EXPLORAÇÃO DE ESTRATÉGIAS PARA A MINERAÇÃO DE EXPERTISE

GABRIEL VALENTIN TIBURCIO[1]

RESUMO

Este projeto de pesquisa individual teve como principais objetivos: o estudo das abordagens de recuperação de informações e mineração de texto, estudo de representações de perfis especializados e estratégias de mineração de *expertise*. A pesquisa explorou a definição de modelos de representação de perfis de competências para professores/pesquisadores da área de Ciência da Computação. Além disso também foram estudadas estratégias que permitem utilizar informações sobre a rede colaborativa de cada professor/pesquisador para compor o perfil de cada um. Este artigo descreve os resultados do projeto e apresenta tecnologias que possibilitam a recuperação de informações e aplica uma técnica de ponderação de termos para estimar o grau de competência de um professor/pesquisador.

PALAVRAS-CHAVES:

Bancos de Dados, Mineração de Dados e Mineração de Competências.

ABSTRACT

The research project described herein had as main objectives: to explore information retrieval and text mining approaches, and to study specialized profile representations and expertise mining strategies. We explored the definition of representation models for expertise profiles for Computer Science researchers. Moreover, we also studied strategies that allow us to use information about the collaborative network of each researcher to compose the profile of each one of them. This article describes the results of the project and presents technologies that make it possible to retrieve information and apply a term-weighting technique to estimate the degree of competence of a researcher.

KEYWORD

Databases, Data Mining and Expertise Mining.

[1] Universidade Federal de Uberlândia, Faculdade de Computação, Programa de Educação Tutorial do curso de Sistemas de Informação, Uberlândia MG.

INTRODUÇÃO

O trabalho de pesquisa desenvolvido durante a iniciação científica teve como objetivo estudar estratégias para mineração de *expertise* no contexto acadêmico. A tarefa de mineração de *expertise* (ou mineração de competências) consiste em encontrar profissionais com determinadas habilidades e conhecimentos [1]. Vários trabalhos científicos têm abordado temas de pesquisa correlatos a esse nos últimos anos. Alguns exemplos desses trabalhos são: o problema de identificar competências de pesquisadores foi investigado em [2]; o trabalho descrito em [3] utiliza uma abordagem baseada em documento para comparar vários tipos de mídia social com o objetivo de identificar *expertise*; um sistema de gerenciamento de *expertise* modelado como um problema de classificação de dados é apresentado em [4]; outro trabalho que define um framework que modela a similaridade profissional entre dois indivíduos com base em informações do LinkedIn sobre suas trajetórias de carreira é descrito em [5]; o workflow proposto em [6] combina estratégias de recuperação de informação, fusão de dados e agrupamento de dados para identificar automaticamente evidencias de *expertise*; entre outros. As abordagens de pesquisa mais recentes têm empregado estratégias para identificar competências de maneira implícita e automática. Essas estratégias, normalmente, consideram perfis que permitem avaliar as competências/habilidades de cada indivíduo. No contexto desse projeto de pesquisa, perfil é definido como um conjunto de características que identificam docentes/pesquisadores no contexto acadêmico. Essas características podem ser de dois tipos: habilidades e conhecimentos individuais e habilidades sociais avaliadas a partir da informação sobre redes de colaboração com outros indivíduos com habilidades correlatas.

As habilidades e conhecimentos individuais podem ser identificado por meio da análise de documentos que descrevam a proposta de projetos e artigos científicos indicados no currículo Lattes do CNPq[2] ou em bibliotecas de teses de universidades. Já as habilidades sociais podem ser obtidas a partir da análise de redes sociais como LinkedIn[3] ou pela análise de dados que podem indicar a existência de relacionamentos implicitamente, como bibliotecas digitais [7].

