Universidade Federal do Amazonas
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Programa de Pós-Graduação em Informática

Modelo Baseado em Hipergrafos para Análise de Apontadores em Coleções Web

Klessius Renato Berlt

Manaus – Amazonas
Abril de 2008

Klessius Renato Berlt

Modelo Baseado em Hipergrafos para Análise de Apontadores em Coleções Web

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Departamento de Ciência da Computação da Universidade Federal do Amazonas, como requisito parcial para obtenção do Título de Mestre em Informática. Área de concentração: Recuperação de Informação.

Orientador: Prof. Dr. Edleno Silva de Moura

Klessius Renato Berlt

Modelo Baseado em Hipergrafos para Análise de Apontadores em Coleções Web

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Departamento de Ciência da Computação da Universidade Federal do Amazonas, como requisito parcial para obtenção do Título de Mestre em Informática. Área de concentração: Recuperação de Informação.

Banca Examinadora

Prof. Dr. Edleno Silva de Moura – Orientador
Departamento de Ciência da Computação – UFAM/PPGI

Prof. Dr. Marco Antonio Pinheiro de Cristo
Departamento de Ciência da Computação – UFAM/PPGI

Profa. Dra. Viviane Moreira Orengo
Instituto de Informática – UFRGS

Manaus – Amazonas
Abril de 2008

Aos meus avos, Atílio, Iracema, Fredolino e Rodolfa.

Agradecimentos

Agradeço primeiramente aos meus pais, Loreni Berlt e Alda Berlt, por sempre incentivarem e investirem na minha educação; à minha irmã, Káren Berlt, pelo companheirismo e amizade sempre presentes e à minha namorada, Lisandra Santos, por estar sempre ao meu lado mesmo nos momentos mais difíceis.

Agradeço ao meu orientador, Edleno Silva de Moura, pela chance oferecida e pelos ensinamentos transmitidos durante este período e a Thierson Couto, Marco Cristo e Nívio Ziviani pela colaboração essencial.

Agradeço aos amigos que me acompanharam durante o mestrado; André Carvalho, Bruno Araújo, Thomas Silva, Karane Vieira, Roberto Oliveira e Felipe Mesquita; pela ajuda prestada nos momentos complicados e pelas boas conversas e momentos de descontração.

Agradeço a todos os amigos que me ajudaram e à Elienai Nogueira e toda a secretaria do Departamento de Ciência da Computação pelo grande auxílio prestado.

Agradeço também ao Universo On Line e à FAPEAM pelo apoio fornecido para a realização deste trabalho.

"É bom sonhar, mas é melhor sonhar e trabalhar."

Thomas Robert Gaines

Resumo

Neste trabalho propomos um modelo para representar a Web através de um hipergrafo direcionado (em vez de um grafo) onde há conexões entre blocos de páginas e páginas. O hipergrafo da Web pode ser derivado do grafo da Web agrupando as páginas em blocos sem intersecção entre si e utilizando os apontadores entre páginas de blocos distintos para criar as hiperarestas. Uma hiperaresta conecta um bloco de páginas a uma única página de outro bloco e tem como objetivo oferecer informação mais confiável para os métodos de análise de apontadores. Criamos versões dos métodos Pagerank e Indegree para o modelo de hipergrafo, chamadas de HiperPagerank e HiperIndegree respectivamente. Comparamos a versão dos algoritmos baseados em hipergrafos com suas versões baseadas em grafos considerando a combinação da reputação estimada pelos algoritmos com o conteúdo textual da página e o texto de âncora. Os resultados experimentais obtidos mostram que o HiperPagerank e o HiperIndegree obtêm melhor desempenho em relação aos algoritmos originais baseados em grafos. Também mostramos que as versões para hipergrafo dos algoritmos são menos afetadas por spam.
Palavras-chave: Recuperação de Informação, Máquinas de Busca, Análise de Apontadores.

Abstract

In this work we propose a model to represent the Web as a directed hypergraph (instead of a graph) where links connect pairs of disjoint sets of pages. The Web hypergraph can be derived from the Web graph by dividing the set of pages into non-overlapping blocks and using the links between pages of distinct blocks to create hyperarcs. A hyperarc connects a block of pages to a single page with the goal of providing more reliable information for link analysis methods. We use the hypergraph model to create the hypergraph versions of the PageRank and InDegree algorithms, referred to as HyperPageRank and HyperInDegree, respectively. We compared the original algorithms with their hypergraph versions, considering the combination of page reputation, textual content of pages and anchor text. Experimental results using three distinct web collections show that the HyperPageRank and HyperInDegree algorithms yield better results when compared to their original graph versions. We also show that the hypergraph versions of the algorithms were slightly less affected by spamming.
Keywords: Information Retrieval, Search engines, Link Analisys.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Máquinas de busca para Web são sistemas de recuperação de informação desenvolvidos para auxiliar usuários a encontrar informação útil em páginas web. Devido à grande quantidade de páginas publicadas atualmente, estes sistemas devem ser extremamente eficientes e precisos ao extrair informação destas páginas. Eles buscam obter o máximo de dados possível sobre cada página para inferir quais delas atendem melhor às necessidades de cada usuário.

Para obter informações sobre uma página, diversas evidências são consideradas, dentre elas destaca-se a reputação da página. Uma página com boa reputação é uma página bem conhecida dentro da comunidade e geralmente de boa qualidade. Em casos onde a página desejada é a página principal de um sítio. Por exemplo, em uma consulta por "banco do brasil", onde o usuário deseja a página principal do Banco do Brasil, a reputação da página é de grande ajuda.

As máquinas de busca atuais utilizam algoritmos que analisam a estrutura de apontadores da Web para estimar a reputação de cada página presente em sua base de dados. Estas técnicas são chamadas de técnicas de análise de apontadores. As técnicas de análise de apontadores atuais utilizam um modelo da Web baseado em grafos direcionados, onde as páginas são os vértices e os apontadores entre elas são as arestas.

Muitos métodos de análise de apontadores propostos na literatura nos últimos anos [2, 3, 14, 13, 16] assumem que um apontador de uma página qualquer A para outra página qualquer B é um voto de A na qualidade de B. O primeiro método a explorar esta intuição foi o Indegree, que considera que a reputação de uma página é proporcional à quantidade de apontadores que ela recebe. Seguindo esta idéia, foram propostos métodos mais sofisticados como Pagerank [3] e HITS [13], que levam em conta não apenas o número de apontadores que uma página recebe,

mas também a qualidade destes apontadores.

Entretanto, nem todos os apontadores representam votos de confiança na qualidade das páginas apontadas. Muitos deles possuem funções distintas que não se enquadram neste conceito. Apontadores criados para fins navegacionais ou até mesmo para gerar spam são comuns na maioria das bases de dados web e podem levar o algoritmo de análise de apontadores a resultados equivocados. Outro fator que pode interferir negativamente no resultado final destes algoritmos é a presença de informação redundante, muitas vezes identificada como votos distintos. Por exemplo, se muitas páginas de um sítio apontam para uma página p, então todas elas irão contribuir para aumentar a reputação de p. Acreditamos que tal influência deve ser amenizada, pois não reflete necessariamente a opinião da comunidade web e sim de um subgrupo específico com opiniões relacionadas entre si. Apontadores de um único sítio não devem ser suficientes para estimar se uma página é popular ou não. Uma possível solução para este problema é a utilização de modelos que reduzam o impacto de apontadores originados em um mesmo sítio ou grupo de páginas co-relacionadas, dando mais importância à diversidade da origem dos apontadores recebidos pela página.

Para extrair melhor a informação presente nos apontadores das páginas da Web, este trabalho propõe uma nova modelagem para a Web, representado-a através de um hipergrafo direcionado em vez de um grafo direcionado. Um hipergrafo direcionado é uma generalização do grafo direcionado, onde as relações estabelecidas deixam de ser exclusivamente entre pares de vértices (arestas) e passam a ser entre pares de conjuntos de vértices (hiperarestas). O hipergrafo que representa um conjunto de páginas web pode ser derivado a partir do grafo que representa este mesmo conjunto de páginas agrupando páginas muito relacionadas em blocos sem intersecção entre si e mapeando arestas para hiperarestas da seguinte forma: sendo um bloco qualquer e p uma página qualquer, existe uma hiperaresta de para p se, somente se, existir pelo menos um apontador de b para p, onde b . O critério adotado para agrupar as páginas em blocos pode variar de acordo com o objetivo final da análise. A principal diferença encontrada entre esta nova modelagem e a tradicional modelagem com grafos é que a reputação da página passa a estar associada mais diretamente à variedade da origem dos apontadores que ela recebe do que à quantidade de apontadores que ela recebe.

Uma característica interessante do modelo de hipergrafo é a possibilidade de definir o nível de relacionamento necessário para que as páginas pertençam a um mesmo bloco. Por exemplo, pode

se definir que os blocos sejam formados por páginas pertencentes a um mesmo domínio, desta forma, as arestas entre páginas serão mapeadas como hiperarestas que conectam domínios a páginas. Neste exemplo, a reputação de uma página terá uma relação direta com a quantidade de domínios distintos que apontam para ela. Esta é uma característica importante porque permite que sejam adotados diferentes critérios para criação dos blocos de acordo com a base de dados utilizada. Enquanto blocos muito grandes causam muita perda de informação, blocos excessivamente pequenos podem não ser suficientes para resolver o problema dos apontadores redundantes ou de baixa qualidade.

Ao representar a Web através de um hipergrafo, novos algoritmos para extrair informação desta estrutura devem ser desenvolvidos. Vários algoritmos tradicionais de análise de apontadores em grafos podem ser adaptados para serem utilizados no hipergrafo. Neste trabalho adaptamos os algoritmos Indegree e Pagerank chamados respectivamente de Hiperindegree e Hiperpagerank.

Neste trabalho realizamos experimentos em duas coleções distintas, onde comparamos os algoritmos tradicionais e suas versões de hipergrafos e verificamos que o Hiperindegree e Hiperpagerank obtêm ganhos significativos em uma coleção, enquanto mantém praticamente os mesmos resultados em outra. Também comparamos a qualidade dos algoritmos em uma base previamente tratada para eliminar ruídos a fim de verificar o impacto dos ruídos na qualidade final da reputação estimada por eles. Por fim, analisamos a suscetibilidade a spam dos algoritmos para hipergrafos em relação aos algoritmos tradicionais.

Este trabalho encontra-se organizado da seguinte forma. No Capítulo 2 são descritos brevemente alguns dos conceitos básicos mais importantes para se obter uma melhor compreensão deste trabalho. O Capítulo 3 descreve um novo modelo para representar a Web baseado em hipergrafos além de alguns novos algoritmos que utilizam este modelo para estimar a reputação de páginas Web. No Capítulo 4 são descritos os experimentos realizados para comprovar a qualidade do modelo proposto. Por fim, o Capítulo 5 apresenta as conclusões obtidas e sugere possíveis direções para trabalhos futuros.

1.1 Trabalhos Relacionados

A análise de apontadores como estratégia para melhorar resultados de máquinas de busca da Web foi proposta inicialmente em [2], onde comenta-se que a quantidade de apontadores que uma página recebe (Indegree) é uma maneira de estimar a sua reputação. Este trabalho foi o primeiro a comentar explicitamente a idéia de que apontadores entre páginas podem ser considerados votos na qualidade das mesmas. O Indegree (ou grau de entrada) considera que a reputação de uma página é igual ao número de apontadores que ela recebe. O Indegree IN de uma página p em um grafo $G = (V; E)$ é calculado da seguinte maneira:

$$IN(p) = \sum_{(q,p)\in E} 1 \tag{1.1}$$

Onde $(q; p)$ é uma aresta que parte do vértice q para o vértice p e também $q; p\ 2\ V$.

O Indegree, apesar de apresentar resultados interessantes, mostrou-se muito suscetível a manipulações. Por exemplo, para aumentar a reputação de uma página p, basta que seja criado um conjunto de apontadores partindo de quaisquer outras páginas para p. Este fato estimulou a criação de métodos de análise de apontadores mais robustos, onde se destacaram o HITS [13] e o Pagerank [3].

No HITS [13], o conceito de reputação é desmembrado em dois conceitos que são definidos recursivamente: autoridade e hub. Boas autoridades são páginas apontadas por bons hubs e bons hubs são páginas que apontam para boas autoridades. Além disso, o autor assume que diferentes assuntos possuem diferentes hubs e autoridades, que devem ser calculados a cada consulta realizada. Desta forma, o HITS utiliza um subgrafo no contexto de cada consulta. Este subgrafo de contexto é formado pela união de três conjuntos de páginas: a) conjunto inicial S, formado pelas páginas da coleção que contém os termos da consulta; b) conjunto de *outlinks*, formado por todas as páginas que são apontadas por pelo menos uma página s, onde $s\ 2\ S$; c) conjunto de *inlinks*, formado por todas as páginas que apontam para qualquer página s, onde $s\ 2\ S$. A união destes três conjuntos de páginas (S, *inlinks* e *outlinks*) forma o subgrafo de contexto que será utilizado pelo HITS para estimar a reputação (hub e autoridade) de cada página.

