FABIANO MAGALHÃES ATALLA DA FONSECA

IMPACTO DO COMPORTAMENTO DINÂMICO DOS PARES NA EFICÁCIA DE MÁQUINAS DE BUSCA PAR-A-PAR

Belo Horizonte
01 de julho de 2008

FABIANO MAGALHÃES ATALLA DA FONSECA
Orientador: Virgílio Augusto Fernandes Almeida

IMPACTO DO COMPORTAMENTO DINÂMICO DOS PARES NA EFICÁCIA DE MÁQUINAS DE BUSCA PAR-A-PAR

Dissertação apresentada ao Programa 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.

Belo Horizonte
01 de julho de 2008

UNIVERSIDADE FEDERAL DE MINAS GERAIS

FOLHA DE APROVAÇÃO

Impacto do Comportamento Dinâmico dos Pares na Eficácia de Máquinas de Busca Par-a-Par

FABIANO MAGALHÃES ATALLA DA FONSECA

Dissertação defendida e aprovada pela banca examinadora constituída por:

Doutor Virgílio Augusto Fernandes Almeida – Orientador
Universidade Federal de Minas Gerais

Doutora Jussara Marques de Almeida
Universidade Federal de Minas Gerais

Doutor Marcos André Gonçalves
Universidade Federal de Minas Gerais

Doutor Nívio Ziviani
Universidade Federal de Minas Gerais

Doutor Wagner Meira Júnior
Universidade Federal de Minas Gerais

Belo Horizonte, 01 de julho de 2008

Agradecimentos

Agradeço a todos aqueles que contribuíram de alguma forma para que eu pudesse chegar até aqui. Em primeiro lugar, agradeço a Deus, meu porto seguro, fonte inesgotável de energia e de inspiração. Agradeço a minha família, em especial a meus pais, avó e irmãs, pelo apoio emocional, moral e financeiro, pela compreensão nos incontáveis momentos em que precisei me ausentar para me dedicar aos estudos. Agradeço a todos os meus amigos que mostraram interesse no trabalho e souberam me apoiar nos momentos de desânimo e participar de meus momentos de empolgação com minha linha de pesquisa; aos colegas de universidade, pelos momentos de descontração e aprendizado. Agradeço às entidades de apoio à pesquisa, CNPq, FAPEMIG e CAPES, pelo suporte financeiro. Agradeço aos professores, desde o ensino básico, passando pelo CEFET, graduação e mestrado na UFMG, pelos ensinamentos. Em particular, agradeço aos professores com quem tive a oportunidade de trabalhar durante a graduação e mestrado: ao professor Wagner Meira Jr., por suas contribuições como orientador de iniciação científica, por sua participação e apoio financeiro para que eu pudesse divulgar e apresentar meu primeiro artigo científico no exterior; aos professores Jussara Marques de Almeida e Marcos André Gonçalves, pelas insubstituíveis contribuições nos artigos científicos escritos durante o mestrado; e finalmente ao meu orientador de mestrado, Virgílio, grande mestre, pelos valiosos ensinamentos, apoio moral e financeiro, que culminaram na divulgação do trabalho em publicações nacionais e internacionais e na consolidação desta dissertação.

Resumo

Na tentativa de ampliar o espectro de busca e atenuar problemas de escalabilidade, redes Par-a-Par (P2P) têm sido apontadas como alternativa para novas gerações de máquinas de busca na Web. No entanto, a eficácia da busca por conteúdo em ambientes P2P pode ser gravemente limitada por características observadas em sistemas P2P reais, tais como a entrada e saída dinâmica de pares no sistema. Nosso estudo analisa o impacto desse aspecto na eficácia de máquinas de busca P2P. De forma a estimar os limites da eficácia, focamos nossa análise em modelos de rede P2P com níveis extremos de conhecimento dos pares sobre os documentos da rede. Nossos resultados revelam que o comportamento dinâmico dos pares pode afetar consideravelmente a eficácia da busca mesmo em cenários otimistas: em redes com altos níveis de conhecimento dos pares sobre os documentos da rede, uma fração significativa de consultas sofre um impacto na qualidade das respostas de pelo menos 26% ainda em cenários muito estáveis. Também confirmamos que o impacto desse aspecto em redes com baixos níveis de conhecimento dos pares pode ser ainda mais grave (75%).

Também avaliamos a replicação de conteúdo como possível forma de atenuar os efeitos do comportamento dinâmico dos pares na eficácia de máquinas de busca P2P. Para tanto, analisamos o efeito de os usuários baixarem algumas páginas listadas na resposta à consulta e as adicionarem à sua coleção local. Observamos que essa estratégia pode melhorar significativamente a eficácia de máquinas de busca P2P. De fato, a qualidade das respostas em redes com níveis muito baixos de conhecimento dos pares sobre os documentos da rede pode melhorar significativamente mesmo em cenários pouco estáveis. Também discutimos os desafios existentes para adoção dessa solução. De fato, considerando a grande autonomia dos pares e a ausência dos benefícios da replicação comuns em sistemas P2P de compartilhamento de arquivos, o desenvolvimento das futuras máquinas de busca P2P pode depender amplamente de novos mecanismos de incentivo que considerem aspectos específicos desse tipo de aplicação.

Abstract

In an attempt to increase the spectrum of searchable information while attenuating scalability issues, Peer-to-Peer (P2P) networks have been viewed as an alternative way to design new Web search engines. However, the effectiveness of P2P Web searching may be severely limited by characteristics commonly observed in real P2P systems such as the dynamics of peer participation (churn). This study analyzes the impact of such issue on the effectiveness of P2P Web search engines. In order to estimate effectiveness boundaries, we focus our analysis on P2P network models with very high and low levels of peer knowledge about documents on the network. Our findings reveal that peer dynamic behavior could strongly affect search effectiveness even in optimistic scenarios: in networks where peers have a high knowledge about documents on the network, a significant fraction of queries suffer an impact on the quality of search of at least 26% still in highly stable scenarios. We also confirm that the impact of such issue in networks where peers have a lower level of knowledge can be even more intense (75%).

We also evaluate content replication as a possible way to attenuate the effects of peer dynamic behavior on the effectiveness of P2P search engines. To this end, we analyze the effect of users downloading some ranked Web pages and adding them to their local collection. We observe that such strategy can significantly improve the effectiveness of P2P Web searching. In fact, the quality of results of networks with a very low level of peer knowledge about documents on the network can be significantly improved even in lowly stable scenarios. We then discuss some imposed challenges for the adoption of such solution. Actually, together with the high autonomy of peers and the absence of file-sharing benefits in replicating documents into the network, effectiveness of P2P Web search engines may strongly depend on new, application-specific incentive mechanisms for the users.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Motivação

Sistemas Par-a-Par (P2P) têm se mostrado particularmente atrativos pelo seu potencial de escalabilidade e autonomia dos usuários, além da crescente popularidade em diversos tipos de aplicação, como voz sobre IP e compartilhamento de arquivos [Yu et al. (2005); Pouwelse et al. (2005)]. Em especial, sistemas P2P têm sido apontados como alternativa para aplicações de máquinas de busca na Web [Bender et al. (2006); Wu et al. (2005); Cuenca-Acuna et al. (2003); Lu e Callan (2003)]. De fato, com o crescimento explosivo da Web, tanto em número de usuários quanto de documentos, as máquinas de busca tradicionais precisam não apenas recuperar e indexar uma enorme quantidade de páginas mas também processar centenas de milhões de consultas de usuários por dia [Ntoulas e Cho (2007)]. De fato, o número de páginas recuperáveis por máquinas de busca excedeu 11 bilhões em 2005 [Gulli e Signorini (2005)]. Além disso, fatores como atualização de páginas, monopólio de informação e interesses comerciais [Bender et al. (2006); Doulkeridis et al. (2006)] sugerem a necessidade de se explorar alternativas de arquitetura eficazes e eficientes para a busca por conteúdo na Web.

Uma máquina de busca P2P é composta de computadores independentes e heterogêneos (i.e., pares), cada um contendo parte dos documentos da rede. Cada par envia consultas conforme interesse dos usuários e responde a consultas (locais e remotas) com uma lista ordenada de documentos relevantes presentes em um repositório local. A aplicação de uma arquitetura distribuída em P2P para a busca na Web pode não apenas amenizar restrições de escalabilidade [Wu et al. (2005)] como também explorar a parte da Web que máquinas de busca centralizadas não indexam – a chamada *Web invisível*, com cerca de bilhões de documentos [Lewandowski e Mayr (2006)].

O desenvolvimento de máquinas de busca P2P é recente. Grande parte dos esforços foca em obter uma eficácia (i.e., recuperação de documentos relevantes) próxima da obtida em máquinas de busca centralizadas. Além disso, em geral considera-se um cenário ideal em que os pares nunca saem da rede. No entanto, estudos de aplicações populares P2P de *compartilhamento*

de arquivos analisaram o comportamento dos pares, concluindo que esses geralmente são muito dinâmicos, entrando e saindo várias vezes da rede [Stutzbach e Rejaie (2006)]. Nesse cenário instável, um usuário pode não obter os documentos mais relevantes para sua consulta se esses documentos pertencerem a pares ausentes no momento em que a consulta for submetida. Logo, o *comportamento dinâmico* dos pares está diretamente relacionado à disponibilidade de conteúdo e pode afetar significativamente a eficácia da busca por conteúdo em ambientes P2P e, em última instância, a qualidade do serviço provido por máquinas de busca P2P.

Acreditamos que o impacto do comportamento dinâmico dos pares na eficácia de máquinas de busca P2P difere daquele visto em aplicações P2P de compartilhamento de arquivos. Geralmente, em aplicações de compartilhamento de arquivos o usuário tem interesse em um objeto em particular; se qualquer par tiver esse objeto, a busca será efetiva. Além disso, o objeto pode estar replicado em vários pares. Por outro lado, em uma máquina de busca P2P, o usuário busca informação sobre um tópico ou assunto; nesse caso, diferentes documentos de diferentes pares podem ser relevantes para o usuário em diferentes níveis. A ausência, no momento em que a consulta é realizada, de um par com muitos documentos relevantes ou muitos pares com poucos documentos relevantes pode ter um grande impacto na qualidade da resposta. Mesmo em casos em que o usuário procura por um único documento na máquina de busca P2P, pode ocorrer que esse documento esteja presente em apenas um par, o que pode levar à falha da busca se o par estiver indisponível. Tais aspectos sugerem que o conjunto de documentos resultantes de uma consulta, assim como o conjunto dos pares selecionados para processá-la, são diferentes para cada uma dessas aplicações de busca, isto é, a eficácia da busca pode ser estritamente dependente da aplicação.

Portanto, o comportamento dinâmico dos pares gera um ambiente instável sobre o qual as máquinas de busca P2P deverão residir. O impacto desse aspecto na eficácia da busca ainda não foi de todo analisado [1]. Ressaltamos ainda que a robustez da aplicação em relação ao comportamento dinâmico dos pares reflete a qualidade do serviço na visão do usuário final e, em última instância, a confiabilidade em relação à aplicação.

1.2 Objetivos

Pretendemos apresentar uma análise da viabilidade de máquinas de busca P2P como alternativa para as máquinas de busca tradicionais. Para tanto, quantificamos, a partir de um ambiente de simulação híbrido, o impacto do comportamento dinâmico dos pares na eficácia da busca por conteúdo em ambientes P2P. Consideramos cenários com diferentes níveis de conhecimento dos pares sobre os documentos da rede e diferentes distribuições de documentos entre os pares.

Além disso, pretendemos avaliar alternativas de se atenuar o potencial impacto do comportamento dinâmico dos pares na eficácia de máquinas de busca P2P. Nessa direção, estendemos

[1] À exceção do trabalho que publicamos abordando o assunto [Atalla et al. (2008)]

nosso ambiente de simulação para considerar cenários em que os pares realizam consultas, baixam uma fração das páginas retornadas pela máquina de busca P2P e as adicionam ao conjunto de documentos local. A cópia local (i.e., *download*) e a adição ao conjunto de documentos local de páginas da Web permitem aumentar a disponibilidade de conteúdo na rede e, conseqüentemente, atenuar os efeitos do comportamento dinâmico dos pares na qualidade da busca.

1.3 Contribuições

Nossos resultados mostram que a eficácia da busca por conteúdo em ambientes P2P sofre significativamente com a instabilidade dos pares mesmo em cenários otimistas, em que pares possuem muito conhecimento sobre os documentos da rede (i.e., muita informação difundida na rede acerca dos documentos existentes e sua relevância). Quando pares estão disponíveis, em média, 75% do tempo de simulação, a eficácia da busca apresenta uma degradação na Precisão Média (AP), i.e., a fração de documentos relevantes presente na resposta a uma consulta, superior a 24% para 20% das consultas, quando comparada à solução centralizada. Além disso, mesmo agregando os resultados da AP sobre todas as consultas (MAP), a degradação chega a 22%. Também mostramos que em cenários mais pessimistas, em que pares possuem menos conhecimento sobre os documentos da rede, o impacto pode ser ainda mais intenso: para 20% das consultas, a degradação é de pelo menos 73% para o mesmo nível de instabilidade. Assim, mostramos que o comportamento dinâmico dos pares pode representar um sério desafio no desenvolvimento de máquinas de busca em ambientes P2P.

No entanto, observamos que a estratégia de replicação dinâmica própria (*owner replication*) [Lv et al. (2002)], adaptada ao contexto de máquinas de busca P2P, pode melhorar consideravelmente a eficácia da busca. De fato, a qualidade dos resultados da busca em redes com níveis muito baixos de conhecimento dos pares sobre os documentos da rede aumenta significativamente mesmo em cenários muito instáveis.

Uma vez que o comportamento dinâmico dos pares é um aspecto intrínseco a redes P2P e logo difícil de se evitar, destacamos a importância de estratégias de incentivo dos usuários. De fato, além da adoção da estratégia de replicação que estudamos, usuários devem ser incentivados a manter a aplicação de busca em execução mesmo quando não estiverem realizando consultas e a não restringir a publicação de documentos do conjunto de documentos local. Essas e outras estratégias podem ser essenciais para viabilizar a adoção de máquinas de busca em ambientes P2P reais.

1.4 Organização do Texto

O texto da dissertação está organizado da seguinte forma: o Capítulo 2 descreve trabalhos relacionados à busca por conteúdo em ambientes P2P. O Capítulo 3 sumariza nossa metodologia de

avaliação de máquinas de busca. Descrevemos nesse capítulo as coleções de teste adotadas e as métricas utilizadas para quantificar a qualidade dos resultados oferecidos pela busca ao usuário. Detalhamos as arquiteturas de máquinas de busca avaliadas nos Capítulos 4 e 5. No Capítulo 4 apresentamos o modelo de máquina de busca centralizada, o qual corresponde à nossa linha de base. No Capítulo 5, diferentes cenários de arquitetura P2P são avaliados em relação à linha de base. Em particular, a Seção 5.5 discute o impacto do comportamento dinâmico dos pares em máquinas de busca P2P e estratégias de replicação de conteúdo propostas para atenuar o problema são avaliadas na Seção 5.6. Estratégias de incentivo dos usuários também são discutidas nesse capítulo. Finalmente, considerações finais e sugestões de trabalhos futuros são enumeradas no Capítulo 6. O Apêndice A discute alguns aspectos de eficiência que foram essenciais para viabilizar a execução do ambiente de simulação.

Capítulo 2

Revisão da Literatura

2.1 Arquiteturas Par-a-Par

Com o interesse cada vez maior da comunidade acadêmica em ambientes Par-a-Par (P2P), especialmente em razão de seu crescimento e popularidade [Crespo e Garcia-Molina (2002); Acosta e Chandra (2005)], diversas arquiteturas de redes de compartilhamento de arquivos têm sido propostas. As arquiteturas de redes P2P mais estudadas são as (des) centralizadas e (não-) estruturadas.

Arquiteturas descentralizadas e não estruturadas são aquelas em que não existe gerenciamento de arquivos e conexões entre pares são definidas arbitrariamente. Versões do Gnutella são exemplos clássicos de aplicações que as utilizam. Nessas arquiteturas não existem pares responsáveis por gerenciamento do conteúdo da rede nem um ambiente centralizado em que consultas possam ser processadas. Isso significa que, sempre que um par precisa localizar um objeto, é necessário que sua requisição seja pesquisada ao longo da rede, em busca do maior número possível de pares com a informação desejada. Lv et al. (2002) propõem alternativas mais eficientes que a busca exaustiva para a localização e acesso a pares em arquiteturas descentralizadas e não estruturadas, como algoritmos de expansão em anel (*expanding ring*) e caminhamento aleatório.

Em arquiteturas centralizadas, um ou mais servidores centrais são responsáveis por armazenar informação sobre pares e conteúdo compartilhado e responder a pedidos de arquivos. Nessas arquiteturas, o conteúdo dos arquivos é armazenado nos pares. Napster é um exemplo de aplicação P2P centralizada. Por centralizar o gerenciamento dos arquivos, essas arquiteturas podem apresentar sérios problemas de escalabilidade e disponibilidade.

Finalmente, arquiteturas descentralizadas e estruturadas são caracterizadas por diluir o gerenciamento dos arquivos por todos os pares da rede. Em outras palavras, ao invés de existirem servidores centrais de gerenciamento de arquivos, cada par fica responsável por localizar uma parte do conteúdo disponível na rede. Isso pode ser feito a partir da utilização de algum protocolo baseado em funções de *hash*. Kademlia [Maymounkov e Mazieres (2002)] é um exemplo de

protocolo utilizado por aplicações como eMule [Kulbak e Bickson (2005)], que utiliza tabelas *hash* distribuídas (DHT). Sempre que um par procura por um objeto, a entrada das palavras-chave referentes ao objeto é acessada na *hash*, a qual informa os pares responsáveis pela localização do conteúdo na rede. Esses pares por sua vez realizam a pesquisa propriamente dita e mantêm tabelas de roteamento para os pares que têm o conteúdo desejado [Stutzbach e Rejaie (2006)].

