Quantum: Uma Ferramenta para Descoberta de Competências e Colaborações Universitárias

Glauco Roberto Munsberg dos Santos, André Guimarães Peil, Daniel Retzlaff, André Alba, Ricardo Matsumura de Araujo, Daniela F. Brauner

Universidade Federal de Pelotas - UFPel - Pelotas/RS - Brasil
grmdsantos, agpeil, dkretzlaff, adhalba, ricardo.araujo@inf.ufpel.edu.br
Universidade Federal do Rio Grande do Sul - Porto Alegre/RS - Brasil
daniela.brauner@ufrgs.br

Resumo. No Brasil, o CNPq mantém uma base de currículos de pesquisadores atravé da Plataforma Lattes. Porém, a ferramenta de consulta oferecida pelo Lattes possui algumas limitações de funcionalidades. O filtro pelas competências de uma única instituição é inviável usando as ferramentas oferecidas. Além disso, a busca textual retorna apenas os resultados que citam explicitamente o referido termo de consulta. Como as ferramentas tradicionais de recuperação de informação utilizam apenas os termos que são mencionados nos currículos para indexar a informação, os usuários precisam ter conhecimento desses termos para recuperar currículos relevantes em suas consultas. Neste contexto, este artigo apresenta o Quantum, uma ferramenta Web que implementa um mecanismo de busca com expansão de termos apoiado por uma base de conhecimento, permitindo uma maior flexibilização no uso dos termos de busca. O objetivo é ampliar os resultados relevantes retornados nas buscas por competências, fornecendo assim uma melhor experiência de uso. Os resultados obtidos mostraram que houve um ganho significativo na aproximação do vocabulário utilizado pela comunidade com as publicações indexadas pelos currículos dos pesquisadores da Universidade Federal de Pelotas, onde a ferramenta foi implantada.

Palavras Chave: expansão de consultas, ontologias, ferramenta de busca, recuperação de informação

1 Introdução

Na Ciência da Computação a Recuperação de Informação (RI) é uma área abrangente que centraliza seus esforços em fornecer ao usuário uma forma fácil de extrair, de um montante maior de informações, as que sejam relevantes. Com o crescente volume de informação gerado pela sociedade, nasceu a necessidade de extrair rapidamente informações de grandes volumes de dados. Com o propósito de auxiliar essa recuperação de informação, surgiram os sistemas de recuperação de informações (SRI). O objetivo principal de um sistema de RI é recuperar os documentos relevantes à necessidade de informações do usuário e, ao mesmo tempo, recuperar o menor número possível de documentos irrelevantes. [1]

O problema central desse trabalho está em como permitir que a comunidade encontre competências de uma universidade, ou de um grupo de acadêmicos, através de uma busca intuitiva em uma base de currículos da instituição. Atualmente a

Plataforma Lattes (PL)[34], mantida pelo CNPq[35] (Conselho Nacional de Pesquisa e Tecnologia), oferece uma base interessante para busca de currículos de pesquisadores, já que tem como finalidade interligar diversas bases de dados como a de currículos, de grupos de pesquisa e de Instituições através de um único sistema. Hoje, a base da PL conta com mais de 3 milhões de currículos cadastrados[36]. A busca e a recuperação dessas informações tornaram-se um processo trabalhoso ao usuário, visto que a plataforma implementa mecanismos tradicionais de busca por termos identificados nos currículos dos pesquisadores. O usuário precisa ter conhecimento especializado sobre os termos a serem utilizados em suas buscas[2]. Sendo assim, uma busca por um conceito ou área pode resgatar apenas 27% do montante esperado para aquela busca [2].

Pensando nisso, a proposta deste artigo apresenta um SRI de currículos Lattes que aproxima os termos usados pela comunidade com aqueles usados pelos pesquisadores. Visto que existe um grande descompasso nos tipos de termos utilizados pela comunidade e pelos pesquisadores, já que a comunidade utiliza termos mais informais na pesquisa, enquanto os pesquisadores utilizam jargões técnico-científicos para descrever seus trabalhos. Atualmente, a ferramenta de busca provida pelo CNPq, aparentemente, não possui mecanismos de expansão de termos ou de consulta. Ademais esses mecanismos tendem a tornar a busca mais flexível a luz do usuário.

Este artigo está organizado da seguinte forma. Na seção 2 é apresentada a ferramenta Quantum[37] e as tecnologias utilizadas. A implantação na UFPel[38] e os resultados dos testes são apresentados na seção 3. Por fim, são apresentadas as conclusões e trabalhos futuros.

2 Ferramenta Quantum

A construção da arquitetura de um SRI está baseada em dois requisitos básicos de software: Eficácia e Eficiência. Sendo que, para o primeiro requisito, quando abordado na área de RI, preocupa-se em prover um mecanismo capaz de recuperar o conjunto mais significativo de documentos para uma determinada consulta do usuário, isso implica na qualidade sobre o sistema. Já o segundo requisito, eficiência, é então esperado que o SRI processe a consulta do usuário o mais rápido possível, implicando assim, no tempo de resposta do sistema[3].

A principal dificuldade para atingir a eficácia está em saber não só como extrair a informação dos arquivos, mas também em como utilizá-la para decidir o quanto ela de fato é relevante. Esse é o principal ponto em RI. Salienta-se ainda que a "relevância" é um julgamento pessoal que está intimamente ligada a tarefa a ser resolvida e o seu contexto. Assim, a relevância pode temporalmente ser modificada, ou seja, um documento que hoje pode ser útil e relevante para um determinado usuário, amanhã o mesmo documento pode não ter a mesma relevância.

