UNIVERSIDADE DE PASSO FUNDO

INSTITUTO DE CIÊNCIAS EXATAS E GEOCIÊNCIAS

PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

RECOMENDAÇÃO DE CARREIRA DE PESQUISADORES: UMA ABORDAGEM BASEADA EM PERSONALIZAÇÃO, SIMILARIDADE DE PERFIL E REPUTAÇÃO

Gláucio Ricardo Vivian

Passo Fundo

2017

UNIVERSIDADE DE PASSO FUNDO

INSTITUTO DE CIÊNCIAS EXATAS E GEOCIÊNCIAS

PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

RECOMENDAÇÃO DE CARREIRA DE PESQUISADORES: UMA ABORDAGEM BASEADA EM PERSONALIZAÇÃO, SIMILARIDADE DE PERFIL E REPUTAÇÃO

Gláucio Ricardo Vivian

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Computação Aplicada na Universidade de Passo Fundo.

Orientador: Prof. Dr. Cristiano Roberto Cervi

Passo Fundo

2017

CIP – Catalogação na Publicação

V858r Vivian, Gláucio Ricardo
Recomendação de carreira de pesquisadores : uma abordagem baseada em personalização, similaridade de perfil e reputação / Gláucio Ricardo Vivian. – 2017.
98 f. : il. color. ; 30 cm.

Orientador: Prof. Dr. Cristiano Roberto Cervi.
Dissertação (Mestrado em Computação Aplicada) – Universidade de Passo Fundo, 2017.

1. Pesquisadores. 2. Pesquisa. 3. Sistemas de computação. I. Cervi, Cristiano Roberto, orientador. II. Título.

CDU: 004.4

Catalogação: Bibliotecário Luís Diego Dias de S. da Silva – CRB 10/2241

UPF Universidade de Passo Fundo | PPGCA Programa de Pós-Graduação em Computação Aplicada
Instituto de Ciências Exatas e Geociências - ICEG

ATA DE DEFESA DO TRABALHO DE CONCLUSÃO DE CURSO DO ACADÊMICO

GLÁUCIO RICARDO VIVIAN

Aos dois dias do mês de outubro do ano de dois mil e dezessete, às 15 horas, realizou-se, no Instituto de Ciências Exatas e Geociências, prédio B5, da Universidade de Passo Fundo, a sessão pública de defesa do Trabalho de Conclusão de Curso **"Recomendação de Carreira de Pesquisadores: Uma abordagem baseada em personalização, similaridade de perfil e reputação"**, de autoria de Gláucio Ricardo Vivian, acadêmico do Curso de Mestrado em Computação Aplicada do Programa de Pós-Graduação em Computação Aplicada – PPGCA/UPF. Segundo as informações prestadas pelo Conselho de Pós-Graduação e constantes nos arquivos da Secretaria do PPGCA, o aluno preencheu os requisitos necessários para submeter seu trabalho à avaliação. A banca examinadora foi composta pelos doutores Cristiano Roberto Cervi, Ana Carolina Bertoletti de Marchi, Alvaro Della Bona e Leandro Krug Wives. Concluídos os trabalhos de apresentação e arguição, a banca examinadora considerou o candidato APROVADO. Foi concedido o prazo de até quarenta e cinco (45) dias, conforme Regimento do PPGCA, para o acadêmico apresentar ao Conselho de Pós-Graduação o trabalho em sua redação definitiva, a fim de que sejam feitos os encaminhamentos necessários à emissão do Diploma de Mestre em Computação Aplicada. Para constar, foi lavrada a presente ata, que vai assinada pelos membros da banca examinadora e pela Coordenação do PPGCA.

Prof. Dr. Cristiano Roberto Cervi
Presidente da Banca Examinadora
(Orientador)

Profa. Dra. Ana Carolina Bertoletti de Marchi - UPF
(Avaliador Interno)

Prof. Dr. Alvaro Della Bona – UPF/PPGOdonto
(Avaliador Externo)

Prof. Dr. Leandro Krug Wives - UFRGS
(Avaliador Externo)

Prof. Dr. Rafael Rieder
Coordenador do PPGCA

AGRADECIMENTOS

Ao orientador Cristiano R. Cervi, obrigado por ter depositado a sua confiança na minha pessoa, procurei dar o melhor possível para corresponder/atender as expectativas. Obrigado por tudo, foi um grande aprendizado ser seu orientando.

Aos meus pais Maria e Ricardo, obrigado pelo apoio familiar durante o período do curso. Desculpem se em algum momento importante não pude estar presente nas suas vidas.

Aos professores do PPGCA Ana Carolina, Rafael, Carlos, Rabello, Willingthon e demais professores do programa, obrigado pelo grande aprendizado durante o período com aluno especial e regular.

Aos colegas do mestrado das turma de 2014, 2015 e 2016, obrigado pelas discussões, auxílios, esclarecimentos e amizade.

As secretárias da graduação e pós-graduação Marcele e Renata, respectivamente. Obrigado pelo auxilio durante todo o curso.

Aos professores Palazzo, Renata e Leandro do Instituto de Informática da UFRGS, obrigado pela coautoria nas publicações realizadas.

Aos professores Ana Carolina (UPF), Alvaro (Odonto-UPF) e Leandro (INF-UFRGS) e demais revisores em anonimato, obrigado pelas suas considerações sobre o trabalho. As mesmas foram indispensáveis para melhorar o trabalho.

Ao Instituto Federal Farroupilha pela oportunidade do afastamento e fomento com os programas PIIQP-BE, PIIQP-AM e PIIQP-AD.

Aos colegas de trabalho Bruno, Tiago, Laine, Aristóteles, Carlos e Paulo, obrigado pela compreensão, apoio e auxilio durante o período de afastamento para curso.

“Se eu vi mais longe, foi por estar sobre ombros de gigantes.”
(Issac Newton.)

RECOMENDAÇÃO DE CARREIRA DE PESQUISADORES: UMA ABORDAGEM BASEADA EM PERSONALIZAÇÃO, SIMILARIDADE DE PERFIL E REPUTAÇÃO

RESUMO

Os Sistemas de Recomendação tradicionais buscam auxiliar os usuários na seleção de produtos e conteúdos. Em um ambiente contemporâneo, com alta oferta de informações, esse auxílio pode ser o diferencial entre o sucesso ou fracasso. No campo da pesquisa científica, a realidade dos pesquisadores está convergindo para um aumento significativo na quantidade e diversidade de produção. Além das tradicionais publicações no formato de artigos científicos, existem inúmeras outras formas de produção que aos poucos estão sendo estimuladas. Dentre muitas, podem ser citadas: patentes, *softwares*, orientações, revisões, editoração, livros, projetos de pesquisa e rede de colaboração. Este paradigma imposto aos pesquisadores, torna mais complexa e árdua a tarefa de traçar planos estratégicos para projeção da carreira do pesquisador. Neste contexto, uma abordagem de recomendação pode apoiar os pesquisadores, buscando orientá-los com estratégias de recomendações eficazes no planejamento da sua carreira. Em outras palavras, uma abordagem de recomendação pode sugerir ao pesquisador o que, como e quando realizar determinada produção. Como resultado, se tem a possibilidade de estar realizando a atividade mais adequada e na ordem cronológica mais apropriada. O objetivo deste trabalho é propor uma abordagem de recomendação para contribuir com a gestão da carreira de pesquisadores, bem como ser um apoio a grupos de pesquisa, programas de pós-graduação e instituições, para que acompanhem a evolução da reputação científica de um pesquisador. Para tanto, foi utilizado a similaridade de perfil e reputação acadêmica como premissa de recomendação. Os experimentos foram realizados com dados de pesquisadores de produtividade do CNPq para as áreas da Ciência da Computação, Odontologia e Economia. Observou-se que a abordagem proposta tem uma boa cobertura na geração de recomendações, sobretudo para os pesquisadores com menor reputação (grupo de teste e níveis iniciais de bolsas do CNPq). Também observou-se uma ótima diversidade nos itens recomendados, o que indica existir baixa repetição de recomendações semelhantes (mesmo item).

Palavras-Chave: abordagem de recomendação, carreira científica de pesquisadores, cientometria, modelagem de perfil, reputação acadêmica.

RECOMMENDATION OF RESEARCHERS PLAN CAREER: AN APPROACH BASED IN PERSONALIZATION, PROFILE SIMILARITY AND REPUTATION

ABSTRACT

Traditional Recommender Systems seek to assist users in the selection of products and content. In a contemporaneous environment, with high offers of informations, this aid can be the differential between success or fail. In the field of scientific research, the reality of researchers are converging to significant increase in the quantity and diversity of production. I addition of the traditional publications in papers formats, there are numerous other forms of production that are gradually being stimulated. Among many, we was possible to cite: patents, softwares, advisory, reviews, publishing, books, research projects and network of co-authorship. This paradigm tax to researchers, makes it more complex and arduous the task of researchers plan career projection. In this context, the recommender approach can support the researchers, seeking to guide them with effective recommendations strategies in the planning of their career. In others words, a recommendations approach may suggest to the researcher what, how and when to perform a particular production. As a result, one has the possibility to be performing the most appropriate activity and in the most appropriate chronological order. The objective of this work is to propose a recommendation approach to contribute to the career management of researchers, as well as support for research groups, post graduate programs and institutions, to follow the evolution of a researcher's scientific reputation. For that, the profile similarity and academic reputation was used as the premise recommendation. The experiments were performed with CNPq Productivity Researchers in the areas of Computer Science, Dentistry and Economics. It was observed that the proposed approach have a good coverage in the generation of recommendations, especially for researchers with lower reputations (test groups and initial levels of CNPq). We also observed an excellent diversity in the recommended items, which indicates a low repetitions of similar recommendations (same item).

Keywords: academic reputation, profile model, recommender approach, researches plan carer, scientometric.

LISTA DE FIGURAS

Figura 1. Trabalhos sobres Sistemas de Recomendação 26
Figura 2. Modelo de referência de Rein para a Modelagem de Reputação 29
Figura 3. Ontologia proposta por Middleton et al. 37
Figura 4. Perfil do usuário construído com a rede de coautoria 39
Figura 5. Quantidade de propostas de SR por ano 40
Figura 6. Etapas do processo de *Text Mining*. 50
Figura 7. Algoritmo de Recomendação Proposto 52
Figura 8. Gráfico de Venn para as três áreas do estudo. 56
Figura 9. Evolução Quantitativa dos Elementos do Rep-Index para Ciência da Computação. 57
Figura 10. Evolução Quantitativa dos Elementos do Rep-Index para Odontologia. 57
Figura 11. Evolução Quantitativa dos Elementos do Rep-Index para Economia. 57
Figura 12. Resultados do Cálculo do Peso do Rep-Index para Ciência da Computação. 59
Figura 13. Resultados do Cálculo do Peso do Rep-Index para Odontologia. 60
Figura 14. Resultados do Cálculo do Peso do Rep-Index para Economia. 61
Figura 15. Quantidade de pesquisadores por Subárea para Ciência da Computação. . . 63
Figura 16. *Precision* e *Recall* para Ciência da Computação. 63
Figura 17. MAE e RMSE para Ciência da Computação. 64
Figura 18. Kappa e MCC para Ciência da Computação. 64
Figura 19. Quantidade de pesquisadores por Subárea para Odontologia. 65
Figura 20. *Precision* e *Recall* para Odontologia. 65
Figura 21. MAE e RMSE para Odontologia. 66
Figura 22. Kappa e MCC para Odontologia. 66
Figura 23. Quantidade de pesquisadores por Subárea para Economia. 67
Figura 24. *Precision* e *Recall* para Economia. 68
Figura 25. MAE e RMSE para Economia. 68
Figura 26. Kappa e MCC para Economia. 68
Figura 27. Grafo de Reputação e Similaridades da Ciência da Computação. 70
Figura 28. *Coverage* de Recomendações de Colaboradores e Grau de Instrução para Ciência da Computação. 71
Figura 29. *Coverage* Média e *Diversity* Média para Ciência da Computação. 71
Figura 30. Recomendações para um pesquisador do nível SR da Ciência da Computação. 72

Figura 31. Grafo para Reputação e Similaridades da Odontologia. 72
Figura 32. *Coverage* de Recomendações de Colaboradores e Grau de Instrução para Odontologia. 73
Figura 33. *Coverage* Média e *Diversity* Média para Odontologia. 73
Figura 34. Recomendações para um pesquisador do nível SR da Odontologia. 74
Figura 35. Grafo para Reputação e Similaridades da Economia. 74
Figura 36. *Coverage* de Recomendações de Colaboradores e Grau de Instrução para Economia. 75
Figura 37. *Coverage* Média e *Diversity* Média para Economia. 75
Figura 38. Recomendações para um pesquisador do nível SR da Economia. 76
Figura 39. Média para a *coverage* em cada nível de área. 77
Figura 40. Normalização da média da *coverage*. 78
Figura 41. Correlação entre a média da *coverage* e a quantidade de elementos com pesos. 78
Figura 42. Gráficos de Dispersão para Ciência da Computação. 97
Figura 43. Gráficos de Dispersão para Odontologia. 98
Figura 44. Gráficos de Dispersão para Economia. 98

LISTA DE TABELAS

Tabela 1. Exemplo de matriz utilizada na Filtragem Colaborativa 25
Tabela 2. Combinação de Características para Recomendações Hibridas 27
Tabela 3. Classificação das métricas de predição 33
Tabela 4. Técnicas usadas no perfil e *dataset* usados nos trabalhos pesquisados 41
Tabela 5. Técnicas de recomendação e avaliação usados nos trabalhos pesquisados . 41
Tabela 6. Categorias e Elementos para Rep-Model com os respectivos pesos no Rep-Index. 45
Tabela 7. Elementos adicionados ao Rep-Model 47
Tabela 8. Pesos para QUALIS por área 48
Tabela 9. Grandes áreas de pesquisa do CNPq 55
Tabela 10. Opções de Pesos do Rep-Index específicos para Ciência da Computação . . 59
Tabela 11. Opções de Pesos do Rep-Index específicos para Odontologia 60
Tabela 12. Opções de Pesos do Rep-Index específicos para Economia 61
Tabela 13. Classes empregadas nos experimentos. 62

LISTA DE SIGLAS

ACM – *Association for Computing Machinery*

AT & T – *American Telephone and Telegraph*

CNPQ – Conselho Nacional de Desenvolvimento Científico e Tecnológico

CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

CSCW – *Computer Supported Cooperative Work*

DCC – Departamento de Ciência da Computação

DBLP – *Digital Bibliography & Library Project*

DMOZ – *Directory* Mozilla

FEA – Faculdade de Economia, Admistração e Contabilidade

FSMA – Faculdade Salesiana Maria Auxiliadora

HITS – *Hyperlink-Induced Topic Search*

IADIS – *Internacional Association for the Development of the Information Society*

IEEE – *Institute of Electrical and Electronics Engineers*

INF – Instituto de Informática

INPI – *Instituto Nacional de Propriedade Intelectual*

ISBN – *International Standard Book Number*

ISSI – *Internation Society for Scientometrics and Informetrics*

ISSN – *International Standard Serial Number*

JSTOR – *Journal Storage*

MAE – *Mean Absolute Error*

MCC – *Matthew's Correlation Coefficient*

MMR – *Maximal Marginal Relevance*

NDCG – *Normalized Discounted Cumulative Gain*

NMAE – *Normalized Mean Average Error*

NRMSE – *Normalized Root of Mean Square Error*

NSTL – *National Science and Technology Library*

OCR – *Optical Character Recognition*

RMSE – *Root of Mean Square Error*

SBC – Sociedade Brasileira da Computação

SR – Sistemas de Recomendação

SR – Sênior

SVD – *Single Value Decomposition*

TF – *Term Frequency*

TF-IDF – *Term Frequency-Inverse Document Frequency*

UFMG – Universidade Federal de Minas Gerais

UFRGS – Universidade Federal do Rio Grande do Sul

UPF – Universidade de Passo Fundo

USA – *United States of American*

USP – Univesidade de São Paulo

XML – *eXtensive Markup Language*

SUMÁRIO

1 INTRODUÇÃO 21

2 FUNDAMENTAÇÃO 23

2.1 CONCEITOS, TIPOS E EVOLUÇÃO DE SISTEMAS DE RECOMENDAÇÃO 23

2.2 SISTEMAS DE RECOMENDAÇÃO PERSONALIZADOS 24

2.2.1 **Filtragem Baseada em Conteúdo** 24

2.2.2 **Filtragem Colaborativa** 25

2.2.2.1 Predição de Avaliações 26

2.2.2.2 A competição Netflix Prize 26

2.2.3 **Recomendações Hibridas** 27

2.2.4 **Outras Estratégias de Recomendações** 27

2.3 MODELAGEM DE PERFIL 28

2.4 MODELAGEM DE REPUTAÇÃO 29

2.5 MEDIDAS DE SIMILARIDADE 30

2.6 MÉTRICAS DE AVALIAÇÃO DE SISTEMAS DE RECOMENDAÇÃO 32

2.6.1 **Métricas de Previsão** 32

2.6.2 **Métricas de Conjunto** 34

2.6.3 **Métricas de Classificação (*Ranking*)** 34

2.6.4 **Métricas de Diversidade** 35

2.7 BOLSAS DE PESQUISA EM PRODUTIVIDADE DO CNPQ 36

2.8 SISTEMAS DE RECOMENDAÇÃO EXISTENTES PARA PESQUISADORES 36

2.8.1 **Resumo Comparativos dos Trabalhos Estudados** 41

2.9 REPUTAÇÃO DE PESQUISADORES 42

2.9.1 **Métricas baseadas no h-index** 42

2.9.2 **Métricas baseadas no PageRank** 44

2.9.3 **Rep-Model e Rep-Index** 45

2.10 CONCLUSÕES DO CAPÍTULO 46

3 ABORDAGEM PROPOSTA 47

3.1 ABORDAGEM PARA RECOMENDAÇÕES NÃO PERSONALIZADAS 49

3.2 ABORDAGEM PARA RECOMENDAÇÕES PERSONALIZADAS 49

3.3 ALGORITMO PARA GERAR AS RECOMENDAÇÕES 52

4 EXPERIMENTOS E RESULTADOS **55**
4.1 DADOS UTILIZADOS NOS EXPERIMENTOS 55
4.2 EXPERIMENTOS REALIZADOS 56
4.2.1 **Experimento 1 - Evolução Quantitativa dos Elementos do Rep-Index entre 2012 e 2016** 56
4.3 DETERMINAÇÃO DOS PESOS ESPECÍFICOS PARA O REP-INDEX 57
4.3.1 **Experimento 2 - Rep-Index Específico para Ciência da Computação** 59
4.3.2 **Experimento 3 - Rep-Index Específico para Odontologia** 60
4.3.3 **Experimento 4 - Rep-Index Específico para Economia** 61
4.4 AVALIAÇÃO DE SIMILARIDADE DE PERFIL PARA SUBÁREA 62
4.4.1 **Experimento 5 - Similaridade de Subárea para Ciência da Computação** 62
4.4.2 **Experimento 6 - Similaridade de Subárea para Odontologia** 65
4.4.3 **Experimento 7 - Similaridade de Subárea para Economia** 67
4.5 AVALIAÇÃO DAS RECOMENDAÇÕES 69
4.5.1 **Experimento 8 - Recomendações para Ciência da Computação** 70
4.5.2 **Experimento 9 - Recomendações para Odontologia** 72
4.5.3 **Experimento 10 - Recomendações para Economia** 74
4.6 ANÁLISE DE RESULTADOS 76

5 CONCLUSÃO **79**
5.1 OBJETIVOS 79
5.2 RESULTADOS 79
5.3 CONTRIBUIÇÕES 80
5.4 PUBLICAÇÕES 80
5.5 SOFTWARES DESENVOLVIDOS 81
5.6 SUGESTÕES DE TRABALHOS FUTUROS 81

REFERÊNCIAS **83**

APÊNDICE A – Tabela de *Stopwords* Utilizadas **91**

APÊNDICE B – Tabela de Sinonímias Utilizadas **93**

APÊNDICE C – Código Fonte LoglikelihoodDistanceMeasure.java **95**

APÊNDICE D – Gráficos de Dispersão do Cálculo dos Pesos do Rep-Index **97**

1. INTRODUÇÃO

A busca pelo desenvolvimento, análise e qualificação das ciências fez surgir as áreas de Cientometria e Bibliometria. Nos últimos tempos vivenciamos uma era onde a produção científica encontra-se em ascensão. Programas de pós-graduação, centros de pesquisa, grupos de pesquisa e agências de fomento buscam otimizar os seus recursos com o objetivo de promover o maior desenvolvimento científico e tecnológico. Neste contexto, foram propostas diversas métricas para mensurar e analisar a produção científica de pesquisadores.

Outro fato vivenciado atualmente trata-se da enorme diversidade de informações produzidas e disponíveis na *internet*. O fenômeno das redes sociais, conteúdos sob demanda, *sites* pessoais e compartilhamentos de arquivos contribuem significativamente para o crescente aumento na oferta dessas informações. De acordo com o site Statistic[1], o número de usuários conectados na internet ultrapassa 2,5 bilhões. Todas essa diversidade de informações representam um enorme desafio computacional quando o foco é encontrar as informações mais relevantes e recomendar algo interessante para o usuário.

No campo da pesquisa científica, a realidade dos pesquisadores também apresenta situação semelhante. De acordo com o Jornal Folha de São Paulo[2], no ano de 2011 foram publicados 49.664 artigos no Brasil. Isto representa 3,5 vezes mais do que os 13.846 publicados em 2001. Os dados dessa pesquisa foram coletados na base de dados aberta Scimago[1] (plataforma Scopus da editora Elsevier). Com relação à produção científica mundial, de acordo com o *site* do Banco Mundial[3], no período compreendido entre os anos 2003 e 2013, a produção total de artigos científicos e técnicos de 248 países foi de mais de 125 milhões. Nestes dados estão computados as áreas de física, biologia, química, matemática, medicina, pesquisa biomédica, engenharia e tecnologia, e ciências da terra e espaciais.

Atualmente, a computação, por meio da área de recuperação de informações, possibilita a existência de diversos portais *on-line* onde as métricas de publicações e citações são divulgadas. Um exemplo deste fato é o *site* Google Scholar[2], onde um pesquisador tem a grande maioria de suas publicações coletadas e utilizadas para computar as métricas de produção. Este fato possibilita a coleta de dados e a elaboração de estudos amplos com o intuito de descobrir as características e tendências das áreas de pesquisa.

Na literatura especializada encontra-se diversos trabalhos com o intuito de auxiliar os pesquisadores na produção de artigos científicos. Pode-se destacar a existência de diversas propostas de Sistemas de Recomendação de artigos científicos, citações e trabalhos relatados. Contudo, sabe-se que além do aumento de produtividade, existe uma tendência de diversificação na produção científica e tecnológica. Dessa forma, observou-se a inexistência de trabalhos com o objetivo principal de orientar o pesquisador a traçar estratégias diversificadas de produção.

[1] http://www.scimagojr.com - Acessado em: 18/04/2016.
[2] http://scholar.google.com.br/

Diante desse contexto, se observa o fato de que na maioria das vezes o próprio pesquisador é o responsável pelo planejamento estratégico de suas ações futuras. Nesse sentido, chegou-se ao seguinte problema de pesquisa: Qual abordagem ou solução computacional pode auxiliar o pesquisador no planejamento de sua trajetória científica para que possa aumentar sua reputação da melhor forma possível? A partir dessa constatação, se propõe uma nova abordagem de recomendação com o intuito de orientar o pesquisador sobre o que, como e quando produzir. Essas três questões devem estar em consonância com o que o pesquisador trabalha. A proposta do presente trabalho justifica-se pelo incremento da reputação científica de um pesquisador, programa de pós-graduação, instituição ou grupo de pesquisa em que o mesmo esteja inserido. Ainda, essas quatro diretrizes poderão apoiar o pesquisador a direcionar seu esforço de pesquisa para projetar sua carreira científica.

A proposta de uma abordagem de recomendação de trajetória científica vem ao encontro com a necessidade de otimizar recursos humanos e financeiros. Além disso, possibilita um incremento no desenvolvimento científico e tecnológico por meio do aumento da produtividade de forma qualificada. A aplicação prática de métricas de bibliometria/cientometria possibilita que o pesquisador realize a sua avaliação de desempenho quase que instantaneamente. Dessa forma, garante-se que o planejamento esteja alinhado com a realidade atual de alta oferta de informações, que demanda precisão e agilidade para identificar as tendências do cenário atual.

Objetivo principal deste trabalho é desenvolver uma abordagem de recomendação baseada na personalização dos dados dos pesquisadores, usando a similaridade de perfil e reputação acadêmica como premissa de recomendação. Os objetivos específicos são: i) Propor um modelo de perfil/reputação dos pesquisadores; ii) Adaptar uma medida de similaridade para comparar o perfil de pesquisadores; iii) Propor uma abordagem para realizar recomendações de atividades a partir do modelo proposto e da medida de similaridade adotada; iv) Realizar experimentos com a solução proposta e avaliar os resultados.

Esta dissertação está organizada da seguinte forma: o capítulo 2 apresenta toda fundamentação teórica sobre Sistemas de Recomendação, Modelagem de Perfil e Reputação acadêmica. O capítulo 3 relata a abordagem proposta nesta dissertação. O capítulo 4 mostra os experimentos realizados para validar a abordagem proposta. Finalmente, o capitulo 5 contem as conclusões, contribuições, resultados obtidos e sugestões de trabalhos futuros.

2. FUNDAMENTAÇÃO

Este capitulo objetiva apresentar uma visão geral sobre os Sistemas de Recomendação. Para tanto, são apresentados os seguintes itens: conceitos, tipos e evolução. Posteriormente é descrito um estudo mais focado em Sistemas de Recomendação Personalizados, onde são estudados técnicas, modelagem de perfil e reputação, medidas de similaridade e métricas de avaliação de sistemas de recomendação.

2.1 CONCEITOS, TIPOS E EVOLUÇÃO DE SISTEMAS DE RECOMENDAÇÃO

Nas relações sociais dos seres humanos, ocorrem frequentemente recomendações sobre os mais diversos temas. De um modo geral, as decisões sobre o que adquirir ou consumir são certamente influenciadas por essas sugestões. O objetivo de Sistemas de Recomendação segundo Brunialti et al.[4] é: "Um sistema de recomendação tem o objetivo de sugerir itens de forma a satisfazer sob algum aspecto as necessidades de um usuário".

O sistema Tapestry, proposto por Goldberg et al.[5] foi um dos primeiros aplicativos experimentais desenvolvidos na área. Ele foi criado por pesquisadores da empresa Xerox no *Palo Alto Research Center*. O mesmo tinha o objetivo de realizar a filtragem de *newsgroups* e apresentou pela primeira vez a técnica de filtragem colaborativa. O termo Sistemas de Recomendação foi introduzido posteriormente por Resnick et al.[6]. Nos sistemas de informação, a recomendação de informação é uma área de estudo com inúmeras pesquisas científicas ativas. Anualmente ocorre o RecSys, trata-se de um congresso internacional promovido pela ACM com o intuito de promover e divulgar as pesquisas nas áreas de Sistemas de Recomendação e afins.

Os websites de comércio eletrônico são os maiores utilizadores dos Sistemas de Recomendação. O principal propósito do seu uso está em aumentar a quantidade de vendas, através da personalização de conteúdos e indicações de itens com alta taxa de aceitação por parte do usuário. Eles também são aplicados a uma enorme diversidade de outros serviços, a destacar: conteúdo sob demanda, notícias, música, vídeo, redes sociais, pesquisa científica, livros, turismo, gastronomia, dentre outros.

Existem basicamente dois tipos de Sistemas de Recomendação com relação a aplicação: i) **Recomendações não personalizadas:** são técnicas que não consideram a especificidade do usuário. Elas são mais simples de serem implementadas por não necessitarem de informações sobre o usuário. Como exemplos pode-se citar a recomendação do item mais vendido, item mais pesquisado, item com melhor avaliação geral, último lançamento, dentre outros. ii) **Recomendações Personalizadas:** geram resultados que em geral são melhor aceitos pelos usuários, pois procuram estabelecer inicialmente um perfil para o mesmo. Neste perfil são considerados os seus gostos pessoais, preferências, avaliações, dentre outros. Dessa forma, a aceitação do item recomendado é maior, pois existe mais probabilidade de agradar e até mesmo surpreender o usuário.

Outros autores como Schafer et al.[7] baseados em exemplos existentes de *e-commerce* definiram uma taxonomia inicial para os Sistemas de Recomendação. Posteriormente Bobadilla et al.[8] apresentam uma nova proposta de taxonomia. A mesma divide os sistemas de Recomendação em duas categorias com relação ao método empregado: i) **Método baseado em Memória:** utilizam apenas a matriz de avaliações sempre atualizada com as últimas avaliações do usuário. As recomendações são executadas em memória (matriz) e usam as métricas de similaridade para obter a distância entre dois usuários, ou dois itens. ii) **Método baseado em Modelo:** usam a matriz de avaliação para aprender um modelo. Este por sua vez é então usado para fazer as predições de avaliações do usuário. Entre os modelos mais utilizados temos: Classificadores Bayesianos[9, 10], Redes Neurais[11], Lógica Fuzzy[12], Algoritmos Genéticos[13], Características Latentes[14, 15, 16], Matriz de Fatoração[17, 18], dentre outros.

2.2 SISTEMAS DE RECOMENDAÇÃO PERSONALIZADOS

O campo de pesquisa em Sistemas de Recomendação Personalizados é bastante difundido. Dessa forma, a maioria dos autores, incluindo Adomavicius[19] os classifica em três tipos: i) Recomendações baseadas em conteúdo. ii) Filtragem Colaborativa. iii) Abordagens Hibridas. Outros autores consideram mais quatro tipos de abordagens: i) Baseada em Aspectos Demográficos. ii) Baseada em Conhecimento. iii) Baseada em Utilidade. iv) Baseada em Aspectos Psicológicos. Nesta seção são apresentadas as três principais técnicas com maior relevância e posteriormente um breve relato das demais classificações.

2.2.1 Filtragem Baseada em Conteúdo

A técnica de Filtragem baseada em Conteúdo busca abstrair as informações mais relevantes de um conjunto muito maior de informações. Como o próprio nome sugere, esta técnica analisa o conteúdo das informações. Muitas vezes essas informações são coletadas em um cenário de sobrecarga, o que faz dela algo muito útil diante da quantidade de informações disponíveis principalmente na Web. A essência da Filtragem Baseada em Conteúdo está nos últimos avanços da área de Recuperação de Informação (filtragem de informação e conteúdo) e modelagem de perfil.

De acordo com Adomavicius e Tuzhilin [19], os Sistemas de Recomendação baseados em Conteúdo ($ContentBasedPRofile(c)$) são formalizados como um perfil de usuário c. Este é pré-definido com o emprego de técnicas de recuperação de informação. A função de utilidade $u(c,s)$ é definida pela equação $u(c,s) = score(ContentBasedProfile(c), Content(s))$.

Sendo $ContentBasedProfile(c)$ e $Content(s)$ definidos como vetores de pesos usando a técnica *Term Frequency–Inverse Document Frequency* (TF-IDF) e representados respectivamente por $\vec{W}_c$ e $\vec{W}_s$. A função utilidade normalmente é computada pelo emprego da medida de Similaridade do Cosseno.

Esta técnica apresenta como limitação a necessidade de grande poder computacional quando o conteúdo das informações é pouco estruturado (ex. video, som, imagens, dentre outros). Uma alternativa com menor processamento é o uso dos metadados (ex. legendas, *tags*, dentro outros) ao invés dos dados propriamente ditos. Outra questão é a superespecialização do sistema ao longo do tempo que dificulta a recomendação de algo novo ao usuário.

Segundo Silva [20], existem outras técnicas que podem ser empregadas para realizar a filtragem de informações. As principais são: i) **Busca booleana:** conjunto de palavras-chave definidas pelo usuário e conectados por operadores booleanos (*and*, *or*, *not*). ii) **Filtragem probabilística:** aplica-se o raciocínio probabilístico para estipular a probabilidade de um documento em satisfazer a necessidade de informação de um usuário. iii) **Consultas com linguagem natural:** interface de consulta onde o usuário deve elaborar as consultas em sentenças naturais.