Máquinas de busca reais recebem milhões de consultas diferentes a cada dia. Isto torna os custos computacionais de métodos *online* como o HITS muito altos, fazendo com que a sua

utilização seja difícil nestes cenários. Desta forma, métodos como o Pagerank, que permitem que a reputação das páginas seja pré-processada e armazenada previamente, ganham popularidade.

O Pagerank [3] é um dos métodos de análise de apontadores de maior sucesso até hoje. Ele propõe-se a modelar o comportamento de um usuário navegando aleatoriamente pelo grafo da Web, seguindo suas arestas (apontadores) para ir de um vértice (página) a outro. Além da possibilidade de seguir os apontadores da página onde ele se encontra, o usuário ainda pode escolher qualquer página aleatoriamente com uma probabilidade constante. Esta probabilidade é chamada de *dampening factor*. O valor de Pagerank de uma página é então a probabilidade deste usuário chegar a esta página seguindo os apontadores que levam a ela ou escolhendo-a aleatoriamente. Este método se diferencia do Indegree por levar em conta a qualidade dos apontadores que apontam para as páginas, não apenas a quantidade de apontadores que apontam para elas. Neste modelo, um apontador originado em uma página com boa reputação tem valor superior ao de um apontador que parte de uma página com má reputação. Desta forma, diferentemente do Indegree, uma página pode ter uma boa reputação mesmo que receba poucos apontadores, desde que eles sejam oriundos de páginas com alto valor de Pagerank. O valor de Pagerank de uma página p em um grafo $G = (V; E)$ é calculado da seguinte maneira:

$$PR(p) = \left[(1 \quad c) \quad \sum_{q \in I(p)} \frac{PR(q)}{|O(q)|} \right] + \frac{c}{|V|} \tag{1.2}$$

onde c é o *dampening factor*, $|O(q)|$ é o número de páginas apontadas pela página q, $I(p)$ é o conjunto das páginas que apontam para p, e $|V|$ é o número de páginas na coleção.

Em [11], os autores buscam utilizar a informação da consulta submetida pelo usuário para computar um valor de reputação para cada consulta. Porém, ao contrário do HITS [13], este valor não é computado durante o processamento da consulta. Um conjunto de tópicos considerados representativos em relação à coleção é elaborado, então a reputação de cada página em relação a cada um destes tópicos é calculada e estes valores são armazenados. Desta forma, cada página possui uma reputação para cada um dos tópicos selecionados. Estes valores são combinados levando em conta a similaridade da consulta com cada um dos tópicos, gerando uma reputação final para cada página. Apesar deste método apresentar bons resultados, a seleção de um conjunto de tópicos suficientemente representativo para a coleção é um problema em aberto que dificulta a sua adoção em máquinas de busca comerciais.

Alguns trabalhos tentam qualificar o apontador de acordo com a sua origem. O Blockrank [5] agrupa as páginas em blocos e classifica os apontadores em duas categorias: *intra-links* (ou apontadores internos), que são apontadores entre páginas do mesmo bloco, e *inter-links* (ou apontadores externos), que são apontadores entre páginas de blocos distintos. O autor atribui pesos para cada apontador de acordo com a sua classificação e então computa uma versão ponderada do Pagerank.

Seguindo esta idéia, o Hierarchical Rank [16] também agrupa as páginas em blocos. Porém, após realizar este agrupamento, adota uma estratégia diferente para estimar as reputações. A reputação de cada bloco é computada através de um Siterank (versão do Pagerank que calcula a reputação de blocos de páginas). Esta reputação então é propagada para cada uma das páginas que compõem o bloco através de um esquema baseado em propagação de calor.

Apesar dos métodos descritos apresentarem bons resultados, nenhum trabalho que represente blocos de páginas e páginas individuais em um mesmo modelo foi encontrado na literatura. Uma das vantagens desta representação é a facilidade para adaptar algoritmos tradicionais desenvolvidos para o modelo de grafos. Podemos também modelar conjuntos de páginas fortemente relacionadas, como comunidades de blogs ou fotologs em blocos únicos, evitando assim que seus apontadores influenciem muito nos resultados. Além disso, este modelo pode utilizar diferentes definições para a criação das partições de páginas. É possível até mesmo assumir que cada página é um bloco, tornando o modelo de hipergrafo idêntico ao modelo de grafos. A modelagem utilizando hipergrafos permite que adaptemos facilmente um grande número de algoritmos de análise de apontadores existentes na literatura.

1.2 Contribuições

Medidas de reputação de páginas web são muito importantes para ajudar usuários a encontrar informações relevantes na Web. Esta importância levou ao desenvolvimento de diversos algoritmos cujo objetivo é estimar a reputação de páginas em coleções web. Entretanto, a grande maioria destes algoritmos modela a Web através de grafos.

Este trabalho propõe um novo modelo para representação da Web através de hipergrafos. Neste modelo, páginas muito relacionadas são agrupadas em blocos e a relação entre estes blocos e páginas individuais da coleção é representada através de hiperarestas.

Para estimar a reputação das páginas de uma coleção web utilizando hipergrafos, dois algoritmos tradicionais de análise de apontadores, Indegree e Pagerank, foram adaptados. Chamamos os novos algoritmos de Hiperindegree e Hiperpagerank. Estes algoritmos são capazes de estimar a reputação de uma página em uma coleção através da análise das conexões presentes no hipergrafo que representa esta coleção.

Como será visto no decorrer da dissertação, algoritmos utilizando a modelagem da Web através de hipergrafos obtêm resultados superiores às suas versões para grafos em diversos cenários.

Capítulo 2

Conceitos Básicos

Neste capítulo apresentaremos alguns conceitos básicos necessários para uma melhor compreensão deste trabalho. Como, por exemplo, o conceito de máquinas de busca e análise de apontadores.

2.1 Máquinas de Busca

Máquinas de busca são ferramentas construídas para auxiliar usuários a encontrar informações na Web. Uma máquina de busca recebe como entrada uma consulta que expressa a informação que o usuário deseja e retorna um conjunto de documentos que devem possuir esta informação. Os documentos retornados devem estar ordenados de acordo com a sua relevância em relação à consulta. Desta forma, os documentos mais relevantes são os primeiros documentos mostrados para o usuário.

Na maioria das máquinas de busca para Web, os documentos são páginas web e as consultas são compostas de termos que descrevem a informação desejada pelo usuário. Descobrir os documentos que atendem às necessidades de cada usuário não é uma tarefa trivial. As consultas submetidas são formas muito reduzidas de expressar a informação desejada, logo oferecem pouca informação sobre as páginas que atenderiam melhor à necessidade de informação do usuário. As máquinas de busca utilizam várias evidências para tentar inferir com a maior precisão possível qual a relevância de cada página para o usuário. Estas evidências podem ser informações presentes nas páginas ou na coleção utilizada. Dentre elas, destacam-se as seguintes:

Texto da Página: todo o texto que pode ser visualizado pelo usuário ao acessar a página.

Excluem-se, por exemplo, tags html.

Texto de Âncora: o texto de âncora de uma página é formado pela concatenação de todos os termos presentes nos apontadores que apontam para ela. Este texto mostra como a página é descrita na coleção.

Apontadores: apontadores (ou *links*) são utilizados para estimar a reputação das páginas na coleção.

Por coleção, entende-se o conjunto de documentos nos quais a máquina de busca irá procurar pela informação desejada.

Os algoritmos utilizados para extrair informações de cada evidência podem variar de acordo com a máquina de busca.

2.1.1 Modelo Vetorial

O Modelo Vetorial [15] é um modelo clássico para representação de documentos através de vetores em um espaço vetorial. Apesar de ter sido proposto há mais de 30 anos, o modelo vetorial ainda é amplamente utilizado com poucas modificações. Mapear os documentos para vetores facilita realização de algumas operações, principalmente o cálculo da similaridade entre dois documentos, que passa a ser representada pelo co-seno entre os seus vetores correspondentes.

Na utilização do modelo vetorial em máquinas de busca para Web, as páginas da coleção e as consultas submetidas são representadas por vetores e a similaridade entre uma consulta q e uma página p é calculada como o co-seno do ângulo formado pelos vetores que representam estas páginas.

É importante observar que o modelo vetorial pode ser utilizado tanto para calcular a similaridade entre uma consulta e o conteúdo de uma página quanto entre uma consulta e o texto de âncora de uma página.

2.1.2 Análise de Apontadores

Métodos de análise de apontadores são utilizados para estimar a reputação das páginas de uma coleção web através dos apontadores existentes nesta coleção. Estes métodos baseiam-se na hipótese de que ao criar um apontador de uma página A para uma outra página B, o autor da página A está fornecendo um indício de que ele acha o conteúdo da página B de alguma forma

interessante e recomenda a sua visitação. Desta forma, a relação entre todas as páginas de uma coleção pode ser analisada para estimar quais são as páginas mais "interessantes"ou "de melhor reputação".

Esta informação sobre a reputação pode ser útil de várias formas. Coletores de máquinas de busca utilizam a reputação da página para decidir a freqüência de atualização das páginas, mantendo as mais interessantes sempre atualizadas. Outra aplicação freqüente é o uso do resultado da análise como um dos elementos para determinar a ordenação de respostas em máquinas de busca. Quando um usuário realiza uma busca, a reputação é utilizada como uma das evidências para estimar o grau de relevância das páginas. As páginas de melhor reputação possuem preferência na ordem em que elas serão mostradas ao usuário, pois tendem a possuir mais qualidade que as de pior reputação. Neste trabalho utilizamos este último cenário para avaliar a qualidade dos métodos.

Métodos tradicionais de análise de apontadores utilizam grafos direcionados para modelar o conjunto de páginas cujas relações serão analisadas. Nestes grafos, os vértices representam as páginas e as arestas representam os apontadores. Geralmente, a reputação das páginas está diretamente ligada à quantidade de apontadores que ela recebe. A análise de apontadores utilizada para auxiliar na ordenação de respostas em máquinas de busca pode ser de dois tipos: local e global.

Na análise de apontadores local [13], um subgrafo de contexto é montado para cada consulta realizada. Este subgrafo é então utilizado para estimar a reputação das páginas no contexto da consulta. O subgrafo de contexto é composto a partir das páginas que possuem os termos presentes na consulta, ampliando-se o conjunto com as páginas apontadas por elas e as páginas para as quais elas apontam. Esta pequena coleção é então analisada buscando estimar a reputação de cada página. Os métodos locais possuem a vantagem de estimar a reputação das páginas no contexto da consulta realizada. Isso permite que páginas que possuam uma reputação boa em um contexto e ruim em outros sejam adequadamente tratadas. Entretanto, os custos computacionais obtidos com a tarefa de criação e processamento do subgrafo de contexto a cada consulta tornam onerosa a utilização de tais métodos em coleções de máquinas de busca reais.

Nos métodos globais [3], o grafo completo da coleção de páginas é processado uma única vez e apenas um valor de reputação é atribuído a cada página, independentemente da consulta realizada. Desta forma, os valores pré-processados podem ser armazenados previamente e

quando a consulta é realizada, o valor de reputação de cada página já está computado. Apesar deste método não levar em conta o tópico pesquisado para estimar a reputação, seu custo computacional faz com que ele seja muito mais viável que os métodos locais.

2.2 Tipos de Consulta

As consultas submetidas pelos usuários a máquinas de busca podem ser classificadas de acordo com o seu propósito [4]. Apesar de existirem outros meios de classificar consultas, esta abordagem é amplamente utilizada dada a sua importância na escolha das métricas de avaliação da qualidade das respostas obtidas em cada sistema. As consultas são classificadas em três categorias distintas:

Consultas Navegacionais: São consultas onde o usuário deseja uma página ou sítio específico como resposta. Por exemplo, ao buscar por "banco do brasil", o usuário provavelmente deseja obter como resposta o sítio do banco do brasil e não informações sobre o Banco do Brasil.

2.3 Métricas de Avaliação

Os resultados fornecidos por máquinas de busca podem atender ou não aos anseios do usuário. Para medir a qualidade de um resultado, diferentes métricas são propostas. A escolha da melhor métrica a ser utilizada depende diretamente do tipo de consulta.

2.3.1 MAP - *Mean Average Precision*

O MAP é uma métrica utilizada para avaliar a qualidade das máquinas de busca quando processam consultas informacionais. O valor de MAP de um sistema é um número real entre 0 e 1 que representa a precisão média deste sistema. O cálculo do MAP utiliza diretamente dois conceitos muito comuns na área de recuperação de informação: precisão e revocação.

A precisão de um sistema é definida como a sua capacidade de mostrar documentos relevantes para o usuário. Um sistema com precisão alta retorna muito mais documentos relevantes para o usuário do que documentos irrelevantes, enquanto que um sistema com precisão baixa retorna um maior número de documentos irrelevantes. Ou seja, a precisão é a quantidade relativa de documentos relevantes à consulta retornados para o usuário, como mostra a Equação 2.1:

$$Precisao = \frac{DocumentosRelevantes \setminus DocumentosRetornados}{DocumentosRetornados} \tag{2.1}$$

Exemplo: Se um sistema retorna 10 documentos, dos quais apenas 7 são relevantes para o usuário, a precisão deste sistema é de 70%.

