UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Um Método para Busca de Competências a Partir de currículos Lattes

AURÉLIO ANDRADE DE MENEZES JÚNIOR

Manaus
2012

Aurélio Andrade de Menezes Júnior

Um Método para Busca de Competências a Partir de currículos Lattes

Dissertação de mestrado apresentada ao Programa de Pós-graduação em Informática da Universidade Federal do Amazonas como requisito para obtenção de Título de Mestre em Informática na área de concentração de Banco de Dados e Recuperação de Informação.

Orientador: Dr. Altigran Soares da Silva

Co-orientadora: Dra. Eulanda Miranda dos Santos

Manaus
2012

Júnior, Aurélio Andrade de Menezes.

59 páginas
Dissertação (Mestrado) - Instituto de Computação da Universidade Federal do Amazonas. Programa de Pós-Graduação em Informática - PPGI.

1. Método para Busca.
2. Competências de Pesquisadores.
3. Currículos Lattes.

I. Universidade Federal do Amazonas. Instituto de Computação. Programa de Pós-Graduação em Informática - PPGI.

Banca:

Prof. Dr. Altigran Soares da Silva

Profa. Dra. Eulanda Miranda dos Santos

Profa. Dra. Célia Regina Simonetti

Prof. Dr. João Marcos Bastos Cavalcante

A Deus e a meus pais Aurélio Menezes e Lindalva Menezes.

Agradecimentos

A realização deste trabalho foi conseguida graças ao apoio direto ou indireto de várias pessoas às quais, nesta oportunidade, manifesto toda a minha gratidão.

Primeiramente, agradeço a minha familia por todo o amor, atenção, incentivo e suporte não somente na minha participação no mestrado mas em todos os momentos de minha vida. Em especial agradeço à minha esposa Silvana pelas sugestões indispensáveis nos momentos de dúvida.

Agradeço aos meus orientadores, professor Altigran Soares e professora Eulanda Miranda, por compartilharem comigo os seus conhecimentos e experiências, e ainda por me proporcionarem diversas oportunidades de aprendizado durante o mestrado. Agradeço também à professora Célia Regina Simonetti por sua participação no processo de avaliação deste trabalho.

Agradeço aos bolsistas do Centro de Biotecnologia da Amazônia (CBA), que participaram no processo de avaliação deste trabalho, pelo cuidado no processo de avaliação de relevância.

Agradecimentos também à Universidade Federal do Amazonas e a todos os professores e colaboradores do Instituto de Computação pelo apoio que me foi dado durante o período do mestrado. Meus sinceros agradecimentos à Fundação de Amparo e Pesquisa do Estado do Amazonas (FAPEAM), pelo suporte financeiro, pois com este benefício, pude me dedicar com tranquilidade aos meus estudos.

Por fim, agradeço a Deus por me protejer de todos os imprevistos e por ter posto em meu caminho pessoas tão especiais.

Resumo

Grandes bases de dados tem sido muito comum hoje em dia e tem permitido o acesso a uma grande quantidade de informação. Por outro lado, esse cenário torna dificil a tarefa de encontrar uma informação específica no meio de uma grande quantidade de informação. Sistemas de Recuperação de Informação (RI) têm sido largamente empregados para a solucionar este tipo de problema. Dentre os problemas ocasionados pela grande quantidade de informação disponível em bases de dados, existe o problema da busca de competências. Este problema ocorre no seguinte contexto, dado um perfil, descrito na forma de um conjunto de competências, procura-se descobrir pesquisadores com perfis similares. Este trabalho descreve um método de RI que fornece apoio à busca de pesquisadores a partir de informações sobre competências extraídas de uma base de currículos Lattes. Assim, dada uma consulta especificando um perfil de competência desejada, são selecionados os currículos com maior grau de similaridade com este perfil. Após a execução de experimentos em três estratégias propostas: Soma de Similaridades, Produção e Contagem de Borda, os resultados indicam o sucesso do método proposto.

Palavras-chave: Método para Busca, Competências de Pesquisadores, currículo Lattes.

Abstract

Large databases have been very common nowadays. These databases allow access to a huge amount of information. However, this scenario leads the task of finding a specific information among such a large amount of information, become very difficult. Systems of Information Retrieval (IR) have been widely used to sole this kind of problem. Among the many problems caused by the large amount of information available on databases, there is the problem related to competence searching. This problem occurs in the following context, given a profile, described as a set of competencies, one looks for finding researchers with similar profiles. In this work, we describe an IR method which provides support to find researchers taking into account competence information retrieved from a database of curriculums Lattes. Thus, given a query specifying a desired competency profile, the proposed method provides the curriculums more similar to the desired profile. The experiments were conducted using three proposed strategies: Sum of Similarities, Production and Borda Count. The results achieved show that the proposed successfully accomplishes its objective.

Keywords: Search Method, Competences Researchers, Curriculum Lattes

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

O advento e a grande utilização da Web, assim como os bancos de dados existentes hoje em empresas públicas e privadas, têm aumentado significativamente a quantidade de dados disponíveis e a quantidade de informação trocada eletronicamente. De acordo com Harzallah et al. (2002), a Web tem revolucionado tanto o acesso a informações pessoais quanto o gerenciamento do conhecimento em instituições. Assim como bancos de dados, que tem sido usados com grande frequência no nosso dia-a-dia nas empresas. Se por um lado, temos a facilidade de ter acesso a grandes quantidades de informação, por outro lado, temos que lidar com o problema de encontrar uma informação específica em meio a uma grande quantidade de informação disponível. Máquinas de busca existem para resolver este tpo de problema, pois facilitam o acesso a informações relevantes por meio da sua recuperação. Esses sistemas são projetados para operar em ambientes onde a quantidade de conteúdo disponível supera a capacidade do usuário de acessá-lo de forma eficiente durante a pesquisa.

A obtenção de informação sobre as competências existentes em uma instituição acadêmica, é um exemplo desse tipo de problema. O conhecimento é o principal produto gerado por uma instituição acadêmica. Segundo Rodrigues et al. (2004), devido ao crescimento destas instituições, muitas vezes em locais geograficamente distantes, e

às novas áreas de conhecimento científico que têm surgido, um novo cenário se coloca para estas instituições: a falta de conhecimento sobre suas próprias competências.

A análise da literatura permite identificar três características principais do conceito de competência. Para Harzallah et al. (2002) competência é o efeito de combinar e colocar em jogo os recursos do individuo (conhecimentos, experiências e comportamentos), em um dado contexto para alcaçar um objetivo ou cumprir uma missão específica. Competências são identificadas e formalmente representadas em currículos. Muitos candidatos a empregos depositam seus currículos na Web, no que é normalmente chamado de *Banco de Currículos*, a fim de tornarem conhecidas suas competências. Harzallah et al. (2002), ressaltam em seu trabalho que os maiores Web sites anunciam mais de 50.000 currículos em suas bases.

O Conselho Nacional de Desenvolvimento Científico (CNPq), mantém um banco de currículos conhecido como Plataforma Lattes. Nessa plataforma, são mantidos de forma integrada, currículos acadêmicos de pesquisadores de instituições públicas e privadas do Brasil. De acordo com e Cesar Junior Roberto Marcondes Mena-Chalco (2009), os currículos Lattes são atualmente considerados um padrão nacional, representando um histórico das atividades científicas, acadêmicas e profissionais de pesquisadores cadastrados. Deste modo, podem ser usados como fonte para identificação de competências de pesquisadores que possuam currículos nesta plataforma. De fato, segundo Nature (2010), a Plataforma Lattes tem sido internacionalmente reconhecida como a mais organizada base de dados com informações sobre cientistas existente atualmente.

Portanto, identificar competências a partir de curriculos Lattes é importante para uma instituição de pesquisa, pois por meio deste conhecimento, é possivel por exem-

plo, identificar pesquisadores para: trabalhar em projetos, participar de bancas de concurso, compor de bancas de mestrado, trabalhar como orientador de mestrado e doutorado e trabalhar como membro de comissão de programa de conferência.

Para Rodrigues et al. (2004), a literatura mostra que a utilização de técnicas de Recuperação de Informação (RI), permitem a descoberta de competências de pesquisadores a partir de publicações científicas. RI é a ciência de pesquisa que possibilita a busca por informações em documentos, busca pelos documentos propriamente ditos e busca em banco de dados, tendo aqui como foco principal a recuperação da informação e não dos dados. Diante desse contexto, no presente trabalho, pretendemos usar técnicas de RI para a descoberta de competências a partir de currículos obtidos da Plataforma Lattes.

Temos, portanto, o seguinte contexto como ponto de partida: de um lado uma grande base de currículos da Plataforma Lattes e de outro, usuários que precisam realizar buscas com o intuito de descobrir pesquisadores com determinados perfis. Conforme mencionado anteriormente, como exemplo de aplicação podemos citar: indicação de orientadores, organização de bancas de concursos, seleção de candidatos a emprego, entre outros.

Por meio de levantamento bibliográfico foi observado que a construção e o uso de uma máquina de busca de competências a partir de currículos Lattes, é uma aplicação nova. Porém, como é discutido no capítulo 2, há trabalhos relacionados que buscam a identificação de competências a partir de documentos científicos e currículos genéricos, isto é, não currículos Lattes. Para a realização dos experimentos deste trabalho, avaliação e posterior validação do projeto, foi utilizada a base de currículos Lattes de professores da UFAM.

1.1 Objetivos

O objetivo principal deste trabalho é desenvolver um método de busca que dê apoio à recomendação de pesquisadores a partir de informações sobre competências extraídas de uma base de currículos Lattes. Assim, dada uma consulta especificando um perfil de competência desejada, são selecionados os currículos com maior grau de similaridade com este perfil.

Os objetivos específicos são os seguintes:

1. Selecionar as características, ou seja, os campos disponíveis nos currículos Lattes mais importantes para a busca de competências.
2. Aplicar técnicas de RI para gerar estratégias capazes de recuperar os currículos mais relevantes para uma especificação de perfil dada como entrada.
3. Submeter o método a um experimento com usuários especializados que permita verificar a relevância dos currículos recuperados.
4. Avaliar os resultados do experimento através da aplicação de métricas de avaliação de RI.

