Universidade Federal do Rio Grande - FURG

Centro de Ciências Computacionais – C3

Sistemas de Informação

Módulo de Critérios de Relevância para o ARGOs*earch*

Caroline Tomasini

Rio Grande

2013

Caroline Tomasini

Módulo de Critérios de Relevância para o ARGO*search*

Trabalho de conclusão de curso de graduação em Sistemas de Informação apresentado como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Orientador: Prof. Msc. André Prisco Vargas

Coorientador: Prof. Msc. Eduardo Nunes Borges

Rio Grande
2013

Este trabalho foi analisado e julgado adequado para obtenção do título de Bacharel em Sistemas de Informação e aprovado em sua forma final pelo orientador.

_____

Prof. Msc. André Prisco Vargas

Banca Examinadora:

Prof. MSc. André Prisco Vargas
Centro de Ciências Computacionais – FURG (orientador)
Prof. MSc. Eduardo Nunes Borges
Centro de Ciências Computacionais – FURG (coorientador)
Prof. Dra. Karina dos Santos Machado
Centro de Ciências Computacionais – FURG
Prof. Dr. Leonardo Emmendorfer
Centro de Ciências Computacionais – FURG

AGRADECIMENTOS

RESUMO

Sistemas de gerenciamento de bibliotecas permitem aos usuários realizar consultas sobre os metadados que descrevem uma coleção. No entanto, o usuário pode não saber quais dos itens retornados na busca são os mais adequados para o seu perfil. Para solucionar este problema foi proposto um sistema de recuperação de informações denominado ARGO*search*, capaz de melhorar a qualidade das consultas em sistemas de gerenciamento de bibliotecas. O sistema é especificado através de uma arquitetura extensível, baseada em critérios de relevância. Desta forma, o objetivo deste trabalho é o desenvolvimento do módulo de critérios de relevância, para o ARGO*search*. Esse módulo corresponde a parte responsável por ordenar os resultados das consultas considerando três tipos de informação: a similaridade textual entre a consulta do usuário e os metadados que descrevem os itens da coleção, estatísticas de uso do sistema proposto e, informações de contexto implícitas extraídas do perfil do usuário. Estas informações são usadas para definir critérios de relevância como, por exemplo, documentos com *maior número de reservas* ou que *pertencem à bibliografia básica das disciplinas em que o usuário está matriculado*. Por fim, os critérios são combinados e utilizados para ordenar os resultados. Deste modo, o *ranking* final dos documentos é personalizado de acordo com o usuário que estiver efetuando a consulta, pois parte do algoritmo de ordenação considera informações implícitas em seu perfil.

Palavras-Chave: recuperação de informações, busca por palavras-chave, gerenciamento de informações, sistemas de gerenciamento de bibliotecas.

ABSTRACT

Library management systems allow users to perform queries on the metadata that describes a collection. However, the user may not know which of the items returned in the search are best suited to his/her profile. To solve this problem we propose a system called ARGOsearch information retrieval, capable of improving the quality of consultation on library management systems. This system is specified through an extensible architecture based on relevance criteria. The objective of this work is the development of the module relevance criteria which is the part responsible for ordering query results considering three types of information: the textual similarity between the user query and the metadata that describes the items in the collection, statistics use of the proposed system, and implicit context information extracted from the user profile. This information is used to define criteria of relevance, for example, documents with more reservations or pertaining to basic bibliography of a discipline in which the user is registered. Finally, the criteria are combined and used to sort results. Thus, the final ranking of the documents is customized according to the user who is performing the query, as part of the sorting algorithm considers implicit information in your profile.

Keywords: information retrieval, keyword search, information management, library management systems.

LISTA DE SIGLAS E ABREVIATURAS

BD – Banco de Dados

FURG – Universidade Federal do Rio Grande

MAP – Mean Avarage Precision

NTI – Núcleo de Tecnologia da Informação

QSL – Quadro de Sequência Lógica

SAB - Sistemas de Administração de Bibliotecas

SAI – Sistema de Informações Acadêmicas

SGBD – Sistema de Gerenciamento de Banco de Dados

SQL – Structured Query Language

SRI – Sistema de Recuperação de Informação

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Bibliotecas de grande porte, como aquelas presentes em universidades, precisam de ferramentas de apoio para localizar em sua coleção itens como livros, mapas, artigos e outros documentos de interesse da comunidade de usuários. Essas ferramentas, as quais estão presentes nos Sistemas de Administração de Bibliotecas (SAB), permitem aos usuários realizar consultas gerais sobre os metadados que descrevem a coleção. Geralmente, as consultas dos usuários são compostas por palavras-chave. Entretanto, não é uma tarefa trivial para o usuário mapear suas necessidades de informação através de uma consulta [Al-Maskari e Sanderson 2011]. Por exemplo, erros simples de grafia no título ou nos nomes dos autores podem omitir dos resultados um documento relevante para o usuário, indicando a ideia de que o documento não existe no acervo.

Algumas consultas retornam centenas de documentos, distribuídos em várias telas ou páginas (em sistemas *Web*). Neste caso, o usuário precisa acessar cada página e verificar item por item para encontrar o documento desejado. Além disso, o usuário que procura por uma informação específica pode não saber exatamente quais dos itens retornados por sua consulta são os melhores de acordo com o seu perfil. Por exemplo, um universitário calouro deveria, em média, priorizar livros-texto ou introdutórios em vez de artigos científicos com temas avançados de pesquisa.

Quanto melhor a qualidade do resultado de uma consulta, mais fácil para os usuários encontrarem documentos de interesse. Neste contexto, o grupo de pesquisa em gerenciamento de informações da FURG projetou o ARGO*search* [Pereira 2011a], um sistema de recuperação de informações [Baeza-Yates e Ribeiro-Neto, 1999] para o SAB. Esse sistema é especificado por meio de uma arquitetura extensível, baseada em múltiplos critérios de relevância, que ordena os resultados das consultas considerando três tipos de informação: a similaridade textual entre a consulta do usuário e os metadados que descrevemos itens da coleção, estatísticas de uso do SAB e por fim, informações de contexto implícitas extraídas do perfil do usuário.

Resultados preliminares foram apresentados em trabalho prévio [Pereira et al., 2012] com o objetivo de validar a arquitetura proposta. Este trabalho de

conclusão de curso especifica o módulo de critérios de relevância do ARGO*search*, apresenta a implementação de uma série de critérios e expande os experimentos realizados anteriormente com o objetivo de analisar o quanto cada critério de relevância proposto contribui para a qualidade do resultados das consultas. Os experimentos também mostram a influência da similaridade textual na recuperação de informações e o ganho de qualidade do ARGO*search* em relação ao modelo espacial vetorial de recuperação de informações [Manning et al. 2008].

O texto está organizado da seguinte forma. O capítulo 2 apresenta a fundamentação teórica sobre recuperação de informação e modelo vetorial. No capítulo 3 são apresentados os trabalhos relacionados. No capítulo 4 são especificados os critérios de relevância e discutidos aspectos técnicos da implementação. No capítulo 5 é apresentada uma avaliação experimental do trabalho proposto. No capítulo 6 são apresentados detalhes sobre a implantação do sistema. Por fim, no capítulo 7 são apresentadas as conclusões e os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Esse capítulo apresenta a fundamentação teórica necessária para o entendimento deste trabalho. Conceitos tais como, recuperação de informação, modelo vetorial e visão relacional são apresentados a seguir.

2.1 RECUPERAÇÃO DE INFORMAÇÃO

Recuperação de informação é uma subárea da ciência da computação que estuda o armazenamento e recuperação automática de documentos. Segundo [Manning et al. 2008], recuperação de informação consiste no ato de encontrar, em grandes coleções, documentos de natureza não estruturada (texto) que satisfaça uma necessidade de informação. Para completar este conceito, Baeza-Yates e Ribeiro Neto [1999] definem um Sistema de Recuperação de Informação (SRI) formalmente como:

$$SRI = [D, Q, M, R(q_i, d_j)]$$

onde:

D = conjunto de documentos

Q = consulta

M = modelo de representação do conjunto de documentos e da consulta

R = função de *ranking* que associa um valor de ordenação a um par composto por uma consulta q_i e um documento d_j

A função de um SRI é extrair as informações nos documentos de uma coleção e ordená-los para o usuário. Seu objetivo é recuperar o menor número possível de documentos não relevantes e ordená-los, sendo que os mais interessantes para o usuário apareçam nas primeiras posições.

Mapear as necessidades do usuário é um problema complexo. Por exemplo, em uma determinada coleção podem existir documentos que falam sobre o tema “abuso de drogas” e que contenham frases como: “encontrados 10kg de maconha” ou “overdose de cocaína”. Entretanto, estes documentos não contêm explicitamente as palavras “drogas” e “abuso”. Tomando por base este exemplo, fica claro que a

recuperação de informação de documentos relevantes está diretamente ligada à interação do usuário com o sistema e a forma de representação dos documentos.

2.2 MODELO VETORIAL

No modelo espacial vetorial [Baeza-Yates e Ribeiro Neto 1999], cada documento *d* é representado por um vetor com *t* dimensões, uma para cada termo do vocabulário de toda a coleção. O peso w_i de cada dimensão *i* é calculado a partir da frequência dos termos e tem a função de quantificar a relevância de cada termo para as consultas e para os documentos. A relevância do documento em relação a uma consulta *q* é dada por uma função de similaridade baseada no cosseno do ângulo formado pelos dois vetores, conforme a Equação 1 [Salton e Bucckley 1988].

$$\mathrm{sim(d,q)} = \frac{\sum_{i=1}^{t} w_{id}\, w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{id}^2}\sqrt{\sum_{i=1}^{t} w_{iq}^2}} \quad (1)$$

Se uma coleção possui *N* documentos e df_i é a quantidade de documentos que possuem o termo t_i, então os pesos w_{id} são calculados através da métrica $tf\ x\ idf$ conforme a Equação 2, onde $freq_{id}$ é a frequência do termo *i* no documento *d* e $\max(freq_{td})$ é a máxima frequência de qualquer termo *t* presente no mesmo documento. Para as consultas, essa frequência normalizada ainda pode ser suavizada através de um fator de amortecimento. Pode-se notar que quanto mais frequente é um termo num documento, maior o peso nessa dimensão. Entretanto, a frequência inversa idf reduz o peso de termos comuns na coleção, ou seja, daqueles que aparecem em muitos documentos.

$$w_{id} = tf_{id}\ idf_i = \frac{freq_{id}}{\max(freq_{td})} log\ \frac{N}{df_i} \quad (2)$$

3 TRABALHOS RELACIONADOS

Neste capítulo são apresentados trabalhos relacionados com o tema proposto nesta monografia: a arquitetura do ARGO*search* , o sistema de administração de bibliotecas ARGO e o sistema de informação acadêmico SIA.

3.1 ARGO*SEARCH*

O ARGO*search* [Pereira et al. 2012] ordena os resultados das consultas considerando similaridade textual, estatísticas de uso do sistema e o perfil do usuário que executa uma consulta. A Figura 3.1 apresenta a arquitetura do sistema.