2.2 Máquinas de Busca Par-a-Par

Vários estudos propõem diferentes modelos para busca na Web em P2P, cobrindo tanto arquiteturas P2P estruturadas quanto não estruturadas.

Bender et al. (2005) apresentam Minerva, uma máquina de busca construída sobre uma arquitetura P2P descentralizada e estruturada que utiliza tabelas DHT. Em seu modelo, cada par apresenta seu próprio conjunto de documentos, indexado por uma lista invertida local para cada termo dos documentos; além disso, descritores de documentos correspondem aos termos que ocorrem nos próprios documentos. Basicamente, quando um par inicia uma consulta, cada termo da consulta é pesquisado na DHT em busca do par responsável. Esse par é em seguida contactado para obtenção de uma lista de pares promissores, isto é, que contenham documentos relevantes para o termo. O par requisitante então agrupa as listas de pares resultantes de alguma forma (os autores sugerem interseção das listas), escolhe os mais promissores e então envia toda a consulta para cada um desses pares. Finalmente, o conjunto de documentos obtidos é classificado e ordenado por relevância utilizando alguma estratégia (os autores sugerem o uso de estatísticas globais presentes na DHT).

Wu et al. (2005) apresentam 6Search (6S), uma máquina de busca em uma arquitetura P2P descentralizada e não estruturada. Inicialmente, as consultas são roteadas aleatoriamente como em modelos mais eficientes de *flooding*, já que não existe informação global (embora distribuída) como em arquiteturas estruturadas. No entanto, o modelo parece utilizar algoritmos de *amostragem baseada em consultas* [Callan e Connell (2001)], em que pares *aprendem* o perfil dos pares da rede na medida em que interagem com os pares vizinhos através de solicitações e processamento de consultas. Com isso, o modelo oferece um ambiente de colaboração que permite que as consultas sejam roteadas dinamicamente conforme perfil dos pares com os quais o par requisitante interagiu anteriormente. Experimentos sugerem que a qualidade da busca usando o modelo 6S é comparável ou mesmo superior àquela obtida por máquinas de busca centralizadas.

Diversos outros modelos de máquinas de busca P2P foram propostos. Por exemplo, Doulkeridis et al. (2006) utilizam a idéia de redes semânticas como alternativa para redução de custo de pesquisa, utilização de banda e latência. No modelo, pares com interesses similares são agrupados. Cada grupo possui pares especiais responsáveis por encaminhar consultas para grupos similares. PlanetP [Cuenca-Acuna et al. (2003)] apresenta-se como um modelo de busca por conteúdo em arquiteturas P2P não estruturadas em que pares possuem uma visão global da classificação dos

documentos da rede. Para tanto, o modelo utiliza o algoritmo de Fofoca (ou *Gossiping*) [Demers et al. (1987)], em que periodicamente os pares trocam mensagens entre si.

2.3 Comportamento Dinâmico dos Pares em Redes Par-a-Par

Alguns trabalhos propõem modelos de comportamento dinâmico de pares em redes P2P. Dunn et al. (2005) propõem um modelo que leva em consideração a *presença* dos pares na rede. Em seu modelo, a disponibilidade de um par atinge seu mínimo apenas quando o par estiver online e nenhum outro par estiver e seu máximo quando o par estiver online e todos os outros pares também estiverem. Os autores definem então a disponibilidade do serviço como uma média ponderada das disponibilidades dos pares. Foi observado que a disponibilidade de serviço da rede Kazaa é bem representada pelo modelo.

Diversos trabalhos avaliam o dinamismo dos pares a partir de *logs* de redes P2P reais. Ripeanu (2001) apresenta um estudo sobre a rede Gnutella, revelando algumas de suas características, como a distribuição de *links* entre os pares seguir Lei de Potência. Nesse estudo o autor destaca a alta disponibilidade de redes P2P não estruturadas – embora não tenham quantificado precisamente – o que está de acordo com a observação de outros autores [Pouwelse et al. (2005)]. Pouwelse et al. (2005) avaliam a rede P2P estruturada BitTorrent em conjunto com o ambiente de pesquisa Suprnova [1]. Dentre outras métricas, os autores avaliam a disponibilidade da rede. Como maior contribuição, os autores concluem que em sistemas P2P existe uma tensão entre disponibilidade e integridade de dados: o gerenciamento centralizado de arquivos em sistemas como BitTorrent beneficia a integridade dos dados, mas torna o sistema vulnerável a falhas. Por outro lado, a descentralização pode melhorar a disponibilidade de documentos, uma vez que há replicação, mas os arquivos são mais suscetíveis à corrupção.

Rhea et al. (2004) mostram que o comportamento dinâmico dos pares pode afetar consideravelmente o projeto e a avaliação de sistemas P2P. Por exemplo, os autores mostram que as atuais implementações de DHT não lidam bem com altas taxas de entrada e saída dos pares. Já Stutzbach e Rejaie (2006) concentram-se na análise da distribuição dos tempos de sessão dos pares, isto é, o intervalo de tempo em que o par permanece na rede até sua próxima saída. Os autores estudaram a entrada e saída dos pares em diferentes aplicações P2P de compartilhamento de arquivos e verificaram que grande parte das sessões dos usuários (tempo *online*) duravam apenas alguns minutos, enquanto outras duravam dias ou mesmo semanas. Os autores concluem que, diferentemente de estudos anteriores, a duração de sessões nas redes Gnutella e Kademlia [Maymounkov e Mazieres (2002)] é bem modelada por uma Log-Normal e a rede BitTorrent por uma Weibull.

No entanto, o comportamento dinâmico dos pares ainda foi pouco explorado em máquinas de busca P2P. A maioria das máquinas de busca P2P assume um ambiente confiável e colaborativo,

[1] http://suprnova.org/

não cobrindo devidamente aspectos de autonomia dos pares. Em especial, a busca por conteúdo sugere que os efeitos do comportamento dinâmico dos pares na eficácia da busca podem ser bem particulares nesse tipo de aplicação. De fato, Stutzbach e Rejaie (2006) mostram que o comportamento dinâmico dos pares é largamente dependente do tipo de aplicação P2P (por exemplo, aplicações P2P de compartilhamento de arquivos ou de distribuição de conteúdo).

2.4 Estratégias de Replicação de Conteúdo em Redes Par-a-Par

Lv et al. (2002) categorizam as estratégias de replicação em redes P2P como *estáticas* e *dinâmicas*.

Na replicação estática, cópias dos objetos (ou documentos) são distribuídas entre os pares conforme alguma regra definida *a priori*. Por exemplo, na replicação estática *uniforme*, cada objeto (ou documento) é replicado em uma mesma quantidade de pares, isto é, M/m pares, onde M corresponde ao número total de objetos (incluindo réplicas) e m é o número de objetos distintos [Lv et al. (2002)]. Os autores afirmam que essa é uma estratégia bastante primitiva, o que é razoável, uma vez que todos os objetos são tratados da mesma forma, mesmo alguns sendo mais populares do que outros. No entanto, a estratégia minimiza o número máximo de *hops* [2] necessários para encontrar um objeto.

Outras estratégias de replicação estática analisadas incluem a replicação proporcional e a replicação *Square-root*. Na replicação proporcional, objetos mais populares são replicados em mais pares, diretamente proporcional. A idéia dessa estratégia é que objetos mais populares sejam mais facilmente encontrados. O preço que se paga é um aumento no número de *hops* na busca dos objetos não tão populares. Os autores propõem a estratégia de replicação *Square-root*, que minimiza o número de *hops*. Nessa estratégia, objetos mais populares são também replicados em mais pares, mas o número de réplicas é proporcional à raiz quadrada da popularidade do objeto, medida pelo número de consultas ao objeto.

Já na replicação dinâmica, cópias dos documentos são distribuídas entre os pares na medida em que requisições aos documentos vão sendo realizadas. Por exemplo, na replicação dinâmica *própria* (*owner replication*), uma cópia do documento encontrado na rede é armazenada no par que fez a consulta, enquanto que na replicação dinâmica *de caminho* (*path replication*), cópias do documento são armazenadas em todos os pares por que a consulta foi repassada até se atingir o par que fez a consulta [Cohen e Shenker (2002)]. Estratégias de replicação dinâmica podem ser úteis em ambientes P2P que possuem pares com alto nível de autonomia.

[2] Distância entre o par requisitante e o par servidor do objeto

Capítulo 3

Metodologia

3.1 Descrição

Avaliamos o impacto do comportamento dinâmico dos pares na eficácia de máquinas de busca P2P a partir de um ambiente de simulação híbrido, composto por milhares de pares independentemente entrando e saindo da rede.

O ambiente de simulação implementa um modelo de máquina de busca P2P inspirado em modelos disponíveis na literatura [Bender et al. (2006); Wu et al. (2005); Cuenca-Acuna et al. (2003); Lu e Callan (2003)]. No entanto, esses modelos não consideram o comportamento dinâmico dos pares e não conhecemos trabalhos que caracterizem esse aspecto em ambientes P2P [1]. Dessa forma, nosso ambiente baseia-se em caracterizações de sistemas P2P reais de compartilhamento de arquivos [Stutzbach e Rejaie (2006)].

Além disso, utilizamos um processador de consultas real, também adotado em trabalhos anteriores [Almeida et al. (2007)], e cargas de trabalhos realistas, construídas a partir de coleções de teste reconhecidas (ver Seção 3.2). Para avaliar os resultados apresentados não apenas pelo modelo de máquina de busca P2P proposto como também pelo processador de consultas real adotado, quantificamos a qualidade dos resultados a partir de métricas usuais da área de Recuperação de Informação [Yilmaz e Aslam (2006)] (ver Seção 3.3).

Nessa direção, pretendemos capturar aspectos cruciais relacionados a eficácia de máquinas de busca P2P reais.

[1] À exceção do trabalho que publicamos abordando o assunto [Atalla et al. (2008)]

3.2 Coleções de Teste

3.2.1 Definição

Uma coleção de teste, ou *coleção*, refere-se à carga de trabalho formada por documentos, consultas e julgamentos de relevância. Em nosso estudo, utilizamos duas coleções, a WBR [Calado (1999)] e a TREC-8 [Voorhees e Harman (1999)] (ou TREC, no decorrer do texto). Obtivemos as coleções já formatadas (e indexadas) pelo laboratório LATIN [2]. Quando mencionarmos WBR e TREC ao longo do texto, estaremos nos referindo às coleções de teste formatadas, a menos que seja explicitamente mencionado o contrário.

3.2.2 Termos

Um *termo* é uma seqüência de caracteres alfanuméricos e acentuados. Não há distinção entre letras maiúsculas e minúsculas.

Tipos especiais de termo, os *conectores* (ou *stopwords*, como conjunções e preposições) podem ser ou não excluídos durante a formatação da coleção. A WBR inclui os conectores enquanto que a TREC não os inclui.

3.2.3 Documentos

Um *documento* é uma seqüência de termos como um texto de jornal ou uma página na Web.

A WBR foi construída a partir de documentos coletados da Web brasileira em novembro de 1999 pela máquina de busca TodoBR.

Por sua vez, a TREC é formada pelos documentos da trilha *ad-hoc task* da conferência TREC-8. Esses documentos foram utilizados pela TREC-8 para avaliação dos algoritmos de Recuperação de Informação submetidos à conferência. O conjunto de documentos inclui notícias de jornais e do governo dos EUA, como o *Los Angeles Times* e *Federal Register*.

3.2.4 Consultas

Uma *consulta* descreve uma necessidade de informação do usuário através de um conjunto de termos.

A WBR oferece 50 diferentes necessidades de informação selecionadas por especialistas dentre as mais freqüentes submetidas por usuários à máquina de busca TodoBR no mês de novembro de 1999. Cada necessidade de informação selecionada é formada por dois campos: o primeiro corresponde ao texto submetido pelo usuário à máquina de busca; o segundo corresponde a uma descrição alternativa elaborada por especialistas (ver Figura 3.1). Na WBR, uma consulta corresponde aos termos encontrados no *primeiro campo* de descrição do tópico, incluindo conectores.

[2] Laboratório para o Tratamento da INformação, DCC/UFMG

Figura 3.1: Necessidades de informação representadas pelas consultas 23 a 27 – WBR

```
...
jornal - sites de jornais ou com listas de jornais online
legião urbana - sites sobre o grupo musical "Legião Urbana"
linux - sites dedicados ao sistema Linux
literatura e novas tecnologias - informação sobre o uso ou influência de novas tecnologias na literatura
machado de assis - sites sobre o escritor "Machado de Assis"
...
```

Para se obter o conjunto de consultas da TREC, especialistas sugerem possíveis necessidades de informação dos usuários (ou *tópicos*). Em seguida, inspecionam manualmente uma amostra de documentos da coleção de forma a estimar o número provável de documentos relevantes por tópico. Os 50 tópicos com maior número provável de documentos relevantes são escolhidos e identificados por números que variam de 401 a 450 [3]. Além do número de identificação, cada tópico é formado por outros três campos: um título, uma descrição sucinta e uma explicação sobre o que seria relevante para o tópico (ver Figura 3.2). Na TREC, uma consulta corresponde aos termos encontrados no campo de descrição do tópico, excluídos os conectores.

Figura 3.2: Necessidade de informação representada pela consulta 12 – TREC

```
<top>

<num> Number: 412
<title> airport security

<desc> Description:
What security measures are in effect or are proposed
to go into effect in airports?

<narr> Narrative:
A relevant document could identify a specific airport
and describe the security measures already in effect
or proposed for use at that airport. Relevant items
could also describe a failure of security that was
cited as a contributing cause of a tragedy which came
to pass or which was later averted. Comparisons between
and among airports based on the effectiveness of the
security of each are also relevant.

<top>
```

3.2.5 Julgamentos de Relevância

Cada consulta possui uma lista de documentos da coleção considerados relevantes. O *julgamento de relevância* de documentos em ambas as coleções foi feito pelo método de *Pooling* [Voorhees e Harman (1999)]. Nesse método, uma seleção de documentos possivelmente relevantes é formada a partir de uma amostra dos documentos retornados por diversos algoritmos de recuperação de informação [4]. Cada um dos documentos da seleção é então julgado manualmente por especialistas.

[3] Na TREC-1 os tópicos foram identificados de 51 a 100, na TREC-2 de 101 a 150 e dessa forma até a TREC-8.
[4] Maiores detalhes sobre os algoritmos utilizados em Voorhees e Harman (1999) e Calado (1999).

Em ambas as coleções, a amostra de cada algoritmo corresponde aos primeiros documentos retornados, isto é, os documentos que o algoritmo considera como mais prováveis de serem relevantes. Os documentos que não foram selecionados pelo método não são julgados e são considerados não relevantes.

Na TREC, a avaliação de relevância é binária, isto é, um documento é julgado como relevante ou não por especialistas. Por outro lado, na WBR o julgamento de relevância possui cinco diferentes pontuações: -1, se o documento for inválido (por exemplo, uma página de erro); 0, se o documento não for relevante para a consulta; 1, se o documento contiver apenas informações relacionadas ou *links*; 2, se o documento apresentar apenas informações relacionadas ou *links*, mas que referenciem documentos relevantes; e 3, se o documento contiver informações relevantes. De forma a padronizar os julgamentos de relevância de ambas coleções, supomos que todo documento da WBR que tiver uma pontuação maior que 0 foi julgado como relevante. Caso contrário, supomos que foi julgado como não relevante.

A Tabela 3.1 sumariza as coleções avaliadas.

Tabela 3.1: Sumário das coleções de teste

Coleção (c)	**# de documentos** (D_c)	**# de termos únicos**	**# de consultas**	**Documentos relevantes por consulta**		
				# máximo	**# mínimo**	**# médio**
WBR	5.751.296	2.669.965	50	96	0	35
TREC-8	528.155	737.833	50	347	6	95

3.3 Métricas de Eficácia de Máquinas de Busca

3.3.1 Definição

Utilizamos o termo *eficácia* para designar a qualidade das respostas de uma máquina de busca a consultas de usuários. Dada uma consulta, uma *métrica de eficácia* estima a similaridade entre o conjunto de documentos retornados pela máquina de busca (em resposta a essa consulta) e o conjunto de documentos previamente julgados como relevantes por especialistas [Baeza-Yates e Ribeiro-Neto (1999)].

3.3.2 Precisão e Revocação

A Precisão (*Precision*) é uma métrica de eficácia relacionada à exatidão das respostas obtidas pela máquina de busca. Essa métrica estima a capacidade que o sistema possui de retornar apenas documentos relevantes [Voorhees e Harman (1999)]. Em outras palavras, corresponde à fração de documentos recuperados que são relevantes. Dados a lista K dos primeiros documentos únicos retornados pela máquina de busca em resposta a uma consulta, isto é, em que as cópias dos documentos que aparecem depois da primeira ocorrência são desconsideradas, e o julgamento binário de relevância x_i para o i-ésimo documento dessa lista, definimos formalmente [Kishida

(2005)] a Precisão P_K como

$$P_K = \frac{1}{|K|} \sum_{i=1}^{|K|} x_i \tag{3.1}$$

$P_K \in [0,1]$, pois $\sum_{i=1}^{|K|} x_i \leq |K|$.

