UMA PROPOSTA DE METODOLOGIA PARA EXPANSÃO DE CONSULTAS BASEADA EM ONTOLOGIA

GLAUCO ROBERTO MUNSBERG DOS SANTOS[1]; DANIELA FRANSCISCO BRAUNER[2]; RICARDO MATSUMURA ARAUJO[3]

[1]*Universidade Federal de Pelotas – grmdsantos@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – danibrauner@inf.ufpel.edu.br*
[3] *Universidade Federal de Pelotas – ricardo@inf.ufpel.edu.br*

1. INTRODUÇÃO

A Plataforma Lattes[1], mantido pelo CNPq (Conselho Nacional de Pesquisa e Tecnologia), tem como finalidade interligar diversas bases de dados como a de Currículos, Grupos de Pesquisa e Instituições, através de um único sistema. Hoje a base do Lattes conta com um enorme volume de informações, até a presente data foram contabilizados mais de 3 milhões de currículos cadastrados[2], naturalmente o mantenedor da plataforma disponibiliza um mecanismo para realizar busca nesse amplo volume de informações. Porém a busca desta informação tornou-se um processo oneroso ao usuário que consulta a base Lattes, dado que o mecanismo de busca disponível exige que o usuário tenha conhecimento do que busca. Sendo assim uma busca por um conceito ou área pode ter um resultado resgatando apenas 27% do montante esperado para aquela busca MEIRELES et al. (2014). Visto que há um alto grau de informalidade nos termos mais frequentes de busca utilizados pelos usuários, enquanto os pesquisadores tendem a usar termos técnicos específicos para descrever seus trabalhos em seus currículos.

Neste contexto e dado que a ferramenta de busca provida pelo CNPq, aparentemente, não possui mecanismos de expansão de consulta, o que tornaria a busca mais flexível a luz do usuário, este trabalho enfoca-se na construção de métricas que permitam a um Sistema de Recuperação de Informação expandir os termos relevantes encontrados dentro dos currículos. Assim, o resultado central deste trabalho, é prover um mecanismo de expansão de termos que aproxime os termos usados pela comunidade com os usados pelos pesquisadores em seus currículos, devolvendo resultados mais relevantes para os usuários. Com esse objetivo propomos a utilização de uma Ontologia Lexical de Português chamada WordNet-Pt[3] para extrair as palavras de sinonímia[4].

Dado que esta importante base de conhecimento lexical permite que se navegue pelos termos através de suas propriedades como a Equivalência, Hiperonímia e Hiponímia. Com o estudo exploratório da versão em Português da WordNet[5] espera-se que seja capaz de atender a aproximação de termos usados entre pesquisadores e comunidade.

2. METODOLOGIA

[1] Plataforma Lattes http://lattes.cnpq.edu.br
[2] http://estatico.cnpq.br/painelLattes/
[3] http://wnpt.brlcloud.com/wn/
[4] Diz-se de palavras que tem o mesmo significado ou sentido, mas escritas com grafias distintas.
[5] http://wordnet-rdf.princeton.edu/

A metodologia empregada neste trabalho foi primeiramente realizado o levantamento de bibliografias sobre expansão de termos e bibliografias de ontologia lexicais. Posteriormente feito um processo experimental para a idealização de um processo de expansão, posteriormente, de forma incremental desenvolveu-se a metodologia abaixo descrito.

3. RESULTADOS E DISCUSSÃO

Baseado nos estudos realizados até então, a metodologia do trabalho resume-se em seis etapas (Figura 1) que descrevem o processo:

I. Construção do Corpus
II. Extração de Campos Relevantes
III. Limpeza e Transformação
IV. Calculo de TF-IDF
V. Expansão de termos
VI. Criação de Pesos

A construção do Corpus é uma etapa de coleta que trata do momento de constituição do *corpus*[6]. conjunto de textos escritos e registros orais em uma determinada língua e que serve como base de análise. No contexto deste trabalho, o corpus é a coleção de documentos (currículos) dos pesquisadores da UFPel que servirão para condução dos experimentos e testes da metodologia proposta. Esta é uma etapa opcional, caso não haja acesso direto aos documentos. Neste trabalho, a etapa de construção do corpus foi realizada com apoio da ferramenta Lattes Extrator disponibilizada pela próprio CNPq, direcionado as instituições de ensino e permite extrair diretamente do banco de dados do CNPq os currículos no formato XML (eXtensible Markup Language).

Figura 1: Processo de expansão de termos

Na etapa de Extração de Campos Relevantes foi realizada a escolha dos campos que melhor descrevem a atuação do autor, para isso foram escolhidos o campo *abstract* que corresponde ao resumo de sua atuação e aos artigos publicados pelo autor. A justificativa está em que estes campos representar o *Status Quo* do autor, assim representa suas práticas mais recentes no meio acadêmico e atualizado com mais frequência

[6] O termo *Corpus* usado na área de recuperação de informação nasce da noção de *Corpus Linguístico. BAEZA-YATES (2013)*

Na etapa de limpeza e transformação, os valores dos campos relevantes foram tratados com o objetivo de tornar os dados ainda mais enxutos e significativos para o processo de indexação. Para isso extraiu-se um conjunto de 100 *stop-words*, também pontuações e as palavras foram radicalizadas.