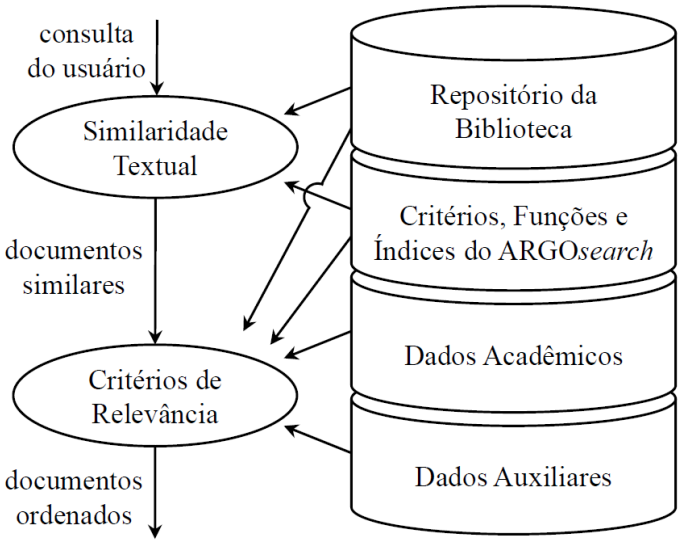

Figura 3.1 - Arquitetura do ARGO*search* detalhando a interação entre os principais componentes e os dados da biblioteca e do usuário.

Em vez de executar a consulta do usuário, o SAB a entrega ao componente de similaridade textual. Este componente busca no repositório da biblioteca por documentos com descritores (metadados) similares aos termos da consulta. A similaridade é calculada usando funções específicas que comparam cadeias de caracteres e retornam um escore de similaridade. Se este escore exceder um determinado limiar de similaridade, serão retornados os documentos cujos descritores sejam suficientemente similares para representar a consulta. O

componente de similaridade textual utiliza o banco de dados do ARGO*search* para armazenar as funções e operadores de similaridade, além dos índices necessários para acelerar as consultas.

Os documentos que satisfazem a condição de similaridade entre os metadados e a consulta são enviados ao componente de critérios de relevância. Nesta fase, o ARGO*search* extrai o perfil do usuário da base de informações acadêmicas analisando dados como o tipo de usuário (estudante, professor ou administrativo), nacionalidade, departamento ao qual está afiliado, disciplinas cursadas e em progresso. A informação extraída é usada para definir alguns dos critérios de relevância como assunto e idioma.

O repositório da biblioteca da FURG fornece registros das transações que descrevem as reservas e os empréstimos de todos os usuários. Estes registros são úteis para extrair alguns critérios de relevância. Por exemplo, o número de reservas e o número de empréstimos por semestre são heurísticas promissoras para determinar a importância de um documento para a comunidade em geral. A arquitetura foi construída para permitir que um especialista possa definir e configurar novos critérios. Os novos critérios também podem ser extraídos de outras bases de dados auxiliares.

Por fim, os critérios de relevância são combinados e utilizados para ordenar os resultados entregues pelo componente de similaridade textual. Nota-se que o ranking final dos documentos depende do usuário que efetua a consulta, porque parte do algoritmo de ordenação considera informações de contexto implícitas em seu perfil.

3.1.1 Similaridade Textual

O componente de similaridade textual utiliza uma métrica de casamento aproximado baseada em trigramas [Angell et al. 1983]. Um trigrama é uma sequência de três caracteres que compõem uma palavra. Seja *Str* o conjunto de todas as cadeias de caracteres e R_1 o conjunto dos números reais no intervalo fechado [0; 1], a função *trigram matching* : { *Str* x *Strg* } → R_1 recebe como parâmetro duas cadeias de caracteres e retorna um escore de similaridade. A Equação 3 especifica a função como a razão entre o número de elementos da

intersecção entre os conjuntos de trigramas A,B que compõem os parâmetros a, b e da união entre os mesmos conjuntos.

$$trigram\ matching\ (a,b) = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

Esta métrica é executada entre a *string* de busca e os valores dos metadados selecionados pelo usuário no momento da consulta (autor, título, assunto, etc.). Neste ponto, o ARGO*search* usa índices específicos para selecionar apenas os registros candidatos ao casamento aproximado, ou seja, que contém pelo menos um trigrama em comum com a *string* de busca, acelerando o tempo da consulta aproximada. A figura 3.2 apresenta um exemplo do cálculo da similaridade entre uma consulta do usuário com erros de grafia e uma instância do metadado autor. Os documentos em que o escore retornado pela função de similaridade for maior que um determinado limiar especificado por um especialista são recuperados e enviados ao componente de critérios de relevância.

Consulta: “Tanebaun”
Metadado Autor: “Tanenbaum”
Trigramas (Tanebaun) = A = {" t", " ta", tan, ane, neb, eba, bau, aun, "un "}
Trigramas (Tanenbaum) = B = {" t", " ta", tan, ane, nen, enb, nba, bau, aum, "um "}

$$|A \cap B| = |\{\text{" t"}, \text{" ta"}, tan, ane, bau\}| = 5$$

$$|A \cup B| = |\{\text{" t"}, \text{" ta"}, tan, ane, neb, nen, eba, enb, nba, bau, aum, aun, \text{"um "}, \text{"un "}\}| = 14$$

$$trigram\ matching\ (Tanebaun, Tanenbaum) = \frac{5}{14} = 0{,}357 \cong 36\%$$

Figura 3.2 - Exemplo do cálculo da função de similaridade *trigram matching* implementada no componente de similaridade textual. Os espaços em branco são adicionados às palavras para compensar os caracteres que aparecem em apenas um trigrama.

Apesar dos modelos tradicionais de recuperação de informações calcularem a similaridade entre os documentos e a consulta do usuário ou a probabilidade do documento ser relevante para esta consulta, os termos são comparados por igualdade. Quando comparado aos modelos vetorial e probabilístico [Manning et al. 2008], o ARGO*search* se destaca porque uma vez que utiliza similaridade textual, aumenta a cobertura dos resultados, retornando um número maior de possíveis

documentos de interesse. Nota-se que os metadados que descrevem os documentos retornados podem conter valores similares aos termos da busca, considerando assim, variações de grafia causadas por erros de cadastramento ou pela ausência de conhecimento do usuário.

É importante ressaltar a função baseada em trigramas adotada pelo componente de similaridade textual pode ser facilmente substituída por outra função específica mais adequada a um determinado contexto [Borges et al. 2011; Cohen et al. 2003].

3.1.2 Implementação

O ARGO*search* foi implementado utilizando a linguagem de programação PHP e Sistema de Gerenciamento de Banco de Dados (SGBD) PostgreSQL para armazenar funções e operadores de similaridade, índices e critérios de relevância. A função *trigram matching* foi incorporada ao banco de dados do ARGO*search* por meio da instalação de uma contribuição da comunidade de usuários do PostgreSQL denominada pg_trgm[1]. Além de calcular a similaridade textual baseada em trigramas, este módulo permite a criação de índices que aceleram a consulta aproximada. Optou-se por utilizar um índice invertido (*Generalized Inverted Index*) porque é mais rápido para leitura e os metadados que descrevem os documentos de uma biblioteca não costumam ser alterados com frequência.

Os critérios de relevância foram implementados por meio de visões armazenadas no banco de dados do ARGO*search*. O capítulo 4 apresenta os detalhes da especificação dos critérios adotados e da implementação dos mesmos relacionando dados do sistema de administração de bibliotecas da FURG e dados acadêmicos dos usuários. Esta é a principal contribuição deste Trabalho de Conclusão de Curso no contexto do projeto ARGO*search* como um todo.

Por fim, o especialista registra o critério de relevância no banco de dados do ARGO*search* e associa a este o peso W_c que pode ser alterado para dar maior ou menor importância ao critério. Além disso, é necessário informar a necessidade de identificação do usuário que realiza uma consulta. Por exemplo, pertence à

[1] Postgresql.org/docs/current/static/pgtrgm.html

bibliografia indicada é um critério que depende do usuário e número da edição modela o comportamento da comunidade como um todo.

3.2 ARGO

ARGO[2] é o sistema de administração de bibliotecas da FURG. O sistema foi desenvolvido dentro da própria Universidade pelo Núcleo de Tecnologia da Informação (NTI) como uma aplicação *Web*. O sistema possui uma interface iterativa e amigável, que possibilita ao usuário realizar consultas, fazer reservas, e renovar os documentos que estão locados por ele. Também é possível consultar multas, visualizar o histórico de locação e criar uma lista de documentos favoritos. Na Figura 3.3 pode ser vista a interface do ARGO.

Figura 3.3 - Interface do ARGO destacando a funcionalidade pesquisa.

A Figura 3.4 apresenta a estrutura simplificada do banco de dados do sistema ARGO. O diagrama apresentado é resultado do estudo minucioso do esquema relacional do banco, já que o NTI havia documentado o sistema na época do desenvolvimento deste trabalho. Ele apresenta apenas as tabelas que foram usadas para o desenvolvimento dos critérios de relevância, descritos no capítulo 4, que ordenam os resultados das consultas. Atualmente, o banco de dados do ARGO possui cerca de 70 tabelas, 22 mil usuários, 4.6 milhões de registros e 1.6 GB.

[2] http://www.argo.furg.br

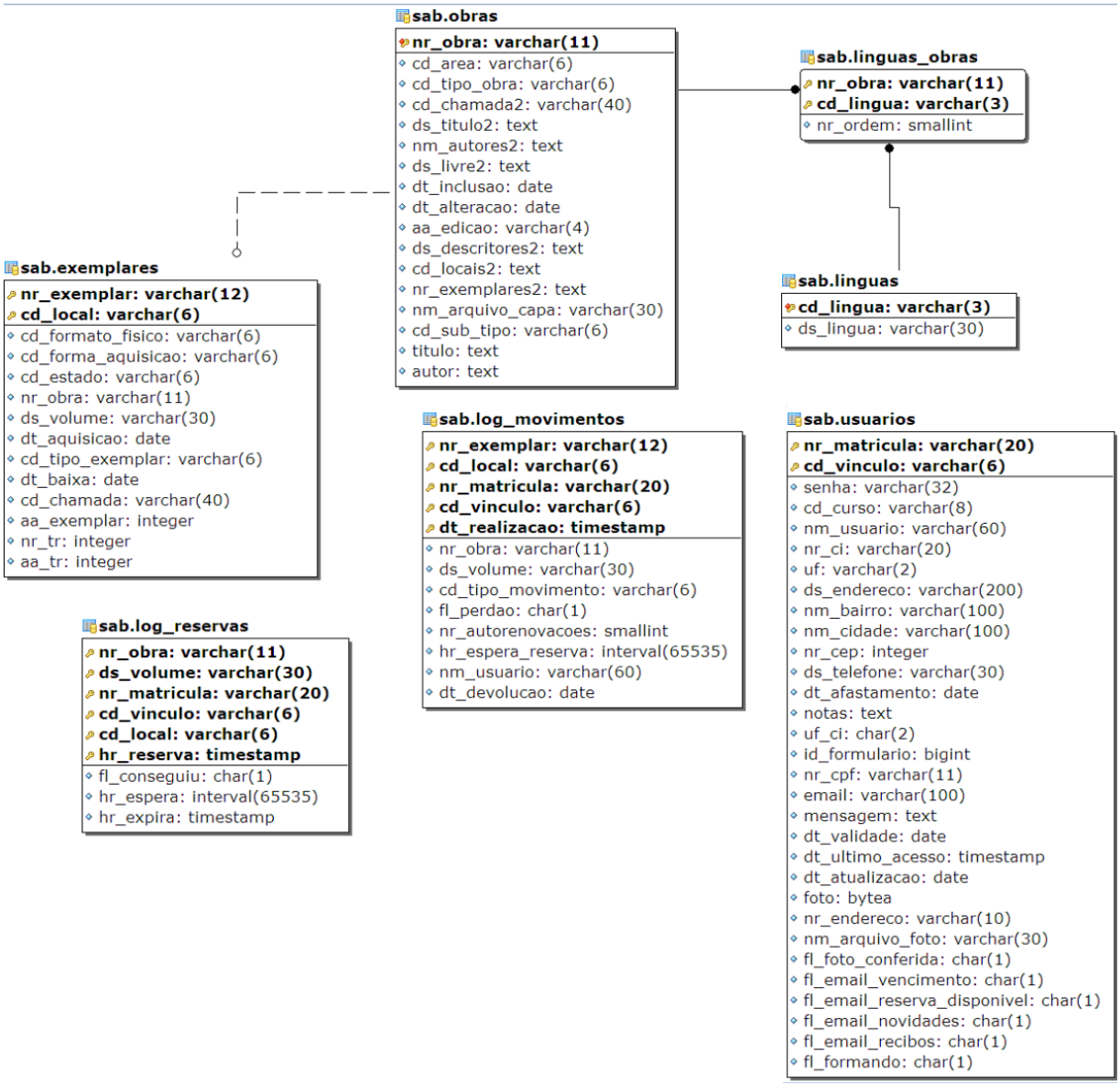

Figura 3.4 - Modelo relacional simplificado do ARGO.

