ALLAN JONES COSTA E SILVA

# Estratégias para a Busca do Texto Completo de Artigos Catalogados em uma Biblioteca Digital

Belo Horizonte
26 de março de 2007

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

# Estratégias para a Busca do Texto Completo de Artigos Catalogados em uma Biblioteca Digital

Dissertação apresentada ao Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

ALLAN JONES COSTA E SILVA

Belo Horizonte
26 de março de 2007

# Resumo

Esta dissertação propõe um processo que utiliza resultados de consultas submetidas a máquinas de busca para encontrar a URL do texto completo correspondente, ou de qualquer outro material relacionado, a artigos catalogados em uma biblioteca digital que não possuem tal informação registrada. Apresentamos um estudo desse processo para investigar diferentes estratégias de consultas aplicadas a três máquinas de busca de propósito geral (Google, Yahoo!, MSN) e a duas especializadas (Scholar e CiteSeer) considerando vários cenários caracterizados por usuários com diferentes níveis de exigências. Especificamente, conduzimos um conjunto de experimentos com artigos provenientes da BDBComp – Biblioteca Digital Brasileira de Computação e da DBLP – Digital Bibliography & Library Project. De acordo com os resultados, Scholar mostrou-se mais eficaz nesta tarefa do que as outras máquinas de busca testadas em todos os cenários estudados. Além disso, nossos experimentos mostraram que estratégias simples para combinação e reordenação fornecem resultados ainda melhores. Nosso estudo também apresenta uma análise do impacto de diferentes fatores na chance de se encontrar o texto completo dos artigos procurados.

# Abstract

This dissertation proposes a process that uses results from queries submitted to search engines for finding the URL of the corresponding full-text, or of any relevant related material, for those articles cataloged in a digital library for which this information is missing. We present a comprehensive study of this process in different situations by investigating different query strategies applied to three general purpose search engines (Google, Yahoo!, MSN) and two specialized ones (Scholar and CiteSeer), considering five user scenarios characterized by distinct requirement levels. Specifically, we have conducted a set of experiments focused on articles taken from BDBComp – Brazilian Digital Library of Computing and DBLP – Digital Bibliography & Library Project. According to the results of these experiments, Scholar has shown to be more effective than the other tested search engines for this task in all considered scenarios. Moreover, our experiments show that a simple combination Scholar-Google with a re-ranking strategy provides even better results. Our study also presents an analysis of the impact of different factors on the likelihood of finding the full-text of the searched articles.

*Viver é buscar incessantemente. Isto explica, em parte, o sucesso delas!*

# Agradecimentos

Agradeço, especialmente, às pessoas que me amam. Espontaneamente, alegram-se quando estou feliz e, sobretudo, manifestam apoio, quando algo me entristece.

Obrigado, mãe! Seu aconselhamento e exemplo de luta são muito importantes. Pai, obrigado por acreditar na minha capacidade e de meus irmãos. Caros irmãos, agradeço pela lição que tenho aprendido com vocês: amar-nos apesar de nossas diferenças.

Agradeço também a meus parentes. Destaco alguns. Márcia, obrigado por me auxiliar na minha transição de Ouro Branco para Belo Horizonte. Niquinho e Vovó, o auxílio de vocês a minha família em momentos de grande dificuldade é algo que está gravado em mim.

Grato também estou às pessoas que se dispõem a partipar de minha vida, amigos. Obrigado aos eternos e aos inesquecíveis: Cristiano, Selma, Carolina e Perini. Obrigado, Karla, Bárbara, Hugo, Rodrygo, Guilherme e a todos os demais companheiros com quem estive envolvido no LBD! Com vocês, o trabalho foi muito mais alegre. Outros amigos, obrigado!

Lembro entidades imprescindíveis: Programa de Moradia Universitária da UFMG, CNPq e CAPES.

Aos orientadores Alberto Laender e Marcos Gonçalves, obrigado pela oportunidade e acompanhamento.

À Deus, obrigado...

# Sumário

# Lista de Figuras

# Lista de Tabelas

# Capítulo 1

# Introdução

A provisão de acesso ao texto completo das publicações é um importante requisito para satisfação das necessidades e expectativas dos usuários de uma biblioteca digital de artigos científicos. Entretanto, em muitas dessas bibliotecas digitais, sobretudo naquelas construídas a partir da agregação de dados de fontes heterogêneas, nem todos os registros de metadados possuem um apontador direto (por exemplo, uma URL) para o texto completo[1] do artigo correspondente. Além do casos em que não há um apontadores para o texto de artigos pertencentes à coleção da biblioteca digital, este problema também muito comum quando se encontram disponíveis no catálogo metadados de artigos referenciados pelos artigos da coleção. O usuário pode se interessar pelos artigos referenciados, entretanto, em muitos casos, a biblioteca digital não fornece qualquer meio para acesso ao texto. Mesmo a presença do apontador no registro de metadados pode não ser de utilidade. Pode ser que, o acesso seja fornecido mediante pagamento e o usuário não esta preparado ou disposto a efetuá-lo, ou ainda, no momento da catalogação do artigo, a URL apontava para o texto do artigo descrito, mas, posteriormente, devido à dinâmica da Web, tornou-se inválida. Nestes casos, um serviço capaz de listar outras fontes que forneçam acesso ao respectivo texto seria de grande utilidade para os usuários de bibliotecas digitais.

Uma alternativa a ser seguida por um usuário que se depara com metadados de um artigo para o qual não se conhece um apontador para o respectivo texto é, a partir dos metadados apresentados, efetuar consultas a máquinas de busca da Web e inspecionar as respostas retornadas até encontrar uma URL que forneça acesso ao texto procurado. Nesta dissertação, propomos um processo para automatizar tal procedimento e, conseqüentemente, diminuir o esforço despendido pelo usuário na tarefa de busca pelo texto do artigo de interesse. Descrevemos e avaliamos experimentalmente diversos aspectos

[1]Nesta dissertação, sempre que referirmos ao texto de um artigo, estaremos nos referindo ao seu texto completo.

desse processo que pode utilizar máquinas de busca de próposito geral e especializadas para recuperação de URLs que forneçam acesso a textos de artigos catalogados em uma biblioteca digital, serviço notadamente útil nas situações em que tal informação não é fornecida, embora existam metadados que descrevam outras propriedades, tais como o título e autores desses artigos.

A idéia é explorar e avaliar a potencialidade das máquinas de busca, já que, mesmo frente ao vasto conteúdo da Web, testes manuais preliminares mostraram que, quando consultadas adequadamente, tendem a fornecer boas respostas a usuários envolvidos com a tarefa que descrevemos. Além disso, há outros aspectos a serem considerados ao aplicar tal abordagem. Pode-se, adicionalmente, encontrar outros documentos relacionados ou com informações potencialmente úteis que satisfaçam parcialmente às necessidades dos usuários. Por exemplo, as máquinas de buscas podem retornar como resposta um apontador para outra publicação (tal como uma tese, dissertação ou artigo semelhante) de um dos autores do artigo procurado ou, ainda, páginas que contenham metadados adicionais (tais como referências ou um resumo) que complementem a informação já presente no catálogo da biblioteca digital e que também podem ser úteis. Outro aspecto a ser considerado é que o conteúdo recuperado pode estar livremente disponível na Web (como, por exemplo, na página pessoal de um dos autores) ou pode ser proveniente de fontes com restrição de acesso (como, por exemplo, o *site* ou biblioteca digital de uma editora), situação na qual o usuário deve estar preparado para pagar por ele.

Dessa forma, o problema tratado nesta dissertação possui diversas nuances ainda não exploradas na literatura. Assim, conduzimos um trabalho cujo relato traz diversas contribuições resumidas neste capítulo introdutório que está organizado da seguinte forma. A Seção 1.1 resume as contribuições geradas por nosso estudo. A Seção 1.2 faz uma revisão dos trabalhos relacionados. Por fim, a Seção 1.3 descreve a organização do resto desta dissertação.

## 1.1 Contribuições

A eficácia do processo que propomos para a busca do texto dos artigos depende de diversos aspectos. A fim de entender como obter bons resultados, estudamos a eficácia da utilização de diferentes estratégias de consultas, submetidas a diferentes máquinas de busca considerando vários cenários caracterizados por usuários de distintos graus de exigência. Especificamente, como primeira etapa, investigamos como especificar, para cada máquina de busca, as consultas mais adequadas para a tarefa em questão, e qual das máquinas de busca seria a mais eficaz em cada cenário.

Além do estudo da eficácia das diferentes estratégias de consulta e do desempenho individual das diferentes máquinas de busca, outro objetivo deste trabalho é propor maneiras de melhorar a qualidade dos resultados obtidos. Dessa forma, também experimentamos estratégias para combinar e reordenar respostas retornadas por diferentes máquinas de busca. Com tais estratégias obtivemos uma melhora significativa na eficácia do processo quando comparado à utilização individual da mais eficaz das máquinas de busca testadas em cada cenário.

O processo proposto pode ser ajustado conforme requisitos de custo de excução das consultas. Fornecemos uma análise que mostra uma diminuição do custo de execução das consulta vem acompanhado de uma diminuição na eficácia das respostas retornadas. Tal observação evidencia um *trade-off* entre os aspectos de eficiência e eficácia do processo proposto.

Além das estratégias e máquinas de busca utilizadas e do custo na busca do texto dos artigos, há outras características que influenciam a chance de o usuário obter sucesso. Assim, analisamos e discutimos o impacto de alguns outros fatores adicionais na chance de se encontrar o conteúdo desejado. Mais especificamente, analisamos a influência do grau de exigência do usuário, o idioma no qual o artigo foi escrito e a data de sua publicação.

Nosso estudo está embasado em experimentos conduzidos com registros do catálogo de metadados da BDBComp – *Biblioteca Digital Brasileira de Computação*[2] – complementados com um conjunto de registros extraídos da DBLP – *Digital Bibliography & Library Project*[3] que correspondem a artigos publicados por pesquisadores brasileiros em conferências internacionais não catalogados na primeira coleção. Selecionamos aleatoriamente um conjunto de registros provenientes dessas duas coleções para os quais não há uma URL do texto, e coletamos e analisamos manualmente as respostas retornadas por consultas de sete tipos diferentes submetidas a três máquinas de busca de propósito geral (Google, Yahoo! e MSN) e a duas especializadas (Scholar e CiteSeer) com o objetivo de recuperar URLs que forneçam acesso aos textos completos correspondentes.

Nossos resultados experimentais demonstram que o processo é efetivo e fornece uma estratégia bem simples para se encontrar textos completos de artigos catalogados em bibliotecas digitais para os quais não há um apontador para o texto dos artigos. Além disso, o processo proposto pode ser utilizado em outras aplicações, como por exemplo, busca de metadados adicionais e mais atualizados dos artigos catalogados. Também é mostrado que, entre as máquinas de busca testadas, o Scholar é a mais efetiva para a tarefa e que, quando suas respostas são combinadas com as do Goo-

[2] http://www.lbd.dcc.ufmg.br/bdbcomp/

[3] http://www.informatik.uni-trier.de/~ley/db/

gle, ganhos significativos são obtidos em todos cenários considerados. É importante ressaltar que, embora os nossos experimentos tenham sido realizados com artigos de Ciência da Computação, área reconhecidamente bem representada na Web, estudos recentes (Walters, 2007) mostram que máquinas de busca, tais como o Scholar, também cobrem de maneira razoável o conteúdo de outras áreas do conhecimento. Portanto, como o processo proposto depende somente dos metadados presentes no catálogo da biblioteca digital, acreditamos que bons resultados também possam ser obtidos para outras áreas desde que possuam uma representatividade na Web a ponto de possuir conteúdo indexado por máquinas de busca.

Em suma, as principais contribuições desta dissertação são (Silva et al., 2006, 2007):

1. A proposta de um processo para encontrar a URL do texto completo, ou conteúdo relacionado relevante, de artigos catalogados em uma biblioteca digital para os quais tal informação não está presente nos registros de metadados.

2. Um estudo desse processo considerando diferentes situações: diferentes estratégias de consultas submetidas a diferentes máquinas de busca considerando usuários de diferentes necessidades e perfis, com diferentes graus de exigência.

3. A proposição de estratégias para combinação do uso de máquinas de busca específicas e para reordenação das respostas retornadas que melhoram os resultados obtidos.

4. Uma análise que relaciona o custo da execução das consultas e eficácia das respostas retornadas evidenciando um *trade-off* entre os apectos eficácia e eficiência do processo proposto.

5. Uma análise do impacto de diferentes fatores na chance de sucesso dos usuários na busca realizada: cenário ao qual os usuários pertencem, idioma em que o artigo procurado foi escrito e data de sua publicação.

## 1.2 Trabalhos Relacionados

Abordagens atuais para busca de documentos ausentes em bibliotecas digitais são baseadas em coletores temáticos. Como caracterizado por Chakrabarti et al. (1999), "um coletor temático busca, coleta, indexa e mantém documentos sobre tópicos específicos, que representam um segmento relativamente pequeno da Web". Essa meta é atingida através da definição de um esquema complexo de prioridades que guia o coletor por documentos relevantes da Web para construir a coleção temática local. Zhuang et al.

(2005) investigam a viabilidade de se usar metadados de publicações para guiar o coletor até páginas pessoais de pesquisadores e coletar documentos ausentes na coleção de uma biblioteca digital.

Diversos algoritmos para coletores temáticos estão discutidos e avaliados na literatura (Menczer et al., 2004; Pant et al., 2004). Coleta temática de alta qualidade é um requisito importante na construção de máquinas de busca especializadas. Tais sistemas oferecem serviços de busca em coleções de documentos da Web sobre tópicos específicos. Porém, usar coletores temáticos para buscar por itens ausentes em uma coleção requer a construção de uma complexa infra-estrutura de *software*. Portanto, advogamos que o uso do vasto conteúdo já indexado pelas máquinas de busca possibilita alcançar resultados satisfatórios com um pequeno esforço, o de formular consultas adequadas a essa tarefa.

Alguns sistemas que fornecem serviços de busca e coleta de publicações científicas estão relatados na literatura. HPSearch e Mops, descritos por Hoff e Mundhenk (2001), são capazes de buscar artigos que estão *próximos de* páginas pessoais de pesquisadores. Paper Search Engine (PaSE), proposto por On e Lee (2004), usa informações de citação para localizar e coletar cópias de artigos na Web. Entretanto, esses trabalhos têm foco em auxiliar na busca de artigos científicos de modo geral. Nossa proposta é a fornecer subsídios para a arquitetura de um serviço que auxilie na busca daqueles que estão catalogados em uma biblioteca digital, mas sem apontador para o respectivo texto completo.

Lawrence (2001) demonstrou que quanto maior o número de citações de um artigo científico, maior a chance de o texto completo desse o artigo estar disponível *online*. Ou seja, a disponibilidade do texto na Web é um dos pré-requisitos fundamentais para que um artigo tenha maior probabilidade de ser citado. Outro pré-requisito importante é que o artigo possa ser encontrado na Web. Máquinas de busca são as principais ferramentas utilizadas para encontrar informação *online*.

Estratégias de recuperação de informação na Web constituem um tópico de pesquisa intenso e essencial para o desenvolvimento das máquinas de busca. Nesta dissertação avaliamos a eficácia de se submeter consultas a máquinas de busca especializadas para busca de itens ausentes comparada com a aplicação da mesma estratégia a máquinas de busca de propósito geral que indexam todo tipo de documento. Dessa forma, tiramos proveito de toda a tecnologia já desenvolvida nas máquinas de busca com o simples esforço de fornecer consultas adequadas a essa tarefa. Entretanto, a especificidade de cada uma das máquinas de busca torna necessário compará-las em termos de adequação à tarefa descrita. São muito freqüentes trabalhos que comparam a eficácia de máquinas de busca de propósito geral para satisfazer necessidades comuns de informação, tais como aqueles descritos por Lawrence e Giles (1999a), Gordon e Pathak (1999), Bharat

e Broder (1998), Lawrence e Giles (1998) e Chu e Rosenthal (1996). Porém, não obtivemos sucesso em encontrar trabalhos que comparem o uso de máquinas de busca especializadas e de propósito geral para a tarefa específica abordada nesta dissertação.

Conforme caracterizado por Gonçalves et al. (2004), o provimento de um serviço de busca é requisito básico para uma biblioteca digital. Normalmente, tais serviços fazem busca restrita ao contexto interno dos catálogos de metadados da bibliotecas digital que os fornece. Em nosso trabalho descrevemos um processo para encontrar no contexto da Web informações ausentes do catálogo, tal como uma URL para o texto completo. Respostas de consultas a máquinas de busca são utilizadas como fontes de apontadores para documentos da Web. Tal abordagem pode ser um caminho interessante para a solução de problemas relacionados ao conteúdo das bibliotecas digitais.