[34] http://lattes.cnpq.edu.br
[35] http://www.cnpq.br
[36] http://estatico.cnpq.br/painelLattes/
[37] http://quantum.indeorum.com
[38] http://quantum.indeorum.com/ufpel

Compreendendo que deve-se construir um SRI que seja tanto eficiente como eficaz e segundo os autores [2] e [3] geralmente os mesmos são arquitetados sobre cinco componentes: a Coleta, Transformação de Dados, Indexação, Ranqueamento e Consulta, veja Fig. 1, a soma deles corresponde por todo o ciclo de indexação ao ranqueamento.

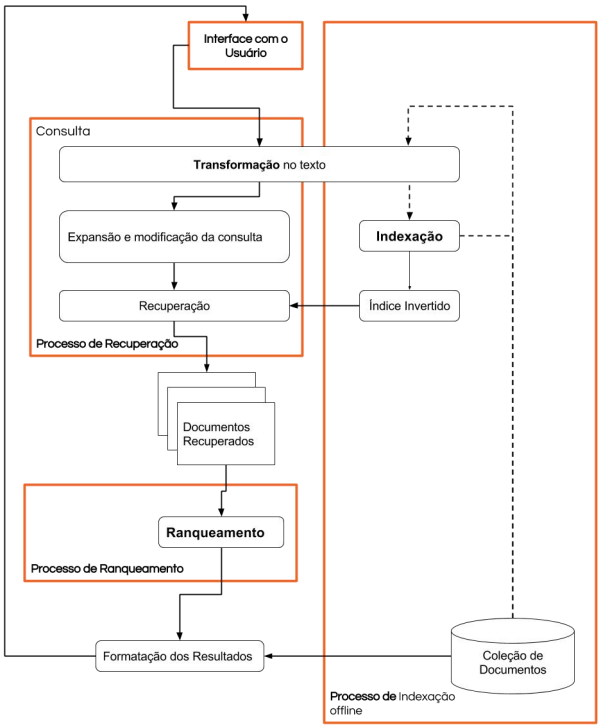

Fig. 1. Ilustração do processo de construção de um SRI.

2.1 Tecnologias

Para a concepção deste trabalho foram utilizados um conjunto de tecnologias com o propósito de auxiliar o desenvolvimento. A escolha delas deu-se por estarem em trabalhos relacionadas ou como recurso oferecido dada a familiarização com estas tecnologias. Elas agrupam-se em dois conjuntos, em que o primeiro conjunto está a linguagem de programação Python[39] utilizado nas etapas anteriores, a etapa de Indexação e Ranqueamento propostas na seção seguinte e a linguagem de

[39] http://www.python.org

Programação Ruby[40] junto ao *framework* Ruby On Rails[41] que servem de subsídio para a interface Quantum. Já o SGBD não relacional MongoDB[42] para a armazenamento dos currículos e informações que permeia todos as etapas do processo. No segundo conjunto estão as tecnologias essenciais para a idealização e concepção do processo aqui apresentado e estão dispostas na Figura 2 para a visualização de como estas interagem com as etapas para solucionar o problema central deste trabalho.

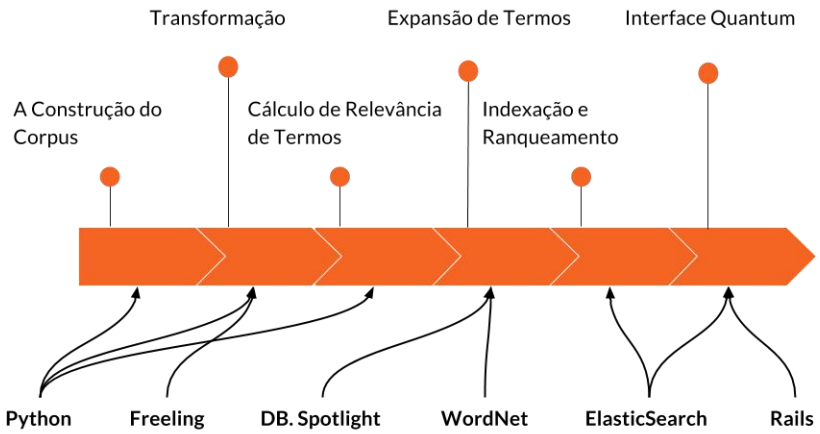

Fig. 2. Tecnologias envolvidas em cada uma das etapas da metodologia adotada

3 Implantação e Testes

Como pode ser visto na Fig. 3 o processo de desenvolvimento como um todo está dividido em 6 etapas, onde a primeira refere-se a construção do *corpus*[43], o segundo passo é então a transformação da informação da base de informação, a terceira está incumbida de calcular qual termo possui relevância para que na próxima etapa, 4º passo, sejam expandido os termos. Já o 5º dedica-se a descrever a indexação e ranqueamento sugerido e por fim o 6º passo está centrado a construção da interface pelo qual o usuário interage com o Quantum.

[40] http://www.ruby-lang.org/
[41] http://rubyonrails.org
[42] https://www.mongodb.org
[43] plural corpora, corpus é o conjunto de textos estruturados

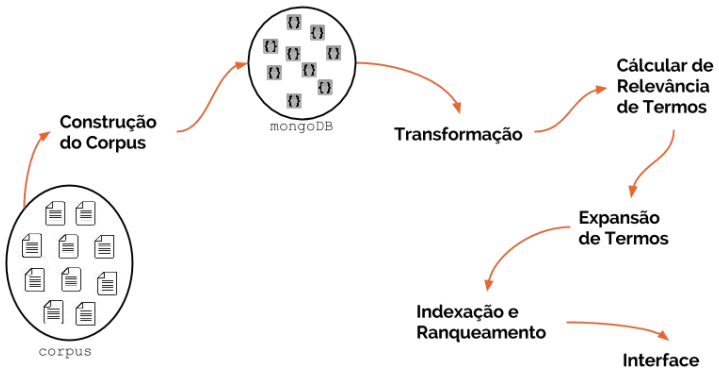

Fig. 3. Ilustração do processo implementado na ferramenta Quantum.