A tabela *obras* armazena todas as informações de uma determinada obra disponível no arquivo da biblioteca. Os principais atributos descrevem uma identificação única, autor, título, edição e número de exemplares. A tabela *linguas* armazena um conjunto de idiomas disponíveis e *línguas_obras* relaciona estes idiomas com as obras. A tabela *exemplares* armazena informações sobre cada exemplar, tais como: data de aquisição e código de chamada. Uma determinada obra pode possuir vários exemplares.

A tabela *usuarios* armazena informações de cada usuário registrado no ARGO. As mais importantes são o número de matrícula e o vínculo na Universidade, que permitem ou restringem determinados acessos e funcionalidades no sistema. Esta tabela também armazena as informações pessoais como nome, endereço e telefone.

As tabelas *log_movimentos* e *log_reservas* armazenam o histórico de empréstimos, renovações, e reservas de um documento. Dessa forma é possível verificar, por exemplo, quem reservou determinado documento em uma data qualquer ou quantos empréstimos um documento sofreu em determinado período.

Todas as tabelas são precedidas pelo qualificador *sab* porque este é o nome do esquema relacional do sistema de administração de bibliotecas.

3.3 SIA

O SIA é o sistema de informações acadêmicas da FURG, desenvolvido e mantido pelo NTI da Universidade. Este sistema implementa uma série de operações administrativas e acadêmicas que dependem do perfil do usuário. Por exemplo, estudantes podem verificar as notas obtidas em disciplinas em andamento ou já cursadas, professores podem preencher os planos de ensino das disciplinas que ministram, incluindo a bibliografia recomendada, coordenadores de curso podem acessar o histórico universitário de qualquer estudante, entre diversas outras operações. A Figura 3.5 apresenta a estrutura simplificada do banco de dados do SIA. O diagrama apresentado é resultado do estudo minucioso do esquema relacional do banco, já que o NTI também não havia documentado este sistema. Assim como o diagrama anterior, ele apresenta apenas as tabelas que foram usadas para o desenvolvimento dos critérios de relevância descritos no capítulo 4. Atualmente, o banco de dados do SIA possui cerca de 200 tabelas, 60 mil usuários, 22 milhões de registros e 2.3 GB.

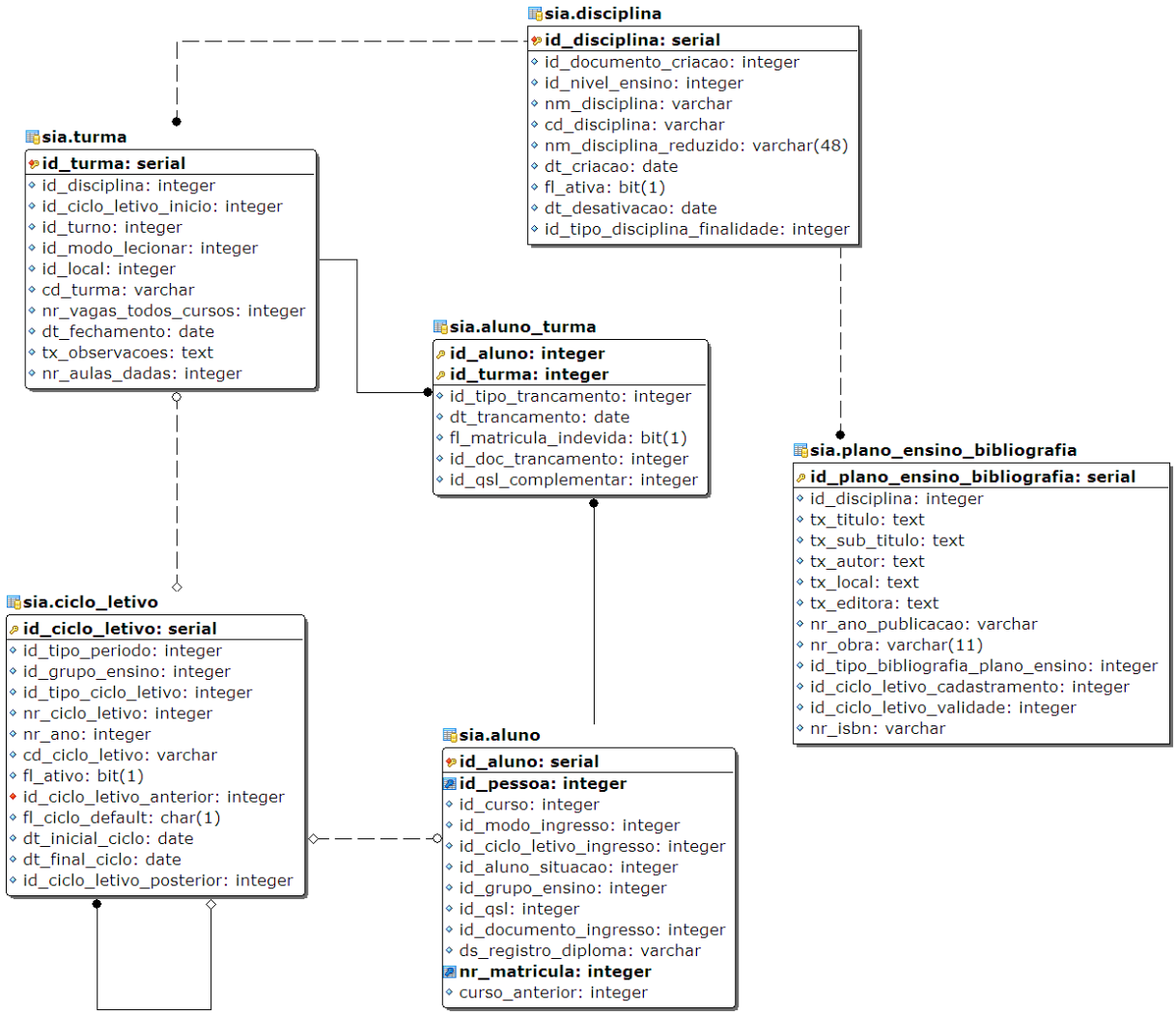

Figura 3.5 - Modelo relacional simplificado do SIA.

A tabela *aluno* armazena informações de cada aluno registrado no SIA. As informações mais relevantes são o número de matrícula e qual seu curso. Esta tabela também armazena as informações pessoais como nome, endereço e telefone e possui um relacionamento com a tabela *aluno_turma*.

A tabela *aluno_turma* é usada como uma tabela intermediária entre o relacionamento de *aluno* e *turma*. A tabela *turma* armazena informações sobre todas as turmas da Universidade, esta tabela também possui relacionamento com a tabela *disciplina* e *ciclo_letivo*.

A tabela *ciclo_letivo* armazena qual é o ciclo letivo corrente e outras informações importantes como, data de inicio e fim do ciclo letivo e também o ciclo letivo anterior.

A tabela *disciplina* armazena informações das disciplinas presentes nos quadros de sequência lógica (QSL) dos cursos. A sua relação com *turma* é dada de maneira que podem existir várias turmas de uma única disciplina.

A tabela *plano_ensino_bibliografia* que se relaciona com a tabela *disciplina* descrita anteriormente é utilizada pelos professores, onde cada professor tem a opção de cadastrar seu plano de ensino e respectiva bibliografia no sistema.

Todas as tabelas são precedidas pelo qualificador *sia* porque este é o nome do esquema relacional do sistema de informações acadêmica.

4 CRITÉRIOS DE RELEVÂNCIA

Este capítulo apresenta o conjunto de critérios de relevância propostos neste trabalho. Os critérios são definidos conceitualmente, pois são independentes de implementação.

4.1 CRITÉRIOS DE RELEVÂNCIA QUE UTILIZAM O PERFIL DO USUÁRIO

Nesta seção, são descritos os critérios que utilizam informações do perfil do usuário, ou seja, informações como o curso do usuário e quais disciplinas ele está matriculado. Estes critérios também podem utilizar valores de metadados e estatísticas de uso do SAB.

4.1.1 Documento pertence à bibliografia básica recomendada

Este critério de relevância privilegia documentos que façam parte da bibliografia básica de uma disciplina em que o usuário esteja matriculado. Esta informação é extraída da base de dados do sistema acadêmico da universidade. Neste sistema, os professores são responsáveis por preencher os planos de ensino de cada disciplina sob sua responsabilidade, incluindo a bibliografia básica e complementar.

Por exemplo, imagine que o usuário João está cursando a disciplina de Sistemas Computacionais I e na bibliografia básica desta disciplina estão cadastrados os livros: Arquitetura e organização de computadores, de Willian Stallings, e Organização estruturada de computadores, de Andrew S. Tanenbaum. Quando o usuário João efetuar uma consulta no ARGO pelo autor Tanenbaum, o sistema irá inferir que o livro intitulado *Organização estruturada de computadores* é mais importante que o livro intitulado *Sistemas operacionais modernos*, ambos do mesmo autor.

4.1.2 Documento locado por colegas de classe no período letivo

Este critério prioriza os documentos que já foram locados por colegas de classe do usuário que está realizando a pesquisa no período letivo atual. Esta

informação, assim como a da bibliografia básica, é extraída da base de dados do sistema acadêmico.

Por exemplo, supõe-se que o usuário João está matriculado na disciplina de Linguagens Formais e Teoria da Computação. Alguns alunos que também cursam esta disciplina estão realizando empréstimos do livro Introdução à teoria das linguagens formais, dos autômatos e da computabilidade. Quando João realizar uma consulta no ARGO pelo título *linguagens formais* o sistema irá inferir que o livro *Introdução à teoria das linguagens formais, dos autômatos e da computabilidade* é mais importante que o livro *linguagens formais e autômatos* pois o primeiro livro tem sido locado por seus colegas e o segundo não.

Desta maneira, com este critério podemos inferir que os livros locados pelos usuários que cursam as mesmas disciplinas que João, serão relevantes para ele.

4.1.3 Documento no idioma nativo

Neste critério, os documentos que forem do idioma nativo do usuário terão prioridade sobre os documentos de outras linguagens. Por exemplo, quando um usuário com nacionalidade brasileira realiza uma busca no sistema ARGO e utiliza este critério, os documentos em português irão aparecer na frente dos documentos de outros idiomas.

