*ajuste*`][`*escala*`]{`*nome do arquivo*`}`
onde *ajuste* é uma distância para deslocar a imagem para baixo
e *escala* é um fator de escala para a imagem. Por exemplo:
`approval=[-2cm][0.9]{`*nome do arquivo*`}`
desloca a imagem 2cm para cima e a escala em 90%.

Dedico este trabalho a todas as pessoas que fazem parte da minha vida.

Agradecimentos

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Pellentesque ornare congue sem, et rhoncus justo fermentum vitae. Sed a mattis ipsum. Aliquam placerat auctor tellus id varius. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec consequat blandit lectus vel tempor. Nulla facilisi. Fusce sed arcu nec ex tempor varius et nec ipsum. In gravida, tortor ut dapibus vestibulum, lectus enim lobortis elit, pretium sagittis nibh mi et nunc. Phasellus fermentum ornare urna, ut tincidunt ex finibus at. Proin felis lectus, tempor vitae metus non, gravida molestie mauris.

Vivamus nisi erat, vestibulum nec ultrices eget, accumsan quis tellus. Integer est ipsum, malesuada ut lectus non, pretium consectetur mi. Nulla facilisi. Nulla ullamcorper viverra nisi, in porttitor est ullamcorper a. Pellentesque quis viverra neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec tempus ultricies porttitor.

Curabitur accumsan risus neque, eget vulputate arcu ultricies at. Etiam placerat blandit iaculis. Cras eget risus vitae nisi maximus pharetra vitae sit amet libero. Phasellus sapien sem, fringilla in vulputate quis, efficitur ac elit. Quisque finibus eu tellus mattis pulvinar. Duis vitae eros finibus, viverra orci eget, placerat massa. Vestibulum a mattis tortor. Ut bibendum eleifend ex quis ornare. Nam at bibendum diam. Fusce vel commodo urna. Maecenas interdum aliquam libero ut molestie. Praesent at congue leo, finibus porta sem. Morbi velit nisl, condimentum non maximus sodales, porta non tellus.

Integer quis mi mi. Mauris nec elit eget ex elementum iaculis. Integer scelerisque lacus magna, vitae sodales ipsum porttitor posuere. Phasellus quis mauris gravida elit mollis commodo. Quisque at maximus neque. Quisque at fringilla ligula, sed maximus diam. Fusce lacinia, leo nec scelerisque interdum, lorem lacus pellentesque metus, vel malesuada erat ante eu mauris. Integer posuere massa quis sagittis porttitor.

Nunc finibus, dolor sit amet egestas vehicula, nisi felis consequat elit, ut euismod quam nulla non sapien. Sed quis scelerisque magna. Duis dignissim, erat vel dictum

tempor, ligula diam aliquet erat, sed ultricies nunc augue vel mi. Sed diam ligula, imperdiet eget felis in, aliquet tincidunt ex. Vivamus justo urna, rhoncus ut laoreet nec, aliquam a nibh. Quisque ac tellus eros. Pellentesque quis aliquet mi. Aliquam lectus nulla, lacinia non blandit ac, blandit eu turpis. Etiam sed molestie elit. Suspendisse luctus justo vel quam hendrerit, sit amet varius ipsum tristique. Phasellus vestibulum tortor mollis, maximus arcu vestibulum, consectetur metus.

Curabitur accumsan risus neque, eget vulputate arcu ultricies at. Etiam placerat blandit iaculis. Cras eget risus vitae nisi maximus pharetra vitae sit amet libero. Phasellus sapien sem, fringilla in vulputate quis, efficitur ac elit. Quisque finibus eu tellus mattis pulvinar. Duis vitae eros finibus, viverra orci eget, placerat massa. Vestibulum a mattis tortor. Ut bibendum eleifend ex quis ornare. Nam at bibendum diam. Fusce vel commodo urna. Maecenas interdum aliquam libero ut molestie. Praesent at congue leo, finibus porta sem. Morbi velit nisl, condimentum non maximus sodales, porta non tellus.

Integer quis mi mi. Mauris nec elit eget ex elementum iaculis. Integer scelerisque lacus magna, vitae sodales ipsum porttitor posuere. Phasellus quis mauris gravida elit mollis commodo. Quisque at maximus neque. Quisque at fringilla ligula, sed maximus diam. Fusce lacinia, leo nec scelerisque interdum, lorem lacus pellentesque metus, vel malesuada erat ante eu mauris. Integer posuere massa quis sagittis porttitor.

Nunc finibus, dolor sit amet egestas vehicula, nisi felis consequat elit, ut euismod quam nulla non sapien. Sed quis scelerisque magna. Duis dignissim, erat vel dictum tempor, ligula diam aliquet erat, sed ultricies nunc augue vel mi. Sed diam ligula, imperdiet eget felis in, aliquet tincidunt ex. Vivamus justo urna, rhoncus ut laoreet nec, aliquam a nibh. Quisque ac tellus eros. Pellentesque quis aliquet mi. Aliquam lectus nulla, lacinia non blandit ac, blandit eu turpis. Etiam sed molestie elit. Suspendisse luctus justo vel quam hendrerit, sit amet varius ipsum tristique. Phasellus vestibulum tortor mollis, maximus arcu vestibulum, consectetur metus.

"Ter um plano ruim é melhor do que não ter um plano."
(Mikhail Chigorin)

Resumo

Conectar candidatos e empregos para promover oportunidades reais de colocação é um dos cenários mais impactantes e desafiadores para os Sistemas de Recomendação (SsR). Uma grande preocupação ao criar SsR para serviços de recrutamento é garantir oportunidades de colocação para todos os candidatos e vagas de empregos o mais rápido possível, evitando perdas financeiras para os dois lados. Nós nos referimos a esses cenários em que candidatos ou empregos sofrem com a ausência de correspondência no sistema como *Problem of Matching Scarcity* (**PMS**). Este trabalho apresenta o PMS, discutindo os motivos pelos quais o consideramos uma ameaça recorrente aos serviços de recrutamento e apresentando novas estratégias para identificar, caracterizar e mitigar o PMS em cenários reais. Avaliações experimentais em dados reais da Catho, empresa líder de recrutamento na América Latina, evidenciaram que os currículos tendem a ser menos bem escritos do que as descrições de cargos, exibindo várias informações irrelevantes. Substituir essas informações adequadamente por outras úteis reduz em até 50% o número de currículos e vagas que sofrem de PMS.

Abstract

Connecting candidates and jobs to promote real placement opportunities is one of the most impacting and challenging scenarios for Recommender Systems (RSs). A major concern when building RSs for recruitment services is ensuring placement opportunities for all candidates and jobs as soon as possible, avoiding financial losses for both sides. We refer to these scenarios where candidates or jobs suffer from the absence of matching in the system as the Problem of Matching Scarcity (**PMS**). This paper introduces the PMS, discussing the reasons we consider it as a recurring threat to recruitment services and presenting novel strategies to identify, characterize, and mitigate the PMS on real scenarios. Experimental assessments on real data from Catho, the leading recruitment company in Latin America, evinced that curricula tend to be more poorly written than job descriptions, exhibiting several irrelevant pieces of information. Replacing these pieces properly by useful ones reduces up to 50% the number of curricula and jobs that suffer from PMS.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Neste capítulo, primeiramente apresentamos uma contextualização acerca do problema de recomendação de vagas de emprego e currículos em ambientes de recrutamento on-line. Em seguida, descrevemos as principais características que diferenciam o problema de recomendação de emprego das recomendações tradicionais (i.e. produtos, filmes, música, etc.). Expomos também os objetivo principal desse trabalho, os objetivos específicos e as hipóteses relacionadas. Finalizamos o capítulo apresentando as principais contribuições deste trabalho.

1.1 Contextualização & Motivação

Com a grande expansão da *Internet* nos últimos anos uma grande variedade de serviços surigiram na *Web* Coffman & Odlyzko [2002]. Junto a esses serviços, os usuários da rede passaram da posição de meros espectadores para terem um papel ativo como produtores de conteúdo Kiesler et al. [2002]. Todos os dias milhões de pessoas e empresas utilizam a *Internet*, produzindo, armazenando e consumindo uma grande quantidade de dados variados. O grande volume de dados disponível gerou um cenário desafiador para as mais variadas aplicações, agora obrigadas a lidarem com uma quantidade massiva de dados. Do outro lado, os usuários passaram a possuir mais opções do que efetivamente podem manipular Adomavicius & Tuzhilin [2005].

Empresas como *Google*[1] são exemplos de companhias que precisam lidar com essa grande quantidade de informação, necessitando processar e armazenar, diariamente, dados na ordem de grandeza de petabytes. Para ilustrar, em 1998, o número de páginas indexadas pelo Google já alcançava o número de um milhão, em 2000 um

[1] www.google.com

bilhão e em 2008 esse número já passava de um trilhão Fan & Bifet [2014]. Outra empresa que também pode ser mencionada é o *YouTube*[2], onde os mais de 800 milhões de usuários mensais enviam para a plataforma mais de uma hora de vídeo por segundo. Outros exemplos que podem ser mencionados no mesmo contexto, são as redes sociais. Empresas como *Facebook*[3] e *Twitter*[4] lidam como uma quantidade enorme de interações dos usuários. O Facebook recebe mais de 10 milhões de fotos por hora, enquanto como o Twitter recebe mais de 250 milhões de postagens diariamente.

Sob uma perspectiva ainda mais comercial, algumas das principais aplicações *Internet-based*, tais como *Amazon*[5], *Deezer*[6], *Netflix*[7], dentre outras, proveem um catálogo de itens com milhares de produtos distintos. Embora a disponibilidade de um amplo conjunto de opções tenha sido um cenário desejado no passado, atualmente representa um desafio primordial. Tal mudança deve-se ao fato deste grande volume de opções estar "sufocando" os usuários, tornando a simples escolha de um objeto de interesse uma tarefa complexa.

Para lidar com os desafios associados a esses cenários, diversos Sistemas de Recomendação (SsR) Burke [2002] v?m sendo propostos e utilizados. SsR são ferramentas adotadas que, considerando o histórico de consumo de um usuário, realizam a filtragem dessa grande variedade de itens indicando quais deles são de potencial interesse Adomavicius & Tuzhilin [2005]. Muitas aplicações, em especial as aplicações WEB, recorrem a SsR para ajudar os seus usuários a encontrarem os produtos de desejo, facilitando a tomada de decisões e aumentando a satisfação dos usuários. A personalização dos serviços fornecidos, por meio do uso de SsR, pode ser uma estratégia decisiva para o sucesso de uma aplicação.

Encontramos na literatura diversas estratégias para recomendar produtos, informações e serviços aos usuários Adomavicius & Tuzhilin [2001]; Abbasse & Mirrokni [2007]. O princípio básico de funcionamento de um SR é simples, no entanto sua implementação apresenta diversos desafios computacionais que vão desde como modelar o comportamento dos usuários, a como utilizar as informações dessa modelagem para prover a recomendação. Por exemplo, o comportamento dos usuários pode ser representado por qualquer subconjunto de itens que ele tenha consumido, ou mesmo por itens ainda não consumidos mas que possam ser relevantes para o sistema ou o usuário, dada uma métrica de interesse.

[2]www.youtube.com
[3]www.facebook.com
[4]www.twitter.com
[5]www.amazon.com
[6]www.deezer.com
[7]www.netflix.com

Dentro desse contexto de serviços digitais, mais recentemente os chamados serviços de recrutamento on-line vêm atraindo a atenção de um número crescente de usuários com demandas urgentes Coffman & Odlyzko [2002]; Spina et al. [2017]. Esses serviços são especializados em gerenciar dois atores principais: (1) candidatos, pessoas com características e competências específicas, em busca de novas oportunidades de emprego; e (2) recrutadores, empresas que oferecem ofertas de emprego e buscam os melhores profissionais do mercado. O objetivo é realizar o *matching* (casamento) entre cada vaga de emprego e seus candidatos ideais, contrastando o perfil dos candidatos (por exemplo, currículo - CV) com as qualificações exigidas pela vaga de trabalho. Esses serviços também contam com a ajuda de SsR para auxiliar candidatos e recrutadores a encontrarem as melhores oportunidades de emprego/recrutamento Fazel-Zarandi & Fox [2009]. O mercado global de recrutamento on-line é extremamente valioso, tendo sido avaliado recentemente como um mercado que movimenta entre $20 a $30 bilhões de dólares anualmente Spina et al. [2017].

Uma das grandes preocupações no processo de construção de SsR para os domínios de recrutamento é garantir oportunidades de *matching* reais para todos os candidatos e vagas de emprego cadastratados no sistema o mais rápido possível. De fato, longos períodos de espera por *matching* implicam em perdas de oportunidades de negócios, causando danos financeiros tanto para candidatos quanto para as empresas. Além disso e como consequ?ncia, a credibilidade dos serviços de recrutamento fica comprometida ao longo do tempo. Nos referimos a esses cenários em que os candidatos ou vagas de emprego sofrem com a ausência de oportunidades de *matching* como o **Problem of Matching Scarcity** (**PMS**). Consideramos o PMS uma ameaça real e recorrente aos serviços de recrutamento, dado os desafios desse cenário: (1) ampla gama de perfis de candidatos e vagas disponíveis; (2) alta dinamicidade, na qual a oferta e a demanda de vagas de emprego podem variar significativamente ao longo do tempo e; (3) desequilíbrios sazonais entre demanda e oferta nas áreas profissionais, fazendo com que existam áreas com mais candidatos do que vagas em diferentes momentos.

1.2 Objetivo

O objetivo principal desse trabalho é propor e avaliar novas estratégias capazes de mitigar o PMS em ambientes de recrutamento on-line. Mais precisamente, nossa meta é que essas estratégias melhorem a qualidade dos currículos e vagas descritas, ao sugerirem alterações pontuais nos mesmos, de forma que possam ser processados pelos SsR de forma mais efetiva. Para isso levantamos quatro hipóteses que serão validadas

no decorrer desse trabalho.

Hipótese 1 (H1): Existem candidatos/vagas que passam longos períodos sem conseguirem uma oportunidade em sistemas de recrutamento on-line.

Hipótese 2 (H2): É possível identificar o conjunto de candadidatos/vagas que sofrem de escassez de oportunidades.

Hipótese 3 (H3): Existe um subconjunto de *features* contidas em curriculos/vagas que trazem maior ganho de informação para SsR em cenários de escassez.

Hipótese 4 (H4): Auxiliar candidatos/recrutadores na identificação dessas *features* que melhor descrevem currículos e vagas é uma forma eficaz de se mitigar cenários de escassez.

A fim de verificar a hipótese H1, propomos uma metodologia de caracterização para avaliar diferentes dimensões e características potencialmente relacionadas ao PMS. Para validarmos as hipóteses H2 e H3, primeiramente formalizamos o próprio PMS, bem como os conceitos intimamente relacionados a ele. Após isso, propomos dois métodos heurísticos *Bayesian-inspired* para (1) identificar candidatos e vagas que sofrem de escassez; (2) identificar características que potencialmente acentuam o PMS. A hipótese H4 é baseada na premissa de que existe um conjunto de *features* utilizadas no processo de descrição do currículo/vaga que são mais informativas e que favorecem a tarefa de correspondência entre um candidato e uma vaga, do que outras *features* irrelevantes que nada contribuem ou até atrapalham. A partir disso, propomos uma terceira heurística *Bayesian-inspired* capaz de propor mudanças no currículo e nas descrições de cargos para mitigar o PMS. Acreditamos que ao substituir adequadamente essas *features* irrelevantes por outras úteis, aproximaremos candidatos que sofrem do PMS a vagas semanticamente relacionadas a eles. Para validar a metodologia e as três heuristicas e, consequentemente, validar todas as hipóteses desse trabalho, utilizamos uma amostra real de dados fornecida pela Catho [8], empresa líder de mercado na América Latina, contendo 376.762 CVs e 115.955 vagas de emprego, obtidas em um período de 13 meses (Janeiro a Janeiro).

1.3 Principais Contribuições

A principal contribuição deste trabalho é ser um guia prático para todos os proprietários de serviços de recrutamento on-line, que lidam constantemente com

[8] www.catho.com.br

a tarefa de encontrar oportunidades de emprego/candidatos adequados para todos os usuários do sistema. Além disso, destaca-se também a avaliação da metodologia de caracterização desses ambientes que, não só ajuda a encontrar oportunidades de melhorias nesse cenários, como também auxilia a idenficar os desequilíbrios entre demanda e oferta que possam levar ao problema do PMS. Em resumo, as principais contribuições deste trabalho são:

1. A formalização e caracterização de um novo problema nos serviços de recrutamento on-line, o *Problem of Matching Scarcity*.
2. A proposta de uma estratégia para identificar automaticamente currículos e vagas de emprego que sofrem de escassez.
3. A proposta de estratégias para mitigar o PMS.
4. A avaliação da metodologia e estratégias propostas em dados reais.

1.4 Organização do Documento

Este trabalho é organizado como se segue. O Capítulo 2 apresenta um referencial teórico da tarefa de recomendação, bem como as estratégias utilizadas para recomendação em sistemas de recrutamento on-line. Ainda neste, abordamos as três classes de trabalhos focados em aprimorar a recomendação de currículos/vagas e mostramos porque as estratégias clássicas não são aplicáveis ao problema do PMS. No Capítulo 3, introduzimos formalmente o PMS e seus conceitos. Aqui também apresentamos a metodologia de caracterização, a qual é dividida em cinco dimensões de análise que podem ser combinadas iterativamente para responder questões reais acerta do cenário de recrutamento. Listamos também dezoito métricas que são dividas entre essas cinco dimensões e que servem para guiar as análises. No Capítulo 4 definimos as estratégias de identificação de currículos/vagas que sofram do PMS, o que consolida nossa a hipótese H2. Neste capítulo, também lidamos com a hipótese H3 ao propormos seis estratégias que para mitigar o PMS. No Capítulo 5 avaliamos a metodologia de caracterização em dados reais e provamos sua utilidade. Essa metodologia nos permite consolidar a hipótese H1 ao verificar a existência de currículos/vagas que sofrem de escassez. Também avaliamos nossas estratégias de identificação de itens escassos, onde para isso definimos duas métricas de avaliação que visam averiguar a efetividade de nossas soluções. É também nesse capítulo que avaliamos nossas soluções para tratamento do PMS. No processo de avaliação cruzamos os resultados das seis estratégias com o intuito de identificar a que melhor se aplica em casos de uso variados, validando por fim a hipótese

H4. Finalmente, o Capítulo 6 apresenta as conclusões deste trabalho, sumarizando os resultados obtidos em cada capítulo anterior. Neste também são apresentados os possíveis trabalhos futuros que a serem realizados para se aprofundar ao PMS.

Capítulo 2

Referencial Teórico

Este capítulo visa apresentar os conceitos relacionados a SsR ao revisar a definição formal da tarefa de recomendação. Apresentamos brevemente as estratégias clássicas de recomendação e suas aplicações. Em seguida abordamos as especificidades dos SsR de ambientes de recrutamento, seus desafios e como eles se diferenciam dos SsR tradicionais. Em seguida, enumeramos as três classes de estratégias pesquisadas com o intuito de aprimorar SsR de candidatos/vagas.

2.1 Sistemas de Recomendação

Sistemas de Recomendação (SsR) são ferramentas cruciais para diversos sistemas na *Web* que possuem uma grande variedade de itens e serviços para oferecer aos seus usuários Abbasse & Mirrokni [2007]; Ramos [2015]; Mavlanova et al. [2016]; Adomavicius & Tuzhilin [2005]. Eles auxiliam encontrar objetos de interesse (i.e., livros, filmes ou músicas) de desejo que estejam em grandes coleções de itens, nas quais seria extremamente exaustivo, ou até mesmo impossível, explorar cada um deles. Portanto, um SR basicamente encontra/prediz, quais itens melhor se encaixam com o perfil do usuário [Mourão, 2014]. Formalizando, em um sistema temos que U representa o conjunto de todos os usuários e I o conjunto de itens a serem recomendados, a tarefa de recomendação consiste em encontrar um subconjunto de itens $R_u \subset I$, de tamanho k (i.e., $|R_u| = k$), que maximiza a função de utilidade $f(u, i)$ para cada usuário $u \in U$ e $i \in R_u$ [Adomavicius & Tuzhilin, 2005].

As estratégias para gerar a recomendação variam conforme o caso de uso. Existem diversos métodos diferentes Bashiri [2018]; Jannach et al. [2010]; Candillier et al. [2009]; Ricci et al. [2011]; Bobadilla et al. [2013] mas em geral, eles se dividem em duas classes: recomendações personalizadas e não personalizadas [Akshita & Smita, 2013]. Enquanto

os métodos personalizados procuram traçar um perfil de cada usuário, entendo quais são suas preferências e anseios, os métodos não personalizados se baseiam no comportamento coletivo dos usuários do sistema. As técnicas personalizadas podem ainda ser divididas em outras três categorias: Filtragem Baseada em Conteúdo (CB), Filtragem Colaborativa (CF) e Métodos Híbridos [Bobadilla et al., 2013; Beel et al., 2016].

2.1.1 Filtragem Baseada em Conteúdo

A estratégia utilizada pela Filtragem Baseada em Conteúdo (CB) é recomendar itens novos a um usuário, baseando-se nas semelhanças entre o novo item e os itens consumidos por ele no passado. Isso é possível assumindo-se a premissa de que usuários possuem uma preferência sistemática correlacionada com os atributos dos itens [Pazzani, 1999; Schafer et al., 2007; Ricci et al., 2011]. Dessa forma, recomendar itens similares aos quais o usuário tenha demonstrado satisfação no passado, é uma forma de encontrar novos itens de potencial interesse. A similaridade de um item é estimada utilizando as características (i.e., atributos) relacionados aos itens. Supondo o cenário de um sistema de filmes, para recomendar um filme f a um usuário u, o recomendador baseado em conteúdo compreende quais são as características entre os filmes que o usuário u mais gostou no passado (gênero, atores, diretor, sinopse e outros) e encontra novos filmes que também compartilhem delas. Filmes que apresentem uma alto grau de similaridade com as preferências passadas do usuário são recomendados. A grande vantagem dessa estratégia é sua capacidade de trabalhar com novos itens inseridos na coleção, uma vez que depende unicamente das características intrínsecas desse item.

