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

Algoritmos Para Avaliação de Confiança em Apontadores Encontrados na Web

Jucimar Brito de Souza

Manaus – Amazonas

2009

Jucimar Brito de Souza

Algoritmos Para Avaliação de Confiança em Apontadores Encontrados na 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 da Informação.

Orientador: Prof. Dr. Edleno Silva de Moura

Jucimar Brito de Souza

Algoritmos Para Avaliação de Confiança em Apontadores Encontrados na 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 da Informação.

Banca Examinadora

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

Prof. Dr. Marcos André Gonçalves
Departamento de Ciência da Computação - UFMG

Prof. Dr. Altigran Soares da Silva
Departamento de Ciência da Computação – UFAM/PPGI

Prof. João Marcos Bastos Cavalcante, Ph.D.
Departamento de Ciência da Computação – UFAM/PPGI

Manaus - Amazonas
2009

Dedicatória

A minha esposa Luciana Souza, aos meus filhos Ana Elisabete e Ian Thiago, aos meus pais Jucimar e Edna Souza pelo incentivo dado para realização deste trabalho.

Agradecimentos

Agradeço primeiramente ao meu Deus por sua infinita bondade e misericórdia em conceder mais esta vitoria ao terminar o Mestrado.

A minha esposa pela força que me deu durante todo o curso, pela ajuda em muitos momentos das atividades do mestrado, pelo apoio em todos os momentos.

Aos meus filhos Ana Elisabete e Ian Thiago por entenderem que muitas vezes o papai não podia fazer os passeios, jogar basquete, ir para piscina em função do meu compromisso com os estudos.

Aos meus pais Jucimar e Edna Souza que sempre me incentivaram a estudar e que sempre vibram com as conquistas dos seus filhos.

Ao meu orientador Edleno Silva de Moura, pela oportunidade e os ensinamentos passados com as suas experiências na área de recuperação da informação que foram de fundamental importância para o sucesso deste trabalho.

A minha amiga Kelen Acquati que me ajudou com a sua experiência nas disciplinas e que esta ajuda foi de fundamental importância para obtenção dos bons resultados conquistados.

Ao Andre Carvalho e Klessius Beltz pelo apoio que me deram para que eu continuasse os trabalhos iniciados por eles e que me ajudaram a concretizar o desenvolvimento dos conceitos dos métodos aqui apresentados.

A Universidade Federal do Amazonas em especial ao Departamento de Ciência da Computação pela oportunidade que me foi dada.

Aos colegas e amigos que direta ou indiretamente me ajudaram para conclusão deste curso.

AGRADEÇO

Como é feliz o homem que acha a sabedoria, o homem que obtém o entendimento, pois a sabedoria é mais proveitosa que a prata e rende mais do que o ouro.

Resumo

Máquinas de busca têm se tornado uma ferramenta imprescindível para os usuários da *Web*. Elas utilizam algoritmos de análise de apontadores para explorar a estrutura dos apontadores da *Web* para atribuir uma estimativa de popularidade a cada página. Essa informação é usada na ordenação da lista de respostas dada por máquinas de busca a consultas submetidas por seus usuários. Contudo, alguns tipos de apontadores prejudicam a qualidade da estimativa de popularidade por apresentar informação ruidosa, podendo assim afetar negativamente a qualidade de respostas providas por máquinas de busca a seus usuários. Exemplos de tais apontadores incluem apontadores repetidos, apontadores resultantes da duplicação de páginas, SPAM, dentre outros. Esse trabalho tem como objetivo detectar ruídos na estrutura dos apontadores existentes em base de dados de máquinas de busca. Foi estudado o impacto dos métodos aqui desenvolvidos para detecção de apontadores ruidosos, considerando cenários nos quais a reputação das páginas é calculada tanto com o algoritmos *Pagerank* quanto com o algoritmo *Indegree*. Os resultados dos experimentos apresentaram melhoria de até 68,33% na métrica *Mean Reciprocal Rank* (MRR) para consultas navegacionais e de até 35,36% para as consultas navegacionais aleatórias quando uma máquina de busca utiliza o algoritmo *Pagerank*.

Palavras-chave: Recuperação da Informação, Máquina de Busca, Análise de Apontadores, Ruído

Abstract

Search engines have become an essential tool for web users today. They use algorithms to analyze the linkage relationships of the pages in order to estimate popularity for each page, taking each link as a vote of quality for pages. This information is used in the search engine ranking algorithms. However, a large amount of links found on the Web can not be considered as a good vote for quality, presenting information that can be considered as noise for search engine ranking algorithms. This work aims to detect noises in the structure of links that exist in search engine collections. We studied the impact of the methods developed here for detection of noisy links, considering scenarios in which the reputation of pages is calculated using Pagerank and Indegree algorithms. The results of the experiments showed improvement up to 68.33% in metric Mean Reciprocal Rank (MRR) for navigational queries and up to 35.36% for randomly selected navigational queries.

Keywords: Information retrieval, search engine, link analysis, noise

Lista de Figuras

Lista de Tabelas

Lista de Gráficos

Lista de Abreviaturas

URL – Uniform Resource Locator

TREC - Text Retrieval Conference

Sumário

Capítulo 1

1. Introdução

Os algoritmos de análise de apontadores em máquinas de busca exploram a estrutura dos apontadores da Web com o objetivo de atribuir uma estimativa de popularidade a cada página. Esta estimativa de popularidade é amplamente utilizada em sistemas de busca na *Web* como fonte de evidência da qualidade de uma página. Sistemas de busca utilizam a estimativa de popularidade como heurística para aferir a qualidade de uma página. Considerando-se que a qualidade de uma página é proporcional a sua popularidade. Essa informação é então utilizada na ordenação de respostas dadas por máquinas de busca a consultas submetidas por seus usuários.

Para alguns sítios é muito importante que sua colocação esteja próxima ao topo das listas de respostas nas consultas em máquinas de busca, estudos passados mostram que 85% dos usuários da Internet utilizam máquinas de busca para localizar informações (KEHOE, et al. 1998). Além disso, muitos usuários de máquina de busca utilizam somente as primeiras respostas da lista apresentadas por máquinas de busca (JOACHISMS, et al. 2005), não levando em consideração os demais resultados na maioria das vezes.

Em muitas situações os primeiros resultados apresentados por máquinas de busca são consideradas ruidosos. Os ruídos podem ser gerados com o intuito de enganar as máquinas de busca, tais como: inclusão de palavras chaves para elevar a relevância da página sobre determinado assunto, inclusão de apontadores repetidos, duplicação de páginas, SPAM, dentre outros.

O SPAM em máquina de busca é uma fonte de ruído muito combatida que ocorre quando evidência são incluídas artificialmente para melhorar a posição de uma página nas respostas de uma máquina de busca. Um dos principais problemas encontrados pelos programas de análise de apontadores é determinar o quão confiável é um apontador para uma determinada página encontrada na Web.

Este trabalho tem como objetivo detectar ruídos na estrutura dos apontadores existentes em base de dados de máquinas de busca. Foram utilizadas como ponto de partida as estratégias de detecção de ruídos propostos por (CARVALHO, et al. 2006), propondo-se modificações para torná-los mais eficazes.

Os algoritmos propostos por Carvalho *et al.* removem os apontadores considerados ruidosos existentes na máquina de busca. Esta remoção depende dos limiares que são estabelecidos através de experimentos, em que um especialista vai especificando valores para o limiar buscando encontrar o melhor valor que maximize os resultados.

Os métodos aqui proposto criam várias escalas, utilizando artifícios matemáticos, que buscam avaliar os relacionamentos existes no grafo que representa a *Web* dando peso aos relacionamentos existente no grafo. Estes pesos são utilizado para indicar o grau de confiança existente em cada relacionamento no grafo.

O impacto dos métodos desenvolvidos para detecção de apontadores ruidosos foi estudado considerando-se cenários onde a reputação de páginas é calculada com os algoritmos *Pagerank* (PAGE, et al. 1998) e *Indegree* (BRAY 1996). Os resultados dos experimentos apresentaram melhoria de 28,41% na métrica *Mean Reciprocal Rank* (MRR) (HAWKING, et al. 1999) para consultas navegacionais em relação aos métodos propostos por (CARVALHO, et al. 2006) e de 68,33% em relação ao *Pagerank*. Utilizando consultas navegacionais

aleatórias, os experimentos mostraram uma melhoria no MRR de 21,47% em relação aos métodos de (CARVALHO, et al. 2006) e de 35,36% em relação ao *Pagerank*.

Esta dissertação deixa como contribuição o uso de uma escala para indicar a confiança nos relacionamentos existentes no grafo da Web evitando a intervenção humana para sua obtenção, avaliação dos experimentos em agrupamentos de domínios e avaliação dos resultados das consultas no cenário em que combina não somente a reputação das páginas calculadas pelos métodos aqui propostos como também o texto de âncora e o texto da página.

1.1 - Trabalhos Relacionados

Um estudo inicial sobre apontadores ruidosos pode ser observado no trabalho proposto por (BHARAT e HENZINGER 1998) que desenvolveram uma análise de relacionamento entre páginas baseado no algoritmo de Kleinberg (KLEINBERG 1998). Este estudo identificou que o algoritmo de Kleinberg falhava por três razões: o relacionamento entre sítios reforçavam-se mutuamente (Mutually Reinforcing Relationships Between Hosts), os apontadores eram gerados automaticamente por alguma ferramenta de desenvolvimento de páginas sem nenhuma intervenção humana (Automatically Generated Links) e quando encontrava uma página sem relevância para um determinado tópico (Non-relevant Nodes). Apesar deste estudo não tratar diretamente o problema de detecção de ruído, estes relacionamentos são fortes candidatos a apontadores ruidosos e foram considerados nos métodos aqui propostos.

O estudo desenvolvido por (GYÖNGYI e GARCIA-MOLINA 2005) descreve como as páginas Web podem ser interconectadas para gerar uma estrutura denominada *link spam farm* (estrutura criada pelos provedores de *SPAM* para interconectar páginas), buscando

aumentar artificialmente a estimativa de popularidade de páginas Web. Este trabalho observa como um *spammer* (criador de SPAM) pode aumentar a estimativa de popularidade de uma página. O trabalho mostra também como um grupo de *spammers* pode colaborar entre si fazendo alianças que interconectam as *spam farms.* Os algoritmos aqui propostos podem auxiliar na identificação de *spams farms*, minimizando o impacto destas estruturas nos algoritmos *Pagerank e Indegree.*

Em (ZHANG, et al. 2004) identificou-se que os valores de Pagerank podem ser fraudados adicionando-se apontadores às páginas, blogs, etc. Os autores criaram uma métrica chamada fator de amplificação, definida por $Out(1/\varepsilon)$ onde ε é a probabilidade de um usuário acessar uma página no *Pagerank* (dump factor). Desta forma, eles propõem que o Pagerank seja calculado com um dump factor diferente para cada página p, gerado automaticamente como uma função do coeficiente de correlação entre $1/\varepsilon$ e o *Pagerank*(p) para diferentes valores de ε, dando ao *Pagerank* mais robustez contra os apontadores. Nesta dissertação foi proposta uma escala de confiança que, ao contrário do fator de amplificação proposto por (ZHANG, et al. 2004), efetua uma atenuação nas relações entre sítios. Assim evita-se que um apontador ruidoso tenha grande impacto na estimativa de popularidade de uma página quando esta é computada com o algoritmo *Pagerank*.

GYÖNGYI et al. (2004) propuseram uma técnica semi-automática usada para separar páginas de boa qualidade de páginas consideradas SPAM. Primeiramente, um pequeno conjunto de páginas, chamadas de sementes, é selecionado e avaliado por especialistas com o objetivo de identificar as páginas consideradas de boa reputação e livres de SPAM, dando a estas o peso um e às demais peso zero. Após a seleção deste conjunto de páginas, a estrutura dos apontadores deste conjunto é usada para encontrar páginas semelhantes às páginas sementes, ou seja, páginas livres de SPAM. Isto é realizado pelo algoritmo TrustRank, que

utiliza este conjunto para gerar um valor de confiança entre zero (para páginas consideradas ruins) e um (para páginas consideradas boas) . Esta dissertação utiliza uma escala de confiança com o mesmo conceito. Porém, seus valores são gerados partindo-se da analise de apontadores realizada pelos algoritmos aqui propostos e que não dependem de especialistas para efetuar qualquer avaliação no que se refere a SPAM.

Os autores do Topical TrustRank (WU, et. al. 2006) identificaram que se as "páginas sementes" selecionadas para o algoritmo TrustRank contiverem mais páginas de um determinado tópico, estas páginas tendem a gerar um valor de confiança maior que os tópicos com menor número de páginas. O algoritmo Topical TrustRank usa as informações de tópicos para separar os conjunto de "páginas sementes" e calcular valores de confiança para cada tópico separadamente. Estes novos valores de confiança são propagados pelas páginas e utilizados para determinar o ranking.

Carvalho et al.(2006) propuseram e avaliaram algoritmos para identificar apontadores ruidosos em coleções *Web*, sejam eles SPAM ou simplesmente relações entre entidades representadas no mundo real por sítios, replicação de conteúdo, etc. Os algoritmos propostos para identificar relacionamentos de reforço mútuo (quando dois sítios trocam muitos apontadores) e suporte anormal (quando um sítio tem o percentual muito alto do total dos apontadores que um sítio recebe) removem da base de dados os apontadores considerados suspeitos durante a estimativa de popularidade das páginas.

Neste trabalho foram propostos novos métodos para detecção de apontadores ruidosos que modificam os métodos propostos por (CARVALHO, et al. 2006). Os novos métodos utilizam uma escala de confiança nos relacionamentos dos sítios da *Web* para aumentar ou diminuir a influência de cada apontador durante o computo da popularidade de uma pagina.

Os experimentos realizados mostram ganhos significativos na qualidade de respostas providas por máquinas de busca quando estas empregam os algoritmos aqui propostos.

1.2 – Organização da Dissertação

Esta dissertação está organizada da seguinte forma: no capítulo 2 são apresentados os conceitos básicos para o melhor entendimento dos assuntos abordados nesta dissertação, no capítulo 3 são apresentados os novos conceitos e métodos aqui desenvolvidos, no capítulo 4 são apresentados os resultados dos experimentos bem como as análises dos resultados e no capítulo 5 são apresentadas as conclusões e trabalhos futuros.

Capitulo 2

2. - Conceitos Básicos

2.1. - Máquinas de Busca

Máquinas ou motores de busca são sistemas desenvolvidos para efetuar consultas na *Web* realizadas através de palavras-chaves fornecidas pelos seus usuários. De acordo com (BRIN e PAGE 1998) maquinas de busca são compostas por três módulos básicos: coletores, indexadores e processadores de consulta.

O processo de coleta geralmente é feito por vários coletores, que são chamados robôs, cuja finalidade é receber um conjunto de URLs sementes e através delas navegam pela *Web* armazenando as páginas encontradas. Cada página recebe um código de identificação único que será usado posteriormente pelo indexador. A cada página lida são identificados os apontadores de saída, os quais são colocados em uma fila para que os robôs possam visitar posteriormente. Este processo é continuo, e os coletores devem ter métodos para atualizar as páginas que são alteradas na *Web* e mantê-las na base de dados do sistema.

O indexador cria um índice para acesso rápido ao conteúdo da base de dados do sistema. Este índice é também conhecido como índice invertido ou lista invertida. Ele é responsável pelo armazenamento dos termos e da quantidade de vezes que o termo ocorreu em cada documento, ou a posição do termo dentro do documento.

Após a indexação, o módulo de processamento de consulta está pronto para ser executado. Geralmente, este módulo fica disponível para ser acessado através da Internet. Seus usuários formulam consultas informando palavras chaves que caracterizam suas

necessidades de informação. O processador de consultas recebe estas palavras e submete aos seus algoritmos que retornam uma lista ordenada de documentos como resposta as consultas. A ordem desta lista é gerada a partir da análise de diversas fontes de evidências sobre a possível relevância dos documentos da base como resposta à consulta formulada pelo usuário.

2.1.1 – Fontes de evidências para consultas em máquinas de busca

Encontrar documentos que satisfaçam a necessidade de informação do usuário não é uma tarefa fácil. A base de dados da maioria das máquinas de busca disponíveis é composta por documentos que representam as páginas W*eb*. Com o intuito de estimar as melhores respostas para as consultas dos usuários, as máquinas de busca utilizam diversas evidências para selecionar as páginas mais relevantes para a consulta em questão. Dentre as diversas evidências utilizadas para estimar a reputação das páginas cita-se:

- Texto de Página: É o texto que compõe uma determinada página. Deste texto são extraídas as palavras (também chamadas de termos), as máquinas de busca utilizam algoritmos para identificar os termos de qualidade e/ou popularidade. As informações que não fazem parte do texto são desconsideradas tais como os códigos HTML (*tags*).
- Texto de Âncora: Texto de âncora são palavras que aparecem em conjuntos com os códigos HTML que trazem o endereço (URL) de uma determinada página. Por exemplo: Toyota. A palavra Toyota está relacionada com o endereço http://www.toyota.com. Este texto é utilizado como uma fonte de evidência nas máquinas de busca. De acordo com (BRIN e PAGE 1998) o texto de âncora descreve melhor o conteúdo da página e ele pode existir para documentos que não foram indexados pela máquina de busca como é o caso de imagens, programas e banco de dados.

- Apontadores: também conhecidos pela palavra inglesa *links*, são usados por diversos algoritmos de análise de apontadores para estimar a reputação de uma página. Os algoritmos desenvolvidos nesta dissertação utilizam esta evidência para dar graus de confiança a estes apontadores.

Estas evidências foram utilizadas na implementação da máquina de busca usada nos experimentos apresentados nesta dissertação. No caso de evidências textuais, é necessário que se utilize um modelo reconhecido na literatura para computar a similaridade entre documentos e consultas. O modelo adotado aqui foi o modelo vetorial (SALTON e YANG 1975).

2.1.2 – Modelo Vetorial

O modelo vetorial é considerado o modelo clássico da recuperação da informação (SALTON e McGILL 1983). Ele representa documentos e consultas em forma de vetores que são dispostos em um espaço n-dimensional, onde n é o número de termos distintos existentes na coleção de documentos.

Documentos e consultas são representados como vetores que têm em suas coordenadas a representação da importância de cada um dos termos no vocabulário da coleção. O documento d_j e a consulta q podem ser representados pelos vetores $\overrightarrow{d_j} = \langle w_{t1j}; w_{t2j}; \ldots; w_{tkj}; \rangle$ e $\vec{q} = \langle w_{t1q}; w_{t2q}; \ldots; w_{tkq}; \rangle$, onde k é o número de termos existente na coleção e w_{tix} é o peso que representa a importância do termo *ti* para o documento ou consulta *x*. O peso do termo no documento é proporcional a freqüência com que ele ocorre no documento e a importância deste termo na coleção. A equação (**1**) apresenta como calcular o peso:

$$w_{ix} = tf_{ix} \times idf_i \quad (1)$$

Onde x é o documento da coleção, i é um termo do documento ou consulta, tf_{ix} é a freqüência do termo i no documento x e idf_i é o inverso da freqüência do termo na coleção de documentos. O idf mede a raridade do termo na coleção, ou seja, quanto maior o idf mais raro é o termo da coleção. O idf é calculado conforme a equação (2)

$$idf_i = \log \frac{N}{n_i} \tag{2}$$

Onde N é o número de documentos da coleção e n_ié o número de documentos onde ocorre o termo i.