4.1.4 Número de renovações

Neste critério, é analisado o histórico de locações do usuário que está fazendo a consulta. Após ter a relação dos livros já locados pelo usuário é analisado o tempo que aquele usuário ficou com os livros emprestados, ou seja, quando um usuário realiza um empréstimo de um livro e renova três vezes significa que ele ficou mais tempo do que se ele estiver com um livro e renovar apenas uma vez. Desta forma, os documentos mais relevantes serão aqueles que o usuário já locou por um maior período de tempo.

4.1.5 Documento locado anteriormente

Este critério infere maior importância para os documentos que o usuário que está fazendo a consulta já tenha locado em algum momento.

4.2 CRITÉRIOS DE RELEVÂNCIA QUE UTILIZAM APENAS ESTATÍSTICAS DE USO DO SAB

Nesta seção, serão descritos os critérios que utilizam dados estatísticos do SAB, como quantidade de empréstimos, quantidade de reservas de um documento, entre outros. Estes critérios modelam o comportamento de toda a comunidade de usuários.

4.2.1 Número de empréstimos

O critério número de empréstimos infere que os documentos que possuírem o maior número de empréstimos desde que eles fazem parte do acervo da biblioteca, serão os documentos mais relevantes.

Por exemplo, o usuário João efetuou uma busca no ARGO, de acordo com sua busca o sistema retornou os documentos com *id* 10, 383 e 500. O documento 10 possuía um total de 30 empréstimos, o 383 um total de 488 empréstimos e o 500 um total de 12 empréstimos. De acordo com este critério o sistema irá inferir que o documento mais importante será o de *id* 383, pois é um documento bastante procurado pelos usuários do ARGO.

4.2.2 Número de exemplares

Semelhante ao critério anterior, os documentos mais relevantes serão os documentos que possuírem maior quantidade de exemplares. Por exemplo, um livro que possuir 10 exemplares será mais relevante que um livro que possuí apenas 3 exemplares. Este critério parte da premissa que são comprados mais exemplares de livros que são mais requisitados.

4.2.3 Número de reservas

Partindo da premissa que um livro possui vários exemplares, neste critério são somadas as reservas de todos os exemplares de um determinado livro e gerado uma listagem com esses valores, identificador do livro e quantidade total de reservas. Com esta listagem, é feito uma ordenação decrescente e no topo será mostrado o livro que tenha a maior quantidade de reservas.

4.2.4 Taxa de reserva pelo período de tempo no acervo

Este critério faz o cálculo da taxa de reserva de cada livro do acervo da biblioteca a partir da data de aquisição, ou seja, divide a quantidade de reservas total do livro pelo tempo que ele existe na biblioteca. Após este cálculo é gerada uma lista com todos os livros e suas respectivas taxas de reserva e a lista é ordenada de forma decrescente. Desta maneira, os livros com maior taxa de reserva serão os do topo da lista.

4.3 IMPLEMENTAÇÃO DOS CRITÉRIOS

Todos os critérios foram implementados utilizando o conceito de visão relacional de banco de dados [Elmasri e Navathe 2005]. Uma visão é uma tabela virtual derivada de uma ou mais tabelas. É um meio para especificar uma tabela que precise ser consultada frequentemente, embora ela não exista fisicamente.

Alguns sistemas de banco de dados permitem que as visões sejam materializadas, ou seja, tornar a visão uma tabela física, garantindo assim que se ocorrerem modificações nas relações reais usadas na definição da visão, a visão também será modificada. Visões materializadas podem ser utilizadas para consolidar dados provenientes de múltiplas fontes. Esta estratégia é frequentemente utilizada para criação de *data marts* e *data warehouses*.

No contexto deste trabalho, as visões permitem que um analista da base de dados possa criar critérios com maior liberdade, relacionando dados do sistema de gerenciamento de bibliotecas, dados acadêmicos ou provenientes de outras fontes

auxiliares. Dados pré-processados por outros sistemas também podem ser relacionados usando visões materializadas.

As visões seguem um padrão simples. Elas são compostas pelo campo *id*, que representa o código de um documento, e pelo campo *escore*, que define o valor atribuído pelo critério de relevância para o documento (S_{dc}). Algumas visões também retornam a identificação única do usuário que realiza a consulta.

As próximas subseções especificam cada critério de relevância implementado através de uma visão SQL.

4.3.1 Documento pertence à bibliografia básica recomendada

Este critério retorna uma distribuição de valores binária, ou seja, escore 1 para os documentos presentes na bibliografia básica de alguma disciplina que o usuário está cursando. Também são retornados o *id* da obra e o número de matricula do usuário.

A Figura 4.1 apresenta a consulta SQL de criação da visão. Nas linhas 2 a 7 é selecionado a obra, seu escore e o identificador do usuário da tabela *aluno_turma*. Nas linhas 8 a 18 é utilizada a cláusula *inner join* para saber qual é o ciclo letivo atual. Nas linhas 19 a 21 são selecionadas apenas as obras que estiverem presentes no plano de ensino de um determinado usuário. Nas linhas 22 a 25 os dados são agrupados e removidas as tuplas repetidas da listagem final.

```
1  CREATE OR REPLACE VIEW argosearch.view_bibliografia_basica AS
2  SELECT
3      peb.nr_obra AS id,
4      1 AS escore,
5      atu.id_aluno AS usuario
6  FROM
7      sia.aluno_turma atu
8  INNER JOIN
9      sia.turma t ON
10         atu.id_turma = t.id_turma
11         AND t.id_ciclo_letivo_inicio = ( SELECT
12                                             cl.id_ciclo_letivo
13                                           FROM
14                                             sia.ciclo_letivo cl
15                                           WHERE
16                                             cl.fl_ciclo_default = 'S'
17                                           LIMIT 1
18                                         )
19 INNER JOIN
20     sia.plano_ensino_bibliografia peb ON
21         t.id_disciplina = peb.id_disciplina
22 INNER JOIN
23     sab.obras o ON peb.nr_obra = o.nr_obra
24 GROUP BY
25     peb.nr_obra, atu.id_aluno;
```

Figura 4.1 - SQL do critério *bibliografia básica*.

4.3.2 Documento locado por colegas de classe no período letivo

Este critério retorna uma distribuição de valores do tipo inteira. São retornados o *id* da obra, o número de matrícula e o escore que será a quantidade de vezes que um determinado documento foi locado por algum colega de classe.

A figura 4.2 apresenta a consulta SQL de criação da *view.* Nas linhas 2 a 7 é selecionado o documento, seu escore e o número de matricula do usuário. Nas linhas 8 a 12 é feito um *join* para que possa ser utilizada informações da tabela *turma*. Nas linhas 16 a 22 é acessada a tabela *disciplinas* para ter a informação de quais são as disciplinas que o usuário esta cursando. Nas linhas 25 a 32 é realizada uma subconsulta para saber qual é o ciclo letivo atual. Na linha 35 restringimos os movimentos para que sejam selecionados apenas aqueles do tipo empréstimo. Nas linhas 36 a 49 é realizada uma subconsulta para selecionar os empréstimos que foram feitos durante o ciclo letivo atual. Nas linhas 50 e 51 as obras e os usuários

são agrupados para que não haja repetição de tuplas. E para finalizar, nas linhas 52 e 53 o *ranking* é ordenado de maneira decrescente.

```
1  CREATE OR REPLACE VIEW argosearch.view_disciplinas AS
2  SELECT
3      lm.nr_obra AS id,
4      count(*) AS escore,
5      a.id_aluno AS usuario
6  FROM
7      sab.log_movimentos lm
8  LEFT JOIN ( SELECT
9                  atu.id_aluno AS usuario,
10                 d.id_aluno
11             FROM
12                 sia.aluno_turma atu
13 INNER JOIN
14     sia.turma t ON
15         atu.id_turma = t.id_turma
16 INNER JOIN ( SELECT
17                 t.id_disciplina,
18                 atu.id_aluno
19             FROM
20                 sia.turma t
21             INNER JOIN
22                 sia.aluno_turma atu ON
23                     atu.id_turma = t.id_turma
24             WHERE
25                 t.id_ciclo_letivo_inicio = ( SELECT
26                                                 cl.id_ciclo_letivo
27                                              FROM
28                                                 sia.ciclo_letivo cl
29                                              WHERE
30                                                 cl.fl_ciclo_default = 'S'
31                                             ) d ON
32                                                t.id_disciplina = d.id_disciplina
33             ) a ON
34                    lm.nr_matricula = a.usuario
35                    AND lm.cd_tipo_movimento = 'E'
36                    AND lm.dt_realizacao >= ( SELECT
37                                                cl.dt_inicial_ciclo
38                                              FROM
39                                                sia.ciclo_letivo cl
40                                              WHERE
41                                                cl.fl_ciclo_default = 'S'
42                                             )
43                    AND lm.dt_realizacao <= ( SELECT
44                                                cl.dt_final_ciclo
45                                              FROM
46                                                sia.ciclo_letivo cl
47                                              WHERE
48                                                cl.fl_ciclo_default = 'S'
49                                             )
50   GROUP BY
51     lm.nr_obra, a.id_aluno
52   ORDER BY
53     count(*) DESC;
```

Figura 4.2 - SQL do critério *documentos locados por colegas de classe no mesmo período letivo*.

4.3.3 Documento no idioma nativo

Este critério retorna uma distribuição de valores binária onde o escore 1 representa que o documento está no idioma português.

A Figura 4.3 apresenta a consulta SQL de criação da *view.* Nas linhas 2 a 4 seleciona-se o id do documento e o escore 1. Nas linhas 5 e 6 indica-se que os dados serão selecionados da tabela *linguas_obras* e nas linhas 7 e 8 é filtra-se o idioma para que sejam selecionados apenas os documentos em português. Este idioma pode ser inferido a partir da nacionalidade do usuário armazenada nos dados pessoais no SIA.

```
1  CREATE OR REPLACE VIEW argosearch.view_livro_por_idioma AS
2  SELECT
3      lo.nr_obra AS id,
4      1 AS escore
5  FROM
6      sab.linguas_obras lo
7  WHERE
8      lo.cd_lingua = 'POR'
```

Figura 4.3 - SQL do critério *idioma nativo*.

4.3.4 Número de renovações

Este critério retorna uma distribuição de valores do tipo inteira, incluindo o *id* da obra, a matrícula do usuário e o escore que será a quantidade de renovações que o usuário realizou.

A figura 4.4 apresenta o SQL de criação da *view.* Nas linhas 2 a 5 seleciona-se o *id* da obra, quantidade de renovações e seu número de matricula. A quantidade de renovações é incrementada em uma unidade para contabilizar a locação do usuário. Nas linhas 6 e 7 é indicado que a seleção será feita da tabela *log_movimentos*. Após esse processo nas linhas 8 a 10 é agrupado por obra e número de matricula.

```
1  CREATE OR REPLACE VIEW argosearch.view_livro_que_aluno_locou_maior_periodo_tempo AS
2  SELECT
3      log_movimentos.nr_obra AS id,
4      sum(log_movimentos.nr_autorenovacoes + 1) AS escore,
5      log_movimentos.nr_matricula AS usuario
6  FROM
7      sab.log_movimentos
8  GROUP BY
9      log_movimentos.nr_obra,
10     log_movimentos.nr_matricula
```

Figura 4.4 - SQL do critério *número de renovações*.

4.3.5 Documento locado anteriormente

Este critério retorna uma distribuição de valores binária, ou seja, o escore será 1 para todos os documentos que o usuário já locou. Além disso, retorna o *id* da obra e o número de matrícula. .