Conforme proposto por Bobadilla et al. [2013], o processo de recomendação baseada em conteúdo pode ser sumarizado em três passos:

1. **Extração dos atributos dos itens do domínio:** identifica-se as características que melhor descrevem os itens do domínio. Utilizando o exemplo de filmes, esses atributos podem ser (mas não se limitando apenas a eles): gênero, lista de atores, ano de lançamento, duração, diretor e outros

2. **Comparação entre os atributos dos itens do domínio e dos itens consumidos pelo usuário:** por meio de uma comparação direta entre os atributos dos itens, encontram-se aqueles que melhor se assemelhem aos itens consumidos pelo usuário.

3. **Recomendação dos itens**: fornece ao usuário a recomendação dos itens identificados como possuidores de atributos de interesse.

O desafio dessa abordagem é determinar o quanto um usuário irá realmente se interessar por um item específico [Van Meteren & Van Someren, 2000]. Para estimar o nível de interesse pode-se utilizar heurísticas tradicionais [Baeza-Yates et al., 1999; Garcia & Amatriain, 2010], algoritmos de classificação como Rocchio [Balabanović & Shohom, 1997; Joachims, 1998], Regras de Indução [Cohen et al., 1995; Kim et al., 2001], Métodos dos Vizinhos Mais Próximos [Billsus et al., 2000] e Métodos Probabilísticos [Park & Tuzhilin, 2008; De Gemmis et al., 2008].

A Filtragem Baseada em Conteúdo está intimamente ligada com a área de Recuperação de Informação, visto quem ambas visam extrair características de itens. A suposição em comum é a de que usuários formalizam consultas que expressam ativamente seus interesses e necessidades em termos de características relacionadas aos itens [Jannach et al., 2010; Ramos et al., 2003]. Os métodos estado-da-arte da recomendação baseada em conteúdo são apresentados por Lops et al. [2011].

No entanto, estratégias dessa classe apresentam alguns desafios, sendo eles:

1. **Análise de conteúdo limitada:** por ser fortemente baseada nos atributos dos itens, esse atributos precisam ser suficientes, relevantes e processáveis. Ainda nesse contexto, dois itens com o mesmo conjunto de atributos, são indistinguíveis.

2. **Superespecialização:** por depender do consumo passado do usuário, o sistema pode sugerir unicamente itens que sejam semelhantes ao que o usuário já conhece. Por exemplo, usuários que tenham apenas visto filmes do gênero drama, jamais teriam algum outro gênero recomendado a ele.

3. **Problema do novo usuário:** um usuário novo não terá consumido itens suficientes para que o sistema seja capaz de entender o seu perfil e realizar recomendações precisas.

2.1.2 Filtragem Colaborativa

A premissa por trás da abordagem Filtragem Colaborativa (CF) é tentar predizer o quanto um usuário se interessará por um novo item, utilizando-se o *feedback* atribuído pelo usuário a itens similares, ou no *feedback* atribuído por usuários similares ao usuário alvo [Ricci et al., 2011]. As técnicas de filtragem colaborativa são separadas em dois grupos Yang et al. [2014]: *memory-based* e *model-based*. Estes dois grupos podem ser orientados segundo o usuário e segundo o item, tal qual sumarizado na tabela 2.1.

1. **Memory-based:** esse método utiliza os *ratings* previamente fornecidos pelos usuários para calcular a similaridade entre usuários e itens [Adomavicius & Tuzhilin,

Tabela 2.1: Classes de estratégias de recomendadores de filtragem colaborativa.

	Classes de Estratégia	
	Memory-based	Model-based
User-based	Combina as preferências dos k usuários mais similares, com características relacionadas.	Explora o histórico de preferências do usuário para treinar modelos de recomendação.
Item-based	Combina as avaliações dos k itens mais similares, considerando todos os usuários.	Explora as avaliações passadas do item para treinar modelos de recomendação.

2005; Ricci et al., 2011]. As técnicas *user-based* agregam os *ratings* atribuídos a um item pelos N usuários mais similares ao usuário alvo da recomendação. Já as técnicas *item-based* agregam os *ratings* recebidos pelos N itens mais similares ao item a ser recomendado. Nessa categoria destacam-se a aplicação de algoritmos baseados em grafos [Huang et al., 2002; Lo & Lin, 2006; Silva et al., 2010] e abordagens relacionadas ao Vizinho Mais Próximo [Deshpande & Karypis, 2004; Dong et al., 2011], como o *UserKNN* e *ItemKNN* [Grčar et al., 2006; Campos et al., 2010].

2. **Model-based:** usam a coleção de *ratings* para gerar um modelo que é usado para fazer predições. Os métodos utilizados para isso podem utilizar diversos algoritmos de aprendizado de máquina: classificadores de redes neuronais [Billsus & Pazzani, 1998]; aprendizagem de regras de indução [Basu et al., 1998]; redes Bayesianas [Horvitz et al., 1998]; e modelos de fatores latentes [Bell & Koren, 2007; Sarwar et al., 2000]. Similarmente aos algoritmos *memory-based*, estas técnicas também pode ser empregadas segundo as modelagens *user-based* ou *item-based*, sobre os mesmos pressupostos anteriores.

As estratégias de Filtragem Colaborativa também apresentam alguns desafios, sendo eles:

1. **Esparsidade:** o número de *ratings* a ser estimado é geralmente muito superior ao número de *ratings* já obtidos. O sucesso da estratégia também depende de haver uma grande quantidade de usuários utilizando o sistema.

2. **Problema do novo item:** uma vez que um novo item seja adicionado ao sistema, o modelo só poderá recomendá-lo quando uma quantidade relevante de usuários já o tiverem classificado.

3. **Problema do novo usuário:** similar ao problema das estratégias de Filtragem Baseada em Conteúdo, para realizar uma predição precisa, o sistema precisa que

um usuário novo tenha classificado itens suficientes para que o modelo possa aprender as preferências dele.

2.1.3 Métodos Híbridos

Os métodos híbridos visam somar as vantagens de ambas as abordagens clássicas [Adomavicius & Tuzhilin, 2005]. Essas estratégias podem ser classficadas em quatro categorias Martin et al. [2014] conforme a listagem abaixo:

- **Combinando Recomendadores Separados**: existem duas formas de implementar essa solução: (1) ambas estratégias são aplicadas separadamente e combina-se os resultados de suas predições através da combinação linear dos seus *ratings*; (2) escolhe-se a estratégia mais adequada para um certo momento, intercalando as duas conforme necessário [Billsus & Pazzani, 2000; Kim et al., 2011].
- **Adicionando características da CB em CF**: incorpora-se características da CB em técnicas de CF. Uma forma de fazer isso é criando-se um perfil de usuário baseado em conteúdo e utilizando-o no cálculo da função de similaridade entre usuários [Melville et al., 2002; Li & Kim, 2003; Hu & Pu, 2010].
- **Adicionando caracterísicas da CF em CB**: incorpora-se características da CF em técnicas CB. A abordagem mais popular nesta categoria é utilizar alguma estratégia em um grupo de perfis de usuários baseados em conteúdo [Mooney & Roy, 2000].
- **Modelo de Recomendação Unificado**: cria-se um único modelo que incorpore características tanto de CF quanto de CB. Um exemplo possível seria utilizar informações baseadas em conteúdo e também as preferências passadas dos usuários [Popescul et al., 2001; De Campos et al., 2010; Choi et al., 2012].

A seguir apresentaremos a importância dos SsR no cenário de recrutamento on-line, como eles se encaixam nesse ambiente, seus desafios e o que vem sendo trabalho para sua melhoria.

2.2 A Recomendação em Sistemas de Recrutamento e seus desafios

A *Internet* mudou significativamente a forma como as empresas e as pessoas encontram oportunidades de emprego hoje em dia. Os sistemas on-line de recrutamento

se tornaram uma ferramenta vital no processo de encontrar um emprego, ou mesmo um candidato ideal para uma vaga. Existem benefícios claros no uso desses sistemas, as pessoas que usam esses serviços são realocadas 25% mais rapidamente do que aquelas que não os usam [Peter & Hani, 2014]. Além disso, as pessoas que encontram seus empregos por meio desses serviços on-line especializados ficam mais tempo em empregos do que outros candidatos Centeno [2004]. Há ainda um outro benefício que é a economia e a eficiência que beneficiam candidatos e recrutadores que fazem uso desses serviços de recrutamento na *web* Lee [2007].

Esses sistemas de recrutamento fazem grande uso de SsR para conectarem candidatos com determinadas competências e expertises, que procuram melhores posições no mercado, com os recrutadores, que procuram os melhores profissionais disponíveis, elencando as atividades que serão desempenhadas por eles. Existem características peculiares a esse cenário que o torna desafiador, por exemplo, não é possível recomendar uma mesma vaga diversas vezes para um mesmo candidato. Além disso, normalmente, as necessidades dos candidatos e recrutadores são urgentes, considerando que o tempo que um indivíduo se mantém desempregado é proporcional aos prejuízos que lhe ocorrem na vida pessoal. Concomitantemente a isso, as empresas podem perder muitas oportunidades de negócio por não conseguirem mão de obra qualificada Alotaibi [2012]. Portanto, consiste em um cenário com seus desafios e fatores que o diferenciam de outros cenários tradicionais de recomendação [Pizzato et al., 2010]. A seguir listamos os desafios dos sistemas de recomendação no cenário de recrutamento.

1. **Recomendações bilaterais:** os sistemas de recrutamento on-line são compostos de recomendações bilaterais sobre as quais é necessário a satisfação mútua tanto do candidato quanto do recrutador [Yu et al., 2011; Malinowski et al., 2006].

2. **Distinção entre habilidades desejadas e habilidades exigidas:** o SR deve ser capaz de compreender quais habilidades são desejadas e quais são exigidas na descrição de uma vaga. É possível recomendar um candidato que não possua todas as habilidades desejadas, no entanto, não é possível recomendar um candidato que não possua alguma das habilidades exigidas [Fazel-Zarandi & Fox, 2009].

3. **Cada indivíduo é único:** ao contrário dos cenários mais comuns de recomendação, cada indivíduo é único e não pode ser recomendado para uma mesma vaga diversas vezes [Keim, 2007; Alotaibi, 2012].

4. **Descrições muito específicos (ou muito genéricas):** trata-se da qualidade e da utilidade das informações inseridas pelos candidatos e recrutadores no sistema [Kureková et al., 2015]. Por exemplo, é comum ver currículos e vagas de emprego

extremamente detalhados e específicos ou mesmo genéricos e inespecíficos, o que dificulta a tarefa de pesquisa e recomendação [Lang et al., 2011].

5. **Cenário dinâmico:** o cenário de empregos é altamente dinâmico, no qual a oferta e a demanda por vagas de emprego podem variar significativamente ao longo do tempo, apresentando desequilíbrios sazonais. Tanto candidatos quanto recrutadores, removem seus perfis tão logo tenham conseguido um *matching*. Junto a isso, há uma grande quantidade de novos candidatos e vagas surgindo diariamente [Cardoso et al., 2019].

A eficácia desses sistemas é limitada a esses desafios, pois a falha em resolvê-los pode levar a longos períodos de espera por uma oportunidade de alocação de um candidato em uma vaga. Existem diversas linhas de pesquisa que visam solucionar ou minimizar os desafios mencionados. A maioria dos trabalhos sobre SsR para serviços de recrutamento ocorrem em três classes. Iremos apresentar essas classes, bem como os trabalhos que vem sendo realizados em cada uma delas.

2.2.1 Classe 1: Adaptando Sistemas de Recomendação Tradicionais

A primeira classe contém trabalhos focados na adaptação de SsR usados em um domínio tradicional, como entretenimento, para recomendações de emprego. A combinação de diferentes estratégias de recomendação também foi avaliada para obter SsR mais robustos, adaptáveis à dinâmica do cenário de recrutamento Lu et al. [2013]; Yang et al. [2017]; Daramola et al. [2010]; Hong et al. [2013].

Em Tran et al. [2017], os autores fazem uma avaliação detalhada das estratégias de recomendação de emprego existentes, apontando suas vantagens e desvantagens. O trabalho Liu et al. [2017] se concentra em gerar boas recomendações para novos graduados e candidatos sem experiência de trabalho, um caso semelhante ao problema do novo usuário em ambientes tradicionais Lika et al. [2014]. Em ul haq Dar & Dorn [2018], os autores realizaram testes com oito classificadores diferentes e compararam os resultados de suas aplicações na classificação de ofertas de emprego. Em Kokkodis [2019], os autores propuseram uma estratégia para minimizar o impacto da supervalorização das pontuações atribuídas nos cenários de recomendação profissional. Por fim,*Changchun Li et al* Li et al. [2019] apresenta uma estratégia de classificação para textos extremamente curtos usando um *bag of words* que pode ser aplicado para currículos e descrições de vagas.

2.2.2 Classe 2: Entendendo o Comportamento de Usuários

A segunda classe de trabalhos concentra-se em uma melhor compreensão da dinâmica do comportamento do usuário nos sistemas de recrutamento on-line para explorar padrões potencialmente úteis para as tarefas de recomendação e pesquisa. Em Paparrizos et al. [2011], os autores pretendem entender a evolução da carreira dos candidatos, fornecendo melhorias nos SsR para que possam detectar quais seriam as próximas vagas de interesse. *Spina et al* Spina et al. [2017] analisa as consultas que os candidatos fazem nesses sistemas para interpretar como eles procuram vagas, concluindo que as métricas e estratégias usadas nos sistemas tradicionais de busca não são eficazes no cenário de recrutamento. Ainda preocupado com o modo como os usuários realizam essas pesquisas, no Salehi et al. [2018] é analisado como pesquisas de emprego mal formuladas podem trazer resultados de baixa qualidade e afetar a satisfação do usuário. Em Pournajaf et al. [2017], os autores procuraram melhorar os resultados de consultas para vagas de emprego na parte final da distribuição. Outros trabalhos visam entender o comportamento dos algoritmos de recomendação e como eles podem ser aprimorados nesse cenário Alotaibi & Ykhlef [2012].

2.2.3 Classe 3: Estratégias de Pré-processamento

Finalmente, a terceira classe compreende esforços no pré-processamento dos dados de entrada para melhor estruturar as informações enviadas pelos usuários, para que possam ser aplicadas posteriormente às estratégias tradicionais de recomendação com mais eficiência Das et al. [2019]; Berti-Equille [2019]. Entre esses trabalhos, destacamos o apresentado por *Turrell et al* Turrell et al. [2018], que apresenta uma abordagem para categorizar as áreas de atividade por meio de um modelo de tópicos latentes, para facilitar a tarefa de correspondência entre um currículo e uma vaga. Na mesma linha, os autores apresentam uma estratégia focada na extração das habilidades mais importantes em sete posições hierárquicas distintas (administrador, analista, desenvolvedor, engenheiro, líder, suporte e testador), destacando as mais desejadas habilidades em cada posição. *Mihuandayani et al* Mihuandayani et al. [2018] adotam dados de redes sociais para agregar informações em perfis de cargos para posterior seleção de candidatos. No Javed et al. [2015], *Javed et al* executam a categorização dos trabalhos por meio de seus títulos, criando assim grupos específicos de áreas profissionais para subsequentemente executar a recomendação das vagas.

2.3 Síntese do Capítulo

Este capítulo se inicia apresentando formalmente os sistemas de recomendação e suas aplicações. Em resumo, trata-se do desafio de encontrar itens de interesse de um usuário, em meio a uma grande coleção de itens. Apresentamos as classes de recomendadores existentes, as quais são divididas em duas categorias, as com SsR focados em cada usuário separadamente (i.e., personalizados) e os que focam em todos usuários, gerando recomendações para todos (i.e., não-personalizados). Para os recomendadores personalizados, descrevemos as três abordagens clássicas que utilizam informações dos usuários para predizer itens.

Posteriormente, apresentamos o cenário de recrutamento on-line e como os sistemas de recomendação são parte fundamental de sua estrutura. Detalhamos os problema inerentes desse domínio, tais como, a alta dinamicidade dos candidatos e vagas, a necessidade de recomendações bilaterais, os problemas com a qualidade das informações providas pelos usuários e a impossibilidade de recomendar um mesmo candidato diversas vezes. Em seguida apresentamos as três classes de trabalhos que visam aprimorar a recomendação de empregos, sendo elas: a adaptação de recomendadores tradicionais, análise do comportamento dos usuários e estratégias de pré-processamento.

Considerando a organização da literatura mencionada, este trabalho está mais relacionado a trabalhos adequados à terceira classe, pois também propomos estratégias que visam preprocessar as informações de vagas e currículos fornecidos pelos usuários. Além disso, a metodologia de caracterização apresentada é muito aderente à segunda classe. No entanto, diferentemente de todos esses esforços, estamos focados no PMS. Conforme discutido na Introdução, argumentamos que, embora as melhorias nos sistemas de recrutamento on-line sejam sempre necessárias, para cenários de escassez elas são ainda mais críticas. Neste trabalho abordamos o problema de escassez de oportunidades e nos referimos a ele como o **Problema of Matching Scarcity** (**PMS**) e, até onde sabemos, é a primeira vez que o PMS é discutido na literatura. No próximo Capítulo introduzimos formalmente o PMS.

Capítulo 3

O Problema da Escassez

Este capítulo, primeiramente, introduz formalmente o *Problem of Matching Scarcity (**PMS**)*. Em seguida, apresenta uma metodologia capaz de caracterizar o PMS em cenários reais, bem como identificar pontos a serem melhorados em sistemas de recrutamento online.

3.1 Definição do Problema

Embora as demandas em domínios de recrutamento sejam geralmente urgentes, atendê-las em um curto período, continua sendo um grande desafio para os SsR. Três características principais pioram o tratamento dessas demandas. Primeiro, existe uma grande diversidade de descrições detalhadas de vagas e candidatos a emprego. Portanto, quantificar a correspondência entre itens distintos se torna complexo, pois nenhuma vaga ou currículo é trivialmente semelhante a outro. Por exemplo, a mesma vaga oferecida pela mesma empresa, mas em locais diferentes, pode não ser semelhante a um determinado candidato. Segundo, o domínio apresenta uma alta dinamicidade. O número de candidatos e vagas de emprego que entram ou saem do sistema pode ser enorme. Por fim, destacamos os desequilíbrios naturais e sazonais entre demanda e oferta nas diferentes áreas. Dada a natureza bilateral dessas demandas, o SsR não pode atender todas elas o tempo todo. Nos referimos a esses períodos em que os candidatos ou vagas de emprego sofrem com a ausência de oportunidades como o ***Problem of Matching Scarcity*** (**PMS**).

Para definir formalmente o PMS, começamos definindo alguns conceitos principais relacionados a ele. Primeiro, este trabalho classifica as *features* (características) existentes [1] em duas classes distintas: **Screening** e **Appendant**. Definimos uma

[1] Uma *feature* é qualquer palavra ou numeral usado em currículos ou vagas para descrever uma

Screening feature da seguinte maneira:

Definição 3.1.1. Screening feature. *Qualquer feature útil para recrutadores ou candidatos tomarem uma decisão ou julgar um currículo ou uma vaga. As Screening features são caracterizadas por três aspectos principais. Primeiro, elas são interpretáveis, o que significa que os atores do domínio as reconhecem como conceitos significativos. Segundo, elas são populares, sendo frequentemente usadas por recrutadores ou candidatos. Terceiro, elas são discriminativas, o que significa que são observadas com mais frequência em pares de <CV, vaga> sorteados aleatoriamente.*

Por outro lado, todos as *features* restantes que não atendem a essa definição são consideradas *Appendant features*. O termo *Appendant* é usado como uma alusão ao tipo de informação presente em um currículo ou vaga, mas desnecessário, não contribuindo ou mesmo prejudicando o processo de recrutamento.

Outro conceito central relacionado ao PMS é o próprio termo **matching**. Neste trabalho, apresentado pela definição 3.1.2. Observe que não estamos assumindo a medida de similaridade a ser empregada. De fato, conforme discutido nos capítulos a seguir, diferentes estratégias de medição devem ser avaliadas.

Definição 3.1.2. Matching. *Estado no qual o valor de similaridade entre o conjunto de features F_v e F_c extraídas, respectivamente, de uma dada vaga v e CV c é acima de um threshold τ.*

Por sua vez, definimos **escassez** conforme apresentado pela definição 3.1.3.

Definição 3.1.3. Escassez. *Estado de uma vaga v ou CV c no qual há menos que μ matchings distintos abrangendo v ou c em um período de tempo T.*

Portanto, definimos o PMS conforme apresentado pela definição 3.1.4. Observe que, apesar do PMS ser um problema personalizado, uma vez que candidatos e recrutadores distintos podem tolerar valores distintos de T_M, na prática, definir T_M representa uma decisão de negócios. O serviço de recrutamento pode determinar um valor capaz de se adequar à necessidade e ao perfil geral de seus usuários.

Definição 3.1.4. Problem of Matching Scarcity (PMS). *Cenários nos quais uma vaga v ou um CV c apresenta escassez de oportunidades por um período contínuo de tempo maior que um threshold T_M.*

---

informação sobre os atores do processo de recrutamento (por exemplo, empresa, contratador ou candidato).

