Sistema de Recomendação de Trabalhos Cientificos e Autores

João Pedro R. D. Saldanha[1], **Fernando Prass**[1]

[1]Ciência da Computação – Universidade Franciscana (UFN)
Rua dos Andradas, 1614 – 97010-032 – Santa Maria – RS – Brasil

{joao.pedro,fernando.prass}@ufn.edu.br

Abstract. *This meta-paper describes the style to be used in articles and short papers for SBC conferences. For papers in English, you should add just an abstract while for the papers in Portuguese, we also ask for an abstract in Portuguese ("resumo"). In both cases, abstracts should not have more than 10 lines and must be in the first page of the paper.*

Resumo. *Este meta-artigo descreve o estilo a ser usado na confecção de artigos e resumos de artigos para publicação nos anais das conferências organizadas pela SBC. É solicitada a escrita de resumo e abstract apenas para os artigos escritos em português. Artigos em inglês deverão apresentar apenas abstract. Nos dois casos, o autor deve tomar cuidado para que o resumo (e o abstract) não ultrapassem 10 linhas cada, sendo que ambos devem estar na primeira página do artigo.*

1. Introdução

No método científico, pesquisadores devem realizar um trabalho criativo sistemático para incrementar o conhecimento na área da pesquisa. Parte fundamental do processo é a busca, dentro do universo dos trabalhos científicos, por embasamento teórico ao tema do trabalho proposto. Além disto, também é importante conhecer e colaborar com pesquisadores desenvolvendo trabalhos relacionados dentro da área de pesquisa. Portanto, é preciso reunir todas as informações pertinentes, pesquisas e resultados anteriores bem como linhas de pesquisa em progresso para não reinventar a roda ou seguir caminhos já trilhados e assim realizar trabalho relevante e produtivo.

A plataforma Lattes é um sistema que integra bases de dados de currículos, em específico de pesquisadores. Ela oferece aos usuários a possibilidade de criar um currículo de maneira gratuita, que é disponibilizado abertamente aos visitantes do site. Segundo [CNPq 2019], o Currículo Lattes *"se tornou um padrão nacional no registro da vida pregressa e atual dos estudantes e pesquisadores do país, e é hoje adotado pela maioria das instituições de fomento, universidades e institutos de pesquisa do País"*.

O universo da pesquisa científica está em constante expansão, tanto no que diz respeito ao conhecimento produzido quanto ao volume de trabalhos e publicações. Estimativas apontavam um valor em torno de 2.5 milhões de artigos científicos publicados por ano, em 2015, com um aumento de 5% ao ano no número de cientistas fazendo publicações [Ware and Mabe 2015]. Pesquisadores não têm o tempo necessário para analisar todos os estudos relacionados à seus próprios trabalhos, mesmo com plataformas como o Lattes, onde estão tais trabalhos estão compilados. Trata-se do problema da sobrecarga de informação, que tem crescido na medida em que sistemas digitais vem

ganhando cada vez usuários e conteúdo. Outro problema decorrente do crescimento do número de pesquisadores e trabalhos é que muitas vezes os pesquisadores não conhecem outros pesquisadores da área e acabam por perder a oportunidade de colaborações ou troca de idéias.

Logo, se faz necessário a filtragem da informação que chega ao pesquisador para maximizar sua eficiência e evitar tempo perdido. A automatização de tal tarefa de filtragem pode ser feita através de sistemas de recomendação, utilizando técnicas de mineração de dados e inteligência artificial para oferecer o conteúdo mais relevante disponível, aumentando a eficiência do acadêmico.

A partir do problema da sobrecarga de informações, nos anos 90 iniciou-se a pesquisa na área de filtragem de conteúdo. O ponto de partida foi a observação que as pessoas usam, no dia-a-dia, dicas de outros para tomar decisões, sendo que as dicas daqueles tidos como especialistas no assunto tem um peso diferenciado. Os primeiros sistemas de recomendação eram algoritmos capazes de analizar tendências dentro de uma certa comunidade e então fazer sugestões aos seus membros. Este método é conhecido como filtragem colaborativa e foi aprimorado desde então, sendo até hoje bastante popular. Além deste, também é bastante difundido o método baseado em conteúdo, no qual novos itens são recomendados baseado no conteúdo consumido pelo usuário no passado [Ricci et al. 2011].