Combinar resultados de várias máquinas de busca em um único conjunto de respostas é uma técnica conhecida como *meta-search* (Selberg e Etzioni (1995); Howe e Dreilinger (1997)). Em nosso trabalho, combinamos resultados de máquinas de busca específicas em um único conjunto de respostas com o intuito de se aumentar a chance de encontrar o texto dos artigos procurados. O trabalho descrito por Lawrence e Giles (1999b) sugere esse caminho, entretanto não são propostas técnicas para automatizar a tarefa.

## 1.3 Organização

Levando em conta os aspectos que discutimos, o resto desta dissertação está organizado como segue. O Capítulo 2 define conceitos básicos de bibliotecas digitais e recuperação de informação utilizados ao longo da dissertação e descreve o processo proposto para recuperação das URLs ausentes no catálogo de uma biblioteca digital. O Capítulo 3 descreve a metodologia experimental seguida para condução e avaliação dos resultados experimentais. O Capítulo 4 relata e discute os resultados experimentais obtidos levando em conta os vários aspectos abordados. O Capítulo 5 analisa o impacto de diversos fatores externos na chance de se encontrar o conteúdo procurado. Por fim, o Capítulo 6 apresenta nossas conclusões e discute trabalhos futuros. Resultados experimentais adicionais cuja análise deixamos para o leitor são apresentados no Apêndice A.

# Capítulo 2

# Processo Proposto

Neste capítulo descrevemos nossa proposta de solução para o problema de busca na Web do texto completo de artigos catalogados em uma biblioteca digital. Para melhor entendimento do leitor, a Seção 2.1 explica alguns conceitos básicos e termos utilizados ao longo do texto. A Seção 2.2 descreve o processo proposto como solução do problema em questão. A Seção 2.3 lista os pontos de pesquisa a serem abordados nos próximos capítulos cuja análise é essencial para uma boa fundamentação de nosso estudo.

## 2.1 Conceitos Básicos e Terminologia

Conforme descrito por Gonçalves et al. (2004), "uma biblioteca digital engloba uma coleção controlada de informações com serviços associados que envolvem comunidades, onde as informações estão armazenadas no formato digital e são acessíveis através de uma rede". Um tipo comum de informação presente em bibliotecas digitais são metadados descritivos. Metadados descritivos são informações utilizadas para descrever os objetos do escopo da coleção de uma biblioteca digital.

```
<oai_dc>
     <title>DEByE - Data Extraction By Example</title>
     <date>2002</date>
     <type>Text</type>
     <creator>Alberto H. F. Laender</creator>
     <creator>Berthier A. Ribeiro-Neto</creator>
     <creator>Altigran Soares da Silva</creator>
     <identifier>dke2002meta402001</identifier>
     <language>eng</language>
     <coverage>February 2002</coverage>
</oai_dc>
```

Figura 2.1: Registro de metadados descrevendo um artigo científico

A Figura 2.1 mostra um registro de metadados que descreve um artigo científico da área de Ciência da Computação de acordo com o padrão Dublin Core[1]. Nele identificamos claramente informações tais como o título, o ano de publicação e os autores do artigo descrito. Essas informações correspondem a metadados descritivos. Os metadados[2] mostrados não são o artigo científico em si, mas somente dados que o descrevem. Note que, entre os metadados do registro mostrado, não há uma URL que aponte para o texto do artigo descrito. Ou seja, por razões tais como o seu desconhecimento no momento da catalogação do metadados, tal apontador não está presente na biblioteca digital. É possível também, que embora exista uma URL presente no registro de metadados, ela não aponte para o texto do artigo, tendo-se tornado inválida devido à dinâmica da Web.

Um conjunto de registros tais como o da Figura 2.1 que descrevam uma coleção de objetos (tal como, um conjunto de artigos científicos) é chamado de catálogo de metadados. Por exemplo, o conjunto de registros que descreve todos os artigos científicos da BDBComp – *Biblioteca Digital Brasileira de Computação* (Laender et al., 2004) é o catálogo de metadados dessa biblioteca. Quando um objeto possui um registro correspondente que o descreva no catálogo de uma biblioteca digital, dizemos que o objeto está catalogado nesta biblioteca digital.

Alguns serviços básicos que devem estar implementados sobre a coleção de uma biblioteca digital são indexação, busca e navegação. Um serviço de indexação gera estruturas de dados necessárias para provimento de serviços de busca sobre as informações da biblioteca digital. Essas estruturas são essenciais para que a busca seja eficiente. No serviço de busca, um usuário é capaz de especificar consultas expressando uma necessidade de informação esperando que, ao serem processadas, sejam retornados apontadores para itens de informação que contemplem a necessidade de informação expressa. Em um serviço de navegação, informações estão ligadas umas às outras de maneira a expressar relações que possuam entre elas. Isso possibilita que os usuários naveguem de uma informação para outra, guiados por tais relações.

Neste contexto, uma situação comum no uso de uma biblioteca digital de artigos científicos é o usuário se deparar com um documento que contenha metadados tais como os mostrados na Figura 2.1 e se interessar pelo texto do artigo descrito. Tal situação é uma aplicação direta da abordagem que propomos neste trabalho: de posse dos metadados que descrevem uma publicação, realizar consultas a máquinas de busca e utilizar as respostas retornadas para sugerir apontadores para o respectivo texto completo. Dessa forma, podemos classificar nossa proposta como um serviço de busca específico

[1] http://dublincore.org

[2] Nesta dissertação, sempre que utilizarmos simplesmente o termo metadados, estaremos nos referindo a metadados descritivos de um objeto tal como um artigo.

para uma tarefa, pois ele tira proveito dos metadados catalogados para expressar consultas sobre o vasto conteúdo indexado por máquinas de busca e selecionar entre as respostas retornadas apontadores com grande chance de satisfazer a necessidade de informação do usuário.

Máquinas de busca são serviços onde é possível expressar uma necessidade genérica de informação através de uma consulta para a qual é retornado como resposta um conjunto de itens que apontam para documentos da Web que possam suprir a necessidade de informação expressa pelo usuário. Diferentemente de uma biblioteca digital, as máquinas de busca oferecem serviço de busca em uma coleção de documentos sobre os quais não se tem controle devido à possibilidade de livre publicação na Web. Em uma biblioteca digital, a coleção tem foco específico e o seu conteúdo é totalmente controlado.

A eficácia de uma estratégia de busca está relacionada à relevância dos itens retornados como resposta quando o usuário submete uma consulta a uma máquina de busca. Devido a especificidades da tecnologia das máquinas de busca, diferentes formas de consulta e diferentes máquinas de busca podem apresentar diferentes níveis de eficácias. Tais diferenças são objeto de comparação no processo de busca proposto.

Para avaliar a eficácia de uma estratégia ao retornar um conjunto de respostas para uma consulta, normalmente necessita-se de uma inspeção manual da relevância dos documentos retornados por ela. São utilizadas métricas padrão da área de Recuperação de Informação para medir quantitativamente a eficácia de uma consulta. Similarmente, podemos avaliar a eficácia para um conjunto de consultas. No caso da tarefa em questão, cada consulta está associada a um registro de metadados sem URL de uma coleção. A eficácia da estratégia em relação à coleção como um todo é estimada através da média das medidas individuais associadas a cada registro da coleção. Assim, para se medir quantitativamente a eficácia de uma estratégia de consulta em relação a uma coleção de registros, deve-se determinar por inspeção manual a relevância dos documentos retornados para todos os registros da coleção. Em nossos experimentos, consideramos somente uma amostra dos registros dos catálogos utilizados. A partir dos resultados obtidos, estimamos um intervalo de valores no qual é provável que o valor da eficácia em relação a todos os registros das coleções se enquadre. Tais intervalos são obtidos a partir de cálculos estatísticos e são chamados de intervalos de confiança (Jain, 1991).

No caso específico tratado nesta dissertação, a relevância de uma resposta para um usuário depende de diversos fatores tais como disponibilidade financeira, grau de exigência e necessidade de momento. Há contextos nos quais, ao buscar pelo texto de um artigo, um usuário somente considera relevantes documentos que forneçam tal texto. Em outros, para a mesma busca, um usuário mais flexível também considera relevante

documentos relacionados, de autoria de um dos autores do artigo procurado, versando sobre o mesmo assunto ou similar, tais como uma tese, dissertação ou outro artigo. A forma de disponibilidade também influencia na relevância. Um texto proveniente de uma biblioteca digital com acesso restrito, na qual os usuários pagam pelos documentos disponíveis, pode ser considerado irrelevante para um usuário. Dessa forma, imaginamos diferentes contextos caracterizados por usuários com determinadas características para comparar as estratégias em cada situação. Chamamos cada uma desses contextos de cenários, que correspondem a uma representação de usuários de determinado perfil.

Figura 2.2: Itens de resposta retornados por uma máquina de busca para a consulta *DEByE – Data Extraction By Example*

A Figura 2.2 mostra itens de resposta retornados por uma máquina de busca ao se efetuar uma consulta com o título do artigo descrito no registro de metadados mostrado da Figura 2.1. Cada item de resposta possui um título e uma URL associados os quais estão destacados na figura. A semelhança do título retornado pela máquina de busca com o título do artigo procurado é uma forte evidência para saber se o documento apontado pela respectiva URL é relevante para o usuário. Documentos com título com pouca semelhança, normalmente, são irrelevantes.

Há diversas maneiras de se especificar uma consulta a partir dos metadados presentes no registro de um catálogo. Por exemplo, para o registro da Figura 2.1, podemos formar uma consulta que utilize somente os termos que compõem o título do artigo, derivando *DEByE – Data Extraction By Example* ou, ainda, uma que utilize o título delimitado por aspas seguido pelo sobrenome do primeiro autor catalogado, formando *"DEByE – Data Extraction By Example" Laender*. Chamamos de tipo de consulta a cada uma das formas possíveis de se formular uma consulta a partir dos metadados presentes nos registros de uma biblioteca digital.

## 2.2 Processo Proposto

Idealizamos um serviço com o intuito de ajudar aos usuários a encontrar, na Web, URLs que levem ao texto completo de artigos catalogados em uma biblioteca digital. O processo seguido para prover tal serviço está esquematizado na Figura 2.3. Ao interagir com a *Interface da Biblioteca Digital*, o usuário visualiza informações dos metadados de um artigo $a$ e aciona o serviço para procurar o seu respectivo texto completo. De posse do registro $m_a$ de *Metadados do Artigo*, a *Interface de Consulta* automaticamente gera e submete uma consulta para uma ou mais máquinas de busca requisitando pelo respectivo texto. *URLs Candidatas* são extraídas das páginas resul-

tantes e compiladas em uma lista ordenada $C_a = [c_1, c_2, \ldots, c_n]$. Cada resultado $c_i$ é uma quádrupla $(t_c, u_c, s, t)$ onde $t_c$ é o título e $u_c$ é a URL de um documento da Web retornado pela máquina de busca $s$ para uma consulta de um tipo $t$ formada a partir do registro de metadados $m_a$. Os resultados em $C_a$ seguem uma ordem que prioriza respostas provenientes de máquinas de busca que possuem maior chance de fornecer URLs de textos de artigos científicos e, para respostas provenientes de uma mesma máquina de busca, privilegia respostas que aparecem primeiro na ordenação fornecida pelas páginas de resposta retornadas pela consulta submetida a ela.

Figura 2.3: Arquitetura do Serviço

A seguir, o *Filtro* seleciona de $C_a$ aqueles elementos com maior chance de possuir conteúdo fortemente relacionado ao artigo $a$ procurado, ou seja, ele tenta remover de $C_a$ URLs candidatas que correspondem a documentos de pouco ou nenhum interesse para o usuário, tais como currículos de autores, textos de artigos que citam $a$ mas tratam de assunto pouco relacionado ao artigo $a$, etc. Além disso, a remoção de URLs de $C_a$ reduz o custo de processamento nos passos subseqüentes do processo. Para tal, adotamos um procedimento simples que consiste em remover de $C_a$ elementos que possuem um título $t_c$ pouco similar ao título do artigo $t_a$. Essa estratégia deve simular o comportamento de usuários que somente examinam resultados cujo título $t_c$ é minimamente similar ao título $t_a$ do artigo cujo texto é procurado. Portanto, o *Filtro* gera uma lista $F_a$ de *URLs Filtradas*.

Em nosso processo, consideramos $t_c$ pouco similar a $t_a$ se o Coeficiente de Jaccard (Tan et al., 2005) $J(t_c, t_a)$[3] computado utilizando os respectivos termos é menor que um limiar $j$. O limiar $j$ é ajustado experimentalmente para maximizar a eficácia das respostas retornadas pelo tipo de consulta $t$ na máquina de busca $s$ para um conjunto de registros de amostra, conforme resultados relatados na Seção 4.2 do Capítulo 4. O Coeficiente de Jaccard foi escolhido pelo fato de poder ser computado de maneira mais simples e rápida do que outras medidas de similaridade tais como a distância do cosseno (Salton et al., 1975).

Finalmente, o *Ordenador* ordena a lista $F_a$ e gera uma nova lista $R_a$ de *URLs Ordenadas* possivelmente usando evidências adicionais tais como o título do documento retornado e a fonte da qual ele é proveniente. O *Ordenador* opera procurando colocar, no topo da lista ordenada, aqueles documentos com maior chance de satisfazer às necessidades do usuário. A lista $R_a$ é então retornada para a *Interface da Biblioteca Digital* e mostrada para o usuário interessado.

[3]Dados dois conjuntos de termos $A$ e $B$, o Coeficiente de Jaccard entre eles é dado por $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$.

## 2.3 Pontos de Pesquisa

Dentre os pontos de pesquisa fundamentais para um serviço dessa natureza destacamos:

- De posse dos metadados contidos nos registros do catálogo, descobrir para cada cenário os tipos de consulta mais adequados para cada máquina de busca.
- Remover dos resultados retornados pelas máquinas de busca, URLs que tenham pouca chance de prover conteúdo relacionado ao artigo procurado.
- Avaliar a adequação de diferentes máquinas de busca ao processo de maneira a utilizar as mais eficazes, desconsiderando aquelas cujos resultados não acrescentam ganho significativo;
- Combinar e reordenar as respostas provenientes de diferentes máquinas de busca de maneira a melhorar a qualidade do conjunto de respostas como um todo.
- Ajustar configuração do serviço conforme requisitos de custo de execução das consultas tempo.

No Capítulo 3 apresentamos a metodologia experimental utilizada para abordar cada um desses pontos. O resultados obtidos estão relatados no Capítulo 4.

# Capítulo 3

# Metodologia Experimental

Neste capítulo descrevemos detalhadamente a metodologia experimental seguida para abordar os pontos de pesquisa que enumeramos no Capítulo 2. A metodologia adotada visa obter resultados experimentais que nos guiem para uma boa configuração do processo proposto para a busca do texto completo de artigos catalogados nas coleções BDBComp e DBLP-Br que descrevemos mais adiante. Assim, neste capítulo, a Seção 3.1 mostra como configuramos os experimentos e a Seção 3.2 descreve a base metodológica adotada para avaliação dos resultados. Os resultados experimentais obtidos são relatados e discutidos no Capítulo 4.

## 3.1 Configuração dos Experimentos

Com o objetivo de encontrar a melhor configuração para o serviço proposto, primeiramente investigamos o comportamento individual das máquinas de busca consideradas na tarefa de buscar o texto completo dos artigos. Testamos cinco máquinas de busca populares na Web: Google[1], Yahoo![2], MSN Search[3], Google Scholar[4] e CiteSeer[5]. As três primeiras são máquinas de busca de propósito geral e estão entre as de maior audiência na Web. As duas últimas são máquinas de busca especializadas que indexam publicações científicas, sendo o CiteSeer focado na área de Ciência da Computação.

Figura 3.1: Ambiente Experimental

As máquinas de busca foram utilizadas conforme mostrado na Figura 3.1. Assim, o ambiente experimental aqui descrito é uma versão simplificada da arquitetura do

[1] http://www.google.com/
[2] http://search.yahoo.com/
[3] http://search.msn.com/
[4] http://scholar.google.com/
[5] http://citeseer.ist.psu.edu/

serviço mostrada na Figura 2.3 do Capítulo 2, onde o *Catálogo de Amostra* contém registros de metadados de artigos da área de Ciência da Computação provenientes de catálogos de duas bibliotecas digitais. Tais registros não possuem URL catalogada que aponte para o texto do respectivo artigo. Cada consulta $q_a$, gerada a partir de um registro de metadados $m_a$, é submetida a uma máquina de busca $M$ com o intuito de encontrar URLs relevantes para $m_a$. Para cada máquina de busca, exceto o CiteSeer, desenvolvemos uma *Interface de Consulta* específica que submete as consultas diretamente para o respectivo processador de consultas e extrai de $P_a$, o conjunto de páginas de resposta retornadas, os títulos e as URLs dos documentos recuperados para criar a lista $C_a$ de URLs candidatas. Devido ao fato de o CiteSeer estar constantemente indisponível no período de nossos experimentos, usamos o Google com a restrição de busca sob o domínio do CiteSeer através da especificação do filtro *site:citeseer.ist.psu.edu*. Adotamos tal alternativa com base nos resultados de um experimento no qual selecionamos aleatoriamente 1060 registros do catálogo do CiteSeer[6] e submetemos consultas para o Google requisitando por conteúdo relacionado. Para 98 $\pm$1 % (intervalo de confiança de 95% (Rea e Parker, 1997)) desses registros, foi possível recuperar pelo menos um documento com título similar ao título presente no registro de metadados.