3.2 Caracterizando PMS

Empresas especializadas em recrutamento on-line utilizam-se de Sistemas de Recomendação com o intuito encontrar as melhores oportunidades de emprego para candidatos e os melhores candidatos para as vagas disponíveis. Esse processo precisa ser ágil e eficaz, pois demoras representam perdas econômicas para ambos as partes. Esses sistemas precisam ser aperfeiçoados e adaptados, constante e automaticamente, a cenários diferentes. A exemplo, temos cenários de escassez de vagas, ausência de um determinado perfil de candidato, desequilíbrio entre oportunidades, áreas e/ou regiões etc.

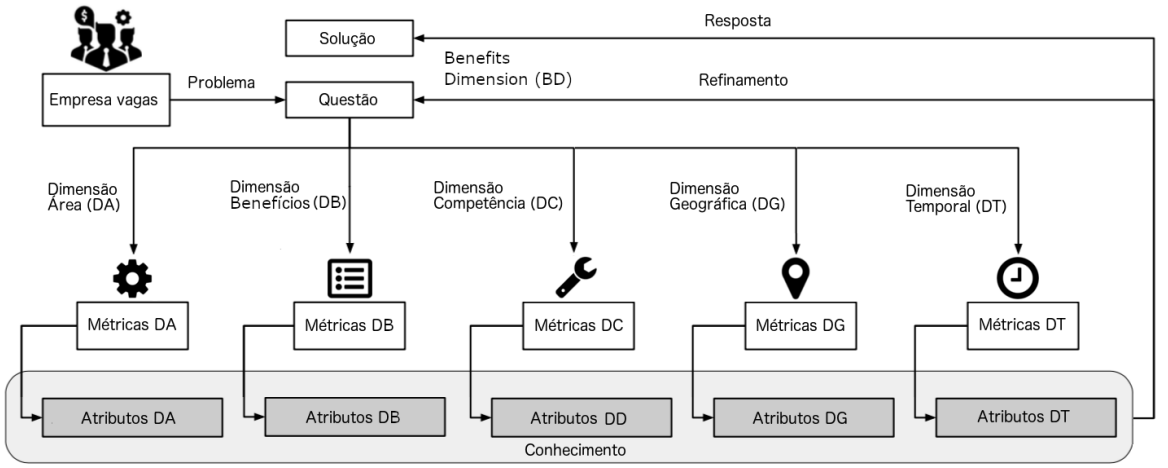


Figura 3.1: Metodologia de Caracterização.

Nesses cenários, a qualidade das recomendações apresentadas aos usuários depende de um processo contínuo e iterativo de ajuste dos SsR, baseado em uma análise constante dos dados disponíveis. Isso porque os usuários estão constantemente gerando novos *logs* de dados e essa análise contínua mantém o sistema com uma visão atualizada do comportamento dos seus usuários. Somente para ilustrar, suponhamos que uma empresa de recrutamento, objetivando responder como está a demanda e oferta de vagas em um determinado momento, interprete que há certas áreas profissionais com uma grande oferta de oportunidades no final do ano. Nessa perspectiva, gera recomendações especializadas para essas áreas a fim de suprir o que o mercado pede. No entanto, essa necessidade pode já não ser necessária no início do ano seguinte, visto que o comportamento dos usuários e do mercado muda ao longo do tempo. Embora seja simples exemplificar, entender e ajustar SsR para o acompanhamento dessas flutuações do mercado não é uma tarefa fácil, momento em que as empresas normalmente precisam de uma estratégia que as ajude a sistematizar o processo de

melhoria contínua. Esta afirmação é especialmente verdadeira quando o objetivo é evitar o PMS. Nesse sentido, discutimos como uma metodologia de caracterização pode ser útil para rastrear e entender melhor o PMS em cenários reais.

A metodologia de caracterização proposta visa assistir empresas especializadas em recrutamento a compreender os diversos fatores que orbitam um SR de vagas de emprego. Conforme ilustrado na figura 3.1, o processo se inicia com uma questão de interesse suscitada pela empresa de recrutamento. Essa questão deve ser encaixada em uma das cinco dimensões suportadas por nossa metodologia: **(1) Dimensão Área (DA):** cujo objetivo é identificar características considerando as áreas de atuação possíveis (i.e., computação, engenharia, saúde, etc.); **(3) Dimensão Benefícios (DB):** avalia as características que descrevem os atributos anseiados pelos candidatos; **(3) Dimensão Competência (DC):** está relacionada às habilidades específicas dos candidatos e/ou ofertas; **(4) Dimensão Geográfica (DG):** identifica características relacionadas à localização dos candidatos e das ofertas; e **(5) Dimensão Temporal (DT):** verifica se um determinado desequilíbrio é apenas sazonal ou recorrente.

Uma vez definidas as dimensões de análise, o próximo passo é aplicar as métricas (descritas em detalhes nas seções 3.2.1 (DA), 3.2.2(DB), 3.2.3(DC), 3.2.4(DG), e 3.2.5(DT)). Essas métricas geram perfis comportamentais e podem ser usadas para: (i) responder a questão elencada e solucionar o problema ou (ii) refinar a questão levantada por meio de uma nova iteração na metodologia, resultando em um perfil mais detalhado. Essa nova iteração pode ser feita por meio de uma nova dimensão e pode utilizar conhecimento gerado previamente por outras iterações. É importante ressaltar que nossa metodologia é um guia e a empresa de recrutamento on-line deve analisar o conhecimento gerado pelas métricas, decidindo se ele é suficiente para responder a questão ou se requer um maior refinamento.

Considerando o exemplo previamente mencionado, em que a empresa percebe uma grande quantidade de vagas de trabalho sendo abertas no final do ano (dimensão temporal), pode-se agora querer iterar novamente na metodologia com o intuito de entender quais setores do mercado estão gerando essa demanda (dimensão área profissional) e assim conseguir gerar recomendações ainda mais adequadas ao mercado. A tabela 3.1 sumariza as métricas definidas por nossa metodologia e a 3.2 ilustra alguns exemplos de *features* para cada dimensão. No entanto, não se trata de um conjunto fechado de métricas e *features*, sendo possível expandí-las conforme necessário.

Tabela 3.1: Métricas descritas por dimensão (DA, DB, DC, DG e DT)

Dimensão	Métricas	Descrição
Área (DA)	DA-1	Para cada área profissional, soma-se o número de vagas de trabalho disponíveis.
	DA-2	Para cada área profissional, somamos o número de candidatos.
	DA-3	Soma-se o número de aplicações separadas por área profissional.
	DA-4	O número de candidatos é dividido pelo número de vagas para cada área profissional.
Benefícios (DB)	DB-1	Para cada item, soma-se o número de vagas de trabalho disponíveis.
	DB-2	Para cada item, soma-se o número de candidatos.
	DB-3	Somamos o número de aplicações separadas por item.
	DB-4	Divide-se o número de candidatos pelo número de vagas em cada item.
Competência (DC)	DC-1	Para cada competência, somamos o número de vagas disponíveis.
	DC-2	Para cada competências, somamos o número de candidatos.
	DC-3	Somamos o número de aplicações separadas por competência.
	DC-4	O número de candidatos é dividido pelo número de vagas de trabalho em cada competência.
Geográfica (DG)	DG-1	Para cada região geográfica, soma-se o número de vagas de trabalho disponíveis.
	DG-2	Para cada região geográfica, somamos o número de candidatos.
	DG-3	Soma-se o número de aplicações separadas por região geográfica.
	DG-4	O número de candidatos é dividido pelo número de vagas de trabalho em cada região geográfica.
Temporal (DT)	DT-1	Soma do número de vagas de trabalho em cada unidade temporal.
	DT-2	Soma do número de candidatos em cada unidade temporal.

Tabela 3.2: Exemplo de features por dimensão

Dimensão	Exemplo de Features
Área (DA)	área geral, subarea, expertise
Benefícios (DB)	salário, horas trabalhadas, ticket refeição
Competência (DC)	skills, idiomas falados, trabalhos anteriores
Geográfica (DG)	cidade, estado, país, região
Temporal (DT)	dia, mês

3.2.1 Dimensão Área (DA)

O objetivo dessa dimensão é avaliar os dados sob a perspectiva das áreas de atuação. A meta é compreender quais setores do mercado possuem mais vagas, quais possuem mais candidatos à procura de uma vaga, ou ainda, quais apresentam excesso ou escassez de oportunidades. Essas informações são importantes para que empresas de recrutamento possam estabelecer estratégias para atender melhor candidatos que atuam em áreas desfavorecidas de oportunidades, bem como ajudar recrutadores a conseguirem colaboradores em áreas com poucos profissionais. Por meio dela, pode-se estabelecer diferentes estratégias de recomendação de acordo com as características das áreas.

(DA-1) Distribuição de vagas por área: fornece uma visão sobre as áreas profissionais com mais e menos vagas.

(DA-2) Distribuição de demandas por área: apresenta as áreas profissionais com mais e menos profissionais em busca de uma realocação no mercado.

(DA-3) Distribuição de aplicações por área: essa métrica tem como objetivo evidenciar quais são as áreas profissionais mais requisitadas por candidatos.

(DA-4) Distribuição candidato/vaga por área: visa entender se existem áreas profissionais com muitas vagas e poucas demandas ou vice-versa.

3.2.2 Dimensão Benefícios (DB)

O objetivo é avaliar as características que compõem uma oferta de emprego. Do ponto de vista do candidato, refere-se aos itens que ele espera encontrar em uma vaga anunciada. Recursos como *horário de trabalho*, *tipo de contrato*, *salário*, *nível de carreira* fazem parte dessa dimensão. A partir das informações fornecidas por essa dimensão, juntamente com as análises da DC (Dimensão Competências), as empresas de recrutamento podem ajudar candidatos e recrutadores a registrar vagas de emprego e currículos com maior probabilidade de sucesso, abordando assim o PMS.

(DB-1) Distribuição de vagas por item: mostra como são as vagas para um tópico específico.

(DB-2) Distribuição de demandas por item: mostra como são as demandas de um tópico específico.

(DB-3) Distribuição de aplicações por item: mostra como os candidatos estão interessados de acordo com um tópico.

(DB-4) Distribuição de candidato/vaga por item: mostra como a distribuição entre currículos e vagas dos candidatos está de acordo com um tópico.

3.2.3 Dimensão Competências (DC)

Essa dimensão considera questões relacionadas às habilidades práticas descritas no currículo do candidato ou exigidas para uma vaga específica. Como alguns exemplos, temos *atendimento ao cliente*, *usinagem*, *programação de computadores* ou mesmo uma habilidade muito específica, como *programação de computadores na linguagem Python*.

(DC-1) Distribuição de vagas por competência: visa entender as vagas de emprego de acordo com as competências exigidas.

(DC-2) Distribuição de demandas de competência: mostra as habilidades cada vez mais comuns entre os profissionais.

(DC-3) Distribuição de aplicações por competência: apresenta em quais competências há mais aplicações a vagas para identificar quais são de maior interesse dos candidatos.

(DC-4) Distribuição de candidato/vaga por competência: mostra quais são as competências mais comuns e menos comuns no mercado. A identificação de habilidades raras pode ajudar os recrutadores a descreverem melhor suas vagas.

3.2.4 Dimensão Geográfica (DG)

Para essa dimensão, consideramos a região geográfica dos currículos e as vagas de emprego. Pode-se considerar diferentes granularidades geográficas, em escala regional, estadual e municipal, ou mesmo qualquer outra granularidade suportada pelo conjunto de dados. A partir dessas informações, uma empresa de recrutamento on-line pode, por exemplo, identificar regiões com oportunidades para expansão de atividades.

(DG-1) Distribuição de vagas por região geográfica: essa métrica apresenta uma visualização da distribuição de vagas de acordo com a região geográfica.

(DG-2) Distribuição de demandas por região geográfica: essa métrica apresenta uma visão da distribuição de demandas de acordo com a região geográfica.

(DG-3) Distribuição de aplicações por região geográfica: Esta métrica mostra uma visão da distribuição de aplicações de acordo com a região geográfica.

(DG-4) Distribuição de candidato/vaga por região geográfica: essa métrica visa identificar regiões com os maiores e menores relacionamentos entre candidato e vaga.

3.2.5 Dimensão Temporal (DT)

Essa dimensão avalia a dinâmica temporal das vagas e das demandas.

(DT-1) Vagas ao longo do tempo: descreve a variação das vagas de emprego ao longo do tempo.

(DT-2) Demandas ao longo do tempo: descreve a variação das demandas de trabalho ao longo do tempo.

3.3 Síntese do Capítulo

Neste capítulo apresentamos formalmente o problema de ausência de oportunidades por longos períodos, que candidatos e recrutadores podem sofrer, o qual nos referimos como *Problem of Matching Scarcity* (PMS). O PMS pode ocorrer devido à grande dificuldade de quantificar *matchings* para diversos itens distintos que possuem descrições detalhadas. Para isso iniciamos com a definição dos conceitos relacionados ao PMS, sendo eles: *Screening feature*; *Appendant feature*; *Matching*; e Escassez. Após isso definimos PMS como: *Cenários nos quais uma vaga v ou um CV c apresenta escassez de oportunidades por um período contínuo de tempo maior que um threshold T_M.*

Após isso apresentamos uma metodologia que pode ser utilizada (mas não se limitando) para caracterizar o PMS. Trata-se de uma metodologia composta por cinco dimensões de análise: Dimensão Área; Dimensão Benefícios; Dimensão Competências; Dimensão Geográfica; e Dimensão Temporal. Cada dimensão possui diferentes métricas. Essas dimensões podem ser combinadas iterativamente, para responderem perguntas relacionadas ao domínio de recrutamento. Para isso inicia-se a análise levantando-se uma questão de negócio por parte de um profissional de recrutamento on-line. Essa questão é então incorporada na metodologia com o objetivo de resolver um problema por meio da resposta obtida. No próximo capítulo definimos as estratégias de identificação de currículos/vagas que sofram do PMS e como identificar *Appendant features* nesses itens. Tambéms apresentamos como aplicar operações de edição na descrição dos currículos e vagas, através de uma das seis estratégias propostas para mitigar o PMS.

Capítulo 4

Solucionando o PMS

Neste capítulo, apresentamos nossas propostas de estratégia para 1) identificar currículos e vagas de empregos que sofrem com o problema de PMS; (2) discriminar *Appendant* e *Screening features*; (3) mitigar o PMS.

4.1 Identificando Itens Escassos

Conforme definido no Capítulo 3, a identificação de currículos e vagas que sofrem do PMS envolve várias etapas. A primeira é encontrar o número mínimo de *matchings* μ usadas para definir quais currículos e vagas sofrem com a escassez. Essa é uma etapa complexa porque o valor de μ é subjetivo e não é diretamente mensurável. Além disso, conforme apresentado no Capítulo 3, um *matching* é definido a partir do valor de similaridade entre os conjuntos de *features* F_v e F_c, extraídos respectivamente de uma determinada vaga v e CV c, acima do *threshold* τ. No entanto, nesta etapa, não sabemos o valor de τ. Assim, para determinar o valor de μ, propomos avaliar o número de contatos que cada CV/vaga obteve no passado. Trata-se de uma suposição consistente uma vez que os contatos representam uma demonstração explícita de interesse de ambos os lados. Além disso, se um item específico obteve n contatos, significa que o sistema de recomendação usado identificou pelo menos n *matchings* para ele.

Nossa solução é baseada em **Inferência Bayesiana** Dempster [2008]. Assim, a partir da observação de como o domínio se comportou no passado, tentamos inferir como ele se comportará no futuro. Nossa variável de observação são os contatos para cada CV/vaga. Essa observação é feita distribuindo as probabilidades de contato entre os currículos/vagas existentes no passsado. Assim, nossa estratégia de inferência é baseada em uma heurística gulosa que procura o valor de μ que representa o comportamento dos dados anteriores. Inicialmente, atribuímos μ como o valor mais baixo observado

no domínio (ou seja, 2 contatos) e obtemos o total de itens considerados escassos. μ aumenta gradualmente e, a cada aumento, é feita uma avaliação do impacto sobre o número de itens identificados como sofrendo de escassez. Quando um determinado aumento no valor de μ não causa mais diferenças significativas no número de itens apontados como escassos, μ é o valor imediatamente anterior. Isso significa que, mesmo que valores mais altos de μ sejam usados, itens considerados escassos permanecem inalterados. Ao avaliar novos currículos/vagas, as informações de contato não estarão presentes. Assim, assumimos que o valor μ obtido de contatos anteriores pode ser usado como *matchings* (itens com similaridade acima de τ) na avaliação de novos currículos/vagas.

Dada o valor de *mu*, podemos utilizá-lo o para se inferir o valor do *threshold* τ que define a similaridade mínima para a ocorrência de um *matching*. O valor de τ é desconhecido e também não é diretamente observável e, consequentemente, complexo de ser calculado. Nossa proposta consiste em encontrar o valor de τ utilizando uma estratégia de **Maximum Likelihood Estimation (MLE)** MENG & RUBIN [1993]. Dado o valor mínimo μ, calculado com base em contatos, a ideia é encontrar o valor de τ que gere uma distribuição de *matchings* (correspondência entre cv e vaga por meio de similaridade) que mais se aproxima da distribuição observada apriori. A estratégia se inicia definindo τ com o menor valor possível (i.e. $tau = 0.10$). De maneira gulosa, o valor de τ é incrementado. A cada novo valor de τ, avaliamos a probabilidade de ocorrência de μ *matchings* na coleção de itens. Quando, para um determinado valor de *tau*, a probabilidade de μ *matchings* diminuir, significa que a aproximação das distribuições de probabilidade observadas apriori e as calculadas por meio de *tau* também diminuiu. Assim, o τ anterior será o valor escolhido.

Tendo definido os valores de τ e μ, propomos uma estratégia para classificação binária de cvs/vagas como sofrendo ou não com o PMS. Essa estratégia é ilustrada pelo Algoritmo 1. Para um conjunto de cvs (C) e vagas (V) que pretendemos avaliar, definimos o período de tempo T no qual o quantitativo de *matchings* será considerado. Por exemplo, para cada cv $c \in C$ (ou $v \in V$), avaliamos quantas vagas (cvs) apresentaram uma similaridade acima de *tau* no período T. Se esse valor for inferior ao *mu*, o cv (vaga) é então considerado sofrer do PMS.

4.2 Identificando *Appendant Features*

Uma vez definido o processo de se encontrar cvs e jobs que sofrem do PMS, a próxima etapa é trabalhar em estratégias que visem mitigar esse problema. Para isso, primeiramente precisamos identificar as características que podem estar relacionadas a escassez de matching desses cvs/jobs. Conforme apresentado no Capítulo 3,

Algorithm 1 Classificador Binário de Escassez

ENTRADA: CV, vagas
SAÍDA: True/False para cenário de escassez

```
μ, τ = getParametros()
totalMatchings = 0, escassez = False
for all vaga ∈ vagas do
    similaridade ← getSimilaridade(CV, vagas)
    if similaridade >= τ then
        totalMatchings ← totalMatchings + 1
    end if
end for
if totalMatchings < μ then
    escassez ← True
end if
Return escassez
```

consideramos que currículos e vagas têm características que podem ser divididas em dois grupos distintos, *Screening* e *Appendant features*. As *screening features* são caracterizadas pela interpretabilidade, popularidade e poder discriminativo. Nesse trabalho, propomos um algoritmo genérico, ilustrado no Algoritmo 2, que visa considerar esses três aspectos para discriminar *screening* e *appendant features*. Conforme podemos observar, o primeiro passo do algoritmo é, a partir de todas as *features* encontradas nos cvs/vagas, filtrar aquelas de maior interpretabilidade e poder discriminativo. Nesse trabalho, instanciamos a filtragem selecionando para o próximo passo apenas features reconhecidas para o domínio, utilizando para isso um conjunto de *thesaurus* e ontologias fornececidas pela Catho. Feito isso, o algoritmo genérico propõe uma estratégia simplificada considerando diretamente a popularidade das características restantes. Nossa premissa é que características muito utilizadas para descrever cvs/vagas são, por consequência, também mais interpretáveis (i.e. muitos usuários a reconhecem como conceitos importantes) e apresentam maiores poderes discriminativos (são observadas com maior frequência em *matchings* de cvs e vagas). A partir disso, nossa estratégia propõe considerar as *features* restantes que apareçam acima de um *threshold* específico como *screening features*. Instanciamos esse passo do algoritmo traçando a distribuição referente à frequência de menções de cada *feature* na coleção de dados e, a partir disso, utilizar a mediana da distribuição como o *threashold*, ou seja, todas as features que se encontram no conjunto acima da mediana, são classificadas *screening features*. Todas as features que não são inseridas como screening, são consideradas *Appendant features*. É importante mencionar que outras dimensões podem ser adicionadas à esse processo,

como por exemplo, encontrar *screening* e *appendant features* por área profissional ou região geográfica. Trata-se de uma primeira proposta de estratégia. Conforme veremos mais a frente, pretendemos aprimorar essa estratégia considerando mais diretamente a interpretabilidade e poder discriminativos das features Cross [2014].