A figura 4.5 apresenta o SQL de criação da *view.* Nas linhas 2 a 5 é selecionado o id da obra, escore e usuário, a cláusula *distinct* é usada para que não haja obras repetidas. Nas linhas 6 e 7 é indicado que a seleção a ser feita será da tabela *obras*. Nas linhas 8 a 12 é realizado um filtro para que ao fazer a seleção das obras retorne apenas as que tiverem o movimento do tipo empréstimo. E para finalizar, nas linhas 13 a 16 a cláusula *inner join* é utilizada para selecionar os usuários referenciados na linha 5.

```
1  CREATE OR REPLACE VIEW argosearch.view_livro_aluno_ja_locou AS
2  SELECT DISTINCT
3      o.nr_obra AS id,
4      1 AS escore,
5      u.nr_matricula AS usuario
6  FROM
7      sab.obras o
8  INNER JOIN
9      sab.log_movimentos lm ON (
10         lm.nr_obra = o.nr_obra
11         AND lm.cd_tipo_movimento = 'E'
12     )
13 INNER JOIN
14     sab.usuarios u ON (
15         u.nr_matricula = lm.nr_matricula
16     )
```

Figura 4.5 - SQL do critério *documento locado anteriormente.*

4.3.6 Número de empréstimos

Este critério retorna uma distribuição de valores inteiros positivos que representa a quantidade de empréstimos de um determinado documento. Também retorna o *id* da obra.

A figura 4.6 apresenta o SQL de criação da *view*. Nas linhas 2 a 6 são selecionadas todas as obras da tabela *log_movimentos* e contados quantos movimentos cada obra possui. Na linhas 7 e 8 utiliza-se um filtro para que apenas os movimentos do tipo empréstimo sejam contados. Nas linhas 9 e 10 agrupa-se as obras que tiverem o mesmo número.

```
1  CREATE OR REPLACE VIEW argosearch.view_numero_de_emprestimos AS
2  SELECT
3      log_movimentos.nr_obra AS id,
4      count(*) AS escore
5  FROM
6      sab.log_movimentos
7  WHERE
8      log_movimentos.cd_tipo_movimento = 'E'
9  GROUP BY
10     log_movimentos.nr_obra
```

Figura 4.6 - SQL do critério *número de empréstimos*.

4.3.7 Número de exemplares

Este critério retorna uma distribuição de valores inteiros positivos que representa será a quantidade de exemplares de um determinado documento, além do *id* da obra.

A figura 4.7 apresenta o SQL de criação da *view*. Nas linhas 2 a 6 são selecionadas todas as obras da tabela *obras* e contados quantos exemplares cada obra possui. Nas linhas 7 a 10 acessamos a tabela *exemplares* já que é nela que se encontra a informação de quantos exemplares cada obra possui. Nas linhas 11 e 12 as obras são agrupadas através de seu identificador.

```
1  CREATE OR REPLACE VIEW argosearch.view_numero_de_exemplares AS
2  SELECT
3      o.nr_obra AS id,
4      count(e.*) AS escore
5  FROM
6      sab.obras o
7  INNER JOIN
8      sab.exemplares e ON (
9          o.nr_obra = e.nr_obra
10     )
11 GROUP BY
12     o.nr_obra
```

Figura 4.7 - SQL do critério *número de exemplares*.

4.3.8 Número de reservas

Este critério retorna uma distribuição valores inteiros postivos que representa a quantidade total de reservas de um determinado documento, além do seu *id*.

A figura 4.8 apresenta o SQL de criação da *view*. Nas linhas 2 a 6 são selecionadas todas as obras da tabela *obras* e contadas quantas reservas cada obra possui. Nas linhas 7 a 10 é feito o acesso a tabela *log_reservas,* pois é nela que consta a informação de quantas reservas uma determinada obra possui. Nas linhas 11 e 12 as obras são agrupadas.

```
1  CREATE OR REPLACE VIEW argosearch.view_numero_de_reservas_total AS
2  SELECT
3      o.nr_obra AS id,
4      count(lr.*) AS escore
5  FROM
6      sab.obras o
7  INNER JOIN
8      sab.log_reservas lr ON (
9          o.nr_obra = lr.nr_obra
10     )
11 GROUP BY
12     o.nr_obra
```

Figura 4.8 - SQL do critério número total de *reservas*.

4.3.9 Taxa de reserva pelo período de tempo no acervo

Este critério retorna uma distribuição de valores reais que representa a taxa de reserva de um determinado documento, além do seu *id*.

A figura 4.9 apresenta o SQL de criação da *view*. Nas linhas 4 a 7 é feito o cálculo da taxa de reserva. Neste cálculo conta-se o total de reservas de uma determinada obra e este valor é dividido pelo tempo em que a obra está presente na biblioteca. Na linha 5 a cláusula *coalesce* possui dois parâmetros: o primeiro é a menor data de aquisição de um exemplar e o segundo é a data de inclusão do exemplar no sistema. Caso o exemplar não possua uma data de aquisição cadastrada, a data de inclusão é usada para efetuar o cálculo.

```
1  CREATE OR REPLACE VIEW argosearch.view_livro_melhor_taxa_reserva AS
2  SELECT
3      o.nr_obra AS id,
4      trunc(count(lr.*)::numeric / ( SELECT
5                                         'now'::text::date - COALESCE(min(e.dt_aquisicao), o1.dt_inclusao)
6                                     FROM
7                                        sab.obras o1
8                                     LEFT JOIN
9                                        sab.exemplares e ON (
10                                         o1.nr_obra = e.nr_obra
11                                    )
12                                    WHERE
13                                         o1.nr_obra = o.nr_obra
14                                    GROUP BY
15                                         o1.nr_obra,
16                                         o1.dt_inclusao
17                                    ), 8) AS escore
18   FROM
19      sab.obras o
20   INNER JOIN
21      sab.log_reservas lr ON (
22          o.nr_obra = lr.nr_obra
23      )
24 GROUP BY o.nr_obra
```

Figura 4.9 - SQL do critério *taxa de reserva pelo período de tempo no acervo*.

5 AVALIAÇÃO EXPERIMENTAL

Uma série de experimentos foram realizados a fim de validar empiricamente e avaliar a qualidade do ARGO*search*. Esses experimentos mostram o impacto do limiar de similaridade adotado e a influência das informações de contexto extraídas do perfil e do sistema de bibliotecas. Os critérios de relevância propostos são avaliados individualmente e em conjunto. Por fim, a ordenação dos resultados é comparada com a métrica *tf x idf* do modelo espacial vetorial, implementado pelo software Lucene[3]. Os resultados apresentados neste capítulo foram sintetizados em um artigo publicado na XXX International Conference of the Chilean Computer Science Society [Borges et al. 2012].

A figura 5.1 exibe a metodologia desta avaliação experimental. Primeiramente, durante um determinado período de tempo, todas as consultas executadas sobre o SAB foram coletadas e armazenadas junto dos resultados apresentados e das transações dos usuários (identificados ou não) aplicadas sobre esses resultados. Após o término da coleta, para cada consulta coletada, os resultados armazenados foram ordenados de acordo com o comportamento simulado do ARGO*search* e do Lucene. Foram gerados diversos rankings a partir de diferentes configurações dos sistemas de recuperação de informações. Esses rankings foram armazenados e comparados em função da qualidade.

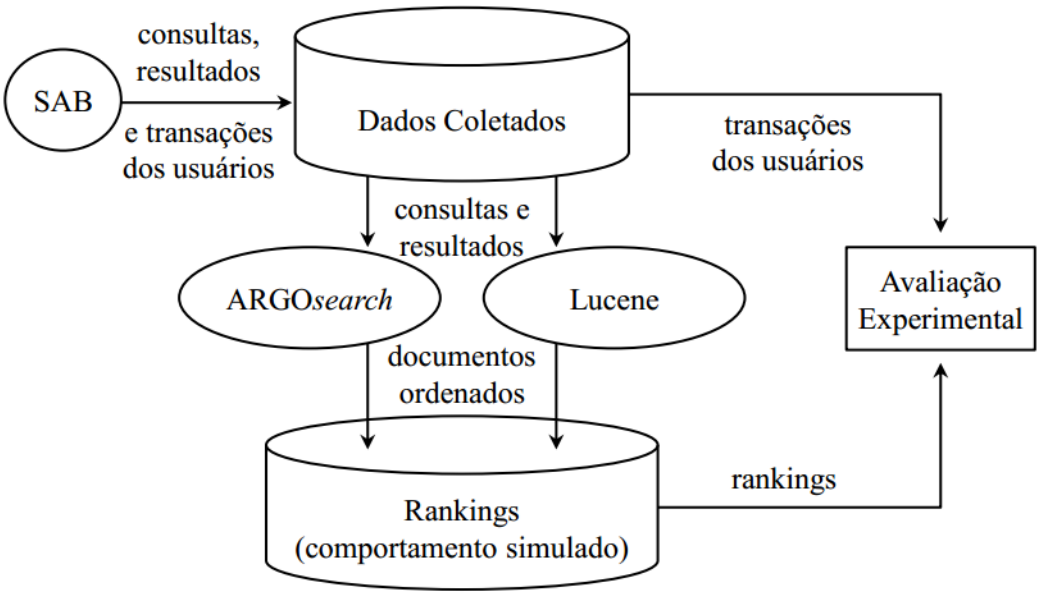

Figura 5.1 - Metodologia da avaliação experimental.

[3] http://lucene.apache.org

As transações dos usuários coletadas anteriormente servem de *feedback* implícito [Joachims et al. 2007] para determinar a importância de um documento retornado. A partir desse *feedback*, a qualidade dos rankings foi medida através da Média das Precisões Médias (*Mean Average Precision* – MAP) [Manning et al. 2008], frequentemente utilizada para avaliar sistemas de recuperação de informações Web, em que não se conhece todo o conjunto de documentos relevantes para cada consulta. Assim como nos motores de busca da Web, um usuário nunca poderá demonstrar interesse por um documento que não tenha sido retornado, portanto não é possível medir a abrangência ou revocação [Manning et al. 2008] dos resultados.

O sistema de administração de bibliotecas da Universidade Federal do Rio Grande – FURG foi utilizado como estudo de caso. A amostra de dados coletada tem 7.758 consultas sobre os metadados autor e título executadas durante 30 dias, os resultados retornados pelo sistema e as transações dos usuários aplicadas sobre esses resultados. Documentos em que o usuário verificou a disponibilidade para empréstimo ou efetuou uma reserva foram considerados relevantes, portanto deveriam aparecer nas primeiras posições dos rankings. É importante ressaltar que a confiança na heurística disponibilidade para empréstimo é bem menor do que nas reservas porque o SAB da FURG apresenta essa informação junto dos metadados que descrevem um documento, após o usuário clicar em um resultado. Portanto, é possível que um usuário tenha investigado os metadados de um documento sem necessariamente estar interessado em verificar se está disponível para empréstimo. Essa baixa confiança é confirmada pelos experimentos apresentados nas subseções a seguir. Em 3.683 consultas (47.5% da amostra) os metadados de pelo menos um dos documentos retornados foram visualizados, mas em apenas 318 destas (4.1% da amostra) os usuários reservaram algum documento. Não existem transações executadas sobre o restante da amostra (4075 consultas, 52.5%), o que pode indicar que o sistema não correspondeu às expectativas do usuário.