3.1 A constituição do *corpus*

A etapa de coleta trata-se então do momento de constituição do *copus*[44], ou seja, da coleção de documentos (currículos) que serão empregados nesse trabalho. Esta é uma etapa opcional, caso não haja acesso direto aos documentos, e no contexto deste trabalho os documentos forma disponibilizados pela UFPel. A coleta ocorreu em Novembro de 2015 e foram adquiridos 1995 currículos Lattes de docentes e técnicos-administrativos da instituição UFPel. Estes arquivos foram então exportados pela API da CNPq no formato XML, este formato de arquivo é largamente reconhecido pela sua estrutura em auxiliar o mapeamento e estruturação da informação e um dos principais meios pelos quais aplicações usam para trocar informação..

Dado o enorme volume de informações que há dentro de cada documento foram selecionados os campos dos currículos que melhor descrevem e definem o pesquisador. Esta restrição tem como objetivo melhorar a fórmula de ranqueamento e formação do *corpus*. Na abordagem evidenciou-se já que o processo de expansão dos termos (ET) não poderia ocorrer de forma arbitrária por todo o currículo, dado que este processo ocasionaria inconsistências, que serão abordados mais adiante na seção de Cálculo de Relevância de Termos. No contexto deste trabalho, selecionamos os 12 campos apresentados na Tabela 1.

[44] O termo Corpus usado na área de recuperação de informação, nasce da noção de Corpus Linguístico que é o conjunto de textos escritos e registros orais em uma determinada língua e que serve como base de análise.

Tabela 1. Campos selecionados para a formulação.

Campos	**Descrição**
lattes-id	Campo Identificador Único de cada um dos pesquisadores composto por 16 dígitos
dados-gerais//nome-completo dados-gerais//nome-em-citacoes-bibliograficas	Dois campos “nome-completo” e “nome-de-citacoes-bibligraficas” permitem a identificação do autor
artigos-publicados//artigo-publicado artigos-aceitos-para-publicacao//artigo-aceito-para-publicacao trabalhos-em-eventos//trabalho-em-eventos	Os três campos compreendem a soma das publicações realizadas pelo autor em periódicos
participacao-em-projeto//projeto-de-pesquisa	Permite adquirir as participação do autor em em projetos
orientacoes-concluidas//orientacoes-concluidas-para-doutorado orientacoes-concluidas//orientacoes-concluidas-para-mestrado orientacoes-concluidas//orientacoes-concluidas-para-pos-doutorado orientacoes-concluidas//orientacoes-concluidas	Permite identificar os trabalhos indiretos realizados pelo pesquisador através de suas orientação
//palavra-chave-	Uma série de 1 a 6 palavras chaves cadastradas em: Produção Bibliográfica, Orientações, Produção Técnica, Livros e Capíıtulos etc.

A fundamentação na escolha dos campos: nome completo, nome de citação, produções bibliográficas, a participação em projetos, orientações e por fim o resumo *currículo vitae* para nos se dá pelo fato de que estes são os campos parecer representam o *Status Quo* do autor, são áreas do currículo Lattes que representam suas práticas mais recentes no meio acadêmico e atualizado com mais frequência pelos seus autores.

Esta escolha demonstrou-se fundamental para que não houvesse a expansão demasiada de termos, como por exemplo, dos registros acadêmicos mais primários do autor. Esta restrição impede a expansão da formação inicial do docente, dado que há

uma propensão na graduação de uma maior volatilidade dos interesses e participação em pesquisas e projetos.

Os arquivos XML obtidos junto a instituição são meios diretos para a extração dos campos da Tabela 1, essa processo é realizado através de uma biblioteca e os campos resultantes armazenado na estrutura com o intuito de servidor como base para a próxima etapa do SRI.

3.2 Transformação

Transformação é a etapa sintetiza-se em um processo de limpeza e radicalização com o objetivo de tornar os dados ainda mais enxutos e significativos para o processo de indexação. O primeiro processo a ser aplicado nos dados, que estão armazenados no Sistema de Gestão de Banco de dados (SGBD), é a de aplicação de caixa baixa (minúsculas). Esta ação é tomada com a finalidade de diminuir a variação e melhorar a contagem e comparação das palavras, evitando que a diferenciação do tipo de caixa (baixa e alta) da palavra implique em duas entradas diferentes. Assim um termo do tipo "CIÊNCIA" terá a mesma referência que "Ciência" e "ciência" ao final do procedimento.

O processo seguinte então trata-se da tokenização onde ocorre a conversão de texto plano em um vetor de palavras. Trata-se de uma tarefa relativamente simples, porém importante para a etapa de análise morfológica onde se aplicam técnicas de *clustering* com o objetivo de determinar limites de morfemas, tratamento de afixos e pesquisas em dicionários para encontrar a sintaxe do termo.

Esta etapa é realizada pelo *Freeling* é uma biblioteca *open-source* que provê serviços básicos de Processmaento de Linguagem Natural entre outras funções para desenvolvedores de aplicações de NLP[45][8]. A sua escolha deu-se por prover o *sense* necessários para a etapa 3.3 (A etapa de casamento entre termo e sentido será melhor descrito mais a adiante no item 3.4). Também observamos que esta ferramenta automatiza a análise morfológica empregada na etapa 3.3 de forma satisfatória.