1.2 Motivação e Justificativa

É possível observar que métodos tradicionais de chamada e oferta de emprego como por exemplo, jornais e revistas, são muito lentos e caros, e deixam a desejar na capacidade de encontrar profissionais de alta qualidade em um curto espaço de tempo,

considerando o moderno mercado de trabalho atual. Além disso, a evolução do mercado de trabalho, baseado no uso de tecnologias de informação, não foi acompanhada pela evolução das ferramentas dedicadas à recuperação e gerenciamento de currículos e ofertas de emprego. Técnicas usadas para analisar fontes de informação relacionadas à seleção de profissionais para preencher vagas para emprego, ainda são rudimentares, segundo Harzallah et al. (2002).

Dentre as diversas aplicações que dependem do levantamento de competências a partir de currículos, destacamos as acadêmicas e científicas. É importante que as instituições de pesquisa possam conhecer a si próprias, suas publicações, pesquisas em andamento e suas competências internas. De acordo com Rodrigues et al. (2004), também é importante conhecer as pesquisas desenvolvidas por outras instituições para que os pesquisadores conheçam os trabalhos correlacionados em outros centros de pesquisa, e assim, possam estabelecer redes de colaboração ou até mesmo comunidades virtuais, que são a base do desenvolvimento científico.

Argumentamos portanto, que a existência de ferramentas para recomendação de competências permite a melhoria do auto-conhecimento das instituições e do conhecimento de outras instituições, além de permitir identificar o que já foi suposto.CORDER

1.3 Contribuições

O resultado do trabalho aqui proposto visa tornar mais eficiente o processo de seleção de profissionais com competências previamente definidas das áreas de pesquisa, para os mais diversos fins, dos quais podemos destacar: trabalhar em projeto, participar de banca de concurso, participar de banca de mestrado, trabalhar como orientador

de mestrado e doutorado, trabalhar como membro de comissão de programa de conferência.

1.4 Organização da Dissertação

A continuação deste trabalho está organizada da seguinte forma: o Capítulo 2 faz uma breve revisão da literatura utilizada para a execução deste trabalho, além de descrever alguns trabalhos relacionados, os problemas por eles abordados e as soluções propostas. O Capítulo 3 lista a sequência de procedimentos para solucionar o problema proposto, e em seguida, o Capítulo 4 descreve a configuração e a execução dos experimentos, além da avaliação dos resultados obtidos. Por fim, são apresentadas as conclusões deste trabalho e propostas para trabalhos futuros.

2 *Referencial Teórico*

Este capítulo tem como objetivo fazer uma revisão de alguns conceitos e ferramentas utilizadas neste trabalho, além de descrever brevemente os principais trabalhos relacionados com busca de competências.

2.1 Conceitos Básicos

O principal conceito sobre o qual este trabalho está fundamentado é o de competência. São vários os conceitos sobre competência encontrados na literatura, dentre os quais podemos destacar a definição de Hongli (2010), que afirma que competência é o estado ou qualidade para realizar uma determinada função. Engloba uma combinação de conhecimentos, habilidades, comportamento e outras características utilizadas para melhorar o desempenho. Segundo Harzallah et al. (2002), o conceito de competência é geralmente associado a outros conceitos tais como conhecimento, destreza, habilidade, experimentação, aptidão, capacidade, característica de personalidade e etc.

Como mencionado na introdução, as competências podem ser encontradas em currículos. Este trabalho utiliza os currículos Lattes dos professores da UFAM como referência, logo, as competências que envolvem conhecimento adquirido pelo pesquisador, suas habilidades, experiências adquiridas e suas atitudes, serão identificados e

associados a campos dos currículos Lattes. Esse processo de associação é mostrado com detalhes no Capítulo 3.

No trabalho de e Cesar Junior Roberto Marcondes Mena-Chalco (2009), os autores destacam que de fato os currículos Lattes dos pesquisadores são considerados atualmente um padrão nacional, os quais representam um histórico das atividades científicas, acadêmicas e profissionais de pesquisadores cadastrados no CNPq.

Um método de busca de competências a partir de currículos Lattes, torna possível que, dado um perfil (conjunto de competências), seja obtido um conjunto de currículos Lattes compatível com o perfil informado. Em seguida, os currículos selecionados podem ser ordenados de acordo com a relevância. Portanto, a chave para o desenvolvimento do método de busca de competência são as ferramentas de RI, descritas na próxima seção.

2.2 Recuperação de Informação

RI é uma área da Informática que lida com o armazenamento de documentos e a recuperação automática de informação associada a eles. Trata-se da ciência de pesquisa que possibilita a busca por informações em documentos, busca pelos documentos propriamente ditos e busca em banco de dados. De acordo com Baeza-Yates and Ribeiro-Neto (1999), RI é a parte da Ciência da Computação que estuda a recuperação de informação, e não de dados, de uma coleção de documentos escritos, e os documentos recuperados visam satisfazer a necessidade de informação do usuário, normalmente expressa em linguagem natural. Este trabalho trata do tipo específico de busca que ocorre em tarefas de recuperação de informação em documentos disponíveis

na Web. Alguns conceitos muito utilizados nesta dissertação são revisados a seguir. Esses conceitos foram obtidos em Baeza-Yates and Ribeiro-Neto (1999).

2.2.1 Consulta

Uma **consulta** é uma expressão da necessidade de informação do usuário em uma linguagem de entrada provida pelo sistema de informação. Baeza-Yates and Ribeiro-Neto (1999) afirmam que o tipo mais comum de linguagem de entrada permite simplesmente a especificação de palavras-chaves e de alguns conectivos Booleanos. As palavras-chaves devem resumir a informação desejada pelo usuário. Uma inconveniência imediata dessa abordagem é que o uso de palavras-chave geralmente introduz uma diferença de semântica entre a intenção do usuário e o conjunto de documentos retornados. Essa diferença de semântica ocorre devido à dificuldade adicional em se lidar com textos em linguagem natural, que nem sempre são bem estruturados e podem ser semanticamente ambíguos.

2.2.2 Relevância

Para ser eficaz na tarefa de satisfazer a necessidade de informação do usuário, os sistemas de RI, devem recuperar e apresentar os documentos de uma coleção por **ordem de relevância** para o usuário. Portanto, como resultado da consulta, os documentos devem ser primeiramente ordenados de acordo com o grau de relevância (uma classificação é gerada) e em seguida, apresentados ao usuário, o qual examina a lista classificada a partir do topo. A presença de documentos (textos) não relevantes para os usuários, entre aqueles recuperados por um sistema de RI, é praticamente certa. Nesse cenário, o principal objetivo dos sistemas de RI é recuperar o maior número possível

de documentos relevantes e o menor número possível de documentos não relevantes. A noção de relevância é um conceito fundamental para a execução e avaliação desta pesquisa em RI, além de ser um componente chave para processar a classificação (ordenação) de documentos em um conjunto de respostas a uma consulta do usuário.

2.2.3 Listas invertidas e ordenação

Entre as principais etapas do processo de busca em RI está a operação de . Um tipo de estrutura de dados bastante utilizada são as **listas invertidas** de termos e documentos. Por fim, a pesquisa, envolve o processo de recuperação de documentos de acordo com a consulta do usuário e a **ordenação**, que é gerada por meio de um grau de similaridade entre o documento e a consulta.

2.2.4 Modelo vetorial

Para ordenar os documentos a partir dos mais relevantes até os menos relevantes, sistemas de RI usualmente adotam um modelo para representar os documentos e a consulta do usuário. Para a execução deste projeto foi utilizado um dos modelos clássicos de RI chamado **Modelo Vetorial**. O modelo de espaço vetorial, ou simplesmente modelo vetorial, representa documentos e consultas como vetores de termos indexados que são ocorrências únicas nos documentos. Os documentos devolvidos como resultado para uma consulta são representados similarmente, ou seja, o vetor resultado para uma consulta é montado através de um cálculo de similaridade entre os termos das consultas e os documentos, para os quais são atribuídos pesos que especificam o tamanho e a direção de seu vetor de representação. Ao ângulo formado por estes vetores dá-se o nome de q. O termo cos(q) determina a proximidade da ocorrência. O cálculo

da similaridade é baseado neste ângulo entre os vetores que representam o documento e a consulta.

Os pesos quantificam a relevância de cada termo para as consultas (W_{iq}) e para os documentos (W_{id}) no espaço vetorial. Para o cálculo dos pesos W_{iq} e W_{id}, utiliza-se uma técnica que faz o balanceamento entre as características do documento, utilizando o conceito de freqüência de um termo num documento. Se uma coleção possui N documentos e n_{ti} é a quantidade de documentos que possuem o termo ti, então o inverso da freqüência do termo na coleção é chamado idf (inverse document frequency).

Este valor é usado para calcular o peso, utilizando a seguinte fórmula: $W_{id} = freq(ti,d)Xidf_i$ ou seja, é o produto da freqüência do termo no documento pelo inverso da freqüência do termo na coleção. As principais vantagens do modelo vetorial são a sua simplicidade, a facilidade que ele provê de se computar similaridades com eficiência e o fato de que o modelo se comporta bem com coleções genéricas.

O modelo de espaço vetorial representa consultas e documentos como vetores em um espaço t-dimensional. Cada dimensão desse espaço é associado com um dos termos da coleção. Para isso, associa-se a cada termo k_i um vetor k_j. Esses vetores de termos são considerados como ortogonais.

$$i \neq j \Longrightarrow \vec{k}_i \bullet \vec{k}_j = 0$$

Isso implica que se considera que os termos da coleção ocorrem de forma independente nos documentos e consultas. Além disso, o modelo de espaço vetorial assinala pesos positivos e não binários para termos nas consultas e documentos. Para calcular tais pesos, um método comum é tentar balancear a importância intra-documento dos

termos (relativo a outras palavras em um mesmo documento) com a importância inter-documento dos termos (relativa a outras palavras em outros documentos). Isso pode ser definido como segue.

