Sistema de Recomendação Baseado em Confiança para Promover a Colaboração em Redes de Pesquisa Científica

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

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

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

Abstract. *This paper presents a proposal for building a recommender system that promotes collaboration between researchers in a publication network. A trust network is abstracted in which researchers are bound by joint publications, which represents a mutual trust statement. Metrics are discussed for computing trust: global trust (from the community's perspective), local trust (subjective to each researcher) or via the concept of distances, seeking "friends of a friend". An architecture is proposed to improve recommendation quality trough profile analysis, as well as validation metrics.*

Resumo. *Este artigo apresenta a proposta da elaboração de um sistema de recomendações para promover a colaboração entre pesquisadores em uma rede de publicações. É abstraída uma rede de confiança na qual pesquisadores são unidos por publicações em conjunto, que constituem um voto de confiança mútua. São discutidas métricas de computação de confiança: confiança global (da comunidade como um todo), local (subjetivo a cada pesquisador) ou com o conceito de distâncias, buscando "amigos de amigos". Foi proposta uma arquitetura para melhorar as recomendações por meio de análise de perfil, além de métricas de validação.*

1. Introdução

No método científico, pesquisadores devem realizar um trabalho criativo sistemático para incrementar o conhecimento na área da pesquisa. Parte importante do processo é a busca, dentro do universo dos trabalhos científicos, por embasamento teórico ao tema do trabalho proposto. Além disto, também é relevante conhecer e colaborar com outros 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 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 ideias.

Logo, se faz necessário a filtragem da informação que chega ao pesquisador para maximizar sua eficiência e evitar tempo perdido. A automação da tarefa de filtragem é feita através de sistemas de recomendação. Utilizando técnicas de mineração de dados e inteligência artificial pode-se oferecer conteúdo mais relevante, 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 analisar 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].

Confiança é um sentimento que pode ser descrito como o ato de contar com a credibilidade e consistência das ações de um indivíduo. Trata-se de um elemento citado por muitos autores como o mais importante para a construção de uma relação de trabalho positiva, por ser essencial para criticismo construtivo e evolução mútua dentro do processo de pesquisa [**?**]. Portanto, esse artigo foca na construção de um sistema de recomendações de pesquisadores baseado em confiança.

Neste trabalho, o foco será sistema de recomendação guiado por estimativa, ou confiança ou heurística. Na estatística, a confiança é um parâmetro de estimativa do intervalo observado [Ekstrand and Konstan 2019], ou seja, qual a porcentagem de confiança da recomendação dada pelo sistema.

1.1. Objetivos

Este trabalho propõe a elaboração de um sistema de recomendações de pesquisadores e trabalhos científicos baseado em confiança a partir de dados da plataforma Lattes. Para tal, é preciso:

- Estudar funções e aplicações de sistemas de recomendação
- Modelar a rede de confiança da comunidade científica
- Estabelecer métricas de confiança para os dados disponíveis
- Estimar a confiança entre os pesquisadores
- Pré-selecionar recomendações
- Filtrar a pré-seleção com a confiança computada
- Avaliar o modelo

A proposta é analisar as relações de confiança entre os pesquisadores do banco e sugerir conteúdo relevante baseado em seus perfis, ponderando o conteúdo com base na confiança estimada. Para solucionar o problema explanado, propõe-se descrever uma rede de colaborações a partir de publicações em conjunto, discutida na Seção 3.2 utilizando técnicas encontradas na literatura para computar a propagação de confiança na rede. A partir disso, é discutida a pré-seleção dos itens com o método baseado em conteúdo (Seção 3.3), proposta uma arquitetura para o sistema de recomendações (Seção 3.4) e uma metodologia para sua validação (Seção 3.5).

2. Revisão Bibliográfica

Nesta seção, são apresentados os principais itens que compõem o embasamento teórico usado como ponto de partida no trabalho e usado para justificar as decisões tomadas.

2.1. 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, por exemplo). 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 itens de seu interesse. Tais sistemas são especificamente úteis quando um usuário encontra dificuldades para analisar os itens ofertados e fazer escolhas. Os itens recomendados pelo Sistema de Recomendação (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 [Ricci et al. 2011].

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 por meio de avaliações, ou implicitamente com 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 [Shani and Gunawardana 2011].

2.1.1. 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 lida-se 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].