Algorithm 2 Classificador de Features

ENTRADA: CV, matchingsPossiveis, conjuntoFeatures
SAÍDA: Screening Feafures, Appendant Features

ListaAppendantFeatures = $listaVazia()$
$dimensoes \leftarrow filtro(conjuntoFeatures)$
$distribuicao \leftarrow getDistribuicao(matchingsPossiveis, dimensoes)$
for all $(feature \in CV) and (feature \in dimensoes)$ **do**
  if $acimaThreshold(feature, distribuicao)$ **then**
    listaScreeningFeatures.$insert(feature)$
  end if
end for
Return $(listaScreeningFeatures, (conjuntoFeatures - listaScreeningFeatures))$

4.3 Aplicando Operações de Edição nas Descrições de Itens

Nesta seção, apresentamos estratégias para mitigar o PMS. Propomos seis soluções distintas, duas delas baseadas na remoção de *features*, duas baseadas na troca de *features* por outras semanticamente semelhantes às removidas e duas baseadas na troca de *features* por outras semanticamente semelhantes ao conjunto das outras *features* presentes no CV/vaga analisados. Para as quatro estratégias semelhantes baseadas na troca de *Appendant features*, propomos construir representações baseadas em *embeddings* Toutanova et al. [2015] para cada *feature* usando o algoritmo Word2Vec Mikolov et al. [2013]. O Word2Vec usa um modelo de n-grama com lacunas (Skip-gram) para prever palavras (ou seja, *features*) que compõem o contexto (descrições de currículos e vagas) de uma determinada *feature* de entrada, capturando relacionamentos de co-ocorrência entre as *features* e seu contexto. As *features* que ocorrem em contextos semelhantes tendem a ter representações vetoriais semelhantes, ou seja, *embeddings*, permitindo relacionamentos semânticos mais ricos e produzindo boas representações do significado geral de uma *feature*. Como veremos abaixo, o que diferencia cada uma das quatro estratégias é como as representações de *embeddings* são construídas e como são

feitos os cálculos de similaridade. Nas quatro estratégias, para cada *features* candidata a ser substituída, encontramos as 10 *features* mais semelhantes. Antes de realizar a troca de *features*, estabelecemos duas regras básicas que devem ser respeitadas: (1) a *feature* a ser inserida no CV/vaga deve ter uma semelhança acima do limite mínimo; e (2) a nova *feature* não pode já estar contido no CV/vaga. Se nenhuma das 10 principais *features* atender a essas duas condições, a *feature* candidata não será substituída.

- **Random Feature Removal (RR):** remove *features* aleatoriamente. O número de *features* a serem removidas é definido de acordo com o total de *Appendant features* encontradas no CV/vaga. Essa estratégia tem como objetivo servir de linha de base para comparação com outras estratégas. O objetivo desta linha de base é demonstrar que uma remoção sem critério de *features* não leva a descrições aprimoradas dos currículos/vagas.

- **Appendant Feature Removal (AR):** esse método remove cada uma das *Appendant features* do currículo/vaga. O objetivo é melhorar o CV/vaga, removendo os pontos fracos e, consequentemente, ressaltar os pontos fortes dos mesmos.

- **Appendant Feature Exchange for Screening Features (AES):** realiza a troca de cada *Appendant feature* de um CV/vaga por uma *Screening feature* semanticamente semelhante à removida. Para isso, utiliza *embeddings* treinados com o conjunto de todas as habilidades provenientes de currículos e vagas. O objetivo é encontrar *features* semelhantes às que serão removidas, porém mais interessantes do ponto de vista de um *matching* entre um currículo e uma vaga.

- **Appendant Feature Exchange for Job Screening Features (AEJS):** esse método troca cada *Appendant feature* por uma *Screening feature* semanticamente semelhante a removida. Para isso, usamos *embeddings* treinados com o conjunto de habilidades que aparecem apenas nas descrições de vagas. A intuição por trás dessa estratégia é que a substituição de uma *Appendant feature* por uma *Screening feature* semanticamente próxima e presente nas vagas tende a melhorar a descrição de um currículo. A premissa, é que as vagas são melhor descritas do que os currículos.

- **Appendant Feature Exchange for Screening Features Contextualized (AESC):** esse método troca cada *Appendant feature* por uma *Screening feature* semanticamente semelhante ao contexto do CV/vaga de forma geral. Nesse caso, mapeamos todo o currículo/vaga no espaço dimensional das representações do *embedding* e calculamos as *features* mais semelhantes a todos os currículos/vagas

analisados. O *embedding* é treinado com um conjunto completo de habilidades presentes nos currículos/vagas. A intuição por trás dessa estratégia é que a adição de recursos contextualizados ao CV/vaga pode melhorar sua descrição, mantendo uma maior consistência com a descrição original.

- **Appendant Feature Exchange for Job Screening Features Contextualized (AEJSC):** esse método é uma combinação dos métodos AEJS e AESC, pois troca cada *Appendant feature* por uma *Screening feature* semanticamente semelhanteao contexto do cv/vaga de forma geral. Para isso, treinamos nosso *embedding* usando o conjunto de habilidades presentes nas descrições de vagas, bem como no AEJS, e a definição de *features* semelhantes é feita com base em toda a descrição de currículo/vaga, bem como no AESC. Assim, essa estratégia visa inserir *features* mais bem descritas presentes em Vagas e, além disso, consistentes com a descrição do CV/vaga alvo.

4.4 Síntese do Capítulo

Nesta capítulo iniciamos apresentando uma forma para identificar CVs e vagas que sofram de escassez. Para isso definimos duas variáveis: o número de *matchings* máximo μ para que um CV/vaga seja considerado como presente em um cenário de escassez; e o *threshold* mínimo de similaridade τ entre um CV e uma vaga para que seja considerado um *matching* entre os dois. CVs e vagas que tiverem menos de μ *matchings* são considerados como sofrendo com o PMS. Propomos uma solução baseada em Inferência Bayesiana para encontrar o valor de μ e uma solução que se utiliza **Maximum Likelihood Estimation (MLE)** para encontrar o valor de τ.

Após isso, mostramos como encontrar as *Appendant features* (e por consequência, as *Screening features*). Para isso utilizamos o conjunto de CVs e vagas classificados em cenário de escassez e traçamos a distribuição de suas *features*. Utilizando a mediana da distruição como ponto de corte, classificamos todas as *features* abaixo da mediana como *Appendant* e as acima como *Screening*. O objetivo de encontrar esses conjuntos de *features* foi empregá-las nas estratégias para mitigação do PMS.

Por fim, apresentamos seis estratégias para mitigar o PMS, sendo elas: *Random Feature Removal (RR)*; *Appendant Feature Removal (AR)*; *Appendant Feature Exchange for Screening Features (AES)*; *Appendant Feature Exchange for Job Screening Features (AEJS)*; *Appendant Feature Exchange for Screening Features Contextualized (AESC)*; e *Appendant Feature Exchange for Job Screening Features Contextualized*

(AEJSC). Cada uma delas utilizando premissas distintas com o intuito de melhorarem a descrição dos currículos e vagas de formas distintas.

Capítulo 5

Avaliação Experimental

Neste capítulo, avaliamos a metodologia de caracterização proposta e as estratégias para mitigar o PMS usando dados reais. Iniciamos apresentando a amostra de dados usada em nossas avaliações. Em seguida, analisamos os resultados da metodologia de caracterização, apontando como que os mesmos auxiliam para detectar e mitigar o PMS. Finalmente, discutimos os resultados das seis estratégias de mitigação do PMS descritas no Capítulo 4.

5.1 Catho Database

A Catho, instalada no Brasil, é uma empresa de recrutamento on-line, que faz parte do grupo Seek[1], a maior companhia multinacional no ramo no mundo. A amostra de dados fornecida é dividida entre duas entidades principais: currículos criados pelos candidatos, sendo que um candidato pode criar vários currículos; e vagas de emprego criadas pelos recrutadores. A amostra possui 376.762 instâncias de currículos e 115.955 instâncias de vagas, distribuídos em um período de 13 meses (janeiro a janeiro). Apesar de ser apenas uma amostra, essa coleção fornece subsídios suficientes para a avaliar a metodologia proposta. Por questões de confidencialidade, não podemos divulgar os dados utilizados nestas análises. Primeiramente vamos fornecer uma análise dos dados da Catho e mostrar como a metodologia proposta no Capítulo 3 pode ser utilizada de forma para responder questões levantadas pela área de negócios das empresas de recrutamento. Posteriormente vamos caracterizar o PMS iterando pelas dimensões alcançar resultados mais refinados.

[1] www.seek.com.au

5.2 Caracterização do Domínio

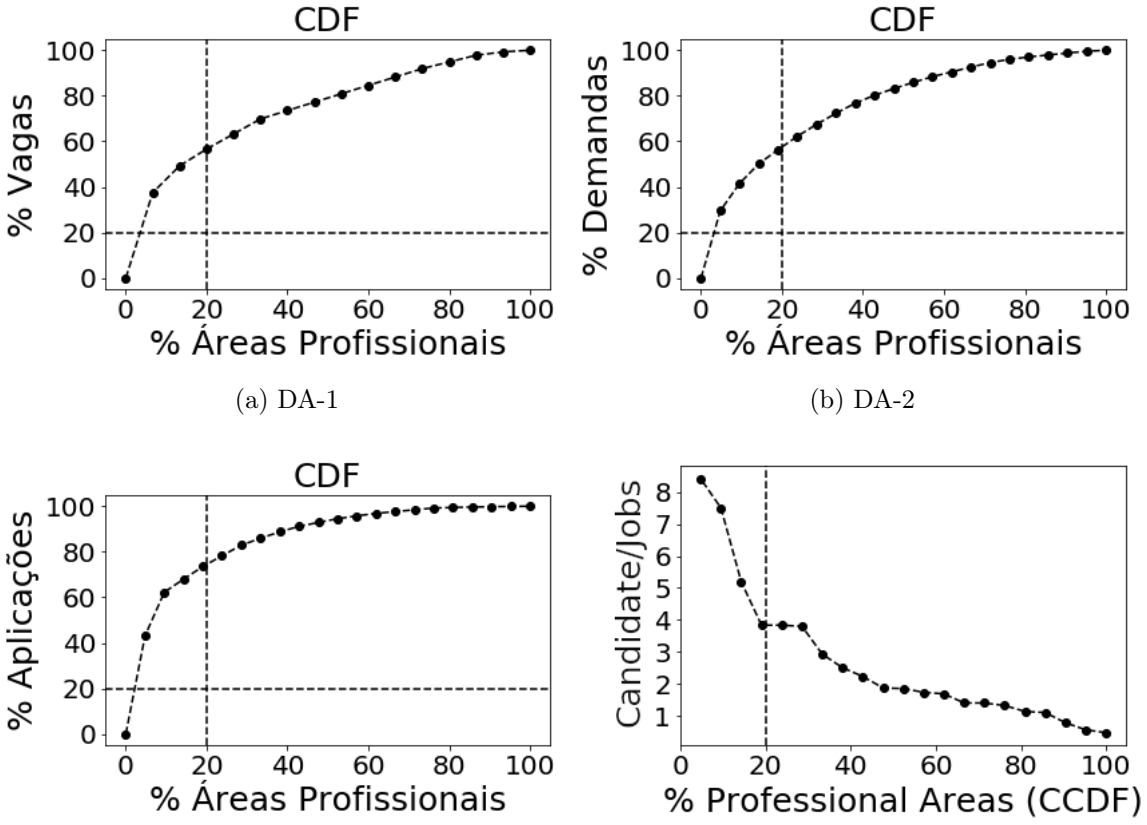
5.2.1 Dimensão Área (DA)

Para essa dimensão provemos uma avaliação mais geral com relação às áreas de atuação cadastradas na Catho com intuito de entender como estão distribuídas as ofertas e vagas de acordo com a essas áreas. No total são 21 áreas distintas, apresentadas na Tabela 5.6.

Tabela 5.1: Áreas Profissionais

ÁREA	CAND/VAGA	ÁREA	CAND/VAGA
Comercial e Vendas	0,4168835	Financeira	2,1250677
Informática	0,4592044	Telecomunicações	3,0686359
Telemarketing	0,7533921	Artes	3,1655855
Hotelaria e Turismo	0,8380434	Arquitetura	3,1655855
Administração	0,8829341	Engenharia	3,8121250
Saúde	0,9144623	Agricultura	6,1834710
Técnica	1,4000209	Veterinária	6,1834710
Suprimentos	1,6320817	Jurídica	7,0540716
Educação	1,7166620	Comércio Exterior	7,0555555
Marketing	1,7409734	Serviços Sociais	12,7073170
Industrial	2,1132189		

Avaliamos e comparamos as distribuições de vagas (DA-1), de demandas (DA-2) e de aplicações a vagas (DA-3) por áreas profissionais, apresentadas nas Figuras 5.1a, 5.1b e 5.1c respectivamente. Podemos observar que há desequilíbrio de demanda e oferta quanto as áreas. Podemos observar que 20% das áreas são responsáveis por 70% das vagas, 60% das demandas e 70% das aplicações. A partir desses exemplos, temos duas observações importantes: (1) considerando esses 20% de áreas mais "populares"sejam as mesmas em todas as análises, existe um desequilíbrio: 10% a mais de vagas em relação às ofertas; (2) 80% das áreas de atuação recebem pouca atenção, tanto de candidatos quanto de recrutadores. Esse desequilíbrio provoca também diferenças na relação candidato/vaga, conforme podemos observar na Figura 5.6b (detalhado na Tabela 5.6). Diante desse cenário, fica claro que a Catho tem um potencial de crescimento enorme e deve elaborar estratégias que conquistem novos usuários, sejam candidatos ou recrutadores, para cobrir as áreas menos populares. Além disso, o processo de realocação de candidatos de áreas de grande concorrência exigem estratégias de recomendação mais especializadas.



(a) DA-1

(b) DA-2

(c) DA-3

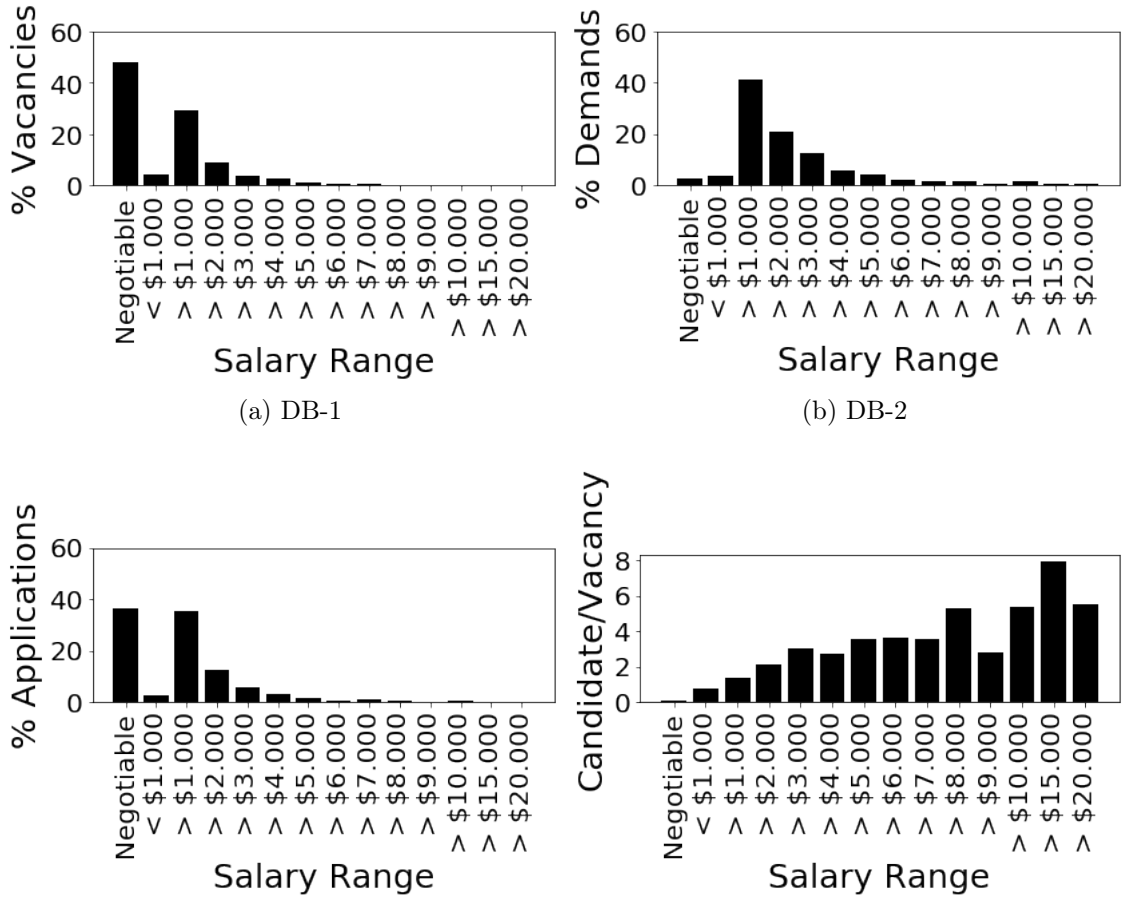
(d) DA-4

Figura 5.1: Dimensão Área (DA).

5.2.2 Dimensão Benefícios (DB)

Para essa dimensão, dentre as várias características descritivas de uma vaga e/ou um currículo, restringimos nossas análises para o item "salário", por questões de espaço. Iniciamos nossas análises avaliando e comparado as métricas DD-1 (Distribuição das vagas por faixa salarial - Figura 5.2a), DD-2 (Distribuição das demandas por faixa salarial - Figura 5.2b) e DD-3 (Distribuição das aplicações por faixa salarial - Figura 5.2c). Podemos observar que 50% das vagas não informam o salário e tratam essa questão como algo definido depois dos primeiros contatos com candidatos. Apesar disso, apenas 35% das aplicações são feitas em vagas com a questão salarial a ser definida posteriormente. Essa é uma informação relevante a ser considerada pelos recrutadores, pois não despertam interesse dos candidatos. Por outro lado, trata-se de uma oportunidade para que a Catho possa melhorar seus SsR. Essa grande quantidade de vagas com salário a combinar acaba por levar há um desalinhamento entre o esperado pelos candidatos e o ofertado pelo mercado. Conforme os salários ficam

mais altos, diminuem as vagas e também as demandas, mostrando que conforme os valores aumentam, existem menos oportunidades, mas também menos candidatos que se enxergam como qualificados para requerer esses salários. Entretanto, avaliando a métrica DD-4 (Distribuição da relação candidato/vaga por faixa salarial - Figura 5.2d) que essa diminuição das vagas e demandas não é proporcional, uma vez que para faixas salariais maiores a relação candidato/vaga aumenta significativamente.



(a) DB-1

(b) DB-2

(c) DB-3

(d) DB-4

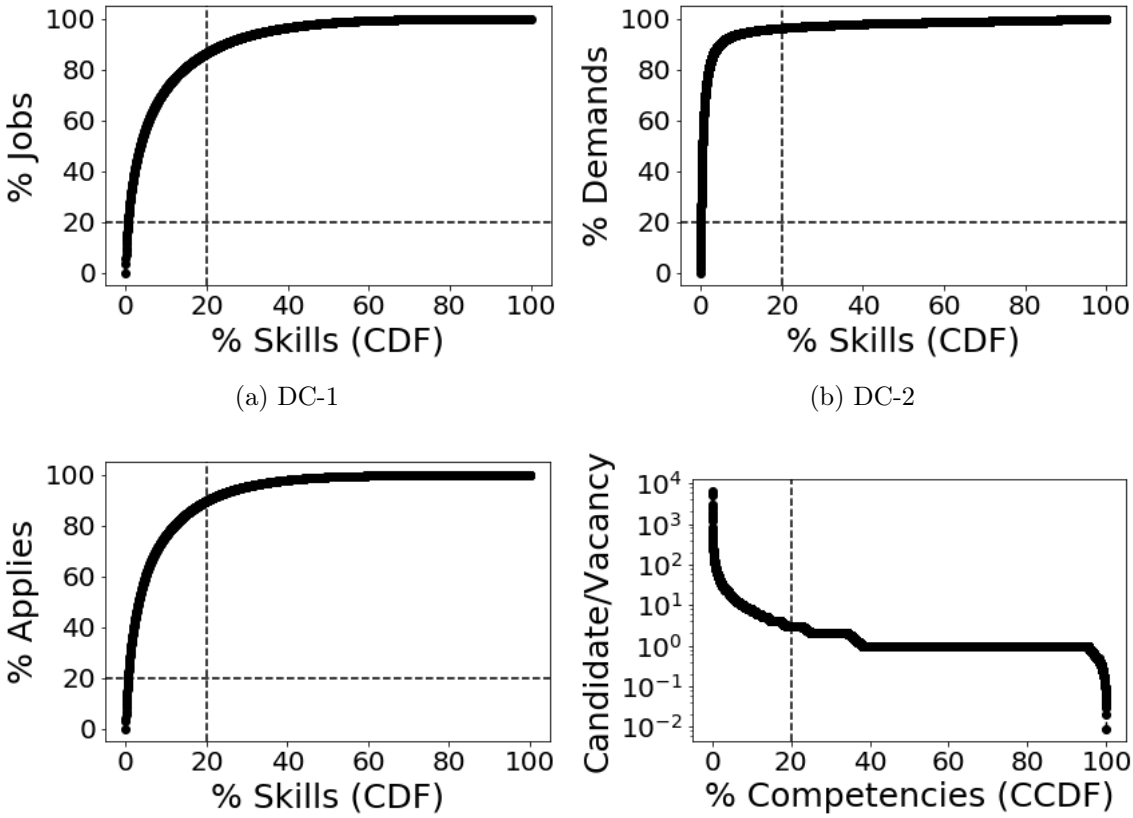
Figura 5.2: Dimensão Benefícios (DB).

Todas essas observações são muito valiosas e mostram que para uma empresa de recrutamento é muito mais difícil realizar a recolocação de candidatos com altas expectativas salariais, visto que a concorrência por eles é maior. Mas isso abre uma oportunidade de adaptar os seus SsR para trabalhar com um leque de faixas salariais desejáveis pelos candidatos. Por exemplo, a partir do currículo de um candidato e suas competências, ranquear esse candidato dentro da área específica e focar em vagas de faixa salariais compatí-

veis com essas competências (quanto maior, melhor), melhorando a qualidade do recrutamento. Do ponto de vista da empresa recrutadora, de forma análoga mas inversa, recrutar um melhor candidato com uma pretensão salarial menor.

5.2.3 Dimensão Competências (DC)

Nessa dimensão realizamos uma avaliação geral sobre as diversas competências (habilidades e/ou requisitos) cadastradas por candidatos e recrutadores. Iniciamos nossas análises por meio das métricas DC-1 (Distribuição de vagas por Competência - Figura 5.3a), DC-2 (Distribuição de Demandas por Competência - Figura 5.3b) e DC-3 (Distribuição de Aplicações por Competência - Figura 5.3c).



