6
Conclusões e Trabalhos Futuros

O presente trabalho tem por objetivo analisar o modelo desenvolvido por *Jon Kleinberg*, o *HITS*, bem como os seus desdobramentos existentes na literatura e, ainda, desenvolver uma nova proposta capaz de proporcionar um melhor desempenho.

As principais contribuições deste trabalho são as seguintes:

- a proposta de um novo modelo estendido do *HITS*, o *XHITS (Extended Hyperlink Induced Topic Search)*;
- a implementação de uma ferramenta de experimentação que permite não só validar os algoritmos propostos, *HITS* e *XHITS*, mas também efetuar testes comparativos.entre os algoritmos;
- a construção de uma ferramenta que permite a implementação e a comparação de quaisquer evoluções dos modelos descritos na alínea anterior.

A próxima seção apresenta as conclusões do presente trabalho e a seção 6.2 traz sugestões para trabalhos futuros motivados a partir da pesquisa desenvolvida nesta dissertação.

6.1.
Conclusões

O foco deste trabalho consiste no estudo e desenvolvimento de modelos que permitam melhorar as classificações existentes nas máquinas de busca na *World Wide Web*. Tal motivação decorre da constatação de um clássico problema na *WWW*: a busca de informações relevantes. Sabemos que a qualidade da busca depende, necessariamente, de um fator intrinsecamente subjetivo: a relevância.

Em decorrência de tal fato, *Jon Kleinberg* observou que a criação de um

hyperlink na *WWW* representa uma indicação concreta de imputação de relevância. Assim, *Jon Kleinberg* apresentou um modelo capaz de buscar e classificar as páginas relevantes para um determinado assunto: *Hubs and Authorities*.

Com base nesse modelo, apresentamos uma extensão que adiciona novas considerações às definições dadas por *Jon Kleingerg* de *hubs* e autoridades: novidades e portais. Assim, formulamos um algoritmo estendido, o *XHITS (Extended Hyperlink Induced Topic Search),* para melhorar a classificação das autoridades do ambiente baseado em *hyperlinks*.

No capitulo 4, que tratou do ambiente de experimentação, vimos que podemos executar na ferramenta desenvolvida, separadamente, cada uma das etapas dos algoritmos *HITS* e *XHITS*, criando um sub-grafo direcionado e efetuando as classificações em cima desse sub-grafo. Vimos também, que o ambiente possui dois módulos implementados, estatística e calibração, para auxiliar na análise do desempenho dos algoritmos *HITS* e *XHITS*.

O módulo de calibração, que efetua a procura dos parâmetros do algoritmo *XHITS* que melhor se ajustam à classificação, subdivide o espaço de solução e o varre por busca exaustiva discretamente.

No módulo de estatística, a partir de uma classificação previamente fornecida, podemos acompanhar a evolução dessas páginas conforme a variação dos parâmetros do algoritmo que estiver em análise por intermédio dos gráficos.

A definição do *benchmark* engendrou um problema delicado, pois o julgamento da relevância das páginas classificadas pelos métodos implementados é subjetivo. Assim, algumas classificações foram consideradas como referência. Foram, pois, definidos dois especialistas para efetuar tais classificações de referência: *Yahoo* e *Google*.

A partir das classificações fornecidas pelas supracitadas máquinas de busca, foram efetuadas comparações de qualidade e, dentro do possível, análises humanas do conteúdo das páginas com melhor classificação.

Analisando os resultados, vimos que em cinqüenta por cento dos tópicos definidos no *benchmark*, o algoritmo *HITS* não conseguiu se aproximar em nada das classificações dos especialistas, resultando num índice de zero por cento em todos os intervalos. Nos demais cinqüenta por cento dos tópicos, o *HITS* conseguiu, no máximo, um índice de trinta por cento nas cinqüenta primeiras páginas do tópico *Harvard*, que possui um baixo índice de concordância.

Entretanto, o *XHITS* possui como seu menor índice dez por cento nas dez e vinte primeiras páginas no tópico *Harvard* que, como antes mencionado, possui um baixo índice de concordância entre os especialistas, qual seja, quarenta por cento. Nos demais intervalos, o *XHITS* possui índices de noventa por cento nas cinqüenta primeiras páginas do tópico Pelé e de sessenta por cento nas dez primeiras páginas do tópico fome.