2.1.2. 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 ideia 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 da *primeira avaliação*: quando há um item novo, sem nenhuma avaliação, como saber se determinado usuário irá avaliar positivamente o mesmo? Nenhum de seus vizinhos fez avaliações [Ricci et al. 2011].

SR 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 itens, tal como o algoritmo **KNN** (*K-Nearest Neighbours*) [da Rosa Furlan et al. 2018].

2.1.3. Método Baseado em Conteúdo

O método baseado em conteúdo parte da ideia de que usuários têm interesse em itens semelhantes àqueles que lhe foram uteis 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. Após, 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, 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].

2.1.4. Método Baseado em Confiança

Conforme [Sinha and Swearingen 2001], 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 e a qualidade das recomendações de amigos superam inclusive as feitas por sistemas de recomendação. 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 SR sociais depende do estabelecimento de uma **rede de confiança**, rede 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 SR usam 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 que a recomendação é geralmente 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].

2.1.5. Métodos Híbridos

Métodos híbridos propõem a combinação de mais de um método de recomendação dentro de um sistema. É necessário para complementar técnicas que podem apresentar problemas em determinadas situações ou para oferecer resultados melhores aos usuários. Furlan [da Rosa Furlan et al. 2018] por exemplo combinou os métodos baseado em conteúdo e filtragem colaborativa para solucionar o problema da primeira avaliação. Já Massa [Massa and Avesani 2004] sugere que um método que leve em consideração a confiança entre usuários pode melhorar a performance de sistemas de filtragem colaborativa.

2.2. Análise de Dados em Redes Sociais

Pode-se pensar na rede de colaborações como sendo um grafo no qual os nodos são pesquisadores e as arestas publicações em conjunto. Além disso 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 por meio da rede. A confiança estimada de determinado pesquisador pode levar em consideração o nível de confiança estimado dos pesquisadores que colaboraram com ele. Outro fator que pode ser considerado é a facilidade de colaboração, levando em consideração "distâncias" na rede: se dois pesquisadores A e B têm laços de confiança com um intermediário C, a colaboração entre A e B tende a ser mais fácil do que se houvesse mais intermediários na rede.

O algoritmo PageRank [Page et al. 1999] 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. Trata-se de um método para computar um *ranking* global de citações, pensado para obter a importância das páginas 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.

Funciona da seguinte forma: 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

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

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*: referências exteriores injetam *ranking* no ciclo, fazendo com que páginas do ciclo acumulem pontuação, porém sem 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 centralidade è uma métrica da teoria dos grafos usada para representar a importância de um nodo na rede. Centralidade de grau è definida como o numero de arestas com as quais um nodo se conecta. A métrica de centralidade apresentada em [Opsahl et al. 2010] se encaixa particularmente bem em casos nos quais o peso da aresta representa a força da conexão, tal qual o problema proposto neste trabalho, por incorporar simultaneamente o grau (número de conexões) e a força (os pesos de cada conexão) dos nodos. O peso pode ser a soma das relevâncias das obras publicadas em conjunto entre os pesquisadores. A fórmula proposta pelos autores faz isso definindo um parâmetro α para ajustar a importância de grau e força:

$$C_D^{w\alpha}(i) = k_i \times \left(\frac{s_i}{k_i}\right)^{\alpha} = k_i^{1-\alpha} \times s_i^{\alpha} \tag{2}$$

No contexto de distância, o algoritmo de Dijikstra [Dijkstra 1959] é definido para calcular distâncias em redes nas quais os pesos representam o custo de travessia. O trabalho de Opsahl e Skvoretz [Opsahl et al. 2010] é definido em redes onde os pesos representam força dos laços, então os autores sugerem que os pesos devem ser invertidos. Além disso, o objetivo do trabalho é considerar também o número de nós intermediários, então os autores propõem novamente o uso de um parâmetro de ajuste, α, que controla o quão importante considera-se o número de nodos intermediários e a força das conexões.

$$d^{w\alpha}(i,j) = min\left(\frac{1}{(w_{ih}^{\alpha})} + \cdots + \frac{1}{(w_{hj}^{\alpha})}\right) \tag{3}$$

Frequentemente na análise de dados é necessário o uso de clusterização. O algoritmo MeanShift é indicado para dados nos quais se espera muitos *clusters* distintos e de tamanhos variáveis [**?**]. A implementação é descrita como uma busca por centróides

baseada em grafo de vizinhos mais próximos. O parâmetro *bandwidth* é usado como estimativa para o tamanho dos clusters.

