Universidade de São Paulo
Instituto de Matemática e Estatística
Bacharelado em Ciência da Computação

Bruna Meneguzzi

Sistema de integração e de recomendação de itens de acervos de bibliotecas brasilianas

São Paulo
Janeiro de 2021

Sistema de integração e de recomendação de itens dos acervos de bibliotecas brasilianas

Monografia final da disciplina
MAC0499 – Trabalho de Formatura Supervisionado

Supervisora: Profa. Dra. Kelly Rosa Braghetto
Coordenadora: Profa. Dra. Nina Sumiko Tomita Hirata

São Paulo
Janeiro de 2021

Resumo

O presente trabalho de conclusão de curso teve como objetivo o desenvolvimento de um sistema de integração e recomendação de itens de acervos de bibliotecas brasilianas, que tratam-se de coleções de livros de autores brasileiros e livros sobre o Brasil escritos por autores estrangeiros. Para isso, criou-se um modelo de banco de dados orientado a grafos para dados catalográficos de itens de acervos e de usuários de bibliotecas baseado em dados organizados no padrão catalográfico MARC 21. A partir da modelagem, definiu-se funções de similaridade entre itens e entre usuários e criou-se um sistema que abrange quatro tipos de recomendações de itens a usuários: a partir do perfil de um usuário, a partir de um item, a partir de termos de busca e a partir de termos de busca em conjunto com os perfis de usuários. Para verificação do sistema, foi utilizado um conjunto de dados de itens do acervo do Instituto de Estudos Brasileiros e dados de empréstimos da Biblioteca Florestan Fernandes, ambos da Universidade de São Paulo. A modelagem do banco de dados apoiou diferentes tipos de buscas e recomendações sofisticadas. O sistema utiliza o formato padronizado MARC 21, adotado em diversos sistemas de gerenciamento de bibliotecas, e define recomendações a partir de *scores* baseados em pesos que podem ser configurados. Assim, o sistema foi projetado para integrar dados de diferentes bibliotecas e se ajustar às necessidades dos seus usuários.

Palavras-chave: biblioteca brasiliana, sistema de integração de dados, itens de acervos, sistema de recomendação, banco de dados orientado a grafos.

Abstract

The present work of conclusion of the course had as objective the development of a system of integration and recommendation of items of collections of Brazilian libraries, which are collections of books by Brazilian authors and books about Brazil written by foreign authors. For this, a graph-oriented database model was created for cataloging data of collection items and library users based on the MARC 21 standard format for bibliographic data. Based on the modeling, similarity functions to quantify the similarity between items and between users were defined. A recommendation system has been created and it covers four types of item recommendations to users: from a user's profile, from an item, from search terms and from search terms together with user profiles. To verificate the system, a set of items data from the collection of the Instituto de Estudos Brasileiros and loan data from the Florestan Fernandes Library were used, both allocated at the University of São Paulo. The database modeling supported different types of searches and sophisticated recommendations. The system use the standardized format MARC 21, adopted by several library automation systems, and define recommendations from scores based on weights that can be configured. Thus, the system was designed to integrate data from different libraries and adjust to the needs of its users.

Keywords: Brazilian library, data integration system, collections items, recommendation system, graph oriented database.

Sumário

Capítulo 1

Introdução

1.1 Contextualização

Este trabalho é parte do projeto intitulado "Brasiliana Inteligente", que se trata de um consórcio de bibliotecas (ainda em formação) liderado pelo Instituto de Estudos Brasileiros (IEB) da Universidade de São Paulo (USP) e que tem a participação também da Biblioteca Brasiliana Guita e José Mindlin.

Uma biblioteca brasiliana, de acordo com Rubens Borba de Moraes, é uma coleção que reúne livros de autores brasileiros, impressos no Brasil e no exterior, bem como livros sobre o Brasil escritos por autores estrangeiros, impressos dentro ou fora do Brasil (Antunes, 2017).

O Instituto de Estudos Brasileiros (IEB), criado em 1962 por Sérgio Buarque de Holanda, é um centro multidisciplinar de pesquisas e documentação sobre a história e as culturas do Brasil. Sua biblioteca é considerada uma das mais ricas em assuntos brasileiros, contendo livros, separatas, teses e periódicos.

O IEB tem sob sua responsabilidade a guarda e a manutenção de um acervo formado, até o momento, por 91 fundos e coleções de artistas e intelectuais brasileiros, além de documentos resultantes de pesquisas. Esse fundos estão distribuídos entre o Arquivo, que reúne 450 mil documentos textuais e audiovisuais, a Biblioteca, que se ocupa de 180 mil livros publicados, e a Coleção de Artes Visuais, que administra 8 mil objetos, sendo eles obras publicadas e objetos tridimensionais. O objetivo da biblioteca é organizar, preservar e divulgar o acervo, visando oferecer conteúdo para a pesquisa de brasileiros e estrangeiros em diversas áreas, como Artes, Literatura e Sociologia, além de subsidiar publicações[1].

A Biblioteca Brasiliana Guita e José Mindlin é um órgão da Pró-Reitoria de Cultura e Extensão Universitária da USP que abriga e integra a coleção brasiliana reunida ao longo de mais de oitenta anos pelo bibliófilo José Mindlin e sua esposa Guita. O acervo doado à USP em 2006 reúne material sobre o Brasil e/ou conteúdo publicado por brasileiros que

[1] http://www.ieb.usp.br

sejam importantes para a compreensão da história e da cultura do país. Até o momento, o arquivo compreende fundos e conjuntos documentais que somam, aproximadamente, 30 mil títulos. O compromisso e objetivo da biblioteca é o de conservar, divulgar e facilitar o acesso de estudantes, pesquisados e do público em geral ao acervo, promovendo o interesse, a pesquisa e a difusão científica de assuntos brasileiros[2].

O projeto "Brasiliana Inteligente" compreende a integração de dados dos acervos das bibliotecas brasilianas e uma iniciativa de recomendação de itens dos acervos aos usuários com base em seus interesses, pesquisas e empréstimos. Em função de existirem diferentes bibliotecas que têm o Brasil como temática ou que se dedicam à manutenção e divulgação de itens de autores brasileiros, um aluno ou pesquisador que deseje encontrar um item pode necessitar consultar (ou se dirigir a) diversas bibliotecas para encontrá-lo, uma vez que não existe um sistema que integre todas elas.

1.2 Objetivos

A existência de bibliotecas brasilianas distribuídas pelo mundo, com seus diferentes sistemas de catalogação, acervos característicos de obras raras e grande quantidade de volumes constrói um cenário em que a integração dos metadados e a busca por itens de interesse são tarefas desafiadoras.

O objetivo deste trabalho é desenvolver um sistema de integração de dados e recomendação de itens de acervos de bibliotecas brasilianas.

Para isso, foi modelado um banco de dados orientado a grafos, próprio para grandes volumes de dados inter-relacionados, para unificar os dados dos acervos. Além disso, o sistema possui um mecanismo de recomendação de itens de usuários com base nas suas preferências, no histórico de empréstimos, na similaridade entre itens ou na similaridade entre usuários. Essa recomendação objetiva incentivar o estudo da cultura brasileira e auxiliar os pesquisadores que já atuam nessa área.

O sistema foi testado com dados de itens do acervo do Instituto de Estudos Brasileiro da Universidade de São Paulo.

1.3 Estruturação da Monografia

O capítulo 2 tratará da fundamentação teórica, introduzindo conceitos importantes para o entendimento do sistema desenvolvido, como bancos de dados de grafos e indexação de textos.

O capítulo 3 apresentará os dados utilizados, suas fontes e como eles se inter-relacionam.

O capítulo 4 mostrará a modelagem dos dados e as métricas de similaridade criadas

[2]https://www.bbm.usp.br

para apoiar a recomendação de itens. O capítulo também apresentará os componentes do sistema de integração e recomendação desenvolvido e explicará detalhadamente sua implementação e as ferramentas utilizadas nela.

O capítulo 5 abordará os experimentos feitos para verificar o sistema implementado e a análise dos resultados deles.

Por fim, o capítulo 7 apresentará as considerações finais e indicará próximos passos que podem contribuir com o projeto "Brasiliana Inteligente".

Capítulo 2

Conceitos

O sistema desenvolvido neste trabalho é, em sua essência, um banco de dados orientado a grafos e um sistema de recomendação. Este capítulo desenvolverá, portanto, conceitos importantes para o entendimento desse sistema.

2.1 Sistema de Gerenciamento de Bancos de Dados

Um sistema de gerenciamento de banco de dados (SGBD) é um conjunto de programas que objetivam armazenar e recuperar coleções de dados inter-relacionados, chamadas de bancos de dados (Macário e Baldo, 2005).

Os SGDBs são projetados para o gerenciamento de grandes quantidades de dados. O gerenciamento envolve a definição de estruturas de armazenamento e manipulação de dados, possibilitando a sua interpretação e posterior construção de informação e conhecimento.

O tipo de SGBD mais usado é o relacional, em que os dados são modelados em tabelas que podem se relacionar entre si.

A Linguagem de Consulta Estruturada, SQL (do inglês, *Structured Query Language*), desenvolvida originalmente pela IBM, é a linguagem padrão para criação, manipulação e consultas a dados em sistemas gerenciadores de bancos de dados relacionais (Macário e Baldo, 2005).

2.1.1 SGBDs NoSQL

Com a crescente demanda de dados e suas diversas fontes, surge um novo modelo de sistema gerenciador, chamado NoSQL que implementa um dos seguintes modelos de dados: chave-valor, de documentos, de família de colunas ou de grafos. Esses modelos não usam a linguagem SQL de consulta nativa.

Os sistemas NoSQL foram propostos para atender a grandes volumes de dados semiestruturados ou não estruturados e que necessitem de alta disponibilidade e escalabilidade,

sendo, dessa forma, conhecidos por atender bem as aplicações *web* (de Souza *et al.*, 2014).

2.1.2 SGBDs de Grafos

O modelo de dados que foi utilizado para o banco de dados deste projeto é o de grafos, um tipo de modelo de dados usado em sistemas NoSQL (de Souza *et al.*, 2014).

Um grafo é uma coleção de vértices e arestas que, em se tratando de bancos de dados, representam entidades ou propriedades e os relacionamentos entre elas.

Os relacionamentos em um banco de dados em grafos são componentes principais, diferentemente do modelo relacional, em que é preciso inferi-los a partir de atributos em comum entre entidades.

Em contraste com os bancos de dados relacionais, em que o desempenho da consulta se deteriora conforme o aumento do conjunto de dados, nos bancos de dados de grafos o desempenho tende a permanecer constante, mesmo com o crescimento do conjunto de dados. Isso ocorre porque as consultas são localizadas em uma parte do gráfico, então o tempo de execução é proporcional apenas ao tamanho dessa parte e não ao tamanho total do grafo.

Dessa forma, esse tipo de banco de dados permite a construção de modelos sofisticados, mas que mapeiam de maneira simples o domínio utilizado, principalmente quando são tratadas grandes quantidades de dados inter-relacionados, como é o caso de dados de bibliotecas.

Além disso, os grafos permitem a adição de novos tipos de relacionamentos e novos nós sem perturbar as consultas existentes, garantindo menos migrações, manutenção e riscos (Robinson *et al.*, 2013).

Muitos sistemas gerenciadores de bancos de dados de grafos (SGBDGs) implementam um modelo de grafo em que nós e arestas podem ter um rótulo (que expressa seu tipo) e atributos. Dois nós ou arestas de um mesmo tipo podem ter atributos diferentes.

A Figura 2.1 mostra um exemplo de modelagem de dados em grafo de dados de acervos. Os círculos representam nós (entidades) do grafo e as setas representam relacionamentos. Cada cor dos nós representa um tipo de nó e dentro do círculo é apresentado um atributo do nós. Na Tabela 2.1, é possível ver a correspondência das cores, dos tipos de nós e do atributo em destaque.