A revocação mede a capacidade do sistema de busca encontrar na coleção os documentos com a informação que o usuário deseja. Sistemas com altos índices de revocação mostram para o usuário a maioria dos documentos relevantes à sua consulta presentes na coleção. A revocação é calculada dividindo-se a quantidade de documentos relevantes retornados pelo total de documentos relevantes da coleção, como descrito na Equação 2.2:

$$Revocacao = \frac{DocumentosRelevantes \setminus DocumentosRetornados}{DocumentosRelevantes} \tag{2.2}$$

Exemplo: Se uma consulta possui 20 documentos relevantes em uma coleção e o sistema retorna 10 deles, a revocação alcançada foi de 50%.

Como a precisão e a revocação são duas medidas distintas que avaliam a qualidade de um

sistema, é comum o uso de métricas que relacionem ambas, como o MAP, para resumir em um único valor a qualidade do sistema. Para calcular o MAP de um sistema, deve-se calcular a precisão média para cada consulta individualmente.

Para isso, primeiramente deve-se calcular para cada resposta oferecida à consulta, qual a precisão e a revocação alcançadas. Assim teremos uma curva de *precisao x revocacao* contendo os níveis de precisão para cada um dos níveis de revocação alcançados. Para padronizar as diferentes curvas geradas, convencionou-se interpolá-las para 11 pontos de revocação (0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% e 100%). Desta forma, a precisão média de uma consulta é o valor médio de precisão nestes 11 pontos. O MAP é então calculado como sendo a média da precisão média de todas as consultas, como descrito na Equação 2.3:

$$MAP(S;Q) = \frac{\sum_{q \in Q} AP(q)}{jQj} \tag{2.3}$$

Onde $MAP(S)$ é o valor de MAP do sistema S ao utilizar o conjunto de consultas Q, $AP(q)$ é o valor de precisão média para a consulta q e jQj é o número de consultas utilizadas para avaliar o sistema. Desta forma, sistemas com valores altos de MAP são sistemas que mantém a precisão alta mesmo quando recuperam um alto número de documentos relevantes.

Métricas como precisão, revocação e MAP são utilizadas para avaliar resultados de consultas informacionais, onde o usuário busca alguma informação e não uma página ou serviço específico, logo quanto mais páginas relevantes apresentadas ao usuário melhor.

2.3.2 MRR - *Mean Reciprocal Rank*

MRR é uma medida utilizada para avaliar resultados onde apenas uma resposta correta já é suficiente para atender às necessidades do usuário. Por isso, é muito utilizada para avaliar a qualidade de sistemas ao processar consultas navegacionais. Máquinas de busca apresentam, como resposta a uma consulta, uma lista de páginas ordenada de acordo com a probabilidade de cada página atender à necessidade de informação do usuário. Boas respostas para consultas navegacionais são listas onde a resposta correta encontra-se próxima do topo. Desta forma, o valor de MRR de sistemas que apresentam mais freqüentemente as repostas corretas próximas ao topo destas listas é maior. O MRR é calculado de acordo com a Equação 2.4:

$$MRR(QS) = \frac{\sum_{\forall q_i \in QS} \frac{1}{PosRespCorreta(q_i)}}{jQSj} \quad (2.4)$$

Onde QS é um conjunto de consultas submetidas a um sistema e $PosRespCorreta(q_i)$ é a posição que encontra-se a resposta correta da consulta i. O valor de MRR de um sistema é um número real entre 0 e 1, sendo 1 o melhor valor de MRR possível para um sistema, obtido apenas quando todas as respostas corretas às consultas submetidas são apresentadas na primeira posição da lista de respostas ordenadas, exibida pelo sistema.

A Tabela 2.1 mostra um exemplo onde o MRR do sistema é igual a 0:5075. Na consulta $q1$ por exemplo, a página desejada pelo usuário é o terceiro resultado da lista de respostas gerada pelo sistema. Isso faz com que o *Reciprocal Rank* desta consulta seja igual a 1=3 ou 0:33. O MRR do sistema é igual à média dos *Reciprocal Ranks* individuais de cada consulta, ou seja, $MRR = (1 + 0:33 + 0; 2 + 0:5)=4 = 0:5075$.

Consulta	Posição da Resposta Correta	Reciprocal Rank
q1	3	1/3 = 1
q2	1	1/1 = 0.33
q3	5	1/5 = 0.2
q4	2	1/2 = 0.5

Tabela 2.1: Tabela com exemplo de MRR.

Esta métrica é mais bem utilizada para avaliar consultas navegacionais ou transacionais, pois neste tipo de consulta apenas uma resposta correta já é suficiente para suprir a necessidade de informação do usuário.

Capítulo 3

Modelo de Hipergrafos para Análise de Apontadores em Coleções Web

Métodos tradicionais de análise de apontadores modelam a Web através de grafos direcionados e baseiam-se na intuição de que o apontador de uma página A para uma página B é um indício fornecido pelo autor da página A de que o conteúdo da página B é interessante de alguma forma. Entretanto, a utilização de grafos para modelar a Web permite que exista uma grande quantidade de apontadores que não obedecem a esta intuição. Exemplos de apontadores deste tipo são:

Apontadores navegacionais: usados para navegação dentro de um conjunto de páginas.

SPAM: apontadores criados para burlar os algoritmos de análise de apontadores. Seu principal objetivo é melhorar a reputação de uma página ou conjunto de páginas.

Apontadores em Redes Sociais: apontadores usados para fortalecer laços sociais em uma comunidade. Por exemplo, apontadores trocados entre páginas são comuns em Blogs e Fotologs.

A ocorrência de apontadores deste tipo faz com que a reputação das páginas estimada pelos algoritmos de análise de apontadores utilizando grafos seja afetada. Este tipo de alteração produz um efeito negativo, uma vez que a reputação estimada pelos algoritmos se distancia da reputação real. Uma forma de contornar este problema é a adoção de um outro modelo que

trate grupos de páginas muito relacionadas entre si como um único grupo, capaz de apontar cada página uma única vez.

Desta forma, podemos agrupar páginas que possuem uma forte relação entre si em blocos, modelando relações entre grupos de páginas e páginas individuais. Assim, não só eliminaríamos a representação das relações entre páginas do mesmo bloco, como também limitaríamos o número de apontadores que cada bloco de páginas fortemente conectadas produz. Ou seja, apontadores que partem de páginas do mesmo bloco e possuam o mesmo destino serão mapeados para um único apontador partindo do bloco.

O efeito obtido ao permitir que cada bloco vote (ou aponte) apenas uma vez em cada página é a valorização de páginas que possuam maior diversidade na origem de seus apontadores. Ou seja, se o conjunto de páginas que aponta para uma página A é composto de páginas de diversos grupos distintos, A irá receber um grande número de votos neste novo modelo. Por outro lado, se este conjunto de páginas que aponta para A for composto de páginas que pertencem apenas a um grupo, A irá receber apenas um voto. Acreditamos que páginas que realmente possuem boa reputação recebem apontadores de diversas páginas não relacionadas entre si.

Tomemos como exemplo um conjunto de 16 páginas de fotologs que criam muitos apontadores entre si baseados em relações de amizade como apresentado na Figura 3.1 (A). Suponha-se que o modelo identifique que há um forte relacionamento entre algumas páginas e as agrupe em dois blocos distintos, como na Figura 3.1 (B). Este agrupamento faz com que as páginas do mesmo bloco não possam mais apontar umas para as outras, o que elimina relações de reforço mútuo. Além disso, cada bloco pode apontar apenas uma vez para cada página da coleção, como mostra a Figura 3.1 (B). A reputação das páginas passa a depender também da variedade da origem dos apontadores e não apenas da sua quantidade.

Neste trabalho, propomos uma nova modelagem onde coleções de páginas web são representadas utilizando hipergrafos direcionados. Nele agruparemos as páginas de acordo com algum critério estabelecido e então representaremos relações entre grupos e páginas através das hiperarestas. Com isso pretendemos eliminar informações redundantes e aumentar a confiabilidade das conexões modeladas.

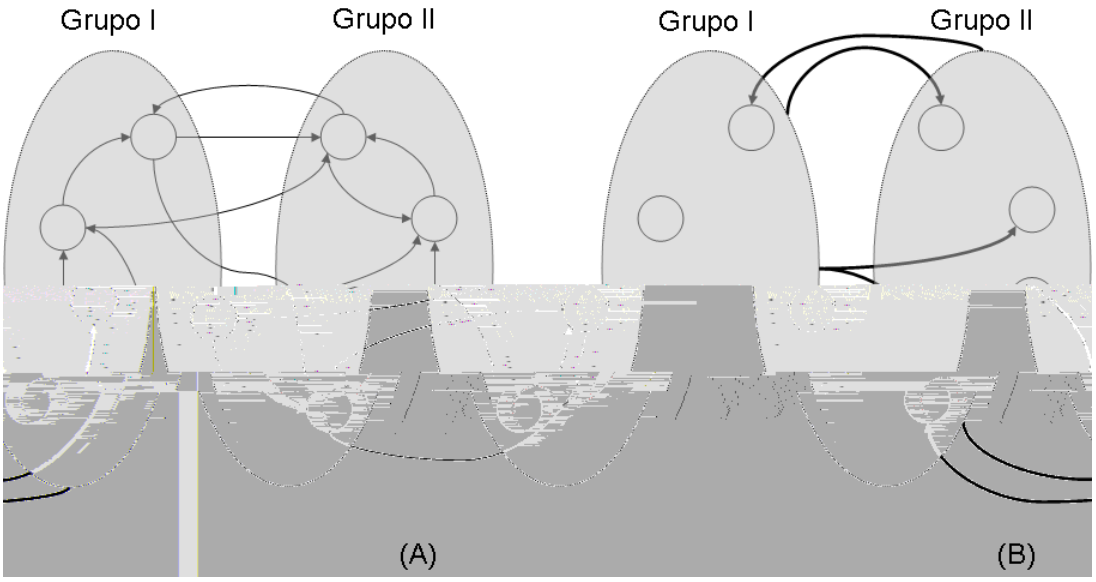

Figura 3.1: Grupo de páginas web modelado de duas formas: através de um grafo (A) e utilizando uma nova abordagem baseada em grupos (B)

3.1 Hipergrafos Direcionados

Um hipergrafo direcionado $H = (V; E)$ consiste de um conjunto de vértices V e um conjunto de hiperarestas E, onde $E \quad 2^{\mathcal{V}} \quad 2^{\mathcal{V}}$. Como desejamos calcular a importância de páginas individuais, nós redefinimos E de um modo mais restritivo, isto é, $E \quad 2^{\mathcal{V}} \quad V$. Desta forma, cada hiperaresta sempre aponta para um único vértice. Também desejamos que para cada hiperaresta $= (G; v)$ 2 E, v $\not\in$ G onde $G \quad 2^{\mathcal{V}}$. Consideramos a representação do hipergrafo sem a última restrição (v $\not\in$ G), que é equivalente a permitir a existência de apontadores internos, mas os resultados finais foram semelhantes aos obtidos na versão com a restrição.

Para modelar a Web, cada página é considerada um vértice no hipergrafo. Particionamos então a coleção, gerando blocos de páginas onde as páginas são agrupadas de acordo com algum critério de afinidade. Cada página deve pertencer a apenas um bloco, ou seja, os blocos não podem possuir intersecção entre si. Um bloco B no hipergrafo aponta para uma página v através de uma hiperaresta $= (B; v)$ se, somente se, houver pelo menos uma página de B que possua um apontador para a página v e v $\not\in$ B. Uma importante diferença entre este modelo e o modelo tradicional que utiliza grafos é que o critério de particionamento determina a granularidade dos blocos e das conexões realizadas através das hiperarestas.

Para efeitos de ilustração, pode-se assumir que o hipergrafo que representa uma coleção pode

ser derivado a partir do grafo que representa esta coleção. A Figura 3.2 ilustra um exemplo de hipergrafo derivado a partir de um grafo em três etapas. A Figura 3.2(A) mostra o grafo inicial. Na Figura 3.2(B) os blocos são criados e os apontadores entre páginas do mesmo bloco são removidos. Por fim, a Figura 3.2(C) mostra o resultado do mapeamento das arestas para hiperarestas.

