ANÁLISE ESTRUTURAL PARA CLASSIFICAÇÃO DE PÁGINAS WEB

Aluno: Iam Jabour
Orientador: Raúl Rentería e Eduardo Laber

Introdução

A *World Wide Web*, ou Web, cresceu nos últimos anos, tornando-se uma importante fonte para análises e pesquisas de opinião, assim como para relações e interações entre diversos grupos [1]. Esse crescimento é observado tanto em seu conteúdo, quanto no número de pessoas acessando-a.

Junto ao crescimento da Web, as ferramentas de análise de conteúdo ganham destaque, pois possibilitam uma visão quantitativa de um vasto domínio de informações, porém é necessária a extração do conteúdo relevante das páginas da Web, para que essa análise possa ser realizada de forma correta.

A extração possibilita a indexação da informação exata, evitando erros, e fornece mais percepção nas análises realizadas. Para a extração é necessária a criação de uma base orientada, que torna possível a aplicação de algoritmos específicos para o conjunto um páginas com características semelhantes, fato o qual aumenta a precisão da extração do conteúdo relevante de forma correta.

A capacidade de agrupar páginas da Web de acordo com os seus objetivos é associada ao problema de classificação e proporciona a base necessária para a evolução de modelos de extração mais precisos. Tal tarefa tem diversas abordagens, como a de Elgersma e Rijke [2], que classifica blogs utilizando informações textuais e elementos como links externos, nome do domínio, dentre outras. Ou ainda, utilizando a análise da estrutura de *hiperlinks* dentre o conjunto de páginas do sítio dando como resultado sua classe, não de apenas uma página, como Amitay et al [3]. É interessante notar, que em alguns estudos, a união de informações extraídas a partir das tags e da análise textuais apresenta "resultados interessantes".

Neste trabalho, buscou-se uma solução onde apenas informações topológicas da página são analisadas, denominada de classificação estrutural, criando independência à análise textual e apresentando um novo ferramental para o problema de classificação de páginas da Web. Também é adotada a análise de páginas individualmente, ao invés da análise do sítio. Essa abordagem barateia o custo da análise, por não ser necessário percorrer diversos documentos para a classificação, e fornece mais informação para as ferramentas que utilizam o resultado dessa classificação, como a de extrair o conteúdo relevante da página.

Trabalhos Relacionados

O problema de classificação de páginas da Web, também conhecido como categorização de páginas Web, é normalmente abordado como um problema de classificação supervisionado [4], onde um conjunto de dados previamente rotulados é utilizado para o treino e teste do modelo de classificação. O problema de classificação pode ser definido como a tarefa de encontrar uma função F, tal que, dado um conjunto de parâmetros x, extraídos a partir de um objeto, F(x) é capaz de retornar à classe do objeto [5].

Figura 1: Classificação é a tarefa de mapear um conjunto de atributos de entrada (x) em uma classe (y).

A função (ou modelo) de classificação pode ser criada de forma a classificar somente uma classe, definindo se um objeto pertence ou não a classe. Pode também classificar múltiplas classes, dando como resposta a classe a qual o documento pertence dentre as classes que ela sabe classificar. A primeira forma é chamada de classificação binária, enquanto que a segunda pode receber o nome de classificação nominal ou ordinal [5].

Existem subgrupos do problema de classificação de páginas da Web que determinam um escopo para o significado da classe, sendo eles: classificação por assunto, classificação por funcionalidade, classificação por sentimento e outros tipos. A classificação por assunto é baseada no título da página, por exemplo, se uma página aborda assunto como esporte, arte ou economia. A classificação por funcionalidade é focada no que a página traz ou o que ela faz, por exemplo, uma página pessoal, de negócios, de notícias ou de vendas. A classificação por sentimento é direcionada para a opinião que a página traz, por exemplo, o que um autor pensa sobre determinado assunto. Exemplos de outros tipos de abordagens podem ser encontrados em [6], que utiliza uma classificação por gênero ou [7], que utiliza um *search engine spam classification*. Neste trabalho, é adotada a classificação por funcionalidade.

Dentre os estudos que utilizam a abordagem da classificação funcional, é possível observar diversas maneiras de se obter atributos para a classificação. Uma observação importante é que existem atributos que podem ser extraídos a partir da análise de um único documento HTML, uma única página, enquanto que outros dependem de análise do sítio.

Elgersma e Rijke [2] utilizam palavras freqüentes para ajudar na classificação de páginas. A formação da URL, assim como, a quantidade de links externos e internos e a existência de imagens são exemplos de informações obtidas analisando-se o conjunto de tags, também utilizadas por ele. Tais atributos são comumente utilizados em estudos da área.

Dentre os atributos que podem ser obtidos a partir da análise de um sítio, os mais utilizados são os que surgem a partir da exploração topológica de links, ligações entre documentos, e da quantidade de páginas no sítio, assim como seu tipo (ASP, PHP, PDF, HTML, DOC). O trabalho de Lindemann e Littig [8] é um bom exemplo de estudo que utiliza essas informações para a classificação de sítios.

É importante ressaltar que a análise textual está presente na maioria dos estudos que abordam a análise de um único documento. A análise textual cria uma dependência adicional na classificação, e essa dependência torna a classificação diretamente ligada ao idioma, à qualidade do texto, à confiança nas palavras e à não existência de spam dentro do texto. Dessa forma, este trabalho apresenta uma análise voltada para as informações topológicas de um único documento, buscando uma independência da análise textual.

Classes e Conjunto de Aferição

Cada sítio da Web possui uma funcionalidade ou utilidade específica, que é recorrente. Sítios com mesmo objetivo apresentam uma familiaridade em sua forma de apresentação e em sua aparência estética, proporcionando aos usuários fácil identificação de seu objetivo. Para exemplificar, alguns tipos de sítios podem ser citados como: sítios de empresas, que tem a finalidade de apresentar seus produtos e suas informações, sítios que trazem notícias na expectativa de informar seu usuário/leitor, sítios que tem o objetivo de vender algum produto e sítios que apresentam textos pessoais. A partir dessas funcionalidades podemos dar origem

ao que chamamos de classes (ou classes funcionais) que ajudam a organizar o conteúdo da Web. Muitas vezes, um sítio pode atender a múltiplas classes, esse recebe o nome de portal.

Um sítio é uma coleção de páginas oriundas de um mesmo domínio, o que torna difícil a identificação de sua classes, e motiva a classificação de páginas individualmente. A abordagem de classificar apenas uma página tende a ser mais precisa e fornece mais informação para a organização e estruturação da Web, criando facilidades para os mecanismos de busca. Ao observar apenas uma página é possível focar em padrões que fornecem "pistas" que facilitam a identificação das características que definem uma classe.

Dentre um universo de classes existentes na Web, foram selecionadas duas para direcionar os estudos de forma clara. A escolha da classe notícia e portal de notícias não é por acaso, já que a classe notícia é importante para a extração de informação da Web e um portal de notícias remete à classe notícia.

Para a criação de um modelo de classificação mais robusto foi introduzida uma última classe chamada de outro, que traz diversas outras classes que não pertencem às duas classes abordadas. As classes serão explanadas na próxima seção.

Classe Notícia

A classe notícia traz funcionalidades já conhecidas e suas características são evidentes na maioria dos casos, porém, algumas páginas dessa classe apresentam características diferentes, dificultando seu reconhecimento. Essa dificuldade dá-se, muitas vezes, pela definição ampla do significado da palavra notícia na Língua Portuguesa.

Para evitar tal dificuldade, é necessário criar uma especificação exata do que essa palavra significa na definição da classe. Por esse motivo, a seguir são fornecidas as principais características da classe notícia abordada neste trabalho:

- título da notícia;
- texto centralizado dentre os demais conteúdos do sítio;
- autor, opcional;
- data, opcional.

Classe Portal de Notícias

A classe portal de notícias tem por objetivo apresentar fragmentos de notícias, similar à primeira página de um jornal impresso. Como explicado anteriormente, a classe portal é destinada a páginas que agrupam outras classes. Logo, definir uma classe chamada portal de notícias significa que o principal objetivo desse portal é apresentar notícias. Porém, por definição, essa página também traz outras informações, como propaganda, menus, comentários, que podem apresentar grande desafio durante a criação de um modelo para sua classificação.

As características que melhor descrevem os documentos pertencentes à classe portal de notícias são:

- vários títulos de notícias ;
- fragmentos de notícias, com hiperlink levando a notícia completa;

Classe Outro

A classe outro traz qualquer página que não seja da classe notícia ou portal de notícias. Essa classe é uma espécie de simulação do mundo real para o ambiente de treino, pois tenta generalizar o conjunto de páginas que podem ser encontrados dentro do universo de treino, tentando dificultar a criação de um modelo de classificação, deixando-o mais robusto.

Conjunto de Aferição

Para os experimentos foram coletadas 100 páginas de cada uma das classes abordadas e 100 páginas da classe outros, resultando no total de 300 páginas. Essas páginas foram armazenadas com o nome "index.html", no formato HTML. Para cada página armazenada foi criado um documento XML, chamado de gabarito que contém as características da página. O conjunto de aferição, chamado de corpus, totaliza 89Mb de dados.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<description>
   <lang>língua do documento</lang>
   <class>classe do documento</class>
   <url>endereço http do documento</url>
</description>
```

Para a criação e evolução do modelo de classificação foram utilizados 80% do conjunto de aferição, 240 documentos, sendo exatamente 80 documentos de cada uma das três classes. Para a validação do modelo apresentado foram reservados 20 documentos de cada classe, totalizando 60 documentos. Esses documentos foram separados no início dos experimentos, não sendo utilizados em nenhum momento durante a evolução do modelo de classificação.

Estrutura e Informação

Utilizando *Document Object Model* (DOM) [9], modelo apresentado pela W3C[1] para a manipulação de um documento HTML ou XML em estrutura de árvore, é possível analisar esses documentos de forma estrutural. Neste trabalho, o modelo DOM é utilizado para a análise e extração de atributos apresentados nas seções subseqüentes.

Em uma árvore DOM, são encontrados diversos tipos de nós. Para o presente trabalho, os tipos *Element* e *Text* se destacam. Toda tag HTML é representada por um nó do tipo *Element*, com exceção daquelas que possuem um tipo especial como comentários e scripts. Já todo o texto existente no documento é representado por um nó do tipo *Text*. Para exemplificar, em "<a> texto do link " o texto "texto do link" será representado por um nó do tipo *Text*, que será filho do nó do tipo *Element* que representa a tag "a".

As análises apresentadas a seguir utilizam os nós dos dois tipos apresentados, ignorando os outros tipos de nós que podem ser encontrados na árvore DOM. É importante salientar que os outros tipos de nós, muitas vezes, não adicionam informações ao documento, pois são nós que apresentam informações como comentários, que são representados pelo tipo *Comment*; códigos javascript, que são representados pelo tipo *CDATASection*; atributos de uma tag, que são representados pelo tipo *Attr*. Ignorar esses tipos de nós não prejudica a análise, apenas a simplifica.

Para facilitar o entendimento dos estudos apresentados, são adotadas as seguintes nomenclaturas:

- documento - página da Web no formato HTML;
- árvore do documento - árvore DOM do documento;
- nó de texto - nós do tipo *Text*, que possuem pelo menos um caractere.

A partir de uma árvore do documento é possível fazer as seguintes análises:

- sobre os tipos de tags utilizadas;
- sobre a utilização de tags seguidas, ou seja o alinhamento de tags;
- sobre o posicionamento dos nós de texto dentro da árvore do documento;

[1] W3C - World Wide Web Consortium - http://www.w3.org

- sobre o balanceamento da árvore do documento, de acordo com a quantidade de caracteres em cada subárvore;
- sobre o tamanho da árvore, ou seja, sua altura e sua quantidade de nós;
- sobre o posicionamento de tags específicas dentro da árvore do documento;

Com essa análise inicial, é possível agrupar os esforços para encontrar padrões que possibilitam a distinção das classes, isto é, sua classificação, em três áreas de estudos principais, sendo elas:

- estudo de nós: busca o entendimento do comportamento dos nós da árvore do documento, sua distribuição, dentre outros. Detalhado na seção ***Estudo de Nós***;
- estudo de tags: busca o entendimento do comportamento das tags, entende-se tipo (nome) da tag, e sua utilização. Detalhado na seção ***Estudo de Tags***;
- estudo de caracteres: busca o entendimento de como os caracteres são dispostos dentro da árvore do documento, em que proporção eles aparecem e qual sua distribuição. Detalhado na seção ***Estudo de Caracteres***.

Dados como média aritmética, variância, desvio padrão e coeficiente de variação são utilizados e constituem a base de informações para análise. Após uma análise individual, esses dados estatísticos são utilizados para a evolução das hipóteses. Essas proporcionam a geração de atributos que são utilizados por modelos clássicos de classificação, com *Support Vector Machine* (SVM). Para validar esses atributos obtidos a partir das linhas de estudo, são realizados experimentos que proporcionam uma visão dos resultados da classificação. Esses resultados podem ser observados na seção ***Experimentos***.

Estudo de Nós

Toda árvore do documento apresenta uma estrutura única, que é influenciada por sua formação. Essa estrutura é reflexo de sua programação, pois durante a criação do documento o programador segue padrões, mesmo que pessoais, que facilitam o entendimento e a manutenção de seus documentos.

O estudo de nós assume que existem padrões estruturais semelhantes para uma mesma funcionalidade, e tenta identificá-los. Essa semelhança é proposta, pois a disposição de nós de um documento HTML interfere em sua forma de apresentação/visualização, assim como o grupo que cria esses documentos interagem entre si, o que homogeneíza os padrões utilizados.

A existência de tags distintas, que são utilizadas com o mesmo objetivo, é eminente. Logo, inicialmente é difícil utilizar uma análise que se baseia apenas nas tags. O estudo dos nós busca eliminar essa variável, simplificando a análise.

É importante frisar que todos os dados fornecidos nessa seção são extraídos dos documentos separados para o treino, vide seção ***Separação do Conjunto de Teste e Treino***. Quando uma metodologia diferente de análise for utilizada, ela será explicitamente descrita.

Dando início ao estudo dos nós, a primeira informação que pode ser observada é a quantidade total de nós e a altura máxima da árvore. Encontram-se os seguintes valores para cada classe:

- classe notícia - média do total de 507 nós, com desvio padrão de 219 e média da altura máxima de 22, com desvio padrão 8;
- classe portal de notícia - média total de 906 nós, com desvio padrão de 819 e média da altura máxima de 23.5, com desvio padrão 8.7;
- classe outro - média do total de 681 nós, com desvio padrão de 916 e média da altura máxima de 17.3, com desvio padrão 7.6.

A tabela 1 apresenta a quantidade de nós entre a altura 40 e 45. Esse dado é apresentado, pois deixa claro que a partir da altura 40 existem poucos nós, pois estamos analisando a média de nós em 80 documentos, e que não existindo mais informação forte para a análise.

Tabela 1: Quantidade de nós na faixa de altura 40 a 45

	Notícia		Portal de Notícias		Outro	
Altura	Quantidade	Média	Quantidade	Média	Quantidade	Média
40	72	0.9	5	0.06	2	0.03
41	0	0	11	0.14	5	0.06
42	0	0	11	0.14	3	0.04
43	0	0	10	0.13	1	0.01
44	0	0	9	0.11	1	0.01
45	0	0	8	0.1	6	0.08

Seguindo a distribuição de nós por altura, é possível observar diferentes estruturas topológicas dentre as classes, fato que motiva o estudo de nós. Nota-se que o número de tags por altura fornece uma característica interessante dos documentos. A figura 2 proporciona a observação das seguintes características, que reforças as hipóteses levantadas:

- até a altura 7 o crescimento de tags da classe notícia e portal de notícias é bem semelhante, já a classe outros apresenta um comportamento diferente, tendo um crescimento muito elevado nesses primeiros níveis;
- a partir da altura 10, até a altura 30, existe uma distinção significativa entre as classes notícias e portal de notícias. No entanto, o desvio padrão elevado desses dados, como o da altura 10 na classe notícia, que chega a 39 de desvio com média 35, ou da altura 12 na classe portal de notícias, que chega a 60 de desvio com média 49.9, torna essa informação menos favorável para a separação dessas classes.

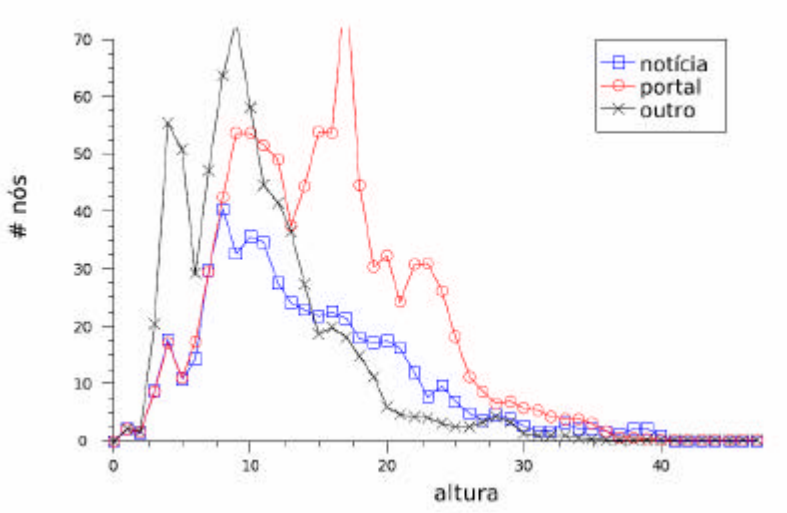

Figura 2: Estudo de nós, média de nós por altura

O crescimento de nós na classe outros não é uniforme, já que essa classe, conforme mencionado na seção \ref{sec: Classes Abordadas}, contém documentos com diversas funcionalidades. Esse crescimento não uniforme cria dificuldades para a classificação.

Continuando a análise de nós por altura é gerada uma nova informação, onde a quantidade de nós em cada nível é dividida pela quantidade total de nós na árvore do documento, chamado de nós proporcionais. A partir da análise dos nós proporcionais, é possível obter a distribuição do crescimento de nós por altura nas classes. A figura 3 mostra como todas as classes apresentam a mesma forma de crescimento, tendo uma pequena

diferença na quantidade de nós de cada nível. É importante fazer observações sobre informações que fornecem marcas nessas estruturas como:

- a classe outros apresenta uma queda acentuada no número de nós por altura, enquanto que, a classe notícia e portal de notícias apresentam um fator de queda menor.
- todas as classes agrupam grande quantidade de nós entre as alturas 8 e 13 e o pico na quantidade de nós proporcionais é encontrado antes da altura 10.

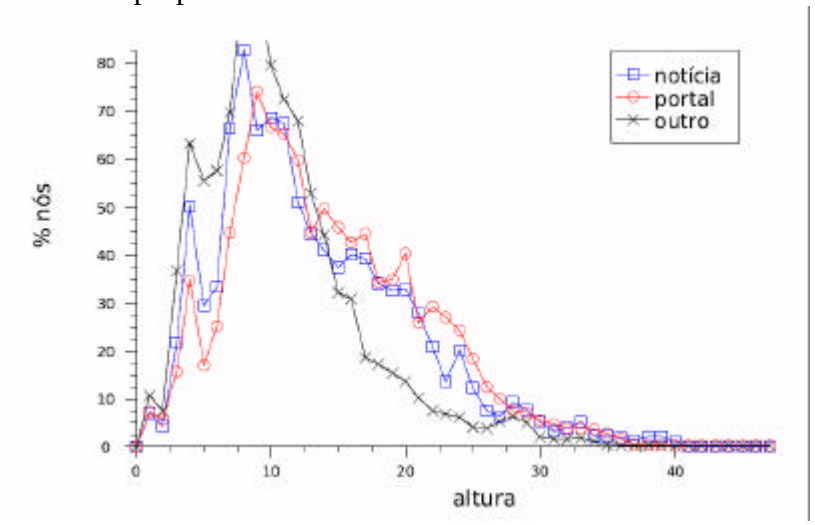

Figura 3: Estudo de nós, média proporcional de nós por altura

Como os dados analisados são obtidos a partir da média dos documentos, sempre é importante acompanhar o desvio padrão. Na tabela 2, pode ser observado que o desvio padrão, muitas vezes, fica acima da média. Tal fato significa que existem documentos onde a quantidade de nós proporcional é muito diferente da média, o que diminui a confiança dessa informação.

Tabela 2: Estudo de nós, desvio padrão da quantidade proporcional de nós na faixa de altura 8 a 11

	Notícia		Portal de Notícias		Outro	
Altura	Média	DP	Média	DP	Média	DP
8	82.43	77.66	60.11	66.03	97.82	110.03
9	65.81	61.18	74.01	76.88	96.99	109.53
10	68.54	67.85	66.39	65.30	79.28	93.51
11	67.38	58.8	65.15	57.84	72.32	83.87

Mesmo com o desvio padrão elevado em grande parte do estudo, é possível afirmar que a distribuição dos nós na árvore do documento fornece informações importantes para a classificação, e são apresentados experimentos com atributos gerados a partir do estudo de nós na seção ***Experimento do Estudo de Nós***.

Estudo de Tags

A tag na linguagem HTML é uma informação importante, por ser o elemento de marcação da linguagem. Estudar a estrutura de um documento HTML e não dedicar esforços na compreensão do uso das tags pode ser encarado como um equívoco.

Como explicado na seção ***Estudo de Nós***, existem diversas tags disponíveis em HTML que podem ser utilizadas para a mesma finalidade. Porém, é razoável esperar que dentro de um grupo de documentos, que atendam a uma mesma funcionalidade, as tags utilizadas tenham alguma relação, assim como, que exista um conjunto de tags básicas que dificilmente são substituídas ou não são utilizadas.

Apresentamos os dados que podem ser obtidos a partir da observação do uso das tags para a classificação dos documentos seguindo a mesma metodologia da seção ***Estudo de Nós***, onde apresentamos estatísticas extraídas dos documentos separados para o treino.

Para uma primeira análise, as dez tags mais utilizadas para cada uma das classes são apresentadas na tabela 3. É interessante perceber que, o conjunto de tags que ocupam as dez primeiras posições no *rank* de utilização são semelhantes, mesmo essas tags não ocupando exatamente a mesma posição.

É importante notar que a média de utilização das dez primeiras tags é distinta entre as classes, pois isso fornecer a geração de um modelo de classificação. No entanto, o desvio padrão aponta uma diferença entre documentos da mesma classe, o que diminui a expectativa sobre a qualidade da classificação, ao utilizar somente essa informação.

Tabela 3: Estudo de tags, as 10 tags mais utilizadas

Notícia			Portal de Notícias			Outro		
Tag	Média	DP	Tag	Média	DP	Tag	Média	DP
a	81.55	54.48	td	140.64	172.82	a	114.98	142.25
td	74.09	71.86	a	138.4	140.8	td	75.14	180.51
tr	47.96	49.36	tr	103.31	157.51	br	68.94	165.91
div	37.91	36.66	div	67.45	66.5	div	50.31	72.81
img	33.49	26.55	img	58.71	50.4	img	46.76	77.85
li	29.55	47.19	br	47.71	53.62	tr	38.31	88.89
br	25.99	25.1	font	41.11	142.63	li	31.68	52.68
script	17.48	11.8	strong	36.41	256.52	p	28.2	77.44
table	17.28	20.13	span	31.81	39.69	span	24.24	44.26
tbody	17.08	20.13	table	29.43	27.11	strong	21.14	116.59

Seguindo a mesma linha do estudo de nós, são apresentados dados de utilização de tags de forma proporcional ao total de tags no documento. Isso torna possível identificar tags que são utilizadas de forma igual entre as classes, o que as destaca como sendo importantes para a estrutura dos documentos. Na tabela 4, são apresentados os valores para as dez tags mais utilizadas.

Tabela 4: Estudo de tags, as 10 tags mais utilizadas proporcionalmente ao numero de tags no documento

Notícia			Portal de Notícias			Outro		
Tag	Média	DP	Tag	Média	DP	Tag	Média	DP
a	155.32	66.3	a	149.8	61.08	a	157.74	89.79
td	133.47	102.23	td	144.8	103.03	td	113.62	114.04
tr	85.56	70.44	tr	102.35	80.54	br	82.57	99.51
div	80.52	70.65	div	88.17	75.92	div	78.15	56.88
img	63.84	36.82	img	68.87	45.76	img	70.55	68.5
li	54.77	75.71	br	58.02	54.88	tr	61.28	67.01
br	53.89	49.48	span	37.89	45.96	li	51.45	65.59
script	38.46	28.58	table	34.99	29.93	p	37.68	63.73
table	30.57	27.15	li	34.83	56.23	span	32.09	45.7
tbody	30.17	27	tbody	34.74	29.8	font	31.09	79.05

A análise das informações, que geram as tabelas 3 e 4, traz a compreensão da necessidade de uma nova informação estatística. Utilizando a média de utilização das tags, proporcional à quantidade de nós no documentos, e o desvio padrão é possível obter o coeficiente de variação, onde o desvio padrão é dividido pela média. Esse dado proporciona a identificação das tags que serão utilizadas na evolução desse estudo, de forma que, as seguintes regras foram utilizadas para separar as tags consideradas importantes:

- ter coeficiente de variação abaixo de 0.8 em pelo menos uma classe. O valor 0.8 foi escolhido por fornecer um universo de tags pequeno, tornando a evolução da análise mais clara.

- ter média de utilização maior que 2 em pelo menos uma classe, tornando necessário uma utilização mínima da tag.

As tags, junto ao seu coeficiente de variação, que atendem essas regras são apresentadas na tabela 5. Essas tags são analisadas a cada nível da árvore do documento e fornecem uma forte descrição da estrutura de cada classe. Essa análise gera uma grande quantidade de dados, o que dificulta sua apresentação em tabelas ou gráficos. Por esse motivo, na seção ***Experimento do Estudo de Tags*** são apresentados apenas os resultados da utilização dos atributos encontrados a partir dessa análise.

Tabela 5: Estudo de tags, coeficiente de variação (CV) das 10 tags mais utilizadas

Notícia		Portal de Notícias		Outro	
Tag	CV	Tag	CV	Tag	CV
a	0.42	a	0.41	a	0.57
div	0.87	div	0.85	div	0.72
form	0.61	form	0.57	form	0.95
img	0.57	img	0.65	img	0.96
script	0.72	script	0.79	script	0.98
td	0.76	td	0.71	td	0.99
tr	0.81	tr	0.78	tr	1.08

Para completar o estudo de tags, são apresentadas também algumas informações sobre os nós da árvore do documento que não são tags. É importante utilizar esses nós, pois não utilizá-los despreza informações importantes que podem ser obtidas a partir da diferenciação de nós de tags dos outros tipos de nós.

Os nós que não são do tipo *Element*, são contabilizados em dois outros tipos. Os nós de texto, do tipo *Text*, são contabilizados com o nome "NodeText". Os demais são contabilizados com o nome "noElement".

Esses dois tipos de nós contabilizados são adicionados às estatísticas obtidas a partir de todas as técnicas apresentadas nessa seção, como se esses tipos fossem tags específicas.

Após analisar os resultados dos experimentos utilizando os atributos que podem ser extraídos a partir das tags, nota-se que essa informação é de extrema relevância e que a capacidade de classificar páginas HTML com a análise estrutural, cada vez mais, apresenta resultados motivadores. Esses resultados podem ser observados na seção ***Experimento do Estudo de Tags***.

Estudo de Caracteres

Para finalizar a linha de análise da estrutura da árvore do documento, o estudo de caracteres busca encontrar e entender o comportamento na utilização dos textos dentro do documento HTML. A hipótese levantada nessa linha é que existem padrões na forma de utilização dos caracteres dentre os documentos que atendem a mesma funcionalidade.

Como apresentado na seção ***Classes Abordadas***, os principais documentos estudados são direcionados à apresentação de textos. Por isso, a utilização do texto do documento para se obter características, que proporcionam a classificação, pode ser feita de diversas formas como:

- busca de palavras chave;
- interpretação do texto, envolvendo a identificação do sujeito e contexto do texto.

Essas formas, apresentadas acima, seguem uma metodologia onde a análise textual é necessária, o que implica em modelos específicos para cada idioma, ou variação do mesmo. Esse fato, torna a extração de atributos onerosa, pois todas as técnicas necessitam interagir sobre o texto do sítio.

Um ponto interessante a ser levantado é que, as ferramentas que utilizam a análise textual precisam identificar qual o texto relevante do documento, pois utilizar todo o texto existente em um documento pode criar deturpações na análise e ocasionar uma grande quantidade de erros. A classificação do documento, normalmente, é um passo anterior à análise textual, pois proporciona que algoritmos específicos sejam utilizados para detectar e interagir sobre o texto do documento, diminuindo a quantidade de erro. Por esses motivos, métodos de análise textual não são utilizados no estudo de caracteres.

A classificação estrutural apresenta novas formas de se obter informações do texto de um documento. A seguir apresentamos a seqüência na evolução da análise dessas informações utilizando os nós do tipo *Text*, que representam os textos na árvore do documento.

Inicialmente são apresentadas informações para análise iguais àquelas já fornecidas no estudo de nós e no estudo de tags. Essas informações foram extraídas do conjunto de documentos separados para o treinamento.

Analisando a quantidade de caracteres existentes nos documentos, é possível apresentar os seguintes valores para as classes:

- classe notícia - média de caracteres de 4032, com desvio padrão de 1685;
- classe portal de notícias - média de caracteres de 5685.9, com desvio padrão de 8400;
- classe outro - média de caracteres de 7922, com desvio padrão de 20652.

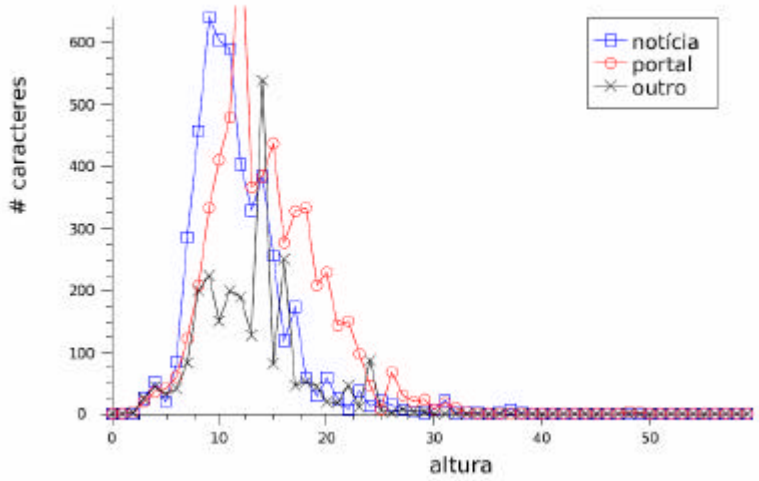

Figura 4: Estudo de caracteres, média de caracteres por altura

Essa informação, de forma bruta, fornece poucas características imediatas, mas é importante para que possamos fazer algumas análises mais adiante.

É importante observar também que, a quantidade de caracteres em cada nível da árvore do documento, ilustrada na figura 4, apresenta uma distribuição interessante onde:

- na classe notícia pode ser encontrada uma quantidade elevadas de caracteres, acima de 300 caracteres, entre as alturas 6 e 16. Já na classe portal de notícias, essa faixa, que contêm a maior quantidade de caracteres, se estende entre as alturas 10 e 18;
- na classe outro pode ser observada uma concentração pequena de caracteres na maioria das alturas;
- a queda na quantidade de caracteres por altura apresenta características diferentes dentre as classes, o que é um fator importante.

A tabela 6 traz o desvio padrão das informações apresentadas na figura 4. É importante reparar os valores elevados do desvio padrão, pois isso prejudica os modelos de classificação encontrados a partir dessa informação.

Tabela 6: Estudo de caracteres, média e desvido padrão (DP) da quantidade de caracteres da altura 10 a 20

	Notícia		Portal de Notícias		Outro	
Altura	Média	DP	Média	DP	Média	DP
10	601.53	2082.81	408.61	651.92	147.7	238.6
11	588.94	1198.2	477.59	701.25	196.96	361.37
12	401.2	841.53	828.58	2985.12	187.66	409.93
13	328.06	713.14	364.15	856.87	125.5	409.47
14	383.06	779.45	381.8	737.68	537.66	4213.61
15	253.55	740.96	435.78	912.08	79.41	283.6
16	117.3	297.96	274.7	507.59	249.49	1725.74
17	170.64	555.58	325.7	1463.5	44.63	143.15
18	56.28	140.27	332.29	1142.44	49.85	254.81
19	27.93	112.2	207.39	490.36	43.1	197.39
20	57.64	220.07	228.15	906.82	17.85	79.88

A quantidade total de caracteres do documento, apresentada na figura 5, é encontrada de forma proporcional à quantidade total de caracteres no documento. Nesses dados, é interessante notar a anormalidade na distribuição dos caracteres proporcionais, fato que motiva o melhor entendimento da distribuição de caracteres dentro de cada classe. Essa nova informação ajuda na busca de atributos eficientes que podem ser utilizados para a classificação dos documentos. O desvio padrão elevado, sendo para muitas alturas dois terços maior que a média, chama a atenção e pode ser observado na tabela 6.

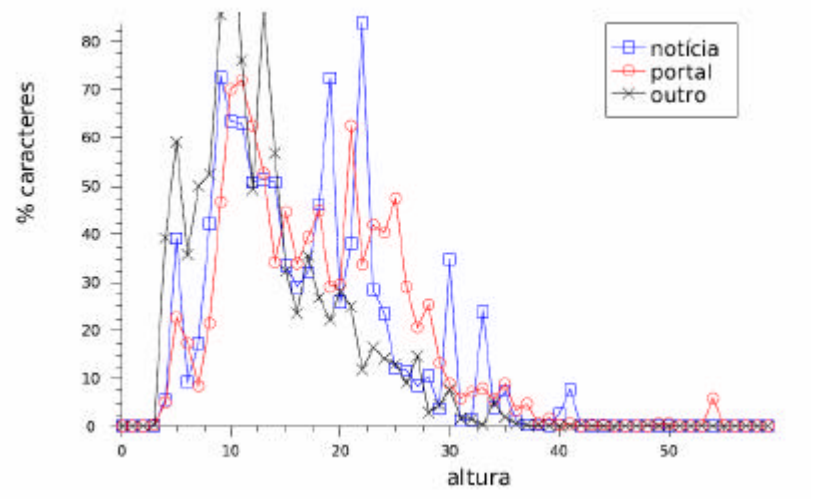

Figura 5: Estudo de caracteres, média proporcional de caracteres por altura

Para entender a variação encontrada na figura 5, é necessário observar a estrutura da árvore do documento de maneira individual, deixando de apenas utilizar informação estatística. Para isso, é utilizado a linguagem *graphML*[2], que utiliza XML para criar de forma declarativa estruturas de grafos. Para a visualização das estruturas declaradas na linguagem *graphML* é utilizado o programa Yed[3], que gera uma saída visual das estruturas declaradas de forma rápida e clara.

As árvores dos documentos foram descritas em documentos graphML de forma que, cada nó do tipo *Element* do documento DOM é represento por um nó na árvore graphML. O conteúdo dos nós do tipo *Text* é transferido para seus pais do tipo *Element*. Com isso, cada nó apresentado no documento *graphML* contêm o nome da tag que ele representa e a quantidade de caracteres associados a ele, que eram de seus filhos imediatos do tipo *Text*.

[2] GraphML - The GraphML File Format - http://graphml.graphdrawing.org
[3] Yed - http://www.yworks.com

Olhar diretamente para essas estruturas não é muito esclarecedor, pois as árvores são, em grande parte, muito grandes, dificultando a identificação de qualquer padrão. Para facilitar a análise dessas árvores, os nós são coloridos seguindo regras, que são extraídas a partir da quantidade de caracteres existente nos documentos. As regras de cores é a seguinte:

- branco - se o nó não contém caracteres;
- amarelo - se o nó contém entre 1 e 182 caracteres, encontra-se no primeiro quartil;
- azul - se o nó contém entre 183 e 265 caracteres, encontra-se no segundo quartil;
- laranja - se o nó contém entre 266 e 628 caracteres, encontra-se no terceiro quartil;
- vermelho - se o nó contém mais de 629 caracteres, encontra-se no quarto quartil.