A Revocação (*Recall*) é uma métrica de eficácia relacionada à abrangência das respostas obtidas pela máquina de busca. Essa métrica estima a capacidade que o sistema possui de retornar todos os documentos relevantes [Voorhees e Harman (1999)]. Em outras palavras, corresponde à fração de documentos relevantes que são recuperados [Baeza-Yates e Ribeiro-Neto (1999)]. Dados a lista K e o julgamento binário x_i já definidos e o conjunto R dos documentos relevantes da coleção por consulta, definimos formalmente a Revocação R_K como

$$R_K = \frac{1}{|R|} \sum_{i=1}^{|K|} x_i \tag{3.2}$$

$R_K \in [0,1]$, pois $\sum_{i=1}^{|K|} x_i \leq |R|$ (R é conhecido *a priori*).

Precisão e Revocação são métricas que não podem ser utilizadas isoladamente para avaliar a eficácia de máquinas de busca [Rijsbergen (1979)]. Note que uma precisão $P_K = 1$ significa que todos os K primeiros documentos únicos recuperados são relevantes, mas não diz nada sobre a fração de documentos relevantes que foram recuperados. Analogamente, uma revocação $R_K = 1$ significa que todos os documentos relevantes encontram-se entre os K primeiros documentos únicos recuperados, mas não diz nada sobre a quantidade de documentos irrelevantes recuperados.

3.3.3 Precisão Média

De forma a sumarizar a eficácia em uma única medida, foram propostas outras métricas que tentam combinar Precisão e Revocação. Dentre as métricas existentes, podemos citar a Precisão Média, a Precisão-R e a Média Harmônica [Baeza-Yates e Ribeiro-Neto (1999)]. Por ser uma métrica amplamente utilizada [Yilmaz e Aslam (2006)], adotamos a Precisão Média em nossa avaliação.

Dados o conjunto R dos documentos relevantes da coleção por consulta, a lista A dos primeiros documentos únicos retornados pela máquina de busca em resposta a uma consulta, isto é, em que as cópias dos documentos que aparecem depois da primeira ocorrência são desconsideradas, a precisão P_K dada pela Equação 3.1, $|K| \leq |A|$, e o julgamento binário de relevância x_K do $|K|$-ésimo documento de A, definimos formalmente [Kishida (2005)] a Precisão Média AP como

$$AP = \frac{1}{|R|} \sum_{K=1}^{|A|} x_K P_K \tag{3.3}$$

$AP \in [0,1]$, pois $P_K \in [0,1]$ e $\sum_{i=1}^{|A|} x_i \leq |R|$ (R é conhecido *a priori*).

Isto é, AP fornece a fração de documentos relevantes, ponderados pela sua posição na lista de documentos recuperados. AP pretende incorporar a exatidão da Precisão e a abrangência da Revocação sem no entanto desprezar a ordem dos documentos recuperados.

Também estimamos a média populacional de AP. Para tanto, agregamos os valores de AP sobre cada consulta em uma única métrica, a média aritmética de AP, ou MAP (*Mean Average Precision*).

Capítulo 4

Máquinas de Busca Centralizadas: a Linha de Base

4.1 Definição

Definimos uma *máquina de busca centralizada* como uma máquina de busca composta pelos módulos de coleta, indexação e processamento de consultas, os quais são gerenciados por uma única entidade.

A definição de máquina de busca centralizada não considera aspectos físicos – por exemplo, se a máquina de busca é implementada com apenas um servidor ou diversos servidores fisicamente distribuídos – mas apenas o fato de que uma única entidade tem todo o controle e responsabilidade sobre a máquina de busca.

4.2 Modelo

A Tabela 4.1 sumariza os parâmetros do modelo.

Tabela 4.1: Parâmetros do modelo de máquina de busca centralizada

Parâmetro	**Descrição**
τ_c	Limiar de similaridade para a coleção de teste c (c=*wbr*, *trec*)
t_q	Tempo médio entre chegadas das consultas à máquina de busca

4.2.1 Coleta

Durante a *coleta*, a máquina de busca centralizada define o conjunto de documentos sobre o qual as consultas serão realizadas. Na Web, esse conjunto é obtido por coletores baseados em *hyperlinks*, a partir dos quais os coletores recuperam uma amostra dos documentos disponíveis na Web.

Em nosso modelo, supomos que os documentos das coleções de teste são resultado de uma coleta. Dessa forma, avaliamos *imagens* de diferentes tipos de repositório de documentos: na WBR, os documentos correspondem a uma imagem da Web brasileira, enquanto que na TREC os documentos correspondem a uma imagem de publicações em revistas e jornais (ver Seção 3.2 para uma descrição mais detalhada).

4.2.2 Indexação

A *indexação* envolve a formatação dos documentos selecionados no módulo de coleta em alguma estrutura eficiente para o processamento das consultas. Também são extraídas informações relevantes dos documentos, como por exemplo o conjunto de termos e estatísticas sobre as ocorrências de termos em documentos.

Em nosso modelo, os documentos são estruturados em uma *lista invertida*. Além disso, o *peso* (importância) do termo em cada um de seus documentos é estimado a partir de estatísticas TF-IDF, uma abordagem tradicional em Recuperação de Informação [Baeza-Yates e Ribeiro-Neto (1999)]. TF é o componente do peso que representa a freqüência do termo no documento e IDF está relacionado ao número de documentos contendo o termo.

O componente TF de um termo i em um documento j é dado pela fórmula

$$tf\,(i,j) = \begin{cases} \frac{freq(i,j)}{freqmax(l,j)}, & se\ freqmax\,(l,j) > 0 \\ 0, & c.c. \end{cases} \tag{4.1}$$

onde $freq\,(i,j)$ é a freqüência (i.e., número de vezes) em que i ocorre em j e $freqmax\,(l,j) = freq\,(l,j) \geq freq\,(k,j)$ é a freqüência do termo l que ocorre mais vezes em j, dentre qualquer outro termo k de j. O componente TF tenta capturar a importância do termo no documento: quanto maior o número de ocorrências do termo no documento, mais representativo é o termo para o documento e portanto maior é o TF.

Por sua vez, o componente IDF de um termo i em um conjunto de documentos D é dado por

$$idf\,(i,D) = \begin{cases} log\,(D/d_i)\,, & se\ D \geq d_i > 0 \\ 0, & c.c. \end{cases} \tag{4.2}$$

onde d_i é o número de documentos com o termo i. O componente IDF tenta capturar a importância do termo no conjunto total de documentos existentes: quanto menor o número de documentos contendo o termo, mais representativo é o termo para todos os documentos e portanto maior é o IDF.

O peso de um termo i em um documento j é então dado pelo produto das Equações 4.1 e 4.2, isto é,

$$W_{ij} = tf\,(i,j) \times idf\,(i,D) \tag{4.3}$$

Em particular, como a importância dos termos de uma consulta não varia de um documento para outro, o peso de um termo i em uma consulta q terá um valor constante, isto é,

$$W_{iq} = \begin{cases} 1, & se\ i\ ocorre\ em\ q \\ 0, & c.c. \end{cases} \tag{4.4}$$

4.2.3 Processamento de Consultas

O *processador de consultas* é o módulo responsável por determinar o conjunto de documentos capazes de atender a uma necessidade de informação do usuário. O módulo deve apresentar o conjunto de documentos ordenados por relevância.

Em nosso modelo, o processador de consultas utiliza o *Modelo de Espaço Vetorial* [Baeza-Yates e Ribeiro-Neto (1999)]. Nesse modelo, a consulta q e o documento j são representados por vetores s-dimensionais, onde s é o total de termos da coleção e cada componente dos vetores corresponde a um peso dado pelas Equações 4.3 e 4.4 [1].

O Modelo de Espaço Vetorial avalia a *similaridade* entre a consulta do usuário, q, e o documento j a partir do cosseno do ângulo entre os vetores, isto é,

$$sim\,(d_j,q) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{s} W_{ij} \times W_{iq}}{\sqrt{\sum_{i=1}^{s} W_{ij}^2} \times \sqrt{\sum_{i=1}^{s} W_{iq}^2}} = \frac{\sum_{i=1}^{|q|} W_{ij}}{\sqrt{\sum_{i=1}^{s} W_{ij}^2} \times \sqrt{|q|}} \tag{4.5}$$

Quanto maior a similaridade, melhor será a posição do documento na resposta. Em nosso modelo, supomos que documentos com similaridade abaixo de um limiar τ_c não são retornados para o usuário.

Além disso, consideramos que os usuários submetem consultas de forma independente. Dessa forma, podemos modelar o tempo entre consultas com uma distribuição exponencial com média t_q [Jain (1991)]. Em nosso modelo, consultas são realizadas por pares escolhidos de forma aleatória.

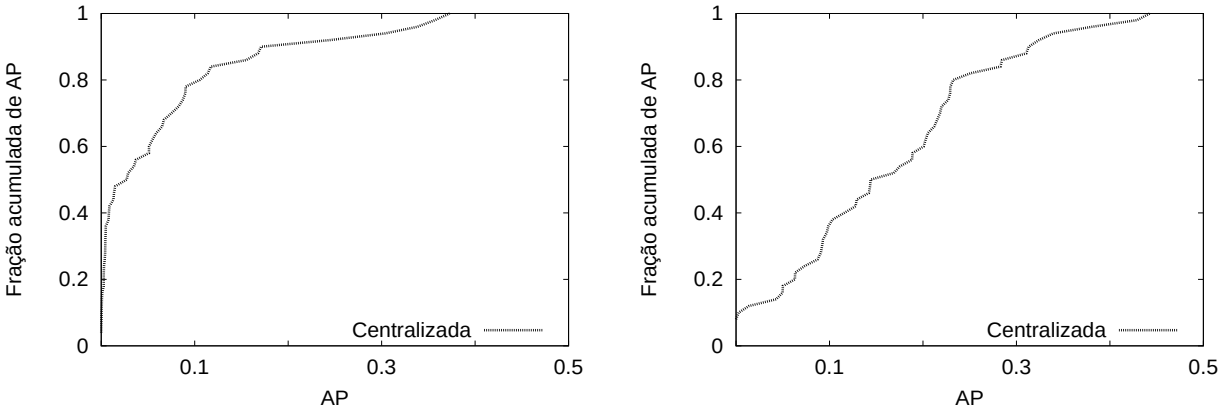
Algumas considerações adicionais: o processador supõe que todos os termos da consulta estão indexados. Além disso, o processador utiliza apenas os termos da consulta, isto é, não faz expansão de consulta ou busca por sinônimos. Finalmente, não existe restrição quanto ao número de termos da consulta presentes no documento; isto é, a similaridade entre a consulta e os documentos é calculada para todos os documentos que tenham pelo menos um termo da consulta.

[1] Estimamos os pesos a partir de estatísticas TF-IDF, mas qualquer peso não binário pode ser utilizado no Modelo de Espaço Vetorial.

4.3 Experimentos

O modelo de máquina de busca adotado apresenta uma eficácia MAP de $0,069$ para a TREC e $0,166$ para a WBR. Esses valores foram obtidos após a normalização da lista final de resultados para o mesmo tamanho médio em ambas coleções. Para tanto, adotamos um limiar de similaridade τ_c específico para cada coleção. Em particular, $\tau_{trec} = 0,10$ e $\tau_{wbr} = 0,32$, de forma que as respostas às consultas possuíssem em torno de 1.000 documentos em média. Essa média corresponde ao tamanho máximo da lista de documentos retornada pelo Google [Gopalakrishnan et al. (2006)].

Note que um resultado de 100% somente seria alcançado se a resposta apresentasse *todos* e *apenas* documentos julgados como relevantes (isto é, máxima Revocação e Precisão, respectivamente – ver Seção 3.3).



(a) TREC (b) WBR

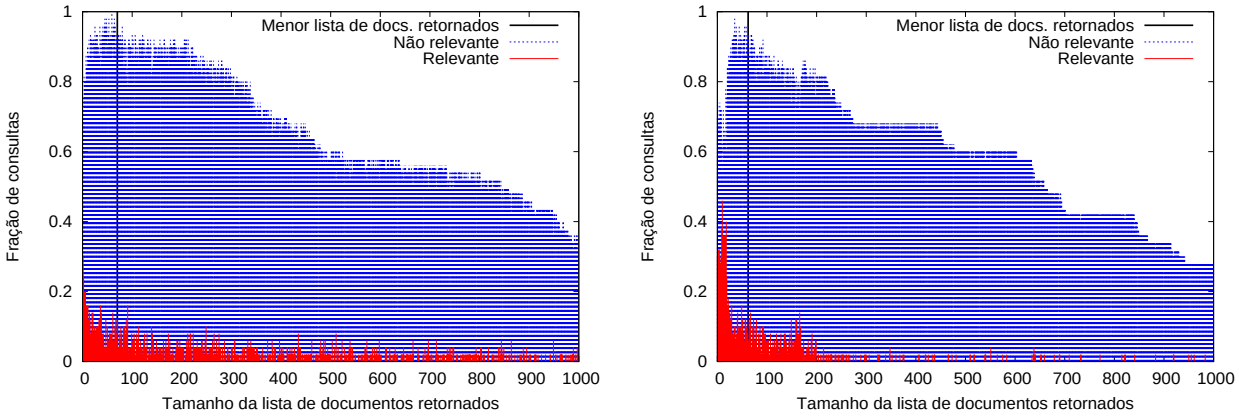
Figura 4.1: Eficácia da máquina de busca centralizada (linha de base)

Embora a métrica MAP seja significativa, também é importante avaliar a degradação das consultas individualmente. A Figura 4.1 mostra a distribuição acumulada da qualidade (i.e., CDF para a métrica AP) de resultados individuais de consultas, para o modelo de linha de base proposto. De fato, as curvas mostram uma grande variabilidade nos valores individuais de AP.

Os resultados deixam claro que a eficácia da máquina de busca centralizada é maior quando avaliada sobre a coleção de teste WBR. De fato, MAP revela que a qualidade dos resultados na WBR é de quase 17% em média, contra cerca de 7% na TREC. Além disso, a Figura 4.1 mostra que, enquanto na WBR uma quantidade significativa de consultas (20%) apresenta respostas com qualidade acima de 25%, na TREC esse número não chega a 12%. Essa diferença entre as coleções de teste pode ser conseqüência dos diferentes tipos de algoritmo de recuperação de informação adotados no julgamento de relevância de seus documentos (ver Seção 3.2), o que pode ter uma forte influência sobre os resultados: na coleção de teste da WBR, utilizou-se um conjunto

bem definido de algoritmos de recuperação de informação também baseados no modelo de espaço vetorial; já na TREC utilizou-se algoritmos baseados em diferentes estratégias e modelos de recuperação de informação.

A Figura 4.2 mostra a concentração de documentos relevantes e não relevantes retornados para as consultas feitas à máquina de busca centralizada. As curvas ilustram a chance de o i-ésimo documento retornado pela máquina de busca ser relevante ou não relevante. Embora em ambas coleções os documentos relevantes estejam mais concentrados entre as primeiras posições, na WBR a concentração é ligeiramente maior, o que contribui para a significativa diferença entre os resultados das coleções.



(a) TREC (b) WBR

Figura 4.2: Relevância das respostas da máquina de busca centralizada

A Figura 4.2 também apresenta um marco vertical em $i = m$ ($m = 70$ na TREC e $m = 62$ na WBR), que indica o número mínimo de documentos retornados por uma consulta. Todas as consultas retornam pelo menos $i = m$ documentos, o que torna as curvas de documentos relevantes e não relevantes complementares até esse marco. Após esse marco, no entanto, o número total de documentos retornados pelas consultas tende a diminuir.

4.4 Desafios

Máquinas de busca centralizadas precisam lidar constantemente com o rápido e explosivo crescimento da Web. Nesse cenário, arquiteturas escaláveis e com ampla cobertura da Web podem ter um impacto significativo na qualidade dos resultados oferecidos aos usuários [Bender et al. (2006); Doulkeridis et al. (2006); Brin e Page (1998)].

Diversos fatores sugerem o alto custo de uma única entidade ser responsável por todo o processo de busca. De fato, Gulli e Signorini (2005) estimaram que a quantidade de documentos que podem ser indexados por máquinas de busca centralizadas – a *Web indexável* – superou 11

bilhões de documentos em 2005. Esses documentos precisam ser localizados, recuperados, indexados e armazenados na menor quantidade de tempo e recursos possível. Para tanto, máquinas de busca centralizadas precisam não apenas dispor de uma enorme capacidade de armazenamento, mas também de tempo e largura de banda. Além disso, a coleta deve ainda ser feita de forma equilibrada para não comprometer a carga dos servidores, documentos da coleção precisam ser constantemente atualizados dado o dinamismo da Web, ao mesmo tempo em que o tempo de resposta exigido pelos usuários da máquina de busca centralizada deve ser respeitado.

O tamanho da Web indexável, embora bastante expressivo, corresponde apenas aos documentos da *superfície* da Web. A grande maioria dos documentos, no entanto, não é indexável por máquinas de busca centralizadas. De acordo com Bergman (2001), o conjunto de documentos pertencentes à chamada *Web invisível* poderia chegar a 100 bilhões ainda em 2001. Esse conjunto inclui documentos que não podem ser recuperados devido a restrições técnicas ou mesmo por critérios de relevância da entidade responsável [Lewandowski e Mayr (2006)]. Documentos que podem não ser recuperados devido a restrições técnicas incluem páginas desconectadas (sem *hyperlinks* de entrada ou saída), documentos com conteúdo com pouca (ou nenhuma) informação textual (e.g. páginas em *Flash* [Millward Brown (2008)]) e páginas geradas dinamicamente, isto é, aquelas que estão por trás de formulários e envolvem consultas a bancos de dados. Documentos que não são recuperados por critérios de relevância da entidade responsável pela máquina de busca podem incluir páginas manipuladas por *Web spammers* [Gyongyi e Garcia-Molina (2005)] ou mesmo páginas que vão contra interesses políticos e comerciais. Nesse último caso, o monopólio da informação pode comprometer a relevância dos resultados apresentados ao usuário com informações tendenciosas.