---

[2] http://lattes.cnpq.br/

[3] https://www.linkedin.com/

A pesquisa descrita aqui considerou que o perfil de um docente pode ser representado por esses dois tipos de habilidades e avaliou diferentes estratégias de mineração de *expertise* nesse contexto. A aplicação de técnicas automáticas para mineração de *expertise* no contexto acadêmico é importante pois pode permitir e aprimorar diferentes tarefas, como: recomendação de colaborações e de artigos acadêmicos, criação de rankings de docentes/pesquisadores considerando diferentes competências, análise da evolução das competências ao longo do tempo etc. [2]

MATERIAL E MÉTODOS

Foram definidas quatro etapas principais para a abordagem do problema: (1) definição do modelo de representação de competências e extração dos dados necessários; (2) definição do perfil do docente/pesquisador; (3) definição das estratégias para o cálculo do grau de competências e (4) exploração e apresentação dos resultados. A Figura 1 ilustra essas etapas que serão detalhas a seguir.

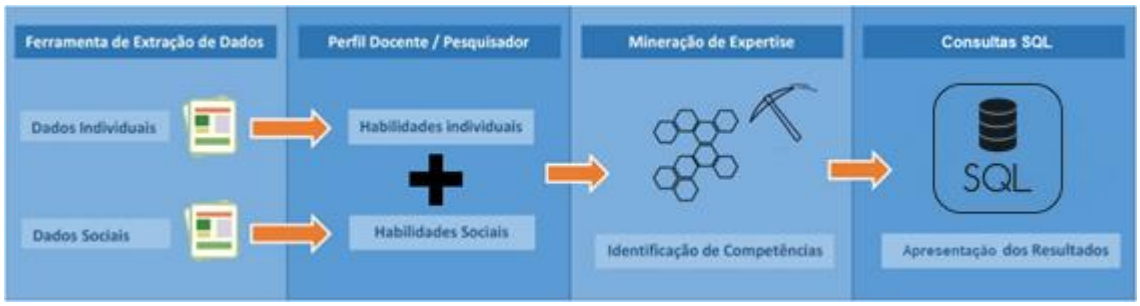

Figura 1. Etapas da abordagem da mineração de *expertise*.

Na etapa 1, o modelo de representação de competências englobou tanto competências individuais quanto competências sociais. As competências individuais dos docentes/pesquisadores foram divididas em duas categorias: habilidades específicas e habilidades gerais (veja Figura 2). A estratégia adotada para definir o perfil de competências dos professores considerou a análise de documentos em língua Inglesa, como artigos científicos, associados aos docentes. As competências específicas e gerais são detalhadas a seguir:

- **Competências Específicas:** *Keywords* e palavras extraídas através de mineração de textos dos resumos dos documentos relacionados aos docentes, definem as competências específicas. Um grau de *expertise* deve ser calculado e associado para cada competência específica relacionada com um docente/pesquisador.

- **Competências Gerais:** Podem ser definidas por meio da associação das competências individuais com uma classificação de áreas de pesquisa de docentes/pesquisadores previamente definida (por exemplo, por uma ontologia). Nas competências gerais, como nas competências específicas, temos um grau de relevância com relação ao pesquisador.