O *Catálogo de Amostra* foi formado a partir da seleção aleatória de 200 registros de metadados com URLs ausentes de duas coleções: a primeira, derivada do catálogo da BDBComp, contém 3969 registros de metadados de artigos publicados em anais de conferências brasileiras relevantes da área de Ciência da Computação, e a segunda, derivada de um subconjunto da DBLP, à qual chamamos de DBLP-Br, com 3181 registros de metadados de artigos publicados, na quase totalidade, por pesquisadores brasileiros em anais de conferências internacionais. A seleção efetuada sobre tais coleções derivou uma composição de 121 registros provenientes da BDBComp e 79 da DBLP-Br para o *Catálogo de Amostra*. Tal conjunto de registros objetiva simular experimentalmente buscas dos textos completos dos artigos dessas coleções efetuadas por usuários dessas bibliotecas digitais. As principais características das coleções estão mostradas na Tabela 3.1. Como podemos ver, os artigos da BDBComp estão distribuídos quase balanceadamente entre os idiomas inglês e português e na coleção DBLP-Br o idioma inglês é predominante. O percentual de artigos da BDBComp sem URL para o texto completo é quase o dobro do percentual da DBLP-Br.

Cada registro de metadados $m_a$ do *Catálogo de Amostra* foi usado para gerar consultas que foram submetidas a cinco máquinas de busca requisitando pelo texto completo dos respectivos artigos. Nesse processo, consideramos sete diferentes tipos de consulta. A sintaxe de cada tipo está descrita e exemplificada a seguir. Os exemplos consideram

[6]Obtido em http://citeseer.ist.psu.edu/oai.html.

Tabela 3.1: Características das coleções

| **Característica** | **BDBComp** | **DBLP-Br** |
|---|---|---|
| # artigos | 3969 | 3181 |
| % artigos sem URL | 65,7% | 35,7% |
| % artigos em português | 48,1% | 2,3% |
| % artigos em inglês | 49,8% | 94,3% |

o artigo *DEByE – Data Extraction By Example* cujos autores, conforme catalogado na DBLP, são Alberto H. F. Laender, Berthier A. Ribeiro-Neto e Altigran S. da Silva.

**AS**: sobrenomes de todos os autores catalogados (exemplo: *Laender Ribeiro-Neto Silva*).

**UT**: título puro (exemplo: *DEByE – Data Extraction By Example*).

**UT+FS**: título puro seguido pelo sobrenome do primeiro autor catalogado (exemplo: *DEByE – Data Extraction By Example Laender*).

**UT+AS**: título puro seguido pelos sobrenomes de todos os autores catalogados (exemplo: *DEByE – Data Extraction By Example Laender Ribeiro-Neto Silva*).

**QT**: título delimitado por aspas (e.g.: *"DEByE – Data Extraction By Example"*).

**QT+FS**: título delimitado por aspas seguido pelo sobrenome do primeiro autor catalogado (exemplo: *"DEByE – Data Extraction By Example" Laender*).

**QT+AS**: título delimitado por aspas seguido pelos sobrenomes de todos os autores catalogados (exemplo: *"DEByE – Data Extraction By Example" Laender Ribeiro-Neto Silva*).

Para cada consulta $q_a$ submetida às máquinas de busca testadas, a *Interface de Consulta* gerou uma lista $C_a$ de URLs candidatas com as $k$ primeiras respostas retornadas. Em nossos experimentos consideramos $k = 40$. A partir dessa lista, o *Filtro* selecionou os elementos cujo título $t_c$ tinha o Coeficiente de Jaccard de similaridade com o título do artigo procurado $t_a$ maior que o limiar[7] de 0,22, compilando a lista $F_a$ de URLs filtradas.

## 3.2 Metodologia de Avaliação

Submetemos as consultas às cinco máquinas de busca no período de 21 de agosto a 6 de setembro de 2006. Para cada registro de metadados de amostra $m_a$ do *Catálogo*

[7] Tal limiar foi escolhido para garantir a viabilidade do experimento, ou seja, manter o número de respostas selecionadas tratável, já que tais respostas foram manualmente classificadas.

*de Amostra*, combinamos as listas $F_a$ em uma única lista de URLs filtradas. Para 172 delas (103 da BDBComp e 69 da DBLP-Br), obtivemos um conjunto não-vazio de URLs. Desse procedimento resultou um total de 3676 pares $(m_a, u)$ onde $m_a$ é o registro de metadados que descreve o artigo $a$ e $u$ é uma URL plausível para o texto (ou documento de conteúdo relacionado) correspondente. Efetuamos então a distribuição desses 3676 pares $(m_a, u)$ entre 27 voluntários (membros de nosso grupo de pesquisa) e pedimos a eles que os enquadrassem em categorias de conteúdo e acessibilidade dos documentos descritas a seguir.

Em relação ao conteúdo dos documentos, consideramos as seguintes categorias:

1. **Texto Completo**: a URL $u$ aponta para o texto completo (ou para um documento que contenha um apontador para ele) do artigo descrito em $m_a$.

2. **Texto Completo Similar**: a URL $u$ aponta para o texto completo (ou para um documento que contenha um apontador para ele) de um documento $d$ similar a $a$ como, por exemplo, um outro artigo relacionado, uma tese ou dissertação de autoria de um dos autores de $a$.

3. **Metadados Úteis**: a URL $u$ não pertence a quaisquer das categorias anteriores e aponta para um documento que contém metadados de $a$ ausentes de $m_a$.

4. **Metadados Similares**: a URL $u$ não pertence a quaisquer das categorias anteriores e aponta para um documento que contém metadados que descrevem um documento $d$ similar a $a$, como, por exemplo, um artigo relacionado, uma tese ou dissertação de autoria de um dos autores de $a$.

5. **Metadados Redundantes**: a URL $u$ não pertence a quaisquer das categorias anteriores e aponta para um documento com metadados já presentes em $m_a$.

6. **Outros**: a URL $u$ não pertence a quaisquer das categorias anteriores.

Sobre acessibilidade, consideramos as seguintes categorias:

a. **Acesso Restrito**: a URL $u$ fornece acesso ao texto completo (ou documento relacionado) por meio de algum tipo de pagamento ou acordo comercial.

b. **Acesso Livre**: a URL $u$ fornece acesso livre ao texto completo (ou documento relacionado).

Tomando como base as categorias acima, derivamos cinco cenários que modelam usuários com diferentes interesses e perfis. Dado um artigo $a$ e uma URL $u$, indicamos a seguir as categorias que se adequam a cada um desses cenários.

**Estrito**: 1. Os usuários deste cenário estão interessados somente no texto completo de $a$, independentemente de como ele possa ser acessado (livremente ou não).

**Estrito & Livre**: 1 e b. Neste cenário, os usuários estão interessados somente no texto completo de $a$, mas somente se ele puder ser livremente acessado.

**Flexível**: 1 ou 2. Neste cenário, os usuários estão interessados no texto completo de $a$ mas também se satisfazem com o texto completo de qualquer documento relacionado, independentemente de como eles possam ser acessados.

**Flexível & Livre**: 1 e b ou 2 e b. Usuários deste cenário possuem as mesmas exigências dos usuários do cenário **Flexível**. Entretanto, o texto completo deve estar livremente acessível.

**Muito Flexível**: 1, 2 ou 3. Adicionalmente ao texto completo de $a$ ou de qualquer documento relacionado, usuários deste cenário também estão interessados em documentos que contêm metadados desconhecidos de $a$, independentemente de como eles possam ser acessados.

Para chegar a uma boa configuração do processo proposto, procedemos conforme a seguir. Primeiramente investigamos, para cada uma das máquinas de busca, os tipos de consulta mais adequados a elas em cada cenário. Conhecidos tais tipos de consulta mais adequados a cada máquina de busca, avaliamos a eficácia dos vários tipos de consulta em cada cenário. Disso resultou uma comparação entre máquinas de busca. Efetuamos testes combinando respostas das máquinas de busca mais adequadas. Tais testes objetivaram melhorar o desempenho individual da máquina de busca mais adequada a cada contexto. O resultados experimentais obtidos são relatados e discutidos no próximo capítulo.

# Capítulo 4

# Resultados Experimentais

Neste capítulo, relatamos e discutimos resultados experimentais essenciais para abordar os pontos de pesquisa levantados no Capítulo 2. Eles estão embasados na metodologia experimental descrita no Capítulo 3. Primeiramente, a Seção 4.1 descreve a métrica utilizada nas comparações experimentais. Na Seção 4.2, analisamos os sete tipos de consulta considerados, avaliando quão eficazes são essas consultas para recuperação do texto completo dos artigos, ao serem submetidas a cada uma das máquinas de busca testadas. Em seguida, na Seção 4.3, efetuamos uma comparação das máquinas de busca testadas considerando, para cada uma, o tipo de consulta de melhor desempenho em cada cenário. Na Seção 4.4, propomos e discutimos formas para melhorar o desempenho da máquina de busca de melhor resultado em cada cenário. A Seção 4.5 analisa o custo envolvido na execução das consultas e discute a relação de tal custo com a eficácia das respostas retornadas. Finalmente, a Seção 4.6 discute as limitações de nossa abordagem mostrando aspectos do processo proposto que podem ser melhorados.

## 4.1 Métrica

A métrica utilizada em nossas comparações é a precisão média nos documentos relevantes (*average precision at seen relevant documents*) (Baeza-Yates e Ribeiro-Neto, 1999). Intuitivamente, ela pondera a satisfação de um usuário ao encontrar documentos relevantes em uma lista ordenada de respostas. É atribuído um maior peso aos documentos relevantes que aparecem nas primeiras posições da lista, premissa razoável sob a hipótese de que os usuários tendem a escolher os primeiros elementos da lista ao inspecionar as respostas retornadas. A seguir definimos formalmente essa métrica.

**Definição 1** *Seja* $D_{ca} = \{d_1, d_2, \ldots, d_n\}$ *um conjunto não vazio de documentos relevantes que satisfazem usuários do cenário c quando procuram por documentos relacionados a um artigo a descrito no registro de metadados* $m_a$*. Seja também* $R_{ts} =$

*$(r_{ts_1}, r_{ts_2}, \ldots, r_{ts_m})$ o vetor de relevância construído a partir de uma lista ordenada de documentos $D_{ts} = [d_{ts_1}, d_{ts_2}, \ldots, d_{ts_m}]$ retornados por uma máquina de busca $s$ ao receber uma consulta de tipo específico $t$ gerada a partir de metadados do registro $m_a$, onde $r_{ts_i} = 1$, se $d_{ts_i} \in D_{ca}$, ou 0 caso contrário. Então:*

a. *$p_{ts_i}$, a precisão do par tipo de consulta-máquina de busca $(t, s)$ no documento $d_{ts_i}$ de $D_{ts}$, é definida como:*

$$p_{ts_i} = \frac{\sum_{j=1}^{i} r_{ts_j}}{i}$$

b. *$P_{ts}$, a precisão média nos documentos relevantes de $(t, s)$, é definida como:*

$$P_{ts} = \frac{\sum_{i=1}^{m} r_{ts_i} \times p_{ts_i}}{n}$$

O conjunto de documentos relevantes $D_{ca}$ para usuários do cenário $c$ é obtido através de uma extensa inspeção manual de todas as respostas retornadas por consultas de todos os tipos de um conjunto $T$ geradas a partir de metadados em $m_a$ e processadas por todas as máquinas de busca de um conjunto $S$. Podemos medir a eficácia de um par $(t, s)$ ao efetuar buscas para todos os registros do catálogo do qual $m_a$ é proveniente por meio da média das precisões médias obtidas para cada registro. Chamamos tal medida de média da precisão média (MAP - *mean average precision*) do par $(t, s)$ em relação à coleção de registros como um todo. Porém, a obtenção da MAP de $(t, s)$ requer a obtenção do conjunto de documentos relevantes para todos os registros. Para amenizar tal inconveniência, em nossa experimentação estimamos um intervalo de valores no qual é provável que o valor real da MAP se enquadre partindo do cálculo da média $\bar{P_{ts}}$ das precisões médias para um subconjunto de registros da coleção. Para tal, utilizamos os registros do *Catálogo de Amostra* descrito no Capítulo 3. Nas próximas seções relatamos resultados que foram obtidos a partir da execução de tal procedimento.

Ao comparar a eficácia das diferentes máquinas de busca também mostramos curvas de Precisão × Revocação (Baeza-Yates e Ribeiro-Neto, 1999). Tais curvas visam evidenciar graficamente as diferenças de eficácia entre as máquinas de busca consideradas.

## 4.2 Análise do Tipo de Consulta

Nesta seção mostramos intervalos de valores para a MAP dos sete tipos de consulta para cada máquina de busca, considerando o cenário **Estrito**. Focamos a análise em tal cenário devido ao fato de seus usuários estarem interessados em encontrar o texto completo de um artigo catalogado em uma biblioteca digital, nosso principal interesse de estudo.

Os intervalos mostrados foram computados com um grau de confiança de 95% conforme descrito em Jain (1991). Tal cálculo dá garantia estatística à afirmação de que um determinado valor de MAP $x$ deve possuir 95% de chance de se enquadrar em um intervalo computado $c \pm e$. Assumimos $c$, o valor central de um intervalo, como o valor representante dele. Esse valor é exatamente a média $\bar{P}_{ts}$ dos valores das precisões médias obtidos para os registros de metadados do *Catálogo de Amostra* por um par $(t, s)$. O limiar de similaridade $j$ do *Filtro* utilizado para cada tipo de consulta foi o que alcançou o mais alto valor de $\bar{P}_{ts}$ para os registros do *Catálogo de Amostra* ao serem submetidas consultas do tipo $t$ à máquina de busca $s$. Para tal, variamos o limiar do *Filtro* entre os valores $0,22$ e $1,0$, escolhendo então $j$ como sendo o valor para o qual se obteve a maior média $\bar{P}_{ts}$. Resultados consideradados significativos estão embasados por testes-$t$ pareados Jain (1991) assumindo nível de confiança também de 95%.

Alertamos, nesta seção, que os intervalos da MAP dos tipos de consulta não são diretamente comparáveis para diferentes máquinas de busca, pois o conjunto de documentos relevantes $D_{ca}$ para um registro $m_a$ foi obtido a partir do conjunto formado pela união das respostas retornadas por uma única máquina de busca ao processar consultas dos sete tipos considerados. Logo, não são válidas comparações relativas aos tipos de consulta para diferentes máquinas de busca. Comparações entre as máquinas de busca são deixadas para a Seção 4.3.

Tabela 4.1: Intervalos da MAP no Google para cada tipo de consulta no cenário **Estrito**

| **Tipo de Consulta** | **Limiar $j$** | **MAP(%)** |
|---|---|---|
| AS | 0,28 | 20,92±5,83 |
| UT | 0,24 | **39,07±6,62** |
| UT+FS | 0,23 | 37,70±6,98 |
| UT+AS | 0,26 | 34,03±7,19 |
| QT | 0,22 | 27,87±6,76 |
| QT+FS | 0,22 | 27,04±6,71 |
| QT+AS | 0,26 | 26,21±6,25 |

Conforme mostrado na Tabela 4.1, as melhores consultas para o Google são aquelas que incluem o título puro (UT, UT+FS, UT+AS). Elas se mostraram significativamente melhores do que as de título delimitado, pois estas últimas tendem a excluir muitos itens relevantes presentes no conjunto final. Isto ocorre devido ao fato de que, em alguns casos, o título presente no registro de metadados de um artigo não coincide com os títulos dos documentos na Web que contém o texto-completo de tal artigo. Detalhes tais como erros de digitação e diferenças em codificação de caracteres especiais prejudicam o desempenho de consultas de título delimitado, que exigem casamento exato to texto da consulta com um trecho contido no documento retornado. Uma situação deste tipo está exemplicada na Figura 2.2 do Capítulo 2. Embora, a primeira

resposta retornada pela máquina de busca seja relevante para a consulta especificada, devido a um erro de digitação, o título da resposta *DEByE - Date extraction by example* não coincide exatamente com o título fornecido na consulta. Os resultados de melhor qualidade foram alcançados pelo tipo de consulta que usa somente o título do artigo (UT), entretanto esse tipo de consulta não apresentou diferença significativa em relação a outros que incluem o título puro (UT+FS, UT+AS). Curiosamente, notamos que consultas que usam somente a lista dos sobrenomes dos autores (AS) não apresentaram desempenho significativamente pior do que as que usam o título delimitado (QT, QT+FS, QT+AS). Em muitos casos, o Google retorna resultados corretos com o simples fornecimento de tal lista.