Os n-gramas resultantes do processos são então submetido ao processo de identificação da classe gramatical a qual pertence e também ao encontro do *word synset* da WordNet. As Wordnets são Ontologias Lexicais dada que as relações de hiperonímia e hiponímia podem ser vistas como categorias de especialização entre os conceitos. Utilizada com inúmeros pesquisadores da área de Processamento de Linguagem Natural a WordNet é um importante mecanismo utilizado para diversas atividades entre elas inclusão de desambiguação de sentido em palavras, sistemas de informação, classificação de textual entre outros[5]. Para este trabalho utilizou-se a openWordnet-PT[46] desenvolvida inicialmente na FGV com colaboradores e que pode ser acessada livremente[47][7].

Vejamos pela Figura 4 que o processo dá-se individualmente a cada título dos campos da Tabela 1, assim o processo resultante permite manter referência da origem da palavra. Posteriormente removidas as palavras de parada *stopwords*, por possuírem uma baixa capacidade de representação.

[45] Natural language processing
[46] https://github.com/own-pt/openWordnet-PT
[47] http://wnpt.brlcloud.com/wn/

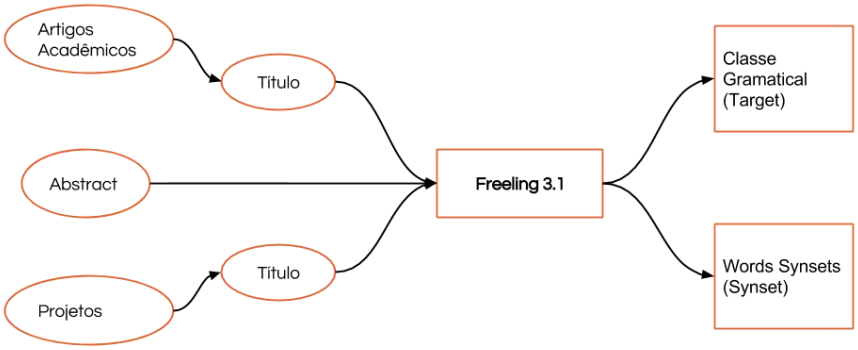

Fig. 4. Os campos são submetidos ao *Freeling* para a classificação

Ao final do processo de transformação, contamos com um *corpus* enxuto para a indexação futura, além do mais contamos com uma série de palavras chaves e sua frequência em cada um dos documentos. Essas tuplas serão úteis para compreender quais n-gramas são importante, sobre o olhar do *corpus*, para que ocorra a expansão do termo.

3.3 Cálculo de Relevância de Termos

Quando observado o conjunto de saída da transformação, não seria interessante a expansão de todos os termos, mesmo que estes estejam já em menor quantidade proveniente do processo de transformação (item 3.2), ainda há palavras de baixa relevância que o processo de ponderação pode verificar.

Portanto a relevância levantada pelo TF-IDF (term frequency–inverse document frequency)[6] é largamente usada na RI dada a sua característica de permitir identificar termos com especificidade mínima[48]. Este mecanismo será usado para verificar quais palavras dentro do *corpus* são relevantes para que ocorra a expansão.

Este processo de corte implica até mesmo na Exaustividade Ótima[49] do documento dado que apenas selecionamos termos, dentro do documento, que o caracterize-o ao mesmo tempo que distingue dos demais.

Isso sugere que o número médio de termos de indexação por documento deve ser otimizado de modo que a probabilidade de relevância de um documento recuperado seja maximizado. Com isso partimos então para a hipótese de que termos que possuem TF-IDF com fator acima de 0.5 são candidatos promissores para que ocorra a Expansão de Termos[50].

Para isso então é calculado o fator TF-IDF para todas as palavras resultantes do processo de Transformação (item 3.2) e consequentemente armazenada esta

[48] Quando o termo ocorre tem todos os documentos do *corpus* então diz-se que este termo tem especificidade mínima, logo não é útil para a recuperação dado que trará todos os documentos [1]

[49] Exaustibilidade da descrição de um documento é interpretada como a abrangência que ela provê para os tópicos principais de um documento [1].

[50] É considerado o intervalo de 0 a 1 com precisão de 6 casas decimais após a virgula

informação para uso no passo seguinte que é aonde ocorre a expansão dos novos termos.

3.3 Expansão de Termos

Terminada a etapa de cálculo de relevância de termo, que serve de subsídio para este passo, então inicia-se a expansão dos termos usando a WordNet. O propósito é obter a expansão de termos de forma a manter a coesão e a exaustividade ótima, porém com novos termos relacionados. Assim optou-se pela criação de uma métrica para prover pesos para as palavras expandidas, com o objetivo de manter a coerência dos pesos em relação aos termos que as originaram.

Compreendida a necessidade de adicionar termos aos currículos, para enriquecer o vocabulário do mesmos, observou-se que as palavras expandidas deveriam então ser originadas a partir de um conjunto de palavras sinônimas[51]. Neste contexto, optou-se por buscar conexões através dos *synsets* da WordNet.

Porém os *synsets* se relacionam através de estruturas que descrevem a relação semântica entre elas. Logo contamos com relações, como por exemplo, de hiperonímia, hiponímia e meronimia (Fig. 5). Com o objetivo de tornar o currículo mais próximo do vocabulário usado pela público alvo, optou-se por realizar expansões dos termos levando apenas em consideração as relações de hiperonímia, equivalência e holonímia.