2.2.2 Filtragem Colaborativa

A técnica de recomendação denominada Filtragem Colaborativa se baseia em utilizar a avaliação de pessoas sobre os itens de seu interesse. Para utilizar as opiniões, inicialmente o sistema agrupa os usuários pela semelhança entre seus perfis. Posteriormente se recomenda itens com maior avaliação dos usuários de perfis semelhantes. Dessa forma, esta técnica permite a geração de recomendações personalizadas. As avaliações dos itens são geralmente realizadas pelo usuário utilizando-se uma escala de 1 a 5. Quanto maior o valor, mais o usuário está satisfeito com o item. Outras escalas também podem ser empregadas, mas o seu uso é menos frequente.

Em geral a filtragem colaborativa é implementada utilizando-se uma matriz, onde os itens são armazenados nas colunas e os usuários nas linhas. Dessa forma, os maiores problemas são a grande esparsidade das informações armazenadas e o alto custo computacional (complexidade assintótica $O(n^2)$) para o processamento e armazenamento. Existem trabalhos que utilizam técnicas de fatoração de matrizes (SVD)[17, 18] e SVD para *Big Data*[21] para criar um modelo que permite a geração de recomendações com baixo custo computacional. Também existe o problema conhecido como do primeiro avaliador: um item novo acrescentado ao conjunto nunca é recomendado enquanto estiver sem avaliações dos usuários. Na Tabela 1 pode-se visualizar um exemplo de matriz utilizada na Filtragem Colaborativa.

Tabela 1. Exemplo de matriz utilizada na Filtragem Colaborativa.

	Item(i_1)	**Item(i_2)**	**Item(i_3)**	**Item(i_4)**	**Item(i_5)**
Usuário(u_1)	1 ⋆	2⋆⋆	5⋆⋆⋆⋆⋆		
Usuário(u_2)	2⋆⋆		3⋆⋆⋆	4⋆⋆⋆⋆	
Usuário(u_3)		5⋆⋆⋆⋆⋆	2⋆⋆	1⋆	
Usuário(u_4)	2⋆⋆		2⋆⋆	3⋆⋆⋆	
Usuário(u_5)	3⋆⋆⋆	5⋆⋆⋆⋆⋆	2⋆⋆	2⋆⋆	

Nesta Tabela estão representados 5 usuários e 5 itens. Em um cenário real, as dimensões dessa matriz são maiores e com mais esparsidade, ou seja, muitos espaços vazios. O item i_5 trata-se de um exemplo do problema do primeiro avaliador. Por não existir avaliações para este item, é impossível para a Filtragem Colaborativa gerar uma recomendação do mesmo.

2.2.2.1 Predição de Avaliações

A partir da matriz de usuários e itens, geralmente emprega-se alguma medida de similaridade para agrupar os perfis semelhantes. Ainda existe a possibilidade de utilização de outras técnicas como agrupamento(*Clustering*), Redes Bayesianas, lógica *fuzzy*, aprendizado de máquina, mineração de dados, dentre outros. Após identificar os perfis semelhantes, ocorre a predição. Trata-se de uma etapa que busca prever qual seria a avaliação do usuário para os seus itens sem avaliação. A predição geralmente é computada com a técnica dos melhores vizinhos, vide equação 1.

$$p_{a,i} = \bar{r}_a + \frac{\sum_{u=1}^{h} (r_{u,i} - \bar{r}_u) * w_{a,u}}{\sum_{u=1}^{h} |w_{a,u}|}. \text{(Adaptado de Silva[20])} \quad (1)$$

Onde, $p_{a,i}$ é a predição do usuário a para o item i, $\bar{r}_a$ é a média de todas as avaliações do usuário a, h é a quantidade considerada dos melhores vizinhos, $r_{u,i}$ é a avaliação do usuário u para o item i, $\bar{r}_u$ é a média de todas as avaliações do usuário u, $w_{a,u}$ é a correlação do usuário a com o usuário u.

Finalmente ocorre um ranqueamento descendente nas predições, onde são recomendados os n primeiros itens com maior predição. A maioria dos experimentos de avaliação de Sistemas de Recomendação busca variar o valor de n a fim de identificar a situação que gera o menor erro possível nas recomendações.

2.2.2.2 A competição Netflix Prize

De acordo com Koren et al.[22] muitos avanços na filtragem colaborativa foram alcançados devido a competição NetFlix Prize[23]. Um dos principais motivos para os avanços foi o acesso a um enorme conjunto de dados que permitiu a comunidade científica realizar seus experimentos. Em 2009 o time Bellkor's ganhou 1 milhão de dólares como prêmio da competição por ser a primeira equipe a conseguir aumentar em 10% a acurácia (RMSE) do sistema de recomendação. Esse feito foi conseguido através da combinação de diversas técnicas algorítmicas, procurando-se beneficiar das características desejáveis e eliminar as indesejáveis. Para maiores detalhes sobre as técnicas empregadas vide [24, 25, 26].

Bobadilla et al. [8] analisaram a quantidade de trabalhos científicos sobres Sistemas de Recomendação entre os anos de 1989 e 2013. Na Figura 1 pode-se visualizar a produção analisada.

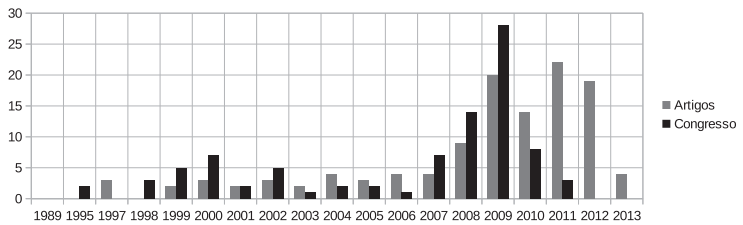

Figura 1. Trabalhos sobres Sistemas de Recomendação (Adaptado de Bobadilla et al.[8]).

Observa-se que no ano de 2009 houve um aumento significativo na quantidade de periódicos e congressos. Foi justamente neste ano que ocorreu a final no NetFlix Prize, momento em que a comunidade provavelmente se voltou intensamente para as pesquisas com a abordagem vencedora. Posteriormente ocorreu uma breve redução das pesquisas, mas sem retornar a patamar anterior.

2.2.3 Recomendações Hibridas

Com o intuito de aumentar a qualidade das recomendações para os usuários, diversos autores propuseram a combinação do resultado de técnicas a fim de se beneficiar das características desejáveis de cada uma. Essa nova abordagem ficou conhecida por Recomendação Hibrida. Em geral utiliza-se duas ou mais técnicas de recomendação, podendo ser inclusive combinando recomendações personalizadas e não personalizadas ou até mesmo dois resultados diferentes gerados com a mesma técnica mas com parâmetros distintos de configuração.

A forma mais simples de fazer uma Recomendação Hibrida é unindo a Filtragem Colaborativa e a Filtragem Baseada em Conteúdo. Essa abordagem procura utilizar as qualidades boas da Filtragem Colaborativa e da Filtragem Baseada em Conteúdo. Na Tabela 2 pode-se visualizar a combinação das características mencionadas.

Tabela 2. Combinação de Características para Recomendações Hibridas.

Filtragem Colaborativa	**Filtragem Baseada em Conteúdo**
Descoberta de novos relacionamentos entre usuários	Bons resultados para os usuários incomuns
Recomendação de itens diretamente relacionados	Precisão independente do número de usuários

2.2.4 Outras Estratégias de Recomendações

Além das estratégias tradicionais citadas anteriormente, alguns autores afirmam a existência de outras abordagens específicas para sistemas de recomendação. De acordo com Silva[20] elas são:

- **Filtragem Baseada em Aspectos Demográficos:** Nesta abordagem o perfil do usuário é definido com base em caraterísticas demográficas do mesmo, tais como: idade, gênero, dentre outros. Dessa forma, as recomendações são geradas para grupos com as características muito próximas.

- **Filtragem Baseada em Conhecimento:** Nesta estratégia, o perfil do usuário é armazenado em conhecimento funcional estruturado e as recomendações são geradas utilizando-se as inferências das preferências do usuário.

- **Filtragem Baseada em Utilidade:** procura-se determinar o valor de utilidade dos itens com base em conhecimento funcional. Os itens mais úteis são recomendados.

- **Filtragem Baseada em Aspectos Psicológicos:** Essa abordagem procura caracterizar o perfil do usuário com adição de aspectos psicológicos da sua personalidade. Características como

emoções, personalidade, identidade são considerados no contexto de um grupo social. Sabe-se que as pessoas interagem mais naturalmente com outras de personalidades semelhantes. Alguns sistemas chegam ao ponto de considerar até mesmo o estado emocional no instante de geração das recomendações. Em geral esta técnica é aplicada em conjunto com a Filtragem Colaborativa/Conteúdo a fim de se obter uma Recomendação Hibrida.

2.3 MODELAGEM DE PERFIL

A modelagem do perfil de um usuário é a parte indispensável do processo de construção de um sistema de recomendação personalizado. A partir desta etapa, são determinados quais atributos do usuário são necessários para o sistema armazenar as características e comportamentos do mesmo. Em outras palavras, pode-se afirmar que o perfil de usuário representa a identidade interna do mesmo no Sistema de Recomendação.

De acordo com Barth[27], o armazenamento das informações do perfil é fundamental para garantir a sua disponibilidade futura. Em geral o mesmo é realizado com uma das seguintes técnicas: lista de compras, histórico de navegação e caixa postal. Outras possibilidades mais avançadas são o emprego de vetores ou árvores de decisão. Existe até mesmo a possibilidade do uso do paradigma de programação em lógica (Prolog) e cláusulas de primeira ordem. Segundo Cazella et al.[28], no campo dos arquivos estruturados (XML) também pode-se empregar uma Ontologia XML para a representação de um perfil.

Ainda segundo Barth[27], o perfil inicial do usuário também é importante para garantir uma primeira experiência satisfatória com o sistema. Existem basicamente quatro técnicas para construir um perfil inicial: i) **Perfil inicial vazio:** a partir das ações do usuário, o sistema realiza a coleta de dados para definir o seu perfil. ii) **Manual:** neste caso o usuário deve responder um questionário ao sistema, o mesmo é utilizado para definir as características do perfil. iii) **Esteriótipos:** o usuário escolhe um esteriótipo inicial que melhor se aproxima do seu perfil real. Essa estratégia possibilita um nível inicial satisfatório de acerto nas recomendações. iv) **Conjunto de treinamento:** utiliza uma coleção de exemplos de interação do usuário com o sistema computacional.

Com exceção do perfil inicial definido por esteriótipos, a grande maioria dos sistemas utilizam técnicas para aprendizado de perfil como forma de construir o mesmo. Essas técnicas são classificadas em três tipos: i) **Extração de informação estruturada:** Utiliza-se técnicas de mineração de texto (*Stopwords*, *Stemming*, TF-IDF, dentre outras) para obter os dados estruturados a partir de informações em formato não estruturado. ii) **Agrupamento:** técnicas de agrupamento (*Clustering*) que procuram agrupar as informações mais próximas em *clusters*. iii) **Classificação:** empregam-se técnicas de classificação como árvores de decisão e aprendizado de máquina[27].

Nos sistemas de recomendação onde existe a necessidade de atualizar o perfil do usuário a fim de considerar mudanças em suas preferências com o passar do tempo, é possível utilizar a técnica de realimentação (*Feedback*). A mesma pode ser: i) **Explícita:** o usuário avalia a qualidade das recomendações de acordo com uma das seguintes formas: gosto/não gosto, ranqueamento na escala

de Likert (ordenação de acordo com a relevância) e comentários textuais. ii) **Implícita:** o sistema busca inferir as preferências do usuário através da análise do monitoramento de suas ações. Em geral são considerados as páginas visitadas, itens comprados, tempo de atividade em cada página, e outras manifestações de interesse como impressão ou salvar o documento. iii) **Híbridas:** Combinação da abordagem explicita com a implícita a fim de obter melhores resultados na atualização do perfil com a realimentação[27].

No trabalho de Ying et al. [29] são descritas diversas técnicas para o desenvolvimento de um perfil de forma prática utilizando a linguagem de programação Java. Os autores detalham técnicas especificas para capturar, representar, armazenar e utilizar os perfis em Sistemas de Recomendação.

2.4 MODELAGEM DE REPUTAÇÃO

A modelagem de reputação atua como um novo ponto de vista em relação ao usuário, enquanto o perfil representa no mundo virtual as características do usuário sob o seu próprio ponto de vista, considerando os seus interesses e preferências. A modelagem de reputação define as suas características sobre o ponto de vista da comunidade onde o mesmo está inserido. Segundo Resnick et al. [30], a reputação pode ser considerada como uma coleção de comentários sobre o comportamento do usuário. Dessa forma, os modelos produzidos pelas duas técnicas podem até mesmo a chegar ao ponto de divergirem completamente em relação ao mesmo usuário[27], pois são pontos de vistas de diferentes focos.

Cazella et al.[28] afirmam que em uma rede de usuários a reputação encoraja os usuários a comportamentos confiáveis, procurando inibir as ações imorais e desonestas. A reputação pode ser definida com base em um modelo de referencia proposto por Rein[31] composto por 10 itens: conhecimento, experiência, credenciais, endosso, contribuidor, conexões, sinais, *feedback*, contexto, valores sociais. Na Figura 2 pode-se visualizar o modelo de referencia de Rein.

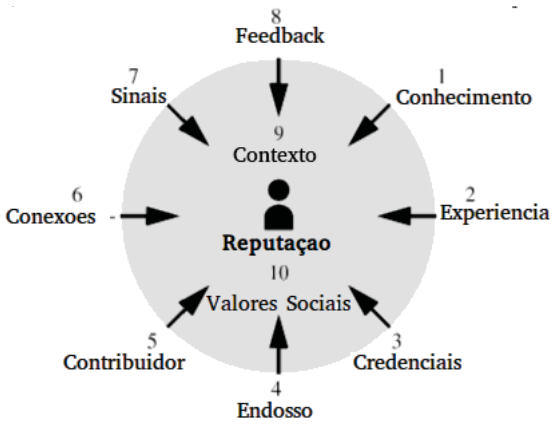

Figura 2. Modelo de referência de Rein para a Modelagem de Reputação (Adaptado de Rein[31]).

O modelo de Rein apresenta as funcionalidades e comportamentos essenciais do ser humano para representar a reputação do usuário de forma explicita. A seguir detalha-se cada um dos itens presentes no modelo.

1. **Conhecimento:** A reputação de um usuário é afetada pelo seu nível de *expertise* ou conhecimento em determinada área.

2. **Experiência:** A reputação de um usuário é afetada pelo seu nível de treinamento e experiência em determinada área.

3. **Credenciais:** A reputação de um usuário é afetada pelas seus títulos, posições na hierarquia e formação em determinada área.

4. **Endosso:** Referências de um usuário com alta reputação pode aumentar a reputação de outro usuário.

5. **Contribuidor:** A reputação de um usuário é diretamente influenciada pelas reputação de seus contribuidores e colaboradores.

6. **Conexões:** Ligações entre um usuário desconhecido de baixa reputação e um com alta reputação pode auxiliar na sua própria reputação.

7. **Sinais:** A reputação de um usuário é afetada pelo cumprimento de algo que anteriormente foi manifestado na forma de um sinal.

8. ***Feedback***: A reputação de um usuário deve ser ajustada de acordo com o resultado de avaliações da comunidade onde o mesmo está inserido.

9. **Contexto:** A reputação de um usuário é interpretada de acordo com o contexto social ou ambiente da sua comunidade. O contexto pode definir importantes elementos da comunidade como idioma, usos e costumes, dentre outros.

10. **Valores Sociais:** Sistemas de Reputação são como Sistemas Sociais que em grande parte são definidos pelos valores da comunidade. Os graus de conhecimento, experiência e credenciais afetam a reputação dependendo do contexto social da comunidade onde estes estão inseridos.

2.5 MEDIDAS DE SIMILARIDADE

Nos Sistemas de Recomendação Personalizados é fundamental encontrar os perfis similares para gerar as recomendações. Existem diversas técnicas empregadas para isso, os coeficientes de similaridade são os mais utilizados. Nesta seção são apresentadas as mais difundidas: Coeficiente de Correlação de Pearson (Eq. 2), Correlação de Spearman (Eq. 3), Distância Euclidiana (Eq. 4), Medida dos Cossenos (Eq. 5), Correlação de Tanimoto (Eq. 6) e log-Likelihood (Eq. 7).

A Correlação de Pearson busca medir o relacionamento linear entre duas variáveis. Os seus valores vão de -1 a $+1$, sendo -1 a situação onde existe o máximo de correlação negativa, 0 não

existe correlação, e +1 existe o máximo de correlação positiva.

$$w_{a,u} = \frac{\sum_{i=1}^{m}(r_{a,i} - \bar{r_a})(r_{u,i} - \bar{r_u})}{\sqrt{\sum_{i=1}^{m}(r_{a,i} - \bar{r_a})^2 \sum_{i=1}^{m}(r_{u,i} - \bar{r_u})^2}}.\text{(Adaptado de Silva[20])} \quad (2)$$

A Correlação de Spearman é o equivalente não paramétrico para a correlação de Pearson. Ela resume o relacionamento linear entre o *ranking* de duas variáveis. Dessa forma, ela é computada usando os *rankings* dos dados em um determinado momento.

$$w_{a,u} = 1 - \frac{6\sum_{i=1}^{m}(p_{a,i} - p_{u,i})^2}{m(m^2 - 1)}.\text{(Adaptado de Spearman[32])} \quad (3)$$

A distância Euclidiana se baseia na distância entre dois pontos no plano cartesiano de duas ou mais dimensões. A origem está diretamente relacionada na aplicação do teorema de Pitágoras no espaço Euclidiano a fim de obter um espaço métrico. Essa medida geralmente é utilizada na técnica de Filtragem Baseada em Conteúdo.

$$w_{a,u} = \sqrt{\sum_{i=1}^{m}(r_{a,i} - r_{u,i})^2}.\text{(Adaptado de Silva[20])} \quad (4)$$

A Medida dos Cossenos busca medir o cosseno dos ângulos existentes entre dois vetores que representam respectivamente os perfis dos usuários no Sistema de Recomendação.

$$w_{a,u} = \frac{\sum_{i=1}^{m}(r_{a,i} * r_{u,i})}{\sqrt{\sum_{i=1}^{m}(r_{a,i})^2}\sqrt{\sum_{i=1}^{m}(r_{u,i})^2}}.\text{(Adaptado de Silva[20])} \quad (5)$$

O coeficiente de correlação de Tanimoto (também chamada de Similaridade de Jaccard) é um método utilizado na teoria de conjuntos para comparar a similaridade de dois conjuntos de dados que armazenam o perfil do usuário em Sistemas de Recomendação.

$$w_{a,u} = \frac{|m_a \cap m_u|}{(|m_a| + |m_u|) - (|m_a \cap m_u|)}.\text{(Adaptado de Silva[20])} \quad (6)$$

A medida de Similaridade Log-Likelihood é um método semelhante a correlação de Tanimoto, onde adicionalmente se calcula o quão provável é a sobreposição dos elementos dos conjuntos. Em outras palavras, esta métrica leva em consideração o quão distinta é a interseção dos itens avaliados.

$$w_a = \sum_i \frac{max(r_{a,i} - d, 0)}{2^{(i-1)/(\alpha-1)}}.\text{(Adaptado de Silva[20])} \quad (7)$$

Onde, a e u são os dois usuários comparados, i é um dado item, m é o número total de itens, $w_{a,u}$ é a correlação do usuário a com o usuário u, $r_{a,i}$ é a avaliação do usuário a para o item i, $r_{u,i}$ é a avaliação do usuário u para o item i, $p_{a,i}$ é a posição no ranking da avaliação do usuário

a para o item i, $p_{u,i}$ é a posição no ranking da avaliação do usuário u para o item i, $\bar{r}_a$ é a média de todas as avaliações do usuário a, $\bar{r}_u$ é a média de todas as avaliações do usuário u, m_a é quantidade de elemento do usuário a,m_u é quantidade de elemento do usuário u, d é a classificação padrão, α é a meia-vida do item na lista, ou seja, é a situação aonde existe 50% de chance que o usuário irá visualizar esse item.

Em um estudo realizado por Silva [20], comparou-se os erros (RMSE) das medidas de similaridades mencionadas utilizando a biblioteca Apache Mahout. Inicialmente foram utilizadas as medidas de similaridades mencionadas para comparar os perfis, posteriormente computou-se as predições e finalmente calculou-se o erro através da métrica. Como resultados LogLikeliHood apresentou o menor erro (0,242), seguido por Tanimoto (0,245), Distância Euclidiana (0,268), Correlação de Spearman (0,384) e Correlação de Pearson (0,425). A medida da similaridade do Cosseno não entrou no experimento, pois a biblioteca não possui a implementação da Filtragem Baseada em Conteúdo.

2.6 MÉTRICAS DE AVALIAÇÃO DE SISTEMAS DE RECOMENDAÇÃO

Após a geração das recomendações, é fundamental que se mensure a qualidade das mesmas para os usuários. A medição possibilita a realização de ajustes no sistema com o objetivo de diminuir os erros e consequentemente melhorar a experiência do usuário com as recomendações. Nesta seção são apresentadas as técnicas empregadas na avaliação da recomendações.

Na literatura encontrou-se diversas propostas para avaliar as recomendações. A maioria dos autores McNee et al.[33], Ekstrand et al.[34], Wang e Blei[35], Beel et al.[36] divide as avaliações em duas categorias: i) *online*: ocorrem quando o sistema de recomendação está em funcionamento e com usuários avaliando as recomendações geradas. As métricas são realizadas entre as predições do sistema e as avaliações das recomendações pelos usuários. ii) *offline*: Esta classificação consiste em avaliar os sistemas sem a interação do usuário. Dessa forma o conjunto total de dados é dividido em duas ou mais partes (*cross-validation*, *Percentage Split*, dentre outros), algumas são utilizadas para realizar as predições do sistema e outras para comparar com os resultados gerados pelo sistema e dessa forma obter as medições.

Segundo Gunawardana e Shani[37], as métricas de predição são classificadas em: i) métricas de previsão. ii) métricas de conjunto. iii) métricas de lista de classificação (*ranking*). iV) métricas de diversidade. Na Tabela 3 pode-se visualizar as métricas mais utilizadas em cada classificação apresentada.

2.6.1 Métricas de Previsão

As métricas de previsão no contexto de sistemas de recomendação são empregadas em situação onde o usuário avalia as recomendações em uma escala de n valores possíveis. Geralmente emprega-se o inteiro 1 como menor aceitação e 5 como maior aceitação. Outras escalas com intervalos diferentes também podem ser empregadas.

Tabela 3. Classificação das métricas de predição. (Adaptado de Silva[20])

Classificação	Métrica
Previsão	MAE (*Mean Absolute Error*)
Previsão	RMSE (*Root of Mean Square Error*)
Previsão	NMAE (*Normalized Mean Average Error*)
Previsão	NRMSE (*Normalized Root of Mean Square Error*)
Conjunto	Precisão (*Precision*)
Conjunto	Revocação (*Recall*)
Conjunto	Cobertura (*Coverage*)
Conjunto	Média harmônica da Precisão e Revocação (*F-Measure*)
Classificação (*Ranking*)	Meia-vida (*half-life*)
Classificação (*Ranking*)	"Desconto de ganho acumulado"
Diversidade	Diversidade
Diversidade	Novidade de itens recomendados

A métrica MAE (*Mean Absolute Error*) mede a média absoluta de desvio na pontuação predita em relação a pontuação atribuída pelo usuário. Na equação 8 pode-se visualizar a métrica MAE.

$$MAE(T) = \frac{\sum_{(u,i)\in T} |\hat{r}_{ui} - r_{ui}|}{N}.(\text{Adaptado de Avazpour et al.[38]}) \quad (8)$$

A métrica RMSE (*Root of Mean Square Error*) mede os quadrados dos desvios. Dessa forma evita que os desvios positivos e negativos se anulem. Na equação 9 pode-se visualizar a métrica RMSE.

$$RMSE(T) = \sqrt{\frac{\sum_{(u,i)\in T} (\hat{r}_{ui} - r_{ui})^2}{N}}.(\text{Adaptado de Avazpour et al.[38]}) \quad (9)$$

Onde, r_{ui} é a atual avaliação do usuário u para o item i, e $\hat{r}_{ui}$ é o valor predito pelo sistema. O valor residual entre duas avaliações é: $(\hat{r}_{ui} - r_{ui})$.

Pode-se afirmar que os valores residuais individuais da métrica MAE são igualmente ponderados, enquanto que em RMSE os grandes erros são mais penalizados que os pequenos. Isso ocorre porque os erros são elevados ao quadrado antes da média ser computada. Dessa forma, a métrica RMSE é mais sensível que a MAE. Normalmente o valor de RMSE é maior que MAE. Baixos valores para MAE e RMSE indicam que o sistema apesenta grande acerto nas recomendações. Caso ambas as métricas sejam iguais, todos os erros apresentam a mesma magnitude.

As duas métricas podem ser normalizadas de acordo com a escala do intervalo de pontuação. Nas equações 10 e 11 pode-se visualizar a normalização.

$$NMAE(T) = \frac{MAE(T)}{r_{max} - r_{min}}.(\text{Adaptado de Avazpour et al.[38]}) \quad (10)$$

$$NRMSE(T) = \frac{RMSE(T)}{r_{max} - r_{min}}.(\text{Adaptado de Avazpour et al.[38]}) \quad (11)$$

Onde r_{max} e r_{min} são respectivamente a máxima avaliação (geralmente 5) e a mínima avaliação (geralmente 1) possíveis.

2.6.2 Métricas de Conjunto

As métricas de conjunto se baseiam na teoria dos conjuntos. Elas são mais empregadas em situações onde sabe-se apenas se o usuário gostou (*like*) ou não gostou (*dislike*) da recomendação.

A métrica Precisão (*Precision*) indica o percentual de itens recomendados relevantes do número total de itens recomendados.

$$Precision = \frac{|R_g \cap R_r|}{R_r}.\text{(Adaptado de Huang et al.[39])} \tag{12}$$

A métrica Revocação (*Recall*) indica o percentual de itens recomendados relevantes em relação ao número de itens relevantes.

$$Recall = \frac{|R_g \cap R_r|}{R_g}.\text{(Adaptado de Huang et al.[39])} \tag{13}$$

Onde, R_g é o conjunto total original, R_r é o conjunto de recomendações e $R_g \cap R_r$ é conjunto de recomendações corretas.

A métrica Cobertura (*Coverage*) verifica a possibilidade de um sistema em gerar recomendações. Em dados momentos, um sistema de recomendação pode não possuir informações suficientes para gerar uma recomendação, isso ocorre principalmente quando são incluídos novos usuários ou itens. A Cobertura refere-se a porção de informações disponíveis do total com a qual pode-se gerar recomendações. Na equação 14 pode-se visualizar o seu cálculo.

$$Coverage = \frac{1}{\#U}\sum_{u \in U}\left(100\frac{\#C_u}{\#D_u}\right).\text{(Adaptado de Bobadilla et al.[8])} \tag{14}$$

Onde, $C_u = \{i \in I | r_{u,i} = \bullet \wedge K_{u,i} \neq \oslash\}$, $D_u = \{i \in I | r_{u,i} = \bullet\}$, U é o conjunto de usuários do sistema de recomendação, I é o conjunto de itens do sistema de recomendação, $r_{u,i}$ a avaliação do usuário u para o item i, $p_{u,i}$ a predição do item i para o usuário u, $\bullet$ a falta de avaliação ($r_{u,i} = \bullet$ significa um item i do usuário u que não possui avaliação.

A métrica da média harmônica entre a Precisão e a Revocação, também chamada de F-Measure é uma forma de combinar ambas em um único valor a fim de simplificar as comparações. Na equação 15 pode-se visualizar o seu cálculo.

$$F - Measure = 2\left(\frac{\text{Precisão} * \text{Revocação}}{\text{Precisão} + \text{Revocação}}\right).\text{(Adaptado de Huang et al.[39])} \tag{15}$$

2.6.3 Métricas de Classificação (*Ranking*)

Em determinadas situações, os Sistemas de Recomendação podem recomendar um grande número de itens ao usuário. Sabe-se contudo, que o maior interesse do usuário está nos itens iniciais

da lista, dessa forma, as medidas de *ranking* buscam apresentar um modo de avaliar os erros considerando que os itens iniciais irrelevantes são mais ponderados que os itens irrelevantes do final da lista de recomendações.

A métrica *half-life*, em tradução livre "meia-vida", pressupõem que ocorre uma redução exponencial de interesse do usuário entre o primeiro item recomendado e o segundo, entre o segundo e terceiro, e assim até o final da lista. Na equação 16 pode-se visualizar o cálculo da métrica Decadência.

$$\text{Decadência} = \frac{1}{\#U} \sum_{u \in U} \sum_{i=1}^{N} \frac{max(r_{u,p_i} - d, 0)}{2^{(i-1)/(\alpha-1)}}.\text{(Adaptado de Silva[20])} \tag{16}$$

A métrica do "Desconto de ganho acumulado", também considera a mesma situação da métrica "meia-vida". Só que neste caso, a decadência deixa de ser exponencial e passa a ser tratada como logarítmica. Na equação 17 pode-se visualizar o cálculo da métrica Desconto.

$$\text{Desconto}^k = \frac{1}{\#U} \sum_{u \in U} (r - u, p1 + \sum_{i=2}^{k} \frac{r_{u,p_i}}{log_2(i)}).\text{(Adaptado de Silva[20])} \tag{17}$$

Onde, k é a avaliação do item avaliado, d é a avaliação padrão, α é o número do item na lista de tal forma que há uma chance de 50% que o usuário irá rever esse item, r_{u,p_i} representa a avaliação verdadeira que o usuário u deu para o item $i(p_i)$.

2.6.4 Métricas de Diversidade

A Diversidade indica o quanto um item recomendado é diferente dos demais itens recomendados. Ou seja, ela realmente mensura a diversidade dos itens recomendados. Na equação 18 pode-se visualizar o cálculo da Diversidade.

$$\text{Diversidade}_{Z_u} = \frac{1}{\#Z_u(\#Z_u - 1)} \sum_{i \in Z_u} \sum_{j \in Z_u, j \neq i} [1 - sim(i,j)].\text{(Adaptado de Silva[20])} \tag{18}$$

A Novidade é uma métrica que procura mensurar o quanto novo é para o usuário um item da lista de recomendações. Na equação 19 pode-se visualizar o cálculo da Novidade.

$$\text{Novidade}_i = \frac{1}{\#Z_u - 1} \sum_{j \in Z_u} [1 - sim(i,j)], i \in Z_u.\text{(Adaptado de Silva[20])} \tag{19}$$

Onde, $sim(i,j)$ é a similaridade de item-item. Segundo Silva[20], não existe um consenso com relação a definição das métricas de diversidade e novidade.

2.7 BOLSAS DE PESQUISA EM PRODUTIVIDADE DO CNPQ

No Brasil as bolsas de pesquisa em produtividade são distribuídas pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq[1]) aos pesquisidores que se destacam entre seus pares. O objetivo principal é valorizar a produção científica por meio do fomento, recursos de bancada e mensalidades. As bolsas são divididas em duas categorias: **Pesquisador 1)** exigência de no mínimo 8 anos após o término do doutorado; **Pesquisador 2)** exigência de no mínimo 3 anos após o enceramento do doutorado.

Para o Pesquisador 1 existem 4 diferentes níveis (A, B, C e D) onde sua produtividade é avaliada considerando os últimos 10 anos. A diferença entre os níveis está nos pesos dos critérios que são definidos de acordo com as características desejadas de seus pesquisadores. Sendo o nível D o menos rigorosos e o nível A o mais exigente. Para o pesquisador 2 avalia-se a produtividade dos últimos 5 anos. A duração da bolsa é de 60 messes para o nível 1A; 48 meses para os níveis 1B, 1C e 1D; e 36 messes para o nível 2.

Além dos tradicionais níveis de bolsas, existe o nível sênior (SR) destinado a pesquisadores que se destacaram entre seus pares como líder e paradigma na sua área de atuação. A mesma é destinada aos que permaneceram por pelo menos 15 anos nos níveis 1A ou 1B; consecutivos ou não. A bolsa SR tem duração igual ao nível 1A de 60 meses.

