ESTUDO DE SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÃO COM CONTEXTO SEMÂNTICO

Aluno: Daniel Schreiber Guimarães
Orientador: Sérgio Lifschitz

Introdução

Neste trabalho de iniciação científica foi feito um estudo do sistema de recuperação de informação Quem@PUC [1] e uma nova versão com aprimoramento de outro sistema de recuperação de informação chamado Busca@NIMA [2], que busca retornar informações com contexto semântico relacionado ao meio ambiente.

Objetivos

Estudar o funcionamento de um sistema de recuperação de informação que se utiliza de um banco de dados relacionado a grafo. Otimizar o sistema existente. Estudar a implementação de uma ferramenta de classificação automática de produções de professores e pesquisadores em função do envolvimento com o tema de meio ambiente. Criar esta ferramenta utilizando uma das possíveis e recomendadas abordagens encontrada durante os estudos e implementar em uma versão do sistema de recuperação de informação existente.

Metodologia

O sistema Quem@PUC (ler “quem na puc”) funciona de maneira similar a um site de buscas comum, mas contemplando apenas produções disponibilizadas no Currículo Lattes dos membros da PUC-Rio, além de disciplinas lecionadas pelos professores da Universidade e de outras informações adicionadas por estes através do próprio sistema, como laboratórios e palavras-chave.

O sistema funciona através de consultas à um banco de dados relacionado a grafo que contém todas as informações previamente carregadas em formato de triplas. Essas consultas são executas todas as vezes que um usuário solicita uma lista de professores e pesquisadores que possuem produções com o termo solicitado, ou se usuário solicita um resumo das produções de um professor ou pesquisador.

Para o contexto semântico, foram experimentadas diversas formas de classificação de produções. Entre as formas estão algoritmos de classificação aplicados a grafos, utilização de algoritmos de análise de sentimento modificados para análise de relação ao tema de meio ambiente, e uma rede neural treinada sobre produções contidas no banco para classificar posteriormente o restante das produções, sendo esta última abordagem a escolhida.

Para obter as produções para o treinamento da rede neural, foram classificadas manualmente por um pequeno grupo de pessoas cerca de mil produções contidas no banco de dados. Feito isso, foram experimentadas diversas configurações de redes neurais para treinamento, e a que melhor obteve resultado foi uma rede neural com uma camada profunda de 20 nós. Além disso, os dados das produções são passados por um pré-processamento de texto para serem convertidos em um vetor de quantidade de palavras para a rede neural utilizar como entrada. O melhor resultado obtido foi uma precisão de acerto de 89%.

Essa rede neural foi primeira desenvolvida em um workflow flexível para os testes iniciais, e depois desenvolvida do zero afim de eliminar qualquer ineficiência do workflow e para haver um melhor controle da rede neural afim de desenvolver uma ferramenta que possa

classificar automaticamente todas as produções do banco de dados utilizando a rede neural treinada.

Essa ferramenta [3] foi desenvolvida de forma que possa ser utilizada tanto para treinar a rede neural utilizando um conjunto de produções pré-classificadas como para classificar todas as produções do banco de dados relacionado a grafo.

Por fim, a informação do contexto semântico de uma produção quanto ao tema de meio ambiente foi armazenada no próprio banco de dados relacionado a grafo e utilizada pelo sistema Busca@NIMA, cujo funcionamento interno é idêntico ao sistema Quem@PUC, porém se utiliza da informação do contexto semântico contido no banco de dados, e exibe as produções somente se esta informação informa que a produção é relacionada ao tema de meio ambiente.

Abaixo estão os resultados dos dois sistemas para a mesma consulta, mostrando a diferença nos resultados devido ao contexto semântico na segunda imagem. Por exemplo, é possível notar que a quantidade de artigos exibidos diminuiu de 46 para 29, e a mesma situação pode ser observada para as biografias, capítulos, disciplinas, livros e orientações.

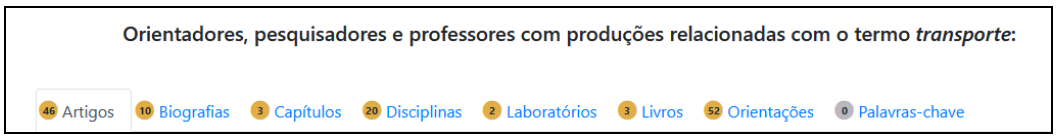

Figura 1 – Busca por "transporte" no sistema Quem@PUC

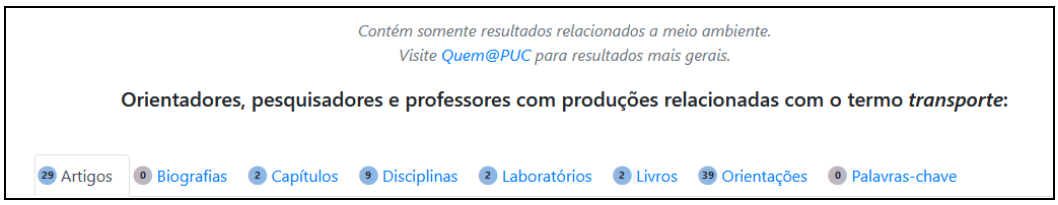

Figura 2 – Busca por "transporte" no sistema Busca@NIMA

Conclusões

Foi possível compreender e otimizar o funcionamento do sistema de recuperação de informação já existente, o Quem@PUC com sucesso.

Também foi possível estudar e criar um algoritmo para classificar produções contidas no banco de dados existente em função da relação com o tema de meio ambiente a partir de produções pré-classificadas.

Por último, foi possível implementar uma ferramenta que aplica a rede neural treinada no banco de dados por completo, classificando todas as produções contidas no banco, e usar essa classificação no sistema Busca@NIMA.

Referências

1 - Quem@PUC. Puc-rio.br. Disponível em: <quempuc.biobd.inf.puc-rio.br>. Acesso em: 20/07/2022.

2 - Busca@NIMA. Puc-rio.br. Disponível em: <nima.biobd.inf.puc-rio.br>. Acesso em: 20/07/2022.

3 - Leinadium/nima-predict: Classificação automática de produções. Produzida para a Iniciação Científica. Github.com. Disponível em: <https://github.com/Leinadium/nima-predict>. Acesso em: 20/07/2022