2

Conceitos Gerais de Classificação de Documentos na Web

2.1.

Páginas, Sites e Outras Terminologias

É importante distinguir uma página de um site. Uma *página web*, ou simplesmente *página*, é um único documento HTML, que pode ter links externos ou não, que possui uma URL e uma série de características estruturais definidas dentro de suas *tags* HTML, e que pode conter texto e código (scripts). Páginas raramente contêm elementos multimídia, como imagens ou vídeos, embutidos; geralmente, são mantidas referências para os mesmos, carregadas por navegadores de forma transparente para o usuário. Toda página pertence a um site.

Um *web-site*, ou *site*, é um conjunto não vazio de páginas web. No caso de possuir mais de uma página, estas podem estar interligadas através de *hiperlinks* (referências externas navegáveis). Um site possui um domínio (".com", ".com.br", ".co.uk", ".gov", ".org", etc.), e uma página de entrada, exibida quando o endereço do site é digitado em um navegador. Os documentos dentro de um site podem ser estáticos, gerados por uma pessoa responsável pelo conteúdo (no caso de uma matéria de jornal ou um post em um blog), ou dinâmicos, gerados automaticamente por programas de computador (como as páginas de respostas de consultas em máquinas de busca).

A alteração do conteúdo de uma página implica na alteração do código HTML da mesma, através de ferramentas específicas. A alteração de um site pode ocorrer através da alteração do conteúdo de suas páginas ou da alteração do conjunto de páginas que o compõe.

O termo "documento" é utilizado na literatura para se referir tanto a páginas quanto a sites, o que pode causar confusão. Neste trabalho, portanto, será utilizado apenas quando sua definição estiver evidente pelo contexto. O termo "material" é utilizado sempre para se referir ao conjunto de páginas e sites.

2.2. A Evolução de Documentos na Web

O dinamismo da Web é uma de suas características fundamentais. Novo conteúdo é constantemente adicionado, sob a forma de arquivos, páginas e sites. Conteúdo existente é alterado, movido de lugar, ou, em raros casos, desaparece totalmente.

Diversos estudos foram realizados para analisar esse caráter evolutivo do conteúdo. Em [14], por exemplo, foi feito um estudo em larga escala que analisou, ao longo de 11 semanas, 150 milhões de páginas distintas. A grande maioria sofreu pelo menos uma modificação, e uma boa parte dessas sofreu um número ainda maior de alterações.

Foi reportado ainda que a frequência de atualização depende do site ao qual as páginas pertencem. As que mais se alteram ao longo do tempo pertencem a sites com domínio “.com” (comerciais), e as que menos se alteram são as governamentais (domínio “.gov”) e de instituições de ensino (domínios “.edu” nos EUA e “.br” no Brasil, por exemplo).

O aparecimento de ferramentas que permitem a atualização de conteúdo sem a necessidade de conhecimentos técnicos aprofundados acelerou este processo evolutivo. Alguns blogs, por exemplo, são atualizados múltiplas vezes por dia, com informações novas sobre assuntos completamente distintos. A adoção destas ferramentas por um número cada vez maior e mais diverso de usuários – de pais solteiros a grandes corporações, praticamente todos tem um blog hoje, e toda empresa possui um wiki interno para o gerenciamento de conhecimento – significa que a velocidade de evolução só tende a aumentar.

Para um usuário da rede que está em busca de informações específicas, a proliferação de conteúdo é, ao mesmo tempo, uma vantagem e uma desvantagem. Se por um lado a quantidade de informação disponível é cada vez maior, por outro, encontrar a informação correta e relevante se torna cada vez mais difícil.

Ao realizar uma busca, este usuário dificilmente procura uma página específica. Segundo um estudo dos logs de máquinas de busca especializadas em blogs [12], usuários quase sempre realizam buscas por algum assunto específico, ou por opiniões de pessoas específicas. De uma forma geral, a maioria das consultas realizadas a máquinas de busca são “informativas”, ou seja, estão atrás

de informações, independente do site ou página onde esta esteja localizada, ao invés de endereços conhecidos.

Desta forma, a evolução constante deve ser levada em consideração em toda e qualquer tarefa de classificação na Web. Determinar a relevância de uma página se torna uma tarefa especialmente complexa, uma vez que conteúdo adicionado em um curto espaço de tempo pode alterar completamente o nível de interesse de um usuário por uma determinada página. Assim, os esquemas de atualização de índices têm de se tornar mais complexos do que o esquema original adotado pelo Google, que revisitava e reindexava todas as páginas em seu índice a cada 30 dias [25].

2.3. Ranqueamento

O ranqueamento de páginas é a funcionalidade central de qualquer máquina de busca, seja na Web ou em qualquer outro ambiente. Dada a natureza não estruturada das consultas, é praticamente impossível alcançar um resultado único e correto. Assim, o ranqueamento das respostas possíveis é fundamental para que o usuário possa encontrar de forma mais fácil a informação desejada.