Figura 2.1: *Exemplo de modelagem de dados de acervos.*

Cor	Tipo de Nó	Atributo
Amarelo	Item do acervo	Título do item
Azul	Autor secundário	Nome do autor secundário
Cinza	Tipo de material	tipo de material
Rosa	Autor primário	Nome do autor primário
Verde	Biblioteca	Nome da biblioteca
Vermelho	Assunto	Assunto

Tabela 2.1: *Cores e tipos de nós referentes à figura 2.1.*

2.2 Sistema de Catalogação

A tecnologia da informação é utilizada em vários domínios, sendo um deles o domínio bibliográfico e catalográfico (Assumpção e Santos, 2015).

O processo de catalogação é um mecanismo de suporte à informação que envolve a representação dos registros bibliográficos de documentos. Ele permite a identificação de obras e sua busca em unidades de informação, levando em consideração a diversidade de bases de dados.

A catalogação é também conhecida como representação descritiva, uma vez que fornece uma descrição única do documento a partir de regras pré-definidas. Dessa forma, ela é um conjunto de normas e procedimentos para a aquisição de uma informação e sua inserção em um catálogo, devendo manter a integridade, a clareza, a precisão, a lógica e a consistência (Fusco, 2012).

Na década de 1970, catálogos foram convertidos ao formato eletrônico e organizados em registros que foram padronizados para serem interpretados pelo computador. Surgiram,

então, os padrões catalográficos. Atualmente, os mais utilizados são a Catalogação Legível por Máquina - MARC 21 (do inglês, *Machine Readable Cataloging*), empregado neste trabalho, o Esquema de Metadados para a Descrição de Objeto - MODS (do inglês, *Metadata Object Description Standard*) e o Esquema de Metadados para a Descrição de Autoridade - MADS (do inglês, *Metadata Authority Description Schema)*). Todos eles foram criados para o uso com XML e os dois últimos são bastante compatíveis com MARC 21, embora não sejam tão específicos (Assumpção e Santos, 2015).

Esses padrões são utilizados pelos softwares de gerenciamento de bibliotecas, como o Aleph, fornecido pela Ex Libris e utilizado pela USP; Sophia, desenvolvido pela empresa Prima e utilizado pela Biblioteca Nacional; Pergamum, utilizado pela Pontifícia Universidade Católica do Rio de Janeiro; Alexandria, utilizado pela Biblioteca Pública de Niterói; e Biblivre, utilizado pela Biblioteca de Hidrogênio da COPPE/UFRJ(Santos, 2016).

2.2.1 MARC 21

Um registro **MARC** é um registro catalogável legível por computador, o que significa que o computador interpreta os dados que estão registrados em formato bibliográfico. O registro catalogável inclui informações sobre um item como descrição, assuntos relacionados, entre outras.

Um registro em formato bibliográfico representa como os dados estariam em um cartão catalográfico. Quando apresentadas no padrão MARC, essas informações são separadas por campos, que são representados por etiquetas de 3 dígitos. Cada campo possui uma regra de como deve ser preenchido, muitas vezes tendo um limite máximo de caracteres ou um tipo único de dado aceitável. Esses campos podem ser subdivididos em um ou mais "subcampos", representados por letras[1].

O exemplo abaixo apresenta dois campos MARC.

245 **|a** *Primeiras Estórias* **|c** *João Guimarães Rosa ; apresentação de Alberto da Costa e Silva ; ensaio de Paulo Rónai*
100 **|a** *Rosa, João Guimarães* **|d** *1908-1967*

A etiqueta '245' representa o campo título. As letras 'a' e 'c' representam, respectivamente, o título e o autor do item. A etiqueta '100' representa o campo autor. As letras 'a' e 'c', nesse caso, representam, respectivamente, o nome do autor e suas datas de nascimento e falecimento.

Por ser uma linguagem padronizada, muitas bibliotecas possuem sistemas que são projetados para funcionar com o formato MARC.

[1] https://www.loc.gov/marc/umb

2.2.2 MARCXML

O *Network Development* e MARC *Standards Office* desenvolveram uma estrutura para trabalhar com dados MARC em formato XML – o MARCXML[2].

O MARCXML, assim como o XML, é uma linguagem desenvolvida em esquema de árvore. Os principais elementos presentes em um arquivo MARCXML são:

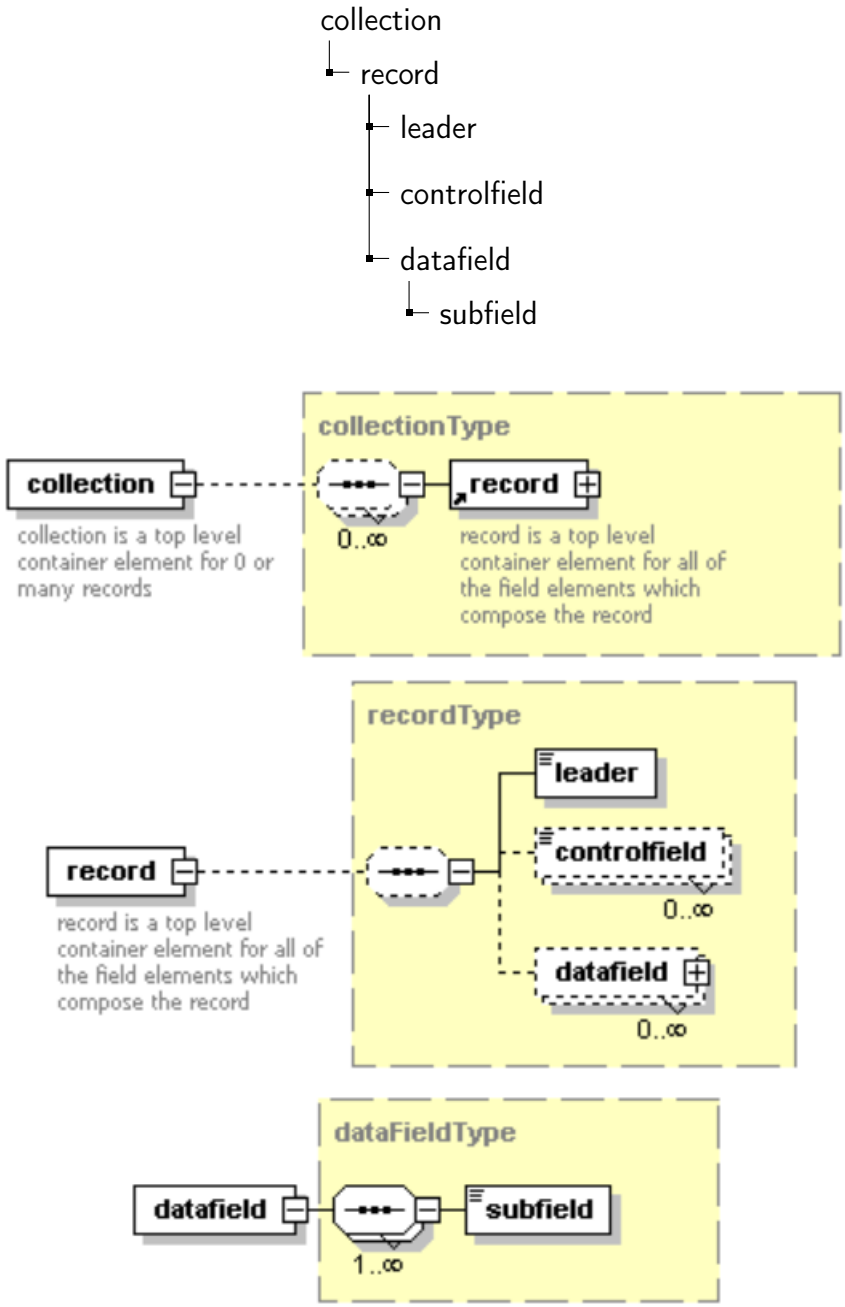

Figura 2.2: *Elementos presentes no MARCXML.*

O código abaixo mostra o exemplo 2.2.1 em código MARCXML. O elemento datafield indica, a partir de uma tag, qual é o campo, enquanto o elemento subfield adiciona as informações necessárias.

[2]https://www.loc.gov/marc/marcxml.html

```
<marc:datafield tag="245" ind2="0" ind1="1">
    <marc:subfield code="a">Primeiras estórias</marc:subfield>
    <marc:subfield code="c">João Guimarães Rosa ;
    apresentação de Alberto da Costa e Silva ;
    ensaio de Paulo Rónai</marc:subfield>
</marc:datafield>

<marc:datafield tag="100" ind2=" " ind1="1">
    <marc:subfield code="a">Rosa, João Guimarães</marc:subfield>
    <marc:subfield code="d">1908-1967</marc:subfield>
</marc:datafield>
```

2.3 Indexação de Texto

Um dos processos de recuperação da informação e que, em um banco de dados, tem como objetivo principal aumentar a eficiência das pesquisas é a indexação de texto, uma técnica de análise de conteúdo que condensa a informação significativa de um texto.

O processo consiste na atribuição de termos para a criação de uma linguagem que facilite a análise do documento pelo usuário, podendo auxiliar, também, em outros processos de análise. Pode ser feito pelo homem (indexação manual) ou pelo computador (indexação automática).

A indexação automática baseia-se na comparação de cada palavra de um documento com uma relação de palavras pré-estabelecidas, seguida pela eliminação de parte do texto e permanência das palavras significativas (Vieira, 1988).

A indexação automática pode ser realizada de diferentes formas, dependendo dos objetivos de cada estudo. Este trabalho emprega o método de índice invertido.

O objetivo principal do método de índice invertido é localizar rapidamente documentos que contenham palavras de uma consulta e, assim, classificar esses documentos por relevância (criando um *score*). O método armazena uma lista dos documentos contendo cada palavra, facilitando e aumentando a velocidade da busca. É criada uma lista invertida, que tem o formato $a \rightarrow X, Z$, ou seja, indicando que o termo a aparece nos documentos X e Z (Gupta *et al.*, 2009).

2.4 Sistema de Recomendação

Um banco de dados permite a execução de consultas e, no caso deste trabalho, busca de itens dos acervos das bibliotecas. Entretanto, para individualizar os resultados dessas pesquisas a fim de apresentar ao usuário itens nos quais ele poderia ter interesse ou que

podem ser relevantes para suas pesquisas, é necessária a elaboração de um sistema de recomendação sofisticado.

O objetivo de um sistema desse tipo é gerar recomendações a usuários sobre itens ou produtos que podem interessá-los, baseado nas suas preferências e na sua interação com o sistema. Sugestões de livros em sites de comércio eletrônico são um exemplo real de operação (Resnick e Varian, 1997).

A tarefa desse tipo de sistema é, portanto, transformar os dados sobre os usuários e sobre os nós e relacionamentos do banco de dados em previsões de possíveis interesses (Lü *et al.*, 2012).

Os tipos de recomendação mais comumente utilizados são:

- **Filtragem baseada em conteúdo (de item para item):** Recomendação de novos itens a partir do histórico de manifestação de interesse em itens já existentes (Sarwar *et al.*, 2001);
- **Filtragem colaborativa (de usuário para usuário):** Recomendação de itens a partir de como o usuário interage com o sistema (Huang *et al.*, 2002).

Em ambos os casos, funções de similaridade são utilizadas para corresponder perfis e itens e serão abordadas nos próximos capítulos do trabalho.

Capítulo 3

Dados de Itens de Acervos

Para amparar o projeto do sistema de integração e recomendação, foi utilizado um conjunto de dados de itens do acervo do Instituto de Estudos Brasileiros e um conjunto de dados de empréstimos de itens da Biblioteca Florestan Fernandes, ambas bibliotecas alocadas na Universidade de São Paulo.

