Uma Proposta de Abordagem de Recomendação para Carreira de Pesquisadores Baseada em Personalização, Similaridade de Perfil e Reputação Acadêmica

Gláucio R. Vivian[1], **Cristiano R. Cervi**[1]

[1]Instituto de Ciências Exatas e Geociências (ICEG)
Universidade de Passo Fundo (UPF) – Passo Fundo – RS – Brazil

{149293,cervi}@upf.br

Abstract. *The Recommendation Systems seek to suggest relevant information to users. In the context of researchers there are numerous approaches proposed to recommend articles and citations. The objective of this work is to present a strategy of recommendation approach focused on the career projection of researchers. As a result, we obtained the best combination for subarea profile similarity with Log-likelihood and Apache Mahout's ClassicAnalyzer class. Regarding the recommendations, two different types of recommendation were generated for several test groups. The results were satisfactory and show that the proposed approach has good coverage in the generation of recommendations.*

Resumo. *Os Sistemas de Recomendação procuram sugerir informações relevantes aos usuários. No contexto dos pesquisadores existem inúmeras abordagens propostas para recomendar artigos e citações. O objetivo deste trabalho é apresentar uma estratégia de abordagem para recomendação com foco na projeção da carreira de pesquisadores. Como resultados, obtivemos a melhor combinação para similaridade de perfil de subárea com Log-likelihood e a classe ClassicAnalyzer do Apache Mahout. Com relação as recomendações, foram geradas dois tipos diferentes de recomendação para diversos grupos de testes. Os resultados foram satisfatórios e demostram que a abordagem proposta tem boa cobertura na geração de recomendações.*

1. Introdução

Os Sistemas de Recomendação (SR) tradicionais buscam auxiliar os usuários na seleção de conteúdos. 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 novo paradigma imposto aos pesquisadores, torna mais complexa e árdua a tarefa de traçar planos estratégicos para projeção da carreira do pesquisador. Neste contexto, os Sistemas de Recomendação podem interagir com os pesquisadores, buscando orientá-los com estratégias de recomendações no planejamento da sua carreira. Em outras palavras, um Sistema de Recomendações pode sugerir ao pesquisador o que, como e quando realizar determinada produção. Como resultado, tem-se a possibilidade de estar realizando a atividade mais adequada e na ordem cronológica mais apropriada.

O objetivo deste trabalho é apresentar uma abordagem de recomendação baseada na personalização dos dados de pesquisadores, usando a similaridade de perfil e reputação acadêmica como premissa de recomendação. A abordagem proposta visa contribuir para o planejamento da carreira do pesquisador, bem como ser um apoio a grupos de pesquisa, programas de pós-graduação e instituições, para que acompanhem a evolução da vida científica de um pesquisador. A proposta 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. Desse modo, 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 onde o pesquisador está inserido.

Este artigo está organizado da seguinte forma: Na seção 2 são analisados alguns trabalhos correlatos. Na seção 3 é exposta a abordagem proposta. Na seção 4 é apresentada a metodologia. Na seção 5 são apresentados os experimentos e resultados encontrados. Finalmente, na seção 6, são apresentadas as conclusões e trabalhos futuros.

2. Trabalhos Correlatos

Esta seção apresenta trabalhos correlatos existentes no contexto de Sistemas de Recomendação para pesquisadores.

O artigo de [Middleton et al. 2004] introduziu a modelagem de perfil para recomendação de artigos científicos com o uso de ontologias. A representação dos artigos foi realizada utilizando-se vetores de termos, computados com a técnica *Term Frequency* (TF) e divididos pelo total de termos. Os experimentos demostraram que a abordagem proposta supera os sistemas apresentados na literatura.

No trabalho de [Ekstrand et al. 2010] foram explorados diversos métodos (177 algoritmos em 5 famílias) para recomendação de artigos científicos baseada em filtragem colaborativa e em conteúdo. O perfil do usuário foi construído com as próprias citações da Web. As medidas de influência foram realizadas com os algoritmos HITS[Kleinberg 1999] e *PageRank*[Page et al. 1999]. Inicialmente, realizou-se testes *offline*, posteriormente se conduziu uma avaliação *online* com pesquisadores. Ao final, demostrou-se que os usuários preferem as recomendações com filtragem colaborativa.