4.5 Alternativas

Apresentamos diferentes estratégias apontadas como alternativa para máquinas de busca centralizadas. De um modo geral, percebe-se que a *colaboração* dos usuários tem sido fundamental para se atenuar os desafios enfrentados por esse tipo de arquitetura.

Nelson et al. (2006) propõem um protocolo que pretende ampliar a cobertura da Web de máquinas de busca centralizadas. Nesse protocolo, os sítios da Web devem disponibilizar meta-dados sobre seus documentos para que possam ser localizados pelos coletores da máquina de busca. Esse tipo de estratégia pretende atenuar as restrições técnicas, relacionadas à Web invisível, listadas anteriormente. *Sitemaps* [Google, Inc., Yahoo, Inc. e Microsoft Corporation (2008)] é um outro exemplo de protocolo, formalmente adotado pelas grandes máquinas de busca atuais.

Uma outra estratégia é conhecida como *busca aliada* (*federated search*) [Luo Si (2005)], que lida com páginas da Web invisível que são geradas dinamicamente, isto é, resultado de buscas em bancos de dados. Nessa estratégia, dada uma necessidade do usuário, a máquina de busca é responsável por submeter a consulta apropriada a cada um de diversos bancos de dados disponíveis

na Web, agregar os resultados obtidos e apresentá-los ao usuário. A busca aliada apresenta-se também como uma solução escalável, já que os módulos de coleta, indexação e processamento de consultas são realizados pelos próprios sítios responsáveis pelos bancos de dados. Porém, esses sítios devem informar *a priori* à máquina de busca detalhes de acesso a seus documentos.

Redes Par-a-Par (P2P) têm sido apontadas como uma arquitetura alternativa para máquinas de busca na Web [Bender et al. (2006); Wu et al. (2005); Cuenca-Acuna et al. (2003); Lu e Callan (2003)]. Nesse tipo de rede, não uma única entidade é responsável por processar consultas dos usuários, além de localizar, coletar, indexar e armazenar enormes quantidades de documentos; cada par é responsável por uma fração mínima de todo o processo de busca. Além disso, os pares não precisam rastrear toda a superfície da Web para indexar documentos e coletores não precisam centralizar documentos em um único repositório – o que pode consumir muito tempo, espaço e recursos de banda. Ao contrário, uma quantidade mínima de coleta é necessária, pois os pares já possuem seus próprios documentos.

Redes P2P podem também contribuir para ampliar a superfície da Web a partir da colaboração dos próprios usuários. Nesse cenário, páginas antes desprezadas por máquinas de busca centralizadas, seja por limitações técnicas ou por critérios particulares de relevância, podem se tornar parte da superfície da Web, já que os usuários têm autonomia para indexá-las à sua maneira.

Capítulo 5

Máquinas de Busca Par-a-Par

5.1 Definição

Uma *máquina de busca Par-a-Par (P2P)* é uma máquina de busca cujos módulos de coleta, indexação e processamento de consultas estão distribuídos em pares em uma rede P2P. A Figura 5.1 ilustra a arquitetura de máquinas de busca P2P, composta por duas camadas: a rede P2P e as máquinas de busca locais executadas sobre ela.

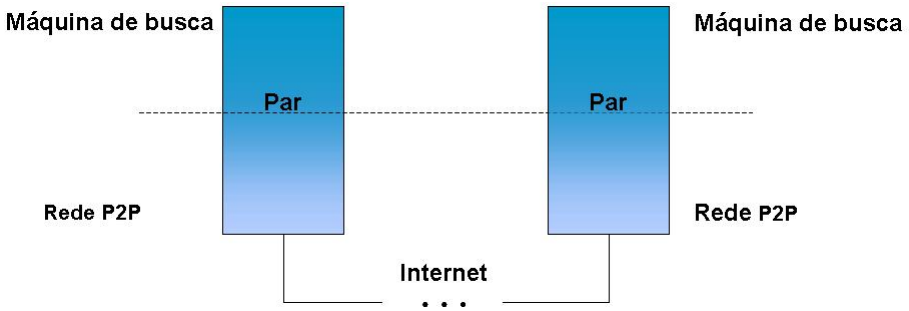


Figura 5.1: Estrutura de uma máquina de busca P2P

Cada par possui sua própria *máquina de busca centralizada* (ver Capítulo 4), capaz de coletar, indexar documentos e processar consultas realizadas localmente por seus usuários de forma independente. Além disso, quando o usuário conecta-se à rede P2P a partir de algum par, pode realizar consultas em outros pares, pesquisando remotamente por documentos com conteúdo relevante à consulta submetida, ampliando assim o espectro de busca. Em troca, o par do usuário deve processar consultas de outros pares, retornando informação sobre documentos locais, cujo conteúdo deve também estar disponível para acesso.

Em nossa definição, não fazemos nenhuma consideração sobre aspectos específicos de arquite-

tura de rede P2P – por exemplo, se a rede P2P é estruturada ou não estruturada. Dessa forma, supomos que *(1)* pares são sempre visíveis entre si, *(2)* todo conteúdo dos pares encontra-se disponível na Web e é sempre localizado e *(3)* o roteamento de consultas e respostas nunca falha. Por último, como a qualidade das respostas da máquina de busca P2P, e não o desempenho, é a métrica de interesse, desconsideramos os tempos de consulta e de resposta.

5.2 Modelo

A Tabela 5.1 sumariza os parâmetros do modelo. Conforme definição, o modelo de máquina de busca P2P *estende* o modelo de máquina de busca centralizada. As próximas subseções descrevem as principais considerações e componentes do modelo.

Tabela 5.1: Parâmetros do modelo de máquina de busca P2P

Parâmetro	**Descrição**
τ_c	Limiar de similaridade para a coleção de teste c (c=*wbr*, *trec*)
t_q	Tempo médio entre chegadas das consultas à máquina de busca
n_c	# de pares da coleção de teste c (c=*wbr*, *trec*)

Nosso modelo pretende capturar inicialmente os efeitos da descentralização de máquinas de busca. Em particular, avaliamos o impacto da falta de conhecimento dos pares sobre os documentos da rede e da forma como os documentos estão distribuídos ao longo da rede conforme interesses dos usuários. Em seguida, aprimoramos o modelo para capturar o comportamento naturalmente dinâmico dos pares de uma rede P2P.

5.2.1 Passos da Busca

A Figura 5.2 ilustra o modelo de máquina de busca P2P adotado: submissão de uma consulta de usuário à máquina de busca P2P; descoberta dos v pares mais promissores (local e/ou remotos), $v \leq n_c$; processamento da consulta em cada um dos v pares; fusão e apresentação dos resultados.

Após o par receber uma consulta submetida pelo usuário, os pares mais promissores para responder à consulta são obtidos pelo *Seletor de Pares*, que obtém os pares que satisfazem os critérios de busca, e pelo *Ordenador de pares*, que ordena os pares que satisfazem os critérios de busca por sua similaridade com a consulta. Em seguida, a consulta é *encaminhada* para os pares selecionados (local e/ou remotos). Cada par então busca os documentos mais relevantes para a consulta no conjunto de documentos local e responde com uma lista ordenada de documentos. Finalmente, os resultados de cada par são fundidos em uma única lista no par requisitante e apresentada ao usuário como um conjunto de *links* ordenados por relevância.

Para determinar os pares mais promissores para processar uma consulta do usuário, o Seletor e o Ordenador de Pares podem aprender o conteúdo de cada par submetendo periodicamente consultas sintéticas à rede P2P e observando os documentos retornados. Essa técnica é conhecida

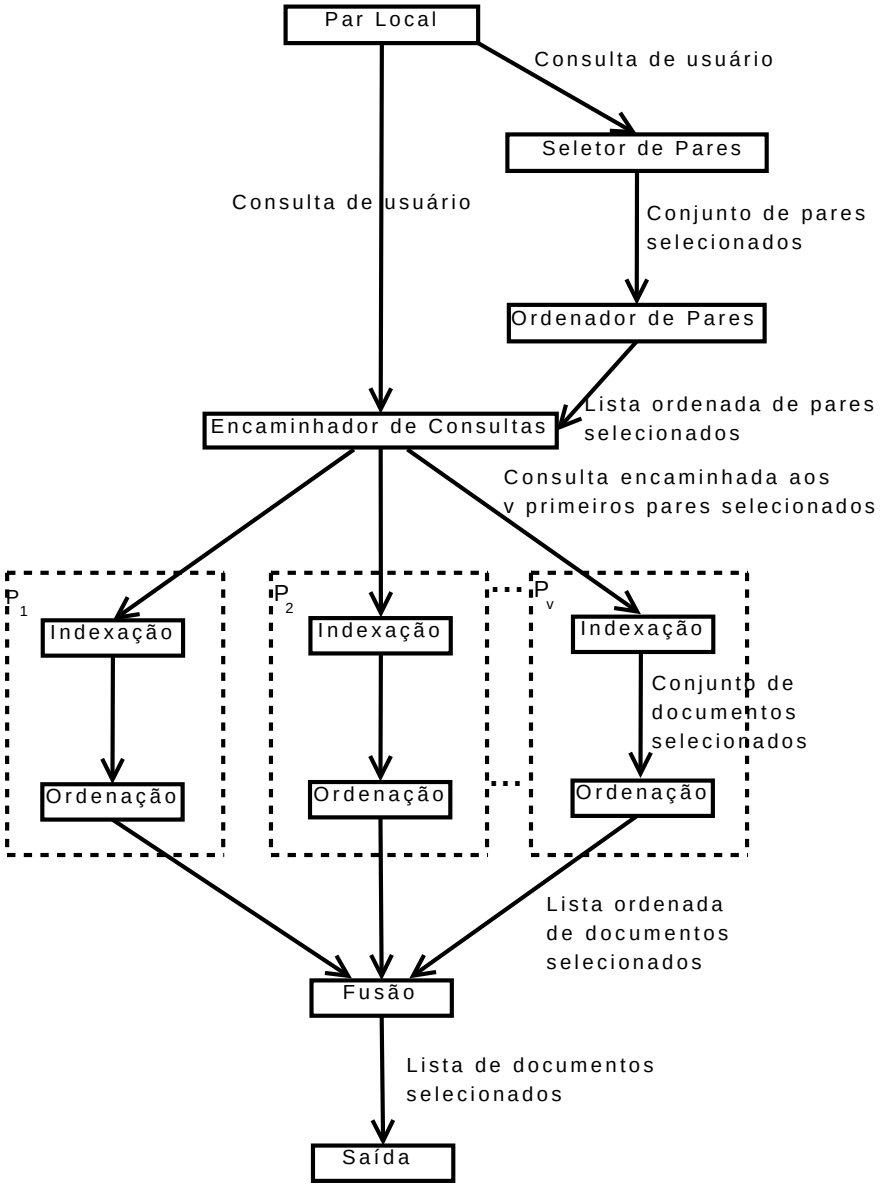


Figura 5.2: Passos de uma consulta em uma máquina de busca P2P

como *amostragem baseada em consultas* [Callan e Connell (2001)] e requer apenas a execução de consultas e recuperação de documentos. No momento em que o usuário submeter uma consulta, supomos que os pares mais promissores já tenham sido obtidos via amostragem.

Alguns autores [Bender et al. (2005)] sugerem considerar como pares promissores apenas aqueles que tenham todos os termos da consulta, ainda que cada termo ocorra em diferentes documentos. No entanto, com o objetivo de obter a máxima eficácia do processador de consultas, optamos por relaxar essa restrição ao supor que um par é promissor se tiver pelo menos um termo da consulta. De fato, experimentos confirmaram que se não relaxarmos tal restrição,

alguns documentos relevantes podem ser excluídos das respostas obtidas pela máquina de busca P2P.

A etapa de Fusão dos resultados envolve a escolha dos documentos mais relevantes retornados por todos os pares promissores. No entanto, cada par possui sua própria máquina de busca centralizada e portanto seus próprios critérios de ordenação dos resultados. Em nosso modelo, todo par possui exatamente o mesmo processador de consultas (ver Seção 4.2.3), mas diferentes conjuntos de documentos, o que pode levar um mesmo documento a ser ordenado de modo completamente diferente por cada par.

Dessa forma, um limiar de similaridade justo deve ser utilizado pela máquina de busca P2P na fusão das listas de documentos retornadas por cada par. Para fins de comparação com a linha de base, adotamos o mesmo limiar τ_c utilizado no modelo de máquina de busca centralizada. Sem perda de generalidade, em nosso modelo podemos supor que os próprios pares retornam uma lista de documentos com limiar acima de τ_c, já que: *(i)* nosso foco é a eficácia, e não o desempenho, de máquinas de busca P2P; e *(ii)* na fusão das listas retornadas por cada par os documentos abaixo do limiar também seriam excluídos.

5.2.2 Níveis de Conhecimento dos Pares

Quando um par possui muito conhecimento sobre a relevância de seus documentos na rede P2P, espera-se que a ordem de seus documentos na lista de resposta aproxime-se daquela obtida em máquinas de busca centralizadas. No entanto, em redes P2P, os pares tendem a se comportar de forma autônoma e independente, podendo ser custoso para um par manter-se atualizado sobre a relevância de seus documentos na rede.

De forma a capturar a eficácia de máquinas de busca P2P, limitamos nossa análise em níveis extremos de conhecimento.

De um lado, em nosso modelo (otimista) de máquina de busca P2P com *alto nível de conhecimento dos pares sobre documentos da rede* (AC), supomos que os pares divulgam informações sobre seus documentos para outros pares, de forma que cada par mantém-se atualizado sobre estatísticas de documentos na rede sem depender de serviços centralizados. Os algoritmos de *fofoca* (ou *Gossiping*) [Demers et al. (1987)] e de *amostragem baseada em consultas* [Callan e Connell (2001)] são soluções propostas na literatura capazes de construir e manter tal ambiente. Nesse nível, o processamento de consultas e a fusão dos resultados, assim como a seleção e a ordenação de pares, são executados como se cada par tivesse pleno acesso às estatísticas sobre os documentos da rede.

De outro lado, em nosso modelo (pessimista) de máquina de busca P2P com *baixo nível de conhecimento dos pares sobre documentos da rede* (BC), além de utilizar seu próprio processador de consultas, cada par processa consultas utilizando apenas conhecimento local. Em outras palavras, nesse ambiente P2P não se espera que os pares recuperem da rede estatísticas sobre

seus documentos. Nesse nível, selecionamos e classificamos os pares da mesma maneira que no nível AC, mas restringimos os processadores de consulta a usar apenas informações presentes no próprio par e, na fusão dos resultados, apenas as similaridades dos documentos informados pelos pares pesquisados.

5.3 Resultados de Simulação

Nesta seção estudamos a eficácia de máquinas de busca P2P em relação a máquinas de busca centralizadas (a linha de base). Para avaliar o modelo, desenvolvemos um ambiente de simulação híbrido em C++. A escolha da linguagem está ligada a aspectos de desempenho e escalabilidade. Utilizamos a biblioteca de simulação SimPack [Fishwick (1992)] para controle de eventos e um processador de consultas baseado em modelo vetorial adotado na literatura [Almeida et al. (2007)].

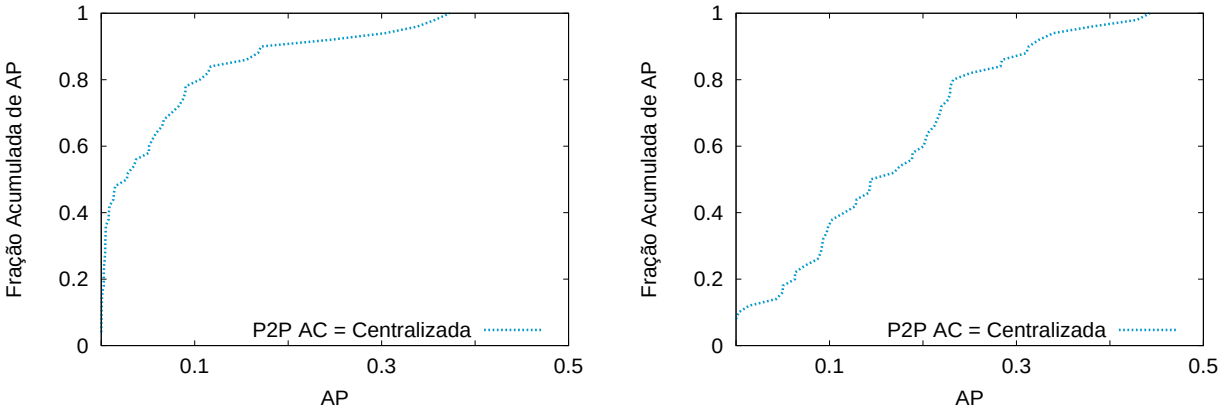
A Tabela 5.2 mostra os parâmetros usados no ambiente de simulação.

Tabela 5.2: Configuração da máquina de busca P2P

n_{trec}	n_{wbr}	τ_{trec}	τ_{wbr}	t_q
880	110.912	0,10	0,32	100

Avaliamos ambas coleções de teste, TREC e WBR, analisando o impacto dos cenários com altos e baixos níveis de conhecimento dos pares (i.e., AC e BC, respectivamente) na eficácia do modelo de máquina de busca P2P. De forma análoga aos experimentos realizados para o modelo de máquina de busca centralizada, apresentamos os resultados em uma distribuição acumulada dos valores de AP de consultas individuais. Cada AP resulta de uma média de 50 simulações.