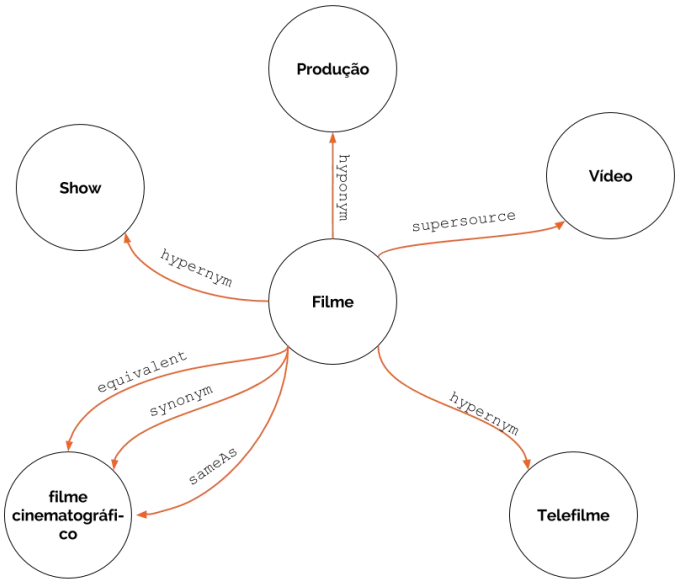

Fig. 5. Relação entre o termo "movie" com outros termos através das relações de sinonímias, hiponímias e hiperonímias. Fonte: [4]

[51] Diz-se de palavras que tem o mesmo significado e sentido, no entanto, são escritas com grafia distinta.

Esta escolha se justifica pela própria definição que se têm de hiperonímia, já que ela é sinônimo de super-ordenado, nome que se dá ao termo cujo sentido inclui aquele (ou aqueles) de um ou de vários outros termos, chamados hipônimos. Assim temos o *synset* "Animal" que é um hiperônimo de "cão", "gato"e "elefante" como exemplo. Esta tipo de relação se demonstra ideal, já que poderemos então expandir o termo "inteligência Artificial" a partir do termo "Redes Neurais", dado a relação de hiperonímia que há da primeira com a segunda[9].

Há outras relações a serem consideradas, como a de equivalência e semelhança(*same as*), por se acreditar que termos equivalentes nem sempre são usais na escrita, destaca-se a utilização desta para que ocorra uma expansão de termos caso o mesmo não esteja presentes no corpo do documento.

Semelhante relação entre palavras que poderia ser usada é a classe de hiponímia, porém esta classe tem com o princípio inverso da hiperonímia, ou seja, nos coloca em um termo mais especializado. A especialização causada pela hiponímia acarreta um inconveniente de identificar se de fato o contexto do termo é o mesmo do termo expandido. Dado o termo "Inteligência Artificial" expandir para "Redes Neurais" poderíamos causar um equívoco no documento, pois nem todos que trabalham com IA trabalham com Redes Neurais, mas a inversa é válida.

Por se tratar de uma rede, e os *synsets* estarem conectado por essas classes de relação, por exemplo, podemos então subir na árvore de hiperonímia, porém identifica-se uma objeção de perda de precisão. Termos hiperônimos possui significado mais abrangente em relação a sua origem, então a escalada na árvore tem como efeito colateral a perda da concisão.

A métrica proposta com o intuito de amortizar essa perda de concisão é o uso de uma progressão de -0.25 sobre o grau em relação ao termo original e este valor é multiplicado ao TF-IDF (peso) do termo original. Esta métrica naturalmente nos coloca um teto de 3 graus sobre o número de vezes que poderá ser expandido um termo.

Com este processo de calculo, o termo é adicionado a lista de termos do documento, junto a frequência igual a 1 e adicionado também a lista global de termos e sua frequência é incrementada. Ao final da expansão de todos os termos do *corpus* que possuem coeficiente superior a 0.5 contamos então com um *corpus'* com os termos originais, seus TF-IDF e também contando com os termos expandidos.

3.4 Indexação e Ranqueamento

Com o término da etapa de expansão de termos parte-se para o passo de indexação, componente este que é responsável pela requisito de eficiência do SRIs. Com a etapa de indexação de currículos finalizada, é necessário definir a etapa de ranqueamento, neste SRI o modelo de ranqueamento faz o uso de um modelo híbrido entre o Modelo Booleano e o Modelo Vetorial.

Para esse propósito e também para que sirva uma solução integrada e simples do projeto, a tecnologia usada pela solução Quantum[52], foi alinhada a tecnologia para a utilização desse modelo híbrido. Com isso propõem-se uma modificação sobre o aspecto de ranqueamento, onde o Boost(impulso) que será usado em cada um dos

[52] O Quantum usa como motor de busca a tecnologia ElasticSearch

campos dos documentos, já indexado pelo SRI, seja diferente dado o grau que se considera importante para a classificação geral.

3.5 Interface Quantum

Com todo o processo desenvolvido para a expansão e motor de busca, então a próxima etapa concentra-se na elaboração de uma interface gráfica que permitisse ao usuário interagir com o sistema. O modelo conceitual da interface apoia-se sobre a experiência prévias dos usuários que notamos sobre como interagem com os principais motores de busca como o Google[53], Bing[54] e Yahoo[55].

Para isso disponibilizamos na página inicial um campo textual centralizado no meio da página com o objetivo de ser o meio único para digitar a consulta (Figura 6). O Quantum disponibiliza o resultado de forma ordenada pela relevância do documento e paginado a cada dez documentos resultantes da consulta (Figura 7).

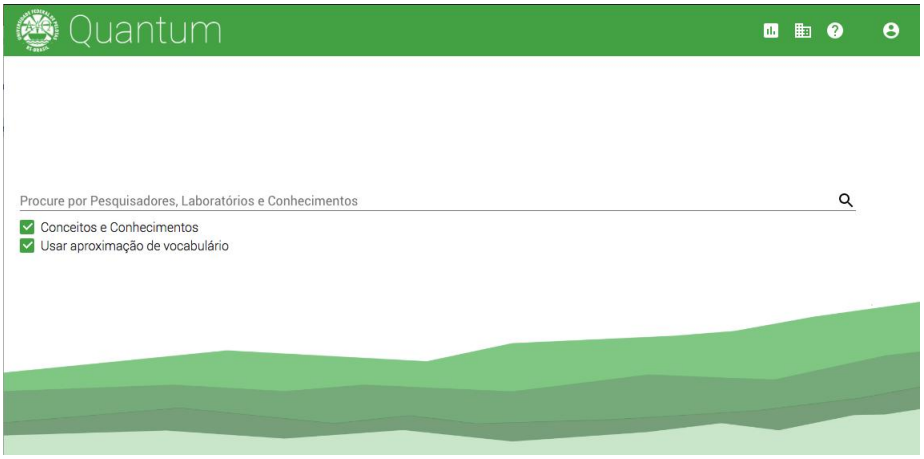

Fig. 6. Página Inicial do Quantum permite que usuário possa pesquisar por um termo, nome ou conhecimento que deseja obter informação com base nos currículos inseridos.