2.8 SISTEMAS DE RECOMENDAÇÃO EXISTENTES PARA PESQUISADORES

Esta seção objetiva apresentar os resultados de uma revisão da área de sistemas de recomendação existentes no contexto de pesquisadores. Para realizar a pesquisa, utilizou-se as seguintes bases: IEEE[2], ACM[3], Springer[4], Google Scholar[5], Periódicos Capes[6], Elsevier[7], Scielo[8]. Foram encontrados trabalhos sobre os temas: recomendação de artigos científicos, recomendação de trabalhos relatados e recomendação de citações.

O *GroupLens Research Project* da Universidade de Minnesota seguramente é um dos pioneiros nas pesquisas sobre sistemas de recomendação em diversas áreas. No trabalho de McNee et al.[33] o *GroupLens* avaliou o uso da tradicional filtragem colaborativa para recomendação de artigos científicos armazenadas em um grafo de citações. O principal objetivo do trabalho foi testar a habilidade da filtragem colaborativa para recomendar citações. Ao todo foram testados seis diferentes algoritmos em um *dataset* com 186 mil artigos do ResearchIndex. Foram realizadas avaliações tanto

[1] http://cnpq.br/bolsas-no-brasil
[2] http://ieeexplore.ieee.org
[3] http://dl.acm.org
[4] http://www.springer.com
[5] http://scholar.google.com.br
[6] http://periodicos.capes.gov.br
[7] http://www.elsevier.com
[8] http://www.scielo.org

online como *offline*. Nos testes *offline*, os melhores resultados foram obtidos com os algoritmos do Google[9] para encontrar trabalhos relatados. Ficando em segundo lugar a técnica Naïve Bayes[9].

O artigo de Middleton et al.[40] introduziu a modelagem de perfil de um sistema de recomendação com o uso de ontologias para uma abordagem híbrida. A Ontologia representa o relacionamento entre 27 classes e os tópicos dos artigos. Em um segundo momento as classes foram aumentadas para 32. As classes foram baseadas na classificação de áreas de pesquisa da Ciência da Computação obtidas diretamente no diretório do projeto dmoz[10]. A representação dos artigos foi realizada utilizando-se vetores de termos, computados com a técnica *Term Frequency* (TF) e divididos pelo número total de termos, representando a frequência normalizada com que as palavras aparecem no artigo. Na Figura 3 pode-se visualizar a Ontologia proposta.

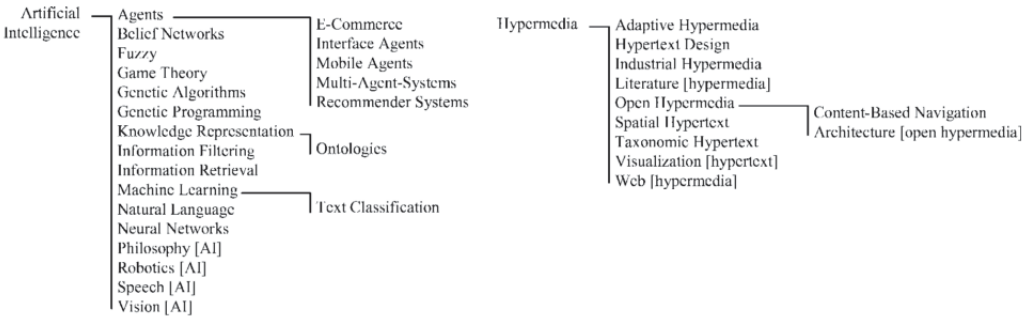

Figura 3. Ontologia proposta por Middleton et al.(Adaptado de Middleton et al.[40]).

No trabalho foram testados dois sistemas experimentais (Quickstep e Foxtrot), além disso foram conduzidos três experimentos avaliativos. Os experimentos realizados demostraram que a abordagem proposta supera os sistemas apresentados na literatura. Ao final se conclui que as ontologias apresentam as seguintes vantagens: i) perfis mais completos, incluindo tópicos de interesse que não são diretamente vistos. ii) aumento na acurácia dos perfis e por consequência nas recomendações.

No trabalho de Ekstrand et al.[34] foram explorados diversos métodos (177 algoritmos em 5 famílias) para recomendação baseada em filtragem colaborativa e filtragem baseada em conteúdo. O perfil do usuário foi construído com base em uma abordagem alternativa, foram utilizadas as próprias citações da Web para computar as similaridades e posteriormente gerar as recomendações. As medidas de influência foram realizadas com os algoritmos HITS[41] e *PageRank*[42]. Inicialmente foram realizados testes *offline* para identificação das melhores propostas, posteriormente se conduziu uma avaliação *online* com a participação de pesquisadores. Os testes *offline* foram realizados com mais de 250 mil artigos da ACM *Digital Library*, coletados em abril de 2010. Ao final, os testes demostraram que os usuários preferiam as recomendações realizadas com a técnica de filtragem colaborativa ao invés das técnicas híbridas de recomendação.

No trabalho de Zhang e Li[43], foi proposto um sistema de recomendação de artigos com o modelo de perfil desenvolvido com o conceito de árvores para ultrapassar os inconvenientes do modelo tradicional de espaço vetorial. A abordagem proposta consiste em criar o perfil do pesquisador

[9]Nota: Os autores não mencionaram, mas provavelmente a técnica seja o algorítimo *PageRank*.
[10]Trata-se de um diretório aberto onde voluntários editam e categorizam páginas da internet.

com base nos trabalhos visualizados. A correlação entre os perfis é computada utilizando a técnica EditDistance[11] para árvores. Finalmente o modelo de ativação disperso é construído para localizar perfis com interesse semelhantes. A distância entre os modelos permite determinar o grau da correlação. Um subconjunto com 60 mil exemplares da *National Science and Technology Library* foi utilizado para realizar os experimentos. Para avaliar os resultados foi utilizada a métrica *Normalized Discounted Cumulative Gain* da área de recuperação de informações. Foram realizadas avaliações para 5, 10, 15, 20, 25 e 30 recomendações. Ao final observou-se que a melhor opção foi a recomendação de 10 artigos.

No trabalho de Wang e Blei[35] é apresentado um algoritmo para realizar a recomendação de artigos científicos para pesquisadores de uma comunidade *online*. A abordagem proposta combina a técnica tradicional de filtragem colaborativa com modelagem probabilística de tópicos. Para realizar as recomendações, foi construída uma matriz com usuários/itens. Nesta matriz está presente a informação se o usuário gosta ou não do artigo científico. Uma característica típica desse tipo de matriz é apresentar alta dispersão, ocorrendo muitos espaços sem avaliação. A proposta dos autores é de utilizar a técnica de predição chamada de *Matrix Fatorization*[17, 18] para preencher completamente a matriz antes de realizar as recomendações por meio da técnica *neighborhood*. Esta técnica não necessita de um perfil definido para o usuário, a mesma realiza as predições por meio de um modelo. A avaliação do algoritmo proposto foi realizada com um subconjunto dos dados do *site* CiteULike[12] através da métrica *recall*. Ao final, a proposta demostrou ser mais efetiva que a filtragem colaborativa tradicional. Como sugestão de melhoramento, os autores propõem utilizar no futuro o perfil de usuário para adequar as recomendações aos interesses particulares.

O trabalho de Ohta, Hachiki e Takasu[44] apresenta uma metodologia para recomendação de trabalhos relatados a partir da análise de um documento antigo digitalizado com a técnica de reconhecimento óptico de caracteres (OCR). A partir da extração de termos técnicos do documento, é aplicada a técnica TF-IDF[45] para analisar a frequência dos termos. Os resultados são então armazenados em um vetor. Em seguida é construído um grafo bipartido com conexões para um conjunto de possíveis trabalhos relatados em função dos seus termos técnicos. O grafo é analisado com o algoritmo HITS[41]. A partir disso se obtêm um *ranking*, os N primeiros trabalhos são então utilizados como recomendação. Para realizar a avaliação da abordagem proposta, os autores utilizaram a métrica da área de recuperação de informação chamada *precision*. Variando N em 5, 10, 15, 30 e 50 foi possível identificar a melhor opção. Ao final, os autores concluem que a melhor precisão obtida foi 0,35 com N igual a 5. Uma limitação do trabalho é o uso apenas do título e resumo para gerar as recomendações.

Sugiyama e Kan[46] propõem uma nova estratégia para recomendação *Serendipity* de artigos a pesquisadores juniores com o intuito de expandir seus horizontes e despertar novos interesses. O conceito de *serendipity* está relacionado com a descoberta de um novo interesse enquanto se está procurando algo já conhecido. A metodologia proposta foi dividida em três etapas: i) construção de um vetor para o perfil básico do pesquisador com tudo que é necessário para gerar a recomendação.

[11] Também conhecida com distância de Levinstein.

[12] http://www.citeulike.org - Acessado em 23/04/2016

Então, a partir desse perfil é definido um novo perfil *serendipity* para as recomendações. ii) construção do vetor de artigos candidatos a recomendação usando as citações e referências. iii) cálculo das similaridades entre os perfis e os artigos candidatos usando a equação da similaridade do cosseno. Ao final as similaridades são ordenadas e as n primeiras são as recomendações. Os melhores resultados dos experimentos com a abordagem dos perfis de usuários dissimilares foram obtidos com $n = 9$ para os pesquisadores juniores e $n = 12$ para os pesquisadores seniores. Ao final os autores expõem que os seus resultados foram melhores que as abordagens baseadas em *Maximal Marginal Relevance* (MMR). Como sugestão de trabalhos futuros fica o melhoramento do perfil a fim de obter melhores resultados, especialmente focando em como selecionar coautores. Na Figura 4 pode-se visualizar o perfil do usuário construído com a rede de coautoria para recomendações *Serendipity*.

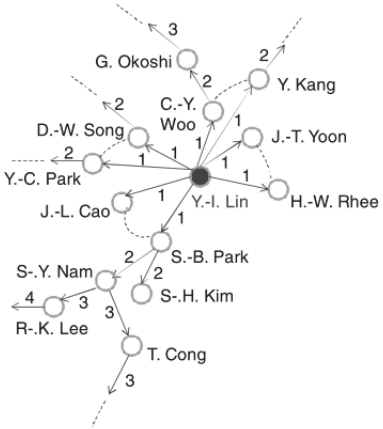

Figura 4. Perfil do usuário construído com a rede de coautoria (Adaptado de Sugiyama et al.[46]).

A proposta de Huang et al.[39] consiste em identificar citações usando palavras explícitas no texto. Posteriormente é proposto um modelo baseado em um dicionário que contem a probabilidade de translação de uma dada referência em uma palavra ou frase para todos os termos da linguagem descritiva. Em seguida é computada a probabilidade de uma dada referência em questão usando as probabilidades da translação. Finalmente as referências passam por um *ranking* e são recomendadas as 20 primeiras. Após uma série de experimentos e comparações, os autores afirmam que a abordagem proposta ultrapassa o estado da arte atual. Nas conclusões os autores apresentam que o emprego do contexto das citações (palavras explícitas) em conjunto com a referência melhoram a qualidade final da recomendação de novas citações. Experimentos em dois *datasets* (CiteSeer e CiteULike) demonstram que a abordagem aumenta as métricas *precision*, *recall* e *f-measure* em pelos menos 5% e 10% respectivamente.

No trabalho de Beel et at.[36], foi realizado um estudo comparativo sobre as abordagens de avaliações *online* e *offline* para Sistemas de Recomendação. Os autores encontrarão contradições entre os resultados das duas formas de avaliação. Em seguida Beel et al.[47] realizaram outro trabalho. Foi feita uma revisão sistemática de 80 abordagens existentes para recomendar artigos científicos. Constatou-se que existe mais de 170 artigos publicados sobre essas abordagens. Após a análise, foi constatado que 21% não foram avaliados. Entre os avaliados, cerca de 19% não foram avaliados em relação ao *baseline*. Com relação ao tipo de avaliação, somente 5 trabalhos (7%) foram avaliados de

forma *online*. A maioria dos experimentos avaliativos (cerca de 69%) foi realizada de forma *offline*. As fontes para avaliações foram obtidas do CiteSeer (29%), ACM (10%), e CiteULike (10%). Ao final foi concluído que não é possível identificar qual abordagem é mais promissora, pois não existem um consenso de qual trabalho representa o estado da arte. Posteriormente, os autores apresentam em outro trabalho[48] um protótipo de um sistema para pesquisar, organizar e criar artigos científicos. Trata-se do *Docear's Research Paper Recommender System*. Os dados dos usuários são armazenados em *Mind Maps* que desempenham o papel de um perfil, e servem para gerar as recomendações. No ano de 2013, o sistema contava com cerca 1,8 milhão de artigos em sua base de dados. O protótipo encontra-se em estágio inicial, os autores esperam melhorar as técnicas empregadas para gerar as recomendações. Na Figura 5 pode-se visualizar a quantidade de artigos encontrados por ano para sistemas de recomendação de artigos.

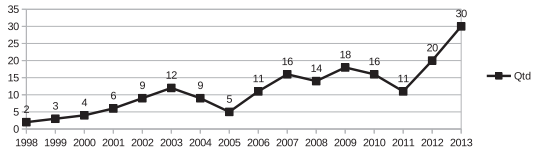

Figura 5. Quantidade de propostas de SR por ano (Adaptado de Beel et al.[47]).

No trabalho de Zhou, Chen e Chen[49] é introduzido o conceito de partição de comunidades de pesquisadores com interesses semelhantes para então gerar as recomendações. As comunidades são armazenadas em um grafo de rede de citações. Os experimentos foram realizados com artigos da cit-HepTh (*High Energy Physics Theory*) compreendidos entre janeiro de 1993 e abril de 2003. A partir dos dados se construiu um grafo de citações e posteriormente aplicou-se o Algoritmo *Greedy Expansion* para encontrar as comunidades. Analisando os artigos de uma comunidade se observou que os mesmos apresentam bastante similaridade em relação a área pesquisada. Posteriormente, selecionou-se cinco comunidades e aplicou-se o algoritmo PaperRank com o intuito de confirmar a convergência dos resultados. Como melhoramentos futuros fica a sugestão dos autores de focar na simplificação da influência computacional e assim melhorar a performance computacional.

Nos trabalhos de Sugiyama e Kan[50, 51] buscou-se construir um sistema de recomendações por meio do potencial de citação de artigos. Inicialmente foi construído um perfil vetorial para os pesquisadores com base nos seus artigos publicados na DBLP[13]. Posteriormente elaborou-se vetores para os artigos presentes na base de conhecimento obtida na ACM *Digital Library*[14]. Finalmente utilizou-se a similaridade do coeficiente de correlação de Pearson entre o perfil e vetor para recomendar os artigos com maior similaridade para os usuários alvo. Após diversos experimentos com a finalidade de ajustar a acurácia em 10% com relação ao *baseline*, os autores afirmam que a abordagem proposta é eficiente em caracterizar artigos candidatos para recomendação com muito alto índice de acurácia.

No trabalho de Ha, Kwon e Kim[52] apresenta-se um sistema de recomendação para sugerir aos pesquisadores trabalhos publicados que ainda não foram citados por outros trabalhos. O objetivo é descobrir novos trabalhos de interesse dos leitores. Os experimentos foram realizados com dados

[13]http://www.informatik.uni-trier.de/ley/db/ - Acessado em 25/04/2016.
[14]http://dl.acm.org/ - Acessado em 25/04/2016.

da DBLP e demostram que a fatoração de matrizes apresentam bons resultados quando comparados com métodos baseados no perfil de usuário e perfil de item.

No trabalho de Alshaikh, Uchyigit e Evans[53] é proposto o uso de árvores normalizadas dinamicamente para modelar o perfil de usuário. Os experimentos foram realizados de forma *offline* com dados de treinamento da ACM *Digital Library* e dados de teste da CiteSeerX *Digital Library*. Os autores utilizaram a média aritmética da *precison* para localizar o melhor conjunto de parâmetros α e *TopN* para recomendação. Os melhores valores para α e *TopN* foram 0,4 e 3, respectivamente. Ao final, a proposta dos autores apresentou melhores resultados quando comparada com outros dois *baselines* previamente estabelecidos (vetores dinâmicos e árvores não normalizadas).

2.8.1 Resumo Comparativos dos Trabalhos Estudados

Nesta subseção foram analisados diversos sistemas de recomendação propostos para auxiliar pesquisadores, tanto na recomendação de novos artigos, como referências e citações. Na Tabela 4 se pode visualizar um resumo sobre a estratégia de modelagem de perfil e *dataset* empregado nos trabalhos estudados.

Tabela 4. Técnicas usadas no perfil e *dataset* usados nos trabalhos pesquisados.

Ano	**Autores**	**Perfil**	***Dataset***
2002	McNee et al.[33]	Não possui perfil, computa similaridade no grafo de citações	ResearchIndex
2004	Middleton et al.[40]	Ontologia	
2010	Ekstrand et al.[34]	Abordagem Alternativa (Similaridade de Citações)	ACM
2010	Zhang e Li[43]	Arvores	NSTL
2011	Wang e Blei[35]	Não possui perfil, utiliza a Fatoração de Matriz (Modelo)	CiteULike
2011	Sugiyama e Kan[46]	Vetores	
2011	Ohta, Hachiki e Takasu[44]	Vetores	
2012	Huang et al.[39]	Modelo de Translação de Referencias	CiteSeer, CiteULik
2013,2015	Sugiyama e Kan[50][51]	Vetores	DBLP,ACM
2013	Beel et al.[47][48]	Modelo baseado em *Mind Maps*	
2014	Zhou, Chen e Chen[49]	Não possui perfil, utiliza comunidades	cit-HepTh
2015	Ha, Kwon e Kim[52]	Não possui perfil, utiliza fatoração de matriz	DBLP
2017	Alshaikh, Uchyigit e Evans[53]	Árvores Normalizadas Dinamicamente	ACM *Digital Library*

Na Tabela 5 pode-se visualizar um resumo sobre as técnicas empregadas nos trabalhos estudados com relação a técnicas utilizada para realizar e avaliar as recomendações.

Tabela 5. Técnicas de recomendação e avaliação usados nos trabalhos pesquisados.

Ano	**Autores**	**Técnica de Recomendação**	**Técnica de Avaliação**
2002	McNee et al.[33]	filtragem colaborativa	
2004	Middleton et al.[40]		
2010	Ekstrand et al.[34]	HITS, pageRank, filt. colaborativa, Hibridas	
2010	Zhang e Li[43]	EditDistance	NDCG
2011	Wang e Blei[35]	filt. colaborativa, *Matrix Fatorization*	*recall*
2011	Sugiyama e Kan[46]	similaridade do cosseno	serendipity
2011	Otha, Hachiki e Takasu[44]	TF-IDF, HITS	*precision*
2012	Huang et al.[39]		*precision, recall, f-measure*
2013,2015	Sugiyama e Kan[50][51]	correlação de Pearson	acurácia
2013	Beel et at.[36]		várias abordagens
2013	Beel et al.[47][48]		
2014	Zhou, Chen e Chen[49]	alg. Greedy Expansion, paperRank	
2015	Ha, Kwon e Kim[52]	SVD	*Precision* e *Recall*
2017	Alshaikh, Uchyigit e Evans[53]		Média da *Precision*

Observou-se diversas abordagens distintas com o uso de técnicas bastante difundidas na literatura como: filtragem colaborativa, filtragem baseada em conteúdo, abordagens híbridas, PageRank, HITS, coeficiente de similaridade de Pearson, TF-IDF. Com relação as fontes de dados utilizadas para os experimentos, a maioria utiliza dados do CiteSeer, CiteULike, ACM e DBLP. Com relação as avaliações dos experimentos, observa-se uma maior predominância de testes *offline*, principalmente com a utilização das métricas da área de recuperação de informações: *precision* e *recall*.

Uma observação importante a destacar é que a maioria dos trabalhos estudados sobre recomendações para pesquisadores exploram apenas a recomendação de artigos, desconsiderando outros elementos presentes na vida acadêmica do pesquisador como orientações, coordenações, revisões, projetos de pesquisa, propriedade intelectual, formação acadêmica e outros elementos.

2.9 REPUTAÇÃO DE PESQUISADORES

Nesta subseção são apresentados os resultados de uma revisão sobre as métricas de avaliação de reputação acadêmica de pesquisadores. Na Cientometria e Bibliometria existem diversas métricas para analisar a produção cientifica e determinar a reputação de pesquisadores. Em essência elas estão divididas em dois grandes grupos. O primeiro origina-se no h-index[54] e todos as propostas de melhoramento como: g-index[55], AR-index [56], e-index[57], hg-index[58], h'-index[59], h_l-index[60], e outras. O segundo grande grupo se baseia no algoritmo PageRank do Google[42]. Neste grupo, pode-se citar o PageRank modificado[61, 62], PageRank ponderado[63], PaperRank[64] e CITEX[65]. O rep-Index[66, 67] é uma nova proposta que procura incorporar outros elementos presentes na vida acadêmica e cientifica.

2.9.1 Métricas baseadas no h-index

Hirsch[54] define o h-index como: "o montante de artigos (quantidade) com o número de citações (qualidade ou visibilidade) maior ou igual a h para cada artigo". A partir de um simples número, é possível avaliar e comparar as citações e publicações cientificas. O número total de citações é sempre maior que h^2 porque artigos não tem o mínimo de quantidade de h citações.

Egghe[55] apresenta o g-index como um melhoramento ao h-index. O fato é que o g-index não considera a quantidade de citações que um artigo teve, em outras palavras, trabalhos com alta quantidade de citações representam uma indicação de qualidade e portanto sua relevância deve ser considerada. A sua definição é: "Para um conjunto de artigos, em ordem descendente de citação, o g-index é o maior número de modo que os g primeiros artigos receberam (somados) ao menos g^2 citações". A partir desta definição é possível afirmar que $g \geq h$. Apesar de o g-index ser uma evolução ao h-index, ele ainda apresenta diversas limitações. O impacto da publicação, auto-citações, obter o total de produção/citações e comparações entre diferentes áreas. Um único trabalho altamente citado pode gerar um falso índice, colocando a reputação do pesquisador acima da média. No trabalho de

Costas e Bordons[68] o g-index é analisado e comparado com o h-index. Este estudo mostrou que o g-index é mais sensível que o h-index para a avaliação seletiva de cientistas.

Jin et al.[56] propõem o AR-index para complementar o h-index. Ele é definido por: a raiz quadrada do somatório da média de citações pelo ano no h-core. A seguinte equação mostra o AR-index.

$$AR = \sqrt{\sum_{j=1}^{h} \frac{cit_j}{a_j}} \text{ (Adaptado de Jin et al.[56])} \tag{20}$$

Onde h é o h-index, cit é a quantidade de citações, a é o número de anos desde a publicação do artigo. A maior contribuição do AR-index é o uso dos anos a partir da publicação para computar o complemento para do h-index. Isto permite compreender melhor o comportamento dos pesquisadores.

Beel e Gipp[69] propõem o i10-index para o site do Google Scholar[15]. É definido na versão padrão como a quantidade de artigos com ao menos 10 citações. Há outra versão dessa métrica, mas computada com os dados dos últimos cinco anos. Outro recurso do *site* é valor do h-index.

Zhang[57] apresenta o e-index como uma alternativa para melhorar dois problemas do h-index. O primeiro é evitar a perda de citações menores que o h-core. Outra limitação é a sua baixa resolução devido ao fato de usar um número inteiro para representar o índice. Então o autor propõem o emprego de números reais como solução ao problema. O e-index é definido pela seguinte equação:

$$e^2 = \sum_{j=1}^{h} (cit_j - h) = \sum_{j=1}^{h} (cit_j - h^2) \text{ (Adaptado de Zhang[57])} \tag{21}$$

Onde cit_j são as citações recebidas pelos j^{th} artigos e e^2 denota o excesso de citações com o $h - core$. Este índice mostra um complemento fundamental para avaliar os autores com altas quantidades de citações. Ele também possibilita a comparação de pesquisadores quando o h-index estiver amarado.

Alonso, Cabrerizo e Herrera[58] apresentam o hg-index. Ele é uma combinação das métricas h-index e g-index. A ideia é combinar os benefícios de ambas em uma simples e melhorada métrica, procurando eliminar as desvantagens com a combinação. O hg-index é calculado pela média geométrica dos dois índices de acordo com a seguinte equação: $hg = \sqrt{h.g}$. Os autores demonstraram que $h \leq hg \leq g$ e $hg - h \leq g - hg$. Esta métrica produz uma visão mais balanceada da produção cientifica.

Zhang[59] propôs o h'-index para melhorar o h-index através da combinação do h-index, e-index e publicações que estão abaixo do h-core (também denominado de h-tail ou t-index). A combinação dessas três áreas do gráfico de citações permite uma completa análise das produções e citações.

[15]http://scholar.google.com.br

Consequentemente, não há perda de informações durante o cálculo do h'-index. O índice em questão é calculado pela seguinte equação.

$$h' = rh = \frac{eh}{t} \text{ (Adaptado de Zhang[59])} \tag{22}$$

Onde e, h e t são e-index, h-index e t-index, respectivamente. Ao final do trabalho, os autores afirmam que h'-index é um número real para avaliar a produção cientifica de modo mais justo e mais razoável.

Zhai, Yan e Zhu[60] propõem o h_l-index para melhorar o h-index baseado na qualidade de citações de artigos. Uma rede de citações foi construída a fim de analisar a qualidade entre os artigos acadêmicos. A sua definição é: "O h_l-index de um conjunto de artigos h, é o maior número inteiro de tal modo que o conjunto de artigos tem, pelo menos, h artigos satisfazendo que o $_l$-index não é menor que h". A análise das propriedades do h_l-index demonstra que: $0 \leq h_l - index \leq h - index$.

2.9.2 Métricas baseadas no PageRank

O sucesso do algoritmo PageRank proposto por Page et al.[42], inspirou diversas propostas de uso em métricas para analisar a reputação cientifica e determinar a reputação de pesquisidores. Chen et al.[61] comparam o PageRank original e PageRank ponderado com o *ranking* de citações, h-index, e medidas centralizadas. Para realizar esta comparação eles selecionaram 108 autores com as maiores citações da área de recuperação de informação entre o período de 1970 até 2008. O intervalo do parâmetro *damping factor* foi variado de 0,05 até 0,95. Para verificar os resultados, a Correlação de Spearman[32] foi utilizada. A principal conclusão é que o *ranking* de citações é similar ao PageRank na rede de co-citação autoral.

No trabalho de Du, Bai e Liu [64] os autores mencionam a similaridade entre a *web* e a rede de citações cientificas. Neste trabalho é proposto o uso do PaperRank como uma extensão ao PageRank original de Page et al.[42], e HITS proposto por Kleinberg[41], para medir a publicação cientifica. A ideia é medir os relacionamentos entre os artigos, computar os relacionamentos de citações indiretas entre os artigos usando o algoritmo de Dijkstra. Os resultados são armazenados em uma matriz de relatividade R. A matriz R é usada para computar a matriz de co-citações e co-referencias. A matriz de associação é construída medindo a relatividade de relacionamentos diretos e indiretos entre dois artigos. O próximo passo é computar a matriz de probabilidades. Finalmente, o *ranking* das pontuações são calculados. Os resultados dos experimentos mostram que PaperRank pode localizar mais artigos dominantes que os outros métodos.

De acordo com os autores Pal e Ruj[65], Citex prove pontuações normalizadas para os autores e artigos a fim determinar seus *rankings*. O problema é modelado com um grafo de publicação e citações, onde os autores e artigos são as arestas, um vértice não dirigido entre dois nodos indica que o autor escreveu o artigo (grafo de publicação). Um vértice de um artigo para outro artigo representa

uma citação (grafo de citação). A maior contribuição deste trabalho é computar as pontuações com um algoritmos específico. Os valores são armazenados em duas colunas de um vetor (uma para os autores e outra para os artigos). Ao final, as pontuações são normalizadas pela divisão da pontuação dos autores/artigos pelo somatório dos autores/artigos, dessa forma, cada pontuação está compreendida entre 0 e 1, e a soma dos scores é sempre 1. A limitação deste trabalho esta em assumir que todos os autores apresentam a mesma contribuição em um artigo.

2.9.3 Rep-Model e Rep-Index

Cervi et al.[66, 67] propuseram o Rep-Index como, um índice para classificar pesquisadores com outros critérios além de artigos e citações. Na proposta está incluído grau de instrução, bancas, orientações, comitês e produção. A métrica é um inteiro positivo entre 1 e 5 (em ordem ascendente). O grande diferencial de outras métricas é a média ponderada, escopo abrangente e adaptabilidade. O Rep-Index é baseado em um perfil específico para pesquisadores, chamado Rep-Model. Ele é um conjunto de elementos que representam o comportamento acadêmico e cientifico dos pesquisadores. Estes elementos não tem foco somente na bibliometria. Na Tabela 6 se pode visualizar a categoria, elementos, abreviações e pesos.

Tabela 6. Categorias e Elementos para Rep-Model com os respectivos pesos no Rep-Index.

Categoria			**Elementos**		
	Abreviação	Peso		Abreviação	Peso
Identificação	ID	15	Grau de Instrução	GI	15
Orientação	ORI	15	Orientação de Pós-doutorado	OP	6
			Orientação de Doutorado	OD	5
			Orientação de Mestrado	OM	4
Banca	BAN	10	Participação em Banca de Doutorado	PBD	6
			Participação em Banca de Mestrado	PBM	4
Comitê	COM	10	Membro de Corpo Editorial de Periódico	MCEP	5
			Revisão de Periódico	RP	3
			Coordenação de Comitê de Conferência	CCC	1
			Membro de Comitê de Conferência	MCC	1
Publicação	PUB	50	Artigo em Periódico	AP	15
			Livro	LIV	8
			Capitulo de Livro	CLIV	5
			Trabalho Completo em Conferência	TCC	8
			H-Index	HI	7
			Rede de Coautoria	RC	3
			Projeto de Pesquisa	PP	2
			Software	SOFT	2
		100			100

O valor decimal para o Rep-Index é computado pela equação 23.

$$Rep-Index_{(R)} = \sum_{i=1}^{c} \left(\sum_{j=1}^{e_i} \frac{(v_j.w_j)}{max(v_j)} \right) \text{ (Adaptado de Cervi et al.[66, 67])} \quad (23)$$

Onde R referencia a reputação do pesquisador procurada, c representa o número total de categorias, i representa o intervalo de 1 até o número total de categorias (c), e_i representa o total de elementos em cada categoria, j refere-se ao intervalo de 1 até o número total de elementos (e_i), v representa o valor do elemento, w_j refere-se ao peso do elemento, $max(v_j)$ representa o maior valor do elemento. Para computar o valor inteiro do Rep-Index, utiliza-se a equação 24.

$$Rep-Index_{(R)} = \begin{cases} 1, & \text{se } Rep-Index_{(R)} \geq 0 \wedge < 20 \\ 2, & \text{se } Rep-Index_{(R)} \geq 20 \wedge < 40 \\ 3, & \text{se } Rep-Index_{(R)} \geq 40 \wedge < 60 \\ 4, & \text{se } Rep-Index_{(R)} \geq 60 \wedge < 80 \\ 5, & \text{se } Rep-Index_{(R)} \geq 80 \wedge \leq 100 \text{ (Adaptado de Cervi et al.[66, 67])} \end{cases} \quad (24)$$

Esta equação classifica a reputação dos pesquisadores em cinco níveis. Valor 1 é o nível inicial e 5 é o nível final. A ideia é que os pesquisadores iniciantes tenham níveis menores e os mais avançados níveis maiores de reputação.

2.10 CONCLUSÕES DO CAPÍTULO

Neste capítulo estudou-se os principais conceitos sobre Sistemas de Recomendação. Iniciou-se a partir da concepção dos Sistemas de Recomendação, evolução, classificações, taxonomia e tipos. Posteriormente focou-se especificamente em Sistemas de Recomendação personalizados com a modelagem de Perfil e Reputação. Ainda estudou-se as três principais técnicas existentes atualmente para Sistemas de Recomendação Personalizados: Filtragem Baseada em Conteúdo, Filtragem Colaborativa e Abordagens Hibridas. Um estudo amplo foi realizado com relação a Sistemas de Recomendação propostos para auxiliar pesquisadores e métricas de reputação acadêmica existentes para os mesmos. Finalmente estudou-se as principais métricas empregadas na avaliação de Sistema de Recomendação bem como a medidas de similaridades existentes para as comparações de perfis dos usuários.