(a) DC-1

(b) DC-2

(c) DC-3

(d) DC-4

Figura 5.3: Dimensão Competências (DC).

Avaliando essas métricas, temos que 20% das competências estão presentes em 90% das vagas, o que mostra que grande parte das vagas apresenta os mesmos requisitos básicos, diferenciando-se por meio de poucos requisitos, que por sua vez tendem a

ser bem específicos (raros). Por outro lado, 10% das competências são comuns a 90% dos candidatos e 20% das competências são comuns a 60% das aplicações, o que mostra que existe um conjunto muito grande de habilidades comuns a todos os candidatos. Essas diferenças refletem na relação candidato/vaga entre as competências. Por exemplo, 60% de competências que possuem apenas 1 candidato/vaga. Além disso, 5% das competências encontradas no sistema podem ser consideradas raras, pois possuem menos de um candidato/vaga. Em contrapartida 20% das habilidades possuem mais de um candidato/vaga e aproximadamente 5% delas possuem um número muito alto de candidatos/vaga. Nas Tabelas 5.2 e 5.3 apresentamos as 10 competências mais raras e mais comuns, respectivamente.

Tabela 5.2: Top 10 competências mais raras.

COMPETÊNCIAS	CAND/VAGA	COMPETÊNCIAS	CAND/VAGA
react	0,0072463	redis	0,0333333
node	0,0202020	jenkins	0,0350877
soap	0,0222223	restful	0,0354609
mongodb	0,0229885	esmaltagem	0,0357142
aparo	0,0270270	calefacao	0,0357143

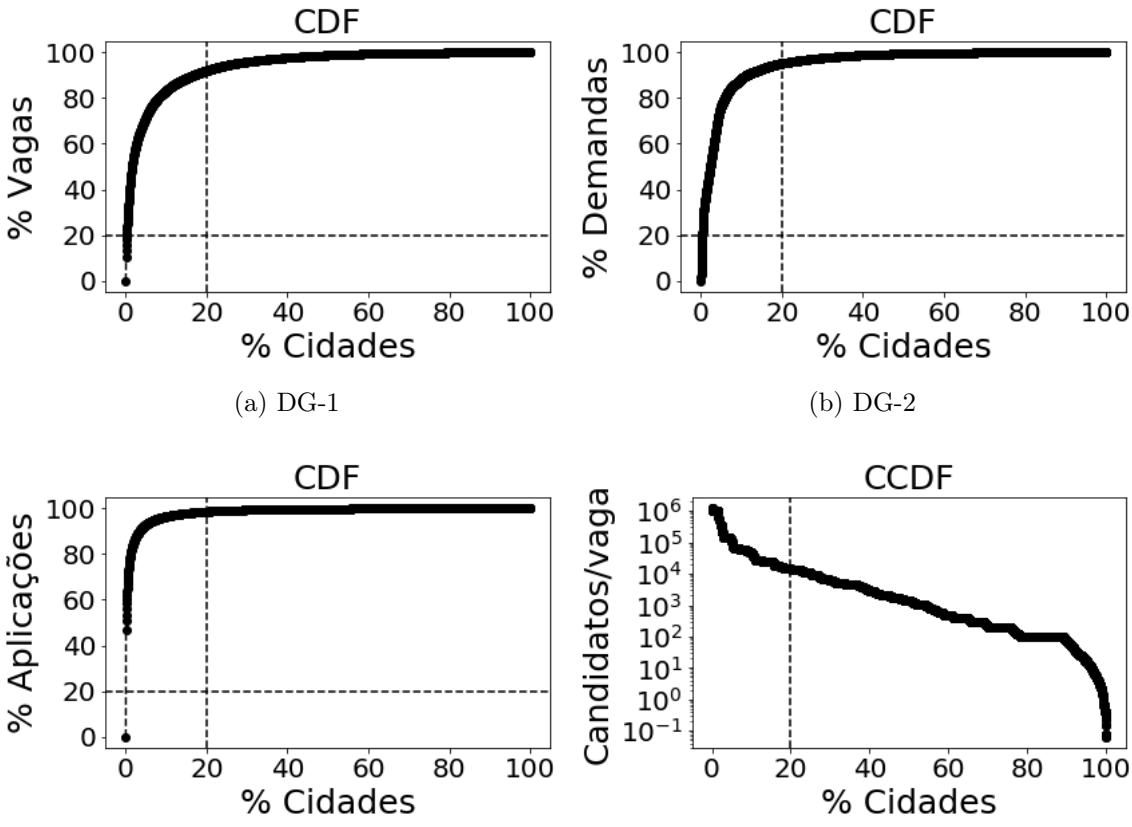
Tabela 5.3: Top 10 competências mais comuns.

COMPETÊNCIAS	CAND/VAGA	COMPETÊNCIAS	CAND/VAGA
educacao medio	20725,0	engenheiro	4625,0
educacao tecnologo	12414,0	sebrae	3751,0
superior completo	9718,0	advogado	2447,0
vender	8087,0	comprador	2369,0
emedio completo	5445,7	at. paciente	2252,0

Todos esses resultados demonstram que, pelo lado do recrutador, utilizar requisitos bem específicos tende restringir significativamente os possíveis candidatos a vaga. Por outro lado, realocar no mercado candidatos com poucas habilidades específicas é um cenário bastante desafiador. Nesse cenário, a Catho poderia criar uma ferramenta de auxílio para os recrutadores, onde a cada competência adicionada como exigência na vaga, indicasse ao recrutador quantos profissionais eles estaria alcançando. Isso levaria os recrutadores a ponderarem sobre suas reais necessidades e flexibilizarem mais suas exigências. Nesse caso, as vagas seriam um pouco mais generalistas e também beneficiariam os candidatos que não possuem as habilidades raras, mas que são competentes e podem até mesmo aprender conforme o trabalho venha a exigir.

5.2.4 Dimensão Geográfica (DG)

Assim como para DA e DC, na Dimensão Geográfica fizemos uma avaliação geral e não por uma região específica. Mais especificamente, utilizamos a granularidade por cidade. Nas Figuras 5.4a, 5.4b, 5.4c e 5.4d apresentamos os resultados relativos às métricas DG-1, DG-2, DG-3 e DG-4 respectivamente. Podemos observar que aproximadamente 90% das vagas estão distribuídas em apenas 20% das cidades. De forma similar, 80% das demandas são para 10% das cidades. Mas são as aplicações para as vagas que geram o maior desequilíbrio com 90% das aplicações concentradas em apenas 10% das cidades. Além disso, temos que 85% das cidades possuem pelo menos 1 candidato/vaga, 20% das cidades possuem pelo menos 100 candidatos/vaga e apenas 5% das cidades possuem pelo menos 1.000 candidatos/vaga. Apenas 15% das cidades possuem mais vagas do que candidatos. Temos um cenário com dois extremos, 20% das cidades possuem um grande número de candidatos/vaga, o que torna a tarefa de conseguir um emprego desafiadora, porém 15% das cidades possuem mais vagas do que emprego.



(a) DG-1

(b) DG-2

(c) DG-3

(d) DG-4

Figura 5.4: Dimensão Geográfica (DG).

Para detalhar melhor essa dimensão, apresentamos nas Tabelas 5.4 e 5.5 as capitais brasileiras com mais e menos vagas, respectivamente. Essas capitais são também as com maior e menor demanda. A cidade de São Paulo é a capital com maior oferta de vagas do país, e também com a maior demanda, seguida por Rio de Janeiro, Belo Horizonte, Curitiba e Porto Alegre, todas concentradas nas regiões sudeste e sul. Por outro lado, as capitais dos estados da região norte, Boa Vista, Rio Branco, Macapá, Porto Velho e Palmas, apresentam as menores ofertas de emprego e também as menores demandas por vagas. Assim, a Catho deve avaliar se a baixa oferta e procura por empregos na região Norte estão relacionadas a questões de desenvolvimento econômico ou a falta divulgação dos serviços oferecidos por ela. Além disso, assim como sugerido na seção anterior, os SsR poderiam trabalhar com um leque maior de possíveis regiões e flexibilizar a recomendação de acordo com esse leque, bem como competências, pretensões salariais entre outras, visando realocar de forma mais rápida e eficaz seus candidatos.

Tabela 5.4: Top 5 capitais com mais ofertas e demandas

CIDADE	OFERTA	CIDADE	DEMANDA
São Paulo	35.959	São Paulo	163.377
Rio de Janeiro	6.574	Rio de Janeiro	20.173
Belo Horizonte	4.690	Belo Horizonte	12.853
Curitiba	3.986	Salvador	8.666
Porto Alegre	3.632	Curitiba	6.414

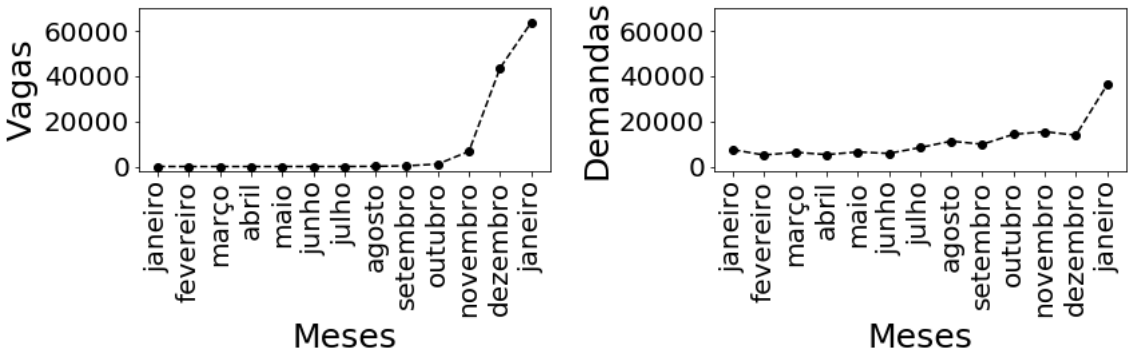
Tabela 5.5: Top 5 capitais com menos ofertas e demandas.

CIDADE	OFERTA	CIDADE	DEMANDA
Boa Vista	82	Boa Vista	43
Rio Branco	98	Rio Branco	70
Macapá	99	Macapá	80
Porto Velho	145	Porto Velho	109
Palmas	213	Palmas	163

5.2.5 Dimensão Temporal (DT)

Para realizar as análises da dimensão temporal, consideramos todo o período de dados disponível (janeiro de 2017 a janeiro de 2018), utilizando uma granularidade mensal. Nas Figuras 5.5a e 5.5b apresentamos os resultados referentes às métricas DT-1 e DT-2, respectivamente. Durante todo ano observamos que o número de vagas é inferior ao número de demandas por parte dos candidatos, com exceção dos últimos meses do ano. Nesse caso, tanto a oferta de vagas quanto a demanda por elas sobrem um aumento em função das contratações temporais que normalmente acontece com as datas

festivas. Para se chegar a conclusões mais objetivas, outras iterações pela metodologia precisariam ser realizadas. Por exemplo, dado esse período de pico de vagas, verificar em quais regiões e áreas de concentram, as competências requeridas, etc. De toda maneira, fica evidente que a empresa de recrutamento possui dois desafios distintos ao longo do ano: períodos de maior demanda e períodos de mais vagas e seus SsR devem considerar essas questões ao realizar recomendações. Na próxima seção apresentamos um exemplo em como a metodologia proposta pode ser utilizada iterativamente.



(a) DT-1 (b) DT-2

Figura 5.5: Dimensão Temporal (DT).

5.3 Caracterizando o PMS

Conforme descrito na Seção 3.2, a metodologia proposta é um processo iterativo orientado a perguntas que nos permite estruturar e sistematizar problemas relevantes de domínios reais. Para caracterizar o PMS, aplicamos essa metodologia em nossa amostra de dados, levantando questões de negócios relacionadas ao PMS e considerando três perspectivas complementares: proprietário do serviço (ou seja, Catho), candidato e recrutador.

5.3.1 Perspectiva da Catho

Ao considerar o proprietário do serviço, em geral, estamos interessados em entender o desequilíbrio entre demandas e oferta de vagas de emprego ao longo do tempo. Por isso, analisamos as seguintes questões comerciais.

Questão 1: *Como as demandas, vagas, e aplicações são distribuídos de acordo com as diferentes área na amostra?*

Análise: Avaliamos e comparamos as distribuições de vagas (DA-1), demandas (DA-2) e aplicação a vagas (DA-3) por áreas profissionais, mostradas na Figura 5.6a.

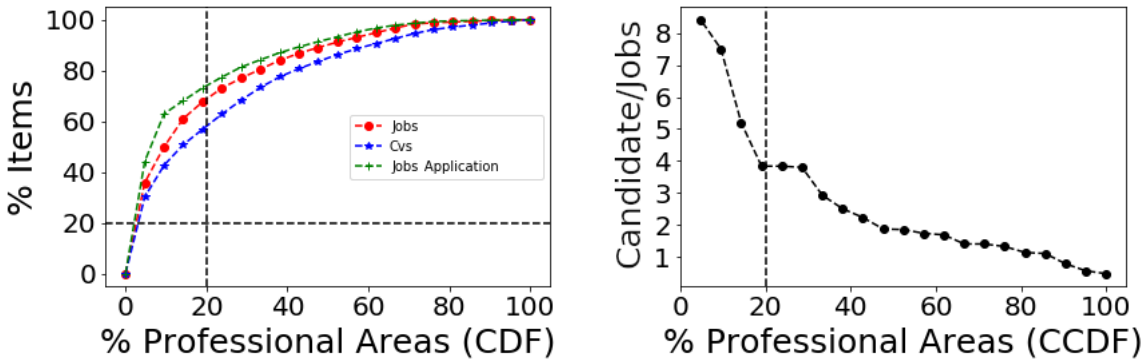
Informação: Na Figura 5.6b, apresentamos a distribuição candidato/vaga (DA-4). Percebe-se que há um desequilíbrio entre demandas e ofertas de emprego em relação às candidaturas nas áreas. Em nossa amostra, 20% das áreas são responsáveis por 60% das vagas, um pouco menos de 60% das demandas, e ela são responsáveis por 70% das aplicações. Também temos 80% das áreas são menos ativas, tanto por candidatos quanto por recrutadores. Esse desequilíbrio também causa diferenças na relação candidato/vaga, como pode ser visto na Figura 5.6b e Tabela 5.6. Nelas, percebe-se que 20% das áreas profissionais possuem maior número de candidatos/vaga, variando entre oito e quatro candidatos. As outras áreas profissionais apresentam um número mais balanceado de candidatos por vaga, variando entre quatro (4) e um pouco menos que (1) candidato.

Conhecimento: Cenários semelhantes aos observados em nossa amostra apresentam potencial de crescimento, por meio de estratégias que atraem novos usuários (ou seja, candidatos ou recrutadores), para cobrir áreas menos populares. Estratégias de marketing, por exemplo, devem agir ativamente para diminuir a lacuna nessa distribuição, impedindo que algumas áreas tenham poucas vagas e algumas outras pouco candidatos. Também é necessário tratar diferentes áreas de maneiras diferentes. Essa análise mostra que existem comportamentos distintoss na amostra usada e deve-se também trabalhar a qualidade da recomendação para previnir que candidatos/recrutadores de algumas áreas sofram com o PMS.

Tabela 5.6: Áreas Profissionais

Área	Candidato/ Vaga	Área	Candidato/ Vaga
IT	0,4732438	Marketing	1,8715728
Business	0,5635589	Industrial	2,2311015
Telemarketing	0,7829192	Financial	2,5044687
Administration	1,1037627	Telecommunication	2,9278688
Hotel Business	1,1369294	Engineering	3,7973395
Health	1,3223934	Agriculture	3,8365758
Arts	1,4028185	Veterinary	3,8365758
Architecture	1,4028185	Social Services	5,1929824
Supplies	1,6877085	International Business	7,5000000
Technical	1,7356205	Law	8,4136125
Educational	1.8529023		

Questão 2: *Como é a distribuição de vagas e demandas ao longo de um mês?*



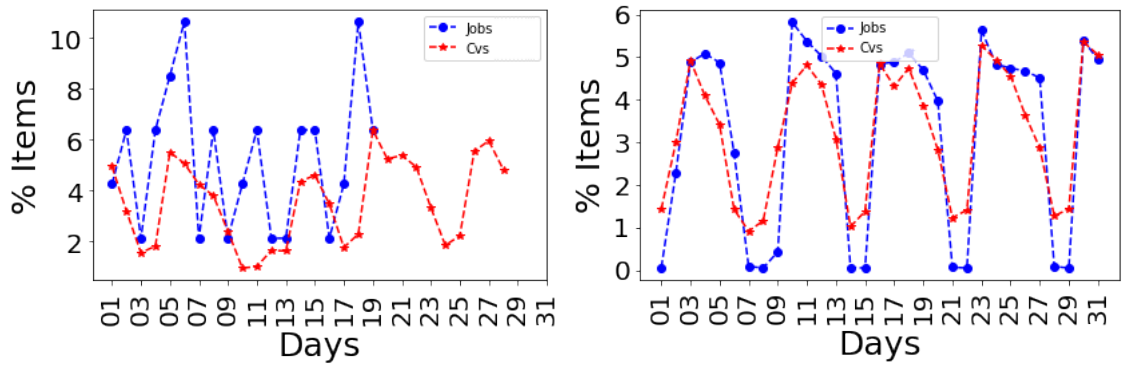
(a) Métricas DA-1, DA-2 e DA-3 (b) Métrica DA-4

Figura 5.6: Dimensão Área (DA)

Análise: Para realizar as análises da Dimensão Temporal, consideramos dois meses distintos, fevereiro e julho, utilizando granularidade diária.

Informação: Nas figuras 5.7a e 5.7b, apresentamos a distribuição das vagas de emprego referentes às métricas DT-1 e DT-2 para Fevereiro e Julho, respectivamente. Como pode-se perceber, na DT-1 de julho, existem alguns períodos de inatividade, onde fica claro o comportamento do recrutamento de funcionários, que tendem a não trabalhar nos fins de semana. Em fevereiro, no entanto, onde ocorreu a festa brasileira carnaval, é possível notar que entre os dias 07 e 15 o número de inscrições de empregos foi menor do que nos outros dias. Também podemos ver os resultados das demandas, referentes à métrica DT-2. Ao contrário dos recrutadores, os candidatos tendem a permanecer levemente ativos nos fins de semana e feriados, como o carnaval. Nos dois casos, as ofertas de emprego e as demandas apresentam uma queda sazonal devido aos finais de semana e feriados.

Conhecimento: É perceptível que a empresa de recrutamento enfrenta desafios relacionados à sazonalidade de vagas e demandas. Avaliando de forma geral, ficou claro que, no período festivo avaliado, tanto a demanda quanto a oferta de vagas de emprego tendem a ser menores. Porém, em outros períodos, como dezembro, por causa do Natal, pode haver um aumento de contratações, principalmente no setor de vendas e serviços. Com mais iterações, analisando outros períodos e outras áreas, é possível entender como o comportamento dos usuários muda ao longo do tempo. Essas variações ao longo do tempo levam candidatos e recrutadores a enfrentarem momentos de escassez devido a variação natural no comportamento de cada setor. Estratégias que tratem essa essa escassez sazonal podem ser úteis para previnir o PMS.



(a) Fevereiro (b) Julho

Figura 5.7: Dimensão Temporal (DT).

5.3.2 Perspectiva do Candidato

Assumimos que os candidatos que enfrentam o PMS possam estar mais interessados em obter mais conhecimento sobre como aumentar suas chances de serem contatados no sistema. Também podem estar interessados em entender quais as faixas salarias que poderiam se candidatar na região desejada, de acordo com suas habilidades.

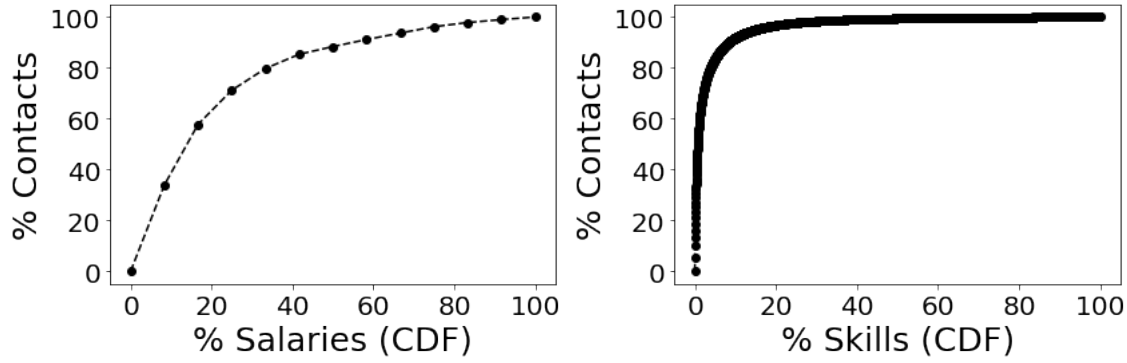
Questão 1: *Como os candidatos do setor de tecnologia da informação deveria preencher o salário e habilidades para que tenham mais contatos de recrutadores?*

Análise: Para esta análise, em primeiro lugar, iteramos a metodologia usando a Dimensão da Área 3.2.1 considerando o setor de TI. Em seguida, realizamos duas análises diferentes: uma focada no salário e outra focada nas habilidades. Portanto, iteramos mais uma vez a Dimensão Benefícios 3.2.2 considerando o item Salário e analisando o número de contatos que o candidato de TI recebe, de acordo com cada faixa salarial desejada. Também fizemos uma segunda análise, iterando uma segunda vez pela Dimensão Competências 3.2.3 avaliando os contatos de acordo com as habilidades descritas no currículo.

