

PESQUISA e DESENVOLVIMENTO DE SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÕES COM CONTEXTO SEMÂNTICO

Aluno: Daniel Schreiber Guimarães

Orientador: Sérgio Lifschitz

Introdução

O sistema de recuperação de informações, intitulado Quem@PUC [1], permite pesquisar competências e conhecimentos dos professores e pesquisadores da PUC-Rio, através de seus artigos publicados, disciplinas lecionadas, livros e capítulos redigidos, laboratórios coordenados, e teses ou dissertações orientadas. Estes dados são obtidos de maneira semiautomática a partir dos currículos Lattes [2] destes professores, informações disponibilizadas pelo Sistema Acadêmico Universitário (SAU) da PUC-Rio, além de dados que o próprio professor pode adicionar no sistema, como palavras-chave adicionais para buscas avançadas ou equipamentos e produtos associados ao seu laboratório na PUC-Rio.

Esta pesquisa consiste em estudar e instanciar o sistema Quem@PUC em uma versão do sistema Busca@NIMA [3] restrito ao contexto de produções acadêmicas e científicas relacionadas ao meio ambiente. Esta versão do sistema possui as mesmas funcionalidades do Quem@PUC, mas com um contexto semântico associado às informações disponibilizadas, de forma que este novo sistema somente exiba produções relacionadas à uma temática específica. O sistema Busca@NIMA faz parte da plataforma do Núcleo Interdisciplinar de Meio Ambiente (NIMA) da PUC-Rio e exibe aos usuários interessados somente os resultados das buscas pertinentes e relevantes aos trabalhos do NIMA, enquanto o sistema Quem@PUC segue disponível pelo *website* do Centro Técnico-Científico (CTC) da PUC-Rio sem filtro de conteúdo.

Objetivos

Esta pesquisa tem como objetivo compreender o funcionamento do sistema Quem@PUC e suas fontes de dados, visando o desenvolvimento de uma solução que permita buscas com contexto semântico. Em particular, foi estudado um algoritmo de classificação de produções quanto ao tema de meio ambiente, que permite realizar o filtro das informações obtidas pelo sistema Quem@PUC, de forma a contextualizar as respostas das buscas.

Descrição do sistema

O sistema de recuperação Quem@PUC possui uma interface de funcionamento simples, similar aos sistemas de pesquisa por palavras-chave encontrados na web: ao acessar o sistema, é solicitado ao usuário um termo de busca que permita ser associado aos professores e pesquisadores nos seus artigos, livros, disciplinas, entre outras produções conhecidas. Esse termo pode ser composto e pode conter caracteres especiais como “*” e “?”, que permitem expandir a busca para além do termo exato.

Após realizar sua consulta, o sistema então fornece uma lista de professores e pesquisadores da PUC-Rio que, de alguma forma, estão associados aquele termo. Por exemplo, na figura abaixo, foram encontrados 34 professores ou pesquisadores que contenham artigos com o termo “meio ambiente”.

Quem@PUC

Descubra as áreas de atuação de **professores-pesquisadores** da PUC-Rio

Pesquise por pessoas, disciplinas, produções ou qualquer termo do seu interesse.

Deseja complementar sua busca com a tradução do termo?

Orientadores, pesquisadores e professores com produções relacionadas com o termo *meio ambiente*:

Artigos
Biografias
Capítulos
Disciplinas
Laboratórios
Livros
Orientações
Palavras-chave

Mostrar 10 orientadores, professores e pesquisadores da PUC-Rio por página
Filtrar:

Nome	Artigos
LEONEL AZEVEDO DE AGUIAR	5
JOSÉ BORZACCHIELLO DA SILVA	5
MARLEY MARIA BERNARDES REBUZZI VELLASCO	3
RACHEL COUTINHO MARQUES DA SILVA	3

Figura 1: Página de resultados do Quem@PUC para a consulta “meio ambiente”

Ao selecionar um dos professores ou pesquisadores da lista de resultados, o sistema gera uma página de perfil do indivíduo, contendo todas as produções (além de disciplinas, laboratórios e palavras-chave adicionadas por este mesmo indivíduo no sistema) encontradas com o termo, além de outras produções recentes publicadas por este.