Seja N o número total de documentos em uma coleção, n_i o número de documentos onde o termo k_i ocorre, e $freq_{ij}$ a freqüência do termo k_i no documento d_j. O fator $freq_{ij}$ quantifica a importância do termo k_i no documento d_j e é usualmente classificada como o fator de freqüência do termo (tf). O fator $\log \frac{N}{n_i}$ quantifica a importância do termo k_i como um fator de discriminação para toda a coleção de documentos e é conhecido como o fator de freqüência inversa do documento.

Uma estratégia popular para calcular o peso termo-documento w_{ij} é $w_{ij} = freq_{ij} \times \log \frac{N}{n_i}$ que é normalmente citada como esquema de peso $\mathsf{tf} - \mathsf{idf}$.

Para o peso termo-consulta w_{iq}, podemos adotar $w_{iq} = freq_{iq} \times \log \frac{N}{n_i}$ onde $freq_{iq}$ é a freqüência do termo k_i no texto associado com a consulta q.

A próxima seção tem o objetivo de descrever as métricas de RI utilizadas nos esperimentos e consequentemente, para avaliar o método desenvolvido neste trabalho.

2.3 Métricas de Avaliação em RI

Avalia-se um sistema de RI através da comparação das respostas geradas por este sistema e o conjunto ideal de respostas. Para isso, o vetor resultante é examinado e comparado com o conjunto ideal. Entre as diversas métricas disponíveis, esta seção descreve o funcionamento daquelas que foram usadas neste trabalho, justificando sua escolha.

2.3.1 Precisão

Seja N o conjunto de resposta ideal, ou seja, de documentos relevantes selecionados por usuários especialistas e R, o vetor resultado, indicando o conjunto de documentos recuperados pelo sistema de RI que foram examinados, então:

$$precisão = \frac{|\, N \cap R \,|}{|\, R \,|}$$

De acordo com Baeza-Yates and Ribeiro-Neto (1999), precisão é a medida de desempenho em RI que quantifica a fração de documentos recuperados que são reconhecidos como relevantes.

2.3.2 MRR

A média de rank reciproco (MRR) é uma estatística para avaliação de qualquer processo que produz uma lista de possíveis respostas a uma consulta, ordenados por probabilidade de acerto. O rank recíproco de uma resposta de consulta é o inverso multiplicativo da posição da primeira resposta correta de uma lista. MRR é a média das listas de resultados recíprocos para uma amostra de consultas Q.

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

onde:

- **rank i** é a posição da primeira resposta correta da lista i.
- **Q** é o número de consultas.

2.3.3 NDCG

Métricas baseadas em Ganho Cumulativo (CG) levam em consideração que existem documentos mais relevantes que outros para os usuários e que um documento relevante distante do topo tem pouca importância para a resposta.

De acordo com Järvelin and Kekäläinen (2002), o Ganho Acumulativo (CG) é a soma dos valores de relevância graduais de todos os resultados em uma lista de resultados de pesquisa. Este cálculo não considera o valor da posição do relevante. O CG em uma posição do rank **p** é definida como:

$$\mathrm{CG_p} = \sum_{i=1}^{p} rel_i$$

onde rel_i é a relevância gradual do resultado na posição i.

Segundo Järvelin and Kekäläinen (2002), o Ganho Cumulativo Descontado (DCG) inclui a ideia de que quanto mais distante do topo, menos importância tem um documento. Portanto, o valor da relevância gradual é reduzido logarítimicamente de forma proporcional à posição do resultado. O DCG acumulado em uma posição particular **p** do rank é dado por:

$$\mathrm{DCG_p} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

O trabalho de Järvelin and Kekäläinen (2002), expõe que para se obter o Ganho Acumulativo Descontado Normalizado (NDCG), é necessário dividir o DCG de cada sistema pelo DCG de um sistema ideal, obtendo valores entre 0 e 1 para cada posição do ranking. Para uma consulta, o NDCG é computado como:

$$\mathrm{NDCG_p} = \frac{DCG_p}{IDCGp}$$

No nosso trabalho foi utilizado o NDCG, o que se justifica por ser uma métrica que leva em consideração os diferentes graus de relevância, além de permitir avaliar a proximidade da lista recuperada real (DCG) da lista ideal (IDCG). Quanto mais próximos os índices melhor será avaliado o resultado. Um importante recurso do NDCG, também usado neste trabalho, é a possibilidade de calcular a média de todas as consultas e assim, avaliar o desempenho em média da máquina de busca, quando se leva em conta a posição e o grau de relevância dos resultados recuperados.

2.4 Contagem de Borda

No trabaho de Wu and McClean (2007), a Contagem de Borda é descrita da seguinte forma: para um conjunto fixo dos candidatos (x) e eleitores (y), cada um dos eleitores classifica estes candidatos em ordem de preferência. Para cada eleitor, ao candidato melhor classificado são dados n pontos, aos candidatos classificados em segundo lugar, são dados n-1 pontos, e assim por diante. Os candidatos são classificados por ordem de total de pontos de todos os eleitores, e o candidato com mais pontos vence a seleção. O objetivo do método, portanto, é o de eleger candidatos com a mais ampla aceitação possível, ao invés dos preferidos pela maioria. A Contagem de Borda tende a favorecer candidatos apoiados por um consenso geral entre os eleitores, ao invés de o candidato que é necessariamente o favorito de uma maioria. Por este motivo, pode ocorrer de o candidato escolhido não ser o preferido da maioria dos eleitores.

Para exemplificar o método, vamos supor uma eleição com 100 eleitores, cujos

resultados são exibidos na Tabela 2.1. O candidato Elias, por exemplo foi o primeiro na preferência de 17 eleitores, o segundo de 15 eleitores, o terceiro de 26 eleitores e o quarto na preferência de 42 eleitores. Em uma eleição majoritária, o resultado teria sido:

1. Primeiro (vencedor): João.
2. Segundo: Pedro.
3. Terceiro: Elias.
4. Quarto: Davi.

	PRIMEIRO	SEGUNDO	TERCEIRO	QUARTO
JOÃO	42	0	0	58
PEDRO	26	42	32	0
DAVI	15	43	42	0
ELIAS	17	15	26	42

Tabela 2.1: Exemplo de eleição majoritaria para 100 eleitores.

Porém, aplicando o método Contagem de Borda, mostrado na Tabela 2.2, os resultados se alteram.

	PRIMEIRO	SEGUNDO	TERCEIRO	QUARTO	PONTUAÇÃO
JOÃO	42x3	0x2	0x1	58x0	126
PEDRO	26x3	42x2	32x1	0x0	194
DAVI	15x3	43x2	42x1	0x0	173
ELIAS	17x3	15x2	26x1	42x0	107

Tabela 2.2: Exemplo do Contagem de Borda para 100 eleitores.

Pedro seria o vencedor, seguido por Davi em segundo, João em terceiro e Elias em quarto lugar.

Como no trabaho de Wu and McClean (2007), a votação de Contagem de Borda pode ser usada aqui para o propósito de avaliação se considerarmos sistemas de recuperação de informações como os candidatos, e os resultados obtidos para cada consulta como eleitores. Estes procedimentos de votação são úteis quando a classificação gerada a partir de todas as consultas as quais são confiáveis, mas a informação pontuação não é confiável ou não é disponível.

2.5 Trabalhos Relacionados

Nesta seção serão descritos e analisados estudos com características similares às do trabalho proposto, especialmente trabalhos com foco na busca por competências em instituições.

No trabalho de Hongli (2010), o Espaço Pessoal de Informação (PSI), que representa todo o conjunto de informações coletadas sobre um funcionario, é utilizado como recurso para a descoberta de tarefas pessoais e conhecimentos necessários correspondentes, que podem ser convertidos para evidências de competências individuais a fim de atualizar o modelo de competência de funcionários. A partir da análise das características dos PSI, é introduzido o método de modelagem de competências automáticas com base neste espaço. O método inclui cinco principais etapas que são: configurar o quadro de descrição de competências, decidir a arquitetura do sistema de modelagem de competência pessoal, extrair informações de experiências do espaço pessoal para construção do modelo de competência, agregar a competência pessoal e avaliar o peso da competência. Esse trabalho descreve um procedimento eficiente para a descoberta de competências, porém, o PSI não deve ser considerado o único ou o melhor ambiente para se encontrar competências dos funcionários, uma vez que seus

currículos são importante fonte de competências.

O trabalho descrito em Rodrigues et al. (2004) trata da dificuldade de instituições de pesquisa formarem comunidades virtuais por não conhecerem suas próprias competências. Os autores propõem uma técnica, para criar e sugerir comunidades Web científicas, com base nas competências dos cientistas. Primeiramente, a competência dos pesquisadores é determinada por mineração de texto em suas publicações científicas e então, é sugerida uma comunidade para o pesquisador se tornar membro. Embora este trabalho resolva o problema de formação de comunidades científicas, a extração de competências é feita sobre publicações científicas e não sobre os currículos Lattes dos pesquisadores que possuem informações muito mais abrangentes sobre as competências, uma vez que não se restringem às publicações científicas.

Em Zhu et al. (2005), é abordado o problema de descoberta de competências a partir de uma base de documentos usando um algoritmo de mineração de conteúdo Web não supervisionado chamado COmmunity Relation Discovery by named Entity Recognition (CORDER). A técnica de Reconhecimento de Entidades Mencionadas (NER), é usado como um passo preliminar para identificar entidades nomeadas (NES) de interesse, tais como nomes de pessoas, nomes de organização e áreas de conhecimento, e assim, resolver parcialmente o problema de tratar dados da Web não-estruturados. Esta abordagem é eficiente, pois utiliza um algoritmo de aprendizagem de máquina não-supervisionado. Entretanto, por utilizar documentos comuns da Web como fonte de dados para a descoberta de competências, não seria, em nossa opinião, a estratégia mais eficiente para descobrir e recomendar competências de pesquisadores. Isso é uma desvantagem pois a base de currículos Lattes tem mais informações sobre competências, além das publicações dos pesquisadores. A utilização da base de

currículos Lattes neste trabalho, pode enriquecer o processo de busca de competências.