Os dados de itens do IEB foram extraídos do Dedalus, que é o Catálogo Online das Bibliotecas da Universidade de São Paulo, um portal de busca integrada que possibilita o acesso de dados de acervos a partir de dados bibliográficos e também viabiliza empréstimos de itens para os alunos, professores e pesquisados da universidade[1]. A descrição de como os dados foram extraídos será apresentada na Seção 4.4.1.

O Dedalus é mantido pela Agência USP de Gestão de Informação Acadêmica (AGUIA), que é o órgão da USP responsável por alinhar a gestão da informação das bibliotecas aos objetivos estratégicos da instituição[2] A AGUIA utiliza um sistema integrado de bibliotecas criado pela empresa ExLibris, o Aleph[3].

No. Registro	001697746
Tipo de material	LIVRO
ISBN	9788525044648
Entrada Principal	Assis, Machado de 1839-1908
Título	**Dom Casmurro** / Machado de Assis ; fixação do texto e notas, Manoel Mourivaldo Santiago Almeida ; prefácio, John Gledson.
Imprenta	São Paulo : Globo, 2008.
Descrição	288 p.. : 21 cm
Série	(Globo Universidade)
Idioma	Português
Assunto	LITERATURA BRASILEIRA
	ROMANCE -- SÉCULO 19 -- BRASIL
Autor Secundário	Santiago-Almeida, Manoel Mourivaldo
	Gledson, John 1945-
Acervo Geral	Todos os itens
Itens na Biblioteca	EACH-Esc. Artes Ciências Hum.
Itens na Biblioteca	FFLCH-Fac. Fi. Let. C. Humanas
Itens na Biblioteca	IEB-Inst. Estudos Brasileiros
Itens na Biblioteca	MP-Museu Paulista

Figura 3.1: *Exemplo de informações exibidas no Dedalus sobre o item "Dom Casmurro", de Machado de Assis.*

[1]https://www.aguia.usp.br/bibliotecas/digitais-sistemicas/catalogo-dedalus
[2]https://www.aguia.usp.br/
[3]http://www.libnets.com/aleph.aspx

Na imagem 3.1, é possível ver as informações mostradas pelo Dedalus sobre um item de acervo a um usuário que é aluno da universidade. Para apoiar a implementação e verificação do sistema, mapeou-se os dados que seriam interessantes na modelagem do banco de dados, explorada no próximo capítulo.

As informações de um item são, portanto:

- Tipo de material (livro, tese ou separata, por exemplo);
- ISBN (International Standard Book Number): número que identifica internacionalmente um item;
- Título;
- Imprenta: nome da editora, cidade e ano de publicação;
- Assunto;
- Idioma;
- País;
- Autor primário: autor principal da obra ou primeiro autor identificado;
- Autor secundário: todos os autores além do principal, assim como tradutores e outras funções atribuídas;
- Descrição física: características físicas como dimensões e número de páginas;
- Nota.

O IEB não realiza empréstimos de itens do seu acervo a usuários, os seus itens não podem ser retirados da biblioteca. Para obter dados de associações entre itens e usuários (para a realização dos testes no sistema de recomendação), foram usados dados anonimizados de empréstimos de itens da Biblioteca Florestan Fernandes, da Faculdade de Filosofia, Letras e Ciências Humanas da USP. Nos dados de empréstimos cedidos, não há dados que permitam identificar a pessoa que fez o empréstimo. Cada registro associa um código numérico de usuário a um item de acervo que ele pegou emprestado.

Assim, para os testes do sistema, foram selecionados dados de itens que existem tanto na biblioteca do IEB quanto na biblioteca Florestan Fernandes. O conjunto de dados contém 1145 itens de acervo, 3881 usuários, 11378 empréstimos e as informações dos itens corresponderam a 1039 autores primários, 879 autores secundários, 1585 assuntos e 2 tipos de material.

Capítulo 4

Sistema de integração e recomendação

O sistema desenvolvido neste trabalho tem três componentes principais: um banco de dados de grafos, um subsistema de integração de dados de acervos de bibliotecas brasilianas e um subsistema de recomendação de itens de acervo a usuários. Este capítulo apresenta a modelagem do banco de dados e a descrição dos subsistemas, assim como as ferramentas utilizadas para a sua implementação e verificação.

4.1 Modelagem do Banco de Dados

Foi projetado um banco de dados para ser um repositório unificado dos principais dados catalográficos de itens de acervos das bibliotecas participantes do consórcio "Brasiliana Inteligente". Na modelagem, os dados foram organizados de modo a agilizar operações de busca de itens e facilitar a identificação de relacionamentos indiretos entre itens e usuários, para apoiar a personalização dos resultados das buscas e a recomendação de itens para os usuários.

O banco de dados não mantém todos os dados catalográficos existentes nos bancos das bibliotecas, uma vez que o propósito dele não é substituí-los. Ele foi criado para apoiar o sistema integrado de busca e recomendações e deverá ser alimentado regularmente com dados dos bancos de dados das bibliotecas participantes do consórcio.

A modelagem de um banco de dados orientado a grafos se dá pela definição dos tipos nós, dos tipos de relacionamentos entre nós e dos seus respectivos atributos (Penteado *et al.*, 2014).

4.1.1 Modelo do Banco de Dados

A partir das informações existentes no banco de dados catalográficos de bibliotecas, criou-se um modelo de banco de dados para o sistema, mostrado na figura 4.1.

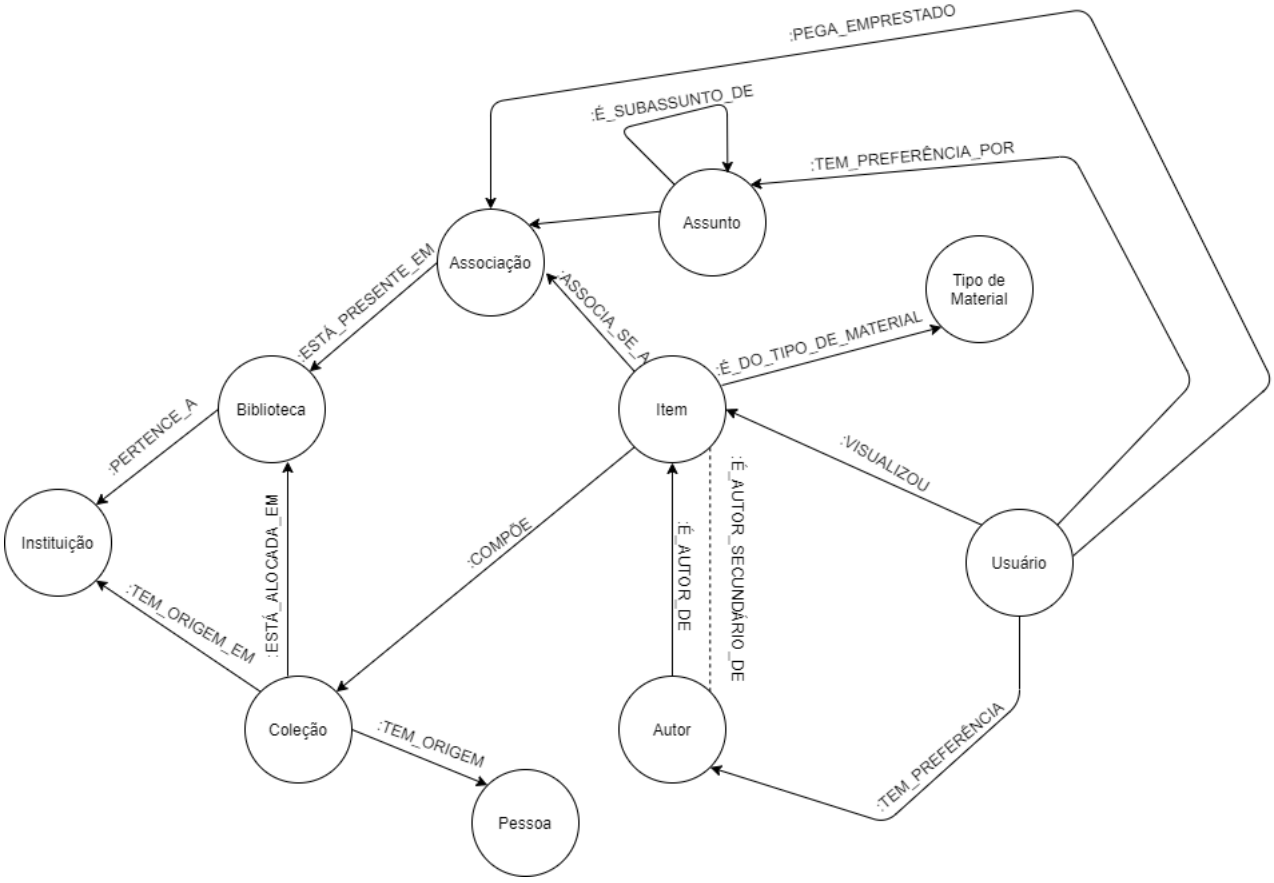

Figura 4.1: *Modelo do banco de dados para o sistema de integração e recomendação de itens de acervo.*

Nós

A tabela 4.1 mostra os tipos de nós definidos na modelagem e seus atributos.

Tipo de Nó	**Atributos**
Item	título, país, imprenta, isbn, nota, descrição física, idioma
Autor	nome, data de nascimento, data de falecimento
Autor Secundário	nome, data de nascimento, data de falecimento
Assunto	assunto
Tipo de material	tipo
Biblioteca	biblioteca
Usuário	usuário
Coleção	nome da coleção
Associação	edição, versão
Instituição	instituição
Pessoa	nome

Tabela 4.1: *Tipos de nós e atributos definidos na modelagem.*

Optou-se por modelar autores, assuntos e tipos de material como nós e não como atributos do tipo de nó Item a fim de estabelecer vínculos entre usuários e itens e entre itens e, assim, favorecer o desempenho de buscas e consultas que serão frequentes na aplicação.

Autores primários e secundários foram modelados em nós diferentes para destacar suas diferenças e facilitar o desempenho das buscas e recomendações, que priorizam buscas a usuários primários.

Devido à existência de itens iguais e itens com versões diferentes alocados em diversas bibliotecas, foi criado um tipo de nó nomeado Associação cujos atributos auxiliariam na identificação do exemplar.

Uma coleção é um conjunto de itens relacionados. Um exemplo de coleção de um mesmo assunto é a Coleção de Artes Visuais do IEB [1]. Uma coleção é adquirida pela biblioteca, seja por compra ou recebida em forma de doação por uma pessoa ou por uma instituição. Esses foram mapeados como nós no grafo. Os relacionamentos TEM_ORIGEM_EM e TEM_ORIGEM representam as doações e suas origens.

Relacionamentos

A partir dos tipos de nós, foi possível definir os tipos de relacionamentos do banco de dados. A tabela 4.4 apresenta os tipos de relacionamentos, os tipo de nós que participam dos relacionamentos e seus atributos.

[1] http://www.ieb.usp.br/sobre-o-ieb/colecao-de-artes-visuais

Tipo de Relacionamento	**Tipos de Nós Participantes**	**Atributos**
ESTÁ_ALOCADO_EM	Coleção, Biblioteca	-
PERTENCE_A	Biblioteca, Instituição	-
TEM_ORIGEM_EM	Coleção, Instituição	-
TEM_ORIGEM	Coleção, Pessoa	-
COMPÕE	Item, Coleção	-
PEGOU_EMPRESTADO	Item, Usuário	-
PERTENCE_AO_ASSUNTO	Item, Assunto	-
TEM_PREFERÊNCIA_POR	Usuário, Assunto	-
TEM_PREFERÊNCIA	Usuário, Autor	-
É_AUTOR_DE	Item, Autor	-
É_AUTOR_SECUNDÁRIO_DE	Item, Autor Secundário	-
É_DO_TIPO_DE_MATERIAL	Item, Tipo de Material	-
VISUALIZOU	Usuário, Item	-
É_SUBASSUNTO_DE	Assunto	-
SE_ASSEMELHA_A	Item	score
PODE_INTERESSAR_A	Item, Usuário	score

Tabela 4.2: *Tipos de relacionamentos, nós participantes e atributos da modelagem.*

4.1.2 Modelo Simplificado do Banco de Dados

A versão inicial do sistema de integração e recomendação implementada neste trabalho não se conecta diretamente aos BDs catalográficos das bibliotecas para a obtenção dos dados. Nela, os dados que alimentaram o BD de grafo foram extraídos do Dedalus por meio de sua interface web de consulta, que disponibiliza para usuários comuns apenas parte dos dados catalográficos dos itens (como descrito no Capítulo 3).