Termos pesquisados: 'meio ambiente'
LEONEL AZEVEDO DE AGUIAR

BIOGRAFIA
ARTIGOS 5
LIVROS 0
CAPÍTULOS 0
ORIENTAÇÕES 0
DISCIPLINAS 0
LABORATÓRIOS 0

ARTIGOS

- [2011] Heurística do Medo: mídia e **MEIO AMBIENTE** na sociedade de risco, Angela Schaun, Leonel Azevedo de Aguiar
- [2008] O discurso sobre **MEIO AMBIENTE** na mídia alternativa: uma análise da revista Ecologia e Desenvolvimento, Leonel Azevedo de Aguiar
- [2007] **MEIO AMBIENTE**: discursos jornalísticos e representações da desordem global, Leonel Azevedo de Aguiar
- [2006] Contribuição para o ensino de jornalismo especializado: um estudo das representações sobre **MEIO AMBIENTE**, Leonel Azevedo de Aguiar
- [2005] Representações da crise do **MEIO AMBIENTE** no jornalismo científico, Leonel Azevedo de Aguiar

Artigos mais recentes:

- [2021] Comunidade interpretativa transterritorial: um mergulho nas rotinas produtivas da RTP, Ana Paula Goulart de Andrade, Leonel Azevedo de Aguiar
- [2021] Entretenimento e Jornalismo: o infotainment no programa Greg News, Júlia Cruz, Leonel Azevedo de Aguiar
- [2021] O populismo digital e a infodemia: reflexos da desordem informacional no discurso da audiência jornalística, Luciana Alcântara Roxo, Leonel Azevedo de Aguiar
- [2021] Nomear a mentira: a estratégia do jornalismo para resgatar seu locus de

Figura 2: Página de perfil do professor Leonel Azevedo, com o termo pesquisado em destaque

Tanto a página de resultado da consulta como a página de perfil do indivíduo selecionado contêm dados provenientes do mesmo banco de dados chamado *AllegroGraph* [4], um banco de dados orientado a grafo. Este banco é preenchido com dados fornecidos pela universidade PUC-Rio, assim como coletados dos currículos Lattes dos professores e pesquisadores da universidade. Os dados são disponibilizados em diferentes formatos (planilhas CSV e arquivos em formato XML) que são processados e convertidos em arquivos RDF que são então inseridos no banco de dados e indexados.

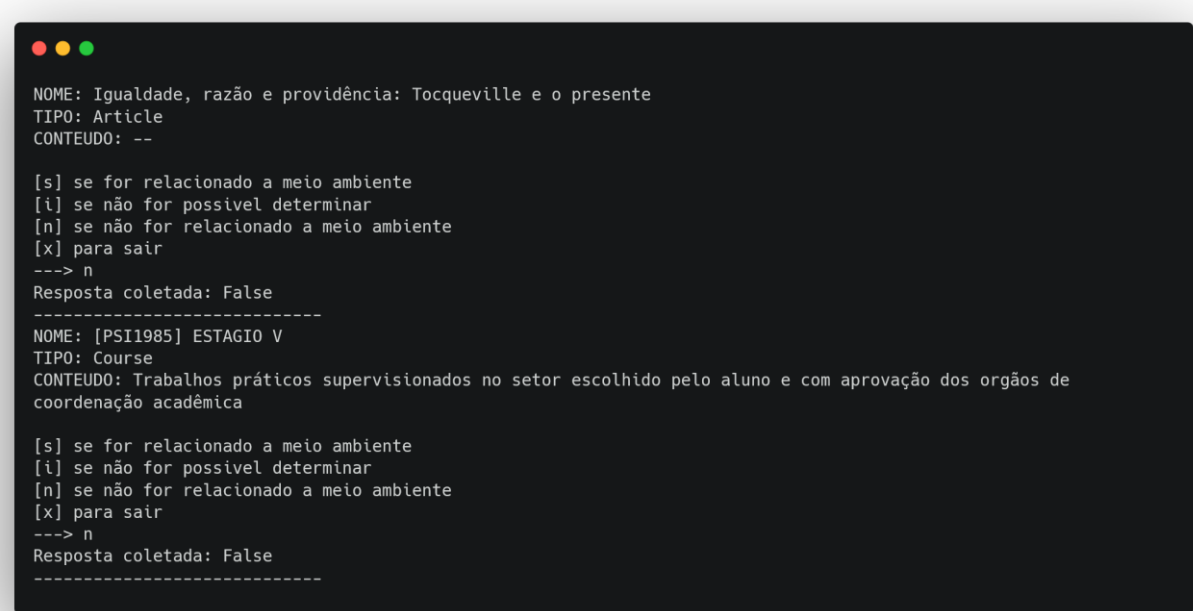
Para fazer as consultas, são utilizadas consultas na linguagem *SPARQL* [5] executadas através da biblioteca em Python [6] de integração com o AllegroGraph, *agraph-python* [7]. Essas consultas retornam os dados em formatos de triplas (*sujeito-predicado-objeto*) convertidas do banco de dados relacionado a grafo, que são usados para preencher a página de resultados e de perfil do professor ou pesquisador.

Classificação manual dos dados

Para desenvolver algum algoritmo de classificação das produções, é necessária uma base de produções que já contenham essa classificação para conseguir validar o resultado do algoritmo, ou servir como treinamento para um algoritmo de aprendizado supervisionado. Como o sistema possui sua própria arquitetura de dados, e a temática do contexto semântico é bem específica, não foi encontrada uma base de produções previamente classificadas. Por isso, é necessário classificar manualmente produções o suficiente para gerar um banco capaz de servir de validação e possivelmente de treinamento.

Para coletar as produções, foi utilizado um script Python juntamente com a biblioteca de integração com o AllegroGraph que coletava centenas de produções. Essas produções foram então armazenadas em um arquivo CSV, totalizando cerca de 1000 produções para serem classificadas.

Essas produções foram então classificadas manualmente por três indivíduos diferentes, denominados “juízes”, que julgavam se determinada produção era relacionada à temática de meio ambiente, se não havia nenhuma relação, ou se era impossível determinar. Foi gerado um script Python que lia o arquivo CSV contendo as 1000 produções e prontificava o juiz com os dados da produção, e armazenada a sua resposta em um arquivo separado.



```
NOME: Igualdade, razão e providência: Tocqueville e o presente
TIPO: Article
CONTEUDO: --

[s] se for relacionado a meio ambiente
[i] se não for possível determinar
[n] se não for relacionado a meio ambiente
[x] para sair
--> n
Resposta coletada: False
-----
NOME: [PSI1985] ESTAGIO V
TIPO: Course
CONTEUDO: Trabalhos práticos supervisionados no setor escolhido pelo aluno e com aprovação dos órgãos de
coordenação acadêmica

[s] se for relacionado a meio ambiente
[i] se não for possível determinar
[n] se não for relacionado a meio ambiente
[x] para sair
--> n
Resposta coletada: False
-----
```

Figura 3: Interface de classificação manual do juiz

Após cada juiz concluir sua classificação individual, todas as classificações eram comparadas, sendo somente consideradas as classificações em que pelo menos dois dos três juízes concordaram. Após fazer essa comparação cruzada, e descartar as classificações marcadas como “indeterminadas”, sobraram cerca de 850 produções com a classificação da temática de meio ambiente. Porém, havia um grande desbalanceamento das classificações positivas (uma classificação positiva significa que essa produção é relacionada à temática de meio ambiente) e negativas, cerca de 8 classificações negativas para 1 positiva. Por isso, os juízes classificaram mais produções de forma a adicionar mais produções positivas na contagem. No final, a base de produções classificadas manualmente possui 1226 produções, com 332 produções positivas e 894 produções negativas.

Aprendizado Supervisionado

As três possíveis implementações consideradas do algoritmo de classificação automática eram um algoritmo de aprendizado supervisionado, um algoritmo de aprendizado não-supervisionado, e um algoritmo de busca de palavras-chave. Dentre as opções possíveis, a escolhida foi a primeira, um algoritmo de aprendizado supervisionado, e as outras opções serão objeto de estudo futuro.

Para testar as diferentes combinações e possibilidades de algoritmos supervisionados, foi criado um *workflow* no Orange [8]. O *workflow* inicia carregando os dados provenientes do arquivo CSV na memória, seleciona as colunas relevantes (como nome, descrição, tipo de produção) e executa um pré-processamento no texto. Esse pré-processamento remove *stopwords* em português e inglês e aplica um algoritmo de *tokenização*, que transforma as palavras em *tokens*. Feito esse pré-processamento, os dados em texto/*tokens* são transformados em um modelo *Bag of Words* (B.O.W.).

As produções carregadas são depois separadas em dois conjuntos: 80% das produções são usadas para o treinamento dos algoritmos de aprendizado supervisionado, e o restante é usado para validação dos algoritmos. Dentre os algoritmos experimentados, estão algoritmos de rede neural profunda (com diferentes quantidades e formatos de camadas ocultas) e SVMs. Após múltiplas experimentações, o algoritmo que apresentou desempenho foi uma rede neural profunda, com três camadas ocultas (50 nós na primeira camada, 70 nós na segunda e 20 nós na terceira camada, e uma precisão de 89.5%, precisão aceitável para o projeto.

Com o modelo definido, o *workflow* no Orange foi adaptado para um script [9] Python que permite baixar produções diretamente do AllegroGraph, transformar em uma tabela CSV, e treinar um modelo de rede neural, armazenando o modelo resultante em um arquivo HDF5 [10], assim como o modelo de *tokenização* em um arquivo Pickle [11]. O mesmo script pode ser utilizado para baixar todas as produções e classificá-las utilizando um modelo já treinado anteriormente, e armazena a classificação no AllegroGraph em uma nova tripla para cada produção. A classificação é um número entre 0 e 1, em que 0 representa nenhuma relação com a temática de meio ambiente, e 1 representa total relação com a temática.

Adaptação do Quem@PUC para o Busca@NIMA

Para adaptar o Quem@PUC para o Busca@NIMA, foram feitas poucas modificações no código fonte. A página inicial passou por pequenas mudanças no design para diferenciação dos dois sistemas, e algumas páginas de explicação e descrição do sistema foram adaptadas conforme necessário.