2. Objetivos

Neste trabalho é proposta a elaboração de um sistema de recomendações de pesquisadores e trabalhos científicos baseado em confiança a partir de dados da plataforma lattes. A ideia é analisar o sentimento da comunidade em relação a cada pesquisador e sugerir conteúdo relevante aos pesquisadores baseado em seus perfis, ponderando o conteúdo com base na confiança estimada da comunidade. Para tal, é preciso:

- Estudar funções e aplicações de sistemas de recomendação
- Escolher uma técnica para ser aplicada
- Planejar a Implementação do Sistema de Recomendações
- Implementar o Sistema De Recomendações
- Avaliar o modelo

3. Revisão Bibliográfica

Nesta seção, será feita a apresentação dos principais itens que compõem o embasamento teórico usado como ponto de partida no trabalho e usado para justificar as decisões tomadas.

3.1. Sistemas de Recomendação

Frequentemente usuários de plataformas digitais se deparam com situações nas quais é necessário escolher entre vários itens ofertados: Produtos, conteúdo ou pessoas. A dificuldade de filtrar o conteúdo encontrado em determinada plataforma tende a aumentar na medida em que o número de itens ofertados cresce, visto que é necessário fazer uma análise individual de tais itens e então compará-los para fazer uma escolha. Para ajudar na tarefa, é comum encontrar sistemas que automatizam o processo de escolha, filtrando o conteúdo com base no perfil do usuário para apresentar seu interesse. Tais sistemas

são especificamente úteis quando um usuário encontra dificuldades para analisar os itens ofertados e fazer escolhas. Os itens recomendados pelos SR podem ser os mais variados, sendo que no geral a recomendação é uma tarefa especializada, ou seja, apenas um tipo de item é recomendado, e a recomendação é relevante para um perfil específico de usuário. Logo, as características do SR, como metodologia usada para sua construção, interface de usuário e critério para ordenar os resultados devem ser adaptados às especificidades da tarefa em questão.

A forma mais simples do resultado de um SR é uma lista de itens ordenada de acordo com a preferência do usuário. A satisfação com as recomendações pode ser coletada explicitamente, como por exemplo através de avaliações, ou implicitamente através de inferências baseadas no comportamento do usuário perante aos itens oferecidos. Para oferecer recomendações, é preciso analisar uma base de conhecimento, que pode ter diversas informações, realizar um trabalho de classificação dos itens ofertados e então coletar algum tipo de feedback perante o resultado que deve ser usado para aprimorar o sistema [Ricci et al. 2011].

3.1.1. Técnicas de Recomendação

O resultado obtido por um SR é dependente da realização de uma **predição**. A predição é fundamental para a qualidade das recomendações: Itens são apresentados ao usuário porque o sistema antecipa que sejam relevantes para ele [Ricci et al. 2011]. Geralmente na elaboração de sistemas de recomendação lidamos com **usuários**, denotados por $u_1, ...u_n \in U$, **itens**, denotados por $i_1, ...i_n \in I$ e **relações**, que associam usuários e itens de diversas maneiras [Ekstrand and Konstan 2019]. As associações podem ser representadas por ontologias [Primo and Loh 2006] ou no caso de relações entre usuários e itens através de uma matriz de associação $|U| \times |I|$. Assume-se a existência no mundo real de uma **função** $f(u, i)$ que retorne um número real representado a utilidade do item i ao usuário u. Em técnicas de filtragem colaborativa, este numero é visto como a avaliação do usuário. A tarefa do SR neste contexto é computar uma função $\hat{f}(u, i)$ que se assemelhe ao máximo à f. Assim, é possível realizar a predição de relevância de um grupo de itens para determinado usuário $\hat{f}(u_n, I)$ e recomendar os itens melhores classificados pelo SR, efetivamente filtrando o conteúdo e oferecendo ao usuário uma seleção personalizada de itens [Ricci et al. 2011].