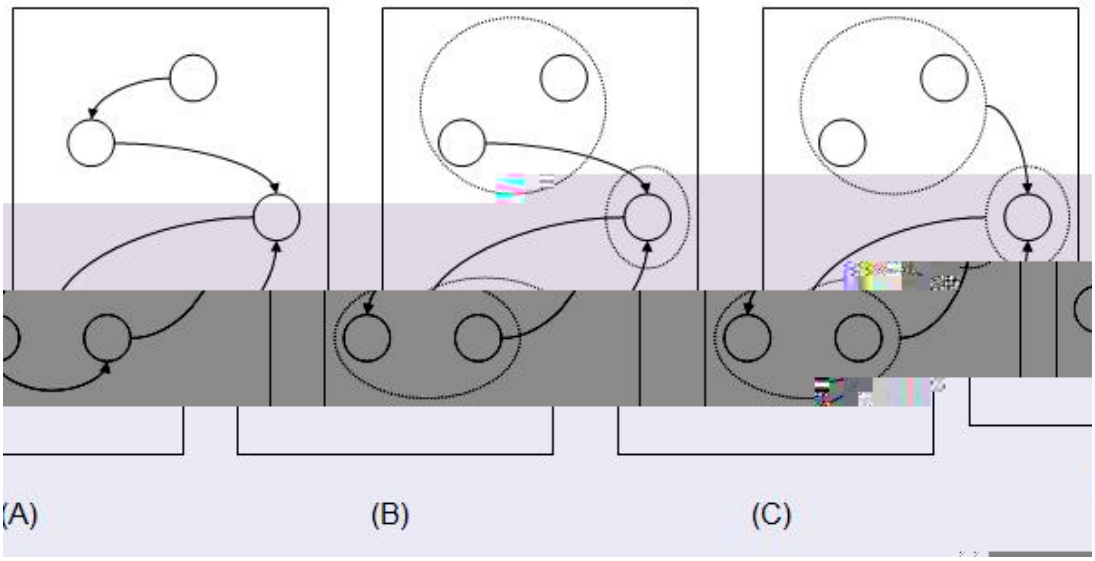

Figura 3.2: Exemplo de criação de um hipergrafo a partir de um grafo em três etapas: (A) grafo original, (B) criação dos blocos e remoção de apontadores internos e (C) mapeamento das arestas para hiperarestas.

3.2 Critérios de Particionamento

O método de particionamento utilizado na criação do hipergrafo é de vital importância, pois irá determinar a qualidade das conexões existentes entre os blocos e as páginas, além de regular o nível de informação que será perdida (blocos muito grandes não eliminam apenas informação redundante, mas também informação útil). Um bom método de particionamento deve encontrar o equilíbrio entre estes dois fatores, ou seja, gerar blocos que garantam uma boa qualidade das conexões ao mesmo tempo em que evita grandes perdas de informação.

Neste trabalho investigamos três critérios de particionamento baseados na hierarquia extraída das URLs das páginas:

Páginas – um bloco é composto de uma única página web. Este critério faz com que a representação utilizando hipergrafo seja igual à representação tradicional da Web.

Domínios – todas as páginas de um bloco pertencem ao mesmo domínio web.

Hosts – todas as páginas de um bloco pertencem ao mesmo host.

A vantagem de particionar a coleção com base apenas nessa hierarquia é o baixo custo, pois para determinar o grupo ao qual a página pertence basta que se processe a URL da mesma.

Nosso modelo é capaz de simular a representação tradicional da Web utilizando partições baseadas em páginas, onde cada página é considerada um bloco completo. Desta forma, ele é capaz de simular o grafo utilizado por métodos tradicionais de análise de apontadores, critério que é representado nos experimentos pela versão de grafos dos métodos estudados.

Nós adotamos também partições baseadas em hosts e domínios porque elas agrupam páginas provavelmente criadas pelo mesmo autor ou por pessoas relacionadas entre si. Este agrupamento foi citado na literatura [2, 9] como uma opção para computar uma variação do Indegree, mas nenhuma avaliação do seu impacto real na ordenação de respostas de máquinas de busca web foi realizada. Consideramos que dois hosts ou dois domínios diferentes criados pelo mesmo autor ou por autores relacionados são mais raros que páginas diferentes criadas pelo mesmo autor ou autores relacionados. Como acreditamos que conexões entre páginas relacionadas entre si não são muito confiáveis, esperamos obter modelos melhores por restringir as conexões àquelas entre blocos e páginas. Desta forma, as hiperarestas irão representar mais fielmente o conceito de reputação do que as arestas do grafo utilizado na modelagem tradicional e a reputação de uma página será proporcional à diversidade de blocos que apontam para ela.

Os três critérios de particionamento utilizados na criação dos blocos foram implementados utilizando a URL das páginas. A partição baseada em páginas é realizada agrupando cada página em um bloco individual. Para particionar o conjunto de páginas em domínios e hosts, processamos os caracteres da URL da página para extrair o *domain name* e o *host name* de cada página. Para definir qual o *host name* de cada página, a sua URL foi processada da seguinte forma:

Removemos os prefixos “http://”, caso existam;

Removemos os prefixos “www.”, caso existam;

Removemos os caracteres encontrados a partir da primeira “/”, caso existam;

Removemos a informação da porta do servidor, caso exista;

A cadeia de caracteres resultante é o *host name* da página.

Na URL "http://dir.yahoo.com:80/esportes´´, por exemplo, o *host name* é "dir.yahoo.com´´ pois o prefixo "http://´´, a string "/esportes´´ e a informação da porta do servidor (":80´´) foram removidas. Após encontrarmos o *host name* de cada página, páginas com o mesmo *host name* são agrupadas no mesmo bloco.

O *domain name* de uma página é encontrado através de uma concatenação de três elementos do *host name*:

Identificador de país: define o país de origem da página. Ex.: ".fr´´ (França), ".br´´ (brasil).

Categoria do servidor: define a categoria na qual a página se encaixa. Ex.: ".com´´ (comercial), ".edu´´ (educacional).

Nome do domínio: é o elemento que precede a categoria do servidor ou o identificador de país, caso o primeiro seja nulo. Ex.: "uol´´, "ufam´´.

A concatenação destes três elementos forma o *domain name*. É importante observar que o identificador de país e a categoria do servidor podem ser nulos, desde que isto não ocorra ao mesmo tempo. Por exemplo, na URL "http://dir.yahoo.com´´, o *domain name* é "yahoo.com´´ (não possui identificador de país, mas possui a categoria do servidor, já a URL "http://www.ufmg.br´´ pertence ao domínio "ufmg.br´´ (sem categoria do servidor, mas com identificador de país). Outros exemplos seguem na Tabela 3.1:

URL	Host name	Domain name
http://www.esportes.uol.com.br/	esportes.uol.com.br	uol.com.br
http://alunos.dcc.ufam.edu.br/klessius	alunos.dcc.ufam.edu.br	ufam.edu.br
http://www.cnn.com/news	cnn.com	cnn.com

Tabela 3.1: Tabela com exemplos de URLs e seus respectivos *domain* e *host names*.

É interessante notar que hosts são conjuntos de uma ou mais páginas ao passo que domínios são conjuntos de um ou mais hosts. Esta relação de hierarquia faz com que a granularidade dos blocos do hipergrafo seja maior ao utilizarmos domínio e menor quando utilizamos as páginas. Esta relação pode ser observada na Figura 3.3. A Figura 3.3(A) mostra o particionamento baseado em páginas (simulando a representação tradicional através de grafos). A Figura 3.3(B),

mostra a partição baseada em hosts. A Figura 3.3(C) ilustra a partição baseada em domínios, onde cada domínio consiste em um conjunto de um ou mais hosts. O número de hiperarestas diminui conforme um número maior de páginas é incluído em cada bloco. A granularidade, ou tamanho dos blocos, deve ser escolhida com cautela para manter o equilíbrio entre o número de hiperarestas (e conseqüentemente a cobertura sobre a quantidade de informação de cada página) e a qualidade da informação fornecida por cada hiperaresta.

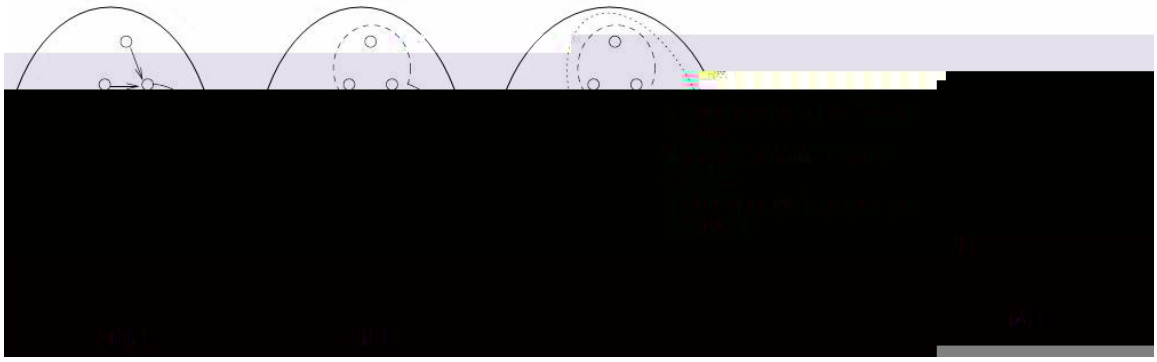

Figura 3.3: O modelo de hipergrafos considerando três diferentes critérios de particionamento: (A) páginas, (B) hosts e (C) domínios.

Nós também estudamos o uso de outras estratégias para o particionamento como, por exemplo, a aplicação de algorítmos de *clustering* para agrupar as páginas. Investigamos o uso de um algoritmo baseado em Árvores Geradoras Mínimas (AGM) [17] para produzir *clusters*. A medida de similaridade empregada em nossos experimentos foi baseada na intersecção entre os apontadores de cada grupo. Assim, dois grupos eram considerados similares se o conjunto de páginas em comum para os quais eles apontavam era superior a certo limiar L. Entretanto, os resultados não foram superiores àqueles obtidos utilizando o critério de particionamento baseado no processamento da URL e os custos computacionais para a criação dos *clusters* foram muito altos. Deste modo, decidimos não incluir estes estudos neste trabalho. Entretanto, outras estratégias para a criação de grupos podem ainda propiciar ganhos superiores aos obtidos atualmente e podem ser estudados no futuro.

3.3 Computando Reputação de Páginas no Modelo de Hipergrafo

Para ilustrar as vantagens do nosso modelo, mostramos neste capítulo exemplos de como dois métodos tradicionais de análise de apontadores, Pagerank e Indegree, podem ser adaptados para o modelo de hipergrafos. É importante observar que outras variações destes métodos ainda podem ser propostas para o hipergrafo. Estes dois exemplos são utilizados apenas para mostrar as potenciais vantagens de utilizar o modelo de hipergrafo para representar a Web. Estas duas adaptações serão chamadas de Hiperpagerank (uma versão do pagerank para o hipergrafo) e Hiperindegree (uma versão do Indegree para o hipergrafo). No próximo capítulo, serão mostrados experimentos para avaliar a desempenho destes dois métodos quando comparados com suas versões originais. Estudos complementares sobre como adaptar estes dois métodos e outros métodos de análise de apontadores devem ser explorados em trabalhos futuros.

Para computar o Hiperpagerank propomos uma maneira de computar a reputação de cada bloco para então computar a reputação (ou valor de HiperPagerank) de cada página. Uma estratégia simples e eficiente é considerar a reputação de todas as páginas no bloco. Dessa forma, cada iteração do Hiperpagerank é computada em dois passos:

1. A reputação de cada bloco de páginas $GR(B)$ é computada como a soma da reputação de cada página p que pertence a ele, como mostra a Equação 3.1.

$$GR(B) = \sum_{p \in \mathcal{B}} HPR(p); \tag{3.1}$$

   onde $PR(p)$ é o valor atual de HiperPagerank da página p.

2. O valor de reputação de cada página depende da sua representação no hipergrafo. Páginas que não recebem hiperarestas possuem valor 0, ou seja, consideramos que ela não possui reputação. Para cada página p que possui hiperarestas apontando para ela, é dado um valor inicial de $1/kVk$ e a reputação $PR(p)$ é computada como:

$$PR(p) = (1 \quad c) \quad \sum_{\mathcal{B} \in I(p)} \frac{GR(B)}{kO(B)k} + \frac{c}{kVk} \tag{3.2}$$

onde c é o dampening factor, $kO(B)k$ é o número de páginas apontadas pelo bloco B, $I(p)$ é o conjunto de blocos de páginas que apontam para a página p, e kVk é o número de páginas que recebem hiperarestas na coleção.

Estes dois passos são repetidos até que os valores de reputação venham a convergir. Observe que, como no Pagerank original, a convergência é garantida. Esta versão de hipergrafo do Pagerank pode ser interpretada como um "site Pagerank´´ com propagação dos valores para as páginas de cada site sendo realizada de acordo com os hiperlinks que cada página recebe.

O Hiperindegree é uma versão intuitiva do indegree, que consiste em contar, para cada página p, o número de hiperarestas que chegam até ela. Desta forma, o valor de Hiperindegree de p é calculado por:

$$HI(p) = \sum_{\mathcal{B} \in I(p)} 1; \tag{3.3}$$

onde I(p) é definido como na Equação 3.2.

Capítulo 4

Experimentos

Realizamos experimentos para verificar o impacto da utilização do modelo de hipergrafo para computar reputação de páginas com o intuito de auxiliar a ordenação dos resultados de máquinas de busca. Os experimentos de busca foram realizados em duas coleções web distintas. Também apresentamos experimentos para estudar o impacto de páginas de spam no modelo proposto. Para tanto, comparamos o comportamento dos algoritmos que utilizam o hipergrafo com o comportamento dos algoritmos que utilizam grafos em uma base com páginas de spam rotuladas.