No trabalho de [Zhang e Li 2010], foi proposta a recomendação de artigos com o modelo de perfil baseado em árvores para ultrapassar os inconvenientes do espaço vetorial. A abordagem proposta cria o perfil do pesquisador com base nos trabalhos visualizados. A correlação entre os perfis é computada utilizando a técnica *Edit Distance* adaptada para árvores. Um modelo de ativação disperso é construído para localizar perfis com interesses semelhantes. Para avaliar foi utilizada a métrica *Normalized Discounted Cumulative Gain* com um subconjunto com 60 mil exemplares da *National Science and Technology Library*. Foram realizadas avaliações de 5 até 30 recomendações. Ao final, a melhor opção foi a recomendação de 10 artigos.

A proposta de [Huang et al. 2012] consiste em definir 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 recomenda-se as 20 primeiras. Ao final dos experimentos, os autores afirmam que a proposta ultrapassa o estado atual da arte.

No trabalho de [Beel et al. 2013] foi feita uma revisão sistemática de 80 abordagens existentes para recomendar artigos científicos. Constatou-se que existem mais de 170 artigos publicados. 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 existe um consenso de qual trabalho representa o estado da arte.

Nos trabalhos de [Sugiyama e Kan 2013, Sugiyama e Kan 2015] buscou-se construir um sistema de recomendações por meio do potencial de citação de artigos. Foi construído um perfil vetorial com base nos artigos publicados na DBLP e na ACM *Digital Library*. Utilizou-se a correlação de Pearson entre o perfil e vetor para recomendar os artigos com maior similaridade para os usuários alvo. Após diversos experimentos com o objetivo de ajustar a acurácia em 10% com relação ao *baseline*, os autores afirmam que a abordagem proposta é eficiente em caracterizar artigos para recomendação.

Como se observou nesta seção, existem diversos trabalhos com o objetivo de auxiliar os pesquisadores com recomendações de artigos, referências e citações. Contudo não localizamos estudos com a finalidade de recomendações para o planejamento de carreira de pesquisadores. A abordagem proposta busca preencher esta lacuna identificada.

3. Abordagem Proposta

Nesta seção definimos uma abordagem com o intuito de gerar recomendações para o planejamento de carreira de pesquisadores. A abordagem proposta para gerar as recomendações deve responder aos seguintes questionamentos: **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 prestígio na mesma subárea de atuação. **ii) Como Fazer?** Descrição de como realizar a atividade recomendada apresentando opções para o pesquisador. **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 abordagem proposta vai ser divida em duas partes: **i) Recomendações Não Personalizadas:** são as que não possuem informações específicas para o usuário, elas se caracterizam por apenas considerarem o elemento do Rep-Index e sua importância para o aumento da reputação. **ii) Recomendações Personalizadas:** são específicas para cada usuário, a similaridade de perfil e reputação dos demais pesquisadores são utilizados para gerar a recomendação.

A etapa inicial consiste na adaptação do modelo do perfil do pesquisador. Optou-se por realizar esta tarefa no Rep-Model proposto por [Cervi et al. 2013b,

Cervi et al. 2013a]. Ele é um conjunto de elementos que representam o comportamento acadêmico e científico dos pesquisadores. O Rep-Model também é utilizado no Rep-Index, trata-se de 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. O grande diferencial de outras métricas é a média ponderada, escopo abrangente e adaptabilidade. As alterações propostas incluem novos elementos no Rep-Model com o intuito de utilizá-lo para esta finalidade. Na Tabela 1 pode-se visualizar a proposta de adição de elementos ao Rep-Model.

Tabela 1. 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

A adição de novos elementos quantitativos ao Rep-Model tem 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 e são importantes para localizar as afinidades entre os pesquisadores. Optou-se por utilizar a(s) subárea(s) de atuação da plataforma Lattes devido ao fato das mesmas serem mais específicas e representarem a(s) sua(s) área(s) de atuação.