No trabalho descrito em Harzallah et al. (2002), foi desenvolvido um sistema de seleção de profissionais para o preenchimento de vagas para emprego e para busca de ofertas de trabalho. A solução proposta foi o projeto Competency©ontology.cv (CommOnCV) que modela as competências dos candidatos a vagas por meio das informações contidas em um banco de currículos. O resultado obtido é a combinação das informações sobre competências extraídas dos currículos com as requeridas no trabalho ofertado. As ideias defendidas no projeto foram implementadas no Curriculum Vitae GENerator (website CvGen). O sucesso do gerenciamento de competências defendido nesse trabalho é baseado no uso de técnicas atualmente desenvolvidas no contexto da área de Web semântica. Porém esse trabalho não trata especificamente da modelagem de competências de pesquisadores, mas sim, de competências genéricas.

Por fim, em e Cesar Junior Roberto Marcondes Mena-Chalco (2009), é feita a descrição do ScriptLattes, um sistema de extração de dados de currículos da Plataforma Lattes. O problema tratado nesse trabalho é a elaboração de relatórios sobre produção científica, supervisões e projetos dos grupos de pesquisa relacionados com instituições de pesquisa, assim como a avaliação de programas de pós-graduação no Brasil a partir da base de currículos Lattes. Essas instituições podem gerar seus relatórios por meio de análises manuais dos dados do currículo Lattes de cada membro do grupo, com o objetivo de obter um completo resumo de todas as produções científicas, supervisões e projetos do grupo. Embora útil, os resultados do scriptLattes não podem ser utilizados diretamente para a recomendação de competências, problema que é o foco de nosso trabalho.

A maioria dos trabalhos relacionados nesta seção destacam-se por utilizarem téc-

nicas inovadoras para a busca e descoberta de competências, entretanto, não utilizam os currículos Lattes ou algo semelhante como fonte de busca de competências. O uso de currículos Lattes para a busca de competências é o diferencial que consideramos fundamental do sucesso deste trabalho. Abordamos a busca de competências de pesquisadores que possuem seus currículos na Plataforma Lattes, pelo fato desta plataforma ser considerada a melhor fonte de busca de competências para instituições acadêmico-científicas. Os campos selecionados dos currículos Lattes para a busca de competências e a forma que eles foram escolhidos são descritos no Capítulo 3.

3 *Busca de competências a Partir de currículos Lattes*

Este capítulo tem como objetivo descrever a execução do projeto desde a escolha dos campos onde se encontram as competências dos pesquisadores à obtenção do Banco de Dados de currículos Lattes, e por fim, à implementação das estratégias do método de busca. A idéia geral e conceitual por trás deste processo é exibida na Figura 3.1. Inicialmente, o usuário define um perfil. Em seguida, é feita a seleção dos currículos Lattes mais similares ao perfil buscado. Os currículos selecionados são, então, ordenados por relevância e exibidos para o usuário.

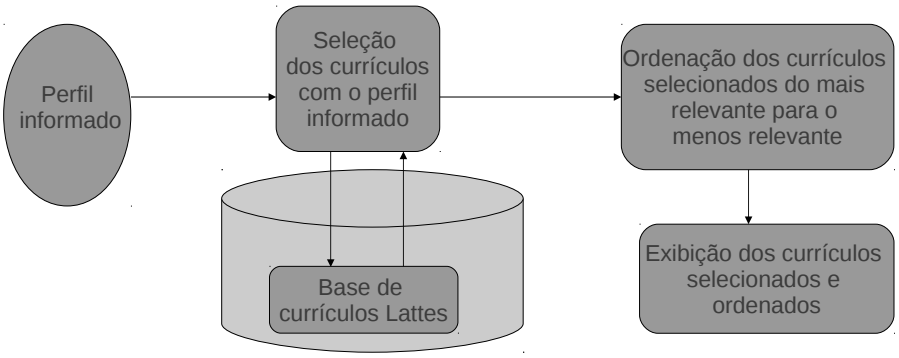


Figura 3.1: Síntese do método de busca de competências a partir de currículos Lattes executado neste trabalho.

3.1 Seleção das Competências

O primeiro problema abordado por este projeto foi descobrir, as características ou áreas de um currículo Lattes que seriam mais importantes para determinar as competências, e a quais competências elas corresponderiam. Após análise dos campos disponíveis nos currículos Lattes, que estão exibidos na figura 3.2, chegou-se à distribuição de campos exibida a seguir. É importante destacar que essa distribuição foi puramente empírica, para cada competência.

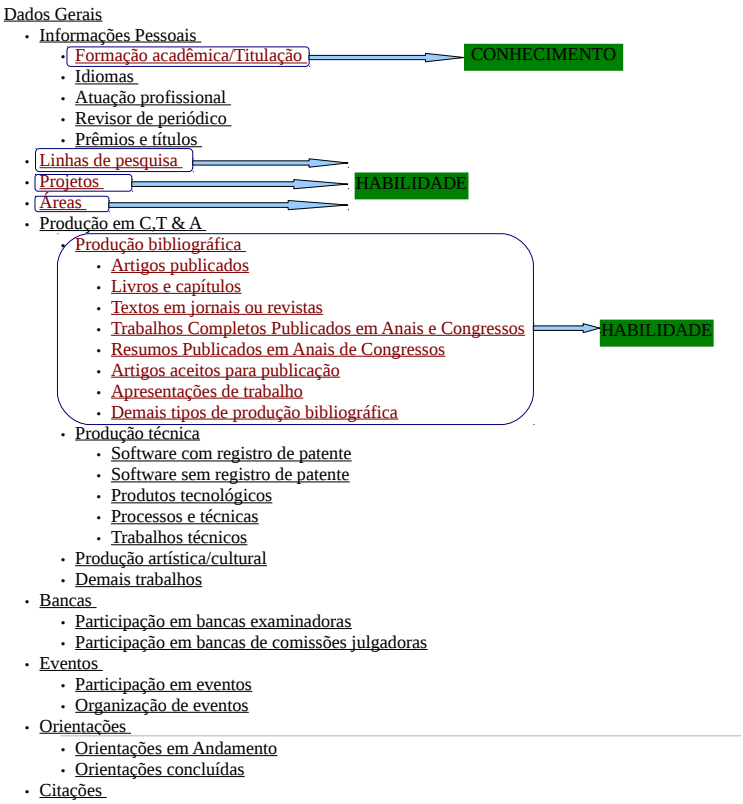


Figura 3.2: Lista de campos disponíveis nos currículos Lattes com seleções destacadas que indicam as competências relacionadas aos campos.

1. **Conhecimento:** Formação do pesquisador.

2. **Habilidade ou experiência:** Publicações Científicas, Projetos de Pesquisa, Linhas de Pesquisa e Área de Atuação.

3.2 Base de currículos Lattes usada neste trabalho

O passo seguinte foi a obtenção de uma base de dados para a realização de testes, e posteriormente dos experimentos. Como solução para este problema foi utilizada a base de currículos Lattes de professores da UFAM. A atividade de obtenção desta base envolveu a aquisição de parte dos currículos dos professores da UFAM disponíveis na base do CNPq.

Após esta etapa, foram feitos alguns ajustes necessários de preparação da base de dados. Estes ajustes tiveram como objetivo a exclusão de áreas dos currículos que não correspondiam às áreas das competências selecionadas na seção anterior. Ao fim, desta etapa, gerou-se um Banco de currículos com um total de 936 pesquisadores com suas características distribuídas de acordo com a Tabela 4.3 mostrada no Capítulo 4.

3.3 O Método de Busca de Competências

Esta seção tem como objetivo descrever as etapas de funcionamento do Método de Busca de Competências e as diferenças entre os algoritmos que implementam as estratégias de recuperação dos currículos Lattes. Estes algoritmos serão chamados a partir deste momento de: Soma das Similaridades, Produção e Contagem de Borda.

3.3.1 Soma das Similaridades

Inicialmente, deve-se digitar um trecho de texto contendo uma ou mais competências a serem pesquisadas pela máquina de busca. O vetor resultado, indicando os currículos Lattes recuperados para uma consulta, é montado através do cálculo da similaridade para cada campo que contém as competências dos pesquisadores (Formação, Projetos de Pesquisa, linhas de Pesquisa, Áreas de Atuação e Produções Científicas). Para o algoritmo Soma das Similaridades, são considerados vencedores os currículos Lattes com a maior pontuação na soma das similaridades de seus campos. Este processo é exibido de forma resumida na Figura 3.3.

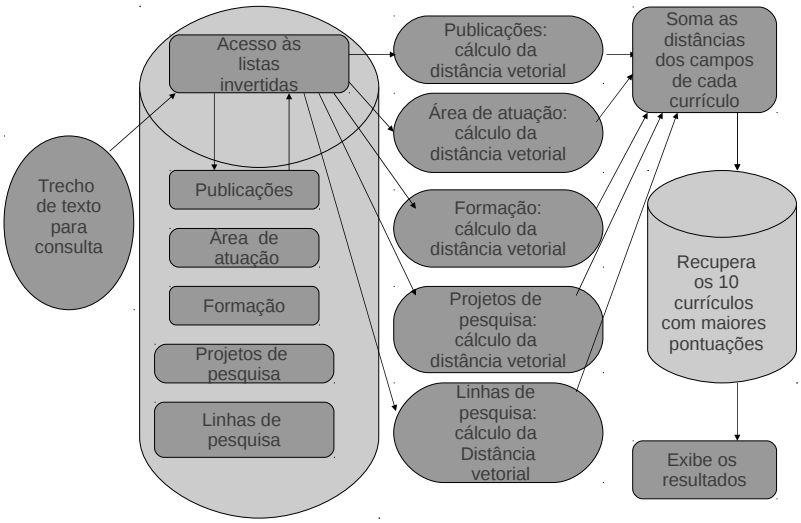


Figura 3.3: Algoritmo Soma das Similaridades.

3.3.2 Produção

No caso do algoritmo Produção, o vetor resultado é obtido através do cálculo da similaridade feita somente no campo de Produções Científicas, onde os currículos com maior similaridade são os vencedores. Este processo é exibido de forma resumida na Figura 3.4.

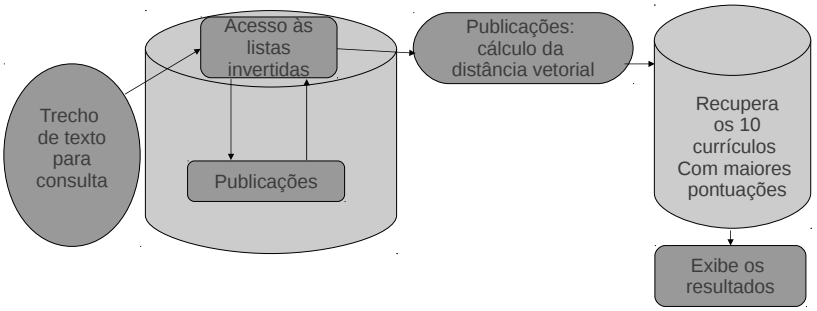


Figura 3.4: Algoritmo Produção.

3.3.3 Contagem de Borda

Por fim, para o algoritmo Contagem de Borda também são considerados todos os campos que correspondem a competências, porém, após o cálculo das similaridades, cada campo de currículo candidato (os 10 currículos com melhor classificação em cada campo), recebe uma pontuação que varia de 10 (maior similaridade) até 1 (menor similaridade). Os currículos vencedores são aqueles com a maior pontuação na soma de cada campo. Por fim, como saída, o algoritmo recupera os 10 currículos mais pontuados por meio de sua estratégia adotada. Este processo é exibido de forma resumida na Figura 3.5.

A próxima etapa deste trabalho foi a preparação de um experimento para avaliar o desempenho de cada algoritmo por meio das métricas de RI, precisão, MRR e NDCG. Este processo é descrito no próximo capítulo.

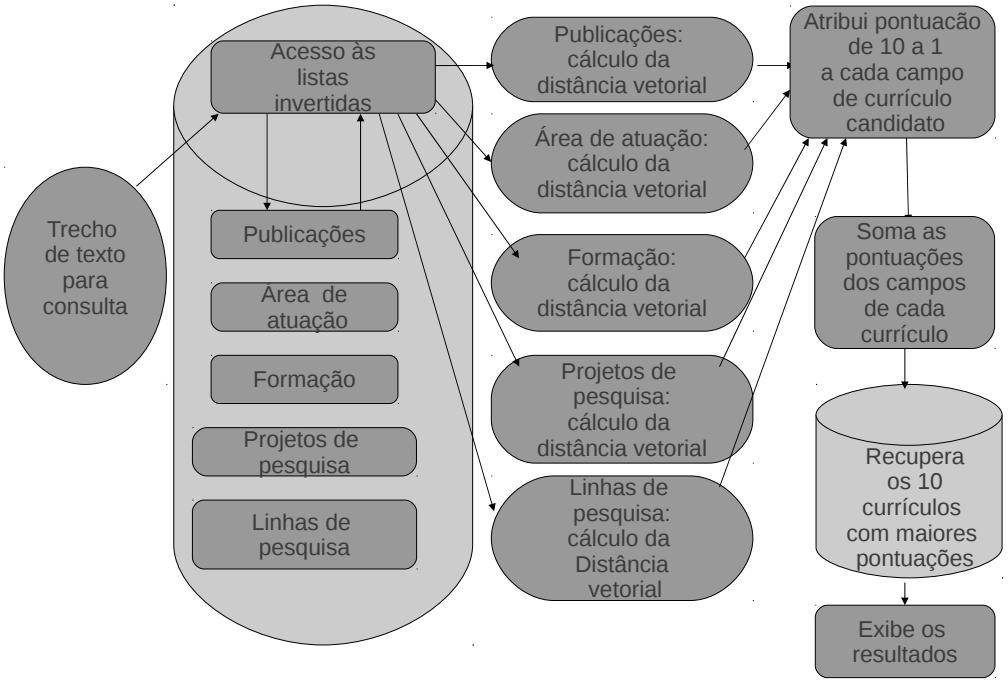


Figura 3.5: Algoritmo Contagem de Borda.

4 Avaliação Experimental

Neste capítulo, apresentamos os experimentos que permitiram avaliar a eficácia dos três algoritmos propostos em termos da relevância de seus resultados. Portanto, são apresentados, os preparativos para o experimento, a execução do experimento propriamente dito, e a aplicação das métricas de avaliação em RI: precisão, MRR e NDCG. O experimento consiste em submeter um conjunto de consultas a cada um dos algoritmos e então solicitar que um grupo de usuários especializados em avaliar competências classifiquem os resultados recuperados por cada algoritmo em: relevante, médio relevante e pouco relevante. A próxima seção descreve a configuração dos experientos, ou seja, como foram feitos os preparativos para que os experimentos fossem executados.

4.1 Protocolo Experimental

Como preparativo para o experimento, solicitamos que sete usuários com experiência no processo de busca de competências em currículos Lattes, criassem 14 pesfis para serem submetidos aos algoritmos de busca utilizando termos que descrevessem a competência desejada em cada perfil. Os perfis foram numerados de um a 14. Na Tabela 4.1 apresentamos os perfis usados.

NÚMERO	PERFIL
01	Pesquisador de Geologia da Amazônia Ocidental
02	Mestre em Matemática, com doutorado em Física
03	Mestre em Propriedade Intelectual, doutorado em Biotecnologia com ênfase em marcadores moleculares
04	Doutor em Antropologia Cultural
05	Graduado em Economia, Mestrado em Desenvolvimento Regional
06	Graduado em Engenharia de Pesca, Mestrado em Ciências Pesqueiras nos Trópicos
07	Bacharel Direito, Especialista em Direito Ambiental
08	Bacharel em Historia com Mestrado e atuação na area de Historia Colonial Amazônica
09	Doutor em Agronomia palmeiras amazônicas açaízeiro
10	Pós-doutorado em Engenharia da Produção
11	Doutor em Física propriedades de sólidos
12	Doutorado em Linguística e Antropologia, linha de pesquisa línguas indígenas
13	Bacharelado em Matemática Aplicada
14	Doutorado em Física, com ênfase em captação de CO2

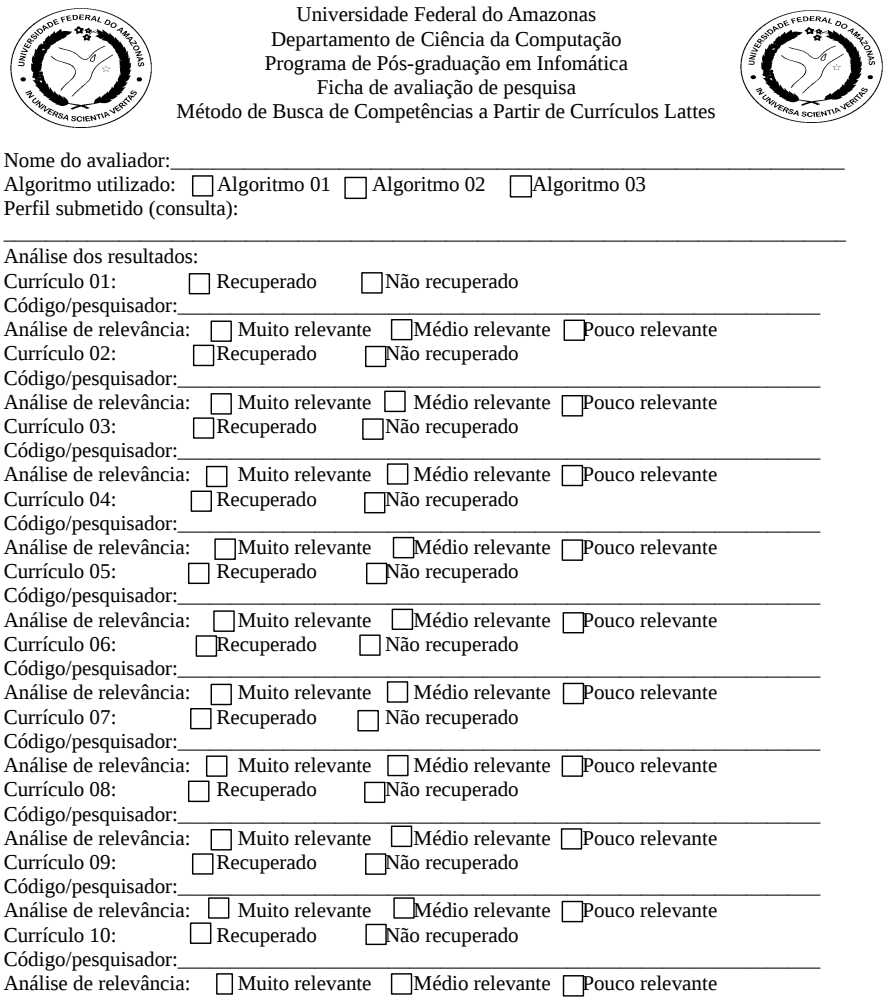
Tabela 4.1: Lista de perfis definidos por usuários com experiência em busca de competências.

Em seguida, foi criada também uma ficha, apresentada na Figura 4.1, onde foram registrados o perfil submetido, o algoritmo utilizado, os dez primeiros resultados (currículos de pesquisadores) recuperados pelo algoritmo, e seus respectivos graus de relevância, os quais seriam classificados de acordo com os seguintes critérios:

1. Muito relevante: currículo recuperado pelo algoritmo onde todas as competências informadas no perfil foram encontradas.

2. Médio relevante: currículo recuperado pelo algoritmo onde somente uma parte das competências informadas no perfil foram encontradas.

3. Pouco relevante: currículo recuperado pelo algoritmo onde nenhuma das competências informadas no perfil foi encontrada.



Universidade Federal do Amazonas
Departamento de Ciência da Computação
Programa de Pós-graduação em Infomática
Ficha de avaliação de pesquisa
Método de Busca de Competências a Partir de Currículos Lattes

Nome do avaliador:_____
Algoritmo utilizado: ☐ Algoritmo 01 ☐ Algoritmo 02 ☐ Algoritmo 03
Perfil submetido (consulta):
_____
Análise dos resultados:
Currículo 01: ☐ Recuperado ☐ Não recuperado
Código/pesquisador:_____
Análise de relevância: ☐ Muito relevante ☐ Médio relevante ☐ Pouco relevante
Currículo 02: ☐ Recuperado ☐ Não recuperado
Código/pesquisador:_____
Análise de relevância: ☐ Muito relevante ☐ Médio relevante ☐ Pouco relevante
Currículo 03: ☐ Recuperado ☐ Não recuperado
Código/pesquisador:_____
Análise de relevância: ☐ Muito relevante ☐ Médio relevante ☐ Pouco relevante
Currículo 04: ☐ Recuperado ☐ Não recuperado
Código/pesquisador:_____
Análise de relevância: ☐ Muito relevante ☐ Médio relevante ☐ Pouco relevante
Currículo 05: ☐ Recuperado ☐ Não recuperado
Código/pesquisador:_____
Análise de relevância: ☐ Muito relevante ☐ Médio relevante ☐ Pouco relevante
Currículo 06: ☐ Recuperado ☐ Não recuperado
Código/pesquisador:_____
Análise de relevância: ☐ Muito relevante ☐ Médio relevante ☐ Pouco relevante
Currículo 07: ☐ Recuperado ☐ Não recuperado
Código/pesquisador:_____
Análise de relevância: ☐ Muito relevante ☐ Médio relevante ☐ Pouco relevante
Currículo 08: ☐ Recuperado ☐ Não recuperado
Código/pesquisador:_____
Análise de relevância: ☐ Muito relevante ☐ Médio relevante ☐ Pouco relevante
Currículo 09: ☐ Recuperado ☐ Não recuperado
Código/pesquisador:_____
Análise de relevância: ☐ Muito relevante ☐ Médio relevante ☐ Pouco relevante
Currículo 10: ☐ Recuperado ☐ Não recuperado
Código/pesquisador:_____
Análise de relevância: ☐ Muito relevante ☐ Médio relevante ☐ Pouco relevante

Figura 4.1: Ficha de Avaliação usada pelos avaliadores.

Os sete usuários que definiram os perfis foram também encarregados de avaliar os resultados dos experimentos. Cada avaliador ficou responsável por seis consultas, duas para cada algoritmo. Esta organização está exibida na Tabela 4.2. Tomou-se o

cuidado de evitar que cada avaliador ficasse responsável por avaliar um mesmo perfil em algoritmos diferentes. Os avaliadores não tiveram acesso ao nome de cada algoritmo de busca, os quais foram identificados unicamente pelos nomes de Algoritmo1, Algoritmo2 e Algoritmo3, que correspondiam à Soma de Similaridades, Produção e Contagem de Borda, respectivamente.

Por fim, os testes foram executados. Cada avaliador submeteu seus perfis ao algoritmo de busca e avaliou os resultados registrando-os na ficha de avaliação.

	Soma de Similaridades	Produção	Contagem de Borda
PERFIL 01	AVALIADOR 01	AVALIADOR 07	AVALIADOR 06
PERFIL 02	AVALIADOR 01	AVALIADOR 07	AVALIADOR 06
PERFIL 03	AVALIADOR 02	AVALIADOR 01	AVALIADOR 07
PERFIL 04	AVALIADOR 02	AVALIADOR 01	AVALIADOR 07
PERFIL 05	AVALIADOR 03	AVALIADOR 02	AVALIADOR 01
PERFIL 06	AVALIADOR 03	AVALIADOR 02	AVALIADOR 01
PERFIL 07	AVALIADOR 04	AVALIADOR 03	AVALIADOR 02
PERFIL 08	AVALIADOR 04	AVALIADOR 03	AVALIADOR 02
PERFIL 09	AVALIADOR 05	AVALIADOR 04	AVALIADOR 03
PERFIL 10	AVALIADOR 05	AVALIADOR 04	AVALIADOR 03
PERFIL 11	AVALIADOR 06	AVALIADOR 05	AVALIADOR 04
PERFIL 12	AVALIADOR 06	AVALIADOR 05	AVALIADOR 04
PERFIL 13	AVALIADOR 07	AVALIADOR 06	AVALIADOR 05
PERFIL 14	AVALIADOR 07	AVALIADOR 06	AVALIADOR 05

Tabela 4.2: Organização de avaliadores e seus algoritmos para o experimento.

A base de dados utilizada para a realização das consultas do experimento foi a de currículos Lattes de professores da UFAM que foi copiada de um Banco de Dados já existente no CPD da UFAM, e que foi previamente extraída da base de dados do CNPq. Esta escolha se deu pelo fato de que a extração dos dados de todos os currículos, diretamente da base do CNPq (que estão disponíveis *online*) iria demandar um tempo razoável e desnecessário para os objetivos do projeto proposto, conforme mencionado

anteriormente. A base de dados do CNPq é atualmente considerada um padrão nacional, representando um histórico das atividades científicas, acadêmicas e profissionais de pesquisadores cadastrados. Portanto, a base de currículos Lattes de professores da UFAM foi usada como fonte para identificação de competências de pesquisadores que possuam currículos nesta plataforma.

Após a etapa de obtenção da base de currículos Lattes de professores da UFAM, foram feitos alguns ajustes necessários de preparação da base de dados. Estes ajustes tiveram como objetivo a exclusão de áreas dos currículos que não representavam áreas de competências relevantes para a realização do experimento, como informações pessoais e atuação profissional. Em seguida, foram excluídos currículos com informações inconsistentes, por exemplo, currículos com Formação e Produções Científicas que não tinham relação com nenhum pesquisador. Ao fim desta etapa gerou-se um banco de currículos de um total de 936 pesquisadores com suas características distribuídas de acordo com a Tabela 4.3.

CAMPO	REGISTROS
Pesquisador	936
Formação	6.213
Área de atuação	2.117
Projetos de Pesquisa	3.393
Linhas de Pesquisa	2.003
Produções Científicas	89.801

Tabela 4.3: Informações relevantes sobre a Base de currículos Lattes de Pesquisadores da UFAM.

4.2 Avaliação

As próximas seções terão por objetivo descrever que procedimentos foram usados, para tabular e classificar os resultados do primeiro experimento, concluindo com uma análise destes resultados na seção 4.3.4. Esta análise motivou a realização de um segundo experimento.

Os resultados são apresentados na seguinte sequência:

1. **Resultados preliminares top 10:** São considerados os 10 primeiros currículos recuperados, de acordo com a execução do experimento.

2. **Resultados top 10 sem titulação:** São considerados os 10 primeiros currículos recuperados, de acordo com a execução do experimento, porém, excluindo-se a titulação dos perfis.

3. **Resultados top 5 sem titulação:** São considerados os 5 primeiros currículos recuperados, ou seja, utilizando o quinto currículo como limiar de corte, e também excluindo-se a titulação dos perfis.

4.2.1 Resultados preliminares top 10

Os resultados obtidos foram tabulados em uma planilha onde, para a avaliação dos três algoritmos, foram aplicadas as métricas avaliação da precisão, MRR e NDCG. Inicialmente as métricas foram aplicadas nos resultados obidos para os dez currículos mais relevantes de cada algoritmo. Os resultados são exibidos nas Tabelas 4.4 e 4.5.

Resultados de precisão e NDCG preliminares top 10						
	Soma		**Produção**		**Borda**	
PERFIL	PRECISÃO	NDCG	PRECISÃO	NDCG	PRECISÃO	NDCG
01	0,3	0,35	0,2	0,33	0,5	1,00
02	0,1	0,43	0,2	0,68	0,1	0,63
03	0,4	0,96	0,2	0,57	0,2	0,66
04	0,5	0,80	0,3	0,74	0,5	0,89
05	0,5	0,91	0,4	0,87	0,7	0,84
06	0,3	0,89	0,5	0,86	0,6	0,87
07	0,3	0,41	0,2	0,44	0,3	0,45
08	0,4	0,53	0,2	0,54	0,4	0,89
09	0,3	0,81	0,0	0,00	0,3	0,44
10	0,3	0,66	0,1	0,50	0,2	0,76
11	0,1	1,00	0,0	0,00	0,3	0,36
12	0,6	0,82	0,2	0,67	0,3	0,82
13	0,5	0,96	0,4	0,92	0,4	0,56
14	0,2	0,72	0,1	1,00	0,2	0,43
MÉDIA	**0,34**	**0,73**	**0,21**	**0,68**	**0,36**	**0,69**

Tabela 4.4: Precisão e NDCG preliminares top 10.

Resultados preliminares MRR top 10			
	Soma	**Produção**	**Borda**
MRR	**0,68**	**0,52**	**0,61**

Tabela 4.5: Mrr preliminar top 10 .

Após uma observação inicial dos resultados, verificou-se que os mesmos não seriam suficientes para serem tomados como prova do sucesso do método de busca desenvolvido, uma vez que os resultados de precisão ficaram muito abaixo do esperado para os três algoritmos testados, o que tornaria o método de busca ineficaz para o uso pela maioria dos usuários. Com uma observação mais detalhada dos perfis elaborados pelos usuários especialistas, um padrão bastante repetitivo foi detectado: a utilização de titulação em todos os perfis (Mestre, Doutor, Especialista e etc). Como a titulação

dos pesquisadores não representa competências, levantou-se a hipótese de que a exclusão da titulação dos perfis elaborados poderia alterar de forma positiva os resultados obtidos. A exclusão da titulação de todos os perfis de fato foi feita e os resultados são mostrados na próxima seção.