4.1 Ambiente Experimental

Adotamos duas coleções web distintas para os experimentos que medem a qualidade das respostas da máquina de busca: WBR03 e WT10g. Para os experimentos com spam, nós utilizamos a coleção WEBSPAM-UK2006.

WBR03

A coleção WBR03 é uma base de dados da máquina de busca real TodoBR[1], composta de 12.020.513 de páginas web coletadas da Web brasileira em 2003. Como mostrado na Tabela 4.1, a coleção WBR03 possui 130.717.004 apontadores válidos entre suas páginas e o tamanho médio do texto de cada documento é 5kb. Este número de apontadores válidos indica que a coleção possui um conjunto de páginas altamente conectado, fornecendo assim, muita informação para métodos de análise de apontadores. Ela representa uma porção consideravelmente conectada

[1]TodoBR é uma marca registrada de Akwan Information Technologies, que foi adquirida pelo Google em Julho de 2005.

resultados para cada consulta e método. As páginas deste conjunto foram classificadas por um conjunto de 15 pessoas, em relevantes ou não relevantes em relação à consulta. O resultado de cada método foi avaliado utilizando duas métricas baseadas em precisão: a Precisão Média (MAP) e a precisão nos primeiros 10 resultados obtidos (P@10).

Processamos ambas as consultas (navegacionais e populares) de acordo com as especificações do usuário extraídas do log: frases, operações AND e operações OR.

WT10g

A segunda coleção adotada para os experimentos de busca é a WT10g, uma coleção adotada na TREC 2001 [1]. Como pode ser visto na Tabela 4.2, a coleção WT10g contém 1.692.096 documentos extraídos de toda a Web. O tamanho médio do texto das páginas é de aproximadamente 4,4 kbytes. O número de apontadores entre páginas da coleção é 2.530.920. Nós observamos que os apontadores entre as páginas na WT10g (aproximadamente 1,5 vezes o número de páginas) é razoavelmente menor que na WBR03 (aproximadamente 11 vezes o número de páginas). Esta diferença pode afetar a escolha do método de análise de apontadores em cada coleção e é importante para garantir a avaliação dos algoritmos em dois cenários distintos.

Outras características da coleção que devem afetar o resultado das técnicas de análise de apontadores é o número de páginas, hosts e domínios de cada coleção. A diferença entre o número de domínios e hosts na WT10g é apenas 10%, o que pode ser considerada uma diferença pequena se comparada com os valores da WBR03. Esta pequena diferença deve afetar o comportamento dos algoritmos baseados em hipergrafos na coleção, pois isto faz com que a diferença entre utilizar partições baseadas em hosts e partições baseadas em domínios seja mínima. Para os experimentos na WT10g, foram utilizadas as primeiras 145 consultas navegacionais da WT10g. Experimentos com consultas informacionais não são mostrados neste trabalho, pois assim como na WBR03, a qualidade dos resultados não foi afetada pelo método de análise de apontadores utilizado.

Os experimentos na coleção WT10g são úteis para estudar o comportamento dos métodos em uma coleção com um número pequeno de apontadores. É importante observar que este é um cenário onde a utilização do hipergrafo é prejudicada, pois a quantidade de hiperarestas é ainda menor que a quantidade de apontadores, fazendo com que o hipergrafo forneça muito pouca informação para ser analisada.

N. de Páginas	1.692.096	N. de Hosts	11.671
N. de Domínios	10.113	N. de Apontadores	2.530.920
N. de Hiperarestas(Host)	1.045	N. de Hiperarestas(Domínio)	952
Tamanho médio do texto das páginas	4,4kb		

Tabela 4.2: Estatísticas sobre a coleção WT10g.

Os experimentos foram realizados em dois cenários distintos: sem combinação e com combinação. No primeiro, avaliamos a qualidade dos métodos implementados sem combinação com outras evidências, ou seja, apenas o valor de reputação computado pelos algoritmos de análise de apontadores foi utilizado para ordenar o conjunto de páginas que possuíam os termos de cada consulta. No segundo, o valor de reputação das páginas foi combinado com o valor de similaridade atribuído pelo modelo vetorial sobre o conteúdo textual da página e também sobre o texto de âncora (o texto de âncora de uma página é formado pela concatenação dos termos encontrados nos apontadores que apontam para ela).

A combinação de diversas evidências não é uma tarefa trivial. Uma boa combinação deve garantir que valores de similaridades obtidos de fontes distintas sejam usados de maneira adequada para representar a similaridade final de cada documento. Neste trabalho utilizamos duas abordagens para combinação de múltiplas evidências (no nosso caso: reputação, conteúdo textual e texto de âncora da página)

A primeira estratégia de combinação utilizada [6] utiliza redes bayesianas para chegar à Fórmula 4.1 para a combinação:

$$sim(q;d) = 1 \quad ((1 \quad sim_{texto}(q;d)) \quad (1 \quad sim_{anchor}(q;d)) \quad (1 \quad reputacao(d))) \qquad (4.1)$$

onde $sim(q;d)$ é a similaridade final da consulta q com o documento d, $sim_{texto}(q;d)$ é a similaridade entre a consulta q e o conteúdo textual do documento d utilizando o modelo vetorial. $sim_{anchor}(q;d)$ é a similaridade entre a consulta q e o texto de âncora do documento d utilizando o modelo vetorial e $reputacao(d)$ é o valor de reputação estimado pelo algoritmo de análise de apontadores para o documento d. Esta combinação será chamada neste trabalho de **BNC** (*Bayesian Network Combination*).

A outra estratégia utilizada para combinar as evidências baseia-se em uma funçao que normaliza [7] o valor de popularidade das páginas. Este valor normalizado é combinado de maneira linear com as outras evidências que serão utilizadas. A Equação 4.2 descreve a função de nor-

malização utilizada:

$$rep_{norm}(d) = w \quad \frac{reputacao(d)^a}{reputacao(d)^a + k^a} \tag{4.2}$$

onde $rep_{norm}(d)$ é a reputação normalizada da página d e $reputacao(d)$ é a reputação não normalizada da página d; w, k e a são constantes cujo valor é definido através de um treinamento. Esta estratégia de combinação obtém resultados melhores, porém necessita de um treinamento baseado em força bruta. Neste trabalho, o treinamento foi realizado variando todos os parâmetros (w, a e k) entre 0.0 e 2.0 e verificando qual combinação de parâmetros propiciou melhores resultados em um conjunto de consultas de treino. A melhor combinação obtida para cada método foi então utilizada em um outro conjunto de consultas chamado conjunto de teste e os resultados obtidos neste último conjunto foram utilizados para comparação e é referenciada neste trabalho por **BFC** (*Brute Force Combination*).

Todos os experimentos adotam o *t-test* para avaliar a significância estatística dos resultados obtidos. O *t-test* é utilizado para avaliar se a diferença entre as respostas de dois sistemas distintos é significante ou não. É importante ressaltar também que apesar dos resultados obtidos no cenário sem combinação serem úteis para evidenciar a diferença existente entre os métodos, a reputação das páginas é uma informação independente da consulta. Logo, ordenar as respostas utilizando apenas esta evidência não é uma estratégia utilizada em máquinas de busca reais. O resultado obtido com os métodos de combinação utilizados se aproxima mais de um cenário real de máquina de busca.

WEBSPAM-UK2006

A base de dados WEBSPAM-UK2006[3] é uma coleção de páginas e apontadores que possui rótulos que classificam os hosts em duas classes: host de spam e host normal. Esta coleção contém 77.741.046 páginas dispostas em 11.402 hosts e 7.650 domínios web, todos pertencentes ao identificador de país UK. Ela possui quase 3 bilhões de apontadores entre páginas, aproximadamente 11 milhões de apontadores entre páginas de hosts diferentes e 8 milhões de apontadores entre páginas de domínios distintos. Esta coleção possui 10.662 hosts rotulados, sendo 8.123 hosts normais e 2.113 hosts de spam. Em nossos experimentos, consideramos que todas

[3] http://yr-bcn.es/webspam/datasets/uk2006/

as páginas pertencentes a um host de spam são páginas de spam. A Tabela 4.3 apresenta mais informações sobre esta coleção.

N. de Páginas	77.741.046	N. de Hosts	11.402
N. de Domínios	7.650	N. de Apontadores	2.951.370.103
N. de Hiperarestas(Host)	11.751.637	N. de Hiperarestas(Domínio)	8.056.314

Tabela 4.3: Estatísticas sobre a coleção WEBSPAM-UK2006.

4.2 Métodos Implementados

Para comparar os resultados obtidos pelos algoritmos que utilizam o modelo de hipergrafo com os resultados obtidos utilizando o modelo tradicional de grafos, implementamos dois métodos de análise de apontadores descritos anteriormente: Pagerank e Indegree. Também implementamos as versões destes algoritmos que utilizam a modelagem da Web através de hipergrafos. Utilizamos dois tipos de hipergrafos: um com os grupos definidos através do host e outro com os grupos definidos através do domínio. Os algoritmos adaptados que utilizam o agrupamento baseado no host são chamados de HiPRHost e HiIndHost (baseados no Pagerank e Indegree respectivamente) e os algoritmos que utilizam o agrupamento baseado no domínio são chamados de HiPRDom e HiIndDom (baseados no Pagerank e Indegree respectivamente).

O Pagerank foi escolhido por ser considerado um bom método de análise de apontadores. Além disso, ele geralmente é utilizado como método básico para comparação com outros algoritmos de análise de apontadores mais recentes propostos na literatura. Como visto anteriormente, ele computa a reputação de uma página como a probabilidade de um surfista aleatório visitar esta página navegando aleatoriamente na Web. Esta navegação é feita através dos apontadores da coleção.

Dada uma página p, o Pagerank desta página é calculado como descrito na Equação 4.3:

$$PR(p) = (1 \quad c) \quad \sum_{q \in I(p)} \frac{PR(q)}{kO(q)k} + \frac{c}{kVk} \tag{4.3}$$

onde c é o dampening factor, $kO(q)k$ é o número de páginas apontadas por q, $I(p)$ é o conjunto de páginas que apontam para p, e kVk é o número de páginas na coleção.

O Indegree simplesmente conta, para cada página p, o número de apontadores que ela recebe. Apesar de este método ser muito suscetível a outros tipos de ruídos encontrados em coleções

web para ser realmente utilizado em máquinas de busca comerciais, ele é útil para que possamos observar a diferença entre a utilização de grafos e hipergrafos e o impacto da escolha de uma boa granularidade nos resultados obtidos utilizando hipergrafos.

Como os métodos que utilizam hipergrafos naturalmente descartam os apontadores entre páginas do mesmo bloco (pois esta informação não é presente no hipergrafo), também implementamos versões do Pagerank e do Indegree que utilizam uma versão do grafo da Web sem estes apontadores. Desta forma, cada um dos métodos para grafos foi implementado utilizando três tipos de grafos:

Grafo original da coleção: este grafo contém todos os apontadores entre páginas da coleção. Os métodos implementados utilizando este grafo são chamados neste trabalho de Indegree e Pagerank.

Grafo sem apontadores entre páginas do mesmo host: neste grafo, foram removidos os apontadores entre páginas que pertenciam a um mesmo host. Os métodos utilizando esta versão do grafo são aqui chamados de PRHost e IndHost.

Grafo sem apontadores entre páginas do mesmo domínio: neste grafo, foram removidos os apontadores entre páginas que pertenciam ao mesmo domínio. Os métodos que utilizam esta versão do grafo são chamados neste trabalho de PRDom e IndDom.

É importante observar que tais variações do grafo são úteis para que analisemos se o ganho obtido pela utilização do modelo de hipergrafo foi alcançado apenas devido à remoção dos apontadores internos ou se deu devido à melhor representação da Web obtida com o modelo.

4.3 Resultados

Os resultados apresentados neste capítulo são divididos em três partes distintas. Na primeira, apresentamos experimentos utilizando a coleção WBR03 com quatro conjuntos distintos de consultas (navegacionais populares, navegacionais aleatórias, informacionais populares e informacionais aleatórias). Ainda na primeira parte dos experimentos, avaliamos o impacto da utilização de um método de remoção de ruídos em coleções web sobre os algoritmos implementados. A segunda parte é composta de experimentos de busca utilizando um conjunto único de consultas navegacionais na coleção WT10g. Na terceira e última parte dos experimentos, estudamos o

impacto da utilização do hipergrafo na coleção WEBSPAM-UK2006, analisando a reputação estimada por diferentes algoritmos de análise de apontadores das páginas de spam presentes nesta coleção.

4.3.1 Experimentos na WBR03

As Tabelas 4.4 e 4.5 mostram o MRR obtido pelas versões do Pagerank e do Indegree respectivamente ao processar o conjunto de consultas navegacionais populares. É importante lembrar que os valores exibidos nas colunas sem combinação das tabelas servem apenas para comparar o desempenho de cada método individualmente, enquanto as colunas BNC e BFC, demonstram o desempenho obtido quando utilizamos estratégias de combinação com outras evidências e representam um comportamento mais próximo do obtido em máquinas de busca reais. Os resultados indicam que em ambos os casos as versões de hipergrafos dos métodos são superiores às versões para grafos neste conjunto de consultas. O *t-test* foi aplicado para verificar se a diferença entre os métodos possuía significância estatística e indicou que todas as diferenças entre os algoritmos de hipergrafos e suas respectivas versões para grafos são significativas.