Todas as máquinas de busca atuais, como Google (www.google.com), Yahoo! (www.yahoo.com) e MSN (www.msn.com), utilizam funções de ranqueamento, tanto estático quanto dinâmico, na ordenação dos resultados. Assim, quanto melhor esse ranqueamento, mais significativos os resultados retornados. A diferença nas páginas retornadas por diferentes máquinas de busca para uma mesma consulta é causada, em parte, pelo método de ranqueamento utilizado.

Apesar de sua evidente importância, as funções de ranqueamento não foram a primeira tecnologia utilizada para métodos de busca na Web. Originalmente, páginas e sites eram classificados manualmente por exércitos de pessoas que avaliavam o seu conteúdo e construíam diretórios massivos. Mesmo nessa época, no entanto, já era percebido que a velocidade de crescimento da Web impossibilitaria a classificação manual de seu conteúdo[10].

Da necessidade de um método automático de classificação de resultados surgiram os primeiros algoritmos de ranqueamento, como o PageRank [11]. Estes algoritmos calculam um ranking para cada página, baseado na sua relevância para

determinadas palavras-chave. O que os diferencia é como esta relevância é calculada.

Existe uma proximidade entre a classificação de documentos da Web e o ranqueamento dos mesmos. Ambos dependem da análise ou de conteúdo ou estrutural dos documentos, e podem utilizar tanto atributos locais quanto de vizinhança. Em alguns casos ainda, o ranqueamento pode fazer uso da classificação para aprimorar seus resultados. Pode-se ainda pensar no ranqueamento como uma espécie de classificação “binária” de uma página, como sendo relevante ou não, à qual é aplicada uma medida de probabilidade de relevância.

A evolução constante do conteúdo afeta diretamente o ranqueamento. A dificuldade de ranqueamento de conteúdo dinâmico se torna ainda mais expressiva sobre subconjuntos de páginas que possuem características estruturais ou topológicas próprias. A *blogosfera* (o universo de todos os blogs), por exemplo, é muito mais interconectada do que a Web como um todo, embora muitas de suas páginas compartilhem de características estruturais e temáticas. Assim, bons métodos de ranqueamento de páginas Web como um todo podem ter dificuldades com ranqueamentos específicos.

2.4. Classificação na Internet

As tarefas de classificação de material da Internet podem ser decompostas ao longo de diversos eixos, relacionados com os tipos de atributos utilizados, quais as classes consideradas, a utilização apenas da página individual ou de sua vizinhança, e outros mais. A seguir, serão apresentados os eixos de segmentação mais importantes.

2.4.1. Alta Granularidade vs. Baixa Granularidade

Segundo Lindemann [1, 2], existem dois tipos distintos de classificação de material presente na Internet, sejam páginas individuais ou sites como um todo. Ao primeiro, relacionado com a utilização de classes funcionais previamente determinadas, ele dá o nome de classificação de “baixa-granularidade” (*coarse-grained classification*).

Exemplos dessa classificação são apresentados em [1] e [2], onde os autores separam sites da Web nas seguintes classes pré-definidas: sites corporativos, acadêmicos, de compras e blogs. Ela sempre exige uma etapa inicial onde é definido, de forma essencialmente arbitrária, o conjunto de classes a ser utilizado.

Esta classificação é útil na melhor apresentação de resultados de máquinas de busca, pois permite o agrupamento de documentos em classes funcionais. Um usuário que faz uma pesquisa pelo termo “basquetebol,” por exemplo, poderia visualizar os resultados segregados em lojas online, sites de federações, sites de times, sites de escolinhas, e assim por diante, facilitando o seu acesso a informação desejada.

Ela também é bastante útil na construção e operação de *crawlers*. Permite, por exemplo, o armazenamento e indexação de uma classe pré-determinada (*crawlers* que indexam apenas notícias ou blogs) ou a utilização de tratamento diferenciado para a revisitação de páginas.

O segundo tipo de classificação, relacionado com a identificação de assuntos e temas abordados dentro dos documentos, foi denominado por Lindemann de classificação de “alta-granularidade” (*fine-grained classification*).

Este tipo permite, por exemplo, a busca por todos os documentos relacionados com um determinado tema. Ele se difere do ranqueamento por palavras-chave na medida em que inclui um elemento semântico na classificação. Em uma busca pela palavra “cachorro,” por exemplo, um ranqueamento simples traria todas as páginas que possuem a palavra “cachorro,” ordenadas por sua relevância. A classificação de alta granularidade traria todas as páginas relacionadas com cachorros, independente da presença do termo em si. Páginas que mencionam raças específicas ou animais de estimação em geral também seriam retornadas. Outra possibilidade que surge com esta classificação é a construção automatizada de diretórios de pesquisa segregados por tema, facilitando o acesso a informações específicas.

Uma vez que o número de classes é gigantesco, e para cada uma é necessário um dicionário de termos relacionados, a classificação de alta-granularidade é mais complexa do que a de baixa-granularidade. O texto das páginas é fundamental para esta classificação, uma vez que é nele que se encontra o tema. Isso significa que a classificação de alta-granularidade é bastante sensível ao idioma das páginas.

A grande maioria dos estudos realizados hoje na Web está relacionada com a classificação de alta granularidade. Conforme descrito acima, esta tarefa é de extrema utilidade para auxiliar usuários em suas buscas. Desta forma, diversas abordagens já foram propostas para este problema [13].

