A natureza distribuída das informações disponíveis na Internet levou à busca constante de maneiras eficientes de executar consultas sobre uma grande coleção de documentos. Os motores de busca para Web realizam essa tarefa, porém, devido ao grande tamanho da coleção, o número de documentos relevantes para uma consulta pode facilmente ter milhares de itens. Devido a essa quantidade de informação, os algoritmos utilizados na área de recuperação de informação tradicional não obtiveram resultados satisfatórios, visto que realizavam uma análise sobre o conteúdo das páginas para montar um ranking. A estrutura de links comum nos documentos HTML fornece uma maneira de estimar quais são as páginas mais "populares", partindo-se da premissa de que quanto maior o número de links que apontam para uma certa página (links esses vindos também de páginas populares), mais popular essa página é.

Os algoritmos que levam em conta essa estrutura de links na montagem do ranking dos resultados de uma pesquisa são chamados de algoritmos de análise de links. O algoritmo de PageRank [Page, Brin, Motwani and Winograd 1998] é o que mais tem destaque nesta família de algoritmos. O termo PageRank é comumente atribuído a um número que diz o quanto uma página Web é relevante na Internet, ou seja, quanto maior esse número, maior será a relevância da página.

No entanto, a velocidade em que a estrutura da Internet se modifica acaba por trazer alguns problemas ao algoritmo de PageRank. Novas versões de uma página Web com alto grau de relevância não irão ter o mesmo grau de relevância da página original, mesmo se tratando de versões de um mesmo documento. Por exemplo, considerando uma página Web p, um fator importante para constituir o ranking pelos motores de busca é a quantidade de outras páginas que apontam para p. Contudo, novas versões desta página terão uma baixa taxa de links que apontam para elas (links entrantes), pois as páginas que apontam para p não sabem da existência de suas novas versões. Nesse contexto, a detecção de versões pode melhorar o posicionamento no ranking de versões baseado nos links que apontam para p.

Neste contexto, o objetivo deste trabalho é especificar uma proposta de um mecanismo para detecção automática de versões de páginas Web, visando melhorar o cálculo do PageRank. Pretende-se fazer com que o grau de relevância de uma página seja computado, considerando o escore de suas versões antigas, mantendo assim, o grau de importância que a página obteve ao longo do tempo. Em outras palavras, isto significa que uma nova versão da página ou a alteração de sua localização não causa prejuízo ao seu posicionamento no ranking de resultados do motor de busca. Para validar a proposta foram realizados experimentos com o motor de busca de código aberto Nutch [Nutch 2007]. A principal contribuição desse trabalho é apresentar uma proposta para o uso de detecção de versões de páginas Web para melhoria do desempenho do algoritmo de PageRank.

O texto deste trabalho está organizado da seguinte forma: na seção 2 é apresentada a base conceitual para a compreensão do funcionamento dos motores de busca para Web. Na seção 3, são apresentados os trabalhos relacionados. Na seção 4, são apresentados os experimentos realizados com o Nutch e uma proposta de detecção de versões de páginas Web para a melhoria de performance do algoritmo de ranking utilizado pelo Nutch. Por fim, a seção 5 traz as conclusões do trabalho e trabalhos futuros.