2.3. 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.

2.3.1. Desenvolvimento de um Sistema de Recomendação para Bibliotecas Digitais

No trabalho de [da Rosa Furlan et al. 2018], é abordado 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.

2.3.2. Técnicas de Recomendação para usuários de Bibliotecas Digitais

Em [Primo and Loh 2006], são apresentadas 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. Além disto, o trabalho apresenta ainda um experimento para ilustrar a importância da opinião de especialistas.

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.

2.3.3. Trust-aware Collaborative Filtering for Recommender Systems

No artigo [Massa and Avesani 2004], sugere-se a possibilidade de melhorar as sugestões em sistemas de recomendação com métricas de confiança, descrevem a modelagem de uma rede de confiança e a necessidade de métricas de propagação de confiança, que consideram ser computável em mais usuários do que a similaridade de perfis. A métrica usada é a distância mínima entre nós para a estimativa de confiança local. Os autores sugerem

ainda a aplicação do algoritmo PageRank [Page et al. 1999] como métrica de confiança global. Pretende-se seguir a arquitetura sugerida no trabalho para a construção de um SR que combina os métodos baseados em conteúdo e confiança, que é composta por módulos substituíveis que representam conceitualmente a aplicação de um algoritmo. A adaptação da arquitetura está descrita na Seção 3.4.

3. Metodologia

Para chegar nas recomendações, é proposto um trabalho em dois momentos: estimar a confiança entre pesquisadores e então selecionar potenciais colaboradores baseando-se no perfil dos pesquisadores. A seleção será filtrada e ordenada de acordo com o nível de confiança estimado dos pesquisadores. O primeiro passo é modelar uma rede de confiança da comunidade científica, descrita por autores e publicações. Pode então ser feita uma seleção dos pesquisadores cadastrados com base no perfil do usuário alvo e ordenar a seleção de acordo com o nível de confiança de cada pesquisador. 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 confiança da comunidade em cada pesquisador. Em termos de ferramentas para implementação, foi escolhida a linguagem Python, que possui riqueza de recursos para trabalhos relacionados a manipulação de dados e computação numérica.

3.1. Dados de análise

Os dados usados no trabalho são provenientes do banco de dados relacional da Plataforma Kennis (www.kennis.com.br), que extraí os dados dos currículos de pesquisadores cadastrados na Plaforma Lattes com o uso de um *parser*, cuja versão inicial é descrita em [Prass et al. 2019]. Optou-se pelo uso dos dados da Plataforma Kennis e não pelos dados originais da Plataforma Lattes, pois durante o processo de *parse* dos currículos, a Plataforma Kennis faz a limpeza e o pré-processamento dos dados, associando pesquisadores e suas publicações, conforme mostra a Figura 1.

A partir do banco descrito, é possível construir uma base de conhecimento focada especificamente no objetivo do presente trabalho.

Uma publicação científica, normalmente, é uma produção de algum **tipo** (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 um conjunto de **palavras-chave**, informadas 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**.

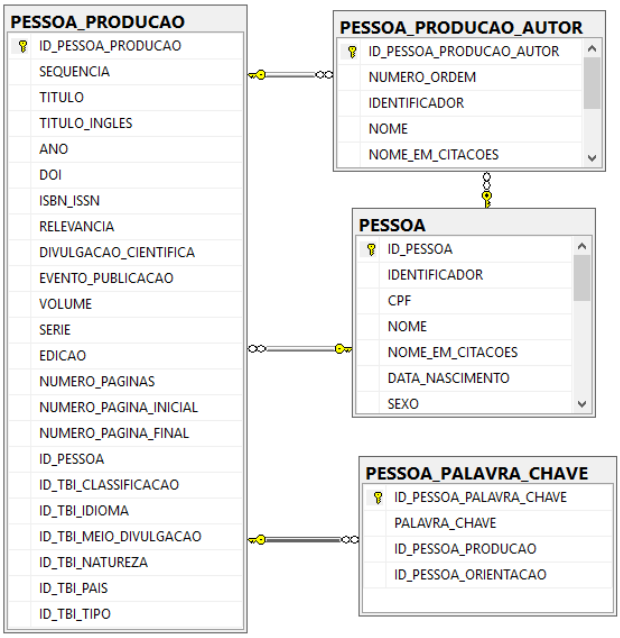

Figura 1. Tabelas Pessoa, Produção e associativa [Prass et al. 2019].