Devido a diferenças significativas entre as coleções (ver Tabela 3.1), atribuímos a cada coleção c uma distribuição de documentos diferente. Na WBR, cada documento está vinculado a uma



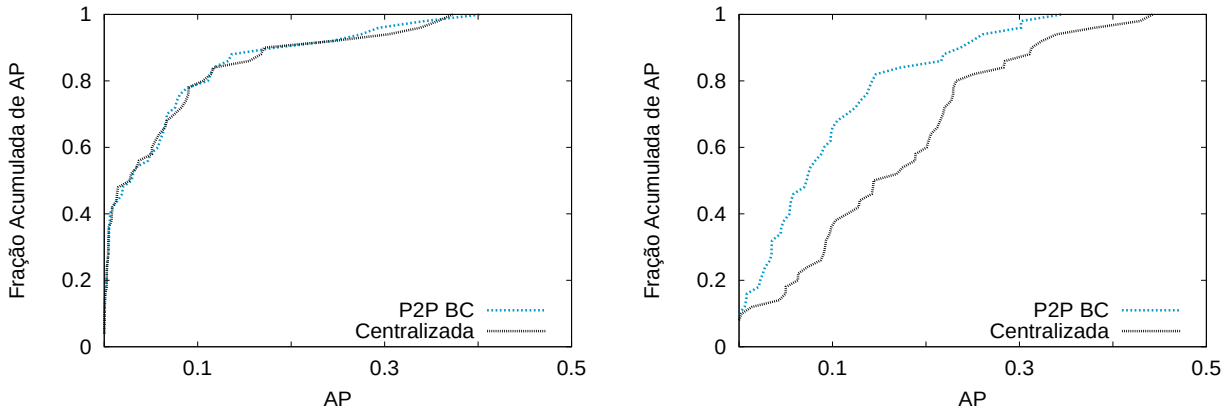
(a) TREC (b) WBR

Figura 5.3: Eficácia da máquina de busca P2P com alto conhecimento dos pares (AC)

URL, o que nos permite fazer uma distribuição direta entre os pares na linha de base: cada uma das n_{wbr} máquinas hospedeiras (i.e., *hosts*) é associada aleatoriamente a um único par e as páginas Web da máquina hospedeira são associadas aos documentos do par correspondente. Por outro lado, como a TREC não possui característica Web, distribuímos uniformemente os documentos pelos pares, de forma que cada par contenha em torno de D_{trec}/n_{trec} documentos, uma amostra da coleção com um nível de confiança de 99% e acurácia em torno de 5%.

Como ilustra a Figura 5.3, os resultados da simulação mostram que no cenário AC, isto é, quando pares têm pleno conhecimento sobre a relevância de seus documentos na rede, nosso modelo de máquina de busca P2P apresenta resultados equivalentes aos de uma máquina de busca centralizada. Isso acontece porque os valores de similaridade entre a consulta e os documentos que um par possui são calculados usando os mesmos algoritmos e as mesmas estatísticas (globais) da máquina de busca centralizada.

Por outro lado, a Figura 5.4 mostra que a eficácia de máquinas de busca P2P com baixos níveis de conhecimento dos pares (BC) pode sofrer um impacto significativo em redes P2P com alta heterogeneidade, pois as similaridades dependem essencialmente de como os documentos da coleção estão distribuídos entre os pares: quando documentos estão distribuídos por máquina hospedeira (*host*), o conjunto de documentos dos pares pode ser muito específico (em torno de um mesmo tópico), o que faz com que a qualidade dos resultados de cada par tenda a ser bastante distinta da linha de base. Por outro lado, os resultados mostram que redes em que documentos estão distribuídos uniformemente (TREC) são muito menos vulneráveis à falta de conhecimento da rede. De fato, os resultados para a TREC mostram que os diferentes níveis de conhecimento da rede pelos pares apresentam resultados bastante similares (quando comparados à WBR), sugerindo que a distribuição uniforme garante aos pares uma amostra representativa dos documentos da rede.



(a) TREC (b) WBR

Figura 5.4: Eficácia da máquina de busca P2P com baixo conhecimento dos pares (BC)

Em suma, nossos resultados mostram que a *autonomia* dos usuários pode ser um aspecto crucial para máquinas de busca P2P. De fato, baixos níveis de conhecimento dos pares sobre a relevância de seus documentos na rede e o potencial de pares disponibilizarem muitos documentos voltados a um mesmo tópico de interesse do usuário parecem comprometer a eficácia de máquinas de busca P2P em relação a máquinas de busca centralizadas.

Nesse sentido, é essencial que pares procurem manter um certo nível de atualização sobre estatísticas de documentos na rede; os algoritmos de fofoca (ou *Gossiping*) [Demers et al. (1987)] e de *amostragem baseada em consultas* [Callan e Connell (2001)] podem ser extremamente úteis. Além disso, técnicas de incentivo para usuários baixarem e indexarem documentos de outros pares podem atenuar os efeitos da disponibilização de conteúdo voltado a um mesmo tópico.

Em nosso estudo, não exploramos os potenciais de escalabilidade e de ampliação da cobertura da Web de máquinas de busca P2P. De fato, nosso foco está na eficácia, e não no desempenho, das máquinas de busca; além disso, para fins de comparação, utilizamos as mesmas coleções de teste na avaliação dos modelos de busca P2P e centralizada. Porém, a eficiência da arquitetura e a maior capacidade de escolha dos usuários podem ser fatores diferenciais oferecidos por máquinas de busca P2P e que também precisam ser avaliados.

5.4 Desafios

Apesar de contar com uma arquitetura escalável, máquinas de busca P2P precisam lidar com problemas ligados à *eficiência* da busca. De fato, cada par encaminha a consulta do usuário à rede para obter os documentos mais relevantes, o que pode exigir uma série de requisições em busca do maior número possível de pares promissores. Apesar de existirem alternativas que amenizem a busca exaustiva ao longo da rede (i.e., *flooding*), como expansão em anel (i.e., *expanding ring*) e caminhamento aleatório, discutidos por Lv et al. (2002), arquiteturas P2P descentralizadas ainda enfrentam problemas sérios de eficiência [Zhu (2006)]. Além disso, a seleção de pares promissores, o encaminhamento e processamento das consultas podem consumir tempo e largura de banda consideráveis.

A *consistência* dos resultados apresentados ao usuário é ainda um grande desafio para arquiteturas P2P descentralizadas. Nesse tipo de rede, pares podem manter diferentes conjuntos de documentos e utilizar diferentes algoritmos de processamento de consultas de acordo com o interesse de seus usuários. Como conseqüência, pares podem apresentar resultados extremamente peculiares e ainda ordenar um mesmo documento em posições completamente diferentes. Durante a fusão dos resultados, o par requisitante precisa adotar um critério de comparação e reordenação dos documentos obtidos que seja ao mesmo tempo justo e consistente. Em redes P2P descentralizadas com grande autonomia dos pares, a escolha desse critério pode ser crucial para a eficiência da busca.

Trabalhos recentes [Gyongyi e Garcia-Molina (2005); Fetterly et al. (2004)] destacam que

as máquinas de busca tradicionais têm sido desafiadas a se proteger contra a *ação maliciosa* de alguns usuários – em particular de *Web spammers*, que visam melhorar a avaliação de certas páginas da Web pela máquina de busca. Em máquinas de busca P2P, o impacto de *Web spamming* na qualidade das respostas pode ser ainda maior [Atalla et al. (2008)]: em um cenário em que a informação está distribuída em pares autônomos, usuários geralmente não têm acesso direto aos documentos dos pares. Assim, ao invés de tentar melhorar a avaliação indiretamente, por exemplo alterando o conteúdo do documento ou informações de *link*, um par *Web spammer* pode alterar diretamente a posição do documento em sua resposta a uma consulta, de forma a forçar a promoção de certas páginas da Web.

Finalmente, o *comportamento dinâmico* dos pares pode comprometer a eficácia de máquinas de busca P2P. Trabalhos anteriores sugerem que a entrada e saída dinâmica dos pares da rede pode ter grande impacto na qualidade de serviço oferecida por sistemas P2P [Rhea et al. (2004); Stutzbach e Rejaie (2006)]. Particularmente em aplicações P2P de busca por conteúdo, o impacto em sua eficácia pode ser significativamente diferente daquele observado em aplicações de compartilhamento de arquivos. De fato, em sistemas de compartilhamento de arquivos, o conteúdo geralmente encontra-se replicado na rede, pois espera-se que os usuários baixem e mantenham cópias dos arquivos desejados localmente. Por outro lado, em máquinas de busca P2P, espera-se que os usuários apenas naveguem pelas páginas retornadas pela consulta, sem precisarem explicitamente manter uma cópia local dessas páginas. Dessa forma, a disponibilidade de conteúdo pode ser significativamente menor em máquinas de busca P2P e, em última instância, a eficácia de máquinas de busca P2P pode ser largamente afetada. Além disso, em aplicações de busca por conteúdo em P2P, diferentes documentos podem satisfazer a consultas dos usuários em diferentes níveis de relevância. Se esses documentos estiverem distribuídos em diferentes pares, a busca será parcialmente efetiva se alguns desses pares não estiverem disponíveis no momento em que a consulta for realizada. Por outro lado, em aplicações de compartilhamento de arquivos, a busca só será efetiva se a cópia exata do arquivo estiver presente em algum dos pares que estiverem disponíveis.

Em nosso estudo, focamos no comportamento dinâmico dos pares. Na próxima seção, estendemos nosso modelo de busca P2P para comportar a entrada e saída dinâmica dos pares, ou *churn*, e discutimos as principais conseqüências desse novo cenário na viabilidade de máquinas de busca P2P.

5.5 Comportamento Dinâmico dos Pares

5.5.1 Definição

O *comportamento dinâmico dos pares* refere-se à freqüente e independente entrada e saída de pares de uma rede P2P. Como em ambientes cliente-servidor, os pares podem eventualmente

abandonar a rede em razão de falhas técnicas [Menascé e Almeida (2001)]. No entanto, em redes P2P, o comportamento dos pares é particularmente determinado pelos interesses de seus usuários, que têm autonomia para decidir o momento mais apropriado para participar da máquina de busca P2P [Stutzbach e Rejaie (2006)].

Máquinas de busca Par-a-Par dinâmicas são máquinas de busca P2P cujos pares apresentam esse tipo de comportamento.

5.5.2 Modelo

Nesta seção, *estendemos* o modelo de máquinas de busca P2P com o comportamento dinâmico dos pares. Os parâmetros do modelo estão sumarizados na Tabela 5.3.

Tabela 5.3: Parâmetros do modelo de máquina de busca P2P dinâmica

Parâmetro	**Descrição**
τ_c	Limiar de similaridade para a coleção de teste c (c=*wbr*, *trec*)
t_q	Tempo médio entre chegadas das consultas à máquina de busca
n_c	# de pares da coleção de teste c (c=*wbr*, *trec*)
λ_{on}, ρ_{on}	Parâmetros da distribuição Weibull para os tempos *online* dos pares
λ_{off}, ρ_{off}	Parâmetros da distribuição Weibull para os tempos *offline* dos pares

O comportamento dinâmico dos pares em máquinas de busca P2P ainda não foi caracterizado na literatura [1], embora tenha sido investigado em diversos trabalhos sobre aplicações P2P de compartilhamento de arquivos. Em particular, uma caracterização recente [Stutzbach e Rejaie (2006)] de três sistemas P2P populares mostra que a duração da sessão dos pares (i.e., tempo *online*) é bem modelada por distribuições Weibull, com função de densidade

$$f(x) = \frac{\rho x^{\rho-1}}{\lambda^{\rho}} e^{-(x/\lambda)^{\rho}} \tag{5.1}$$

e parâmetros de forma $\rho \approx 0,44$ e de escala $\lambda \approx 35,20$ unidades de tempo, em média. Note que $\rho < 1$ modela sistemas com uma taxa decrescente de saída de pares [Schroeder e Gibson (2006)] enquanto altos (ou baixos) valores de λ ampliam (ou reduzem) a duração das sessões. Embora relevante, esse estudo não apresenta dados similares para os tempos *offline*, i.e., o tempo entre duas sessões consecutivas de um mesmo par, outro aspecto de suma importância para a eficácia da busca em P2P.

Devido à falta de caracterização do comportamento dinâmico dos pares em máquinas de busca P2P assim como sua caracterização completa em outros tipos de aplicação em P2P, optamos por modelar os tempos *online* e *offline* dos pares por distribuições Weibull com parâmetros de forma ρ_{on} e ρ_{off} e parâmetros de escala λ_{on} e λ_{off}, respectivamente.

[1] À exceção do trabalho que publicamos abordando o assunto [Atalla et al. (2008)]

5.5.3 Resultados de Simulação

Nesta seção quantificamos o impacto do comportamento dinâmico dos pares na eficácia de máquinas de busca P2P. Nossos resultados mostram que redes com diferentes níveis de conhecimento dos pares sobre a rede e com diferentes formas de disponibilização de conteúdo podem sofrer de forma diferente com a instabilidade dos pares.

Para avaliar o impacto de diferentes níveis de comportamento dinâmico, utilizamos a média da distribuição Weibull, isto é,

$$\mu = \lambda \times \Gamma\left(1 + 1/\rho\right) \tag{5.2}$$

onde Γ é a função Gamma e ρ e λ são os parâmetros de forma e escala descritos anteriormente. Fixamos então a distribuição dos tempos *online* dos pares com uma média de $\mu_{on} = 91,84$ unidades de tempo, dada pelos parâmetros ρ_{on} e λ_{on} iguais a 0,44 e 35,20 respectivamente. Além disso, fixamos $\rho_{off} = \rho_{on}$ enquanto variamos λ_{off}, produzindo diferentes tempos médios *offline* μ_{off}.

De forma a capturar a estabilidade do sistema, adotamos a definição de *disponibilidade dos pares* como sendo o percentual do tempo (simulado) que os pares participam da rede, em média. A disponibilidade dos pares, A, pode ser aproximada [Menascé e Almeida (2001)] por

$$A = \frac{\mu_{on}}{\mu_{on} + \mu_{off}} \times 100 \tag{5.3}$$

A Tabela 5.4 sumariza os diferentes níveis de disponibilidade avaliados. Quando todos os pares estão disponíveis ($A = 100\%$), os resultados da simulação são equivalentes àqueles obtidos do modelo de máquina de busca P2P. Quando todos os pares estão indisponíveis ($A = 0\%$), AP é sempre zero.

Tabela 5.4: Parâmetros de disponibilidade dos pares (ρ_{off}=ρ_{on}=0,44,λ_{on}=35,20)

$A\%$	0	25	50	75	100
λ_{off}	∞	105,60	35,20	11,73	0
μ_{off}	∞	275,52	91,84	30,61	0

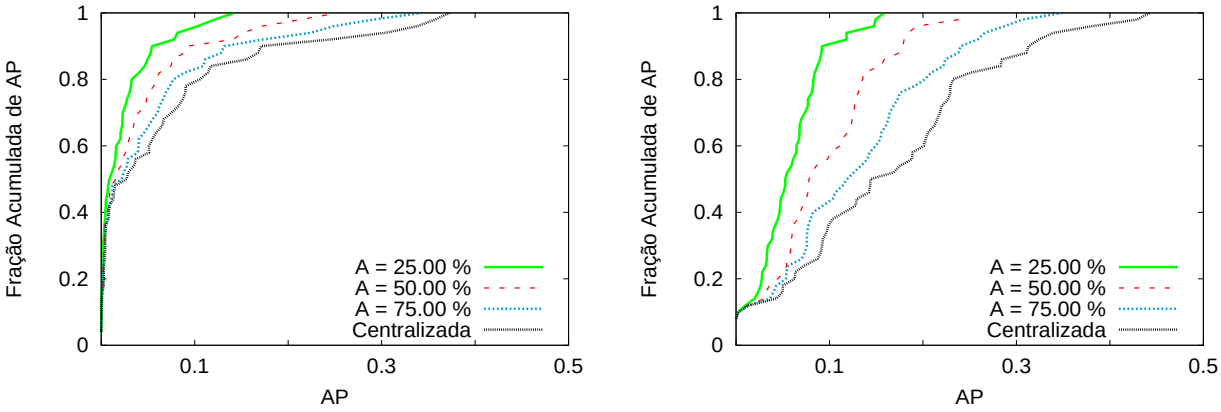
Os resultados principais estão sumarizados na Tabela 5.5. A Tabela mostra o percentual de degradação para os valores de MAP, quando comparados à linha de base (LB), para todos os cenários analisados, em ambos níveis alto (AC) e baixo (BC) de conhecimento dos pares sobre os documentos da rede P2P. A Tabela ainda mostra os valores de LB e os valores de MAP entre parênteses. Mesmo quando os pares têm um alto nível de conhecimento sobre os documentos da rede (AC) e a maioria dos pares está disponível em média ($A = 75\%$), o MAP reduz em torno de 22% em ambas coleções. No entanto, conforme observamos no modelo de máquina de busca P2P, em ambientes com baixo conhecimento dos pares (BC) o impacto pode ser ainda mais intenso na coleção WBR. Nesse cenário, a degradação da eficácia da busca em relação à LB pode chegar a

54% também para $A = 75\%$.

Tabela 5.5: Impacto do comportamento dinâmico dos pares sobre MAP (MAP entre parênteses)

	TREC		**WBR**	
A %	**AC %**	**BC %**	**AC %**	**BC %**
LB	0 (0,069)		0 (0,166)	
75	22 (0,054)	23 (0,053)	22 (0,129)	54 (0,076)
50	43 (0,039)	45 (0,038)	43 (0,094)	66 (0,056)
25	67 (0,023)	67 (0,023)	65 (0,058)	78 (0,036)

De forma similar à análise realizada para a linha de base, também avaliamos a degradação das consultas individualmente. As Figuras 5.5 e 5.6 mostram a distribuição acumulada da métrica AP de resultados individuais de consultas, para os vários parâmetros de disponibilidade dos pares, incluindo o cenário LB. Novamente, as curvas mostram uma grande variabilidade nos valores individuais de AP.