3.1.2. Filtragem Colaborativa

Técnicas de filtragem colaborativa analisam o **perfil** do usuário e sua **avaliação** dos itens previamente acessados para chegar em recomendações. Procura-se analisar o perfil do **usuário alvo** para então achar um *cluster* de usuários com perfis similares (**vizinhos**). A idéia é que os itens bem avaliados pelos vizinhos serão também avaliados positivamente pelo usuário alvo, já que os perfis são semelhantes. Um problema encontrado na técnica é o problema da primeira avaliação: Quando temos um item novo, sem nenhuma avaliação, como saber se determinado usuário irá avaliar positivamente o item? Nenhum de seus vizinhos fez avaliações [Ricci et al. 2011].

SRs baseados em filtragem colaborativa são os mais populares na área e vêm

sido pesquisados há mais tempo. [Ricci et al. 2011] É comum utilizar métodos baseados em vizinhança, nos quais um algoritmo de clusterização é usado para determinar grupos de usuários ou ítens, tal como o algoritmo **KNN** (*K-Nearest Neighbours*) [da Rosa Furlan et al. 2018].

3.1.3. Método Baseado em Conteúdo

O método baseado em conteúdo parte da idéia de que usuários têm interesse em itens semelhantes àqueles que lhe foram ùteis no passado [Ricci et al. 2011]. No caso, é importante determinar a **semelhança entre itens** para então recomendar para determinado usuário itens semelhantes aos que foram **previamente bem avaliados** por ele. Nesse método, é preciso estabelecer estratégias para descrever itens bem como para montar o perfil dos usuários, descrevendo os tipos de itens que ele tem interesse em. Então, deve ser feito o **comparativo** dos itens com o perfil do usuário para predizer seu interesse em tais itens. Geralmente procura-se dividir o universo dos itens, I, em categorias: Relevantes ou irrelevantes ao usuário, por exemplo. Para construir a classificação dos itens, é possível usar uma série de algoritmos que realizam trabalho de **classificação estatística**, como por exemplo **árvores de decisão** [Pazzani and Billsus 2007].

3.1.4. Método Baseado em Confiança

Estudos indicam que os usuários têm a tendência de **valorizar mais as recomendações de amigos** do que aquelas feitas por outros usuários com perfil semelhante, porém desconhecidos. A qualidade das recomendações de amigos superam inclusive as feitas por sistemas de recomendação [Sinha and Swearingen 2001]. A partir deste conceito, com a grande aderência de usuários à **redes sociais** um novo método para a construção de sistemas de recomendação está sendo estudado. Trata-se do método baseado em confiança, ou sistema de recomendação social (*social recommender system*) [Ricci et al. 2011].

A construção de SRs sociais depende do estabelecimento de uma **rede de confiança**, ou seja, uma rede social que descreve o nível de confiança entre seus membros. Assim, o usuário recebe recomendações de itens avaliados positivamente por usuários em sua rede de confiança. Estes SRs precisam usar o conceito de **agregação e dissipação de confiança**, ou seja, dados um grupo de usuários $u_1 \dots u_n$, calcular o nível de confiança entre u_1 e u_n considerando usuários intermediários $u_2 \dots u_{n-1}$ que possuem alguma relação de confiança, direta ou indireta, com u_1 e u_n (**dissipação**) ou combinar uma série de estimativas de confiança em um valor final (**agregação**) [Victor et al. 2011].

Um ponto fraco de tais sistemas é mais previsível e pode facilmente ser inundada por itens que o usuário já conhece, enquanto técnicas mais usuais de recomendação podem apresentar resultados mais inesperados, mas relevantes ao usuário [Sinha and Swearingen 2001].

3.1.5. Métodos Hibridos

3.2. Trabalhos Correlatos

Os trabalhos correlatos foram escolhidos utilizando como critério a contemporaneidade e semelhança com o presente trabalho, de forma a trazer um embasamento atualizado das metodologias usadas para a resolução de problemas semelhantes.