Os autores são o centro dos dados da base de conhecimento e são também responsáveis por cadastrar todas as informações lá encontradas. 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.

3.2. Computando confiança

A Figura 2 representa a rede de confiança desta proposta (discutida na Seção 2.2, ilustrando os pesos das arestas (discutidos na Seção 3.2.1).

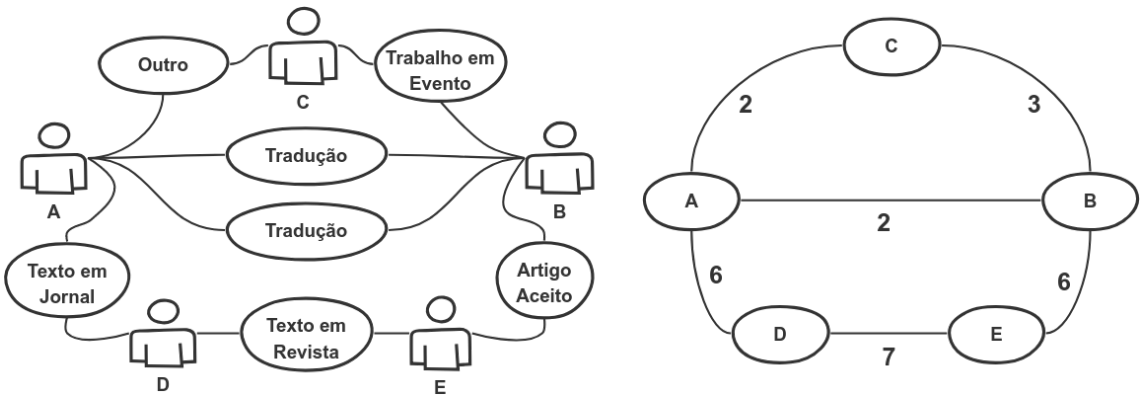

Figura 2. Rede de confiança

A aplicação de PageRank (discutido na Seção 2.2) a 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

do autor (**centralidade** do nodo). Enquanto esta métrica é relevante, perde-se a ideia inicial: não é considerada a confiança dos colaboradores, somente o valor total de colaborações. Além disso, dois conceitos importantes não são levados em consideração: a relevância e o número de colaborações entre os pesquisadores. No caso da **relevância**, a importância da publicação é uma dica para o nível de confiança mútua entre os pesquisadores: colaborações em publicações importantes requerem maior confiança. O total de **colaborações em conjunto** entre um par de pesquisadores, por sua vez, indica uma relação mais duradoura, com mais confiança mútua. É importante distinguir total bruto de publicações de determinado pesquisador e o número de colaborações entre dois pesquisadores, pois há mais confiança quando observa-se frequentes colaborações. Uma vez estabelecida uma heurística para a importância de determinada colaboração, pode-se somar as importâncias para avaliar a força dos laços de confiança.

3.2.1. Relevância, Centralidade e Distância

As publicações não são iguais entre si. Para construir uma heurística ou estimativa que defina a relevância de uma publicação deve-se considerar as características descritas na representação da mesma, atribuindo pesos aos seus atributos. Aqui, a heurística considerada é a natureza da publicação, com pesos atribuídos conforme a Tabela 1.

Tabela 1. Heurística de Relevância.

Tipo de Publicação	Peso
Livro Publicado ou Organizado	9
Capítulo de Livro Publicado	8
Artigo Publicado	8
Artigo Aceito para Publicação	7
Texto em Jornal/Revista	6
Trabalho em Evento	3
Outra Produção Bibliográfica	2
Prefácio/Posfácio	2
Tradução	1

Vale ressaltar que a heurística neste caso é relativa por haver muitos fatores que influenciam a relevância de uma publicação: artigos em certas publicações de prestígio podem valer mais que capítulos de livro, e o mesmo pode ser verdade para textos em jornais ou revistas. Da mesma maneira, publicações mais recentes podem valer mais do que publicações mais antigas, porém o contrário pode ser verdade para linhas de pesquisa na área da história, por exemplo. É possível também considerar mais do que a relevância e quantidade das publicações para descrever os laços de confiança entre pesquisadores.