[53] https://www.google.com.br
[54] https://www.bing.com.br
[55] https://br.yahoo.com/

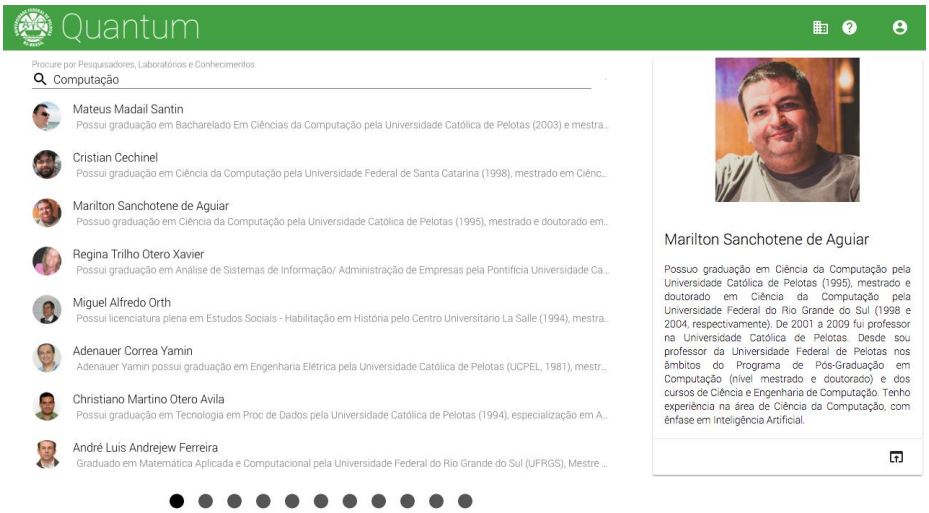

Fig. 7. Página de retorno para uma pesquisa por "computação" retorna um resultado de 109 currículos diferentes.

Ao clicar em um dos resultados da lista é possível visualizar um resumo do currículo recuperado pelo Quantum. Nele é há informações sobre o abstract do colaborador, os artigos científicios publicados e ordenados por ano, há também os projetos (Figura 8) e orientações também ordenados por ano e todos eles também disponíveis em forma de gráfico para facilitar a leitura da evolução das pesquisas e participações do currículo em questão.

Fig. 8. Com o resultado do Quantum é possível visualizar o perfil de cada um dos

colaboradores retornados. É possível visualizar a informação sobre Artigo, Projeto, Orientações e outros.

O trabalho de processamento dos termos (ítens 3.2,3.3 e 3.4) criou-se uma rede de termos semânticos encontrados e expandidos dentros dos currículos. Então a partir de cada currículo é possível navegar pelas palavras chaves vinculadas ao currículo e descobrir outros pesquisadores que possuem aquela palavra relevante dentro do seu currículo (Figura 9).

Fig. 9. Cada Currículos existe uma série de palavras chaves que identificam o autor. O Quantum permite navegar entre elas, encontrando outras pessoas com essas competências.

A compilação destas informações pelo Quantum permitiu também obter indicadores e gráficos que demostram a evolução como um todo dos currículos inseridos na ferramenta. O resultado é chamado de "Indicadores" e disponibiliza informações como a evolução da produção, projetos e patentes dentro da instituição. No caso da Figura 10 é possível ver que como produziram os 1995 colaboradores da UFPel.

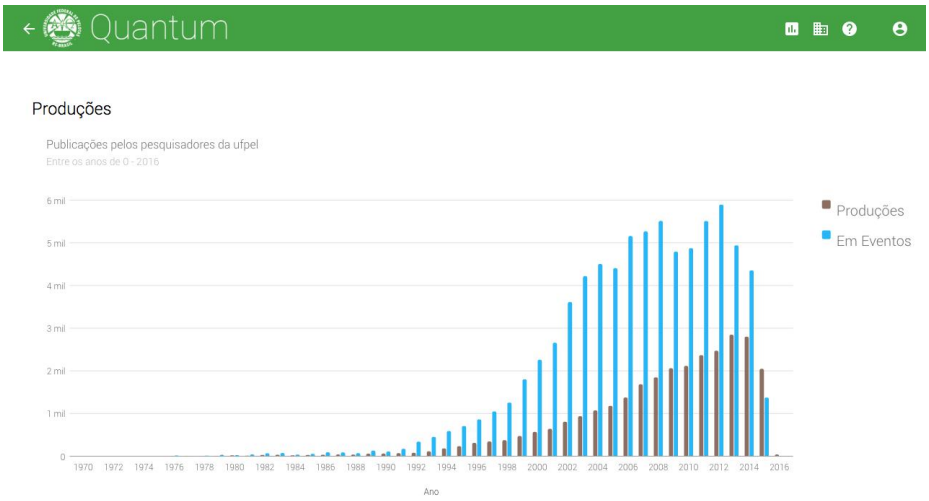

Fig. 10. A ferramenta permite visualizar de forma dos currículos como um todo e visualizar como está a produção de todos os colaboradores cadastrados no Quantum.

