Sistemas de Recuperação de Informações Aplicados à Produções Acadêmicas

Mariana D. A. Salgueiro, Veronica dos Santos, André L. C. Rêgo, Daniel S. Guimarães, Edward H. Haeusler, Jefferson de B. Santos, Marcos V. Villas, Sérgio Lifschitz

[1]Departamento de Informática – (PUC-Rio) - Rio de Janeiro - RJ

{msalgueiro, vdsantos, hermann, jsantos, villas, sergio}@inf.puc-rio.br

{andrerego, danielg}@aluno.puc-rio.br

Resumo. *Este trabalho apresenta o projeto e a construção de Sistemas de Recuperação de Informações que permitem a identificação de projetos de pesquisa e/ou desenvolvimento, e as competências existentes em laboratórios e departamentos, coordenados por integrantes do quadro de professores-pesquisadores da PUC-Rio, a partir da busca por uma ou uma lista de palavras-chave. As fontes de informação que compõem o banco de dados do projeto são convertidas para o formato RDF usando ontologias de domínio, e são armazenadas em uma base NoSQL que suporta indexação de texto livre nativamente. Os resultados da busca incluem nomes, produções científicas diversas, atividades de ensino e links para contato. Ilustramos nossa solução com dois sistemas em desenvolvimento: Busc@NIMA*[1] *e Quem@PUC*[2]*.*

1. Introdução

Como podemos encontrar pesquisadores da PUC-Rio que trabalham com temas vinculados ao meio ambiente? Ou ainda, mais genericamente, como podemos encontrar pesquisadores da PUC-Rio que trabalham com qualquer tema? Estas perguntas podem surgir de diversas maneiras, como por exemplo, quando existe a necessidade de se consultar um especialista para uma entrevista em um jornal ou até para alunos que desejam aprofundar seus estudos e gostariam de encontrar aqueles que possam lhes orientar.

Este artigo descreve o projeto e a construção de dois Sistemas de Recuperação de Informação [Cardoso 2000], denominados Busc@NIMA e Quem@PUC, que permitem a descoberta de pesquisadores e professores de uma instituição de ensino e pesquisa, no caso a PUC-Rio, a partir de uma lista de palavras-chave. Por meio destas ferramentas, disponíveis através de navegadores na Web, é possível divulgar as atividades de pesquisa da comunidade PUC-Rio para serem utilizadas pela sociedade em geral, considerando o interesse particular de outros pesquisadores e também de jornalistas.

2. Contexto e Motivação

O primeiro Sistema de Recuperação de Informações desenvolvido, denominado Busc@NIMA, diz respeito a uma demanda do Núcleo Interdisciplinar de Meio Ambiente (NIMA), órgão da PUC-Rio que incentiva e desenvolve projetos e pesquisas sobre

[1]https://buscanima.biobd.inf.puc-rio.br/

[2]https://quemnapuc.biobd.inf.puc-rio.br/

assuntos relacionados ao meio ambiente. Os coordenadores do NIMA desejavam facilitar e ampliar o acesso do público às atividades de professores e pesquisadores na área de meio ambiente. Assim, a ideia da ferramenta Busc@NIMA surgiu para permitir aos interessados identificar as competências na área de meio ambiente existentes na PUC-Rio, em seus laboratórios e departamentos, coordenados por integrantes do quadro de professores e pesquisadores da Universidade. A partir disso, resolvemos desenvolver também o Quem@PUC (”quem na PUC”). Ao invés de retornar somentes professores/pesquisadores que trabalham em temas relacionados com a área de meio ambiente, quisemos ampliar as possibilidades de acesso ao público para fazer buscas a partir de qualquer tópico de interesse. Essa demanda surgiu naturalmente a partir do momento que o Busc@NIMA foi colocado em produção.

Principal diferença entre os sistemas (ainda em desenvolvimento)	
Busc@NIMA	Quem@PUC
Identificar membros da comunidade PUC-Rio que trabalham com temas **vinculados ao meio ambiente**	Identificar membros da comunidade PUC-Rio que trabalham com temas **diversos**

Tabela 1. Diferença entre os sistemas.

Como indicado na tabela 1, o Busc@NIMA necessita que o escopo da base de dados seja reduzido somente a temática de meio ambiente, o que ainda está em processo de desenvolvimento. Portanto, hoje em dia, nos dois sistemas, é possível fazer buscas de qualquer palavra-chave relacionada a qualquer tema. Além da motivação de construção dos sistemas em si, que traz consigo a possibilidade de conhecimento técnico de uma série de tecnologias recentes, é objeto de estudo a forma na qual será possível reduzir o escopo do projeto somente a área de meio ambiente. Alguns exemplos para redução de escopo são: fazer um pré-processamento com uma lista de palavras relacionadas a meio ambiente, o que ainda não é viável porque não possuímos tal lista, ou poderíamos inserir todas as informações relacionadas somente a professores da PUC-Rio que sabe-se que atuam na área de meio ambiente, o que, como revés, excluiria professores que eventualmente fizeram trabalhos relacionados a esta temática.