Por essa razão, foi criado um modelo simplificado do BD de grafo da figura 4.1, para ser usado nesta versão inicial do sistema. O modelo simplificado pode ser visto na figura 4.2.

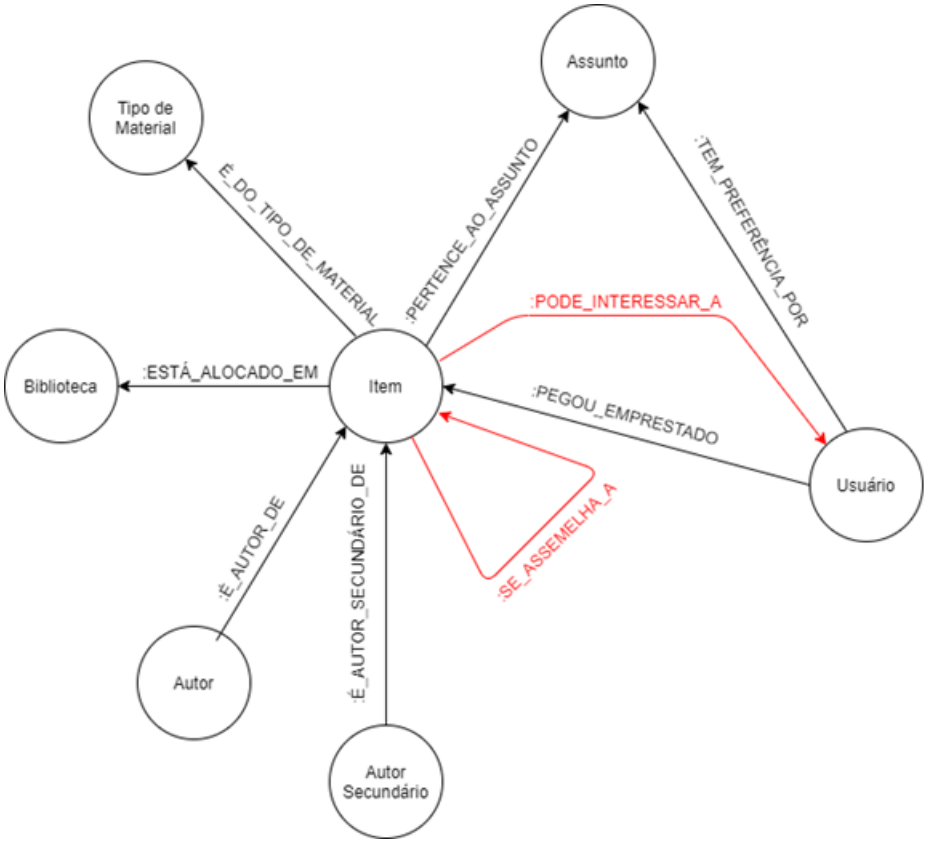

Figura 4.2: *Modelo simplificado do BD do sistema de integração e recomendação de itens de acervo. Os relacionamentos em preto são permanentes e, em vermelho, são temporários.*

Nós

Na tabela 4.3, verifica-se os tipos de nós que foram mantidos e que tiveram alterações no modelo simplificado.

Tipo de Nó	**Atributos**
Item	título, país, imprenta, isbn, nota, descrição física, idioma
Autor	nome
Autor Secundário	nome
Assunto	assunto
Tipo de material	tipo
Biblioteca	biblioteca
Usuário	usuário

Tabela 4.3: *Tipos de nós e atributos definidos no modelo simplificado do BD.*

Relacionamentos

Na tabela 4.4, verifica-se os tipos de relacionamentos que foram mantidos e os que tiveram alterações na versão simplificada do modelo inicial.

Para apoiar recomendações, foram incluídos no grafo do BD dois tipos de relacionamentos de caráter temporário. Os relacionamentos desses tipos são criados apenas durante a execução das recomendações e, após a exibição dos resultados, são apagados. A similaridade entre usuários não foi mapeada através de um relacionamento, pois é facilmente inferida através dos relacionamentos entres os usuários e os assuntos em comum.

Tipo de Relacionamento	**Nós Participantes**	**Caráter**	**Atributos**
ESTÁ_ALOCADO_EM	Item, Biblioteca	Permanente	-
PEGOU_EMPRESTADO	Item, Usuário	Permanente	-
PERTENCE_AO_ASSUNTO	Item, Assunto	Permanente	-
TEM_PREFERÊNCIA_POR	Usuário, Assunto	Permanente	-
É_AUTOR_DE	Item, Autor	Permanente	-
É_AUTOR_SECUNDÁRIO_DE	Item, Autor Secundário	Permanente	-
É_DO_TIPO_DE_MATERIAL	Item, Tipo de Material	Permanente	-
SE_ASSEMELHA_A	Item	Temporário	score
PODE_INTERESSAR_A	Item, Usuário	Temporário	score

Tabela 4.4: *Tipos de relacionamentos, nós participantes, caráter dos relacionamentos e atributos definidos no modelo simplificado do BD.*

4.2 Subsistema de Integração

Criou-se um sistema baseado em um banco de dados orientado a grafos para o armazenamento e manipulação de informações de itens de acervos das bibliotecas brasilianas.

O subsistema de integração é responsável por receber os dados de itens de acervos provenientes das bibliotecas e inseri-los ou atualizá-los no banco de dados de grafo. Ele recebe cada lote de dados em um arquivo no formato MARCXML, processando-o de maneira a extrair cada um dos registros catalográficos para inserção ou atualização do registro no banco de dados de grafos.

Um único registro é decomposto em nós e relacionamentos. Para que um mesmo registro não seja incluído mais de uma vez e para que nós existentes sejam utilizados, é necessário verificar a existência deles ou sua necessidade de criação. Essa verificação ocorre através dos atributos. Por exemplo, utiliza-se o ISBN para verificar itens iguais.

4.2.1 Formato de Entrada de Dados no Sistema

Os dados dos acervos devem ser disponibilizados ao sistema de integração no formato MARCXML com as informações dos itens dos acervos. Esse formato foi selecionado em função de ser padronizado e utilizado por muitos sistemas gerenciadores de biblioteca. Os itens devem conter informações mínimas para sua identificação, como o ISBN.

4.2.2 Inserção dos Dados no BD de Grafos

Os dados dos registros em MARCXML são processados através de um *parser* XML e as informações extraídas são adicionadas ao banco de dados através de um conector que cria e atualiza nós e relacionamentos.

4.3 Subsistema de Recomendação

Para apoiar as recomendações idealizadas para o sistema, foram definidas funções de similaridade entre itens e entre usuários, com base nos dados armazenados no banco de dados de grafo.

4.3.1 Tipos de Similaridade

Similaridade entre Usuários

Dois usuários são considerados similares quando têm interesse pelos mesmos assuntos. Um exemplo é mostrado na figura 4.3, em que é possível ver os usuários "7798" e "2127", que são similares por terem interesse em comum pelos assuntos "MODERNISMO", "ARTE MODERNA" e "ARTES".

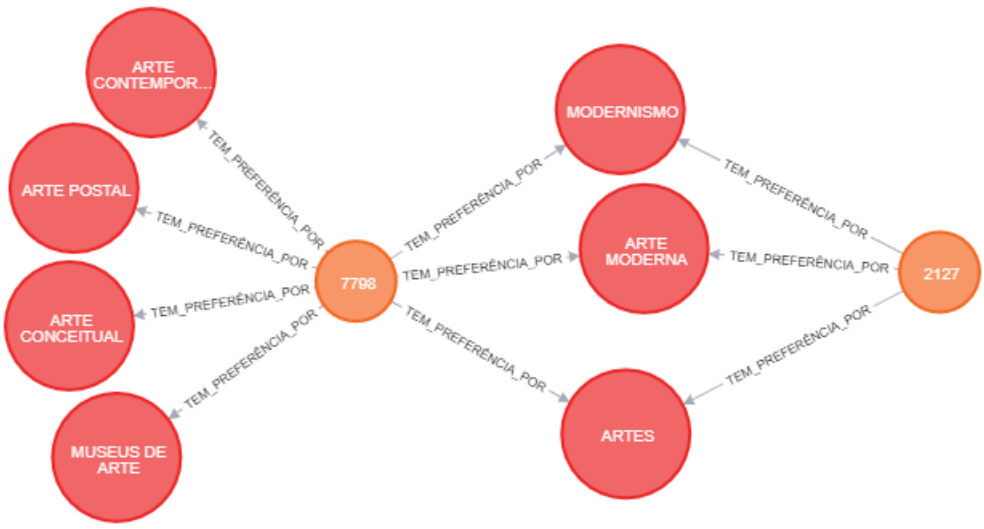

Figura 4.3: *Usuários "7798" e "2127".*

Similaridade entre Itens

Dois itens são considerados similares quando têm em comum:

- Autor primário;
- Autor(es) secundário(s);
- Assunto(s);

- Tipo de material;
- Biblioteca;
- Indexação de título;
- Indexação de nota;

Assim, os relacionamentos considerados nessa similaridades são:

- Autor É_AUTOR_DE Item;
- Autor Secundário É_AUTOR_SECUNDÁRIO_DE Item;
- Item PERTENCE_AO_ASSUNTO Assunto;
- Item É_DO_TIPO_DE_MATERIAL Tipo de material;
- Item ESTÁ_ALOCADO_EM Biblioteca;

Um exemplo é mostrado na figura 4.4, em que os itens de títulos "Teoria da vanguarda" e "Poéticas do processo arte conceitual no Museu" são similares por ambos pertencerem ao assunto "ARTES", serem do tipo de material "LIVRO" e estarem alocados na biblioteca "IEB Instituto de Estudos Brasileiros".

Figura 4.4: *Itens de títulos "Teoria da vanguarda" e "Poéticas do processo arte conceitual no Museu".*

4.3.2 Cálculo do Grau de Similaridade

Uma vez definidos os tipos de similaridade, um questionamento surge: como definir graus de similaridade?

Para expressar o grau de similaridade, foram definidos *scores* que, uma vez calculados, são guardados no BD de grafo como atributos de relacionamentos temporários que ligam os pares de nós considerados similares.

Usuários são considerados similares quando têm preferência por pelo menos um assunto em comum. Não são definidos graus de similaridade entre usuários, mas utiliza-se essa similaridade para outros cálculos.

Score da Similaridade entre um Par de Itens

Além do *score* calculado a partir do conceito de similaridade entre itens, foram calculados, também, *scores* dos graus de associação entre um usuário e um item, para auxiliar nas recomendações.

Pode haver diferentes formas de ligar um item a outro por um caminho com apenas duas arestas, como visto exemplo da Figura 4.4. Por isso, foram definidos, a partir de testes, diferentes pesos para os tipos de relacionamentos de item.

Dessa forma, o cálculo do *score* foi definido como:

$$Score = \frac{8 * SA1 + 3 * SA2 + 3 * SA + 1 * SM + 1 * SB + 5 * ST + 1 * SN}{22},$$

em que:
SA1 = score de autor primário (1 se os itens tiverem autor primário em comum e 0, caso contrário);
SA2 = score de autor(es) secundários (1 se os itens tiverem autor secundário em comum e 0, caso contrário);
SA = score de assuntos (quantidade de assuntos em comum); SM = score de tipo de material (1 se os itens forem do mesmo tipo de material e 0, caso contrário);
SB = score de biblioteca (1 se os itens estiverem alocados na mesma biblioteca e 0, caso contrário);
ST = score da indexação dos títulos dos itens;
SN = score da indexação das notas dos itens.