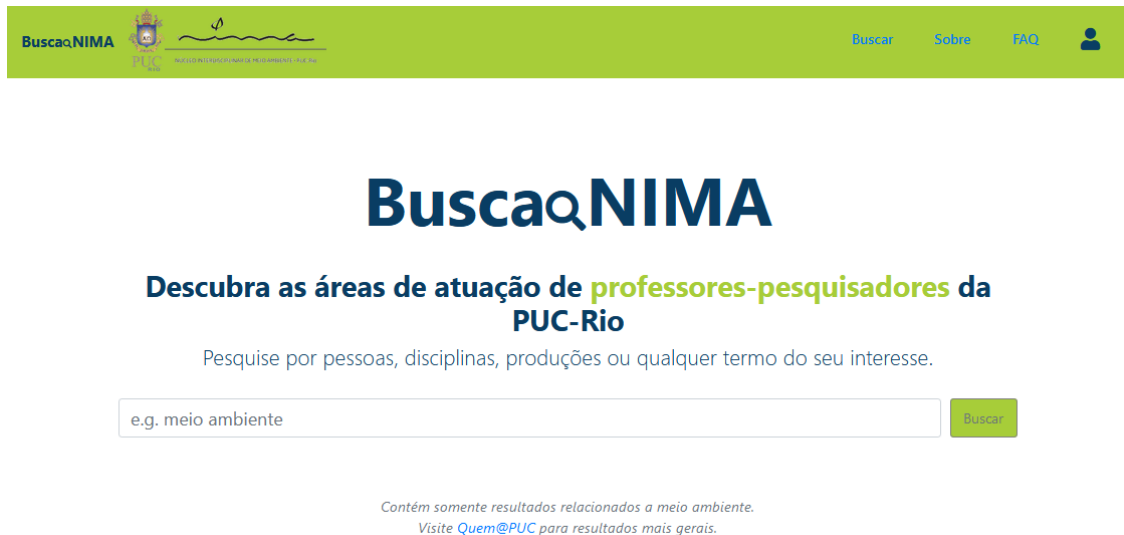


Figura 4: Página inicial do sistema Busca@NIMA

A única modificação efetuada no funcionamento das consultas foi a inserção de uma nova linha nas consultas SPARQL. No sistema Busca@NIMA, as produções retornadas devem possuir um valor de relação de meio ambiente acima do configurado. A imagem a seguir mostra uma das quatro consultas que são modificadas, adicionando duas linhas que filtram todos os resultados para mostrar somente produções que possuem o índice de relação com a temática de meio ambiente acima de certo valor limite, sendo usado o valor 0,5 no exemplo a seguir.

```
SELECT ?Producao ?tipo ?titulo ?autor ?param_a ?param_b
WHERE {
  # (...)
  VALUES ?tipo { bibo:Thesis bibo:Article bibo:Book bibo:Chapter } .
  {
    ?Producao
      rdf:type ?tipo ;
      dcterms:isReferencedBy/bibo:identifier ?Lattes ;
      dc:title ?titulo ;
      dcterms:issued ?data_producao ;
      dc:creator/foaf:name ?autor

    # linhas adicionadas no sistema para o Busca@NIMA:
    ?Producao cad-puc:RelacionadoMeioAmbiente ?meio_ambiente .
    FILTER ( ?meio_ambiente > 0.5 ) .

    # (...)
  }
}
```

Figura 5: Trecho da consulta SPARQL modificada

Como as mudanças no código fonte do sistema Quem@PUC são poucas, foi optado por uma implementação dinâmica. Ambos os sistemas possuem exatamente o mesmo código. Porém, uma variável de ambiente define o comportamento do sistema, por exemplo se o filtro na consulta deve ser adicionado e se as páginas e textos devem ser modificados.

O funcionamento do filtro pode ser ilustrado com a busca de um mesmo termo nos dois sistemas. As figuras 6 e 7 abaixo apresentam o resultado da consulta ao termo “toxicidade” tanto no Quem@PUC quanto no Busca@NIMA, respectivamente.

The screenshot shows the search interface of Quem@PUC. At the top, there is a search bar with the text 'toxicidade' and a 'Buscar' button. Below the search bar, a blue banner asks 'Deseja complementar sua busca com a tradução do termo?' with a text input field containing 'e.g. toxicity'. The main heading reads 'Orientadores, pesquisadores e professores com produções relacionadas com o termo toxicidade:'. Below this, there is a horizontal list of filters: Artigos (2), Biografias (0), Capítulos (1), Disciplinas (1), Laboratórios (0), Livros (0), Orientações (1), and Palavras-chave (0). A 'Mostrar' dropdown is set to '10' and the text 'orientadores, professores e pesquisadores da PUC-Rio por página' is visible. A 'Filtrar:' input field is also present. The results are displayed in a table with two columns: 'Nome' and 'Artigos'.

Nome	Artigos
RICARDO QUEIROZ AUCÉLIO	1
TACIO MAURO PEREIRA DE CAMPOS	1

Figura 6: Resultado da consulta usando o termo “toxicidade” no Quem@PUC

The screenshot shows the search interface of Busca@NIMA. It has a similar layout to Quem@PUC, with a search bar containing 'toxicidade' and a 'Buscar' button. The blue banner also asks 'Deseja complementar sua busca com a tradução do termo?' with 'e.g. toxicity' in the input field. However, the main heading is 'Orientadores, pesquisadores e professores com produções relacionadas com o termo toxicidade:'. Below this, the filter list is: Artigos (1), Biografias (0), Capítulos (0), Disciplinas (1), Laboratórios (0), Livros (0), Orientações (0), and Palavras-chave (0). The 'Mostrar' dropdown is set to '10' and the text 'orientadores, professores e pesquisadores da PUC-Rio por página' is visible. A 'Filtrar:' input field is also present. The results are displayed in a table with two columns: 'Nome' and 'Artigos'.

Nome	Artigos
TACIO MAURO PEREIRA DE CAMPOS	1

Figura 7: Resultado da consulta usando o termo “toxicidade” no Busca@NIMA

É possível observar que no Quem@PUC foram encontrados mais artigos relacionados a esse termo que no Busca@NIMA, onde há o contexto semântico de meio ambiente.

Por exemplo, o professor Ricardo Aucélio possui um artigo com o título “*Complexos de Mn(II) e Co(II) de nitro-tiossemicarbazonas: estudo das propriedades espectroscópicas e toxicidade frente a Artemia sp*”, que não possui relação com a temática de meio ambiente. Por isso, esse artigo aparece no sistema Quem@PUC, mas não no Busca@NIMA. Por outro lado, o professor Tacio Mauro aparece em ambos os sistemas pois o seu artigo “*Influência da salinidade na toxicidade de sedimentos dragados na Lagoa Rodrigo de Freitas e Baía de Guanabara (RJ): Efeitos tóxicos em minhocas*” está relacionado ao meio ambiente.

