Universidade de Pernambuco - UPE
Campus Garanhuns
Licenciatura em Computação

DÉBORA DA CONCEIÇÃO ARAÚJO

DESENVOLVIMENTO E ANÁLISE DE UM SISTEMA DE RECOMENDAÇÃO DE ARTIGOS CIENTÍFICOS

Trabalho de Conclusão de Curso

Garanhuns
Dezembro, 2016

DÉBORA DA CONCEIÇÃO ARAÚJO

DESENVOLVIMENTO E ANÁLISE DE UM SISTEMA DE RECOMENDAÇÃO DE ARTIGOS CIENTÍFICOS

Monografia apresentada como requisito parcial para obtenção do diploma de Licenciado em Computação pela Universidade de Pernambuco - Campus Garanhuns.

ORIENTADOR: JOÃO FAUSTO LORENZATO DE OLIVEIRA

Garanhuns
Dezembro, 2016

Toda honra e toda glória ao Senhor.

AGRADECIMENTOS

A Deus, meu Senhor e meu Guia.

Aos meus pais, Edinaldo e Vanda, e à minha irmã, Dara Vitória, por todo amor, apoio e esforços dedicados.

Ao namorado e amigo, Iago Ferreira, por cada conselho, por toda paciência (rs) e incentivo.

Às primas e amigas, Jaciara, Jadielma, Edilene, Jaqueline e Mayra, por estarem sempre presentes, pela paciência (de novo), por todo carinho e apoio.

Aos amigos, Marciele, Aleandro, Dorghisllany e Izabele, que estiveram presentes durante toda jornada acadêmica, e antes mesmo dela, pela paciência (de novo), pelas cobranças (rs) e amizade.

Aos amigos que a graduação me presenteou, Vangéssica, Máverick, Leonardo, Paulo, Alisson, Poliana e todos os outros que participaram dessa caminhada, tornaram mais felizes os dias e contribuíram com meu crescimento pessoal e intelectual.

A todos os professores que fazem o curso de Licenciatura em Computação, por todos os ensinamentos e apoio dedicados. Em especial, agradeço ao professor Fausto Lorenzato pela orientação deste trabalho, pela paciência e confiança em mim depositada.

Por fim, agradeço a todos que de forma direta ou indireta tornaram possível a elaboração deste trabalho, a vocês meu reconhecimento e gratidão!

“UM PASSO À FRENTE E VOCÊ NÃO ESTÁ MAIS NO MESMO LUGAR’”

CHICO SCIENCE

RESUMO

Com o advento da internet uma grande quantidade de informação passa a ser produzida diariamente. No cenário acadêmico é possível verificar milhares de artigos, teses e dissertações disponíveis nas diversas bibliotecas digitais. Nesse sentido, esse trabalho apresenta um sistema de recomendação de artigos científicos que pretende tornar mais rápido e preciso o processo de busca por produções acadêmicas. Para montar o perfil do pesquisador foram utilizadas técnicas de mineração de texto, além das distâncias euclidianas para realização do cálculo da similaridade com os demais *papers*. Para validação das recomendações foi utilizada a métrica de avaliação *Normalized Discounted Cumulative Gain*. De acordo com a NDCG as recomendações foram bem sucedidas em relação a alguns pesquisadores e com baixa acurácia em relação a outros.

Palavras-chave: sistemas de recomendação; *overloading*; recomendação de artigos.

ABSTRACT

With the advent of the internet a great amount of information happens to be produced daily. In the academic scenario it is possible to verify thousands of articles, theses and dissertations available in the various digital libraries. In this sense, this work presents a system of recommendation of scientific articles that intends to make the process of search for academic productions faster and more accurately. In order to set up the profile of the researcher, text mining techniques were used, in addition to the Euclidean distances to calculate the similarity with the other papers. To validate the recommendations, the Normalized Discounted Cumulative Gain evaluation metric was used. According to NDCG the recommendations were successful in relation to some researchers and with low accuracy in relation to others.

Keywords: recommendation systems; overloading; recommendation of articles

UNIVERSIDADE DE PERNAMBUCO - UPE
CAMPUS GARANHUNS
LICENCIATURA EM COMPUTAÇÃO
DISCIPLINA: TRABALHO DE CONCLUSÃO DE CURSO II

VERSÃO FINAL DA MONOGRAFIA

Avaliação do Presidente da Banca

Discente: Débora da Conceição Araújo

Título: Desenvolvimento e Análise de um Sistema de Recomendação de Artigos Científicos

Orientador(a): João Fausto Lorenzato de Oliveira

Presidente da Banca: Emanoel Francisco Spósito Barreiros

Data: 12/12/2016

Após a apresentação da versão final monografia ao Presidente da Banca da Defesa, a mesma foi considerada:

(X) Aprovada ( ) Reprovada

Comentários

Presidente da Banca: Emanoel Francisco Spósito Barreiros

SUMÁRIO

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

CAPÍTULO 1

Neste capítulo será apresentada uma introdução ao tema do trabalho, bem como o que justifica a realização do mesmo, juntamente aos seus objetivos.

1.1 INTRODUÇÃO

Com a popularização das TIC e, mais recentemente, da internet, é possível observar uma grande quantidade de conteúdos aos quais as pessoas possuem acesso. São "dezenas ou centenas de canais de TV, milhares de filmes, milhões de CDs e livros e bilhões de documentos on-line" (Sampaio, 2006, p. 2).

Mesmo sendo algo muito vantajoso, existe um problema relacionado a essa grande quantidade de informações, conhecido como "*information overload*", ou sobrecarga de informação (Sugiyama e Kan, 2013)[1]. Imersos nesse cenário de *overloading*, torna-se perceptível o fato de que nossas buscas na *Web* possuem cada vez mais resultados, no entanto, grande parte dos resultados encontrados são irrelevantes para o objetivo dos internautas, dessa forma, a busca por conteúdos relevantes acaba se tornando uma tarefa exaustiva.

Nas relações humanas, usamos diversos artifícios para filtrar informações importantes, como cartas de recomendação, opiniões de revisores sobre filmes e livros, impressos de jornais, entre outros (Reategui e Cazella, 2005). Para tratar dessa sobrecarga em um cenário muito maior, o da *Web*, e dessa forma tornar o processo de seleção de conteúdos mais eficaz, vem se disseminando abordagens relacionadas à recomendação personalizada de conteúdos por meio de "sistemas computacionais que automatizam ou facilitam esse processo de seleção" (Sampaio, 2006, p. 2). Esses sistemas computacionais se tornaram conhecidos como Sistemas de Recomendação (SR).

Os SR, definidos por Cazella *et al.* (2009) como sistemas que visam auxiliar o usuário na busca e seleção de um conteúdo focado em seu perfil, são amplamente usados por empresas de e-commerce para recomendação personalizada de produtos e serviços de acordo com as buscas e preferências de seus clientes. A Amazon.com, por exemplo, faz uso dessa tecnologia desde 1999 (Schafer *et al.*, 1999). O stream de vídeo, Netflix, também faz uso de um sistema

[1] Traduzido pelo autor.

de recomendação que se baseia nas pesquisas de cada usuário e de usuários que possuem um perfil similar. Os SR também estão presentes nas redes sociais indicando possíveis conhecidos, páginas do interesse do usuário, em parceria com *e-commerces* – recomendando produtos e serviços.

Assim como em e-commerces e redes sociais, sistemas que realizam recomendações personalizadas também se tornam relevantes no meio acadêmico, tendo em vista que o cenário de *overloading* também está relacionado aos conteúdos acadêmicos (livros, artigos e teses). Nesse sentido (Sugiyama e Kan, 2013)[2] apontam ao fato de que os pesquisadores modernos possuem acesso a uma quantidade de informações sem precedentes.

Ao analisar possíveis soluções para o problema de busca por conteúdos acadêmicos relevantes, é possível destacar o papel dos repositórios que permitem a indexação de textos, as DLs, do inglês *Digital Libraries* (Bibliotecas Digitais). Através desses repositórios o pesquisador pode filtrar suas buscas por meio de mecanismos como palavras-chave, por autor, título, entre outros. Alguns repositórios com grande acervo de conteúdos são: o *Directory of Open Access Journals*[3] (DOAJ), que "indexa cerca de 2.200 publicações periódicas científicas" (Blattmann e Bomfá, 2006); e a *Scientific Electronic Library Journal*[4] (ScieLo), que se destaca no que diz respeito à publicações brasileiras. Outros exemplos de DLs que possuem um acervo considerável de conteúdos são: a *Elsevier*[5] (*Direct Science*); e o *IEEE Xplore*[6].