O próximo passo da abordagem é a definição da utilização dos elementos definidos para o modelo de perfil (Rep-Model adaptado). 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 vetor $Rep - \vec{I}ndex$. Os elementos 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, optou-se por empregar técnicas de recomendação baseada em conteúdo. Na Figura 1 pode-se visualizar as etapas do processo de *Text mining*.

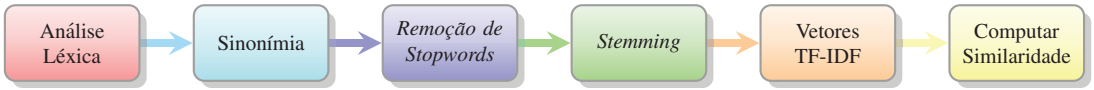

Figura 1. Etapas do processo de *Text Mining*.

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 listas de classes de palavras sem

relevância ou que podem gerar falsas similaridades. Posteriormente, utiliza-se a técnica denominada de *stemming*, a mesma procura reduzir as palavras ao seu radical por meio da supressão de sufixos. Finalmente, aplica-se a técnica de vetorização denominada TF-IDF (*Term Frequency-Inverse Document Frequency*). A mesma simplifica o processamento das informações textuais por meio da representação em vetores esparsos com a frequência de ocorrência e relevância dos seus termos.

O resultado da técnica de vetorização é aplicado às funções de correlação de Pearson, Spearman e Kendall Tau; similaridade Fuzzy; distância Euclidiana, Canberra, Tanimoto, Log-likelihood, Manhattan, Minkowski, Chebyshev, Coseno e EarthMovers. Cada uma dessas funções terá como resultado final uma matriz triangular $M_{sim_{(U_{m,n})}}$ onde são armazenadas as similaridades entre os perfis dos pesquisadores. No caso das distâncias os valores foram todos convertidos em similaridades (normalizados) por meio da Equação: $s = \frac{1}{1+d}$. Onde: d representa a distância e s a similaridade obtida no intervalo de valores entre 0 e 1, inclusive.

As **recomendações personalizadas** sobre "o que fazer" para um usuário U são realizadas pela análise do vetor $Rep - \vec{Index}_{U(i)}$, 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. A partir do conjunto de n possíveis recomendações ao usuário U, deve-se computar o quanto cada uma incrementa na sua reputação.

As **recomendações não personalizadas** são geradas a partir da simulação do aumento do Rep-Index do pesquisador em questão. O aumento da reputação (denotado por 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. O cálculo do aumento da reputação pode ser realizado pela equação: $\Delta_{(R)} = Rep - Index\ Novo_{(R)} - Rep - Index\ Atual_{(R)}$. Esta equação 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 atual. Pode-se simplificar a mesma e corrigir a situação acima mencionada. A Equação 1 apresenta uma proposta melhorada para o cálculo em questão.

$$\Delta_{(R)} = \begin{cases} 0, & \text{se } inc \geq max_{(R_i)} \\ \frac{inc * w_{(R_i)}}{max_{(i)}}, & \text{senão} \end{cases} \quad (1)$$

Onde, $\Delta_{(R)}$ é o aumento na reputação do pesquisador R, inc 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 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.

4. Metodologia

Para realização dos experimentos, foram utilizados dados de pesquisadores da área da Ciência da Computação. Os dados foram coletados da plataforma Lattes[1] e Google Scholar[2] entre novembro de 2016 e janeiro de 2017. No trabalho de [Vivian e Cervi 2016a] pode-se encontrar detalhadamente os passos para recuperação das informações e criação do XML *Dataset* utilizado para realizar os experimentos. As consultas dos dados foram realizadas com auxílio da linguagem XQuery por meio do *software* Basex[3]. O intercâmbio de informações entre formatos foi realizado utilizando-se o *software* Xml2Arff[4] proposto por [Vivian e Cervi 2016b].

A avaliação da abordagem proposta foi realizada com os seguintes experimentos: **i)** Encontrar o melhor conjunto de sinonímia, *Stopwords* e *Stemming*. **ii)** Encontrar a melhor correlação / similaridade / distância. **iii**) Avaliar as recomendações geradas. Em cada item previsto para os experimentos, foi utilizado o conjunto apropriado de dados bem como as métricas / métodos mais utilizadas para realizar o experimento.

5. Experimentos e Resultados

Esta seção apresenta os experimentos e resultados das três etapas da abordagem proposta anteriormente.

5.1. Pesos Específicos para o Rep-Index

A determinação dos pesos específicos do Rep-Index para cada área de estudo foi realizada anteriormente utilizando-se o complemento para o Rep-Index proposto por [Vivian et al. 2016]. O mesmo utiliza técnicas de mineração de dados e aprendizado de máquina para computar cinco opções de pesos. A avaliação da melhor opção é realizada pela correlação de Spearman. Dessa forma, o Rep-Index é adequado para classificar os pesquisadores o mais próximo possível da realidade da área (pesquisadores do CNPq).

5.2. Similaridade entre as Subáreas

Para localizar a melhor combinação de técnicas, deve-se inicialmente agrupar os pesquisadores em categorias (subáreas do currículo Lattes). Utilizou-se o elemento Área de Atuação (AA) do Rep-Model modificado para esta finalidade. Em um grupo formado por 398 pesquisadores da área de Ciência da Computação, inicialmente foram localizadas 959 subáreas de atuação. Essa quantidade se justifica no fato de que a maioria dos pesquisadores apresenta 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 apenas um único pesquisador. As classes com apenas um pesquisador foram retiradas dos experimentos.

A partir das matrizes de similaridades (uma para cada função) entre os pesquisadores e as categorias, aplica-se o algoritmo do vizinho mais próximo (*Nearest Neighbor*)

[1] http://lattes.cnpq.br
[2] https://scholar.google.com.br
[3] http://basex.org
[4] https://github.com/grvivian/xml2arff

com o parâmetro n (indica quantos vizinhos devem ser selecionados) igual ao número de pesquisadores existentes na categoria em questão. Dessa forma, seleciona-se os n pesquisadores mais afins à categoria. Isto é obtido pela ordenação decrescente das similaridades do pesquisador, seguido pela seleção dos n primeiros pesquisadores. Após localizar os pesquisadores mais afins para cada categoria, basta comparar com a definição original das categorias e encontrar os verdadeiros positivos (TP), falsos positivos (FP), verdadeiro negativo (TN) e falso negativo (FN). Com essas informações, constrói-se a matriz de confusão e aplica-se as métricas de avaliação.

Os experimentos foram realizados no Apache Mahout[5] versão 1.12.2. 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* e sinonímia. Também criou-se a classe `LoglikelihoodDistanceMeasure.java` para computar esta medida de distância com vetores esparsos, uma vez que o Mahout possui o `Log-Likelihood` apenas para filtragem colaborativa. As palavras com frequência relativa (TF) abaixo de 2 foram desconsideradas. Na Tabela 2 pode-se visualizar as classes utilizadas.

Tabela 2. Classes empregadas nos experimentos.

Classe	Análise léxica	*Stopwords*	Sinonímia	*Stemming*	Dicionário
ClassicAnalyzer (CL)	ClassicTokenizer, ClassicFilter, LowerCaseFilter	33 (Inglês)	Não	Não	44.723
StandardAnalyzer (ST)	StandardTokenizer, StandardFilter, LowerCaseFilter	33 (Inglês)	Não	Não	44.360
EnglishAnalyzer (EN)	StandardTokenizer, StandardFilter, LowerCaseFilter	33 (Inglês)	Não	PorterStemFilter	35.379
BrazilianAnalyzer (BR)	StandardTokenizer, StandardFilter, LowerCaseFilter	128 (Português)	Não	BrazilianStemFilter	33.290
MyBrazilianAnalyzer (MY)	StandardTokenizer, StandardFilter, LowerCaseFilter	1234 (Português)	147*	BrazilianStemFilter	32.842

*Usou-se a Sinonímia para Traduzir termos da área em língua Inglesa para língua Portuguesa.

Observa-se que a classe `ClassicAnalyzer` foi a que gerou o maior dicionário de palavras. Ficando em segundo lugar a classe `StandardAnalyzer`, em seguida `EnglishAnalyzer`, `BrazilianAnalyzer` e `MyBrazilianAnalyzer`.

A partir das similaridades apresentadas e das classes da Tabela 2 realizou-se as avaliações com as métricas: *F-Measure* (média harmônica entre *precision* e *recall*), RMSE e *Kappa*. Utilizou-se a configuração *full-training set* para avaliar o modelo. Os valores foram obtidos através da média ponderada entre o resultado de cada classe e o seu número de pesquisadores. Na Figura 2 pode-se visualizar as métricas acima mencionadas.

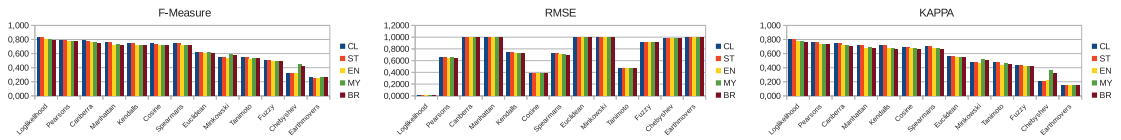

Figura 2. Métricas *F-Measure*, RMSE e *Kappa*.

Observa-se que a classe `ClassicAnalyzer` (maior dicionário) em conjunto com a técnica `Log-Likelihood` foi a que obteve a maior *F-Measure* (0,831). Além disso, esta combinação apresentou os menores erros para RMSE (0,0012). As correlações

[5]http://mahout.apache.org

de Pearson e Kendall, similaridade Fuzzy, distância do Cosseno e Tanimoto apresentam RMSE entre 0,4 e 0,85. No entanto, as demais funções, por apresentarem domínio positivo infinito acabaram ficando com os valores muito próximos de zero quando convertidas em similaridades pela equação mencionada na abordagem. Dessa forma, o RMSE delas está muito próximo de 1,0. O resultado da métrica estatística Cohen's *Kappa* (0,806) também obteve a melhor colocação. A medida estatística *Kappa* está no intervalo de 0,80 até 0,90 (resultado forte). Com relação as demais classes, verifica-se que as mesmas apresentam resultados ligeiramente inferiores a `ClassicAnalyzer`. Um fato também observado é a enorme diferença nos erros (RMSE) entre a técnica `Log-Likelihood` e as demais. Isto indica um alto grau de precisão nas similaridades, ou seja, valores muito próximos de 1,0 devido ao fato da mesma ser uma função monótona crescente.

5.3. Recomendações

Para realizar os experimentos com as recomendações, utilizou-se o grupo total de 398 pesquisadores 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 composto por 80 docentes do grupo INF da UFRGS[6] e 63 docentes do DCC da UFMG[7]. Destes 64 são bolsistas de produtividade do CNPq e 79 não possuem bolsas do CNPq. Solicitou-se a geração de recomendações para rede de colaboradores e para Grau de Instrução. Na Figura 3 pode-se visualizar a métrica *coverage* adaptada à abordagem com o parâmetro n variando de 1 até 50 recomendações. Não se utilizou a *recall* e *precision* pois só podem ser empregadas em situações onde pode-se predizer se o usuário gostou ou não da recomendação.

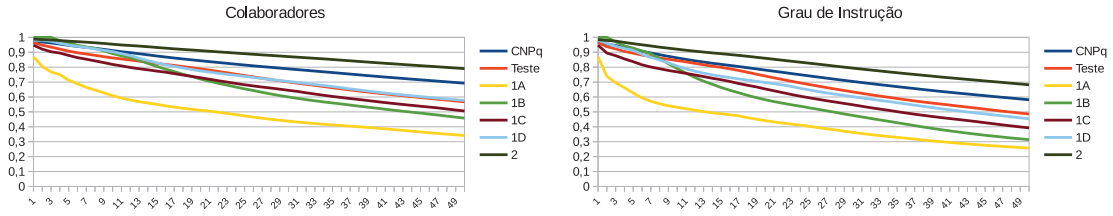

Figura 3. *Coverage* de Recomendações para Colaboradores e Grau de Instrução.

Nas Figuras anteriores pode-se observar que nunca ocorreu a situação onde não existe o que recomendar (valores zerados). Também fica evidente que os grupos iniciais do CNPq, ou seja, os grupos 2 e 1D possuem valores maiores de *coverage* do que os grupos mais avançados (1C, 1B e 1A). Isto se justifica pelo fato que os grupos finais têm maior reputação no Rep-Index e Grau de Instrução do que os grupos iniciais e de teste. As recomendações geradas estão disponíveis em um repositório[8] como material suplementar. Os pesquisadores estão identificados apenas pela id da plataforma Lattes.

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 foi utilizado 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

[6]http://www.inf.ufrgs.br/site/pessoas/corpo-docente/
[7]http://www.dcc.ufmg.br/dcc/?q=pt-br/professores
[8]https://github.com/grvivian/ERBD2017/

valores para os pesos do Rep-Index foram ignorados e portanto não serão recomendados. Foi computada a média da métrica *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. A Figura 4 apresenta as mesmas.

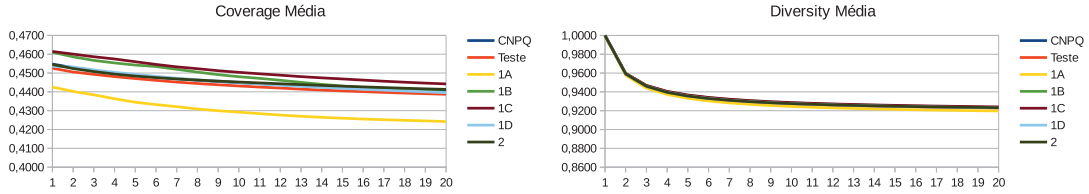

Figura 4. *Coverage* Média e *Diversity* Média para Ciência da Computação.

Observa-se um comportamento semelhante para a *coverage* da Figura 3 porem com valores menores. Com relação a *diversity*, pode-se constatar que quando $n = 1$ ocorre o máximo valor, a partir de $n \geq 2$ a mesma decresce para 0,92 quando $n = 20$. Em todos as situações a diversidade de elementos recomendados foi satisfatória. Na Tabela 3 pode-se visualizar as recomendações geradas para um pesquisador do nível SR.

Tabela 3. Recomendações geradas.

Recomendação	Tipo	Peso	Máx.	Inc.	Aumento
Aumente o item: Orientação de Doutorado (PTA) para 6	REC_PTA	16,789	46	1	0,365
Aumente o item: Orientação de Pós-doutorado (PA) para 1	REC_PA	5,273	19	1	0,278
Aumente o item: Membro de Corpo Editorial de Periódico (EBM) para 6	REC_EBM	4,569	18	1	0,254
Aumente o item: Livro (BP) para 7	REC_BP	5,187	57	1	0,091
Aumente o item: Orientação de Mestrado (MDA) para 18	REC_MDA	9,328	114	1	0,082
Aumente o item: H-Index (HI) para 18	REC_HI	8,609	116	1	0,074
Aumente o item: Artigo em Periódico (ASJ) para 24	REC_ASJ	17,420	246	1	0,071
Aumente o item: Revisão de Periódico (RJ) para 1	REC_RJ	5,146	77	1	0,067
Aumente o item: Participação em Banca de Mestrado (PEBPT) para 56	REC_PEBPT	8,434	127	1	0,066
Aumente o item: Prêmios (PREM) para 6	REC_PREM	3,812	59	1	0,065
Aumente o item: Capítulo de Livro (BCP) para 12	REC_BCP	3,644	62	1	0,059
Amplie a sua Rede de colaboração com: 5554254760869075, similaridade: 0,999155 Rep-Index: 29,65	REC_NC	4,387	158	1	0,028
Aumente o item: Trabalho Completo em Conferência (CWPCP) para 71	REC_CWPCP	7,401	470	1	0,016

Na tabela anterior observa-se que as recomendações estão relatadas em ordem de relevância para a reputação, isto responde ao questionamento de “quando fazer”. As próprias recomendações respondem ao questionamento ”do que fazer” e as personalizações respondem ”como fazer”.