Considerando as heurísticas discutidas, é possível alcançar um fluxo de confiança mais interessante na rede modelada. O conceito de agregação e dissipação de confiança pode seguir a ideia do algoritmo *PageRank* aplicando-se um algoritmo para o cálculo da centralidade dos nodos (Equação 2). No caso, a relevância das publicações e a quantidade de colaborações devem ser levadas em consideração na distribuição de confiança.

Ao incorporar o número 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.

Na Tabela 2 são apresentados os diferentes valores computados para a centralidade de cada nodo aplicando-se a Equação 2 na rede exemplificada na Figura 2, variando o parâmetro α. No caso, as recomendações seriam ordenadas do maior para o menor valor computado. Para $\alpha = 0$ a centralidade de cada pesquisador seria igual ao número de colaborações, aumentando-se α é atribuída maior importância para a heurística de relevância das publicações.

Tabela 2. Centralidade

Nodo	$C_D^{w\alpha}(i)$			
	$\alpha = 0.00$	$\alpha = 0.50$	$\alpha = 1.00$	$\alpha = 1.50$
A	3.00	5.47	10.00	18.25
B	3.00	5.74	11.00	21.06
C	2.00	3.16	5.00	7.90
D	2.00	5.09	13.00	33.00
E	2.00	5.09	13.00	33.00

A centralidade do pesquisador, ponderada pelo número de colaborações e suas relevâncias se mostra em teoria uma forte métrica de confiança global. O cálculo de confiança local, porém, oferece uma estimativa da **confiança subjetiva** de um usuário em relação aos membros da rede. Outra métrica valiosa neste contexto é a distância entre os pesquisadores, isto é, identificar **amigos de amigos** e pesquisadores próximos na rede é uma maneira de promover a colaboração entre pesquisadores: pesquisadores próximos na rede podem encontrar mais facilidade para realizar colaborações. Para tal, a relevância e quantidade de colaborações (pesos das arestas - confiança) deve ser um fator positivo e o número de nodos intermediários entre os autores um fator negativo (maior distância).

O conceito de distância (discutido na Seção 2.2 e ilustrado na Equação 3), aplicado à rede de colaborações, significa que a distância entre os pesquisadores levará em consideração o nível de confiança das conexões bem como o número de pesquisadores intermediários. Para $\alpha < 1$, caminhos com maior número de intermediários são considerados mais distantes, enquanto $\alpha > 1$ vai considerar mais importante a força das relações de confiança, atribuindo menores distâncias para caminhos onde há fortes relações de confiança entre os pesquisadores, podendo estes ter mais intermediários [Opsahl et al. 2010]. Na Tabela 3 são apresentadas as distâncias calculadas entre os pesquisadores A e B, conforme a rede exemplificada na Figura 2. As recomendações seriam ordenadas da menor para a maior distância.

3.3. Pré-seleção de Recomendações

Com as métricas de confiança aqui propostas, é possível estabelecer níveis de confiança da comunidade, uma rede subjetiva do pesquisador alvo ou distâncias ponderadas por confiança e usar as predições para oferecer sugestões de colaboradores para cada membro

Tabela 3. Distâncias

Caminho	$d^{w\alpha}(i,j)$			
	$\alpha = 0.00$	$\alpha = 0.50$	$\alpha = 1.00$	$\alpha = 1.50$
{A, B}	1.00	0.81	0.50	0.35
{A, C, B}	2.00	1.53	0.83	0.54
{A, D, E, B}	3.00	1.72	0.47	0.19

da rede em diferentes contextos. Todavia confiança apenas pode não ser suficiente para recomendações de qualidade. Considerar também a linha de pesquisa do pesquisador no momento e os perfis dos potenciais colaboradores pode aumentar a qualidade das recomendações: mesmo que a confiança em um pesquisador seja alta, a sugestão de colaboração pode não fazer sentido caso as linhas de pesquisa não se encaixem.

Para aprimorar as recomendações, é proposta uma pré-seleção dos itens utilizando um método baseado em conteúdo. A técnica consiste no cálculo de correspondência de palavras chave em um modelo de espaço vetorial (MEV), visto que esta é a mais comum em sistemas de recomendação baseados em conteúdo [Ricci et al. 2011]. No MEV proposto, a ideia é estabelecer um perfil geral dos pesquisadores através de palavras chaves ponderadas por relevância, extraídas dos títulos e palavras chaves das publicações do qual o autor fez parte. O perfil do pesquisador é representado por um vetor em um espaço *n-dimensional*: $d_j = w_{1,j}, w_{2,j}, \ldots, d_{n,j}$, no qual $w_{i,j}$ representa o quanto o termo i é relevante dentro do trabalho do pesquisador j. Pode-se pensar em uma matriz na qual as linhas são pesquisadores, conforme descrito e as colunas representam os termos-chave extraídos do universo das publicações (*corpus*), removendo palavras vazias - "ou", "de", "para" ... - tanto em português quanto em inglês.

Tal matriz é construída por meio da técnica de vetorização *TF-IDF*, na qual considera-se termos importantes aqueles que aparecem com frequência relacionados á um item específico e com menor frequência nos outros itens do *corpus* [Pazzani and Billsus 2007]. A partir disso é preciso computar a semelhança entre termos. Para tal, a proposta é o uso da similaridade de cossenos por ser a técnica mais comumente aplicada [Ricci et al. 2011]. Para a pré-seleção dos itens, a proposta é basear-se em uma *query* que representa o trabalho sendo desenvolvido pelo pesquisador no presente. A partir desta, pode-se calcular as similaridades dos perfis dos autores cadastrados com a *query*, construindo assim a pré-seleção.

3.4. Arquitetura

Até este ponto do texto foram descritas técnicas que podem ser usadas para a computação de confiança entre membros de uma comunidade, resultando em vetores de confiança global e local, bem como distâncias ponderadas por confiança. Também foi discutida a possibilidade de aprimorar recomendações baseadas em confiança usando o conteúdo dos itens disponíveis. Agora, detalha-se a descrição geral de como é possível combinar as metodologias descritas para obter um resultado.

Em [Massa and Avesani 2004] é sugerida uma arquitetura de SR combinando filtragem colaborativa e método baseado em confiança. O sistema é descrito em módulos substituíveis e portanto pode ser usado para combinar os métodos baseado em conteúdo e em confiança. Basicamente, conforme apresentado na Figura 3, a saída do método usado

para pré-seleção é usada para filtrar e ordenar as recomendações a partir das confianças computadas.

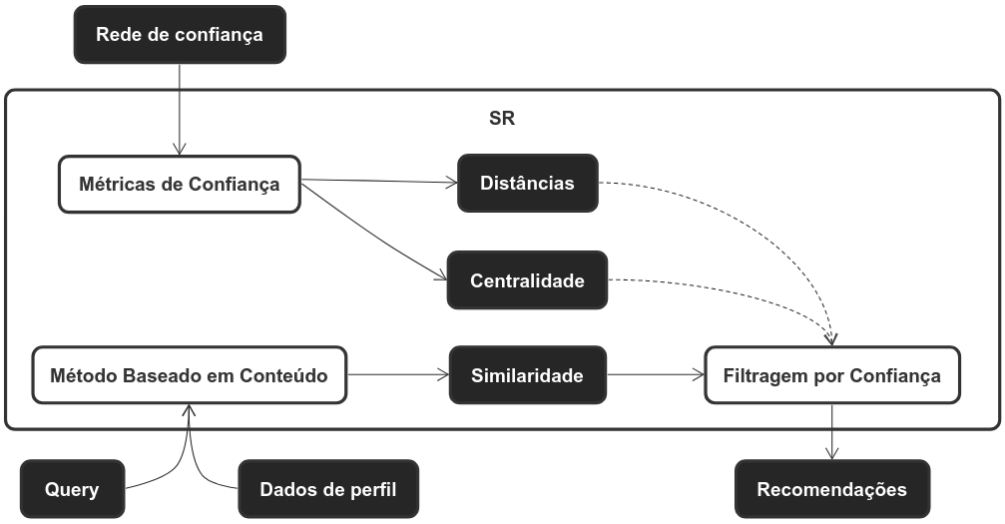

Figura 3. Arquitetura do Sistema de Recomendação proposto.