4.2.2 Resultados top 10 sem titulação

Para verificar a válidade da hipótese descrita na seção anterior, foram corrigidas as competências dos perfis elaborados pelos usuários especialistas, e utilizou-se um novo procedimento para avaliar a relevância dos currículos recuperados. Este novo procedimento consistiu em:

1. Retirar dos perfis palavras que indiquem titulação (Mestre, Doutor ou especialista).
2. Refazer o processo de classificação de currículos relevantes. Esta nova classificação foi feita por um avaliador não especializado.
3. Tabular os dados.
4. Avaliar novamente os algoritmos usando as métricas de precisão, MRR e NDCG.
5. Comparar os novos resultados com os dados obtidos anteriormente.

Os resultados destes experimentos estão exibidos nas Tabelas 4.6 e 4.7.

Resultados de precisão e NDCG top 10 sem titulação						
	Soma		**Produção**		**Borda**	
PERFIL	PRECISÃO	NDCG	PRECISÃO	NDCG	PRECISÃO	NDCG
01	0,3	0,34	0,3	0,34	0,6	0,89
02	0,2	0,72	0,2	0,72	0,5	0,43
03	0,7	0,98	0,7	0,98	0,8	0,85
04	1,0	0,99	1,0	0,99	0,8	0,91
05	0,9	1,00	0,9	1,00	0,9	0,91
06	0,6	0,92	0,6	0,92	0,7	0,87
07	0,8	0,83	0,8	0,83	0,4	0,65
08	0,9	0.90	0,9	0.90	0,9	0,95
09	0,5	0,84	0,5	0,84	0,5	0,91
10	0,9	0,97	0,9	0,97	0,9	0,96
11	0,1	1,00	0,1	1,00	0,3	0,35
12	0,8	0,73	0,8	0,73	0,7	0,92
13	0,7	0,90	0,7	0,90	0,9	0,90
14	0,6	0,80	0,6	0,80	0,6	0,98
MÉDIA	**0,64**	**0,85**	**0,64**	**0,85**	**0,68**	**0,82**

Tabela 4.6: Precisão e NDCG top 10 sem titulação.

Resultados MRR top 10 sem titulação			
	Soma	**Produção**	**Borda**
MRR	**0,87**	**0,87**	**0,84**

Tabela 4.7: Mrr top 10 sem titulação.

A observação de que os currículos mais relevantes concentram-se entre os cinco primeiros recuperados, motivou a criação de um procedimento para a reavaliação dos resultados considerando-se somente os cinco primeiros currículos recuperados. Este procedimento é descrito na seção a seguir.

4.2.3 Resultados top 5 sem titulação

Por fim, com o objetivo de melhorar a análise dos resultados obtidos na avaliação anterior, uma nova estatégia, listada abaixo, foi utilizada:

1. Reaproveitar a última classificação de relevantes, considerando-se o limiar de corte no quinto currículo recuperado.
2. Avaliar novamente o resultado usando precisão, MRR e NDCG.
3. Comparar os novos resultados, com os dados obtidos na avaliação anterior.

Os resultados são exibidos nas Tabelas 4.8 e 4.9.

Resultados de precisão e NDCG top 5 sem titulação						
	Soma		**Produção**		**Borda**	
PERFIL	PRECISÃO	NDCG	PRECISÃO	NDCG	PRECISÃO	NDCG
01	0,0	0,00	0,0	0,00	0,5	0,89
02	0,4	0,72	0,4	0,72	0,0	0,00
03	0,8	1,00	0,8	1,00	0,4	1,00
04	1,0	1,00	1,0	1,00	0,4	0,94
05	1,0	1,00	1,0	1,00	0,5	0,91
06	0,8	0,94	0,8	0,94	0,4	0,93
07	0,8	0,86	0,8	0,86	0,3	0,65
08	1,0	1,00	1,0	1,00	0,5	1,00
09	1,0	0,84	1,0	0,84	0,4	0,93
10	1,0	0,98	1,0	0,98	0,4	0,98
11	0,2	1,00	0,2	1,00	0,0	0,00
12	0,8	0,72	0,8	0,72	0,3	0,97
13	0,8	0,91	0,8	0,91	0,4	0,93
14	0,8	0,79	0,8	0,79	0,4	0,99
MÉDIA	**0,74**	**0,90**	**0,74**	**0,90**	**0,70**	**0.93**

Tabela 4.8: Precisão e NDCG top 5 sem titulação.

Resultados MRR top 5 sem titulação			
	Soma	**Produção**	**Borda**
MRR	**0,86**	**0,86**	**0,82**

Tabela 4.9: Mrr top 5 sem titulação.

4.2.4 Análise dos Resultados

Após o cálculo da precisão nos resultados iniciais verificou-se em média os resultados de 0,36 para Contagem de Borda 0,34 para a Soma das Similaridades e 0,21 para Produção. Embora com resultados similares para Contagem de Borda e para Soma das Similaridades, observou-se individualmente uma alternâncias de resultados entre um e outro. De um modo geral, nesta etapa nenhum dos algoritmos obteve resultados significativos.

A hipótese levantada para justificar este resultado inexpressivo, seria de que todos os perfis informados consideravam fortemente o título como competência. Porém, como a titulação aparece na composição do campo Formação dos currículos Lattes, isto prejudicou a eficiência da máquina de busca para os três algoritmos, uma vez que, para Soma das Similaridades e Contagem de Borda são considerados, além da Formação, Produções Científicas, Área de atuação, Linhas de Pesquisa e Projetos de Pesquisa. O algoritmo Produção, com os piores resultados, só considera as Produções Científicas. Alguns casos patológicos também foram identificados, como por exemplo, consultas que buscam competências na área de Física traziam Currículos na área de Educação Física.

Após desconsiderar a titulação dos perfis e refazer a avaliação, conseguimos resultados mais expressivos. O valor de precisão nos algoritmos Soma das Similaridades e

Produção aumentou para 0,64 e o vencedor Contagem de Borda ficou em 0,68. Porém, os algoritmos Soma das Similaridades e Produção têm melhores resultados de MRR e NDCG. Estes resultados levaram à conclusão que o Contagem de Borda recupera mais currículos relevantes entre os dez primeiros recuperados do que os outros dois métodos, porém, os outros algoritmos concentram seus relevantes mais próximos do topo, o que justificou uma avaliação somente dos cinco primeiros currículos recuperados.

Após considerar os cinco primeiros recuperados por cada algoritmo, os resultados da precisão melhoraram, sendo Soma das Similaridades e Produção os melhores com 0,74 e por último, Contagem de Borda com 0,68. Estes resultados reforçam a observação já feita anteriormente de que os algoritmos Soma das Similaridades e Produção, concentram seus resultados mais relevantes mais próximos do topo do que o Contagem de Borda. Os resultados são sintetizados nas Tabelas 4.10, 4.11 e 4.12.

	Soma	Produção	Contagem de Borda
PRECISÃO PRELIMINAR TOP 10	0,34	0,21	0,36
PRECISÃO TOP 10 SEM TITULAÇÃO	0,64	0,64	0,68
PRECISÃO TOP 5 SEM TITULAÇÃO	0,74	0,74	0,70

Tabela 4.10: Resultados de precisão em média do primero experimento.

	Soma	Produção	Contagem de Borda
NDCG PRELIMINAR TOP 10	0,73	0,68	0,69
NDCG TOP 10 SEM TITULAÇÃO	0,85	0,85	0,82
NDCG TOP 5 SEM TITULAÇÃO	0,90	0,90	0,93

Tabela 4.11: Resultados de NDCG em média do primeiro experimento.

	Soma	Produção	Contagem de Borda
MRR PRELIMINAR TOP 10	0,68	0,52	0,61
MRR TOP 10 SEM TITULAÇÃO	0,87	0,87	0,84
MRR TOP 5 SEM TITULAÇÃO	0,86	0,86	0,82

Tabela 4.12: Resultados de Mrr do primeiro experimento.

Por fim, os resultados da avaliação anterior, que consideraram os perfis sem a titulação e o limiar de corte no quinto currículo recuperado, mostraram-se muito mais expressivos em relação aos primeiros. Esta observação nos encorajou a repetir o experimento com os usuários especialistas, sendo que desta vez, estes teriam o cuidado de elaborar novos perfis sem a titulação. Os detalhes deste novo experimento são exibidos na seção a seguir.

4.3 Reavaliação

Como preparação para este novo experimento os usuários especialistas tiveram o cuidado de elaborar novos perfis sem a titulação e tendo como foco as competências a serem pesquisadas nos currículos. Os novos perfis elaborados estão listados na Tabela 4.13.

NÚMERO	PERFIL
01	Linguística Indígena
02	Divulgação Científica
03	História da Amazônia
04	Comunicação – Comunicação Empresarial
05	Arte Comtemporânea e Dança
06	Semiótica da Cultura Peirciana
07	Biologia Celular
08	Matemática Aplicada
09	Desenvolvimento Regional e Economia
10	Engenharia e Engenharia Mecânica
11	Agronomia
12	Serviço Social Indígena
13	Comunicação Cultura Indígena
14	Direito, Patentes e Eletrônica

Tabela 4.13: Lista de novos perfis elaborados pelos usuários especialistas.

As próximas seções terão por objetivo descrever que procedimentos foram usados, para tabular e classificar os resultados do segundo experimento, concluindo com uma análise destes resultados.

Os resultados são apresentados na seguinte sequência:

1. **Resultados top 10 do segundo experimento:** Considera os 10 primeiros currículos recuperados, de acordo com a execução do experimento.

2. **Resultados top 5 do segundo experimento:** Considera os 5 primeiros currículos recuperados, ou seja, utilizando o quinto currículo como limiar de corte.

4.3.1 Resultados top 10 do segundo experimento

Os resultados obtidos foram tabulados em uma planilha onde, para a avaliação dos três algoritmos, foram aplicadas as métricas de avaliação de precisão, MRR e NDCG. Inicialmente, as métricas foram aplicadas nos resultados obidos para os dez currículos mais relevantes de cada algoritmo. Os resultados são exibidos nas Tabelas 4.14 e 4.15.