Tabela 4.2: Intervalos da MAP no Yahoo! para cada tipo de consulta no cenário **Estrito**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,22 | 26,99±8,57 |
| UT | 0,23 | 58,01±9,36 |
| UT+FS | 0,22 | **62,66±9,60** |
| UT+AS | 0,30 | 54,65±10,04 |
| QT | 0,22 | 56,45±9,80 |
| QT+FS | 0,22 | 53,82±9,86 |
| QT+AS | 0,30 | 49,91±10,07 |

Em relação ao Yahoo! (veja Tabela 4.2), o melhor desempenho foi alcançado por consultas do tipo UT+FS, ou seja, consultas que usam o título puro do artigo seguido pelo sobrenome do primeiro autor catalogado. Contrariamente ao observado no Google, não há uma diferença significativa na eficácia das consultas desse tipo quando comparadas às que incluem o título puro (UT, UT+FS e UT+AS) e as respectivas versões que utilizam o título delimitado. Consultas que incluem o título puro mostraram um desempenho levemente superior, entretanto estatisticamente insignificante. Também notamos que, diferentemente do caso do Google, consultas que incluem somente a lista de sobrenomes dos autores (AS) não retornaram bons resultados no Yahoo!.

Tabela 4.3: Intervalos da MAP no MSN para cada tipo de consulta no cenário **Estrito**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,22 | 30,43±19,27 |
| UT | 0,27 | 24,49±18,18 |
| UT+FS | 0,22 | 55,43±20,08 |
| UT+AS | 0,22 | 64,13±19,24 |
| QT | 0,22 | **76,09±17,09** |
| QT+FS | 0,22 | **76,09±17,09** |
| QT+AS | 0,22 | 66,30±19,45 |

O número de artigos para os quais houve recuperação de alguma resposta relevante ao consultar o MSN (aproximadamente 10% do total) não foi suficiente para fazer afirmações conclusivas sobre a eficácia dos tipos de consulta considerados na tarefa em questão. Os resultados obtidos estão mostrados na Tabela 4.3. Aparentemente, consultas que incluem o título delimitado tendem a trazer melhores resultados do que as que usam o título puro, entretanto não podemos afirmar isso conclusivamente. Também deve ser notado nessa máquina de busca a reduzida eficácia das consultas que utilizam somente a lista de sobrenomes dos autores (AS) e daquelas que utilizam o título puro (UT) quando comparadas a outros tipos de consulta.

Tabela 4.4: Intervalos da MAP no Scholar para cada tipo de consulta no cenário **Estrito**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,32 | 30,80±8,50 |
| UT | 0,37 | 55,02±9,19 |
| UT+FS | 0,36 | **71,36±7,51** |
| UT+AS | 0,36 | 66,08±8,12 |
| QT | 0,30 | 66,05±8,28 |
| QT+FS | 0,30 | 65,84±8,25 |
| QT+AS | 0,30 | 63,04±8,52 |

Para o Scholar (veja Tabela 4.4), as consultas mais eficazes foram as que incluíram o título delimitado seguido pelo sobrenome do primeiro autor catalogado (UT+FS). Entre as consultas com título delimitado, as que não incluem informação adicional (QT) alcançaram o melhor desempenho. Consultas que incluem somente a lista de sobrenomes dos autores (AS) são significativamente piores do que as demais. Aparentemente, esse tipo de consulta não é capaz de recuperar bons resultados nessa máquina de busca quando comparadas com aquelas que incluem o título do artigo. Elas também tendem a recuperar documentos relacionados, o que, no caso do cenário **Estrito**, não é considerado relevante.

Tabela 4.5: Intervalos da MAP no CiteSeer para cada tipo de consulta no cenário **Estrito**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,42 | 60,74±14,76 |
| UT | 0,24 | 65,96±13,29 |
| UT+FS | 0,23 | **67,08±13,48** |
| UT+AS | 0,29 | 58,67±15,19 |
| QT | 0,22 | 53,51±15,56 |
| QT+FS | 0,22 | 44,58±15,92 4 |
| QT+AS | 0,22 | 50,22±16,17 |

Como mostrado na Tabela 4.5, para o CiteSeer as consultas mais eficazes foram aquelas que incluíram o título puro dos artigos e o sobrenome do primeiro autor catalogado (UT+FS). As consultas menos eficazes foram as que incluíram o título delimitado (QT, QT+FS e QT+AS). Entretanto, relembramos que consultas ao CiteSeer foram submetidas ao Google restringindo a busca ao domínio do CiteSeer, o que pode explicar a menor eficácia das consultas de título delimitado que não são as mais eficazes no Google. Isso também pode ter sido a razão para o fato de consultas incluindo somente a lista de sobrenomes de todos os autores (AS) não terem apresentado desempenho significativamente pior do que consultas com o título do artigo. Esse tipo de consulta, quando submetida ao Google, muitas vezes retorna resultados corretos.

Analisando os resultados como um todo, podemos notar que, no caso geral, consultas do tipo UT+FS parecem ser a melhor opção entre as sete alternativas experimentadas na tarefa. Apesar de que intuitivamente fosse esperado melhores resultados para consultas com o título delimitado, parece que tais consultas tendem a excluir muitos resultados relevantes.

Resultados relativos aos demais cenários estão relatados no Apêndice A. Deixamos para o leitor a interpretação dos dados. Adiantamos que, embora se trate de cenários diferentes, os resultados são parecidos com os que apresentamos nesta seção.

## 4.3 Comparação entre as Máquinas de Busca

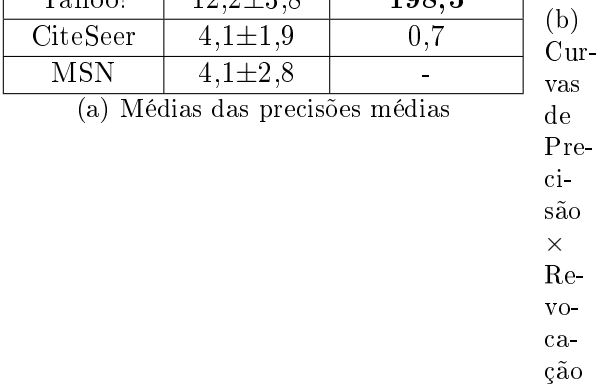
A comparação entre as máquinas de busca foi realizada através da seleção para cada cenário do tipo de consulta mais eficaz em cada uma delas. Por exemplo, no cenário **Estrito**, os tipos de consulta selecionados foram aqueles cujos valores da MAP nas tabelas da Seção 4.2 estão em negrito. Para o caso do MSN, em que os tipos de consulta QT e QT+FS apresentaram o mesmo intervalo para a MAP, selecionamos o tipo QT+FS, pois inclui o título delimitado QT.

O conjunto de documentos relevantes para um registro de metadados em um cenário específico foi obtido a partir da seleção entre todas as respostas retornadas, ou seja, a união dos resultados das consultas de todos os tipos considerados submetidas às cinco máquinas de buscas testadas. A análise dos resultados é mostrada a seguir. Deve ser notado que valores da MAP são comparáveis somente em um cenário específico, ou seja, não podemos comparar um valor de MAP de uma máquina de busca para um cenário com valores para quaisquer outros cenários.

A Figura 4.1 compara a eficácia das máquinas de busca para o cenário **Estrito**. A Tabela 4.1(a) mostra os respectivos valores de MAP para cada máquina de busca desse cenário. Resultados são mostrados dos maiores para os menores valores centrais

| **Máquina** | **MAP(%)** | **Ganho (%)** |
|---|---|---|
| Scholar | 30,3±5,8 | **76,5** |
| Google | 17,1±3,6 | 40,6 |
| Yahoo! | 12,2±3,8 | **198,5** |
| CiteSeer | 4,1±1,9 | 0,7 |
| MSN | 4,1±2,8 | - |

(a) Médias das precisões médias




(b) Curvas de Precisão × Revocação

Figura 4.1: Comparação das máquinas de busca no cenário **Estrito**

dos intervalos. Para cada máquina de busca $r$, mostramos também o ganho ($100 \times \frac{\bar{P}_r - \bar{P}_s}{\bar{P}_s}$) sobre o valor central da MAP da máquina de busca $s$ imediatamente abaixo. Valores em negrito mostraram ser ganhos estatisticamente significantes em testes-$t$ pareados (Jain, 1991). Para o cenário **Estrito**, a superioridade do Scholar sobre todas as outras máquinas de busca é evidente. Um ganho significativo acima de 76% é apresentado sobre o segundo lugar, o Google. O Google, por sua vez, supera o Yahoo!, mas o ganho não é significativo. Apesar da superioridade do Google em relação ao valor central do intervalo da MAP, houve alguns casos nos quais o Yahoo! superou o Google. MSN e CiteSeer não mostraram ser uma boa alternativa em tal cenário. O ganho significativo do Yahoo! sobre o CiteSeer alcançou 198%. A Figura 4.1(b) mostra curvas de Precisão × Revocação retratando graficamente as diferenças de eficácia entre as máquinas de busca mostradas na Tabela 4.1(a) neste cenário. Note que todos os comentários de comparação entre as eficácias das máquinas de busca considerando os dados da Tabela 4.1(a) são confirmados pelas curvas da Figura 4.1(b).

Como pode ser visto na Figura 4.2, o Scholar também obteve o melhor desempenho para o cenário **Estrito & Livre**. A vantagem do Google sobre o Yahoo! não é significativa, sendo que as duas máquinas de busca apresentam desempenho muito parecido. A diminuição do ganho do Google sobre o Yahoo!, quando comparada ao desempenho mostrado no cenário **Estrito**, pode ser atribuída à grande quantidade de conteúdo restrito indexado pelo Google, irrelevante neste cenário. Novamente, CiteSeer e MSN mostraram um baixo desempenho.

No cenário **Flexível**, onde os usuários também consideram relevantes textos completos de documentos relacionados, o Scholar mostrou-se mais efetivo novamente, mas com um ganho menor sobre o segundo lugar, Google (veja Figura 4.3). Por outro lado,

| Máquina | MAP(%) | Ganho (%) |
|---|---|---|
| Scholar | 29,5±7,0 | **82,9** |
| Google | 16,1±4,9 | 6,7 |
| Yahoo! | 15,1±5,0 | **144,1** |
| CiteSeer | 6,2±2,6 | 31,5 |
| MSN | 4,7±3,3 | - |

(a) Médias das precisões médias

(b) Curvas de Precisão × Revocação

Figura 4.2: Comparação das máquinas de busca no cenário **Estrito & Livre**

| Máquina | MAP(%) | Ganho (%) |
|---|---|---|
| Scholar | 28,4±4,9 | **64,5** |
| Google | 17,3±3,6 | **58,9** |
| Yahoo! | 10,9±3,5 | **104,6** |
| CiteSeer | 5,3±2,6 | 60,7 |
| MSN | 3,3±2,4 | - |

(a) Médias das precisões médias

(b) Curvas de Precisão × Revocação

Figura 4.3: Comparação das máquinas de busca no cenário **Flexível**

Google obteve um aumento significativo da sua vantagem sobre o Yahoo!. CiteSeer e MSN, novamente, repetiram um desempenho bem inferior às demais.

No cenário **Flexível & Livre** (veja Figura 4.4), mesmo considerando conteúdo restrito como não relevante para os usuários desse cenário, o Scholar ainda aparece na primeira posição. Nesse cenário, o Google parece ser tão eficaz quanto o Yahoo!. Esse fato reforça a observação de que a vantagem do Google sobre o Yahoo! somente aparece quando conteúdo restrito é considerado relevante. CiteSeer e MSN, novamente, não se mostraram como boas alternativas.

Finalmente, no cenário **Muito Flexível** (veja Figura 4.5), a vantagem do Scholar sobre o segundo lugar, o Google, é perto de 50%. A vantagem do Google sobre o Yahoo! é bastante significativa. O Yahoo!, por sua vez, supera o CiteSeer por uma

| **Máquina** | **MAP(%)** | **Ganho (%)** |
|---|---|---|
| Scholar | 22,6±5,0 | **53,5** |
| Google | 14,7±3,8 | 8,1 |
| Yahoo! | 13,6±4,3 | 59,1 |
| CiteSeer | 8,6±3,6 | 113,6 |
| MSN | 4,0±2,7 | - |

(a) Médias das precisões médias

(b) Curvas de Precisão × Revocação

Figura 4.4: Comparação das máquinas de busca no cenário **Flexível & Livre**

| **Máquina** | **MAP(%)** | **Ganho (%)** |
|---|---|---|
| Scholar | 29,4±5,0 | **49,0** |
| Google | 19,7±4,0 | **88,8** |
| Yahoo! | 10,4±3,3 | 37,1 |
| CiteSeer | 7,6±3,0 | **122,0** |
| MSN | 3,4±2,3 | - |

(a) Médias das precisões médias

(b) Curvas de Precisão × Revocação

Figura 4.5: Comparação das máquinas de busca no cenário **Muito Flexível**

vantagem próxima de 37%. O fato de considerar documentos contendo metadados desconhecidos também como relevantes ajuda o CiteSeer a diminuir sua desvantagem. MSN novamente obteve o pior desempenho.

Em suma, podemos concluir que o Scholar é a melhor alternativa para a tarefa de encontrar o texto completo, ou documentos relacionados, correspondentes a artigos catalogados em bibliotecas digitais, em todos os cenários. O Google pode ser considerado como uma segunda alternativa, mas com desempenho equivalente ao Yahoo! nos cenários onde os usuários consideram relevantes somente documentos livremente acessíveis. CiteSeer e MSN não se mostraram como boas alternativas nessa tarefa em particular.

## 4.4 Melhorias

Nossos experimentos mostraram que o Scholar é a melhor opção para se encontrar URLs ausentes de artigos catalogados em uma biblioteca digital. Entretanto, dados os resultados moderados obtidos (valores de MAP entre 17,6 e 36,5%), levanta-se uma questão natural: "Como podemos melhorar os resultados alcançados pelo Scholar?" Combinar respostas de diferentes máquinas de busca poderia ser uma alternativa. De fato, observações sobre as respostas retornadas por diferentes máquinas de busca evidenciaram o que a literatura (Bharat e Broder, 1998) já endossa: diferentes máquinas de busca possuem diferentes coberturas da Web. Mesmo para máquinas de buscas mantidas por uma mesma corporação, tais como o Google e o Scholar, não podemos afirmar que o conjunto de URLs cobertas por uma seja subconjunto das URLs cobertas pela outra. Sob essa premissa, testamos estratégias para combinar respostas retornadas[1] pelo Scholar e pelo Google. Descrevemos na Seção 4.4.1 duas propostas de combinação que efetivamente melhoram a qualidade obtida pelo Scholar separadamente em todos os cenários analisados.

Outra fonte de melhoria em potencial é o procedimento de ordenação a ser utilizado para obtenção do conjunto de respostas final. A Seção 4.4.2 especifica um método baseado em prioridades capaz de melhorar significativamente os resultados das combinações descritas na Seção 4.4.1.

### 4.4.1 Estratégias para Combinação de Respostas

Testamos duas estratégias para combinação das respostas provenientes das máquinas de busca Scholar e Google. Para a primeira, à qual nos referiremos como CSG', são feitas consultas às duas máquinas de busca para todo o registro de metadados. Uma primeira consulta é submetida ao Scholar e uma segunda ao Google, removendo da segunda lista de respostas URLs que já aparecem na lista do Scholar. Portanto, a lista final de resultados candidatos é formada pelos elementos do conjunto de respostas do Scholar seguidos dos elementos da lista de respostas do Google. A outra estratégia, referida como CSG", é similar à primeira, entretanto a segunda consulta ao Google somente é efetuada no caso de não haver respostas selecionadas entre os candidatos retornados pelo Scholar. Assim, se a primeira consulta retorna resultados, estes serão considerados como o conjunto de resultados candidatos. Caso contrário, são consideradas as respostas retornadas pela segunda consulta.

[1] Testamos também combinações com outras máquinas de busca seguindo os mesmos procedimentos. Entretanto, não foram obtidas melhoras significativas.

Como pode ser visto na quarta e sexta colunas (**G(%)**) da Tabela 4.6, os ganhos de CSG' e CSG", respectivamente, sobre o Scholar foram significativos em todos os cenários analisados para as duas propostas de combinação. Esses ganhos podem ser atribuídos à existência de URLs relevantes cobertas pelas consultas ao Google não cobertas pelas consultas efetuadas ao Scholar. CSG' e CSG" alcançam maior melhoria em eficácia no cenário **Muito Flexível** (aproximadamente 55% e 23%, respectivamente), contexto no qual tal situação ocorre com maior freqüência. Notamos ainda que os ganhos apresentados por CSG' são bem superiores aos de CSG". Tal superioridade se deve ao fato de a lista de respostas formada por CSG', na maioria dos casos, fornecer mais URLs relevantes do que a lista de CSG".