3.2.1. Desenvolvimento de um Sistema de Recomendação para Bibliotecas Digitais

Este trabalho também aborda o problema da sobrecarga de informações dos pesquisadores baseando-se no perfil do currículo lattes. O trabalho busca recomendações de produções científicas utilizando o motor de buscas google acadêmico e traz uma combinação das técnicas de filtragem colaborativa e baseado em conhecimento. A metodologia para gerar recomendações utilizada neste trabalho será usada no presente trabalho como referência para a elaboração do SR, levando em consideração os pontos fracos e fortes da abordagem descrita no trabalho. Em particular, será considerada a maneira com que o trabalho propôs solucionar o problema da avaliação inicial de um SR de filtragem colaborativa através do método baseado em conteúdo [da Rosa Furlan et al. 2018].

3.2.2. Técnicas de Recomendação para usuários de Bibliotecas Digitais

Trabalho que apresenta algumas das mais populares técnicas de recomendação, bem como a justificativa e contexto para a correta implementação dos mesmos. O trabalho descreve diversas abordagens para a elaboração de um SR de obras literárias em bibliotecas digitais, usando as técnicas de filtragem colaborativa e baseado em conteúdo bem como uma abordagem híbrida. O contexto do sistema de recomendação descrito no trabalho se assemelha ao do presente trabalho por ter como alvo uma biblioteca digital, sendo que as obras literárias da biblioteca podem ser comparadas aos artigos encontrados na plataforma lattes. O comparativo das metodologias usadas serve como referência para a elaboração do SR descrito no presente trabalho. As relações entre usuário e items (no caso, obras literárias) é descrita através de uma ontologia na qual conceitos são definidos pelos termos que os definem e organizados em uma hierarquia. A ontologia serve para descrever as relações entre item e usuário e serve como referência para a modelagem das relações do SR desenvolvido neste trabalho. O trabalho também apresenta um experimento para ilustrar a importância da opinião de especialistas [Primo and Loh 2006].

3.2.3. The PageRank Citation Ranking: Bringing Order to the Web

[Page et al. 1999]

3.2.4. Trust-aware Collaborative Filtering for Recommender Systems

4. Materiais & Métodos

Para chegar nas recomendações, é proposto um trabalho em dois momentos: Estimar a confiança entre pesquisadores e então selecionar publicações baseando-se no perfil dos pesquisadores. A seleção será ordenada de acordo com o nìvel de confiança estimado das publicações. O primeiro passo é modelar uma rede de confiança da comunidade científica, descrita por autores e publicações. Pode então ser feita uma seleção das publicações cadastradas com base no perfil do usuário e ordenar a seleção de acordo com o nível de confiança de cada item. A confiança pode ser local ou global, sendo que local diz respeito à confiança estimada de um pesquisador específico em seus colegas e a global corresponde à a confiança da comunidade em cada pesquisador. Em termos de ferramentas para implementação, foi escolhida a linguagem python, que possui riqueza de ferramentas e recursos para trabalhos relacionados a manipulação de dados e computação numérica.

4.1. Os Dados da Plataforma Lattes

Os dados usados no trabalho são provenientes de um banco de dados relacional extraído do currículo lattes dos pesquisadores com o uso de *web crawling*. A partir dos arquivos *XML* exportados da plataforma, o *crawler* faz uma análise a partir da qual são montadas relações entre os itens presentes no banco e os novos itens, relacionando autores às suas publicações cadastradas por terceiros, ou seja, inserindo uma linha em uma tabela associativa entre autor e publicação. A partir do banco descrito, é possível construir uma base de conhecimento focada especificamente no objetivo do presente trabalho, que é descrita a seguir.

4.1.1. Publicações

Uma publicação científica é uma produção de alguma **natureza** (artigo, livro, trabalho) que passou por um processo de revisão por pares e foi aceito como sendo uma contribuição válida, de autoria de um ou mais pesquisadores. A maioria das publicações da base de conhecimento possuem alguma **palavra-chave**, informada pelos autores, que servem como uma pista dos assuntos abordados. As publicações são feitas originalmente em um **idioma** e **país** específicos e obrigatoriamente possuem um **título** no idioma original bem como uma possível tradução para o inglês. O **ano** da publicação é considerado relevante devido à característica progressiva do conhecimento científico, onde observa-se constante introdução de novos dados e fatos. Em alguns casos é possível inferir a **abrangência** da publicação, por exemplo pode ser regional, nacional ou internacional. Algumas tuplas podem também conter a **natureza**, que deve ser vista como o nível de cobertura do assunto discutido alcançado pelo trabalho: Completo, resumo e assim por diante. A ordem dos nomes dos autores geralmente é indicativo da importância do autor para a publicação: O **autor principal** geralmente é o primeiro nome, e o **orientador** do trabalho o último. Entre eles, os colaboradores **intermediários**.