Ainda sobre os resultados, podemos observar que as vinte primeiras páginas retornadas pelo algoritmo *XHITS*, mesmo que os especialistas tenham baixo índice de concordância, são de boa qualidade, conforme se depreende da classificação do tópico Rio de Janeiro. Nesse caso, conseguimos buscar as páginas oficiais do estado do Rio de Janeiro, da Prefeitura do Rio de Janeiro, do Museu de Arte Moderna do Rio de Janeiro, das duas principais universidades do estado (Universidade Estadual do Rio de Janeiro e Universidade Federal do Rio de Janeiro) e páginas de turismo com eventos e dicas sobre hotéis.

Com o objetivo aprender com a topologia dos grafos de consulta foi utilizado, como base do treinamento, o conjunto inicial de oito consultas que foram efetuadas as calibrações. O cerne do treinamento foi buscar uma configuração para os parâmetros do *XHITS* que refletisse um bom desempenho em consultas quaisquer, indicando a capacidade do modelo extrair mais informações da topologia dos grafos.

Em outras palavras, com o treinamento conseguimos buscar fatores aditivos que modificam e interferem na classificação das máquinas de busca comerciais, apenas atuando na topologia, fornecendo pesos diferentes para as novas relações definidas pelo *XHITS*.

A acurada análise dos testes feitos após o treinamento indica que o desempenho das soluções por histograma média e modal é superior ou, em

poucos casos, igual ao do *HITS*. Exemplificando, em quase metade das consultas de teste o *HITS* não apresenta nenhuma porcentagem de acerto. Tais fatos indicam que os pesos diferenciados para as relações foram determinantes para o bom desempenho alcançado em comparação ao *HITS*, que não possui nenhum parâmetro de ajuste, e que o treinamento foi eficaz na definição dos parâmetros das soluções.

As avaliações realizadas e seus resultados demonstram a importância do modelo estendido proposto, o *XHITS*, que foi capaz de proporcionar índices de relevância bem superiores aos apresentados pelo *HITS*. Mais ainda, avaliando superficialmente as classificações dos tópicos, identificamos uma boa qualidade de resultados, o que reforça a sua relevância.

6.2. Trabalhos Futuros

Para trabalhos futuros, podemos mencionar o aprimoramento da inteligência no módulo de calibração, de forma que este seja capaz de aprender e inferir valores melhores para os parâmetros do *XHITS*. Como visto nos resultados experimentais, um ajuste preliminar dos parâmetros do *XHITS* gerou resultados satisfatórios.

Tal fato indica boas perspectivas no estudo de melhores heurísticas de treinamento sobre as matrizes de adjacência. Isto para aprender mais sobre o comportamento da topologia e encontrar os padrões que minimizem o erro e produzam os melhores parâmetros para o *XHITS*.

Outro ponto importante é a inserção dos novos conceitos no modelo de *Jon Kleinberg*. Podemos sugerir uma generalização do modelo estendido, *XHITS*, incorporando outros conceitos, tais como novidades e portais, ampliando a influência dos fatores nas ordenações.

Em decorrência, haverá um aumento na flexibilidade dos ajustes. Conforme exposto no decorrer deste trabalho, a complexidade do algoritmo para percorrer o espaço de solução é $O(n^5)$, e a potência de n aumenta de acordo com o número de parâmetros utilizados. Em conjunto com a complexidade,

incumbe avaliar até que ponto a inserção de um novo parâmetro contribui para a ordenação e de que forma essa influi. Em termos de proposta futura:

- estudar o impacto computacional versus o benefício na ordenação;
- estudar e demonstrar a existência um limite cuja inserção de um novo conceito não seja perceptível na ordenação.

A abordagem do *HITS* e do *XHITS* é baseada, principalmente, na topologia, ao passo que uma terceira proposta de abordagem será baseada na inserção do texto das páginas no cálculo dos parâmetros. Entretanto, uma forma cautelosa de integração e avaliação do conteúdo textual das páginas deve ser observada, uma vez que o núcleo inicial dos dois modelos é formado por páginas que já foram classificadas textualmente. Sendo assim, o estudo da proximidade das ordenações puramente textuais com a do *XHITS*, conforme inoculado os pesos textuais, será de grande valia.

Por fim, procurar parcerias com as entidades que possuam bases de dados da *WWW* já classificadas por sua relevância, de forma a aumentar a qualidade dos resultados. O dinamismo da *WWW* acaba gerando distorções nos resultados finais, visto que os grafos são gerados num determinado instante e todo o treinamento pode não reproduzir mais a realidade num momento posterior. Sendo assim faz-se mister uma base estática e devidamente classificada para consolidar os resultados obtidos.