3.6 Teste do Quantum

Os testes foram idealizados com o objetivo de coletar o quanto as expansões realizadas impactariam os resultados das buscas. Para isso então foram coletadas uma série de informações de forma anônima aos os usuários, mas que identifica-se como eles realizavam as pesquisas e interagiram com o resultado. As informações coletas foram:

I. As palavras chaves nas consultas, o dia em que ocorreu a busca e o navegador de origem do usuário;
II. O identificador do currículo lattes que foi clicado;
III. A posição em que o resultado clicado estava na lista geral de resultados retornadas pelo motor de busca;
IV. Se alguma posição da lista visível ao usuário tinha a expansão da palavra buscada;
V. Se o resultado em que o usuário clicou havia ele sido retornado por causa do termo expandido naquele documento.

Também foram coletadas informações de *feedback* através de um formulário que esporadicamente aparecia para os usuários. Neste formulário foram coletadas informações sobre a experiência que o usuário obteve. Foram feitas as seguintes perguntas:

I. "O que você estava buscando no Quantum" com múltiplas escolha;
II. "Qual é o grau de satisfação com a(s) busca(s) realizada(s)" em uma escala de 0 a 5;
III. "Você identifica-se como" com múltiplas escolhas;
IV. "Como você chegou até o Quantum" com múltiplas escolhas;
V. "Ajude a melhorar a Ferramenta descrevendo sua experiência" múltipla escolha;

Com o objetivo de mensurar o impacto das expansões o sistema foi liberado para acesso ao público no dia 07 de Novembro de 2015 e foram realizadas as coletas até o dia 25 de Novembro de 2015. Contabilizando assim 18 dias que foram coletadas as informações que dão base a aos resultados abaixo descritos. Foram realizadas um total de 1,063 consultas no sistema pelos usuários, sendo que deste montante, 604 resultados foram clicados para visualizar mais informações sobre o currículo. Estes 604 resultados estão distribuídos sobre 280 lattes dos 1995 currículos cadastrados no sistema.

Do formulário foram contabilizadas 51 participações, destas participações 24 foram de discentes, 11 de professores internos e externos à UFPel e 9 participação de pessoas em empresas privadas e 5 participações em ouras categorias. O grau de satisfação de 1 de 5 obtivemos uma média de 3,66. Vejamos que em 71,2% das buscas contavam com a necessidade de encontrar Conhecimento e Competências e que do montante 28,8% procuravam uma pessoa específica (Figura 11).

Fig. 11. Resultado: Busca por tipo de informação.

Com o cuidado de preservar a informação de cada clique para que pode-se ser feita a catalogação e qualificação delas culminou nos resultados abaixo descrito. Primeiramente analisamos a questão de distribuição dos cliques por posição, ou seja, o quanto bem posicionado estava o resultado esperado para o usuário. Assim entre as 10 melhores posições encontrou-se 562 cliques, ou seja, 93% das buscas foram realizadas e o esperado estava na primeira página, dado que o Quantum listava os 10 primeiros resultados e paginava os demais onde apenas 7% precisou ir buscar o resultado em outras páginas que não fosse a primeira (Figura 12).

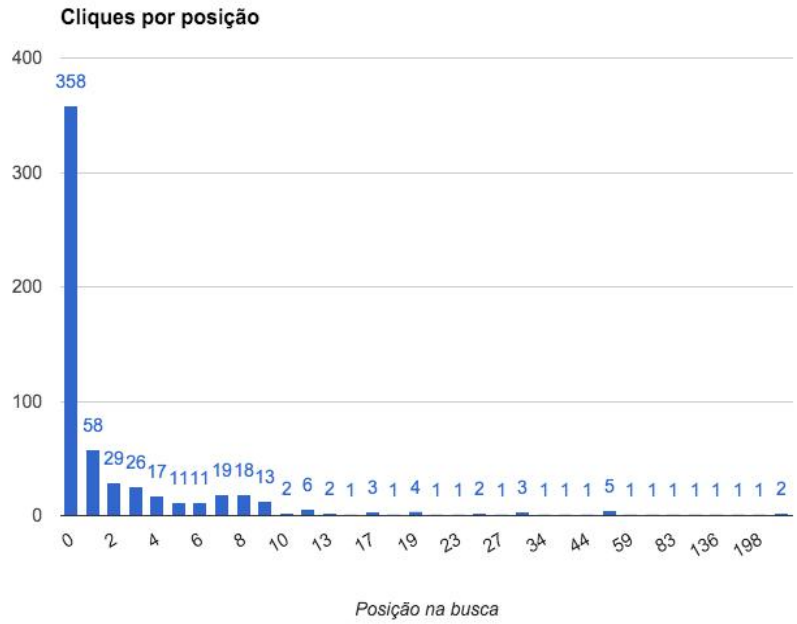

Fig. 12. Distribuição dos cliques por posição.

A Figura 13 demonstra a distribuição por posição dos 131 cliques que foram realizados unicamente por causa da expansão de busca. As posições que não receberam nenhum clique foram omitidos no gráfico.