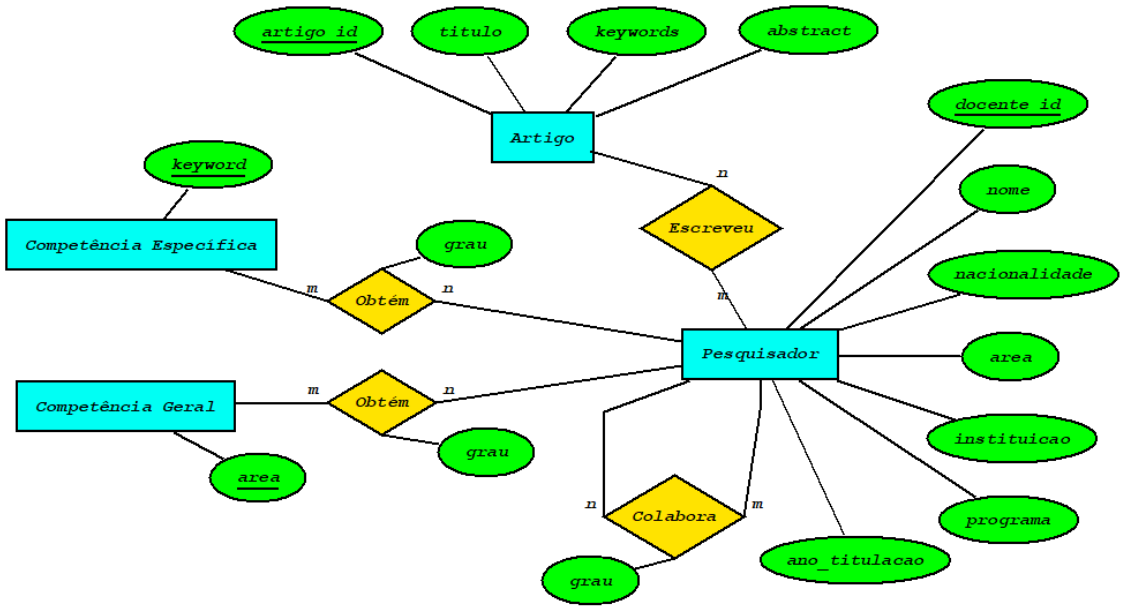

Figura 2. Diagrama Entidade Relacionamento do banco de dados de *expertise* dos docentes/pesquisados da Ciência da Computação no Brasil.

Para a realização do trabalho descrito aqui foram considerados os docentes/pesquisadores da área de Ciência da Computação ligados a algum programa de pós-graduação. Para realizar a coleta dos dados necessários para povoar o banco de dados com o perfil dos pesquisados (veja Figura 2) foi necessário coletar: os dados e afiliação dos pesquisadores (por exemplo, nome, formação acadêmica, área de atuação e instituição. Essas informações são importantes pois podem auxiliar na etapa de busca possibilitando a filtragem de resultados.) e os artigos científicos associados a cada pesquisador.

Foram utilizadas as informações de [8] e [9] para obter a lista de docentes e utilizar a biblioteca digital DBLP (*Digital Bibliography & Library Project*) como repositório para obter os artigos científicos de cada docente. Essa biblioteca foi escolhida por se tratar de uma base de dados bibliográficos totalmente voltada a área de Ciência da Computação e disponibilizar um XML constantemente atualizado. Foi adotada uma estratégia semelhante à adotada em [10]. Em janeiro de 2019, a DBLP

indexava mais de 4,4 milhões de publicações de mais de 2,2 milhões de autores [11]. A Figura 3 ilustra metadados bibliográficos obtidos a partir da DBLP.

```
<article mdate="2017-05-28" key="journals/acta/Saxena96">
        <author>Sanjeev Saxena</author>
        <title>Parallel Integer Sorting and Simulation Amongst CRCW Models.</title>
        <pages>607-619</pages>
        <year>1996</year>
        <volume>33</volume>
        <journal>Acta Inf.</journal>
        <number>7</number>
        <url>db/journals/acta/acta33.html#Saxena96</url>
        <ee>https://doi.org/10.1007/BF03036466</ee>
</article>
```

Figura 3. Representação de um registro na DBLP do tipo artigo.

Para a coleta dos dados dos artigos científicos associados a cada pesquisador foi necessário implementar um *crawler* a partir da listagem dos artigos na DBLP de cada pesquisador. Para tanto, o trabalho descrito em [10] foi adaptado para se adequar ao escopo do trabalho descrito aqui. O objetivo geral do *crawler* era realizar um *parsing* na página XML da DBLP de cada pesquisador para acessar e coletar as informações necessárias dos artigos científicos dos pesquisadores (que são usadas para a definição das competências dos pesquisadores).