Tabela 4.6: MAPs e comparação do Scholar com as combinações CSG' e CSG" em cada cenário

| | **Scholar** | **CSG'** | | **CSG"** | |
|---|---|---|---|---|---|
| **Cenário** | **MAP(%)** | **MAP(%)** | **G(%)** | **MAP(%)** | **G(%)** |
| **Estrito** | 30,3±5,8 | 42,2±5,7 | **39,5** | 33,3±5,6 | **10,1** |
| **Estrito & Livre** | 29,5±7,0 | 38,8±7,1 | **31,6** | 32,9±7,1 | **11,5** |
| **Flexível** | 28,4±4,9 | 41,9±5,3 | **47,4** | 33,1±5,0 | **16,4** |
| **Flexível & Livre** | 22,6±5,0 | 33,8±5,8 | **49,7** | 27,3±5,4 | **20,6** |
| **Muito Flexível** | 29,4±5,0 | 45,5±5,3 | **55,0** | 36,1±5,3 | **23,0** |

### 4.4.2 Estratégias para Reordenação de Respostas

Testamos um método baseado em prioridades para ser utilizado como procedimento de reordenação dos conjuntos de respostas de CSG' e CSG". Dado um artigo com um título $t_a$ e uma lista ordenada $F_a = [f_1, f_2, \ldots, f_n]$, onde $f_i = (t_i, u_i, s_i)$ é uma tripla formada pelo título $t_i$, pela URL $u_i$ do documento retornado e pelo identificador $s_i$ da máquina de busca da qual é proveniente tal resposta, procedemos da seguinte maneira para gerar uma nova lista reordenada $R_a$. Dados dois resultados $f_i, f_j \in F_a$, as condições para $f_i$ preceder $f_j$ na nova lista $R_a$ são, na seguinte ordem de prioridade:

1. $conf(s_i) > conf(s_j)$;
2. $sim(t_a, t_i) > sim(t_a, t_j)$;
3. $inf(u_i) > inf(u_j)$;
4. $j > i$.

A ordem de condiçoes especificada acima foi a que obteve os melhores resultados. Descrevemos a seguir o significado e a intuição de cada delas.

A função $conf(s)$ retorna o grau de confiança que um usuário deposita em uma máquina de busca identificada por $s$ ao procurar por textos completos de artigos científicos. A intuição é que, ao priorizar uma URL indexada por máquinas de busca com maior confiança, há uma menor chance de listar conteúdo irrelevante para os usuários nas primeiras posições da lista exibida. Nos resultados apresentados nesta seção assumimos que o Scholar possui maior confiança em tal tarefa do que o Google. Um procedimento razoável para distigüir máquinas de busca com respeito a confiança é efetuar algumas poucas consultas procurando textos de artigos científicos e estimar a proporção das respostas retornadas que efetivamente trazem apontadores para textos completos. A máquina de busca com maior proporção é a de maior confiança. Tal procedimento pode ser efetuado manualmente quando o número de máquinas de busca envolvidas é reduzido.

Para os casos de respostas com o mesmo grau de confiança (por exemplo, provenientes de uma mesma máquina de busca), a função da segunda condição $sim : string \times string \rightarrow [0,1]$ é distingüir as respostas a partir da similaridade entre duas cadeias de caracteres, neste caso, o título do artigo e o título da resposta. Seguimos a intuição de que quanto maior a similaridade entre tais títulos, maior a chance de a respectiva URL apontar para um documento muito relacionado a $a$. Há muitas funções para medir similaridade entre cadeias de caracteres. Os resultados mostrados baseiam-se no valor do **cosseno** no qual os pesos para os termos, considerados como fontes de evidência no processo proposto, foram calculados utilizando o tradicional esquema TF-IDF (Salton et al., 1975) a partir dos títulos presentes nos registros de todos os artigos das coleções usadas.

A função $inf : url \rightarrow [0,1]$ é uma medida que calcula o grau de infreqüência para uma URL $u$ baseando-se no respectivo domínio[2] $d_u$. Nossa hipótese é a de que quanto mais alto o grau de infreqüência de uma fonte de textos completos, mais alta a probabilidade de se ter acesso livre ao texto completo em $d_u$. A função de infreqüência pode ser calculada conforme especificado a seguir.

**Definição 2** *Seja $M = \{m_1, m_2, \ldots, m_n\}$ uma amostra de registros de metadados de artigos onde, para $j \neq k$, $m_j$, $m_k$ descrevem um par de artigos distintos $a_j$, $a_k$, respectivamente. Suponha que os resultados de uma consulta $q$ gerada a partir dos metadados do registro $m \in M$ componham uma lista ordenada $F_m = (f_1, f_2, \ldots, f_l)$, onde $f_i = (t_i, u_i)$ é um par formado pelo título $t_i$ e a URL $u_i$ de um documento retornado. Seja $d_u$ uma função que retorna o domínio de uma URL $u$. Finalmente, seja $i(m, d)$ uma*

[2]Neste trabalho, uma cadeia de caracteres tipicamente no formato de três níveis "servidor.organização.tipo", usada para identificar um computador, uma organização, outra entidade na Internet.

*função que recebe um registro de metadados m e um domínio d, e é definida como 0, se existe um par* $(t, u) \in F_m$ *tal que* $d_u = d$*, ou 1 caso contrário. Assim:*

*$inf(u)$, o grau de infreqüência de uma URL u, é definido como*

$$inf(u) = \frac{\sum_{m \in M} i(m, d_u)}{n}$$

Note que, dada uma amostra de registros de metadados $M$, o grau de infreqüência de uma URL $u$ depende somente de seu domínio. Portanto, podemos pré-calcular e armazenar as infreqüências dos domínios das URLs retornadas ao efetuar consultas com os registros de metadados do conjunto pré-selecionado $M$. As infreqüências são consideradas como fonte de evidência adicional no processo proposto.

Em nossos experimentos, computamos o grau de infreqüência utilizando uma amostra de 200 registros de metadados sem URL para o texto completo selecionados aleatoriamente das coleções de teste. Simulamos a submissão de consultas dos tipos mais eficazes ao Scholar e ao Google no cenário **Estrito** para gerar os graus de infreqüência dos domínios das URLs retornadas para as respectivas combinações CSG' e CSG".

Em suma, nossa hipótese é que, quando dois títulos de respostas com o mesmo grau de confiança são igualmente similares a $t_a$, o que vier do domínio mais infreqüente, como a página pessoal de um pesquisador ou de um grupo de pesquisa, possui maior chance de fornecer acesso livre ao texto completo de artigos do que aquele proveniente de um domínio altamente freqüente, como a biblioteca digital de uma editora. O último critério para desempate é a posição dos documentos na lista de entrada $F_a$ considerando a ordem original gerada pela máquina de busca da qual as respostas são provenientes.

Tabela 4.7: MAPs e comparação das combinações CSG' e CSG" com as respectivas versões com respostas reordenadas RCSG' e RCSG"

| | **CSG'** | **RCSG'** | | **CSG"** | **RCSG"** | |
|---|---|---|---|---|---|---|
| **Cenário** | **MAP(%)** | **MAP(%)** | **G(%)** | **MAP(%)** | **MAP(%)** | **G(%)** |
| **Estrito** | 42,2±5,7 | 43,9±5,8 | 4,1 | 33,3±5,6 | 35,6±5,9 | **7,0** |
| **Estrito & Livre** | 38,8±7,1 | 42,7±7,3 | **10,0** | 32,9±7,1 | 37,0±7,5 | **12,7** |
| **Flexível** | 41,9±5,3 | 42,0±5,4 | 0,4 | 33,1±5,0 | 33,0±5,1 | -0,003 |
| **Flexível & Livre** | 33,8±5,8 | 36,1±6,0 | **6,7** | 27,3±5,4 | 29,2±5,6 | **7,3** |
| **Muito Flexível** | 45,5±5,3 | 45,0±5,2 | -1,1 | 36,1±5,3 | 35,9±5,2 | -0,6 |

A Tabela 4.7 mostra resultados obtidos pelo procedimento de reordenação quando aplicado aos conjuntos de respostas das combinações CSG' e CSG". Resultados na coluna RCSG' dizem respeito à aplicação do método à combinação CSG' e os da coluna RCSG" à combinação CSG". O procedimento é capaz de melhorar significativamente os resultados nos cenários **Estrito & Livre** e **Flexível & Livre** para ambas as propostas de combinação de respostas testadas. A quarta e sétima colunas (**G(%)**) mostram para

cada cenário os ganhos do procedimento de reordenação em relação à eficácia alcançada ao se assumir a ordem original dos conjuntos de respostas das combinações CSG' e CSG", respectivamente. As perdas apresentadas em alguns cenários são irrizórias.

Tabela 4.8: MAPs e comparação da ordenação original do Scholar e do Google com a ordenação obtida nestas máquinas de busca ao aplicar o procedimento de reordenação para cada cenário

| | **Scholar** | **RScholar** | | **Google** | **RGoogle** | |
|---|---|---|---|---|---|---|
| **Cenário** | **MAP(%)** | **MAP(%)** | **G(%)** | **MAP(%)** | **MAP(%)** | **G(%)** |
| **Estrito** | 30,3±5,8 | 32,9±6,2 | **8,7** | 17,1±3,6 | 16,5±3,2 | -3,5 |
| **Estrito & Livre** | 29,5±7,0 | 33,8±7,4 | **14,7** | 16,1±4,9 | 14,4±4,3 | -10,6 |
| **Flexível** | 28,4±4,9 | 28,7±5,0 | 1,0 | 17,3±3,6 | 17,2±3,6 | -0,6 |
| **Flexível & Livre** | 22,6±5,0 | 25,1±5,4 | **11,2** | 14,7±3,8 | 14,8±3,8 | 0,7 |
| **Muito Flexível** | 29,4±5,0 | 29,7±5,0 | 1,1 | 19,7±4,0 | 17,9±3,6 | **-9,1** |

Com o intuito de entender melhor os ganhos apresentados em alguns cenários pelo procedimento de reordenação, aplicamos tal estratégia ao Scholar e ao Google separadamente. Mostramos na Tabela 4.8 os resultados obtidos (colunas RScholar e RGoogle) comparando-os com os valores alcançados pela ordenação original retornada por essas máquinas de busca. Notamos que a reordenação se adequa melhor ao Scholar e que, quando aplicada ao Google, compromete os resultados em alguns cenários.

Os ganhos no Scholar são devidos ao fato de que, ao receber consultas dos tipos considerados, são privilegiadas URLs de textos de artigos com maior número de citações do que o artigo procurado, segundo estimado por esta máquina de busca. Tal prioridade é inadequada quando consideramos uma consulta formada por metadados de um artigo na qual o usuário claramente explicita o interesse pelo respectivo texto. Um contexto mais adequado para tal ordenação é quando a consulta é formada por termos genéricos que expressam interesse do usuário em determinado assunto. Nestes casos, de fato, faz sentido priorizar os artigos mais citados e, conseqüentemente, mais relevantes sobre o assunto de interesse. Além disso, há preferência por URLs provenientes de bibliotecas digitais de editoras quando o mesmo texto pode ser obtido de forma livre. Tal prioridade é razoável se consideramos que o conteúdo proveniente dessas bibliotecas digitais é mais confiável. Entretanto, tal preferência não faz sentido quando conteúdo com acesso restrito é considerado irrelevante.

Por outro lado, as respostas do Google já aparecem razoavelmente ordenadas, de maneira que a introdução do procedimento prejudica a eficácia consideravelmente em alguns casos. Isso mostra que o procedimento ainda pode ser mais estudado a fim de ser ajustado para se adequar melhor a casos como este.

Estas observações sugerem que é melhor aplicar o procedimento de reordenação somente às respostas do Scholar do que aplicá-lo a todo o conjunto de respostas das

combinações CSG' e CSG". Assim, efetuamos um acerto no procedimento de reordenação de maneira que somente o conjunto de respostas retornado pelo Scholar é reordenado, permanecendo as respostas do Google na ordem original. Mostramos os resultados na Tabela 4.9. Note que há um pequeno ganho quando comparamos com os resultados apresentados na Tabela 4.7.

Tabela 4.9: MAPs e comparação das combinações CSG' e CSG" com as respectivas versões com respostas reordenadas RCSG' e RCSG" em que somente resultados do Scholar são reordenados

| | **CSG'** | **RCSG'** | | **CSG"** | **RCSG"** | |
|---|---|---|---|---|---|---|
| **Cenário** | **MAP(%)** | **MAP(%)** | **G(%)** | **MAP(%)** | **MAP(%)** | **G(%)** |
| **Estrito** | 42,2±5,7 | 44,8±5,9 | **6,2** | 33,3±5,6 | 35,9±6,0 | **7,9** |
| **Estrito & Livre** | 38,8±7,1 | 43,1±7,3 | **11,2** | 32,9±7,1 | 37,2±7,5 | **13,2** |
| **Flexível** | 41,9±5,3 | 42,3±5,3 | 1,1 | 33,1±5,0 | 33,5±5,1 | 1,3 |
| **Flexível & Livre** | 33,8±5,8 | 36,4±5,9 | **7,7** | 27,3±5,4 | 29,9±5,6 | **9,6** |
| **Muito Flexível** | 45,5±5,3 | 46,0±5,4 | 1,0 | 36,1±5,3 | 36,6±5,4 | 1,3 |

Há uma parte considerável dos registros para os quais são retornadas somente URLs de acesso livre ou somente URLs de acesso restrito ao texto completo dos artigos procurados. Nestes casos, não há sentido aplicar um procedimento de reordenação que privilegie URLs de acesso livre. Na Tabela 4.10, mostramos resultados de experimentos para os cenários **Estrito & Livre** e **Flexível & Livre** nos quais aplicamos o procedimento de reordenação somente para os casos em que são fornecidas URLs de ambos os tipos de acesso, restrito e livre, para um dado registro. Conforme pode ser visto, os ganhos nessas situações são maiores do que os ganhos obtidos considerando toda a amostra.

Tabela 4.10: MAPs e comparação do impacto do procedimento de reordenação sobre as respostas das combinações CSG' e CSG" para os casos de acesso restrito e livre

| | **CSG'** | **RCSG'** | | **CSG"** | **RCSG"** | |
|---|---|---|---|---|---|---|
| **Cenário** | **MAP(%)** | **MAP(%)** | **G(%)** | **MAP(%)** | **MAP(%)** | **G(%)** |
| **Estrito & Livre** | 45,7±10,6 | 56,6±10,7 | **24,0** | 40,8±10,7 | 51,8±11,3 | **26,9** |
| **Flexível & Livre** | 39,8±7,1 | 47,1±7,4 | **18,4** | 33,9±7,4 | 41,2±8,0 | **21,6** |

Da mesma forma, mostramos na Tabela 4.11 resultados para os mesmos casos e cenários considerados na Tabela 4.10, aplicando o procedimento de reordenação sem levar em conta o grau de infreqüência das URLs. Os ganhos são bem reduzidos e não significativos, demonstrando que, nos casos em que são fornecidas URLs de ambos os tipos de acesso, o grau de infreqüência faz grande diferença na melhoria dos resultados. Isto reforça o fato de que o método de reordenação testado é capaz de dar maior prioridade a um documento de livre acesso quando um documento similar de acesso restrito também está presente na lista de resultados proveniente do Scholar.

Tabela 4.11: MAPs e comparação do impacto do procedimento de reordenação sobre as respostas das combinações CSG' e CSG" sem considerar o grau de infreqüência das URLs para os casos de acesso restrito e livre

| | **CSG'** | **RCSG'** | | **CSG"** | **RCSG"** | |
|---|---|---|---|---|---|---|
| **Cenário** | **MAP(%)** | **MAP(%)** | **G(%)** | **MAP(%)** | **MAP(%)** | **G(%)** |
| **Estrito & Livre** | 45,7±10,6 | 47,9±10,2 | 4,8 | 40,8±10,7 | 43,0±10,5 | 5,4 |
| **Flexível & Livre** | 39,8±7,1 | 40,5±6,7 | 1,8 | 33,9±7,4 | 34,6±7,1 | 2,1 |

## 4.5 Análise do Custo das Consultas

Conforme pode ser visto na arquitetura idealizada para o serviço mostrada na Figura 2.3 do Capítulo 2, o tempo gasto para efetuar uma consulta está, em grande parte, relacionado ao número de resultados extraídos das páginas de respostas de cada uma das máquinas de busca consultadas. Dada uma lista de URLs candidatas $C_a$ com $n$ elementos, mostramos a seguir uma descrição sucinta das operações executadas em cada etapa e a ordem de complexidade de cada delas até a exibição ao usuário da lista resultados final $R_a$. Para simplificação da análise assumimos que não há intercessão entre os conjuntos de respostas de diferentes máquinas de busca e que toda a lista de respostas das combinações é submetida ao procedimento de reordenação. Essa análise revela os seguintes custos:

**Extração de resultados**: Das páginas de respostas retornadas são extraídos na *Interface de Consulta* os resultados para geração da lista de candidatos $C_a$. Portanto, esta etapa tem custo $O(n)$.

**Filtragem**: No *Filtro*, as respostas presentes em $C_a$ são verificadas individualmente a fim de excluir resultados pouco interessantes. Portanto, a geração de uma nova lista de elementos filtrados $F_a$ com $m$ elementos ($m \leq n$) possui um custo $O(n)$ .