Após a representação de todos os documentos como vetores é possível efetuar a similaridade entre os documentos e a consulta. Salton propôs o uso da similaridade entre cossenos, ou seja, o valor do cosseno do ângulo formado entre o vetor da consulta e o vetor do documento gera um valor de similaridade entre eles. A similaridade entre o documento d_j e a consulta q pode ser obtida através da equação (3).

$$Sim(d_j, q) = \cos\theta = \frac{\sum_{i=1}^{k} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{k} w_{ij}^2} \times \sqrt{\sum_{i=1}^{k} w_{iq}^2}} \tag{3}$$

Onde k é o número de termos existe na coleção, $\sqrt{\sum_{i=1}^{k} w_{ix}^2}$ é a norma da consulta ou do documento.

A norma representa a quantidade de informação presente em um documento. Uma de suas funções, no modelo vetorial, é penalizar os documentos maiores e beneficiar os menores. Como o modelo vetorial é baseado na freqüência dos termos, logo documentos mais longos possuiriam uma vantagem já que possuem um vocabulário maior e conseqüentemente a

freqüência destes termos é maior se comparada com documentos pequenos. Portanto é conveniente dar pesos diferentes de acordo com a quantidade de informação presente.

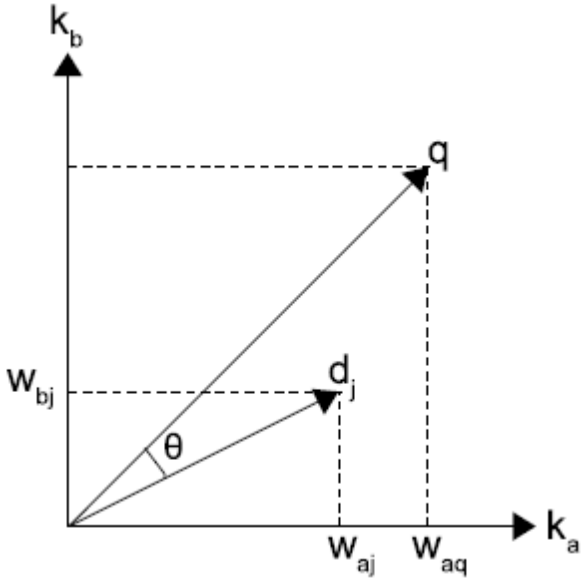

Figura 1 – Similaridade entre d_j e q no modelo vetorial

A Figura 1 apresenta um documento d_j composto por dois termos k_a e k_b e uma consulta q. A similaridade entre a consulta q e d_j é representada pelo cosseno de θ.

2.2 - Análise de Apontadores

A informação de conectividade entre páginas *Web* é uma das evidências mais importantes nas máquinas de busca. É através dela que é formada a estrutura dos apontadores entre as páginas. Os algoritmos de análise de apontadores partem da hipótese de que um apontador de uma página A para uma página B indica que o autor da página A recomenda a página B como de qualidade. Assim, é possível analisar toda coleção de páginas e estimar quais são as páginas mais interessantes ou de melhor reputação.

Esta identificação é importante principalmente para ordenação das respostas em cada consulta, pois a reputação é utilizada como evidência para estimar o grau de relevância das páginas. As páginas de melhor reputação têm preferência na ordem em que as respostas são

mostradas para os usuários, pois estima-se que o topo da lista de respostas é composto por páginas de melhor qualidade.

Nestes algoritmos, a *Web* é representada como um grafo dirigido $G = (V, E)$, onde V é conjunto de vértices que representam as páginas da Web, e E o conjunto de arestas direcionadas que representam os apontadores que saem das páginas, formando os pares ordenados $(p, q) | \, p, q \in V$ indicando que existe um apontador de p para q. Logo $In(p)$ representa o conjunto de páginas que apontam para p e $Out(p)$ o conjunto de páginas apontadas por p.

Algumas estratégias foram propostas na literatura para efetuar analise de apontadores (BRAY, 1996; BRIN e PAGE,1998; KLEINBERG, 1998; XUE, et al. 2005), as quais têm como idéia principal identificar o quanto o apontador de uma página para outra representa um voto de confiança.

2.2.1 - *Indegree*

O algoritmo *Indegree* foi proposto em (BRAY 1996) e baseia-se na hipótese de que se uma página aponta para outra página**,** isto é uma evidência de que a página destino tem um bom conteúdo e um voto de qualidade. O valor do *Indegree* pode ser calculado conforme a equação (4):

$$Indegree(p) = \|In(p)\| \tag{4}$$

Onde $\|In(p)\|$ é o total de apontadores que apontam para p.

Neste algoritmo a qualidade da página é diretamente proporcional à quantidade de páginas que apontam para ela. Este método tem um bom desempenho. Porém, por ser muito simples ele é muito susceptível a ruídos.

2.2.2 - *Pagerank*

O algoritmo *Pagerank* proposto por (PAGE, et al. 1998) parte da seguinte hipótese: "Um apontador vindo de uma página popular é uma fonte de evidência mais forte do que um apontador vindo de uma página desconhecida".

O *Pagerank* tenta estimar a probabilidade de um usuário navegando pela Internet chegar a uma página durante um caminhamento aleatório. No primeiro momento o usuário tem a probabilidade de visitar qualquer uma das páginas de V, portanto a probabilidade é dada pela equação (5).

$$PR(p) = \frac{1}{\|V\|} \tag{5}$$

Onde $|V|$ é o número de vértices existente no grafo que representa a *Web* e $PR(p)$ é o *Pagerank* da página p.

Ao acessar a página o usuário tem duas opções: visitar uma das páginas apontadas pela página corrente ou ir para outra página aleatoriamente. Se ele seguir pelos apontadores da página, a probabilidade de visitar qualquer página apontada por p é dada pela equação (6):

$$PR(np) = \frac{PR(p)}{\|Out(p)\|} \tag{6}$$

Onde $Out(p)$ é o total de apontadores que saem da página p.

Observa-se que a probabilidade da página p é propagada para nova página $np \in Out(p)$ que acontece seguindo os apontadores da página $Out(p)$. Uma demonstração simplificada deste cálculo pode ser visualizada na Figura 2. Pode-se dizer que quanto maior a probabilidade da página corrente maior será a influencia desta na probabilidade de chegar às páginas que ela aponta.

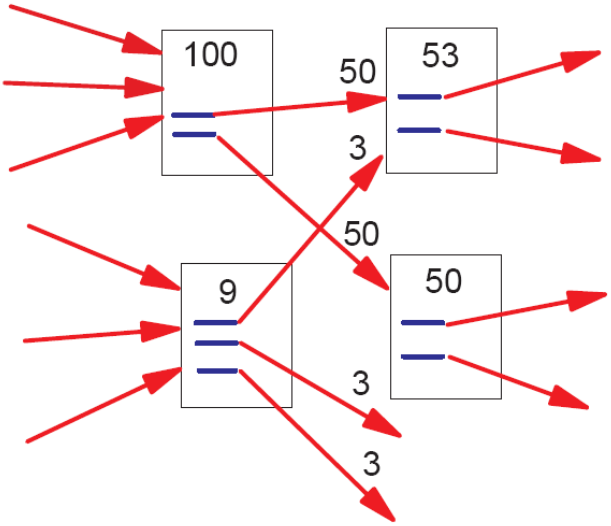

Figura 2 – Calculo simplificado do Pagerank

Fonte: (PAGE, et al. 1998)

O valor do *Pagerank* de uma página é dado pela probabilidade de um usuário navegando na Web chegar àquela página, conforme a equação (7).

$$PR(p) = (1 - c) \times \sum_{q \in In(p)} \frac{PR(q)}{\|Out(p)\|} + \frac{c}{\|V\|} \quad (7)$$

Onde c é a probabilidade do usuário escolher uma página de V ao invés de seguir os apontadores da página corrente. O valor de 0,15 para c foi considerado por (PAGE, et al. 1998) como satisfatório.

2.3 – Tipos de Consulta

Identificar a necessidade de informação que esta por trás de uma consulta não é uma tarefa simples. Os estudos realizados por (BRODER 2002) propôs uma taxonomia para consultas na *Web*. Que as classificou as consultas de acordo com a intenção do usuário, são elas:

- **Informacionais:** ocorrem quando o usuário está em busca de uma informação sobre um determinado tópico ou assunto. Por exemplo: o usuário ao consultar "sincronizar celular com PC", deseja informações de como realizar o sincronismo do computador pessoal com o celular e qualquer informação sobre o assunto pode ser considerada uma boa resposta.
- **Navegacionais**: ocorrem quando o usuário está em busca de um endereço de uma determinada página, que pode ser uma empresa, órgão governamental, instituição e outros. Por exemplo: o usuário ao consultar "CEFET-AM" provavelmente está procurando o sítio do Centro Federal de Educação Tecnológica do Amazonas e para ele não interessa informações sobre o CEFET-AM e sim a resposta com o endereço do sítio.
- **Transacionais**: ocorrem quando o usuário tem o objetivo de efetuar uma transação, ou seja, quando o usuário pretende compra um produtos, procurar arquivos para *download*, pesquisa pacotes de viagem, etc. Por exemplo: o usuário ao consultar "Comprar celular N95" indica que ele está a procura de sítios que vendam o celular N95 da Nokia e como resposta ele espera receber uma lista de sítios que vendam este produto.

2.4 – Métricas de avaliação

Os métodos propostos nesta dissertação foram avaliados com a métrica do MRR (HAWKING, et al. 1999) que é a mais indica para as consultas navegacionais. A métrica MRR tem como finalidade de avaliar os resultados das consultas submetidas a máquinas de busca com somente uma resposta correta. Por este motivo foi adotada pela *Text Retrieval Conference* (TREC[1]) para avaliar a qualidade dos sistemas no processamento de consultas navegacionais. Boas respostas para as consultas navegacionais são aquelas que aparecem próximo ao topo da lista de respostas. Assim os valores do MRR são maiores para os sistemas que acertam com mais freqüência as respostas para as consultas navegacionais. A equação (8) apresenta a forma de calcular o MRR:

$$MRR(QS) = \frac{\sum_{\forall\, q_i \in QS} \frac{1}{PosRespCorreta(q_i)}}{|QS|} \tag{8}$$

Onde QS é um conjunto de consultas avaliadas e $PosRespCorreta(q_i)$ é a posição em que se encontra a resposta correta da consulta i. O resultado do MRR é um número real entre 0 e 1, sendo 1 o melhor resultado possível para um sistema. Isto indica que em todas as consultas submetidas a máquinas de busca a resposta certa estava na primeira posição da lista.

[1] http://trec.nist.gov/

2.5 – Tipos de Apontadores Ruidosos

A existência de um apontador de uma página A para uma página B na *Web* é tratada por algoritmos de análise de apontadores como sendo um voto do criador da página A a favor da popularidade da página B. Entretanto, por diversas razões há apontadores que não podem ser considerados como votos, ou cujo voto não é plenamente confiável. Alguns exemplos de situações que criam esses tipos de apontadores são:

- Trocas de Apontadores: Um caso típico de troca de apontadores acontece quando os autores de sítios se conhecem, e por amizade realizam esta troca de apontadores. Muitas vezes, o conteúdo de suas páginas não se relaciona, mas mesmo assim a troca é realizada.
- Apontadores Navegacionais: Estes apontadores ocorrem com freqüência em páginas tipo *menu* de opções, usadas para navegar entre páginas do mesmo sítio, onde não existe a menor preocupação com qualidade de página ou algo do gênero. Os apontadores navegacionais que apontam para páginas do mesmo sítio são facilmente detectados e podem ser removidos no momento da execução dos algoritmos de análise de apontadores. Entretanto, os apontadores navegacionais para sítios distintos, como é o caso de grandes corporações que têm suas páginas apontando para sítios das filiais, são difíceis de serem detectados e são tratados pelos métodos aqui propostos.
- SPAM: Ocorrem quando fontes de evidências são incluídas artificialmente para melhorar a posição de uma página na ordenação das respostas de uma máquina de busca. Uma das técnicas usadas para criação de SPAM é a inclusão de um número anormal de páginas apontando para uma única página.

Desta forma os algoritmos de análise de apontadores identificam e contam este número de apontadores, classificando esta página mais próxima do topo da lista de respostas nas consultas (GYÖNGYI et al., 2004). Os criadores de SPAM investem seus esforços procurando criar estruturas que não sejam detectadas facilmente por algoritmos de análise de apontadores.

2.6 – Detecção de Apontadores Ruidosos

Nesta dissertação buscou-se identificar os mesmos relacionamentos entre sítios propostos por (CARVALHO, et al. 2006) para detectar apontadores ruidosos, que são os seguintes:

- Reforço Mútuo – Ocorre quando muitos apontadores são trocados entre dois sítios.
- Suporte Anormal – Ocorre quando um sítio é responsável por um percentual muito alto do total de apontadores que outro sítio recebe.
- Aliança de Apontadores – Ocorre quando uma estrutura complexa de apontadores é criada para melhorar os valores do Pagerank de suas páginas.

A identificação destes tipos de relacionamento entre sítios é um indicativo de que os seus apontadores podem ser ruidosos e, neste caso, optou-se em penalizá-los através do uso de pesos de acordo com a confiabilidade calculada pelos métodos aqui propostos. Essa medida de confiabilidade é a principal contribuição desta dissertação em relação aos método propostos por Carvalho et al. (2006). No capítulo 3 são apresentados os dois novos métodos para estimar a confiabilidade de apontadores em base de dados de máquina de busca bem como suas combinações com o método SLLA, proposto por Carvalho et al..

2.8 – Bi-Directional Mutual Site Reinforcement (BMSR)

O algoritmo BMSR busca identificar os relacionamentos de reforço mútuo identificando os apontadores que são trocados entre as páginas de dois sítios. Pode-se dizer que duas páginas *p1* e *p2* trocam apontadores entre si, se existirem apontadores de *p1* para *p2* e apontadores de *p2* para *p1*. O BMSR conta o número de trocas de apontadores existentes entre páginas de grupos distintos que podem ser *Host* ou Domínios, eliminado as trocas encontradas quando estas ultrapassam um limiar previamente estabelecido.

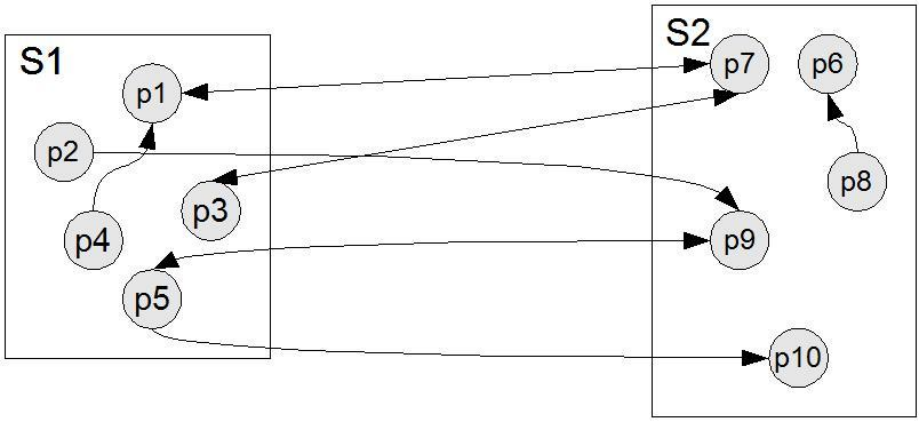

Figura 3 – Exemplo de Troca de Sítios detectado pelo Trust-BMSR

A Figura 3 ilustra o cálculo do BMSR. Existe troca entre as paginas p1-p7, p3-p7,p5-p9. Neste caso o número de trocas entre S1 e S2 é 3.

2.9 – Site Level Abnormal Support (SLAbS)

O método SLAbS proposto por (CARVALHO, et al. 2006) busca identificar os relacionamentos de suporte anormal entre sítios que ocorrem quando um único sítio é responsável por um percentual muito alto do total de apontadores que o outro sítio recebe. Esta situação forma uma cadeia de apontadores que pode ser detectada facilmente dentro de coleções na Internet. Estas cadeias podem ser geradas artificialmente pelos criadores de

SPAM através da criação de sítios como pontos de apoio para formar estruturas conforme a Figura 4.

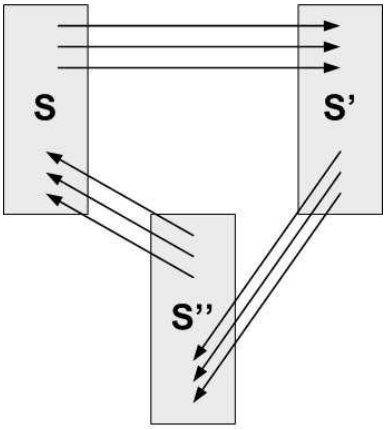

Figura 4 - Exemplo de cadeia de apontadores entre sítios

Para identificar ruído neste tipo de relacionamento, o algoritmo parte do seguinte princípio: ”O total de apontadores para um sítio, ou seja, a soma dos apontadores para suas páginas, não deve ser fortemente influenciado pelas relações que recebe a partir de algum outro sítio”. Em outras palavras, dado um sítio *s*, o algoritmo conta, para cada sitio s_i que aponta para s, o percentual do total de apontadores que *s* recebe vindo de s_i. Se este percentual é menor que um determinado limiar o relacionamento é considerado suspeito. Quando um par de sítios (*s,si)* é considerado ruidoso, todos os apontadores entre eles são removidos da base.

A Equação para o cálculo do percentual de suporte pode ser visto em (9):

$$PercSuporte(S_{Ori}, S_{Dest}) = \frac{TotOutLink(S_{Ori})}{TotInLink(S_{Dest})} \quad (9)$$

Onde $TotOutLink(S_{Ori})$ é o total de apontadores que saem do sítio S_{Ori} para o S_{Dest}, e $TotInLink(S_{Dest})$ corresponde ao total de apontadores que o S_{Dest} recebe.

A Figura 5 mostra um exemplo do calculo do suporte entre sítios: o sítio S_0 recebe 10 apontadores ($TotInLink(S_0)$). O sítio S_1 tem 3 apontadores saindo para S_0 ($TotOutLink(S_1)$), portanto:

$$PercSuporte(S_1, S_0) = \frac{3}{10}$$

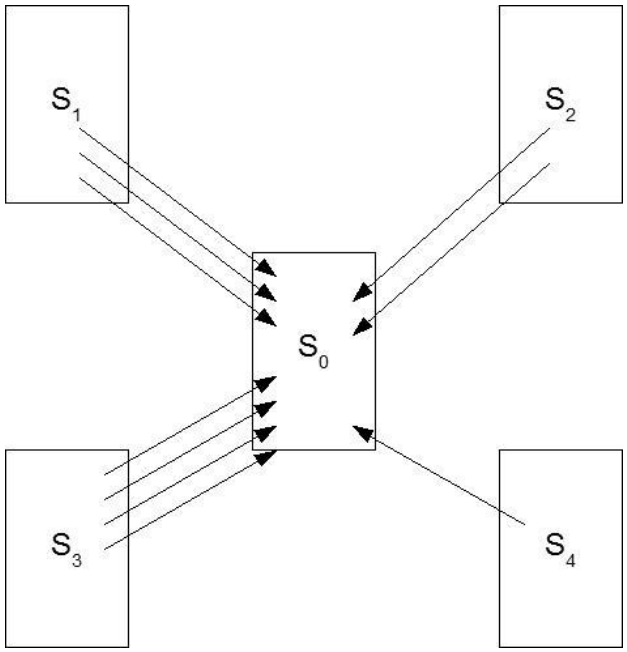

Figura 5 – Cálculo do suporte anormal com SLAbS

2.10 –Site Level Link Aliances (SLLA)

Este algoritmo proposto por (CARVALHO, et al. 2006) considera que a popularidade de um sítio não pode ser influenciada por um grupo fechado de sítios fortemente conectados (CHAKRABARTI 2001). A idéia é que um sítio na Web é popular quanto mais diverso e independente são os apontadores que o conectam.