3. Projeto e Especificação do Sistema

A construção dos sistemas foi feita com o uso da linguagem de programação Python em conjunto com o microframework Flask, que permite uma rápida prototipação de aplicações Web. As duas ferramentas não exigem autenticação e podem ser acessadas livremente, bastando ter acesso à internet e a um browser da Web.

As fontes principais que compõem a base de dados, comum aos dois sistemas, são os currículos Lattes dos professores da PUC-Rio, combinadas com suas páginas pessoais em websites departamentais, e as disciplinas normalmente oferecidas. Como estas informações estão espalhadas em diversas fontes de dados, um primeiro desafio deste projeto foi reuni-las em um só local. Em seguida, permitir retornar, através de palavras-chave, os resultados de forma estruturada e organizada, de maneira eficiente.

3.1. Bancos de Dados

Considerando a finalidade dos sistemas, dois requisitos não-funcionais foram identificados na fase inicial de desenvolvimento que nortearam o design da arquitetura do sistema: (1) Suportar esquema de dados flexível e (2) Suportar o carregamento de dados de arquivos em diferentes formatos, como XML, RDF, CSV e planilhas Excel. Um modelo de dados sem esquema rígido permite que os dados tenham variedade na estrutura. Portanto, decidimos que os dados integrados seriam armazenados em um armazenamento NoSQL com esquema flexível.

O SGBD NoSQL AllegroGraph foi selecionado para o armazenamento centralizado de dados originados de outros sistemas. Trata-se de um *triplestore* que permite a manipulação de triplas RDF e a visualização, em estruturas de grafo. O AllegroGraph permite o maior número de triplas (cinco milhões) por repositório em comparação com outras opções disponíveis em sua versão gratuita, oferece suporte à linguagem SPARQL (uma linguagem de consulta padrão para *linked data*) e, também, possui suporte nativo à indexação de dados textuais (*Freetext Indexing*), permitindo mapear rapidamente palavras e frases para as triplas do banco de dados.

Já o SGBD PostgreSQL foi escolhido para registro de eventos como forma de analisar o comportamento dos usuários e os termos buscados. Este log de dados tem como intuito permitir a geração de estatísticas como, por exemplo, quais termos foram os mais buscados, quais professores foram os mais selecionados, quais os horários em que os usuários mais fazem buscas, entre outros. Em caso de eventuais problemas, o log também fornece informações para que eles sejam corrigidos rapidamente. Como os dados a serem cadastrados estavam bem definidos e não sujeitos a constantes mudanças no esquema, optou-se por adotar um SGBD Relacional para atender a este requisito.

3.2. Conversão das Diferentes Fontes de Dados

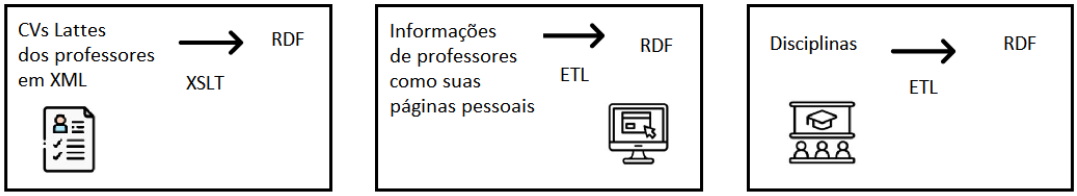

Figura 1. Conversão das diferentes fontes de dados

A plataforma Lattes, mantida pelo CNPq, é responsável por armazenar e disponibilizar os dados dos Currículos Lattes dos pesquisadores brasileiros. Para extração destes dados a plataforma permite a geração de um arquivo XML. Um script XSLT transformou cada arquivo XML em RDF usando ontologias selecionadas. Entretanto, no ínicio do projeto, este script de conversão XML para RDF foi feito somente mapeando produções científicas (artigos em conferências e revistas, livros, orientações de teses, capítulos de livros etc.) e para as nossas ferramentas atualmente é interessante a disponibilização de outras informações relevantes, como projetos de pesquisa e desenvolvimento. A geração de um novo XSLT que mapeie estes novos elementos dos currículos está sendo feita.

A ferramenta de ETL *Linked Pipes*[3] foi utilizada para a conversão das informações sobre os professores e das disciplinas oferecidas na PUC-Rio. É uma ferramenta leve, especializada em triplificar, ou seja, especializada no processo de transformar dados de qualquer formato, estruturado ou semiestruturado (relacional, XML, CSV), em formato de tripla, sendo apoiada pelo uso de ontologias de domínio, necessárias para a transformação.