Como se fez possível perceber, as bibliotecas digitais chegaram para facilitar a vida dos pesquisadores, organizando os conteúdos e indicando-os de acordo com as pesquisas realizadas em seus sistemas. No entanto, esses sistemas possuem algumas limitações. Uma das principais limitações observadas, dentre as bibliotecas supracitadas, está relacionada ao fato de que as mesmas não possuem uma funcionalidade que recomende os *papers*[7] de forma personalizada, de acordo com o perfil dos pesquisadores. Dessa forma, a cada vez que o pesquisador precise buscar novos artigos, ele precisa redefinir

[2] Traduzido pelo autor;

[3] Saiba mais em https://doaj.org/

[4] Saiba mais em http://www.scielo.org/

[5] Saiba mais em http://www.sciencedirect.com/

[6] Saiba mais em http://ieeexplore.ieee.org/

[7] *Papers*: "artigos científicos", em livre tradução.

manualmente as palavras-chave ou demais critérios de busca, para que tenha acesso às pesquisas com o tema desejado.

Tendo em vista que o processo de busca pode ser bastante cansativo e demorado, além de impossibilitar uma busca bem sucedida, fazendo com que diversos trabalhos não sejam encontrados, Chen (2003, p. 17) indica a utilização de Sistemas de Recomendação, definindo-os como "um paradigma emergente para dispensar usuários cotidianos de configurar e instruir manualmente seus sistemas computacionais".

Diante destas questões, a presente pesquisa pretende resolver a seguinte problematização: como tornar mais rápido e preciso o processo de busca e seleção de artigos científicos, tendo em vista o grande número de *papers* disponibilizados na *Web*?

Como possível solução para essa problematização, este trabalho apresenta um modelo para recomendação de artigos científicos personalizados de acordo com o perfil do usuário. O sistema, que será apresentado de forma detalhada no Capítulo 3, foi baseado no apresentado por Sugiyama e Kan (2010), e divide os pesquisadores em dois grupos gerais de usuários: pesquisadores juniores (Jr.) e pesquisadores seniores (S.^{or}). Os usuários que se enquadram no grupo de pesquisadores Jr. representam os pesquisadores jovens, que possuem apenas uma publicação e que ainda não são citados em outros *papers*. Em contrapartida, os usuários que fazem parte do grupo de pesquisadores S.^{or} são aqueles que possuem muitas publicações e que são citados em outras produções.

Para fazer as recomendações para os pesquisadores juniores, o sistema deve levar em consideração os interesses dos usuários, baseando-se em sua publicação e nos *papers* que foram referenciados em seu trabalho. No caso das recomendações para pesquisadores S.^{or}, além dos *papers* do pesquisador e dos que ele referencia, também serão utilizadas as demais produções que o citam. Esse processo será detalhado na seção 3.3, bem como as respectivas equações utilizadas para realização das recomendações.

Alguns relatos acerca de sistemas de recomendação no âmbito acadêmico/educacional podem ser encontrados em (Cazella *et al.*, 2012; Aguiar

et al., 2015; Carvalho *et al.*, 2014), porém a ênfase desses trabalhos está voltada à recomendação de Objetos de Aprendizagem[8].

O sistema proposto neste trabalho monográfico acrescenta ao apresentado em Sugiyama e Kan (2010) a soma das distâncias euclidianas, como uma técnica de aprendizagem de métrica para realizar a recomendação dos *papers*, para dessa forma, propiciar uma análise comparativa das duas versões do sistema e, assim, encontrar a versão que recomenda os *papers* de forma mais rápida e precisa.

1.2 MOTIVAÇÃO E JUSTIFICATIVA

As Bibliotecas Digitais, segundo Sugiyama e Kan (2010)[9], apesar de estarem em sua era de ouro, não aproveitam totalmente o contexto do usuário, tendo em vista que não realizam previsões de acordo com seu histórico, apenas geram resultados levando em consideração pesquisas realizadas manualmente.

Utilizando-se de Sistemas de Recomendação é possível solucionar essa lacuna apontada nas DLs, tendo em vista que os SR se utilizam de informações dos usuários para aumentar a capacidade e eficácia do processo de indicação de conteúdos (Resnick e Varian, 1997).

Dessa forma, torna-se relevante a criação de um sistema que recomende artigos científicos de acordo com o perfil dos pesquisadores, para facilitar o processo de busca por produções do interesse do usuário.

Ao analisar o procedimento realizado por Sugiyama e Kan (2010), observou-se que, dentre as técnicas de aprendizagem de métricas adotadas, a que realizou as melhores recomendações foi a *Cosine Similarity*, uma técnica com características lineares. Assim, o presente trabalho irá adicionar a técnica das distâncias euclidianas, apontada por Huang (2008) como uma das mais utilizadas para problemas dessa natureza, para, ao fim, analisar as recomendações realizadas por ambas as técnicas.

Dessa maneira, se torna possível uma análise das duas versões do sistema: a primeira, proposto por Sugiyama e Kan (2010); e a segunda, com a adição das distâncias euclidianas como medida para calcular a similaridade e realizar as

[8] Objetos de Aprendizagem podem ser definidos como quaisquer entidades, digitais ou não, que possam ser utilizadas, reutilizadas ou referenciadas no ensino assistido por tecnologia (Wiley 2002).

[9] Traduzido pelo autor.

recomendações dos artigos. Ao final da análise, os resultados acerca da qualidade das recomendações dos artigos poderão ser comparados, para se verificar a versão do sistema que realiza as recomendações mais precisas.

1.3 OBJETIVOS

1.3.1 Objetivo Geral

- Projetar e desenvolver um sistema que recomende artigos científicos de acordo com o perfil do pesquisador, utilizando técnicas de Inteligência Artificial para tornar as recomendações mais rápidas e precisas.

1.3.2 Objetivos Específicos

- Implementar o SR proposto na literatura por Sugiyama e Kan (2010) para fins de comparação;
- Desenvolver um sistema computacional inteligente, utilizando como cálculo de similaridade dos *papers* a soma das distâncias euclidianas, para resolução do problema de recomendação personalizada de artigos científicos;
- Avaliar os resultados dos sistemas utilizando a métrica de desempenho/avaliação *Normalized Discounted Cumulative Gain* (NDCG).

CAPÍTULO 2

Este capítulo irá apresentar uma revisão da literatura, onde serão evidenciados trabalhos relacionados ao tema da pesquisa, além de explicitações acerca dos sistemas de recomendação, bem como referentes aos conceitos de esquemas de ponderação e métricas de desempenho que permitem a realização/avaliação das recomendações do sistema.

2.1 TRABALHOS RELACIONADOS

Tendo um sistema de recomendação como um sistema que funciona, de forma geral, utilizando as informações de seus usuários e comparando essas informações entre si, a fim de selecionar usuários com perfis similares e, assim, recomendar conteúdos de acordo tanto com a navegação de cada pessoa em específico quanto com a navegação de pessoas com perfis similares, Silva Filho e Cazella (2005) apresentam o STAR, um framework para o problema de recomendação de artigos científicos.

Silva Filho e Cazella (2005) detalham o STAR como um sistema que se utiliza tanto de informações coletadas na plataforma Lattes[10] quanto de informações coletadas por meio de questionário. Dentre as perguntas que contém no questionário, estão perguntas acerca das áreas de interesse do pesquisador, além do tempo de experiência que o mesmo possui em cada área.

Em Sugiyama e Kan (2010) também é proposto um sistema para recomendação de artigos científicos. Os autores dividem os pesquisadores em dois grupos (Jr e S.or), como proposto no presente trabalho, e realizam recomendações de acordo com as produções dos pesquisadores e de acordo com as produções referenciadas pelos mesmos.

O presente trabalho se diferencia das pesquisas encontradas na literatura por não necessitar que o pesquisador responda questionamentos de forma direta, levando em consideração dados indexados de forma automática, como as produções do próprio pesquisador, as referências usadas por ele e as produções que os citam. Também se difere da pesquisa descrita por Sugiyama e Kan (2010)

[10] Saiba mais em <http://lattes.cnpq.br/>.

por fazer adicionar ao modelo proposto o esquema das distâncias euclidianas para verificar a similaridade dos *papers* e, assim, realizar as recomendações.

2.2 SISTEMAS DE RECOMENDAÇÃO

Cazella *et al.* (2010), define sistemas de recomendação inteligente, como um sistema ao qual as pessoas fornecem informações como entradas e essas informações são processadas de modo a considerar potenciais interesses do indivíduo.