3. ABORDAGEM PROPOSTA

Este capitulo objetiva apresentar a abordagem de recomendação proposta, destacando as suas características, escopo, tipos e algoritmo. A abordagem proposta para gerar as recomendações deve responder aos três questionamentos mencionados anteriormente no trabalho:

i) **O que Fazer?** Recomendar o que os pesquisadores com maior reputação da mesma subárea (consonância com o que produz) realizaram. Em outras palavras, essa abordagem sugere que se siga os passos de outros pesquisadores com mais reputação na mesma subárea de atuação.

ii) **Como Fazer?** Descrição de como realizar a atividade recomendada. Esta descrição também deve apresentar opções para o pesquisador. É possível realizar vários tipos de recomendações, uma para cada elemento quantitativo previsto no modelo. Também pode-se combinar elementos na mesma recomendação. Exemplos: “Amplie a sua Rede de colaboração com o Pesquisador B.”, “Aumente o H-Index para 29.”

iii) **Quando Fazer?** Fazer por primeiro o que tiver maior impacto na reputação do pesquisador para que ele possa evoluir na carreira. A abordagem proposta deve apresentar ao pesquisador os itens recomendados em ordem decrescente de relevância para a reputação do mesmo.

A etapa inicial da abordagem consiste na adaptação do modelo do perfil do pesquisador. Neste caso, se optou por realizar esta tarefa no Rep-Model[66, 67] conforme especificado no tópico 2.9.3. As alterações propostas incluem novos elementos no Rep-Model com o intuito de utilizá-lo para esta finalidade, bem como eliminar elementos que não se adaptam ao contexto proposto. Na Tabela 7 pode-se visualizar a proposta de adição de elementos ao Rep-Model, os novos pesos serão ajustados posteriormente.

Tabela 7. Elementos adicionados ao Rep-Model.

Rep-Model Original			**Adições ao Rep-Model**	
Port.	Ing.	Elemento	Port.	Elemento
NM	NM	Nome	CP	Cultivar Protegida
INST	INST	Instituição	CR	Cultivar Registrada
GI	ED	Grau de Instrução	DI	Desenho Industrial
OP	PA	Orientação de Pós-doutorado	MARC	Marca
OD	PTA	Orientação de Doutorado	PAT	Patente
OM	MDA	Orientação de Mestrado	TCI	Topografia Circuíto Integrado
PBM	PEBPT	Participação em Banca de Mestrado	PRODTEC	Produto Tecnológico
PBD	PEBMD	Participação em Banca de Doutorado	PROCTEC	Processo ou Técnicas
MCEP	EBM	Membro de Corpo Editorial de Periódico	TT	Trabalho Técnico
RP	RJ	Revisão de Periódico	PREM	Prêmios
CCC	CCC	Coordenação de Comitê de Conferência		
MCC	CCM	Membro de Comitê de Conferência		
AP	ASJ	Artigo em Periódico	**Adições Textuais ao Rep-Model**	
LIV	BP	Livro	TPB	Títulos Produção Bibliográfica
CLIV	BCP	Capítulo de Livro	TPT	Títulos Produção Técnica
TCC	CWPCP	Trabalho Completo em Conferência	RC	Resumo do Currículo
HI	HI	H-Index	AA	Áreas Atuação
RC	NC	Rede de Coautoria	TO	Títulos de Orientações
PP	RP	Projeto de Pesquisa	TB	Títulos de Bancas
SOFT	SOFT	Software	TOA	Títulos Orientações Andamento

Observa-se a adição de novos elementos quantitativos ao Rep-Model com a finalidade de contemplar a diversidade de produção anteriormente mencionada. Quanto aos elementos textuais, os mesmos são utilizados para construir um perfil de subárea de atuação durante o processo de mineração de texto (TF-IDF) e possibilitam localizar as afinidades (similaridades) entre as áreas de pesquisa dos pesquisadores.

Afim de considerar a qualidade dos artigos publicados em não somente a sua quantidade, transformou-se o elemento ASJ em decimal e por meio de uma multiplicação entre o seu valor unitário e o percentual definido no documento de área da CAPES para cada nível Qualis. O Qualis Periódicos foi obtido diretamente da plataforma Sucupira[1] para o período compreendido entre 2010 e 2016. Na tabela 8 pode-se visualizar os percentuais utilizados para cada área.

Tabela 8. Pesos para QUALIS por área.

Qualis	Ciência da Computação	Odontologia	Economia
A1	100%	100%	100%
A2	85%	85%	80%
B1	70%	70%	60%
B2	50%	55%	40%
B3	20%	40%	25%
B4	10%	15%	15%
B5	5%	5%	5%
C	0%	0%	0%

A partir da análise dos 28 elementos do Rep-Model utilizados para gerar as recomendações, observa-se que alguns podem incluir informações relevantes ao contexto da recomendação. Dessa forma, a abordagem proposta é divida em duas partes: **i) Abordagem para Recomendações Não Personalizadas:** são as recomendações que não possuem informações especificas para o usuário, elas se caracterizam por apenas considerarem o elemento do Rep-Index e sua importância para o aumento da reputação. Em outras palavras, são geradas por meio de um simples cálculo de reputação que simula o incremento hipotético de uma unidade em um elemento do Rep-Index. Nesta categoria pode-se citar como exemplos: “Aumente o item: Artigo em Periódico (ASJ) para 36”, “Aumente o item: Trabalho Completo em Conferência (CWPCP) para 180”, “Aumente o item: Membro de Corpo Editorial de Periódico (EBM) para 18”. **ii) Abordagem para Recomendações Personalizadas:** são específicas para cada usuário. Elas são geradas por um processo mais complexo que o anterior, onde a similaridade de perfil e reputação dos demais pesquisadores são utilizadas para gerar a recomendação. Como exemplos se pode citar: “Amplie a sua Rede de colaboração com o Pesquisador: 3286329883412205, similaridade: 0,999154 Rep-Index: 18,91”.

Ao final do processo, as duas abordagem são combinadas e recomendadas de forma única ao pesquisador. A partir do conjunto de n possíveis recomendações (personalizadas e genéricas) ao usuário U, deve-se computar o quanto cada uma incrementa na sua reputação, posteriormente aplica-se uma ordenação em ordem decrescente de valor. Como estratégia para responder a pergunta

[1] http://sucupira.capes.gov.br

“quando fazer?”, recomenda-se primeiro os itens que mais incrementam a sua reputação (maiores valores).

3.1 ABORDAGEM PARA RECOMENDAÇÕES NÃO PERSONALIZADAS

As recomendações não personalizadas são geradas a partir da simulação do aumento do Rep-Index do pesquisador em questão. O aumento da reputação (representado pela letra Δ) para um pesquisador pode ser calculado pela diferença entre a nova reputação e a sua atual. A nova reputação deve ser computada considerando o incremento hipotético de uma unidade no elemento desejado. Na equação 25 pode-se visualizar o referido cálculo.

$$\Delta_{(R)} = \left(\text{Rep-Index Novo}_{(R)}\right) - \left(\text{Rep-Index Atual}_{(R)}\right) \tag{25}$$

A equação anterior é funcional para a maioria das situações, contudo não considera a situação do valor máximo (teto) do elemento em questão. Além deste fato, existe a necessidade de computar todos os elementos do Rep-Index para obter a reputação nova e a atual. Pode-se simplificar o mesmo e corrigir a situação acima mencionada. A equação 26 apresenta uma proposta melhorada para o cálculo em questão.

$$\Delta_{(R)} = \begin{cases} 0, & \text{se } n \geq max_{(R_i)} \\ \frac{n * w_{(R_i)}}{max_{(i)}}, & \text{senão} \end{cases} \tag{26}$$

Onde, $\Delta_{(R)}$ é o aumento na reputação do pesquisador R, n representa o incremento desejado para o elemento, i indica o elemento do Rep-Index em questão para o pesquisador R, e $max_{(i)}$ é o valor máximo do elemento i no grupo formado pelos pesquisadores do CNPq para a área em questão. Observa-se que enquanto o incremento for menor que o valor máximo, o aumento da reputação é diretamente proporcional ao peso do elemento e inversamente proporcional ao valor máximo do elemento para o grupo dos pesquisadores do CNPq da área.

3.2 ABORDAGEM PARA RECOMENDAÇÕES PERSONALIZADAS

Ao contrário das recomendações não personalizadas, as personalizadas necessitam de várias etapas e no geral são mais complexas. A etapa inicial da abordagem é a definição da utilização dos elementos definidos para o modelo de perfil (Rep-Model adaptado). A Equação 27 apresenta uma estrutura de decisão que direciona as ações de acordo com o tipo do elemento.

$$Elemento_{(U_n)} = \begin{cases} \text{Calcular Rep-Index}_{(Elemento)}, & se\ Tipo_{(Elemento)} = Inteiro \\ \text{Text mining}_{(Elemento)}, & se\ Tipo_{(Elemento)} = Texto \end{cases} \quad (27)$$

Onde: $Elemento_{(U_n)}$ é uma estrutura de decisão, U é um usuário em análise no momento e n o número do elemento. Os elementos do tipo inteiro do Rep-Model são utilizados para computar o Rep-Index. Os resultados de cada usuário são armazenados em uma posição do seguinte vetor.

$$\overrightarrow{Rep - Index_{U(i)}} = \left[\; U_1 \quad \ldots \quad U_i \; \right] \quad (28)$$

Os elementos que forem textuais, exceto NN e INST, são submetidos a uma etapa mais complexa que os quantitativos. Nesta fase é inferido um perfil com base nas informações textuais. Para isso, se optou por empregar técnicas de recomendação baseada em conteúdo (recuperação de informações). Na Figura 6 pode-se visualizar as etapas do processo de *Text mining*.

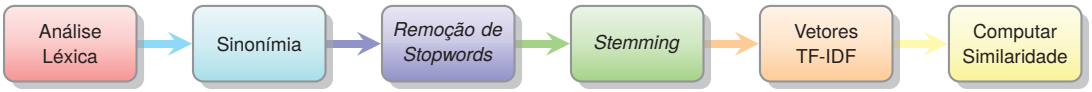

Figura 6. Etapas do processo de *Text Mining*.

A primeira etapa é a análise léxica que tem por objetivo separar as informações em palavras. A segunda etapa trata-se da sinonímia, a mesma busca por meio de um dicionário de sinônimos aproximar textos semanticamente semelhantes. Em seguida ocorre a remoção de *stopwords*, ela busca a eliminação de classes de palavras sem relevância ou que podem gerar falsas similaridades. A maioria das listas existentes é composta por artigos, pronomes, preposições, numerais, conjunções, consoantes, advérbios, entre outras classes gramaticais ou conjuntos específicos de palavras. Posteriormente, pretende-se utilizar a técnica denominada de *stemming*, a mesma procura reduzir as palavras ao seu radical por meio da supressão de sufixos. Dessa forma, garante-se, por exemplo, que um verbo apresente sempre o mesmo radical independente do modo e tempo verbal que esteja conjugado no texto. Finalmente, aplica-se a técnica de vetorização denominada TF-IDF. A mesma visa simplificar o processamento das informações textuais por meio de sua representação em vetores esparsos com a frequência de ocorrência e relevância dos seus termos.

O resultado da técnica de vetorização é aplicado as funções de similaridades, correlação ou distância que estão disponíveis no Apache Mahout. Na Equação 29 pode-se visualizar todas as

técnicas utilizadas.

$$Sim(U_{m,n}) = \begin{cases} \text{Correlação de Pearson}(U_{m,n}) \\ \text{Correlação de Spearman}(U_{m,n}) \\ \text{Correlação de Kendall Tau}(U_{m,n}) \\ \text{Similaridade Fuzzy}(U_{m,n}) \\ \text{Distância Euclidiana}(U_{m,n}) \\ \text{Distância de Canberra}(U_{m,n}) \\ \text{Distância de Tanimoto}(U_{m,n}) \\ \text{Distância de Log-likelihood}(U_{m,n}) \\ \text{Distância de Manhattan}(U_{m,n}) \\ \text{Distância de Minkowski}(U_{m,n}) \\ \text{Distância de Chebyshev}(U_{m,n}) \\ \text{Distância do Coseno}(U_{m,n}) \\ \text{Distância de EarthMovers}(U_{m,n}) \end{cases} \tag{29}$$

Cada uma dessas funções terá como resultado final uma matriz triangular $M_{sim_{(U_{m,n})}}$ onde são armazenas as similaridades entre os perfis dos pesquisadores. Na Equação 30 pode-se visualizar a representação da matriz de similaridades.

$$M_{sim_{(U_{m,n})}} = \begin{bmatrix} U_{1,1} & \cdots & U_{n,m} \\ \vdots & \ddots & \vdots \\ U_{m,n} & \cdots & U_{m,m} \end{bmatrix} \tag{30}$$

No apache Mahout, as distâncias são convertidas em similaridades por meio da equação: $\text{similaridade} = \frac{1}{1+\text{distância}}$. A abordagem utilizada pelo Apache Mahout tem como objetivo principal promover uma simplificação neste processo e obter ganhos de desempenho computacional. Contudo, observou-se discrepâncias ao computar as métricas MAE e RMSE. Por isso, se propõem utilizar a equação 31 que considera inicialmente a normalização das distâncias e posteriormente a sua conversão em similaridades.

$$s = 1 - \left(\frac{d - d_{min}}{d_{max} - d_{min}} \right) \tag{31}$$

Onde: d representa a distância obtida na equação 29, min e max representam os valores mínimo e máximo respectivamente, e s a similaridade obtida no intervalo de valores entre 0 e 1, inclusive.

A partir das matrizes de similaridades entre os pesquisadores e as categorias, aplica-se o algoritmo do vizinho mais próximo (*Nearest Neighbor*) com o parâmetro n (indica quantos vizinhos

devem ser selecionados) igual ao número de pesquisadores existentes na categoria atual. Dessa forma, seleciona-se os n pesquisadores mais afins à categoria. Isto é obtido pelo *ranking* decrescente dos valores.

A recomendação personalizada sobre "o que fazer" para um usuário U é realizada pela análise do vetor $\overrightarrow{Rep-Index}_{U(i)}$, onde são localizados os pesquisadores que possuem maior reputação que U. A matriz $M_{sim_{(U_{m,n})}}$ também é utilizada para localizar os perfis mais semelhantes com relação a subárea de atuação. O cálculo do incremento na reputação é realizado com a equação 26 proposta anteriormente.

3.3 ALGORITMO PARA GERAR AS RECOMENDAÇÕES

A partir do vetor $\overrightarrow{Rep-Index}$ preenchido com as reputações e da matriz M_{sim} de similaridades de perfis, computada apenas com a função que apresentar os melhores resultados, aplica-se o algoritmo da figura 7 para gerar as recomendações personalizadas e não personalizadas.

```
1.  função recomendacao(p, max⃗, RepIndex⃗, Msim, n)
2.    requer n ≥ 1
3.    para cada i de p.elem faça
4.      r_i.delta ← 0
5.      se (eRecomendacaoPersonalizada(p.elem_i)) então
6.        reput⃗ ← localizarMaioresReputacoes(RepIndex, p.repIndex)
7.        perfis⃗ ← localizarPerfisSemelhantes(Msim, p, 0.99005)
8.        se ((reput.size() ≥ 0) e (perfis.size() ≥ 0) e ((p.elem_i.valor + n) ≤ max_i)) então
9.          r_i.delta ← (n * p.elem_i.peso) / max_i
10.         r_i.texto ← "Aumente o item:  " + p.elem_i.tipo + " para " + (p.elem_i.valor + n) + msgPers(reput, perfis)
11.       fim se
12.     senão
13.       se ((p.elem_i.valor + n) ≤ max_i) então
14.         r_i.delta ← (n * p.elem_i.peso) / max_i
15.         r_i.texto ← "Aumente o item:  " + p.elem_i.tipo + " para " + (p.elem_i.valor + n)
16.       fim se
17.     fim se
18.   fim para
19.   r ← ordenaDeltaDecrecenteRemoveDeltaZero(r)
20.   retornar r
21. fim função
```

Figura 7. Algoritmo de Recomendação Proposto

O algoritmo proposto apresenta como parâmetros de entrada: p é o pesquisador e seus atributos, max é vetor com os valores máximos (teto) da área para cada elemento do Rep-Model, $Rep-Index$ é um vetor com o Rep-Index computado para cada pesquisador, $Msim$ é a matriz de similaridades de perfis e n indicando o aumento hipotético em um elemento do Rep-Model. O algoritmo utiliza as seguintes funções auxiliares:

- **eRecomendacaoPersonalizada(elemento)**: retorna verdadeiro se o parâmetro elemento do Rep-Model é uma recomendação personalizada. Nesta abordagem apenas os elementos ED e NC são recomendações personalizadas.
- **localizarMaioresReputacoes(Rep-Index, limiar)**: localiza no vetor Rep-Index os pesquisadores com reputação maior que o parâmetro limiar. No algoritmo proposto, o valor para o limiar é igual ao Rep-Index do pesquisador.

- **localizarPerfisSemelhantes(Msim, pesquisador, limiar)**: localiza na matriz Msim os pesquisadores com maior similidade que o parâmetro limiar para o pesquisador.
- **msgPers(reputações, perfis)**: gera uma mensagem de recomendação personalizada com base nos parâmetros de reputação e perfil.
- **ordenaDeltaDecrecenteRemoveDeltaZero(recomendacoes)**: ordena o vetor em ordem decrescente pelo valor de delta e remove os que apresentarem valor zero.

Ao final, o algoritmo retorna um vetor em ordem decrescente de incremento de reputação, contendo todas as informações relevantes sobre a recomendação.

4. EXPERIMENTOS E RESULTADOS

Este capítulo apresenta os experimentos realizados com o intuito de avaliar a abordagem de recomendação proposta. Os mesmos foram realizados individualmente para cada uma das áreas de estudo. Ao final apresenta-se uma análise sobre os resultados obtidos.

A avaliação da abordagem proposta foi realizada com os seguintes experimentos em cada área do estudo: **i)** Avaliar a evolução temporal dos elementos dos Rep-Index para o período entre 2012 e 2016. (Experimento: 4.2.1); **ii)** Encontrar a melhor combinação de pesos para o Rep-Index. (Experimentos: 4.3.1, 4.3.2 e 4.3.3); **iii)** Encontrar a melhor combinação de técnicas para similaridade de subárea (Análise Léxica, Sinonímia, *Stopwords*, *Stemming*, correlação/similaridade/distância). (Experimentos: 4.4.1, 4.4.2 e 4.4.3); **iv)** Gerar e avaliar as recomendações. (Experimentos: 4.5.1, 4.5.2 e 4.5.3).

4.1 DADOS UTILIZADOS NOS EXPERIMENTOS

O CNPq[1] define uma classificação em três grandes áreas: i) Engenharias, Ciências Exatas e da Terra; ii) Ciências Humanas e Sociais Aplicadas; iii) Ciências da Vida. Cada grande área é composta por diversas áreas. A CAPES[2] também possui uma definição semelhante ao CNPq, onde existem três grandes colégios: i) Colégio de Ciências da Vida; ii) Colégio de Ciências Exatas, Tecnológicas e Multidisciplinar; iii) Colégio de Humanidades. Cada um dos colégios é composto por três grandes áreas. Para realização dos experimentos, escolheu-se três grandes áreas de acordo com as definições do CNPq, o que contempla tanto a classificação do CNPq como da CAPES. Dessa forma, se obtém resultados mais amplos para a abordagem proposta, o que comprova a sua adaptabilidade. Na Tabela 9 pode-se visualizar a classificação definida pelo CNPq e as áreas utilizadas nos experimentos.

Tabela 9. Grandes áreas de pesquisa do CNPq.

Engenharias, Ciências Exatas e da Terra	**Ciências Humanas e Sociais Aplicadas**	**Ciências da Vida**
Ciência da Computação	Economia	Odontologia

Os dados foram coletados na plataforma Lattes[3], Microsoft Academic Search[4], Google Scholar[5] e Publish or Perish[6]. No trabalho de Vivian e Cervi[70] pode-se encontrar detalhadamente os passos para recuperação das informações e criação do XML *Dataset* utilizado para realizar os experimentos. As consultas e intercambio dos dados é realizada com auxílio da linguagem XQuery por meio do *software* Xml2Arff[71]. A análise dos dados/resultados é realizada com o auxílio do *software* Weka[72] nos casos de mineração de dados, classificação e aprendizado de máquina. As análises

[1] http://plsql1.cnpq.br/divulg/RESULTADO_PQ_102003.curso
[2] http://www.capes.gov.br/avaliacao/sobre-as-areas-de-avaliacao
[3] http://lattes.cnpq.br
[4] http://academic.research.microsoft.com
[5] https://scholar.google.com.br
[6] http://www.harzing.com/pop.htm

estatísticas e gráficas são realizadas utilizando o *software* R[73]. Os grafos são gerados com auxílio do *software* Gephi[74] por meio do algoritmo de Fruchterman-Reingold[75], aplicou-se um limiar de 0,99905 nas similaridades (igual ao algoritmo da figura 7).

4.2 EXPERIMENTOS REALIZADOS

O experimento 4.2.1 analisa a evolução quantitativa dos elementos do Rep-Index entre o período de 2012 e 2016. Os experimentos 4.3.1, 4.3.2, 4.3.3 apresentam o cálculo dos pesos do Rep-Index específicos para cada uma das três áreas do estudo. Os experimentos 4.4.1, 4.4.2, 4.4.3 apresentam o cálculo da melhor técnica para encontrar as similaridades das subáreas de atuação dos pesquisadores das três áreas do estudo. Finalmente, os experimentos 4.5.1, 4.5.2, 4.5.3 apresentam as recomendações e as suas avaliações para as áreas de estudo.

4.2.1 Experimento 1 - Evolução Quantitativa dos Elementos do Rep-Index entre 2012 e 2016

Elaborou-se um experimento comparativo entre os dados do Rep-Index original de 2012 (utilizados no trabalho de Cervi [76]) e os dados atualizados ao final do ano de 2016. O objetivo deste experimento é apenas verificar a evolução temporal dos elementos e com isso assegura que os mesmos foram coletados de forma satisfatória. Na figura 8 pode-se visualizar os diagramas de Venn com as quantidades de pesquisadores em 2012 (círculo da esquerda) e 2016 (círculo da direita). No ponto de intersecção é apresentada a quantidade que permaneceu com bolsa de produtividade em pesquisa nos dois períodos.

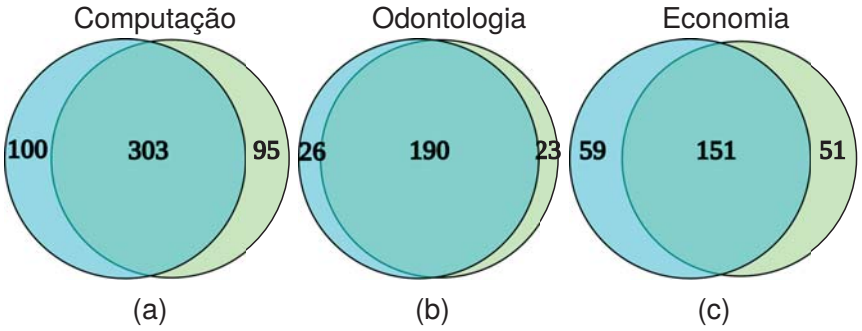

Figura 8. Gráfico de Venn para as três áreas do estudo.

Na figura 8 (a) pode-se visualizar o diagrama para Ciência da Computação, em (b) para Odontologia e em (c) para Economia. Observa-se que a Odontologia manteve praticamente o mesmo grupo durante o período. Já os grupos de Economia e Ciência da Computação alteram um percentual maior de bolsas.

Também elaborou-se gráficos com os elementos quantitativos do Rep-Index para constatar a evolução temporal e verificar possíveis inconsistências na obtenção dos mesmos. As barras azuis (escala da esquerda) representam a média aritmética dos elementos coletados no ano de 2012 e as vermelhas em 2016. As linhas amarelas (escala da direita) representam o que está acima do limiar

de 1,0 e portanto indicam um crescimento no período. Por outro lado, as linhas verdes (limiar $\leq$ 1,0) indicam que houve decréscimo durante o período. Os itens que apresentam apenas barras vermelhas são os novos elementos adicionados ao Rep-Model durante este trabalho, logo não possuem série histórica para realizar a comparação. Nas Figuras 9, 10, 11 pode-se visualizar os resultados.

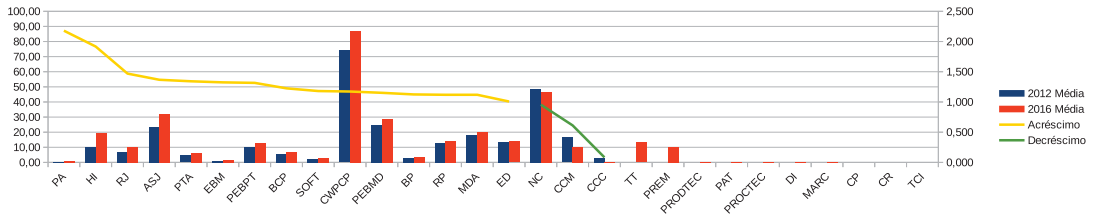

Figura 9. Evolução Quantitativa dos Elementos do Rep-Index para Ciência da Computação.

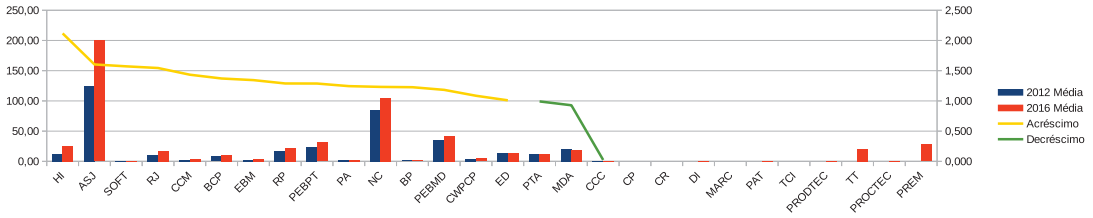

Figura 10. Evolução Quantitativa dos Elementos do Rep-Index para Odontologia.

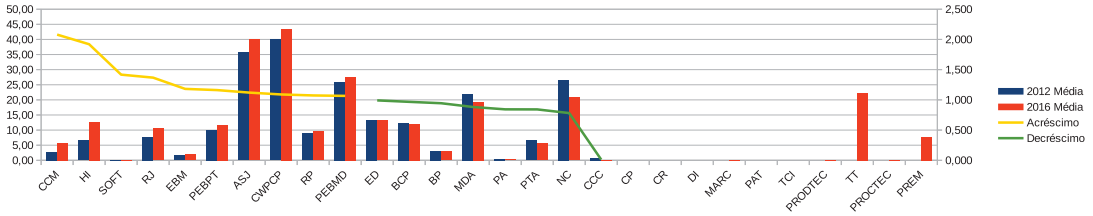

Figura 11. Evolução Quantitativa dos Elementos do Rep-Index para Economia.

Observa-se que a grande maioria dos elementos encontra-se acima do limiar de 1,0. Isto indica que apesar de o grupo de pesquisadores ter sofrido mudanças (bolsas e pesquisadores) as três áreas em geral evoluíram positivamente durante o período. Tal constatação é fundamental para validar os dados coletados em 2016 e que serão utilizados nos próximos experimentos. O elemento AJS não está considerando os pesos do QUALIS. Nas três áreas do estudo o elemento CCC apresentou uma significativa redução devido a mudança na metodologia de mensuração do mesmo. As demais diminuições são consequências da troca de pesquisadores. A Odontologia, que apresentou a menor alteração de pesquisadores, também demostra a menor diminuição de valores. Já a Economia que possui a maior troca de pesquisadores demonstra as maiores alterações.

4.3 DETERMINAÇÃO DOS PESOS ESPECÍFICOS PARA O REP-INDEX

Além de adaptar o Rep-Model para ser o modelo do usuário, é fundamental que os pesos do Rep-Index sejam ajustados devido ao incremento de novos elementos, bem como representar cada

uma das áreas o mais próximo possível da realidade da mesma. A determinação dos pesos específicos do Rep-Index é realizada utilizando-se o complemento para o Rep-Index proposto por Vivian, Cervi e Rovadosky[77]. Este complemento utiliza técnicas de mineração de dados e aprendizado de máquina presente no *software* Weka para computar cinco opções de pesos a partir das pontuações geradas. Entre as técnicas, três são diretamente baseadas na teoria da entropia: `GainRatio`[78] (Equação 32), `InfoGain`[79] (Equação 33) e `SymmetricalUncert`[80] (Equação 34). A técnica `ChiSquared`[81] é desenvolvida a partir da técnica estatística homônima e computa o seu valor com relação ao nível do CNPq. O método `ReliefF`[82, 83, 84] avalia o valor de um elemento, amostrando repetidamente uma instância e considerando o valor do mesmo para a instância mais próxima do mesmo e diferente nível do CNPq.

$$\textit{Gain Ratio}(\text{CNPq, elemento}) = \frac{H(CNPq) - H(CNPq|elemento)}{H(elemento)} \tag{32}$$

$$\textit{Info Gain}(\text{CNPq, elemento}) = H(CNPq) - H(CNPq|elemento) \tag{33}$$

$$\textit{Symmetrical Uncert}(\text{CNPq, elemento}) = 2\Big(\frac{H(CNPq) - H(CNPq|elemento)}{H(CNPq) + H(elemento)}\Big) \tag{34}$$

Onde H representa a entropia da informação definida por Shannon[85] e pode ser calculada por $H(n) = -\sum_{i=1}^{n}(p_i \ \log_2(p_i))$, $CNPq$ é o nível da bolsa de produtividade em pesquisa do CNPq e $elemento$ é o item do Rep-Model em análise pela técnica.

O cálculo do peso a partir da pontuação obtida na etapa anterior é realizado pela Equação 35.

$$w_i = \frac{p_i.100}{\sum_{j=1}^{n}(p_j)} \tag{35}$$

Onde w é o novo peso para o elemento i, p é a pontuação obtida pelo elemento na etapa anterior e n é a quantidade de elementos do Rep-Index.

A avaliação da melhor opção é realizada por meio do cálculo da correlação de Spearman[32], Pearson[86] e Kendall[87] entre o Rep-Index computado com o conjunto proposto de pesos e o nível do CNPq convertido em número inteiro. O nível 2 tem valor 1, nível 1D tem valor 2, nível 1C tem valor 3, nível 1B tem valor 4 e nível 1A tem valor 5. No trabalho de Kozak el al.[88] indica-se como sendo mais adequada a correlação de Spearman para este tipo de comparação, por esse motivo, adotou-se a mesma como critério principal de classificação. Dentre as seis possibilidades de pesos, o conjunto que obtiver o maior valor para correlação é considerada como os novos pesos da área. Dessa forma, o Rep-Index é adequado para classificar os pesquisadores o mais próximo possível da realidade da área (pesquisadores do CNPq).

4.3.1 Experimento 2 - Rep-Index Específico para Ciência da Computação

Este experimento objetiva calcular os pesos do Rep-Index específicos para área da Ciência da Computação devido ao acréscimo de novos elementos. Utilizou-se o complemento proposto por Vivian et al. [77] para tal atividade. Na Tabela 10 pode-se visualizar as opções de pesos.

Tabela 10. Opções de Pesos do Rep-Index específicos para Ciência da Computação.

