Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Recomendação de Artigos Científicos: um Foco na Integração de Perfis de Usuários

Jônathas José de Magalhães

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Modelos Computacionais e Cognitivos

Evandro de Barros Costa
Joseana Macêdo Fechine Régis de Araújo
(Orientadores)

Campina Grande, Paraíba, Brasil

Setembro de 2013

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

M118r Magalhães, Jônathas José de.

Recomendação de artigos científicos : um foco na integração de perfis de usuários / Jônathas José de Magalhães. – Campina Grande, 2013.

63 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2013.

"Orientação: Prof. Dr. Evandro de Barros Costa, Profª. Drª. Joseana Macêdo Fechine Régis de Araújo".

Referências.

1. Sistemas de Recomendação. 2. Modelagem de Usuário. 3. Sistema de Recomendação de Artigos. 4. Integração de Perfis de Usuário. 5. Filtragem Baseada em Conteúdo. I. Costa, Evandro de Barros. II. Araújo, Joseana Macêdo Fechine Régis de. III. Título.

CDU 004.5(043)

"RECOMENDAÇÃO DE ARTIGOS CIENTÍFICOS: UM FOCO NA INTEGRAÇÃO DE PERFIS DE USUÁRIOS"

JÔNATHAS JOSÉ DE MAGALHÃES

DISSERTAÇÃO APROVADA EM 20/09/2013

EVANDRO DE BARROS COSTA, D.Sc, UFAL
Orientador(a)

JOSEANA MACÊDO FECHINE RÉGIS DE ARAÚJO, D.Sc, UFCG
Orientador(a)

NAZARENO FERREIRA DE ANDRADE, D.Sc, UFCG
Examinador(a)

GUILHERME ATAÍDE DIAS, Phd, UFPB
Examinador(a)

CAMPINA GRANDE - PB

Resumo

Os Sistemas de Recomendação personalizada surgiram como uma possível solução para o problema da sobrecarga de informação. Entretanto, sua qualidade está relacionada ao perfil de usuário e gerar um perfil de qualidade não é uma tarefa trivial. Consequentemente, o usuário que não recebe boas recomendações poderá perder o interesse e confiança no sistema. A pesquisa ora apresentada trata deste problema propondo uma abordagem para Sistemas de Recomendação de artigos científicos com foco na integração de perfis de usuário. Os perfis foram construídos a partir de três fontes: CV Lattes, Mendeley e LinkedIn. A integração de perfis de usuário foi realizada por meio de combinação linear, propondo-se três estratégias: (i) importância igual (Igual); (ii) quantidade de itens (Quant); e (iii) atividade do usuário na fonte (Ativ). Para validar os modelos de perfis, foi realizado um experimento em que os participantes analisaram a relevância de 50 artigos, sendo utilizada a métrica NDCG@5. Foram realizadas duas avaliações, a primeira apenas no Lattes, utilizando como fator a estratégia de construção de perfil, tendo sido avaliadas as seguintes estratégias: termos (LT); conceitos (LC) e estratégia de Lopes. As estratégias propostas proporcionaram os melhores resultados, conforme o teste de Wilcox (α = 0,05): Hipótese Alternativa (H_A) = $LT > Lopes$ (p-$valor$ = 0,01543) e $H_A = LC > Lopes$ (p-$valor$ = 0,04292). Na segunda avaliação, com os perfis integrados, foram utilizados dois fatores: representação do perfil (termos e conceitos) e estratégia de integração (Igual; Quant; Ativ). Os perfis integrados não proporcionaram resultados melhores que os perfis não integrados, conforme o teste de Friedman (α = 0,05): $H_A =$ Existe diferença (p-$valor$ = 0,9971). De posse dos resultados, pôde-se concluir que o modelo proporcionou resultados satisfatórios na plataforma Lattes, o que pode ser caracterizado como uma contribuição importante, dada a importância desta plataforma para os pesquisadores brasileiros. Em se tratando da integração de perfis, não foram alcançados os resultados esperados. Neste sentido, verifica-se que o modelo de integração precisa ser investigado com mais aprofundamento, seja realizando um experimento com mais fatores ou buscando uma amostra maior de usuários.

Palavras-chave: Sistemas de Recomendação, Modelagem de Usuário, Sistema de Recomendação de Artigos, Integração de Perfis de Usuário, Filtragem Baseada em Conteúdo.

Abstract

The personalized Recommender Systems have emerged as a possible solution to the information overload problem. However, their quality is related to the user profile and generate a profile with quality is not a trivial task. Consequently, the user that does not receive good recommendations may lose interest and confidence in the system. Our research presented here addresses this problem by proposing an approach to paper Recommendation Systems focusing on the integration of user profiles. The profiles were constructed from three sources: CV Lattes, Mendeley and LinkedIn. The integration of user profiles was performed by linear combination and we proposed three strategies: (i) equal importance (Igual); (ii) quantity of items (Quant); and (iii) user activity on the source (Ativ). To validate the profile models, we performed an experiment in which the participants evaluated the relevance of 50 papers, we used the metric NDCG@5. We performed two evaluations, the first only in Lattes, we used the strategy of building profile as a factor and evaluated the following strategies: terms (LT); concepts (LC) and Lopes strategy. The proposed strategies provided the best results, according to the Wilcox's test ($\alpha = 0.05$): Alternative Hypothesis (H_A) = $LT > Lopes$ ($p\text{-}value$ = 0.01543) and H_A = $LC > Lopes$ ($p\text{-}value$ = 0.04292). In the second evaluation, with the integrated profiles, we used two factors: profile representation (terms and concepts) and integration strategy (Igual; Quant; Ativ). The integrated profiles did not provide better results than non-integrated profiles, according to the Friedman's test (α = 0.05): $H_A =$ There is difference ($p\text{-}value$ = 0.9971). Based on the results, we can conclude that the model provided satisfactory results in the Lattes platform, which can be characterized as an important contribution, given the importance of this platform for Brazilian researchers. Concerning the profiles integration, we did not achieved the expected results. In this sense, we verify that the integration model needs further investigation, whether conducting an experiment with more factors or with a larger sample of users.

Keywords: Recommender Systems, User Modeling, Paper Recommender Systems, User Profiles Integration, Content-based Filtering.

Agradecimentos

Agradeço primeiramente a Deus por me iluminar nos momentos mais obscuros da vida, ora me dando força e inspiração para enfrentar os desafios, ora me dando respostas aos vários questionamentos que surgem ao longo do caminho. Acredito que a Ciência seja uma forma de desvendar os mistérios da vida e de se aproximar de Deus.

Em segundo lugar, obviamente, agradeço a minha mãe, dona Fátima, que sempre me apóia e me dá conselhos valiosos para a vida; sem falar que desta vez a exploração foi grande, de forma que ela até me ajudou na revisão deste trabalho. Agradeço também a meu pai, seu Jonas, por sua camaradagem e sua espontaneidade.

Ao grande amigo que fiz nesse mestrado, meu orientador Evandro, pelos direcionamentos e debates acadêmicos, pelas discussões não-acadêmicas e os muitos conselhos que guardarei para a vida. À minha orientadora Joseana, que é uma espécie de "dona Fátima acadêmica", sempre com uma paciência infinita, conselhos importantes e puxões de orelha nos momentos oportunos.

Agradeço ao Cleyton, Heitor e Marlos que foram como irmãos acadêmicos que fiz nesse mestrado. Ao Cleyton pela amizade formada desde o início do mestrado, pelas discussões sempre produtivas e pelas cobranças, ele pode ser considerado como um orientador honorário deste trabalho. Ao Marlos e o Heitor por sempre emprestarem seus ouvidos para ouvir minhas teorias e análises, um tanto peculiares, que faço sobre a vida e a ciência. Ao pessoal do laboratório, em especial Priscylla, Joyse, Tarsys, Anderson e Manu pela amizade e cumplicidade no trabalho.

Ao grupo TIPS, pelas discussões geradas que muito contribuíram para a formação do eu acadêmico. Agradeço ao Baldoino e ao Márcio, que com suas sugestões, muito contribuiram para este trabalho. Agradeço também ao Vinnicyus por ter me ajudado e muito com a implementação do sistema utilizado na validação.

No âmbito não-acadêmico, agradeço ao meu parceiro de apartamento, meu primo João Paulo, pela companhia nos momentos mais solitários. Agradeço também ao Everaldo Cig e a Sarah pelos momentos de descontração nos poucos momentos de pausa no trabalho. Ao Rodolfo, Edvaldo, Diogo, Dário e Angélica pelas conversas sobre a vida, futuro e amizade.

Em memória de Alfredo, por sempre acreditar em mim e fazer de seu local de trabalho a minha casa enxadrística. Em memória de vó Laura, por seu carinho, acolhimento e me adotar como neto.

Às amizades que fiz em Campina Grande, em especial o André, pelos bons momentos de distração e compartilhamento de experiências. Ao Paulo Barbosa que me apresentou ~~aos capivaras~~ ao público enxadrístico de Campina Grande, Joca, Robson, Jovany e Evandro.

Aos professores membros da banca examinadora da proposta de Mestrado, Leandro e Nazareno pela importante contribuição para este trabalho. Aos professores membros da banca de Mestrado, Nazareno e Guilherme pela disponibilidade e pelos comentários valiosos a respeito deste trabalho.

Por fim, aos voluntários que doaram seu valioso tempo para participar do experimento, nossa Senhora da Ciência vos abençoe.

PS. Se alguém se considera esquecido neste texto de agradecimentos, por favor mande-me e-mail que eu considerarei te incluir nos agradecimentos da tese de doutorado.

Conteúdo

1 Introdução **1**

1.1 Contextualização 1

1.2 Problemática 3

1.2.1 Problema de Negócio 3

1.2.2 Problema Técnico 4

1.3 Objetivos e Relevância da Pesquisa 4

1.4 Organização da Dissertação 4

2 Fundamentação Teórica **6**

2.1 Sistemas de Recomendação 6

2.1.1 Tipos de Sistemas de Recomendação 7

2.1.2 Avaliação de Sistemas de Recomendação 12

2.2 Discussão 13

3 Trabalhos Relacionados **14**

3.1 Integração de Perfis de Usuário 14

3.2 SR de Artigos Científicos 15

3.2.1 Abordagem *artigo-artigo* 16

3.2.2 Abordagem *usuário-artigo* 17

3.2.3 Análise Comparativa entre os Trabalhos 19

3.3 Sobre o Diferencial deste Trabalho 20

3.4 Discussão 20

4 Modelo para Integração de Perfis de Usuário **21**

4.1 Definições Preliminares 21

4.1.1 Determinando o Conhecimento de Domínio 22
4.2 A Construção dos Perfis de Usuário numa Fonte 23
4.3 Integração de Perfis de Usuário 25
4.4 Indexando Itens 26
4.5 Recomendando Itens 27
4.6 Discussão 27

5 Sistema que Instancia o Modelo de Integração 29
5.1 Visão Geral do Sistema 29
5.2 Arquitetura do Sistema 33
5.2.1 Módulo *Crawler* 34
5.2.2 Módulo Analisador 36
5.2.3 Módulo Construtor de Conhecimento 37
5.2.4 Módulo Indexador de Artigos 38
5.2.5 Módulo Construtor de Perfis 39
5.2.6 Módulo Recomendador 40
5.3 Discussão 41

6 Validação 42
6.1 Metodologia 42
6.2 Analisando a Plataforma Lattes 43
6.2.1 Resultados e Discussão 43
6.3 Analisando os Perfis Integrados 45
6.3.1 Resultados e Discussão 46
6.4 Definindo um Modelo de Recomendação 48
6.4.1 Estrutura e Nós da RB 48
6.4.2 Definindo os Pesos da RB 51
6.4.3 Utilizando o Modelo para Recomendação 51
6.5 Ameaças à Validade 55

7 Considerações Finais 57
7.1 Caracterização Geral da Pesquisa 57

7.2 Principais Contribuições 58
7.3 Sugestões para Trabalhos Futuros 58

A Artigo Publicado no 3rd SRS 2012 **63**

Lista de Símbolos

ACM - *Association for Computing Machinery.*

API - *Interface de Programação de Aplicativos, do inglês, Application Programming Interface*.

CNPq - *Conselho Nacional de Desenvolvimento Científico e Tecnológico*.

DCG - *Desconto do Ganho Cumulativo, do inglês, Discounted Cummulative Gain*.

EQR - *Erro Quadrado Regularizado*.

FBC - *Filtragem Baseada em Conteúdo*.

FC - *Filtragem Colaborativa*.

FM - *Fatoração de Matrizes*.

IDF - *Frequência Inversa do Documento, do inglês, Inverse Document Frequency.*

IEEE - *Instituto de Engenheiros Eletricistas e Eletrônicos*.

NDCG - *Desconto Normalizado do Ganho Cumulativo, do inglês, Normalized Discounted Cummulative Gain*.

NLTK - *Natural Language Toolkit*.

PLN - *Processamento de Linguagem Natural*.

RB - *Redes Bayesianas*.

RI - *Recuperação de Informação*.

SR - *Sistema de Recomendação*.

TF - *Frequência do Termo, do inglês, Term Frequency.*

TF-IDF - *Frequência do Termo/Frequência Inversa do Documento, do inglês, Term Frequency/Inverse Document Frequency.*

UFAL - *Universidade Federal de Alagoas*.

UFCG - *Universidade Federal de Campina Grande*.

UFMG - *Universidade Federal de Minas Gerais*.

XML - *eXtensible Markup Language*.

Lista de Figuras

4.1 Processo de construção do conhecimento. 23
4.2 Processo de construção do perfil de termos do usuário u. 24
4.3 Processo de construção do perfil de conceitos do usuário u. 24

5.1 Tela inicial do sistema. 31
5.2 Tela de exibição das áreas de conhecimento em Ciência da Computação, em que o usuário pode especificar seus interesses em cada área. 31
5.3 Parte da tela de exibição dos artigos, em que os usuários podem inserir a relevância nos mesmos. 32
5.4 Arquitetura geral do sistema, módulos e interações. 33
5.5 Etapas do pré-processamento textual. 36
5.6 Arquitetura do módulo construtor de conhecimento. 37
5.7 Arquitetura do módulo indexador de artigos. 38
5.8 Arquitetura do módulo construtor de perfis. 40
5.9 Arquitetura do módulo recomendador. 40

6.1 Avaliação da recomendação utilizando diferentes estratégias de construção de perfil de usuário na plataforma Lattes. 44
6.2 Avaliação da recomendação utilizando diferentes estratégias de construção de perfil integrado de usuário nas plataformas Lattes, Mendeley e LinkedIn. 47
6.3 Modelo de recomendação utilizando Redes Bayesianas. 49
6.4 Inferência no modelo: recomendando um método de recomendação para usuários que possuam informação apenas no Lattes. Imagem da ferramenta Netica. 53

6.5 Inferência no modelo: recomendando um método de recomendação para usuários que possuam informação no Lattes e no LinkedIn. Imagem da ferramenta Netica. 54

6.6 Inferência no modelo: recomendando um método de recomendação para usuários que possuam informação no Lattes e no Mendeley. Imagem da ferramenta Netica. 54

6.7 Inferência no modelo: recomendando um método de recomendação para usuários que possuam informação em todas as fontes (Lattes, Mendeley e LinkedIn). Imagem da ferramenta Netica. 55

Lista de Tabelas

2.1 Exemplo de uma matriz $R_{|U|\times|D|}$ de avaliações de um SR. 7

3.1 Comparação entre os trabalhos relacionados. 19

5.1 Os níveis de interesse e relevância de acordo com a marcação de estrelas. . 32

5.2 Mapeamento dos dados externos dos usuários para o metamodelo utilizado pelo sistema. O símbolo "+" indica a concatenação de caracteres. 35

5.3 Quantidade de artigos por cada subárea de Inteligência Artificial e Engenharia de Software. 39

6.1 Média e mediana do NDCG@5 das estratégias na plataforma Lattes. 43

6.2 Testes de hipótese realizados na comparação das estratégias na plataforma Lattes. 44

6.3 Perfis de usuário e suas respectivas siglas. 46

6.4 Média e mediana do NDCG@5 das estratégias de perfil integrado nas plataformas Lattes, Mendeley e LinkedIn. 47

6.5 Testes de hipótese realizados na comparação das estratégias de perfis integrados nas plataformas Lattes, Mendeley e LinkedIn. 48

6.6 Estados e mapeamento dos nós $lattes\_bibliografia$, $lattes\_tecnica$, $linkedin$, $mendeley$ e $lattes\_projetos$. 51

6.7 Conjunto de casos utilizados no aprendizado dos pesos do modelo. 52

Capítulo 1

Introdução

O trabalho ora apresentado está inserido nas áreas de Modelagem do Usuário e Sistemas de Recomendação (SR). Neste capítulo, será apresentado um panorama geral da pesquisa, iniciando-se com uma contextualização e a motivação do estudo. Em seguida, serão discutidos a problemática e os objetivos vinculados ao trabalho. Por fim, será descrita a organização do documento.

1.1 Contextualização

Nas últimas décadas, tem ocorrido um aumento no desenvolvimento da pesquisa científica, aumentando, consequentemente, o número de trabalhos científicos publicados (TORRES et al., 2004), segundo Torres et al. (2004), cada ano, desde 1986, tem havido acréscimo de 1% no número de artigos científicos publicados. Anualmente, acontecem centenas e talvez milhares de eventos científicos em todo planeta. Com a popularização da internet, diferente do que acontecia há 20 anos, os artigos publicados nessas conferências, simpósios, workshops, dentre outros, são compartilhados na Web e ficam disponíveis para qualquer um e em qualquer lugar. Um exemplo bem conhecido da comunidade científica são as bibliotecas digitais, algumas de acesso proprietário e.g., *Association for Computing Machinery* (ACM[1]) e Instituto de Engenheiros Eletricistas e Eletrônicos (IEEE[2]), e outras de acesso aberto e.g.,

[1] <http://portal.acm.org/>
[2] <http://ieeexplore.ieee.org/>

arXiv[3] e CEUR[4], que disponibilizam aos seus usuários um vasto acervo de artigos científicos; tal acervo tende a aumentar constantemente seu volume, devido à periodicidade e ao surgimento de novos veículos de publicação. Em decorrência desses fatos, torna-se cada vez mais difícil para um usuário encontrar material relevante e de qualidade para sua pesquisa em tempo hábil. De forma geral, esse fato é caracterizado como o *problema da sobrecarga da informação* (do inglês, *information overload problem*) (BAEZA-YATES; RIBEIRO-NETO, 2011).

As bibliotecas digitais, em sua maioria, disponibilizam ferramentas de busca de forma a ajudar seus usuários a encontrar conteúdo relevante. No entanto, sob a perspectiva do usuário, construir a busca com os termos condizentes ao seu interesse nem sempre é uma tarefa simples. Isso pode ocorrer por diferentes motivos: o usuário pode ainda ser inexperiente no domínio a ser pesquisado ou então artigos relevantes podem não aparecer no resultado por não possuírem os termos exatos da busca (SUGIYAMA; KAN, 2010).

Os Sistemas de Recomendação têm sido utilizados para reduzir este problema e, nesse contexto, ajudar os usuários a encontrar artigos[5] relevantes (ADOMAVICIUS; TUZHILIN, 2005). Na literatura, as técnicas de recomendação de artigos científicos são divididas em duas categorias (JIANG et al., 2012): (i) *artigo-artigo*, em que a recomendação é baseada na similaridade entre artigos, analisando-se as citações de um dado artigo ou de um conjunto de artigos (e.g., McNee et al. (2002), Gori e Pucci (2006), Nascimento et al. (2011)); e (ii) *usuário-artigo*, em que o objetivo é recomendar artigos com base nas preferências do usuário, por meio de uma análise do conteúdo por ele consumido[6] (e.g., Lopes, Souto e Wives (2007), Sugiyama e Kan (2010), Goossen et al. (2011), Magalhães et al. (2012)). Essas preferências do usuário são utilizadas para criação de um perfil de usuário.