4.1.2. Pesquisadores

Os autores são o centro dos dados da base de conhecimento e são também responsáveis por cadastrar todas as informações lá encontradas. Cada pesquisador possui um **texto descritivo** que pode ser redigido manualmente ou gerado automaticamente pela plataforma lattes. O perfil do pesquisador é composto por sua **formação**, **atuação profissional**, e **publicações** das quais fez parte. A formação é representada por um título, que diz respeito ao nível de ensino, por exemplo doutorado ou mestrado.

4.2. Computando confiança

Pode-se pensar na rede de confiança como sendo um grafo no qual os nodos são pesquisadores e as arestas publicações em conjunto. Considerando que uma colaboração em publicações é um voto de confiança entre os pesquisadores envolvidos, a matriz de adjacência pode ser usada para computar a propagação de confiança através da rede. A confiança estimada de determinada publicação é a soma da confiança estimada de cada pesquisador envolvido no trabalho.

O algoritmo PageRank foi inspirado em parte por estudos realizados em redes de citações acadêmicas, nas quais a relevância de um artigo era descrita por contagem de citações, por exemplo [Page et al. 1999]. Trata-se de um método para computar um ranking global de citações, pensado para computar a importância das paginas web. O ranking R de uma página é definido como a soma dos rankings das páginas que oferecem links para ela, ponderada pelo total de links encontrados nas páginas [Page et al. 1999].

É definido para cada página u um conjunto F_u de páginas as quais u referencia e um conjunto B_u de páginas que fazem referência à u. Sendo $\hat{A}$ a matriz de adjacência da web, tal que

$$\hat{A}_{i,j} = \begin{cases} 1 & \text{se há links de i para j} \\ 0 & \text{se não há links de i para j} \end{cases}$$

A matriz A deve ser obtida dividindo todas as linhas de $\hat{A}$ por $|F_u|$ (o grau do nodo u). Assim, PageRank pode ser definido como $R = c(AR + E)$, sendo c um fator de normalização.

Quando ocorrem ciclos no fluxo de referência, nos quais duas páginas se referenciam mutuamente e não fazem referência a nenhuma outra página, pode ocorrer o chamado *rank sink* quando há referencias exteriores injetando ranking no ciclo, assim as páginas do ciclo acumulam ranking, porém não há distribuição. Para solucionar, foi introduzido o vetor E, que no modelo de PageRank é o conceito de um *random surfer*, ou seja, uma probabilidade de um usuário da internet aleatoriamente mudar a página, sem seguir nenhum de seus links. [Page et al. 1999]

A aplicação de PageRank à um grafo não-direcionado gera um vetor R estatisticamente similar à distribuição de grau dos nodos da rede [Perra and Fortunato 2008]. Isto é, aplicando diretamente o algoritmo ao problema proposto, no final das contas a confiança seria proporcional ao número de publicações em conjunto (**centralidade** do nodo). Enquanto esta métrica é relevante, perde-se a idéia inicial: Não é considerada a confiança

dos colaboradores, somente o valor total de colaborações. Além disso, alguns conceitos importantes não são levados em consideração: A relevância da publicação e a expertise dos acadêmicos. No caso da **relevância**, a importância da publicação é uma dica para o nível de confiança mútua entre os pesquisadores. Já o nível de **expertise** do pesquisador é indicativo da qualidade de seu julgamento.

4.2.1. Relevância

Nem todas as publicações são feitas iguais. É possível considerar as características descritas na representação de publicação usada na base de conhecimento para construir uma heurística sobre a relevância da publicação. Um valor é atribuído a cada uma das características relevantes da publicação, quando presentes e a heurística é a soma das pontuações. A pontuação é considerada conforme a tabela.