Os termos alta- e baixa-granularidade podem ser confusos. Segundo o pensamento de Lindemann, o número de classes que tratam de temas específicos sempre será maior do que o número de classes funcionais dentro das quais podemos agrupar os documentos. Por mais refinadas que sejam as classes funcionais determinadas, sempre é possível alcançar um número maior de classes temáticas. Portanto, baixa-granularidade se refere à classificação funcional, enquanto alta-granularidade se refere à classificação temática.

Por outro lado, é necessário considerar que os temas em si sempre podem ser refinados. Pode-se definir o tema esportes, mas dentro de cada esporte existem uma série de modalidades, cada uma com sua linguagem específica. Dentro de cada modalidade, existem diferentes categorias, e assim por diante. Sendo assim, por mais alta a granularidade definida, é sempre possível se chegar a uma granularidade maior, o que torna os termos confusos.

Uma nomenclatura melhor, que será adotada ao longo deste trabalho, é a que utiliza a definição dos tipos de classificação determinados por Lindemann, ou seja, classificação funcional, ou em classes funcionais (para o que ele chama de baixa-granularidade) e classificação temática, ou em temas (para o que ele chama de alta-granularidade).

2.4.2. Contexto vs. Conteúdo

Attardi, Gulli e Sebastiani [10] foram os primeiros a explicitar o contraponto inerente entre a “classificação por conteúdo” e a “classificação por contexto.” O contexto de um documento na Web é o conjunto de documentos que apontam para ele. Assim, enquanto na classificação por conteúdo utilizamos a informação do próprio documento, na por contexto utilizamos informações presentes no contexto do documento.

Tanto o algoritmo PageRank quanto o algoritmo proposto em [10], por exemplo, utilizam a classificação por contexto para realizar o ranqueamento. Já o algoritmo apresentado em [20] é um exemplo de classificação de páginas por

conteúdo. A classificação por contexto pode ser meramente topológica ou similar a realizada por Fürnkranz [22], fazendo uso do conteúdo dos vizinhos.

A distinção entre classificação por conteúdo ou por contexto é independente da distinção de classificação em temas ou em classes funcionais. Documentos podem ser separados em classes funcionais através de seu conteúdo, de seu contexto, ou de ambos, e o mesmo vale para a classificação em temas.

2.4.3. Atributos Estruturais vs. Atributos de Texto

Outra distinção possível na classificação está nas características dos atributos utilizados. Ela pode ser realizada apenas com o texto dos documentos, apenas com a estrutura, ou através de combinações dos dois.

Em [19], Cohen sugere que o que diferencia a busca e classificação de documentos HTML da de documentos tradicionais é a estrutura inerente dos primeiros. Assim, a utilização de elementos desta estrutura é fundamental para a melhor classificação. De fato, em [18], o próprio Cohen apresenta melhorias de até 50% na performance de classificadores através da utilização de atributos estruturais dos documentos.

Novamente, a distinção do tipo de atributos é independente do tipo de classificação. Em [23], por exemplo, é realizado um trabalho de classificação funcional, de contexto (utiliza informações de conectividade entre diferentes sites e páginas) e que utiliza principalmente atributos estruturais. Por outro lado, [20] é um trabalho de classificação funcional binária (blog ou não), de conteúdo (utiliza informações da própria página, ignorando o contexto), e que utiliza somente atributos estruturais.

Grande parte da informação estrutural de uma página está contida dentro de tags HTML específicos, como “*meta*” ou “*title*.” Diversos algoritmos de classificação e ranqueamento de páginas utilizam informações contidas nessas tags. Ainda assim, segundo [21], a grande maioria das páginas existentes na rede em 2001 não continha nenhuma informação nestas. Com a difusão do entendimento de como os algoritmos de ranqueamento funcionam e das informações que podem ser colocadas para melhorar os resultados dentro de máquinas de busca, esta situação tem se revertido.

2.4.4. Estado da Arte

Embora poucos trabalhos lidem diretamente com a classificação funcional de páginas web, em especial das classes utilizadas neste trabalho, é possível encontrar na literatura trabalhos similares, que podem ser utilizados como balizadores dos resultados apresentados aqui.

A classificação temática de páginas apresenta grande variação nos resultados obtidos, devido a diversidade de abordagens e temas possíveis. Alguns trabalhos de classificação de tema de blog posts apresentam acurácia de 80%, e classificadores tópicos de páginas em geral chegam a apresentar acurácia de 91%. Por outro lado, alguns classificadores temáticos apresentam acurácia inferior a 20% [24].

A classificação funcional de páginas, exploradas em uma série de trabalhos como Fürnkranz[17] e de Rijke[20], apresenta acurácia variando de 80% a 95%. No entanto, todos estes trabalhos consideram a classificação binária (em blogs e não-blogs, por exemplo), o que simplifica o trabalho de classificação. Os trabalhos com melhores resultados utilizam uma combinação de atributos estruturais com atributos construídos manualmente, como a página estar contida ou não dentro de um determinado domínio.