A indexação de título e nota é realizada com o auxílio de uma ferramenta de indexação de textos (o Apache Lucene) a partir de lista invertida. Logo, o score das indexações é utilizado conforme retornado pela ferramenta.

Score de Associação entre um Usuário e um Item

Foram definidos dois tipos de *score* para esse tipo de associação, baseados nos dois tipos de recomendação definidos: recomendação de item a partir do perfil do usuário e recomendação de item a partir de termos em conjunto ao perfil do usuário.

- Score para a recomendação de item a partir do perfil do usuário (atributo do relacionamento PODE_INTERESSAR_A)

  No exemplo na Figura 4.3, nota-se que a distância entre um usuário e seu assunto de preferência é de uma aresta (relacionamento TEM_PREFERÊNCIA_POR). Por outro lado, a distância entre um usuário e um usuário similar a ele são duas arestas do relacionamento TEM_PREFERÊNCIA_POR. Apesar de as arestas serem

direcionadas, o banco de dados em grafos permite a navegação em qualquer sentido do relacionamento.

O cálculo do *score* baseou-se nessas distâncias, dando um peso maior para a menor distância. Assim, a equação para esse cálculo foi definida como:

$$\text{Score} = 2 * \text{AC} + 1 * \text{US},$$

em que:
AS = quantidade de assuntos relacionados ao mesmo tempo ao usuário e ao item;
US = quantidade de usuários similares ao usuário que pegaram emprestado o item.

- Score para a recomendação de item a partir de termos e a partir do perfil do usuário

  Esse cálculo leva em consideração:

  - Assuntos relacionados ao item e ao usuário;
  - Itens que o usuário pegou emprestados;
  - Itens que usuários similares ao usuário pegaram emprestados.

  Dessa forma, o *score* é calculado da seguinte forma:

$$\text{Score} = 2 * \text{AC} + 1 * \text{SA1} + 1 * \text{SA2} + 2 * \text{TC} + 1 * \text{US},$$

  em que:
  AC = quantidade de assuntos em comum entre o item e as preferências do usuário;
  SA1 = score de autor primário (1 se o item tiver autor primário em comum com algum item pego emprestado pelo usuário e 0, caso contrário);
  SA2 = score de autor(es) secundários (1 se o item tiver autor secundário em comum com algum item pego emprestado pelo usuário e 0, caso contrário);
  TC = quantidade de termos em comum entre o título do item e títulos de itens pegos emprestado pelo usuário;
  US = quantidade de usuários similares que pegaram emprestado o item.

4.3.3 Tipos de Recomendações

Conforme mencionado no Capítulo 2, o sistema de recomendação desenvolvido neste trabalho emprega duas estratégias de recomendação: filtragem baseada em conteúdo e filtragem colaborativa. Elas foram utilizadas tanto separadamente quanto em conjunto.

Foram criados quatro tipos de recomendação para os usuários, que são descritos a seguir.

Recomendação a Partir do Perfil do Usuário

A recomendação baseada nas preferências dos usuários foi idealizada para ser utilizada quando um usuário nunca pegou emprestado um item ou tem interesse de conhecer itens. Para isso, a recomendação é baseada nos relacionamentos entre usuários e itens e entre usuários e assuntos (que indicam suas preferências). São considerados também, portanto, usuários similares.

Dois relacionamentos geram insumo à recomendação:

- Usuário PEGOU_EMPRESTADO Item;
- Usuário TEM_INTERESSE_POR Assunto.

Recomendação a Partir de um Item

Essa recomendação ocorre quando o usuário visualiza um item. A partir dele, outros itens são indicados como de possível interesse.

A recomendação leva em consideração relacionamentos que esse item tem no banco de dados. É uma função que retorna uma lista de dez itens relacionados em ordem decrescente de score, que foi definido a partir da similaridade de itens, abordada anteriormente. A quantidade de dez itens retornados pela recomendação foi definido em função da quantidade de itens alocados no banco de dados. Entretanto, esse número pode ser modificado, pois não há nenhuma restrição de implementação ligada a ele.

Busca a Partir de Termos

Quando um usuário faz uma busca a partir de termos, a recomendação leva em consideração a indexação de texto feita pela ferramenta de indexação.

Foram implementados quatro tipos de busca, que baseiam-se nas seguintes indexações:

- Indexação por título;
- Indexação por autor;
- Indexação por biblioteca;
- Indexação sem definir campo de busca.

Quando um usuário realiza uma busca sem definir um campo de pesquisa, o sistema executa os três outros tipos de busca (na ordem por título, por autor e por biblioteca) e concatena os seus resultados como saída, retornando um corte de 10 itens (definido conforme explicado anteriormente), mas possibilitando a mudança desse número. Caso a saída não tenha 10 itens, são adicionados a ela itens alocados na mesma biblioteca que os outros da lista.

Busca a Partir de Termos Baseada no Perfil do Usuário

O sistema também realiza um outro tipo de busca a partir de termos que considera as preferências e empréstimos do usuário, para devolver respostas mais personalizadas.

A busca baseada no perfil do usuário utiliza como base a saída da busca simples e reorganiza a lista de itens da resposta conforme os relacionamentos entre os itens e o usuário. Dessa forma, quando a lista de itens é retornada pela busca simples, a busca complexa calcula um score entre os itens e o usuário baseado nos seguintes relacionamentos:

- Se o item tem algum assunto em comum com as preferências do usuário;
- Se o item tem algum autor em comum com algum item que o usuário pegou emprestado (tanto autor primário quanto secundário);
- Se o título do item tem algum termo que aparece no título de um item que o usuário pegou emprestado;
- Se o item foi pego emprestado por usuários similares ao usuário.

4.4 Ferramentas Utilizadas

4.4.1 Aquisição de Dados para a Verificação do Sistema

Os dados do IEB usados nos experimentos para verificação do sistema foram coletados de forma automatizada do site do Dedalus a partir de um *crawler*, que é um programa criado para automatizar a análise do código fonte de uma página *web* e dela extrair informações. O *crawler* foi implementado como um *script* Python e utilizou a biblioteca Selenium (Pypi, 2018). O código desenvolvido possibilitou o download de arquivos com informações de itens de acervos no formato MARC.

As informações sobre a alocação de itens não é apresentada no formato MARC, então, quando houver inclusão de itens por uma biblioteca, ela deve se identificar para que essa informação seja guardada no banco de dados.

Para converter os dados em MARC em arquivos MARCXML (formato de entrada do subsistema de integração de dados deste trabalho), foi criado um código Python com a biblioteca PyMarc.

Na verificação do sistema de recomendação, também foram usados dados anonimizados de empréstimos da Biblioteca Florestan Fernandes da Faculdade de Filosofia, Letras e Ciências Humanas, cedidos pela divisão de apoio tecnológico da AGUIA-USP.

4.4.2 Leitura dos Dados

No subsistema de integração, os dados no formato MARCXML são lidos através do *parser* XML e adicionados em um dicionário.

Os dados em MARCXML são processados usando a API xml.etree.ElementTree do Python. Essa API é um *parser* XML e implementa uma estrutura em árvore.

Dessa forma, foi criado um *script* que, com auxilio do módulo citado, lesse informações e as guardasse para a criação dos nós e relacionamentos.

A leitura do arquivo MARCXML foi feita através da função *parse*. O elemento *collection* é a raiz da árvore criada pelo módulo. Cada filho da árvore é um registro (*record*) da biblioteca, ou seja, um item do acervo.

Dentro de cada *record*, existem elementos leader, controlfield e datafield com tags. As tags representam informações do item. Por exemplo, a tag "100" representa o autor da obra.

A partir da função *attrib.get* foi possível identificar as *tags* e, a partir da função *find* foi possível encontrar os valores atribuídos a esses campos. Esses valores foram armazenados em um dicionário para serem utilizados posteriormente na criação dos nós e relacionamentos.

Inserção dos Dados

Os dados armazenados em dicionários são inseridos no banco de dados orientado a grafos a partir da biblioteca Python Py2neo, ideal para trabalhar com o gerenciador Neo4j dentro de aplicações Python (Py2neo, 2020).

4.4.3 Gerenciamento do Banco de Dados

O sistema de gerenciamento de bancos de dados de grafos usado no trabalho é o Neo4j. Para bancos de dados de grafos, ele é o sistema mais usado na atualidade. O Neo4j é um sistema transacional compatível com modelo ACID. Um sistema transacional ACID é aquele que garante às transações que gerencia quatro propriedades – Atomicidade, Consistência, Isolamento e Durabilidade (do inglês *Atomicity, Consistency, Isolation, Durability*). Esse é o modelo transacional implementado pelos sistemas gerenciadores de bancos de dados relacionais[2].

Indexação de Texto

Apache Lucene é um projeto de código aberto e uma biblioteca para pesquisa de texto completo com alto desempenho, escrito em Java. (Lucene Apache, 2010) Ele é o provedor base de indexação do Neo4j e, por isso, foi utilizado nesse trabalho (Santos e Silva, 2013).

[2]https://neo4j.com/docs/cypher-manual/current/administration/indexes-for-search-performance

4.5 Repositório dos códigos implementados

Os códigos implementados na construção do sistema estão disponíveis em https://github.com/Brun

Capítulo 5

Experimentos e análises

Para verificar o sistema implementado, foram realizados experimentos com um conjunto de dados do acervo do Instituto de Estudos Brasileiros extraídos da base catalográfica do Dedalus e dados de empréstimos da Biblioteca Florestan Fernandes. A estrutura desses dados e suas características foram descritas no capítulo 3.

5.1 Integração e Consultas

Foram levantadas algumas informações sobre os dados do banco a partir de consultas. A figura 5.1 apresenta os dez assuntos mais relacionados a itens no grafo.

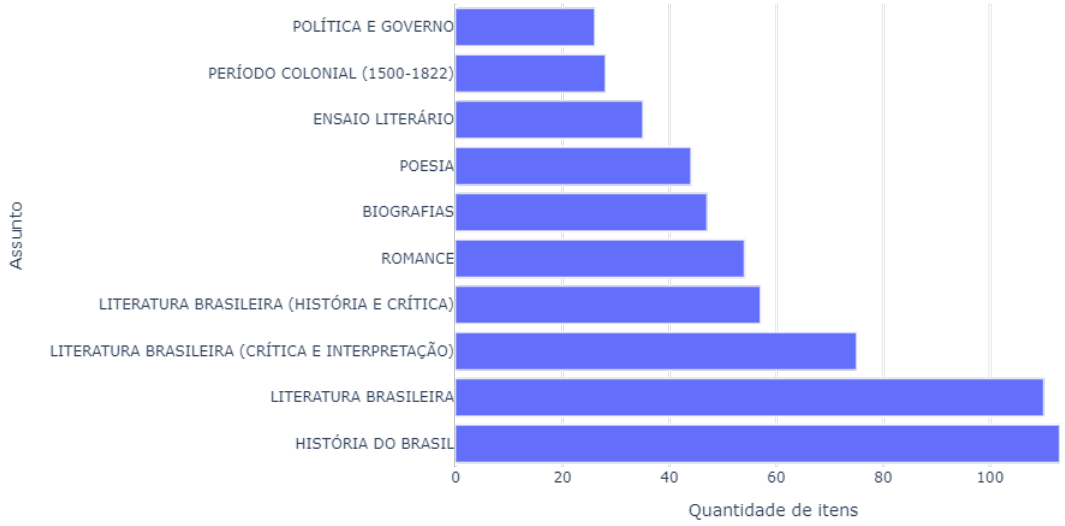

Figura 5.1: *Os dez assuntos mais relacionados a itens no grafo.*

É possível ver que os assuntos mais presentes no grafo estão relacionados à história e literatura brasileira, o que é esperado, uma vez que estamos tratando do acervo de uma biblioteca brasiliana.