Informação: A Figura 5.8a mostra que 20% das faixas salariais representam aproximadamente 60% dos contatos dos recrutadores. A Figura 5.8b apresenta uma curva com 5% das habilidades possíveis, respondendo por 80% de todos os contatos. Isso mostra que, embora possa-se observar várias faixas salariais e várias habilidades distintas, os contatos são mantidos em torno de alguns conjuntos muito específicos. Esses dados sugerem que, para aumentar as chances de contato dos recrutadores, o candidato deve conter pelo menos algumas das habilidades mais importantes ou faixa salarial.

Conhecimento: A principal aplicação seria usar essas informações na criação de

um sistema de ajuda que faça recomendações durante o registro dos currículos (recomendação pré-registro). Este sistema pode indicar a um candidato quais as chances de ser contatado por um recrutador de acordo com o salário e as habilidades que ele possui, por exemplo. O sistema também poderia perguntar quais habilidades relevantes o candidato possui. Vamos assumir o seguinte cenário, o candidato adiciona ao seu currículo a habilidade *Especialista em J2EE*. Essa é uma habilidade muito específica e atinge um conjunto menor de possíveis posições de trabalho. Esse sistema de recomendação pré-registro pode sugerir a troca dessa habilidade por *Programador Java*. Ao mesmo tempo, notificaria o candidato qual porcentagem de vagas seria alcançada com a alteração sugerida, apresentando um tipo de *feedback* ao candidato em tempo real. Essa ferramenta também pode sugerir alterações nas descrições das vagas. Os setores de mercado que possuem muitas habilidades técnicas, como TI, podem se beneficiar dessa ferramenta. Além disso, a estratégia de substituir *features* muito específicas por outras mais abrangentes pode ser uma forma eficaz de tratar o PMS.



(a) Contatos por Faixa Salarial (b) Contatos por Habilidade

Figura 5.8: Análise de contatos por faixa salarial e habilidades

Questão 2: *Como as habilidades exigidas e os salários oferecidos variam de região para região ao considerar o setor de TI?*

Análise: Nesta análise, iteramos inicialmente através da Dimensão da área 3.2.1 selecionando os dados de acordo com a área de TI. A partir desse ponto, fizemos duas análises diferentes. Para a primeira iterada pela Dimensão Geográfica 3.2.4, consideramos as cidades de São Paulo, Rio de Janeiro e Belo Horizonte (as três maiores cidades do Brasil). Em seguida, iteramos pela Dimensão das Competências 3.2.3 e extraímos as 10 habilidades mais procuradas pelos recrutadores. A segunda análise foi feita de maneira semelhante, iterando pela Dimensão dos Benefícios 3.2.2 considerando o item Salário. Como resultado final, apresentamos a lista com as 10 habilidades mais procuradas por

região e as 10 mais procuradas de acordo com a faixa salarial. Por fim, aplicamos a **Distância de Jaccard** Niwattanakul et al. [2013] entre esses itens para descobrir como habilidades semelhantes são buscadas entre regiões e entre faixas salariais.

Informação: Na Distância de Jaccard, quanto mais semelhantes forem dois conjuntos de itens, mais próximo de 1 será o resultado. Avaliando os resultados mostrados nas Tabelas 5.7 e 5.8, observamos uma grande semelhança entre os requisitos de diferentes regiões. Considerando os resultados para uma faixa salarial adjacente, os requisitos nas faixas salariais mais baixas são muito semelhantes. A semelhança aumenta à medida que avançamos nas faixas salariais intermediárias, caindo novamente conforme caminha às faixas salariais mais altas.

Conhecimento: Esses resultados podem ajudar na ferramenta mencionada anteriormente para sugestões de alterações em currículos (vagas). Por exemplo, de acordo com as habilidades inseridas pelo usuário, é possível sugerir a ele que selecione cidades diferentes para procura de emprego. Além disso, observando que existem conjuntos de habilidades específicas para faixas salariais diferentes, seria interessante usar as habilidades inseridas pelos usuários para categorizá-las em uma faixa salarial específica, e o sistema de recomendação poderia recomendar vagas nessa categoria. Além disso o sistema poderia sugerir a alteração de algumas habiliaddes por outras semelhantes que o candidato possa ter e que sejam mais interessantes para conseguir a vaga desejada. Esses dados, portanto, servem como subsídio para a criação de um sistema de pré-registro e para a melhoria dos SsR. Outra ferramenta interessante que poderia ser adicionada a esses serviços de recrutamento on-line seriam aplicativos que podem ajudar os candidatos a evoluir em suas carreiras, apresentando aprimoramentos de aprendizado e habilidades relacionadss à faixa salarial. Todas essas sugestões são úteis para ajudar os candidatos a saírem de um cenário de escassez que possam se encontrar.

Tabela 5.7: Distância Entre Habilidades para Regiões

CAPITAL 1	CAPITAL 2	DISTÂNCIA DE JACCARD
São Paulo	Rio de Janeiro	0,7
São Paulo	Belo Horizonte	1,0
Rio de Janeiro	Belo Horizonte	0,7

5.3.3 Perspectiva do Recrutador

Da mesma forma que a perspectiva do candidato, o recrutador que sofre da PMS pode querer entender como descrever suas vagas de maneira mais atraente. A seguir, apresentamos dois exemplos de pergunta alinhadas com esse objetivo.

Tabela 5.8: Distância Entre Habilidades para Faixas Salariais

Faixa Salarial 1	Faixa Salarial 2	Distância de Jaccard
< 1.000,00	> 1.000,00	0,2
> 1.000,00	> 2.000,00	0,2
> 2.000,00	> 3.000,00	0,5
> 3.000,00	> 4.000,00	0,7
> 4.000,00	> 5.000,00	0,1
> 5.000,00	> 6.000,00	0,1
> 6.000,00	> 7.000,00	0,7
> 7.000,00	> 8.000,00	0,6
> 8.000,00	> 9.000,00	0,4
> 9.000,00	> 10.000,00	0,2
> 10.000,00	> 15.000,00	0,4
> 15.000,00	> 20.000,00	0,0

Questão 1: *Como os recrutadores devem preencher o salário ofertado para o setor de tecnologia da informação de forma a aumentar o número de candidatos interessados na vaga?*

Análise: Para responder a essa pergunta, usamos várias iterações por meio de nossa metodologia. A primeira iteração foi feita através da Dimensão Área 3.2.1, usando apenas a área de TI novamente. Em segundo lugar, iteramos pela Dimensão Benefícios 3.2.2 duas vezes seguidas, a primeira usando o atributo Salário e a segunda, o atributo Posição. As Posições selecionadas para análise foram as de Desenvolvedor Júnior, Desenvolvedor Pleno, Desenvolvedor Sênior e Gerente.

Informação: Na Figura 5.9, observamos o percentual de inscrições de candidatos, de acordo com a posição. Há uma alta porcentagem de solicitações por vagas em que a faixa salarial não é descrita. Esses dados estão diretamente relacionados ao número de vagas em aberto, que não possuem um salário descrito. Os recrutadores preferem não preencher o requisito salarial. Acreditamos que seja por uma das duas razões: tentar maximizar o número de candidatos, incentivando o interesse em uma provável negociação salarial; ou, para poder avaliar melhor o candidato para posteriormente fazer uma oferta salarial adequada. Os gráficos 5.10a, 5.10b, 5.10c e 5.10d representam o número médio de candidatos que se aplicam a cada posição de acordo com cada faixa salarial. Nesses gráficos, temos a probabilidade de os candidatos se candidatarem a vagas com salários declarados. Considerando apenas as posições dos desenvolvedores, temos aquelas que exigem menos experiência (desenvolvedor júnior e desenvolvedor pleno) com uma média maior de candidaturas em comparação com a posição de desenvolvedor sênior. Essas posições iniciais também tendem a ter mais aplicações em faixas salariais iniciais/intermediárias. Os desenvolvedores seniores tendem a procurar

por salários mais altos. Os cargos de gerente são os que apresentam a maior média de aplicações, especialmente para as faixas salariais mais altas.

Conhecimento: Os dados mostram que um sistema de ajuda também pode ser criado para os recrutadores, fornecendo informações úteis ao registrar a vaga (recomendação pré-inscrição). Suponhamos que um recrutador precise urgentemente de um candidato para uma vaga de Desenvolvedor Sênior. Qual salário o recrutador deve oferecer para que o cargo/vaga tenha o número máximo de aplicações? A partir dos dados analisados, observa-se que, para obter um número maior de aplicações, é necessário oferecer um salário em torno de $8.000,00. Fornecer uma estimativa em tempo real do quão interessante é a vaga de acordo com a faixa salarial relatada é um mecanismo de *feedback* potencialmente útil para os recrutadores. Os SsR também podem se beneficiar com essas informações. É possível quantificar quantos usuários diferentes a mesma vaga pode ser recomendada, considerando o número médio de aplicações por faixa salarial. Por exemplo, vagas de emprego com um salário que não leva a uma média satisfatória de candidatos interessados, podem ser recomendadas para um número maior de usuários do que as outras vagas de emprego. Ao dar mais visibilidade às vagas de emprego com poucas inscrições, o sistema ajudaria solucionar o PMS e a satisfazer a necessidade do recrutador em obter um número adequado de candidatos.

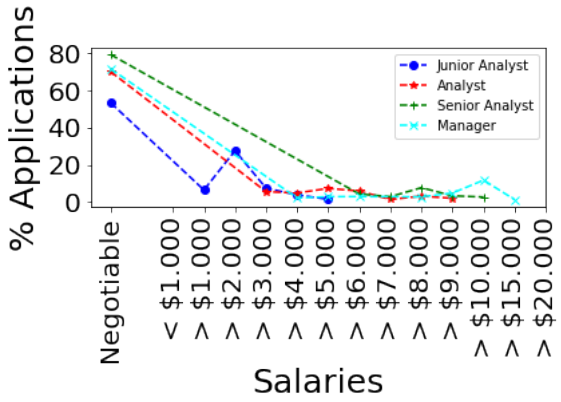
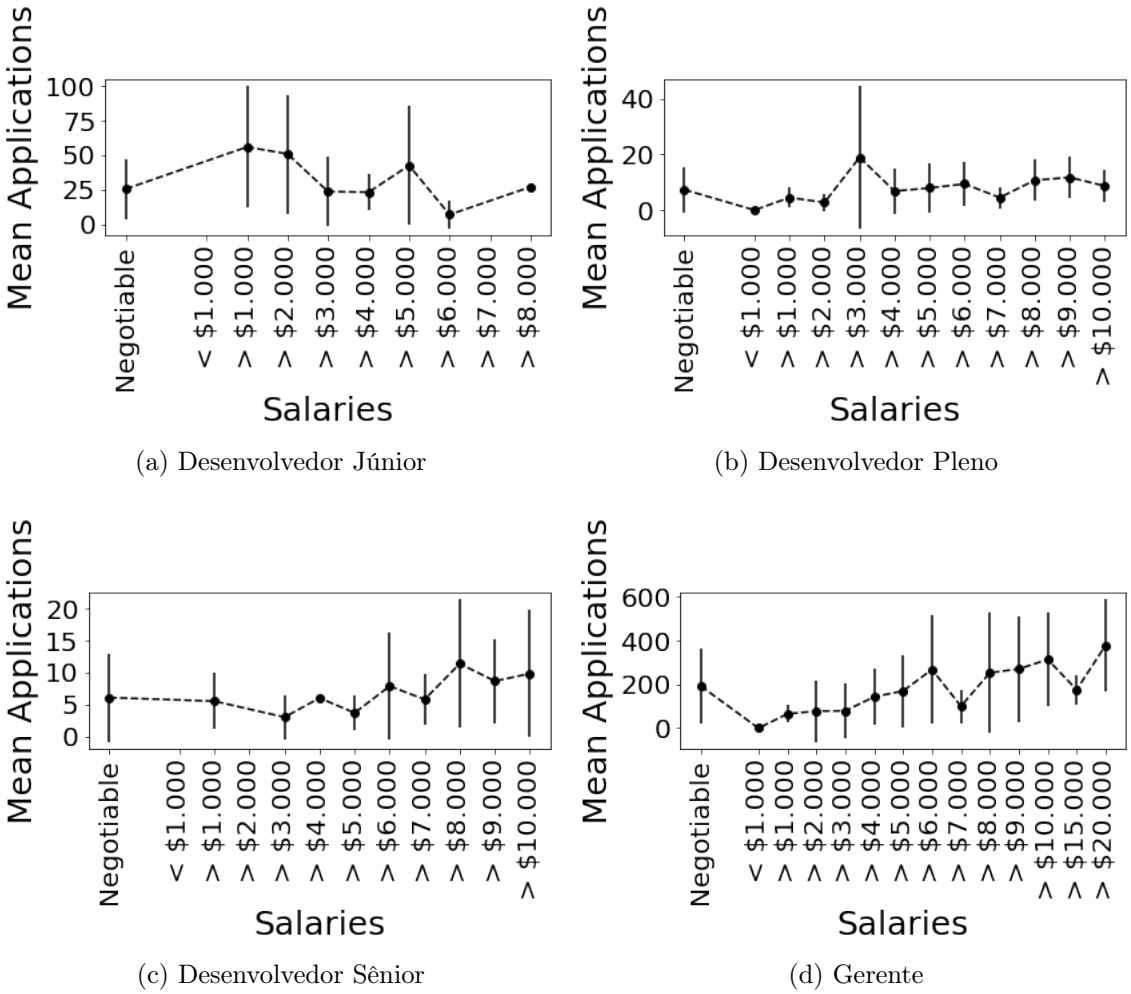


Figura 5.9: Porcentagem de aplicações por faixa salarial

Questão 2: *Quão atraente é um anúncio de emprego para um candidato quando comparado a outros anúncios para a mesma posição no departamento de TI?*

Análise: Neste análise, nós realizamos três iterações na metodologia. A primeira iteração foi feita através da Dimensão Área 3.2.1 usando a área Informática. Posteriormente, iteramos por meio da Dimensão Benefícios 3.2.2 duas vezes, na primeira usando o atributo Salário e na segunda, o atributo Posição/Vaga. As posições selecionadas



(a) Desenvolvedor Júnior

(b) Desenvolvedor Pleno

(c) Desenvolvedor Sênior

(d) Gerente

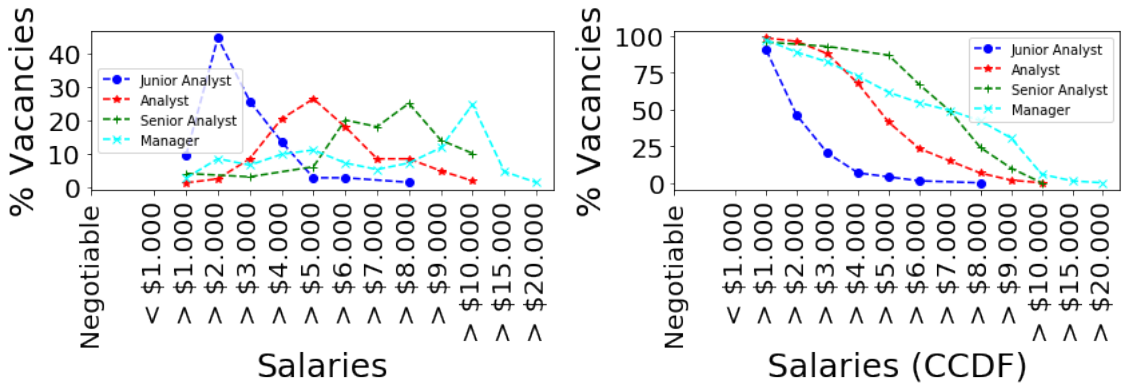
Figura 5.10: Média de aplicações por faixa salarial

para análise foram as de Desenvolvedor Júnior, Desenvolvedor Pleno, Desenvolvedor Sênior e Gerente.

Informação: A Figura 5.11a mostra a porcentagem de vagas de emprego para cada cargo. Em cada uma das curvas do gráfico, podemos ver que há um pico. Para os cargos de Desenvolvedor Júnior, esse pico ocorre na faixa salarial acima de $2.000. Em Desenvolvedor Pleno, está acima de $5.000. Em Desenvolvedor Sênior, o pico ocorre em cerca de $8.000. Quanto aos cargos de Gerente, o pico de ofertas de vagas aparece na faixa salarial acima de $10.000. Quanto maior o salário, mais atraente a vaga se torna. Essa atratividade pode ser medida através da Figura 5.11b, onde podemos ver para cada um dos salários oferecidos, quanto à frente das outras vagas a oportunidade em questão está localizada. Por exemplo, considerando a posição de Desenvolvedor Pleno, um salário oferecido acima de $7.000,00 está à frente de 75% das outras vagas,

uma vez que está entre os 25% mais altos da curva mostrada no gráfico.

Conhecimento: Juntos, esses gráficos ajudam o sistema a dizer ao recrutador quão bom é o benefício ofertado. Benefícios mais altos tendem a atrair mais candidatos, especialmente para cargos que exigem habilidades mais refinadas, como Desenvolvedor Sênior e Gerente. Portanto, o sistema pode indicar que, para se destacar entre as vagas de Desenvolvedor Júnior, um salário superior a $3.000 é suficiente. Para um Gerente, no entanto, será necessário oferecer um salário superior a $15.000 por mês. Dessa forma ajudaria recrutadores a descreverem melhor as vagas que estejam sofrendo com escassez de usuários. Por outro lado, essas informações podem ser úteis para os SsR atenderem aos usuários mais exigentes e os ajudar a se aplicarem vagas satisfatórias. Anúncios mais atraentes podem ser estritamente recomendados para usuários muito seletivos que buscam posições interessantes de acordo com suas demandas. Isso aumentaria a satisfação do cliente e, ao mesmo tempo, impediria que ele se mantivesse sofrendo com o PMS ao receber apenas recomendações de vagas desinteressantes.



(a) (b)

Figura 5.11: Vagas por Faixa Salarial

5.3.4 Discussão

Além de todas as questões levantadas, outras poderiam ter sido discutidas nesta seção. O conjunto de perguntas abordadas aqui serve de exemplo para a aplicação de nossa metodologia e como ela responde a perguntas reais que podem ser muito úteis na identificação de oportunidades de melhoria em diferentes pontos de vista. Observamos que muitas áreas profissionais apresentam um desequilíbrio entre demanda e oferta de oportunidades (ou seja, 60% das vagas estão relacionadas a apenas 20% das diferentes áreas profissionais). No que diz respeito à avaliação das características

de CVs e vagas de emprego, observamos que existem algumas características que ajudam os recrutadores a gerar as descrições de vagas mais atraentes e que os CVs tendem a ser mal escritos em comparação com as vagas de emprego. Todos esses resultados consistem em maneiras de mitigar o problema da escassez de *matchings*, maximizando as chances de recrutamento de candidatos e vagas, o que corresponde ao nosso principal objetivo neste trabalho.

5.4 Avaliando a Identificação de Itens Escassos e Appendant features

Nesta seção, apresentamos o processo usado para: (1) definir os valores limite para a similaridade τ e o número de *matchings* μ; (2) classificar itens escassos; e (3) identificar *Appendant Features*. Além disso, apresentamos um experimento para validar todas essas estratégias.

Iniciamos apresentando o processo para definir μ e τ, bem como a classificação de itens escassos. Para isso, consideramos os quatro passos descritos abaixo:

1 Separamos nossa amostra de dados em conjunto de treinamento (janeiro a agosto) e testes (setembro e outubro). Consequentemente, o valor considerado para a variável T, descrita na Seção 4.1, é 2.

2 Considerando o conjunto de treinamento, aplicamos a estratégia apresentada na Seção 4.1 para definir μ. Encontramos o valor 3 em nossa amostra.

3 Usando o conjunto de treinamento novamente, aplicamos a estratégia apresentada na Seção 4.1 para definir τ. Nesse caso, encontramos o valor 0,6 em nossa amostra;

4 Finalmente, com base nos valores de μ e τ, aplicamos o Classificador Binário de Escassez (Algoritmo 1) no conjunto de testes, classificando cada item (CV e vaga) como estando em cenário de escassez ou não.

Uma vez definidos os CVs e as vagas que sofrem do PMS, identificamos as *Appendant features* aplicando a estratégia apresentada em 4.2. Mais especificamente, para cada item no conjunto de testes (CV ou vaga), extraímos suas *features* e classificamos cada um delas como *Screening* ou *Appendant* usando o Classificador de *Features* (Algoritmo 2). A tabela 5.9 apresenta *Screening* e *Appendant features* encontradas em currículos na área de informática.

Tabela 5.9: Exemplos de Screening e Appendant Features

SCREENING FEATURES	APPENDANT FEATURES
Administrador	Logística
PHP	Motorista
Wordpress	Corel
Javascript	PowerBI Básico
Desenvolvimento de Software	Autocad Intermediário
CSS	Outlook
Graduação Universitária	Espanhol

Avaliar a qualidade das estratégias propostas é uma tarefa complexa, pois não temos informações no conjunto de testes a respeito de quais CVs/vagas sofrem do PMS, nem informações de quais *features* são *Screening* ou *Appendant*. Portanto, para avaliar as estratégias propostas, geramos conjuntos de dados sintéticos nos quais podemos controlar essas informações. Mais especificamente, com base na classificação fornecida pelo nosso Algoritmo 1 nos CVs e vagas do conjunto de testes, bem como na classificação de *features* fornecida pelo Algoritmo 2, geramos a probabilidade distribuição das *Screening features* e *Appendant features* para seis cenários: (a) vagas em geral; (b) vagas classificadas como escassas; (c) vagas classificadas como não escassas; (d) CVs em geral; (e) CVs classificados como escassos; e (f) CVs classificados como não escassos. Com base nessas distribuições, criamos quatro conjuntos de dados de itens, conforme descrito:

- **Vagas Escassas**: 50 vagas, sinteticamente criadas, compostas por *features* selecionadas aleatoriamente e seguindo a distribuição para cenários (a) e (b);
- **Vagas Não Escassas**: 50 vagas, sinteticamente criadas, compostas por *features* selecionadas aleatoriamente e seguindo a distribuição para cenários (a) e (c);
- **CVs Escassos**: 50 CVs, sinteticamente criados, compostos por *features* selecionadas aleatoriamente e seguindo a distribuição para cenários (d) e (e);
- **CVs Não Escassos**: 50 CVs, sinteticamente criados, compostos por *features* selecionadas aleatoriamente e seguindo a distribuição para cenários (d) e (f);

Portanto, temos um total de 100 vagas e 100 currículos para os quais sabemos se eles são escassos ou não. Aplicamos o Classificador Binário de Escassez (Algoritmo 1) a todos esses itens. A tabela 5.10 apresenta os resultados alcançados pelo nosso algoritmo. Como podemos observar, nosso classificador alcançou altos valores de acurácia para todo o conjunto de itens, demonstrando coerência na identificação de vagas e currículos criados sinteticamente como escassos ou não. Esses resultados apontam

que o processo para definir os parâmetros μ e τ e a classificação de *features* são consistentes e podem ser generalizados. Finalmente, na próxima seção, apresentamos os experimentos realizados para avaliar nossas estratégias para mitigar o PMS.