Quando observado o conjunto de saída após a etapa de limpeza e transformação, não julgamos interessante a expansão de todos os termos, mesmo que estes estejam já em menor quantidade proveniente do processo, ainda há conectivos e palavras de baixo relevância que um processo de ponderação pode verificar. Portanto a relevância levantada pelo TF-IDF foi escolhida sabendo que é largamente usados na Recuperação de Informação dado a sua característica de permitir identificar termos com especificidade mínima[7] (BAEZA-YATES, 2013). Este mecanismo será usado para verificar quais palavras dentro do *corpus* são relevantes para que ocorra a expansão.

Terminada a etapa de calculo de TF-IDF, que serve de subsídio para este passo, então evocou-se a utilização da WordNet com o propósito de obter as Expansões de Termos de forma a manter a coesão e a exaustividade ótima, porém agora havendo 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 junto os termos que as originaram

Compreendida a necessidade que há de adicionar termos aos currículos para enriquecer o vocabulário do mesmo, observou-se que as palavras expandidas deveriam então ser originárias a partir de um conjunto de palavras que possuem algum sentido de sinônima. Sendo assim valendo-se das conexões através dos *synsets* da WordNet para resolver esse problema.

Porém os *synsets* se relacionam através de estruturas que descrevem a subordinação entre elas. Logo contamos com relações, como por exemplo, de hiperonímia, hiponímia e meronímia. Com o objetivo de tornar o currículo mais próximo do vocabulário usado pelo público alvo, optou-se por realizar expansões dos termos levando apenas em consideração as relações de hiperonímia.

Esta escolha se justifica pela própria definição que se têm de hiperonímia, já que ela é sinônimo de superordenado, nome que se dá ao termo cujo sentido inclui aquele (ou aqueles) de um ou de vários outros termos, chamados hipônimos (FELLBAUM, 2009). Assim temos o *synset* "Animal" que é um hiperônimo de "cão", "gato" e "elefante" como exemplo. Este tipo de relação se demostra 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.

Há outras relações que poderiam ser consideradas, como a de equivalência e semelhança (*same as*), porém por se acreditar que termos equivalentes são usuais na escrita para evitar a repetição do termo no próprio nome de artigos, descartou-se a utilização desta para que não ocorra uma expansão de termos já presentes no corpo do documento.

Na etapa de expansão 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 (Figura 2).

[7] 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 *BAEZA-YATES (2013)*

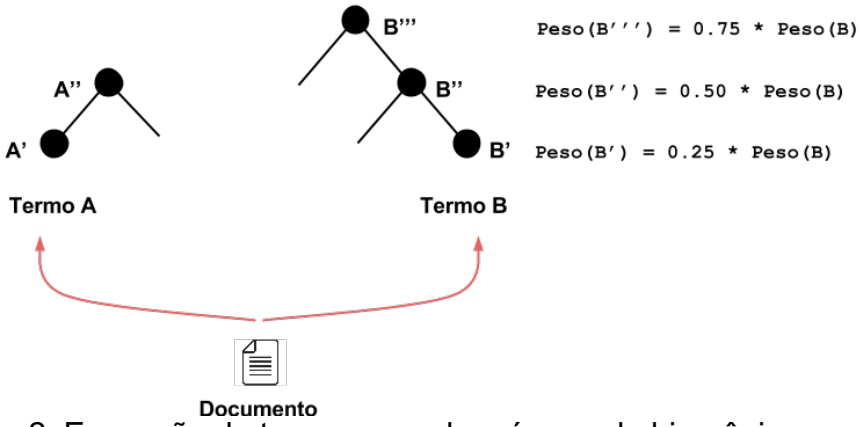

Figure 2: Expansão de termos usando a árvore de hiperônimos

O trabalho atualmente encontra-se na etapa de automatização da expansão dos termos onde evidencia-se que o suíte Freeling[8] permite um alto grau de precisão de vinculação do *synset* correto ao termo encontrado no *corpus* (GARCIA, 2010).

Como forma de avaliação do trabalho espera-se comparar os resultados entre dois *corpus*: Onde o primeiro caracteriza-se pelos termos extraídos dos lattes retornados pelo Lattes Extrator e o segundo *corpus* será dado pelo primeiro *corpus* porém com expansões realizadas pela métrica proposta e assim validar a sua eficiência.

4. CONCLUSÕES

O trabalho apresenta uma abordagem para expansão **DE CONSULTAS BASEADA EM ONTOLOGIA,** utilizando a Wordnet-PT. Como resultado, propõe-se uma metodologia para expansão de termos que aproxima os termos usados pela comunidade com os usados pelos pesquisadores em seus currículos, devolvendo resultados mais relevantes para os usuários numa busca por informações sobre pesquisadores.

5. REFERÊNCIAS BIBLIOGRÁFICAS

MEIRELES, G. S. **Currículo Lattes: Uma abordagem de busca explorando a recuperação de informação**. 2014. – Curso de Ciência da Computação, Universidade Federal de Pelotas.

SHEKARPOUR K., S. H., Keyword query expansion on linked data using linguistic and semantic features. In: SHEKARPOUR, S. K. **Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on**. Irvine, CA: IEEE, 2013. p. 191-197.

BAEZA-YATES, R. **Recuperação da Informação: Conceitos e Tecnologia das Máquinas de busca**. Porto Alegre: Bookman, 2013. 2v

FELLMAN, C. **WordNet: An Electronic Lexical Database**. Bradford Books, 2009.

GARCIA, M, G. P. **Análise Morfossintáctica para Português Europeu e Galego: Problemas, Soluções e Avaliação.** Linguamática 2.2, 2010.

[8] Freeling é um software OpenSource módulos de tokenização, segmentação de orações, reconhecimento de entidades e anotação morfossintáctica e pode ser acessada pelo link http://nlp.lsi.upc.edu/freeling/