A figura 5.3 apresenta os dez autores primários mais relacionados a itens no grafo.

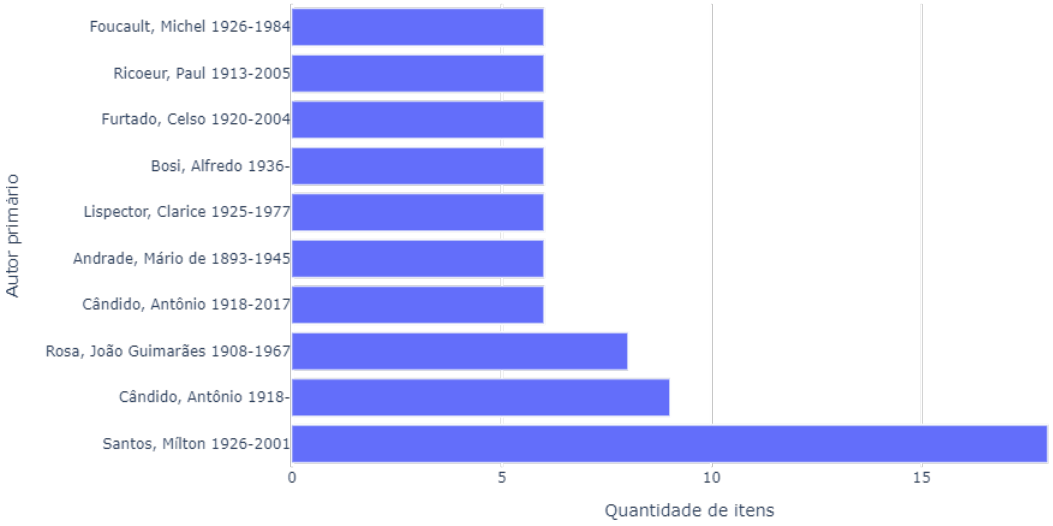

Figura 5.2: *Os dez autores primários mais relacionados a itens no grafo.*

A figura 5.3 apresenta os dez autores secundários mais relacionados a itens no grafo.

Figura 5.3: *Os dez autores secundários mais relacionados a itens no grafo.*

5.2 Recomendação de Itens a Partir do Perfil do Usuário

Para demonstrar o resultado da recomendação de itens a partir do perfil do usuário, foi selecionado o usuário 7514, que pegou emprestado três itens e têm preferência por 13 assuntos. A figura 5.4 mostra o usuário e seus relacionamentos no banco de dados.

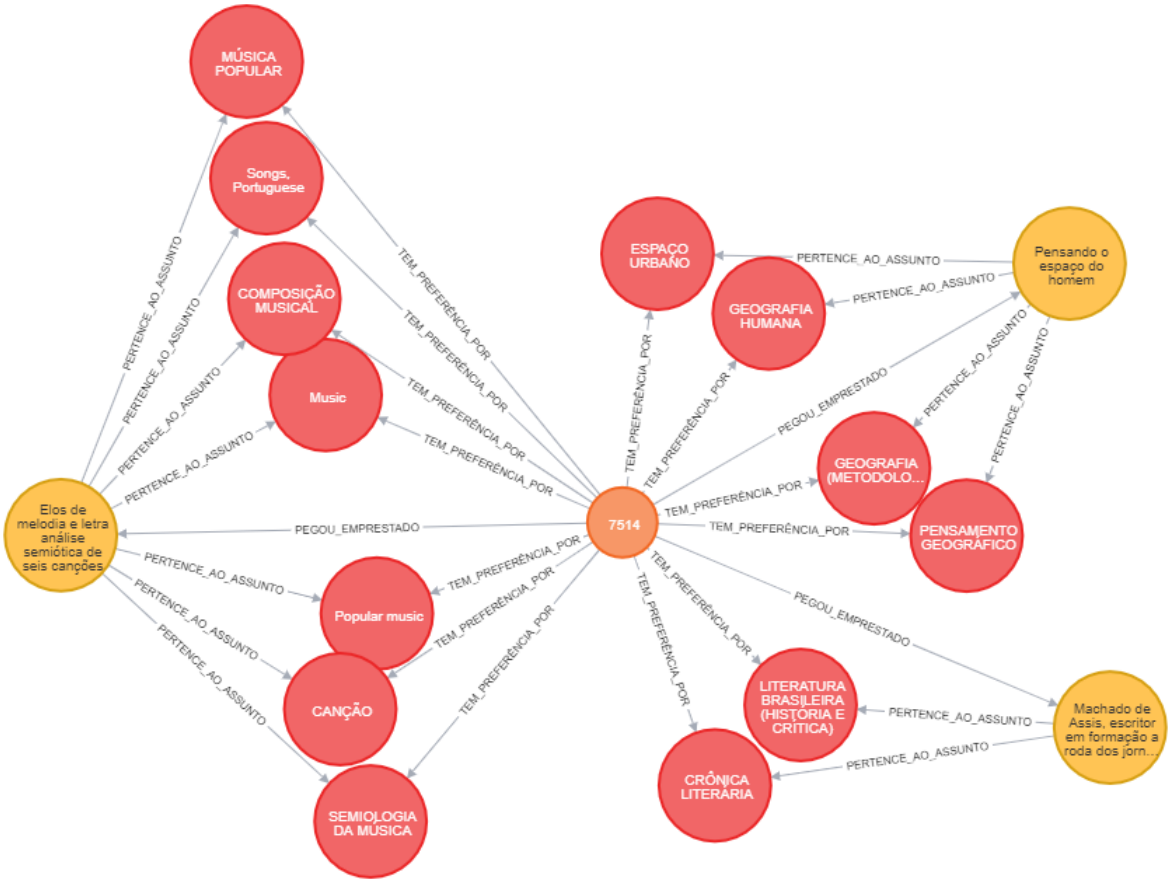

Figura 5.4: *Usuário 7514 e seus relacionamentos.*

O resultado da recomendação de itens para esse usuário é mostrado na tabela 5.1:

Índice	Título	Score
1	Machado de Assis o enigma do olhar	209
2	1930 a crítica e o Modernismo	73
3	Elos de melodia e letra análise semiótica de seis canções	50
4	A caminho do encontro uma leitura de Contos novos	45
5	Espaço e método	42
6	Pensando o espaço do homem	42
7	Metamorfoses do mal uma leitura de Clarice Lispector	39
8	Iniciação à literatura brasileira	36
9	O cacto e as ruínas a poesia entre outras artes	32
10	Folhetim uma história	29

Tabela 5.1: *Resultado da recomendação a partir do perfil do usuário 7514.*

Na figura 5.5, é possível ver que os dois itens mais recomendados ao usuário estão igualmente relacionados ao assunto "LITERATURA BRASILEIRA (HISTÓRIA E CRÍTICA)". Entretanto, existem 207 usuários similares a 7514 que pegaram emprestado o item de índice 1 e apenas 71 usuários similares que pegaram emprestado o item de índice 2.

Uma vez que o score leva em consideração os assuntos que estão relacionados ao usuário e ao item e a quantidade de usuários similares ao usuário que pegaram emprestado o item, o item "Machado de Assis o enigma do olhar" é mais recomendado ao usuário em questão do que "1930 a crítica e o Modernismo".

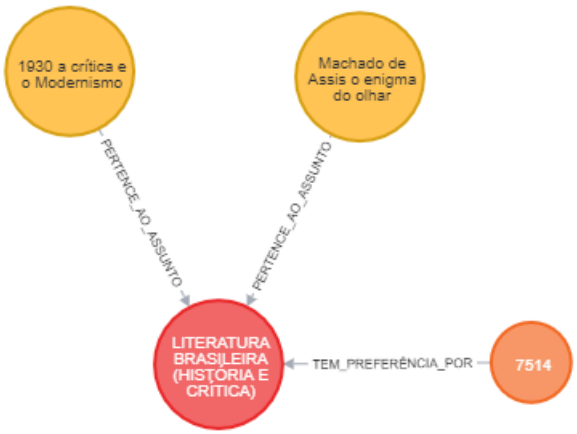

Figura 5.5: *Usuário 7514 e seus relacionamentos com os itens de índices 1 e 2 do resultado.*

5.3 Recomendação a Partir de um Item

Para demonstrar essa recomendação a partir de um item, foi selecionado o item de título "Machado de Assis o enigma do olhar". É possível verificar o nó desse item na figura 5.6.

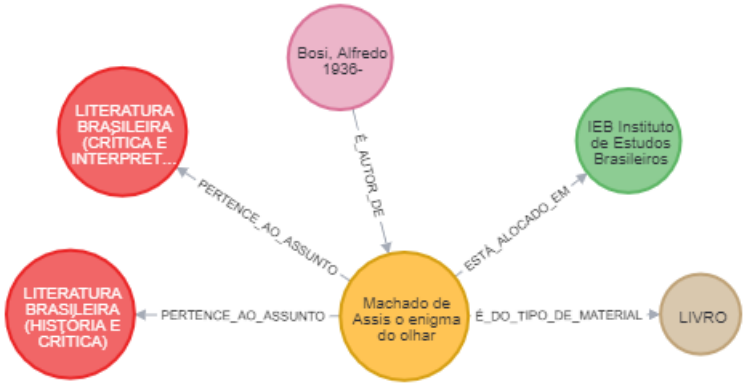

Figura 5.6: *Item de título "Machado de Assis o enigma do olhar"*

Foram criados relacionamentos temporários "SE_ASSEMELHA_A" relacionando todos os itens do banco de dados com o item em questão. Cada relacionamento possui um score que é calculado levando em consideração os relacionamentos existentes no banco, conforme apresentado no Capítulo 4.

O resultado é, portanto, uma lista de 10 itens ordenados de forma decrescente conforme o score. É possível ver o resultado na tabela 5.2.

Índice	Título	Score
1	Machado de Assis um gênio brasileiro	1.864
2	Machado de Assis e a crítica internacional	1.728
3	Três vezes Machado de Assis	1.591
4	Por um novo Machado de Assis ensaios	1.591
5	Machado e Borges e outros ensaios sobre Machado de Assis	1.591
6	Machado de Assis, historiador	1.591
7	A mulher e a cidade imagens da modernidade brasileira em quatro escritoras paulistas	1.591
8	Fábrica de contos ciência e literatura em Machado de Assis	1.455
9	Natureza e cultura no Brasil (1870-1922)	1.455
10	Machado de Assis, escritor em formação a roda dos jornais	1.455

Tabela 5.2: *Resultado da recomendação a partir do item de título "Machado de Assis o enigma do olhar".*

Na figura 5.7, é possível ver os diferentes relacionamentos entre o item e os itens de título "Machado de Assis um gênio brasileiro", que possui o maior score (1.864), e "História econômica e social do estado de São Paulo", que possui o menor score (0.062) e, por isso, não entrou nos 10 primeiros resultados.