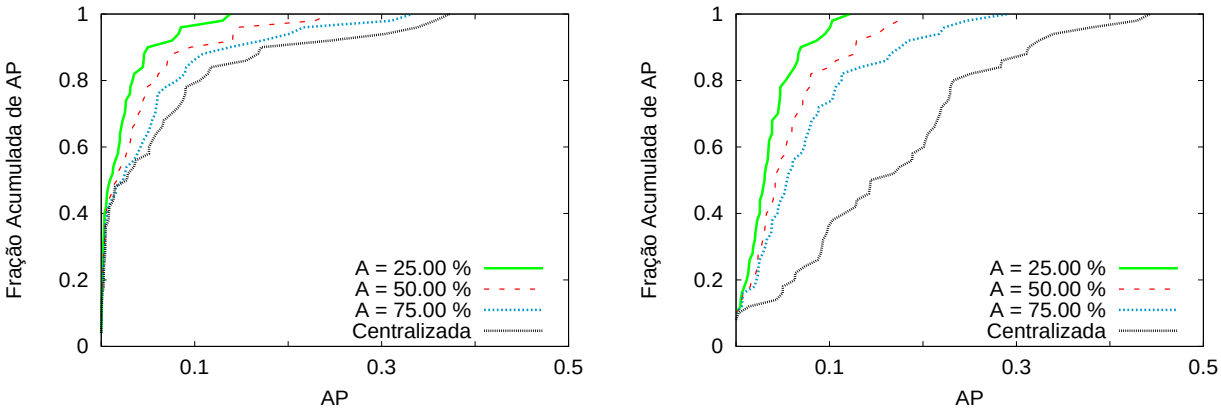


(a) TREC (b) WBR

Figura 5.5: Eficácia da máquina de busca P2P dinâmica com alto conhecimento dos pares (AC)

As curvas ainda mostram que o impacto do comportamento dinâmico dos pares aumenta com a métrica AP. Isso acontece porque altos valores de AP são conseqüência do grande número de documentos relevantes presentes na rede (em alguns pares) e informados na resposta a uma consulta. Quanto maior esse número, maior é a chance de que documentos presentes em pares indisponíveis no momento em que a consulta é realizada sejam relevantes para a consulta.

O impacto do comportamento dinâmico dos pares pode ser ainda maior do que sugerido pelos resultados agregados (MAP). De fato, comparado à linha de base, para $A = 75\%$ e com alto nível de conhecimento dos pares sobre os documentos da rede (AC), 20% das consultas sofrem uma degradação acima de 24% na métrica AP em ambas coleções. As conseqüências de cenários com menor nível de conhecimento dos pares (BC) são ainda mais severas na eficácia da busca:



(a) TREC (b) WBR

Figura 5.6: Eficácia da máquina de busca P2P dinâmica com baixo conhecimento dos pares (BC)

para 20% das consultas, o impacto é maior que 29% e 73%, para as coleções TREC e WBR, respectivamente ($A = 75\%$).

Em suma, o comportamento dinâmico dos pares pode ter grande impacto na eficácia de máquinas de busca P2P, mesmo em ambientes razoavelmente estáveis (A=75%) e alto nível de conhecimento dos pares sobre os documentos da rede (AC). Os resultados também sugerem que os efeitos da entrada e saída dinâmica dos pares podem ser ainda mais intensos em redes com menos conhecimento dos pares (e mais reais), em que se espera que os pares processem consultas usando quase que apenas informações locais e indexem seus próprios documentos.

5.5.4 Soluções

A *replicação* de documentos tem se mostrado uma alternativa bastante promissora para atenuar os efeitos do comportamento dinâmico dos pares em aplicações de redes P2P descentralizadas [Pouwelse et al. (2005)]. Estratégias de replicação estáticas (ver Capítulo 2) apresentam soluções promissoras para aumentar a disponibilidade de documentos populares (relevantes). No entanto, a definição *a priori* de regras de distribuição de documentos entre pares pode não ser viável para máquinas de busca P2P, em que os usuários possuem autonomia para coletar e indexar documentos localmente à sua maneira. Por outro lado, estratégias de replicação dinâmicas podem ser soluções promissoras e viáveis para máquinas de busca P2P, já que os pares só precisariam armazenar documentos que atendessem às necessidades de informação de seus usuários.

Em nosso estudo, focamos em estratégias de replicação dinâmicas. Na próxima seção, estendemos nosso modelo para incluir a possibilidade de os pares baixarem e indexarem uma fração de documentos retornados em resposta às consultas de seus usuários. Também discutimos as principais implicações dessa solução na viabilidade de máquinas de busca P2P.

5.6 Replicação Dinâmica por Similaridade

5.6.1 Definição

A *replicação dinâmica por similaridade* é uma estratégia de replicação dinâmica que propomos para o contexto de máquinas de busca P2P. Nessa estratégia, alguns documentos apresentados pela máquina de busca P2P em resposta a uma consulta são copiados no par requisitante, de forma a aumentar a disponibilidade de conteúdo no sistema.

5.6.2 Modelo de Simulação

Nesta seção, *estendemos* o modelo anterior com a replicação dinâmica por similaridade. Os parâmetros do modelo estão sumarizados na Tabela 5.6.

Tabela 5.6: Parâmetros do modelo de máquina de busca P2P dinâmica com replicação

Parâmetro	**Descrição**
τ_c	Limiar de similaridade para a coleção de teste c (c=*wbr*, *trec*)
t_q	Tempo médio entre chegadas das consultas à máquina de busca
n_c	# de pares da coleção de teste c (c=*wbr*, *trec*)
λ_{on}, ρ_{on}	Parâmetros da distribuição Weibull para os tempos *online* dos pares
λ_{off}, ρ_{off}	Parâmetros da distribuição Weibull para os tempos *offline* dos pares
r	# de documentos replicados dinamicamente por similaridade

Pretendemos avaliar a replicação por conteúdo como estratégia para aumentar a disponibilidade de conteúdo e, em última instância, a qualidade do serviço oferecido por máquinas de busca P2P. Nessa direção, adaptamos a estratégia de replicação dinâmica própria [Cohen e Shenker (2002)] ao cenário de máquinas de busca P2P. Em nosso modelo, após a consulta ser respondida, o par requisitante armazena localmente cópias dos r primeiros documentos da lista; se o par requisitante já possuir k documentos dentre os r primeiros, apenas os $r - k$ documentos não existentes no par serão copiados. O conjunto de documentos local é então reindexado.

Em nosso modelo, supomos que *(1)* os pares não possuem restrições de armazenamento, isto é, cópias locais nunca são removidas; *(2)* as cópias dos documentos são sempre consistentes, isto é, os pares mantém as versões mais recentes de seus documentos na rede; e *(3)* o número de documentos únicos da rede é constante e dado pelas coleções de teste, isto é, não há criação ou exclusão de documentos. Essas suposições nos permitem avaliar o ganho potencial da replicação de conteúdo como estratégia para atenuar o impacto de diferentes níveis de conhecimento dos pares sobre os documentos da rede na eficácia de máquinas de busca P2P dinâmicas, considerando mudanças no conjunto de documentos de cada par ao longo do tempo.

5.6.3 Modelo Teórico

Consideremos um cenário de máquina de busca P2P em que todo par apresenta uma amostra representativa dos documentos da rede P2P e que é capaz de replicar localmente os r primeiros documentos obtidos em resposta a uma consulta. Nesse cenário, como todo par sempre baixa os primeiros r documentos de uma consulta, em algum momento sempre haverá um par disponível que possua esses documentos. Seja t o número total de documentos obtidos pela máquina de busca centralizada. Podemos supor que a qualidade máxima da busca, MAP_r, corresponde no mínimo à qualidade dos resultados obtidos por uma máquina de busca centralizada para os $r \leq t$ primeiros documentos, LB_r, isto é,

$$MAP_r \geq LB_r \tag{5.4}$$

Como os $t - r$ documentos restantes nunca são replicados, se apenas $A\%$ dos pares estiverem disponíveis no momento em que a consulta é realizada, em média, então a resposta final deverá ter aproximadamente essa mesma fração de documentos. Nesse cenário, como os documentos da rede encontram-se uniformemente distribuídos entre os pares, espera-se que a qualidade da resposta determinada pelos documentos restantes, $LB_{t-r} = LB_t - LB_r$, seja reduzida por um fator em torno de $A\%$ em relação à qualidade obtida pela máquina de busca centralizada.

Portanto, a qualidade máxima de uma máquina de busca P2P dinâmica com replicação pode ser estimada pela métrica MAP quando os r primeiros documentos obtidos em resposta à consulta são replicados localmente, isto é,

$$MAP_r = LB_r + A \times LB_{t-r} = A \times LB_t + (1 - A) \times LB_r \tag{5.5}$$

Note que MAP_0 e MAP_t correspondem aos limites teóricos inferior e superior, respectivamente, para a eficácia da máquina de busca.

O modelo teórico pode ser utilizado não apenas para validar os resultados obtidos pelo modelo de simulação, como para antecipar a eficácia máxima da busca P2P. Exemplos de aplicação incluem cenários cuja avaliação experimental pode ser inviável devido ao tempo de convergência da simulação ou mesmo devido a restrições de recursos computacionais.

5.6.4 Resultados

Nesta seção avaliamos a replicação dinâmica por similaridade como forma de atenuar o impacto do comportamento dinâmico dos pares na eficácia de máquinas de busca P2P. Consideramos diferentes níveis de replicação de documentos, focando em cenários pessimistas em que a maioria dos pares não está disponível no momento em que a consulta é realizada, isto é, $A = 25\%$. Para tanto, a cada iteração do simulador, um par disponível é sorteado para selecionar e submeter à rede uma consulta escolhida aleatoriamente dentre aquelas fornecidas pela coleção de teste. Os r primeiros documentos apresentados pela máquina de busca P2P em resposta à consulta são

adicionados automaticamente ao conjunto de documentos do par requisitante.

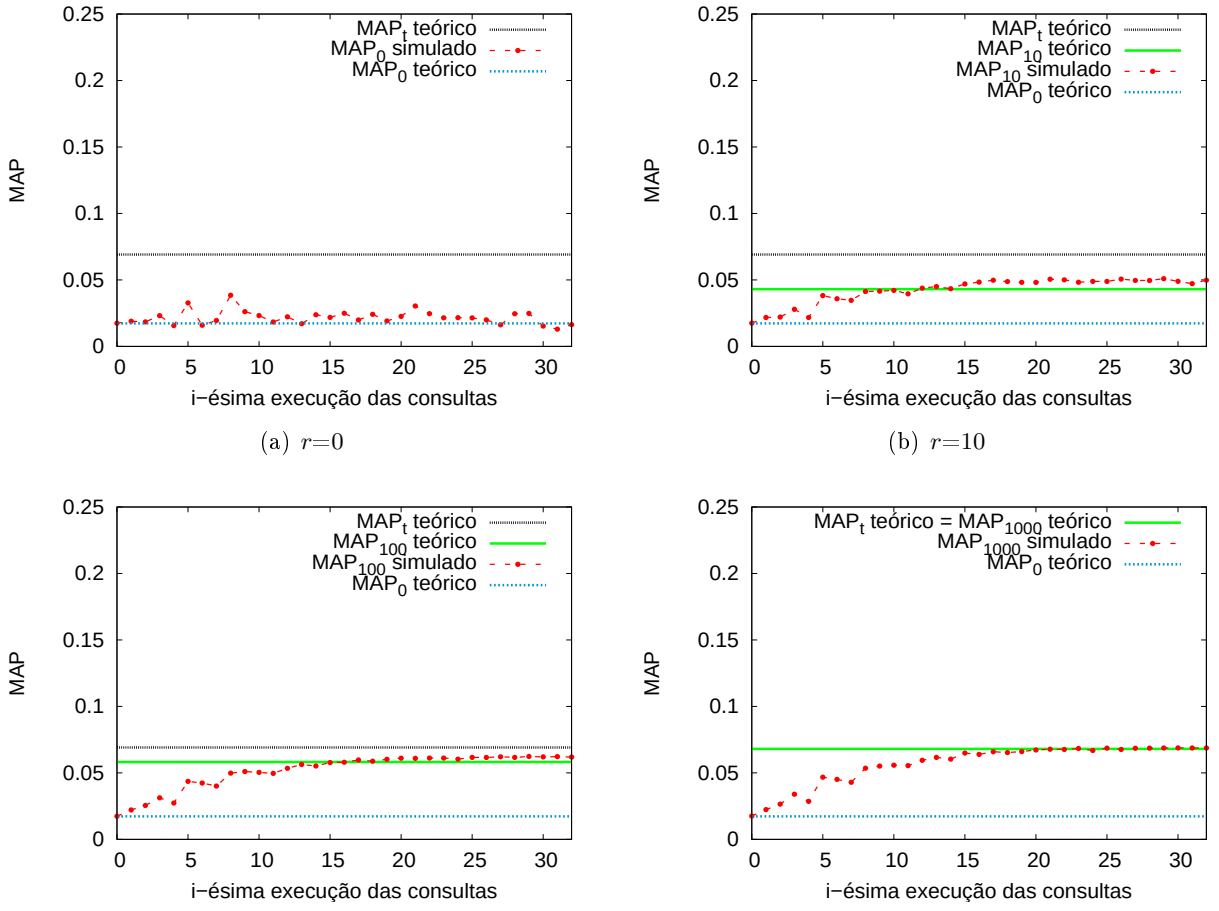
Tabela 5.7: Parâmetros da eficácia máxima esperada (MAP_r): LB_r, A=25%

	LB_r	
r	**TREC**	**WBR**
t	0,069	0,166
1.000	0,068	0,166
100	0,055	0,149
10	0,034	0,059
0	0	0

Os resultados principais são apresentados nas Figuras 5.7-5.10. As Figuras ilustram a melhoria nos resultados agregados (MAP) na medida em que documentos vão sendo replicados pela rede ao longo do tempo. O i-ésimo valor de MAP das curvas de simulação foi obtido na primeira iteração em que todas as consultas foram executadas pelo menos i vezes. Já os valores esperados de MAP foram calculados a partir dos parâmetros sumarizados na Tabela 5.7. Os diferentes parâmetros LB_r correspondem aos valores de MAP obtidos no cenário de busca centralizada em que os r primeiros documentos retornados pela máquina de busca são avaliados. Em particular, LB_t corresponde ao valor obtido no Capítulo 4, em que a avaliação da eficácia foi feita sobre todos os documentos retornados.

Inicialmente avaliamos o ganho potencial da replicação dinâmica sem considerar possíveis limitações no conhecimento dos pares sobre os documentos da rede. As Figuras 5.7 e 5.8 mostram a melhoria na métrica MAP no cenário AC para as coleções de teste TREC e WBR, respectivamente. Em particular, os resultados da simulação mostram que mesmo se os pares baixarem apenas os 10 primeiros documentos apresentados pela máquina de busca P2P em resposta à consulta, a degradação na qualidade das respostas provocada pela indisponibilidade dos pares pode reduzir de 68% para 28% e de 66% para 32% nas coleções TREC e WBR, respectivamente (Figuras 5.7(b) e 5.8(b)). Como esperado, quando os pares baixam ainda mais documentos, a melhoria na qualidade da busca pode ser ainda maior: por exemplo, a replicação dos 100 primeiros documentos pode reduzir a degradação para apenas 10% e 4% na TREC e WBR, respectivamente, enquanto que a adição dos primeiros 1000 documentos praticamente elimina os efeitos da indisponibilidade de conteúdo. No entanto, embora seja possível eliminar a degradação, a melhoria na eficácia da busca não é tão expressiva *por documento baixado*. Isso acontece porque geralmente as máquinas de busca relacionam a maioria dos documentos relevantes para a consulta entre as primeiras posições da lista de resposta. Dessa forma, mesmo a cópia local de apenas os primeiros 10 documentos pode ser suficiente para garantir a disponibilidade do conteúdo mais relevante da rede.

As Figuras 5.7 e 5.8 também sugerem que nosso modelo teórico é *pessimista*. De fato, em todos os cenários com alto nível de conhecimento dos pares (AC), os resultados da simulação mostraram-se ligeiramente superiores àqueles previstos pela eficácia máxima esperada MAP_r.