Tabela 5.10: Resultados da Classificação de Escassez de CVs/vagas sinteticamente criados

Dataset	Acurácia
Vagas Escassas	86%
Vagas Não Escassas	75%
CVs Escassos	78%
CVs Não Escassos	77%

5.5 Aplicando Operações de Edição nas Descrições de Itens

5.5.1 Métricas de Avaliação

Como mencionado anteriormente, até onde sabemos, esse é o primeiro esforço para abordar o PMS na literatura. Portanto, também é necessário propor métricas para avaliar as estratégias propostas neste trabalho. Nosso objetivo é avaliar como cada estratégia melhora a qualidade dos currículos e vagas de emprego. Para isso, seguimos as quatro etapas abaixo:

1 Para cada vaga de trabalho ou CV em nosso conjunto de teste, o qual está classificado em um cenário de escassez, calculamos a média das distâncias semânticas para os três *matchings* mais próximos no conjunto de treino;

2 Aplicamos nossas estratégias para mitigar o PMS em CVs e vagas de trabalho do conjunto de teste;

3 Mais uma vez, para cada CV e vaga de trabalho modificada por nossas estratégias, calculamos a média das distâncias semânticas para os três *matchings* mais próximos do conjunto de treino;

4 Realizamos a comparação entre a média das distâncias semânticas para os três *matchings* mais próximos antes e depois de aplicar nossas estratégias, considerando duas métricas distintas detalhadas abaixo, avaliando as melhorias alcançadas.

As métricas propostas para comparar a média das distâncias semânticas mencionadas na etapa 4 são:

Distância Relativa: Essa medida é a diferença absoluta entre a média das distâncias semânticas antes e depois das mudanças sugeridas por nossas estratégias, divididas pela distância semântica antes das mudanças. O objetivo é entender o percentual de melhoria (ou piora) que ocorreu com cada CV e vaga. Essa métrica é apresentada na Equação 5.1 e, quanto maior o valor, melhor a qualidade alcançada.

$$\frac{mediaDistanciaDepois - mediaDistanciaAntes}{mediaDistaciaAntes} * 100 \tag{5.1}$$

Perturbação: Essa métrica avalia a porcentagem de habilidades que foram modificadas após a aplicação de nossas estratégias para mitigar o PMS. O objetivo é entender o quanto cada estratégia precisa alterar os currículos e vagas para que haja melhorias. Idealmente, não seria necessário sugerir um grande número de alterações para os usuários. Portanto, há um *tradeoff* entre melhoria e perturbação. Essa métrica é apresentada na Equação 5.2 e, quanto menor o valor, melhor a qualidade do método usado.

$$1 - \frac{numeroHabilidadesNaoAlteradas}{numeroTotalHabilidades} \tag{5.2}$$

5.5.2 Resultados

Nesta seção, apresentamos os resultados alcançados pelas estratégias propostas para mitigar o PMS. Primeiro, comparamos as seis estratégias propostas na Seção 4.3: 1) *Random Feature Removal* (RR); 2) *Appendant Feature Removal* (AR); 3) *Appendant Feature Exchange for Screening Features* (AES); 4) *Appendant Feature Exchange for Job Screening Features* (AEJS); 5) *Appendant Feature Exchange for Screening Features Contextualized* (AESC); e 6) *Appendant Feature Exchange for Job Screening Features Contextualized* (AEJSC). Avaliamos as estratégias para as descrições de currículos e vagas, considerando as duas métricas apresentadas na seção anterior: 1) Distância Relativa; e 2) Perturbação. Para ambas as métricas, apresentamos os resultados por meio de distribuições probabilísticas.

A Figura 5.12 mostra os resultados alcançados pelas estratégias propostas, considerando todas as áreas profissionais presentes no conjunto de testes, considerando a métrica de distância relativa, para vagas de emprego e currículos. No gráfico relacionado às vagas, é possível observar duas tendências distintas nos resultados. Por um lado, temos duas estratégias que apresentaram piora ou pequenas alterações na qualidade. Por outro, quatro estratégias com melhorias significativas. As estratégias do primeiro grupo

são RR e AES. Como esperado, o RR apresenta uma ligeiro piora na qualidade das descrições dos empregos. A estratégia da AES não apresenta melhorias relevantes na qualidade das descrições. As quatro estratégias do segundo grupo melhoraram a qualidade das descrições de vagas na seguinte ordem de ganhos: AEJSC, AEJS, AESC e AR. Para todos elas, pelo menos 70% das descrições de vagas obtiveram melhorias, nas quais podemos destacar a estratégia AEJSC com a curva de ganho mais acentuada. No caso da AEJSC, os ganhos representam uma redução de 56% das vagas que sofrem de PMS.

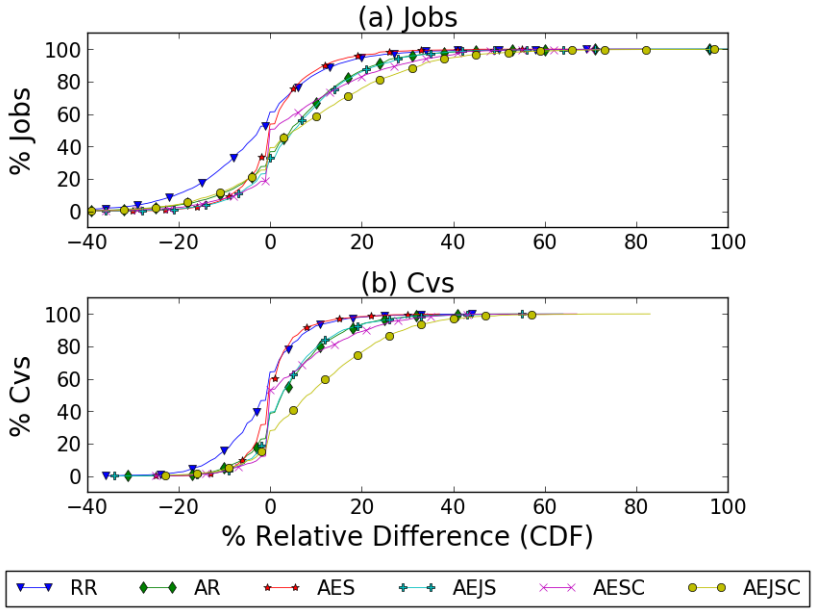


Figura 5.12: Distância Relativa para Todas as Estratégias

Considerando o gráfico relacionado aos currículos, apenas a estratégia RR apresenta piora nas descrições dos currículos, enquanto a AES mostra uma leve melhora. As quatro estratégias restantes melhoraram a qualidade dos currículos: AESC com 55% das descrições de currículos alcançando melhorias; AR e AEJS com mais de 70% dos currículos apresentando ganho na qualidade de suas descrições; AEJSC com 75% das descrições de currículo alcançando melhorias. Mais uma vez, a estratégia AEJSC apresenta uma curva de ganho mais acentuada. Esses resultados significam uma redução de 53% dos currículos que sofrem de PMS. Considerando essa estratégia como a melhor para os dois casos, podemos observar que, tanto para a vaga quanto para as descrições de currículos, que aproximadamente 60% dos itens apresentam melhorias de até 10%. Além disso, podemos ver que 20% dos itens apresentam melhorias de até 25%.

Uma observação importante baseada nos resultados descritos acima é que a simples troca de *Appendant features* por *Screening features* que são semanticamente

próximas (considerando o espaço vetorial criado por *embeddings*) não aprimora as descrições para aproximar as vagas de emprego e os currículos. Por outro lado, observamos ganhos muito significativos quando as *Appendant features* são trocadas por *features* presentes nas descrições de vagas, considerando a descrição original completa de um currículo/vaga. Além disso, podemos ver os currículos se beneficiam mais das estratégias propostas do que as vagas. Portanto, podemos concluir que os CVs tendem a ser mais mal escritos do que as vagas de emprego, uma vez que quase todas as estratégias oferecem alguma melhoria na qualidade. Isso nos leva a crer que os currículos têm um excesso de habilidades distintas (os candidatos tendem a inserir o maior número possível de *features* na tentativa de abranger o maior número possível de vagas), as quais podem agregar pouco, ou até mesmo nada para a ocorrência de um *matching* com uma vaga de emprego. As ofertas de emprego, por outro lado, são elaboradas por profissionais, consistindo em textos curtos e focados nas habilidades necessárias para o candidato realizar o trabalho proposto, apresentando melhorias apenas com estratégias específicas e mais elaboradas. Isso explica por que as estratégias que alteram as *Appendant features* para a *Screening features* nas descrições de vagas são aquelas que se destacam em nossos resultados.

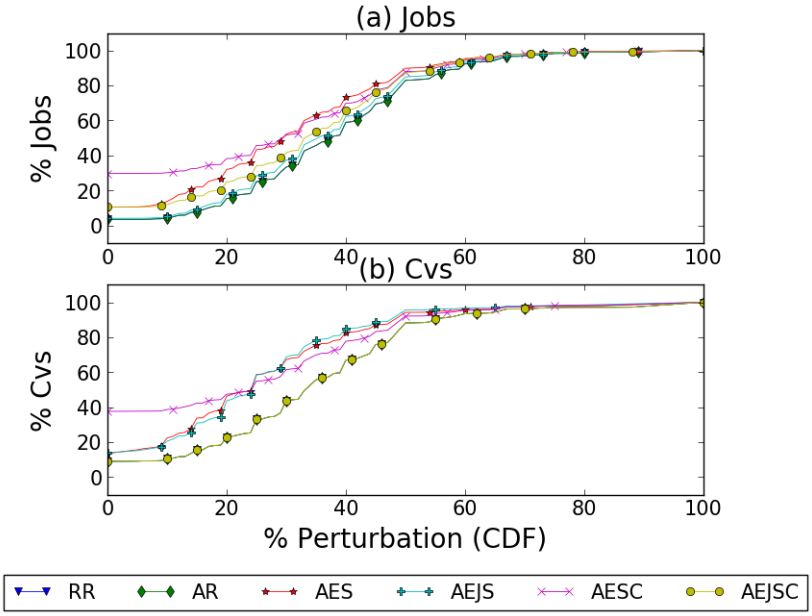


Figura 5.13: Perturbação para Todas as Estratégias

A perturbação causada pela aplicação de nossas estratégias pode ser observada na Figura 5.13 (a) e na Figura 5.13 (b). Quanto menor a perturbação, menor o número de alterações feitas nas descrições de um CV/vaga. No entanto, é importante

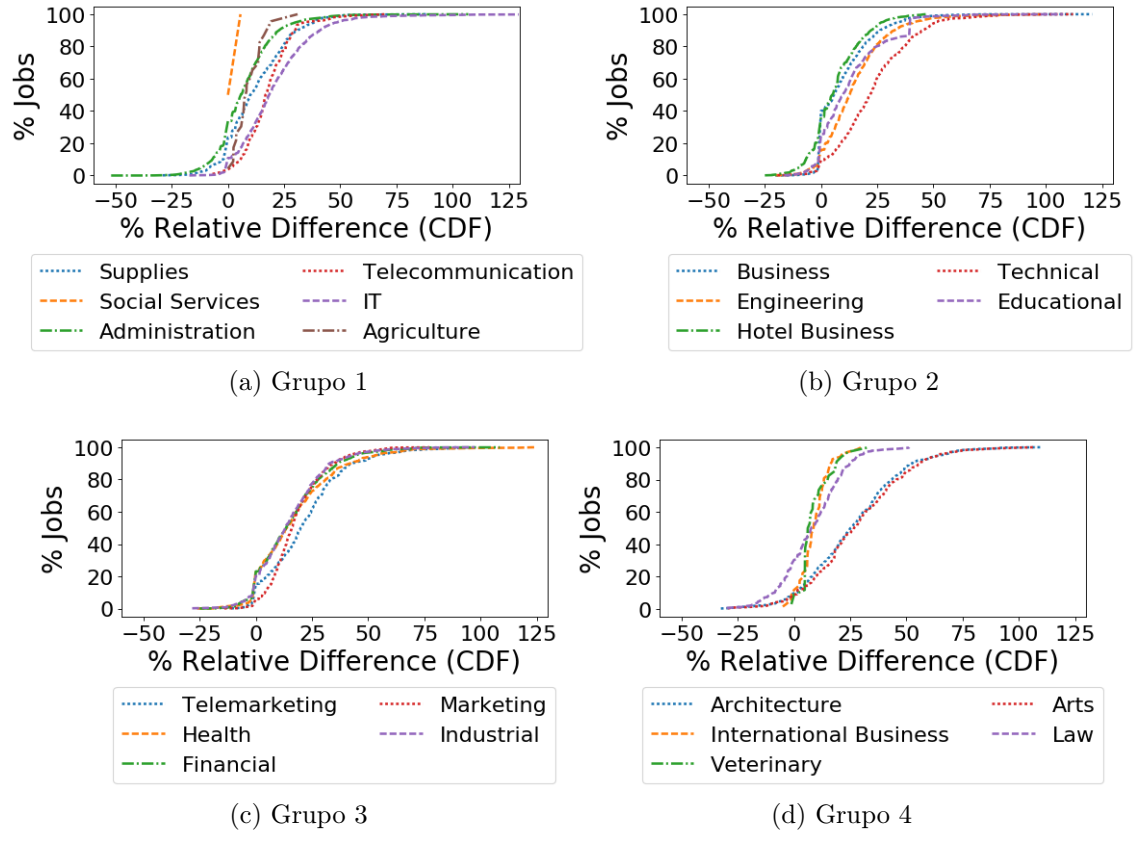
ressaltar que poucas alterações podem não ser suficientes para melhorar a descrição dos itens. Portanto, é essencial entender o comportamento das estratégias, a fim de encontrar o melhor equilíbrio entre perturbação e qualidade, maximizando a qualidade das descrições com o mínimo de alterações possível. Outro ponto a ser observado é que as estratégias de *Random Feature Removal* e *Appendant Feature Removal* possuem um comportamento semelhante em relação ao número de *features* removidas em cada CV/ vaga. Consequentemente, eles têm a mesma perturbação e suas distribuições se sobrepõem nos gráficos. As outras estratégias exibem comportamentos diferentes por selecionarem novos recursos de maneiras diferentes.

Iniciamos nossa análise pelos resultados relacionados às vagas Figura 5.13 (a). Como podemos ver, a solução que apresenta os melhores resultados em termos de distância relativa, a *Appendant Feature Exchange for Job Screening Features Contextualized*, é a terceira solução que causa menos perturbações nas descrições das vagas, com 80% delas com menos de 50% de perturbação. A estratégia que causa a menor perturbação foi a *Appendant Feature Exchange for Job Screening Features*, com 80% das vagas de emprego com uma perturbação menor que 45%. Apesar disso, como observamos anteriormente, esta estratégia não possui uma porcentagem relevante de melhorias nas descrições de cargos. Concluímos também que as descrições de vagas tendem a ser melhor descritas, portanto, são necessárias alterações adicionais para aumentar substancialmente a qualidade das descrições sem comprometer a consistência entre as *features*. No entanto, é interessante notar que as estratégias que causam a maior perturbação, a *Random Feature Removal* e *Appendant Feature Removal*, também não alcançam ganhos em termos de Distância Relativa. Portanto, esses resultados mostram que uma estratégia eficiente precisa ser capaz de entender o contexto de uma vaga para sugerir mudanças relevantes.

Agora, concentrando nossa análise nas perturbações causadas pelas estratégias nas descrições de CVs, podemos ver que a *Appendant Feature Exchange for Job Screening Features* atingiu a quarta menor perturbação de todas as estratégias, com cerca de 80% dos CVs com menos de 48% de perturbação em suas descrições. Enquanto isso, a estratégia *Appendant Feature Exchange for Job Screening Features Contextualized* causa a menor perturbação com 80% dos CVs com menos de 35% de perturbação em suas descrições. Observando os resultados da Distância Relativa, temos que essa estratégia também alcançou resultados interessantes em termos de qualidade dos CVs alterados. A partir desses resultados, podemos concluir que, para os CVs, como geralmente são mal escritos, poucas alterações já resultam em grandes melhorias na qualidade de suas descrições. Assim, a troca de *Appendant features* por *Screening features* de vagas foi o melhor *tradeoff* entre a qualidade final dos CVs e a perturbação.

No entanto, é importante enfatizar que o objetivo das estratégias propostas e avaliadas deve ser usado como uma ferramenta auxiliar para os usuários quando eles estiverem cadastrando seus currículos/vagas. Portanto, considerando a perspectiva do potencial de aprimoramentos na qualidade das descrições de CVs, a estratégia *Appendant Feature Exchange for Job Screening Features Contextualized* produz os melhores resultados. Portanto, na próxima seção, na qual apresentamos uma avaliação separada para cada área profissional em nosso conjunto de dados, consideramos a estratégia *Appendant Feature Exchange for Job Screening Features Contextualized*.

5.5.3 Distância Relativa por Área Profissional



(a) Grupo 1

(b) Grupo 2

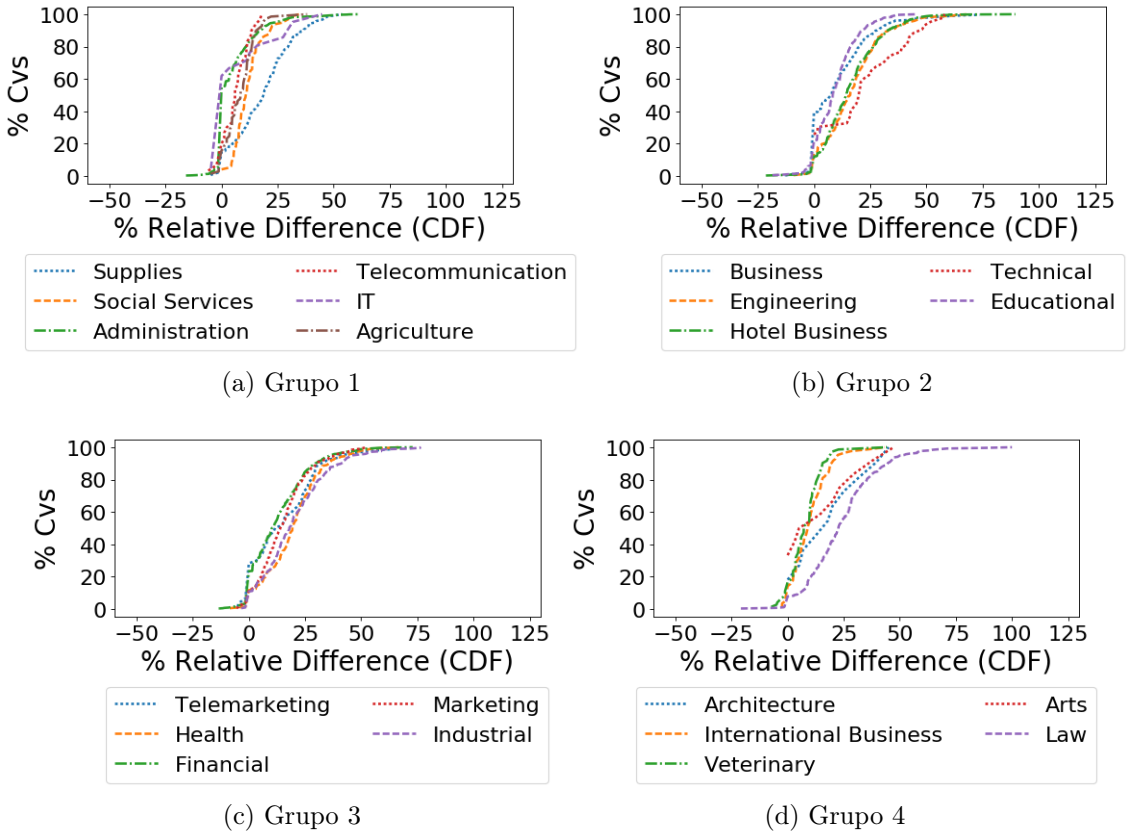
(c) Grupo 3

(d) Grupo 4

Figura 5.14: Distância Relativa por Área Profissional para Vagas

Em nosso conjunto de dados, encontramos 21 áreas profissionais distintas. Inserir a análise de todos elas no mesmo gráfico seria muito confuso. Portanto, dividimos a análise em quatro gráficos para vagas (Figura 5.14) e quatro gráficos para CVs (Figura 5.15). É possível comparar como a melhor estratégia previamente identificada (ou seja, Appendant Feature Exchange for Job Screening Features Contextualized)

afeta as diferentes áreas. Podemos observar que, para os vagas, as áreas com maior impacto positivo nas descrições de empregos são: Telecomunicações, Informática, Técnica, Telemarketing, Arquitetura e Artes, todas com 60% das vagas com mais de 15% de melhorias. Por outro lado, para CVs, as áreas com maiores ganhos positivos são: Suprimentos, Técnica, Saúde, Industrial e Jurídica, todas com 60% de CVs com mais de 12% de melhoria.