O conceito de independência por diversidade, usado neste método, parte da hipótese de que um conjunto de sítios é independente se eles não tiverem apontadores trocados entre si.

Na Figura 6 pode ser visto um conjunto de sítios totalmente dependente, onde uma comunidade densamente conectada tem em seus sítios apontadores que apontam para outros sítios comuns neste conjunto. Esta relação teria um percentual de independência próxima de zero de acordo com esta abordagem. Portanto, quanto maior o número de apontadores para outros sítios diferentes deste conjunto, maior o percentual de independência.

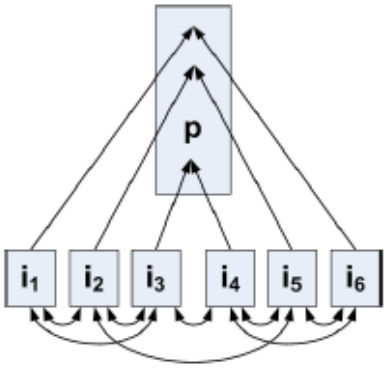

Figura 6 - Exemplo de uma aliança de sítios

A equação (10) mostra como é calculada a independência por diversidade:

$$Independência(C) = \frac{\sum_{l \in Out(\,In(C))} independente(l, In(C))}{|Out(\,In(C))|} \quad (10)$$

Onde C é o conjunto de sítios, $Out(\,In(C))$ é o conjunto de páginas apontadas por páginas que pertencem ao conjunto C, $In(C)$ é o conjunto de páginas que apontam para C e $independente(l, In(C))$ tem seu valor de acordo com a equação (11).

$$independente(l, In(C)) = \begin{cases} 1, & l \notin In(C) \\ 0, & l \in In(C) \end{cases} \quad (11)$$

2.10.1 – Cálculo da independência das páginas

O Algoritmo 1 foi utilizado para calcular a independência de cada página:

1. **Para** cada página *p* **faça**
2. **Faça** *Tot* contar o número de apontadores de todas as páginas $q \in In(p)$
3. **Faça** *TotIn* contar o número de apontadores de todas as páginas $q \in In(p)$ e que apontem para outras páginas $\in In(p)$
4. **Para** cada página $q \in In(p)$ **faça**
5. **Para** cada página $t \in Out(q)$ **faça**
6. $Tot = Tot + 1$;
7. **Se** $t \in In(p)$ **então**
8. $TotIn = TotIn + 1$;
9. Independência(p) = $\frac{Tot - TotIn}{Tot}$

Algoritmo 1 - Cálculo da independência das páginas

2.10.2 – Adaptação do valor de independência para escala de confiança

A informação de independência foi adaptada ao método de análise de apontadores *Pagerank*. Optou-se por calcular para cada sítio existente na base de dados, a independência do conjunto de sítios que apontam para eles. A independência de cada sítio *s* é calculada desta forma:

$$Independência(s) = Independência(\ In\ (\ s\ ))\qquad (12)$$

Onde *In(s)* é o conjunto de sítios que possuem páginas com apontadores para *s*.

Com os valores de independência de cada sítio calculados, utilizam-se destes valores para dar pesos aos apontadores de acordo com os sítios que os recebem. Os apontadores que apontam para uma página *p* de um determinado sítio têm um peso proporcional ao valor de

independência deste sítio. Assim, as páginas que recebem apontadores de um conjunto de sítios dependentes tendem a ter seu peso menor se comparados aos de uma página que recebem apontadores de um conjunto de sítios mais independentes.

Observou-se que este método possui uma escala de confiança onde estes pesos foram utilizados para efetuar o calculo do *Pagerank*. Este método serviu de base para o desenvolvimentos de escalas de confiança apresentados no próximo capítulo.

Capítulo 3

3 - Escalas de Confiança para Apontadores da Web

Carvalho et al. propuseram métodos interessantes para a remoção de apontadores ruidosos em máquinas de busca em um trabalho recente (CARVALHO, et al. 2006). Os métodos apresentados estudam os relacionamentos entre sítios *Web* para detectar e remover apontadores ruidosos, obtendo ganhos significativos na qualidade de sistemas de busca executados sobre a base após a remoção de apontadores ruidosos em relação aos mesmos sistemas utilizando todos os apontadores da base. Contudo, cada método apresentado por Carvalho et al. depende de ajustes de parâmetros usados para classificar apontadores de forma binária como ruidoso ou não ruidoso. Apesar dos resultados apresentados serem interessantes, o esforço operacional para ajustar parâmetros para os métodos era muito grande. Pois os parâmetros propostos precisavam ser ajustados a cada coleção onde os métodos eram aplicados.

Este processo poderia ser inviável para utilização em máquinas de busca disponíveis na Web, devido à atualização constante da suas bases de dados e à morosidade do processo para obtenção dos parâmetros.

O método Bi-Directional Mutual Site Reinforcement (BMSR) proposto por Carvalho et al. identifica o número de troca de apontadores existente entre sítios com o intuito de encontrar relacionamentos que se reforçam mutuamente com intuito de melhorar a posição de suas páginas na ordem das respostas das máquina de busca. Este método propõe um limiar

para estas trocas de tal forma que os pares de sítios que atingem este limiar suas arestas são removidas do grafo da *Web*.

Por exemplo: Quando o limiar do BMSR é configurado para 20 trocas, os pares de sítios com 20 trocas ou mais são considerados ruidosos e seus apontadores são removidos da base, mas um par de sítios com 19 trocas não é penalizado. Este exemplo mostra que o intervalo para considerar uma página ruidosa ou não ruidosa é muito tênue e que existe a possibilidade de perda de informação que pode ser relevante nos relacionamentos eliminados.

Com o intuito de diminuir o esforço para encontrar os limiares dos métodos e evitar perda de informação na remoção das arestas do grafo da *Web*, propõe-se nesta dissertação uma escala que estabelece valores de confiança através dos relacionamentos existente entre os sítios. Os valores de confiança ficam entre 0.0 (não é confiável) e 1.0 (completamente confiável), indicando o quão confiáveis são os relacionamentos entre os sítios. Estes valores serão usados como pesos na adaptação realizada no cálculo do Pagerank e Indegree.

3.1- Construção das escalas de confiança

Os métodos BMSR e SLAbS foram adaptados aqui para gerar estimativas de confiança em apontadores com base em estatísticas computadas a respeito do relacionamento entre sítios. Os métodos adaptados para usar escalas de confiança serão denominados aqui como Trust-BMSR e Trust-SLAbS. As métricas para obter a escala de confiança experimentadas para cada método foram: Razão de valores, média de valores, probabilidade de valores e entropia de valores.

Estas métricas foram experimentadas no calculo do valor de confiança dos métodos Trust-BMSR e Trust-SLAbS propostos nesta dissertação. Para facilitar a representação das fórmulas utilizadas nesta seção, definiu-se:

- RES (Relação entre Sítios) como o número de troca entre sítios para o Trust-BMSR ou o percentual de suporte para o Trust-SLAbS.

3.1.1 – Escala de Razão de Valores

A razão pode ser expressa na forma de divisão entre duas grandezas de algum sistema de medidas. Ao usar este recurso objetivou-se identificar o quanto um valor de uma relação representa em relação ao maior valor existente no grafo da *Web*. Por isso foi identificado, primeiramente, qual o maior valor existente nas relações entre sítios no grafo da Web e construiu-se uma escala proporcional ao maior valor de relação entre sítios dos métodos. A equação (13) apresenta como pode ser obtido o valor da maior relação. Este valor foi usado no cálculo do valor de confiança para que os pares de sítios com maior número de relações tivessem um peso menor que os sítios com poucas relações.

$$MaiorRelação = \max([RES(S_{Ori}, S_{Dest})]) \quad (13)$$

Onde $[RES(S_{Ori}, S_{Dest})]$ representa um conjunto com todos os pares de sítios que efetuam trocas no caso Trust-BMSR ou que tenham percentual de suporte no caso do Trust-SLAbS no grafo da Web.

O valor de *MaiorRelação* foi usado para gerar a escala de confiança da razão das relações que um par de sítios representa em relação ao todo:

$$ValorConfiançaDaRazão(S_{Ori}, S_{Dest}) = 1 - \frac{RES(S_{Ori}, S_{Dest})}{MaiorRelação} \quad (14)$$

3.1.2 – Escala da Média de Valores

A média é uma das medidas de posição mais utilizadas na estatística. As medidas de posição têm por objetivo indicar um valor que represente melhor um conjunto de dados. A média tem um comportamento fortemente influenciado por valores discrepantes. Entretanto, esta discrepância pode representar o número de relações que existe na maioria dos sítios, e neste caso, o valor da média estaria próximo deste valor.

Os percentuais de suporte do Trust-SLAbS assumem muitos valores distintos. Para representar a distribuição dos valores dos percentuais de suporte foi conveniente agrupá-los em intervalos de classes de percentual. Os valores existente entre 0 e 0,01 ficam agrupados na classe de 0,01 (1%) permitindo a utilização da média populacional para quantitativos discretos conforme a equação (15) usada na estatística.

$$MédiaRelações = \frac{1}{TotSítios} \sum_{i=0}^{NoMaxRelações} NumRelações_i . FreqSítios_i \quad (15)$$

Onde $NumRelações_i$ é o número de relações existentes entre dois sítios (número de trocas no Trust-BMSR e percentual de suporte no Trust-SLAbS) e $FreqSítios_i$ é a freqüência dos sítios que tem este número de relações e TotSítios é a soma de todos os sítios que tem estas relações.

O cálculo do valor de confiança média é calculado conforme a equação (16).

$$ValorConfiançaMédia(S_{Ori}, S_{Dest}) = 1 - \frac{RES(S_{Ori}, S_{Dest})}{MediaRelações} \quad (16)$$

Onde $RES(S_{Ori}, S_{Dest})$ é o número de relações existentes entre o sítio de origem e o sítio de destino (total de trocas para o Trust-BMSR e percentual de suporte para o Trust-SLAbS).

Os valores de $ValorConfiançaMédia$ ficam negativos quando $RES(S_{Ori}, S_{Dest})$ é maior que $MediaRelações$ neste caso estes valores são zerados para evitar erros no cálculo do *Pagerank* e *Indegree.*

3.1.3 –Escala de Probabilidade de Valores

Esta escala foi criada baseada nos conceitos de probabilidade. A composição para os valores da escala parte da identificação da probabilidade de um par de sítios possuir *k* ou mais relacionamentos em um grafo que representa a Web. Neste caso o espaço amostral (Ω) foi composto por pares de sítios que tinham uma ou mais relações e cada ponto amostral é da forma (*N*, *R*), sendo *N* o numero de relações entre sítios e *R* a quantidade de pares de sítios que possuem N relações. A partir desta definição foram realizados os cálculos de probabilidade para as relações existentes no grafo.

Para calcular a probabilidade de um par de sítios de ter *k* ou mais relações utilizou-se a equação (17):

$$P(k) = \frac{1}{\sum_{i=1}^{n} R_i} \sum_{j=k}^{n} R_j \quad (17)$$

Onde R_x é o número de relações que fazem *x* trocas.

Por exemplo: Seja o espaço amostral do número de relações existente entre os sítios representado por $\Omega = \{(1,30),(2,20),(3,15),(4,10)\}$. Para calcular o total de relações soma-se 30+20+15+10. Portanto, o total de relações é igual a 75. A probabilidade de um par de sítios efetuar 2 ou mais trocas será:

$$P(2) = \frac{20 + 15 + 10}{30 + 20 + 15 + 10} = \frac{45}{75} = \frac{9}{15}$$

Analisando a Equação (17), observa-se que há uma probabilidade maior de haver pares de sítios com um número menor de relações. Os valores de confiança baseados na probabilidade são calculados conforme a equação (18):

$$ValorConfiançaProbabilidade(S_{Ori}S_{Dest}) = P(RES(S_{Ori}, S_{Dest})) \quad (18)$$

Os valores do $ValorConfiançaProbabilidade$ serão usados para o cálculo do *Pagerank* e *Indegree* nos métodos do Trust-BMSR e Trust-SLAbS.

3.1.4 – Escala de Entropia de Valores

Na teoria da informação, entropia é uma medida de quantidade de informação que a ocorrência de um símbolo de um alfabeto carrega em uma fonte de dados (SHANNON 1948). Símbolos que ocorrem com maior probabilidade na fonte têm entropia menor e símbolos que ocorrem com menor probabilidade tem entropia maior. A escala de entropia adotada aqui toma o número de relações entre sítios como símbolos cujas freqüência é dada pelo número de pares existentes com tal número de relações. Com esta representação pode-se então calcular a Entropia de um dado número de relações conforme a equação (19).

$$Entropia(i) = \log_2\left(\frac{1}{P_i}\right) \quad (19)$$

Onde *i* é o número de relações existentes entre dois sítios e P_i é a probabilidade de um par de sítios ter *i* relações (trocas para o Trust-BMSR ou suporte para Trust-SLAbS).

Para o cálculo dos valores de confiança baseados na entropia foi encontrado o valor de maior entropia no grafo (20) e este valor foi usado para obtenção da escala.

$$EntropiaMáxima = \max_{0 \geq i \geq n} (Entropia(i)) \quad (20)$$

Onde *n* é o maior valor encontrado nas relações, ou seja, a maior troca no Trust-BMSR ou a maior percentual de suporte no Trust-SLAbS, existente no grafo da Web.

O cálculo dos valores da escala de confiança retira da informação existente o valor de entropia de cada relação, conforme abaixo:

$$ValorConfiançaEntropia(S_{Ori}, S_{Dest}) = 1 - \frac{Entropia(NumRelações(S_{Ori}, S_{Dest}))}{EntropiaMáxima} \quad (21)$$

A idéia com esta modelagem é fazer com que relações que normalmente ocorram sejam consideradas confiáveis por "serem eventos que carregam pouca informação" e as relações mais raras, logo mais atípicas, sejam consideras menos confiáveis.

3.2 – Cálculo da troca de sítios para o Trust-BMSR

Para calcular os valores do Trust-BMSR é necessário percorrer todo o grafo da Web e identificar os relacionamentos de reforço mútuo, somando as trocas existentes entre os sítios de origem (*So)* e destino (*Sd)* . Isto pode ser observado no Algoritmo 2.

```
1.  Faça V o grafo que representa a Web
2.  Faça trocasBMSR[So,Sd]) denotar a quantidade de trocas de apontadores entre os
    sítios So e Sd.
3.  Faça S o conjunto de todos os sítios que existem em V
4.  Para cada Sítio So ∈ S faça
5.      Para cada Sítio Sd ∈ S  faça
6.          trocasBMSR[So,Sd]=0;
7.  Para cada página p ∈ V, p residindo nos sítios S
8.      So=sitio(p);
9.      Para cada página q ∈ Out( p ), q do sítio Sd ≠ So
10.         Se p ∈ Out( q ) então
11.             Sd=sitio(q);
12.             trocasBMSR[So,Sd])= trocasBMSR[So,Sd]+1;
```

Algoritmo 2 - Cálculo das trocas entre sítios BMSR

O método sítio(p) tem a finalidade de identificar a que sítio pertence a página p e *Out(p)* representa os apontadores que saem da página *p*.

Com os valores das *trocasBMSR* definidos pode-se utilizar esta base para calcular o valor proporcional, média, probabilidade e entropia das trocasBMSR.

3.3 – Cálculo do percentual de suporte entre sítios para o Trust-SLAbS

O Algoritmo 3 apresenta a forma para calcular o percentual do suporte entre sítios:

1. *trocasBMSR[So,Sd]*=0;
2. **Para** cada página $p \in V$, p residindo nos sítios S
3. Sd=sitio(p);
4. **Para** cada página $q \in In(p)$, q do sítio $Sd \neq So$
5. So=sitio(q);
6. totInLink[Sd,So]= totInLink[Sd,So]+1;
7. totInLinkSitio[Sd]= totInLinkSitio[Sd]+1;
8. **Para** cada Sítio $Sd \in S$ **faça**
9. **Para** cada Sítio $So \in$ In(Sd) **faça**
10. suporteSLAbS[*Sd,So]*= $\frac{\text{totInLink [Sd,So]}}{\text{totInLinkSitio [Sd]}}$;

Algoritmo 3 – Calculo do percentual de suporte entre sítios do método Trust-SLAbS

O método *sítio(p)* retorna o sítio ao qual a página *p* pertence, *In(p)* retorna os apontadores que a página *p* recebe e *In(Sd)* retorna os sítios que apontam para *Sd*.

3.4 – Cálculo da escala de confiança para os métodos

Com os valores das relações entre sítios dos métodos Trust-BMSR e Trust-SLAbS calculados pode-se obter a escala de confiança, de acordo com as formas apresentadas na seção 3.3.

Para simplificar a descrição dos algoritmos será adotado o termo totalDeRelações(So,Sd) tanto para representar o total das trocas existentes entre os sítios So e Sd (Trust-BMSR) como o percentual de suporte existente entre os sítios Sd e So (Trust-SLAbS).

3.5 – Adaptação e cálculo do *Indegree* e *Pagerank* para uso de escala de confiança

Para efetuar o cálculo dos métodos de análise de apontadores *Indegree* e *Pagerank* foi feita uma adaptação, onde as escalas de confianças calculadas são utilizadas como pesos nos apontadores recebidos de cada uma das páginas do grafo. Os apontadores recebidos em uma página têm seus pesos reduzidos de acordo com o valor de confiança calculado entre as relações existentes nos sítios das páginas de origem e de destino.

3.5.1 – Cálculo e adaptação do *Indegree*

Neste trabalho foi feita uma adaptação no cálculo do *Indegree*, onde o valor do *Indegree* representa a soma dos valores de confiança dos apontadores para uma determinada página. A equação (22) apresenta a adaptação para o cálculo do *Indegree*.

$$Indegree(p) = \sum_{\forall\, l \in In(p)} ValorDeConfiança[l,p] \tag{22}$$

Onde $ValorDeConfiança(l,p)$ é o valor de confiança entre o sitio da página l e o sítio da página p que foi calculado nos métodos anteriormente apresentados e $In(p)$ é o conjunto de apontadores para p.

Esta adaptação feita no algoritmo *Indegree* pode ser vista no Algoritmo 4.