**Reordenação**: Considerando o método proposto para reordenação, no *Ordenador*, para cada um dos $m$ elementos $(t_f, u_f, s_f)$ de $F_a$, são obtidos o grau de confiança da máquina de busca $s_f$, os pesos dos termos presentes no título $t_f$ para cálculo de sua similaridade com o título do artigo procurado $t_a$ e o respectivo grau de infreqüência da URL $u_f$. Com base nessas evidências e na ordem original das respostas, é feita a ordenação dos elementos de $F_a$, gerando uma nova lista ordenada $R_a$. Assim, o passo para obtenção dos valores tem custo $O(m)$ e a ordenação final pode ser computada com custo $O(mlog(m))$. Logo, como $m \leq n$, o custo dominante desta etapa $O(nlog(n))$.

**Exibição**: Na *Interface da Biblioteca Digital*, cada resultado presente em $R_a$ é exibido ao usuário. Logo, o custo desta etapa é $O(m)$.

Assim, o custo total de processamento de uma consulta é dominado por $O(nlog(n))$ onde $n$ é o número de elementos extraídos das páginas de resposta das máquinas de busca. Assim, $n$, o número de elementos da lista de candidatos, tem influência direta no tempo de processamento de uma consulta.

Nas propostas de combinação RCSG' e RCSG", ao extrair no máximo os 40 primeiros resultados presentes nas páginas de resposta do Scholar e do Google, limitamos $n$ a ser inferior ou igual à 80. No entanto, na média bem menos respostas são extraídas.

Tabela 4.12: Número médio de respostas extraídas para as combinações RCSG' e RCSG"

| **Cenário** | $\bar{n}$ (RCSG") | $\bar{n}$ (RCSG') | **O(%)** |
|---|---|---|---|
| **Estrito** | 17,2 | 47,7 | 177,0 |
| **Estrito & Livre** | 15,5 | 41,0 | 164,5 |
| **Flexível** | 17,2 | 47,7 | 177,0 |
| **Flexível & Livre** | 17,2 | 47,7 | 177,0 |
| **Muito Flexível** | 15,5 | 41,0 | 164,5 |

A Tabela 4.12 mostra as médias dos números de candidatos $n$ para as combinações RCSG' e RCSG" para cada cenário considerando consultas relativas aos registros do *Catálogo de Amostra*. Mostramos na quarta coluna (**O(%)**) o acréscimo médio em termos de elementos da lista de candidatos de CSG' em relação à de CSG". Para todos os cenários considerados, o acréscimo médio é maior do que 150%. Tal acréscimo no custo de processamento nos motiva a investigar se realmente o ganho de eficácia de RCSG' sobre RCSG" é compensador em relação à estratégia de sempre efetuar uma segunda consulta ao Google para todos os artigos procurados.

Definimos um ganho compensador como o acréscimo em eficácia ao efetuar uma segunda consulta ao Google quando nenhuma URL relevante é encontrada na primeira consulta ao Scholar. Outros tipos de acréscimo, tais como novas URLs relevantes da segunda consulta quando já há conteúdo relevante retornado na primeira, são considerados ganhos não-compensadores.

A quarta coluna (**G(%)**) da Tabela 4.13 mostra os ganhos de RCSG' sobre RCSG" para todos os cenários considerados. Na quinta coluna (**CG(%)**) mostramos a porcentagem desses ganhos proveniente de acréscimos compensadores para cada cenário. Notamos que, embora tenham ocorrido aumentos maiores do que 150% no número de URLs candidatas extraídas, os acréscimos compensadores são inferiores a 20% do ganho global. Logo, a combinação mais adequada em termos de eficácia e tempo e custo de processamento é RCSG", a qual adotamos como a combinação a ser utilizada na imple-

Tabela 4.13: MAPs e comparação entre as combinações RCSG" e RCSG' em relação a ganhos compensadores

| | **RCSG"** | **RCSG'** | | |
|---|---|---|---|---|
| **Cenário** | **MAP(%)** | **MAP(%)** | **G(%)** | **CG(%)** |
| **Estrito** | 35,9±6,0 | 44,8±5,9 | **24,7** | 9,7 |
| **Estrito & Livre** | 37,2±7,5 | 43,1±7,3 | **15,9** | 16,9 |
| **Flexível** | 33,5±5,1 | 42,3±5,3 | **26,3** | 1,9 |
| **Flexível & Livre** | 29,9±5,6 | 36,4±5,9 | **22,0** | 7,3 |
| **Muito Flexível** | 36,6±5,4 | 46,0±5,4 | **25,7** | 1,9 |

mentação de um serviço deste tipo na BDBComp e nos referiremos como combinação Scholar-Google.

Caso o tempo de processamento seja um problema crítico, podemos diminuir o número máximo de candidatos extraídos. Analisamos para o cenário **Estrito** a variação da eficácia conforme a variação do número máximo de resultados extraídos para a combinação Scholar-Google ao procurar pelos textos dos artigos catalogados no *Catálogo de Amostra*.

Figura 4.6: Eficácia *versus* Número máximo de respostas extraídos

Como pode ser verificado na Figura 4.6, para valores menores do número máximo de respostas extraídas das máquinas de busca ($k \leq 8$), conforme este número aumenta há um aumento significativo na eficácia atingida. Entretanto, para valores maiores o incremento em eficácia é menor. Isso indica que as primeiras respostas extraídas são determinantes para a eficácia da estratégia proposta. Testes-$t$ pareados mostraram que quando utilizamos o limite de 24 respostas, a eficácia obtida não é significativamente inferior à obtida com as 40 primeiras respostas para a combinação Scholar-Google, cenário **Estrito**. Entretanto, conforme mostrado na Figura 4.7, na qual mostramos o número médio de respostas extraídas em relação ao respectivo número máximo de respostas que poderiam ser extraídas das páginas de respostas de cada máquina de busca, ao considerarmos somente as 24 primeiras respostas o número médio de respostas extraídas é 14, enquanto que com as 40 primeiras tal número médio é 17,2. Ou seja, há uma redução pequena de cerca de 18% no número médio de elementos candidatos na lista $C_a$ sem ocorrer diminuição significativa na eficácia obtida. Isto é uma evidência forte de que a diminuição do custo está fortemente relacionada com a diminuição da eficácia das consultas.

Figura 4.7: Número Médio de Respostas Extraídas *versus* Número Máximo de Respostas

Mostramos na Tabela 4.14, para a combinação Scholar-Google, os menores números máximos de respostas (coluna $k'$) para os quais não há diminuição significativa na eficácia em relação à eficácia obtida ao extrair as 40 primeiras respostas do Scholar e do Google, quando consultado, considerando cada cenário estudado. Também mostramos os números médios de respostas extraídas para o limite de 40 respostas (coluna $\bar{n}_{k=40}$) e para o menor limite (coluna $\bar{n}_{k=k'}$) para os quais não ocorrem prejuízos significativos aos resultados. Mostramos também na coluna **D(%)** a respectiva diminuição percentual ($100 \times \frac{(\bar{n}_{k=40})-(\bar{n}_{k=k'})}{\bar{n}_{k=40}}$) no número médio de respostas extraídas quando adotado o novo limite $k'$.

Tabela 4.14: Números médios de respostas extraídas para o limite $k = 40$ e para o limite $k = k'$ para os quais não ocorrem diminuição significativa na eficácia

| **Cenário** | $k'$ | $\bar{n}_{k=40}$ | $\bar{n}_{k=k'}$ | **D(%)** |
|---|---|---|---|---|
| **Estrito** | 24 | 17,2 | 14,0 | 18,6 |
| **Estrito & Livre** | 27 | 15,5 | 13,6 | 12,3 |
| **Flexível** | 31 | 17,2 | 15,8 | 8,1 |
| **Flexível & Livre** | 31 | 17,2 | 15,8 | 8,1 |
| **Muito Flexível** | 24 | 15,5 | 12,9 | 16,8 |

Notamos que a diminuição do número médio de candidatos sem prejuízos significativos aos resultados não ultrapassa 20% em quaisquer dos cenários. Limitações maiores ao número máximo de respostas do que as mostradas podem ser feitas, diminuindo ainda mais o número médio de canditados e, conseqüentemente, o custo de processamento das consultas. No entanto, neste caso, haverá prejuízos significativos na eficácia em relação ao desempenho alcançado com a extração das 40 primeiras respostas das máquinas de busca.

## 4.6 Limitações da Abordagem

Conforme analisado nas Seções 4.4 e 4.5, escolhemos como mais adequada ao processo proposto a combinação Scholar-Google que consulta primeiro o Scholar e extrai as 40 primeiras respostas como *URLs Candidatas*, que são então submetidas aos procedimentos de filtragem e reordenação antes de serem exibidas ao usuário. Caso nenhuma resposta seja selecionada para exibição ao usuário, uma segunda consulta é enviada ao Google, efetuando-se os procedimentos de extração e filtragem, e assumindo, entre as respostas selecionadas, a mesma ordem proveniente do Google para exibição ao usuário. Nesta seção levantamos algumas limitações de nossa abordagem que podem ser fontes de outras melhorias.

Em primeiro lugar, analisamos como a combinação se comporta em termos de número de registros para os quais ela é capaz de encontrar URLs relevantes. Para cada

cenário considerado, mostramos na coluna **C/URL Rel.** da Tabela 4.15 o número de registros do *Catálogo de Amostra* para os quais foi possível encontrar alguma URL relevante considerando a classificação manual efetuada nas URLs retornadas por consultas dos sete tipos considerados submetidas às cinco máquinas de busca testadas. Na coluna **Não Rec.(%)** mostramos a percentagem destes registros para os quais a combinação especificada não recuperou qualquer conteúdo relevante. Para o cenário **Muito Flexível** houve um percentual significativo de cerca de 10% de artigos para os quais não houve resposta relevante. Para os outros cenários tal percentual não ultrapassou 6%, o que indica que não encontramos evidências de que os tipos de consulta escolhidos para cada cenário e o processo de filtragem de candidatos da combinação impossibilite a recuperação de respostas relevantes para uma parcela significativa dos artigos procurados. Por outro lado, para uma parcela significativa de registros, mesmo com o retorno de URLs, não houve conteúdo relevante recuperado. A coluna **C/Resp.** mostra, para cada cenário, o número de registros para os quais a combinação retornou alguma URL como resposta e a coluna **S/URL Rel.(%)** mostra o percentual desses para os quais todas as URLs de resposta são irrelevantes. Neste caso, exceto para o cenário **Muito Flexível**, tal percentual foi maior do que 20%. Isso demonstra que a filtragem pode ser significativamente melhorada nestes outros cenários.

Tabela 4.15: Análise da capacidade de recuperação da combinação Scholar-Google em relação ao número de artigos cuja URL não foi recuperada e em relação a registros para os quais somente conteúdo irrelevante foi recuperado

| Cenário | C/URL Rel. | Não Rec.(%) | C/Resp. | S/URL Rel.(%) |
|---|---|---|---|---|
| **Estrito** | 109 | 1,8 | 147 | 27,2 |
| **Estrito & Livre** | 92 | 2,2 | 133 | 32,3 |
| **Flexível** | 126 | 5,6 | 152 | 21,7 |
| **Flexível & Livre** | 113 | 5,3 | 152 | 29,6 |
| **Muito Flexível** | 132 | 9,1 | 135 | 11,1 |

Analisamos também o procedimento de reordenação das respostas. Comparamos na Tabela 4.16 o procedimento de reordenação de acordo com as propostas de combinação (colunas **RCSG'** e **RCSG"**) com um procedimento de ideal reordenação, no qual todas as respostas relevantes aparecem nas primeiras posições da lista a ser exibida ao usuário, aplicado sobre as mesmas combinações (colunas **IRCSG'** e **IRCSG"**) As colunas **G(%)** mostram os respectivos ganhos do prodedimento ideal sobre o procedimento de reordenação adotado. Notamos que para todos os cenários a eficácia do procedimento adotado foi significativamente inferior à eficácia do procedimento ideal. As maiores melhorias possíveis dizem respeito à combinação RCSG' chegando a cerca de 20% em alguns cenários. Portanto, há espaço para melhorias no procedimento de reordenação em trabalhos futuros.

Tabela 4.16: MAPs e comparação das combinações RCSG' e RCSG" com as combinações de reordenação ideal IRCSG' e IRCSG" nas quais todas as respostas relevantes estão nas primeiras posições da lista final

| | **RCSG'** | **IRCSG'** | | **RCSG"** | **IRCSG"** | |
|---|---|---|---|---|---|---|
| **Cenário** | **MAP(%)** | **MAP(%)** | **G(%)** | **MAP(%)** | **MAP(%)** | **G(%)** |
| **Estrito** | 44,8±5,9 | 53,6±6,2 | **19,5** | 35,9±6,0 | 39,9±6,2 | **11,1** |
| **Estrito & Livre** | 43,1±7,3 | 50,7±7,5 | **17,6** | 37,2±7,5 | 40,3±7,6 | **8,4** |
| **Flexível** | 42,3±5,3 | 48,2±5,6 | **13,9** | 33,5±5,1 | 36,4±5,4 | **8,7** |
| **Flexível & Livre** | 36,4±5,9 | 44,4±6,5 | **21,8** | 29,9±5,6 | 33,3±6,1 | **11,6** |
| **Muito Flexível** | 46,0±5,4 | 49,1±5,3 | **6,7** | 36,6±5,4 | 37,8±5,4 | **3,1** |

# Capítulo 5

# Análise da Cobertura

Neste capítulo, analisamos alguns fatores que podem afetar a probabilidade de se encontrar a URL do texto de um artigo. Para isso, usaremos a noção de cobertura, que é a proporção de registros de uma coleção para os quais se encontra, dentre as respostas retornadas por uma máquina de busca ao se submeter consultas de um certo tipo, documentos relevantes a usuários de determinado grau de exigência. Os resultados relatados neste capítulo foram obtidos sob a combinação Scholar-Google considerando os melhores tipos de consulta no cenário **Estrito** para as máquinas de busca dessa combinação conforme relatado no Capítulo 4. Neste contexto, este capítulo está organizado da seguinte forma. A Seção 5.1 define formalmente o conceito de cobertura e descreve como efetuamos comparações utilizando este conceito. A Seção 5.2 analisa a chance de sucesso do usuário conforme o grau de exigência com as respostas. A Seção 5.3 compara as coberturas de artigos escritos nos idiomas português e inglês. Por fim, a Seção 5.4 apresenta resultados em relação a diferentes períodos de tempo em que os artigos foram publicados.

## 5.1 Conceito de Cobertura

Dada uma coleção de registros descrevendo um conjunto de artigos, entendemos como cobertura de uma máquina de busca sobre tal coleção a taxa desses registros para os quais é possível encontrar pelo menos uma URL relevante para usuários de um cenário submetendo consultas de determinado tipo a ela. Quando o número de registros da coleção é suficientemente grande, a cobertura está fortemente relacionada à probabilidade de se achar conteúdo relevante para um artigo da coleção descrita. A seguir definimos formalmente o conceito de cobertura.

**Definição 3** *Seja* $M = (m_1, m_2, \ldots, m_n)$ *uma lista de registros de metadados descrevendo uma coleção de artigos. Seja também* $R_{cts} = (r_{cts_1}, r_{cts_2}, \ldots, r_{cts_n})$ *um vetor de*

*cobertura, onde $r_{cts_i}$ é igual a 1, se pelo menos um documento relevante é encontrado para usuários do cenário c ao se submeter uma consulta q do tipo t à máquina de busca s, ou igual a 0, caso contrário. Assim:*

> *$C_{cts}$, a cobertura de M, no cenário c, para o tipo de consulta t e máquina de busca s, é definida como:*
>
> $$C_{cts} = \frac{\sum_{i=1}^{n} r_{cts_i}}{n}$$

O custo de se computar a cobertura aumenta de acordo com o tamanho do conjunto de registros. Devido à necessidade de inspeção manual das URLs recuperadas para cada registro, obter a cobertura para grandes coleções é inviável. Entretanto, podemos estimar um intervalo de variação para a cobertura utilizando amostragem aleatória. Diferentemente dos resultados apresentados no Capítulo 4, onde comparamos a eficácia de uma estratégia de busca em relação a outras com base na média das precisões médias associadas aos registros do *Catálogo de Amostra*, aqui estimamos a cobertura sobre uma coleção através de um único valor, a cobertura sobre uma amostra de seus registros.

Comparamos intervalos de variação da cobertura segundo o método estatístico descrito por Rea e Parker (1997), muito utilizado em pesquisas de opinião pública nas quais, com base em entrevistas a um pequeno grupo de pessoas, é predita a preferência de uma população ao ser questionada sobre diferentes alternativas. A representatividade dos resultados obtidos para o grupo de pessoas em relação à população como um todo é testada por meio de intervalos de valores cujos limites são obtidos usando-se técnicas estatísticas para cálculo de variações atribuídas ao erro de amostragem. Para tal, é necessário que o grupo de pessoas seja selecionado de maneira a reletir as características da população. Para atingir tal objetivo ao estudarmos a cobertura, efetuamos uma seleção aleatória de registros das coleções estudadas. Dessa forma, estimamos as coberturas sobre as coleções como um todo considerando amostras representativas dos respectivos registros. Tal como no Capítulo 4, os intervalos relatados possuem grau de confiança de 95%.