(a) Grupo 1

(b) Grupo 2

(c) Grupo 3

(d) Grupo 4

Figura 5.15: Distância Relativa por Área Profissional para Currículos

Embora todas as áreas profissionais apresentem ganhos, o que vemos é que os ganhos relacionados a cada uma delas são diferentes, com algumas áreas apresentando melhores resultados do que outras. Curiosamente, exceto a área Técnica, as áreas com maiores ganhos para vagas são diferentes das áreas com maiores ganhos para CVs. Isso se deve ao fato de que algumas áreas tendem a ter vagas de emprego descritas com menor eficiência (tornando o trabalho pouco atraente ou muito exigente para os candidatos), enquanto outras podem ter CVs mais mal escritos (dificultando a adequação do candidato ao emprego). De qualquer forma, esses resultados nos mostram que uma linha de trabalho futura é analisar as descrições de vagas e CVs separadamente por área. É claro

que é possível obter mais aprimoramentos com estratégias especializadas em subconjuntos de dados (áreas profissionais em nosso exemplo), além de propor estratégias focadas apenas em vagas ou currículos (já que a qualidade descritiva tende a ser diferente)

5.6 Síntese do Capítulo

Este capítulo iniciou caracterizando o conjunto de dados da Catho por meio das cinco dimensões de análise da metodologia. A seguir realizamos a caracterização do PMS ao elencar questões de negócios sob as perspectivas da Catho, dos candidatos e dos recrutadores. Dentre outros resultados, mostramos que há desequilíbrio na relação candidato/vaga em diversas áreas profissionais. Essas áreas podem ainda vir a sofrer com o PMS de forma sazonal, estando sujeitas a épocas do ano e feriados. Usamos esses resultados para ressaltar possibilidades de melhoria nos SsR e como isso ajudaria a mitigar o problema de escassez de ocorrência de *matchings*.

Finalmente, aplicamos no conjunto de dados a metodologia de classificação de CVs/vagas em escassos ou não escassos. Com isso separamos os itens classificados como sofrendo do PMS para serem utilizados como teste das estratégias de melhoria. Devido ao fato do PMS ser um problema ainda não estudado na literatura, definimos duas métricas de avaliação para mensurarmos o quanto nossas estratégias foram úteis. A primeira métrica é a Distância Relativa, cujo objetivo é entender o percentual de melhoria (ou piora) que ocorreu em cada CV e vaga. A segunda métrica é a Perturbação, utilizada para avaliar o quanto cada estratégia precisa alterar os currículos e vagas para que haja melhorias. Por fim, aplicamos as seis estratégias propostas e obtivemos ganhos expressivos, principalmente ao aplicarmos a *Appendant Feature Exchange for Job Screening Features Contextualized* que se mostrou a solução mais sofisticada. Esses resultados mostram que as estratégias definidas para mitigar o PMS podem ser aplicadas em cenários reais, o que é discutido no próximo capítulo.

Capítulo 6

Conclusões & Trabalhos Futuros

Neste trabalho apresentamos uma metodologia iterativa de caracterização para serviços de recrutamento on-line. Foram apresentadas diversas métricas, divididas em cinco dimensões de análise: Área, Competência, Descritiva, Geográfica e Temporal. Nosso objetivo é identificar, analisar e compreender as características intrínsecas de cenários de recomendação de empregos e candidatos, utilizando-se das informações obtidas para melhorar os SsR dessas empresas. Avaliamos nossa metodologia em uma amostra de dados reais fornecida pela Catho. Os resultados mostraram que, para a amostra avaliada, há um grande desequilíbrio entre oferta e demanda, considerando diversas características, tais como área de atuação, diferentes regiões geográficas, competências exigidas, entre outras. Além disso, apresentamos um exemplo de iteração sobre a metodologia, demonstrando os vários níveis de informação que podem ser obtidos a partir da correlação entre as diversas dimensões.

Como trabalhos futuros, nosso objetivo é aplicar iterativamente nossa metodologia considerando distintos períodos ao longo do tempo e prover uma avaliação da evolução temporal dessas aplicações. Além disso, nossa meta é propor novos SsR de empregos capazes de atingirem um número muito maior de pessoas, alocando-as nas melhores vagas possíveis, mesmo em setores com escassez de oportunidades.

Neste trabalho, apresentamos o **Problem of Matching Scarcity** (**PMS**) bem como os cenários nos quais candidatos e recrutadores sofrem com a ausência de oportunidades de *matchings* nos sistemas de recrutamento on-line. Primeiro, formalizamos e caracterizamos o PMS considerando uma amostra de dados reais fornecida pela Catho. Segundo, propusemos e avaliamos estratégias para identificar automaticamente candidatos e vagas de emprego que sofrem de PMS com até 75% de precisão. Terceiro, propusemos seis estratégias diferentes para mitigar o PMS. Nossas estratégias consistem na introdução de alterações nos currículos e nas descrições das

vagas, para aproximar os candidatos que sofrem do PMS a vagas semanticamente relacionadas a eles e vice-versa. A melhor estratégia foi capaz de reduzir em até 50% o número de currículos e vagas que sofrem de PMS antes de aplicá-la em nossa amostra. Até onde sabemos, esse é o primeiro esforço para abordar a PMS na literatura.

Como trabalho futuro, pretendemos analisar descrições de vagas e currículos separadamente por áreas profissionais. Em nossos resultados de caracterização, observamos que algumas áreas sofrem mais com a PMS do que outras. Isso é um indicativo de que é possível obter melhorias mais altas empregando estratégias distintas especializadas em diferentes áreas profissionais. Além disso, nosso objetivo é realizar avaliações on-line de todas as estratégias com SsR para medir o impacto real de nosso trabalho.

Referências Bibliográficas

Abbasse, Z. & Mirrokni, V. (2007). A recommender system based on local random walks and spectral methods. Em *Proceedings of the 9th WebKDD and 1st SNA-KDD*, pp. 102--108. Springer.

Adomavicius, G. & Tuzhilin, A. (2001). Expert-driven validation of rule-based user models in personalization applications. *Data Mining and Knowledge Discovery*, 5(1):33--58.

Adomavicius, G. & Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, pp. 734–749.

Akshita & Smita (2013). Recommender system: Review. *International Journal of Computer Applications*.

Alotaibi, S. (2012). A survey of job recommender systems. *International Journal of the Physical Sciences*, 7.

Alotaibi, S. & Ykhlef, M. (2012). Job recommendation systems for enhancing e-recruitment process. *Proceedings of the information and knowledge engineering conference*.

Baeza-Yates, R.; Ribeiro-Neto, B. et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

Balabanović, M. & Shohom, Y. (1997). Content-based, callaborative recommendation. *Communications of the ACM*, 40(3).

Bashiri, P. (2018). Recommender systems: Survey and possible extensions.

Basu, C.; Hirsh, H.; Cohen, W. et al. (1998). Recommendation as classification: Using social and content-based information in recommendation. Em *Aaai/iaai*, pp. 714--720.

Beel, J.; Gipp, B.; Langer, S. & Breitinger, C. (2016). Paper recommender systems: A literature survey. *International Journal on Digital Libraries*, 17(4):305--338.

Bell, R. M. & Koren, Y. (2007). Lessons from the netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, 9(2):75--79.

Berti-Equille, L. (2019). Learn2clean: Optimizing the sequence of tasks for web data preparation. Em *The World Wide Web Conference*, WWW '19, pp. 2580--2586, New York, NY, USA. ACM.

Billsus, D. & Pazzani, M. J. (1998). Learning collaborative information filters. Em *Icml*, volume 98, pp. 46--54.

Billsus, D. & Pazzani, M. J. (2000). User modeling for adaptive news access. *User modeling and user-adapted interaction*, 10(2-3):147--180.

Billsus, D.; Pazzani, M. J. & Chen, J. (2000). A learning agent for wireless news access. Em *Proceedings of the 5th international conference on Intelligent user interfaces*, pp. 33--36. ACM.

Bobadilla, J.; Ortega, F.; Hernando, A. & Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems*, 46:109--132.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331--470.

Campos, P. G.; Bellogín, A.; Díez, F. & Chavarriaga, J. E. (2010). Simple time-biased knn-based recommendations. Em *Proceedings of the Workshop on Context-Aware Movie Recommendation*, pp. 20--23. ACM.

Candillier, L.; Jack, K.; Fessant, F. & Meyer, F. (2009). State-of-the-art recommender systems. *Collaborative and Social Information Retrieval and AccessTechniques for Improved User Modeling*.

Cardoso, A.; MourÃ£o, F. & Rocha, L. (2019). A characterization methodology for candidates and recruiters interaction in online recruitment services. pp. 333–340.

Centeno, M. (2004). The match quality gains from unemployment insurance. *The Journal of Human Resources*, 39(3):839--863. ISSN 0022166X.

Choi, K.; Yoo, D.; Kim, G. & Suh, Y. (2012). A hybrid online-product recommendation system: Combining implicit rating-based collaborative filtering and sequential pattern analysis. *Electronic Commerce Research and Applications*, 11(4):309--317.

Coffman, K. G. & Odlyzko, A. M. (2002). Handbook of massive data sets. capítulo Internet Growth: Is There a "Moore's Law"for Data Traffic?, pp. 47--93. Kluwer Academic Publishers, Norwell, MA, USA.

Cohen, W.; Prieditis, A. & Russell, S. (1995). Proceedings of the twelfth international conference on machine learning.

Cross, R. (2014). *Duns Scotus's Theory of Cognition*. Oxford Scholarship.

Daramola, J. O.; Oladipupo, O. O. & Musa, A. G. (2010). A fuzzy expert system tool for online personnel recruitments. *Int. J. Bus. Inf. Syst.*, 6(4):444--462. ISSN 1746-0972.

Das, P.; Barua, K.; Pandey, M. & Routaray, S. S. (2019). Context level entity extraction using text analytics with big data tools. Em Abraham, A.; Dutta, P.; Mandal, J. K.; Bhattacharya, A. & Dutta, S., editores, *Emerging Technologies in Data Mining and Information Security*, pp. 357--367, Singapore. Springer Singapore.

De Campos, L. M.; Fernández-Luna, J. M.; Huete, J. F. & Rueda-Morales, M. A. (2010). Combining content-based and collaborative recommendations: A hybrid approach based on bayesian networks. *International Journal of Approximate Reasoning*, 51(7):785--799.

De Gemmis, M.; Lops, P.; Semeraro, G. & Basile, P. (2008). Integrating tags in a semantic content-based recommender. Em *Proceedings of the 2008 ACM conference on Recommender systems*, pp. 163--170. ACM.

Dempster, A. P. (2008). *A Generalization of Bayesian Inference*, pp. 73--104. Springer Berlin Heidelberg, Berlin, Heidelberg.

Deshpande, M. & Karypis, G. (2004). Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)*, 22(1):143--177.

Dong, W.; Moses, C. & Li, K. (2011). Efficient k-nearest neighbor graph construction for generic similarity measures. Em *Proceedings of the 20th international conference on World wide web*, pp. 577--586. ACM.

Fan, W. & Bifet, A. (2014). Mining big data: Current status, and forecast to the future. *ACM SIGKDD Explorations Newsletter*, 16:1–5.

Fazel-Zarandi, M. & Fox, M. (2009). Semantic matchmaking for job recruitment: An ontology-based hybrid approach. *Proceedings of the 8th International Semantic Web Conference*, 525.

Garcia, R. & Amatriain, X. (2010). Weighted content based methods for recommending connections in online social networks. Em *Workshop on Recommender Systems and the Social Web*, pp. 68--71.

Grčar, M.; Fortuna, B.; Mladenič, D. & Grobelnik, M. (2006). knn versus svm in the collaborative filtering framework. *Data Science and Classification*, pp. 251--260.

Hong, W.; Zheng, S.; Wang, H. & Shi, J. (2013). A job recommender system based on user clustering. *JCP*, 8:1960–1967.

Horvitz, E.; Breese, J.; Heckerman, D.; Hovel, D. & Rommelse, K. (1998). The lumiere project: Bayesian user modeling for inferring the goals and needs of software users. Em *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pp. 256--265. Morgan Kaufmann Publishers Inc.

Hu, R. & Pu, P. (2010). Using personality information in collaborative filtering for new users. *Recommender Systems and the Social Web*, 17.

Huang, Z.; Chung, W.; Ong, T.-H. & Chen, H. (2002). A graph-based recommender system for digital library. Em *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pp. 65--73. ACM.

Jannach, D.; Zanker, M.; Felfernig, A. & Friedrich, G. (2010). *Recommender systems: an introduction*. Cambridge University Press.

Javed, F.; Luo, Q.; McNair, M.; Jacob, F.; Zhao, M. & Kang, T. S. (2015). Carotene: A job title classification system for the online recruitment domain. Em *Proceedings of the 2015 IEEE First International Conference on Big Data Computing Service and Applications*, BIGDATASERVICE '15, pp. 286--293, Washington, DC, USA. IEEE Computer Society.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98*, pp. 137--142.

Keim, T. (2007). Extending the applicability of recommender systems: A multilayer framework for matching human resources. pp. 169 – 169.

Kiesler, S.; Kraut, R.; CUMMINGS, J.; BONEVA, B.; HELGESON, V. & CRAWFORD, A. (2002). Internet evolution and social impact. *Information Technology - IT*, 1.

Kim, H.-N.; Alkhaldi, A.; El Saddik, A. & Jo, G.-S. (2011). Collaborative user modeling with user-generated tags for social recommender systems. *Expert Systems with Applications*, 38(7):8488--8496.

Kim, J. W.; Lee, B. H.; Shaw, M. J.; Chang, H.-L. & Nelson, M. (2001). Application of decision-tree induction techniques to personalized advertisements on internet storefronts. *International Journal of Electronic Commerce*, 5(3):45--62.

Kokkodis, M. (2019). Reputation deflation through dynamic expertise assessment in online labor markets. Em *The World Wide Web Conference*, WWW '19, pp. 896--905, New York, NY, USA. ACM.

Kureková, L. M.; Beblavý, M. & Thum-Thysen, A. (2015). Using online vacancies and web surveys to analyse the labour market: a methodological inquiry. *IZA Journal of Labor Economics*, 4(1):18. ISSN 2193-8997.

Lang, S.; Laumer, S.; Maier, C. & Eckhardt, A. (2011). Drivers, challenges and consequences of e-recruiting: A literature review. Em *Proceedings of the 49th SIGMIS Annual Conference on Computer Personnel Research*, SIGMIS-CPR '11, pp. 26--35, New York, NY, USA. ACM.

Lee, I. (2007). An architecture for a next-generation holistic e-recruiting system. *Commun. ACM*, 50(7):81--85. ISSN 0001-0782.

Li, C.; Ouyang, J. & Li, X. (2019). Classifying extremely short texts by exploiting semantic centroids in word mover's distance space. Em *The World Wide Web Conference*, WWW '19, pp. 939--949, New York, NY, USA. ACM.

Li, Q. & Kim, B. M. (2003). Clustering approach for hybrid recommender system. Em *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on*, pp. 33--38. IEEE.

Lika, B.; Kolomvatsos, K. & Hadjiefthymiades, S. (2014). Facing the cold start problem in recommender systems. *Expert Syst. Appl.*, 41(4):2065--2073. ISSN 0957-4174.

Liu, R.; Rong, W.; Ouyang, Y. & Xiong, Z. (2017). A hierarchical similarity based job recommendation service framework for university students. *Front. Comput. Sci.*, 11(5):912--922. ISSN 2095-2228.

Lo, S. & Lin, C. (2006). Wmr–a graph-based algorithm for friend recommendation. Em *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 121--128. IEEE Computer Society.

Lops, P.; De Gemmis, M. & Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. Em *Recommender systems handbook*, pp. 73--105. Springer.

Lu, Y.; El Helou, S. & Gillet, D. (2013). A recommender system for job seeking and recruiting website. Em *Proceedings of the WWW 2013*, pp. 963--966, New York, NY, USA. ACM.

Malinowski, J.; Keim, T.; Wendt, O. & Weitzel, T. (2006). Matching people and jobs: A bilateral recommendation approach. Em *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, volume 6, pp. 137c–137c. ISSN .

Martin, R.; Walid, M.; Robert, W. & Thomas, Z. (2014). *Recommendation Systems in Software Engineering*. Springer.

Mavlanova, T.; Benbunan-Fich, R. & Lang, G. (2016). The role of external and internal signals in e-commerce. *Decision Support Systems*, 87:59--68.

Melville, P.; Mooney, R. J. & Nagarajan, R. (2002). Content-boosted collaborative filtering for improved recommendations. Em *Aaai/iaai*, pp. 187--192.

MENG, X.-L. & RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278. ISSN 0006-3444.

Mihuandayani; Utami, E. & Luthfi, E. T. (2018). Profiling analysis based on social media for prospective employees recruitment using SVM and chi-square. *Journal of Physics: Conference Series*, 1140:012043.

Mikolov, T.; Chen, K.; Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mooney, R. J. & Roy, L. (2000). Content-based book recommending using learning for text categorization. Em *Proceedings of the fifth ACM conference on Digital libraries*, pp. 195--204. ACM.

Mourão, F. H. d. J. (2014). A hybrid recommendation method that combines forgotten items and non-content attributes.

Niwattanakul, S.; Singthongchai, J.; Naenudorn, E. & Wanapu, S. (2013). Using of jaccard coefficient for keywords similarity.

Paparrizos, I.; Cambazoglu, B. B. & Gionis, A. (2011). Machine learned job recommendation. Em *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pp. 325--328, New York, NY, USA. ACM.

Park, Y.-J. & Tuzhilin, A. (2008). The long tail of recommender systems and how to leverage it. Em *Proceedings of the 2008 ACM conference on Recommender systems*, pp. 11--18. ACM.

Pazzani, M. J. (1999). A framework for collaborative, content-based and demographic filtering. *Artificial intelligence review*, 13(5-6):393--408.

Peter, K. & Hani, M. (2014). Is internet job search still ineffective? *The Economic Journal*, 124(581):1213–1233.

Pizzato, L.; Rej, T.; Chung, T.; Koprinska, I. & Kay, J. (2010). Recon: A reciprocal recommender for online dating. Em *Proc. of the ACM RecSys 2010*, pp. 207--214, New York, NY, USA. ACM.

Popescul, A.; Pennock, D. M. & Lawrence, S. (2001). Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. Em *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pp. 437--444. Morgan Kaufmann Publishers Inc.

Pournajaf, L.; Aljadda, K. & Korayem, M. (2017). Long tail query enrichment for semantic job search. Em *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 215–220. ISSN .

Ramos, E. (2015). *E-commerce*. Editora FGV.

Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. Em *Proceedings of the first instructional conference on machine learning*, volume 242, pp. 133--142.

Ricci, F.; Rokach, L. & Shapira, B. (2011). *Introduction to recommender systems handbook*. Springer.

Salehi, B.; Spina, D.; Moffat, A.; Sadeghi, S.; Scholer, F.; Baldwin, T.; Cavedon, L.; Sanderson, M.; Wong, W. & Zobel, J. (2018). A living lab study of query amendment in job search. Em *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pp. 905--908, New York, NY, USA. ACM.

Sarwar, B.; Karypis, G.; Konstan, J. & Riedl, J. (2000). Application of dimensionality reduction in recommender system-a case study. Relatório técnico, DTIC Document.

Schafer, J. B.; Frankowski, D.; Herlocker, J. & Sen, S. (2007). Collaborative filtering recommender systems. Em *The adaptive web*, pp. 291--324. Springer.

Silva, N. B.; Tsang, R.; Cavalcanti, G. D. & Tsang, J. (2010). A graph-based friend recommendation system using genetic algorithm. Em *Evolutionary Computation (CEC), 2010 IEEE Congress on*, pp. 1--7. IEEE.

Spina, D.; Maistro, M.; Ren, Y.; Sadeghi, S.; Wong, W.; Baldwin, T.; Cavedon, L.; Moffat, A.; Sanderson, M.; Scholer, F. & Zobel, J. (2017). Understanding user behavior in job and talent search: An initial investigation. Em *Proceedings of the SIGIR 2017 Workshop On eCommerce co-located with the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Toutanova, K.; Chen, D.; Pantel, P.; Poon, H.; Choudhury, P. & Gamon, M. (2015). Representing text for joint embedding of text and knowledge bases. Em *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1499--1509, Lisbon, Portugal. Association for Computational Linguistics.

Tran, M.; Nguyen, A.; Nguyen, Q. & Huynh, T. (2017). A comparison study for job recommendation. Em *2017 International Conference on Information and Communications (ICIC)*, pp. 199–204. ISSN .

Turrell, A.; Thurgood, J.; Djumalieva, J.; Copple, D. & Speigner, B. (2018). Using online job vacancies to understand the UK labour market from the bottom-up. Relatório técnico.

ul haq Dar, E. & Dorn, J. (2018). Classification of job offers of the world wide web. Em *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pp. 1–8. ISSN .

Van Meteren, R. & Van Someren, M. (2000). Using content-based filtering for recommendation. Em *Proceedings of the Machine Learning in the New Information Age: MLnet/ECML2000 Workshop*, pp. 47--56.

Yang, S.; Korayem, M.; AlJadda, K.; Grainger, T. & Natarajan, S. (2017). Combining content-based and collaborative filtering for job recommendation system. *Know.-Based Syst.*, 136(C):37--45. ISSN 0950-7051.

Yang, X.; Guo, Y.; Liu, Y. & Steck, H. (2014). A survey of collaborative filtering based social recommender systems. *Computer Communications*, 41:1--10.

Yu, H.; Liu, C. & Zhang, F. (2011). Reciprocal recommendation algorithm for the field of recruitment. 8:4061–4068.