Resultados de precisão e NDCG top 10						
	Soma		**Produção**		**Borda**	
PERFIL	PRECISÃO	NDCG	PRECISÃO	NDCG	PRECISÃO	NDCG
01	0,5	0,98	0,9	0,99	0,5	0,76
02	0,3	0,55	0,4	0,55	0,0	1,00
03	0,2	0,57	0,5	0,81	0,5	0,92
04	0,9	0,95	0,7	0,74	0,4	0,79
05	0,2	0,75	0,6	0,96	0,2	0,67
06	0,0	1,00	0,3	0,37	0,2	0,39
07	1,0	0,83	0,5	0,53	0,6	0,83
08	0,5	0,88	0,5	0,93	0,3	0,67
09	0,6	0,97	0,9	0,92	0,7	1,00
10	0,8	0,78	0,4	0,45	0,4	0,95
11	0,9	0,99	0,9	0,99	0,8	0,96
12	0,5	0,79	1,0	0,97	0,8	0,89
13	1,0	0,91	0,7	0,89	1,0	0,99
14	0,1	0,50	0,1	0,33	0,7	0,73
MÉDIA	**0,54**	**0,82**	**0,60**	**0,71**	**0,51**	**0,82**

Tabela 4.14: Precisão e NDCG top 10 do segundo experimento.

Resultados MRR top 10			
	Soma	**Produção**	**Borda**
MRR	**0,81**	**0,70**	**0,81**

Tabela 4.15: Mrr top 10 do segundo experimento.

4.3.2 Resultados top 5 do segundo experimento

Por fim, com o objetivo de melhorar os resultados obtidos na avaliação anterior, uma nova estatégia, listada abaixo, foi utilizada.

1. Reaproveitar a última classificação de relevantes, considerando-se o limiar de corte no quinto currículo recuperado.

2. Avaliar novamente usando precisão, MRR e NDCG.

3. Comparar os novos resultados, com os dados obtidos anteriormente, onde foram considerados os dez primeiros resultados.

Os resultados são exibidos nas tabelas 4.16 e 4.17.

Resultados de precisão e NDCG top 5						
	Soma		**Produção**		**Borda**	
PERFIL	PRECISÃO	NDCG	PRECISÃO	NDCG	PRECISÃO	NDCG
01	0,8	1,00	1,0	1,00	0,6	0,78
02	0,2	1,00	0,2	1,00	0,0	1,00
03	0,2	1,00	0,6	0,81	1,0	0,92
04	1,0	1,00	0,8	0,82	0,6	0,81
05	0,4	0,75	0,8	1,00	0,2	1,00
06	0,0	1,00	0,0	1,00	0,2	0,43
07	1,0	0,83	0,2	0,63	0,8	0,82
08	0,8	1,00	0,8	0,93	0,4	0,72
09	0,8	0,99	1,0	0,92	1,0	1,00
10	0,6	0,73	0,2	0,50	0,6	1,00
11	1,0	1,00	1,0	1,00	1,0	0,99
12	0,6	0,81	1,0	0,99	1,0	1,00
13	1,0	0,91	1,0	0,89	1,0	1,00
14	0,2	0,50	0,0	1,00	0,6	0,84
MÉDIA	**0,61**	**0,89**	**0,61**	**0,87**	**0,64**	**0,87**

Tabela 4.16: Precisão e NDCG top 5 do segundo experimento.

Resultados MRR top 5			
	Soma	**Produção**	**Borda**
MRR	**0,81**	**0,70**	**0,81**

Tabela 4.17: Mrr top 5 do segundo experimento.

4.3.3 Análise dos Resultados

Os resultados do segundo experimento mostraram-se significativamente superiores aos resultados do primeiro experimento, o que já era esperado, uma vez que neste último não foi usada a titulação dos pesquisadores nos perfis criados pelos usuários especialistas.

Considerando os resultados com limiar de corte no décimo currículo recuperado, o algoritmo de Produções teve o melhor resultado usando a métrica de precisão, o que demonstra que esse algoritmo possui maior eficiência em recuperar mais currículos relevantes do que os outros algoritmos. Embora os outros algoritmos tenham perdido na métrica de precisão, tiveram resultados mais expressivos de MRR e NDCG, mostrando uma eficiência maior em recuperar currículos relevantes mais próximo do topo da lista de resultados, o que serviu de indicador para uma análise mais detalhada, considerando o limiar de corte no quinto currículo recuperado.

Após a análise dos resultados com o limiar de corte no quinto currículo recuperado, verificaram-se ganhos percentuais em relação aos resultados com limiar de corte no décimo currículo, nas métricas de precisão e NDCG para todos os algoritmos, o que demonstra a ocorrência mais frequente de currículos relevantes entre os cinco primeiros recuperados.

Ainda considerando o limiar de corte no quinto currículo, o algoritmo Contagem de Borda se mostrou o mais eficiente quando aplicada a métrica de precisão, ou seja, é o algoritmo que traz maior quantidade de currículos relevantes entre os cinco primeiros recuperados. Quando aplicada a métrica NDCG, os melhores resultados foram para o algoritmo Soma das Similaridades, significando que este recupera currículos

relevantes mais próximos do topo da lista que os outros dois algoritmos. Os principais resultados são sintetizados nas Tabelas 4.18, 4.19 e 4.20.

	Soma	Produção	Contagem de Borda
PRECISÃO TOP 10	0,54	0,60	0,51
PRECISÃO TOP 5	0,61	0,61	0,64

Tabela 4.18: Resultados de precisão em média do segundo experimento.

	Soma	Produção	Contagem de Borda
NDCG TOP 10	0,82	0,71	0,82
NDCG TOP 5	0,89	0,87	0,87

Tabela 4.19: Resultados de NDCG em média do segundo experimento.

	Soma	Produção	Contagem de Borda
MRR TOP 10	0,81	0,70	0,81
MRR TOP 5	0,81	0,70	0,81

Tabela 4.20: Resultados de MRR do segundo experimento.

Com os resultados numéricos obtidos de precisão, NDCG e MRR, podemos observar o sucesso dos experimentos, fato que indica que a recomendação de competências pode ser usada para os mais diversos propósitos práticos, conforme é discutido na conclusão do trabalho.

5 Conclusões e Trabalhos Futuros

Nesta dissertação, foi proposto um método de busca de competências a partir de currículos Lattes, para a recomendação de pesquisadores a partir das competências recuperadas. A inovação proposta neste trabalho, e que o diferencia de outros similares na literatura, está na utilização e seleção dos campos mais relevantes da base dos currículos Lattes para executar a busca de competências.

Além disso, os algoritmos de busca Soma das Similaridades, Produções e Contagem de Borda, cada um utilizando uma estratégia diferente, foram submetidos a dois experimentos para testar sua eficácia na recuperação dos currículos mais relevantes.

A eficácia e a aplicabilidade do nosso método para busca de competências foi demonstrada por um experimento que obteve resultados de precisão que alcançaram média a 64%, para o algoritmo Contagem de Borda. Para alguns perfis testados, 100% dos currículos recuperados foram relevantes, para os três algoritmos testados. Em média, obtivemos precisão de 61% para os algoritmos Soma da Similaridades e Produção, o que nos leva a concluir que estes dois algoritmos são menos efetivos em recuperar currículos relevantes.

Portanto, os resultados do experimento foram satisfatórios e serviram para demonstrar a relevância deste trabalho. Suas aplicações práticas envolvem, portanto, a

busca e seleção de profissionais da área de pesquisa para as seguintes atividades: trabalhar em projeto, participar de banca de concurso, participar de banca de mestrado, trabalhar como orientador de mestrado e trabalhar como membro de comissão de programa de conferência. Além disso, poderá ser utilizado para o auto conhecimento institucional.

Outras aplicações deste trabalho são possíveis para outros domínios, por exemplo seleção de candidatos para emprego ou o uso para classificados *on line*.

Como sugestão para trabalhos futuros, podemos investigar os resultados do experimento para diferentes seleções de características. Por exemplo, após uma nova verificação cuidadosa sobre quais partes dos currículos Lattes encontram-se a maioria das competências pesquisadas pelos usuários, pode ser feito um novo método de busca que utilize somente estes campos.

Uma abordagem para alcançar melhores resultados, seria a utilização de novos métodos de votação, como por exemplo, o método Condorcet, e em seguida, comparar os seus resultados com os já obtidos neste trabalho.

Referências

Baeza-Yates, R. A. B. Ribeiro-Neto (1999). *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

e Cesar Junior Roberto Marcondes Mena-Chalco, J. P. (2009, 12). Scriptlattes: an open-source knowledge extraction system from the lattes platform. *Journal of the Brazilian Computer Society 15*, 31–39.

Harzallah, M., M. Leclère, and F. Trichet (2002). Commoncv: modelling the competencies underlying a curriculum vitae. In *Proceedings of the 14th international conference on Software engineering and knowledge engineering*, SEKE '02, New York, NY, USA, pp. 65–71. ACM.

Hongli, C. (2010, nov.). A study on the employees' competence modeling based on the personal space of information. In *Information Management, Innovation Management and Industrial Engineering (ICIII), 2010 International Conference on*, Volume 4, pp. 539–542.

Järvelin, K. and J. Kekäläinen (2002, October). Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst. 20*(4), 422–446.

Nature, J. L. (2010, March). Let's make science metrics more scientific. *Nature 464*(7288), 488–489.

Rodrigues, S., J. Oliveira, and J. M. de Souza (2004, July). Competence mining for virtual scientific community creation. *Int. J. Web Based Communities 1*, 90–102.

Wu, S. and S. McClean (2007, oct.). Several methods of ranking retrieval systems with partial relevance judgment. In *Digital Information Management, 2007. ICDIM '07. 2nd International Conference on*, Volume 1, pp. 13–18.

Zhu, J., A. Goncalves, V. Uren, E. Motta, and R. Pacheco (2005). Mining web data for competency management. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pp. 94–100.