O presente trabalho concentra-se na abordagem *usuário-artigo*, fazendo uma análise da produção bibliográfica do usuário e considerando que esse conteúdo é uma fonte de dados[7] que fornece elementos para recomendar artigos relevantes ao usuário (LOPES; SOUTO; WIVES, 2007; SUGIYAMA; KAN, 2010; MAGALHÃES et al., 2012). No entanto, um

[3] <http://arxiv.org/>

[4] <http://ceur-ws.org/>

[5] Neste trabalho, o termo "artigo" é considerado sinônimo de "publicações científicas".

[6] O termo "consumir" é utilizado para designar o ato do usuário criar um conteúdo ou utilizá-lo de alguma forma.

[7] O termo "Fonte de dados" ou apenas "fonte" caracteriza-se como sendo um lugar virtual que permite que os seus usuários consumam conteúdo, e.g. blogs, páginas pessoais, currículos online, redes sociais, etc.

problema bastante conhecido desta estratégia ocorre quando o usuário não possui conteúdo suficiente para gerar a recomendação adequada, que é o caso de alunos de iniciação científica ou pesquisadores no início de sua carreira acadêmica. Na área de SR, este problema é conhecido como o problema do novo usuário (do inglês, *cold-start problem*) (SCHEIN et al., 2002).

Uma estratégia para lidar com o problema do novo usuário consiste na Filtragem Colaborativa (FC). No entanto, na FC são utilizadas apenas as avaliações dos itens e não se faz uso de informações adicionais sobre os itens e usuários, o que inviabiliza a sua utilização em determinados contextos (BERKOVSKY; KUFLIK; RICCI, 2008). Outra estratégia para lidar com esse problema consiste em analisar o conteúdo consumido pelo usuário em outras fontes de dados com o objetivo de melhorar o seu perfil, fazendo uma integração de perfis de usuário (BERKOVSKY; KUFLIK; RICCI, 2008; SAHEBI; WONGCHOKPRASITTI; BRUSILOVSKY, 2010; MAGALHÃES et al., 2012). Tal estratégia é fundamentada no fato de os usuários disponibilizarem, com bastante frequência, informações pessoais sobre suas habilidades e preferências em múltiplos ambientes virtuais, os quais funcionam como fontes de dados (WANG; ZHANG; VASSILEVA, 2010).

Diante do contexto apresentado, este trabalho propõe um modelo para integração de perfis de usuário no intuito de aprimorar o vetor de informação de suporte a um SR de artigos científicos, possuindo a hipótese de que o perfil integrado de usuário resultará num perfil mais acurado do ponto de vista da qualidade da recomendação[8].

1.2 Problemática

A problemática deste trabalho está dividida em: *problema de negócio*, que consiste nas dificuldades observadas com respeito ao público alvo deste trabalho; e *problema técnico*, que envolve os desafios técnicos que serão enfrentados para se resolver o problema de negócio.

1.2.1 Problema de Negócio

No contexto de Sistemas de Recomendação personalizada, a qualidade da recomendação está relacionada com o perfil de usuário. No entanto, gerar um perfil de usuário com qualidade,

[8] Qualidade da recomendação no sentido da relevância do conteúdo recomendado para o usuário.

isto é, que seja automático, personalizado e que represente seus interesses atuais não é uma tarefa trivial. Dessa forma, o usuário que não possui um perfil de qualidade provavelmente não receberá boas recomendações, e poderá perder o interesse e a confiança no sistema.

1.2.2 Problema Técnico

O problema técnico consiste em definir um perfil de usuário automático, personalizado e que seja representativo de seus interesses atuais. Em decorrência desse problema surgem alguns desafios técnicos, quais sejam: (i) quais dados extrair do usuário nas diferentes fontes de dados em que ele está presente; (ii) como extrair os dados de usuário nas diferentes fontes de dados em que ele está presente; (iii) como representar e construir o perfil de usuário numa única fonte de dados; e (iv) como fazer a integração dos perfis de usuário provindos de diversas fontes de dados.

1.3 Objetivos e Relevância da Pesquisa

O objetivo geral deste trabalho consiste em definir e validar um modelo de integração de perfis de usuário provindos de diferentes fontes de dados. O perfil de usuário resultante do modelo servirá como entrada para um SR de artigos científicos. Assim, espera-se que, por meio da integração de perfis, seja obtido um perfil de usuário mais acurado, e que isto tenha impacto positivo na qualidade da recomendação.

Como objetivos específicos pode-se enumerar: (i) realizar um levantamento bibliográfico sobre SR de artigos científicos e integração de perfis de usuário; iii) elaborar um modelo de integração de perfis; iv) planejar um experimento envolvendo o modelo construído; e v) validar o modelo proposto.

1.4 Organização da Dissertação

O restante desta dissertação está organizado conforme descrição a seguir.

- Capítulo 2 – **Fundamentação Teórica**, apresenta o embasamento teórico relacionado a este trabalho. São descritos alguns conceitos que envolvem sistemas de recomendação e modelagem de usuário;

- Capítulo 3 – **Trabalhos Relacionados**, são apresentados trabalhos relacionados ao presente, com exposição das características, vantagens e desvantagens desses trabalhos;

- Capítulo 4 – **Modelo de Integração de Perfis de Usuário Proposto**, apresenta detalhes sobre o modelo de integração de perfis;

- Capítulo 5 – **Sistema que Instancia o Modelo de Integração**, apresenta o sistema desenvolvido com vistas à validação do modelo proposto;

- Capítulo 6 – **Validação do Modelo**, apresenta a metologia adotada e os resultados obtidos para o modelo proposto;

- Capítulo 7 – **Considerações Finais**, apresenta as principais conclusões, além de indicar possíveis desdobramentos imediatos do trabalho.

Capítulo 2

Fundamentação Teórica

Neste capítulo, são descritos os principais conceitos relevantes ao contexto do trabalho ora apresentado.

2.1 Sistemas de Recomendação

Os SR surgiram como uma possível solução para o problema da sobrecarga da informação. A principal tarefa de um SR é prever qual é a utilidade de um determinado item para um determinado usuário. Formalmente, um SR pode ser definido da seguinte forma (ADOMAVICIUS; TUZHILIN, 2005): seja U um conjunto de usuários e seja D um conjunto de itens que podem ser recomendados, os usuários U e os itens D estão relacionados por uma função de utilidade $util$ que mede a utilidade de um item $d \in D$ para um usuário $u \in U$. Esta definição pode ser representada conforme Eq. 2.1.

$$util : U \times D \rightarrow R_{|U|\times|D|}, \tag{2.1}$$

em que R é uma matriz de avaliações (do inglês, *rating matrix*) de ordem $|U| \times |D|$ e, r_{ij} representa a avaliação do usuário u_i sobre o item d_j. As avaliações podem ser, por exemplo, um número natural ou um número real no intervalo $[0, 1]$. Na Tabela 2.1 é apresentado um exemplo de uma matriz $R_{|U|\times|D|}$ de avaliação, em que cada avaliação consiste em um número inteiro de 1 a 5, em que 5 significa a melhor avaliação possível. O símbolo $\varnothing$ significa que o item ainda não foi avaliado pelo usuário, por exemplo, $r_{1,3} = \varnothing$ significa que o usuário u_1

não avaliou o item d_3.

Tabela 2.1: Exemplo de uma matriz $R_{|U|\times|D|}$ de avaliações de um SR.

$R_{\|U\|\times\|D\|}$	d_1	d_2	d_3	d_4	...	$d_{\|D\|}$
u_1	2	3	$\varnothing$	$\varnothing$	...	4
u_2	1	$\varnothing$	2	$\varnothing$	...	5
...	...	...	...	...	...	...
$u_{\|U\|}$	$\varnothing$	1	$\varnothing$	3	...	2

Dados um usuário u_i e um conjunto de itens que ele ainda não tenha avaliado, representado por $D'_{u_i} \subseteq D = \{d_j \in D | r_{i,j} = \varnothing \wedge j = 1...|D|\}$, o objetivo de um SR é recomendar itens $d \in D'_{u_i}$ de forma a maximizar a utilidade do item para o usuário u_i, conforme Eq. 2.2.

$$D^{rec}_{u_i} = \underset{d \in D'_{u_i}}{\operatorname{argmax}^{n}} util(u_i, d), \tag{2.2}$$

em que a função argmax retorna o conjunto $D^{rec}_{u_i}$ que contém os n itens que maximizam a função de utilidade $util$ em relação ao usuário u_i. Entretanto, a utilidade dos itens $d \in D'_{u_i}$ *a priori* é desconhecida, então a tarefa principal de um SR consiste em estimá-las.

2.1.1 Tipos de Sistemas de Recomendação

Os SRs podem ser classificados de acordo com o tipo de dado que é utilizado para prover a recomendação, por exemplo, as características de usuários e dos itens que serão recomendados. Desta forma, os SRs são baseados em Filtragem Colaborativa (FC) ou, Filtragem Baseada em Conteúdo (FBC) ou em abordagens híbridas.

Filtragem Colaborativa (FC)

Na Filtragem Colaborativa (HERLOCKER et al., 1999), um usuário obterá recomendação de itens que foram bem avaliados por usuários com preferências similares as dele. Sua função de utilidade é representada de acordo com a Eq. 2.3 (BERKOVSKY; KUFLIK; RICCI, 2008).

$$util_{CF} : U_{id} \times D_{id} \rightarrow R_{|U|\times|D|}, \tag{2.3}$$

em que U_{id} representa os identificadores únicos dos usuários e D_{id} representa os identificadores únicos dos itens. Geralmente, na FC não são utilizadas informações adicionais sobre as características dos itens e dos usuários, o que inviabiliza a sua utilização em determinados contextos (BERKOVSKY; KUFLIK; RICCI, 2008). Basicamente, os algoritmos de FC são divididos em duas categorias (BOBADILLA et al., 2013):

- *Baseados em memória* – são algoritmos que atuam apenas na matriz de avaliações e que geralmente utilizam métricas de similaridade para obter a distância entre dois usuários ou entre dois itens. Um dos exemplos mais conhecidos dessa categoria são os algoritmos baseados em vizinhança que são executados em três passos (HERLOCKER et al., 1999):

  1. Calcular a similaridade do usuário ativo em relação aos outros usuários, nesse sentido pode ser utilizada a similaridade do cosseno, conforme Eq. 2.4 (ADOMAVICIUS; TUZHILIN, 2005).

$$sim(u,v) = \cos(\vec{u},\vec{v}) = \frac{\vec{u}\cdot\vec{v}}{\parallel \vec{u} \parallel \times \parallel \vec{v} \parallel} = \frac{\sum_{i=1}^{|D|} r_{u,i} r_{v,i}}{\sqrt{\sum_{i=1}^{|D|} r_{u,i}^2}\sqrt{\sum_{i=1}^{|D|} r_{v,i}^2}}; \qquad (2.4)$$

  em que $\|\cdot\|$ denota a norma Euclidiana padrão e $\cos(\vec{u},\vec{v})$ retorna um número real no intervalo $[0,1]$, quanto mais próximo de 1 mais os usuários serão similares.

  2. Selecionar o subconjunto de usuários que será utilizado como conjunto preditor, esse conjunto V_u será composto pelos k vizinhos mais próximos ao usuário ativo u, como apresentado na Eq. 2.5.

$$V_u = \underset{v \in U - \{u\}}{\overset{k}{\operatorname{argmax}}}\, sim(u,v), \qquad (2.5)$$

  em que $sim(u,v)$ é a função de similaridade apresentada na Eq. 2.4.

  3. Normalizar as avaliações e computar a predição baseando-se numa combinação das avaliações do conjunto preditor. Nesse intuito pode ser utilizada a média das avaliações do conjunto V_u, de acordo com a Eq. 2.6.

$$util(u,d) = \frac{1}{|V_u|}\sum_{v \in V_u} r(v,d). \qquad (2.6)$$

Mais detalhes sobre FC utilizando métodos baseados em memória estão disponíveis em (HERLOCKER et al., 1999).

- *Baseados em modelo* – são métodos que utilizam a informação do SR, neste caso, a matriz de avaliações, para construir um modelo que gera recomendações. Um dos métodos que mais tem chamado atenção da comunidade científica é a Fatoração de Matrizes (FM) (LUO; XIA; ZHU, 2012), que foi publicado pela primeira vez por Webb (2006). O método consiste em representar a matriz de avaliações R como um produto de duas matrizes de valores latentes e aprender esses valores por meio de técnicas de otimização. Primeiramente são definidos dois parâmetros de valores latentes com dimensão f, a matriz $P \in \mathbb{R}^{|U| \times f}$ que representa os usuários, em que cada linha representa um determinado usuário e a matriz $Q \in \mathbb{R}^{f \times |D|}$ que representa os itens, em que cada coluna representa um determinado item. A avaliação $\hat{r}_{u,i}$ de um usuário u em determinado item d é estimada pelo produto interno das características do usuário e das características do item, conforme Eq. 2.7.

$$\hat{r}_{u,i} = p_u q_i. \tag{2.7}$$

Dessa forma, a tarefa principal do modelo é estimar adequadamente os parâmetros P e Q. Na FM, o treinamento dos parâmetros ocorre por meio da aplicação do gradiente descendente estocástico no erro sobre o conjunto de treinamento $T \subset R$, que é computado sobre o Erro Quadrado Regularizado (EQR) como apresentado na Eq. 2.8.

$$EQR = \sum_{u,i \in T} (r_{u,i} - p_u^T q_i) + \lambda(\|p_u\|^2 + \|q_i\|^2), \tag{2.8}$$

em que o parâmetro λ representa o termo de regularização de Tikhonov, que é utilizado para evitar o *overfitting* do modelo. Uma das formas de calcular os parâmetros é inicializar P e Q com valores aleatórios numa escala pré-definida e atualizá-los simultaneamente para cada exemplo do treinamento de acordo com a seguinte regra:

$$(P,Q) = \underset{p,q}{\operatorname{argmin}}\, EQR \Rightarrow \begin{cases} \frac{\partial}{\partial p_u} EQR = -2(r_{u,i} - p_u q_i)q_i + 2\lambda p_u \\ \frac{\partial}{\partial q_i} EQR = -2(r_{u,i} - p_u q_i)q_u + 2\lambda q_i \end{cases} \Rightarrow \begin{cases} p_u \leftarrow p_u + \eta((r_{u,i} - p_u q_i)q_i - \lambda p_u) \\ q_i \leftarrow q_i + \eta((r_{u,i} - p_u q_i)p_u - \lambda q_i) \end{cases}, \tag{2.9}$$

em que η é a taxa de aprendizado.

Filtragem Baseada em Conteúdo (FBC)

A Filtragem baseada em Conteúdo consiste em recomendar itens a um usuário, baseando-se em informações extraídas dos itens a ele associados. Sua função de utilidade é representada conforme Eq. 2.10 (BERKOVSKY; KUFLIK; RICCI, 2008).

$$util_{BC} : U_{id} \times D_{inf} \rightarrow R_{|U|\times|D|}, \tag{2.10}$$

em que U_{id} representa os identificadores únicos dos usuários e D_{inf} representa as informações relativas aos itens. Em relação à construção do perfil de usuário, existem duas abordagens principais (ADOMAVICIUS; TUZHILIN, 2005):

- *Perfil explícito* – o usuário definirá diretamente quais são seus interesses e também o que não lhe interessa, podendo esta informação ser adquirida por meio de algum questionário;
- *Perfil implícito* – o perfil de usuário será inferido de forma implícita, baseando-se em itens que foram previamente por ele avaliados[1].

Um fator primordial na FBC é a representação do espaço de itens e do perfil de usuário (ADOMAVICIUS; TUZHILIN, 2005). Um dos modelos mais utilizados na representação é o modelo vetorial que foi herdado da área de Recuperação de Informação (RI). No modelo vetorial (SALTON; BUCKLEY, 1988), o conteúdo dos itens[2] é baseado em análise

[1] A ação do usuário avaliar um item pode ser entendida como a existência de uma relação entre usuário e item, por exemplo, o usuário pode ter criado ou interagido com o item.

[2] O termo "item" é utilizado como sinônimo do termo "documento", que é mais comum na área de RI.

textual, sendo os itens descritos pelos principais termos presentes em sua descrição (no caso de documentos não textuais) ou conteúdo (no caso de documentos textuais).

No modelo vetorial, cada documento é representado por um vetor de pesos, em que cada peso desse vetor será referente à importância de um termo retirado da análise textual. Então, dado o conjunto de termos $T = \{t_1, ..., t_{|T|}\}$, deseja-se ponderar a importância $w_{d,t}$ de cada termo $t \in T$ para o item d. Uma das formas mais conhecidas de computar esse peso é a métrica *Frequência do Termo/Frequência Inversa do Documento* (do inglês, *Term Frequency/Inverse Document Frequency* (TF-IDF)) (SALTON; BUCKLEY, 1988) que é calculado pelo produto da *Frequência do Termo* (do inglês, *Term Frequency* (TF)) pela *Frequência Inversa do Documento* (do inglês, *Inverse Document Frequency* (IDF)).

O TF-IDF é computado da seguinte forma: TF(d, t) é a *Frequência do Termo* que computa a frequência do termo t no item d, definida pela Eq. 2.11.

$$\text{TF}(d, t) = \frac{\text{q}(d, t)}{\text{q}(d, T)}, \tag{2.11}$$

em que q(d, t) é a quantidade de vezes que o termo t aparece no item d e q(d, T) representa o número total de termos presentes em d. O IDF(t, D) computa a relevância do termo t em relação ao conjunto de itens D, calculado de acordo com a Eq. 2.12.

$$\text{IDF}(D, t) = \log\left(\frac{|D|}{|D^t|}\right), \tag{2.12}$$

em que $D^t \subseteq D$ é o conjunto de itens, cujo termo t ocorre pelo menos uma vez. O peso $w_{d,t}$ é computado por TF-IDF(D, d, t) que corresponde ao produto entre o TF e o IDF, conforme a Eq. 2.13.

$$w_{d,t} = \text{TF-IDF}(D, d, t) = \text{TF}(d, t) * \text{IDF}(D, t). \tag{2.13}$$

Por fim, cada item $d \in D$ é representado por um vetor de pesos $\vec{d}$:

$$\vec{d} = (w_{d,1}, ..., w_{d,|T|}). \tag{2.14}$$

Na FBC, o perfil de usuário é obtido por meio de heurísticas que analisam os itens previamente avaliados pelo usuário e, o vetor resultante é representado por:

$$\vec{p_u} = (w_{u,1}, ..., w_{u,|T|}), \tag{2.15}$$

em que o peso $w_{u,i}$ indica a importância do termo t_i para o usuário u.

Após computados os perfis dos documentos e o perfil do usuário, a utilidade de um documento d para o usuário u é computada pela similaridade entre o perfil do documento $\vec{d}$ e o perfil do usuário $\vec{u}$, sendo geralmente utilizada a similaridade do cosseno Eq. 2.4.

2.1.2 Avaliação de Sistemas de Recomendação

As avaliações de SR são classificadas em duas categorias: avaliação *online* e avaliação *offline* (SHANI; GUNAWARDANA, 2011). Na avaliação *offline*, são utilizados dados já existentes contendo informações sobre os usuários, itens e avaliação. Essa forma de avaliação é amplamente utilizada devido a sua fácil reprodutibilidade. Na avaliação *online*, a recomendação é computada e o resultado é apresentado ao usuário, sendo que esse tipo de avaliação é mais fidedigno à preferência atual do usuário, entretanto, não é de fácil reprodução, ou seja, não é facilmente replicável.