A tabela 5.1 apresenta os critérios de relevância usados para configurar o ARGO*search*, organizados de acordo coma fonte de informação extraída. São priorizados na construção dos rankings os documentos em que o metadado idioma é definido como português. A partir das estatísticas de uso do SAB são extraídos de cada documento o número de empréstimos, de reservas e de exemplares,

considerando todos os usuários e todo o tempo de existência dos documentos no acervo. Para evitar que livros mais antigos e, consequentemente, com maior número de reservas e empréstimos fossem demasiadamente priorizados, outro critério considera a razão entre o número de reservas e o tempo do documento no acervo. Entre as informações extraídas do perfil do usuário, são considerados se o usuário já locou o documento anteriormente, e em caso positivo, o número de vezes em que ele renovou o empréstimo. Este critério de relevância prioriza documentos em que o usuário já mostrou forte interesse através do empréstimo e da renovação. Ainda são extraídos outros dois critérios de relevância dos dados acadêmicos, focando no perfil dos estudantes. Os planos de ensino das disciplinas em que o usuário está matriculado são investigados e se um documento estiver presente na bibliografia recomendada, então ele é priorizado na ordenação. Por fim, são analisados os empréstimos de outros estudantes matriculados nas mesmas disciplinas do usuário no mesmo período letivo.

Tabela 5.1 - Descrição dos critérios de relevância utilizados nos experimentos, organizados de acordo com a fonte de informação.

Fonte de Informação	Critério de Relevância
Metadados do repositório (sistema de bibliotecas)	Documento no idioma nativo
Estatística de uso (sistema de bibliotecas)	Número de Empréstimos
	Número de Reservas
	Número de Exemplares
	Taxa de reserva pelo período de tempo no acervo
Perfil do usuário (sistema de bibliotecas)	Documento locado anteriormente
	Número de renovações
Perfil do usuário (dados acadêmicos)	Documento pertence à bibliografia básica recomendada
	Documento locado por colegas de classe no período letivo

5.1 COLETOR

O coletor foi desenvolvido no contexto do projeto como parte de outro trabalho de conclusão de curso [Pereira 2011a]. O banco de dados foi modelado pelo autor do trabalho e a implementação foi contribuição do NTI. O esquema do coletor foi adicionado ao banco de dados do ARGO e permanece rodando diariamente e armazenando todos os movimentos de consulta realizados pelos usuários. A figura 5.2 apresenta o esquema do banco de dados relacional do coletor.

Em relação às consultas os seguintes dados são armazenados no banco:

- A *string* de consulta – entrada do usuário ao realizar a consulta;
- Metadados – campo da consulta como autor, título, edição, etc;
- Data e hora da consulta;
- Endereço IP do computador que o usuário está realizando a consulta;
- Resultado – lista das obras retornadas ao usuário e sua respectiva posição no *ranking;*

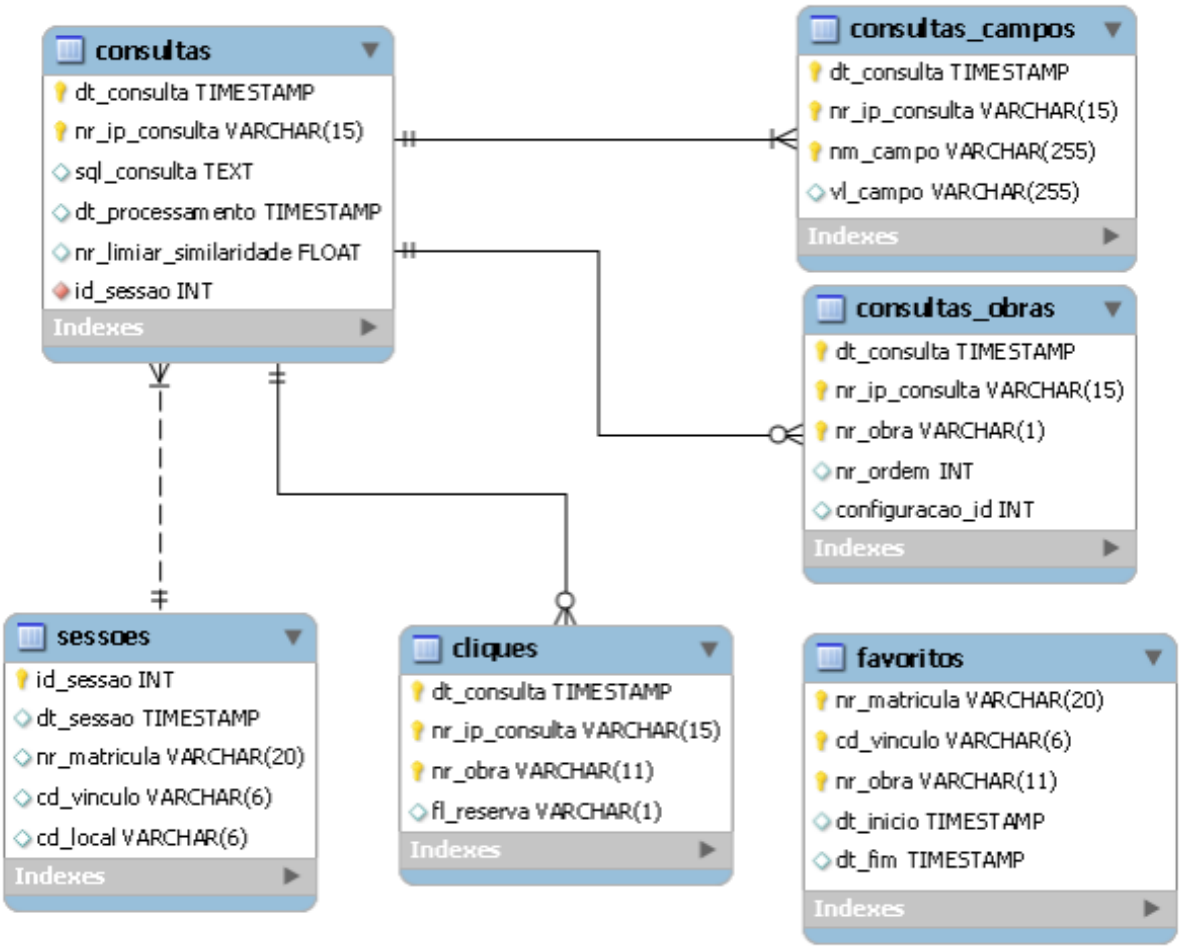

Figura 5.2 - Modelo relacional do Coletor.

5.2 ANÁLISE DO IMPACTO DO LIMIAR DE SIMILARIDADE

A Figura 5.3 apresenta o impacto do limiar de similaridade na qualidade dos rankings gerados pelo ARGO*search*. O sistema foi configurado para ordenar os resultados considerando apenas os escores de similaridade, ou seja, sem considerar os critérios de relevância apresentados na Tabela 5.1. Os gráficos apresentam a Média das Precisões Médias para cada limiar de similaridade, considerando o *feedback* implícito tanto na visualização dos metadados e, consequentemente, na disponibilidade para empréstimo (superior), quanto na reservados documentos (inferior).

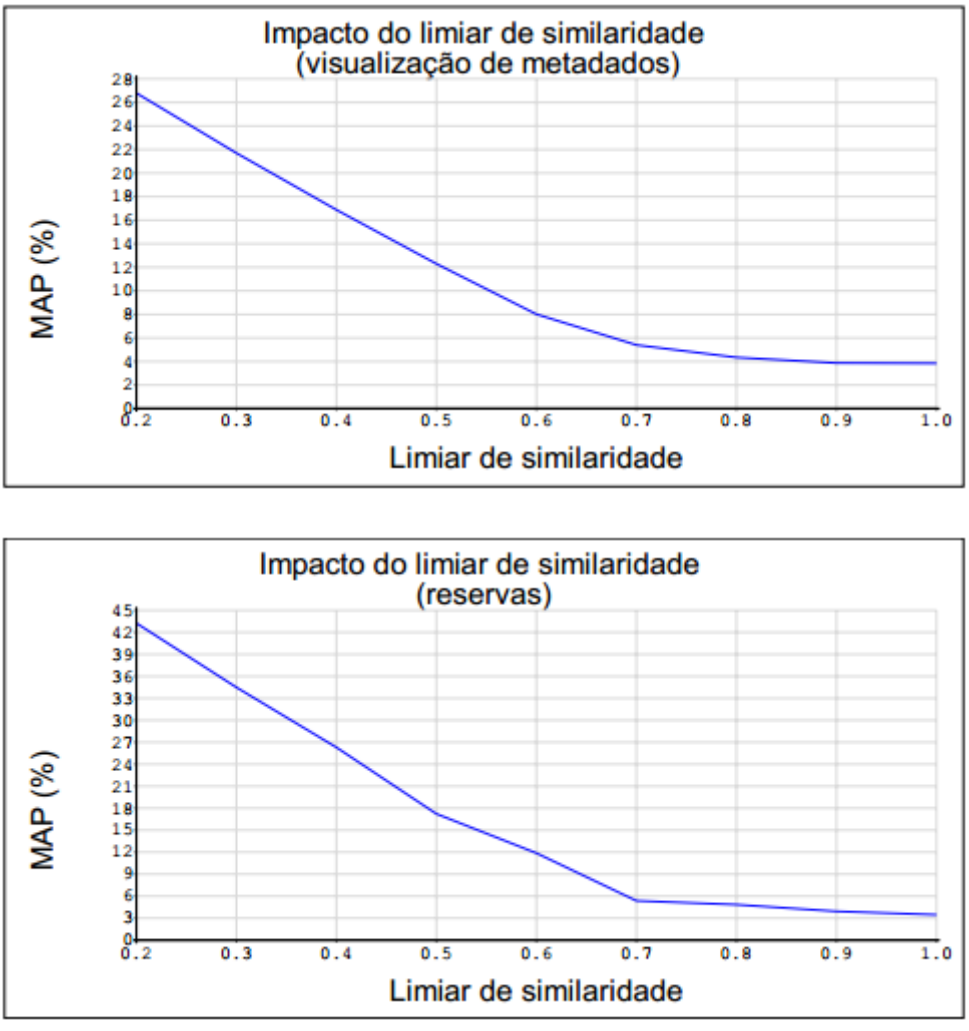

Figura 5.3 - Impacto do limiar de similaridade na qualidade dos rankings considerando o *feedback* implícito na visualização dos metadados (superior) e nas reservas (inferior).

Em geral, os resultados foram consideravelmente melhores para as reservas. Perceba que quanto maior o limiar de similaridade adotado, menor a qualidade da ordenação. A qualidade decresce linearmente até atingir um nível crítico onde a *MAP* = 6%. Este comportamento ocorre porque alguns documentos de interesse do usuário não são recuperados pelo módulo de similaridade e, consequentemente, não são enviados ao módulo de critérios de relevância para serem ordenados. O melhor resultado obtido sem que o custo computacional fosse alterado consideravelmente foi utilizando limiar = 0.2. Os experimentos apresentados nas próximas seções utilizam este valor de limiar.

5.3 INVESTIGANDO OS CRITÉRIOS DE RELEVÂNCIA PROPOSTOS

A Tabela 5.2 apresenta o comportamento do ARGO*search* quando configurado com cada um dos critérios de relevância propostos individualmente, ou seja, quando o peso w_c = 1 para um determinado critério e w_c = 0 para os demais. A qualidade dos rankings gerados é analisada por meio da Média das Precisões Médias considerando o *feedback* do usuário implícito na visualização dos metadados (MAP-VM) e nas reservas (MAP-R). Os resultados foram ordenados de acordo com a MAP-VM.