Versões do Pagerank (consultas navegacionais populares)			
Método	**sem combinação**	BNC	BFC
Pagerank	0,2834	0,4610	0,4553
PRHost	0,3987	0,5036	0,5794
PRDom	0,4888	0,5785	0,6361
HiPRHost	0,5535	0,5819	0,6174
HiPRDom	0,6378	0,7150	0,7978

Tabela 4.4: Valores de MRR (*Mean Reciprocal Rank*) para consultas navegacionais populares na coleção WBR03, ao modelar a Web como um grafo (Pagerank, PRHost e PRDom) e como um hipergrafo (HiPRHost e HiPRDom).

Versões do Indegree (consultas navegacionais populares)			
Método	**sem combinação**	BNC	BFC
Indegree	0,5109	0,5890	0,6726
IndHost	0,5413	0,6101	0,6565
IndDom	0,6510	0,7122	0,7617
HiIndHost	0,5964	0,6241	0,7166
HiIndDom	0,7273	0,7706	0,8547

Tabela 4.5: Valores de MRR (*Mean Reciprocal Rank*) para consultas navegacionais populares na coleção WBR03, ao modelar a Web como um grafo (Indegree, IndHost e IndDom) e como um hipergrafo (HiIndHost e HiIndDom).

Na maioria dos casos para consultas navegacionais, os resultados foram melhorados quando

utilizamos os métodos baseados em domínios ou hosts, confirmando a nossa hipótese inicial de que a remoção dos apontadores internos gera melhores resultados. Um exemplo de mudança nos resultados pode ser visto ao processar a consulta "BOL"[4], onde os dois primeiros resultados fornecidos pelo Indegree são blogs apontados por muitos outros blogs, enquanto que a resposta correta à consulta aparece na terceira posição. O método HiIndDom não foi afetado neste caso porque a maioria das páginas que alavancaram a posição dos blogs na lista de respostas são outros blogs que pertencem a um pequeno grupo de domínios distintos. Desta forma, a quantidade de domínios distintos que aponta para a página inicial do BOL é muito maior que a quantidade de domínios distintos que aponta para os blogs que anteriormente ocupavam as primeiras posições. Ou seja, a página inicial do BOL é apontada por um conjunto mais diverso de páginas e, por isso, possui melhor reputação do que os blogs que ocupavam as primeiras posições anteriormente e são apontados quase que exclusivamente por outros blogs.

As diferenças entre os resultados obtidos pelos métodos PRDom e HiPRDom são úteis para mostrar que a superioridade dos resultados do método baseado no hipergrafo não se dá apenas devido à remoção dos apontadores internos que ocorre no hipergrafo, mas sim como conseqüência de uma melhor representação das conexões fornecidas pelo modelo. Por exemplo, o algoritmo PRDom foi a versão do Pagerank que obteve melhores resultados no conjunto de consultas navegacionais populares, o que indica que a remoção dos apontadores entre páginas do mesmo domínio melhora os resultados obtidos na coleção WBR03. Entretanto, o método HiPRDom obteve um ganho de 30.48% comparado com o PRDom sem utilizar nenhuma outra evidência além da informação de reputação das páginas. Além disso, apresentou um ganho de 23.60% ao combinar outras evidências utilizando o BNC e 25.42% ao utilizar o BFC. Estes resultados indicam que este desempenho não ocorre simplesmente devido à remoção dos apontadores entre páginas do mesmo bloco (domínio ou host), mas sim pela melhor representação da Web obtida através da modelagem utilizando hipergrafos.

As Tabelas 4.4 e 4.5 também mostram que os resultados obtidos ao utilizar domínios como partições são superiores aos resultados obtidos ao utilizar hosts. Ao examinar as partições criadas utilizando hosts encontramos algumas razões para este fraco desempenho obtido. Uma delas é o fato de que partições baseadas em hosts acabam criando vários blocos conectados devido à replicação de *host names* ou por pertencerem a um mesmo portal. Por exemplo, os hosts "es-

[4]BOL (http://www.bol.uol.com.br/)

portes.uol.com.br" e "games.uol.com.br"pertencem ao mesmo portal (http://www.uol.com.br) e quando utilizamos o modelo de hipergrafo, muitas hiperarestas conectando um bloco às páginas do outro bloco são criadas. Estes casos são muito comuns na Web e estas hiperarestas não correspondem ao conceito de voto baseado na qualidade da página. Como conseqüência, elas reduzem a qualidade dos resultados obtidos ao utilizarmos o host como critério de particionamento.

Em todos os casos, as diferenças entre os métodos são atenuadas quando a ordenação dos resultados leva em consideração outras evidências. Entretanto, as versões de algoritmos que utilizam hipergrafos ainda apresentam melhores resultados quando comparadas com as versões dos mesmos algoritmos utilizando grafos. Os resultados obtidos no conjunto de consultas navegacionais populares indicam que o modelo proposto é especialmente útil para este tipo de consulta, pois neste cenário estão os maiores ganhos obtidos pelo método. Esta informação pode ser utilizada para que o método seja adotado em um sistema de máquina de busca que classifique as consultas submetidas a ele de acordo com seu propósito e popularidade.

As Tabelas 4.6 e 4.7 mostram os resultados obtidos pelos métodos implementados com um conjunto de consultas navegacionais selecionadas aleatoriamente. Este conjunto de consultas visa mostrar o comportamento esperado dos algoritmos em sistemas de máquinas de busca que utilizam uma única estratégia de ordenação de resultados para todos os tipos de consultas, tanto populares quanto aleatórias. Como pode ser visto na Tabela 4.6, as versões de hipergrafo do Pagerank, HiPRHost e HiPRDom, ainda obtêm melhores resultados quando comparadas com suas versões para grafos. Ao comparar as versões do Indegree para grafos com as versões para hipergrafos na Tabela 4.7, podemos observar que os resultados são próximos, com uma pequena vantagem para os algoritmos de hipergrafos. Quando nenhuma outra evidência é utilizada, os algoritmos para hipergrafos obtêm sempre os melhores resultados.

Versões do Pagerank (consultas navegacionais aleatórias)			
Método	**sem combinação**	BNC	BFC
Pagerank	0,2823	0,5729	0,5280
PRHost	0,3599	0,5614	0,6442
PRDom	0,4642	0,6216	0,6967
HiPRHost	0,4899	0,6144	0,6834
HiPRDom	0,5562	0,6889	0,7856

Tabela 4.6: Valores de MRR (*Mean Reciprocal Rank*) para consultas navegacionais aleatórias na coleção WBR03, ao modelar a Web como um grafo (Pagerank, PRHost e PRDom) e como um hipergrafo (HiPRHost e HiPRDom).

Também realizamos um experimento de validação cruzada pra verificar o impacto da escolha

Versões do Indegree (consultas navegacionais aleatórias)			
Método	**sem combinação**	BNC	BFC
Indegree	0,4353	0,6258	0,7468
IndHost	0,4540	0,5832	0,7470
IndDom	0,5256	0,6587	0,7916
HiIndHost	0,5236	0,5863	0,7841
HiIndDom	0,6343	0,6784	0,8391

Tabela 4.7: Valores de MRR (*Mean Reciprocal Rank*) para consultas navegacionais aleatórias na coleção WBR03, ao modelar a Web como um grafo (Indegree, IndHost e IndDom) e como um hipergrafo (HiIndHost e HiIndDom).

do conjunto de consultas utilizado na fase de treino da combinação BFC. Um conjunto de 75 consultas (15 para treino e 60 para teste) foi dividido em cinco subgrupos, cada um contendo 15 consultas. Quando um subgrupo foi utilizado na fase de treino, as 60 consultas dos 4 subgrupos restantes formaram o conjunto de teste. Os resultados encontram-se nas Tabelas 4.8 e 4.9.

Consultas Populares						
Método/Conjunto de Treino	1	2	3	4	5	Média
Indegree	0,6726	0,6205	0,5169	0,6066	0,5236	0,5880
IndHost	0,6565	0,5853	0,5069	0,6152	0,6095	0,5947
IndDom	0,7617	0,7572	0,7241	0,7655	0,6412	0,7299
HiIndHost	0,7166	0,6516	0,5622	0,6649	0,6388	0,6468
HiIndDom	**0,8547**	**0,8548**	**0,7964**	**0,8115**	**0,8391**	**0.8313**
Método/Conjunto de Treino	1	2	3	4	5	Média
Pagerank	0,4553	0,5041	0,5094	0,5322	0,4848	0,4972
PRHost	0,5794	0,5428	0,5169	0,5783	0,5121	0,5459
PRDom	0,6361	0,6251	0,6507	0,6748	0,6208	0,6415
HiPRHost	0,6174	0,6325	0,5373	0,6332	0,5595	0,5960
HiPRDom	**0,7978**	**0,7861**	**0,7888**	**0,7802**	**0,6271**	**0,7560**

Tabela 4.8: Teste de validação cruzada utilizando MRR para consultas navegacionais populares. O melhor valor de MRR para cada conjunto de treino está destacado.

As Tabelas 4.8 e 4.9 mostram que, entre as versões do Pagerank, o HiPRDom é o método que apresenta os melhores resultados gerais, atingindo um desempenho 17.85% melhor que o PRDom em consultas populares e 16.59% superior em consultas navegacionais aleatórias. A versão do Indegree para hipergrafos, HiIndDom, também atinge os melhores valores de MRR entre todas as versões do Indegree, com um ganho médio de 13.89% sobre o IndDom em consultas populares e 5% em consultas aleatórias.

As Tabelas 4.10 e 4.11 mostram os resultados obtidos com o conjunto de consultas informacionais. Os resultados do teste de significância aplicado (*t-test*) para comparar os resultados obtidos na versão de hipergrafo dos algoritmos com os resultados obtidos na versão de grafos dos

Consultas Navegacionais Aleatórias						
Método/Conjunto de Treino	1	2	3	4	5	Média
Indegree	0,7468	0,6501	0,7535	0,6277	0,6434	0,6843
IndHost	0,7470	0,6131	0,7315	0,6160	0,6333	0,6682
IndDom	0,7916	0,7964	0,8273	0,6953	**0,6914**	0,7604
HiIndHost	0,7841	0,6735	0,7185	0,6446	0,5539	0,6749
HiIndDom	**0,8391**	**0,8309**	**0,8613**	**0,7963**	0,6648	**0.7984**
Método/Conjunto de Treino	1	2	3	4	5	Média
Pagerank	0,5280	0,4824	0,4827	0,5701	0,5515	0,5229
PRHost	0,6442	0,4774	0,6122	0,5809	0,5538	0,5737
PRDom	0,6967	0,6614	0,6215	0,6557	0,6800	0,6631
HiPRHost	0,6834	0,6864	0,7242	0,6770	0,6089	0,6760
HiPRDom	**0,7856**	**0,7534**	**0,8238**	**0,7539**	**0,7491**	**0,7731**

Tabela 4.9: Teste de validação cruzada utilizando MRR para consultas navegacionais aleatórias. Os melhores valores de MRR para cada conjunto de treino estão destacados.

Versões do Pagerank (consultas informacionais populares)						
Método	sem comb.		BNC		BFC	
	MAP	P@10	MAP	P@10	MAP	P@10
Pagerank	0,064	0,334	0,105	0,456	0,428	0,643
PRHost	0,058	0,300	0,095	0,412	0,489	0,757
PRDom	0,053	0,298	0,098	0,422	0,487	0,757
HiPRHost	0,067	0,370	0,099	0,434	0,481	0,753
HiPRDom	0,057	0,312	0,093	0,410	0,498	0,777

Tabela 4.10: Valores de MAP e P@10 para o conjunto de consultas informacionais populares na coleção WBR03, modelando a Web como um grafo (Pagerank, PRHost e PRDom) e como um hipergrafo (HiPRHost e HiPRDom).

Versões do Indegree (consultas informacionais populares)						
Método	sem comb.		BNC		BFC	
	MAP	P@10	MAP	P@10	MAP	P@10
Indegree	0,058	0,316	0,108	0,486	0,488	0,763
IndHost	0,053	0,302	0,087	0,394	0,473	0,737
IndDom	0,056	0,306	0,081	0,368	0,487	0,763
HiIndHost	0,056	0,318	0,099	0,452	0,473	0,736
HiIndDom	0,066	0,364	0,105	0,428	0,495	0,780

Tabela 4.11: Valores de MAP e P@10 para o conjunto de consultas informacionais populares na coleção WBR03, modelando a Web como um grafo (Indegree, IndHost e IndDom) e como um hipergrafo (HiIndHost e HiIndDom).

mesmos indica que não há diferença significativa entre eles. O fato da variação dos resultados obtidos com as consultas informacionais ser menor do que a variação encontrada nas consultas navegacionais sugere que nas consultas informacionais o usuário está interessado apenas no conteúdo das páginas retornadas (não importa qual será a página retornada, importa apenas que ela possua o conteúdo desejado), o que torna este conteúdo mais importante para consultas

informacionais que para consultas navegacionais. Logo, mudar apenas a estratégia de análise de

Versões do Indegree			
Consultas Populares			
Método	**sem combinação**	BNC	BFC
IndDom	0,6940	0,7256	0,7617
HiIndDom	**0,7357**	**0,7592**	**0,8547**
Consultas Aleatórias			
Método	**sem combinação**	BNC	BFC
IndDom	0,5738	0,6277	0,7096
HiIndDom	**0,5891**	**0,6446**	**0,7591**

Tabela 4.12: Valores de MRR (*Mean Reciprocal Rank*) para consultas navegacionais na coleção WBR03 quando o método é utilizado após remoção de apontadores ruidosos.