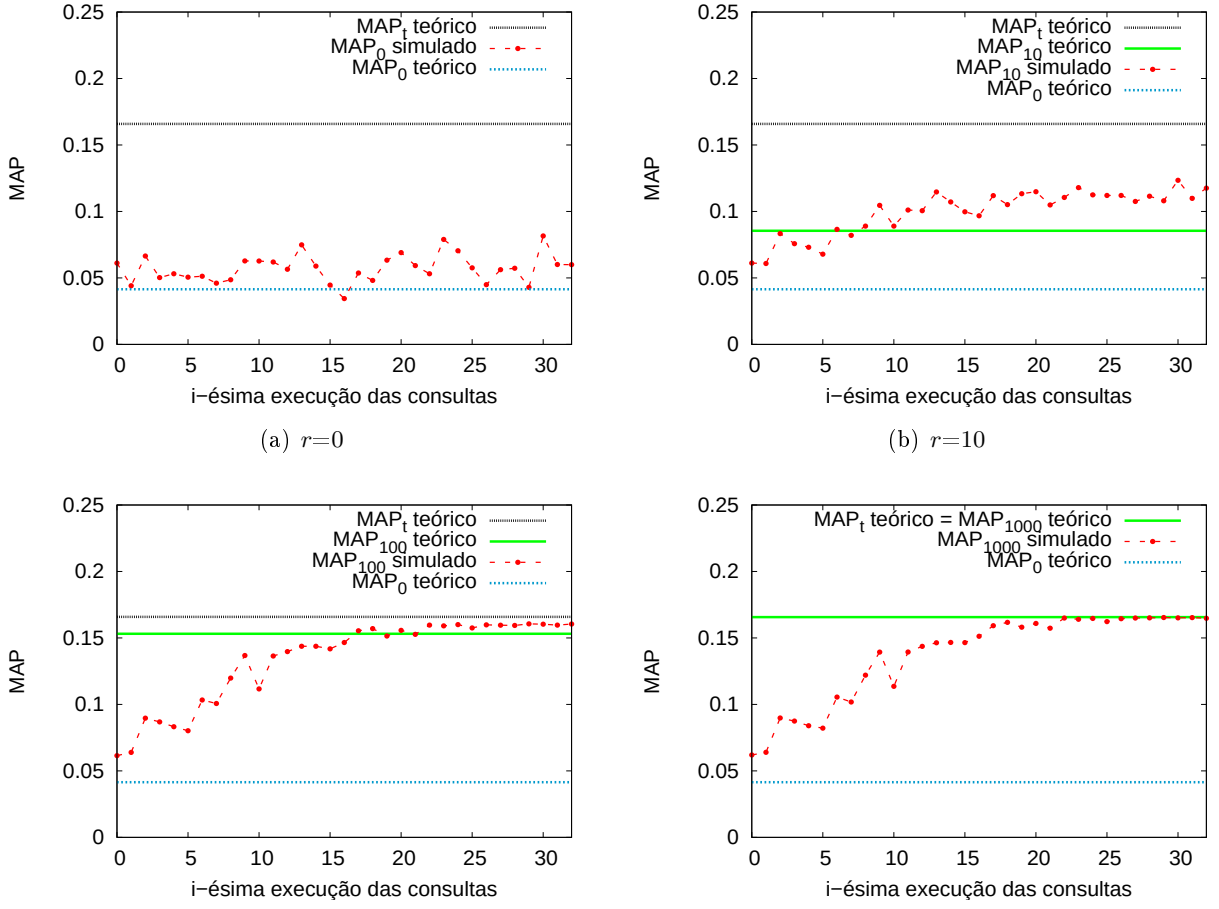


(a) r=0 (b) r=10

(c) r=100 (d) r=1000

Figura 5.7: Eficácia da máquina de busca P2P dinâmica com replicação e alto conhecimento dos pares (AC) – TREC

Isso acontece porque o modelo teórico supõe que os documentos relevantes estão uniformemente distribuídos na lista de documentos retornados pela máquina de busca, quando na prática os documentos julgados como relevantes ficam mais concentrados entre as primeiras posições da lista (ver Figura 4.2). De fato, quando $r = 0$, em cenários com distribuição uniforme de documentos como a TREC, espera-se que cada par contribua com uma fração semelhante de documentos na resposta da máquina de busca a uma consulta. Dessa forma, quando $A = 25\%$, uma fração semelhante de documentos deve ser removida uniformemente da resposta. Como a maioria dos documentos da resposta não é relevante, a maior parte dos documentos removidos da resposta também não será relevante, o que pode contribuir para um aumento na qualidade da busca. Além disso, como a estratégia de replicação por similaridade propõe que os r primeiros documentos sejam baixados da resposta, na prática a quantidade de conteúdo relevante replicado poderá ser ainda maior que o esperado.



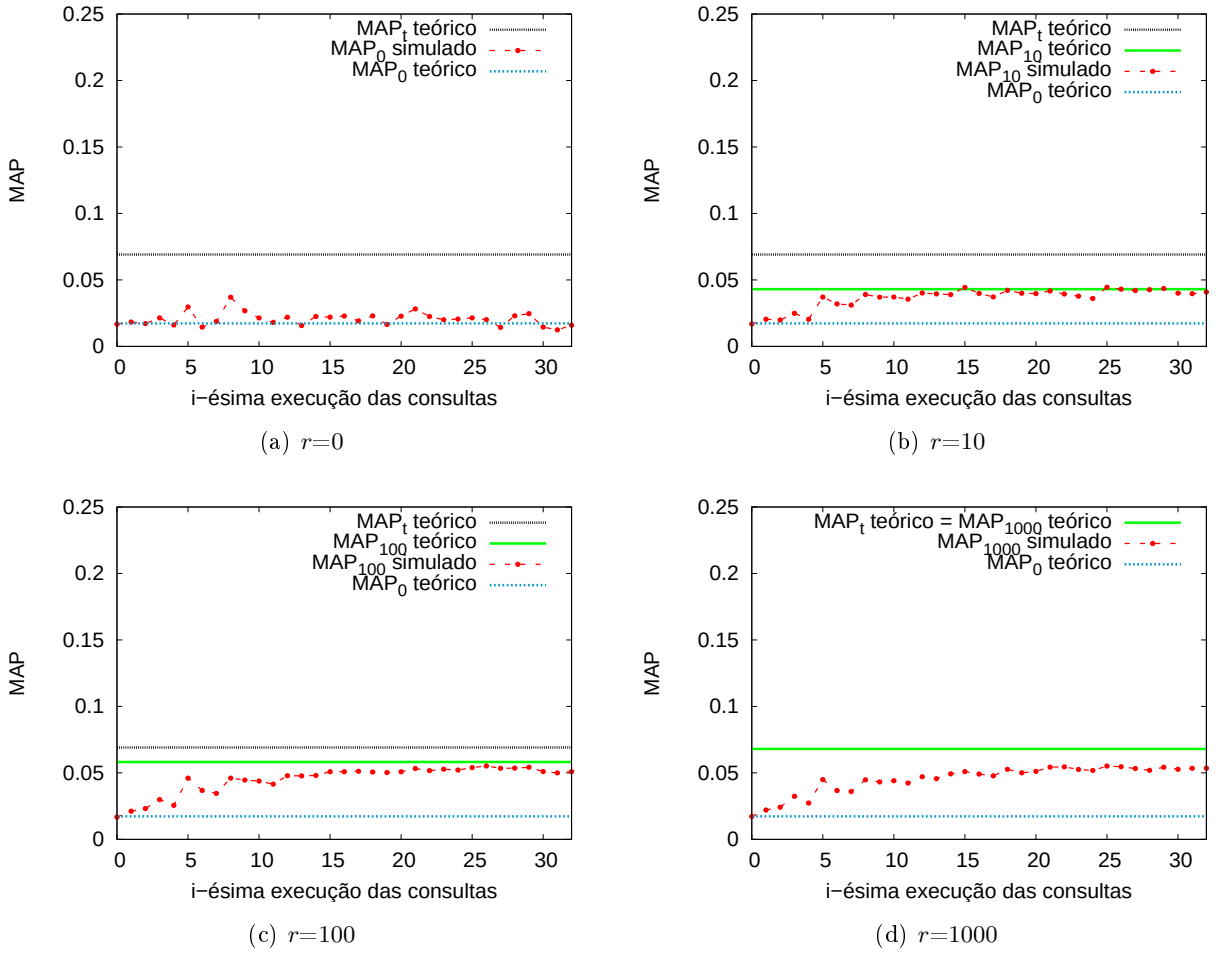
(a) r=0 (b) r=10

(c) r=100 (d) r=1000

Figura 5.8: Eficácia da máquina de busca P2P dinâmica com replicação e alto conhecimento dos pares (AC) – WBR

No entanto, as Figuras 5.7 e 5.8 mostram que na medida em que o número r de documentos baixados da lista de resposta aumenta, a diferença entre as curvas teórica e de simulação tende a diminuir. Na medida em que r aumenta, a fração de documentos relevantes entre os documentos baixados diminui, já que a maior parte dos documentos das coleções de teste é julgado como não relevante (ver Seção 3.2). Por exemplo, na TREC a diferença entre as curvas reduz significativamente, passando de 23% para 14% e 1%, quando $r = 0$, $r = 10$ e $r = 1000$, respectivamente.

Em um segundo momento, avaliamos o ganho potencial da replicação dinâmica considerando possíveis limitações no conhecimento dos pares sobre os documentos da rede. As Figuras 5.9 e 5.10 mostram a melhoria na métrica MAP no cenário BC para as coleções de teste TREC e WBR, respectivamente. Mesmo quando pares possuem pouco conhecimento sobre os documentos da rede e a maioria dos pares encontra-se indisponível no momento da consulta ($A = 25\%$), MAP_{10}



(a) r=0 (b) r=10

(c) r=100 (d) r=1000

Figura 5.9: Eficácia da máquina de busca P2P dinâmica com replicação e baixo conhecimento dos pares (BC) – TREC

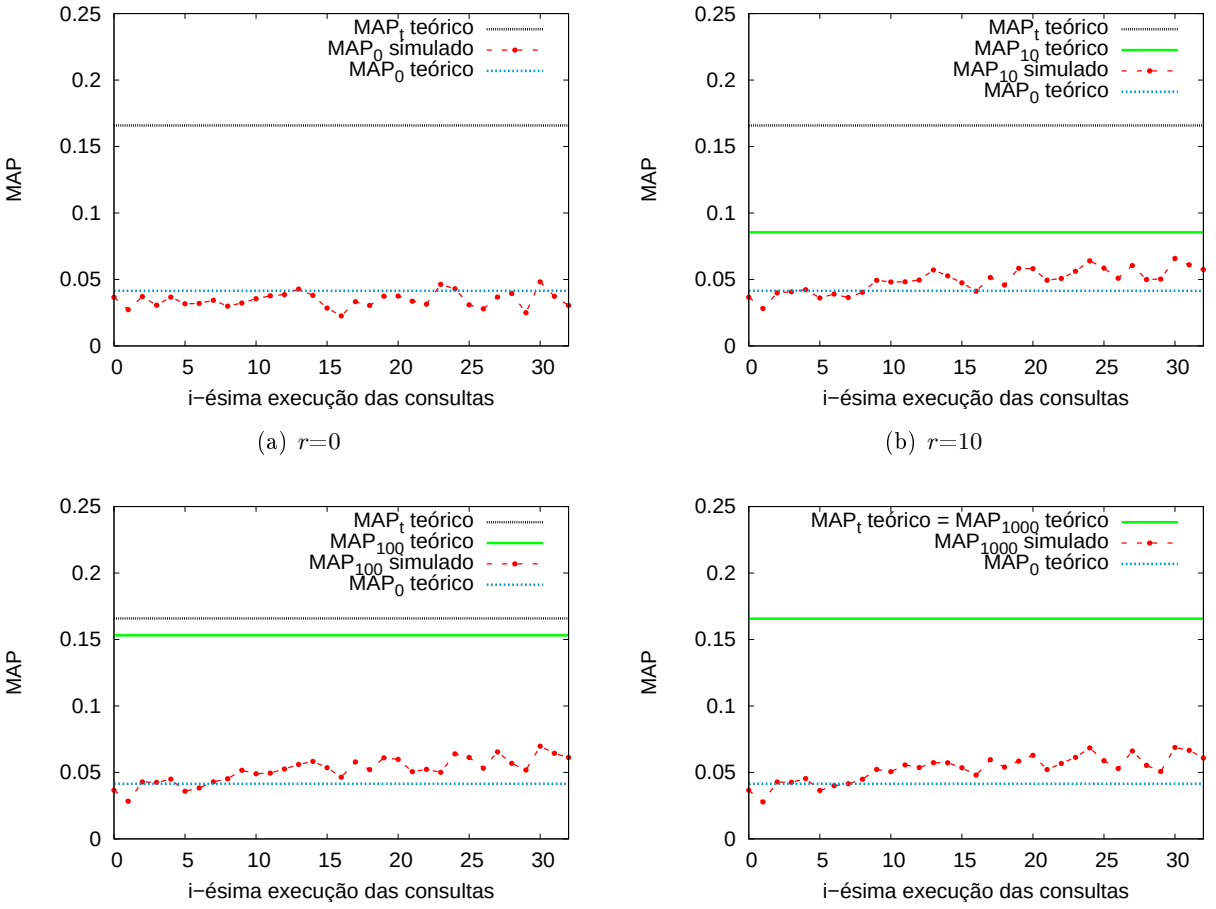
sugere que a degradação na qualidade da busca pode reduzir significativamente, de 75% para 38% e de 75% para 49% nas coleções TREC e WBR, respectivamente.

A Figura 5.9 sugere que a *topicidade* [2] dos documentos replicados pode ter um grande impacto na eficácia da estratégia de replicação proposta. Os documentos apresentados pela máquina de busca em resposta a uma consulta são geralmente muito específicos, isto é, em torno de um mesmo tópico de interesse (ou assunto). Dessa forma, a estratégia de replicação por similaridade pode aumentar substancialmente a topicidade dos pares especialmente quando o número r de documentos é relativamente alto (Figura 5.9(d)). Mesmo em cenários em que os pares contém uma amostra representativa dos documentos da rede (TREC), a primeira consulta feita pelo par já é capaz de tornar o conjunto de documentos local muito específico. Em conseqüência, como o processador de consultas adotado utiliza estatísticas TF-IDF, as próximas consultas respondidas

[2]Cobertura de um conjunto de documentos sobre um determinado tópico de interesse

pelo par e que forem relacionadas ao mesmo tópico de interesse poderão apresentar documentos com similaridade menor, uma vez que o componente IDF é largamente reduzido quando muitos documentos abordam um mesmo tópico de interesse (ver Equação 4.2).

O efeito da topicidade pode ser ainda mais intenso quando os pares já possuem documentos voltados para um mesmo tópico de interesse. De fato, a Figura 5.10 mostra que mesmo se apenas os $r = 10$ primeiros documentos da lista de resposta são replicados localmente, o tempo de convergência pode ser extremamente alto, uma vez que para o mesmo número de consultas executadas no cenário AC, os resultados da simulação ainda não alcançaram a qualidade máxima esperada (MAP_{10}). Como esperado, para valores de r maiores, o tempo de convergência é ainda maior.



(a) r=0 (b) r=10

(c) r=100 (d) r=1000

Figura 5.10: Eficácia da máquina de busca P2P dinâmica com replicação e baixo conhecimento dos pares (BC) – WBR

Portanto, os resultados de simulação sugerem que o tempo de convergência depende amplamente do nível de conhecimento dos pares sobre os documentos da rede, o que pode levar a

diferentes números de consultas necessárias para se alcançar a qualidade máxima esperada pela simulação.

De fato, no cenário de alto nível de conhecimento dos pares (AC) estudado anteriormente, pela teoria da confiabilidade de componentes [Menascé e Almeida (2001)], quando por exemplo apenas 3 pares têm certo documento, a chance de o documento ser encontrado em algum par disponível no momento da consulta já é de quase 60%. Isto é, o limite teórico pode ser rapidamente atingido, pois basta que esses 3 pares façam cada qual uma consulta que retorne o documento e com isso cada um tenha uma réplica do documento. Os documentos restantes obtidos pela máquina de busca não interferem na qualidade dos documentos replicados pelo par.

No entanto, no cenário de baixo nível de conhecimento dos pares (BC), os documentos restantes obtidos pela máquina de busca interferem, pois como os documentos são similares entre si, a qualidade do documento diminui localmente. Isso faz com que os mesmos 3 pares não apenas tenham que ter cada um uma réplica do documento, mas também tenham que fazer diversas consultas diferentes para que documentos não similares sejam baixados e com isso a estatística local do documento melhore. Depois de um número suficiente de consultas disparadas na rede, o limite teórico deverá ser atingido.

Dessa forma, o modelo teórico proposto pode ser utilizado em cenários com diferentes níveis de conhecimento dos pares sobre os documentos da rede (AC e BC), embora o tempo de convergência possa ser substancialmente diferente. De fato, os resultados de simulação mostram que no cenário AC o limite teórico pode ser atingido muito mais rapidamente que no cenário BC: no primeiro, alguns pares precisam executar uma única consulta, enquanto que no último, esses mesmos pares precisam executar diversas consultas diferentes.

Em suma, a replicação dinâmica por similaridade apresenta-se como uma solução em potencial para atenuar o impacto do comportamento dinâmico dos pares na eficácia de máquinas de busca P2P. De fato, nossos resultados mostram que o impacto do comportamento dinâmico pode ser reduzido significativamente quando pares baixam apenas os primeiros $r = 10$ documentos retornados pela máquina de busca em resposta a uma consulta, mesmo em cenários pessimistas em que a maioria dos pares não está disponível e possuem um baixo nível de conhecimento sobre os documentos da rede. Além disso, a execução de consultas voltadas a diferentes tópicos de interesse pode contribuir para atenuar rapidamente a degradação na eficácia da busca de máquinas de busca P2P não estruturadas.

5.6.5 Desafios

Embora a estratégia de replicação dinâmica por similaridade potencialmente contribua para a eficácia de máquinas de busca P2P, alguns desafios devem ser considerados antes de se adotar essa solução.

Primeiro, o tempo de convergência dos resultados da busca para o valor máximo esperado

pode ser crítico para a viabilidade de máquinas de busca P2P. Particularmente em ambientes P2P descentralizados e não estruturados, os usuários possuem autonomia para definir o conjunto de documentos de seus pares conforme seus próprios interesses. Isso pode levar a situações como aquelas ilustradas na Figura 5.10, em que cada par precisaria submeter diversas consultas diferentes de forma a eliminar a topicidade de seu conjunto de documentos. No entanto, uma vez uniforme a distribuição de documentos entre os pares, a qualidade máxima esperada deve ser rapidamente atingida (ver Figura 5.9).

Uma alternativa para reduzir o tempo de convergência é adicionar um pouco de estrutura à rede P2P. Por exemplo, pares com mais recursos computacionais e maior disponibilidade poderiam ser eleitos *superpares*. Esses pares representariam um subconjunto de pares da rede P2P e seriam responsáveis, por exemplo, por redistribuir dinamicamente os documentos entre os pares ou mesmo por calcular e fornecer estatísticas IDF ao grupo. Essas estratégias poderiam ser adotadas por máquinas de busca P2P para reduzir de maneira mais eficiente os efeitos da topicidade dos pares.