Observe que qualquer um dos critérios de relevância propostos melhoram os valores de MAP em relação a ordenação considerando apenas os escores de similaridade. O melhor critério de relevância proposto é a taxa de reserva pelo período de tempo no acervo, cuja configuração atingiu MAP de 31 a 64,7%. Este critério está intimamente ligado às outras estatísticas de uso do SAB. Os documentos mais reservados são os de maior procura e, consequentemente, acabam sendo os mais emprestados. Para evitar de pagar multa, os usuários costumam renovar os empréstimos desses documentos com bastante frequência. Além disso, a política de compra da biblioteca favorece a aquisição de documentos frequentemente indisponíveis, ou seja, com maior número de reservas. Portanto, estes mesmos documentos costumam ter um número maior de exemplares.

Tabela 5.2 - Avaliação dos critérios de relevância propostos considerando o feedback implícito na visualização dos metadados (VM) e nas reservas (R).

Critério de Relevância	MAP-VM(%)	MAP-R(%)
Taxa de reserva pelo período de tempo no acervo	31.0	64.7
Número de empréstimos	30.5	55.8
Número de reservas	30.3	58.8
Número de exemplares	29.2	52.2
Número de renovações	28.2	56.3
Documento no idioma português	27.6	45.5
Documento pertence à bibliografia recomendada	27.3	44.8
Documento locado anteriormente	27.3	46.1
Documento locado por colegas no período letivo	27.0	44.0

É importante ressaltar que a alta qualidade dos critérios mencionados (MAP > 50%) deve-se ao fato de que a transação do usuário realizada sobre o resultado das consultas para demonstrar interesse é a reserva. Entretanto, estes são os melhores critérios considerando o *feedback* implícito na visualização dos metadados.

5.4 INFLUÊNCIA DAS INFORMAÇÕES DE CONTEXTO

A Figura 5.4 mostra a influência das informações de contexto extraídas do perfil do usuário e das estatísticas de uso do SAB usadas no algoritmo de ranking do ARGO*search*. Os gráficos apresentam a Precisão Média (*avg P*) de cada consulta em que os metadados de pelo menos um documento retornado foram visualizados (superior) ou em que o usuário realizou uma reserva (inferior). É importante mencionar que as consultas estão ordenadas em termos de *avg P* para facilitar a comparação da qualidade dos rankings.

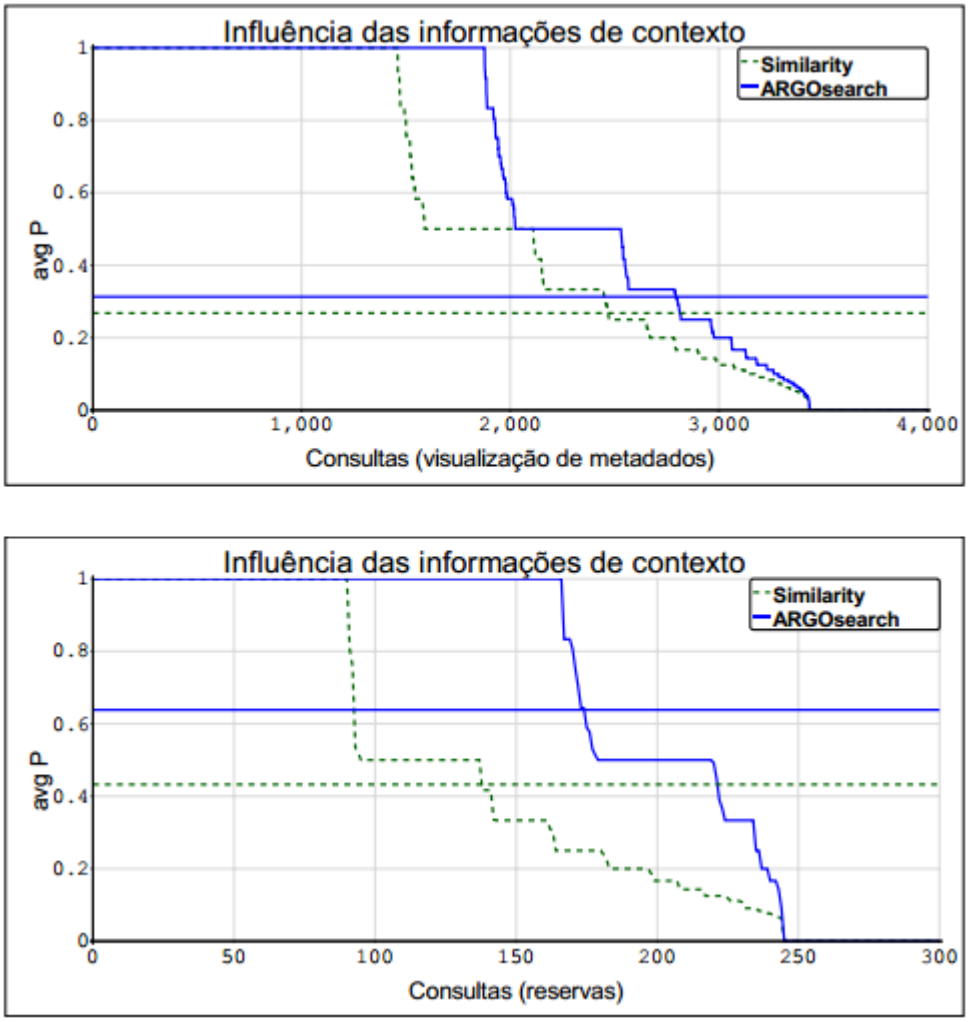

Figura 5.4 - Influência das informações de contexto na qualidade dos rankings considerando o *feedback* implícito na visualização dos metadados (superior) e nas reservas (inferior).

Nos gráficos da figura 5.4 a série representada por linha tracejada corresponde ao ARGO*search* configurado para que utilize apenas o escore de similaridade para ordenar os documentos. Esta configuração atingiu *MAP* = 26.8 e 43.3% (retas horizontais) considerando visualização de metadados e reservas respectivamente. Na série representada pela linha cheia, são utilizados a similaridade e os demais critérios apresentados na Tabela 5.2 com W_C = 1, ou seja, são levadas em conta as informações de contexto com a mesma importância para cada critério. Esta configuração atingiu *MAP* = 31.3% (visualização de metadados) e 63.8% (reservas).

É importante destacar a contribuição dos critérios de relevância que consideram o contexto, principalmente para as reservas, onde a *MAP* aumentou 47%. Das 318 consultas em que algum documento foi reservado, 167 atingiram *avg P* = 100%, o que representa um aumento de 83% quando comparada às 91 consultas da configuração que utiliza apenas a similaridade.

Apesar dos rankings gerados pela combinação dos critérios de relevância terem maior qualidade, a consulta aproximada é essencial para aumentar a abrangência dos resultados e lidar com as variações de grafia na forma como os nomes de autores são representados em referências bibliográficas [Borges et al. 2011]. Esta hipótese é apoiada pela informação apresentada na figura 5.3, a qual mostra que quanto maior o limiar de similaridade da consulta, menor a cobertura dos resultados e a qualidade dos rankings.

5.5 COMPARAÇÃO COM O *BASELINE*

A Figura 5.5 apresenta uma comparação entre o sistema ARGO*search* e o modelo espacial vetorial, implementado pelo software Lucene. Os gráficos apresentam as mesmas informações da Figura 5.4 (*avg P* por consulta).

O Lucene, representado pela série com linha tracejada, atingiu *MAP* = 25.6 e 40.4% considerando, respectivamente, visualização de metadados e reservas. O ARGO*search* (linha cheia) marcou *MAP* = 31.3 e 63.8%, o que significa um acréscimo na qualidade de 22 a 58% dependendo da heurística adotada como *feedback* do usuário.

Analisando as 318 consultas em que algum documento foi reservado, 167 atingiram *avg P* = 100%, o que representa um aumento significativo quando comparadas às 77 consultas do Lucene.

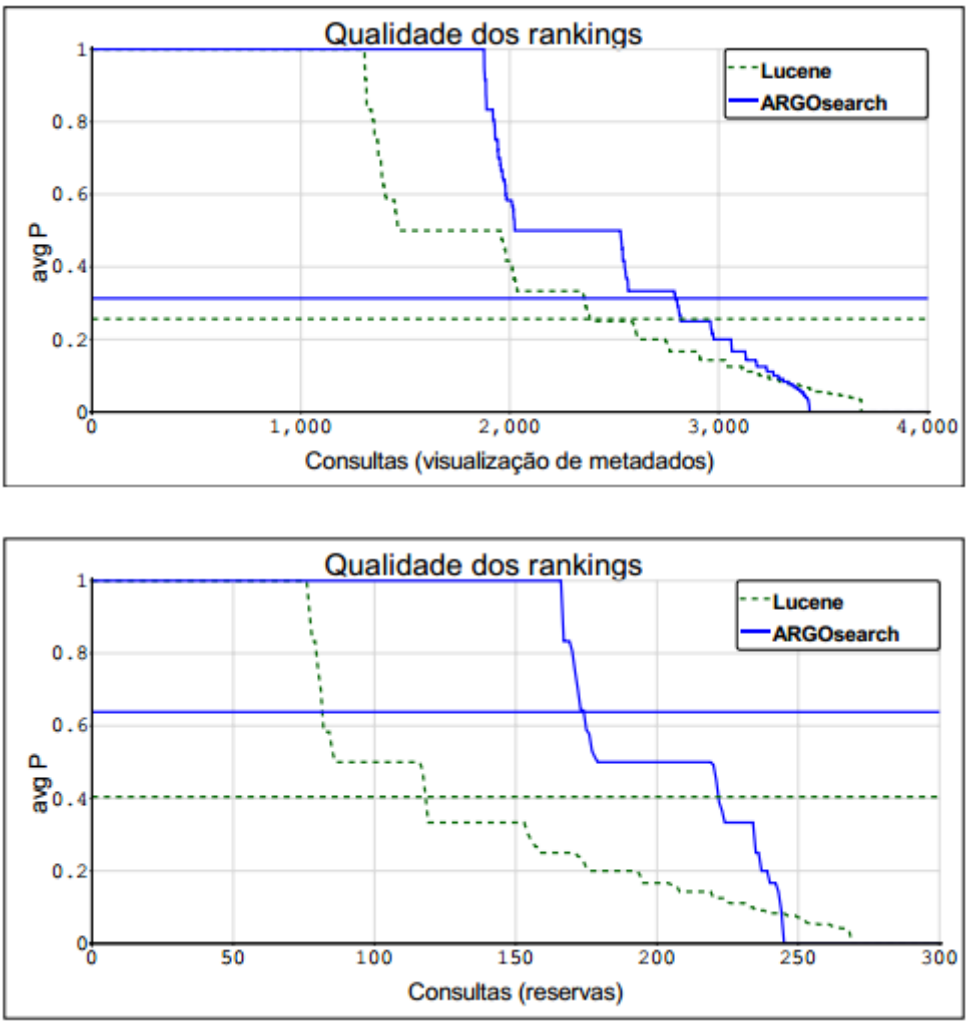

Figura 5.5 - Qualidade dos rankings gerados pelo ARGO*search* e pelo Lucene considerando o *feedback* implícito na visualização dos metadados (superior) e nas reservas (inferior).