3.3. Linguagem de Consulta

As consultas de correspondência de subgrafos foram especificadas na linguagem SPARQL para recuperar recursos de interesse sobre pesquisadores/professores. A correspondência de palavras-chave nas consultas foi realizada usando o operador (*magic property*) *fti:match*, que permite que a consulta use o índice *freetext* criado.

Para que a busca pela palavra-chave indicada retorne um resultado que envolva todos os repositórios considerados, foi necessário usar o conceito de Federação. Este recurso permite que o AllegroGraph automaticamente distribua as consultas SPARQL entre os repositórios e combine os resultados de forma transparente para a aplicação.

3.4. Implantação

Todo o sistema está dividido em dois servidores: um contendo a ferramenta de ETL (ETL Server), onde são feitas todas as conversões RDF necessárias e o outro contendo o servidor de aplicação (Web Server + DB Server), onde os bancos de dados estão hospedados e as ferramentas de busca são executadas, gerenciadas pela ferramenta Apache.

3.5. Modelo de Domínio

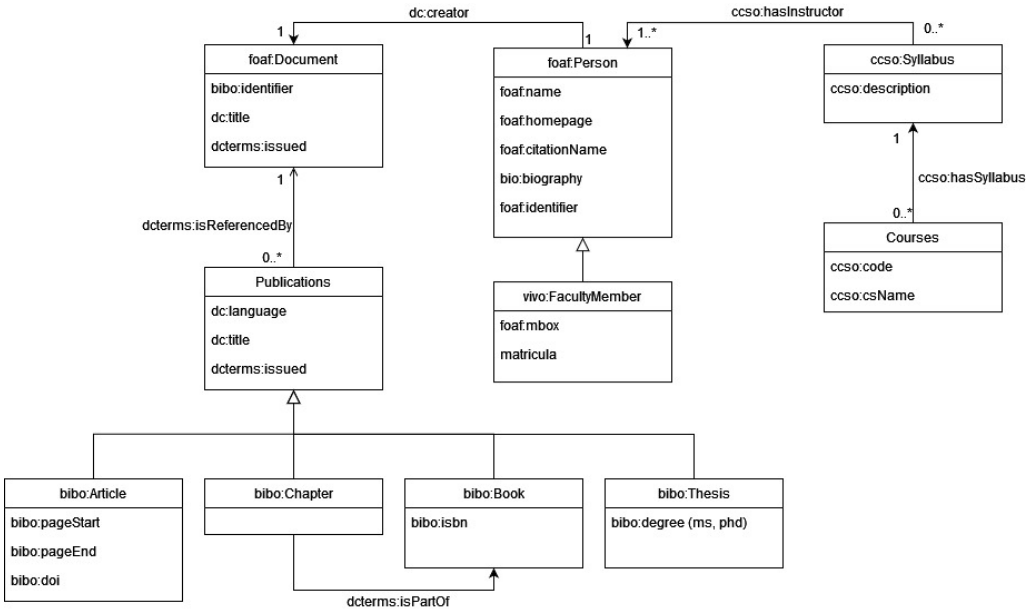

Figura 2. Diagrama de Modelo de Domínio

4. Ferramentas

O resultado inicial de uma busca nas duas ferramentas é uma lista de professores/pesquisadores que de alguma maneira estão relacionados com o termo buscado, como

[3] `https://etl.linkedpipes.com/`

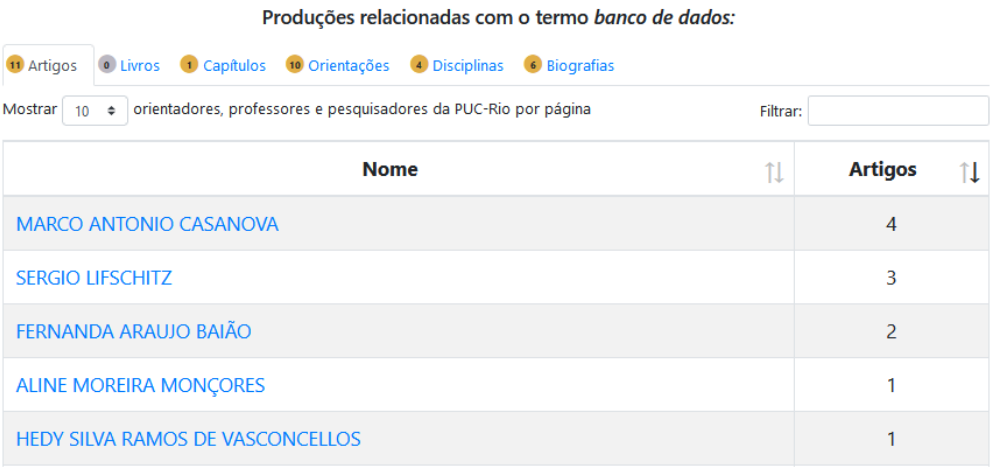

Figura 3. Exemplo de Busca na Ferramenta

mostra a figura 3. Esta lista é ranqueada de acordo com um critério de relevância [Baeza-Yates and Ribeiro-Neto 1999] que definimos ser a maior quantidade de produções encontradas por professor. A partir desta lista é possível selecionar um professor, e através desta seleção, ler sua biografia, acessar sua página para contato, seu CV Lattes e são apresentadas as produções e as disciplinas que ele leciona, caso existam. Ao selecionar qualquer uma das categorias de produção, os itens que contenham a palavra-chave buscada aparecem primeiro e, em seguida, aparecem os últimos itens lançados nos anos de 2019, 2020 e 2021 ordenados de forma cronológica inversa, com um limite de até as 5 mais recentes, como mostra a figura 4.

Figura 4. Retorno da Ferramenta ao Selecionar um Professor

Os sistemas hoje, ainda não são capazes de expandir a abrangência de suas buscas por não conseguirem lidar com a ambiguidade em uma pesquisa. Eles estão focados nos termos envolvidos nas buscas (sintaxe) e não em seus significados (semântica). Os sistemas mostram resultados de correspondência exata a partir das palavras-chave

de entrada, ou seja, na lista de professores/pesquisadores encontrados a partir de uma busca, teremos os itens associados a eles que contenham exatamente a(s) palavra(s) informada(s). Nesse contexto, é necessário incorporar elementos semânticos em seus mecanismos com o intuito de melhor compreender as intenções dos usuários por trás de suas buscas [Rozsa et al. 2019]. Como uma maneira inicial de contornar esta questão, os sistemas já são capazes de distinguir quando a palavra-chave buscada se refere a um nome de professor ou a um termo contido em algum título, como mostra a figura 5.

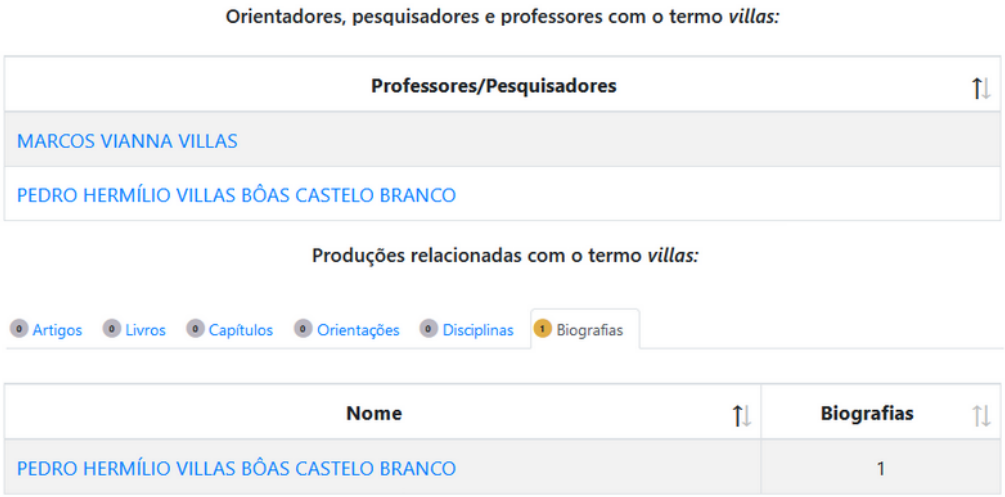

Figura 5. Desambiguação de termo: nome de professor ou título?

5. Conclusão

As ferramentas Busc@NIMA e Quem@PUC, então, são protótipos de Sistema de Recuperação de Informações que identificam projetos e trabalhos de pesquisa e/ou desenvolvimento de professores na PUC-Rio, com o Busc@NIMA sendo voltado somente para a área de meio ambiente. O objetivo inicial do projeto era desenvolver as ferramentas contemplando os CVs Lattes, as informações de professores (tais como suas páginas pessoas em websites departamentais) e os dados sobre as disciplinas oferecidas na PUC-Rio e os professores envolvidos. Destas três fontes principais, todas foram contempladas.

Referências

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Modern information retrieval. In *Modern Information Retrieval*, pages 1–2. Pearson, Addison-Wesley.

Cardoso, O. N. P. (2000). Recuperação de informação. *INFOCOMP Journal of Computer Science*, 2(1):33–38.

Rozsa, V., Godoy Viera, A. F., and Dutra, M. (2019). Aplicação de tecnologias da web semântica em motores de busca na internet. *Investigación bibliotecológica*, 33(78):165–191.