Métricas de Avaliação Utilizadas

A métrica utilizada neste trabalho foi o Desconto Normalizado do Ganho Cumulativo (do inglês, *Normalized Discounted Cummulative Gain* – NDCG). O NDCG é utilizado para avaliar a relevância de um conjunto de itens recomendados, levando-se em conta a ordem na qual eles aparecem e atribuindo mais peso aos itens que estão no topo da lista de recomendação. O NDCG é calculado pela normalização do Desconto do Ganho Cumulativo (*Discounted Cummulative Gain* – DCG), conforme a Eq. 2.16 (JÄRVELIN; KEKÄLÄINEN, 2002).

$$NDCG = \frac{DCG - DCG_{min}}{DCG_{max} - DCG_{min}}, \tag{2.16}$$

em que DCG é definido por:

$$DCG = \sum_{j=1}^{n} \frac{2^{r_j} - 1}{\log_2(1 + j)}, \tag{2.17}$$

em que n é a quantidade de itens recomendados, r_j é a relevância atribuída (pode ser utilizada, por exemplo, uma escala de 0 a 4, em que 4 é o mais relevante) a um item na j-ésima posição na lista de recomendação, e DCG_{max} é o valor máximo de DCG (todos os itens com a maior avaliação possível) e DCG_{min} é o valor mínimo do DCG (todos os itens com a menor avaliação possível).

No trabalho, como será recomendada uma quantidade n de itens a um usuário, será utilizada a representação NDCG@n, que indica a relevância dos n itens recomendados, e.g. NDCG@5 indica o NDCG para os 5 primeiros itens da lista de recomendação.

2.2 Discussão

Neste Capítulo foram apresentados os principais conceitos e modelos que serviram de base ao presente trabalho. Nesse sentido foi explanada a tecnologia de Sistemas de Recomendação, cujo objetivo consiste em prever qual é a utilidade de um determinado item para um determinado usuário. Logo após, seus principais modelos foram abordados: Filtragem Colaborativa, em que a recomendação para um determinado usuário é baseada nas avaliações de usuários com preferências similares as dele e; Filtragem Baseada em Conteúdo, em que a recomendação é baseada no conteúdo extraído dos itens associados ao usuário. Por fim, foram apresentados aspectos relativos à avaliação de SR, bem como as métricas que serão utilizadas na avaliação da proposta.

No próximo Capítulo desta dissertação serão discutidos os trabalhos relacionados à mesma.

Capítulo 3

Trabalhos Relacionados

Neste capítulo, são apresentados os trabalhos relacionados ao ora apresentado. O capítulo está dividido em duas partes: (i) trabalhos que abordam a integração de perfis de usuário; e (ii) trabalhos sobre SR de artigos científicos.

3.1 Integração de Perfis de Usuário

No âmbito de SR, a tarefa de integrar perfis de usuário de forma a obter um perfil único no intuito de prover recomendações melhores é chamada de *Mediação de Perfis de Usuário* (do inglês *User Model Mediation*) (BERKOVSKY; KUFLIK; RICCI, 2008). Em outras palavras, esta estratégia consiste em complementar um perfil de usuário por meio da integração de perfis utilizando sua informação contida noutros sistemas. Berkovsky, Kuflik e Ricci (2008) identifica os principais desafios envolvidos na tarefa de mediação de perfis:

1. **Limitação de negócio** – esta tarefa diz respeito à competição existente entre sistemas comerciais, o que dificulta a cooperação entre eles, impossibilitando o compartilhamento de informações sobre seus usuários;

2. **Privacidade do usuário** – em sua maioria, os sistemas possuem termos de compromisso com seus usuários e estabelecem políticas de privacidade sobre seus dados, o que não permite que informação do usuário seja disponibilizada para outros sistemas.

3. **Considerações técnicas** – este desafio é referente às questões técnicas que envolvem a mediação de perfis de usuário, principalmente na tarefa de prover comunicação entre

os diferentes sistemas, levando-se em conta as limitações de recursos.

4. **Heterogeneidade de dados** – esta tarefa refere-se à heterogeneidade da natureza dos dados existentes em SR. De forma geral, podem-se destacar dois aspectos ligados à heterogeneidade: (i) *heterogeneidade de domínio*, os SR são construídos e orientados para domínios específicos, de tal forma que nem sempre é possível adequar determinado SR para outros contextos ou utilizar informações de outro sistema e (ii) *heterogeneidade de perfil de usuário*, dependendo da técnica de recomendação utilizada, as preferências de determinado usuário serão armazenadas de diferentes formas, além do que as informações provindas de outros sistemas podem ser conflitantes ou ultrapassadas.

Este trabalho tem como foco o quarto desafio apresentado, heterogeneidade de dados. A seguir são apresentados alguns trabalhos relevantes que tratam desta questão.

O trabalho de Sahebi et al. (SAHEBI; WONGCHOKPRASITTI; BRUSILOVSKY, 2010) faz a mediação de perfis de usuário no intuito de recomendar colóquios acadêmicos no CoMet. O CoMet é um sistema colaborativo construído com o objetivo de compartilhar informação sobre colóquios acadêmicos de pesquisadores de duas universidades de Pittsburgh: Carnegie Mellon University[1] e University of Pittsburgh[2]. Nesse trabalho, são utilizadas três fontes de dados: (i) *tags* do usuário no CoMet, um sistema colaborativo para compartilhar informação sobre colóquios acadêmicos; (ii) os artigos do CiteULike[3] (abstract e título); e (iii) as *tags* do CiteULike.

3.2 SR de Artigos Científicos

A literatura de SR de artigos científicos pode ser classificada de acordo com a forma como a recomendação é realizada. Basicamente, existem duas abordagens principais (JIANG et al., 2012):

- *artigo-artigo* – a recomendação é baseada na similaridade entre artigos analisando-se as citações de um dado artigo ou de um conjunto de artigos (e.g., (MCNEE et al.,

[1] <http://www.cmu.edu/>
[2] <http://www.pitt.edu/>
[3] <http:///www.citeulike.org>

2002; TORRES et al., 2004; GORI; PUCCI, 2006; NASCIMENTO et al., 2011)). Essa abordagem tem a característica de não ser personalizada, ou seja, a recomendação não será diferente de usuário para usuário.

- *usuário-artigo* – o objetivo é recomendar artigos com base nas preferências do usuário, por meio de uma análise do conteúdo por ele consumido[4] (e.g. (LOPES; SOUTO; WIVES, 2007; SUGIYAMA; KAN, 2010; GOOSSEN et al., 2011; MAGALHÃES et al., 2012)). Diferentemente da abordagem *artigo-artigo*, essa abordagem proporciona recomendação de forma personalizada.

Desta forma, a literatura referente a este assunto será apresentada de acordo com estas duas principais abordagens.

3.2.1 Abordagem *artigo-artigo*

A abordagem artigo-artigo trata de trabalhos de recomendação de artigos que não são focados em perfil de usuário, ou seja, não são algoritmos personalizados. No geral, esses algoritmos criam modelos de recomendação utilizando as citações dos artigos, o objetivo pode ser enunciado da seguinte forma: dado um artigo ou um conjunto de artigos como gerar citações relevantes. Uma desvantagem desse método é que se um artigo relevante não possuir citações, seja por diversos motivos (e.g. um artigo muito recente na área), ele dificilmente será recomendado.

Nesta linha, o trabalho de McNee et al. (2002) propõe uma abordagem para recomendar citações de artigos científicos por meio de Filtragem Colaborativa, sua proposta consiste em construir uma matriz binária de avaliações, em que as linhas são os artigos e as colunas são suas citações. Desta forma, essa informação pode ser mapeada no framework dos algoritmos de Filtragem Colaborativa. Seguindo esta mesma abordagem, o trabalho de Gori e Pucci (2006) cria um grafo de citação e faz uma adaptação do algoritmo PageRank (PAGE et al., 1998). No entanto, os autores não fazem uma otimização do fator de decaimento do algoritmo. O trabalho de Huang et al. (2004) apresenta um algoritmo de aprendizagem por reforço para recomendar artigos relacionados a um dado artigo. Esta abordagem utiliza

[4]O termo "consumir" é utilizado para designar o ato do usuário criar um conteúdo ou utilizá-lo de alguma forma.

informações de citação, co-citação de artigos em diferentes contextos de citação (citação do mesmo problema ou citação do mesmo método utilizado). No entanto, os autores não fazem uso de informação de conteúdo como abstract e títulos dos artigos. Diferentemente do trabalho de Nascimento et al. (2011) em que dado um artigo, o sistema o analisa de quatro formas diferentes: (i) apenas o título, (ii) apenas o abstract, (iii) apenas o corpo e; (iv) todos os campos do artigo.

Estes trabalhos supracitados possuem duas limitações principais: (i) se um artigo relevante não possuir citações, ele dificilmente será recomendado; e (ii) alguns artigos citados não são diretamente relacionados ao artigo em questão, muitas vezes servindo como suporte a algum método utilizado. Uma possível solução para este problema consiste em utilizar a informação de conteúdo dos artigos, como por exemplo: título, abstract e veículo de publicação. Com este intuito, o trabalho de Torres et al. (2004) apresenta abordagens híbridas incorporando informação de conteúdo (título e abstract) no modelo de Filtragem Colaborativa. No entanto, as abordagens híbridas não obtiveram resultados satisfatórios em relação às abordagens puras. Os autores argumentam que os algoritmos de recomendação puros não são desenvolvidos no intuito de receber entrada de outro SR.

Com relação à análise da utilidade da recomendação, em (MCNEE; KAPOOR; KONSTAN, 2006) os autores analisam diversos algoritmos de recomendação. Por meio de um experimento online, eles avaliam os artigos recomendados em diferentes perspectivas, como por exemplo: familiaridade do usuário, personalização e relevância da recomendação. No entanto, é necessária mais pesquisa no intuito de definir como atualizar o perfil de usuário de acordo com esse *feedback*.

3.2.2 Abordagem *usuário-artigo*

A abordagem usuário-artigo consiste em, a partir de um dado conjunto de artigos relacionados a um indivíduo, criar um perfil de usuário e recomendar artigos similares a este perfil. Uma estratégia comum dessa abordagem consiste em compor o perfil de usuário utilizando sua produção bibliográfica, e.g. (LOPES; SOUTO; WIVES, 2007; SUGIYAMA; KAN, 2010). O trabalho de Lopes, Souto e Wives (2007) constrói o perfil de usuário utilizando informação da produção bibliográfica capturada de seu currículo Lattes. O perfil de usuário é representado por termos, em que o peso de cada termo depende de três fatores: (i) se o

termo faz parte de palavra-chave ou do título; (ii) da proficiência linguística do usuário e (iii) do ano em que o termo foi observado, ou seja, é feita uma análise temporal do item no perfil do usuário. O perfil de usuário $\vec{p_u}$ é representado pelo Modelo do Espaço Vetorial (BAEZA-YATES; RIBEIRO-NETO, 2011), $\vec{p_u} = (w_1, w_2, ..., w_{|K|})$, em que $w_k \in [0, 1]$ representa a importância do termo k para o usuário u. O peso w_k é obtido pela Eq. 3.1.

$$w_k = w_{keyword\_or\_title} * w_{language} * w_{year}, \tag{3.1}$$

em que $w_{keyword\_or\_title}$ leva em consideração o tipo do termo (se é obtido de uma "palavra-chave" ou "título"), $w_{language}$ que considera o idioma da publicação que o termo foi utilizado e w_{year} que pondera sobre o ano em que as produções foram publicadas. Os pesos $w_{keyword\_or\_title}$, $w_{language}$ e w_{year} são calculados conforme a Eq. 3.2.

$$w_j = 1 - (j - 1)\left(\frac{1 - w_{min}}{v - 1}\right), \tag{3.2}$$

em que os parâmetros usados em w_j variam de acordo com o tipo do peso. No caso da variável $w_{keyword\_or\_title}$, são utilizados $w_{min} = 0.95$, e $i = 1$ para termos presentes nas palavras-chave e $i = 2$ para termos presentes nos títulos do trabalho. Para $w_{language}$, $w_{min} = 0.60$ e $i = 1$ se a proficiência na linguagem é "bem", $i = 2$ para "razoavelmente" e $i = 3$ para "pouco". Para a variável w_{year}, são utilizados $w_{min} = 0,95$ e i varia de 1 até n, em que n é o intervalo de anos considerado, sendo 1 para o maior e n para o menor, Lopes, Souto e Wives (2007) utilizam o intervalo entre 2003 e 2006.

No entanto, o trabalho de Lopes, Souto e Wives (2007) possui algumas limitações: (i) não faz uso de informação valiosa que está presente no Lattes (por exemplo, projetos, currículo resumido, etc.), (ii) não existe uma otimização dos pesos dos termos no perfil do usuário (iii) não faz utilização de conceitos; e (iv) não existe solução para o problema do usuário novo.

O trabalho de Sugiyama e Kan (2010) também utiliza a publicação passada do usuário para compor um perfil de usuário. No entanto, eles analisam também os artigos citados e os artigos referenciados. Por meio da utilização dessa informação extra, é possível obter um perfil de usuário mais robusto e prover recomendação para os novos usuários, isto é, se pelo menos o usuário possuir uma publicação no seu perfil.

3.2.3 Análise Comparativa entre os Trabalhos

Os trabalhos apresentados nas Subseções 3.2.1 e 3.2.2 são comparados conforme as características definidas a seguir:

- **Abordagem (Abord)** – Se a abordagem de recomendação utilizada é do tipo artigo-artigo (*a-a*) ou usuário-artigo (*u-a*);
- **Personalizado (Pers)** – Se a recomendação é personalizada, isto é, se pode ser diferente de usuário para usuário;
- **Temporal (Temp)** – Se existe algum decaimento temporal na construção de perfil de usuário;
- **Novo** – Se o modelo oferece alguma solução para o problema do novo usuário;
- **Representação (Repres)** – Se o perfil de usuário é representado por termos (*termos*) ou por conceitos (*conceitos*), caso o trabalho não analise o conteúdo do item, e verifique apenas o identificador do artigo, essa característica é marcada pelo valor *id*;
- **Validação (Valid)** – Se o trabalho apresenta uma validação online, offline, ou ambas validações (*online e offline*).

Na Tabela 3.1 é apresentada a comparação entre os trabalhos relacionados.

Tabela 3.1: Comparação entre os trabalhos relacionados.

Trabalho	Abord	Pers	Temp	Novo	Repres	Valid
Lopes, Souto e Wives (2007)	u-a	sim	sim	não	termos	online
Sugiyama e Kan (2010)	u-a	sim	sim	sim	termos	offline
Gori e Pucci (2006)	a-a	não	não	não	id	offline
Huang et al. (2004)	a-a	não	não	não	termos e id	offline
Nascimento et al. (2011)	a-a	não	não	não	termos	online
McNee et al. (2002)	a-a	não	não	não	termos e id	online e offline
Torres et al. (2004)	a-a	não	não	não	termos e id	online e offline
McNee, Kapoor e Konstan (2006)	a-a	não	não	não	termos e id	online

3.3 Sobre o Diferencial deste Trabalho

O presente trabalho possui diferencial em relação aos demais trabalhos nos seguintes aspectos: (i) com relação à recomendação baseada no currículo Lattes, são analisados mais aspectos do currículo, como projetos, produção técnica, etc., é demonstrado que essa estratégia proporciona resultados melhores do que o estado da arte da recomendação baseada no Lattes (trabalho de Lopes, Souto e Wives (2007)); (ii) em se tratando da integração de perfis de usuário, é realizada uma mensuração da importância de cada fonte de dados para cada usuário, o que inclusive foi um trabalho futuro proposto por Berkovsky, Kuflik e Ricci (2008); (iii) é realizada uma análise do impacto da utilização de conceitos na representação de perfis de usuário e indexação de artigos; e (iv) com relação ao SR, adota-se uma estratégia diferente dos demais, é proposto um chaveador de métodos de recomendação, ou seja, é escolhido o método mais adequado de acordo com as características do usuário.

3.4 Discussão

No presente Capítulo foram abordados os trabalhos relacionados a esta dissertação. Primeiramente, foram apresentados trabalhos que abordam a integração de perfis de usuário, em que foram discutidos os principais desafios existentes nesta área de pesquisa. Logo após, foram explanados e comparados os trabalhos que propõem SR de artigos científicos. Os trabalhos sobre este tema foram divididos de acordo com a abordagem utilizada: (i) *artigo-artigo* – trabalhos que recomendam baseando-se na similaridade entre artigos, por meio da análise das citações de um dado artigo ou de um conjunto de artigos; e (ii) *usuário-artigo* – trabalhos que recomendam artigos com base nas preferências do usuário, por meio de uma análise do conteúdo por ele consumido. Por fim, foram apresentados alguns pontos que ressaltam o diferencial deste trabalho.

No próximo Capítulo, a proposta deste trabalho será apresentada, que trata de um modelo para integração de perfis de usuário.

Capítulo 4

Modelo para Integração de Perfis de Usuário

Neste capítulo, é apresentada a proposta de modelo para integração de perfis de usuário. O capítulo está dividido em cinco partes, a saber: (i) apresentação das notações utilizadas e definição do conhecimento de domínio; (ii) apresentação da forma como o perfil do usuário é construído para uma dada fonte de dados (Seção 4.2); (iii) explanação de como ocorre a integração de perfis de usuário em múltiplas fontes (Seção 4.3); (iv) descrição da forma como os itens são indexados (Seção 4.4); e por fim (v) como é realizada a recomendação de itens (Seção 4.5).

4.1 Definições Preliminares

O conjunto $S = \{s_1, ..., s_{|S|}\}$ representa o conjunto de fontes de dados[1] e $U = \{u_1, ..., u_{|U|}\}$ o conjunto de usuários, em que uma fonte $s \in S$ é caracterizada como um lugar que permite que os seus usuários criem e/ou consumam itens[2], por exemplo: blogs, páginas pessoais, currículos online.

O conjunto $D = \{d_1, ..., d_{|D|}\}$ representa os itens disponíveis aos usuários $u \in U$. Cada item $d \in D$ possui uma descrição textual representada por m_d. O conjunto $D_{u,s} \subseteq D$ representa os itens d que o usuário u consumiu na fonte s. Além disto, o rótulo $y_{u,s,d}$ representa

[1] Doravante o documento, por motivos de simplicidade, será utilizado apenas o termo *fonte* para designar uma *fonte de dados*.

[2] No presente trabalho *item, conteúdo* e *documento* são considerados sinônimos.

um rótulo temporal, que indica o ano em que o documento d foi consumido pelo usuário u na fonte s. Neste modelo, os itens podem ser: vídeos, músicas, artigos, postagens, etc., desde que possuam um texto que os representem.

O conjunto $T = \{t_1, ..., t_{|T|}\}$ denota um conjunto de termos e o conjunto $C = \{c_1, ..., c_{|C|}\}$ denota um conjunto de conceitos; os termos e os conceitos são utilizados para indexar e representar os documentos e perfis de usuário. Cada documento d é representado por dois vetores: (i) utiliza termos $\vec{dt} = (w_{d,1}, ..., w_{d,|T|})$, em que o peso $w_{d,t} \in [0, 1]$ representa a importância do termo t para o documento d; e (ii) utiliza conceitos $\vec{dc} = (w_{d,1}, ..., w_{d,|C|})$, em que o peso $w_{d,c} \in [0, 1]$ representa a importância do conceito c para o documento d.