6 IMPLANTAÇÃO

Neste capítulo são descritas as etapas realizadas para a implantação do sistema junto ao NTI da Universidade. Em um primeiro momento, foi realizada uma reunião onde estavam presentes os autores do projeto ARGO*search* e o analista do NTI Carlos Alberto Madsen, responsável pelo sistema de biblioteca (ARGO) da FURG. O ARGO*search* foi apresentado ao analista para que ele pudesse entender e se familiarizar com a estrutura proposta. Ao final da apresentação do sistema e do código-fonte, foram definidos as seguintes etapas da implantação: adaptação do código-fonte utilizando o framework de desenvolvimento padrão do NTI, materialização das visões que implementam os critérios, implementação de uma ferramenta de gerenciamento e configuração do ARGO*search* e a disponibilização do sistema para teste.

O código-fonte do ARGO*search* foi remodelado para a estrutura do *framework* CASCA[4]. Os testes do novo código revelaram um problema de desempenho que poderia ser facilmente resolvido com visões materializadas na implementação dos critérios de relevância. Desta forma, obtemos um ganho de *performance*, não sendo necessário fazer determinados cálculos uma vez que o resultado já estaria materializado no banco de dados.

Tomando por base a decisão de utilizarmos as visões materializadas, foi preciso realizar a criação de *scripts* que seriam executados com determinada periodicidade, variando de acordo com a necessidade, para materializar os dados no banco. Por exemplo, para o critério *documento pertence à bibliografia básica recomendada*, bastaria materializar o critério a cada novo semestre porque esta é a frequência de alteração dos planos de ensino das disciplinas. Já o critério *número de reservas* precisa ser materializado diariamente porque ocorrem muitas reservas neste período considerando todos os usuários.

A aplicação para gerenciamento do sistema construída permite criar novos critérios e configurar os pesos de cada um. São especificadas novas configurações compostas de pesos para a similaridade textual e para cada um dos critérios implementados. O peso *zero* desabilita um critério, passando a não ser considerado

[4] http://www.cpd.furg.br/bin/doc_casca/index2.html

no cálculo da relevância. A criação desta aplicação foi realizada utilizando a linguagem *PHP*, SGBD PostgreSQL e o *framework* CASCA.

Na Figura 6.1 pode ser visto a tela da aplicação criada com a aba configurações aberta. Nesta aba é possível ver a listagem de todas as configurações criadas, adicionar uma nova configuração, editar, excluir ou abrir uma determinada configuração.

Figura 6.1 - Tela da aplicação criada no *Framework* CASCA

Ao selecionar uma determinada configuração e clicar na pasta (botão para abrir a configuração) é apresentada a tela da Figura 6.2, onde pode ser visto o nome da configuração, o valor do limiar de similaridade utilizado, o peso do limiar de similaridade e se esta configuração está ativa ou não no sistema. Estes parâmetros são os mesmos que são preenchidos ao criar uma nova configuração. Logo abaixo, na aba critérios, é possível saber quais são os critérios que estão presentes nesta configuração e seus respectivos pesos.

Por fim, foi criado um domínio no servidor de produção do NTI para testar o sistema. Este está disponível para testes em http://argoteste.furg.br

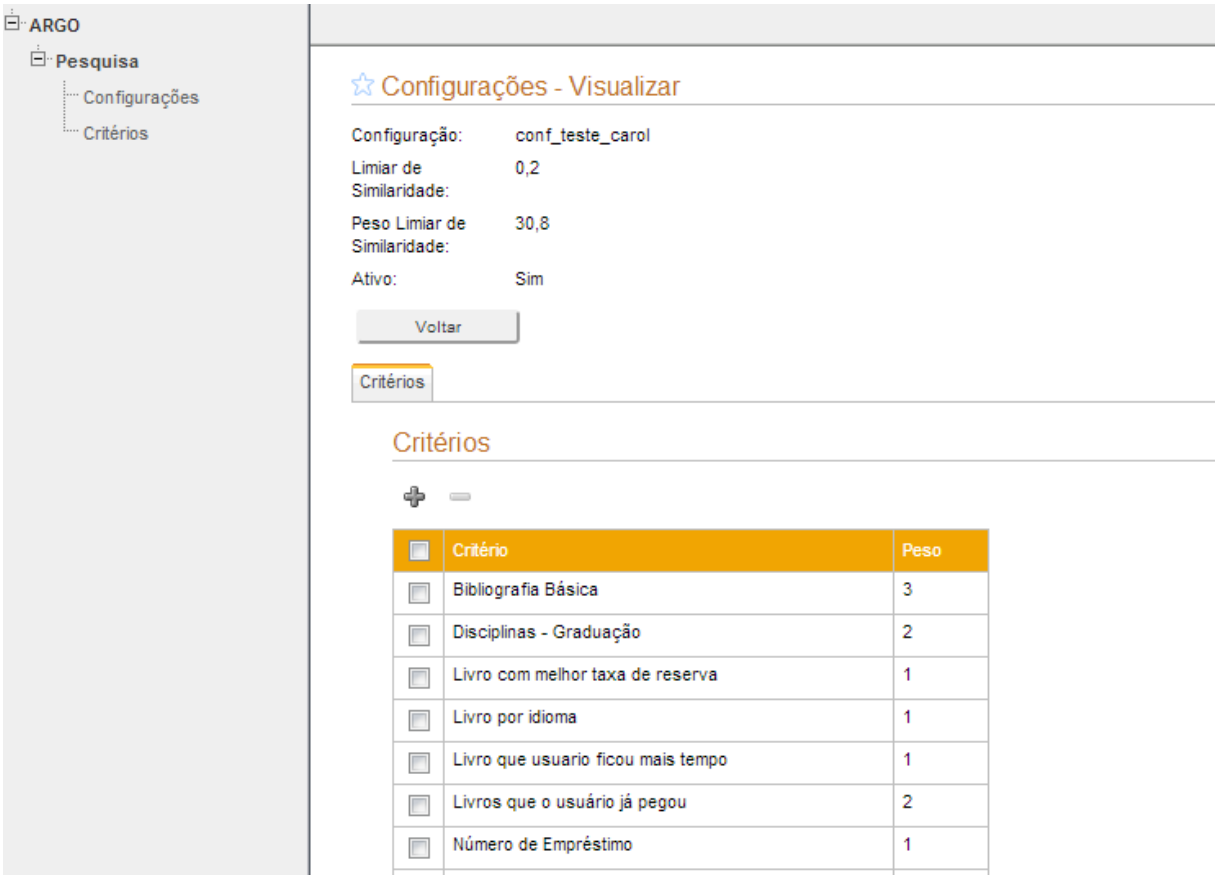

Figura 6.2 - Tela da aplicação quando visualizamos uma determinada configuração

7 CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho apresenta o módulo de critérios de relevância para o ARGO*search*, um sistema de recuperação de informações baseado em similaridade e critérios de relevância extensíveis que melhora a qualidade do resultado das consultas realizadas no ARGO. Os critérios de relevância são modelados como funções que participam do processo de ordenação de um conjunto de documentos. Eles podem ser definidos a partir dos metadados descritivos da coleção, de estatísticas de uso do ARGO e de informações de contexto extraídas do perfil do usuário. Um administrador da biblioteca pode especificar novos critérios de relevância sem conhecimento de programação de computadores.

Foram definidas duas métricas para ordenação dos resultados. O potencial de relevância normaliza os escores retornados pelos critérios para que valores com distribuições heterogêneas possam ser comparáveis. A estimativa de relevância pondera os potenciais de relevância de acordo com a importância de cada critério e os combina em um único valor usado para ordenar os documentos e gerar o *ranking* final.

Por ser uma arquitetura genérica baseada em critérios de relevância extensíveis, o ARGO*search* pode ser adotado em quaisquer sistemas que necessitem ordenar os resultados de uma consulta de acordo como perfil do usuário que a executa, tendo como vantagem adicional considerar variações de grafia já que realiza consultas por similaridade. O ARGO*search* aumenta a cobertura dos resultados quando comparado aos modelos tradicionais de recuperação de informações, retornando um número maior de possíveis documentos de interesse do usuário. Os experimentos realizados mostram que, quando comparado ao modelo vetorial de recuperação de informações implementado pelo software Lucene, considerando a métrica MAP, o ARGO*search* melhorou em até 58% a qualidade dos *rankings*.

A estratégia adotada pelas métricas de ordenação permite que um especialista faça um ajuste fino do ARGO*search*, vinculando os pesos adequados a cada critério de relevância. Futuramente pretende-se implementar um novo componente para o sistema ARGO*search* que utilize aprendizagem de máquina para determinar os melhores parâmetros da sintonia fina.

A implantação definitiva do ARGO*search* no ARGO da FURG foi conduzida em conjunto com o NTI. Futuramente, pretende-se analisar o custo computacional envolvido tanto no cálculo dos critérios de relevância quanto na busca por similaridade. Se necessário, serão aplicadas as técnicas de otimização dos algoritmos para reduzir esse custo. Além disso, novas análises poderão ser realizadas usando como *feedback* do usuário outras transações como o tempo em que permaneceu visualizando os metadados e a rolagem de página.

REFERÊNCIAS

A. Al-Maskari and M. Sanderson, "*The effect of user characteristics on search effectiveness in information retrieval*," Inf. Process. Manage., vol. 47, no. 5, pp. 719–729, 2011.

Angell, R. C.; Freund, G. E. and Willett, P. "Automatic spelling correction using a trigram similarity measure," Inf. Processing and Management, vol. 19, no. 4, p. 255-261, 1983.

Baeza-Yates, R. A. and Ribeiro-Neto, B. A., Modern Information Retrieval. ACM Press / Addison-Wesley, 1999.

Borges, E. N.; de Carvalho, M. G.; Galante, R.; Gonçalves, M. A. and Laender, A. H. F. "An unsupervised heuristic-based approach for bibliographic metadata deduplication," Inf. Processing and Management., vol. 47, no. 5, pp. 706-718, 2011.

Borges, E. N.; Pereira, I. A.; Tomasini C. and Prisco, A. "ARGOserach: an information retrieval system based on text similarity and extensible relevance criteria", in XXXI International Conference of the Chilean Computer Science Society (SCCC), Valparaíso. p. 47-53, 2012.

Cohen, W. W.; Ravikumar, P. and Fienberg, S. E. "A comparison of string distance metrics for name-matching tasks," in IIWeb, 2003, pp. 73-78

Elmasri, R. E. e Navathe, S. "Sistemas de Banco de Dados", AddisonWesley, 2005.

Joachims, T.; Graka, L.; Pan, B.; Hembrooke, H.; Radlinski, F. and Gay, G. "Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search." ACM Trans. Inf. Syst., vol. 25, no. 2, 2007.

Manning, C. D.; Raghavan, P. and Schutze H. Introduction to Information Retrieval. Cambridge University Press, 2008.

Pereira, I. A. ARGOfeedback – uma Ferramenta de apoio à decisão para o ARGOsearch. Monografia de Graduação em Engenharia de Computação, FURG, Rio Grande, 2011.

Pereira, I. A. Uma Arquitetura para Pesquisa Por Relevância Baseada em Critérios para o ARGOsearch. Monografia de Graduação em Tecnologia em Análise e Desenvolvimento de Sistemas, FURG, Rio Grande, 2011.

Pereira, I. A.; Tomasini, C.; Prisco, A. e Borges, E. N. "Um mecanismo de busca para sistemas de gerenciamento de bibliotecas baseado em critérios de relevância extensíveis", in Anais da Conferência Sul em Modelagem Computacional, pp. 7-12, 2012.

Salton, G. and Bucckley, C. Term-weighting approaches in Automatic Retrieval. In Information Processing & Management, vol. 24, no. 5, p. 513-523, 1988.