O trabalho descrito aqui se concentrou na obtenção de informações a respeito de dois tipos de documentos associados aos pesquisadores: artigos científicos publicados em conferências e artigos científicos publicados em periódicos. Foram coletadas as seguintes informações para cada artigo: título, palavras chaves (*keywords*) e abstract.

O *crawler* foi implementado utilizando a linguagem *Python* com o auxílio da biblioteca *webdriver* do *Selenium* para realizar a navegação na página de cada artigo. O algoritmo funciona de maneira simples, ele espera o conteúdo desejado ser visível na DOM da página web e então, utilizando funções da própria biblioteca, extrai-os e armazena-os em um arquivo criando um registro que será utilizado posteriormente.

Uma dificuldade encontrada no desenvolvimento do trabalho foi a necessidade de se criar um algoritmo diferente para cada tipo de página navegada, pois a DOM difere de uma página para outra. Para tentar contornar esse problema, a escolha dos sites foi limitada aqueles que continham um número razoável de artigos, entre eles os sites da *IEEE, ACM* e *ELSEVIER*. Essa decisão também se justifica pelo tempo elevado que

seria necessário para realizar a extração do conteúdo de todas as páginas. Em trabalhos futuros a coleta de dados de outros sites pode ser expandida.

Após a etapa de coleta de dados e afiliação dos pesquisadores e os artigos científicos associados a cada pesquisador, o banco de dados de expertise dos docentes/pesquisadores da Ciência da Computação no Brasil foi povoado. O diagrama Entidade-Relacionamento mostrado na Figura 2 ilustra a estrutura pensada para esse banco de dados. Foram criadas as entidades "Competência Específica" e "Competência Geral" para armazenar as habilidades coletadas/mineradas. Essas entidades estão associadas à entidade "Pesquisador" por meio de relacionamentos M:N. Esses relacionamentos possuem o atributo "grau" que permite o armazenamento do grau de *expertise* de cada pesquisador em relação a uma competência específica ou geral. O auto relacionamento "Colabora" irá permitir o armazenamento das competências sociais que poderão ser obtidas, por exemplo, a partir das informações das colaborações (coautoria) em artigos científicos.

Para fazer o povoamento do banco de dados foi necessário a construção de um algoritmo simples em *Python* que realiza a leitura dos registros obtidos da etapa de extração e armazena as tabelas do banco de dados relacional que correspondem ao mapeamento das entidades do diagrama ER ilustrado na Figura 2 (veja a Figura 4).

Pesquisador *(docente_id, nome, nacionalidade, area, instituicao, programa, ano_titulacao)*

Artigo *(artigo_id, titulo,keywords, abstract)*

Competencia Especifica *(keyword)*

Competecia Geral *(area)*

GrauCE *(docente_id (Pesquisador.docente_id), keyword (Competencia Especifica.keyword), grau)*

GrauCG *(docente_id (Pesquisador.docente_id), area (Competencia Geral.area), grau)*

Colabora *(docente1 (Pesquisador.docente_id), docente2 (Pesquisador.docente_id), grau)*

Escreveu *(artigo_id (Artigo.artigo_id), autor_id (Pesquisador.docente_id))*

Figura 4. Modelo de dados relacional do banco de dados de expertise dos docentes/pesquisados da Ciência da Computação no Brasil.

Após a finalizadas as etapas (1) e (2), na etapa (3) foi necessário definir as estratégias para o cálculo do grau de competências com base nos dados coletados nas etapas anteriores.

Em um primeiro momento, decidiu-se trabalhar apenas com as habilidades e conhecimentos individuais de cada docente/pesquisador. Várias abordagens para a definição do grau *expertise* foram encontradas em trabalhos correlatos como [2] e [6]. No trabalho descrito o grau das competências específicas de um docente foi calculado, com base na ponderação das palavras-chaves extraídas dos documentos associados a ele (armazenados na relação "Artigo", veja Figura 4). O cálculo da ponderação se deu com base em uma variação da ponderação TF-IDF que consiste em cálculo de frequência de palavras-chaves em um documento e na base de dados.