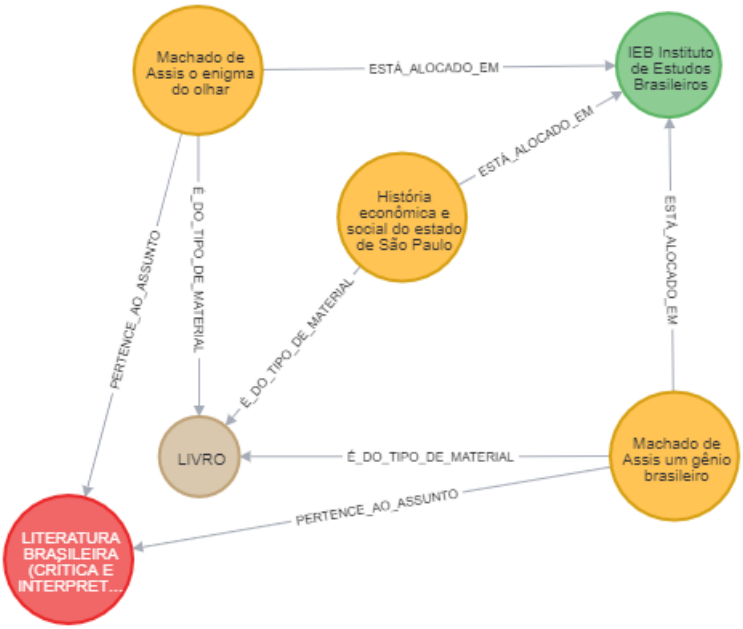

Figura 5.7: *Item de título "Machado de Assis o enigma do olhar e seus relacionamentos com os itens de maior e menor score".*

Verifica-se na imagem que, ambos itens de maior e menor score estão relacionados ao item em questão por serem do mesmo tipo de material (livro) e estarem alocados na

mesma biblioteca (IEB Instituto de Estudos Brasileiros). Entretanto, o item de maior score também está relacionado ao item em questão pelo assunto "LITERATURA BRASILEIRA (CRÍTICA E INTERPRETAÇÃO)". Além disso, percebe-se que os títulos têm mais termos em comum entre si.

5.4 Busca a Partir de Termo

A busca a partir de termo pode ser feita de quatro formas diferentes, selecionando um campo para a pesquisa ou não. A seguir, é possível ver como cada busca performa. Quando um campo para pesquisa é selecionado, a busca baseia-se na indexação dos itens e retorna todos os itens que possuem score produzido pelo Lucene. Quando nenhum campo para pesquisa é selecionado, são feitos os três tipos de busca (a partir do título, do autor e da biblioteca), concatenando-os até formarem 10 itens no resultado. Entretanto, se houver menos de 10 itens recomendados, é selecionada uma biblioteca em que esteja alocado um dos itens recomendados e são adicionados ao resultado itens alocados nela.

5.4.1 Busca a Partir do Título

Para demonstrar essa busca, utilizou-se o termo "Alencar" e especificou-se que a busca deveria ser feita a partir do título.

Os resultados da busca são mostrados na tabela 5.3.

Índice	**Título**	**Score**
1	José de Alencar o poeta armado do século XIX	2.362
2	A fonte subterrânea José de Alencar e a retórica oitocentista	2.362
3	Paraísos artificiais o romantismo de José de Alencar e sua recepção crítica	2.206
4	Mulheres de papel um estudo do imaginário em José de Alencar e Machado de Assis	2.069
5	O inimigo do rei uma biografia de José de Alencar, ou, a mirabolante aventura de um romancista que colecionava desafeto, azucrinava d. Pedro II e acabou inventando o Brasil	1.327

Tabela 5.3: *Resultado da busca a partir do termo "Alencar" no título.*

Na tabela 5.3, é possível ver que o score diminui conforme o aumento do tamanho título, uma vez que a indexação do texto leva em consideração a quantidade de termos em comum.

5.4.2 Busca a Partir do Autor

Para demonstrar essa busca, utilizou-se o termo "Alencar" e especificou-se que a pesquisa deveria ser feita a partir do autor.

Os resultados da busca são mostrados na tabela 5.4:

Índice	Título
1	Cartas a favor da escravidão
2	O guarani
3	Ao correr da pena

Tabela 5.4: *Resultado da busca a partir do termo "Alencar" especificando a pesquisa por autor.*

A busca relaciona o termo "Alencar" aos autores presentes no banco e retorna todos os itens relacionados a esses autores. Dessa forma, não são calculados scores.

5.4.3 Busca a Partir da Biblioteca

Assim como nos exemplos anteriores, buscou-se o termo "Alencar", mas desta vez especificando que a pesquisa deveria ser feita a partir de bibliotecas. Entretanto, não há nenhuma biblioteca que contenha no nome o termo "Alencar" no banco de dados e, por isso, não houve nenhum item no resultado.

Então, para demonstrar essa busca, utilizou-se o termo "Brasileiros". O resultado foi a lista de todos os itens presentes na biblioteca do IEB (Instituto de Estudos Brasileiros).

5.4.4 Busca a Partir de Termo sem Especificar a Pesquisa

Como explicado anteriormente, a busca sem especificar título, autor ou biblioteca leva em consideração as três buscas.

O resultado obtido para o termo "Alencar" está na tabela 5.5.

Índice	Título
1	José de Alencar o poeta armado do século XIX
2	A fonte subterrânea José de Alencar e a retórica oitocentista
3	Paraísos artificiais o romantismo de José de Alencar e sua recepção crítica
4	Mulheres de papel um estudo do imaginário em José de Alencar e Machado de Assis
5	O inimigo do rei uma biografia de José de Alencar, ou, a mirabolante aventura de um romancista que colecionava desafeto, azucrinava d. Pedro II e acabou inventando o Brasil
6	Cartas a favor da escravidão
7	O guarani
8	Ao correr da pena
9	Política de botinas amarelas o MDB-PMDB paulista de 1965 a 1988
10	Clientelismo e política no Brasil do século XIX

Tabela 5.5: *Resultado da busca a partir do termo "Alencar" sem especificar campo para a pesquisa.*

É possível identificar, na tabela 5.5, os resultados a partir do título nos índices 1 a 5 e os resultados a partir do autor nos índices 6 a 8. Os itens 9 e 10 são itens alocados no IEB como os outros itens do resultado (eles foram adicionados porque o resultado teria menos de 10 itens).

5.5 Busca de um Termo Baseada no Perfil de um Usuário

Conforme explicado no Capítulo 4, a busca de um termo baseada no perfil de um usuário leva em consideração a busca do termo especificado e seu resultado de 10 itens. A partir dessa lista, a recomendação ordena os 10 itens conforme um score calculado a partir dos relacionamentos do usuário com os itens em questão.

Para essa recomendação, são criados relacionamentos temporários entre item e usuário denominado "PODE_INTERESSAR_A" em que o score calculado é atributo.

Para demonstrar essa recomendação, será comparado o resultado da busca baseada no termo "Romance brasileiro moderno" (tabela 5.6) com o resultado da mesma busca baseada no usuário 22388.

A tabela 5.7 mostra o resultado da busca pelo termo "Romance brasileiro moderno".

Índice	Título
1	Escritos sobre arte e modernismo brasileiro
2	Poesia concreta brasileira as vanguardas na encruzilhada modernista
3	Poesia, mito e história no modernismo brasileiro
4	A tradição regionalista no romance brasileiro, 1857-1945
5	Uma história da música popular brasileira das origens à modernidade
6	Agosto romance
7	Lendas e romances
8	Modernismo
9	Moderno e Brasileiro a história de uma nova linguagem na arquitetura (1930-60)
10	A mulher e a cidade imagens da modernidade brasileira em quatro escritoras paulistas

Tabela 5.6: *Resultado da busca a partir do termo "Romance brasileiro moderno" sem especificar a pesquisa.*

Quando o mesmo termo é buscado para um usuário que pegou emprestado 2 itens e possui interesse por 6 assuntos (visto na figura 5.8), o resultado obtido é o apresentado na tabela 5.7.

Índice	Título	Score
1	Agosto romance	15
2	Lendas e romances	4
3	Poesia concreta brasileira as vanguardas na encruzilhada modernista	2
4	A tradição regionalista no romance brasileiro, 1857-1945	2
5	Escritos sobre arte e modernismo brasileiro	1
6	A mulher e a cidade imagens da modernidade brasileira em quatro escritoras paulistas	1
7	Poesia, mito e história no modernismo brasileiro	0
8	Uma história da música popular brasileira das origens à modernidade	0
9	Modernismo	0
10	Moderno e Brasileiro a história de uma nova linguagem na arquitetura (1930-60)	0

Tabela 5.7: *Resultado da busca a partir do termo "Alencar".*

A imagem 5.8 mostra o usuário e seus relacionamentos.

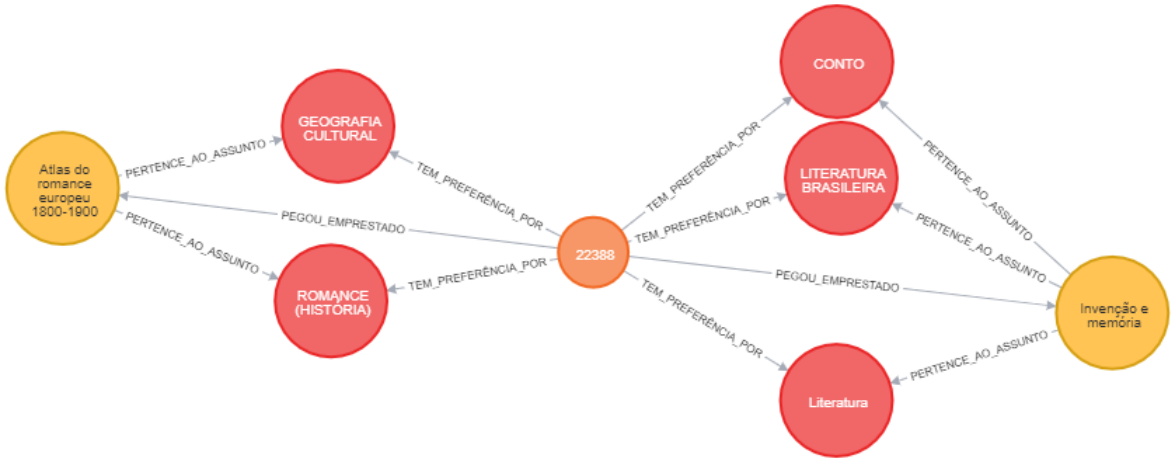

Figura 5.8: *Usuário 22388 e seus relacionamentos.*

A partir do resultado da tabela 5.7, será apresentado item a item e o cálculo de seus scores.

Item de índice 1: Agosto romance

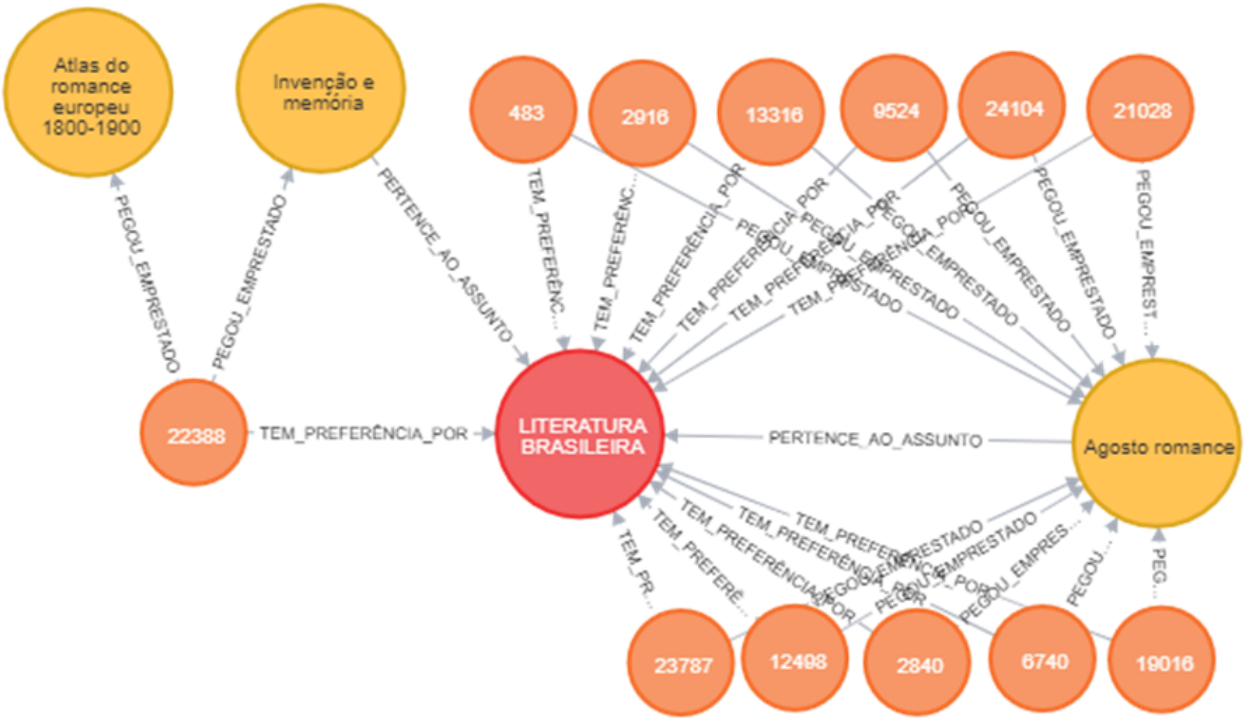

Figura 5.9: *Usuário 22388 e item de título "Agosto romance".*

Na figura 5.9, é possível verificar os relacionamentos que influenciaram no cálculo do score. São eles:

- Assunto "LITERATURA BRASILEIRA" como preferência de usuário e relacionado ao item;

- 11 usuários similares ao usuário 22388 (possuem interesse em comum pelo assunto "LITERATURA BRASILEIRA") que pegaram emprestado o item;
- Termo "romance" existente no título do item e também no título de um item que o usuário 22388 pegou emprestado.

Item de índice 2: Lendas e romances

Figura 5.10: *Usuário 22388 e item de título "Lendas e romances".*

Na figura 5.10, é possível verificar os relacionamentos que influenciaram no cálculo do score. São eles:

- Assunto "LITERATURA BRASILEIRA" como preferência de usuário e relacionado ao item;
- Termo "romance" existente no título do item e também no título de um item que o usuário 22388 pegou emprestado.

Item de índice 3: Poesia concreta brasileira as vanguardas na encruzilhada modernista

Figura 5.11: *Usuário 22388 e item de título "Poesia concreta brasileira as vanguardas na encruzilhada modernista".*

Na figura 5.12, é possível verificar os relacionamentos que influenciaram no cálculo do score. São eles:

- Assunto "LITERATURA BRASILEIRA" como preferência de usuário e relacionado ao item.

Item de índice 4: A tradição regionalista no romance brasileiro, 1857-1945

Figura 5.12: *Usuário 22388 e item de título "A tradição regionalista no romance brasileiro, 1857-1945".*

Na figura 5.12, é possível verificar os relacionamentos que influenciaram no cálculo do score. São eles:

- Termo "romance" existente no título do item e também no título de um item que o usuário 22388 pegou emprestado.

Item de índice 5: Escritos sobre arte e modernismo brasileiro

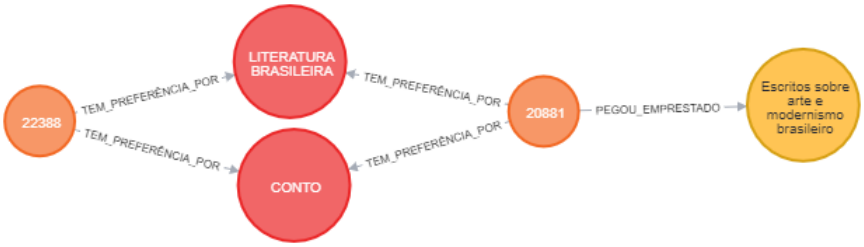

Figura 5.13: *Usuário 22388 e item de título "Escritos sobre arte e modernismo brasileiro".*

Na figura 5.13, é possível verificar os relacionamentos que influenciaram no cálculo do score. São eles:

- Um usuário similar ao usuário 22388 (possui interesse em comum pelos assuntos "LITERATURA BRASILEIRA" e "CONTO") que pegou emprestado o item.

Item de índice 6: A mulher e a cidade imagens da modernidade brasileira em quatro escritoras paulistas

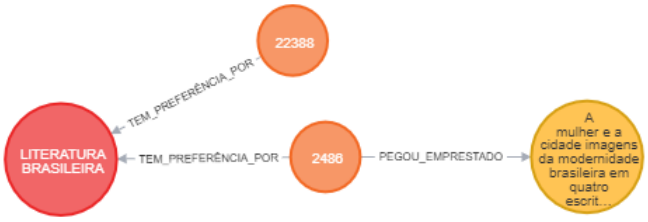

Figura 5.14: *Usuário 22388 e item de título "A mulher e a cidade imagens da modernidade brasileira em quatro escritoras paulistas".*

Na figura 5.14, é possível verificar os relacionamentos que influenciaram no cálculo do score. São eles:

- Um usuário similar ao usuário 22388 (possui interesse em comum pelo assunto "LITERATURA BRASILEIRA") que pegou emprestado o item.

Os itens de índices 7, 8, 9 e 10 não possuem nenhuma associação com o usuário e, por isso, seus relacionamentos temporários têm score igual a zero.

Capítulo 6

Conclusões

No presente trabalho foi desenvolvido um sistema que possibilita a integração das bibliotecas brasilianas presentes na Universidade de São Paulo e bibliotecas externas cujos sistemas empreguem o formato padronizado MARC.

Para validação do sistema, foram feitos experimentos a partir de dados coletados do sistema de catalogação Dedalus no formato MARC pertencentes ao Instituto de Estudos Brasileiros. As informações de empréstimo, entretanto, foram provenientes da Biblioteca Florestan Fernandes, da Faculdade de Filosofia, Letras e Ciências Humanas, em que foram selecionados apenas itens que existiam em ambas as bibliotecas para o entrelaçamento dos dados.

Os arquivos coletados passaram por transformações necessárias para permitir sua leitura e inserção em um banco de dados de grafos. As bibliotecas do Python utilizadas se mostraram suficientes para esses processamentos.

A modelagem do banco baseou-se nos dados disponíveis no sistema Dedalus que possui as informações mais pertinentes para a identificação dos itens dos acervos e para a construção de relacionamentos importantes. Dessa forma, ela apoia diferentes tipos de buscas e recomendações sofisticadas.

O banco de dados orientado a grafos foi escolhido devido à quantidade de itens dos acervos e dos relacionamentos que surgiram conforme a modelagem avançava. O Neo4j foi o sistema gerenciador selecionado para esse trabalho e se mostrou viável para a implementação do sistema proposto e para acomodar os dados selecionados.

Sendo assim, o sistema se mostrou flexível, uma vez que o padrão MARC que foi utilizado permite que o banco de dados receba dados de diferentes tipos de sistemas de catalogação. Além disso, as recomendações se baseiam em *scores* que são facilmente modificados, podendo suportar novos nós e relacionamentos que surjam de outras bibliotecas. A modelagem dos dados em grafos comporta a adição de novos nós e relacionamentos sem perturbar as consultas existentes.

Dessa forma, o sistema suporta a sua expansão para outras bibliotecas brasilianas além das que foram estudadas nesse trabalho.

Como os experimentos envolveram um conjunto de dados limitado, são interessantes e necessários novos experimentos com dados provenientes de fontes heterogêneas, principalmente no caso de expansão do sistema para outras bibliotecas fora da Universidade de São Paulo.

6.1 Trabalhos Futuros

Conforme os resultados desse trabalho, algumas ideias foram idealizadas para projetos futuros.

A primeira é a validação do sistema com usuários das bibliotecas.

Outra sugestão de continuação desse projeto é prover mecanismos para que o sistema possa ser diretamente conectado a sistemas de catalogação das bibliotecas da Universidade de São Paulo.

Além disso, um trabalho futuro é, conforme planejado pelo projeto "Brasiliana Inteligente", a expansão desse sistema para outras bibliotecas brasilianas.

Referências Bibliográficas

Antunes(2017) Cristina Antunes. *Biblioteca Brasiliana Guita e José Mindlin BBM - USP*. Citado na pág. 1

Assumpção e Santos(2015) Fabrício Silva Assumpção e Plácida Leopoldina Ventura Amorim da Costa Santos. Representação no domínio bibliográfico: um olhar sobre os Formatos MARC 21. *Perspectivas em Ciência da Informação*, 20:54 – 74. ISSN 1413-9936. URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362015000100054&nrm=iso. Citado na pág. 6, 7

de Souza *et al.***(2014)** Alexandre Morais de Souza, Edmir P. V. Prado, Violeta Sun e Marcelo Fantinato. Critérios para seleção de sgbd nosql: o ponto de vista de especialistas com base na literatura. *In: Anais Principais do X Simpósio Brasileiro de Sistemas de Informação.*, páginas 149–160. doi: https://doi.org/10.5753/sbsi.2014.6109. Citado na pág. 5

Fusco(2012) Elvis Fusco. Modelos conceituais de dados como parte do processo da catalogação: perspectiva de uso dos frbr no desenvolvimento de catálogos bibliográficos digitais. Citado na pág. 6

Gupta *et al.***(2009)** Vishal Gupta, Gurpreet S Lehal *et al.* A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, 1(1): 60–76. Citado na pág. 9

Huang *et al.***(2002)** Zan Huang, Wingyan Chung, Thian-Huat Ong e Hsinchun Chen. A graph-based recommender system for digital library. *Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries - JCDL 02*, página 65–73. doi: 10.1145/544220.544231. Citado na pág. 10

Lucene Apache(2010) Lucene Apache. *Apache Lucene-Overview*. URL https://svn-us.apache.org/repos/asf/lucene/java/site/docs/index.pdf. Acessado em: 12/12/2020. Citado na pág. 25

Lü *et al.***(2012)** Linyuan Lü, Matúš Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang e Tao Zhou. Recommender systems. *Physics Reports*, 519(1):1 – 49. ISSN 0370-1573. doi: https://doi.org/10.1016/j.physrep.2012.02.006. URL http://www.sciencedirect.com/science/article/pii/S0370157312000828. Recommender Systems. Citado na pág. 10

Macário e Baldo(2005) Carla Geovana do N. Macário e Stefano Monteiro Baldo. O modelo relacional. *In: Anais da VII Escola Regional de Informática de Goiás.* Trabalho como parte da avaliação do curso MO-410 (curso: Introdução a Banco de Dados)-Instituto de Computação da Unicamp. Citado na pág. 4

Penteado *et al.***(2014)** Raqueline R. M. Penteado, Rebeca Schroeder, Diego Hoss, Jaqueline Nande, Ricardo M. Maeda, Walmir O. Couto e Carmem S. Hara. Um estudo sobre bancos de dados em grafos nativos. *X ERBD-Escola Regional de Banco de Dados.* Citado na pág. 13

Py2neo(2020) Py2neo. The Py2neo Handbook, 2020. URL https://py2neo.org/2020.1. Acessado em: 12/12/2020. Citado na pág. 25

Pypi(2018) Pypi. selenium 3.141.0 - Project description, 2018. URL https://pypi.org/project/selenium. Acessado em: 12/12/2020. Citado na pág. 24

Resnick e Varian(1997) Paul Resnick e Hal R. Varian. Recommender systems. *Communications of the ACM.*, páginas 56–58. Citado na pág. 10

Robinson *et al.***(2013)** Ian Robinson, Jim Webber e Emil Eifrem. *Graph databases.* Citado na pág. 5

Santos e Silva(2013) Erik Santos e Marcos Silva. Abordagem ao banco de dados orientado a grafos neo4j em um nível empresarial. Citado na pág. 25

Santos(2016) Jéssica Ferreira Santos. Avaliação de critérios de seleção de software especializado para automação de unidades de informação: estudo comparativo entre os softwares sophia, pergamum, alexandria, aleph e biblivre. Citado na pág. 7

Sarwar *et al.***(2001)** Badrul Sarwar, George Karypis, Joseph Konstan e John Riedl. Item based collaborative filtering recommendation algorithms. *Proc. 10th International World Wide Web Conference.* Citado na pág. 10

Vieira(1988) Simone Bastos Vieira. Indexação automática e manual: revisão de literatura. páginas 43–57. Citado na pág. 9