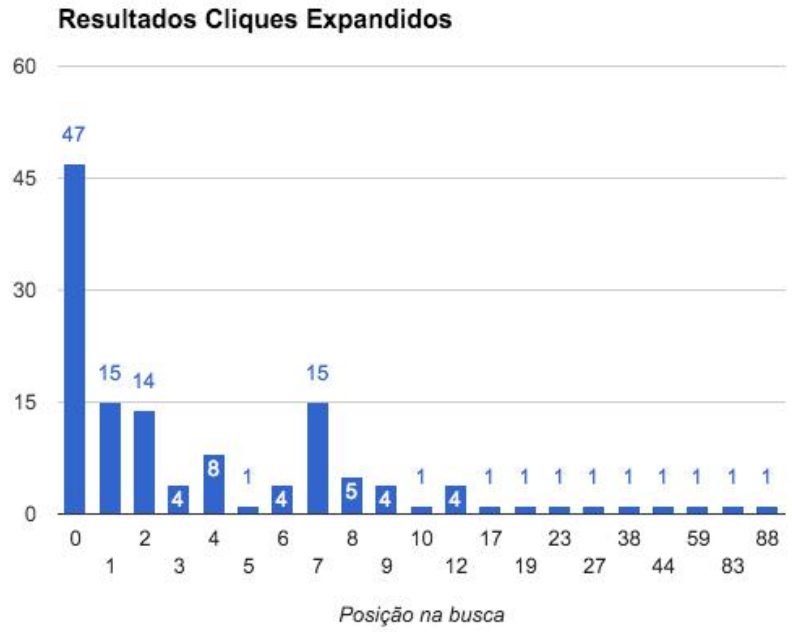

Fig. 13. Distribuição dos cliques por posição

4 Conclusão

Este trabalho teve como objetivo promover a expansão de termos realizados pela base de conhecimento lexical WordNet para aproximar o vocábulo dos docentes ao da comunidade. E também prover um meio pelo qual o público poderia encontrar competências, conhecimentos e pessoas através dos currículos da plataforma lattes.

Os resultados obtidos no motor de busca, com o sem expansões de termos, mostraram que houve um ganho significativo na aproximação do vocabulário entre o utilizado pela comunidade e pelas publicações indexadas. Já observando que 23.1% das consultas (202) realizadas contaram com uma expansão e que estas consultas 131 delas foram clicadas pelo usuário demonstra, que para esse conjunto de texto, houve uma relevância significativa para o motor de busca.

Porém os resultados sobre os cliques nos parece ainda carecer de maiores avaliações para compreender se o clique realizado foi efetuado em um resultado realmente desejado ou apenas clicou-se por estar em uma das primeiras posições do ranqueamento.

Com o propósito de melhorias na sequência são apresentados algumas propostas que poderão ser realizados para melhorar a ferramenta:

Há espaço para o refinação dos métodos de avaliação, assim por exemplo, realizar testes com amostragens temporais maiores com grupos de usuário com e sem expansão sobre o(s) mesmo(s) termo(s) de busca. Isto ajudaria a compreender como a expansão de busca está impactando a distância entre a resposta esperada do usuário com o ranqueamento retornado pelo SRI;

Ainda sobre a perspetiva de melhoraria do ranqueamento, pensa-se na utilização do fator H para melhorar o ranqueamento. Este é um modelo fortemente indicado, dado o Fator H é um cálculo para compreender quantas citações têm o artigo de um determinado autor. A incorporação deste fator poderá alavancar os currículos que possuem mais trabalhos com citações externas a base. Porém o Fator H é apenas uma das abordagens possíveis para a incorporações de informações externas para o ranqueamento, o número de *links* que há externamente ao currículo do Lattes nos parece também uma abordagem interessante;

Compreendemos ao final do trabalho que a realimentação de relevância nos parece uma abordagem interessante a se melhorar o ranqueamento do documentos, assim quando um documento é clicado ele ganha uma maior relevância nas próximas consultas, tendo invista que ao realizar novamente a mesma consulta o documento já terá melhorado sua posição;

Uma abordagem interessante a ser empregada a partir deste momento é a realimentação por meio de cliques[1] com o objetivo de aumentar o ranqueamento dos documentos mais populares do *corpus.*

Agradecimentos

Os autores deste trabalho gostariam de expressar o mais profundo agradecimento a Instituição de Ensino Superior Universidade Federal de Pelotas (UFPel) pelo apoio e disponibilidade da informação para a concepção desta ferramenta.

Referências

1. Baeza-yates R., R.-N. B. Recuperação de informação: Conceitos e tecnologias das máquinas de busca. 2.ed. [S.l.]: Porto Alegre: Bookman, 2013.

2. Souza Meireles, G. de. Currículo Lattes: Uma abordagem de busca explorando a recuperação de informação. 2014 — CDTec/Universidade Federal de Pelotas.

3. Croft, W. B; Metzler, D.; Strohman, T. Search engines: Information retrieval in practice. [S.l.]: Addison-Wesley Reading, 2010.

4. Padró, L.;Stanilovsky, E. FreeLing3.0:Towards Wider Multilinguality.In:Language Resources and evaluation conference (LREC 2012), 2012, Istan- bul, Turkey. Proceedings... [S.l.: s.n.], 2012.

5. Fellbaum, C. WordNet: An Electronic Lexical Database. 1.ed. [S.l.]: Bradford Books, 2009.

6. Büttcher, S.; Clarke, C.L.; Cormack, G.V. Informationretrieval:Implementing and evaluating search engines. [S.l.]: Mit Press, 2010.

7. Paiva, V. de; Rademaker, A.; MELO, G. de. OpenWordNet-PT: An Open Brazilian WordNet for Reasoning. In: International Conference on Computational Linguistics, 24., 2012. Proceedings... [S.l.: s.n.], 2012. See at http://www.coling2012-iitb.org (Demonstration Paper). Published also as Techreport http://hdl.handle.net/10438/10274.

8. Carreras, X.; CHAO, I.; Padró, ,L.;Padró, M.FreeLing: An Open-SourceSuite of Language Analyzers. In: LREC, 2004. Anais... [S.l.: s.n.], 2004.

9. Shekarpour, S.; Hoffner, K.; Lehmann, J.; Auer, S. Keyword query expansion on linked data using linguistic and semantic features. In: Semantic Computing(ICSC), 2013 IEEE Seventh Internacional Conference, 2013. Anais... [S.l.: s.n.], 2013. p.191–197.