Elemento	ChiSquared	GainRatio	InfoGain	ReliefF	SymmetricalUncert
ED				19,7434	
PA	5,6033	6,9728	5,2674	4,3599	5,9250
PTA	21,1304	15,2218	21,4138	11,3965	18,8653
MDA	9,5015	10,2270	10,3244	6,9956	10,4815
PEBPT	10,0755	7,8075	10,5657	4,1758	9,4770
PEBMD				3,9939	
EBM	4,9825	5,3275	4,8737	5,2548	5,1343
RJ	7,4801	4,6426	6,5921	2,3386	5,7823
CCC				0,8845	
CCM				1,0667	
ASJ	6,9494	8,4816	6,4072	7,1888	7,2071
BP	5,3881	5,7576	5,6970	2,8259	5,8285
BCP	3,8800	4,0312	4,0115	3,4382	4,0952
CWPCP	7,2874	11,9256	6,8153	5,5254	8,3161
HI	8,8787	9,4538	9,5194	3,5747	9,6739
NC	4,7184	5,8527	4,3648	5,3237	4,9299
RP				2,3097	
SOFT				1,6115	
CP					
CR					
DI				0,1026	
MARC				-0,0066	
PAT				1,1270	
TCI					
PRODTEC				1,0624	
TT				0,7124	
PROCTEC				0,5936	
PREM	4,1249	4,2982	4,1477	4,4011	4,2840
Total	100,00	100,00	100,00	100,00	100,00

A partir das cinco opções de pesos geradas pelo complemento do Rep-Index e mais os pesos originais propostos por Cervi et al.[66, 67], calculou-se o Rep-Index para todos os pesquisadores da área e comparou-se os resultados com o nível de bolsa de produtividade em pesquisa do CNPq. Essa comparação foi realizada por meio da Correlação de Pearson, Spearman e Kendall. Na Figura 12 pode-se visualizar os resultados obtidos.

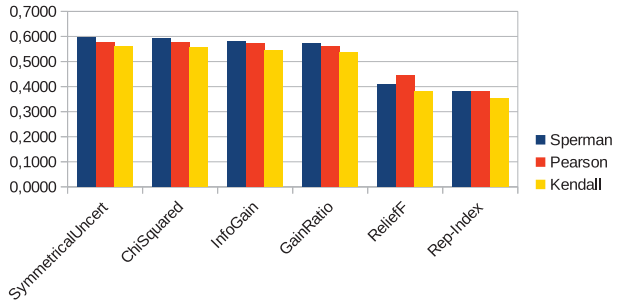

Figura 12. Resultados do Cálculo do Peso do Rep-Index para Ciência da Computação.

Observa-se que o conjunto de pesos gerados com a técnica `SymmetricalUncert` obteve o maior valor para correlação de Spearman 0,5981. Seguido por: `ChiSquared` com 0,5915; `InfoGain` com 0,5818; `GainRatio` com 0,5727; `ReliefF` com 0,4083 e Rep-Index original com 0,3806. Para maiores detalhes vide os gráficos de dispersão no Apêndice D figura 42. O conjunto de pesos obtido com o algoritmo `SymmetricalUncert` será utilizado para computar o Rep-Index dos pesquisadores da área.

4.3.2 Experimento 3 - Rep-Index Específico para Odontologia

Este experimento objetiva calcular os pesos do Rep-Index específicos para área da Odontologia devido ao acréscimo de novos elementos. Utilizou-se o complemento proposto por Vivian et al. [77] para tal atividade. Na Tabela 11 pode-se visualizar as opções de pesos.

Tabela 11. Opções de Pesos do Rep-Index específicos para Odontologia.

Elemento	ChiSquared	GainRatio	InfoGain	ReliefF	SymmetricalUncert
ED				2,2866	
PA	12,9471	12,3876	13,2555	9,7323	12,9736
PTA	15,6858	14,9644	15,9613	12,2492	15,6389
MDA				3,0677	
PEBPT	13,9709	14,4910	12,6576	9,6348	13,2135
PEBMD				4,6135	
EBM				4,2949	
RJ				1,6821	
CCC				0,9238	
CCM				0,2740	
ASJ	17,5878	18,1355	17,2590	13,4697	17,5519
BP				3,8597	
BCP				2,6148	
CWPCP				0,2870	
HI	25,8128	26,3864	26,4079	6,6367	26,4275
NC	13,9956	13,6351	14,4587	7,7493	14,1947
RP				4,1285	
SOFT				0,7450	
CP					
CR					
DI				0,0405	
MARC					
PAT				1,8314	
TCI					
PRODTEC				1,5749	
TT				0,8660	
PROCTEC				0,3226	
PREM				7,1152	
Total	100,00	100,00	100,00	100,00	100,00

A partir das cinco opções de pesos geradas pelo complemento do Rep-Index e mais os pesos originais propostos por Cervi et al.[66, 67], calculou-se o Rep-Index para todos os pesquisadores da área e comparou-se os resultados com o nível de bolsa de produtividade em pesquisa do CNPq. Essa comparação foi realizada por meio da Correlação de Pearson, Spearman e Kendall. Na Figura 13 pode-se visualizar os resultados obtidos.

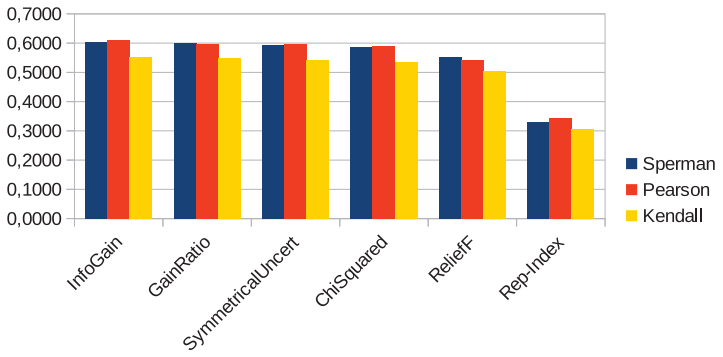

Figura 13. Resultados do Cálculo do Peso do Rep-Index para Odontologia.

Observa-se que o conjunto de pesos gerados com a técnica `InfoGain` obteve o maior valor para correlação de Spearman 0,6044. Seguido por: `GainRatio` com 0,5998; `SymmetricalUncert` com 0,5932; `ChiSquared` com 0,5861; `ReliefF` com 0,5515 e Rep-Index original com 0,3300. Para maiores detalhes vide os gráficos de dispersão no Apêndice D figura 43. O conjunto de pesos obtido com o algoritmo `InfoGain` será utilizado para computar o Rep-Index dos pesquisadores da área.

4.3.3 Experimento 4 - Rep-Index Específico para Economia

Este experimento objetiva calcular os pesos do Rep-Index específicos para área da Economia devido ao acréscimo de novos elementos. Utilizou-se o complemento proposto por Vivian et al. [77] para tal atividade. Na Tabela 12 pode-se visualizar as opções de pesos.

Tabela 12. Opções de Pesos do Rep-Index específicos para Economia.

Elemento	ChiSquared	GainRatio	InfoGain	ReliefF	SymmetricalUncert
ED				8,8500	
PA				1,7840	
PTA	40,8650	38,1721	38,8748	6,3392	38,6281
MDA	36,8250	36,2527	37,1363	8,2571	36,8212
PEBPT	22,3100	25,5752	23,9888	8,0488	24,5508
PEBMD				7,5589	
EBM				4,9931	
RJ				6,1183	
CCC				0,0060	
CCM				3,6780	
ASJ				4,8307	
BP				3,0544	
BCP				10,4093	
CWPCP				4,8478	
HI				5,3339	
NC				1,7069	
RP				1,5887	
SOFT				0,9364	
CP					
CR					
DI					
MARC				0,3971	
PAT					
TCI					
PRODTEC				0,9661	
TT				3,8772	
PROCTEC				0,0277	
PREM				6,3903	
Total	100,00	100,00	100,00	100,00	100,00

A partir das cinco opções de pesos geradas pelo complemento do Rep-Index e mais os pesos originais propostos por Cervi et al.[66, 67], calculou-se o Rep-Index para todos os pesquisadores da área e comparou-se os resultados com o nível de bolsa de produtividade em pesquisa do CNPq. Essa comparação foi realizada por meio da Correlação de Pearson, Spearman e Kendall. Na Figura 14 pode-se visualizar os resultados obtidos.

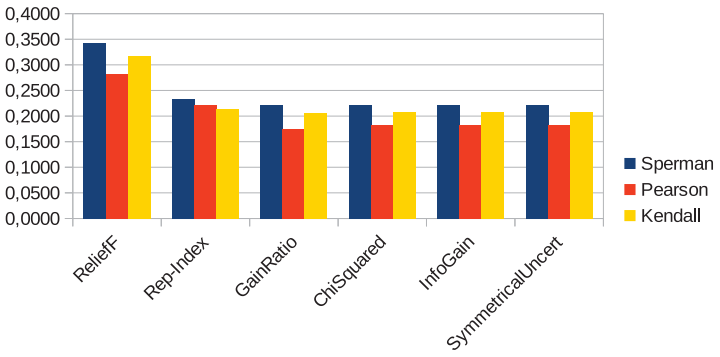

Figura 14. Resultados do Cálculo do Peso do Rep-Index para Economia.

Observa-se que o conjunto de pesos gerados com a técnica `ReliefF` obteve o maior valor para correlação de Spearman 0,3427. Seguido por: Rep-Index original com 0,2325; `GainRatio` com 0,2225; `ChiSquared`, `InfoGain` e `SymmetricalUncert` com 0,2222. Para maiores detalhes vide os gráficos de dispersão no Apêndice D figura 44. O conjunto de pesos obtido com o algoritmo `ReliefF` será utilizado para computar o Rep-Index dos pesquisadores da área.

4.4 AVALIAÇÃO DE SIMILARIDADE DE PERFIL PARA SUBÁREA

Para localizar a melhor combinação de técnicas para a similaridade de subárea, deve-se inicialmente agrupar os pesquisadores em categorias (definidas pelas subáreas do currículo Lattes). Utilizou-se o elemento Área de Atuação (AA) do Rep-Model modificado para esta finalidade.

Os experimentos foram realizados no Apache Mahout[7] versão 1.12.2 executando em apenas um único nodo do Apache Hadoop[8]. O ambiente já possui classes Java prontas para diversos idiomas, cada uma com as regras pré determinadas de análise léxica, *stemming* e *stopwords*. Devido ao fato de que as informações textuais estarem escritas principalmente nos idiomas Português e Inglês, optou-se por realizar experimentos com ambas as classes. Além das existentes, criou-se uma nova classe em Java denominada `MyBrazilianAnalyzer.java`, a qual possui regras personalizadas de *stopwords* (vide Apêndice A) e sinonímia (vide Apêndice B). Também criou-se a classe `LoglikelihoodDistanceMeasure.java` (vide Apêndice C) para computar esta medida de distância com vetores esparsos, uma vez que o Mahout possui o `Log-Likelihood` apenas para filtragem colaborativa por meio da biblioteca Taste/Apache Mahout Math. Nos experimentos utilizou-se as classes da Tabela 13.

Tabela 13. Classes empregadas nos experimentos.

Classe	Análise léxica	*Stopwords*	Sinonímia	*Stemming*
`ClassicAnalyzer` (CL)	`ClassicTokenizer`, `ClassicFilter`, `LowerCaseFilter`	33 (Inglês)	Não	Não
`StandardAnalyzer` (ST)	`StandardTokenizer`, `StandardFilter`, `LowerCaseFilter`	33 (Inglês)	Não	Não
`EnglishAnalyzer` (EN)	`StandardTokenizer`, `StandardFilter`, `LowerCaseFilter`	33 (Inglês)	Não	`PorterStemFilter`**
`BrazilianAnalyzer` (BR)	`StandardTokenizer`, `StandardFilter`, `LowerCaseFilter`	128 (Português)	Não	`BrazilianStemFilter`***
`MyBrazilianAnalyzer` (MY)	`StandardTokenizer`, `StandardFilter`, `LowerCaseFilter`	1234 (Português)	147*	`BrazilianStemFilter`***

*Usou-se a Sinonímia para Traduzir termos da área em língua Inglesa para língua Portuguesa. Vide Apêndice B.
**Nota: utiliza a classe `PorterStemmer` do projeto Snollball/Apache Lucene.
***Nota: utiliza a classe `PortugueseStemmer` do projeto Snollball/Apache Lucene.

Após localizar os pesquisadores mais afins para cada categoria, basta realizar a comparação com a definição original das categorias e encontrar os verdadeiros positivos e falsos positivos. De pose dessas informações, constrói-se a matriz de confusão. A partir das similaridades apresentadas na Equação 29 e das classes da Tabela 13 realizou-se as avaliações e optou-se por utilizar as métricas: *precision*, *recall*, MAE, RMSE, estatística Cohen's Kappa[89, 90, 91] e coeficiente de correlação de Matthews[92]. Utilizou-se as mesmas por elas estarem presente na biblioteca do Weka. Os valores são obtidos através da média ponderada entre o resultado individual de cada classe com o número de pesquisadores na mesma.

4.4.1 Experimento 5 - Similaridade de Subárea para Ciência da Computação

Este experimento tem por objetivo localizar a melhor combinação de técnicas para realizar a mineração de texto com os elementos textuais do Rep-Model, e dessa forma localizar os pesquisadores mais afins em sua(s) subárea(s) de atuação. Em um grupo formado por 398 pesquisadores da área de

[7] http://mahout.apache.org
[8] http://hadoop.apache.org

Ciência da Computação, inicialmente foram localizadas 959 subáreas de atuação. Essa quantidade se justifica no fato de que a maioria dos pesquisadores apresentam mais de uma subárea de atuação, em geral uma área clássica da Ciência da Computação e algumas áreas de pesquisas mais atuais ou mesmo multidisciplinares. Ao final foram obtidas 219 categorias distintas e mais uma categoria denominada *Empty* para os casos sem o elemento AA. Entre as 220 categorias, 57 (25,90%) possuem mais de um pesquisador e 163 (74,09%) são formadas por um único pesquisador. Na Figura 15 pode-se visualizar as subáreas da Ciência da Computação com mais de um pesquisador.

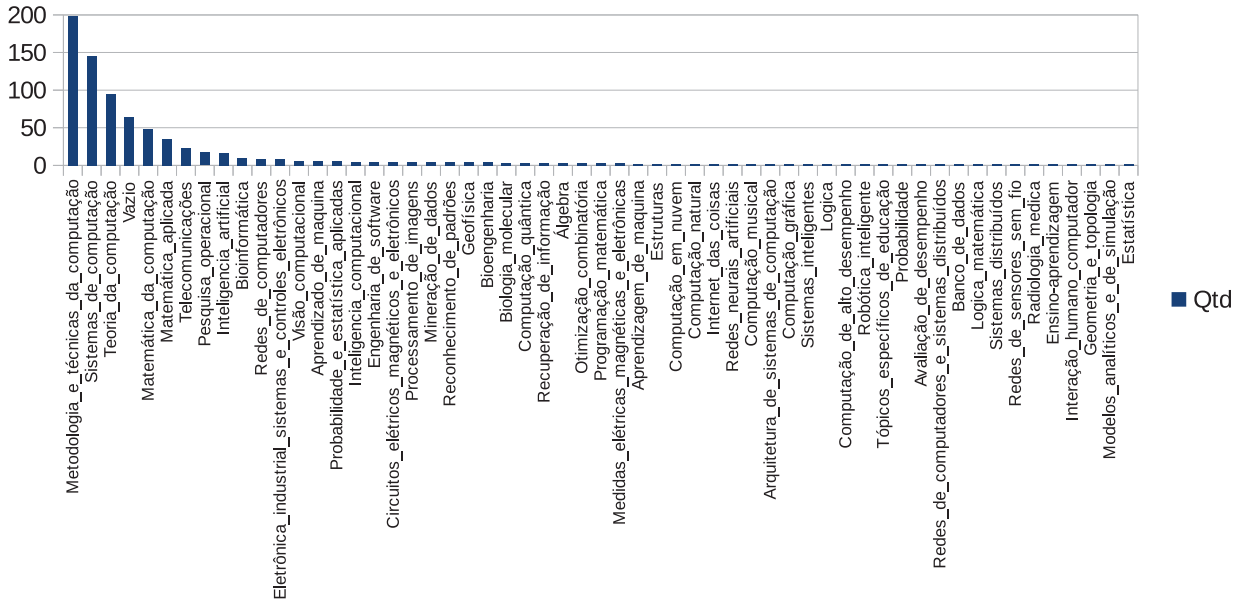

Figura 15. Quantidade de pesquisadores por Subárea para Ciência da Computação.

Utilizou-se cinco combinações (CL, ST, EN, MY e BR) diferentes para a análise léxica, sinonímia, remoção de *stopwords* e *steeming*. Utilizou-se três técnicas de correlação, uma técnica de similaridade e nove de distância para computar as similaridades entre os vetores resultantes da etapa de TF-IDF. Nas Figuras 16, 17 e 18 pode-se visualizar os resultados mensurados com as métricas de recuperação de informações, medições de erros, estatísticas e correlação específicas.

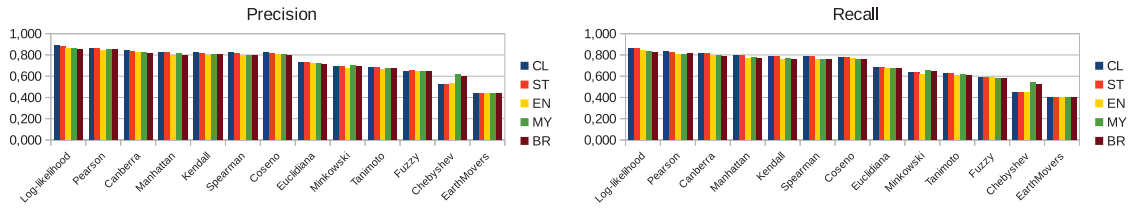

Figura 16. *Precision* e *Recall* para Ciência da Computação.

Na Figura 16 verifica-se que a combinação `Log-Likelihood` e `ClassicAnalyzer` obteve a maior *precision* (0,884), seguido por `StandardAnalyzer` com 0,88; `EnglishAnalyzer` com 0,864; `MyBrazilianAnalyzer` com 0,861 e `BrazilianAnalyzer` com 0,854. Com relação a *recall* os resultados também apresentam a mesma ordem. Ficando `ClassicAnalyzer` com 0,858;

StandardAnalyzer com 0,856; EnglishAnalyzer com 0,837; MyBrazilianAnalyzer com 0,832 e BrazilianAnalyzer com 0,827.

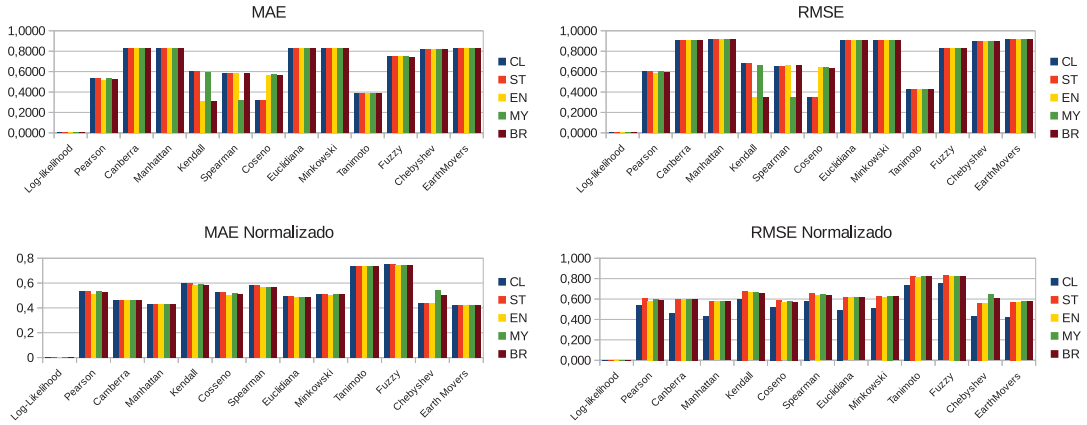

Figura 17. MAE e RMSE para Ciência da Computação.

Na Figura 17 apresenta-se as métricas relativas aos erros. Observa-se que Log-Likelihood apresentou erros (MAE e RMSE) muitos insignificante com relação as demais funções utilizadas. Para a ClassicAnalyzer foi encontrado o valor de 0,0008 para MAE e 0,0011 para RMSE. Este fato indica um alto grau das similaridades, ou seja, valores muito próximos de 1,0.

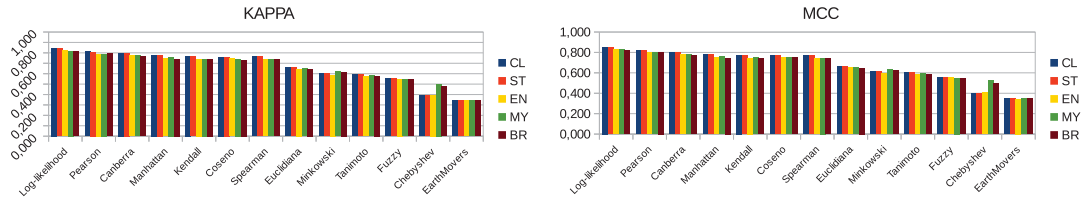

Figura 18. Kappa e MCC para Ciência da Computação.

Ainda utilizou-se a estatística Cohen's *Kappa* para avaliar a distribuição dos elementos na matriz de confusão. Observa-se que a mesma também apresentou a mesma ordem das anteriores, ficando ClassicAnalyzer com 0,845; StandardAnalyzer com 0,842; EnglishAnalyzer com 0,822; MyBrazilianAnalyzer com 0,816 e BrazilianAnalyzer com 0,81. Incluiu-se também o Coeficiente de Correlação de Matthew, uma correlação específica para avaliar classificadores e sistemas afins. Mais uma vez os resultados se repetem, ficando ClassicAnalyzer com 0,853; StandardAnalyzer com 0,85; EnglishAnalyzer com 0,831; MyBrazilianAnalyzer com 0,825 e BrazilianAnalyzer com 0,818. Ao final dos experimentos para encontrar a melhor combinação de técnicas para localizar os pesquisadores mais afins pela(s) subárea(s) de atuação, observou-se que a combinação da função Log-Likelihood e ClassicAnalyzer é sem sombra de dúvidas a melhor opção. Os resultados de todas as avaliações utilizadas demostraram isso.

4.4.2 Experimento 6 - Similaridade de Subárea para Odontologia

Este experimento tem por objetivo localizar a melhor combinação de técnicas para realizar a mineração de texto com os elementos textuais do Rep-Model, e dessa forma localizar os pesquisadores mais afins em sua(s) subárea(s) de atuação.

Em um grupo formado por 214 pesquisadores da área de Odontologia, inicialmente foram localizadas 596 subáreas de atuação. Essa quantidade se justifica no fato de que a maioria dos pesquisadores apresentam mais de uma subárea de atuação, em geral uma área clássica da Odontologia e algumas áreas de pesquisas mais atuais ou mesmo multidisciplinares. Ao final foram obtidas 134 categorias distintas e mais uma categoria denominada *Empty* para os casos sem o elemento AA. Entre as 135 categorias, 43 (31,85%) possuem mais de um pesquisador e 92 (68,15%) são formadas por um único pesquisador. Na Figura 19 pode-se visualizar as subáreas da Odontologia com mais de um pesquisador.

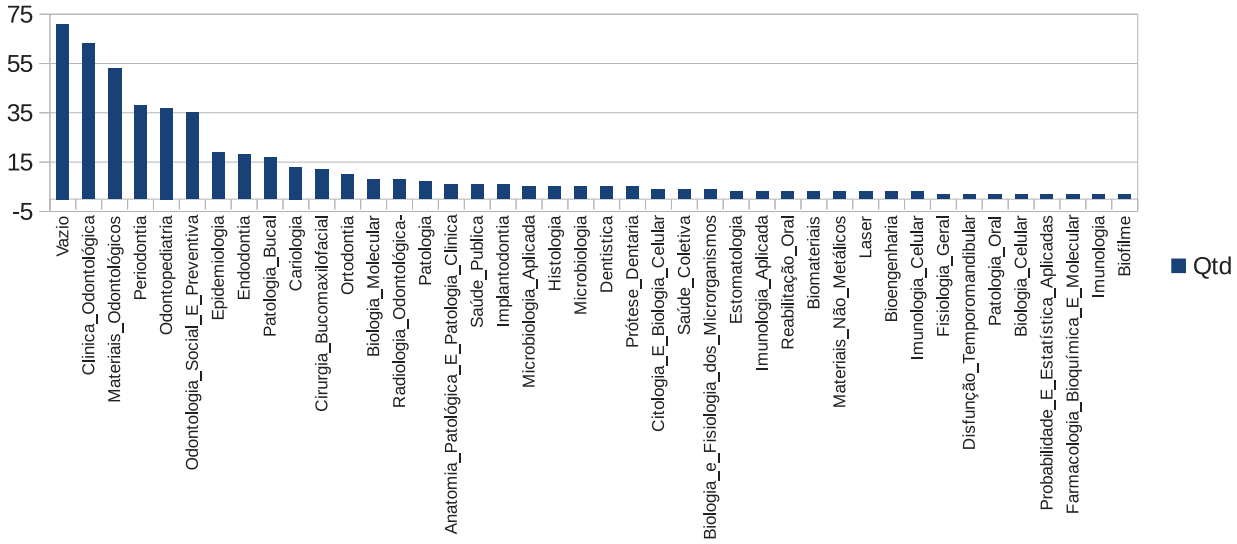

Figura 19. Quantidade de pesquisadores por Subárea para Odontologia.

Utilizou-se cinco combinações (CL, ST, EN, MY e BR) diferentes para a análise léxica, sinonímia, remoção de *stopwords* e *steeming*. Utilizou-se três técnicas de correlação, uma técnica de similaridade e nove de distância para computar as similaridades entre os vetores resultantes da etapa de TF-IDF. Nas Figuras 20, 21 e 22 pode-se visualizar os resultados mensurados com as métricas de recuperação de informações, medições de erros, estatísticas e correlação específicas.

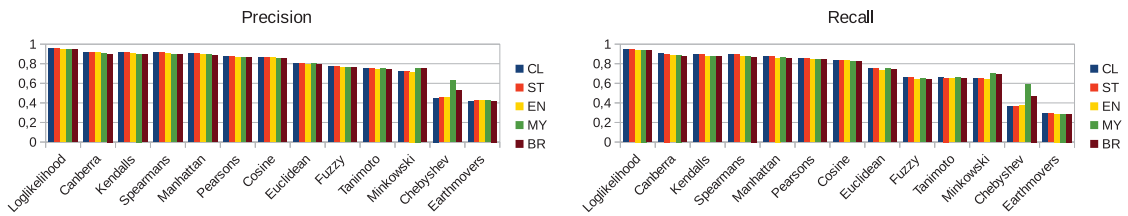

Figura 20. *Precision* e *Recall* para Odontologia.

Na Figura 20 verifica-se que a combinação `Log-Likelihood` e `ClassicAnalyzer` obteve a maior *precision* (0,953), seguido por `StandardAnalyzer` com 0,953; `EnglishAnalyzer` com 0,949; `MyBrazilianAnalyzer` com 0,948 e `BrazilianAnalyzer` com 0,944. Com relação a métrica *recall* os resultados também apresentam a mesma ordem. Ficando `ClassicAnalyzer` com 0,946; `StandardAnalyzer` com 0,946; `EnglishAnalyzer` com 0,940; `MyBrazilianAnalyzer` com 0,940 e `BrazilianAnalyzer` com 0,933.

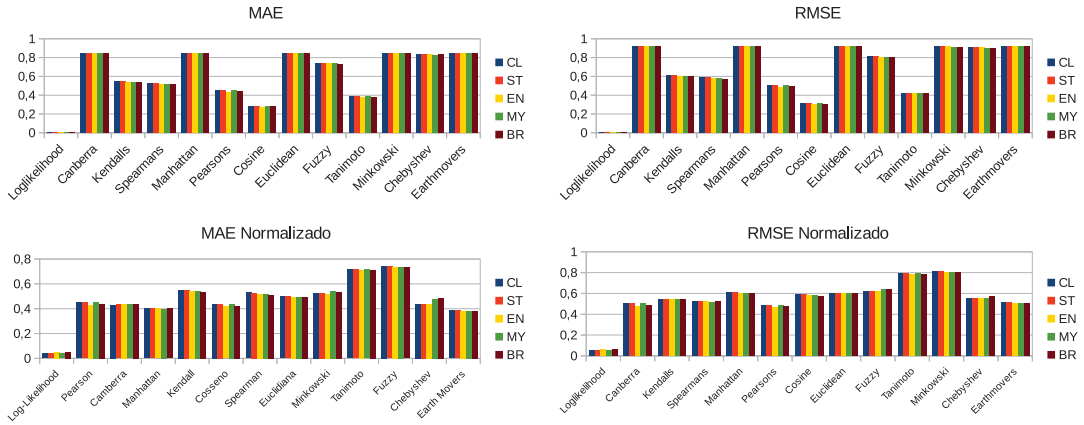

Figura 21. MAE e RMSE para Odontologia.

Na Figura 21 apresenta-se as métricas relativas aos erros. Observa-se que `Log-Likelihood` apresentou erros (MAE e RMSE) muitos insignificante com relação as funções utilizadas. Para a `ClassicAnalyzer` obteve-se o valor de 0,0003 para MAE e 0,0005 para RMSE. Este fato indica um alto grau de precisão nas similaridades, ou seja, valores muito próximos de 1,0.

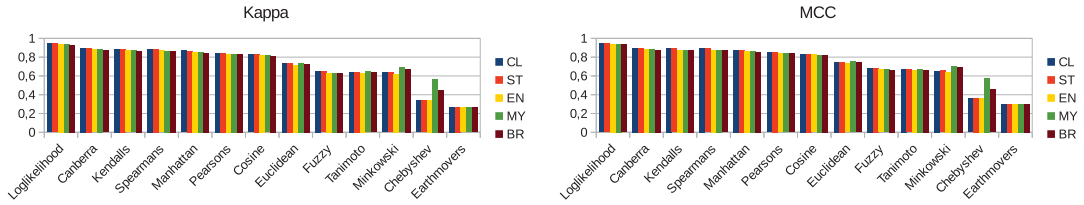

Figura 22. Kappa e MCC para Odontologia.

Ainda utilizou-se a métrica estatística Cohen's *Kappa* para avaliar a distribuição dos elementos na matriz de confusão. Observa-se que a mesma também apresentou a mesma ordem das anteriores, ficando `ClassicAnalyzer` com 0,944; `StandardAnalyzer` com 0,944; `EnglishAnalyzer` com 0,937; `MyBrazilianAnalyzer` com 0,937 e `BrazilianAnalyzer` com 0,929. Incluiu-se também o Coeficiente de Correlação de Matthew, uma correlação específica para avaliar classificadores e sistemas afins. Mais uma vez os resultados se repetem, ficando `ClassicAnalyzer` com 0,945; `StandardAnalyzer` com 0,945; `EnglishAnalyzer` com 0,939; `MyBrazilianAnalyzer` com 0,939 e `BrazilianAnalyzer` com 0,933. Ao final dos experimentos para encontrar a melhor combinação de técnicas para localizar os pesquisadores mais afins pela(s) subárea(s) de atuação, observou-se que da mesma forma como a área da Computação, a combinação da função `Log-likelihood` e

ClassicAnalyzer é sem sombra de dúvidas a melhor opção. Os resultados de todas as avaliações utilizadas demostraram isso.

4.4.3 Experimento 7 - Similaridade de Subárea para Economia

Este experimento tem por objetivo localizar a melhor combinação de técnicas para realizar a mineração de texto com os elementos textuais do Rep-Model, e dessa forma localizar os pesquisadores mais afins em sua(s) subárea(s) de atuação.

Em um grupo formado por 203 pesquisadores da área de Economia, inicialmente foram localizadas 576 subáreas de atuação. Essa quantidade se justifica no fato de que a maioria dos pesquisadores apresentam mais de uma subárea de atuação, em geral uma área clássica da Economia e algumas áreas de pesquisas mais atuais ou mesmo multidisciplinares. Ao final foram obtidas 113 categorias distintas e mais uma categoria denominada *Empty* para os casos sem o elemento AA. Entre as 114 categorias, 33 (28,95%) possuem mais de um pesquisador e 81 (71,05%) são formadas por um pesquisador. Na Figura 23 pode-se visualizar as subáreas com mais de um pesquisador.

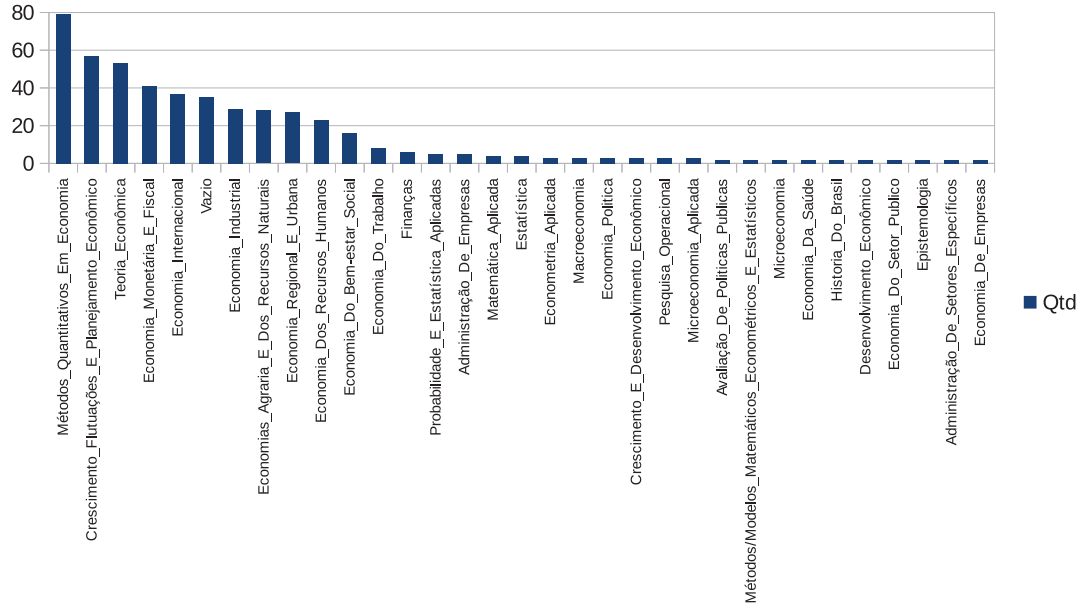

Figura 23. Quantidade de pesquisadores por Subárea para Economia.

Utilizou-se cinco combinações (CL, ST, EN, MY e BR) diferentes para a análise léxica, sinonímia, remoção de *stopwords* e *steeming*. Utilizou-se três técnicas de correlação, uma técnica de similaridade e nove de distância para computar as similaridades entre os vetores resultantes da etapa de TF-IDF. Nas 24, 25 e 26 pode-se visualizar os resultados mensurados com as métricas de recuperação de informações, medições de erros, estatísticas e correlação específicas.

Na Figura 24 verifica-se que a combinação Log-Likelihood e StandardAnalyzer obteve a maior *precision* (0,932), seguido por ClassicAnalyzer com 0,930; EnglishAnalyzer com 0,920; MyBrazilianAnalyzer com 0,912 e BrazilianAnalyzer com 0,908. Ainda com relação a métrica *precision* os resultados apresentam uma inversão de posição insignificante entre ClassicAnalyzer

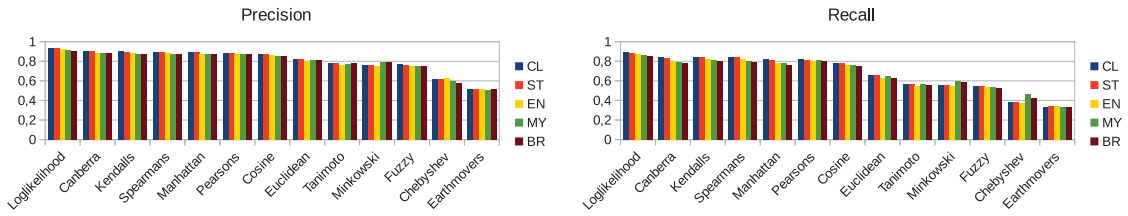

Figura 24. *Precision* e *Recall* para Economia.

e `StandardAnalyzer`. No caso da *recall* a situação volta ao normal, ficando `ClassicAnalyzer` com 0,891; `StandardAnalyzer` com 0,887; `EnglishAnalyzer` com 0,870; `MyBrazilianAnalyzer` com 0,858 e `BrazilianAnalyzer` com 0,856.

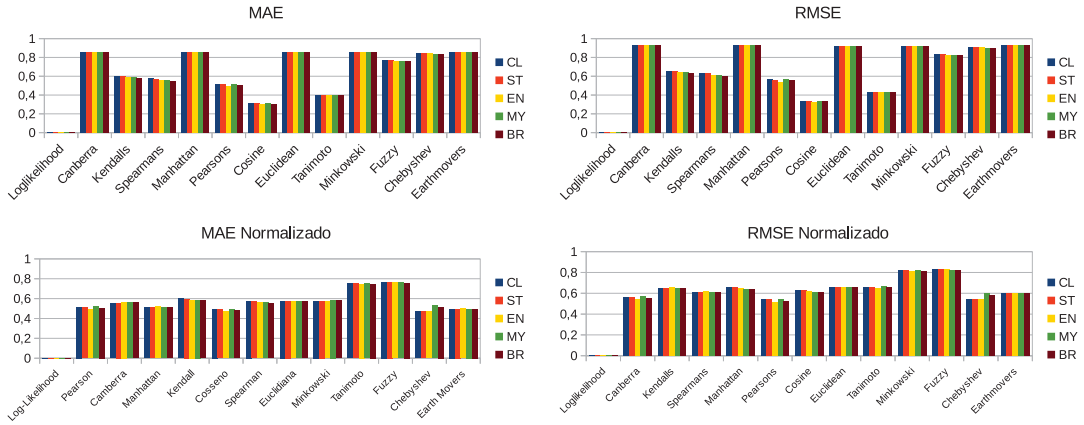

Figura 25. MAE e RMSE para Economia.

Na Figura 25 apresenta-se as métricas relativas aos erros. Observa-se que `Log-Likelihood` apresentou erros (MAE e RMSE) muitos insignificante com relação as demais funções. Para as classes `ClassicAnalyzer` e `StandardAnalyzer` encontrou-se o valor de 0,0006 para MAE e 0,0008 para RMSE. Este fato indica um alto grau de precisão nas similaridades (valores próximos de 1,0).

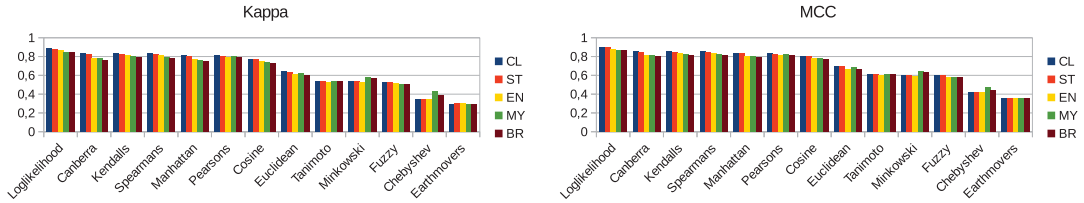

Figura 26. Kappa e MCC para Economia.

Ainda utilizou-se a métrica estatística Cohen's *Kappa* para avaliar a distribuição dos elementos na matriz de confusão. Observa-se que a mesma também apresentou a mesma ordem das anteriores, ficando `ClassicAnalyzer` com 0,884; `StandardAnalyzer` com 0,880; `EnglishAnalyzer` com 0,862; `MyBrazilianAnalyzer` com 0,849 e `BrazilianAnalyzer` com 0,847. Incluiu-se também o Coeficiente de Correlação de Matthew, uma correlação específica para avaliar classificadores e sistemas afins. Mais uma vez os resultados se repetem, ficando `ClassicAnalyzer` com 0,898;

`StandardAnalyzer` com 0,897; `EnglishAnalyzer` com 0,881; `MyBrazilianAnalyzer` com 0,869 e `BrazilianAnalyzer` com 0,866. Ao final observou-se que da mesma forma como nas áreas da Computação e Odontologia, a combinação de `Log-Likelihood` e `ClassicAnalyzer` é sem sombra de dúvidas a melhor opção. Os resultados de todas as avaliações demostraram isso, a única com exceção foi *precision* da classe `ClassicAnalyzer` que ficou em segundo lugar por apenas 0,01.

4.5 AVALIAÇÃO DAS RECOMENDAÇÕES

As recomendações personalizadas foram avaliadas utilizando-se a métrica *coverage* com o parâmetro n variando de 1 até 50 recomendações solicitadas. Adaptou-se a *coverage* da equação 14 ao contexto da abordagem proposta conforme a equação 36.

$$Coverage = \frac{\sum_{i=1}^{p}(C)}{p * n} \tag{36}$$

Onde C é a quantidade total de recomendações geradas para um único tipo de recomendação, p representa o número de pesquisadores e n é a quantidade de recomendações solicitadas.

A métrica *diversity* da equação 18 também está adaptada ao contexto da abordagem proposta, a mesma se caracteriza por ser mais eficiente em termos computacionais devido a adoção de uma generalização. Propôs-se a generalização da mesma para situações onde não se considera diretamente a similaridade entre os itens recomendados. Ao invés disso, utiliza-se a contagem de repetições dos tipos de recomendações (elemento do Rep-Model) definida na equação 37.

$$rep = \sum_{i=1}^{r}(\frac{r_i}{2})(r_i - 1) \tag{37}$$

Onde r_i é um vetor com a contagem de repetição de cada elemento. A nova proposta também necessita da quantidade de elementos em uma matriz triangular (apenas um lado e sem a diagonal principal) definida pela equação 38.

$$t = (\frac{C}{2})(C - 1) \tag{38}$$

Na equação 39 pode-se visualizar o cálculo da *diversity* proposto.

$$diversity = \frac{t - rep}{t} \tag{39}$$

Onde rep é calculado pela equação 37 e t é calculado pela equação 38.

O conjunto total das recomendações (personalizadas e não personalizadas) é avaliado com a média aritmética da métrica *coverage* e a média aritmética da métrica *diversity* com o parâmetro n variando de 1 até 20 recomendações solicitadas.

4.5.1 Experimento 8 - Recomendações para Ciência da Computação

Para realizar os experimentos com as recomendações, utilizou-se o grupo total de 398 pesquisadores com bolsa de produtividade em pequisa do CNPq, os grupos individuais de bolsa 1A (23), 1B (22), 1C (38), 1D (50) e 2 (264); e mais um grupo de teste com 143 pesquisadores. O mesmo é composto por 80 docentes do grupo INF da UFRGS[9] e 63 docentes do DCC da UFMG[10]. Destes, 64 são bolsistas de produtividade do CNPq e 79 não possuem bolsa. Na Figura 27 pode-se visualizar o grafo de similaridades de subárea e reputação da Ciência da Computação.

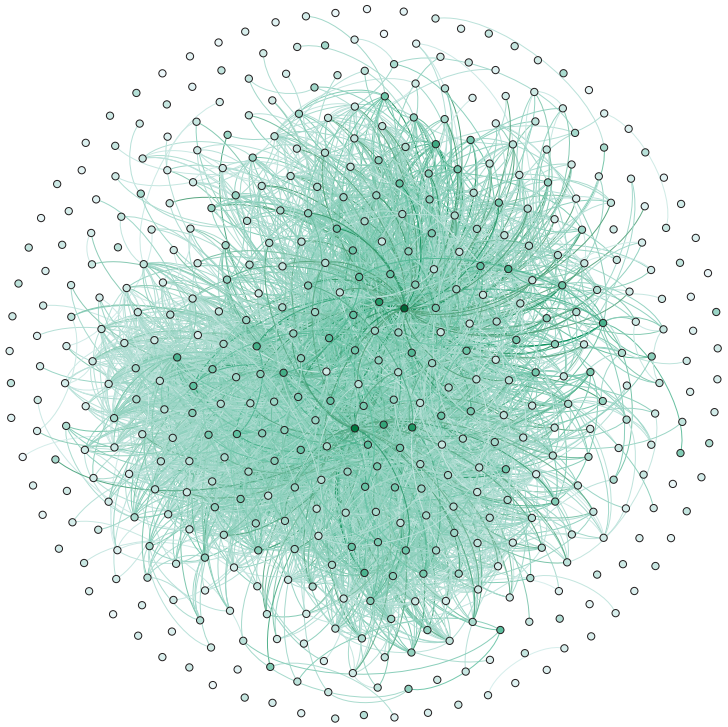

Figura 27. Grafo de Reputação e Similaridades da Ciência da Computação.

No grafo os nodos representam os pesquisadores do CNPq e o grupo de teste. A cor verde interna de cada nodo representa o valor decimal do Rep-Index (experimento 4.3.1), tons mais claros representam valores mais próximos de zero; e tons mais escuros próximos de 100 (valor decimal do Rep-Index, equação 23). As arestas representam a similaridade (experimento 4.4.1) entre as áreas de atuação dos pesquisadores. O tom de cor neste caso representa o valor da similaridade, a única diferença está no intervalo, que neste caso é entre 0 e 1.0, inclusive.

[9] http://www.inf.ufrgs.br/site/pessoas/corpo-docente/
[10] http://www.dcc.ufmg.br/dcc/?q=pt-br/professores

Solicitou-se a geração das recomendações personalizadas para rede de colaboradores e para Grau de Instrução. Na figura 28 pode-se visualizar a métrica *coverage* para n variando de 1 até 50 recomendações.

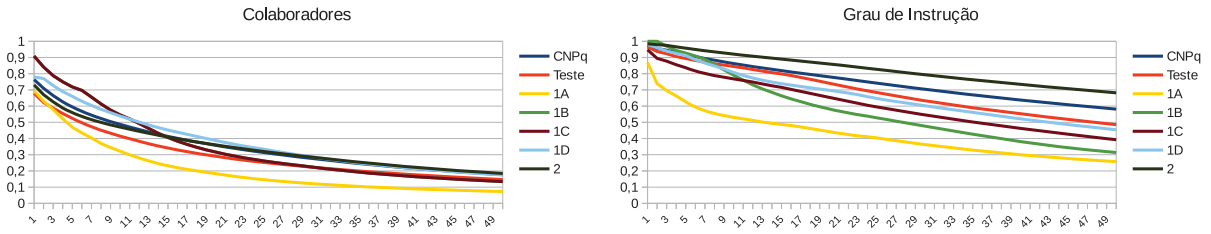

Figura 28. *Coverage* de Recomendações de Colaboradores e Grau de Instrução para Ciência da Computação.

Na figura observa-se que nunca ocorreu a situação onde não existe o que recomendar (valores zerados). Também fica evidente que os grupos iniciais do CNPq, ou seja, grupo 2 e 1D possuem valores maiores de *coverage* do que os grupos mais avançados (1B e 1A). Isto se justifica pelo fato que os grupos finais tem na maioria das vezes maior reputação no Rep-Index do que os iniciais. Ao final calculou-se a média aritmética da métrica *coverage*. A mesma busca mensurar a capacidade de um recomendador em gerar um número n de recomendações. Como resultados, o grupo de pesquisadores do CNPq do nível 2 obteve o maior valor (0,8896); CNPq com 0,8206; 1D com 0,7548; Teste com 0,7444; 1C com 0,6921; 1B com 0,6897 e 1A com 0,5012.

Ao final, gerou-se o conjunto total das 28 possíveis recomendações diferentes (personalizadas e não personalizadas), uma para cada elemento do tipo inteiro do Rep-Model. Neste experimento utilizou-se o limiar de 0,99905 para limitar a similaridade entre os pesquisadores. As recomendações que não incrementam a reputação (Rep-Index) do pesquisador foram desconsideradas, isto significa que os elementos que não tiveram valores para os pesos no experimento 4.3.1 foram ignorados e portanto não são recomendados. Foi computada a média da métrica *coverage* e da *diversity* com n variando de 1 até 20. Na Figura 29 pode-se visualizar as mesmas.

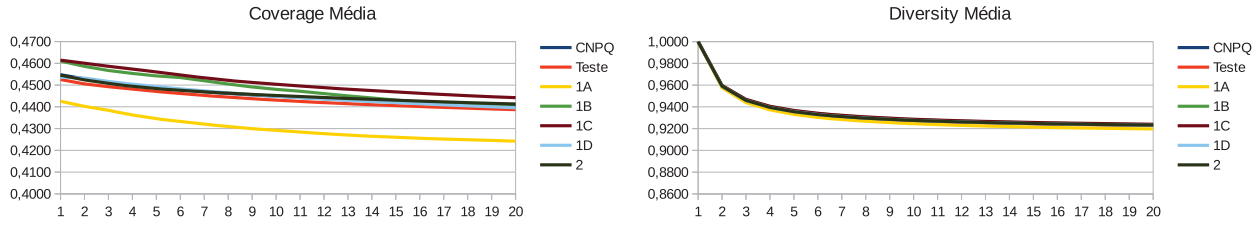

Figura 29. *Coverage* Média e *Diversity* Média para Ciência da Computação.

Observa-se que a média *coverage* decresce lentamente até $n = 20$. A mesma é maior para os níveis iniciais e menor para o grupo 1A. Com relação a média da *diversity*, pode-se constatar que quando $n = 1$ ocorre o máximo de diversidade, para $n = 2$ a mesma cai para aproximadamente 0,96. Quando $n \geq 3$ a mesma começa a crescer positivamente chegando a 0,99 para $n = 20$. Em todos as situações a diversidade de elementos recomendados é satisfatória.

Na Figura 30 pode-se visualizar o conjunto de recomendações geradas com $n = 1$ para um pesquisador do nível SR da Ciência da Computação.

Recomendações - Coverage: 0,4643 - Diversity: 1,0000

Recomendação	Tipo	Peso	Rep-Index	Máximo	Inc.	Aumento
Aumente o item: Orientação de Doutorado (PTA) para 6	REC_PTA	18,865	14,982	46	1	0,41
Aumente o item: Orientação de Pós-doutorado (PA) para 1	REC_PA	5,925	14,982	19	1	0,312
Aumente o item: Membro de Corpo Editorial de Periódico (EBM) para 6	REC_EBM	5,134	14,982	18	1	0,285
Aumente o item: Artigo em Periódico (ASJ) - Qualis A1 para 5.6000000000000005	REC_ASJ	7,207	14,982	41,35	1	0,174
Aumente o item: Livro (BP) para 7	REC_BP	5,829	14,982	57	1	0,102
Aumente o item: Orientação de Mestrado (MDA) para 18	REC_MDA	10,482	14,982	114	1	0,092
Aumente o item: H-Index (HI) para 18	REC_HI	9,674	14,982	116	1	0,083
Aumente o item: Revisão de Periódico (RJ) para 1	REC_RJ	5,782	14,982	77	1	0,075
Aumente o item: Participação em Banca de Mestrado (PEBPT) para 56	REC_PEBPT	9,477	14,982	127	1	0,075
Aumente o item: Prêmios (PREM) para 6	REC_PREM	4,284	14,982	59	1	0,073
Aumente o item: Capítulo de Livro (BCP) para 12	REC_BCP	4,095	14,982	62	1	0,066
Amplie a sua Rede de colaboração com o Pesquisador: 5554254760869075, similaridade: 0,999155 Rep-Index: 33,03	REC_NC	4,93	14,982	158	1	0,031
Aumente o item: Trabalho Completo em Conferência (CWPCP) para 71	REC_CWPCP	8,316	14,982	470	1	0,018

Figura 30. Recomendações para um pesquisador do nível SR da Ciência da Computação.

4.5.2 Experimento 9 - Recomendações para Odontologia

Para realizar os experimentos com as recomendações, utilizou-se o grupo total de 213 pesquisadores com bolsa de produtividade em pequisa do CNPq, os grupos individuais de bolsa 1A (19), 1B (24), 1C (23), 1D (27) e 2 (118); e mais um grupo de teste com 84 pesquisadores. O mesmo é composto por docentes da Faculdade de Odontologia da UFRGS[11]. Destes, 5 são bolsistas de produtividade do CNPq e 79 não possuem bolsa. Na Figura 31 pode-se visualizar o grafo de similaridades de subárea e reputação da Odontologia.

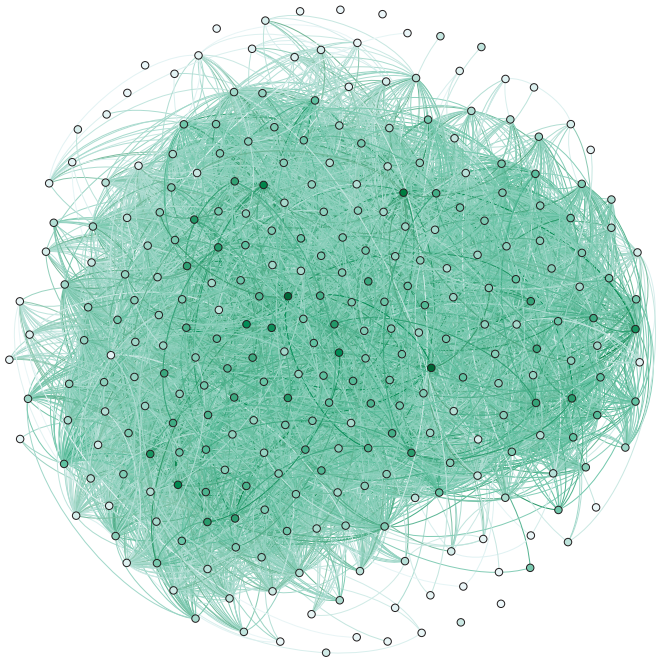

Figura 31. Grafo para Reputação e Similaridades da Odontologia.

[11] http://www.ufrgs.br/odontologia/faculdade/Corpo Docente

No grafo os nodos representam os pesquisadores do CNPq e o grupo de teste. A cor verde interna de cada nodo representa o valor decimal do Rep-Index (experimento 4.3.2), tons mais claros representam valores mais próximos de zero; e tons mais escuros próximos de 100 (valor decimal do Rep-Index, equação 23). As arestas representam a similaridade (experimento 4.4.2) entre as áreas de atuação dos pesquisadores. O tom de cor neste caso representa o valor da similaridade, a única diferença está no intervalo, que neste caso é entre 0 e 1.0, inclusive.

Solicitou-se a geração das recomendações personalizadas para rede de colaboradores e para Grau de Instrução. Na figura 32 pode-se visualizar a *coverage* para n variando de 1 até 50.

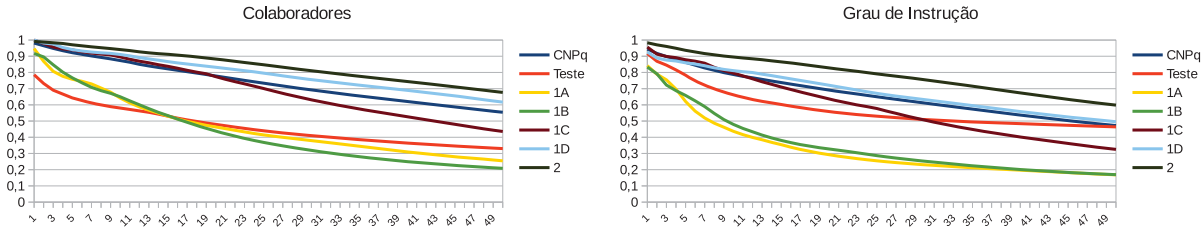

Figura 32. *Coverage* de Recomendações de Colaboradores e Grau de Instrução para Odontologia.

Na figura observa-se que nunca ocorreu a situação onde não existe o que recomendar (valores zerados). Também fica evidente que os grupos iniciais do CNPq, ou seja, grupo 2 e 1D possuem valores maiores de *coverage* do que os grupos mais avançados (1B e 1A). Isto se justifica pelo fato que os grupos finais tem na maioria das vezes maior reputação no Rep-Index do que os iniciais. Ao final calculou-se a média aritmética da métrica *coverage*. A mesma busca mensurar a capacidade de um recomendador em gerar um número n de recomendações. Como resultados, o grupo de pesquisadores do CNPq do nível 2 obteve o maior valor (0,8896); CNPq com 0,8206; 1D com 0,7548; Teste com 0,7444; 1C com 0,6921; 1B com 0,6897 e 1A com 0,5012.

Ao final, gerou-se o conjunto total das 28 possíveis recomendações diferentes (personalizadas e não personalizadas), uma para cada elemento do tipo inteiro do Rep-Model. Neste experimento utilizou-se o limiar de 0,99905 para limitar a similaridade entre os pesquisadores. As recomendações que não incrementam a reputação (Rep-Index) do pesquisador foram desconsideradas, isto significa que os elementos que não tiveram valores para os pesos no experimento 4.3.2 foram ignorados e portanto não são recomendados. Foi computada a média da métrica *coverage* e da *diversity* com n variando de 1 até 20. Na Figura 33 pode-se visualizar as mesmas.

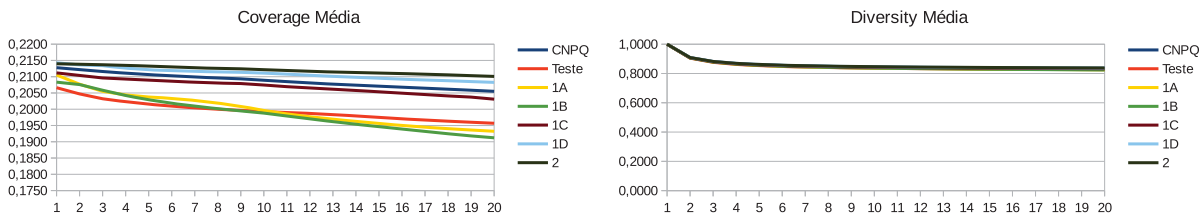

Figura 33. *Coverage* Média e *Diversity* Média para Odontologia.

Observa-se que a média da *coverage* decresce lentamente até $n = 20$. Os níveis 1A e 1B apresentam valores menores que os demais grupos. Com relação a média da *diversity*, pode-se constatar que quando $n = 1$ ocorre o máximo de diversidade, para $n = 2$ a mesma cai para aproximadamente 0,90. Quando $n \geq 3$ a mesma começa a crescer positivamente chegando a 0,98 para $n = 20$. Em todos as situações a diversidade de elementos recomendados é satisfatória.

Na Figura 34 pode-se visualizar o conjunto de recomendações geradas com $n = 1$ para um pesquisador do nível SR da Odontologia.

Recomendações - Coverage: 0,2143 - Diversity: 1,0000

Recomendação	Tipo	Peso	Rep-Index	Máximo	Inc.	Aumento
Aumente o item: Orientação de Pós-doutorado (PA) para 1	REC_PA	13,256	19,986	18	1	0,736
Aumente o item: Orientação de Doutorado (PTA) para 18	REC_PTA	15,961	19,986	45	1	0,355
Aumente o item: H-Index (HI) para 23	REC_HI	26,408	19,986	135	1	0,196
Aumente o item: Artigo em Periódico (ASJ) - Qualis A1 para 10.900000000000006	REC_ASJ	17,259	19,986	143,25	1	0,12
Aumente o item: Participação em Banca de Mestrado (PEBPT) para 39	REC_PEBPT	12,658	19,986	118	1	0,107
Amplie a sua Rede de colaboração com o Pesquisador: 5172418667327780, similaridade: 0,999594 Rep-Index: 47,48	REC_NC	14,459	19,986	314	1	0,046

Figura 34. Recomendações para um pesquisador do nível SR da Odontologia.

4.5.3 Experimento 10 - Recomendações para Economia

Para realizar os experimentos com as recomendações, utilizou-se o grupo total de 202 pesquisadores com bolsa de produtividade em pequisa do CNPq, os grupos individuais de bolsa 1A (13), 1B (13), 1C (12), 1D (40) e 2 (123); e mais um grupo de teste com 51 pesquisadores. O mesmo é composto por docentes do grupo FEA da USP[12]. Destes, 14 são bolsistas de produtividade do CNPq e 37 não possuem bolsas. Na Figura 35 visualiza-se o grafo de similaridades de subárea e reputação da Economia.

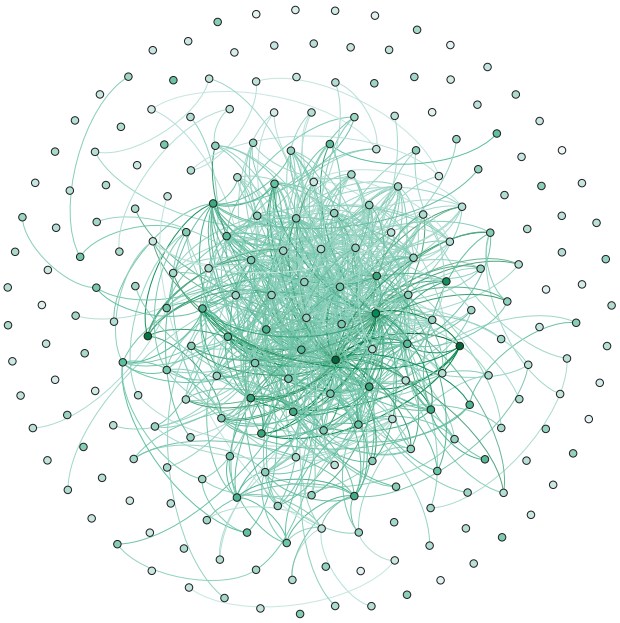

Figura 35. Grafo para Reputação e Similaridades da Economia.

[12]http://www.fea.usp.br/economia/pessoas/corpo-docente

No grafo os nodos representam os pesquisadores do CNPq e o grupo de teste. A cor verde interna de cada nodo representa o valor decimal do Rep-Index (experimento 4.3.3), tons mais claros representam valores mais próximos de zero; e tons mais escuros próximos de 100 (valor decimal do Rep-Index, equação 23). As arestas representam a similaridade (experimento 4.4.3) entre as áreas de atuação dos pesquisadores. O tom de cor neste caso representa o valor da similaridade, a única diferença está no intervalo, que neste caso é entre 0 e 1.0, inclusive.

Solicitou-se a geração das recomendações personalizadas para rede de colaboradores e para Grau de Instrução. Na figura 36 pode-se visualizar a *coverage* para n variando de 1 até 50.

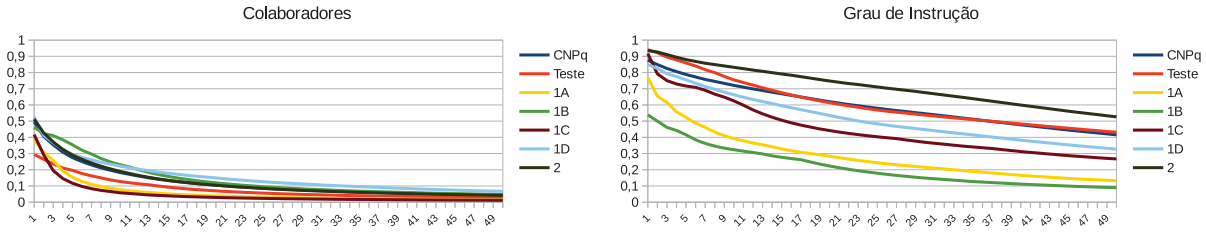

Figura 36. *Coverage* de Recomendações de Colaboradores e Grau de Instrução para Economia.

Na figura observa-se que nunca ocorreu a situação onde não existe o que recomendar (valores zerados). Também fica evidente que os grupos iniciais do CNPq, ou seja, grupo 2 e 1D possuem valores maiores de *coverage* do que os grupos mais avançados (1B e 1A). Isto se justifica pelo fato que os grupos finais tem na maioria das vezes maior reputação no Rep-Index do que os iniciais. Ao final calculou-se a média aritmética da métrica *coverage*. A mesma busca mensurar a capacidade de um recomendador em gerar um número n de recomendações. Como resultados, o grupo de pesquisadores do CNPq do nível 2 obteve o maior valor (0,8896); CNPq com 0,8206; 1D com 0,7548; Teste com 0,7444; 1C com 0,6921; 1B com 0,6897 e 1A com 0,5012.

Ao final, gerou-se o conjunto total das 28 possíveis recomendações diferentes (personalizadas e não personalizadas), uma para cada elemento do tipo inteiro do Rep-Model. Neste experimento utilizou-se o limiar de 0,99905 para similaridade dos pesquisadores. As recomendações que não incrementam a reputação (Rep-Index) do pesquisador foram desconsideradas, isto significa que os elementos que não tiveram valores para os pesos no experimento 4.3.3 foram ignorados e portanto não são recomendados. Foi computada a média da *coverage* e da *diversity* com n variando de 1 até 20. A primeira indica a capacidade da abordagem em gerar recomendações e a última avalia a diversidade apenas do conjunto de itens recomendados. Na Figura 37 pode-se visualizar as mesmas.