Além disso, a replicação de documentos em máquinas de busca P2P não é natural para o usuário. Diferentemente de aplicações P2P de compartilhamento de arquivos, não se espera que usuários voluntariamente repliquem documentos quando estiverem utilizando uma máquina de busca P2P, mas que apenas naveguem pelas páginas apresentadas pela máquina de busca em resposta a uma consulta. Assim, uma vez que o comportamento dinâmico dos pares é um aspecto intrínseco a redes P2P e logo difícil de se evitar, a viabilidade de máquinas de busca P2P poderá depender fortemente do consentimento do usuário para a replicação local de páginas da Web.

Uma alternativa seria tornar a replicação de documentos transparente para o usuário. Idealmente, os usuários deveriam utilizar uma máquina de busca P2P da mesma forma que costumam utilizar uma máquina de busca tradicional, isto é, a partir de um navegador Web. Navegadores Web normalmente adotam alguma estratégia de *caching* com o intuito de melhorar a experiência de navegação do usuário. Para tanto, os navegadores armazenam cópias de páginas visitadas anteriormente pelo usuário para reduzir o tempo de resposta. Como usuários se sentem confortáveis com a estratégia de *caching* dos navegadores atuais, possivelmente também se sentirão confortáveis se os navegadores baixarem algumas das páginas obtidas em resposta às consultas à máquina de busca P2P em troca da oportunidade de obterem resultados de busca com maior qualidade.

Capítulo 6

Conclusões e Trabalhos Futuros

Com o crescente interesse na adoção de sistemas P2P como arcabouço para as futuras máquinas de busca na Web, entender os efeitos do comportamento dinâmico dos pares na eficácia da busca mostra-se essencial para avaliação da viabilidade desse novo tipo de aplicação. Nessa direção, analisamos o impacto desse aspecto na qualidade das respostas obtidas na busca na Web em P2P com diferentes níveis de conhecimento e autonomia dos pares. Nossos resultados indicam que o dinamismo dos pares revela-se um sério desafio para o desenvolvimento de máquinas de busca P2P reais.

Também observamos que a replicação de documentos obtidos pelo par em resposta a uma consulta (replicação dinâmica por similaridade) apresenta-se como uma solução em potencial para atenuar os efeitos da indisponibilidade dos pares. No entanto, enfatizamos a necessidade de mecanismos de incentivo específicos para aplicações de busca na Web em P2P. Tais mecanismos devem considerar que, diferentemente de aplicações P2P de compartilhamento de arquivos, não se espera que usuários baixem e compartilhem páginas Web de outros pares, mas que apenas naveguem por elas. Além disso, diferentemente de máquinas de busca centralizadas, não se espera que usuários acessem livremente o conjunto de documentos dos pares, mas apenas os documentos retornados à consulta, que o usuário mantenha a aplicação constantemente em execução e que as cópias locais estejam sempre acessíveis para a rede P2P.

Além da análise de novos mecanismos de incentivo, a *consistência* dos resultados e a *ação maliciosa* dos usuários (ver Seção 5.4), assim como o *tempo de convergência* dos resultados (ver Seção 5.6.5) também são aspectos cruciais de serem explorados para análise da viabilidade de máquinas de busca P2P, especialmente em redes P2P de grande escala.

Além disso, aspectos específicos de arquitetura de rede P2P (ver Seção 5.1) devem ser considerados, especialmente na avaliação de estratégias de replicação de conteúdo em cenários pessimistas. Estratégias eficientes de *caching* e atualização de documentos devem ser avaliadas em ambientes de busca P2P não estruturados, em que pares possuem restrições de armazenamento, tempo e uso de banda para manter versões consistentes dos documentos da rede. Estratégias de

poda de pares pouco promissores para satisfazer a uma consulta podem ser exploradas para reduzir o número de documentos irrelevantes da resposta. Essa estratégia pode não apenas melhorar a qualidade dos resultados, mas também aumentar a eficiência da busca, ao diminuir o número de pares promissores em que a consulta deve ser processada.

Apêndice A

Aspectos de Eficiência do Ambiente de Simulação

A.1 *Caching* de Listas Invertidas

Apresentamos algumas estratégias que se mostraram essenciais para garantir a eficiência do simulador. Avaliamos o tempo de execução do simulador para a WBR, em um cenário de máquina de busca P2P dinâmica com replicação e nível de conhecimento local dos pares sobre documentos da rede (BC). Cada simulação é realizada a partir de uma única consulta executada seqüencialmente em 5 pares escolhidos aleatoriamente. Cada par baixa e adiciona localmente um conjunto de documentos apresentados na resposta da consulta. Os resultados apresentados são uma média do tempo de execução (relógio) e de memória utilizada pelas diferentes simulações. O ambiente de simulação é executado sobre um sistema Linux Debian 2.6, que utiliza um processador Intel(R) Core(TM) 2 Quad, 2,40 GHz, 4096 KB de cache e 8 GB de memória RAM.

Devido a limitações do ambiente de simulação, apenas uma parte das coleções de teste pôde ser mantida em memória durante a reindexação. De fato, restrições de endereçamento da memória do sistema fazem com que no máximo 4 GB de memória sejam endereçáveis por processo em sistemas de 32 bits ($2^{32bits} = 4$ GB). Além disso, 25% desses 4 GB são reservados para o sistema (Linux). Considerando que apenas o simulador chega a ocupar cerca de 1 GB de memória em alguns cenários e que o tamanho das listas invertidas da coleção de teste WBR somam 3,6 GB no HD, optamos então por carregar as listas invertidas em memória sob demanda, até que o limite físico de memória fosse atingido. As listas subseqüentes são acessadas do HD mas não são mantidas em memória. No cenário de teste, observamos um tempo de simulação (de relógio) de cerca de 5 horas.

Avaliamos então algumas políticas para reposição das listas invertidas em memória. Consideramos inicialmente uma *cache* global, em que não há distinção entre documentos de diferentes pares. Avaliamos inicialmente a política de reposição LRU, em que as listas invertidas utiliza-

das mais recentemente são mantidas em memória. Essa política não se mostrou eficiente. Em seguida, avaliamos a política de reposição LFU, em que as listas invertidas utilizadas com mais freqüência são mantidas em memória. A adoção dessa política nos permitiu reduzir o tempo de simulação em 40%.

Em seguida, optamos por manter em memória apenas as listas invertidas do par sendo reindexado. Essa melhoria nos permitiu uma redução adicional de 25% no tempo de simulação.

Finalmente, consideramos que muita informação presente nas listas invertidas não precisaria ser carregada em memória. De fato, considerando a melhoria anterior, listas invertidas podem ser reduzidas em memória para conter apenas informação do par sendo reindexado. Além disso, reduzimos o número de acessos ao HD ao prevenir a replicação de listas invertidas em memória e ao aproveitar um mesmo acesso ao HD para extrair mais estatísticas sobre a coleção (TF e IDF). Essas melhorias mostraram-se tão efetivas que todas as listas invertidas utilizadas pelo par puderam ser carregadas em apenas 50% da memória disponível. Além disso, obtivemos uma redução adicional de 50% no tempo de simulação.

No entanto, mesmo diante de uma melhoria final no tempo de simulação de 81% no cenário de teste, o tempo total de execução do simulador é ainda crítico em cenários com um número mais significativo de execuções de consultas. Para avaliarmos um único cenário da WBR com 3000 execuções de consultas escolhidas aleatoriamente – em que em média cada uma das 50 consultas da coleção de teste é executada 60 vezes – o tempo de simulação exigido é em torno de 25 dias. Note que existem ainda diversos cenários que devem ser avaliados: para cada coleção de teste, TREC e WBR, avaliamos diferentes níveis de replicação em cenários com diferentes níveis de conhecimento dos pares (ver Seção 5.6).

A.2 Geração *Offline* de Índices Locais

Simulamos os índices de documentos dos pares a partir de um único índice (global) que possui todos os documentos da coleção. Assim, ao reindexar o conjunto de documentos de um par, todo o índice precisa ser percorrido em busca dos documentos pertencentes ao par, o que pode dispender muito tempo de simulação.

Uma alternativa mais eficiente seria converter o índice centralizado em N índices, onde N corresponde ao número de pares da máquina de busca P2P avaliada. Dessa forma, ao reindexar os documentos de um par, muito menos dados precisariam ser lidos do HD, já que as listas invertidas possuiriam apenas os documentos que efetivamente pertencessem ao par.

No entanto, projeções de uso de HD não recomendam a geração offline dos índices locais. De fato, experimentos mostraram que quando índices são gerados para os documentos de apenas 1.000 termos da coleção, o espaço ocupado no HD chegou a 1,7 GB. Quando aumentamos o número de termos para 10.000, o espaço ocupado chegou a 11 GB. No entanto, o índice centralizado da coleção WBR contém muito mais termos (2.669.965) e ocupa apenas 3,6 GB no HD.

Esse aumento considerável no espaço ocupado pelos índices locais acontece porque um mesmo termo pode ocorrer em diversos documentos e portanto em diversos índices locais. O crescimento explosivo no espaço ocupado pelos índices locais no HD parece então inviabilizar a adoção dessa solução.

Referências Bibliográficas

Acosta, W. e Chandra, S. (2005). Unstructured Peer-to-Peer Networks - Next Generation of Performance and Reliability. *IEEE INFOCOM Posters*, Miami, EUA.

Almeida, H.; Gonçalves, M.; Cristo, M. e Calado, P. (2007). A Combined Component Approach for Finding Collection-Adapted Ranking Functions Based on Genetic Programming. *30th International ACM Special Interest Group on Information Retrieval (SIGIR)*, Amsterdã, Países Baixos.

Atalla, F.; Miranda, D.; Almeida, J.; Gonçalves, M. A. e Almeida, V. (2008). Analyzing the Impact of Churn and Malicious Behavior on the Quality of Peer-to-Peer Web Search. *23rd Annual ACM Symposium on Applied Computing (ACM SAC)*, Fortaleza, Brasil.

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

Bender, M.; Michel, S.; Triantafillou, P.; Weikum, G. e Zimmer, C. (2006). P2P Content Search: Give the Web back to the People. *5th International Workshop on Peer-to-Peer Systems (IPTPS)*, Santa Bárbara, EUA.

Bender, M.; Michel, S.; Weikum, G. e Zimmer, C. (2005). Challenges of Distributed Search Across Digital Libraries. *8th DELOS Workshop: System Architecture & Information Access*, Dagstuhl, Alemanha.

Bergman, M. K. (2001). The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing (JEP)*, 7(1).

Brin, S. e Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *7th International Conference on World Wide Web (WWW)*, Brisbane, Austrália.

Calado, P. (1999). The WBR-99 Collection: Description of the WBR-99 Web Collection Data-Structures and File Formats. Technical report, Laboratory for Treating Information (LATIN), Belo Horizonte, Brasil.

Callan, J. e Connell, M. (2001). Query-based Sampling of Text Databases. *ACM Transactions on Information Systems (TOIS)*, 19(2).

Cohen, E. e Shenker, S. (2002). Replication Strategies in Unstructured Peer-to-Peer Networks. *ACM Special Interest Group on Data Communication (SIGCOMM)*, Pitsburgo, EUA.

Crespo, A. e Garcia-Molina, H. (2002). Semantic Overlay Networks for P2P Systems. Technical report, Universidade de Stanford, Palo Alto, EUA.

Cuenca-Acuna, F. M.; Peery, C.; Martin, R. P. e Nguyen, T. D. (2003). PlanetP: Using Gossiping to Build Content Addressable Peer-to-Peer Information Sharing Communities. *12th IEEE International Symposium on High-Performance Distributed Computing (HPDC)*, Seattle, EUA.

Demers, A.; Greene, D.; Hauser, C.; Irish, W.; Larson, J.; Shenker, S.; Sturgis, H.; Swinehart, D. e Terry, D. (1987). Epidemic Algorithms for Replicated Database Maintenance. *6th ACM Symposium on Principles of Distributed Computing*, Vancouver, Canadá.

Doulkeridis, C.; Norvag, K. e Vazirgiannis, M. (2006). The SOWES Approach to P2P Web Search Using Semantic Overlays. *15th World Wide Web Conference (WWW)*, Edinburgo, Escócia.

Dunn, R. J.; Zahorjan, J.; Gribble, S. D. e Levy, H. M. (2005). Presence-Based Availability and P2P Systems. *5th IEEE International Conference on Peer-to-Peer Computing (P2P)*, Konstanz, Alemanha.

Fetterly, D.; Manasse, M. e Najork, M. (2004). Spam, Damn Spam, and Statistics: Using Statistical Analysis to Locate Spam Web Pages. *7th WebDB Workshop*, Paris, França.

Fishwick, P. A. (1992). SimPack: Getting Started with Simulation Programming in C and C++. *24th Winter Simulation Conference (WSC)*, Arlington, VA, USA.

Google, Inc., Yahoo, Inc. e Microsoft Corporation (2008). The Sitemaps Protocol. `http://www.sitemaps.org/`. Página visitada em 07 de junho de 2008.

Gopalakrishnan, V.; Bhattacharjee, B. e Keleher, P. (2006). Distributing Google. *2nd IEEE NetDB Workshop*.

Gulli, A. e Signorini, A. (2005). The Indexable Web is more than 11.5 Billion Pages. *14th World Wide Web Conference (WWW), Special Interest Tracks and Posters*, Chiba, Japão.

Gyongyi, Z. e Garcia-Molina, H. (2005). Web Spam Taxonomy. *1st AIRWeb Workshop*, Chiba, Japão.

Jain, R. (1991). *The Art of Computer Systems Performance Analysis*. John Wiley & Sons, São Francisco, EUA.

Kishida, K. (2005). Property of Average Precision and its Generalization: an Examination of Evaluation Indicator for Information Retrieval Experiments. Technical report, National Institute of Informatics (NII), Tóquio, Japão.

Kulbak, Y. e Bickson, D. (2005). The Emule Pprotocol Specification. Technical report, Hebrew University of Jerusalem, Jerusalém, Israel.

Lewandowski, D. e Mayr, P. (2006). Exploring the Academic Invisible Web. *Library Hi Tech Journal*, 24(4).

Lu, J. e Callan, J. (2003). Content-based Retrieval in Hybrid Peer-to-Peer Networks. *12th Conference on Information and Knowledge Management (CIKM)*, Nova Orleans, EUA.

Luo Si, J. C. (2005). Modeling Search Engine Effectiveness for Federated Search. *28th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brasil.

Lv, Q.; Cao, P.; Cohen, E.; Li, K. e Shenker, S. (2002). Search and Replication in Unstructured Peer-to-Peer Networks. *16th ACM International Conference on Supercomputing (ICS)*, Nova Iorque, EUA.

Maymounkov, P. e Mazieres, D. (2002). Kademlia: A Peer-to-Peer Information System Based on the XOR Metric. *1st International Workshop on Peer-to-Peer Systems (IPTPS)*, Cambridge, EUA.

Menascé, D. A. e Almeida, V. A. F. (2001). *Capacity Planning for Web Services: Metrics, Models, and Methods*. Prentice Hall.

Millward Brown (2008). Flash Player Penetration Survey. `http://www.adobe.com/products/player_census/flashplayer/`. Página visitada em 07 de junho de 2008.

Nelson, M. L.; Smith, J. A. e del Campo, I. G. (2006). Efficient, Automatic Web Resource Harvesting. *8th annual ACM International Workshop on Web Information and Data Management (WIDM)*, Arlington, EUA.

Ntoulas, A. e Cho, J. (2007). Pruning Policies for Two-Tiered Inverted Index with Correctness Guarantee. *30th International ACM Special Interest Group on Information Retrieval (SIGIR)*, Amsterdã, Países Baixos.

Pouwelse, J. A.; Garbacki, P.; Epema, D. H. J. e Sips, H. J. (2005). The Bittorrent P2P File-Sharing System: Measurements and Analysis. *4th International Workshop on Peer-to-Peer Systems (IPTPS)*, Ithaca, EUA.

Rhea, S. C.; Geels, D.; Roscoe, T. e Kubiatowicz, J. (2004). Handling Churn in a DHT. *USENIX Annual Technical Conference*, Boston, EUA.

Rijsbergen, C. J. V. (1979). *Information Retrieval*. Butterworth-Heinemann.

Ripeanu, M. (2001). Peer-to-Peer Architecture Case Study: Gnutella Network. *First International Conference on Peer-to-Peer Computing*, Linköping, Suécia.

Schroeder, B. e Gibson, G. A. (2006). A Large-Scale Study of Failures in High-Performance Computing Systems. *International Conference on Dependable Systems and Networks (DSN)*, Filadélfia, EUA.

Stutzbach, D. e Rejaie, R. (2006). Understanding Churn in Peer-to-Peer Networks. *6th ACM SIGCOMM Internet Measurement Conference (IMC)*, Rio de Janeiro, Brasil.

Voorhees, E. M. e Harman, D. (1999). Overview of the Eighth Text REtrieval Conference (TREC-8). *8th Text REtrieval Conference (TREC-8)*, Gaithersburg, EUA.

Wu, L.; Akavipat, R. e Menczer, F. (2005). 6S: Distributing Crawling and Searching Across Web Peers. *Web Technologies, Applications, and Services (WTAS)*, Calgary, Canadá.

Yilmaz, E. e Aslam, J. A. (2006). Estimating Average Precision with Incomplete and Imperfect Judgments. *15th ACM International Conference on Information and Knowledge Management (CIKM)*, Arlington County, EUA.

Yu, W.; Chellappan, S. e Xuan, D. (2005). P2P/Grid-based Overlay Architecture to Support VoIP Services in Large-Scale IP Networks. *International Journal of Future Generation Computer Systems (FGCS)*, 21(1).

Zhu, Y. (2006). Enhancing Search Performance on Gnutella-Like P2P Systems. *IEEE Transactions on Parallel and Distributed Systems*, 17(12).