TF-IDF é uma abreviação do inglês *Term Frequency – Inverse Document Frequency*, esse peso é uma medida estatística usada para avaliar a importância de uma palavra para um documento em uma coleção ou corpus. A importância aumenta proporcionalmente ao número de vezes que uma palavra aparece no documento, mas é compensada pela frequência da palavra no corpus [12]. Abaixo a Equação 1 define o cálculo do TF-IDF de um documento, onde W_{ij} representa o peso de um termo K_i em um documento D_j, N representa o número total de documentos na base, f_{ij} representa a frequência do termo K_i no documento D_j e n_i representa o número de documentos com o termo K_i [13].

$$w_{ij} = \begin{cases} (1 + \log(f_{ij})) \times \log\left(\frac{N}{n_i}\right), & se\ f_{ij} \geq 1 \\ 0, & caso\ contrário \end{cases} \quad (1)$$

No trabalho apresentado aqui, houve uma variação no cálculo do TF-IDF, pois como as palavras-chaves não são repetidas no mesmo documento, foi necessário tratar o cálculo do TF de modo a considerar a repetição de uma palavra-chave na coleção de documentos do próprio autor. Todo o cálculo da ponderação foi realizado em um algoritmo na linguagem *Python* e seus resultados foram usados para definir o valor do atributo “grau” da associação entre as competências especificas de cada docente no banco de dados.

RESULTADOS E DISCUSSÃO

Inicialmente foi proposta a realização de não só uma estimativa do grau de conhecimento de um docente em determinada área de estudo, mas também tentar estimar um grau de colaboração de docentes que realizam pesquisas e tem afinidade nas mesmas áreas. No projeto discutido aqui foi possível estimar, inicialmente, um grau de

conhecimento em assuntos específicos da área de computação dos docentes do Brasil que já publicaram algum artigo listado na DBLP.

Ao todo tivemos um total de 6.490 artigos da DBLP com seus conteúdos extraídos da web, sendo 1.788 classificados como *articles* (artigos publicados em periódicos) e 4.702 classificados como *inproceeding* (publicados em conferências), de um total de 21.000 artigos do Brasil contidos na DBLP. Essa limitação da quantidade de artigos extraídos foi devida ao fato de o tempo de extração do conteúdo de cada página ser bastante elevado. Mesmo assim ainda foi possível obter um total de 9.291 palavras-chaves, cada uma representando uma competência específica de um docente com seu respectivo grau de relevância.

Na etapa (4) do projeto foi possível realizar consultas nos dados obtidos, para explorar diferentes tipos de informação relacionadas as competências dos docentes/pesquisadores. A seguir são apresentados alguns exemplos por meio da realização de consultas em SQL em um banco de dados criado no SGBD *MySQL*.

Consulta 1. A consulta abaixo apresenta as dez competências específicas com maior grau de relevância de um docente selecionado aleatoriamente.

```
SELECT keyword, grau
FROM `GrauCE`
WHERE docente_id = 18
ORDER BY grau DESC LIMIT 10
```

KEYWORD	GRAU
Pattern matching	3.11327469246435
Open source software	2.812244696800369
Vegetation	2.733063450752744
Documentation	2.666116661122131
Knowledge based systems	2.636153437744688

Maintenance engineering	2.4900254020664496
Partitioning algorithms	2.450516860782776
Software architecture	2.3650866654581497
Robustness	2.3070947184804633
Java	2.0962413531655697

Tabela 1. Top 10 competências de um docente aleatório.

Na consulta é realizada uma busca pelo nome da competência específica (*keyword*) e seu grau na tabela "GrauCE", aplicando um filtro para o id do autor (docente_id) igual a 18, ordenando os resultados pelo maior grau de relevância e filtrando para serem apresentados somente os 10 primeiros resultados.