## 5.2 Análise dos Cenários

Iniciamos analisando a influência do grau de exigência dos usuários nos resultados. A Tabela 5.1 mostra os intervalos para a cobertura para os cenários em estudo com base na análise manual das respostas obtidas a partir dos registros da amostra utilizada. Para fins de completude, incluímos na análise três cenários adicionais. Esses cenários estão descritos a seguir com base nas mesmas categorias de conteúdo e de acessibilidade descritas no Capítulo 3. Listamos para cada um deles as categorias às quais as

URLs recuperadas devem pertencer para serem consideradas relevantes e uma descrição sucinta do grau de exigência dos usuários de cada cenário:

> **Estrito & Restrito**: 1 e a. Os usuários deste cenário estão interessados somente no texto completo de um artigo $a$ proveniente de fontes com acesso restrito.
>
> **Pelo Menos Metadados**: 1, 2, 3, 4 ou 5. Adicionamente a todos os documentos considerados relevantes no cenário **Muito Flexível**, neste cenário os usuários também estão interessados em documentos que contenham metadados redundantes que descrevem o artigo procurado $a$ ou metadados descrevendo outro documento $d$ relacionado ao artigo $a$, independente da maneira na qual podem ser acessados.
>
> **Sem Exigências**: Quaisquer das categorias. Os usuários deste cenário não possuem qualquer exigência. Qualquer tipo de conteúdo retornado ao se procurar por um artigo $a$ é considerado relevante, independente da maneira por meio da qual tal conteúdo possa ser acessado.

Na Tabela 5.1, mostramos os valores mínimo, médio e máximo dos intervalos de cobertura da coleção para cada cenário, ordenando-os do cenário de maior (menos restritivo) para o de menor (mais restritivo) valor médio. Notamos que todos os cenários abaixo do **Pelo Menos Metadados** possuem valores significativamente inferiores aos valores do cenário **Sem Exigências**. Tal diferença pode ser vista como uma evidência de que a cobertura das respostas retornadas para esses cenários menos restritivos do que o cenário **Pelo Menos Metadados** pode ser significativamente melhorada em relação a um cenário no qual os usuários consideram relevantes quaisquer das URLs recuperadas. Ou seja, há uma parcela significativa de registros para os quais somente conteúdo irrelevante é recuperado para os usuários desses outros cenários, evidenciando que o processo de filtragem pode ser melhorado.

Um outra conclusão interessante é a de que, para as coleções testadas, o percentual de artigos para os quais é possível obter acesso livre ao texto completo é significativamente maior que o de artigos com acesso restrito. Esse fato é uma evidência de que uma parte significativa dos registros catalogados nas coleções consideradas não possui texto completo disponível nas bibliotecas digitais de editoras, mas possui texto livremente acessível em alguma fonte da Web indexada pelo Scholar ou Google.

Observando, ainda, os dados em relação a um usuário do cenário **Estrito**, notamos que o percentual dos artigos das coleções BDBComp e DBLP-Br para os quais é possível recuperar o texto completo sob a combinação Scholar-Google é de no máximo cerca de 51%. Ou seja, há no mínimo uma parcela de 49% dos artigos com textos não recuperados pela nossa abordagem. Se o usuário flexibiliza um pouco as exigências e

Tabela 5.1: Comparação da cobertura para usuários de diferentes cenários

| Cenário | Min(%) | Médio(%) | Max(%) |
|---|---|---|---|
| **Sem Exigências** | 67,5 | 73,5 | 79,5 |
| **Pelo Menos Metadados** | 56,0 | 62,5 | 69,0 |
| **Muito Flexível** | 48,8 | 55,5 | 62,2 |
| **Flexível** | 52,0 | 52,0 | 58,8 |
| **Estrito** | 37,3 | 44,0 | 50,7 |
| **Flexível & Livre** | 32,4 | 39,0 | 45,6 |
| **Estrito & Livre** | 27,1 | 33,5 | 39,9 |
| **Estrito & Restrito** | 14,1 | 19,5 | 24,9 |

aceita conteúdo relacionado (cenário **Flexível**) tais porcentagens se alteram para cerca de 59% e 41%, respectivamente. É provável que essa parcela de artigos de no mínimo 41% de registros sem conteúdo recuperado possa não estar presente na Web ou não foi indexada pelo Scholar ou Google.

## 5.3 Análise por Idiomas

Também comparamos os valores de cobertura para coleções de registros de artigos em diferentes idiomas. Selecionamos da amostra inicial artigos em português e em inglês obtendo duas outras amostras. Obtivemos os intervalos de cobertura de cada uma das amostras para fins de comparação.

A Tabela 5.2 mostra os intervalos de cobertura para essas duas amostras para usuários do cenário **Estrito**. Como pode ser visto, os resultados obtidos para artigos em inglês são muito melhores. O valor médio da porcentagem coberta de artigos em inglês ultrapassa duas vezes a mesma taxa para artigos em português. Uma explicação razoável para isso é que artigos em inglês são publicados em conferências de maior exposição do que as conferências locais brasileiras nas quais foram publicados a quase totalidade dos artigos em português. Tais conferências possuem maior chance de possuírem fontes para o texto dos respectivos artigos indexadas por máquinas de busca tais como o Scholar e o Google.

| Idioma | Min(%) | Médio(%) | Max(%) |
|---|---|---|---|
| **Inglês** | 53,8 | 59,4 | 65,0 |
| **Português** | 18,0 | 26,9 | 35,8 |

Tabela 5.2: Comparação das coberturas para os idiomas inglês e português para um usuário do cenário **Estrito**

## 5.4 Análise Temporal

Similarmente, particionamos nossa amostra original por períodos de tempo para estudar o impacto desse fator na cobertura. O período de 1972 a 1994 diz respeito a artigos das coleções testadas publicados antes da Web começar a ser utilizada substancialmente como um meio para troca de informação. O segundo período (de 1995 a 2000) corresponde a uma fase de transição e expansão da Web para sua consolidação, a qual consideramos efetivamente concretizada a partir de 2001.

| Período | Min(%) | Médio(%) | Max(%) |
|---|---|---|---|
| **[1972, 1994]** | 11,2 | 22,4 | 33,6 |
| **[1995, 2000]** | 52,7 | 66,0 | 79,3 |
| **[2001, 2005]** | 34,9 | 44,2 | 53,5 |

Tabela 5.3: Comparação da cobertura em relação a períodos de tempo

A Tabela 5.3 mostra os respectivos intervalos de cobertura para cada período. Embora reduzido, o valor médio de cerca de 22% para o primeiro período mostra que é possível encontrar textos de artigos publicados antes de 1994 que, conforme mostramos na Tabela 5.4, é o período que possui os mais altos índices de registros sem URL. Para o período de transição da Web se alcançou os maiores valores mínimo, médio e máximo do intervalo de cobertura. O valor médio ficou em 66% em tal período, caindo no período da Web consolidada para aproximadamente 44%.

| | DBLP-Br | | | BDBComp | | |
|---|---|---|---|---|---|---|
| **Período** | # | **WF(%)** | **Médio** | # | **WF(%)** | **Médio** |
| **[1972, 1994]** | 397 | 91% | 0,48 | 326 | 100% | 0,00 |
| **[1995, 2000]** | 887 | 49% | 0,68 | 1171 | 55% | 0,63 |
| **[2001, 2005]** | 1898 | 23% | 0,45 | 2532 | 68% | 0,44 |

Tabela 5.4: Comparação da cobertura por períodos de tempo. Note que # indica o número total de artigos do período, **WF(%)**, a porcentagem de artigos sem URL para o texto completo, e **Médio** é o valor médio do intervalo de cobertura.

Para tentar verificar o efeito do aumento da cobertura no período de transição e possível queda no período posterior, subdividimos as amostras de cada período em subcoleções com registros provenientes da coleção BDBComp e da DBLP-Br. A Tabela 5.4 mostra algumas características e as respectivas coberturas para os artigos de cada período. Quando consideramos a coleção DBLP-Br, notamos que a porcentagem de artigos sem apontador para o texto completo tem diminuído. Tal taxa diminuiu de 91% no período de [1972, 1994] para 23% no período de [2001, 2005]. Notamos que o valor médio para cobertura no período [1995, 2000] (aproximadamente 68%) é maior do que em [1972, 1994] (aproximadamente 48%). Isso pode ser devido à tendência

de os pesquisadores tornar disponível o texto dos artigos na Web e à pouca rigidez da política de manutenção das URLs na DBLP neste período. Entretanto, depois de 2000, o valor diminui (para cerca de 45%), indicando que o pequeno percentual de artigos sem texto completo desta época seja um conjunto de trabalhos realmente mais difíceis de se encontrar o texto na Web, pois a política de catalogação dos metadados desta biblioteca tem se tornado mais restritiva.

Ao considerar a coleção BDBComp, notamos que artigos sem URL para o texto completo são mais comuns do que na coleção DBLP-Br, especialmente depois de 2000. Isso provavelmente também pode ser atribuído a mudanças na política de manutenção ao longo do tempo nesta coleção. Também é notado o impacto da transição da Web. Textos completos de artigos publicados anteriormente a 1995 são mais difíceis de serem encontrados do que os publicados no período [1995, 2000]. Por sua vez, também é observada a queda do valor médio no período da Web já consolidada. Portanto, o efeito do aumento e queda da cobertura se sustenta mesmo com a nova divisão dos registros. Finalmente, notamos que o valor médio do intervalo da cobertura é sempre menor na coleção da BDBComp do que na DBLP-Br, o que indica que os textos completos dos artigos dessa coleção sejam mais difíceis de serem encontrados na Web do que os da DBLP-Br, já que geralmente correspondem a artigos publicados em anais de conferências com menor exposição internacional.

# Capítulo 6

# Conclusões e Trabalhos Futuros

Propusemos nesta dissertação um processo que utiliza respostas de consultas submetidas a máquinas de busca para encontrar a URL do respectivo texto completo, ou de qualquer material relacionado, de artigos catalogados em uma biblioteca digital. Tal processo pode ser utilizado na implementação de um serviço de utilidade para usuários nos casos em que a biblioteca digital não fornece qualquer apontador para o texto completo, situação muito comum quando são fornecidos metadados de artigos referenciados por outros, ou ainda, quando o apontador fornecido é irrelevante, como por exemplo, quando a URL fornecida já não mais aponta para o texto desejado ou o acesso ao texto é permitido, mas de forma restrita, mediante pagamento, e o usário não está preparado ou disposto a efetuá-lo. Além disso, o processo proposto pode ser utilizado em outros contextos, como, por exemplo, busca de metadados adicionais e mais atualizados relacionados aos artigos catalogados na biblioteca digital.

Tabela 6.1: MAPs e comparação do Scholar com a combinação Scholar-Google em cada cenário

| | **Scholar** | **Scholar-Google** | |
|---|---|---|---|
| **Cenário** | **MAP(%)** | **MAP(%)** | **G(%)** |
| **Estrito** | 30,3±5,8 | 36,0±6,0 | **18,8** |
| **Estrito & Livre** | 29,5±7,0 | 37,0±7,5 | **25,7** |
| **Flexível** | 28,4±4,9 | 33,5±5,1 | **17,9** |
| **Flexível & Livre** | 22,6±5,0 | 29,8±5,6 | **31,7** |
| **Muito Flexível** | 29,4±5,0 | 36,6±5,4 | **24,6** |

Apresentamos um estudo que investiga diferentes estratégias de consulta aplicadas a diferentes máquinas de busca considerando diferentes cenários caracterizados por usuários de diferentes exigências. De acordo com experimentos efetuados com registros de metadados de artigos da área de Ciência da Computação, cujos resultados relatamos no Capítulo 4, concluímos que, para artigos de tal tipo, o Scholar é a melhor alternativa para a tarefa em questão, em todos os cenários estudados. O Google pode ser conside-

rado como uma segunda alternativa, mas apresentou eficácia equivalente à do Yahoo! nos cenários em que os usuários consideram relevantes somente documentos livremente acessíveis. Para essas três máquinas de busca, consultas que utilizam o título puro sem incluí-lo entre aspas mostraram-se mais efetivas do que as que utilizam aspas como delimitadores. MSN e CiteSeer não se mostraram como boas alternativas para esta tarefa em particular. Além disso, mostramos que, ao utilizar estratégias para combinar e reordenar resultados do Scholar e do Google, a qualidade das URLs recuperadas pode ser significativamente melhorada. Conforme mostra a Tabela 6.1 que compara a eficácia do uso da combinação Scholar-Google descrita na Seções 4.4 e 4.5 com a utilização individual do Scholar, tal combinação pode atingir um ganho de até 31,7% sobre o Scholar, dependendo do cenário considerado.

Dessa forma, sugerimos que a arquitetura do serviço descrito para uma biblioteca digital de artigos científicos da área de Ciência da Computação seja conforme mostrado na Figura 6.1. Ao interagir com a *Interface da Biblioteca Digital*, o usuário visualiza informações dos metadados de um artigo $a$ e ativa o serviço para procurar o respectivo texto completo. De posse do registro $m_a$ de *Metadados do Artigo*, a *Interface de Consulta* automaticamente gera e submete ao Scholar uma primeira consulta do tipo UT+FS formada pelo título e pelo sobrenome do primeiro autor catalogado em $m_a$, conforme descrito na Seção 3.1 do Capítulo 3. *URLs Candidatas* são então extraídas das páginas resultantes e compiladas em uma lista ordenada $C_a = [c_1, c_2, \ldots, c_n]$. Por meio do *Filtro*, *URLs Candidatas* cujo título $t_c$ seja pouco similar ao título do artigo $t_a$ são removidas, gerando uma nova lista ordenada de respostas $F_a$. Para tal sugerimos que sejam removidas URLs tais que o Coeficiente de Jaccard $J(t_c, t_a)$ computado utilizando os respectivos termos de $t_c$ e $t_a$ seja menor que um limiar $j$ entre 0,2 e 0,4; conforme mostrado experimentalmente. Caso restem elementos, a lista $F_a$ é submetida ao *Ordenador* que executa um procedimento de reordenação baseado em evidências conforme descrito na Seção 4.4. A nova lista ordenada de respostas $R_a$ é então exibida ao usuário. Caso a lista $F_a$ seja vazia, ou seja, não há resposta a ser exibida ao usuário, uma segunda consulta do tipo UT com somente o título do artigo ou do mesmo tipo usado no Scholar (UT+FS) é submetida ao Google, sendo as respostas extraídas e filtradas tal como efetuado para URLs do Scholar. Entretanto, após o procedimento de filtragem, os resultados devem ser exibidos ao usuário sem que seja realizado qualquer procedimento de reordenação, pois a ordem das respostas fornecida pelo Google já é razoável para o problema tratado. O número máximo de respostas extraídas do Scholar e do Google deve ser ajustado conforme requisitos de tempo de execução da consulta. Entretanto, conforme mostrado na Seção 4.5, salientamos que tal número tem influência direta na eficácia das respostas.

Outros resultados experimentais também mostraram que, para as coleções testadas,

Figura 6.1: Arquitetura do serviço para uma biblioteca digital de artigos científicos com foco na área de Ciência da Computação

o percentual de artigos para os quais é possível encontrar URLs de livre acesso ao texto completo dos artigos é significativamente maior do que o mesmo percentual para artigos que possuem acesso restrito, e que artigos em inglês são mais facilmente encontrados do que os escritos em português. Além disso, encontrar o texto completo de artigos publicados antes de 1995 é mais difícil do que para aqueles publicados depois desse ano. Entretanto, foi observada uma maior dificuldade em se encontrar textos de artigos recentes publicados a partir de 2001 quando comparamos com artigos publicados no período intermediário de 1995 a 2000.

Como um primeiro trabalho futuro, sugerimos a implementação de um serviço baseado no processo proposto para ajudar aos usuários da BDBComp a encontrar as URLs dos textos dos artigos catalogados nesta biblioteca digital. Mesmo para os casos nos quais já exista uma URL catalogada, tal serviço pode ser de grande valor, pois URLs podem se tornar inválidas com o decorrer do tempo devido à dinâmica da Web. Além disso, tal implementação permitiria testar a viabilidade de materialização do processo proposto evidenciando pontos que careçam de tratamento mais elaborado.

Outro trabalho importante seria abordar as limitações mostradas na Seção 4.6 do Capítulo 4. Uma delas é a melhoria do processo de filtragem, pois, em um número considerável de cenários, os experimentos mostraram que, para uma parcela significativa dos registros, foram selecionadas para exibição ao usuário somente URLs irrelevantes. Mostramos também, que o procedimento de reordenação também pode ser melhorado significativamente. Além disso, tal procedimento de reordenação pode ser adaptado de maneira a se adequar automaticamente a um conjunto de respostas razoavelmente bem ordenado como, por exemplo, o conjunto de respostas fornecido pelo Google (veja Seção 4.4.2).

Também propomos a experimentação de técnicas para fusão de múltiplas respostas tal como proposto por Gauch et al. (1996). A aplicação de tais técnicas tem o intuito de possibilitar ao usuário lidar melhor com as múltiplas cópias dos textos dos artigos.

Finalmente, apesar de resultados recentes mostrarem que máquinas de busca tais como o Scholar também cobrem de maneira significativa o conteúdo de outras áreas do conhecimento (Walters, 2007), seria importante conduzir novos experimentos com coleções de outras áreas para avaliar a generalidade do processo proposto. Uma área a ser explorada é a de saúde. Experimentos para buscar o texto de artigos catalogados na MEDLINE[1], uma coleção com mais de 15 milhões de registros de artigos relacionados

[1] http://www.nlm.nih.gov/

a biomedicina, podem trazer resultados interessantes. Uma parcela significativa desses registros possui URL catalogada. Entretanto, o acesso ao texto é restrito, situação na qual pode ser útil achar uma URL na Web que forneça livre acesso ao texto.

# Apêndice A

# Resultados Adicionais

## A.1 Cenário Estrito & Livre

Tabela A.1: Intervalos da MAP no Google para cada tipo de consulta no cenário **Estrito & Livre**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,32 | 17,09 ± 7,00 |
| UT | 0,31 | 36,62 ± 8,30 |
| UT+FS | 0,23 | **39,26 ± 8,97** |
| UT+AS | 0,26 | 30,62 ± 8,26 |
| QT | 0,22 | 28,56 ± 8,91 |
| QT+FS | 0,22 | 27,74 ± 8,85 |
| QT+AS | 0,22 | 25,70 ± 7,74 |

Tabela A.2: Intervalos da MAP no Yahoo! para cada tipo de consulta no cenário **Estrito & Livre**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,30 | 17,17 ± 7,96 |
| UT | 0,23 | 57,66 ± 9,69 |
| UT+FS | 0,22 | **58,45 ± 10,76** |
| UT+AS | 0,30 | 49,03 ± 11,16 |
| QT | 0,22 | 54,06 ± 10,83 |
| QT+FS | 0,22 | 51,12 ± 10,83 |
| QT+AS | 0,30 | 46,25 ± 10,87 |

Tabela A.3: Intervalos da MAP no MSN para cada tipo de consulta

| **Tipo de Consulta** | **Limiar $j$** | **MAP(%)** |
|---|---|---|
| AS | 0,22 | 27,50 ± 20,76 |
| UT | 0,27 | 23,17 ± 19,19 |
| UT+FS | 0,22 | 61,25 ± 21,63 |
| UT+AS | 0,22 | 71,25 ± 19,83 |
| QT | 0,22 | **75,00 ± 19,36** |
| QT+FS | 0,22 | **75,00 ± 19,36** |
| QT+AS | 0,22 | 63,75 ± 21,96 |

Tabela A.4: Intervalos da MAP no Scholar para cada tipo de consulta no cenário **Estrito & Livre**

| **Tipo de Consulta** | **Limiar $j$** | **MAP(%)** |
|---|---|---|
| AS | 0,32 | 24,17 ± 8,88 |
| UT | 0,37 | 51,39 ± 10,30 |
| UT+FS | 0,36 | **67,14 ± 8,84** |
| UT+AS | 0,36 | 59,77 ± 9,29 |
| QT | 0,30 | 62,29 ± 9,59 |
| QT+FS | 0,30 | 62,73 ± 9,60 |
| QT+AS | 0,30 | 59,47 ± 9,42 |

Tabela A.5: Intervalos da MAP no CiteSeer para cada tipo de consulta no cenário **Estrito & Livre**

| **Tipo de Consulta** | **Limiar $j$** | **MAP(%)** |
|---|---|---|
| AS | 0,42 | 60,74 ± 14,76 |
| UT | 0,24 | 65,96 ± 13,29 |
| UT+FS | 0,23 | **67,08 ± 13,48** |
| UT+AS | 0,29 | 58,67 ± 15,19 |
| QT | 0,22 | 53,51 ± 15,56 |
| QT+FS | 0,22 | 44,58 ± 15,92 |
| QT+AS | 0,22 | 50,22 ± 16,17 |

## A.2 Cenário Flexível

Tabela A.6: Intervalos da MAP no Google para cada tipo de consulta no cenário **Flexível**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,23 | 22,00 ± 5,53 |
| UT | 0,23 | **37,84 ± 5,88** |
| UT+FS | 0,23 | 36,25 ± 5,70 |
| UT+AS | 0,23 | 34,54 ± 6,18 |
| QT | 0,22 | 21,19 ± 5,31 |
| QT+FS | 0,22 | 20,73 ± 5,28 |
| QT+AS | 0,22 | 20,15 ± 5,05 |

Tabela A.7: Intervalos da MAP no Yahoo! para cada tipo de consulta no cenário **Flexível**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,22 | 34,56 ± 8,62 |
| UT | 0,23 | 53,14 ± 8,42 |
| UT+FS | 0,22 | **59,87 ± 8,48** |
| UT+AS | 0,22 | 51,60 ± 8,86 |
| QT | 0,22 | 43,26 ± 8,92 |
| QT+FS | 0,22 | 41,64 ± 8,84 |
| QT+AS | 0,22 | 39,14 ± 8,80 |

Tabela A.8: Intervalos da MAP no MSN para cada tipo de consulta no cenário **Flexível**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,22 | 29,17 ± 14,65 |
| UT | 0,23 | 21,64 ± 13,55 |
| UT+FS | 0,22 | 48,94 ± 15,67 |
| UT+AS | 0,22 | 52,10 ± 15,74 |
| QT | 0,22 | **52,50 ± 15,65** |
| QT+FS | 0,22 | **52,50 ± 15,65** |
| QT+AS | 0,22 | 46,39 ± 16,22 |

Tabela A.9: Intervalos da MAP no Scholar para cada tipo de consulta no cenário **Flexível**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,22 | 33,85 ± 7,90 |
| UT | 0,24 | 55,31 ± 8,29 |
| UT+FS | 0,23 | **72,24 ± 6,80** |
| UT+AS | 0,23 | 64,77 ± 7,38 |
| QT | 0,30 | 50,33 ± 8,29 |
| QT+FS | 0,30 | 50,14 ± 8,25 |
| QT+AS | 0,30 | 47,87 ± 8,23 |

Tabela A.10: Intervalos da MAP no CiteSeer para cada tipo de consulta no cenário **Flexível**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,22 | **64,39 ± 11,74** |
| UT | 0,24 | 53,91 ± 12,30 |
| UT+FS | 0,22 | 52,73 ± 12,51 |
| UT+AS | 0,29 | 44,88 ± 12,85 |
| QT | 0,22 | 34,80 ± 12,32 |
| QT+FS | 0,22 | 28,06 ± 11,75 |
| QT+AS | 0,22 | 32,77 ± 12,68 |

## A.3 Cenário Flexível & Livre

Tabela A.11: Intervalos da MAP no Google para cada tipo de consulta no cenário **Flexível & Livre**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,23 | 19,70 ± 6,60 |
| UT | 0,23 | **34,93 ± 6,79** |
| UT+FS | 0,23 | 34,65 ± 6,89 |
| UT+AS | 0,23 | 30,02 ± 6,68 |
| QT | 0,22 | 21,19 ± 6,78 |
| QT+FS | 0,22 | 20,74 ± 6,78 |
| QT+AS | 0,22 | 19,52 ± 6,20 |

Tabela A.12: Intervalos da MAP no Yahoo! para cada tipo de consulta no cenário **Flexível & Livre**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,22 | 24,81 ± 8,46 |
| UT | 0,23 | 51,76 ± 8,39 |
| UT+FS | 0,22 | **56,75 ± 9,31** |
| UT+AS | 0,22 | 47,38 ± 9,38 |
| QT | 0,22 | 39,80 ± 9,32 |
| QT+FS | 0,22 | 38,09 ± 9,18 |
| QT+AS | 0,22 | 34,59 ± 9,00 |

Tabela A.13: Intervalos da MAP no MSN para cada tipo de consulta no cenário **Flexível & Livre**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,22 | 27,27 ± 15,25 |
| UT | 0,23 | 20,58 ± 13,88 |
| UT+FS | 0,22 | 51,87 ± 16,59 |
| UT+AS | 0,22 | **55,32 ± 16,58** |
| QT | 0,22 | 49,70 ± 16,60 |
| QT+FS | 0,22 | 49,70 ± 16,60 |
| QT+AS | 0,22 | 43,03 ± 17,09 |

Tabela A.14: Intervalos da MAP no Scholar para cada tipo de consulta no cenário **Flexível & Livre**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,22 | 24,86 ± 8,21 |
| UT | 0,24 | 50,67 ± 9,15 |
| UT+FS | 0,23 | **64,38 ± 7,99** |
| UT+AS | 0,23 | 55,93 ± 8,30 |
| QT | 0,22 | 45,40 ± 9,03 |
| QT+FS | 0,22 | 45,80 ± 9,07 |
| QT+AS | 0,30 | 43,67 ± 8,88 |

Tabela A.15: Intervalos da MAP no CiteSeer para cada tipo de consulta no cenário **Flexível & Livre**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,22 | **64,39 ± 11,74** |
| UT | 0,24 | 53,91 ± 12,30 |
| UT+FS | 0,22 | 52,73 ± 12,51 |
| UT+AS | 0,29 | 44,88 ± 12,85 |
| QT | 0,22 | 34,80 ± 12,32 |
| QT+FS | 0,22 | 28,06 ± 11,75 |
| QT+AS | 0,22 | 32,77 ± 12,68 |

## A.4 Cenário Muito Flexível

Tabela A.16: Intervalos da MAP no Google para cada tipo de consulta no cenário **Muito Flexível**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,23 | 24,58 ± 5,46 |
| UT | 0,23 | 41,72 ± 5,78 |
| UT+FS | 0,23 | **41,74 ± 5,60** |
| UT+AS | 0,23 | 37,43 ± 5,99 |
| QT | 0,22 | 26,57 ± 5,52 |
| QT+FS | 0,22 | 25,34 ± 5,46 |
| QT+AS | 0,22 | 24,04 ± 5,24 |

Tabela A.17: Intervalos da MAP no Yahoo! para cada tipo de consulta no cenário **Muito Flexível**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,22 | 33,61 ± 8,44 |
| UT | 0,23 | 53,78 ± 8,30 |
| UT+FS | 0,22 | **60,83 ± 8,38** |
| UT+AS | 0,22 | 52,96 ± 8,73 |
| QT | 0,22 | 46,33 ± 8,95 |
| QT+FS | 0,22 | 44,78 ± 8,88 |
| QT+AS | 0,22 | 39,99 ± 8,69 |

Tabela A.18: Intervalos da MAP no MSN para cada tipo de consulta no cenário **Muito Flexível**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,22 | 26,32 ± 13,50 |
| UT | 0,23 | 20,24 ± 12,91 |
| UT+FS | 0,22 | 47,68 ± 15,51 |
| UT+AS | 0,22 | 50,67 ± 15,13 |
| QT | 0,22 | **57,63 ± 15,08** |
| QT+FS | 0,22 | **57,63 ± 15,08** |
| QT+AS | 0,22 | 46,58 ± 15,33 |

Tabela A.19: Intervalos da MAP no Scholar para cada tipo de consulta no cenário **Muito Flexível**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,22 | 37,66 ± 8,14 |
| UT | 0,24 | 56,32 ± 8,20 |
| UT+FS | 0,23 | **73,48 ± 6,72** |
| UT+AS | 0,23 | 66,86 ± 7,33 |
| QT | 0,30 | 53,22 ± 8,21 |
| QT+FS | 0,30 | 53,22 ± 8,21 |
| QT+AS | 0,30 | 51,06 ± 8,22 |

Tabela A.20: Intervalos da MAP no CiteSeer para cada tipo de consulta no cenário **Muito Flexível**

| **Tipo de Consulta** | **Limiar** $j$ | **MAP(%)** |
|---|---|---|
| AS | 0,22 | **74,13 ± 8,95** |
| UT | 0,24 | 57,25 ± 11,36 |
| UT+FS | 0,24 | 55,34 ± 10,85 |
| UT+AS | 0,29 | 50,25 ± 10,94 |
| QT | 0,22 | 39,99 ± 10,43 |
| QT+FS | 0,22 | 34,32 ± 10,05 |
| QT+AS | 0,22 | 36,54 ± 10,96 |

# Referências Bibliográficas


Baeza-Yates, R. e Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley, New York, NY, USA.

Bharat, K. e Broder, A. (1998). A technique for measuring the relative size and overlap of public web search engines. *Computer Networks and ISDN Systems*, 30(1-7):379–388.

Chakrabarti, S.; van den Berg, M. e Dom, B. (1999). Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networks*, 31(11-16):1623–1640.

Chu, H. e Rosenthal, M. (1996). Search engines for the World Wide Web: A comparative study and evaluation methodology. In *Proceedings of the Annual Conference of the American Society for Information Science*, pp. 127–135, Baltimore, MD, USA.

Gauch, S.; Wang, G. e Gomez, M. (1996). ProFusion: Intelligent Fusion from Multiple, Distributed Search Engines. *Journal of Universal Computing*, 2(9):637–649.

Gonçalves, M. A.; Fox, E. A.; Watson, L. T. e Kipp, N. A. (2004). Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries. *ACM Trans. Inf. Syst.*, 22(2):270–312.

Gordon, M. e Pathak, P. (1999). Finding information on the world wide web: The retrieval effectiveness of search engines. *Information Processing and Management*, 35(2):141–180.

Hoff, G. e Mundhenk, M. (2001). Finding scientific papers with homepagesearch and mops. In *Proceedings of the 19th Annual International Conference on Computer Documentation*, pp. 201–207, Sante Fe, New Mexico, USA. ACM Press.

Jain, R. (1991). *The Art of Computer Systems Performance Analysis*. John Wiley and Sons, Inc., New York, NY, USA.

Laender, A. H. F.; Gonçalves, M. A. e Roberto, P. A. (2004). BDBComp: Building a Digital Library for the Brazilian Computer Science Community). In *Proceedings*

*of the 2004 ACM/IEEE Joint Conference on Digital Libraries*, pp. 23–24, Tucson, Arizona, USA.

Lawrence, S. (2001). Online or invisible? *Nature*, 411(6837):521.

Lawrence, S. e Giles, C. L. (1998). Searching the World Wide Web. In *Science, 280(5360):98.*

Lawrence, S. e Giles, C. L. (1999a). Accessibility of information on the web. *Nature*, 400:107–109.

Lawrence, S. e Giles, C. L. (1999b). Searching the web: General and scientific information access. *IEEE Communications*, 37(1):116–122.

Menczer, F.; Pant, G. e Srinivasan, P. (2004). Topical web crawlers: Evaluating adaptive algorithms. *ACM Transactions on Internet Technology*, 4(4):378–419.

On, B.-W. e Lee, D. (2004). Pase: Locating online copy of scientific documents effectively. In *Proceedings of the 7th annual International Conference of Asian Digital Libraries*, pp. 408–418, Shanghai, China.

Pant, G.; Srinivasan, P. e Menczer, F. (2004). Crawling the Web. In Levene, M. e Poulovassilis, A., editores, *Web Dynamics: Adapting to Change in Content, Size, Topology and Use*, pp. 153–177. Springer-Verlag, Berlin, Germany.

Rea, L. M. e Parker, R. A. (1997). *Designing and Conducting Survey Research: A Comprehensive Guide.* Jossey-Bass, San Francisco, California, USA.

Salton, G.; Wong, A. e Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Selberg, E. e Etzioni, O. (1995). Multi-service search and comparison using the MetaCrawler. In *Proceedings of the 4th International World Wide Web Conference*, Boston, MA, USA.

Silva, A. J. C.; Gonçalves, M. A.; Laender, A. H. F.; Modesto, M. A. B. e Cristo, M. (2007). Finding what is missing from a digital library: A case study in the computer science field. *Submetido para publicação.*

Silva, A. J. C.; Modesto, M. A. B.; Gonçalves, M. A.; Cristo, M.; Laender, A. H. F. e Ziviani, N. (2006). Busca pelo texto completo de artigos catalogados em uma biblioteca digital. In *Anais do II Workshop de Bibliotecas Digitais*, pp. 71–80, Florianópolis, Santa Catarina.

Tan, P.-N.; Steinbach, M. e Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Walters, W. H. (2007). Google scholar coverage of a multidisciplinary field. *Information Processing and Management*, 43(4):1121–1132.

Zhuang, Z.; Wagle, R. e Giles, C. L. (2005). What's there and what's not?: focused crawling for missing documents in digital libraries. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 301–310, Denver, CO, USA.