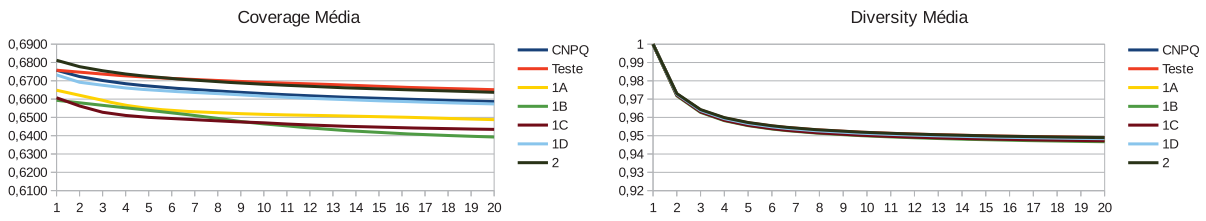

Figura 37. *Coverage* Média e *Diversity* Média para Economia.

Observa-se que novamente a *coverage* apresentou valores menores para o níveis finais do CNPq. Com relação a média da *diversity*, pode-se constatar que quando $n = 1$ ocorre o máximo de diversidade, para $n = 2$ a mesma cai para aproximadamente 0,97. Quando $n \geq 3$ a mesma começa a crescer positivamente chegando a 0,995 para $n = 20$. Em todos as situações a diversidade de elementos recomendados é satisfatória.

Na Figura 38 pode-se visualizar o conjunto de recomendações geradas com $n = 1$ para um pesquisador do nível SR da Economia.

Recomendações - Coverage: 0,7500 - Diversity: 1,0000

Recomendação	Tipo	Peso	Rep-Index	Máximo	Inc.	Aumento
Aumente o item: Marca (MARC) para 1	REC_MARC	0,397	32,408	1	1	0,397
Aumente o item: Orientação de Pós-doutorado (PA) para 1	REC_PA	1,784	32,408	7	1	0,255
Aumente o item: Produto Tecnológico (PRODTEC) para 1	REC_PROD...	0,966	32,408	4	1	0,242
Aumente o item: Membro de Corpo Editorial de Periódico (EBM) para 4	REC_EBM	4,993	32,408	21	1	0,238
Aumente o item: Software (SOFT) para 1	REC_SOFT	0,936	32,408	5	1	0,187
Aumente o item: Artigo em Periódico (ASJ) - Qualis A1 para 6.149999999999998	REC_ASJ	4,831	32,408	26,1	1	0,185
Aumente o item: Capítulo de Livro (BCP) para 33	REC_BCP	10,409	32,408	66	1	0,158
Aumente o item: Prêmios (PREM) para 12	REC_PREM	6,39	32,408	47	1	0,136
Aumente o item: Participação em Banca de Mestrado (PEBPT) para 37	REC_PEBPT	8,049	32,408	75	1	0,107
Aumente o item: Revisão de Periódico (RJ) para 1	REC_RJ	6,118	32,408	58	1	0,105
Aumente o item: Orientação de Doutorado (PTA) para 28	REC_PTA	6,339	32,408	80	1	0,079
Aumente o item: H-Index (HI) para 36	REC_HI	5,334	32,408	74	1	0,072
Aumente o item: Membro de Comitê de Conferência (CCM) para 1	REC_CCM	3,678	32,408	54	1	0,068
Aumente o item: Livro (BP) para 12	REC_BP	3,054	32,408	53	1	0,058
Aumente o item: Orientação de Mestrado (MDA) para 59	REC_MDA	8,257	32,408	144	1	0,057
Aumente o item: Participação em Banca de Doutorado (PEBMD) para 29	REC_PEBMD	7,559	32,408	177	1	0,043
Aumente o item: Trabalho Completo em Conferência (CWPCP) para 63	REC_CWPCP	4,848	32,408	186	1	0,026
Aumente o item: Trabalho Técnico (TT) para 1	REC_TT	3,877	32,408	270	1	0,014
Aumente o item: Processo ou Técnicas (PROCTEC) para 1	REC_PROC...	0,028	32,408	3	1	0,009
Aumente o item: Projeto de Pesquisa (RP) para 1	REC_RP	1,589	32,408	230	1	0,007
Aumente o item: Coordenação de Comitê de Conferência (CCC) para 1	REC_CCC	0,006	32,408	1	1	0,006

Figura 38. Recomendações para um pesquisador do nível SR da Economia.

4.6 ANÁLISE DE RESULTADOS

A partir dos 10 experimentos realizados, foi possível ajustar, avaliar e validar a abordagem proposta. O experimento 4.2.1 verificou a evolução e consistência dos dados utilizados para os próximos experimentos dessa abordagem. Observou-se que a grande maioria dos elementos do modelo tiveram incremento no seu quantitativo, indicando que a produção geral das três áreas apresentou crescimento mesmo com a alteração de pesquisadores e bolsas.

Os experimentos 4.3.1, 4.3.2 e 4.3.3 possibilitaram ajustar os pesos do Rep-Index para que o índice reflita o mais próximo possível a realidade da área. Ao final dos experimentos, as três áreas apresentaram correlações entre os níveis do Rep-Index e níveis de bolsa de produtividade do CNPq maiores do que a proposta de pesos original do Rep-Index. Para a Ciência da Computação os pesos propostos com o algoritmo `SymmetricalUncert` obtiveram os melhores resultados, para Odontologia foi o `InfoGain` e para a Economia o `ReliefF`. Observou-se que no caso específico da Economia o melhor resultado foi de apenas 0,3427 para `ReliefF`, tal valor é ligeiramente inferior ao encontrado para Ciência da Computação (0,5948) e Odontologia (0,5909). Tal fato justifica-se devido aos dados da

Economia apresentarem uma distribuição estatística mais heterogênea e dependências fracas entre os elementos do Rep-Model.

Os experimentos 4.4.1, 4.4.2 e 4.4.3 permitiram identificar qual o melhor conjunto de técnicas para realizar a mineração de texto (análise léxica, *stemming*, *stopwords*, sinonímia e função de similaridade) nos elementos textuais do Rep-Model, dessa forma é possível localizar os pesquisadores com as maiores afinidades de subárea de atuação. A combinação `Log-Likelihood` e `ClassicAnalyzer` do Apache Mahout apresentou os melhores resultados para as três áreas nos experimentos. Portanto, adota-se a mesma como padrão de fato para a abordagem proposta. Cabe destacar que a maioria dos autores considera o Cosseno como função padrão para a TF-IDF, neste sentido pode-se afirmar que as métricas *Precision* e *Recall* do `Log-Likehood` são estatisticamente minimamente superior ao *baseline* estabelecido.

Ainda sobre as avaliações da mineração de texto, a equação 31 proposta, corrigiu as discrepâncias observadas anteriormente nas métricas MAE e RMSE. A mesma considera inicialmente a normalização das distâncias e posteriormente a sua conversão em similaridades. A classe em Java para `Log-Likelihood` (Apêndice C) adaptada para vetores esparsos do Apache Mahout também é outro melhoramento proposto pelo presente trabalho.

A geração das recomendações e a validação da abordagem foram realizadas nos experimentos 4.5.1, 4.5.2 e 4.5.3. Para gerar as recomendações utilizou-se os grupos de pesquisadores do CNPq e um grupo de teste formado por pesquisadores de um ou mais programas de pós-graduação da área. Utilizou-se a métrica *coverage* e *diversity* para avaliar as recomendações. Sobre a métrica *diversity*, a mesma apresentou valores altos em todas as situações de testes, situação considerada plenamente satisfatória. Na figura 39 pode-se observar a média para a *coverage* em cada nível de área.

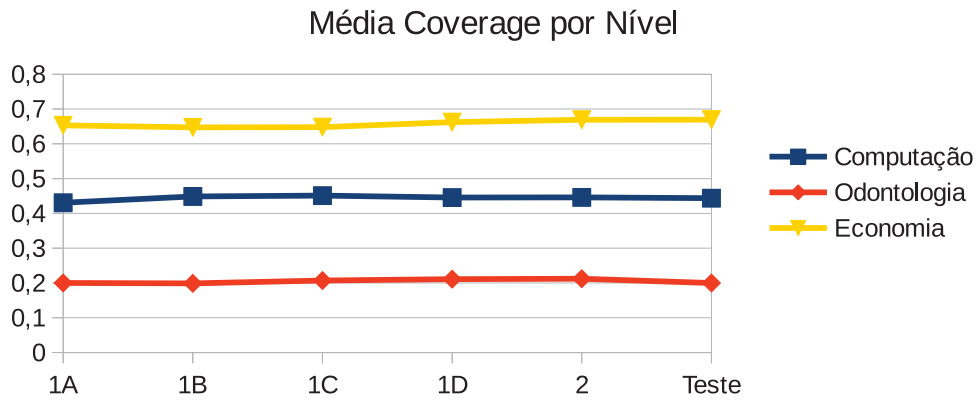

Figura 39. Média para a *coverage* em cada nível de área.

Observa-se que os valores são praticamente constantes para cada área, dessa forma, optou-se por normalizar os dados com o intuito de realizar uma melhor comparação dos mesmos. Na figura 40 pode-se visualizar a alteração proposta.

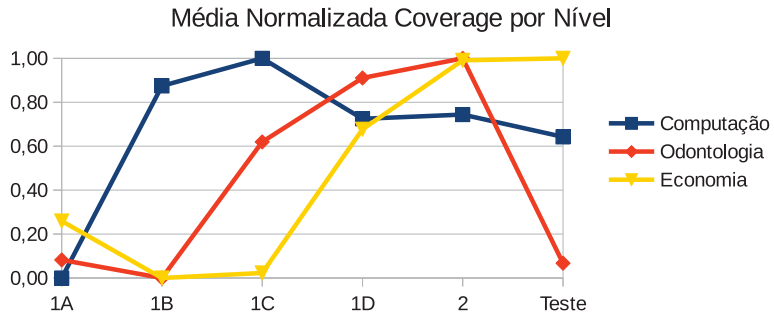

Figura 40. Normalização da média da *coverage*.

Observa-se que a média da *coverage* é no geral maior nos nível iniciais: 2, 1D e 1C e Teste. Isto indica que estes níveis tem maior número de recomendações e portanto maiores benefícios com a abordagem proposta. O grupo de teste da Odontologia apresentou baixa *coverage*, isto indica que o grupo selecionado para os testes apresenta alta reputação em comparação aos demais grupos de testes. Além disso, os níveis 1B e 1C da Ciência da Computação apresentaram *coverage* maior que os níveis 1D e 2. Isto novamente indica uma reputação maior dos níveis iniciais em relação aos níveis 1B e 1C.

Durante a avaliação das recomendações, observou-se uma propriedade na abordagem. Trata-se da forte correlação existente entre a média da *coverage* e a quantidade de elementos com pesos. Na figura 41 pode-se visualizar a mesma.

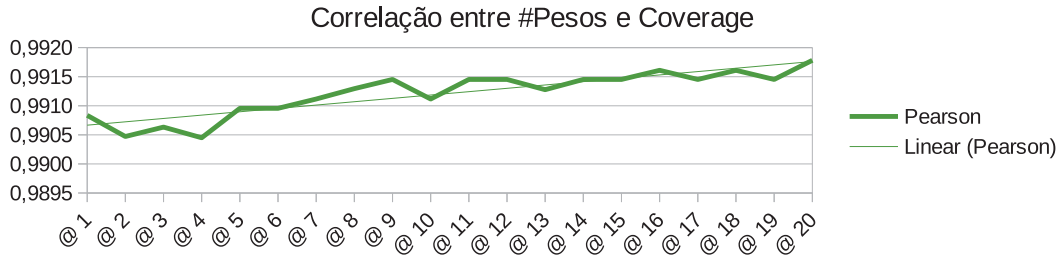

Figura 41. Correlação entre a média da *coverage* e a quantidade de elementos com pesos.

A partir da análise da figura anterior, pode afirmar que a média da *Coverage* é diretamente proporcional a quantidade de elementos com pesos. Portanto, para obter maiores valores de *coverage* é necessário que a quantidade de pesos para o Rep-Index seja a maior possível. Esta propriedade enfatiza ainda mais a relevância da etapa de cálculo dos pesos para a abordagem proposta neste trabalho.

Cabe ainda ressaltar que caso deseja-se aumentar a *coverage*, pode-se utilizar o conjunto de pesos propostos com a técnica `ReliefF`, este algoritmo apresenta com característica não descartar atributos, dessa forma todos os elementos do modelo possuem pesos. Nas comparações entre os conjuntos de pesos optou-se pelo melhor resultado, dessa forma o `ReliefF` foi utilizado apenas para Economia, contudo obteve melhores resultados nas três áreas em comparação com o conjunto original de pesos para o Rep-Index.

5. CONCLUSÃO

Neste capítulo, são apresentadas as conclusões, destacando-se os objetivos, as contribuições e os resultados alcançados por esta abordagem. Além disso, são apresentadas as publicações e *softwares* que foram desenvolvidos durante o curso de mestrado. Por fim, são discutidas algumas sugestões de trabalhos futuros identificados ao longo do desenvolvimento desta abordagem.

5.1 OBJETIVOS

O objetivo deste trabalho foi de apresentar uma abordagem para recomendação de plano de carreira de pesquisadores com base na personalização, similaridade de perfil e reputação acadêmica.

Inicialmente, se propôs um modelo de perfil/reputação dos pesquisadores. Posteriormente, adaptou-se uma medida de similaridade para comparar o perfil de pesquisadores. Implementou-se uma solução computacional para realizar as recomendações de atividades a partir do modelo proposto e da medida de similaridade adotada. As recomendações foram geradas por dois processos distintos, e ao final combinadas e avaliadas em experimentos específicos.

Utilizou-se dados dos bolsistas de produtividade em pesquisa do CNPq para as áreas da Ciência da Computação, Odontologia e Economia. A escolha dessas três áreas baseou-se na classificação das grandes áreas do CNPq e dos três grandes colégios da CAPES. A avaliação com áreas distintas do conhecimento é fundamental para constar a adaptabilidade e flexibilidade da abordagem proposta. Além dos bolsistas do CNPq, utilizou-se grupos de testes formados por pesquisadores de programas de pós-graduação das três áreas citadas anteriormente.

5.2 RESULTADOS

Os experimentos para ajustar os pesos do Rep-Index destacaram a relevância do complemento proposto por Vivian, Cervi e Rovadosky[77]. Obteve-se uma nova combinação de pesos para as três áreas com classificações melhores do que Rep-Index original e enfatizando as características de cada área. Este resultado advêm principalmente do fato que o Rep-Index original ter sido ajustado para as três áreas em conjunto, por outro lado, neste trabalho, procurou-se torná-lo mais específico com o emprego de técnicas de aprendizado de máquina e mineração de dados. Observou-se a inclusão do elemento prêmio (PREM) para a Ciência da Computação e Economia. Além disso, marca (MARC), produto tecnológico (PRODTEC), trabalho técnico (TT) e processo técnico (PROCTEC) foram incluídos para Economia. Observou-se também que a quantidade de elementos com pesos influência diretamente a *coverage* da abordagem de recomendação proposta.

No âmbito da mineração de texto, os treinamentos foram conduzidos com a sub área de atuação do currículo Lattes, os resultados foram semelhantes para as três áreas em estudo. A combinação `Log-likelihood` e `ClassicalAnalizer` do Apache Mahout obteve os melhores resultados para

precison, *recall*, *kappa*, MCC, MAE e RMSE. Posteriormente se calculou as similaridades dos pesquisadores a partir dos elementos textuais (TPB, TPT, RC, AA, TO, TB e TOA) adicionados ao Rep-Model neste trabalho. Além disso, a proposta de normalizar as distâncias e posteriormente convertê-las em similaridades possibilitou analisar adequadamente os resultados das métricas de erro MAE e RMSE.

Com relação as recomendações, gerou-se o conjunto das 28 possíveis recomendações previstas no Rep-Model. As recomendações para ED e NC utilizaram a abordagem personalizada, as demais são todas não personalizadas. As recomendações de elementos que não possuem peso e que portanto não aumentam a reputação foram desconsideradas. Além dos pesquisadores do CNPq, utilizou-se grupos de testes formados por pesquisadores de programas de pós graduação. Os resultados apresentaram boa cobertura e ótima diversidade. Em geral, os pesquisadores dos níveis iniciais e de testes obtiveram resultados ligeiramente maiores pois possuem no geral menor reputação quando comparados com os pesquisadores de níveis mais avançados do CNPq.

Nas três áreas, os elementos CP, CR, DI, PAT, TCI não aumentam a reputação dos pesquisadores. No caso da Ciência da Computação, inclui-se os elementos ED, PEBMD, CCC, CCM, RP, SOFT, MARC, PRODTEC, TT e PROCTEC. Na Odontologia inclui-se o elementos ED, MDA, PEBMD, EBM, RJ, CCC, CCM, BP, BCP, CWPCP, RP, SOFT, MARC, PRODTEC, TT, PROCTEC e PREM.

5.3 CONTRIBUIÇÕES

Pode-se destacar como principais contribuições deste trabalho os seguinte itens:

- Uma abordagem inédita para recomendação de carreira pesquisadores.
- Nova proposta para a determinação dos pesos do Rep-Index utilizando técnicas de aprendizado máquina e estatística.
- Adição de novos elementos ao Rep-Model e determinação de seus pesos para a área.
- Adição do Qualis Periódicos para o elemento ASJ do Rep-Index.
- Proposta de normalização das distância em similaridades para corrigir discrepâncias nas métricas MAE e RMSE.
- Adição do método `Log-Likelihood` para vetores esparsos no Apache Mahout.
- Nova proposta para calcular a diversidade utilizando a contagem de repetição de tipos ao invés da similaridade de itens.

5.4 PUBLICAÇÕES

Apresenta-se a seguir as publicações desenvolvidas durante o curso de mestrado e que possuem relação com os temas estudados durante este trabalho:

- VIVIAN, G. R.; CERVI, C. R. xml2arff: Uma ferramenta automatizada de extração de dados em arquivos xml para data science com weka e r. XII Escola Regional de Informática de Banco de Dados, p. 159–162, 2016. ISSN 2177-4226.[71]

- VIVIAN, G. R.; CERVI, C. R. Utilizando técnicas de data science para definir o perfil do pesquisador brasileiro da área de ciência da computação. XII Escola Regional de Informática de Banco de Dados, p. 108–117, 2016. ISSN 2177-4226.[70]

- VIVIAN, G. R.; CERVI, C. R.; ROVADOSKY, W. Using selection attribute algorithms from data mining to complement the rep-index. IADIS International Journal on WWW/Internet, IADIS, v. 15, p. 219–226, 2016. ISSN 1645-7641.[77]

- VIVIAN, G. R.; CERVI, C. R. MahoutGUI: Uma Interface Gráfica para Gerar Recomendações com o Apache Mahout Diretamente de Banco de Dados usando Mapeamento Objeto-Relacional. XIII Escola Regional de Informática de Banco de Dados, p. 79–82, 2017. ISSN 2177-4226.[93]

- VIVIAN, G. R.; CERVI, C. R. Uma Proposta de Abordagem de Recomendação para Carreira de Pesquisadores Baseada em Personalização, Similaridade de Perfil e Reputação Acadêmica. XIII Escola Regional de Informática de Banco de Dados, p. 7–16, 2017. ISSN 2177-4226.[94]

5.5 SOFTWARES DESENVOLVIDOS

Apresenta-se a seguir os *softwares* desenvolvidas durante o curso de mestrado e que possuem relação com os temas estudados durante este trabalho:

- Xml2Arrf, Registrado no INPI, processo Nº: BR 51 2016 001072-0, data expedição: 10 de janeiro de 2017.

- MahoutGUI, Registrado no INPI, processo Nº: BR 51 2017 000696-2, data expedição: 04 de junho de 2017.

5.6 SUGESTÕES DE TRABALHOS FUTUROS

Sugere-se com trabalhos futuros as seguintes possibilidades:

- Pode-se experimentar outras técnicas para melhorar os resultados do cálculo dos pesos para o Rep-Index. A regressão linear logística e as redes neurais são ótimos candidatos para obter melhores classificações dos pesquisadores. Contudo, o paradigma da reputação deixa de ser representado por apenas um único número e passa a ser por um conjunto de equações ou modelos computacionais abstratos.

- Pode-se ampliar os experimentos para todas as áreas de pesquisa definidas pelo CNPq e CAPES.

- Pode-se propor novos tipos de recomendações personalizadas, que no geral, são mais apreciadas pelos usuários.
- Pode-se integrar as recomendações com uma plataforma *online* ou rede social de pesquisadores. Deste modo, os pesquisadores terão acesso imediato as recomendações. A avaliação das recomendações também pode ser executada de forma *online*.
- Pode-se dividir (*split*) de forma temporal os dados em duas partes, a partir disso gerar as recomendações com base na primeira parte e comparar com segunda parte usando a *precision* e *recall*.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] STATISTICS. Statistics and market data about the internet. *The Statistics Portal*. Disponível em: <http://www.statista.com/markets/424/internet/>. Acessado em: 18/04/2016.

[2] RIGHETTI, S. Brasil cresce em produção científica, mas índice de qualidade cai. *Folha de São Paulo*, abril 2013. Disponível em: <http://www1.folha.uol.com.br/ciencia/2013/04/1266521-brasil-cresce-em-producao-cientifica-mas-indice-de-qualidade-cai.shtml>. Acessado em: 18/04/2016.

[3] BANK, T. W. Scientific and technical journal articles. *The World Bank*, abril 2016. Disponível em: <http://data.worldbank.org/indicator/IP.JRN.ARTC.SC/countries?display=default>. Acessado em: 18/04/2016.

[4] BRUNIALTI, L. F. et al. Aprendizado de máquina em sistemas de recomendação baseados em conteúdo textual: Uma revisão sistemática. *XI Brazilian Symposium on Information System, Goiânia, GO, May 26-29*, 2015.

[5] GOLDBERG, D. et al. Using collaborative filtering to weave an information tapestry. *Communications of the Association of Computing Machinery*, ACM, v. 35, n. 12, p. 61–70, 1992.

[6] RESNICK, P.; VARIAN, H. R. Recommender systems. *Communications of the ACM*, ACM, v. 40, n. 3, p. 56–58, 1997.

[7] SCHAFER, J. B.; KONSTAN, J.; RIEDL, J. Recommender systems in e-commerce. In: ACM. *Proceedings of the 1st ACM conference on Electronic commerce*. [S.l.], 1999. p. 158–166.

[8] BOBADILLA, J. et al. Recommender systems survey. *Knowledge-Based Systems*, Elsevier B.V., v. 46, p. 109–132, 2013. ISSN 09507051. Disponível em: <http://dx.doi.org/10.1016/j.knosys.2013.03.012>.

[9] FRIEDMAN, N.; GEIGER, D.; GOLDSZMIDT, M. Bayesian network classifiers. *Machine learning*, Springer, v. 29, n. 2-3, p. 131–163, 1997.

[10] PARK, M.-H.; HONG, J.-H.; CHO, S.-B. Location-Based Recommendation System Using Bayesian User's Preference Model in Mobile Devices. In: ____. *Ubiquitous Intelligence and Computing: 4th International Conference, UIC 2007, Hong Kong, China, July 11-13, 2007. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 1130–1139. ISBN 978-3-540-73549-6. Disponível em: <http://dx.doi.org/10.1007/978-3-540-73549-6_110>.

[11] ROH, T. H.; OH, K. J.; HAN, I. The collaborative filtering recommendation based on {SOM} cluster-indexing {CBR}. *Expert Systems with Applications*, v. 25, n. 3, p. 413–423, 2003. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417403000678>.

[12] YAGER, R. R. Fuzzy logic methods in recommender systems. *Fuzzy Sets and Systems*, v. 136, n. 2, p. 133–149, 2003. ISSN 0165-0114. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0165011402002233>.

[13] HO, Y.; FONG, S.; YAN, Z. A hybrid ga-based collaborative filtering model for online recommenders. *ICE-B 2007: Proceedings of the Second International Conference on e-Business*, n. October, p. 200–203, 2007.

[14] HOFMANN, T. Latent semantic models for collaborative filtering. *ACM Trans. Inf. Syst.*, ACM, New York, NY, USA, v. 22, n. 1, p. 89–115, jan. 2004. ISSN 1046-8188. Disponível em: <http://doi.acm.org/10.1145/963770.963774>.

[15] LANGSETH, H.; NIELSEN, T. D. A latent model for collaborative filtering. *International Journal of Approximate Reasoning*, v. 53, n. 4, p. 447–466, 2012. ISSN 0888-613X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0888613X11001654>.

[16] ZHONG, J.; LI, X. Unified collaborative filtering model based on combination of latent features. *Expert Systems with Applications*, v. 37, n. 8, p. 5666–5672, 2010. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417410000837>.

[17] KOREN, Y.; BELL, R.; VOLINSKY, C. Matrix factorization techniques for recommender systems. *Computer*, IEEE, n. 8, p. 30–37, 2009.

[18] GUO, M.-j.; SUN, J.-g.; MENG, X.-f. A neighborhood-based matrix factorization technique for recommendation. *Annals of Data Science*, Springer, p. 1–16, 2015.

[19] 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*, v. 17, n. 6, p. 734–749, jun 2005. ISSN 10414347. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1423975>.

[20] SILVA, E. Q. D. *Um método evolucionário para combinar resultados das técnicas de sistemas de recomendação baseado em filtragem colaborativa*. Tese (Doutorado) — UNIVERSIDADE FEDERAL DE GOIÁS, 2014.

[21] WU, H. et al. Content embedding regularized matrix factorization for recommender systems. In: *2017 IEEE International Congress on Big Data (BigData Congress)*. [S.l.: s.n.], 2017. p. 209–215.

[22] KOREN, Y.; BELL, R. Advances in collaborative filtering. In: *Recommender Systems Handbook*. [S.l.]: Springer, 2015. p. 77–118.

[23] BENNETT, J.; LANNING, S. The netflix prize. In: *Proceedings of KDD cup and workshop*. [S.l.: s.n.], 2007. p. 35.

[24] BELL, R. M.; KOREN, Y.; VOLINSKY, C. *The BellKor solution to the Netflix prize*. 2007.

[25] BELL, R. M.; KOREN, Y.; VOLINSKY, C. The bellkor 2008 solution to the netflix prize. *Statistics Research Department at AT&T Research*, 2008.

[26] KOREN, Y. The bellkor solution to the netflix grand prize. 2009. Disponível em: <http://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf>.

[27] BARTH, F. J. Modelando o perfil do usuário para a construção de sistemas de recomendação: um estudo teórico e estado da arte. *Revista de Sistemas de Informação da FSMA*, v. 6, p. 59–71, 2010.

[28] CAZELLA, S. C.; NUNES, M.; REATEGUI, E. A ciência da opinião: Estado da arte em sistemas de recomendação. *JAI Jorn. Atualização em Informática da SBC. Rio Janeiro, RJ PUC Rio*, p. 161–216, 2010.

[29] YING, A. T.; ROBILLARD, M. P. Developer profiles for recommendation systems. In: *Recommendation Systems in Software Engineering*. [S.l.]: Springer, 2014. p. 199–222.

[30] RESNICK, P. et al. Reputation systems. *Commun. ACM*, v. 43, n. 12, p. 45–48, 2000. ISSN 0001-0782. Disponível em: <http://dx.doi.org/10.1145/355112.355122>.

[31] REIN, G. L. Reputation information systems: A reference model. In: IEEE. *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*. [S.l.], 2005. p. 26a–26a.

[32] SPEARMAN, C. The proof and measurement of association between two things. *The American journal of psychology*, JSTOR, v. 15, n. 1, p. 72–101, 1904.

[33] 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 CSCW 02*. New York, New York, USA: ACM Press, 2002. p. 116. ISBN 1581135602. ISSN 1581135602.

[34] EKSTRAND, M. D. et al. Automatically Building Research Reading Lists. *RecSys2010*, p. 159–166, 2010. Disponível em: <http://files.grouplens.org/papers/reading-lists.pdf>.

[35] WANG, C.; BLEI, D. M. Collaborative topic modeling for recommending scientific articles. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*. New York, New York, USA: ACM Press, 2011. p. 448. ISBN 9781450308137. ISSN 14710072. Disponível em: <http://dl.acm.org/citation.cfm?doid=2020408.2020480>.

[36] BEEL, J. et al. A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. In: *RecSys RepSys 2013: Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*. New York, New York, USA: ACM Press, 2013. p. 7–14. ISBN 9781450324656. ISSN 15525996.

[37] GUNAWARDANA, A.; SHANI, G. A survey of accuracy evaluation metrics of recommendation tasks. *J. Mach. Learn. Res.*, JMLR.org, v. 10, p. 2935–2962, dez. 2009. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=1577069.1755883>.

[38] AVAZPOUR, I. et al. Recommendation systems in software engineering. In: ____. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. cap. Dimensions and Metrics for Evaluating Recommendation Systems, p. 245–273. ISBN 978-3-642-45135-5. Disponível em: <http://dx.doi.org/10.1007/978-3-642-45135-5_10>.

[39] HUANG, W. et al. Recommending citations. *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*, p. 1910, 2012. Disponível em: <http://dl.acm.org/citation.cfm?doid=2396761.2398542>.

[40] MIDDLETON, S. E.; SHADBOLT, N. R.; De Roure, D. C. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems*, v. 22, n. 1, p. 54–88, jan 2004. ISSN 10468188. Disponível em: <http://portal.acm.org/citation.cfm?doid=963770.963773>.

[41] KLEINBERG, J. M. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, ACM, v. 46, n. 5, p. 604–632, 1999.

[42] PAGE, L. et al. The pagerank citation ranking: bringing order to the web. *Stanford InfoLab*, 1999.

[43] ZHANG, Z.; LI, L. A research paper recommender system based on spreading activation model. In: *The 2nd International Conference on Information Science and Engineering*. IEEE, 2010. p. 928–931. ISBN 978-1-4244-7616-9. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5689417>.

[44] OHTA, M.; HACHIKI, T.; TAKASU, A. Related paper recommendation to support online-browsing of research papers. In: *4th International Conference on the Applications of Digital Information and Web Technologies, ICADIWT 2011*. IEEE, 2011. p. 130–136. ISBN 9781424498246. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6041413>.

[45] ZHANG, W.; YOSHIDA, T.; TANG, X. A comparative study of tf*idf, lsi and multi-words for text classification. *Expert Systems with Applications*, Elsevier, v. 38, n. 3, p. 2758–2765, 2011. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2010.08.066>.

[46] SUGIYAMA, K.; KAN, M.-Y. Serendipitous Recommendation for Scholarly Papers Considering Relations Among Researchers. In: *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*. New York, New York, USA: ACM Press, 2011. p. 307–310. ISBN 9781450307444. ISSN 15525996. Disponível em: <http://portal.acm.org/citation.cfm?doid=1998076.1998133>.

[47] BEEL, J. et al. Research paper recommender system evaluation: A Quantitative Literature Survey. In: *Proceedings of the International Workshop on Reproducibility and Replication in Recommender*

Systems Evaluation - RepSys '13. New York, New York, USA: ACM Press, 2013. p. 15–22. ISBN 9781450324656. Disponível em: <http://dl.acm.org/citation.cfm?id=2532508.2532512>.

[48] BEEL, J. et al. Introducing Docear's Research Paper Recommender System. In: *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries - JCDL '13*. New York, New York, USA: ACM Press, 2013. p. 459. ISBN 9781450320771. ISSN 15525996. Disponível em: <http://dx.doi.org/10.1145/2467696.2467786 http://dl.acm.org/citation.cfm?doid=2467696.2467786>.

[49] ZHOU, Q.; CHEN, X.; CHEN, C. Authoritative scholarly paper recommendation based on paper communities. In: *Proceedings - 17th IEEE International Conference on Computational Science and Engineering, CSE 2014, Jointly with 13th IEEE International Conference on Ubiquitous Computing and Communications, IUCC 2014, 13th International Symposium on Pervasive Systems,*. [S.l.]: IEEE, 2015. p. 1536–1540. ISBN 9781479979813.