Consulta 2. A consulta abaixo apresenta os três docentes/pesquisadores mais bem ranqueados (com maior grau de relevância) na competência específica de *cloud computing*.

```
SELECT Pesquisador.nome, GrauCE.grau
FROM `GrauCE`, `Pesquisador`
WHERE Pesquisador.docente_id = GrauCE.docente_id
AND GrauCE.keyword = 'cloud computing'
ORDER BY GrauCE.grau DESC LIMIT 3
```

DOCENTE	**GRAU**
Carlos Becker Westphall	2.809896256499121
Raquel Vigolvino Lopes	2.1517503576162373
Judith Kelner	2.1517503576162373

Tabela 2. Top 3 docentes com maior grau de *expertise* em *cloud computing*.

A consulta 2 efetua a busca dos nomes dos docentes/pesquisadores e seu grau de relevância nas tabelas Pesquisador e GrauCE, filtrando os resultados para somente docentes/pesquisadores que contenham algum grau de relevância na competência específica de "*cloud computing*", ordenando os resultados pelo maior grau de relevância e filtrando para retornar apenas os 3 primeiros resultados.

Consulta 3. A consulta abaixo lista as principais competências específicas da Universidade Federal do Rio Grande do Sul (UFRGS) e seu grau de relevância.

```
SELECT GrauCE.keyword, GrauCE.grau
FROM `GrauCE`, `Pesquisador`
WHERE GrauCE.docente_id = Pesquisador.docente_id
AND Pesquisador.instituicao = 'UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL (UFRGS)'
ORDER BY GrauCE. grau DESC
```

KEYWORD	GRAU
Graph Transformation	5.142908481410281
Power Comsumption	4.741832060029271
Mapping	4.176546593441344
Reconfigurable Architectures	4.176546593441344
Network-on-chip	3.8603468020128315
Segmentation	3.8603468020128315
Abbrief	3.812244696800369
Active Clause Coverage (acc) criteria	3.812244696800369
Adders	3.812244696800369

Aggregation and Reduction Algorithms	3.812244696800369

Tabela 3. Principais competências específicas da UFRGS.

A consulta 3 realiza a busca das competências específicas (*keyword*) e seu grau de relevância nas tabelas "GrauCE" e "Pesquisador", filtrando a consulta para retornar somente as competências específicas da UFRGS, ordenando-os pelo maior grau de relevância. A consulta no banco de dados, retornou de 407 competências específicas, porém para demonstração no trabalho, foram apresentadas somente as dez primeiras competências específicas.

Consulta 4. A consulta a seguir apresenta a listagem de competências específicas com o menor grau de relevância de todo o banco de dados.

```
SELECT DISTINCT (keyword), grau
FROM `GrauCE`
ORDER BY grau ASC
```

KEYWORD	GRAU
Software	1.4543098497999152
Computational Modeling	1.4962743513434513
Computer Architecture	1.5177784706387762
Visualization	1.5569721916970631
Context	1.6000570923964113
Monitoring	1.6219129986300775
Training	1.6304011088555965
Cloud Computing	1.6538822047051194
Feature Extraction	1.6599563524173127

Measurement	1.7330634507527443

Tabela 4. Competências específicas com menor grau de relevância.

A consulta 4 efetua a busca do nome da competência específica (*keyword*) e seu grau de relevância na tabela "GrauCE", ordenando os resultados do menor grau de relevância para o maior.

É importante salientar que no banco de dados há um total de 1.914 docentes/pesquisadores registrados, porém quando é observado o número de docentes/pesquisadores em que foi realizado o *crawler* de documentos, é constatado a presença de somente 544 desses docentes/pesquisadores. Esse fato se dá pelo mesmo motivo citado anteriormente, o tempo de extração do conteúdo das páginas web serem elevados.

CONCLUSÃO

O gerenciamento de competências de docentes/pesquisadores em universidades e centros de pesquisa é importante para a interação não só entre pesquisadores da mesma universidade, mas também entre outras universidades e instituições públicas e privadas.