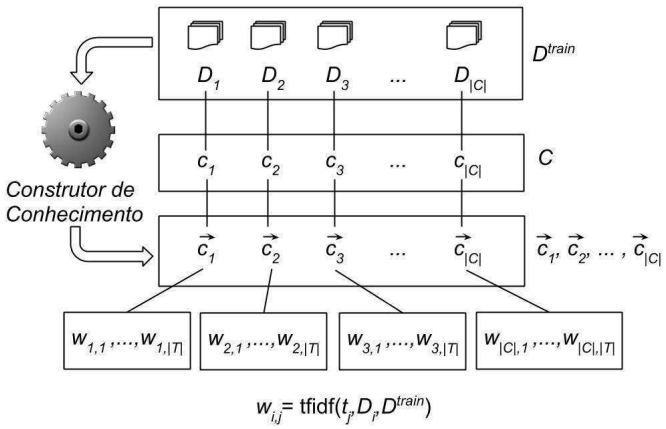
Assim como os documentos, os usuários também são representados por termos e conceitos, de tal forma que o usuário u na fonte s possui dois perfis: (i) um perfil no nível de termos denotado pelo vetor $\vec{pt}_{u,s} = (w_{u,s,1}, ..., w_{u,s,|T|})$, em que $w_{u,s,t}$ representa a importância do termo t no perfil do usuário u na fonte s; e (ii) um perfil no nível de conceitos denotado por $\vec{pc}_{u,s} = (w_{u,s,1}, ..., w_{u,s,|C|})$, em que $w_{u,s,c}$ representa o peso do conceito c para o usuário u na fonte s. Os perfis de termos do usuário nas diferentes fontes $\vec{pt}_{u,1}, ..., \vec{pt}_{u,|S|}$ são integrados, resultando num único perfil de termos denotado por $\vec{pt}_u$, da mesma forma ocorre com os perfis de conceitos $\vec{pc}_{u,1}, ..., \vec{pc}_{u,|S|}$ que são integrados num único perfil representado por $\vec{pc}_u$.

4.1.1 Determinando o Conhecimento de Domínio

O conhecimento de domínio tem a função de auxiliar na construção dos perfis de conceitos de usuário e na indexação dos artigos a serem recomendados. É utilizada uma abordagem baseada em ontologia para representar o domínio, nesse sentido, foi seguida a abordagem proposta por Loh et al. (2006). O conjunto $C = \{c_1, ..., c_{|C|}\}$ representa os conceitos associados ao domínio, em que cada conceito $c \in C$ é um nó na ontologia O. O conjunto $T = \{t_1, ..., t_{|T|}\}$ denota os termos associados ao domínio.

Na Figura 4.1 é apresentado o processo de construção do conhecimento de domínio. Observando a Figura 4.1, verifica-se que cada conceito $c \in C$ é representado por um vetor de pesos, $\vec{c} = (w_{c,1}, ..., w_{c,|T|})$, em que o peso $w_{c,t} \in \vec{c}$ representa a importância do termo t para o conceito c. Os pesos w de cada vetor $\vec{c}$ são calculados estatisticamente utilizando um

Figura 4.1: Processo de construção do conhecimento.



conjunto de treinamento de documentos $D^{train} \subseteq D$. O conjunto de treinamento é denotado por $D^{train} = \{D_1, ..., D_{|C|}\}$, em que $D_c \subset D^{train}$ representa o conjunto de treinamento para o conceito $c \in C$. O peso $w_{c,t}$ é computado conforme a Eq.4.1.

$$\text{TF-IDF}(t, D_c, D^{train}), \tag{4.1}$$

em que o TF-IDF é obtido conforme a Eq. 2.13.

4.2 A Construção dos Perfis de Usuário numa Fonte

O perfil de termos $\vec{pt}_{u,s}$ do usuário u na fonte s é obtido a partir do conjunto de itens $D_{u,s}$ que o usuário u consumiu na fonte s. O peso $w_{u,s,t} \in \vec{pt}_{u,s}$ é definido pela Eq. 4.2.

$$w_{u,s,t} = \sum_{d \in D_{u,s}} \text{TF}(t, d) * temp(y_{u,s,d}), \tag{4.2}$$

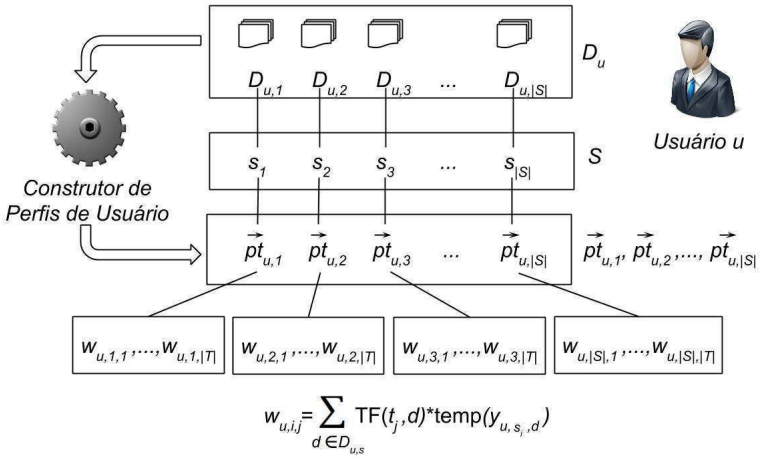
em que a função $temp(y_{u,s,d})$ faz uma calibração do peso de acordo com o tempo que o item d foi consumido, de tal modo que o objetivo é que os itens mais novos possuam maior importância para o perfil do usuário do que os itens mais antigos. A função $temp(y_{u,s,d})$ é definida conforme a Eq. 4.3.

$$temp(y_{u,s,d}) = -\frac{\Delta y}{v} + 1, \tag{4.3}$$

em que $v \in \mathbb{N}^*$ é o intervalo de anos em que um item é considerado no perfil do usuário e Δy é o intervalo entre o ano atual y_{now} e o ano em que o item foi consumido $y_{u,s,d}$.

Na Figura 4.2 é apresentado o processo de construção dos perfis de termos do usuário. Nota-se que o usuário u possui $|S|$ perfis $\vec{pt}_{u,s}$, um para cada fonte de dados.

Figura 4.2: Processo de construção do perfil de termos do usuário u.

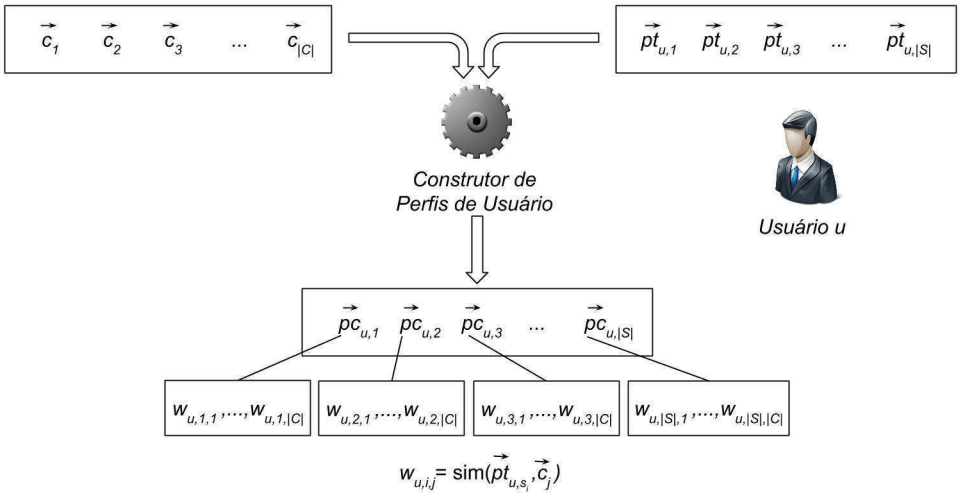


O processo de construção do perfil de conceitos é apresentado na Figura 4.3, em que o perfil de conceitos $\vec{pc}_{u,s}$ do usuário u na fonte s é definido combinando-se o perfil $\vec{pt}_{u,s}$ com vetores de conceitos. O peso $w_{u,s,c} \in \vec{pc}_{u,s}$ é definido pela Eq. 4.4.

$$w_{u,s,c} = sim(\vec{pt}_{u,s}, \vec{c}), \tag{4.4}$$

em que a função sim calcula a similaridade do cosseno, conforme Eq. 2.4.

Figura 4.3: Processo de construção do perfil de conceitos do usuário u.



4.3 Integração de Perfis de Usuário

A integração de perfis de usuário, tanto para o perfil integrado de termos $\vec{pt}_u$ quanto para o perfil integrado de conceitos $\vec{pc}_u$, é realizada por meio de combinação linear (MAGALHÃES et al., 2012), conforme as Eq. 4.5 e 4.6.

$$\vec{pt}_u = \sum_{s \in S} \vec{pt}_{u,s} \cdot a_{u,s}, \tag{4.5}$$

$$\vec{pc}_u = \sum_{s \in S} \vec{pc}_{u,s} \cdot a_{u,s}. \tag{4.6}$$

O peso $a_{u,s}$ nas Eq. 4.5 e 4.6 representa a importância da fonte s para o usuário u. Quanto maior o valor de $a_{u,s}$, mais importante a fonte s será para construção dos perfis integrados $\vec{pt}_u$ e $\vec{pc}_u$, em que os pesos são normalizados de forma que $\sum_{s \in S} a_{u,s} = 1$. A seguir são apresentadas diferentes estratégias utilizadas para contabilizar este peso.

- **Fontes com Importância Igual (Igual)** – Este método consiste em atribuir importância igual a todas as fontes, sendo considerado como *baseline* justamente por não fazer distinção entre as fontes. O peso é atribuído de acordo com a quantidade de fontes, conforme a Eq. 4.7.

$$a_{u,s}^{igual} = \frac{1}{|S|}. \tag{4.7}$$

- **Importância de Acordo com a Quantidade de Itens (Quant)** – Neste esquema, quanto mais itens o usuário tiver consumido numa fonte, mais importante esta será na integração de perfis. O peso é definido conforme Eq. 4.8.

$$a_{u,s}^{quant} = \frac{|D_{u,s}|}{|D_u|}. \tag{4.8}$$

- **Utilizando a Atividade do Usuário na Fonte (Ativ)** – Nesta estratégia, a importância de uma fonte para um usuário dependerá de sua atividade na referida fonte. Dado que cada item possui uma marca temporal $y_{u,s,d}$, que indica o ano em que o documento d foi consumido pelo usuário u na fonte s, a atividade pode ser calculada utilizando a média da latência entre as marcas temporais, conforme a Eq. 4.9 (SOUZA; MAGALHÃES;

COSTA, 2011).

$$a_{u,s}^{ativ} = \frac{y_{now} - y_{u,s,|D_{u,s}|} + \sum_{j=1}^{|D_{u,s}|-1}(y_{u,s,j+1} - y_{u,s,j})}{|D_{u,s}| + 1}, \tag{4.9}$$

em que y_{now} é o tempo presente. Na Eq. 4.9, quanto menor for o valor de $a_{u,s}^{ativ}$, mais ativo o usuário será. Por isto, os valores são normalizados, de tal forma que quanto maior for o valor de a_u^s, mais ativo o usuário será na fonte de dados s.

4.4 Indexando Itens

Após definidos os perfis de usuário, é necessário definir como os itens disponíveis para recomendação são indexados. Assim como os perfis, os itens são indexados utilizando termos e conceitos. Então, o primeiro passo é definir o conjunto de itens disponíveis para recomendação D^{rec} que será um subconjunto do conjunto dos itens, logo $D^{rec} \subset D$. Cada item $d \in D^{rec}$ é representado e indexado por dois vetores:

- Vetor de termos $\vec{dt} = (w_{d,1}, ..., w_{d,|T|})$, cujo peso $w_{d,t} \in \vec{dt}$ é dado pelo esquema TF-IDF, logo:

$$w_{d,t} = \text{TF-IDF}(t, d, D^{rec}), \tag{4.10}$$

  em que o peso $w_{d,t} \in [0, 1]$ representa a importância do termo t para o item d e o TF-IDF é obtido pela Eq. 2.13;

- Vetor de conceitos $\vec{dc} = (w_{d,1}, ..., w_{d,|C|})$, cujo peso $w_{d,c} \in \vec{dc}$ é obtido pela similaridade entre o perfil de termos do item e o vetor do conceito c, portanto:

$$w_{d,c} = sim(\vec{dt}, \vec{c}), \tag{4.11}$$

  em que o peso $w_{d,c} \in [0, 1]$ representa a importância do conceito c para o item d e a função sim calcula a similaridade do cosseno Eq. 2.4.

4.5 Recomendando Itens

A recomendação para o usuário u é feita analisando-se o conjunto $D^{rec} - D_u$, que representa os itens disponíveis para recomendação retirando-se os itens que já foram consumidos pelo usuário u. A recomendação é gerada de acordo com a Eq. 4.12.

$$D_u^{rec} = \underset{d \in D^{rec} - D_u}{\overset{n}{\operatorname{argmax}}} \; util(u, d), \tag{4.12}$$

em que a função argmax retorna o conjunto dos n itens d mais relevantes em relação ao usuário u. A função $util(u, d)$ retorna a informação do quão relevante o documento d é para o usuário u.

São definidas duas abordagens para computar esta relevância, quais sejam:

- Similaridade entre perfis de termos, computada de acordo com a Eq. 4.13:

$$util(u, d) = sim(\vec{pt}_u, \vec{dt}). \tag{4.13}$$

- Similaridade entre perfis de conceitos, computada de acordo com a Eq. 4.14:

$$util(u, d) = sim(\vec{pc}_u, \vec{dc}). \tag{4.14}$$

Nas Eq. 4.13 e 4.14, a função de similaridade sim é computada pelo cosseno entre os dois vetores, conforme Eq. 2.4, apresentada no Capítulo 2 desta dissertação.

4.6 Discussão

Neste Capítulo foi apresentada a proposta deste trabalho que consiste num modelo de integração de perfis de usuário. O capítulo foi dividido em cinco partes, primeiramente, foram apresentadas as notações utilizadas na representação dos itens e perfis de usuários, utilizando-se dois métodos de representação: por termos e por conceitos. Daí foi apresentado como o conhecimento de domínio é definido. Logo após, foi explanado como o perfil de usuário é construído para uma dada fonte de dados, de forma a possuir um perfil para cada fonte. Por conseguinte, foi apresentado como ocorre a integração de perfis de usuário,

em que os perfis são integrados por meio de combinação linear. Nesse sentido foram apresentadas três estratégias para computar os pesos dos perfis em cada fonte: (i) fontes com importância igual; (ii) importância de acordo com a quantidade de itens; e (iii) utilizando a atividade do usuário na fonte. O Capítulo prosseguiu com a descrição da forma como os itens são indexados e finalizou apresentando como é realizada a recomendação.

O próximo Capítulo abordará os detalhes do sistema implementado com vistas à realizar a validação da presente proposta.

Capítulo 5

Sistema que Instancia o Modelo de Integração

Neste capítulo, são apresentadas as questões de implementação dos modelos de perfil de usuário e recomendação de artigos propostos na presente dissertação (Capítulo 4). O capítulo está dividido em duas partes: (i) inicialmente, será apresentada uma visão geral do sistema, apresentando suas telas e a forma de interação do usuário com o sistema; e por fim (ii) será apresentada a arquitetura do sistema e os detalhes de seus módulos.

5.1 Visão Geral do Sistema

O presente trabalho diz respeito a um SR de artigos científicos com foco na área de Ciência da Computação[1], tendo sido desenvolvido na linguagem Python[2], utilizando o framework Web Django[3]. O objetivo principal consistiu em criar uma base de dados no intuito de avaliar os modelos propostos, de forma que essa base possa ser utilizada por futuros trabalhos relacionados ao presente.

O sistema permitiu que os usuários voluntários do experimento se cadastrassem, que eles permitissem acesso às fontes de dados nas quais estivessem presentes e dessem retorno sobre suas áreas de interesse e sobre a relevância de um conjunto de artigos. Nesse experimento foram utilizadas três fontes de dados, que foram escolhidas de acordo com sua popularidade

[1] Acessível em <http://recsalt.com/>
[2] <http://www.python.org/>
[3] <https://www.djangoproject.com/>

no meio acadêmico e pela facilidade de acessibilidade aos seus dados.

A seguir são apresentados detalhes sobre as fontes escolhidas:

- CV Lattes[4] – foi criado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)[5] no intuito de padronizar a informação sobre a comunidade científica Brasileira, essa informação é utilizada por agências governamentais do país para avaliar pesquisadores, projetos, programas de pós-graduação, entre outros. De acordo com o site PainelLattes (LATTES, 2013), no dia 30 de Junho de 2013, o CV Lattes alcançou a marca de 2.601.696 currículos cadastrados, os quais englobam pesquisadores, estudantes e profissionais de diversas áreas do conhecimento. Dentre esse conjunto de currículos, 1.009.318 são de estudantes.

- Mendeley[6] – é um gerenciador de referências e uma rede social acadêmica que é utilizado por pesquisadores para organizar sua pesquisa, colaborar com outros pesquisadores e na descoberta de novas publicações. Em questão de números, em 10 de Outubro de 2013, o site do Mendeley exibia os seguintes dados: 2.694.756 usuários, 263.625 grupos de pesquisa, além de 498.509.731 documentos de usuários (MENDELEY, 2013).

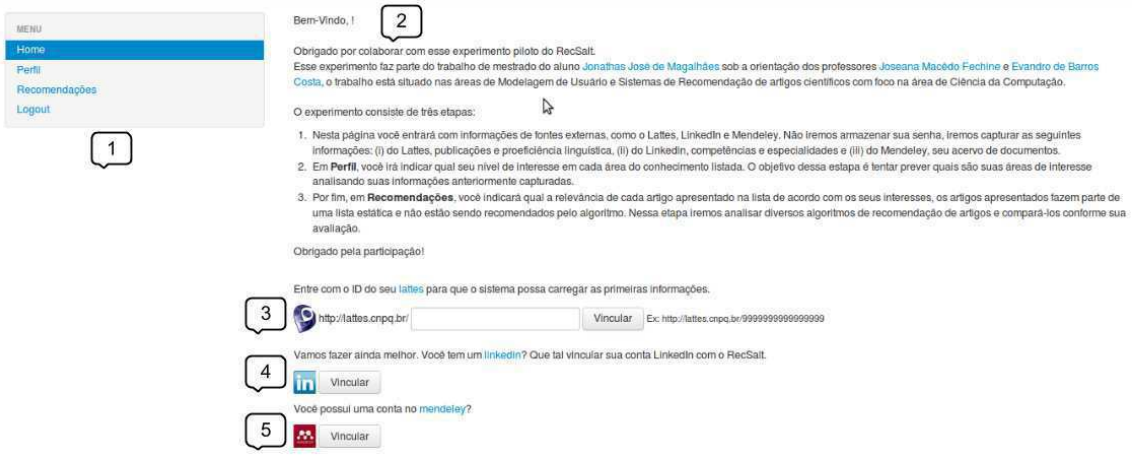
- LinkedIn[7] – é uma rede social profissional, cujo usuários podem construir um currículo voltado para empresas, além de poder se conectar com outros profissionais, ter acesso à ofertas de emprego e notícias profissionais, entre outros. De acordo com seu site, atualmente, o LinkedIn contém cerca de 225 milhões de usuários em mais de 200 países e territórios (LINKEDIN, 2013).

A seguir, são apresentados os passos da participação do usuário no experimento. Na Figura 5.1 é apresentada a tela inicial do sistema o usuário pode ler as instruções do experimento[8] e permitir acesso às fontes de dados em que ele está presente. Pela numeração, tem-se: 1) menu do sistema; 2) instruções para participar do experimento; 3) campo para o usuário inserir seu identificador do Lattes; 4) botão de vinculação do LinkedIn e 5) botão de

[4] <http://lattes.cnpq.br/>
[5] <http://www.cnpq.br/>
[6] <http://www.mendeley.com/>
[7] <https://www.linkedin.com/>
[8] As informações também ficaram disponíveis em <http://goo.gl/SwxATs>.

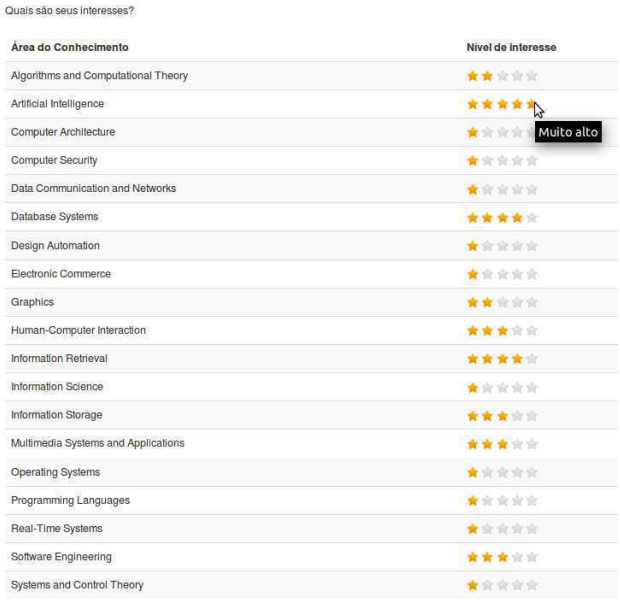
vinculação do Mendeley. Ao clicar no botão de vinculação do LinkedIn ou do Mendeley, o usuário foi direcionado para uma tela de login da respectiva fonte de dados no intuito de permitir que o sistema acesse seus dados.

Figura 5.1: Tela inicial do sistema.



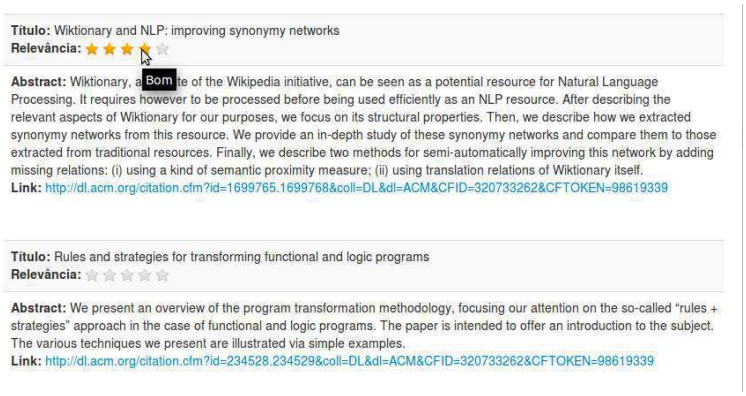
Por conseguinte, o usuário foi instruído a clicar na aba *Perfil* para ser direcionado para a tela de atribuição de perfil apresentada na Figura 5.2. A partir desta página, o usuário deveria informar seu nível de interesse, pela quantidade de estrelas, em cada área do conhecimento do domínio de Ciência da Computação. O mapeamento do nível de interesse e a quantidade de estrelas está apresentado na Tabela 5.1.

Figura 5.2: Tela de exibição das áreas de conhecimento em Ciência da Computação, em que o usuário pode especificar seus interesses em cada área.



Por fim, para completar sua participação, na aba *Recomendações*, o usuário analisaria uma lista com 50 artigos, marcando pela quantidade de estrelas a relevância para ele de cada artigo. Na Figura 5.3 é apresentada uma parte da tela de marcação de relevância nos artigos. Os artigos foram exibidos seguindo os seguintes passos: (i) foram criados 10 grupos rotulados de 0 a 9 com 5 artigos cada; (ii) o primeiro grupo de artigos a ser exibido para um usuário foi o que estava rotulado com o último número de seu identificador; e (iii) os outros grupos de artigos foram exibidos numa sequência circular na ordem do rótulo. Por exemplo, para o usuário com identificador 14, os grupos de artigos foram exibidos na seguinte ordem: 4-5-6-7-8-9-0-1-2-3. Este procedimento foi adotado para minimizar o efeito da ordem de exibição dos artigos nos resultados.

Figura 5.3: Parte da tela de exibição dos artigos, em que os usuários podem inserir a relevância nos mesmos.



O nível de interesse do usuário em relação às áreas e o nível de relevância em relação aos artigos foram identificados pela quantidade de estrelas marcadas, estando o significado descrito na Tabela 5.1.

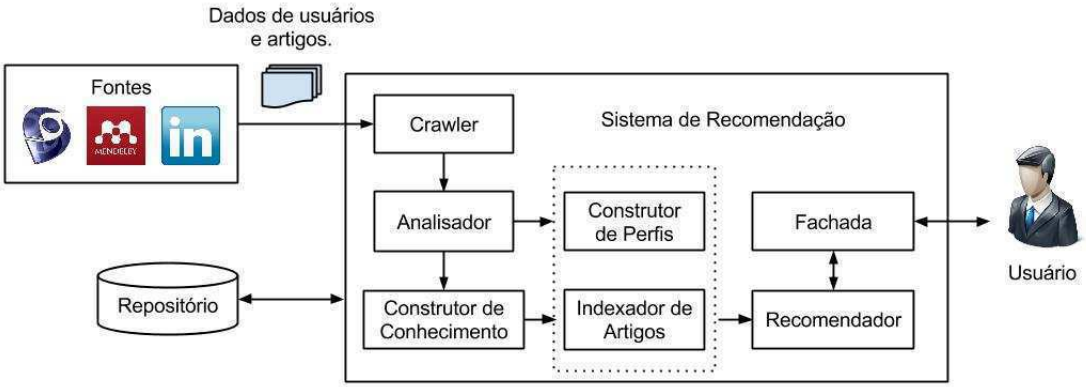
Tabela 5.1: Os níveis de interesse e relevância de acordo com a marcação de estrelas.

Número de Estrelas	Nível de Interesse (áreas)	Nível de Relevância (artigos)
1	Não há	Inadequado
2	Baixo	Ruim
3	Médio	Médio
4	Alto	Bom
5	Muito alto	Excelente

5.2 Arquitetura do Sistema

A arquitetura geral do sistema é apresentada na Figura 5.4, o sistema é composto por seis módulos: *Crawler*, Analisador, Construtor de Conhecimento, Indexador de Artigos, Construtor de Perfis e Recomendador. A recomendação é executada por meio dos seguintes processos:

Figura 5.4: Arquitetura geral do sistema, módulos e interações.



1. **Captura dos Dados** – O módulo *Crawler* captura os dados dos usuários e dos artigos nas fontes de dados, nessa etapa os dados ainda não sofreram nenhum processamento, e por isso os dados após esta etapa são chamados de *dados brutos*.

2. **Processamento dos Dados** – Os dados brutos são tratados pelo módulo Analisador, nesta etapa são aplicadas técnicas de Processamento de Linguagem Natural (PLN) nos dados brutos, após terminado este processo dá-se aos dados a nomenclatura de *dados pré-processados*.

3. **Construção da Base de Conhecimento** – O módulo Construtor de Conhecimento utiliza os dados pré-processados para criar a base de conhecimento na área de Ciência da Computação. Dessa forma, pode-se dizer que esse processo consiste em transformar os dados pré-processados em *informação* com respeito aos artigos acadêmicos da área de estudo.

4. **Construção dos Perfis de Usuário** – O perfil de usuário é construído pelo módulo Construtor de Perfis utilizando para isso os dados pré-processados do usuário e a informação contida na base de conhecimento. A construção do perfil de usuário pode ser

considerada como a transformação dos dados pré-processados em informação relativa ao usuário.

5. **Indexação dos Artigos** – Cada artigo é indexado utilizando os dados pré-processados pelo módulo Analisador utilizando o texto de seu título e abstract. No final do processo de indexação, cada perfil de artigo conterá informação relacionada ao mesmo.

6. **Recomendação de Artigos** – O processo de recomendação consiste em predizer as relevâncias entre o perfil de usuário e os artigos indexados, em que essa relevância consiste no cálculo da similaridade entre o perfil de usuário e os artigos.

7. **Exibição da Recomendação e Captura do Feedback** – Por fim os artigos recomendados são exibidos aos usuários pela fachada do sistema, além de que o *feedback* do usuário nos artigos também é capturado.

Em cada subseção que se segue são apresentados detalhes sobre cada módulo do sistema.

5.2.1 Módulo *Crawler*

Módulo responsável por capturar os dados brutos dos usuários nas fontes de dados em que ele está presente, como também os dados brutos dos artigos disponíveis para recomendação. Os dados foram capturados das fontes por meio de *Interface de Programação de Aplicativos* (do inglês, *Application Programming Interface* (API). Dessa forma, nesta versão do sistema, foram definidos três crawlers:

- Mendeley *Crawler* – faz a captura dos dados provenientes do Mendeley via Mendeley API[9]. Este processo consiste em capturar os documentos de leitura do usuário. Cada documento possui um identificador, um título e o ano em que foi adicionado à base pelo usuário.

- LinkedIn *Crawler* – faz a captura dos dados provenientes do LinkedIn via LinkedIn API[10]. A informação capturada do usuário diz respeito às competências e interesses informados pelo mesmo.

[9] <http://apidocs.mendeley.com/>
[10] <http://developer.linkedin.com/apis>

- Lattes *Crawler* – a plataforma Lattes não disponibiliza API para desenvolvedores. Deste modo, uma alternativa consiste em solicitar ao usuário o arquivo XML (*eXtensible Markup Language*) contendo informações relacionadas a seu perfil. No entanto, tal processo é pouco prático do ponto de vista do usuário. Por este motivo foi desenvolvido um *parser* que analisa a página Lattes do usuário e retorna os seguintes dados:

  - Resumo – currículo resumido do usuário;
  - Formação – dados sobre a formação acadêmica do usuário: graduação, mestrado e doutorado;
  - Projetos – projetos de pesquisa que o usuário participou ou participa;
  - Produção Técnica – produção técnica do usuário, dados de sofware, de patente e demais tipos de produção técnica;
  - Produção Bibliográfica – dados sobre a produção técnica do usuário: artigos em eventos, artigos em periódicos, livros, capítulos de livro, prefácio, posfácio, artigos aceitos para publicação e demais tipos de produção bibliográfica.

A Tabela 5.2 apresenta o mapeamento dos dados externos dos usuários para o metamodelo utilizado pelo sistema, no campo "ano do item", o valor "ano agora" significa o ano atual.

Tabela 5.2: Mapeamento dos dados externos dos usuários para o metamodelo utilizado pelo sistema. O símbolo "+" indica a concatenação de caracteres.

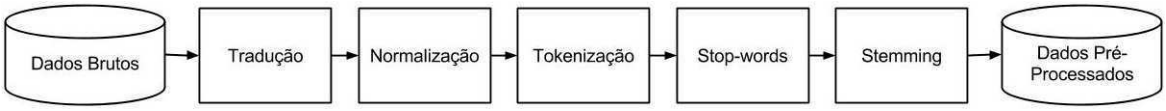
Fonte de dados	**Tipo de dado**	**Metamodelo**	
		Descrição do Item	**Ano do Item**
Mendeley	Documento	título + abstract	ano de inserção
LinkedIn	Competências e Interesses	nome	ano agora
Lattes	Resumo	descrição	ano agora
	Formação	título	ano
	Projeto	título + descrição	ano de conclusão
	Produção técnica	título + descrição + palavras-chave	ano
	Produção bibliografia	título + descrição + palavras-chave	ano

5.2.2 Módulo Analisador

Módulo responsável por analisar e pré-processar os dados brutos dos documentos textuais. Para tanto, foi utilizada a ferramenta NLTK[11] (Natural Language Toolkit) que é destinada a auxiliar desenvolvedores que necessitem realizar tarefas de PLN, como classificação, tokenização, *stemming*, entre outras.

As etapas do processo de pré-processamento de dados textuais são apresentadas na Figura 5.5 (BAEZA-YATES; RIBEIRO-NETO, 2011).

Figura 5.5: Etapas do pré-processamento textual.



O processo tem como entrada os dados brutos capturados pelo módulo *Crawler* e tem como saída os dados pré-processados. A seguir, são explanados detalhes sobre esse processo (BAEZA-YATES; RIBEIRO-NETO, 2011):

1. **Tradução** – essa etapa consiste em fazer a tradução dos documentos em Português para o Inglês, isso foi realizado devido aos artigos disponíveis para recomendação estarem descritos em Inglês, para isso foi utilizada a ferramenta online Google Translate[12];

2. **Normalização** – são removidos os caracteres especiais do texto (ex.: dígitos e pontuação) e o texto é colocado em caixa baixa;

3. **Tokenização** – são extraídos os unigramas do texto e é gerado uma lista de *tokens*;

4. **Stop-words** – as *stop-words*[13] e as palavras com tamanho menor que dois caracteres são removidas;

5. **Stemming** – aplicação de Stemming de Lancaster (PAICE, 1990).

A seguir é apresentado um exemplo do pré-processamento textual, o título do artigo (MAGALHÃES et al., 2012): "Improving a recommender system through integration of user

[11] <http://nltk.org/>

[12] <http://translate.google.com.br/>

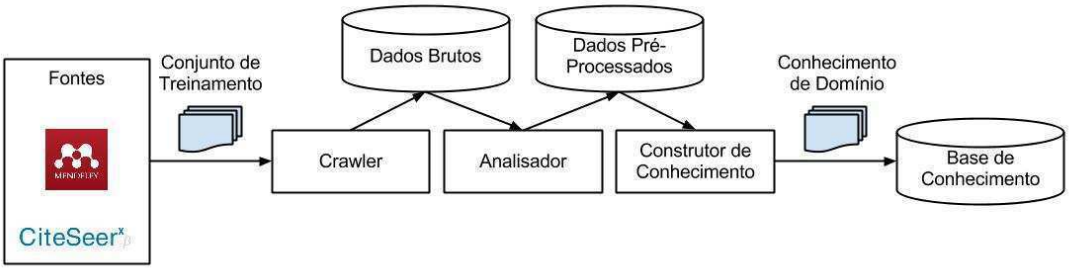
[13] As stop-words consideradas neste trabalho estão disponíveis em <nltk.googlecode.com/svn/trunk/nltk_data/packages/corpora/stopwords.zip>

profiles: a semantic approach.", entra como dado bruto no módulo Analisador e após o processo de pré-processamento é retornado o seguinte dado pré-processado: "[u'improv', u'recommend', u'system', u'integr', u'us', u'profil', u'sem', u'approach']", que consiste numa lista de termos.

5.2.3 Módulo Construtor de Conhecimento

Este módulo é responsável por realizar a construção do conhecimento, isto é, tem como entrada um conjunto de treinamento de artigos e a saída é uma base de conhecimento. Na Figura 5.6 é apresentada a arquitetura do construtor de conhecimento.

Figura 5.6: Arquitetura do módulo construtor de conhecimento.



Primeiramente, foi definida uma ontologia de Ciência da Computação baseada na taxonomia utilizada pelo Mendeley, contendo 19 conceitos, a saber: Algoritmos e Teoria da Computação, Inteligência Artificial, Arquitetura de Computadores, Segurança de Computadores, Redes de Computadores, Banco de Dados, Automação, Comércio Eletrônico, Computação Gráfica, Interação Humano-Computador, Recuperação de Informação, Ciência da Informação, Armazenamento de Informação, Sistemas Multimídia, Sistemas Operacionais, Linguagens de Programação, Sistemas em Tempo Real, Engenharia de Software, e Teoria do Controle.

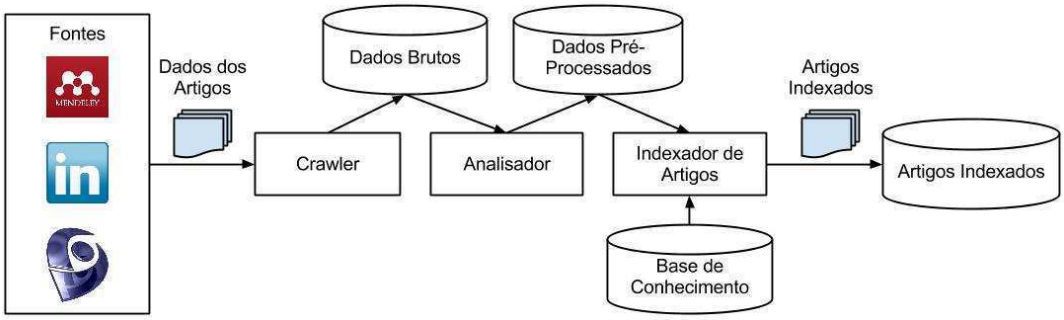
Após definido os conceitos, o próximo passo é a captura do conjunto de treinamento, em que para cada conceito foi definido um conjunto de treinamento de 1.000 artigos. Os artigos foram extraídos da plataforma Mendeley pelo módulo *Crawler*, utilizando a Mendeley API e pré-processados pelo módulo Analisador. Por fim, a base de conhecimento foi construída pelo construtor de conhecimento, conforme apresentado na Subseção 4.1.1. Na Figura 5.6, o

CiteSeerX[14] aparece como uma complementação ao Mendeley, o que seria de fácil inserção devido à flexibilidade da arquitetura.

5.2.4 Módulo Indexador de Artigos

Este módulo é responsável por indexar os artigos que são recomendados aos usuários. Na Figura 5.7 é apresentada a arquitetura do módulo Indexador. Inicialmente, os artigos são capturados das fontes, como Mendeley, CiteSeerX, ou outra biblioteca digital, utilizando o módulo *Crawler*. Em seguida, os artigos são analisados por sua descrição (título e abstract) pelo módulo *Analisador* e são indexados analisando-se por termos e conceitos, conforme apresentado na Seção 4.4.

Figura 5.7: Arquitetura do módulo indexador de artigos.



Para este primeiro teste, foram selecionados 50 artigos[15], 25 de Inteligência Artificial e 25 de Engenharia de Software, por especialistas, Dr. Evandro de Barros Costa[16] (Inteligência Articial) e Dr. Baldoino Fonseca dos Santos Neto[17] (Engenharia de Software). A Tabela 5.3 apresenta a especificação dos artigos utilizados neste teste inicial em relação às subáreas de Inteligência Artificial e Engenharia de Software, as subáreas foram definidas com base na taxonomia da ACM de 1998[18].

Os artigos foram selecionados a partir de conferências bem cotadas no intuito de isolar o fator qualidade do artigo no experimento, ou seja, espera-se evitar que um artigo que o usuário considera relevante ao seu trabalho receba uma nota baixa devido a sua qualidade.

[14] <http://citeseerx.ist.psu.edu>

[15] Os artigos estão disponíveis em <https://docs.google.com/file/d/0B8WYIdggJz4aaG5oYlRDTDNvSGc/edit?usp=sharing>

[16] <http://lattes.cnpq.br/5760364940162939>

[17] <http://lattes.cnpq.br/0306751604362704>

[18] A taxonomia está disponível em <http://www.acm.org/about/class/ccs98-html>

Tabela 5.3: Quantidade de artigos por cada subárea de Inteligência Artificial e Engenharia de Software.

Sub-área de Inteligência Artificial	**Quantidade de Artigos**
Aprendizado	5
Representação do Conhecimento – Formalismos e Métodos	4
Processamento de Linguagem Natural	2
Dedução, Provas de Teoremas e Processamento de Conhecimento	2
Resolução de Problemas, Métodos de Controle e Busca	3
Inteligência Artificial Distribuída	3
Web Semântica	2
Robótica	4
Total – Inteligência Artificial	**25**
Sub-área de Engenharia de Software	**Quantidade de Artigos**
Técnicas e Ferramentas de Design	4
Gerenciamento	4
Distribuição, Manutenção e Aprimoramento	3
Métricas/Medição	3
Teste e Depuração	2
Arquiteturas de Software	2
Software/Verificação de Programa	3
Requisitos/Especificações	2
Ferramentas e Técnicas de Codificação	2
Total – Engenharia de Software	**25**
Total de Artigos	**50**

Por fim, os artigos foram apresentados a cada usuário conforme apresentado na Figura 5.3, com o objetivo de capturar a preferência dos usuários para cada artigo.

5.2.5 Módulo Construtor de Perfis

Este módulo é responsável pela criação dos perfis de usuário, cuja arquitetura é apresentada na Figura 5.8. O processo de construção de perfil de usuário funciona da seguinte forma. Inicialmente, os dados dos usuários são capturados pelo Crawler e são pré-processados pelo Analisador. Em seguida, os perfis de usuário são criados e atualizados de forma automática pelo Construtor de Perfis e armazenados no repositório. O processo de construção de perfis consiste em combinar os dados de usuário e a base de conhecimento de acordo com os modelos apresentados na Seção 4.2.

Todavia, o usuário também poderá inspecionar seu perfil, como apresentado na Figura 5.8, no intuito de prover um *feedback* sobre sua consistência, podendo desta forma, ser atualizado pelo sistema. Uma forma de exibição do perfil do usuário é apresentada na

Figura 5.8: Arquitetura do módulo construtor de perfis.

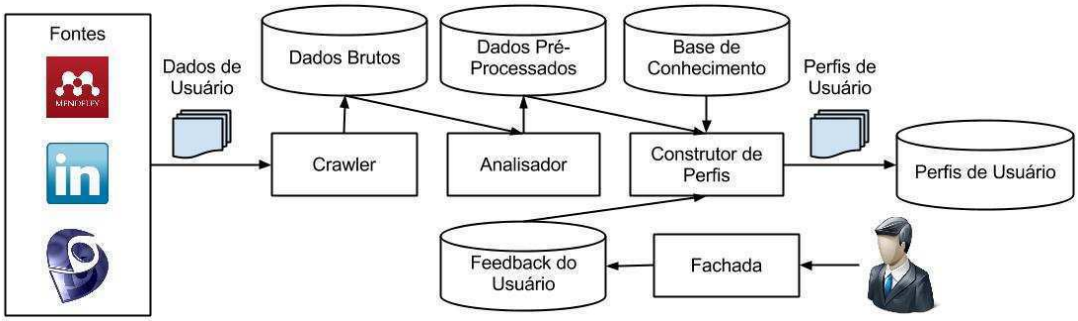


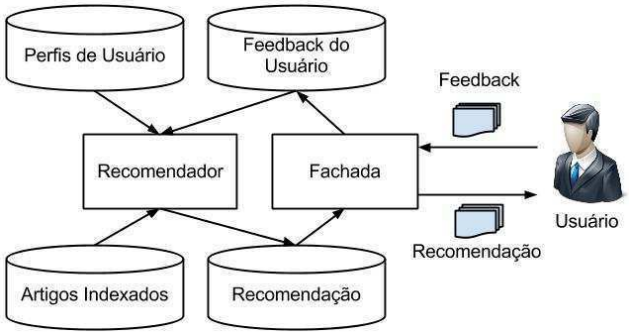
Figura 5.2, em que os interesses do usuário em cada área do conhecimento são expostos por meio de estrelas.

5.2.6 Módulo Recomendador

Este módulo é responsável por recomendar artigos ao usuário por meio dos algoritmos de recomendação descritos na Seção 4.5. Os artigos mais similares ao usuário são recomendados de acordo com algum limiar, podendo ser exibidos com certa periodicidade ao usuário, por exemplo, uma compilação semanal de artigos relevantes.

Na Figura 5.9 é apresentada a arquitetura deste módulo, sendo a recomendação realizada computando-se a relevância dos artigos em relação ao perfil de usuário e são apresentadas ao usuário.

Figura 5.9: Arquitetura do módulo recomendador.



O usuário pode dar *feedback* na recomendação segundo sua relevância (Figura 5.3). Assim, é possível o módulo recomendador utilizar esta informação no objetivo de prover recomendação mais relevante.

5.3 Discussão

Este Capítulo apresentou o sistema implementado de forma a possibilitar a validação da presente proposta. Primeiramente foi apresentada uma visão geral do sistema, sua arquitetura e as telas de apresentação ao usuário. Em seguida, foram apresentados detalhes sobre cada módulo do sistema. O sistema foi dividido em seis módulos: (i) *Crawler* – responsável por capturar conteúdo dos usuários nas fontes de dados e dos artigos disponíveis para recomendação, neste trabalho foram utilizadas três fontes de dados: Mendeley, LinkedIn e Lattes. Por este motivo foram definidos três *crawlers*, um para cada fonte. Além disto, foi indicada qual informação foi capturada de cada fonte e como a mesma foi armazenada. (ii) Analisador – responsável pelo pré-processamento textual, que consiste em seis passos: tradução, normalização, tokenização, remoção de *stop-words* e *stemming*. (iii) Construtor de Conhecimento. (iv) Indexador de Artigos. (v) Construtor de Perfis e; (vi) Recomendador.

O próximo Capítulo abordará questões relativas à validação da proposta.

Capítulo 6

Validação

Neste capítulo, é apresentada a validação do modelo proposto na presente dissertação. O seu objetivo consiste em apresentar as evidências empíricas coletadas que demonstram a eficácia da proposta.

6.1 Metodologia

O sistema apresentado no Capítulo 5 foi utilizado na validação, sendo divulgado em listas de email das seguintes Universidades: Universidade Federal de Alagoas (UFAL), campus Maceió e campus Arapiraca, Universidade Federal de Campina Grande (UFCG) e Universidade Federal de Minas Gerais (UFMG). Um total de 73 usuários se cadastrou no sistema, dos quais serão utilizados na avaliação apenas os que marcaram todos os artigos, o que totalizou 29 usuários.

A qualidade de cada estratégia de perfil de usuário foi medida pela métrica NDCG@5, Eq. 2.16, que foi computada pela comparação da lista de recomendação gerada pelo perfil e a marcação do usuário nos artigos. Por exemplo, considerando que os cinco melhores artigos marcados pelo usuário receberam as seguintes relevâncias: $[5, 4, 4, 3, 3]$, isto significa que essas são as notas máximas que um algoritmo pode alcançar; e supondo que os cinco artigos recomendados obtiveram relevância: $[4, 4, 3, 3, 3]$. O NDCG@5 é obtido de acordo com a Eq. 6.1.

$$NDCG = \frac{DCG - DCG_{min}}{DCG_{max} - DCG_{min}} = \frac{12,68 - 3,56}{14,31 - 3,56} = 0,84, \tag{6.1}$$

em que DCG é computado utilizando a lista [4, 4, 3, 3, 3] retornada pelo algoritmo, DCG_{max} é computado utilizando a lista ideal [5, 4, 4, 3, 3] e $NDCG_{min}$ é o pior caso, em que a lista recebe as piores notas possíveis, ou seja, [1, 1, 1, 1, 1].

6.2 Analisando a Plataforma Lattes

Na primeira investigação, foi analisada apenas a plataforma Lattes, utilizando como fator a estratégia de construção de perfil de usuário, tendo os seguintes níveis: (i) perfil de termos utilizando informação do Lattes (Lattes-Termos LT); (ii) perfil de conceitos utilizando informação do Lattes (Lattes-Conceitos LC); e (ii) estratégia de Lopes, Souto e Wives (2007) ($Lopes$). Nesta análise, pretende-se responder as seguintes perguntas:

- Q_1 – O perfil de conceitos LC obteve uma qualidade de recomendação comparável ao perfil de termos LT?
- Q_2 – Qual perfil escolher dentre (LT, LC ou $Lopes$)?

6.2.1 Resultados e Discussão

Na Figura 6.1 é apresentado o Beanplot da comparação entre as estratégias e na Tabela 6.1 são apresentadas as médias e as medianas das estratégias.

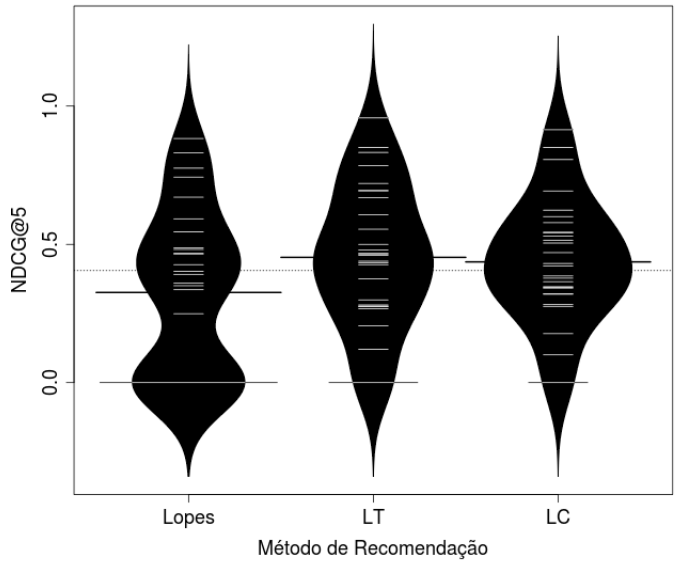
Tabela 6.1: Média e mediana do NDCG@5 das estratégias na plataforma Lattes.

	Lopes	**Lattes-Termos**	**Lattes-Conceitos**
Média	0,3262	**0,4530**	0,4367
Mediana	0,3596	**0,4600**	0,4212

Observando-se o Beanplot na Figura 6.1, verifica-se que a estratégia LT obteve melhores resultados em relação às outras estratégias, o que também é verificado pelos resultados das médias e medianas na Tabela 6.1.

Para dar mais representatividade ao estudo, foram rodados testes de hipóteses, comparando as estratégias estatisticamente. Inicialmente, foi aplicado um teste de normalidade, analisando se as amostras são provenientes de uma população normal. Esta análise é necessária para decidir qual o teste a ser aplicado. Para tanto, foi utilizado o teste de Shapiro-Wilk

Figura 6.1: Avaliação da recomendação utilizando diferentes estratégias de construção de perfil de usuário na plataforma Lattes.



(α=0,05), obtendo-se os seguintes resultados: $Lopes$ ($p\text{-}value$ = 0,002186), LT ($p\text{-}value$ = 0,579) e LC ($p\text{-}value$ = 0,6785). Logo, apenas o conjunto de dados $Lopes$ não é distribuído normalmente. Por isto, nas comparações envolvendo este conjunto foi utilizado o teste de Wilcox, nos outros casos, foi utilizado o teste t de Student. Na Tabela 6.2 são apresentados os testes realizados e os resultados da análise. No decorrer do texto, um teste estatístico será citado de acordo com seu id.

Tabela 6.2: Testes de hipótese realizados na comparação das estratégias na plataforma Lattes.

Id do Teste	**Teste realizado**	**Pareado**	**Hipótese Alternativa**	**P-value**	**Significado**
T_1	Teste de Wilcox (α=0,05)	✓	$LT > Lopes$	0,01543	Hipótese nula rejeitada
T_2	Teste de Wilcox (α=0,05)	✓	$LC > Lopes$	0,04292	Hipótese nula não rejeitada
T_3	Teste t de Student (α=0,05)	✓	$LT > LC$	0,3411	Hipótese nula não rejeitada
T_4	Teste t de Student(α=0,05)	✓	$LT! = LC$	0,6822	Hipótese nula não rejeitada

De acordo com Testes T_1 e T_2, os perfis LT e o LC foram estatisticamente superiores ao perfil $Lopes$. Isto se deve principalmente ao fato dessas estratégias analisarem mais informação de usuário proveniente do Lattes como, por exemplo, resumo do currículo, projetos e produção técnica.

Comparando-se o perfil de termos LT e o de conceitos LC (Testes T_3 e T_4), verificou-se

que a estratégia LT não obteve significância estatística. Este fato é creditado à qualidade da base de conhecimento construída. Com isto, é possível responder a Questão Q_1: *Sim, o perfil de conceitos LC obteve uma qualidade de recomendação comparável ao perfil de termos LT.*

Para responder à questão 2, que diz respeito à qual perfil deve ser escolhido, é necessário analisar outros aspectos, além da qualidade da recomendação.

A seguir são apresentadas vantagens e desvantagens entre os perfis LT e LC:

- LT – Perfil de rápida construção e fácil adaptação para outros domínios. No entanto, dependendo da quantidade de termos no perfil e nos itens indexados, o cálculo de similaridade torna-se um gargalo na recomendação. Outra desvantagem é que por ser formado por termos, não é um perfil muito intuitivo, o que torna difícil de ser inspecionado por um usuário.

- LC – Perfil que envolve uma engenharia de conhecimento para construir sua base de conhecimento, tanto na definição dos conceitos que serão utilizados quanto na definição do conjunto de treinamento. Este fato implica numa construção mais demorada e trabalhosa de perfil, além de ter uma difícil adaptação a outros domínios. Entretanto, após realizado este procedimento de conhecimento e comprovada sua eficácia, este perfil possui algumas vantagens, tais como: a computação mais rápida da recomendação por se tratar de um vetor de pesos com menor dimensão e ser mais intuitivo para o usuário inspecioná-lo.

Após esta análise dos perfis, é possível responder a questão Q_2: *O melhor perfil a ser utilizado para o domínio estudado é o Lattes-Conceitos LC*. No entanto, é importante frisar que o perfil LC é composto por apenas 19 conceitos, sendo necessária a investigação de outras formas de obtenção de conhecimento de domínio e a utilização de um número mais elevado de conceitos.

6.3 Analisando os Perfis Integrados

Nesta segunda análise, são analisados os perfis integrados, ou seja, informações externas ao Lattes são adicionadas aos perfis de usuário conforme os modelos de integração apresentados

na Seção 4.3. Nesta etapa, são utilizados apenas os usuários que possuem informação no Linkedin ou no Mendeley, o que totalizou 17 usuários. Foram utilizados dois fatores:

- Representação do perfil, com os seguintes níveis:
  - Perfil formado por termos (Termos);
  - Perfil formado por conceitos (Conceitos).
- Tipo de integração, com os seguintes níveis:
  - Fontes com importância igual (Igual);
  - Importância de acordo com a quantidade de itens (Quant);
  - Importância dada pela Atividade do Usuário na Fonte (Ativ).

Por conseguinte, fazendo-se as combinações possíveis, totalizam-se seis formas diferentes de computar o perfil de usuário. Na Tabela 6.3 são apresentadas as combinações e as siglas utilizadas para representar cada perfil.

Tabela 6.3: Perfis de usuário e suas respectivas siglas.

Representação de Perfil	**Tipo de Integração**		
	Igual	Quant	Ativ
Termos	EIT	QIT	AIT
Conceitos	EIC	QIC	AIC

Com relação à análise dos perfis integrados, pretende-se responder as seguintes perguntas:

- Q_3 – Os perfis de conceitos EIC, QIC e AIC obtiveram uma qualidade de recomendação comparável ao seus respectivos perfis de termos EIT, QIT e AIT?
- Q_4 – Qual perfil escolher dentre os integrados (EIC, QIC, AIC, EIT, QIT e AIT)?

6.3.1 Resultados e Discussão

Na Figura 6.2 é apresentado o Beanplot da comparação entre as estratégias e na Tabela 6.4 são apresentadas as médias e as medianas das estratégias.

Figura 6.2: Avaliação da recomendação utilizando diferentes estratégias de construção de perfil integrado de usuário nas plataformas Lattes, Mendeley e LinkedIn.

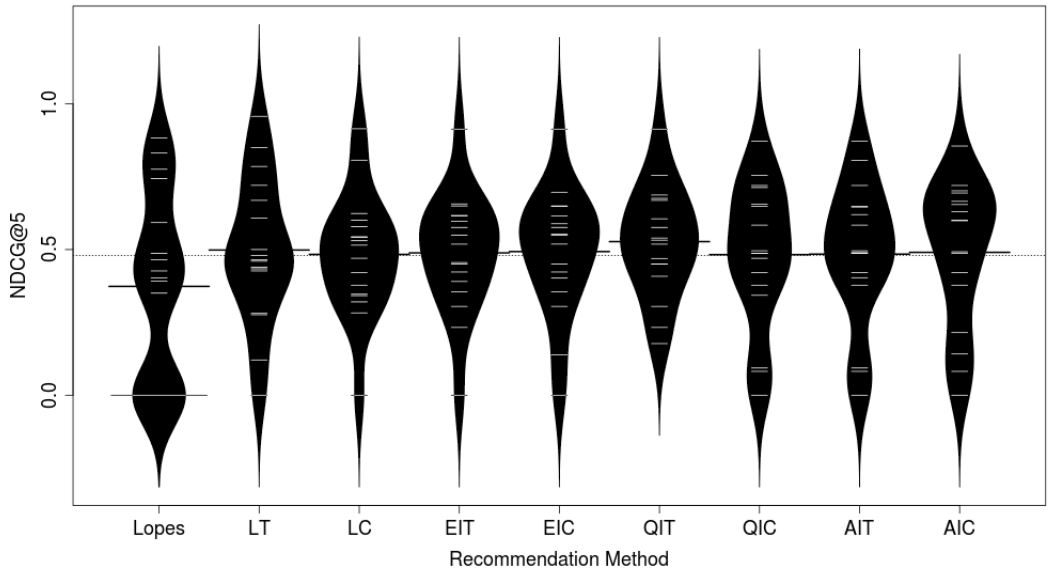


Tabela 6.4: Média e mediana do NDCG@5 das estratégias de perfil integrado nas plataformas Lattes, Mendeley e LinkedIn.

	Lopes	**LT**	**LC**	**EIT**	**EIC**	**QIT**	**QIC**	**AIT**	**AIC**
Média	0,3735	0,4983	0,4833	0,4931	0,4843	**0,5273**	0,4904	0,4826	0,4826
Mediana	0,4026	0,4654	0,5150	0,5492	0,4894	0,5320	**0,5988**	0,4872	0,4872

Analisando a Tabela 6.4, verifica-se que as estratégias de integração baseadas na quantidade, QIT e QIC, proporcionaram, respectivamente, a melhor média e mediana. No entanto, observando-se o Beanplot na Figura 6.2, verifica-se que não há diferenças significativas entre as estratégias.

Para confirmar este fato estatisticamente, verificou-se, inicialmente, a normalidade dos conjuntos dos dados por meio do teste de Shapiro-Wilk (α=0,05), sendo obtidos os seguintes resultados: $Lopes$ ($p\text{-}value$ = 0,02301), LT ($p\text{-}value$ = 0,8914), LC ($p\text{-}value$ = 0,6615), EIT ($p\text{-}value$ = 0,5787), EIC ($p\text{-}value$ = 0,2844), QIT ($p\text{-}value$ = 0,9846), QIC ($p\text{-}value$ = 0,129), AIT ($p\text{-}value$ = 0,6383) e AIC ($p\text{-}value$ = 0,3485). Assim, dentre todos os conjuntos, apenas o $Lopes$ não foi originado de uma distribuição normal.

Para responder a questão Q_3, foram aplicados testes de hipótese, comparando os perfis de conceitos com seus respectivos perfis de termos. Na Tabela 6.5 são apresentados os

testes realizados e os resultados obtidos. Como toda comparação entre perfis proporcionou $p\text{-}value > 0,05$, é possível responder a questão Q_3: *Sim, os perfis de conceitos obtiveram qualidade comparável aos perfis de termos*. Este fato tende a aumentar a certeza sobre a conclusão discutida na Seção anterior, cuja qualidade dos perfis de conceitos foi creditada à base de conhecimento gerada.

Tabela 6.5: Testes de hipótese realizados na comparação das estratégias de perfis integrados nas plataformas Lattes, Mendeley e LinkedIn.

Id do Teste	Teste realizado	Pareado	Hipótese Alternativa	P-value	Significado
T_5	Teste t de Student (α=0,05)	✓	$QIC! = QIT$	0,4768	Hipótese nula rejeitada
T_6	Teste t de Student (α=0,05)	✓	$AIC! = AIT$	0,923	Hipótese nula não rejeitada
T_7	Teste t de Student (α=0,05)	✓	$EIC! = EIT$	0,8837	Hipótese nula não rejeitada

Para responder a questão Q_4, foi aplicado o teste de Friedman com a hipótese alternativa de que existe diferença entre os conjuntos. Foi obtido ($p\text{-}value$ = 0,9971), logo, não existe diferença significativa entre os conjuntos de dados. Com isto, não é possível responder a questão Q_4, ou seja, não houve uma estratégia de perfil integrado que se sobressaísse em relação às demais.

6.4 Definindo um Modelo de Recomendação

Nesta seção, é apresentado um modelo de recomendação que faz o chaveamento de qual método de recomendação utilizar baseado nas características do usuário. A ideia não é encontrar o melhor método para o caso geral, mas encontrar qual o melhor método para cada situação. O modelo escolhido foi o de Redes Bayesianas (RB), por possuir uma forte base matemática, além de possuir uma fácil evolução por sua natureza dinâmica.

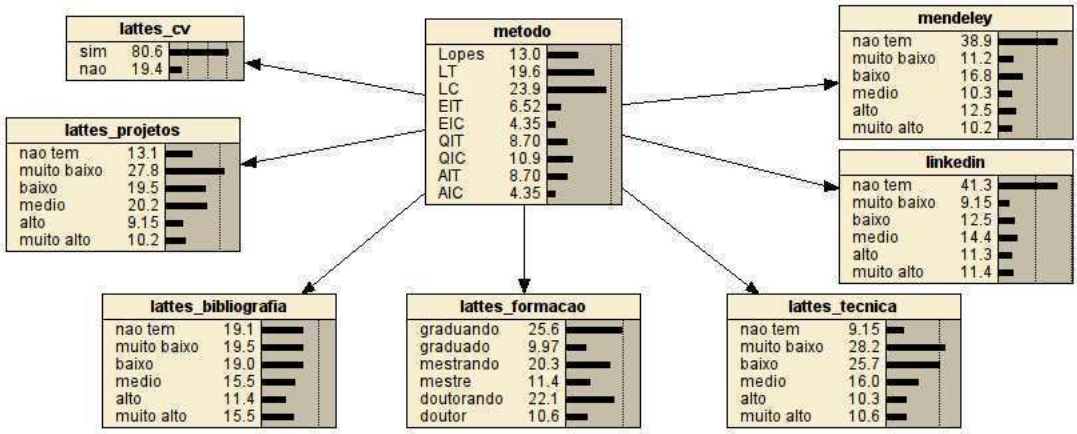
6.4.1 Estrutura e Nós da RB

Na Figura 6.3 é apresentada a RB, a qual foi desenvolvida com a ferramenta Netica[1], sendo composta por oito nós: $metodo$, $lattes\_cv$, $lattes\_projetos$, $lattes\_bibliografia$,

[1] <http://www.norsys.com/netica.html>

$lattes\_formacao$, $lattes\_tecnica$, $linkedin$ e $mendeley$. A seguir, são apresentados detalhes sobre cada nó, seu significado, seus estados e como a informação do usuário é mapeada para rede.

Figura 6.3: Modelo de recomendação utilizando Redes Bayesianas.



- **Nó metodo** – Nó-pai dos demais nós da rede, representa os possíveis métodos de recomendação disponíveis para utilização e, consequentemente cada estado desse nó será um método de recomendação:
  - $Lopes$ – recomendação proposta por Lopes, Souto e Wives (2007);
  - LT – recomendação utilizando o perfil Lattes-Termos;
  - LC – recomendação utilizando o perfil Lattes-Conceitos;
  - EIT – recomendação utilizando o perfil Integrado-Igual-Termos;
  - EIC – recomendação utilizando o perfil Integrado-Igual-Conceitos;
  - QIT – recomendação utilizando o perfil Integrado-Quantidade-Termos;
  - QIC – recomendação utilizando o perfil Integrado-Quantidade-Conceitos;
  - AIT – recomendação utilizando o perfil Integrado-Atividade-Termos;
  - AIC – recomendação utilizando o perfil Integrado-Atividade-Conceitos.
- **Nó lattes_cv** – Nó que indica se o usuário possui o currículo resumido no Lattes, possuindo dois estados: *sim* – se o usuário possui currículo e *não* – caso contrário.

- **Nó lattes_formacao** – Este nó indica a formação acadêmica do usuário, possuindo seis estados:
  - $graduando$ – se o usuário ainda não possui uma graduação;
  - $graduado$ – se o usuário já terminou algum curso de graduação;
  - $mestrando$ – se o usuário está cursando o mestrado;
  - $mestre$ – se o usuário terminou o mestrado;
  - $doutorando$ – se o usuário está cursando o doutorado;
  - $doutor$ – se o usuário terminou o doutorado.
- **Nós lattes_projetos, lattes_bibliografia, lattes_tecnica, linkedin e mendeley** – Estes nós computam a quantidade de conteúdo do usuário e possuem os mesmos estados, a diferença entre eles está no tipo de conteúdo que cada um representa e na forma como a informação do usuário é mapeada. A seguir, é apresentado o tipo de informação que cada nó trata.
  - Nó $lattes\_projetos$ – representa a quantidade de projetos (de pesquisa, de extensão ou de desenvolvimento) dos quais o usuário participou ou está participando.
  - Nó $lattes\_bibliografia$ – representa a quantidade de produção bibliográfica (trabalhos em eventos, artigos, livros, capítulos de livro, posfácio e prefácio, artigos aceitos para publicação e outras publicações) do usuário.
  - Nó $lattes\_tecnica$ – representa a quantidade de produção técnica (softwares, patentes e demais tipos de produção técnica) do usuário.
  - Nó $linkedin$ – representa a quantidade de conteúdo (competências e interesses) do usuário no LinkedIn.
  - Nó $mendeley$ – representa a quantidade de conteúdo (artigos) do usuário no Mendeley.

  Na Tabela 6.6, são apresentados os possíveis estados dos nós e as regras de mapeamento das informações do usuário.

Tabela 6.6: Estados e mapeamento dos nós $lattes\_bibliografia$, $lattes\_tecnica$, $linkedin$, $mendeley$ e $lattes\_projetos$.

Estado do Nó	**Nó**		
	$lattes\_bibliografia, lattes\_tecnica, linkedin$	$mendeley$	$lattes\_projetos$
nao_tem	$quant = 0$	$quant = 0$	$quant = 0$
$muito\_baixo$	$0 < quant < 5$	$0 < quant < 50$	$0 < quant < 3$
$baixo$	$5 \leq quant < 10$	$50 \leq quant < 100$	$3 \leq quant < 5$
$medio$	$10 \leq quant < 15$	$100 \leq quant < 150$	$5 \leq quant < 7$
$alto$	$15 \leq quant < 20$	$150 \leq quant < 200$	$7 \leq quant < 9$
$muito\_alto$	$quant \geq 20$	$quant \geq 200$	$quant \geq 9$

6.4.2 Definindo os Pesos da RB

Os pesos da RB, probabilidades *a priori* e probabilidades condicionais, foram definidas a partir de aprendizado por meio da inserção de casos na ferramenta Netica.

1. Para cada usuário foram criados n casos, cujo número n significa a quantidade de métodos que proporcionaram o melhor desempenho de acordo com a métrica NDCG@5 para o usuário, de tal forma que $n > 1$ apenas em casos em que ocorreram empates entre os métodos;

2. Cada usuário teve suas informações mapeadas conforme apresentado na Subseção anterior, sendo o campo $metodo$ preenchido com o método que obteve melhor resultado.

Na Tabela 6.7, é apresentado o conjunto de casos formado para o aprendizado dos pesos da RB. A rede compilada com os casos é exibida na Figura 6.3.

6.4.3 Utilizando o Modelo para Recomendação

O processo de recomendação utilizando o modelo consiste em duas etapas:

1. Inserção de informações relativas à quantidade de conteúdo que o usuário possui nas fontes de dados (Lattes, LinkedIn e Mendeley);

2. Inferência de qual método é o mais apropriado para aquele usuário, ou seja, é retornado o método de maior probabilidade; em caso de empate entre métodos, opta-se pelo método com menor custo de recomendação. Por exemplo, em caso de empate entre

Tabela 6.7: Conjunto de casos utilizados no aprendizado dos pesos do modelo.

cv	projetos	bibliografia	formacao	tecnica	linkedin	mendeley	metodo
sim	baixo	muito_baixo	graduado	baixo	muito_alto	muito_baixo	EIT
sim	muito_baixo	nao_tem	graduando	muito_baixo	nao_tem	baixo	LC
sim	medio	alto	doutorando	alto	baixo	alto	Lopes
sim	baixo	medio	doutorando	baixo	medio	nao_tem	LT
sim	baixo	baixo	doutorando	baixo	nao_tem	nao_tem	LC
sim	muito_baixo	medio	doutorando	medio	nao_tem	alto	QIC
sim	muito_baixo	nao_tem	graduando	muito_baixo	nao_tem	baixo	LC
sim	muito_baixo	nao_tem	graduando	muito_baixo	nao_tem	baixo	AIT
sim	muito_baixo	nao_tem	graduando	muito_baixo	nao_tem	baixo	EIT
sim	muito_baixo	nao_tem	graduando	muito_baixo	nao_tem	baixo	QIT
sim	baixo	baixo	graduando	muito_baixo	alto	muito_baixo	Lopes
sim	muito_baixo	muito_baixo	mestrando	medio	baixo	nao_tem	AIT
sim	medio	baixo	graduando	baixo	nao_tem	nao_tem	Lopes
sim	medio	baixo	graduando	baixo	nao_tem	nao_tem	LT
sim	medio	baixo	graduando	baixo	nao_tem	nao_tem	LC
sim	baixo	medio	mestrando	baixo	nao_tem	nao_tem	LC
sim	muito_baixo	muito_baixo	mestrando	muito_baixo	nao_tem	nao_tem	LC
sim	muito_baixo	muito_baixo	graduando	muito_baixo	baixo	nao_tem	Lopes
sim	muito_baixo	muito_alto	doutorando	baixo	nao_tem	muito_alto	QIC
sim	medio	muito_alto	doutorando	medio	medio	nao_tem	AIC
sim	medio	muito_alto	doutorando	medio	medio	nao_tem	EIC
sim	medio	muito_alto	doutorando	medio	medio	nao_tem	QIC
sim	medio	muito_alto	doutorando	medio	medio	nao_tem	LC
sim	medio	baixo	mestrando	baixo	nao_tem	baixo	AIT
sim	medio	baixo	mestrando	baixo	nao_tem	baixo	QIT
sim	nao_tem	muito_baixo	mestrando	muito_alto	nao_tem	nao_tem	LT
sim	muito_baixo	medio	doutorando	baixo	nao_tem	nao_tem	LT
sim	nao_tem	baixo	mestre	muito_baixo	nao_tem	medio	Lopes
nao	muito_baixo	nao_tem	mestrando	muito_baixo	nao_tem	nao_tem	LT
sim	nao_tem	muito_alto	doutor	baixo	nao_tem	nao_tem	LT
sim	muito_baixo	muito_baixo	graduando	muito_baixo	nao_tem	nao_tem	LC
sim	muito_baixo	nao_tem	graduando	muito_baixo	nao_tem	nao_tem	LT
nao	muito_alto	alto	mestre	muito_baixo	nao_tem	alto	QIC
sim	baixo	muito_baixo	doutorando	baixo	nao_tem	nao_tem	LC
sim	baixo	medio	mestrando	medio	alto	nao_tem	QIT
sim	muito_baixo	nao_tem	graduando	muito_baixo	muito_alto	nao_tem	LT
sim	baixo	muito_baixo	mestrando	muito_baixo	nao_tem	nao_tem	LC

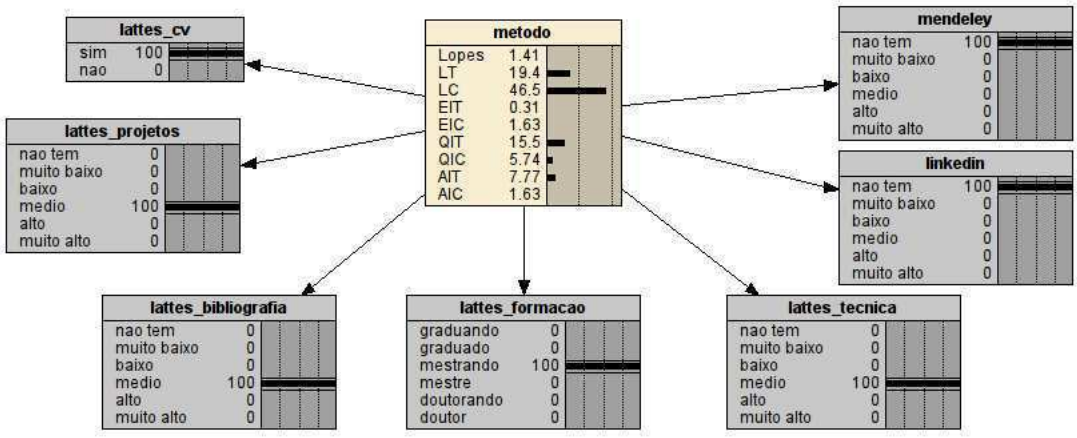
um método que utilize apenas o Lattes e um método integrado, opta-se pelo que utilize o Lattes, pois não há o custo de analisar outras fontes.

Para demonstrar o funcionamento do modelo foram definidos quatro casos, demonstrados a seguir.

Caso 1: Usuários que Possuem Informação Apenas no Lattes

Neste caso, é esperado que a saída do modelo seja um modelo de recomendação que utilize apenas o Lattes em sua concepção. Na Figura 6.4 é apresentado o modelo configurado para este caso. Percebe-se que o método retornado foi o LC - Lattes-Conceitos com 46,5%, o que está coerente com a saída esperada.

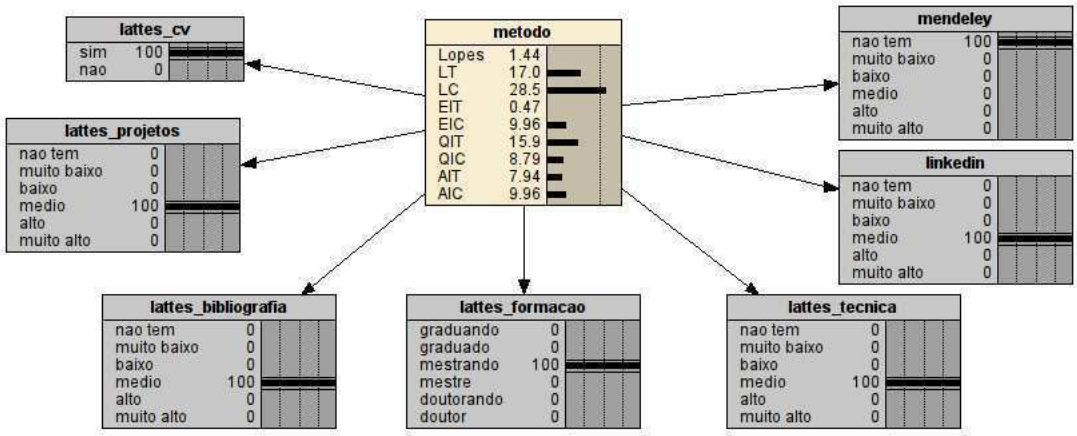
Figura 6.4: Inferência no modelo: recomendando um método de recomendação para usuários que possuam informação apenas no Lattes. Imagem da ferramenta Netica.



Caso 2: Usuários que Possuem Informação no Lattes e no LinkedIn

Para este caso em que se procura recomendar para usuários que possuem informação em outra fonte, Lattes + LinkedIn, existem duas possibilidades: (i) ou a informação do LinkedIn melhorará o perfil de usuário, de forma que alguma recomendação integrada será recomendada; ou (ii) o LinkedIn não é um bom agregador e alguma recomendação utilizando apenas o Lattes será considerada a melhor. Na Figura 6.5, é apresentado o modelo para este caso. Percebe-se que o método de recomendação retornado foi o LC - Lattes-Conceitos com 28,5%, o que significa que a informação do LinkedIn não melhorou o perfil de usuário.

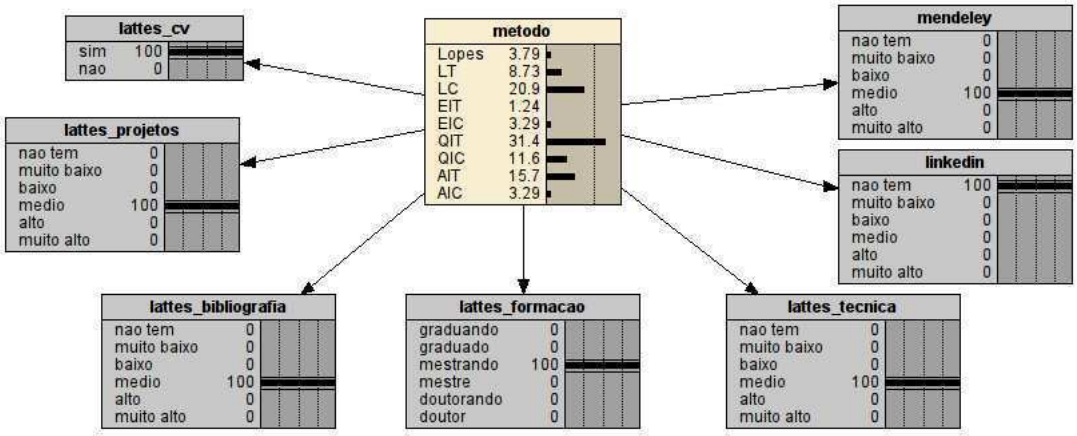
Figura 6.5: Inferência no modelo: recomendando um método de recomendação para usuários que possuam informação no Lattes e no LinkedIn. Imagem da ferramenta Netica.



Caso 3: Usuários que Possuem Informação no Lattes e no Mendeley

Este caso é semelhante ao anterior, no entanto considera-se que o usuário possui informação no Lattes e no Mendeley. Da mesma forma, pode haver duas possibilidades: (i) ou o perfil de usuário é melhorado adicionando-se informação do Mendeley e alguma recomendação integrada será recomendada; ou (ii) a informação do Mendeley não é relevante para o perfil de usuário e alguma recomendação utilizando apenas o Lattes será recomendada. Na Figura 6.6, é apresentado o modelo de recomendação para este caso, em que o método de recomendação retornado foi o que utiliza o perfil integrado QIT - Integrado-Quantidade-Termos com 31,4%, o que significa que a informação do Mendeley agregou qualidade ao perfil de usuário.

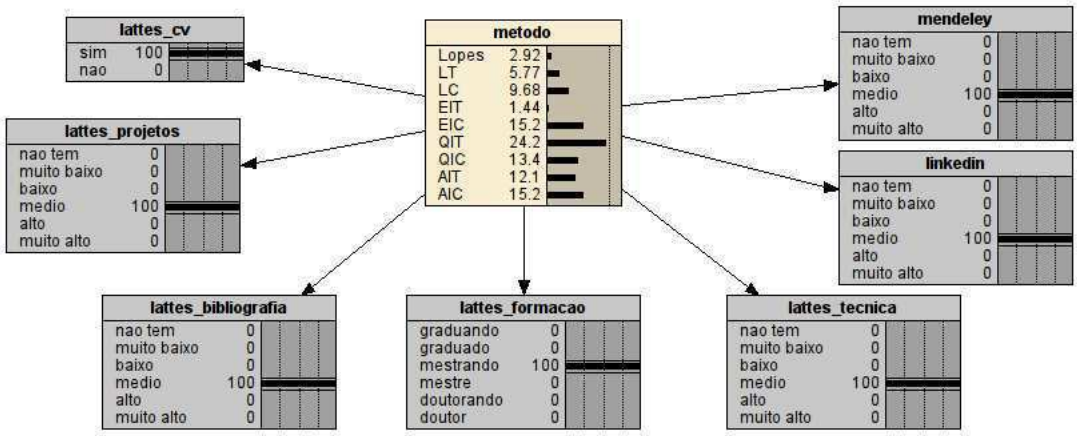
Figura 6.6: Inferência no modelo: recomendando um método de recomendação para usuários que possuam informação no Lattes e no Mendeley. Imagem da ferramenta Netica.



Caso 4: Usuários que Possuem Informação em Todas as Fontes (Lattes, Mendeley e LinkedIn)

Neste caso, deseja-se investigar qual método de recomendação melhor se adequa aos usuários que possuem informação em todas as fontes de dados (Lattes, Mendeley e LinkedIn). Pelo que foi apresentado no Caso 3, em que a agregação do Mendeley ocasionou no modelo recomendar o método de recomendação integrado, espera-se que algum perfil integrado seja recomendado. Na Figura 6.7 é apresentado o modelo para este caso, percebendo-se que o método recomendado foi do tipo integrado, QIT - Integrado-Quantidade-Termos com 24,2%, o que demonstra que o modelo está coerente com a saída esperada.

Figura 6.7: Inferência no modelo: recomendando um método de recomendação para usuários que possuam informação em todas as fontes (Lattes, Mendeley e LinkedIn). Imagem da ferramenta Netica.



6.5 Ameaças à Validade

A seguir são apresentadas as ameaças à validade que foram identificadas neste trabalho. Em relação à validade de conclusão pode-se citar o número limitado de usuários, apenas 29, o que implica num baixo poder estatístico das conclusões. Isto ocorreu devido ao usuário necessitar marcar 50 artigos para efetivar sua participação, o que requeria um tempo considerável.

À respeito da validade interna pode-se citar três fatores: (i) ao final do processo o usuário poderia estar cansado de marcar os artigos, e com isso poderia atribuir notas incoerentes aos últimos artigos da lista; para amenizar esse efeito foi realizada uma randomização na forma como os artigos foram apresentados a cada usuário; (ii) no processo de seleção dos artigos

não foi verificado se o usuário já conhecia algum artigo, no entanto, acredita-se que este fato, ao invés de atrapalhar, aumente a confiança da relevância do usuário num artigo; além disso, no intuito de excluir o fator qualidade do artigo do experimento, optou-se por escolher somente artigos de qualidade e de boas conferências; e (iii) não foi verificado se os usuários eram das áreas utilizadas no experimento, Inteligência Artificial e Engenharia de Software, o que poderia influenciar o julgamento dos usuários nos artigos.

Em relação à validade externa, verifica-se que é difícil de generalizar os resultados obtidos, pois o estudo foi realizado no domínio de Ciência da Computação, além de que os artigos escolhidos foram das áreas de Engenharia de Software e Inteligência Artificial. No entanto, é importante frisar que a adequação dos modelos em outros domínios não seria uma tarefa difícil.

Capítulo 7

Considerações Finais

Neste capítulo são apresentadas as considerações finais à pesquisa ora apresentada, englobando as principais conclusões. Primeiramente é exposta uma caraterização geral do presente trabalho, logo após são discutidas as principais contribuições. Por fim, são apresentados possíveis desdobramentos futuros.

7.1 Caracterização Geral da Pesquisa

Nesta pesquisa foi proposto, conforme apresentado no Capítulo 4, um modelo para Integração de Perfis de Usuário em múltiplas fontes de dados. Em que o modelo de perfil de usuário foi construído de forma a atender os requisitos de ser implícito e automático. Com vistas à testar a utilização do modelo, o mesmo foi instanciado num Sistema de Recomendação Personalizada de artigos científicos como apresentado no Capítulo 5. Os perfis de usuário foram construídos utilizando de dados capturados de três fontes de dados: Currículo Lattes, LinkedIn e Mendeley. Os perfis foram construídos para cada fonte de dados e integrados por meio de combinação linear.

Para validar os modelos de perfis e o Sistema de Recomendação, foi realizado um experimento envolvendo 29 usuários com o objetivo de analisar a qualidade da recomendação, conforme apresentado no Capítulo 6. Na primeira avaliação, envolvendo apenas o Lattes, verificou-se que os métodos propostos superaram a estratégia de Lopes, Souto e Wives (2007) com significância estatística. O que pode ser caracterizado como uma importante contribuição, levando-se em consideração a importância do Lattes para os pesquisadores

brasileiros. Uma segunda avaliação foi realizada com o intuito de avaliar os perfis integrados, para isso foram utilizados apenas os usuários que possuíam informação no Linkedin ou no Mendeley, o que totalizou 17 usuários. No entanto, nesta análise não foram alcançados os resultados esperados, pois os perfis integrados não obtiveram melhores resultados do que os perfis sem integração. Portanto, a modelo utilizado na integração de perfis de usuário necessita ser melhor investigado, seja realizando um experimento com mais fatores ou buscando uma amostra maior de usuários.

7.2 Principais Contribuições

Como contribuições deste trabalho, pode-se enumerar: i) avanço no estado da arte de Sistemas de Recomendação de Artigos utilizando-se o currículo da plataforma Lattes; ii) proposição de modelos de integração de perfis de usuário, em que se buscou mensurar a importância de uma fonte de dados para um usuário; iii) criação de uma base de dados, contendo informações de usuários em três fontes de dados, de forma a possibilitar que trabalhos futuros possam ser comparados ao presente; iv) criação de uma base de conhecimento no domínio de Ciência da Computação que poderá ser utilizada por outros trabalhos; v) proposição de um selecionador de métodos de recomendação; e (vi) um artigo publicado no 3rd International Workshop on Social Recommender Systems 2012, conforme apresentado no Apêndice A.

7.3 Sugestões para Trabalhos Futuros

Para trabalhos futuros, sugere-se os seguintes desdobramentos:

- Visualização de perfil – Pesquisar e desenvolver técnicas de visualização de perfil de usuário e validá-las por meio de experimento;
- Base de conhecimento – Melhoria da obtenção de conhecimento de domínio por meio do aprendizado de ontologias, utilização de conceitos diferentes, como a nova taxonomia proposta pela ACM[1];

[1] <http://www.acm.org/about/class/2012>

- Análise do artigo a ser recomendado – Incorporar atributos dos artigos no Sistema de Recomendação, por exemplo, qualidade do evento em que foi publicado (por exemplo o Qualis[2]), quantidade de citações, dentre outros;

- Recomendação de outros recursos acadêmicos – Analisar como adequar e modelar a recomendação de outros tipos de recursos acadêmicos, tais como pesquisadores, conferências, etc.;

- Recomendação de outros tipos de itens – Analisar o modelo proposto em outros tipos de SR, tais como filmes, livros, etc.;

- Caracterização e análise dos usuários do Lattes – Esta etapa pode ser realizada por meio de algoritmos de agrupamento, no intuito de definir tipos de perfil de usuário;

- Outras áreas do conhecimento – Adequar o sistema a outras áreas do conhecimento, por exemplo, Medicina, Engenharias, etc., verificando a interdisciplinaridade entre elas;

- Recomendação de recursos educacionais – Adequar o modelo para atender a fins educacionais, o objetivo é que o modelo ofereça suporte a um curso e modele as necessidades do estudante;

- Outras fontes de dados – Analisar a utilização e o impacto de outras fontes de dados no perfil de usuário, por exemplo, CiteULike[3], ArnetMiner[4], Facebook[5], etc.;

- Recomendação para grupos – Prover recomendação para grupos de pesquisa e para laboratórios de pesquisa. O desafio neste caso está em como modelar o grupo e fazer o balanceamento entre as necessidades de cada usuário do grupo;

- Melhorar o algoritmo de recomendação – Incorporar outros elementos ao SR, tais como: (i) atualização do perfil do usuário de acordo com o seu *feedback* nos artigos e nas áreas de domínio; (ii) analisar as citações e as referências existentes nos artigos do usuário; e (iii) incorporar métodos de FC na recomendação.

[2] <http://qualis.capes.gov.br/webqualis/principal.seam>
[3] <http://www.citeulike.org/>
[4] <http://arnetminer.org/>
[5] <https://www.facebook.com/>

Bibliografia

ADOMAVICIUS, G.; TUZHILIN, A. 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*, IEEE Educational Activities Department, v. 17, n. 6, p. 734–749, 2005. ISSN 10414347. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1423975>.

BAEZA-YATES, R. A.; RIBEIRO-NETO, B. A. *Modern Information Retrieval - The Concepts and Technology behind search, Second edition*. [S.l.]: Pearson Education Ltd., Harlow, England, 2011. ISBN 978-0-321-41691-9.

BERKOVSKY, S.; KUFLIK, T.; RICCI, F. Mediation of user models for enhanced personalization in recommender systems. *User Modeling and User-Adapted Interaction*, Kluwer Academic Publishers, Hingham, MA, USA, v. 18, n. 3, p. 245–286, ago. 2008. ISSN 0924-1868. Disponível em: <http://dx.doi.org/10.1007/s11257-007-9042-9>.

BOBADILLA, J. et al. Recommender systems survey. *Knowledge-Based Systems*, v. 46, n. 0, p. 109 – 132, 2013. ISSN 0950-7051. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0950705113001044>.

GOOSSEN, F. et al. News personalization using the cf-idf semantic recommender. In: *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*. New York, NY, USA: ACM, 2011. (WIMS '11), p. 10:1–10:12. ISBN 978-1-4503-0148-0. Disponível em: <http://doi.acm.org/10.1145/1988688.1988701>.

GORI, M.; PUCCI, A. Research paper recommender systems: A random-walk based approach. In: *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*. Washington, DC, USA: IEEE Computer Society, 2006. (WI '06), p. 778–781. ISBN 0-7695-2747-7. Disponível em: <http://dx.doi.org/10.1109/WI.2006.149>.

HERLOCKER, J. et al. An algorithmic framework for performing collaborative filtering. In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999. p. 230–237. Disponível em: <http://portal.acm.org/citation.cfm?id=312682>.

HUANG, S. et al. Tssp: A reinforcement algorithm to find related papers. In: *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*. Washington, DC, USA: IEEE Computer Society, 2004. (WI '04), p. 117–123. ISBN 0-7695-2100-2. Disponível em: <http://dx.doi.org/10.1109/WI.2004.145>.

JÄRVELIN, K.; KEKÄLÄINEN, J. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, ACM, New York, NY, USA, v. 20, n. 4, p. 422–446, out. 2002. ISSN 1046-8188. Disponível em: <http://doi.acm.org/10.1145/582415.582418>.

JIANG, Y. et al. Recommending academic papers via users' reading purposes. In: *Proceedings of the sixth ACM conference on Recommender systems*. New York, NY, USA: ACM, 2012. (RecSys '12), p. 241–244. ISBN 978-1-4503-1270-7. Disponível em: <http://doi.acm.org/10.1145/2365952.2366004>.

LATTES, P. *Painel Lattes – Estatísticas da Base de Currículos da Plataforma Lattes*. 2013. Acessado em 10/10/2013. Disponível em: <http://estatico.cnpq.br/painelLattes/>.

LINKEDIN. *LinkedIn – About us*. 2013. Acessado em 10/10/2013. Disponível em: <http://www.linkedin.com/about-us>.

LOH, S. et al. Constructing Domain Ontologies for Indexing Texts and Creating Users' Profiles. In: *Work. on Ontologies and Metamodeling in Software and Data Engineering, Brazilian Symp. on Databases, UFSC, Florianópolis*. [s.n.], 2006. p. 72–82. Disponível em: <http://paginas.ucpel.tche.br/~loh/pdfs/womsde.pdf>.

LOPES, G.; SOUTO, M.; WIVES, L. Personalizing bibliographic recommendation under semantic web perspective. In: *International Workshop on Web Information Systems Modeling, WISM'07; CAISE'07*. Trondheim, Norway: [s.n.], 2007. p. 779–790. Disponível em: <http://people.few.eur.nl/frasincar/workshops/wism2007/Papers/wism2007-6.pdf>.

LUO, X.; XIA, Y.; ZHU, Q. Incremental collaborative filtering recommender based on regularized matrix factorization. *Knowledge-Based Systems*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 27, p. 271–280, mar. 2012. ISSN 0950-7051. Disponível em: <http://dx.doi.org/10.1016/j.knosys.2011.09.006>.

MAGALHÃES, J. et al. Improving a recommender system through integration of user profiles: a semantic approach. In: *3rd International Workshop on Social Recommender Systems – UMAP Workshops*. [S.l.]: CEUR-WS.org, 2012. (CEUR Workshop Proceedings, v. 872).

MCNEE, S. M. et al. On the recommending of citations for research papers. In: *Proceedings of the 2002 ACM conference on Computer supported cooperative work*. New York, NY, USA: ACM, 2002. (CSCW '02), p. 116–125. ISBN 1-58113-560-2. Disponível em: <http://doi.acm.org/10.1145/587078.587096>.

MCNEE, S. M.; KAPOOR, N.; KONSTAN, J. A. Don't look stupid: avoiding pitfalls when recommending research papers. In: *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. New York, NY, USA: ACM, 2006. (CSCW '06), p. 171–180. ISBN 1-59593-249-6. Disponível em: <http://doi.acm.org/10.1145/1180875.1180903>.

MENDELEY. *Mendeley – It's time to change the way we do research*. 2013. Acessado em 10/10/2013. Disponível em: <http://www.mendeley.com/>.

NASCIMENTO, C. et al. A source independent framework for research paper recommendation. In: *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*. New York, NY, USA: ACM, 2011. (JCDL '11), p. 297–306. ISBN 978-1-4503-0744-4. Disponível em: <http://doi.acm.org/10.1145/1998076.1998132>.

PAGE, L. et al. *The PageRank Citation Ranking: Bringing Order to the Web*. [S.l.], 1998. Disponível em: </brokenurl\#http://publication.wilsonwong.me/load.php?id=233281827>.

PAICE, C. D. Another stemmer. *SIGIR Forum*, ACM, New York, NY, USA, v. 24, n. 3, p. 56–61, nov. 1990. ISSN 0163-5840. Disponível em: <http://doi.acm.org/10.1145/101306.101310>.

SAHEBI, S.; WONGCHOKPRASITTI, C.; BRUSILOVSKY, P. Recommending research colloquia: a study of several sources for user profiling. In: *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*. New York, NY, USA: ACM, 2010. (HetRec '10), p. 32–38. ISBN 978-1-4503-0407-8. Disponível em: <http://doi.acm.org/10.1145/1869446.1869451>.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, v. 24, n. 5, p. 513–523, 1988.

SCHEIN, A. I. et al. Methods and metrics for cold-start recommendations. In: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2002. (SIGIR '02), p. 253–260. ISBN 1-58113-561-0. Disponível em: <http://doi.acm.org/10.1145/564376.564421>.

SHANI, G.; GUNAWARDANA, A. Evaluating recommendation systems. In: RICCI, F. et al. (Ed.). *Recommender Systems Handbook*. Springer US, 2011. p. 257–297. ISBN 978-0-387-85819-7. Disponível em: <http://dx.doi.org/10.1007/978-0-387-85820-3_8>.

SOUZA, C.; MAGALHÃES, J.; COSTA, E. A Formal Model To The Routing Questions Problem In The Context Of Twitter. In: *IADIS International Conference WWW/Internet (ICWI 2011)*. [S.l.: s.n.], 2011.

SUGIYAMA, K.; KAN, M.-Y. Scholarly paper recommendation via user's recent research interests. In: *Proceedings of the 10th annual joint conference on Digital libraries*. New York, NY, USA: ACM, 2010. (JCDL '10), p. 29–38. ISBN 978-1-4503-0085-8. Disponível em: <http://doi.acm.org/10.1145/1816123.1816129>.

TORRES, R. et al. Enhancing digital libraries with techlens+. In: *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*. New York, NY, USA: ACM, 2004. (JCDL '04), p. 228–236. ISBN 1-58113-832-6. Disponível em: <http://doi.acm.org/10.1145/996350.996402>.

WANG, Y.; ZHANG, J.; VASSILEVA, J. Personalized Recommendation of Integrated Social Data across Social Networking Sites. In: *Workshop on Adaptation in the Social Semantic Web*. Hawaii, USA: [s.n.], 2010. Disponível em: <http://users.dimi.uniud.it/~antonina.dattolo/papers/2010/book/Dattolo-sasweb2010.pdf\\#page=25>.

WEBB, B. *Netflix Update: Try This at Home*. 2006. Acessado em 10/10/2013. Disponível em: <http://sifter.org/~simon/journal/20061211.html>.

Apêndice A

Artigo Publicado no 3rd SRS 2012

Improving a Recommender System Through Integration of User Profiles: a Semantic Approach

Jonathas Magalhães
Federal University of Campina Grande
Campina Grande - PB - Brazil
jonathas@copin.ufcg.edu.br

Cleyton Souza
Federal University of Campina Grande
Campina Grande - PB - Brazil
cleyton.caetano.souza@gmail.com

Priscylla Silva
Federal University of Alagoas
Maceió - AL - Brazil
pmss@ic.ufal.br

Evandro Costa
Federal University of Alagoas
Maceió - AL - Brazil
evandro@ic.ufal.br

Joseana Fechine
Federal University of Campina Grande
Campina Grande - PB - Brazil
joseana@dsc.ufcg.edu.br

ABSTRACT
The users are present in multiple social networks/virtual communities and each one can be considered as a source of information about this user. In face to this question it is important a mechanism to integrate the user profiles. Through the integration of user profiles it is possible identifier more accurately their interests analyzing other data sources that they are present, possible reducing the cold-start problem. In this context, we present a semantic approach to help integrate data from multiple sources, for the construction and maintenance of user profiles that will be used to improve the quality of a recommender system. To integrate data from multiple sources, we defined a heuristic that quantifies the importance of each data source for a given user. To validate our approach, we perform a case study, where the solution was coupled into a recommender system of papers focused in Software Engineering domain. The user profiles were built extracting their information from the Brazilian Curriculum Vitae database named CV-Lattes, an academic platform, and Linkedin, a business network. We compared the quality of the recommendation based on the profiles integrated and non-integrated. The results show the superior quality of the recommendation based on integrated profile.

Author Keywords
User profile, building user profile, integration of user profile, maintaining user profile.

ACM Classification Keywords
H.3.3 Information systems Search and Retrieval: [Information Search and Retrieval; Retrieval models]