Os sistemas de recomendação, em geral, comparam perfis de usuários, tendo em vista pessoas que possuem perfis/interesses parecidos. Tomando como exemplo um cenário de recomendação de filmes, duas pessoas costumam assistir filmes do mesmo gênero (terror, suspense), assim, as escolhas de uma dessas pessoas influenciam no que será recomendado para a outra. Além deste, tem-se diversos aspectos sendo considerados para a realização de uma recomendação, como as buscas do usuário, por exemplo, esses aspectos variam conforme o problema que se deseja solucionar.

2.3 ESQUEMAS DE PONDERAÇÃO

Ao lidar com grandes bases de dados, é possível notar que nem todos os termos existentes são igualmente úteis para determinado experimento. Nesse sentido, (Baeza-Yates e Ribeiro-Neto, 2013, p. 31) pontuam que "decidir a importância de um termo na sumarização dos conteúdos de um documento não é uma tarefa trivial".

Recorrentes pesquisas apresentam abordagens que visam automatizar esse processo de ponderação da relevância de termos [Spark Jones, 1974; Salton e Yang, 1973]. Assim, duas formas de ponderação são amplamente utilizadas: a ponderação da frequência dos termos em cada documento; e a ocorrência de dado termo em todos os documentos presentes na base.

A ponderação da frequência dos termos foi proposta por Luhn (1957), baseando-se na suposição de que o peso de um termo k_i que ocorre em um documento d_j é proporcional a frequência do termo f_{ij} (Baeza-Yates e Ribeiro-

Neto, 2013). É preciso levar em consideração que termos com uma frequência elevada podem descrever o tema ao qual se refere o documento. Assim, tem-se a frequência de termo (TF), calculada para cada termo de um documento d_j através da Equação 1.

$$\mathrm{TF} = \frac{Número\ de\ vezes\ em\ que\ um\ termo\ aparece\ no\ documento}{Número\ total\ de\ termos\ presentes\ no\ documento} \qquad (1)$$

Como mostra a Equação 1, o esquema de ponderação de termos TF é calculado a partir da divisão entre a quantidade de vezes que um determinado termo aparece no documento e a quantidade total de termos que o documento possui. Ou seja, levando em consideração que o documento X possui 10 termos no total, e é necessário calcular o TF em relação ao termo "IA", que aparece 2 (duas) vezes em X, tem-se: TF = $\frac{2}{10}$.

O esquema TF é de grande relevância quando se trata da normalização de termo em um determinado documento, tendo em vista que cada documento possui um tamanho diferente, o que altera o número de frequência de um termo. Assim, dividindo a frequência de cada termo pelo número total de termos no documento, é possível normalizar os valores para utilizá-los relacionando com demais documentos.

A segunda forma de ponderação apresentada é o esquema de ponderação IDF (*Inverse Document Frequency*), o mesmo é utilizado para atribuir pesos a um termo de acordo com a sua frequência em uma coleção de documentos (Baeza-Yates e Ribeiro-Neto, 2013). Com o TF, todos os termos são considerados igualmente importantes, tendo seus valores normalizados conforme apresenta a Equação 1, no entanto, dado uma coleção de documentos, alguns termos podem ser mais relevantes à pesquisa que outros, para encontrar tais termos utiliza-se o IDF, conforme Equação 2.

$$\mathrm{IDF} = 1 + \log_e \left(\frac{Número\ total\ de\ documentos}{Quantidade\ de\ documentos\ que\ apresentam\ determinado\ termo}\right) \qquad (2)$$

Assim, levando em consideração uma coleção de 100 (cem) documentos, onde o termo "IA" aparece em apenas 3 (três), para calcular o IDF do termo "IA", tem-se: IDF = 1 + $\log_e$ ($\frac{100}{3}$). Atualmente a ponderação IDF fornece a base para os esquemas de ponderação modernos e é usada por quase todos os sistemas modernos de Recuperação de Informação-RI (Baeza-Yates e Ribeiro-Neto, 2013).

Salton e Yang (1973) propuseram um esquema que combina a frequência dos termos, apresentada no esquema TF, e à relevância ao se considerar uma coleção, apresentada no esquema IDF. Dessa combinação foi montado o esquema TF-IDF, representado pela Equação 3.

$$\text{TF-IDF} = \text{TF} * \text{IDF} \qquad (3)$$

Assim, através do esquema TF-IDF é possível considerar tanto aspectos relativos à frequência de um termo em relação a um determinado documento, quanto a relevância desse termo em relação a uma coleção de documentos.

2.4 MÉTRICAS DE DESEMPENHO/AVALIAÇÃO

As Métricas de Desempenho, também conhecidas por métricas de avaliação, são, de acordo com Matos *et al.* (2009), medidas utilizadas para analisar resultados gerados a partir da mineração de textos. Ainda segundo os mesmos autores, essas métricas são utilizadas para avaliar a qualidade dos resultados em diversas áreas do conhecimento. Dentre as medidas de desempenho existente, este trabalho destaca a *Normalized Discounted Cumulative Gain* (NDCG) ou, em livre tradução, ganho normalizado, acumulado com desconto.

A NDCG, ou DCG normalizada, foi escolhida como métrica de desempenho para o problema de recomendação de artigos, pois, de acordo com Sugiyama e Kan (2010), é uma medida que dá maior precisão (peso) para

formulação de um ranking de documentos, além de incorporar diferentes níveis de relevância[11].

Para calcular a NDCG, neste trabalho, utilizou-se a mesma noção binária aplicada em Sugiyama e Kan (2010), tendo em vista a comparação das recomendações de ambos os trabalhos, descrita na seção 4.1 de resultados. Assim, tem-se a relevância dos *papers* classificada em 1, para relevantes ao usuário/pesquisador, e 0 (zero), para irrelevantes ao usuário/pesquisador.

As Equações 4 e 5 descrevem o cálculo de NDCG, onde DCG é uma medida que verifica a utilidade, ou ganho , de um documento com base em sua posição na lista de resultados e NDCG realiza essa verificação em relação a todas as consultas, para obter uma medida do desempenho médio do algoritmo de classificação (Wang *et al.* 2013):

$$\mathrm{DCG} = \sum \frac{2^{rel}-1}{log_2(i+1)} \qquad (4)$$

$$\mathrm{NDCG(i)} = \mathrm{DCG(i-1)} + \frac{rel(i)}{\log(i)} \qquad (5)$$

Nas Equações 4 e 5 tem-se *rel* representando a lista de documentos relevantes e *i* variando entre 0 (zero) e o tamanho de *rel.*

Sugiyama e Kan (2010) apontam para a utilização do DCG normalizado, em relação a todos os usuários, para demonstrar a precisão das recomendações. Neste trabalho foram submetidos ao DCG normalizado os *papers* selecionados em dois experimentos, à serem detalhados na seção 4.2 de resultados e discussões.

[11] Traduzido pelo autor.

CAPÍTULO 3

Este capítulo apresenta os procedimentos metodológicos adotados para o desenvolvimento da presente pesquisa.

3.1 ELABORAÇÃO DO MÉTODO PROPOSTO

Segundo Lakatos e Marconi (2003), a presente pesquisa pode ser caracterizada como bibliográfica devido a busca feita acerca de trabalhos relacionados ao seu objetivo; documental devido ao levantamento feito, em documentos oficiais da base de dados *NUS Computing*, da *National University of Singapore*, para a realização das recomendações feitos pelo sistema; experimental devido ao experimento desenvolvido e; quantitativa devido aos métodos adotados para medição da eficiência do sistema proposto.

3.2 BASE DE DADOS

Para montar o perfil do usuário foram levados em consideração dados disponibilizados na base *Nus Computing*, da National University of Singapure. Tais dados mantêm organizadas as palavras presentes em cada artigo de 15 (quinze) pesquisadores juniores e de 13 (treze) pesquisadores seniores. Essa divisão ocorre, pois os pesquisadores Jr possuem apenas um artigo e as referências do mesmo, enquanto os S.^{or} possuem mais artigos, as referências destes, além de pessoas que o citam. A base conta ainda com as palavras de cada um dos artigos referenciados pelos pesquisadores e, no caso dos pesquisadores seniores, também é possível encontrar as palavras presentes em cada um dos artigos que citam tais pesquisadores.

3.3 MODELO PROPOSTO

Como é possível observar na seção 3.2 deste trabalho, a base de dados utilizada para realização desta pesquisa divide os pesquisadores em juniores e seniores. Os pesquisadores classificados como juniores são aqueles que possuem apenas um artigo publicado e não possuem produções que citam seu trabalho. Desse modo, para realização das recomendações para este tipo de

pesquisador, as referências utilizadas por ele para construção do seu *paper* são consideradas pelo modelo.