Otimização dos sistemas

Durante o estudo do sistema previamente implementado no sistema Quem@PUC, foi possível observar que as consultas SPARQL realizadas poderiam ser otimizadas se fossem reescritas de outra maneira. Anteriormente todas as consultas dependiam do nome do

professor ou pesquisador, que são campos não indexados na base de dados AllegroGraph. Se as consultas dependessem do identificador Lattes, que são identificadores de nodes no AllegroGraph, as consultas poderiam encontrar os resultados mais rapidamente.

Para isso, as consultas foram modificadas para usarem como base esse identificador. A seguir estão comparações de algumas consultas executadas no sistema em Abril de 2022, período em que o sistema utilizava as consultas não otimizadas, com as mesmas consultas realizadas no mesmo banco, mas com as consultas otimizadas.

Consulta	Antes	Depois	Eficiência atual
Termo “ algoritmo ”	17.83 s	2.77 s	643% mais eficiente
Termo “ meio ambiente ”	57.70 s	2.95 s	1955% mais eficiente
Termo “ dados ”	216.82 s	5.18 s	4185% mais eficiente
Termo “ daniel ”	19.55 s	3.20 s	610% mais eficiente
Termo “ eficiente ”	13.40 s	3.20 s	418% mais eficiente
Termo “ anali* ”	597.38 s	7.73 s	7728% mais eficiente
Termo “ dado? ”	223.10 s	5.24 s	4259% mais eficiente
Perfil “ Marcos Villas ”	23.43 s	2.33 s	1005% mais eficiente
Perfil “ Sérgio Lifschitz ”	21.35 s	4.56 s	468% mais eficiente

Tabela 1: Comparação das consultas

Conclusões

A implementação de um algoritmo de aprendizado supervisionado para a classificação de produções quanto à temática de meio ambiente foi implementada com sucesso, através de uma rede neural com três camadas ocultas e um pré-processamento de texto. Existem outras opções que podem trazer uma maior eficácia, mas a implementação utilizada era suficiente para o projeto.

Também foi possível adaptar e otimizar o sistema existente, Quem@PUC, em um novo sistema sem muitas modificações do código fonte, permitindo que futuramente mudanças em um sistema podem ser adaptadas no outro sistema de maneira fácil e rápida.

Referências

- 1 - **Quem@PUC**. Puc-rio.br. Disponível em: <quempuc.biobd.inf.puc-rio.br>. Acesso em: 20/07/2022.
- 2 – **Plataforma Lattes**. CNPq. Disponível em: <lattes.cnpq.br>. Acesso em: 29/08/2022.
- 3 - **Busca@NIMA**. Puc-rio.br. Disponível em: <nima.biobd.inf.puc-rio.br>. Acesso em: 20/07/2022.
- 4 – **AllegroGraph**. Franz Inc. Disponível em: <allegrograph.com>. Acesso em 29/08/2022.
- 5 – **SPARQL Query Language for RDF**. W3C. Disponível em <https://www.w3.org/TR/rdf-sparql-query>. Acesso em 29/08/2022.
- 6 – **Welcome to Python.org**. Python Software Foundation. Disponível em: <python.org>. Acesso em 29/08/2022.
- 7 – **franzinc/agraph-python: AllegroGraph Python API**. Github.com Disponível em <https://github.com/franzinc/agraph-python>. Acesso em: 29/08/2022.
- 8 – **Orange Data Mining**. University of Ljubljana. Disponível em <https://orangedatamining.com/>. Acesso em: 29/08/2022

9 - **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.

10 – **The HDF5© Library & File Format – The HDF Group.** The HDF Group. Disponível em: <<https://www.hdfgroup.org/solutions/hdf5/>>. Acesso em: 29/08/2022.

11 – **pickle – Python object serialization.** Python Software Foundation. Disponível em: <<https://docs.python.org/3.10/library/pickle.html>>. Acesso em: 29/08/2022