Dentro desse contexto, diferentes modelos de representação de competências foram estudados e investigados no trabalho descrito aqui. Com base no estudo da literatura correlata foi definido um perfil de competências de um docente/pesquisador, além do desenvolvimento de um algoritmo capaz de navegar nas páginas web indexadas na DBLP e recuperar as informações relevantes para as análises e definições de grau de *expertise* de um docente.

Dentre os trabalhos futuros estão a definição das competências gerais e da colaboração entre docentes com afinidade nas mesmas áreas de conhecimento e seus graus de *expertise* a partir dos dados coletados a respeito dos docentes/pesquisadores da área de Ciência da Computação no Brasil. Além disso, também deverão ser exploradas alternativas para a aplicação de técnicas de mineração texto dos resumos dos documentos associados aos docentes e a apresentação visual dos dados minerados.

REFERÊNCIAS

[1] Krisztian Balog and Maarten de Rijke. Determining expert profiles (with an application to expert finding). In IJCAI 2007, Proceedings of the 20th International

Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007, pages 2657-2662, 2007.

[2] Kelly Hannel. Qualificação de pesquisadores por Área da ciência da computação com base em uma ontologia de perfil. Dissertação de Mestrado, UFRGS/PósGraduação em Computação, Porto Alegre-RS, 98 p, 2008.

[3] Ido Guy, Uri Avraham, David Carmel, Sigalit Ur, Michal Jacovi, and Inbal Ronen. Mining expertise and interests from social media. In 22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, pages 515-526, 2013.

[4] Kush R. Varshney, Vijil Chenthamarakshan, Scott W. Fancher, Jun Wang, DongPing Fang, and Aleksandra Mojsilovic. Predicting employee expertise for talent management in the enterprise. In The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014, pages 1729-1738, 2014.

[5] Ye Xu, Zang Li, Abhishek Gupta, Ahmet Bugdayci, and Anmol Bhasin. Modeling professional similarity by mining professional career trajectories. In The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014, pages 1945-1954, 2014.

[6] Raya Horesh, Kush R. Varshney, and Jinfeng Yi. Information retrieval, fusion, completion, and clustering for employee expertise estimation. In 2016 IEEE International Conference on Big Data, BigData 2016, Washington DC, USA, December 5-8, 2016, pages 1385-1393, 2016.

[7] Michele A. Brandão, Pedro O. S. Vaz de Melo, and Mirella M. Moro. STACY: um novo algoritmo para automaticamente classificar a força dos relacionamentos ao longo dos anos. In XXXII Simpósio Brasileiro de Banco de Dados, Uberlandia, MG, Brazil, October 4-7, 2017., pages 136-147, 2017.

[8] Site do Prof. Palazzo. Computação, Tecnologia & Humanismo. Avaliação dos cursos de computação – CAPES 2013. Disponível em: <https://www.palazzo.pro.br/Wordpress/?p=253>. Acesso em: 30/10/2018.

[9] Plataforma Sucupira. Disponível em: <https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/docente/listaDocente.jsf>. Acesso em: 07/11/2018.

[10] Laécio Pioli Junior. Redes Complexas para Análise de Influência entre Pesquisadores. In Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas, Juiz de Fora, MG, Brazil, November, 2017.

[11] DBLP Computer Science Bibliography. What is dblp? Disponível em: <https://dblp.uni-trier.de/faq/What+is+dblp>. Acesso em: 19/03/2019.

[12] TF-IDF. Disponível em: <http://www.tfidf.com/>. Acesso em: 23/06/2019.

[13] Ricardo Baeza-Yates, Berthier Ribeiro-Neto; Recuperação de Informação - Conceitos e Tecnologia das Máquinas de busca. Tradução técnica: Leandro Krug Wives, Viviane Pereira Moreira. – 2. Ed. – Dados eletrônicos. – Porto Alegre: Bookman, 2013.

Aluno: **Gabriel Valentin Tiburcio**

Professor Orientador: **Maria Camila Nardini Barioni**