No caso dos pesquisadores seniores, além de possuir mais *papers* publicados, a base de dados conta também com trabalhos que citam os artigos do pesquisador, bem como com as referências que ele utilizou para embasar suas produções. Além destes, vale ressaltar que no caso dos pesquisadores seniores o modelo ainda trata da variável tempo, tendo em vista que um pesquisador pode mudar sua área de interesse (SUGIYAMA e KAN, 2010), assim artigos mais recentes são tratados com maior relevância.

Dessa forma, para realizar as recomendações, o sistema que será detalhado na seção próxima, 3.4, faz uso desses dados para montar um perfil para cada pesquisador, similar ao proposto em Sugiyama e Kan (2010), diferenciando-se por inserir a técnica das distâncias euclidianas no cálculo da similaridade. Esta técnica é apontada por Huang (2008) como uma técnica bastante utilizada para este tipo de problema.

3.4 PROCESSO DE DESENVOLVIMENTO

Mesmo com o modelo apresentado contando com dois tipos de usuários – pesquisadores seniores e pesquisadores juniores -, o sistema desenvolvido e analisado nesta pesquisa priorizou o cenário dos pesquisadores juniores, pois os mesmos, de acordo com Sugiyama e Kan (2010), são mais complexos no que se refere a recomendação personalizada de artigos, tendo em vista a pouca quantidade de informações em relação aos pesquisadores seniores.

No que tange o desenvolvimento do sistema, o primeiro passo consiste na criação de um perfil de usuário referente a cada pesquisador. Esse perfil se dá por meio da criação de uma matriz que comporta todas as palavras dos *papers* do pesquisador, bem como a frequência de cada uma dessas palavras. Após a criação deste vetor, os dados são normalizados por meio do esquema de ponderação TF-IDF, apresentado na seção 2.3 deste trabalho. Assim tem-se, para cada pesquisador, uma matriz, conforme apresenta a Figura 1.

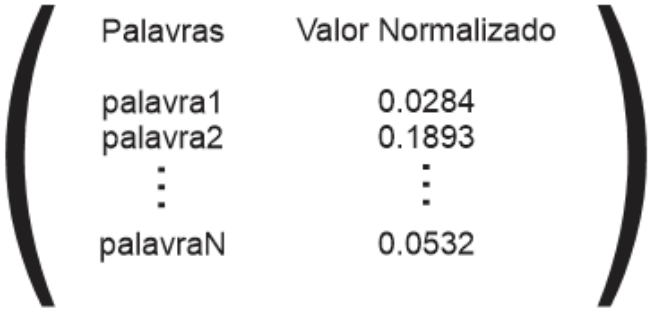

Figura1. Matriz com Perfil do Usuário

Após a normalização dos dados, foram implementados dois esquemas de ponderação para verificar a similaridade dos *papers*. O *cosine similarity*, utilizado por Sugiyama e Kan (2010), que foi implementado para fins de comparação, e a técnica das distâncias euclidianas, proposta nesta pesquisa, tendo em vista que esta técnica realiza o cálculo da distância entre dois vetores, formando perfis de usuários, assim este cálculo apresenta a similaridade entre dois perfis (COSTA *et al*. 2013).

Tendo em vista que o valor da distância está no intervalo [0, ∞[, quanto mais próximo este valor estiver de 0 (zero), maior a similaridade dos *papers* e, consequentemente, quanto mais distante de 0 (zero) o valor estiver, menor a similaridade.

Quando se tem o vetor referente aos pesquisadores (ver figura1), os mesmos passos são repetidos em relação às referências, assim, cada referência tem seus dados organizados em matrizes nx2, sendo a primeira coluna composta pelas palavras, e a segunda coluna pela frequência (TF-IDF) de cada termo. Assim, tem-se a quantidade de linhas relacionadas a quantidade de palavras existentes em cada artigo de referência.

O próximo passo consiste em calcular os pesos, por meio do *cosine similarity* (Sugiyama e Kan, 2010), e das distâncias euclidianas apresentados neste trabalho. O cálculo se dá entre o vetor que representa o perfil do usuário e cada um dos vetores que representam cada referência.

$$Cosine\ Similarity = \sum\left(\frac{vetorPesquisador_{i1} * vetorReferência_{i1}}{|vetorPesquisador_{i1}| * |vetorReferência_{1}|}\right) \qquad (6)$$

A Equação 6 é válida para as palavras em comum entre ambos os *papers*, assim, entende-se que palavras importantes, ou seja, com maior frequência, ficarão com peso maior.

Para calcular a soma das distâncias euclidianas, o primeiro passo consistiu em deixar os dois vetores (*paper* do pesquisador + referência em evidência) com o mesmo tamanho. Para tal, adotou-se um esquema em que as palavras que existiam em um vetor, mas não existiam no outro, passavam a estar nos dois vetores, com valores 0 (zero) sendo atribuídos a elas. Como mostra a sequência de figuras 2 e 3.

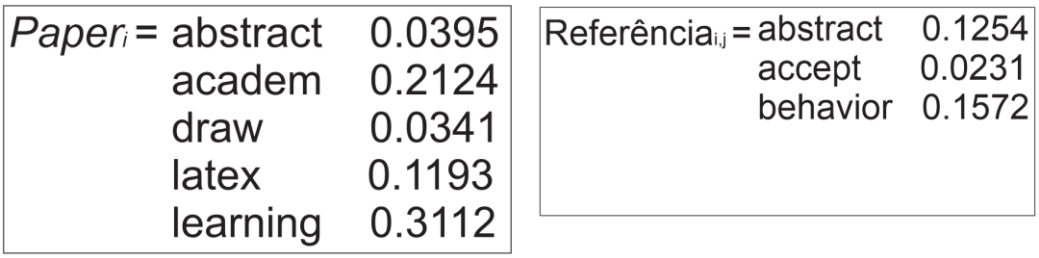

Figura 2. Palavras presentes nos *papers* do pesquisador e da referência.

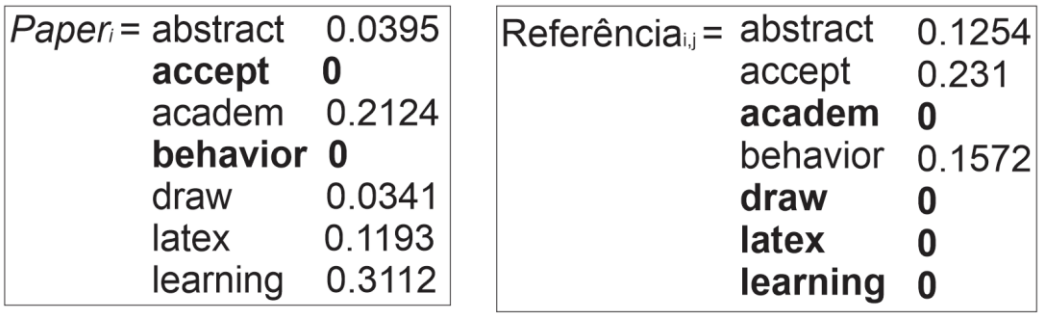

Figura 3. Vetores com mesmo tamanho

Tendo os dois vetores com o mesmo tamanho, o próximo passo consiste no cálculo das distâncias euclidianas, representado na Equação 7.

$$\text{Distância\_Euclidiana} = \sqrt{\sum(a_{i,j} - b_{i,j})^2} \qquad (7)$$

Na Equação 6, “a” representa as palavras que estão contidas no artigo do pesquisador, enquanto “b” representa as palavras presentes na referência. Vale salientar que esses passos se repetem em relação a todas as referências de cada pesquisador.

Para finalizar o cálculo e obter a similaridade realizada pela técnica das distâncias euclidianas, tem-se:

$$\text{Similaridade}(\textit{pesquisador, referência}) = \frac{1}{Distância\_Euclidiana} \qquad (8)$$

Assim, tem-se dois perfis de usuários para cada pesquisador, sendo um perfil montado considerando a similaridade do cosseno e outro perfil montado considerando a soma das distâncias euclidianas.

Com o perfil do pesquisador montado, foi calculada a similaridade, por meio do *Cosine Similarity* e das Distâncias Euclidianas, entre o perfil do pesquisador e cada um dos *papers* que poderiam ser recomendados. Este trabalho contou com um total de 597 artigos disponíveis para recomendação.

Assim, na seção 4.1 são apresentados os resultados acerca da qualidade das recomendações de ambas as técnicas (*Cosine Similarity* e Distância Euclidiana), qualidade, esta, testada por meio da *Normalized Discounted Cumulative Gain* (NDCG), detalhada na seção 2.4 deste trabalho.

CAPÍTULO 4

Neste capítulo serão apresentados os resultados alcançados ao final do projeto, juntamente com as contribuições que o mesmo deve trazer à comunidade acadêmica.

4.1 RESULTADOS E DISCUSSÕES

Ao término do desenvolvimento do sistema a acurácia[12] dos *papers* recomendados pôde ser mensurada por meio da métrica de desempenho *Normalized Dicounted Cumulative Gain*, detalhada na seção 2.4 deste trabalho.

Nesta seção serão apresentados os trabalhos recomendados para cada pesquisador, bem como os trabalhos que deveriam ser recomendados, de acordo com a base de dados utilizada.

Vale ressaltar que mesmo esse trabalho tendo priorizado os 10 artigos com maior similaridade, muitos dos pesquisadores possuem mais de 10 *papers* para recomendação e apenas dois pesquisadores, o *Research* 2 e o *Research* 3 (ver tabelas 2 e 3), possuem menos, estando, respectivamente, o Pesquisador 2 com 4 e o Pesquisador 3 com 8 artigos com possibilidade de recomendação. Assim, tem-se o cálculo do NDCG realizado utilizando todos os artigos que a base contém marcados com possibilidade de recomendação, mesmo para os casos em que os pesquisadores possuíam mais de 10 artigos com tal possibilidade.

Todas as tabelas apresentadas nesta seção estão organizadas de modo a contemplar sempre na primeira coluna os *papers* recomendados pelo sistema proposto, utilizando as Distâncias Euclidianas para o cálculo da similaridade, enquanto a segunda coluna traz os artigos que estavam marcados na base com possibilidade de recomendação para cada pesquisador.

Ao final, tem-se os resultados dos valores da DCG normalizada encontrados a partir da similaridade pelas distâncias euclidianas comparados aos valores presentes na base de dados *Nus Computing*.

[12] Acurácia: proximidade entre o valor obtido experimentalmente e o valor verdadeiro na medição de uma grandeza física.

Tabela 1. Recomendações R*esearch* 1

Artigos Recomendados pelo Sistema	Artigos Recomendados na Base de Dados
P06-1074_recfv.txt	P00-1002.txt
P05-1062_recfv.txt	P00-1044.txt
P01-1026_recfv.txt	P00-1062.txt
P04-1028_recfv.txt	P00-1064.txt
P04-1012_recfv.txt	P01-1026.txt
P06-1128_recfv.txt	P01-1064.txt
P01-1049_recfv.txt	P03-1028.txt
P05-1008_recfv.txt	P03-1029.txt
P00-1010_recfv.txt	P04-1056.txt
P01-1070_recfv.txt	P05-1058.txt

A tabela 1. mostra os *papers* referentes ao Pesquisador 1. Ao todo haviam 12 artigos com possibilidade de recomendação, desses, a similaridade encontrada pelas Distâncias Euclidianas apontou uma acurácia com valor 0.094. Este valor foi obtido por meio da medida de desempenho NDCG, detalhada na seção 2.4.

Tabela 2. Recomendações R*esearch* 2

Artigos Recomendados pelo Sistema	Artigos Recomendados na Base de Dados
P04-1059_recfv.txt	P00-1034.txt
P01-1005_recfv.txt	P02-1001.txt
P04-1033_recfv.txt	P03-1030.txt
P04-1078_recfv.txt	P03-1033.txt
P00-1063_recfv.txt	P06-1061.txt
P03-1035_recfv.txt	
P06-1074_recfv.txt	
P06-1129_recfv.txt	
P04-1012_recfv.txt	
P03-1061_recfv.txt	

Considerando o Pesquisador 2, é possível verificar, através da tabela 2. que não há artigos iguais entre os recomendados pelo sistema e os presentes na base de dados. Tem-se então um valor de acurácia 0 (zero), de acordo com a DCG normalizada.

Tabela 3. Recomendações R*esearch* 3

Artigos Recomendados pelo Sistema	Artigos Recomendados na Base de Dados
P02-1047_recfv.txt	P00-1002.txt
P06-1049_recfv.txt	P00-1010.txt
P05-1008_recfv.txt	P00-1023.txt
P05-1061_recfv.txt	P00-1038.txt
P05-1070_recfv.txt	P00-1039.txt
P01-1040_recfv.txt	P00-1043.txt
P03-1069_recfv.txt	P00-1051.txt
P00-1043_recfv.txt	P00-1052.txt
P04-1087_recfv.txt	P01-1005.txt
P00-1074_recfv.txt	P01-1008.txt

A tabela 3 apresenta os *papers* relacionados ao Pequisador 3. Vale ressaltar que mesmo a tabela 3 apresentando apenas os 10 primeiros artigos, o Pesquisador 3 possui um total de 36 artigos com possiblidade de recomendação. Para este, o valor de acurácia, de acordo com a DCG normalizada foi de 0.1, este valor, assim como ocorreu com o Pesquisador 1 (ver tabela 1.), pode ser considerado baixo.

Tabela 4. Recomendações R*esearch* 4

Artigos Recomendados pelo Sistema	Artigos Recomendados na Base de Dados
P06-1121_recfv.txt	P00-1002.txt
P01-1068_recfv.txt	P00-1008.txt
P05-1062_recfv.txt	P00-1062.txt
P02-1019_recfv.txt	P00-1064.txt
P03-1061_recfv.txt	P00-1071.txt
P06-1126_recfv.txt	P01-1009.txt
P06-1132_recfv.txt	P01-1021.txt
P01-1017_recfv.txt	P01-1028.txt
P05-1057_recfv.txt	P01-1033.txt
P02-1021_recfv.txt	P01-1047.txt

O Pesquisador 4 tem seus artigos, recomendados e para recomendação, descritos na tabela 4. Para este pesquisador o valor de acurácia foi 0(zero), tendo em vista que o sistema não recomendou os artigos que foram apontados na base para sua recomendação.

Tabela 5. Recomendações R*esearch* 5

Artigos Recomendados pelo Sistema	Artigos Recomendados na Base de Dados
P05-1057_recfv.txt	P00-1041.txt
P01-1017_recfv.txt	P00-1046.txt
P00-1061_recfv.txt	P00-1059.txt
P04-1015_recfv.txt	P02-1013.txt
P04-1066_recfv.txt	P02-1039.txt
P02-1026_recfv.txt	P03-1031.txt
P03-1035_recfv.txt	P03-1034.txt
P03-1066_recfv.txt	P03-1066.txt
P05-1063_recfv.txt	P04-1015.txt
P04-1006_recfv.txt	P05-1007.txt

Para o Pesquisador 5 haviam 14 artigos com possibilidade de recomendação, destes, os 10 primeiros estão descritos na tabela 5, bem como os 10 artigos que foram recomendados pelo sistema. Para este pesquisador a DCG normalizada apontou uma acurácia de 0.128.

Tabela 6. Recomendações R*esearch* 6

Artigos Recomendados pelo Sistema	Artigos Recomendados na Base de Dados
P05-1062_recfv.txt	P00-1006.txt
P06-1132_recfv.txt	P00-1017.txt
P04-1078_recfv.txt	P00-1032.txt
P03-1020_recfv.txt	P00-1035.txt
P00-1075_recfv.txt	P00-1043.txt
P04-1001_recfv.txt	P00-1046.txt
P05-1008_recfv.txt	P00-1055.txt
P06-1105_recfv.txt	P00-1056.txt
P02-1061_recfv.txt	P00-1067.txt
P06-1145_recfv.txt	P01-1005.txt

Na tabela 6. é possível visualizar os dados referentes ao Pesquisador 6. Este pesquisador obteve um valor de acurácia 0(zero), referente as recomendações. Isso ocorre, pois os artigos recomendados pelo sistema não fazem referência aos artigos dispostos na base de dados para recomendação deste pesquisador.

Tabela 7. Recomendações R*esearch* 7

Artigos Recomendados pelo Sistema	Artigos Recomendados na Base de Dados
P00-1022_recfv.txt	P00-1014.txt
P02-1014_recfv.txt	P00-1015.txt
P06-1006_recfv.txt	P00-1016.txt
P04-1017_recfv.txt	P00-1017.txt
P03-1022_recfv.txt	P00-1018.txt
P04-1020_recfv.txt	P00-1019.txt
P05-1021_recfv.txt	P00-1022.txt
P05-1053_recfv.txt	P00-1024.txt
P05-1008_recfv.txt	P00-1072.txt
P06-1079_recfv.txt	P01-1006.txt

O Pesquisador 7, com dados dispostos na tabela 7, obteve um valor de acurácia de 0.289. Vale ressaltar que dos 10 artigos recomendados pelo sistema, 7 estavam marcados com possiblidade de recomendação, de acordo com a base de dados, porém como este pesquisador possui 47 artigos com possibilidade de recomendação, o cálculo de NDCG retornou um valor abaixo do esperado.

Tabela 8. Recomendações R*esearch* 8

Artigos Recomendados pelo Sistema	Artigos Recomendados na Base de Dados
P01-1005_recfv.txt	P00-1064.txt
P04-1038_recfv.txt	P00-1066.txt
P06-1057_recfv.txt	P00-1069.txt
P06-1058_recfv.txt	P01-1005.txt
P03-1058_recfv.txt	P02-1033.txt
P05-1005_recfv.txt	P02-1044.txt
P00-1054_recfv.txt	P03-1007.txt
P00-1077_recfv.txt	P03-1058.txt
P04-1059_recfv.txt	P04-1036.txt
P04-1033_recfv.txt	P04-1037.txt

A tabela 8 apresenta os dados relativos ao Pesquisador 8. Para este pesquisador o valor de acurácia do sistema foi de 0.418. Vale ressaltar que dos 10 artigos recomendados pelo sistema, 6 estavam marcados pela base com possibilidade de recomendação. No entanto o valor o valor do NDGC encontrado se justifica pelo cálculo que considera os 23 artigos marcados para recomendação do pesquisador.

Tabela 9. Recomendações R*esearch* 9

Artigos Recomendados pelo Sistema	Artigos Recomendados na Base de Dados
P03-1023_recfv.txt	P00-1022.txt
P02-1045_recfv.txt	P00-1023.txt
P06-1006_recfv.txt	P00-1042.txt
P00-1022_recfv.txt	P00-1044.txt
P04-1059_recfv.txt	P00-1051.txt
P04-1020_recfv.txt	P00-1065.txt
P03-1022_recfv.txt	P00-1070.txt
P06-1071_recfv.txt	P01-1005.txt
P01-1020_recfv.txt	P01-1006.txt
P01-1005_recfv.txt	P01-1014.txt

O Pesquisador 9 tem seus artigos - recomendados pelo sistema e recomendados através da base, apresentados na tabela 9 -, no entanto, se faz importante esclarecer que este pesquisador possui 45 artigos com possibilidade de recomendação. Assim, tem-se um valor de acurácia de 0.2 para suas recomendações. No entanto, vale ressaltar que dos 10 artigos recomendados pelo sistema, 5 estão entre os interesses do pesquisador de acordo com a base utilizada.

Tabela 10. Recomendações R*esearch* 10

Artigos Recomendados pelo Sistema	Artigos Recomendados na Base de Dados
P06-1017_recfv.txt	P04-1053.txt
P01-1005_recfv.txt	P04-1054.txt
P05-1001_recfv.txt	P04-1055.txt
P03-1025_recfv.txt	P05-1051.txt
P01-1015_recfv.txt	P05-1052.txt
P04-1060_recfv.txt	P05-1053.txt
P02-1065_recfv.txt	P05-1061.txt
P06-1103_recfv.txt	P06-1015.txt
P06-1079_recfv.txt	P06-1016.txt
P05-1066_recfv.txt	P06-1017.txt

Os dados dispostos na tabela 10 são referentes ao Pesquisador 10. Este pesquisador possui apenas 11 artigos, no total, marcados para seu interesse. O cálculo do NDCG apontou uma acurácia de 0.2 para as recomendações deste pesquisador.

Tabela 11. Recomendações R*esearch* 11

Artigos Recomendados pelo Sistema	Artigos Recomendados na Base de Dados
P06-1049_recfv.txt	P00-1038.txt
P06-1042_recfv.txt	P00-1039.txt
P05-1070_recfv.txt	P00-1040.txt
P02-1058_recfv.txt	P00-1041.txt
P04-1049_recfv.txt	P01-1014.txt
P02-1047_recfv.txt	P01-1016.txt
P05-1008_recfv.txt	P01-1036.txt
P00-1041_recfv.txt	P01-1049.txt
P03-1069_recfv.txt	P01-1064.txt
P00-1054_recfv.txt	P02-1012.txt

O Pesquisador 11 tem seus artigos - recomendados e com possibilidade de recomendação, de acordo com a base, - detalhados na tabela 11. Ao todo, são 33 artigos marcados para recomendação desse pesquisador. Dos 10 escolhidos pelo sistema, 4 estão entre os marcados para recomendação na base, assim tem-se um valor de acurácia das recomendações de 0.142.

Tabela 12. Recomendações R*esearch* 12

Artigos Recomendados pelo Sistema	Artigos Recomendados na Base de Dados
P00-1010_recfv.txt	P00-1010.txt
P05-1061_recfv.txt	P00-1065.txt
P06-1145_recfv.txt	P00-1066.txt
P06-1017_recfv.txt	P01-1039.txt
P04-1074_recfv.txt	P01-1042.txt
P06-1047_recfv.txt	P02-1029.txt
P06-1095_recfv.txt	P02-1032.txt
P00-1060_recfv.txt	P02-1041.txt
P04-1052_recfv.txt	P03-1002.txt
P06-1050_recfv.txt	P03-1028.txt

O Pesquisador 12, com dados disponíveis na tabela 12, atingiu um valor de acurácia de 0.352. Vale salientar que tal pesquisador possui 34 artigos na base com possibilidade de recomendação e entre os 10 recomendados, 7 foram classificados corretamente.

Tabela 13. Recomendações R*esearch* 13

Artigos Recomendados pelo Sistema	Artigos Recomendados na Base de Dados
P01-1040_recfv.txt	P00-1065_recfv.txt
P00-1054_recfv.txt	P02-1032_recfv.txt
P03-1068_recfv.txt	P04-1055_recfv.txt
P00-1075_recfv.txt	P05-1006_recfv.txt
P04-1033_recfv.txt	P05-1061_recfv.txt
P06-1069_recfv.txt	P05-1072_recfv.txt
P05-1073_recfv.txt	P05-1073_recfv.txt
P03-1066_recfv.txt	P06-1044_recfv.txt
P06-1058_recfv.txt	
P05-1053_recfv.txt	

O sistema atingiu um valor de acurácia de 0.079 em relação ao Pesquisador 13, com dados disponíveis na tabela 13. Esse pesquisador possui apenas 8 artigos com possibilidade de recomendação e o sistema teve um desempenho baixo em relação as recomendações para ele.

Tabela 14. Recomendações R*esearch* 14

Artigos Recomendados pelo Sistema	Artigos Recomendados na Base de Dados
P00-1074_recfv.txt	P00-1008.txt
P05-1061_recfv.txt	P00-1033.txt
P00-1054_recfv.txt	P00-1058.txt
P00-1077_recfv.txt	P02-1007.txt
P03-1061_recfv.txt	P02-1025.txt
P04-1066_recfv.txt	P02-1055.txt
P06-1049_recfv.txt	P02-1056.txt
P02-1047_recfv.txt	P03-1004.txt
P05-1053_recfv.txt	
P05-1062_recfv.txt	

O sistema não recomendou nenhum artigo que estava marcado na base para recomendação do Pesquisador 14, como é possível visualizar na tabela 14. Portanto o valor de acurácia foi de 0 (zero) para este pesquisador.

Tabela 15. Recomendações R*esearch* 15

Artigos Recomendados pelo Sistema	Artigos Recomendados na Base de Dados
P05-1001_recfv.txt	P00-1065.txt
P04-1033_recfv.txt	P00-1069.txt
P04-1001_recfv.txt	P01-1049.txt
P04-1080_recfv.txt	P03-1007.txt
P06-1017_recfv.txt	P03-1017.txt
P05-1049_recfv.txt	P03-1044.txt
P00-1063_recfv.txt	P03-1058.txt
P05-1008_recfv.txt	P04-1003.txt
P06-1046_recfv.txt	P04-1032.txt
P05-1005_recfv.txt	P04-1037.txt

Por fim, para o Pesquisador 15 o valor de acurácia foi de 0.2. Na tabela 15 estão dispostos os artigos referentes a este pesquisador, Vale salientar que haviam 23 artigos com possibilidade de recomendação para o Pesquisador 15.

A hipótese levantada acerca da utilização das distâncias euclidianas para cálculo de similaridade entre os *papers e*, assim, realizar recomendações mais precisas não atingiu bons resultados em todos os casos, como é possível visualizar nas tabelas 2, 4, 6 e 14, onde as recomendações tiveram 0 (zero) como valor de acurácia.

Os melhores desempenhos foram encontrados ao considerar as recomendações dos Pesquisadores 7, 8, 9 e 12. Dos 10 artigos, 7 foram recomendados de forma correta, considerando o contexto dos Pesquisadores 7 e 12; 6 foram recomendados corretamente considerando o Pesquisador 8 e; 5 foram recomendados corretamente considerando o Pesquisador 12.

Com a análise das recomendações, em especial dos pesquisadores 7 e 12, tornou-se possível perceber que o cálculo do NDCG retornava um valor abaixo do esperado por conta da comparação que era realizada com todos os artigos presentes na base marcados para recomendação, por exemplo, para o Pesquisador 7, tinha-se os 10 artigos recomendados pelo sistema comparados aos 47 presentes em sua base com possibilidade de recomendação. Dessa forma, mesmo que os 10 artigos fossem classificados corretamente, o NDCG não retornaria um valor maior que 0.4.

Assim, mais um experimento foi realizado sem eleger um número específico de artigos para recomendar, mas considerando a quantidade de *papers* marcados na base de dados para recomendação. Os valores de DCG normalizada encontrados nessa segunda etapa do experimento estão dispostos na tabela 16.

Tabela 16. NDCG relacionado a cada Pesquisador

Pesquisador (P)	NDCG com top10	NDCG (experimento II)
P1	0.094	0.094
P2	0.0	0.0
P3	0.1	0.171
P4	0.0	0.029
P5	0.128	0.128
P6	0.0	0.069
P7	0.289	0.353
P8	0.418	0.547
P9	0.198	0.382
P10	0.197	0.248
P11	0.142	0.294
P12	0.352	0.418
P13	0.079	0.079
P14	0.0	0.0219
P15	0.121	0.278

Ao igualar o número de artigos recomendados ao número de artigos com possibilidade de recomendação presentes na base, foi possível constatar um aumento nos valores de NDCG. Na recomendação baseada em um top10, 4 pesquisadores tiverem 0 (zero) como valor do NDCG, com a mudança na quantidade de artigos à serem recomendados, apenas o Pesquisador 2 manteve o valor 0(zero).

Os pesquisadores P1, P5 e P13 mantiveram o valor de acurácia em ambos os testes. Com exceção de P1, P2, P5 e P13, todos os pesquisadores passaram a ter recomendações mais precisas ao passo que a quantidade de artigos recomendados foi igualada a quantidade de artigos com possibilidade de recomendação.

Após a realização dos experimentos, com a similaridade dos artigos verificada por meio dos pesos retornados pelas Distâncias Euclidianas, a média aritmética[13] do NDCG de todos os pesquisadores foi considerada, para, assim, comparar com as recomendações realizadas por meio de similaridade verificada a partir do *Similarity Cosine*, utilizado por Sugiyama e Kan (2010). Assim, tem-se o sistema proposto no presente trabalho atingindo um valor médio de acurácia nas recomendações no valor de 0.208, considerando a média

[13] Calcula-se a média aritmética determinando-se a soma dos valores do conjunto e dividindo-se esta soma pelo número de valores no conjunto (Stevenson, 1986).

aritmética, enquanto o sistema proposto por Sugiyama e Kan (2010) atinge uma acurácia no valor de 0.459, considerando seu experimento que realiza um top10.

Para avaliação/interpretação dos resultados estatísticos obtidos através da média aritmética, Garcia (1989) recomenda a exploração de todas as informações disponíveis para o pesquisador, por meio das técnicas de desvio padrão (DV) e coeficiente de variação (CV). Para o experimento proposto em Sugiyama e Kan (2010) não é possível a utilização destas técnicas, pois os dados referentes a cada pesquisador não estão disponíveis no artigo publicado. Dessa forma, as Equações 8 e 9 apresentam o cálculo de desvio padrão e coeficiente de variação respectivamente, calculados em relação aos dados obtidos na presente pesquisa.

$$\mathrm{DV} = \frac{\sum(NDCG_{Pi} - média\_aritimética)}{quantidade\ de\ pesquisadores} \qquad (8)$$

$$\mathrm{CV} = \frac{DV}{média\_aritmética} * 100 \qquad (9)$$

Na Equação 8 p representa cada pesquisador, com i variando entre 0 (zero) e a quantidade total de pesquisadores. Assim, o DV e o CV foram calculados considerando os dados obtidos por meio do sistema que considera as distâncias euclidianas para o cálculo de similaridade. Obteve-se um valor de DV igual a 0.162 e um coeficiente de variação de 77%. De acordo com Garcia (1989) valores de CV acima de 30% são classificados como muito altos, assim, é possível perceber que os valores de acurácia das recomendações possuem característica heterogenia. É possível verificar isso por meio da comparação entre os dois pesquisadores que estão nos extremos, sendo o *Pesquisador 2* o que obteve o menor valor de acurácia, 0 (zero), e o *Pesquisador 8* o que obteve o maior valor de acurácia das recomendações, sendo 0.547. A heterogeneidade do valor de acurácia das recomendações se justifica pela quantidade de dados disponíveis para criação do perfil de cada pesquisador, ou seja, quanto mais dados disponíveis melhores os resultados das recomendações.

Por fim, é possível perceber que a similaridade do cosseno atinge um desempenho melhor para o problema de recomendação de artigos científicos para pesquisadores com o perfil dos pesquisadores juniores. Mesmo tendo um desempenho melhor, comparado à média aritmética dos valores de acurácia das recomendações realizadas utilizando as distâncias euclidianas, não é possível verificar se há homogeneidade nas recomendações realizadas com o *cosine*

similarity ou não, tendo em vista que não estão presentes no artigo de Sugiyama e Kan (2010) os dados referentes a cada pesquisador de forma individual, para o cálculo de DV e CV.

4.2 CONCLUSÕES

Com o crescente número de artigos, teses, livros, dentre outras produções científicas disponibilizadas na *Web*, está se tornando cada vez mais difícil aos pesquisadores encontrar produções relevantes às suas áreas de estudo.

Como alternativa ao problema de *overloading* de conteúdos acadêmicos e com o intuito de tornar mais rápida e precisa a busca por informações relevantes às pesquisas dos acadêmicos, o presente trabalho destinou-se a implementação de um sistema que recomenda artigos científicos considerando o perfil dos pesquisadores.

O sistema proposto foi baseado no trabalho realizado por Sugyiama e Kan (2010) que considera a técnica *Similarity Cosine* para o cálculo do peso de similaridade para realização das recomendações, no entanto os autores apontaram dificuldades para adquirir conhecimento sobre os pesquisadores juniores. Assim, o sistema aqui apresentado considera as Distâncias Euclidianas para cálculo dos pesos, pois essa técnica é apontada por Huang (2008) como uma das principais para este tipo de problema. O perfil dos pesquisadores juniores foi montado de acordo com seu artigo e as referências que eles utilizaram para construção do *paper*. Técnicas de mineração de texto foram utilizadas para realização das recomendações e montagem do perfil dos pesquisadores.

Como forma de avaliar o desempenho do sistema, foi utilizada a métrica *Normalized Discounted Cumulative Gain* que consegue mensurar a eficácia do sistema e é vastamente utilizada em problemas de recomendação.

As informações acerca dos pesquisadores foram coletadas através da base de dados *Nus Computing*, da *National University of Singapore*. A base conta com os artigos dos pesquisadores, as referências que eles utilizaram na construção dos artigos e produções que citam tais pesquisadores.

Dois experimentos foram realizados para análise das recomendações. O primeiro experimento teve por objetivo apontar os 10 *papers* com maior similaridade em relação ao perfil do pesquisador. Tendo em vista que a base de dados disponibiliza os artigos com maior possibilidade de recomendação para cada pesquisador, o segundo experimento teve por objetivo recomendar a mesma quantidade de artigos que foi apontada pela base de dados com probabilidade de recomendação. Ao final da pesquisa, a média aritmética da acurácia das recomendações de ambos os testes foram comparadas. Assim, ficou constatado que o segundo experimento teve uma maior acurácia nas recomendações, sendo um valor de 0.208. Vale ressaltar que o cálculo do NDCG, utilizado para verificar a acurácia das recomendações, varia entre [0, 1].

Além da média aritmética das recomendações, vale ressaltar dois casos em específico, um que obteve a maior acurácia e outro que obteve o menor valor de acurácia nas recomendações. Assim, tem-se o Pesquisado 2 que em ambos os experimentos obteve um valor de NDCG igual a 0(zero), pois os artigos recomendados pelo sistema não faziam parte dos que estavam com probabilidade de recomendação, de acordo com a base. Atribui-se esse resultado do Pesquisador 2 a pouca quantidade de informações que o sistema tinha sobre ele, tendo em vista que tal pesquisador possui apenas um artigo e esse artigo conta com somente 4 referências. E, no caso de maior valor de NDCG, o Pesquisador 8 que atingiu um valor de 0.547 de acurácia nas recomendações. Para o Pesquisador 8 a base conta com um artigo publicado por ele e 23 artigos de referência, o que justifica a criação de um perfil mais fidedigno as preferências do pesquisador.

Mesmo com bons resultados nas recomendações de alguns pesquisadores, de modo geral, o peso para o cálculo de similaridade realizado juntamente com as Distâncias Euclidianas não permitiram bons valores de acurácia das recomendações, de modo geral.

Ao serem comparados os valores médios (média aritmética) alcançados pelo sistema proposto no presente trabalho e os valores alcançados pelo sistema proposto por Sugiyama e Kan (2010), tem-se um valor de acurácia maior atingido pelo segundo sistema, com NDCG de 0.459. Ambos os valores (0.208 e 0.459) podem ser considerados baixos, dentro do universo entre [0, 1].

Para análise dos valores encontrados pelas médias aritméticas, o cálculo do desvio padrão e do coeficiente de variação foi realizado em relação as recomendações que consideram as distâncias euclidianas. Assim, alcançou-se um valor de DV de 0.162 e um valor de CV de 77%. Esses valores mostram que os resultados das recomendações não foram homogêneos, o que já era esperado, tendo em vista que quanto menos informações o pesquisador possuir, menor será a acurácia das recomendações, enquanto que quanto mais informações o pesquisador possuir, maior será a acurácia das recomendações, pois será possível a montagem de um perfil mais fidedigno aos interesses do pesquisador. Vale ressaltar que não foi possível realizar os cálculos de DV e CV em relação ao sistema proposto por Sugiyama e Kan (2010), tendo em vista que os mesmos não disponibilizam em seu artigo os dados de cada pesquisador de forma individual.

Assim, como trabalhos futuros projeta-se a utilização de técnicas não lineares para o cálculo de similaridade, tendo em vista que as técnicas utilizadas no presente trabalho, mesmo com bons desempenhos em alguns casos, atingiram um desempenho baixo. Projeta-se ainda continuar o estudo com recomendações para os pesquisadores seniores, utilizando as novas técnicas, tendo em vista que este trabalho priorizou a recomendação para pesquisadores juniores, pois estes possuem menos informações tornando-se, assim, mais difícil de se montar um perfil para a tarefa de recomendação.

REFERÊNCIAS

AGUIAR, J. J. B.; FECHINE, J. M.; COSTA, Evandro. B. **Recomendação de Objetos de Aprendizagem baseada na Popularidade dos Objetos e nos Estilos de Aprendizagem dos Alunos**. Disponível em <http://www.br-ie.org/pub/index.php/sbie/article/view/ 5438/3797> Acesso em abril, 2016.

BLATTMANN, Ursula; BOMFÁ, Cláudia Regina Ziliotto. **Gestão de conteúdos em bibliotecas digitais: acesso aberto de periódicos científicos eletrônicos**. Revista Brasileira de Biblioteconomia e Documentação, v. 2, p. 41-56, 2006.

CARVALHO, V. C. de; DORÇA, F. A.; CATELLAN, R. G.; ARAÚJO, R. D. **Uma Abordagem para Recomendação Automática Dinâmica de Objetos de Aprendizagem Baseada em Estilos de Aprendizagem**. Disponível em <http://www.br-ie.org/pub/index.php/sbie/article/view/3065/2573> Acesso em abril, 2016.

CAZELLA, S. C.; REATEGUI, Eliseo.; MACHADO, M.; BARBOSA, J. 2009. **Recomendação de Objetos de Aprendizagem Empregando Filtragem Colaborativa e Competências**. In: Simpósio Brasileiro de Informática na Educação (SBIE), Florianópolis.

CAZELLA, S. C.; Nunes, M. A.; REATEGUI, Eliseo. **A Ciência da Opinião: Estado da arte em Sistemas de Recomendação.** In: André Ponce de Leon F. de Carvalho; Tomasz Kowaltowski. (Org.).. (Org.). JAI: Jornada de Atualização em Informática da SBC. Rio de Janeiro: Editora da PUC Rio, 2010, v., p. 161-216.

CAZELLA, S. C.; BEHAR, P.; SCHNEIDER, D.; SILVA, K. K. A.; FREITAS, R. **Desenvolvendo um Sistema de Recomendação de Objetos de Aprendizagem baseado em Competências para a Educação: relato de experiências**. In: Simpósio Brasileiro de Informática na Educação, 2012, Rio de Janeiro. Anais do Simpósio Brasileiro de Informática na Educação, 2012.

CHEN, H. (2003). **An Intelligent Broker for Context-Aware Systems**. Disponível em

<https://www.researchgate.net/publication/2567594_An_Intelligent_Broker_Architecture_for_Context-Aware_Systems > Acesso em maio, 2016.

COSTA, Evandro; AGUIAR, Janderson; MAGALHAES, J. J. **Sistemas de Recomendação de Recursos Educacionais: conceitos, técnicas e aplicações**. In: Melo, A. M.; Borges, M. A. F.; Silva, C. G. (Org.). Anais da II Jornada de Atualização em Informática na Educação (JAIE 2013). 1ed.: Sociedade Brasileira de Computação (SBC), 2013, v. , p. 57-78.

GARCIA, C. H. **Tabelas para classificação do coeficiente de variação**. In: Instituto de Pesquisas Florestais (IPEF), 1989.

HUANG, A. (2008). **Similarity measures for text document clustering**. In New Zealand Computer Science Research Student Conference, pp. 49–56.

LAKATOS, Maria Eva. MARCONE, Marina de Andrade. Fundamentos de metodologia científica, 5. ed., São Paulo: Atlas 2003.

LUHN, H. P. (1957). **A statistical approach to mechanized encoding and searching of literaryinformation**. IBM Journal of Research and Development, 1 (4), 309–317.

MATOS, P. F.; LOMBARDI, L. O.; CIFERRI, R. R.; PARDO, T. A. S. CIFERRI, C. D. A.; VIEIRA, M. T. P. **Relatório Técnico "Métricas de Avaliação"**. Disponível em <http://www.icmc.usp.br/~taspardo/TechReportUFSCar2009a-MatosEtAl.pdf> Acesso em jun., 2016.

REATEGUI, Eliseo.; CAZELLA, Sílvio César. Mini-Curso: **Sistemas de Recomendação**. In: IV Encontro Nacional de Inteligência Artificial, 2005, São Leopoldo, RS, Brasil. Anais do ENIA. Porto Alegre: SBC, 2005. p. 306-348.

RESNICK, P.; VARIAN, H. R. 1997. **Recommender Systems**. Communications of the ACM, New York, v.40, n.3, pp. 55-58, Mar.

SAMPAIO, I. A. 2006. **Aprendizagem Ativa em Sistemas de Filtragem Colaborativa**. Dissertação de Mestrado, Universidade Federal de Pernambuco.

SCHAFER, J. B.; KONSTAN, J.; RIEDL, J. **Recommender Systems in E-Commerce**. Eletronic Commerce: Proceeding of the 1st ACM conference on Electronic commerce. p. 158-166, 199. Disponível em

<http://dl.acm.org/citation.cfm?id=336992 &picked=prox> Acesso em maio, 2016.

Sidorov , Grigori; Gelbukh , Alexander; Gómez-Adorno, Helena; Pinto, David. **Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model**. Computación y Sistemas 18(3):491–504.

SILVA Filho, W. D. da.; CAZELLA, S. C. **STAR: um Framework para recomendação de artigos científicos baseado na relevância da opinião dos usuários e em filtragem colaborativa**. Disponível em < http://www.lbd.dcc.ufmg.br/colecoes/ enia/2005/042.pdf > Acesso em maio, 2016.

STEVENSON, W. J. (1981). **Estatística Aplicada à Administração**; tradução Alfredo de Farias. Harper & Raw do Brasil, São Paulo, SP, Brasil.

SUGIYAMA, Kazunari.; KAN, Min-Yen. 2010. **Scholarly Paper Recommendation via User's Recent Research Interests**. Disponível em <http://dl.acm.org/citation.cfm?id= 1816129> Acesso em maio, 2016

SUGIYAMA, Kazunari.; KAN, Min-Yen. 2013. **Exploiting Potential Citation Papers in Scholarly Paper Recommendation**. Disponível em <http://dl.acm.org/citation.cfm ?id=2467701> Acesso em maio, 2016.

WANG, Y.; WANG, L.; LI, Y.; HE, D.; CHEN, W.; LIU, Tie-Yan. **A Theoretical Analysis of Normalized Discounted Cumulative Gain (NDCG) Ranking Measures**. In: Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013).

WILEY, D. A. **Connecting learning objects to instructional design theory: a definition, a metaphor, and a taxonomy**. In: WILEY, D. A. (Ed.). The instructional use of learning objects: Online Version. 2002. Disponível em: <http://reusability.org/read> Acesso em maio, 2016.