[50] SUGIYAMA, K.; KAN, M.-Y. Exploiting Potential Citation Papers in Scholarly Paper Recommendation. *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries - JCDL '13*, p. 153, 2013. ISSN 14325012. Disponível em: <http://dl.acm.org/citation.cfm?doid=2467696.2467701>.

[51] SUGIYAMA, K.; KAN, M. Y. A comprehensive evaluation of scholarly paper recommendation using potential citation papers. *International Journal on Digital Libraries*, Springer Verlag, v. 16, n. 2, p. 91–109, jun 2015. ISSN 14321300. Disponível em: <http://link.springer.com/10.1007/s00799-014-0122-2>.

[52] HA, J.; KWON, S.-H.; KIM, S.-W. On recommending newly published academic papers. In: *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. New York, NY, USA: ACM, 2015. (HT '15), p. 329–330. ISBN 978-1-4503-3395-5. Disponível em: <http://doi.acm.org/10.1145/2700171.2791047>.

[53] ALSHAIKH, M. A.; UCHYIGIT, G.; EVANS, R. A research paper recommender system using a dynamic normalized tree of concepts model for user modelling. In: *2017 11th International Conference on Research Challenges in Information Science (RCIS)*. [S.l.: s.n.], 2017. p. 200–210.

[54] HIRSCH, J. E. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, National Acad Sciences, v. 102, n. 46, p. 16569–16572, 2005.

[55] EGGHE, L. An improvement of the h-index: The g-index. *ISSI newsletter*, v. 2, n. 1, p. 8–9, 2006.

[56] JIN, B. et al. The r-and ar-indices: Complementing the h-index. *Chinese science bulletin*, Springer, v. 52, n. 6, p. 855–863, 2007.

[57] ZHANG, C.-T. The e-index, complementing the h-index for excess citations. *PLoS One*, Public Library of Science, v. 4, n. 5, p. e5429, 2009.

[58] ALONSO, S. et al. hg-index: A new index to characterize the scientific output of researchers based on the h-and g-indices. *Scientometrics*, Akadémiai Kiadó, co-published with Springer Science+ Business Media BV, Formerly Kluwer Academic Publishers BV, v. 82, n. 2, p. 391–400, 2009.

[59] ZHANG, C.-T. The h'-index, effectively improving the h-index based on the citation distribution. *PloS one*, Public Library of Science, v. 8, n. 4, p. e59912, 2013.

[60] ZHAI, L.; YAN, X.; ZHU, B. The h l-index: improvement of h-index based on quality of citing papers. *Scientometrics*, Springer, v. 98, n. 2, p. 1021–1031, 2014.

[61] CHEN, P. et al. Finding scientific gems with google's pagerank algorithm. *Journal of Informetrics*, Elsevier, v. 1, n. 1, p. 8–15, 2007.

[62] KRAPIVIN, M.; MARCHESE, M.; CASATI, F. Exploring and understanding scientific metrics in citation networks. In: *Complex Sciences*. [S.l.]: Springer, 2009. p. 1550–1563.

[63] DING, Y. et al. Pagerank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, Wiley Online Library, v. 60, n. 11, p. 2229–2243, 2009.

[64] DU, M.; BAI, F.; LIU, Y. Paperrank: a ranking model for scientific publications. In: IEEE. *Computer Science and Information Engineering, 2009 WRI World Congress on*. [S.l.], 2009. v. 4, p. 277–281.

[65] PAL, A.; RUJ, S. Citex: A new citation index to measure the relative importance of authors and papers in scientific publications. In: IEEE. *Communications (ICC), 2015 IEEE International Conference on*. [S.l.], 2015. p. 1256–1261.

[66] CERVI, C. R.; GALANTE, R.; OLIVEIRA, J. P. M. d. Comparing the reputation of researchers using a profile model and scientific metrics. *in: XIII IEEE International Conference on Computer and Information Technology(CIT)*, 2013. Sydney, Australia.

[67] CERVI, C. R.; GALANTE, R.; OLIVEIRA, J. P. M. d. Application of scientific metrics to evaluate academic reputation in different research areas. *in: XXXIV International Conference on Computational Science(ICCS) 2013*, 2013. Bali, Indonesia.

[68] COSTAS, R.; BORDONS, M. Is g-index better than h-index? an exploratory study at the individual level. *Scientometrics*, Akadémiai Kiadó, co-published with Springer Science+ Business Media BV, Formerly Kluwer Academic Publishers BV, v. 77, n. 2, p. 267–288, 2008.

[69] BEEL, J.; GIPP, B. Google scholar's ranking algorithm: an introductory overview. In: RIO DE JANEIRO (BRAZIL). *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*. [S.l.], 2009. v. 1, p. 230–241.

[70] VIVIAN, G. R.; CERVI, C. R. Utilizando técnicas de data science para definir o perfil do pesquisador brasileiro da area de ciência da computação. *XII Escola Regional de Informática de Banco de Dados*, p. 108–117, 2016. ISSN 2177-4226.

[71] VIVIAN, G. R.; CERVI, C. R. xml2arff: Uma ferramenta automatizada de extração de dados em arquivos xml para data science com weka e r. *XII Escola Regional de Informática de Banco de Dados*, p. 159–162, 2016. ISSN 2177-4226.

[72] HALL, M. et al. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, ACM, v. 11, n. 1, p. 10–18, 2009.

[73] TEAM, R. C. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2015. Disponível em: <https://www.R-project.org/>.

[74] BASTIAN, M.; HEYMANN, S.; JACOMY, M. Gephi: An open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*, 2009. Disponível em: <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>.

[75] FRUCHTERMAN, T. M.; REINGOLD, E. M. Graph drawing by force-directed placement. *Software: Practice and experience*, Wiley Online Library, v. 21, n. 11, p. 1129–1164, 1991.

[76] CERVI, C. R. *Rep-Index - Uma Abordagem Abrangente e Adaptável Para Identificar Reputação Acadêmica*. Tese (Doutorado) — Universidade Federal do Rio Grande do Sul, Porto Alegre/RS – BR, dec 2013. Programa de Pós-Graduação em Computação.

[77] VIVIAN, G. R.; CERVI, C. R.; ROVADOSKY, D. N. Using selection attribute algorithms from data mining to complement the rep-index. In: . [S.l.]: IADIS, 2016. v. 15, p. 219–226. ISBN 978-989-8533-57-9.

[78] HALL, M.; WEKA. *GainRatioAttributeEval*. 2017. Disponível em: <http://weka.sourceforge.net/doc.dev/weka/attributeSelection/GainRatioAttributeEval.html>.

[79] HALL, M.; WEKA. *InfoGainAttributeEval*. 2017. Disponível em: <http://weka.sourceforge.net/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html>.

[80] HALL, M.; WEKA. *SymmetricalUncertAttributeEval*. 2017. Disponível em: <http://weka.sourceforge.net/doc.dev/weka/attributeSelection/SymmetricalUncertAttributeEval.html>.

[81] FRANK, E.; WEKA. *ChiSquaredAttributeEval*. 2017. Disponível em: <http://weka.sourceforge.net/doc.stable/weka/attributeSelection/ChiSquaredAttributeEval.html>.

[82] KIRA, K.; RENDELL, L. A. A practical approach to feature selection. In: SLEEMAN, D. H.; EDWARDS, P. (Ed.). *Ninth International Workshop on Machine Learning*. [S.l.]: Morgan Kaufmann, 1992. p. 249–256.

[83] KONONENKO, I. Estimating attributes: Analysis and extensions of relief. In: BERGADANO, F.; RAEDT, L. D. (Ed.). *European Conference on Machine Learning*. [S.l.]: Springer, 1994. p. 171–182.

[84] ROBNIK-SIKONJA, M.; KONONENKO, I. An adaptation of relief for attribute estimation in regression. In: FISHER, D. H. (Ed.). *Fourteenth International Conference on Machine Learning*. [S.l.]: Morgan Kaufmann, 1997. p. 296–304.

[85] SHANNON, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, v. 27, n. 3, p. 379–423, July 1948. ISSN 0005-8580.

[86] PEARSON, K. Mathematical contributions to the theory of evolution.–on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the royal society of london*, The Royal Society, v. 60, n. 359-367, p. 489–498, 1896.

[87] KENDALL, M. G. A new measure of rank correlation. *Biometrika*, JSTOR, v. 30, n. 1/2, p. 81–93, 1938.

[88] KOZAK, M.; BORNMANN, L. A new family of cumulative indexes for measuring scientific performance. *PloS one*, Public Library of Science, v. 7, n. 10, p. e47679, 2012.

[89] COHEN, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, Sage Publications, v. 20, n. 1, p. 37–46, 1960.

[90] GALTON, F. *Finger prints*. [S.l.]: Macmillan and Company, 1892.

[91] SMEETON, N. C. *Early history of the kappa statistic*. [S.l.]: JSTOR, 1985.

[92] MATTHEWS, B. W. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, Elsevier, v. 405, n. 2, p. 442–451, 1975.

[93] VIVIAN, G. R.; CERVI, C. R. Mahoutgui: Uma interface gráfica para gerar recomendações com o apache mahout diretamente de banco de dados usando mapeamento objeto-relacional. *XIII Escola Regional de Informática de Banco de Dados*, p. 79–82, 2017. ISSN 2177-4226.

[94] VIVIAN, G. R.; CERVI, C. R. Uma proposta de abordagem de recomendação para carreira de pesquisadores baseada em personalização, similaridade de perfil e reputação acadêmica. *XIII Escola Regional de Informática de Banco de Dados*, p. 7–16, 2017. ISSN 2177-4226.

APÊNDICE A – TABELA DE *STOPWORDS* UTILIZADAS

a	anything	by	december	either	get	interdisciplinar	may
0	anyway	c	decima	el	get	into	maybe
1	anyways	ç	decimo	ela	gets	inward	me
2	anywhere	cada	definido	elas	getting	is	mean
3	ao	calmamente	definir	ele	give	isto	meanwhile
4	aonde	came	definitely	eles	given	it	mediante
5	aos	came	dela	else	gives	its	melhor
6	apart	campi	dele	elsewhere	giving	itself	membro
7	aperfeiçoamento	campus	deles	em	go	j	menos
8	após	câmpus	demais	en	goes	ja	merely
9	appear	can	depois	en	going	já	mesma
a	appreciate	cannot	depressa	encontro	gone	jamais	mesmas
à	appropriate	cant	des	end	got	jan	mesmo
â	approximately	capítulos	described	enough	gotten	janeiro	mesmos
â	apr	cause	descrição	ensino	graduação	january	mestrado
á	april	causes	desde	entao	Grande	jul	mestre
able	aquela	cedo	despite	então	grande	julho	metodologia
about	aquelas	cem	desta	entirely	greetings	july	meu
about	aquele	centesimo	deste	entre	h	jun	meus
above	aqueles	certa	devagar	era	ha	june	mi
abr	aqui	certain	deveras	escolar	had	junho	might
abril	aquilo	certainly	deverás	especialista	han	just	might
absolutamente	are	certamente	dez	especialização	hand	just	mil
academica	área	certas	dezembro	especializado	happens	k	milesimo
acadêmicas	áreas	certo	dezenove	especially	hardly	keep	milesimos
acaso	aren	certos	dezesseis	essa	has	keeps	milhao
according	around	changes	dezessete	essas	have	kept	milionesimo
accordingly	as	chi	dezoito	esse	having	keywords	milionesimos
across	às	ciencias	di	esses	he	keywords	mim
actually	aside	ciências	did	esta	hello	kg	min
ad	ask	científica	different	estas	help	km	mine
add	asking	científico	discente	este	hem	know	ml
after	assaz	científicos	dispoe	estes	hence	known	mm
afterwards	assim	cinco	dispoem	estudante	her	knows	mon
again	associated	cinquenta	dissertação	estudos	here	l	monday
against	at	clearly	dissertações	et	hereafter	la	more
ago	ate	co	diversa	etc	hereby	lá	moreover
agora	até	colóquio	diversas	eu	herein	largely	most
agosto	atual	com	diversos	even	hereupon	las	mostly
ah	au	come	do	ever	hers	last	much
ai	aug	come	docentes	every	herself	lately	muita
ainda	august	comes	does	everybody	hi	later	muitas
albeit	autor	comigo	does	everyone	him	latter	muito
alem	aux	como	doing	everything	himself	latterly	muitos
algo	available	concerning	dois	everywhere	his	le	multidisciplinar
alguem	avaliador	Concluído	dom	ex	hither	least	must
algum	avante	concluído	domingo	exactly	hoje	les	my
alguma	away	concurso	don	example	hopefully	less	myself
algumas	awfully	congresso	don't	except	how	lest	n
alguns	b	conosco	done	experimento	howbeit	let	na
ali	bacharelado	consequently	dos	experimentos	however	levemente	nacional
all	basta	consider	doutor	f	html	lhe	nada
allow	bastante	considering	doutorado	faculdade	hum	lhes	name
allows	be	consigo	down	far	i	lie	namely
almost	became	contain	downwards	faz	í	like	nao
alo	because	containing	doze	feb	ì	liked	não
alone	become	contains	dr	february	î	likely	nas
along	becomes	contigo	dra	fev	ï	little	Natureza
already	becoming	contra	du	fevereiro	ĩ	ll	nd
also	been	contudo	ducentesimo	few	ie	lo	near
although	before	convosco	due	fielmente	if	logo	nearly
alto	beforehand	coordenação	dum	fifth	ignored	look	necessary
aluno	behind	Coordenador	duma	financiador	ih	looking	need
always	being	coordenador	dumas	first	image	looks	needs
am	believe	coragem	duns	five	immediate	los	neither
amanha	below	corresponding	durante	flowers	in	low	nem
amanhã	bem	corretamente	during	foi	inasmuch	ltd	nenhum
ambas	beside	could	duzentos	followed	inc	m	nenhuma
ambos	besides	course	e	following	indeed	ma	nenhumas
among	best	cuja	é	follows	indicate	má	nenhuns
amongst	better	cujas	ê	for	indicated	made	nesse
an	between	cujo	ẽ	formação	indicates	mai	neste
and	beyond	cujos	è	former	inner	mainly	never
andamento	big	currently	ë	formerly	insofar	maio	nevertheless
ano	bilhao	currículo	each	forth	instead	mais	new
another	bilionesimo	curso	each	found	institucional	make	next
ante	bilionesimos	cursos	edu	four	instituição	mal	nine
anteontem	bis	d	educação	fri	instituições	many	ninguem
antes	both	da	educadores	friday	instituto	mar	no
any	bravo	das	efetivamente	from	integrado	marco	nobody
anybody	breve	de	eg	further	integrante	março	nogentesimo
anyhow	brief	debalde	eia	furthermore	Integrantes	mars	non
anyone	but	dec	eight	g	integrantes	mas	nona

nonagesimo	over	quatorze	sep	tal	trezentos	w
none	overall	quatro	september	talvez	tried	want
nono	own	quatrocentos	septingentesimo	tambem	tries	want
noone	oxala	que	septuagesimo	tanta	trigesimo	wants
nor	p	quem	ser	tantas	trinta	was
normally	par	quer	sera	tanto	truly	way
nos	para	qui	serious	tantos	try	we
nós	participantes	quica	seriously	tao	trying	wed
nosso	particular	quingentesimo	sessenta	tarde	tu	wednesday
nossos	particularly	quinhentos	set	te	tua	welcome
not	pedagógicas	quinquagesimo	sete	tecnologicas	tuas	well
nothing	pela	quinta	setecentos	tell	tudo	went
nov	pelas	quinta-feira	setembro	tem	tue	were
nove	pelo	quinto	setenta	tends	tuesday	what
novecentos	pelos	quinze	setima	ter	tutor	what's
novel	per	quite	setimo	terca	tutores	whatever
november	perante	qv	seu	terça	twice	whatsoever
novembro	perhaps	r	seus	terca-feira	two	when
noventa	Pesquisa	rather	seven	terça-feira	u	whence
now	pesquisa	rd	several	terceira	û	whenever
nowhere	pesquisador	re	sex	terceiro	ũ	whensoever
num	pesquisadores	realizado	sexagesimo	teu	ú	where
numa	pior	really	sexcentesimo	teus	ù	whereafter
numas	placed	realmente	sexta	th	ũ	whereas
nunca	plataforma	reasonably	sexta-feira	than	ue	whereat
nuns	plataformas	regarding	sexto	thank	uh	whereby
o	please	regardless	shall	thanks	ui	wherefrom
ô	plus	regards	she	thanx	ultimamente	wherein
õ	pois	regional	should	that	um	whereinto
ó	por	reitoria	show	that's	uma	whereof
ò	por	relatively	showed	thats	umas	whereon
ö	porem	respectively	shown	the	un	whereto
oba	porém	resultados	shows	their	under	whereunto
obtain	porque	resulted	si	theirs	under	whereupon
obtained	portanto	resulting	significant	them	une	wherever
obviously	porventura	revisor	significantly	them	unfortunately	wherewith
octingentesimo	possible	right	silencio	themselves	universidade	whether
october	possivelmente	s	sim	then	universidades	which
octogesimo	pouca	sab	simposio	thence	unless	whichever
of	poucas	sáb	simpósio	ther	unlikely	whichsoever
off	pouco	sabado	since	there	uns	while
often	poucos	sábado	sir	thereafter	until	whilst
oh	pre	said	Situação	thereby	unto	whither
oitava	presumably	same	situação	therefore	up	who
oitavo	previously	sao	six	therein	upon	whoever
oitenta	primeira	são	so	theres	url	whole
oito	primeiramente	sat	só	thereupon	us	whom
oitocentos	primeiro	saturday	sob	these	use	whomever
ok	probably	saw	sobre	these	use	whomsoever
okay	processo	say	some	thet	used	whose
ola	professor	saying	somebody	they	useful	whosoever
old	professora	says	somehow	think	uses	whreas
on	professores	se	someone	third	using	why
once	programa	second	something	this	uso	will
onde	projetos	secondly	sometime	thorough	uso	willing
one	propios	see	sometimes	thoroughly	usos	wish
ones	proprio	seeing	somewhat	those	usually	with
only	provavelmente	seem	somewhere	though	uucp	within
ontem	provides	seemed	soon	three	v	without
onto	psit	seeming	sorry	through	value	wonder
onze	psiu	seems	specified	through	vamos	workshop
opa	publicações	seen	specify	throughout	varia	would
or	put	seg	specifying	thru	varias	x
orientação	puxa	segunda	sr	thu	vario	y
orientações	q	segunda-feira	still	thurday	varios	yes
os	qua	segundo	su	thus	various	yet
other	quadragesimo	sei	sua	ti	ve	you
others	quadrigentesimo	seis	suas	to	very	yours
otherwise	quais	seiscentesimo	sub	toda	very	yourself
ou	quaisquer	seiscentos	sub	todas	via	yourselves
ought	qual	seja	subárea	todo	via	z
our	qualquer	self	such	todos	vigesimo	zero
ours	quando	selves	such	together	vinte	
ourselves	quanta	sem	suggest	too	viva	
out	quantas	seminário	sun	took	viz	
outra	quanto	sempre	sunday	toward	voce	
outras	quantos	sendo	sup	towards	você	
outrem	quao	senhor	sure	tras	vos	
outro	quarenta	senhora	t	trecentesimo	vós	
outrora	quarta	senhoria	ta	treinamento	vossa	
outros	quarta-feira	senhorita	tais	tres	vosso	
outside	quarto	sensible	take	três	vossos	
outubro	quase	sent	taken	treze	vs	

APÊNDICE B – TABELA DE SINONÍMIAS UTILIZADAS

Palavra	Sinonímia	Palavra	Sinonímia	Palavra	Sinonímia
access	acesso	informations	informações	scientific	científica
aerospace	aeroespaciais	inspired	inspirada	security	segurança
analytics	analíticos	instrumentation	instrumentação	semiotic	semiótica
analysis	análise	intelligence	inteligente	sensors	sensores
applied	aplicada	interaction	interação	signals	sinais
artificial	artificiais	interfaces	interfaces	simulation	simulação
automation	automação	interval	intervalar	smart	inteligente
automotive	automotivas	knowledge	conhecimento	statistic	estatística
autonomous	autônomo	large	larga	structures	estruturas
base	banco	learning	aprendizado	surveillance	vigilância
based	baseados	linguistics	linguística	system	sistema
biology	biologia	logic	lógica	systemic	sistêmico
biomedical	biomédica	logistics	logística	systems	sistemas
centered	centrado	machine	máquina	team	time
cloud	nuvem	magnetic	magnéticas	tecnics	técnicas
cognitive	cognitiva	maintenance	manutenção	temporal	temporais
collections	acervos	manufacturing	fabricação	theory	teoria
combinatorial	combinatória	math	matemática	things	coisas
commerce	comercio	measures	medidas	tolerance	tolerância
complex	complexas	mechatronics	mecatrônica	ubiqua	ubíqua
computational	computacional	medicine	médica	usability	usabilidade
computer	computador	meteorology	meteorologia	user	usuário
computers	computadores	methodology	metodologia	vehicles	veiculares
computing	computação	methods	métodos	verified	verificada
community	comunidades	mining	mineração	vision	visão
concurrency	concorrência	mobile	móvel	visualization	visualização
control	controle	modeling	modelagem	wireless	sem fio
cultural	culturais	models	modelos		
data	dados	multimedia	multimídia		
deep	profundo	music	musical		
defined	definido	natural	naturais		
design	projeto	network	rede		
detection	detecção	networks	redes		
development	desenvolvimento	neural	neurais		
discreet	discreta	operation	operacional		
distributed	distribuído	optical	ópticas		
education	educação	optimization	otimização		
electronic	eletrônicas	parallel	paralela		
electric	elétricas	patterns	padrões		
electronic	eletrônico	performance	desempenho		
encryption	criptografia	plain	planejamento		
end	final	prediction	previsão		
engineering	engenharia	preservation	preservação		
entomology	entomologia	probability	probabilidade		
evolution	evolução	process	processo		
extract	extração	processing	processamento		
fail	falhas	product	produto		
forensic	forense	programming	programação		
fundamentals	fundamentos	quantity	quântica		
geographic	geográfica	reality	realidade		
geoprocessing	geoprocessamento	recognition	reconhecimento		
graphics	gráfico	recommendation	recomendação		
high	alto	research	pesquisa		
hypermedia	hipermídia	resilience	resiliência		
human	humano	resources	recursos		
hybrids	híbridos	retrieval	recuperação		
images	imagens	risk	risco		
improvement	melhoria	robotics	robótica		
informatics	informática	scale	escala		
information	informação	science	ciência		

APÊNDICE C – CÓDIGO FONTE LOGLIKELIHOODDISTANCEMEASURE.JAVA

```
 1.  package org.apache.mahout.common.distance;
 2.
 3.  import java.util.HashSet;
 4.  import java.util.Iterator;
 5.  import java.util.Set;
 6.  import org.apache.mahout.math.RandomAccessSparseVector;
 7.  import org.apache.mahout.math.Vector;
 8.  import org.apache.mahout.math.stats.LogLikelihood;
 9.
10.  public class LoglikelihoodDistanceMeasure extends WeightedDistanceMeasure {
11.
12.    @Override
13.    public double distance(Vector vector0, Vector vector1) {
14.      Set<Integer> intersection = new HashSet<>(vector0.size());
15.
16.      int numUsers = vector0.size(); //Size of tf-idf dictionary
17.      int preferring1 = vector0.getNumNonZeroElements();
18.      int preferring2 = vector1.getNumNonZeroElements();
19.
20.      Iterator<Vector.Element> iter = vector0.nonZeroes().iterator();
21.      while (iter.hasNext()) {
22.        Vector.Element feature = iter.next();
23.        int i = feature.index();
24.        if (!intersection.contains(i)) { intersection.add(i); }
25.      }
26.
27.      int intersec = 0;//Number of vector0 and vector1 intersection
28.
29.      iter = vector1.nonZeroes().iterator();
30.      while (iter.hasNext()) {
31.        Vector.Element feature = iter.next();
32.        int i = feature.index();
33.        if (intersection.contains(i)) { intersec++; }
34.      }
35.
36.      if (intersec == 0) { return Double.NaN; }
37.
38.      LogLikelihood.logLikelihoodRatio(intersec,
39.              preferring1 - intersec,
40.              preferring2 - intersec,
41.              numUsers - preferring1 - preferring2 + intersec);
42.
43.      return 1.0 / (1.0 + l);
44.    }
45.  }
```

APÊNDICE D – GRÁFICOS DE DISPERSÃO DO CÁLCULO DOS PESOS DO REP-INDEX

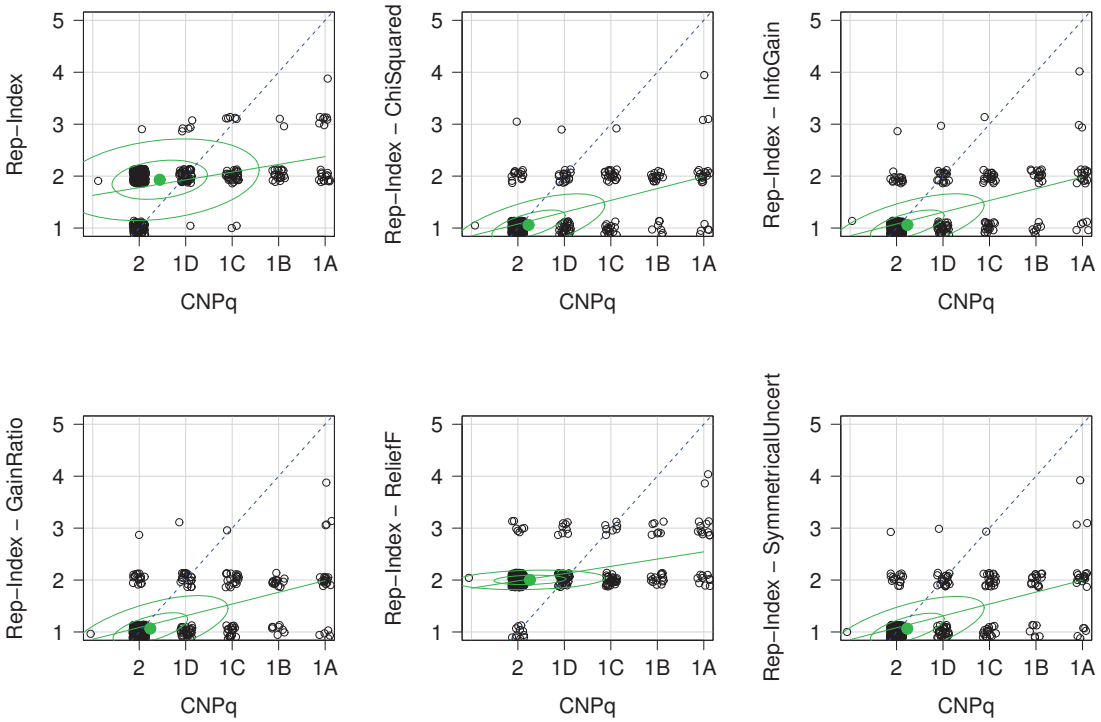

Figura 42. Gráficos de Dispersão para Ciência da Computação.

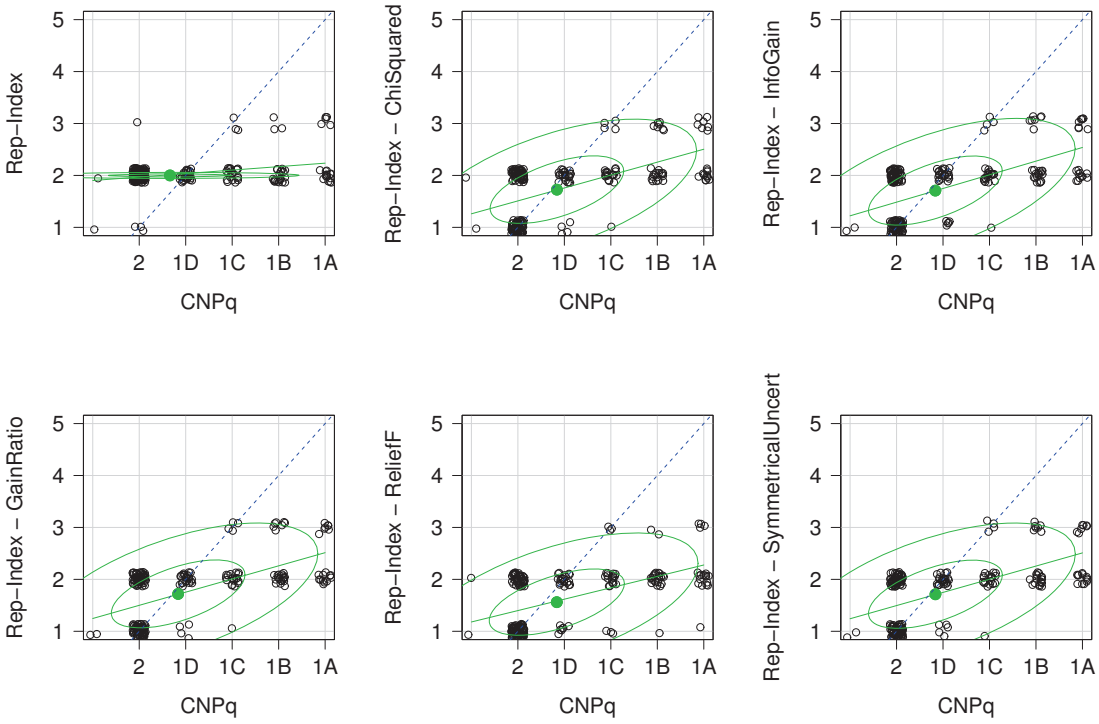

Figura 43. Gráficos de Dispersão para Odontologia.

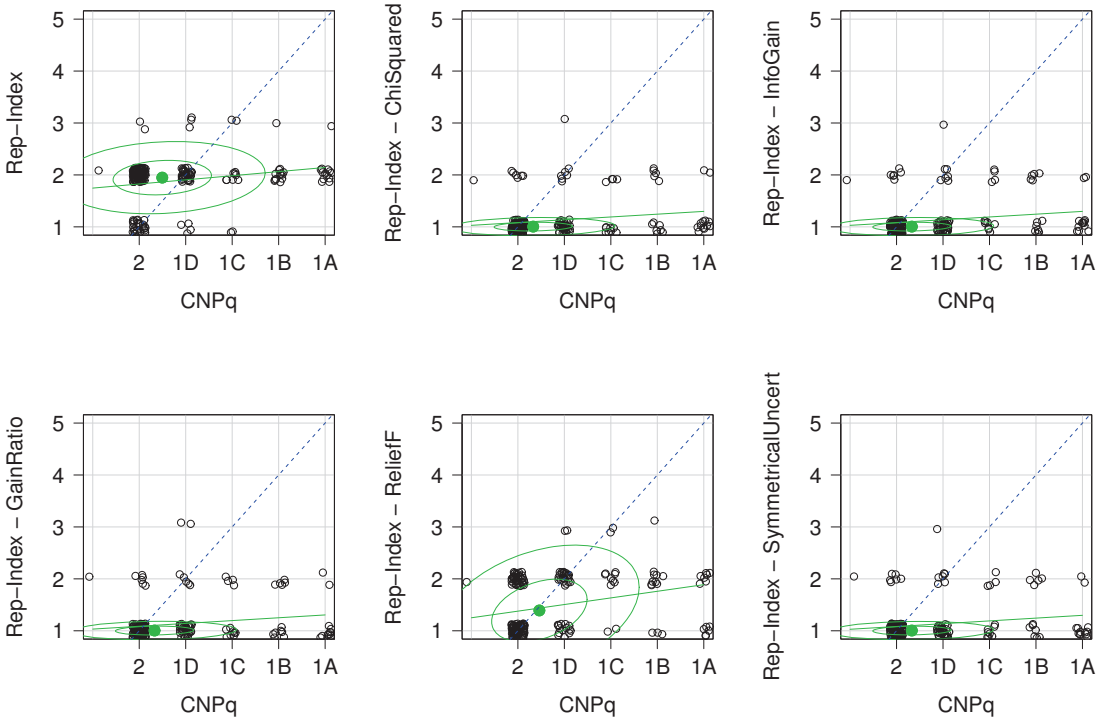

Figura 44. Gráficos de Dispersão para Economia.