1. **Faça** *V* denotar o grafo da Web
2. **Faça** *In(p)* o conjunto de páginas que apontam para *p*
3. **Para** cada página $p \in V$ **faça**
4. **Para** cada página $q \in In(p)$ **faça**
5. $Indegree[p] = Indegree[p] + ValorDeConfiança[sítio(q), sítio(p)]$

Algoritmo 4 – Adaptação do algoritmo do *Indegree*

3.5.2 – Cálculo e adaptação do *Pagerank*

Para utilizar a escala de confiança para o cálculo do *Pagerank* foi necessário efetuar uma adaptação. Os apontadores recebidos em uma determinada página *p* que faz parte de um sítio *s* têm seu peso proporcional ao valor de confiança existente nas relações entre o sítio de origem S_o e o sítio de destino S_d. Desta forma, o valor do *Pagerank* será menor para uma determinada página quanto menor for o valor de confiança dos apontadores que chegam até ela (23).

$$PR(p) = (1 - c). \sum_{q \in In(p)} \frac{PR(q).ValorDeConfiança[sítio(q), sítio(p)]}{Out(q)} + \frac{c}{\|V\|} \quad (23)$$

Onde *V* representa o grafo da Web, *c* é a probabilidade do usuário escolher uma página de *V* ao invés de seguir os apontadores da página corrente, $sítio(q)$ retorna o sítio da página q e $ValorDeConfiança[sítio(q), sítio(p)]$ é o valor de confiança gerado pelos algoritmos de análise de confiança apresentados nesta dissertação.

O valor removido do *Pagerank* com a aplicação do *ValorDeConfiança* é acumulado para que na próxima iteração do cálculo, este seja distribuído uniformemente em todas as páginas do grafo da Web. Este recurso foi necessário para manter a convergência da cadeia de

Markov que está associada ao grafo da Web, ou seja, isto garante que a soma das probabilidades de transição de cada estado da cadeia continue sendo um. Esta adaptação pode ser verificada no Algoritmo 5:

1. **Faça** $PR(i) = \frac{1}{\|V\|}, \forall\, i \in \{1, 2, 3, \ldots, \|V\|\}$
2. **Repita** até convergir
3. **Para** cada página $p \in V$
4. $PR(p) = (1 - c).\sum_{q \in In(p)} \frac{PR(q).ValorDeConfian\ \ ça[sítio(q), sítio(p)]}{Out(q)} + \frac{c}{\|V\|}$
5. $Resíduo = Resíduo +$
6. $(1 - c).\sum_{q \in In(p)} \frac{PR(q).(1 - ValorDeConfian\ \ ça[sítio(q), sítio(p)])}{Out(q)} + \frac{c}{\|V\|}$
7. **Para** cada página $p \in V$
8. $PR(p) = PR(p) + \frac{Resíduo}{\|V\|}$

Algoritmo 5 – Adaptação do algoritmo do *Pagerank* para os métodos propostos

3.6– Combinação dos Métodos

Nesta dissertação, cada método possui uma escala de confiança que atribui valores para os relacionamentos entre os sítios. Observou-se que na maioria dos casos os três métodos atribuem valores de distintos de confiança para o relacionamento entre um mesmo par de sítios. Buscando-se em extrair os valores mais confiáveis para combinar os métodos, foram propostas e estudadas as seguintes formas de combinação:

- Menor confiança: Nesse caso optou-se por usar o menor valor de confiança para penalizar as relações entre os sítios. Neste caso considera-se que um único método indique que o relacionamento é suspeito, então deve-se confiar pouco no relacionamento.

- Maior confiança: Funciona de maneira oposta a anterior. Neste caso optou-se por usar o maior de valor confiança atribuindo-se credibilidade ao valor de confiança gerado pelo método, que indica que os relacionamentos entre estes sítios são confiáveis.
- Média da confiança: Neste caso optou-se por utilizar a média dos valores de confiança. Assim, o valor final fica sendo uma mistura da confiança obtida por cada método.
- Ou Probabilístico da Confiança: Neste caso optou-se em combinar os valores de confiança dos métodos utilizando a fórmula do operador OU, onde valores de confiança são nomeados como valores de probabilidade e a combinação é dada por:

$$TrustValue = 1 - \big((1 - TrustA).(1 - TrustB).(1 - TrustB)\big) \quad (24)$$

Onde *TrustA*, *TrustB* e *TrustC* representa os valores de confiança dos métodos propostos.

Capitulo 4

4 - Experimentos

Os métodos propostos foram avaliados utilizando-se a coleção WBR03 e os algoritmos *Pagerank* e *Indegree*. Foram realizados também experimentos com os algoritmos propostos por (CARVALHO, et al. 2006) com o intuito de medir a eficácia dos métodos aqui propostos.

4.1 – Ambiente Experimental

Os experimentos foram realizados utilizando-se a coleção WBR03. A coleção WBR03 é uma base de dados da máquina de busca real TodoBR[2] composta por 12.020.513 páginas extraídas da *Web* brasileira, estas páginas são conectadas por 130.717.004 apontadores válidos, 1.001.070 host, 141.284 domínios e o tamanho médio das páginas é de 5Kb. O número de apontadores mostra que existe um conjunto de páginas altamente conectado, que fornecem muita informação para os métodos aqui propostos. Um fator decisivo para escolha desta coleção é que ela representa uma parte significativa da *Web* brasileira, que é tão diversa quanto a *Web* mundial.

Os experimentos foram realizados com as consultas navegacionais extraídas do *log* da maquina de busca TodoBR, composto por 11.246.351 consultas submetidas a esta máquina. Estas consultas foram divididas em dois tipos:

[2] TodoBR é uma marca registrada de Akwan Information Tecnologies, que foi adquirida pela Google em julho de 2005.

- Consultas Populares: Foram escolhidas as consultas navegacionais mais freqüentes no *log* da máquina de busca TodoBR. Com estas consultas foi possível observar o impacto dos métodos aqui propostos na remoção de apontadores ruidosos.
- Consultas Aleatórias: Foram selecionadas consultas navegacionais aleatoriamente no *log* da máquina de busca. Observou-se o comportamento dos métodos nestas consultas, verificando o impacto causado em páginas não populares.

Foram avaliadas 60 consultas de cada tipo com a métrica do MRR (utilizada pela TREC para consultas navegacionais). Efetuando-se a comparação entre os métodos aqui propostos e os métodos de (CARVALHO, et al. 2006) e também os algoritmos *Pagerank* e *Indegree*. Os resultados obtidos com o MRR foram submetidos ao teste-T de *student* com uma confiança de 95%.

As consultas informacionais não foram incluídas nos experimentos por sofrer pouco impacto com alterações nas estruturas dos apontadores (SILVA, et al. 2008).

Os experimentos foram realizados em dois cenários distintos de processamento de consulta: sem combinação e com combinação. No primeiro caso, avaliou-se a qualidade dos métodos implementados sem combinação com outras evidências, utilizou-se somente os valores de reputação calculados pelos algoritmos de análise de apontadores para ordenar as páginas que possuíam os termos da consulta (nesta dissertação será chamado de "sem combinação"). No segundo caso, a reputação das páginas foi combinada com os valores de similaridade calculados pelo modelo vetorial sobre o conteúdo das páginas e o texto de âncora. A combinação destas evidências (reputação, texto de âncora e conteúdo textual) foi

realizada utilizando-se a formula de *ranking* proposta em (CALADO, et al. 2003), conforme a equação (25):

$$sim(q,d) = 1 - \Big(\big(1 - sim_{texto}\,(q,d)\big) * \big(1 - sim_{anc\,hor}(q,d)\big) * \big(1 - reputação(d)\big)\Big) \quad (25)$$

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 contexto textual do documento d utilizando o modelo vetorial, $sim_{anc\,hor}(q,d)$ é a similaridade entre a consulta q e o texto de âncora do documento d utilizando o modelo vetorial e $reputação(d)$ é o valor de reputação estimado pelo algoritmo de análise de apontadores o documento d. Esta combinação foi chamada nesta dissertação de BNC (Bayesian Network Combination).

Finalmente, foram estudados dois cenários de análise de apontadores utilizando o algoritmo *Pagerank* e outro utilizando o algoritmo *Indegree*. Para resultados identificados nas tabelas, o termo "Com Pesos Iguais" indica que a máquina de busca para estes experimentos utilizou o valor calculado pelos algoritmos *Indegree* e *Pagerank* aplicando o peso de valor um para todos os relacionamentos, aceitando integralmente o valor da reputação calculada por eles.

4.1.1 – Critério de agrupamento no grafo da *Web*

Para o teste dos métodos durante a realização dos experimentos foi necessário agrupar as páginas em sítios. Com o objetivo de identificar o impacto dos métodos em grupos com poucas páginas e em grupos com um número maior de páginas, foram utilizados dois critérios de agrupamento no grafo da *Web*. São eles:

- Agrupamento por host: cada grupo é composto por páginas que pertencem ao mesmo *host*.
- Agrupamento por domínios: cada grupo é composto por paginas que pertencem ao mesmo domínio.

Para identificar o nome do *host* e o nome do domínio de cada página foram processados os caracteres que compõem sua URL, seguindo o mesmo critério proposto por (BERLT, et al. 2007). Para definir qual o nome *host* da página foram desenvolvidos os seguintes passos:

1. Remove-se os prefixos “http:// “, caso existam;
2. Remove-se os prefixos www. , caso existam;
3. Remove-se os caracteres após a primeira “/”, caso existam;
4. Remove-se as informações da porta, caso existam.
5. A cadeia de caracteres resultante é o nome do *host* da página.

Como exemplo pode-se citar a URL ” http://www.cefetam.edu.br:8080/Audionews “ tem como nome de host “cefetam.edu.br “ , pois foram removidos ” http:// “, www.” , “:8080” e “/Audionews” .

O nome do domínio de uma página é composto pela concatenação de três elementos do nome do *host,* são eles:

- Identificador do país: indica o país de origem da página. Ex: “il” refere-se a paginas de Israel, “br” refere-se ao Brasil.
- Categoria do servidor: indica a categoria a que a pagina se encaixa. Ex: “.com” para páginas comerciais, “.edu” para instituições de ensino.

- Nome do domínio: indica o nome do domínio, que está localizado antes da categoria do servidor ou do identificador do pais caso, caso o primeiro seja nulo. Ex: “cefetam”, “gmail”

Por exemplo, na URL “http://noticias.yahoo.com.br/s/11032009” o nome do host é noticias.yahoo.com.br e o nome do domínio é yahoo.com.br.

É importante observar que o host é composto por uma ou mais páginas e que um domínio é composto por um ou mais *hosts*.

4.2 – Resultados

Os resultados aqui apresentados estão divididos em quatro partes distintas. Na primeira são apresentados os resultados dos experimentos realizados com os métodos Trust-BMSR e Trust-SLAbS com os dois conjuntos de consultas (navegacionais populares e navegacionais aleatórias) comparando-os com os resultados obtidos com um sistema de busca quando estes métodos foram aplicados conforme descritos no capítulo 3. Os resultados são comparados com os obtidos pelo mesmo sistema sem aplicação dos métodos. Na segunda parte são apresentados os resultados dos experimentos com os métodos combinados. Na terceira parte serão apresentados os melhores resultados dos métodos efetuando comparações com a métodos proposto por Carvalho et al.(2006). Os experimentos realizados até este ponto foram feitos com agrupamento de *host,* isto foi necessário para manter o mesmo grafo dos experimentos realizados por Carvalho *et al.* Na quarta parte são apresentados os melhores resultados destes métodos quando agrupado por domínios.

4.2.1 – Resultado com os métodos individuais

4.2.1.1 – Resultados com o Trust-BMSR

As Tabelas 1 e 2 apresentam os valores de MRR obtidos com as consultas navegacionais populares nos cenários onde o sistema de busca utiliza o algoritmo *Pagerank* e o algoritmo *Indegree* respectivamente. Na Tabela 1 pode-se observar que a estratégia da média obteve melhores resultados com ganho de 6,17% para o cenário sem combinação e 3,33% para o cenário combinado (BNC) quando comparados os resultados quando um sistema de busca utiliza o *Pagerank*.

Esquema de Pesos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,230478	-	-	0,417476	-	-
Razão Max Confiança	0,231025	0,24%	0,1563	0,418342	0,21%	0,1606
Média Confiança	0,244704	6,17%	0,0409	0,431375	3,33%	0,0632
Probabilidade da Conf	0,241906	4,96%	0,0891	0,423720	1,50%	0,0562
Entropia Confiança	0,241502	4,78%	0,0966	0,423905	1,54%	0,0508

Tabela 1 : Valores de MRR para consultas navegacionais populares com Trust-BMSR aplicando as quatro métricas de confiança estudadas quando um sistema de busca utiliza o Pagerank

Quando o método foi comparado com o *Indegree* observou-se um ganho constante. Ao verificar os valores dos relacionamentos calculados entre os sítios nas escalas calculadas, muitas vezes as diferenças entre elas eram de ordem de centésimo e milésimo. Fazendo com que o cálculo do *Indegree* ao efetuar a soma dos valores dos apontadores para a página alvo não produzissem ganhos diferentes.

Tipos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,402264	-	-	0,535635	-	-
Razão Max Confiança	0,402264	0,00%	Não	0,535635	0,00%	Não
Média Confiança	0,419839	4,37%	0,0020	0,531344	-0,80%	0,3313
Probabilidade da Conf	0,419839	4,37%	0,0020	0,531344	-0,80%	0,3313
Entropia Confiança	0,419839	4,37%	0,0020	0,531344	-0,80%	0,3313

Tabela 2 : Valores de MRR para consultas navegacionais populares com Trust-BMSR aplicando as quatro métricas de confiança estudadas quando um sistema de busca utiliza o Indegree

As Tabelas 3 e 4 apresentam os resultados de MRR para consultas navegacionais aleatórias. Observa-se que para estas consultas, os resultados não apresentaram ganhos relevantes. Os valores do MRR estão muito próximos com uma variação de 1% positivo ou negativo que o torna não recomendado para ser utilizado invidualmente em uma máquina de busca.

Esquema de Pesos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,241811	-	-	0,507483	-	-
Razão Max Confiança	0,242439	0,26%	0,0867	0,507483	0,00%	Nao
Média Confiança	0,239411	-0,99%	0,0172	0,501927	-1,09%	0,2993
Probabilidade da Conf	0,239541	-0,94%	0,2526	0,499149	-1,64%	0,2357
Entropia Confiança	0,242841	0,43%	0,2767	0,501927	-1,09%	0,2658

Tabela 3: Valores de MRR para consultas navegacionais aleatórias com Trust-BMSR aplicando as quatro métricas de confiança estudadas quando um sistema de busca utiliza o Pagerank

A Tabela 4 apresenta os resultados quando um sistema da busca utiliza o algoritmo *Indegree*. Os quais também mantiveram um valor constante de ganho para todos os métodos, de forma semelhante aos experimentos realizados nas consultas navegacionais populares.

Esquema de Pesos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,343364	-	-	0,547397	-	-
Razão Max Confiança	0,343364	0,00%	Não	0,547397	0,00%	Não
Média Confiança	0,358359	4,37%	0,0031	0,545036	-0,43%	0,2025
Probabilidade da Conf	0,358359	4,37%	0,0031	0,545036	-0,43%	0,2025
Entropia Confiança	0,358359	4,37%	0,0031	0,545036	-0,43%	0,2025

Tabela 4: Valores de MRR para consultas navegacionais aleatórias com Trust-BMSR aplicando as quatro métricas de confiança estudadas quando um sistema de busca utiliza o Indegree

A métrica da média obteve melhores resultados para consultas populares nos cenários sem combinação e BNC. Por isto selecionou-se esta métrica para ser utilizada na combinação com o Trust-SLAbS e SLLA apresentados a seguir.

4.2.1.2 – Resultados com o Trust-SLAbS

As Tabelas 5 e 6 apresentam os resultados das consultas navegacionais populares utilizando o método Trust-SLAbS nos cenários onde o sistema de busca utiliza o algoritmo *Pagerank* e o algoritmo *Indegree* respectivamente. Observa-se que os ganhos apresentados na Tabela 5 são significantes com 13,79% para a média confiança no cenário sem combinação e 9,86 % para a confiança baseada na entropia no cenário com combinação.

Esquema de Pesos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,230478	-	-	0,417476	-	-
Razão Max Confiança	0,244366	6,03%	0,0492	0,420932	0,83%	0,4416
Média Confiança	0,262257	13,79%	0,0001	0,433603	3,86%	0,0004
Probabilidade da Conf	0,243744	5,76%	0,4273	0,429047	2,77%	0,0144
Entropia Confiança	0,254625	10,48%	0,1092	0,458646	9,86%	0,0106

Tabela 5: Valores de MRR para consultas navegacionais populares com Trust-SLAbS aplicando as quatro métricas de confiança estudadas quando um sistema de busca utiliza o Pagerank

Os experimentos feitos com o Indegree são apresentados na Tabela 6. Obtendo 2,33% de ganho para média confiança no cenário individual e 0,88% para confiança baseado na entropia com os valores de significância estatística baixos, havendo perdas em alguns casos.

Esquema de Pesos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,402264	-	-	0,535635	-	-
Razão Max Confiança	0,402264	0,00%	0,1606	0,535635	0,00%	0,1606
Média Confiança	0,411621	2,33%	0,3100	0,533711	-0,36%	0,4437
Probabilidade da Conf	0,396889	-1,34%	0,3281	0,540126	0,84%	0,3075
Entropia Confiança	0,403353	0,27%	0,4678	0,540357	0,88%	0,3144

Tabela 6:Valores de MRR para consultas navegacionais populares com Trust-SLAbS aplicando as quatro métricas de confiança estudadas quando um sistema de busca utiliza o Indegree

As Tabelas 7 e 8 apresentam os resultados dos experimentos realizados com as consultas aleatórias. Observa-se na Tabela 7 que a maioria dos resultados não apresentou ganho quando comparado com sistemas de busca que utilizam o algoritmo *Pagerank*. O

método baseado na probabilidade apresentou ganho de 1,59% com o valor de significância muito baixo.

Esquema de Pesos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,241811	-	-	0,507483	-	-
Razão Max Confiança	0,240509	-0,54%	0,0104	0,505731	-0,35%	0,3061
Média Confiança	0,232431	-3,88%	0,1729	0,504897	-0,51%	0,2993
Probabilidade da Conf	0,241786	-0,01%	0,0119	0,515538	1,59%	0,1733
Entropia Confiança	0,226252	-6,43%	0,0664	0,498905	-1,69%	0,3125

Tabela 7:Valores de MRR para consultas navegacionais aletórias do Trust-SLAbS aplicando as quatro métricas de confiança estudadas quando um sistema de busca utiliza o Pagerank

Os resultados das consultas aleatórias quando comparadas com o *Indegree* podem ser visualizados na Tabela 8. Para estas consultas este método apresentou 5,34% para média confiança no cenário sem combinação e 2.11% para confiança baseado na entropia para o cenário BNC. As consultas aleatórias podem ser compostas por páginas que não sejam populares e com isso terem poucos apontadores para estas páginas fazendo com que os valores do suporte sejam maiores, mas insuficientes para causar impacto no cálculo do *Indegree*.

Esquema de Pesos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Esquema de Pesos	0,343364	-	-	0,547397	-	-
Razão Max Confiança	0,343364	0,00%	0,1606	0,547397	0,00%	Não
Média Confiança	0,361715	5,34%	0,1692	0,533415	-2,55%	0,2455
Probabilidade da Conf	0,358322	4,36%	0,1660	0,552304	0,90%	0,3705
Entropia Confiança	0,361276	5,22%	0,1330	0,558971	2,11%	0,1792

Tabela 8:Valores de MRR para consultas navegacionais aletórias no Trust-SLAbS aplicando as quatro métricas de confiança estudadas quando um sistema de busca utiliza o Indegree

A métrica da média também obteve os melhores resultados para consultas populares nos cenários sem combinação e ganho menor no cenário BNC quando comparado com a métrica de confiança baseado em entropia. Entretanto os resultados com o *Indegree,* tanto para consultas populares como aleatórias, foram melhores com a média confiança. Por este

motivo, esta métrica também foi selecionada para ser utilizada na combinação com os Trust-BMSR e SLLA.

4.2.2 – Resultados com os métodos combinados

4.2.2.1 – Resultados com o método SLLA

Para combinar os métodos também foi utilizado o método SLLA proposto por (CARVALHO, et al. 2006), neste trabalho o SLLA obteve o maior ganho individual. A Tabela 9 apresenta os resultados do MRR com as consultas navegacionais populares e aleatórias utilizadas nos experimentos desta dissertação.

Tipos	Sem Combinação				BNC			
	Pagerank	SLLA	Ganho(%)	Signif.	Pagerank	SLLA	Ganho(%)	Signif.
Populares	0,230478	0,280734	21,81%	0,01621	0,417476	0,522667	25,20%	0,00008
Aleatorias	0,241811	0,280004	15,79%	0,04082	0,507483	0,569116	12,14%	0,00599

Tabela 9: Valores do MRR para consultas navegacionais populares e aleatórias aplicando o método SLLA estudado quando um sistema de busca utiliza o Pagerank

Os valores no cenário sem combinação comportaram-se de forma semelhante aos valores de experimentos já realizados anteriormente (CARVALHO, et al. 2006). Avaliou-se também os resultados dos experimentos no cenário BNC onde os ganhos foram significantes, mostrando que este método é uma boa opção para ser adicionado em máquinas de busca reais.

A Tabela 10 apresenta os valores de MRR para as consultas navegacionais populares e aleatórias comparadas com o *Indegree*. Observa-se ganhos com baixa significância dos resultados. Apesar de Carvalho *et al.* não terem realizados experimentos comparando-os com o *Indegree*, optou-se em usá-lo para estudar o comportamento do método quando combinado com os melhores resultados dos métodos propostos nesta dissertação.

Tipos	Sem Combinação				BNC			
	Indegree	MRR	Ganho(%)	Signif.	Indegree	MRR	Ganho(%)	Signif.
Populares	0,402264	0,402416	0,04%	0,1606	0,535635	0,525933	0,06%	0,1606
Aleatorias	0,343364	0,343515	0,04%	0,1606	0,547397	0,547397	0,00%	Não

Tabela 10: Valores do MRR para consultas navegacionais populares e aleatórias aplicando o método SLLA estudado quando um sistema de busca utiliza o Indegree

4.2.2.2– Resultados combinados dos métodos *Trust*-BMSR e *Trust*-SLAbS

As Tabelas 11 e 12 apresentam os resultados dos experimentos da combinação do *Trust*-BMSR com o *Trust*-SLAbS para as consultas navegacionais populares nos cenários onde o sistema de busca utiliza o algoritmo *Pagerank* e o algoritmo *Indegree* respectivamente. Os resultados mostram ganhos de 12,34% quando combinados com o ou probabilístico para o cenário sem combinação, e o cenário BNC com ganho de 9,27% usando a menor confiança quando comparados com o *Pagerank.*

Forma de Combinar os Métodos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos iguais	0,230478	-	-	0,417476	-	-
Menor Confiança	0,256275	11,19%	0,09492	0,456125	9,27%	0,01564
Maior Confiança	0,258275	12,06%	0,07132	0,429933	2,98%	0,21316
Media Confiança	0,255102	10,68%	0,00001	0,453897	8,72%	0,01730
Ou Probabilistico	0,258924	12,34%	0,06700	0,439119	5,18%	0,04800

Tabela 11: Valores de MRR para consultas navegacionais populares dos métodos Trust-BMSR e Trust-SLAbS aplicando as combinações estudadas quando um sistema de busca utiliza o algoritmo Pagerank

O resultado da combinação dos métodos obteve ganhos de 9,27% os quais são superiores aos obtidos individualmente para os métodos Trust-SLAbS e Trust-BMSR (Tabelas 1 e 5). Apesar de penalizar as relações com o uso da menor confiança entre os métodos, o resultado da combinação melhorou em relação aos resultados individuais.

A Tabela 12 apresenta os resultados quando um sistema de busca utiliza o algoritmo *Indegree*. Observa-se que não existem ganhos na maioria das combinações realizadas com estes dois métodos. Os valores de MRR ficaram inferiores aos resultados dos métodos quando usados individualmente.

Forma de Combinar os Métodos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,402264	-	-	0,535635	-	-
Menor Confiança	0,401327	-0,23%	0,4812	0,532877	-0,51%	0,4212
Maior Confiança	0,402416	0,04%	0,4521	0,535635	0,00%	Não
Media Confiança	0,410705	2,10%	0,3286	0,533711	-0,36%	0,4437
Ou Probabilistico	0,401327	-0,23%	0,4812	0,532877	-0,51%	0,4212

Tabela 12: Valores de MRR para consultas navegacionais populares dos métodos combinados Trust-BMSR e Trust-SLAbS aplicando as combinações estudadas quando um sistema de busca utiliza o algoritmo Indegree

Os experimentos para as consultas navegacionais aleatórias nesta combinação não apresentaram resultados significativos quando comparado com os resultados de um sistema de busca utilizando o algoritmo *Pagerank.* Como pode ser visto na Tabela 13, o resultado das consultas com esta combinação não melhorou os resultados individuais como ocorreu nas consultas populares.

Formas de Combinar os Métodos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,241811	-	-	0,507483	-	-
Menor Confiança	0,228997	-5,30%	0,1088	0,498786	-1,71%	0,3112
Maior Confiança	0,225754	-6,64%	0,0571	0,496108	-2,24%	0,2302
Media Confiança	0,227957	-5,73%	0,0895	0,497953	-1,88%	0,2906
Ou Probabilistico	0,230040	-4,87%	0,1190	0,496961	-2,07%	0,2470

Tabela 13: Valores de MRR para consultas navegacionais aleatórias dos métodos combinados Trust-BMSR e Trust-SLAbS aplicando as combinações estudadas quando um sistema de busca utiliza o algoritmo Pagerank

Na Tabela 14 são apresentados os resultados de um sistema de busca que utiliza o algoritmo *Indegree* cujos ganhos obtidos não foram significantes.

Forma de Combinar os Métodos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,343364	-	-	0,547397	-	-
Menor Confiança	0,347219	1,12%	0,4193	0,532523	-2,72%	0,1826
Maior Confiança	0,347219	1,12%	0,4193	0,532523	-2,72%	0,1826
Media Confiança	0,355574	3,56%	0,2585	0,541563	-1,07%	0,3762
Ou Probabilístico	0,347219	1,12%	0,4193	0,532523	-2,72%	0,1826

Tabela 14: Valores de MRR para consultas navegacionais aleatórias dos métodos combinados Trust-BMSR e Trust-SLAbS aplicando as combinações estudadas quando um sistema de busca utiliza o algoritmo Indegree

Nas consultas navegacionais populares, os melhores resultados foram apresentados pela combinação dos métodos com o maior valor de confiança no cenário sem combinação, e o menor valor de confiança no cenário BNC quando o sistema de busca utiliza o algoritmo do *Pagerank*. Esta combinação melhorou os resultados dos métodos quando comparados aos utilizados individualmente em um sistema de busca. Não houve melhoria nos resultados referentes às consultas navegacionais aleatórias. Semelhante ao que aconteceu quando da utilização do algoritmo *Indegree,* em que a maioria dos resultados apresentaram baixa significância estatística.

A melhor formar de efetuar a combinação destes métodos foi a seleção do maior valor que será utilizada para ser comparadas com os métodos propostos por (CARVALHO, et al. 2006).

4.2.2.3 – Resultados combinados dos métodos SLLA e Trust-BMSR

As Tabelas 15 e 16 apresentam os resultados referentes a combinação do método SLLA com o Trust-BMSR para as consultas navegacionais populares, nos cenários onde o sistema de busca utiliza o algoritmo *Pagerank* e o algoritmo *Indegree* respectivamente. A Tabela 15 demonstra que a combinação apresentou bons resultados que podem ser observados com o uso da maior confiança entre os métodos obtendo ganho de 32,77% para o cenário sem combinação e de 26,05% para o cenário BNC. Este resultado mostra que o Trust-BMSR

complementa o SLLA selecionando o relacionamento entre sítios mais confiável dentre os pares.

Forma de Combinar os Métodos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos iguais	0,230478	-	-	0,417476	-	-
Menor Confiança	0,280734	21,81%	0,0162	0,522667	25,20%	0,000078
Maior Confiança	0,306016	32,77%	0,0086	0,526245	26,05%	0,000047
Media Confiança	0,304463	32,10%	0,0048	0,516410	23,70%	0,000062
Ou Probabilistico	0,275928	19,72%	0,0260	0,522667	25,20%	0,000078

Tabela 15: Valores de MRR para consultas navegacionais populares da combinação do SLLA com Trust-BMSR aplicando as combinações estudadas quando um sistema de busca utiliza o algoritmo Pagerank

Na Tabela 16 são apresentados os valores de MRR quando um sistema de busca utiliza o algoritmo *Indegree*,os resultados foram significantes como é o caso da combinação com o menor valor de confiança com 4,41% de ganho.

Forma de Combinar os Métodos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,402264	-	-	0,535635	-	-
Menor Confiança	0,419991	4,41%	0,0108	0,531642	-0,75%	0,3149
Maior Confiança	0,402264	0,04%	0,3313	0,535635	0,00%	Não
Media Confiança	0,412975	2,66%	0,0165	0,524193	-2,14%	0,0980
Ou Probabilistico	0,402416	0,04%	0,1606	0,535933	0,06%	0,1609

Tabela 16: Valores de MRR para consultas navegacionais populares da combinação do SLLA com Trust-BMSR aplicando as combinações estudadas quando um sistema de busca utiliza o algoritmo Indegree

As Tabelas 17 e 18 apresentam os resultados dos experimentos para as consultas navegacionais aleatórias. Destaca-se a menor confiança dos métodos com ganho de 15,79% sem combinação e de 15,32% com BNC e com alta significância. Comparando com os resultados individuais dos métodos, observa-se que o uso do menor valor utiliza somente a confiança calculada no SLLA. No cenário BNC observa-se que o Trust-BMSR complementa o SLLA melhorando o resultado individual.

Forma de Combinar os Métodos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,241811	-	-	0,507483	-	-
Menor Confiança	0,280004	15,79%	0,0408	0,569116	12,14%	0,0059
Maior Confiança	0,255897	5,83%	0,3117	0,585234	15,32%	0,0081
Media Confiança	0,270032	11,67%	0,1123	0,577761	13,85%	0,0108
Ou Probabilístico	0,278756	15,28%	0,0458	0,569116	12,14%	0,0059

Tabela 17: Valores de MRR para consultas navegacionais aleatórias da combinação do SLLA com Trust-BMSR aplicando as combinações estudadas quando um sistema de busca utiliza o algoritmo Pagerank

Os resultados com o *Indegree* seguiram o comportamento das consultas populares apresentando ganhos pequenos e alguns casos significantes como pode ser vistos na Tabela 18.

Forma de Combinar os Métodos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,343364	-	-	0,547397	-	-
Menor Confiança	0,358511	4,41%	0,0230	0,545036	-0,43%	0,3956
Maior Confiança	0,343515	0,04%	0,3850	0,547397	0,00%	Não
Media Confiança	0,354420	3,22%	0,0164	0,547582	0,03%	0,1606
Ou Probabilístico	0,343515	0,04%	0,1606	0,547397	0,00%	Não

Tabela 18: Valores de MRR para consultas navegacionais aleatórias da combinação do SLLA com Trust-BMSR aplicando as combinações estudadas quando um sistema de busca utiliza o algoritmo Indegree

As melhores combinações foram o uso do maior valor de confiança para consultas populares e menor valor para consultas aleatórias. Estes resultados serão utilizados e apresentados comparando-os com os resultados dos métodos propostos por (CARVALHO, et al. 2006).

4.2.2.4 – Resultados combinados dos métodos SLLA e Trust-SLAbS

As Tabelas 19 e 20 apresentam os resultados dos experimentos para as consultas navegacionais populares nos cenários onde o sistema de busca utiliza o algoritmo *Pagerank* e o algoritmo *Indegree* respectivamente. Na Tabela 19 os resultados mostram ganhos significativos nesta combinação destacando o uso da maior confiança no cenário sem combinação com 30,92% e no cenário BNC com 26,52%. Analisando os valores individuais

dos métodos observa-se que Trust-SLAbS melhora o ganho em 41,68% em relação ao SLLA individualmente. Mostrando que utilizar o maior valor de confiança na relação é uma boa estratégia de combinação.

Forma de Combinar os Métodos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,230478	-	-	0,417476	-	-
Menor Confiança	0,289177	25,47%	0,1382	0,510253	22,22%	0,2718
Maior Confiança	0,301753	30,92%	0,0132	0,528176	26,52%	0,00002
Media Confiança	0,276422	19,93%	0,0169	0,488917	17,11%	0,00025
Ou Probabilístico	0,246573	6,98%	0,0011	0,426408	2,14%	0,2452

Tabela 19: Valores de MRR para consultas navegacionais populares da combinação do SLLA com Trust-SLAbS aplicando as combinações estudadas quando um sistema de busca utiliza o algoritmo Pagerank

Os valores do *Indegree* apresentaram ganhos pequenos e significantes para o cenário Sem Combinação como pode ser visto na Tabela 20, destacando-se o uso da média confiança com 6.79% de ganho. Observando os valores de MRR nos métodos individuais constata-se que o método do SLLA combinado com a média melhorou o resultado do Trust-SLAbS que tinha o ganho individual de 2.33%.

Forma de Combinar os Métodos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,402264	-	-	0,535635	-	-
Menor Confiança	0,426022	5,91%	0,1368	0,531159	-0,84%	0,3977
Maior Confiança	0,416834	3,62%	0,0025	0,538479	0,53%	0,4127
Media Confiança	0,429585	6,79%	0,0171	0,537480	0,34%	0,4457
Ou Probabilístico	0,416031	3,42%	0,0045	0,546812	2,09%	0,1245

Tabela 20: Valores de MRR para consultas navegacionais populares da combinação do SLLA com Trust-SLAbS aplicando as combinações estudadas quando um sistema de busca utiliza o algoritmo Indegree

As Tabelas 21 e 22 apresentam os resultados dos experimentos realizados para as consultas navegacionais aleatórias. Na Tabela 21 observa-se ganhos significantes destacando a combinação com a menor confiança dos métodos com ganhos de 13,11% para o cenário sem combinação e 13,40% para o cenário BNC quando comparados com os resultados de um sistema de busca que utiliza o algoritmo do *Pagerank*. Ao analisar os valores individuais do

SLLA detecta-se uma queda no ganho mostrando que a combinação não é indicada para ser utilizada em máquinas de busca.

Forma de Combinar os Métodos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,241811	-	-	0,507483	-	-
Menor Confiança	0,273504	13,11%	0,0040	0,575505	13,40%	0,000043
Maior Confiança	0,266521	10,22%	0,1163	0,561604	10,66%	0,0315
Media Confiança	0,240017	-0,74%	0,4455	0,539658	6,34%	0,0938
Ou Probabilístico	0,237382	-1,83%	0,0065	0,501961	-1,09%	0,3922

Tabela 21: Valores de MRR para consultas navegacionais aleatórias da combinação do SLLA com Trust-SLAbS aplicando as combinações estudadas quando um sistema de busca utiliza o algoritmo Pagerank

Os resultados dos experimentos para estas consultas quando comparados com o *Indegree* também obtiveram ganhos com baixa significância destacando-se a combinação da menor confiança conforme a Tabela 22. Observa-se que esta combinação com o SLLA melhorou o resultado dos métodos quando comparados individualmente.

Forma de Combinar os Métodos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,343364	-	-	0,547397	-	-
Menor Confiança	0,366235	6,66%	0,1280	0,536638	-1,97%	0,2939
Maior Confiança	0,347554	1,22%	0,3339	0,552258	0,89%	0,2972
Media Confiança	0,358829	4,50%	0,1667	0,558971	2,11%	0,1792
Ou Probabilístico	0,346654	0,96%	0,3609	0,552258	0,89%	0,2972

Tabela 22: Valores de MRR para consultas navegacionais aleatórias da combinação do SLLA com Trust-SLAbS aplicando as combinações estudadas quando um sistema de busca utiliza o algoritmo Indegree

A combinação destes métodos também apresentou bons resultados nos experimentos destacando-se a combinação pela maior confiança para consultas navegacionais populares e menor confiança para consultas navegacionais aleatórias quando o algoritmo *Pagerank* é utilizado em uma máquina de busca. Quando foi utilização do algoritmo *Indegree* na máquina de busca a melhor combinação para consultas aleatórias foi a média das confianças para o cenário sem combinação e a combinação com o ou probabilístico para o cenário BNC e para consultas aleatórias a melhor combinação foi o uso da menor confiança no cenário sem

combinação e média confiança para o cenário BNC. Estes resultados serão utilizados para comparar com os resultados propostos por Carvalho *et al.*

4.2.2.5 – Resultados combinados dos métodos SLLA , Trust-SLAbS e Trust-BMSR

As Tabelas 23 e 24 apresentam os resultados dos experimentos realizados com a combinação dos métodos SLLA, *Trust*-BMSR e *Trust*-SLAbS nas consultas navegacionais populares nos cenários onde o sistema de busca utiliza o algoritmo *Pagerank* e o algoritmo *Indegree* respectivamente. A estratégia de combinação utilizando a menor confiança demonstrou ganhos expressivos e significantes com 68,33% para o cenário Sem Combinação e de 27,23% para o cenário BNC conforme apresentado na Tabela 23.

Forma de Combinar os Métodos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,230478	-	-	0,417476	-	-
Menor Confiança	0,387961	68,33%	0,0001	0,531159	27,23%	0,0003
Maior Confiança	0,289177	25,47%	0,00003	0,510253	22,22%	0,0001
Media Confiança	0,305313	32,47%	0,0001	0,524662	25,67%	0,0003
Ou Probabilistico	0,247770	7,50%	0,0472	0,422382	1,18%	0,3360

Tabela 23: Valores de MRR para consultas navegacionais populares da combinação do SLLA, Trust-SLAbS e Trsut BMSR aplicando as combinações estudadas quando um sistema de busca utiliza o algoritmo Pagerank

Os experimentos desta combinação com o *Indegree* mostraram ganhos moderados e significantes no cenário sem combinação. Isto pode ser observado na combinação com a média das escalas que obteve ganhos de 9,57% conforme apresentado na Tabela 24.

