4
Coleta de Informações, Extração de Atributos e Métodos de Classificação

4.1.
Coleta de Informações

O processo de coleta de informações foi realizado em três estágios. No primeiro, que durou 4 meses, foi realizada a coleta de 2000 documentos HTML, 500 de cada uma das classes descritas no capítulo 3.

O prazo de 4 meses de coleta foi fundamental para evitar que alguns termos ganhassem importância indevida durante o processo de classificação. Durante uma extração inicial de atributos de texto, realizada sobre 500 documentos coletados ao longo de 1 semana, foi possível observar que termos relacionados com os principais acontecimentos da semana ("olimpíadas", "guerra" – referente a guerra entre Rússia e Geórgia – "eleições", dentre outros) apareciam em uma quantidade significativa (acima de 60%) das notícias.

Embora fosse possível realizar a filtragem destes termos através de listas específicas, não existe nenhuma garantia de que uma lista criada arbitrariamente removeria todos os termos com uma importância artificial. Assim, dado o tempo disponível para o trabalho, tomou-se a decisão de prolongar o processo de coleta, aproveitando a própria natureza temporal da informação para evitar que termos ganhassem importância indevida.

Este corpus inicial foi dividido aleatoriamente em dois conjuntos de páginas, denominados "**Treino**" e "**Teste1**". O conjunto "Treino" possui 983 páginas, e o conjunto "Teste1" possui 1017 páginas. O primeiro foi utilizado, conforme o nome indica, para realizar o treinamento dos classificadores nos diferentes experimentos, enquanto o segundo foi utilizado para testes.

Em seguida, foi realizada, ao longo de uma semana, a coleta de um corpus adicional de 200 documentos HTML, 100 da classe blogs e 100 da classe blog posts. Neste corpus, os blog posts foram escolhidos especificamente por

possuírem muitos comentários, de forma a permitir o estudo da complexidade de diferenciação entre as duas classes representadas no caso de posts com número grande de comentários.

Intuitivamente, o problema se torna mais difícil tanto estrutural quanto textualmente, uma vez que os comentários podem incluir links, representam partes de texto adicionais, aumentam o tamanho do texto como um todo, e não seguem o estilo de escrita que encontramos nos posts em si. Este corpus deu origem ao conjunto de testes rotulado de "**Posts+Comments**".

Por último, foi utilizado um corpus de teste previamente coletado e classificado. Este corpus possui algumas características específicas que o diferenciam dos outros conjuntos de páginas coletados. Primeiro, ele possui páginas em diversas línguas, o que impossibilitou a extração imediata de atributos de texto, mas permite um estudo das diferenças estruturais entre páginas em diferentes línguas. Outra característica importante é que neste conjunto, os blog posts foram novamente selecionados por possuírem muitos comentários, de forma a dificultar a classificação.

Este corpus deu origem a dois conjuntos de testes distintos e disjuntos. O primeiro, rotulado de "**Teste2_Gr**", corresponde ao conjunto de páginas do corpus que são de qualquer língua menos a inglesa. O segundo, rotulado de "**Teste2_Ing**", corresponde ao conjunto de páginas em inglês dentro deste corpus. O conjunto "Teste2_Gr" possui 318 instâncias, e o "Teste2_Ing" 259. A quantidade de blog posts nos dois conjuntos ficou desbalanceada. O primeiro possui poucos posts em relação ao resto das páginas, enquanto o segundo possui muitos.

A seguir, é apresentada uma tabela com a distribuição das páginas por classe para cada um dos conjuntos.

	Blogs	Blog Posts	Notícias	Portais de Notícias
Treino	247	241	250	245
Teste1	253	259	250	255
Posts+Comments	100	100	0	0
Teste2_Gr	87	45	95	91
Teste2_Ing	57	110	42	50

Tabela 1 – Quantidade de páginas por tipo para cada um dos diferentes conjuntos de informações utilizados no decorrer dos experimentos. Tanto o conjunto "Teste2_Ing"

quanto o “Teste2_Gr” apresentam quantidades desproporcionais de posts, o que pode impactar os resultados obtidos.

Os conjuntos “Treino”, “Teste1”, “Posts+Comments” e “Teste2_Ing” possuem apenas páginas de língua inglesa. Os documentos de todos os conjuntos foram armazenados em formato texto com a codificação ASCII. Os documentos foram coletados manualmente através da visita a página, visualização do código fonte, e armazenamento do mesmo em um disco rígido. Ao código fonte HTML de cada página foi adicionado uma *tag* especial “URL”, dentro da qual está contida a URL original da página coletada. Todas as páginas foram classificadas manualmente durante sua coleta.

O processo de coleta dos blogs e dos portais de notícias foi realizado com o uso de máquinas de busca e sites agregadores já existentes. Para a coleta de blogs, foram utilizados o Google Blog Search (http://blogsearch.google.com/) e o Technorati (http://technorati.com/). Para a coleta de portais de notícias, foram utilizados o Google (www.google.com), MSN (www.msn.com), Yahoo! (www.yahoo.com) e Newspapers.com (www.newspapers.com).