Versões do Pagerank			
Consultas Populares			
Método	**sem combinação**	BNC	BFC
PRDom	0,5556	0,6242	0,6361
HiPRDom	**0,6639**	**0,7178**	**0,7978**
Consultas Aleatórias			
Método	**no combination**	BNC	BFC
PRDom	0,4669	0,6366	0,6626
HiPRDom	**0,5181**	**0,6407**	**0,7475**

Tabela 4.13: Valores de MRR (*Mean Reciprocal Rank*) para consultas navegacionais na coleção WBR03 quando o método é utilizado após remoção de apontadores ruidosos.

4.3.2 Experimentos na WT10g

Para avaliar o desempenho dos algoritmos propostos em um cenário distinto, realizamos experimentos com uma coleção web de pequeno porte, a coleção WT10g. Foram utilizadas 145 consultas navegacionais da própria coleção. O fato desta coleção não possuir log disponível, impossibilitou a divisão destas consultas em populares e aleatórias, como foi realizado na WBR03.

As Tabelas 4.14 e 4.15 mostram que todos os métodos apresentam resultados insatisfatórios quando utilizados como única fonte de informação sobre a página e os algoritmos de hipergrafos aparentemente produzem resultados piores quando comparados com suas versões para grafos. Isto pode ser explicado pela perda de informação sofrida pelo hipergrafo ao substituir as arestas por hiperarestas. Entretanto, esta diferença qualitativa entre os métodos não está presente quando os mesmos são combinados com outras evidências. Os resultados obtidos tanto na combinação BNC quanto na BFC são muito similares para todos os métodos, isso ocorre porque esta coleção fornece muito pouca informação de apontadores (apenas 1,5 apontadores por página em média). Isto torna o resultado final gerado pelos algoritmos de análise de apontadores seja de baixa qualidade, fazendo com que as outras evidências sejam as maiores responsáveis pela

qualidade das respostas.

Versões do Pagerank			
Método	**sem combinação**	BNC	BFC
Pagerank	0,0840	0,2547	0,3224
PRHost	0,0762	0,2883	0,3196
PRDom	0,0762	0,2883	0,3196
HiPRHost	0,0218	0,2874	0,3196
HiPRDom	0,0218	0,2868	0,3188

Tabela 4.14: Valores de MRR (Mean Reciprocal Rank) para consultas navegacionais na coleção WT10g, modelando a Web como um grafo (Pagerank, PRHost e PRDom) e como um hipergrafo (HiPRHost e HiPRDom).

Versões do Indegree			
Método	**sem combinação**	BNC	BFC
Indegree	0,0616	0,1901	0,3299
IndHost	0,0776	0,2698	0,3147
IndDom	0,0776	0,2698	0,3147
HiIndHost	0,0776	0,2698	0,3188
HiIndDom	0,0776	0,2698	0,3188

Tabela 4.15: Valores de MRR (Mean Reciprocal Rank) para consultas navegacionais na coleção WT10g, modelando a Web como um grafo (Indegree, IndHost e IndDom) e como um hipergrafo (HiIndHost and HiIndDom).

Os experimentos realizados nesta coleção indicam que o uso do modelo de hipergrafo não resulta em uma mudança na qualidade da ordenação final das respostas da máquina de busca em um cenário real (combinando a reputação estimada com outras evidências). Isso acontece porque em coleções com pouca informação de apontadores, os métodos de análise de apontadores têm dificuldade de extrair informação útil o suficiente para estimar a reputação das páginas de forma satisfatória.

4.3.3 SPAM

Técnicas para melhorar artificialmente a posição de páginas em resultados de máquinas de busca são atualmente muito utilizadas na Web. Isto pode ser feito de diversas maneiras, como através da criação de informação artificial de apontadores para alterar a reputação estimada pelos algoritmos de análise de apontadores [10] para determinada página ou grupo de páginas. Estas páginas são chamadas de páginas de Web spam.

Como algoritmos de análise de apontadores podem ser afetados por estas técnicas, é importante estudar a resistência a técnicas de spam oferecida pelos novos algoritmos propostos. Por

exemplo, se a reputação das páginas de spam estimada pelo HiPRDom for superior à reputação destas páginas estimada pelo Pagerank tradicional, isto pode ser um indício de que os métodos derivados do modelo de hipergrafo são mais suscetíveis a spam. Em contrapartida, se o modelo de hipergrafo penaliza páginas com spam, isto pode ser um indicativo de que parte do ganho de qualidade obtido por este modelo nos experimentos anteriores é uma conseqüência dessa penalização.

Os experimentos seguintes mostram o comportamento do modelo de hipergrafo em relação a páginas de spam. Para cada método avaliado, realizamos o seguinte procedimento: calculamos o valor da reputação de cada página da coleção WEBSPAM-UK2006; computamos então, para cada página comum, quantas páginas de spam possuíam valor de reputação superior a ela. Por fim, as páginas foram ordenadas em ordem decrescente de reputação como mostram as Figuras 4.1, 4.2, 4.3 e 4.4.

Estes experimentos foram realizados utilizando apenas os métodos que obtiveram os melhores resultados nos experimentos anteriores, ou seja, os métodos que utilizam domínios como grupo (IndDom, HiIndDom, PRDom, HiPRDom). Cada método de hipergrafo foi comparado com sua versão para grafos (IndDom vs. HiIndDom e PRDom vs. HiPRDom). Para cada um destes pares, dois experimentos foram realizados: um mostrando os resultados para todas as páginas da coleção que receberam pelo menos um apontador e outro mostrando apenas as 100.000 páginas com maior reputação estimada por cada algoritmo. Este último experimento é útil para mostrar como as páginas de melhor reputação são afetadas pelo spam.

A Figura 4.1 mostra o resultado considerando apenas as 100.000 páginas de maior reputação em cada método. O HiIndDom supera o IndDom, relacionando um menor número de páginas de spam entre as páginas com maior reputação estimada.

Na Figura 4.2 são exibidos os resultados quando levamos em conta todas as páginas da coleção que recebem ao menos um apontador. Estes resultados mostram que os métodos IndDom e HiIndDom possuem comportamento similar. Isto nos leva à conclusão de que nesta coleção web os resultados obtidos pelo HiIndDom são superiores aos resultados obtidos pelo IndDom em termos de vulnerabilidade a spam. O método de hipergrafo apresenta resultados similares quando toda a coleção é levada em conta, porém as páginas de melhor reputação de acordo com o método são menos afetadas pelo spam.

A Figura 4.3 mostra os resultados obtidos quando consideramos apenas as 100.000 páginas

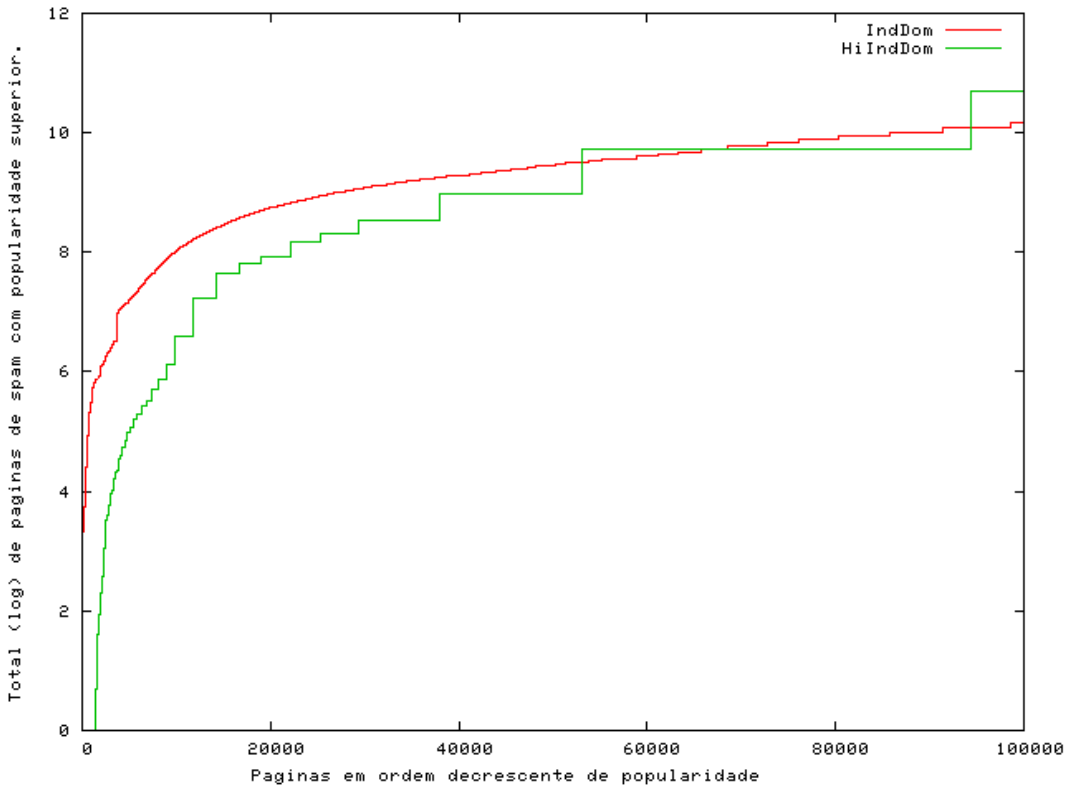

Figura 4.1: Número (log) de páginas de spam com reputação estimada superior a cada uma das 100.000 páginas comuns (não spam) com maior reputação estimada por cada um dos dois métodos.

com maior reputação estimada pelos métodos HiPRDom e PRDom. Nesta figura pode-se observar que o desempenho do HiPRDom é superior à do PRDom para todas as páginas analisadas. Isso significa que para as páginas com maior reputação estimada, o algoritmo baseado em hipergrafos é menos afetado pelo spam do que o algoritmo baseado em grafos.

O resultado dos métodos baseados no Pagerank para todas as páginas da coleção que recebem apontadores é exibido na Figura 4.4. Nesta figura, podemos ver que o HiPRDom obtém resultados superiores nas 500.000 páginas de melhor reputação, obtendo performance semelhante ao PRDom no restante da curva. Com isso concluímos que o HiPRDom, é menos afetado por spam do que o algoritmo PRDom nesta coleção, pois no geral as páginas de spam possuem menor reputação no algoritmo baseado em hipergrafos do do que no algoritmo baseado em grafos.

Os resultados exibidos indicam que o uso do modelo de hipergrafo em substituição ao modelo tradicional de grafo pode gerar melhoras na qualidade dos resultados no que diz respeito ao tratamento de páginas de spam. Entretanto é importante ressaltar que este ganho não é

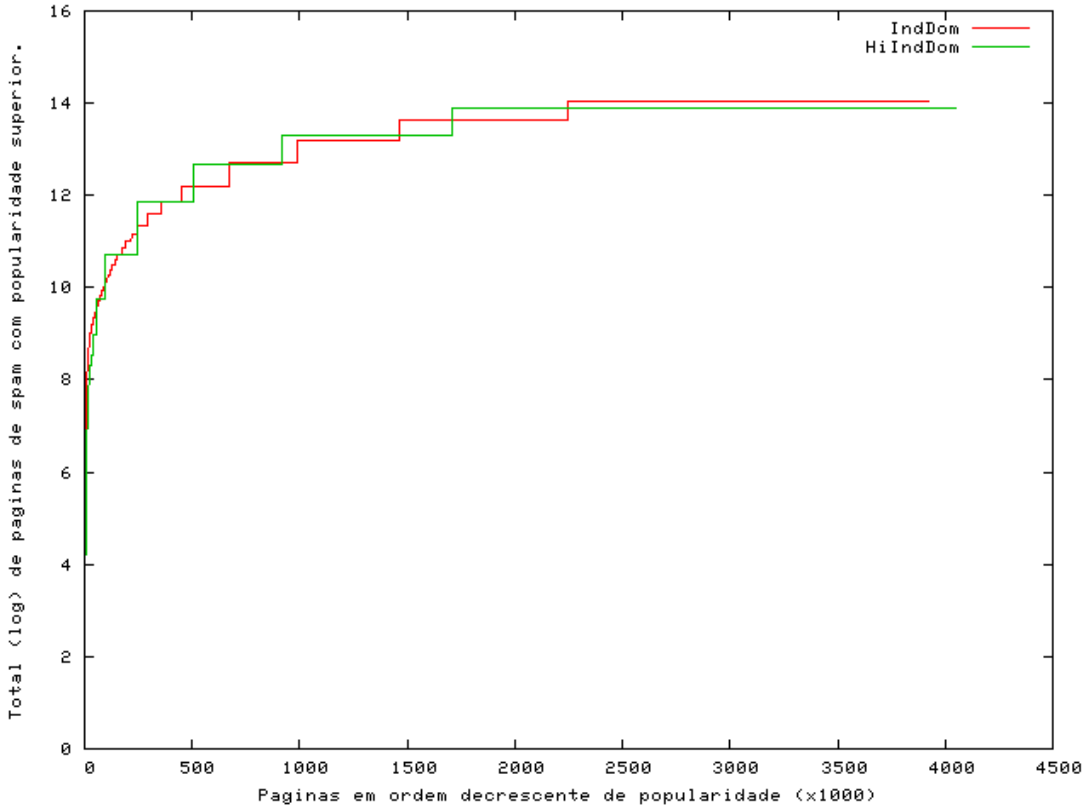

Figura 4.2: Número (log) de páginas de spam com reputação estimada superior a cada uma das páginas da coleção que recebe pelo menos um apontador.

substancial, dado que a diferença entre as curvas é pequena.

