1
Introdução

A Internet é uma fonte de informações que cresce a cada dia. Uma "presença" na Web, através de sites, notícias e outros documentos, é imprescindível para qualquer empresa moderna. Sites sociais, como Orkut e Facebook, possuem milhões de usuários, cada um deles com a sua própria página. Todo tipo de pessoa, de jornalistas e escritores a donas de casa, adotam os blogs como uma ferramenta de comunicação com o resto do mundo. Wikis, páginas cujo conteúdo é construído de forma colaborativa, como a Wikipedia (www.wikipedia.org), vem se transformando em referências confiáveis cada vez mais utilizadas.

É fundamental para um usuário encontrar rapidamente informações na web através de consultas simples por palavras-chave. Esta funcionalidade é provida hoje pelas máquinas de busca (*search engines*) como Google (www.google.com), Yahoo! (www.yahoo.com) e MSN (www.msn.com). Através delas, qualquer um pode encontrar todo o tipo de informação sobre os mais diversos temas, mesmo sem conhecer nenhum site específico sobre o assunto pesquisado. Elas vêm se tornando, portanto, a porta de entrada dos usuários para a Internet.

Por trás da interface simples apresentada por estas máquinas de busca existe uma vasta arquitetura tecnológica que permite a coleta, processamento e indexação de bilhões de páginas, sites e documentos. Através desta indexação, as páginas são ranqueadas de acordo com a sua relevância para diferentes termos de pesquisa. Durante uma pesquisa, páginas relevantes são recuperadas e um rankeamento dinâmico é aplicado sobre elas, gerando uma ordem final na qual os resultados são apresentados para o usuário.

Durante estes dois processos, de coleta e de exibição de resultados, a tarefa de classificação de conteúdo das páginas se torna importante. Para a exibição de resultados, ter o conteúdo pré-classificado permite o agrupamento dos documentos em conjuntos lógicos de exibição, abrindo novas possibilidades de visualização das informações e facilitando o acesso à informação.

Durante a coleta, a classificação das diferentes páginas permite uma série de abordagens especializadas. No caso de uma máquina de busca específica para blogs, como o Technorati (www.technorati.com), ela permite a distinção do que são ou não blogs, segregando quais documentos devem ser retornados aos usuários. Outra estrutura possível é o armazenamento e indexação apenas das páginas classificadas como blogs (ou outra classe qualquer), reduzindo requisitos de hardware e, conseqüentemente, custos operacionais destas empresas.

Ainda durante o processo de coleta, a classificação do conteúdo dos documentos da Web permite a construção de *crawlers* "focados", que busquem por um assunto específico ou por um tipo de página pré-determinado.

Por último, a classificação permite a adoção de estratégias de visitação por parte dos *crawlers* diferenciadas por tipos de páginas. Blogs e wikis têm um potencial de mudança grande, ou seja, tendem a mudar significativamente em espaços curtos de tempo. Por outro lado, páginas governamentais mudam com uma freqüência muito menor. Como a utilização do conteúdo mais recente é fundamental para um bom rankeamento, um *crawler* pode ser configurado para visitar com maior freqüência páginas de um tipo com propensão maior a mudança, conforme identificadas pela classificação.

Dada a sua grande importância, não é nenhuma surpresa que a classificação de páginas web seja um tema tão abordado em trabalhos de pesquisa e artigos. Existe uma variedade enorme de problemas de classificação, desde a segmentação de sites em classes distintas (educativos, comerciais, etc.) até a classificação do "sentimento" transmitido por uma página (positivo, negativo, indiferente, etc.).

O objetivo deste trabalho é realizar uma investigação de diferentes abordagens e métodos para um problema específico de classificação. O problema em questão é a classificação funcional em 4 classes distintas (*blogs*, *blog posts*, *notícias* e *portais de notícias*) de diferentes documentos HTML.

Estas 4 classes foram escolhidas porque sua identificação permite uma série de aplicações diretas. Uma empresa que deseja coletar na Web páginas que se refiram aos seus produtos, por exemplo, pode utilizar a classificação em blog posts ou notícias para diferenciar as páginas que possuem matérias jornalísticas de páginas que apresentam opiniões pessoais, utilizando cada uma para fins diferentes. Um *crawler* focado em coletar notícias, para um serviço de agregação, por exemplo, pode ter como semente inicial de busca um portal de notícias e, ao

longo da navegação, se deparar com blogs, blog posts, outros portais, notícias, e outros tipos de páginas. No entanto, para a agregação, apenas as notícias seriam processadas e indexadas.

Ao longo dos experimentos realizados, foram testados diferentes conjuntos de atributos, tanto estruturais quanto textuais, para a classificação, assim como métodos de classificação distintos: árvores de decisão, redes neurais, e *Support Vector Machines* (SVMs).

Através dos experimentos realizados, foi possível observar que os melhores resultados foram obtidos com a combinação de atributos estruturais e atributos textuais. A utilização apenas de atributos estruturais se mostrou competitiva com a abordagem híbrida, apresentando a vantagem de ser independente de idioma. Dentre os três métodos utilizados, as árvores de decisão se mostraram o método de classificação mais estável, ou seja, com menor variação nos resultados encontrados para os diferentes conjuntos de atributos.

Este trabalho está divido em 5 capítulos além desta introdução. No segundo capítulo, são apresentados conceitos gerais sobre documentos da Web, seus atributos e características, que serão utilizados ao longo do resto do texto, e alguns trabalhos relacionados. No terceiro, são apresentadas as classes selecionadas para a tarefa de classificação e suas características principais. No quarto é realizada uma exposição sobre o método de coleta de informações utilizado, dificuldades e pontos importantes encontrados ao longo do processo, e ainda sobre a extração dos diferentes conjuntos de atributos. No quinto capítulo, os diferentes experimentos realizados são descritos, seguidos de breves sumários dos resultados obtidos. Finalmente, no capítulo 6, são apresentadas as principais conclusões e resultados obtidos, as limitações implícitas do estudo realizado e direções futuras de trabalho.