Alguns problemas interessantes se apresentaram durante a coleta de informações. Uma série de sites hospedeiros de blogs oferecem ferramentas de escrita e publicação de conteúdo próprias. Essas ferramentas causam a padronização estrutural das páginas. Para reduzir o potencial impacto sobre os classificadores desenvolvidos, durante o processo de coleta evitou-se armazenar muitas páginas que houvessem sido criadas com as mesmas ferramentas (limitando a 100 o número de páginas por ferramenta).

Algo similar ocorre com a classe de portais de notícias. Algumas empresas oferecem pacotes de construção de sites jornalísticos, através de uma plataforma de software especial. Muitos portais utilizam a mesma plataforma de software, resultando em padronização. Assim, o mesmo cuidado foi tomado na coleta de portais, para evitar a concentração de páginas similares dentro do corpus, e, por conseqüência, a distorção dos resultados.

A coleta de blog posts e notícias foi realizada de forma mais simples. Para cada blog selecionado, todos os blog posts (mínimo de 2 e máximo de 10) disponíveis na página foram acessados e coletados. A mesma estratégia foi

adotada para notícias, selecionando-se artigos de portais de notícias. Desta forma, é possível garantir uma diversidade de documentos nestas classes e, ao mesmo tempo, uma semelhança estrutural entre eles e suas fontes. Assim, pode ser realizado um estudo dos atributos que diferenciam blogs dos seus blog posts, e os portais de suas notícias.

4.2. Conjuntos de Atributos e Processos de Extração

Uma das partes integrantes das estratégias de classificação é o conjunto de atributos utilizados. Ao longo deste trabalho, portanto, foram utilizados uma série de diferentes conjuntos de atributos, derivados dos conjuntos básicos apresentados a seguir.

4.2.1. Atributos Estruturais

O primeiro conjunto básico de atributos utilizados no trabalho foi um conjunto de atributos estruturais extraído das páginas coletadas.

Os atributos estruturais estão listados na tabela a seguir.

	Atributos Estruturais
1	URL Length
2	URL Depth
3	Tag Count
4	Average Tag Depth
5	Meta Tag Count
6	Anchor Tag Count
7	Link Tag Count
8	Style Tag Count
9	Script Tag Count
10	Has ATOM Feed
11	Has RSS Feed
12	Uses CSS
13	Image Tag Count
14	Total Text Pieces
15	Total Text Length
16	Average Text Length
17	Total Anchor Text Length
18	Average Anchor Text Length
19	Uses External Components
20	Comment Count
21	Total Script Pieces
22	Total Script Length
23	Average Script Length

Tabela 2 – Atributos Estruturais extraídos dos documentos web

O atributo 2, de profundidade da URL, foi medido contando-se o número de caracteres "/" dentro da URL do documento armazenado, descartando-se a parte do protocolo (o "http://"). Assim, um documento com a URL *http://www.msn.com*, por exemplo, tem profundidade zero, enquanto um documento como *http://www.msn.com/money* tem profundidade um, e assim por diante. De uma forma intuitiva, espera-se que blogs e news portals apresentem sempre uma profundidade inferior aos blog posts e news, uma vez que esses últimos estão contidos dentro dos primeiros.

Os atributos 3 e 4 estão diretamente relacionados com a árvore DOM do documento HTML sendo processado. O atributo 3 mede o número de nós dentro desta árvore, enquanto que o 4 mede a profundidade média destes nós. Os atributos 5, 6, 7, 8, 9 e 13 são medidas diretas de tipos de tags específicas dentro do código HTML do documento.

Os atributos 10, 11 e 12 estão diretamente relacionados com o tipo de tecnologia utilizada na geração das páginas sendo analisadas. Sua extração foi realizada através da análise das tags contidas dentro do documento. *Feeds* RSS e ATOM são tecnologias de distribuição de conteúdo para programas de leitura especializados, que permitem o recebimento de informações da web sem a necessidade de se visitar diretamente a página de interesse. A presença ou não de um feed RSS ou ATOM dentro de uma página pode ser extraída através da interpretação de tags de link presentes dentro da página de interesse. Se a página possuir algum link apontando para um feed RSS ou ATOM, ela é marcada como possuindo um feed. A utilização de CSS por uma página pode ser extraída diretamente das tags de estilo (style) do código HTML. A motivação por trás da extração destes atributos é que páginas como blogs e blog posts tendem a fazer uso mais pesado de tecnologias como RSS e ATOM para a distribuição de seu conteúdo, uma vez que não dependem de visitas para garantir rentabilidade.

Os atributos 14, 15 e 16 estão relacionados com os blocos de texto presentes em um documento HTML. O atributo 14 informa o número de blocos de texto distintos existentes dentro de um documento, o 15 informa o comprimento total destes blocos quando condensados, e o atributo 16 mostra a média do comprimento dos blocos. No caso destes atributos, a idéia é de que um blog tenha diversos blocos de texto de um tamanho intermediário, enquanto um blog post

apenas um, ou alguns poucos, blocos de texto mais curtos. Da mesma forma, espera-se que portais de notícias apresentem dezenas de blocos de texto extremamente curtos, enquanto uma notícia específica apresente um único bloco contínuo de texto.