**** tabela ****

4.2.2. Expertise

Expertise é o conceito de quanto conhecimento determinado pesquisador possui em sua área de atuação. Para estabelecer uma heurística de expertise, é usado um modelo semelhante ao usado para estimar a relevância de um artigo: Para cada ponto considerado relevante à expertise do autor, é atribuído um valor e então a heurística é o somatório dos valores dos campos que possuem dados.

**** tabela ****

4.2.3. Centralidade

Considerando as heurísticas definidas, é possível alcançar um fluxo de confiança mais interessante na rede modelada. O conceito de agregação e dissipação de confiança pode seguir a idéia do algoritmo *PageRank* aplicando-se um algoritmo para o cálculo da centralidade dos nodos. Assim, a relevância das publicações e a expertise dos pesquisadores podem ser levadas em consideração na distribuição de confiança.

Várias métricas foram propostas na literatura para o cálculo da centralidade, em especial a proposta em [Opsahl et al. 2010] se encaixa particularmente bem no problema proposto por incorporar simultaneamente o grau (número de conexões) e a força (os pesos de cada conexão) dos nodos. Aqui, o peso pode ser uma combinação da expertise dos autores com a relevância da obra. A formula proposta faz isso definindo um parâmetro α para ajustar a importância de grau e força:

$$C_D^{w\alpha}(i) = k_i \times \left(\frac{s_i}{k_i}\right)^{\alpha} = k_i^{1-\alpha} \times s_i^{\alpha}$$

Ao incorporar a expertise dos autores e a relevância das publicações como um peso para as arestas da rede, é reintroduzido o conceito de considerar a confiança da comunidade nos colaboradores que depositaram confiança em determinado autor através de publicações em conjunto para o cálculo da confiança estimada de tal autor. Assim, é

possível chegar em um fluxo de confiança apurado levando em conta informações sobre obras e autores que são relevantes para considerar a confiança compartilhada entre os membros da rede.

4.2.4. Segunda métrica de propagação de confiança (local)

Explicar o algoritmo para computar a cofiança local (alternativa ao pagerank)

4.2.5. Algoritmo baseado em conteudo

Apos computar a confiança, apliar um algoritmo baseado em conteudo simples para selecionar artigos para recomendação. a seleção será ordenada conforme a confiança.

4.3. Validação

Referências

CNPq (2019). Sobre a plataforma lattes. `http://memoria.cnpq.br/web/portal-lattes/sobre-a-plataforma`. Acesso em: Abril/2019.

da Rosa Furlan, L. A., de Oliveira Zamberlan, A., Vieira, S. A. G., and Canal, A. P. (2018). Desenvolvimento de um sistema de recomendação para bibliotecas digitais. *Disciplinarum Scientia— Naturais e Tecnológicas*, 19(1):87–104.

Ekstrand, M. D. and Konstan, J. A. (2019). Recommender systems notation: Proposed common notation for teaching and research. *arXiv preprint arXiv:1902.01348*.

Opsahl, T., Agneessens, F., and Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social networks*, 32(3):245–251.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Pazzani, M. J. and Billsus, D. (2007). Content-based recommendation systems. In *The adaptive web*, pages 325–341. Springer.

Perra, N. and Fortunato, S. (2008). Spectral centrality measures in complex networks. *Physical Review E*, 78(3):036107.

Primo, T. and Loh, S. (2006). Técnicas de recomendação para usuários de bibliotecas digitais. *Simpósio Brasileiro de Sistemas de Informação. Curitiba, PR*.

Ricci, F., Rokach, L., and Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender systems handbook*, pages 1–35. Springer.

Sinha, R. R. and Swearingen, K. (2001). Comparing recommendations made by online systems and friends. In *DELOS*.

Victor, P., De Cock, M., and Cornelis, C. (2011). Trust and recommendations. In *Recommender systems handbook*, pages 645–675. Springer.

Ware, M. and Mabe, M. (2015). The stm report. *International Association of Scientific, Technical and Medical Publishers*, page 5.