6. Conclusões e Trabalhos Futuros

Este trabalho apresentou uma abordagem para gerar recomendações de plano de carreira de pesquisadores utilizando a similaridade de perfil e reputação acadêmica. A similaridade de perfil foi definida com base nas informações textuais do Rep-Model por meio da técnica de TF-IDF. Foi realizado um experimento com o objetivo de localizar a melhor técnica para esta tarefa. Criou-se para isso, diversas categorias que possuem todas as informações textuais de seus pesquisadores e comparou-se as mesmas com o conjunto total de pesquisadores. Experimentou-se três funções de correlação, nove de distância e uma de similaridade, bem como cinco classes de pré-processamento. Ao final, observou-se que a combinação `Log-likelihood` e `ClassicAnalyzer` obteve os melhores resultados. Ao final dos experimentos, gerou-se as recomendações personalizadas e não personalizadas para diversos grupos de teste. As recomendações foram avaliadas com base na métrica *coverage* e *diversity* e apresentaram resultados satisfatórios. Por se tratar de uma abordagem inédita não temos referências para comparar com o *baseline*.

Referências

Beel, J., Langer, S., Genzmehr, M., Gipp, B., Breitinger, C., e Nürnberger, A. (2013). 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*, number April, páginas 15–22, New York, New York, USA. ACM Press.

Cervi, C. R., Galante, R., e Oliveira, J. P. M. d. (2013a). Application of scientific metrics to evaluate academic reputation in different research areas. *in: XXXIV International Conference on Computational Science(ICCS) 2013*. Bali, Indonesia.

Cervi, C. R., Galante, R., e Oliveira, J. P. M. d. (2013b). Comparing the reputation of researchers using a profile model and scientific metrics. *in: XIII IEEE International Conference on Computer and Information Technology(CIT)*. Sydney, Australia.

Ekstrand, M. D., Kannan, P., Stemper, J. A., Butler, J. T., Konstan, J. A., e Riedl, J. T. (2010). Automatically Building Research Reading Lists. *RecSys2010*, páginas 159–166.

Huang, W., Kataria, S., Caragea, C., Mitra, P., Giles, C. L., e Rokach, L. (2012). Recommending citations. *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*, página 1910.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632.

Middleton, S. E., Shadbolt, N. R., e De Roure, D. C. (2004). Ontological user profiling in recommender systems. *ACM Transactions on Information Systems*, 22(1):54–88.

Page, L., Brin, S., Motwani, R., e Winograd, T. (1999). The pagerank citation ranking: bringing order to the web. *Stanford InfoLab*.

Sugiyama, K. e Kan, M.-Y. (2013). Exploiting Potential Citation Papers in Scholarly Paper Recommendation. *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries - JCDL '13*, página 153.

Sugiyama, K. e Kan, M. Y. (2015). A comprehensive evaluation of scholarly paper recommendation using potential citation papers. *International Journal on Digital Libraries*, 16(2):91–109.

Vivian, G. R. e Cervi, C. R. (2016a). 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*, páginas 108–117.

Vivian, G. R. e Cervi, C. R. (2016b). 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*, páginas 159–162.

Vivian, G. R., Cervi, C. R., e Rovadosky, D. N. (2016). Using selection attribute algorithms from data mining to complement the rep-index. *IADIS International Journal on WWW/Internet*, 15:219–226.

Zhang, Z. e Li, L. (2010). A research paper recommender system based on spreading activation model. In *The 2nd International Conference on Information Science and Engineering*, páginas 928–931. IEEE.