Os atributos 17 e 18 são similares ao 15 e 16, mas lidam apenas com blocos de texto de âncora, ou seja, o texto contido dentro das tags de âncora (<a>). Intuitivamente espera-se que blogs e blog posts apresentem textos de âncora mais descritivos e longos do que os textos de âncora encontrados em veículos de informação tradicionais.

Por fim, os atributos 19, 20, 21, 22 e 23 estão relacionados com a utilização de *javascript* dentro do documento HTML sendo analisado. O atributo 19 indica se o javascript sendo utilizado encontra-se dentro do próprio código-fonte da página ou em algum arquivo externo. A expectativa é de que news portals e news, por pertencerem a organizações maiores, tendem a ser mais estruturadas e modularizadas, utilizando arquivos externos. O atributo 20 mede o número de comentários dentro do código javascript contido dentro da página. Ele só será superior a zero quando o atributo 19 for verdadeiro. Já os atributos 21, 22 e 23 realizam uma métrica similar a apresentada anteriormente para blocos de texto, mas agora para o código javascript presente na página. Novamente, a expectativa é de que news portals e news sejam páginas mais sofisticadas, utilizando assim uma quantidade maior de javascript.

A extração destes atributos estruturais foi realizada através de um parser HTML desenvolvido por terceiros (www.yunqa.de), integrado com um programa que realizou os tratamentos especiais e contagens necessárias.

4.2.2. Atributos de Texto

O segundo conjunto básico de atributos utilizados para a classificação foram atributos de texto, especificamente as diferentes palavras dentro do texto das páginas coletadas. As páginas foram representadas através de uma estrutura de "bag-of-words" [29].

Em [27], os autores realizam uma comparação de diferentes medidas de tratamento de atributos de texto, com o objetivo de verificar qual delas seria a

mais bem sucedida, e avaliar a eficiência de cada uma com relação ao seu tempo de cálculo. Nesse trabalho, eles chegam a conclusão de que a medida ***D****ocument* ***F****requency* (DF) é tão boa quanto qualquer outra para a extração das palavras mais importantes dentro de um documento, e possui ainda um dos menores tempos de cálculo.

Considerando estes resultados, a medida de DF foi utilizada para realizar a seleção das palavras representativas de uma classe. Com o objetivo de se comparar conjuntos diferentes de atributos de texto, foram utilizados dois níveis diferentes de DF das palavras para sua seleção como atributos. No primeiro conjunto de atributos de texto, foram utilizados como atributos todos os termos que apareciam em pelo menos 80% dos documentos de uma mesma classe. No segundo conjunto, foram utilizados os termos que apareciam em pelo menos 50% dos documentos. Todas as seleções foram realizadas em cima do conjunto "Treino".

Essa abordagem, evoluída a partir das sugestões de [27], permite a seleção das palavras que são atributos mais significativos para cada uma das diferentes classes selecionadas, e possui a vantagem de ser facilmente expandida para acomodar novas classes. Por outro lado, conforme o conjunto de documentos de treino é alterado, a seleção de palavras significativas precisa ser reconstruída. No entanto, isso é verdade para qualquer estratégia de seleção de palavras como atributos.

A partir dos 4 conjuntos de palavras mais significativas, um dicionário global foi construído, e os documentos foram transformados em vetores numéricos, onde cada posição é a freqüência daquela palavra dentro do documento.

Blogs e blog posts não seguem nenhum padrão formal de escrita. Pelo contrário, a linguagem utilizada dentro deles é em grande parte informal e direta, e utiliza uma série de termos e formas de endereçamento que não são utilizadas em veículos de comunicação tradicionais. Palavras como "eu", "você", "nós" e outras são freqüentes dentro de blogs e posts, mas quase não aparecem em notícias. Ao mesmo tempo, quando um processo de remoção de stopwords tradicional é utilizado, essas palavras tendem a ser removidas, uma vez que não agregam nada em termos de classificação do conteúdo das páginas. Como o objetivo deste

trabalho não é a classificação de conteúdo, mas sim a classificação funcional das páginas, não foi utilizado nenhum processo de remoção de stopwords.

O dicionário advindo do conjunto de termos que aparecem em pelo menos 80% das páginas de cada classe possui 78 termos, enquanto que o dicionário resultante da análise dos termos que aparecem em pelo menos 50% possui 307 termos. Assim, o primeiro conjunto possui 78 atributos, e o segundo possui 307. A utilização de um dicionário de termos significa que os classificadores baseados em atributos de texto são diretamente dependentes da língua das páginas, o que representa uma desvantagem marcada com relação à utilização de atributos estruturais, que é independente de linguagem.

Pode ser possível remover esta dependência de linguagem com um conjunto suficientemente grande de páginas, onde houvessem páginas representando um grande número de línguas diferentes. No entanto, isso implicaria também em um enorme conjunto de atributos para classificação, o que seria inviável.

Após a extração, os arquivos com os atributos foram convertidos para o formato ARFF, permitindo a sua utilização na ferramenta WEKA (www.cs.waikato.ac.nz/ml/weka/).