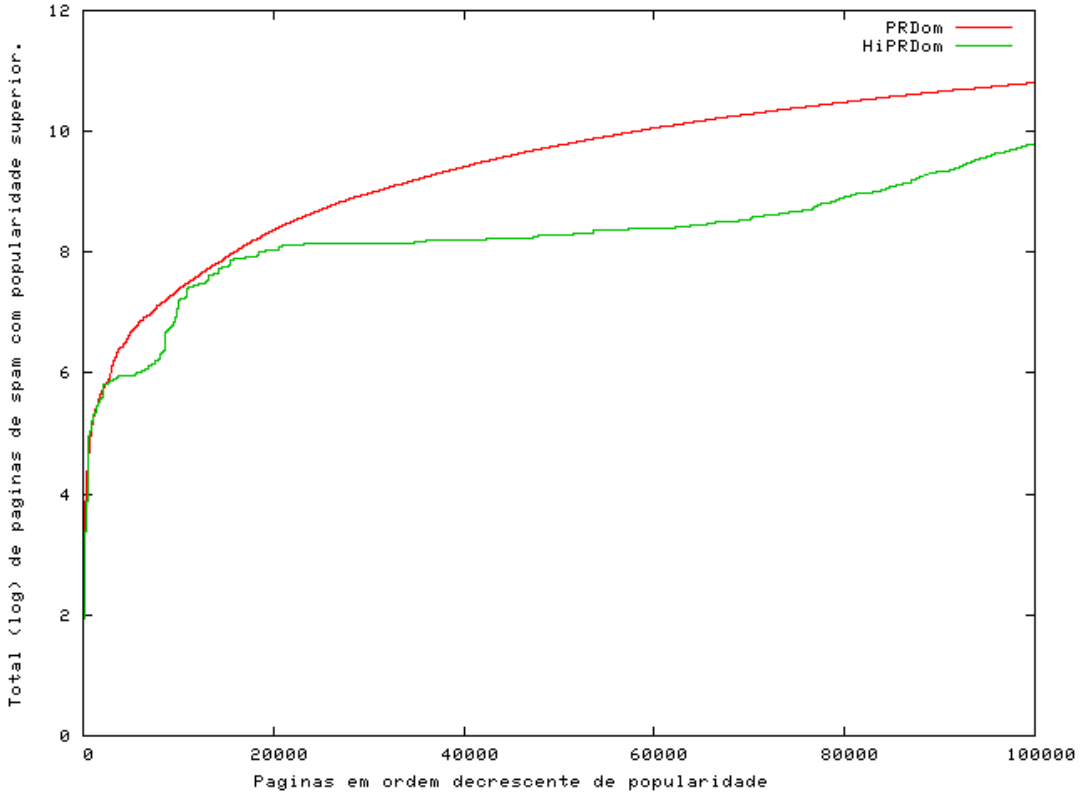

Figura 4.3: Número (log) de páginas de spam com reputação estimada superior a cada uma das 100.000 páginas comuns (não spam) com maior reputação estimada por cada um dos dois métodos.

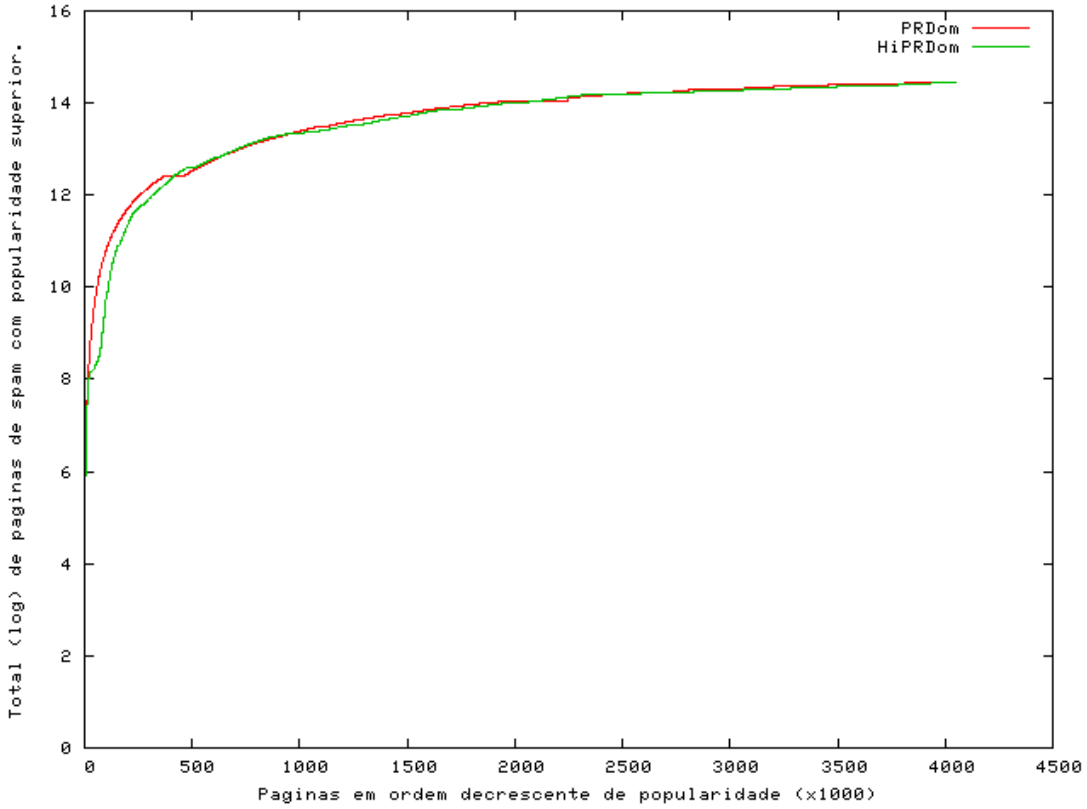

Figura 4.4: Número (log) de páginas de spam com reputação estimada superior a cada uma das páginas da coleção que recebe pelo menos um apontador.

Capítulo 5

Conclusão e Trabalhos Futuros

A principal vantagem de utilizar um modelo de hipergrafo para computar a reputação das páginas é permitir uma modelagem que controle a qualidade das conexões entre páginas web representadas. Este controle é alcançado através da definição apropriada do critério de particionamento adotado para criar as hiperarestas. Tal flexibilidade na definição do critério de particionamento cria possibilidades de futuros estudos que determinem o impacto da definição de diferentes critérios e permitam que os desenvolvedores de máquinas de busca escolham a abstração de hipergrafo mais apropriada para cada coleção utilizada.

Os experimentos apresentados mostram como o modelo de hipergrafo pode ser utilizado para fornecer uma melhor estimativa da reputação das páginas e melhorar a ordenação dos resultados de máquinas de busca. Apresentamos estudos com três métodos de particionamento distintos derivados da hierarquia da URL: páginas, hosts e domínios. Também mostramos exemplos de como adaptar algoritmos tradicionais de análise de apontadores para o modelo de hipergrafo proposto. Nos experimentos realizados com uma base de dados de uma máquina de busca real (a coleção WBR03), os algoritmos de análise de apontadores utilizando hipergrafos apresentaram melhores resultados para consultas navegacionais do que os algoritmos que utilizam grafos. Quando analisamos o desempenho em relação às consultas informacionais, tanto os algoritmos que utilizam grafos quanto os algoritmos que utilizam hipergrafos foram equivalentes. Com isto podemos concluir que ao utilizar algoritmos baseados em hipergrafos com uma modelagem adequada, os resultados obtidos de uma maneira geral são superiores.

Uma desvantagem da utilização do modelo de hipergrafo é que o número de hiperarestas tende a ser menor que o número de arestas na coleção. Deste modo, em coleções com poucos

apontadores, a representação através de hipergrafos pode produzir hipergrafos muito esparsos, com poucas conexões entre os elementos. Nestas situações o modelo de hipergrafo pode causar perda na qualidade dos resultados obtidos. Entretanto, nos experimentos realizados em uma coleção com baixo grau de conectividade (WT10g), o uso do modelo de hipergrafo em tarefas de busca resultou em um desempenho final similar ao obtido quando representamos a Web através de grafos.

Finalmente, realizamos experimentos para analisar os efeitos na reputação de páginas de spam causados pela adoção do modelo de hipergrafo com o domínio como critério de particionamento. Os resultados indicam que a adoção deste modelo causou uma pequena redução na importância dada a páginas de spam por dois algoritmos distintos, HiperIndegree e Hiper-Pagerank em relação ao Indegree e Pagerank respectivamente.

Em trabalhos futuros pretendemos investigar a possibilidade da utilização de algoritmos de agrupamento como estratégia para o particionamento. A idéia é determinar as partições de acordo com as propriedades que se deseja que as mesmas possuam no hipergrafo, tais como forte relacionamento entre suas páginas, em vez de utilizar a hierarquia da URL. Uma outra direção é o estudo das possíveis relações entre o melhor critério de particionamento e as características da coleção utilizando outras coleções web para realizar os experimentos.

Referências Bibliográficas

[1] Peter Bailey, Nick Craswell, and David Hawking. Engineering a multi-purpose test collection for web retrieval experiments. *Information Processing and Management*, 39(6):853–871, 2003.

[2] Tim Bray. Measuring the web. In *Proceedings of the 5th International World Wide Web Conference on Computer Networks and ISDN Systems*, pages 993–1005, Amsterdam, The Netherlands, The Netherlands, 1996. Elsevier Science Publishers B. V.

[3] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International World Wide Web Conference*, pages 107–117.

[4] Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[5] Deng Cai, Xiaofei He, Ji-Rong Wen, and Wei-Ying Ma. Block-level link analysis. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 440–447, New York, NY, USA, 2004. ACM Press.

[6] Pável Pereira Calado, E. S. de Moura, Berthier Ribeiro-Neto, Ilmério Silva, and Nivio Ziviani. Local versus global link information in the web. *ACM Transactions on Information Systems (TOIS)*, 21(1):42–63, 2003.

[7] Nick Craswell, Stephen Robertson, Hugo Zaragoza, and Michael Taylor. Relevance weighting for query independent evidence. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 416–423, New York, NY, USA, 2005. ACM Press.

[8] André; Luiz da Costa Carvalho, Paul Alexandru Chirita, Edleno Silva de Moura, Pável Calado, and Wolfgang Nejdl. Site level noise removal for search engines. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 73–82, New York, NY, USA, 2006. ACM Press.

[9] Cathal Gurrin and Alan Smeaton. Replicating web structure in small-scale test collections. *Information Retrieval*, 7(3-4):239–263, 2004.

[10] Zoltán Gyöngyi and Hector Garcia-Molina. Link spam alliances. In *Proceedings of the 31st International Conference on Very Large Data Bases*, pages 517–528. VLDB Endowment, 2005.

[11] Taher Haveliwala. Topic-sensitive pagerank. In *Proceedings of the 11th International Conference on World Wide Web*, pages 517–526, New York, NY, USA, 2002. ACM.

[12] David Hawking, Ellen Voorhees, Nick Craswell, and Peter Bailey. Overview of the trec8 web track. In *8th Text REtrieval Conference*, 1999.

[13] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677.

[14] Ronny Lempel and Shlomo Moran. Salsa: The stochastic approach for link-structure analysis. *ACM Transactions on Information Systems (TOIS)*, 19(2):131–160, 2001.

[15] Gerard Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. Technical report, Ithaca, NY, USA, 1974.

[16] Gui-Rong Xue, Qiang Yang, Hua-Jun Zeng, Yong Yu, and Zheng Chen. Exploiting the hierarchical structure for link analysis. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 186–193, New York, NY, USA, 2005. ACM Press.

[17] C. T. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers*, 20(1):68–86, 1971.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo