1
Introdução

A *World Wide Web* (*WWW*) cresce desordenadamente, sem nenhum controle global ou planejamento, tornando-se estruturalmente complexa. Esse ambiente congrega milhões de participantes em tempo real, com interesses divergentes, criando e atualizando as estruturas dos hiperlinks e, por conseguinte, inserindo novas informações e conhecimento.

A estrutura global da *WWW* pode ser observada como uma enorme rede de dados interconectados, em que os nós da rede são as páginas e as conexões são os hiperlinks, formando uma densa estrutura de informação correlacionada.

Essa grande malha de conexões tem sido objeto de vasto estudo, pois permite fazer ilações sobre como a informação na *WWW* está organizada.

Basicamente, os hiperlinks possuem três aspectos importantes que têm sido estudados em relação aos sistemas de Recuperação da Informação (RI), quais sejam:

- Descoberta de novas páginas: ao construir as suas páginas os autores as conectam, através dos hiperlinks, a outras páginas que os mesmos sabem existir. Esse conhecimento do autor sobre a existência de uma página pode advir de diversas fontes de informação de fora da *WWW*. Assim, novos recursos à *WWW* são adicionados e descobertos através dos hiperlinks;
- Medida subjetiva de importância: quando um autor de uma página julga outra como importante, provavelmente vai criar um hiperlink para essa página. Desta forma, o hiperlink é visto como um julgamento de relevância e a partir da quantidade de referências que uma página possui, podem ser estabelecidas métricas e métodos para qualificá-la e classificá-la;
- Relação entre conteúdo: nesse caso, o autor de uma página pode ter referenciado a outra, pois considerou que o conteúdo de ambas possui

alguma relação.

Assim, boa parte da pesquisa em torno das máquinas de busca na *WWW* objetiva melhorar a qualidade dos seus resultados utilizando as características antes descritas e como isso pode ser feito é de grande valia. Quando todos esses aspectos forem compreendidos, podemos determinar quando e como utilizá-los para filtrar resultados de busca de baixa qualidade.

Nesse ambiente vem à tona um segmento de algoritmos que utiliza a estrutura de hiperlinks para classificar um conjunto de páginas. Esses são denominados algoritmos baseados em hiperlinks e são divididos basicamente em dois grupos:

- os que utilizam apenas as quantidades de hiperlinks de entrada e saída para classificar as páginas;
- e, os que utilizam os hiperlinks para propagar pesos de uma página para a outra.

Os primeiros classificam as páginas de acordo com a contagem direta do número de hiperlinks de entrada (*in-degree*) e de saída (*out-degree*), podendo haver variações que combinam as duas estratégias. Os segundos, que são os mais relevantes para o objeto deste trabalho, valem-se da estrutura de hiperlinks para propagar pontuações entre as páginas com o objetivo de classificá-las.

Quanto a esses, podemos citar o *PageRank* (PAGE et al., 1999), o *HITS* (*Hyperlink-Induced Topic Search*) (KLEINBERG, 1999; KLEINBERG et al., 1999) e o *SALSA* (*Stochastic Approach for Link-Structure Analysis*) (LEMPEL; MORAN, 2001) como os mais importantes e populares.

O *PageRank* atribui os valores às páginas utilizando a estrutura de hiperlink como uma forma de identificar e contextualizar a relevância da página com o todo. De forma simplificada, a ideia principal é atribuir valores altos a páginas que possuam páginas que as referenciam com pontuações altas. Esse princípio cobre tanto a situação em que a página possui muitas referências, como a de poucas referências de páginas com valores altos de pontuação.

O *HITS* (*Hyperlink-Induced Topic Search*), por sua vez, é um algoritmo que calcula a autoridade de uma página para um determinado tópico de consulta. Basicamente, as autoridades relevantes presentes possuem não só um alto grau de hiperlinks que as apontam, como também possuem grupos de páginas que as apontam em comum.

Essas páginas são denominadas *hubs* e são responsáveis por vincular as autoridades comuns excluindo as páginas que possuem um alto grau de hiperlinks de chegada e não são relevantes para o assunto. Desta forma, autoridades e *hubs* exibem uma relação de interdependência: uma boa autoridade será uma página apontada por bons *hubs* e um bom *hub* será uma página que aponta para boas autoridades. A partir dessas definições, o algoritmo iterativamente calcula esse reforço mútuo até atingir a convergência.

O *SALSA* (*Stochastic Approach for Link-Structure Analysis*), por fim, é similar ao *HITS* e possui uma análise ponderada sobre a estrutura de hiperlink. O SALSA efetua uma caminhada aleatória na estrutura em que as transições consistem em trajetos através de dois hiperlinks, um hiperlink para frente e um para trás, utilizando a ideia da autoridade e do *hub*.

Existem muitas propostas diferentes para efetuar buscas e classificação de páginas na *WWW* (CRASWELL; SZUMMER, 2007; RAFIEI; MENDELZON, 2000; FLAKE; LAWRENCE; GILES, 2000; AGOSTI; PRETTO, 2005; COHN; CHANG, 2000; MIZZARO; ROBERTSON, 2007; AGICHTEIN; BRILL; DUMAIS, 2006; KAO et al., 2002).

O presente trabalho enfoca, contudo, os algoritmos de propagação e, em particular, o *HITS*.

Como é cediço, o modelo proposto por Kleinberg prevê apenas duas categorias de páginas, autoridades e *hubs*. Em 2005, fizemos uma primeira extensão do *HITS*, denominada de *Extended Hyperlink Induced Topic Search*(*XHITS*) (FILHO, 2005), que inseriu duas novas categorias de páginas *Web*, quais sejam, novidades e portais. Essas novas categorias foram inseridas de forma ponderada através de diversos parâmetros que passaram a integrar o modelo.

Esses parâmetros desempenharam um papel fundamental no *XHITS*, pois dependendo dos seus valores a contribuição de uma nova categoria é maior ou menor e, por consequência, a classificação das páginas se altera. Por isso, a compreensão do ajuste desses novos parâmetros se tornou um importante objeto de estudo.

Com o escopo de entender a capacidade do modelo *XHITS*, é utilizado, em (FILHO, 2005), um método de busca exaustiva para calibrar os parâmetros que investigavam uma região do espaço de solução definida entre 0 e 1 para cada parâmetro, com passo entre 0,01 e 0,5. Sabendo que para duas novas categorias de páginas cinco parâmetros foram inseridos, a complexidade do algoritmo de força bruta é $O(n^5)$.

Filho, Rentería e Milidiú (2009) propuseram o primeiro processo de aprendizagem para calibrar o *XHITS*, chamado de aprendizagem aproximada por gradiente descendente. Nessa abordagem, os parâmetros foram ajustados através da investigação dos seus valores pelo método do gradiente descendente. Porém, no processo de derivação algumas aproximações foram necessárias, abrindo espaço para mais investigações. Neste trabalho é proposta uma generalização do modelo que permite inserir k novas categorias de páginas e sua corretude.

Entretanto, Filho, Rentería e Milidiú (2011) publicaram uma nova abordagem de aprendizagem para o *XHITS*, dessa vez sem aproximações, o que demonstra ser mais promissora que a abordagem anterior e que será detalhadamente apresentada no presente trabalho.

Sem prejuízo, apresentamos também, nesta tese, alguns resultados experimentais utilizando a coleção *ClueWeb09* que indicam que a abordagem aqui proposta é promissora, colocando *XHITS*, após comparação com os resultados obtidos pelos participantes da *2010 Web TREC Track*, entre os seis primeiros lugares.

1.1 Objetivo

À luz do que foi exposto, vê-se que os objetivos deste trabalho são a generalização do modelo desenvolvido por Jon Kleinberg, aqui denominado *XHITS*, e a proposição de uma metodologia de aprendizagem para o modelo geral capaz de demonstrar as suas potencialidades.

1.2 Contribuições

Como contribuições da presente tese, podemos citar:

- O modelo *HITS* generalizado, *XHITS*, com a inserção de k novas categorias, permitindo que sejam extraídas mais informações das estruturas de hiperlink das páginas *Web*, melhorando as suas classificações. Também, a prova formal da sua corretude e convergência (FILHO; RENTERÍA; MILIDIÚ, 2009);
- Um processo de aprendizagem eficiente capaz de ajustar os pesos gerados

pelas inserções das novas k categorias (FILHO; RENTERÍA; MILIDIÚ, 2011), mostrando que o *XHITS* é uma proposta competitiva.

1.3 Organização do Trabalho

Após a contextualização do problema no presente capítulo, no capítulo 2 é feita um revisão dos trabalhos publicados considerados mais relevantes para essa tese. No capítulo seguinte, 3, são apresentadas a reformulação do modelo *HITS* e a sua generalização. Em seguida, no capítulo 4, é apresentado um método de aprendizagem para o modelo generalizado baseado em conceitos de decomposição em valores singulares (SVD). Segue-se, após, o capítulo 5 que trata dos experimentos realizados e, por fim, a conclusão proposta neste trabalho.