Forma de Combinar os Métodos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,402264	-	-	0,535635	-	-
Menor Confiança	0,405418	0,78%	0,0048	0,520424	-2,84%	0,1631
Maior Confiança	0,342123	-14,95%	0,1483	0,527052	-1,60%	0,0593
Media Confiança	0,440780	9,57%	0,0064	0,507408	-5,27%	0,3836
Ou Probabilistico	0,437196	8,68%	0,0901	0,509789	-4,83%	0,0676

Tabela 24: Valores de MRR para consultas navegacionais populares da combinação do SLLA, Trust-SLAbS e Trsut BMSR aplicando as combinações estudadas quando um sistema de busca utiliza o algoritmo Indegree

As Tabelas 25 e 26 apresentam os resultados dos experimentos com as consultas navegacionais aleatórias. A estratégia de combinação utilizando o menor valor de confiança das escalas também apresentou melhores resultados para estas consultas, proporcionando um ganho de 25,36% para o cenário sem combinação e 9,98% para o BNC com significância alta.

Forma de Combinar os Métodos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,241811	-	-	0,507483	-	-
Menor Confiança	0,327317	35,36%	0,0012	0,558144	9,98%	0,0148
Maior Confiança	0,273504	13,11%	0,0908	0,575505	13,40%	0,0052
Media Confiança	0,288515	19,31%	0,0003	0,582237	14,73%	0,0101
Ou Probabilistico	0,246049	1,75%	0,2687	0,502331	-1,02%	0,3697

Tabela 25: Valores de MRR para consultas navegacionais aleatórias da combinação do SLLA, Trust-SLAbS e Trsut BMSR aplicando as combinações estudadas quando um sistema de busca utiliza o algoritmo Pagerank

Esta combinação para consultas navegacionais aleatórias quando comparado com os resultados de um sistema de busca que utiliza o algoritmo *Indegree* também apresentou ganhos moderados e significativos. Observa-se que a combinação com o ou probabilístico também obteve bons resultados como pode ser visto na Tabela 26. Os experimentos feitos no cenário BNC não obtiveram ganhos, semelhantes ao obtidos nas consultas populares.

Forma de Combinar os Métodos	Sem Combinação			BNC		
	MRR	Ganho(%)	Significância	MRR	Ganho(%)	Significância
Pesos Iguais	0,343364	-	-	0,547397	-	-
Menor Confiança	0,343878	0,15%	0,0559	0,553138	1,05%	0,3174
Maior Confiança	0,259493	-24,43%	0,2945	0,525519	-4,00%	0,2590
Media Confiança	0,386294	12,50%	0,0182	0,469941	-14,15%	0,0083
Ou Probabilistico	0,402874	17,33%	0,0103	0,501543	-8,38%	0,0579

Tabela 26: Valores de MRR para consultas navegacionais aleatórias da combinação do SLLA, Trust-SLAbS e Trsut BMSR aplicando as combinações estudadas quando um sistema de busca utiliza o algoritmo Indegree

Esta combinação foi a que obteve os melhores resultados de uma máquina de busca quando utiliza o algoritmo *Pagerank*. Para consultas populares os melhores resultados foram

obtidos combinando com o menor valor de confiança dos métodos e para consultas aleatórias os melhores resultados foram apresentados combinando com a média confiança dos métodos.

Quando o cenário da máquina de busca foi alterado utilizando o algoritmo *Indegree*, o método com a média confiança apresentou melhores resultados para as consultas navegacionais populares, para consultas navegacionais aleatórias o método que efetua a combinação com ou probabilístico apresentou melhor resultado no cenário sem combinação. Os resultados com o cenário BNC não apresentaram ganhos significantes.

4.2.3 – Melhores resultados com agrupamentos de *Host*

Nesta seção são apresentados os melhores resultados obtidos com métodos aqui propostos, juntamente com os métodos propostos por (CARVALHO, et al. 2006) que estão identificados nas tabelas com o termo "Remoção de Apontadores". Nas tabelas são apresentados os valores de MRR referente as consultas populares e aleatórias quando um sistema de busca utiliza os algoritmos *Pagerank*, *Indegree* e Remoção de apontadores.

A Tabela 27 apresenta os melhores resultados para consultas navegacionais populares com o agrupamento de host. Observa-se que o método SLLA apresentou os melhores resultados nos métodos individualmente, com ganhos 21,81% para o cenário sem combinação e 25,20% para o cenário combinado (BNC) quando um sistema de busca utiliza o algoritmo *Pagerank*. Este método quando combinado com os métodos aqui propostos, Trust-BMSR e Trust-SLAbS, apresenta ganhos significantes com 68,33% no cenário sem combinação e 27,23% no cenário BNC.

Métodos	Cenário	MRR			Remoção de Apontadores		PageRank	
		Pesos Iguais	Remoção de Apontadores	Trust	% Ganho	Signif.	%Ganho	Signif.
Trust-SLAbS	Sem	0,23048	0,25766	0,26226	1,79%	0,42243	13,79	0.00013
	BNC	0,41748	0,43030	0,45865	6,59%	0,43736	9,86%	0,01060
Trust-BMSR	Sem	0,23048	0,23841	0,24470	2,64%	0,02059	6,17%	0,05524
	BNC	0,41748	0,42043	0,43138	2,60%	0,00951	3,33%	0,00587
SLLA	Sem	0,23048	0,28073	-	-	-	21,81%	0,01621
	BNC	0,41748	0,52267	-	-	-	25,20%	0,00008
Trust-BMSR+ Trust-SLAbS	Sem	0,23048	0,26151	0,25892	-0,99%	0,15560	12,34%	0,06700
	BNC	0,41748	0,43038	0,45389	5,46%	0,01347	8,72%	0,01730
SLLA+Trust-BMSR	Sem	0,23048	0,28073	0,30602	9,01%	0,09350	32,77%	0,00865
	BNC	0,41748	0,52168	0,52625	0,88%	0,37854	26,05%	0,00005
SLLA+Trust-SLAbS	Sem	0,23048	0,30297	0,30175	-0,40%	0,47550	30,92%	0,01320
	BNC	0,41748	0,51188	0,52818	3,18%	0,22576	26,52%	0,00002
SLLA+Trust-BMSR +Trust-SLAbS	Sem	0,23048	0,30214	0,38796	28,41%	0,00002	68,33%	0,00010
	BNC	0,41748	0,51188	0,53116	3,77%	0,18943	27,23%	0,00004

Tabela 27: Melhores resultados dos métodos estudados com os valores de MRR das consultas populares quando um sistema de busca utiliza o algoritmo Pagerak e grafo com os apontadores ruidosos removidos

Comparando os resultados obtidos em relação aos métodos de remoção de apontadores da base, constata-se ganho de 28,41% no cenário sem combinação com significância alta e ganho de 3,77% no cenário BNC com significância baixa. Apesar dos ganhos dos métodos propostos nesta dissertação serem pequenos, quando comparados aos de Carvalho *et al.* no cenário BNC, isto não inviabiliza sua utilização já que o cálculo das escala de confiança não depende de intervenção humana.

Na Tabela 28 são apresentados os melhores resultados dos experimentos nas consultas navegacionais aleatórias. O método SLLA também apresentou os melhores resultados para as consultas navegacionais aleatórias quando avaliado individualmente. Na combinação do SLLA com Trust-BMSR e Trust-SLAbS os ganhos também foram altos e significativos. No cenário sem combinação o ganho em relação ao método de remoção de apontadores foi de 21,47% e com os novos métodos 35,36%. Quando comparado no cenário BNC os métodos

aqui propostos têm uma pequena perda de 1,69% quando comparados com os resultados dos métodos com a remoção de apontadores.

Métodos	Cenário	MRR			Remoção de Apontadores		PageRank	
		Pesos Iguais	Remoção de Apontadores	Trust	% Ganho	Signif.	%Ganho	Signif.
Trust-SLAbS	Sem	0,24181	0,22555	0,23243	3,05%	0,33994	-3,88%	0,17771
	BNC	0,50748	0,48979	0,50490	3,09%	0,25226	-0,51%	0,29930
Trsut-BMSR	Sem	0,24181	0,24076	0,24244	0,70%	0,33078	0,26%	0,24083
	BNC	0,50748	0,49859	0,50748	1,78%	0,24795	0,00%	0,32064
SLLA	Sem	0,24181	0,28000	-	-	-	15,79%	0,04082
	BNC	0,50748	0,56912	-	-	-	12,14%	0,00599
Trust-BMSR+ Trust-SLAbS	Sem	0,24181	0,23248	0,23004	-1,05%	0,01417	-4,87%	0,11900
	BNC	0,50748	0,48946	0,49878	1,90%	0,00345	-1,71%	0,31120
SLLA+Trust-BMSR	Sem	0,24181	0,27982	0,28000	0,07%	0,11800	15,79%	0,04083
	BNC	0,50748	0,55801	0,58523	4,88%	0,03471	15,32%	0,00811
SLLA+Trust-SLAbS	Sem	0,24181	0,26946	0,27350	1,50%	0,40093	13,11%	0,00409
	BNC	0,50748	0,56775	0,57551	1,37%	0,38023	13,40%	0,00004
SLLA+Trust-BMSR +Trust-SLAbS	Sem	0,24181	0,26946	0,32732	21,47%	0,01245	35,36%	0,00120
	BNC	0,50748	0,56775	0,58224	2,55%	0,35353	14,73%	0,08838

Tabela 28: Valores de MRR das consultas aleatórias com os melhores resultados quando um sistema de busca utiliza o algoritmo *Pagerank*

O Gráfico 1 apresenta os valores do MRR de todos os métodos aqui avaliados no cenário sem combinação para consultas navegacionais populares e aleatórias.

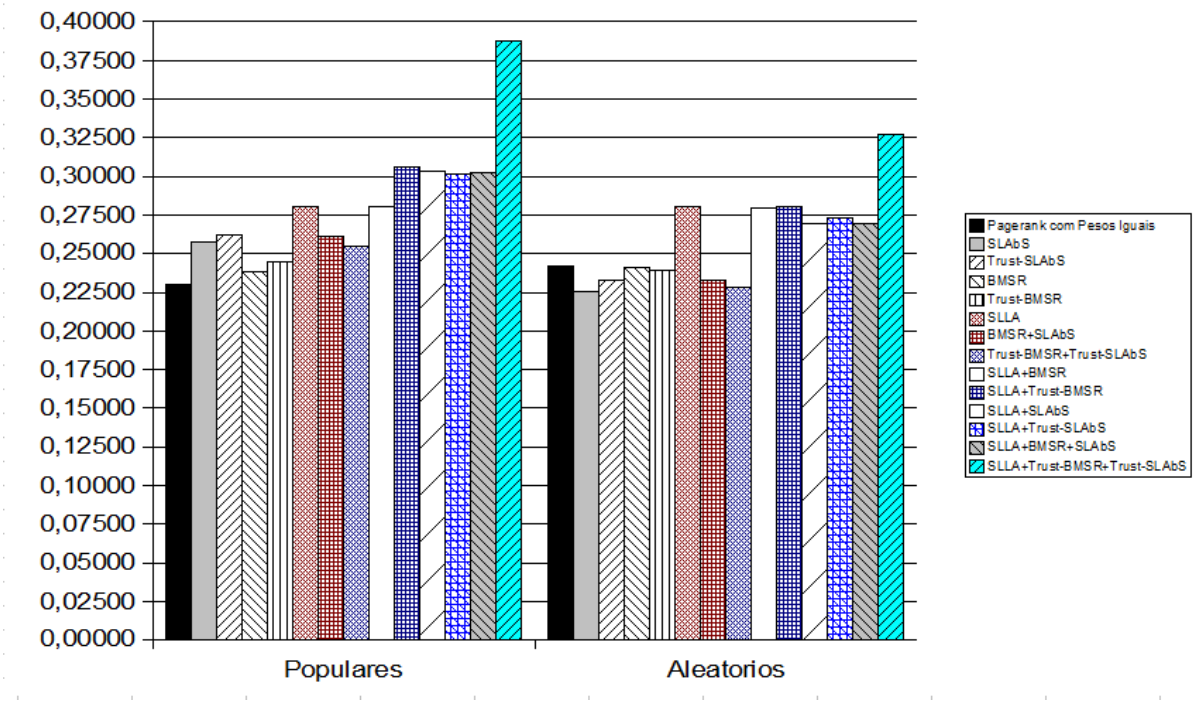

Gráfico 1: Melhores resultados sem combinação para todos os métodos avaliados quando um sistema de busca utiliza o algoritmo *Pagerank*

Constata-se que os resultados individuais dos métodos aqui propostos superaram os resultados dos métodos com remoção de apontadores com ganhos pequenos. Nos métodos combinados, os valores de MRR estão muito próximos destacando-se a combinação dos três métodos propostos nesta dissertação. Outro ponto a ser observado é que o método SLLA tem o valor de MRR maior que todos os métodos combinados para as consultas aleatórias, excetuando a combinação dos três métodos aqui propostos.

O Gráfico 2 apresenta os resultados dos métodos avaliados no cenário BNC para consultas aleatórias e populares.

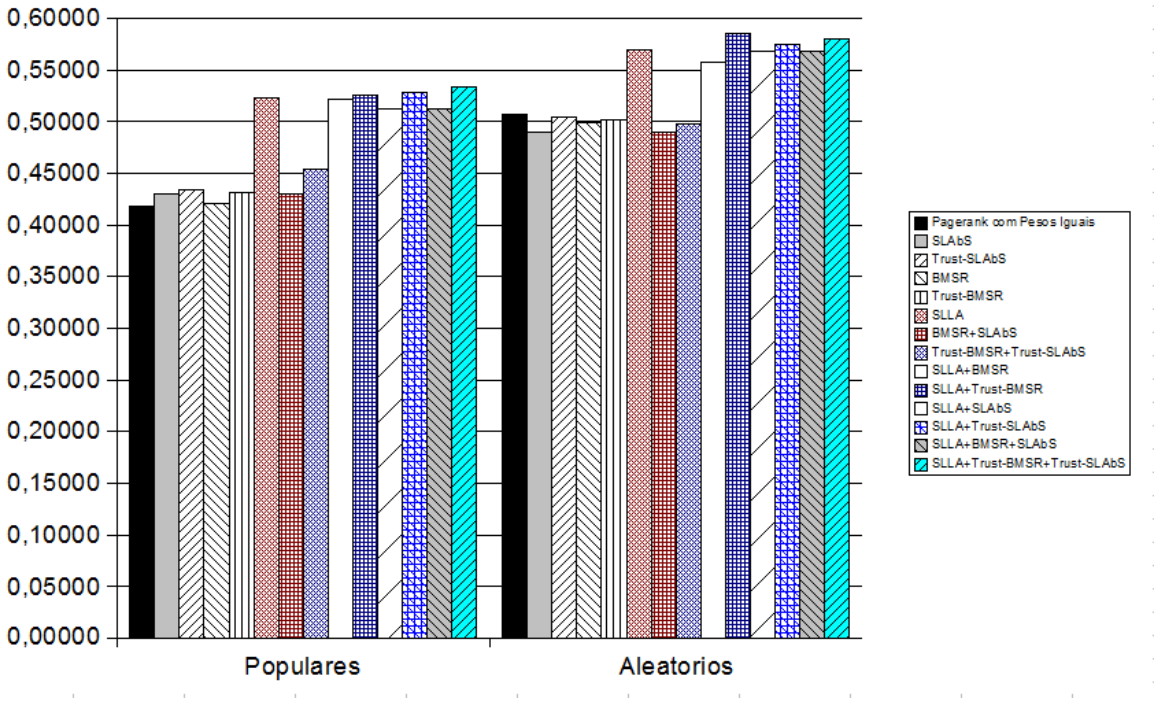

Gráfico 2: Melhores resultados com combinação BNC para todos os métodos avaliados quando um sistema de busca utiliza o algoritmo *Pagerank*

Comparando o resultado com os métodos individuais, o SLLA também destaca-se como o melhor resultado para as consultas aleatórias ficando próximo dos métodos combinados. Um ponto a ser destacado, é que os resultados dos métodos combinados obtiveram valores de MRR muito próximos, deixando os resultados inconclusivos para indicar a melhor opção para implementar em uma máquina de busca real.

A Tabela 29 apresenta os melhores resultados dos métodos para consultas navegacionais populares comparados com o *Indegree*. Observa-se que o melhor resultado obtido foi a combinação do SLLA, Trust-SLAbS e Trust-BMSR com ganho de 9,57% para o cenário sem combinação. Os resultados no cenário combinado não apresentaram resultados significativos tendo o método Trust-.SLAbS como melhor resultado com ganho de 0,88%.

Métodos	Cenário	MRR			
		Pesos Iguais	Trust	% Ganho	Signif.
Trust-BMSR	Sem	0,402264	0,419839	4,37%	0,00200
	BNC	0,535635	0,531344	-0,80%	0,33130
Trust-SLAbS	Sem	0,402264	0,411621	2,33%	0,31000
	BNC	0,535635	0,540357	0,88%	0,31440
SLLA	Sem	0,402264	0,40242	0,04%	0,16060
	BNC	0,535635	0,53593	0,06%	0,16060
Trust-BMSR +TrustSLAbS	Sem	0,402264	0,410705	2,10%	0,32860
	BNC	0,535635	0,533711	-0,36%	0,44370
SLLA+Trust-BMSR	Sem	0,402264	0,412975	2,66%	0,01650
	BNC	0,535635	0,524193	-2,14%	0,09800
SLLA+Trust-SLAbS	Sem	0,402264	0,429585	6,79%	0,01710
	BNC	0,535635	0,537480	0,34%	0,44570
SLLA+Trust-BMSR +Trust-SLAbS	Sem	0,402264	0,440780	9,57%	0,00640
	BNC	0,535635	0,527052	-1,60%	0,05930

Tabela 29 : Melhores resultados do MRR para consultas navegacionais populares quando um sistema de busca utiliza o algoritmo Indegree

A Tabela 30 apresenta os melhores resultados dos métodos para as consultas navegacionais aleatórias comparadas com o *Indegree*. Observa-se ganhos pequenos na maioria dos casos destacando-se a combinação do SLLA, Trust-BMSR e Trust-SLAbS com ganhos de 17,33% no cenário sem combinação. Os melhores resultados no cenário combinado foram os métodos Trust-SLAbS e o combinado SLLA com o Trust-SLAbS que obtiveram ganhos da ordem de 2,10% porém não foram significativos.

Métodos	Cenário	MRR			
		Pesos Iguais	Trust	% Ganho	Signif.
Trust-BMSR	Sem	0,343364	0,358359	4,37%	0,00310
	BNC	0,547400	0,547397	0,00%	Não
Trust-SLAbS	Sem	0,343364	0,361715	5,34%	0,16920
	BNC	0,547400	0,558911	2,10%	0,17920
SLLA	Sem	0,343364	0,343520	0,04%	0,16060
	BNC	0,547400	0,547400	0,00%	Não
Trust-BMSR +TrustSLAbS	Sem	0,343364	0,355574	3,56%	0,25850
	BNC	0,547400	0,541563	-1,07%	0,37620
SLLA+Trust-BMSR	Sem	0,343364	0,354420	3,22%	0,01649
	BNC	0,547400	0,547582	0,03%	0,16069
SLLA+Trust-SLAbS	Sem	0,343364	0,366235	6,66%	0,12800
	BNC	0,547400	0,558971	2,11%	0,17920
SLLA+Trust-BMSR +Trust-SLAbS	Sem	0,343364	0,402874	17,33%	0,01030
	BNC	0,547400	0,553138	1,05%	0,31740

Tabela 30: Melhores resultados do MRR para consultas navegacionais aleatórias quando um sistema de busca utiliza o algoritmo Indegree

O Gráfico 3 apresenta os resultados dos experimentos no cenário sem combinação comparando com o método *Indegree* para as consultas navegacionais aleatórias e populares.

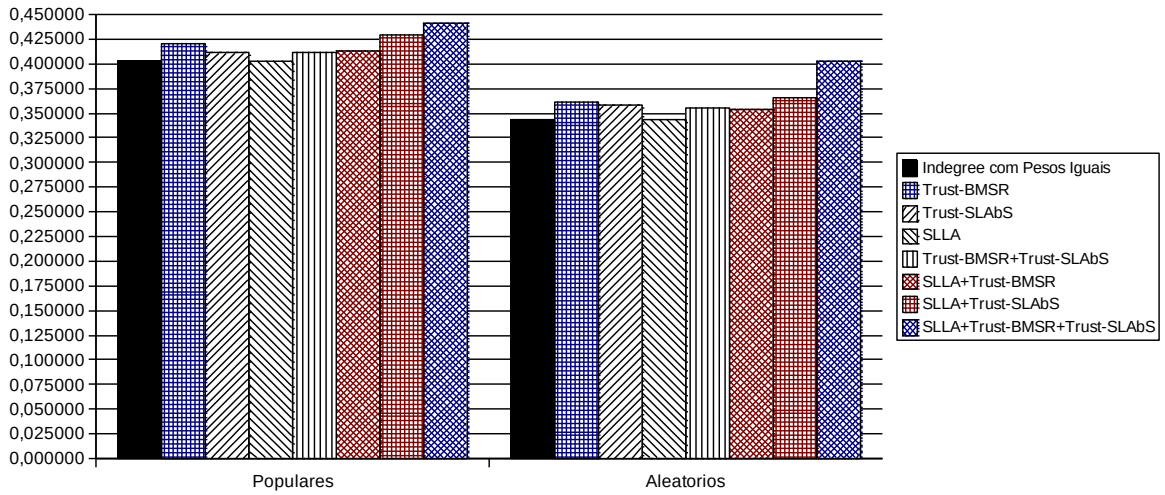

Gráfico 3: Melhores resultados dos métodos avaliados com Indegree no cenário sem combinação quando um sistema de busca utiliza o algoritmo *Indegree*

Observa-se um pequeno destaque para os ganhos dos métodos que utilizaram as combinações dos três métodos. Entretanto, os valores de MRR tanto dos métodos aqui propostos, quanto dos métodos com remoção de apontadores, quando comparados com os métodos com peso iguais têm seus valores muito próximos, levando a conclusão de que estes métodos não são indicados para serem implementados em um sistema de busca que utiliza o algoritmo *Indegree*.

O Gráfico 4 apresenta os melhores resultados dos experimentos para o cenário combinado (BNC) para as consultas navegacionais e aleatórias comparando os resultados quando um sistema de busca utiliza o algoritmo *Indegree*.

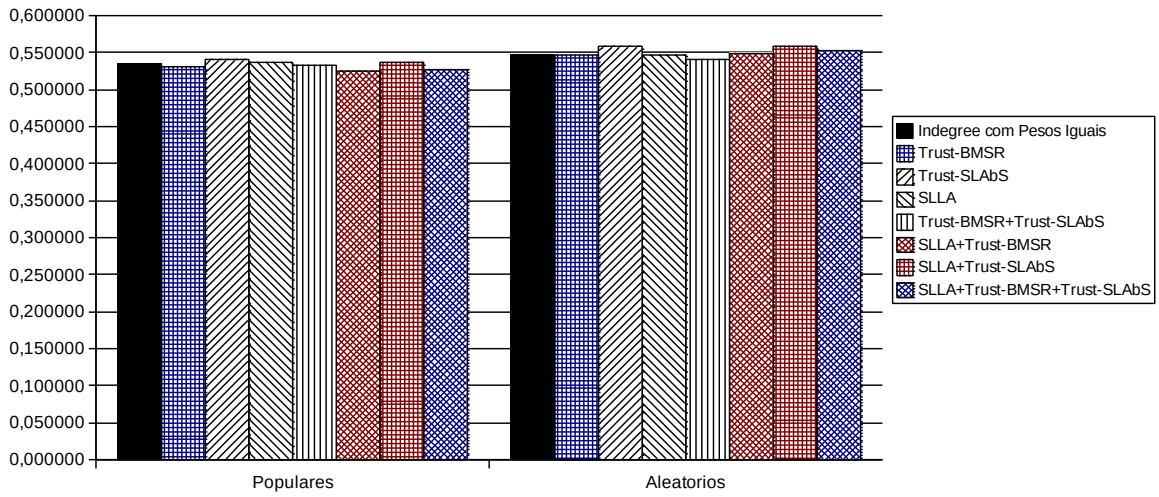

Gráfico 4: Melhores resultados dos métodos avaliados com Indegree no cenário combinado quando um sistema de busca utiliza o algoritmo Indegree

Nesta seção foram apresentados os melhores resultados dos valores de MRR para consultas navegacionais populares e aleatórias com agrupamentos de *host*. Os resultados mostraram que a combinação do método SLLA com os métodos aqui propostos, Trust-BMSR e Trust-SLAbS, apresentaram ganhos altos e significativos quando comparados com os métodos propostos por Carvalho *et al.* em um sistema de busca utilizando o algoritmo *Pagerank* no cenário sem combinação. Quando o cenário utilizado na máquina de busca é o BNC, os resultados não são significativos. Entretanto, os métodos aqui propostos apresentam-se como alternativa aos métodos de remoção de arestas, uma vez que estes não removeram arestas do grafo original que representa a Web e não houve a necessidade de intervenção humana para identificação definição de limiares.

4.2.4 – Melhores resultados com agrupamento por Domínios

Os experimentos realizados com *host* apresentaram bons resultados. Por este motivo optou-se em avaliar os resultados dos métodos aqui propostos com agrupamento feitos com domínios. A Tabela 31 apresenta os valores de MRR dos experimentos para consultas navegacionais populares comparadas com os métodos que removem os apontadores dos relacionamentos suspeitos quando sistemas de busca utilizam o algoritmo *Pagerank*.

Métodos	Cenário	MRR			Remoção de Apontadores		Pesos Iguais	
		Pesos Iguais	Remoção de Apontadores	Trust	% Ganho	Signif.	%Ganho	Signif.
Trust-SLAbS	Sem	0,23048	0,27792	0,27789	-0,01%	0,49935	20,57%	0,02992
	BNC	0,41748	0,46748	0,52987	13,35%	0,00304	26,92%	0,00008
Trust-BMSR	Sem	0,23048	0,27742	0,29830	7,52%	0,18276	29,43%	0,00648
	BNC	0,41748	0,45939	0,50639	10,23%	0,20698	21,30%	0,00780
SLLA	Sem	0,23048	0,27471	-	19,19%	0,02898	-	-
	BNC	0,41748	0,51261	-	22,79%	0,00023	-	-
Trust-BMSR+ Trust-SLAbS	Sem	0,23048	0,28806	0,26794	-6,99%	0,00000	16,25%	0,00001
	BNC	0,41748	0,48217	0,47362	-1,77%	0,00100	13,45%	0,06920
SLLA+Trust-BMSR	Sem	0,23048	0,32377	0,32937	1,73%	0,41485	42,91%	0,00108
	BNC	0,41748	0,54129	0,60259	11,33%	0,00949	44,34%	0,00000
SLLA+Trust-SLAbS	Sem	0,23048	0,33217	0,30945	-6,84%	0,17253	34,26%	0,00460
	BNC	0,41748	0,52828	0,56583	7,11%	0,04524	35,54%	0,00001
SLLA+Trust-BMSR+Trust-SLAbS	Sem	0,23048	0,28981	0,44750	54,41%	0,00002	94,16%	0,00001
	BNC	0,41748	0,48256	0,65808	36,37%	0,00002	57,63%	0,00001

Tabela 31: Valores de MRR das consultas navegacionais populares com os melhores resultados dos métodos quando um sistema de busca utiliza o algoritmo Pagerank

Observa-se que na maioria dos resultados os métodos aqui propostos obtiveram ganhos altos e significantes quando comparados com os métodos com remoção de apontadores. Individualmente, o método SLLA destaca-se como o melhor resultado para o cenário sem combinação com 19,19% e 22,79% para o cenário combinado (BNC). Nos métodos combinados, os melhores resultados foram apresentados pela combinação dos métodos SLLA, Trust-BMSR e Trust-SLAbS com ganhos na ordem de 54,41% no cenário

sem combinação e 36,37% no cenário BNC em relação aos métodos com remoção de apontadores.

A Tabela 32 apresenta os resultados das consultas navegacionais aleatórias.

Métodos	Cenário	MRR			Remoção de Apontadores		Pesos Iguais	
		PageRank	Remoção de Apontadores	Trust	% Ganho	Signif.	%Ganho	Signif.
Trust-SLAbS	Sem	0,24181	0,25217	0,26003	3,12%	0,31659	7,53%	0,17491
	BNC	0,50748	0,51743	0,54903	6,11%	0,07197	8,19%	0,09960
Trust-BMSR	Sem	0,24181	0,29110	0,29843	2,52%	0,11713	23,42%	0,00207
	BNC	0,50748	0,55753	0,56031	0,50%	0,39299	10,41%	0,01698
SLLA	Sem	0,24181	0,27794	-	14,94%	0,04963	-	-
	BNC	0,50748	0,55969	-	10,29%	0,02205	-	-
Trust-BMSR+ Trust-SLAbS	Sem	0,24181	0,25657	0,26429	3,01%	0,00001	9,30%	0,00001
	BNC	0,50748	0,51978	0,54195	4,27%	0,00055	6,79%	0,00230
SLLA+Trust-BMSR	Sem	0,24181	0,33954	0,30747	-9,44%	0,17275	27,15%	0,03547
	BNC	0,50748	0,59928	0,64230	7,18%	0,04523	26,57%	0,00040
SLLA+Trust-SLAbS	Sem	0,24181	0,28564	0,27237	-4,64%	0,17762	12,64%	0,11555
	BNC	0,50748	0,57476	0,57502	0,04%	0,49422	13,31%	0,03268
SLLA+Trust-BMSR +Trust-SLAbS	Sem	0,24181	0,26191	0,34656	32,32%	0,00999	43,32%	0,00141
	BNC	0,50748	0,52256	0,62225	19,08%	0,00918	22,61%	0,00775

Tabela 32: Valores de MRR das consultas aleatórias com os melhores resultados quando um sistema de busca utiliza o algoritmo Pagerank

Semelhante aos resultados das consultas populares o SLLA destacou-se como o melhor método individual com 14,94% para o cenário sem combinação e 10,41% de ganho para o cenário BNC com alta significância estatística. A combinação dos métodos SLLA, Trust-SLAbS e Trust-BMSR obteve 32,32 % para cenário sem combinação e método combinado do SLLA+Trust-BMSR obteve ganho de 19,08% em relação aos métodos que removem os apontadores do grafo.

O Gráfico 5 apresenta os resultados dos experimentos realizados com todos os métodos utilizando o agrupamento por domínios para consultas navegacionais populares e aleatórias utilizando o cenário sem combinação no processador de consultas.

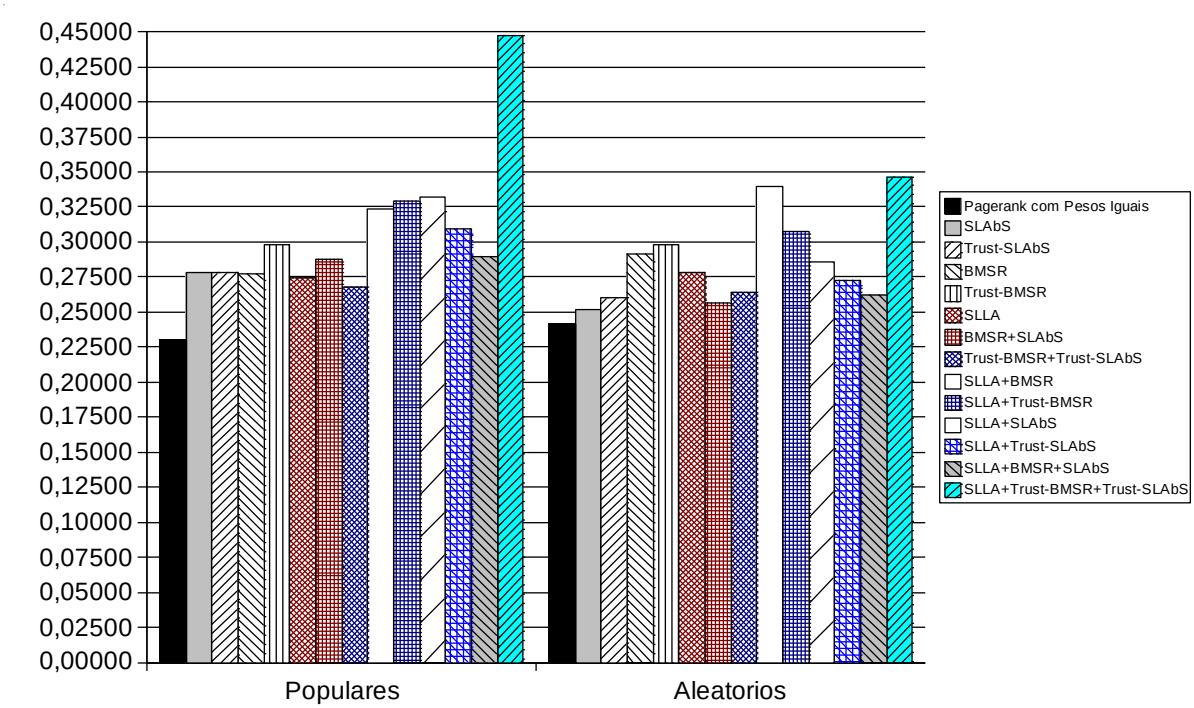

Gráfico 5: Valores de MRR gerados no cenário sem combinação para as consultas populares e aleatórias com agrupamento de domínios quando um sistema de busca utiliza o algoritmo *Pagerank*

Observa-se que a combinação dos três métodos destaca-se para consultas populares. Nas consultas aleatórias o método com remoção de apontadores SLLA+BMSR obtém o valor de MRR bem próximo da combinação do SLLA, Trust-BMSR e Trust-SLAbS.

O Gráfico 6 apresenta os resultados dos experimentos realizados com todos os métodos utilizando o agrupamento por domínios para consultas navegacionais aleatórias utilizando o cenário BNC.

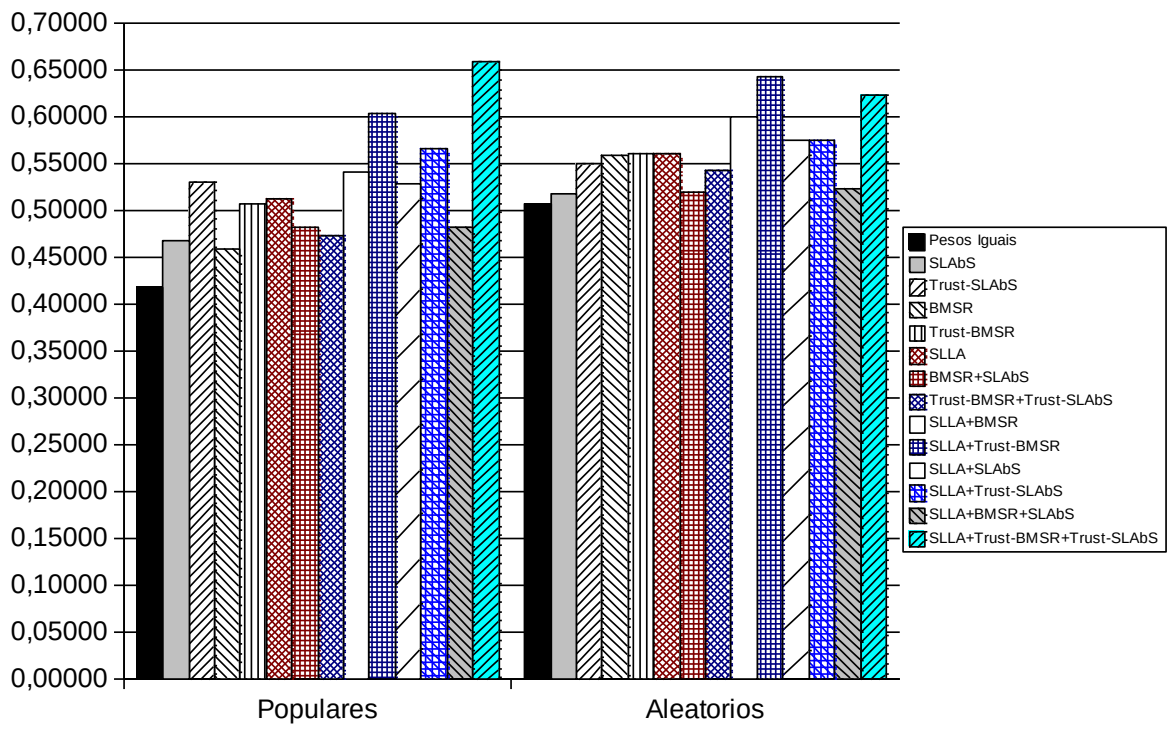

Gráfico 6: Valores de MRR gerados no cenário combinado (BNC) para as consultas populares e aleatórias com agrupamento de domínios quando um sistema de busca utiliza o algoritmo *Pagerank*

Os métodos quando avaliados com agrupamento de domínios obtiveram resultados melhores que os de *host* quando uma máquina de busca utiliza o algoritmo *Pagerank*. Isto ocorre pois os domínios têm um número de grupos menores e um número menor de apontadores trocados entre si. Isto eleva os valores de confiança melhorando o cálculo do *Pagerank* para consultas navegacionais.

A Tabela 33 apresenta os valores de MRR referentes aos resultados dos experimentos realizados com as consultas navegacionais populares quando um sistema de busca utiliza o algoritmo *Indegree*.

Métodos	Cenário	MRR			
		Pesos Iguais	Trust	% Ganho	Signif.
Trust-BMSR	Sem	0,402264	0,406865	1,14%	0,31549
	BNC	0,535635	0,519564	-3,00%	0,09925
Trust-SLAbS	Sem	0,402264	0,409021	1,68%	0,00235
	BNC	0,535635	0,520075	-2,90%	0,47466
SLLA	Sem	0,402264	0,402416	0,04%	0,16060
	BNC	0,535635	0,53593	0,06%	0,16060
Trust-BMSR +TrustSLAbS	Sem	0,402264	0,410561	2,06%	0,31269
	BNC	0,535635	0,520075	-2,90%	0,17143
SLLA+Trust-BMSR	Sem	0,402264	0,420032	4,42%	0,09798
	BNC	0,535635	0,538421	0,52%	0,43900
SLLA+Trust-SLAbS	Sem	0,402264	0,448748	11,56%	0,00332
	BNC	0,535635	0,544294	1,62%	0,32666
SLLA+Trust-BMSR +Trust-SLAbS	Sem	0,402264	0,451210	12,17%	0,00221
	BNC	0,535635	0,544690	1,69%	0,31935

Tabela 33: Valores de MRR das consultas navegacionais populares com os melhores resultados dos métodos com agrupamento de domínios quando um sistema de busca utiliza o algoritmo *Indegree*

Pode-se observar que no agrupamento de domínio houve um ganho de 12,17% na combinação dos métodos SLLA, Trust-BMSR e Trust-SLAbS para o cenário sem combinação. Observa-se que no agrupamento de domínios os métodos obtiveram ganhos melhores que os ganhos obtidos com *host*.

Nas consultas navegacionais aleatórias os resultados do MRR mostram ganhos pequenos e significativos. Pode-se observar que o método combinado SLLA e Trust-BMSR obteve um ganho de 14,86% para o cenário sem combinação e ganho de 5,54% para o cenário BNC com o método SLLA e Trust-SLAbS.

Métodos	Cenário	MRR			
		Pesos Iguais	Trust	% Ganho	Signif.
Trust-BMSR	Sem	0,343364	0,370550	7,92%	0,04627
	BNC	0,547400	0,539249	-1,49%	0,28837
Trust-SLAbS	Sem	0,343364	0,353559	2,97%	0,00000
	BNC	0,547400	0,550313	0,53%	0,33000
SLLA	Sem	0,343364	0,34352	0,04%	0,16060
	BNC	0,547400	0,54740	0,00%	Não
Trust-BMSR +TrustSLAbS	Sem	0,343364	0,359009	4,56%	0,19444
	BNC	0,547400	0,547536	0,03%	0,49576
SLLA+Trust-BMSR	Sem	0,343364	0,394382	14,86%	0,00823
	BNC	0,547400	0,569665	4,07%	0,12234
SLLA+Trust-SLAbS	Sem	0,343364	0,372126	8,38%	0,05065
	BNC	0,547400	0,577721	5,54%	0,04508
SLLA+Trust-BMSR +Trust-SLAbS	Sem	0,343364	0,378851	10,34%	0,02347
	BNC	0,547400	0,569388	4,02%	0,13443

Tabela 34: Valores de MRR das consultas navegacionais aleatórias com os melhores resultados dos métodos com agrupamento de domínios quando um sistema de busca utiliza o algoritmo *Indegree*

O Gráfico 7 apresenta os resultados dos métodos utilizando o cenário sem combinação para as consultas navegacionais populares e aleatórias com agrupamento de domínios.

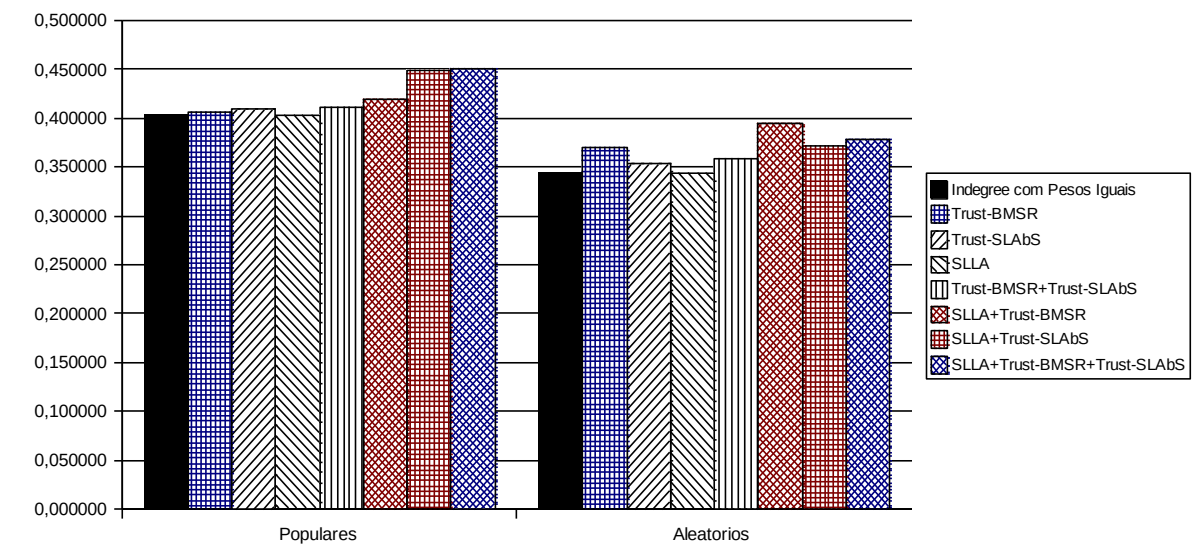

Gráfico 7: Valores de MRR gerados no cenário combinado sem combinação para as consultas populares e aleatórias com agrupamento de domínios quando um sistema de busca utiliza o algoritmo *Indegree*

Observa-se que os ganhos obtidos para consultas aleatórias são muito próximos dos valores obtidos quando o algoritmo quando o sistema de busca utiliza do algoritmo *Indegree*.

O Gráfico 8 apresenta os resultados dos métodos no cenário combinado (BNC) utilizando a métrica do *Indegree* com agrupamento de domínios.

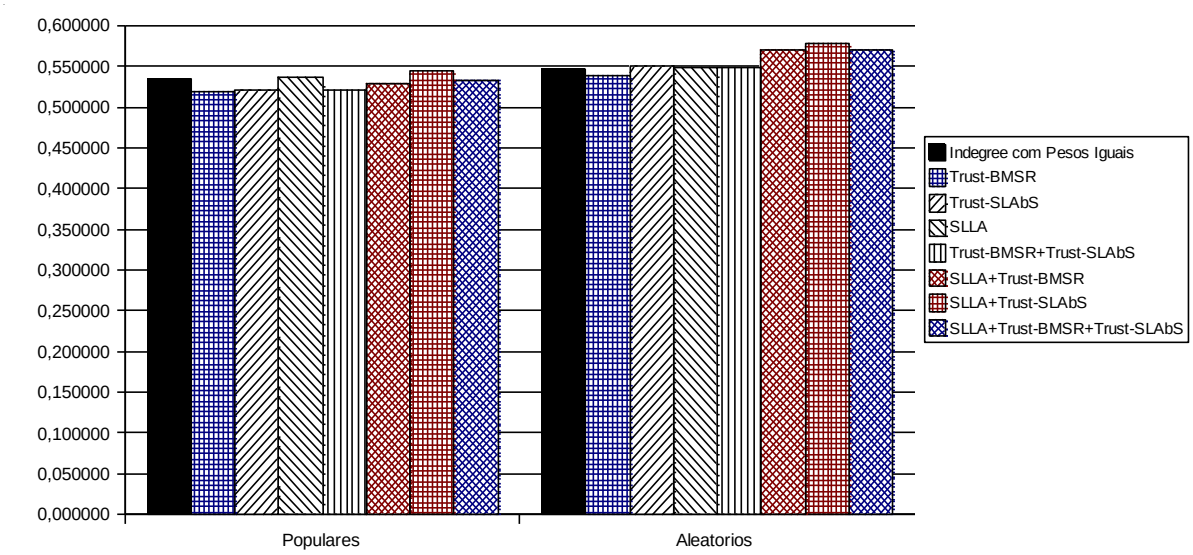

Gráfico 8: de MRR gerados no cenário combinado (BNC) para as consultas populares e aleatórias com agrupamento de domínios quando um sistema de busca utiliza o algoritmo *Indegree*

Avaliando os resultados dos métodos estudados em um sistema de busca quando utiliza o algoritmo *Indegree*, observam-se ganhos pequenos e um pouco melhores que no agrupamento de host. Isto porque os valores de confiança melhoraram em função do agrupamento ser maior, e conseqüentemente com um número menor de apontadores, levando a um valor de confiança maior. Apesar desta melhoria, os métodos aqui desenvolvidos não são indicados para serem usados em uma máquina de busca que utilizam o algoritmo *Indegree*.

4.3 – Avaliação sobre os resultados dos experimentos

A adaptação do algoritmo Pagerank no cenário BNC com os métodos aqui propostos obteve melhores resultados quando o agrupamento foi feito com domínios. Observa-se ganhos de 36,37% quando comparado com os melhores resultados dos métodos de remoção de

apontadores para as consultas navegacionais populares e ganhos de 19,08% com as consultas navegacionais aleatórias. Analisando-se os resultados com agrupamentos de *host* no mesmo cenário, observa-se que os ganhos são menores, com 3,77% para as consultas navegacionais populares e 2,55% para as consultas navegacionais aleatórias.

Com estes resultados pode-se concluir que os métodos aqui propostos são uma alternativa interessante para substituição dos métodos de remoção de apontadores. Os resultados apresentaram ganhos quando foram realizados agrupamentos com domínios e se comportaram de maneira semelhante quando o agrupamento foi realizado com *host*. Alem disso os métodos não precisam de intervenção humana para encontrar os melhores limiares de corte para remoção de apontadores.

Capítulo 5

5- Conclusões e Trabalhos Futuros

Nesta dissertação foram realizados estudos sobre técnicas para detecção de apontadores ruidosos em base de dados de máquinas de busca. Foram propostos e avaliados métodos para identificar relacionamentos suspeitos como o reforço mútuo, suporte anormal e aliança de apontadores. Os métodos aqui propostos basearam-se nos relacionamentos existentes entre sítios para detectar aqueles considerados ruidosos.

Os métodos propostos por Carvalho et al. (2006) necessitam que para cada base em que os métodos serão utilizados, um especialista execute experimentos continuamente com diferentes valores de limiares buscando atingir o melhor desempenho de seus métodos. Estes limiares servem de base para indicar que os relacionamentos que estiverem acima destes valores são considerados ruidosos e devem ser removidos do grafo que representa a web. Este trabalho demanda um esforço muito grande principalmente se a base de dados for muito grande que é caso da *Internet*.

Diferentemente de Carvalho et al., a abordagem utilizada nos métodos aqui proposto não remove do grafo da Web os relacionamentos considerados suspeitos, mas propõe um escala de confiança para estes relacionamentos com pesos que variam entre zero (para relacionamentos não confiáveis) e um (para relacionamentos totalmente confiáveis).

O uso desta escala resultou em ganhos em todos os cenários experimentados. Quando os agrupamentos no grafo são feitos com *host*, os ganhos apresentados em relação aos

métodos de remoção de apontadores foram de 27,35% para consultas navegacionais populares e 21,47% para consultas navegacionais aleatórias. Isto foi obtido quando um processador de consultas utiliza somente a reputação das páginas calculadas pelos métodos (cenário sem combinação). Quando o processador de consultas foi configurado para considerar, além da reputação das páginas, o texto de âncora e o conteúdo textual (cenário BNC), os ganhos foram pequenos para consultas navegacionais populares, 3,77%, e houve perda de 1,69% para consultas navegacionais aleatórias. Como esse cenário é o mais importante na prática e testes estatísticos indicaram que os resultados nesse cenário deram diferenças não significativas, pode-se concluir que não houve melhoria de *performance* nos experimentos considerando a partição das páginas em *hosts*.

Nos resultados dos experimentos comparados com o *Indegree*, destacou-se apenas a combinação dos métodos SLLA, *Trust*-SLAbS e *Trust*-BMSR para as consultas navegacionais aleatórias no cenário sem combinação com ganho de 17,33% quando o agrupado por *host*. Os resultados quando avaliados no cenário BNC não apresentaram ganhos significativos mostrando que os métodos aqui desenvolvidos não são indicados para serem usados com este algoritmo de análise de apontadores.

Foram realizados testes utilizando a mesma coleção com o agrupamento de domínios onde os resultados obtidos foram melhores que no agrupamento por *host*. Os ganhos nesse cenário para o uso de *Pagerank* foram de 54,41% em relação aos métodos de remoção de apontadores para consultas navegacionais populares e 32,32% nas consultas navegacionais aleatórias quando o processador utiliza o cenário sem combinação de evidências. Os ganhos foram de 36,37% para consultas populares e 19,08% para consultas aleatórias para o cenário BNC. Estes resultados mostram que os métodos aqui desenvolvidos apresentam ganhos quando utilizado o agrupamento de domínios para o grafo que representa a Web utilizando o

ambiente BNC, que está mais próximo das máquinas de busca reais, indicando que estes métodos podem ser facilmente implementado nestas máquinas.

Quando foi utilizado o algoritmo *Indegree*, os experimentos indicaram ganhos no cenário sem combinação, com 12,17% para consultas navegacionais populares e 14,86% para consultas aleatórias. Semelhante ao agrupamento de *host*, os resultados no cenário combinado (BNC) não deram ganhos significativos. Confirmando que estes métodos não são indicados para serem usados com este algoritmo.

Os estudos e experimentos realizados demonstram que o uso de pesos atribuídos como graus de confiança nos relacionamentos existentes no grafo da *Web* apresentaram bons resultados e melhoraram a qualidade das respostas na maior parte dos cenários experimentados. Outro ponto importante, sendo este uma das contribuições desta dissertação, é a inexistência da intervenção humana para configurar valores de parâmetros nos métodos.

Como trabalhos futuros, propõem-se aplicar estes métodos em outras coleções para dar mais segurança aos estudos aqui realizados. Pretende-se ainda adaptar os métodos desta dissertação a um cenário onde a Web é modelada como um hipergrafo (BERLT, et al. 2007), onde os autores propõem a modelagem da Web como um hipergrafo para computar a reputação de páginas, permitindo um controle da qualidade das conexões entre páginas da *Web*.

Bibliografia

BENCZUR, Andras A., Karoly CSALOGANY, Tamas SARLOS, e Mate UTHER. "Spamrank - fully automatic link spam detection. In First International Workshop ." *Adversarial Information Retrieval on the Web*, 2005.

BERLT, Klessius, Edleno Silva DE MOURA, André CARVALHO, Marco CRISTO, Nivio ZIVIANI, e Thierson COUTO. "A hypergraph model for computing page reputation on web." *Anais do Simpósio Brasileiro de Banco de Dados*, 2007: 35-49.

BHARAT, Krishna, e Monika R. HENZINGER. "Improved algorithms for topic distillation in a hyperlinked environment." *The 21st ACM International SIGIR Conference on Research and Development in Information Retrieval*, 1998: 104-111.

BRAY, T. "Measuring the web." *The 5th Internacional Word Wide Web Conference on Computer Networks and ISDN Systems.* Amesterdam, The Netherlands: Elsevier Science Publishers B.V., 1996. 993-1005.

BRIN, Sergey, e Lawrence PAGE. "The Anatomy of a Large-Scale Hypertextual." *7th International Word Wide Web Conference*, 1998: 107-117.

BRODER, Andrei Z. "A taxonomy of web search." *SIGIR Forum*, 2002: 3 -10.

CALADA, Pável, Berthier RIBEIRO-NETO, Nivio ZIVIANI, Edleno MOURA, e Ilmério SILVA. "Local Versus Global Link Information in the Web." *ACM Transactions on Information Systems*, 2003: 1-22.

CALADO, Pável Pereira, Edleno S. DE MOURA, Berthier RIBEIRO-NETO, Ilmério SILVA, e Nivio ZIVIANE. "Local versus global link information in the web." *ACM Transactions on Information (TOIS)*, 2003: 42-63.

CARVALHO, Andre Luiz da Costa, Paul Alexandru CHIRITA, Edleno Silva de MOURA, Pável CALADO, e Wolfgang Nejdl. “Site level noise removal for search engines.” *15th International Conference on World Wide Web*, 2006: 73-82.

CHAKRABARTI, Soumen. “Integrating the document object model with hyperlinks for enhanced.” *The 10th International Conference on World Wide Web*, 2001: 211-220.

—. *Mining the Web Discovering Knowlegdge from Hypertext Data.* San Francisco: Morgan Kaufmann Publishers, 2003.

GYÖNGYI, Zoltan, e Hector GARCIA-MOLINA. “Link Spam Alliances.” *The 31st Internacional VLDB Conference on Very Large Data Bases*, 2005: 527-528.

GYÖNGYI, Zoltán, e Hector GARCIA-MOLINA. “Web Spam Taxonomy.” 2004.

GYÖNGYI, Zoltán, Hector GARCIA-MOLINA, e Jan PEDERSON. “Combating web spam with trustrank.” 2004: 576-587.

HAWKING, David, Ellen VOORHEES, Nick CRASWELL, e Peter BAILEY. “Overview of the trec8 web track.” *Eighth Text REtrieval Conference*, 1999.

JOACHISMS, Thorsten, Laura GRANKA, Bing PAN, Helene HEMBROOKE, e Geri GAY. “Accurately Interpreting Clickthrough Data as Implicit.” *The 28th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 15-19 de 08 de 2005.

KEHOE, Colleen, Jim PITKOW, Kate SUTTON, Gaurav AGGARWAL, e Juan D. ROGERS. “GVU’s 10th WWW User Survey.” *Graphic, Visualization, and Usability Center.* outubro de 1998. http://www.cc.gatech.edu/gvu/user_surveys/survey-1998-10/graphs/use/q52.htm (acesso em janeiro de 2009).

KLEINBERG, Jon M. “Authoritative sources in a hyperlinked environment.” 1998: 668– 677.

LAWRENCE, Steve, e C. Lee GILE. “Accessibility of information on the Web.” *Intelligence Vol. 11,* 2000: 32-39.

LEMPEL, R., e S. MORAN. “The stochastic approach for link-structure analysis (SALSA) and the TKC effect.” *Computer Networks,* 2000: 387-401.

PAGE, L., S. BRIAN, R. MOTWANI, e T. WINOGRAD. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical report, Stanford University, 1998.

SALTON, G., e C. S. YANG. “A vector space model for automatic indexing.” *Communication of the ACM,* Novembro de 1975.

SALTON, Gerard, e M. J. McGILL. *Introduction to Modern Information Retrieval.* 1a. Mc Graw Hill, 1983.

—. *Introduction to Modern Information Retrieval.* 1a. Mc Graw Hill, 1983.

SHANNON, Claude Elwood. “A Mathematical Theory of Communication.” *The Bell System Technical Journal,* julho de 1948: 379-423.

SILVA, Thomaz P. C., Edleno S. MOURA, João M. B. CAVALCANTE, Altigran S. SILVA, Moisés G. CARVALHO, e Marcos A. GONÇALVES. “An evolutionary approach for combining different sources of evidence in search engines.” *Information Systems,* Julho 2008: 276-289.

WU, Baoning, e Brian D. DAVISON. “Identifying Link Farm Spam Pages.” *International World Wide Web Conference Committee,* 05 de 2005: 10-14.

WU, Baoning, V. GOELI, e Brian D. DAVISON. “Topical TrustRank: using topicality to combat web spam.” *The 15th International Conference on World Wide Web,* 2006.

XUE, Gui-Rong., Qiang YANG, Hua-Jun ZENG, Yong YU, e Zheng CHEN. “Exploiting the hierarchical structure for link analysis.” *the 28th Annual International ACM*

SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2005. 186–193.

ZHANG, Hui, Ashish GOEL, Ramesh GOVINDAN, Kahn MASON, e Benjamin VAN ROY. "Improving eigen vector based reputation systems against collusions." *The 3rd Workshop on Web Graph Algorithms*, 2004.