É preciso então definir um método para oferecer ao usuário uma amostra seleta com os pesquisadores mais relevantes seguindo as métricas de confiança (filtragem por confiança) a partir de uma lista de itens ordenada por similaridade de cossenos e vetores de confiança estimada. Neste ponto, é possível usar alguma técnica para combinar as métricas de confiança, porém é importante as aplicar individualmente para observar os resultados. As linhas em pontilhado representam que os objetos são intercaláveis: pode-se usar um ou outro ou combinações. Após a pré-seleção, o custo computacional da aplicação das distâncias, por exemplo, diminui consideravelmente pois é preciso apenas computar distâncias entre o pesquisador alvo e uma amostra pequena de pesquisadores com o perfil compatível com a *query*.

Portanto, uma vez aplicado o método por conteúdo, pode-se aplicar a métrica de distâncias ponderadas por confiança em uma amostra reduzida da rede, visto que a métrica em questão em teoria é mais indicada para promover colaboração entre os membros da rede.

3.4.1. Tecnologias

- ARQUIVOS .CSV
- LINGUAGEM
- FRAMEWORK
- BANCO...
- pandas
- scikit-learn
- numpy
- COMPUTAÇÃO EM NUVEM - QUAL SERVIÇO
- DESCRIÇÃO DA BASE DE DADOS (LINHAS, ARTIGOS, PESQUISADORES)

Os dados foram exportados através de uma consulta ao banco de dados da plataforma Kennis, resultando em um arquivo *csv - comma separated value* com 16793 linhas, cada uma representando uma publicação cadastrada por seu autor. Contém 169 autores e 15821 títulos de obra distintos. É importante notar que os títulos distintos não representam publicações distintas, é possível que dois pesquisadores cadastrem a mesma publicação com leves variações no título (erros de caligrafia, supressão de artigos, distinções de acentuação e de letras maiúsculas ou minúsculas), o que pode causar com que a mesma publicação não seja reconhecida comparando seus títulos diretamente. Por isso, se faz necessário agrupar as publicações em comum cadastradas por pesquisadores distintos.

A biblioteca pandas foi utilizada para facilitar a descoberta de conhecimento em bancos de dados, oferecendo objetos que podem encapsular uma base de dados e facilitar sua manipulação. Aliado com a biblioteca scikit-learn, que implementa diversos algoritmos comumente usados na mineração de dados e computação científica, tais como vetorização TF-IDF e o algoritmo de clusterização MeanShift [**?**]. Também foi utilizada a biblioteca numpy para computação numérica, que possibilita a manipulação dos dados discutidos no artigo.

3.5. Validação

Várias métricas são sugeridas para medir a *performance* de um Sistema de Recomendação (SR). Quando a tarefa do sistema gira em torno de predizer avaliações, é possível realizar predições e calcular o erro diretamente, de maneira supervisionada. Em [Shani and Gunawardana 2011], os autores sugerem que é importante considerar o contexto da aplicação de cada SR, pois como os objetivos de diferentes sistemas variam, é natural que variem também as métricas de validação. É preciso então avaliar cada caso e estabelecer quais as propriedades que influenciam o sucesso do SR.

Considerando que o principal fator de sucesso do SR aqui proposto é a promoção da colaboração entre os pesquisadores da rede de publicações, a métrica mais completa para a validação do sistema é subjetiva ao usuário:

- Estaria ele interessado em colaborar com o pesquisador recomendado ?
- Qual a facilidade de iniciar a colaboração ?
- Qual a confiança do usuário no pesquisador responsável ?

Logo pode-se seguir por dois caminhos para a validação: implantar o sistema em alguma aplicação e acompanhar seus resultados no mundo real, hipótese na qual considera-se uma recomendação de sucesso aquela que resultar em colaborações, ou realizar recomendações para uma amostra seleta de pesquisadores e recolher seus *feedbacks* sobre as recomendações.

No contexto da implementação proposta neste artigo, somente é possível a segunda opção. Portanto foi elaborado um questionário em dois momentos. A proposta é primeiramente perguntar a um grupo de pesquisadores qual a linha de pesquisa estão desenvolvendo, nas seguintes palavras: *"No que você está trabalhando no momento?"* - A resposta para tal pergunta será usada como a *query*, conforme ilustrado na Figura 3. A partir disso serão elaboradas as recomendações conforme a Seção 3.4. A validação virá no segundo momento do questionário, no qual os pesquisadores serão apresentados às

sugestões elaboradas e indagados sobre sua qualidade. É considerado que três perguntas são suficientes para a avaliação:

- *"Você acha relevante a colaboração com este pesquisador?"*
  - Resposta: sim/não
- *"Em uma escala de um a dez, o quão fácil você considera realizar uma colaboração com este pesquisador?"*
  - Resposta: Valor no intervalo [1, 10], 1 representa facilidade mínima e 10 representa máxima facilidade
- *"Em uma escala de um a dez, qual o seu nível de confiança nesse pesquisador?"*
  - Resposta: Valor no intervalo [1, 10], 1 representa confiança mínima e 10 representa confiança total.

4. Resultados

Para a base de dados da plataforma Kennis, foram feitos testes com o parâmetro *bandwidth* variando de 0.3 até 1.0, fazendo a comparação dos resultados

AS IMAGENS

5. Análise e Discussão dos Resultados

TRECHO DE ALGUM CÓDIGO E O RESULTADO GERADO.....

TABELAS DE RESULTADOS....

Ações ou tarefas realizadas para "limpeza"ou processamento da base fornecida pela plataforma Kenis:

- clusterização dos títulos, pois não havia um identificador dos artigos. Com essa clusterização foi possível agrupar por título os artigos;
- na geração da matriz TFIDF (), foi preciso gerar a tabela com todos os termos que apareciam nos títulos dos artigos, entretanto, foi necessário excluir as palavras ou expressões como artigos, preposições (palavras vazias). Essa matriz tinha as 17 mil linhas da base original, mas 20 mil colunas, referentes aos termos dos títulos. O valor da linha coluna era a importância do termo para o artigo.

Alguns detalhes de implementação e processamento da base:

- o tempo de processamento para a clusterização foi em média de 10 horas, e varia de acordo com o parâmetro bandwidth do algoritmo MINSHIFT;
- no serviço de processamento em nuvem utilizado, era fornecido 35 threads/núcleos, mas devido à memória, somente 7 núcleos;
- o porquê de não ter sido processado localmente

6. Considerações Finais

Neste trabalho buscou-se apresentar sistemas de recomendação e promover a colaboração entre membros de uma comunidade de pesquisadores. Foram discutidos conceito de Sistema de Recomendação, as técnicas mais utilizadas e uma tendência mais recente de recomendações baseadas em confiança. Sugeriu-se um modelo de rede de confiança na qual pesquisadores são conectados por publicações em conjunto, considerado um

voto de confiança. Também foram discutidas três métricas de propagação e agregação de confiança na rede, usando conceitos do algoritmo PageRank [Page et al. 1999] e de métricas de centralidade e distância entre nodos em grafos não-direcionados ponderados [Opsahl et al. 2010].

A partir disso, foi proposta a recomendação por meio da pré-seleção de perfis baseada em conteúdo seguida de filtragem dos perfis via distância ponderada por confiança, seguindo a arquitetura proposta por Massa e Avesani [Massa and Avesani 2004] para um SR híbrido utilizando o método baseado em conteúdo.

Para a avaliação do modelo, foi apresentado o fator de sucesso que justifica a validação por meio de questionários respondidos pelo usuário mediante sugestões elaboradas pelo SR.

É possível estender o trabalho proposto pensando em melhores heurísticas para a relevância das publicações, considerando por exemplo a data de publicação, com diferentes pesos para diferentes épocas. É possível também incorporar mais métodos discutidos na seção 3.2 para melhorar as recomendações, levando em consideração por exemplo métricas locais e globais, além das distâncias.

Referências

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

[da Rosa Furlan et al. 2018] 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.

[Dijkstra 1959] Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271.

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

[Massa and Avesani 2004] Massa, P. and Avesani, P. (2004). Trust-aware collaborative filtering for recommender systems. In *OTM Confederated International Conferences: On the Move to Meaningful Internet Systems*, pages 492–508. Springer.

[Opsahl et al. 2010] 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 et al. 1999] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

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

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

[Prass et al. 2019] Prass, F. S., Matheus Boijink, F., and de Oliveira Zamberlan, A. (2019). Parser e leitura aautomatizada de currículos da plataforma lattes para extração de in-

dicadores acadêmicos e tecnológicos. In *Comunicação, Mídias e Educação*, pages 492–508. Atena Editora.

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

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

[Shani and Gunawardana 2011] Shani, G. and Gunawardana, A. (2011). Evaluating recommendation systems. In *Recommender systems handbook*, pages 257–297. Springer.

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

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

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