É importante esclarecer que, o cálculo dos quartis foi obtido da união da quantidade de caracteres das três classes, na tentativa de encontrar uma informação que se aplique na separação das classes.

Analisando essas estruturas já coloridas, é possível observar que dentro da classe notícia existem muitos documentos com nós da cor laranja e vermelho e quando existem apenas nós azuis ou laranjas, esses estão próximos, ou seja, possuem um ancestral comum a poucos níveis de distância.

Já na classe portal de notícias, é possível observar vários nós laranjas ou azuis, mais que na classe notícia, e a distância entre esses nós é proporcionalmente maior que os encontrados na classes notícia.

Na classe outro, dificilmente se encontram documentos que tenham nós na cor vermelha ou laranja. Quando encontrados esses nós, com quantidade elevada de caracteres, esses são únicos ou aparecem isolados, ou seja, existe uma grande distância entre eles.

Para os resultados apresentados nos experimentos, foi utilizada uma outra forma de medir a distância. Essa distância foi calculada utilizando o desvio padrão entre os ids, numerados em um percurso pré-ordem, de nós que contêm mais que 250 caracteres, o teto de 265 foi reduzido por fornecer uma melhor faixa de corte.

Essa informação, obtida a partir da distância dos nós com mais de 250 caracteres, parece ser de extrema importância e gera resultados que podem ser observados na seção ***Experimento do Estudo de Caracteres***.

Na tabela 7 são apresentados alguns dados que tornam possível a visualização dessa separação nas classes.

Tabela 7: Estudo de caracteres, quantidade de nós com mais de 250 caracteres e desvio padrão de seus ids (DP) para todos os documentos do conjunto de treino

	Notícia		Portal de Notícias		Outro	
documento	# Nós	DP	# Nós	DP	# Nós	DP
1	3	10.42	0	0	1	0
2	2	0.5	0	0	0	0
10	5	4.5	1	0	0	0
11	1	0	5	308.03	0	0
12	1	0	0	0	1	0
23	2	0.5	2	106.5	0	0
24	1	0	0	0	0	0
25	4	2.17	0	0	7	188.7
58	8	6	0	0	0	0
75	6	3.7	0	0	7	72.48

O estudo de caracteres mostra que existem muitas informações na estrutura do documento. Essas informações podem ser exploradas antes da necessidade de realizar a análise textual. Porém, interrompemos os estudos em um ponto inicial, não aprofundando a busca de dados obtidos a partir da informação existente na estrutura do documento, por

considerar que as análises, aqui apresentadas, oferecem um conjunto de informações estruturais suficiente para a classificação dos documentos.

Os experimentos na seção ***Experimentos*** apresentam os testes da utilização de atributos obtidos a partir das metodologias apresentadas, sem utilizar qualquer tipo de análise numérica para a geração de atributos mais eficientes. Porém, mesmo com o resultado que não ultrapassa 75% de acurácia, podemos afirmar que esses atributos, em sua forma bruta, oferecem uma boa quantidade de informação para a classificação dos documentos.

Técnicas Utilizadas

Geração dos Atributos

Os atributos foram gerados a partir de classes, chamadas de coletores, especializadas em obter informações específicas, sendo essas informações referentes a cada linha de estudo apresentada na seção ***Estrutura e Informação***. Essas classes geram dados para cada documento, que são salvos em um arquivo no formato CSV (*common separeted values*), um arquivo separado por vírgulas, utilizado pelo framework de classificação WEKA.

Todos os coletores, assim como o código que dá suporte à sua execução são codificados em Python. A biblioteca externa libxml2[4] é utilizada, pois implementa um *parser* DOM de documentos HTML e esta é disponibilizada na linguagem C, para ser utilizada em Python é necessário o pacote Python libxml2dom[5] que implementa uma API para a libxml2, sendo considerada um *wrapper*.

Para o entendimento da complexidade da coleta de atributos, é fornecida uma breve explicação sobre o processamento executado:

- para cada documento HTML é gerada sua árvore DOM;
- cada coletor é executado interagindo sobre a árvore DOM do documento. Todos os coletores têm complexidade O(n) e seu resultado é mantido em memória;
- após a execução de todos os coletores sobre todos os documentos, é gerado o arquivo de saída CSV, onde são salvas as informações obtidas pelos coletores.

Ferramenta WEKA

Para a realização dos experimentos, é utilizada a ferramenta *Waikato Environment for Knowledge Analysis*[6] (WEKA), desenvolvida pela Universidade neozelandesa Waikato, em 1993, com o objetivo de agrupar diversos algoritmos de preparação de dados, mineração de dados e validação de resultados. Nessa ferramenta, são encontradas técnicas de mineração de dados como NaiveBayes, Linear Regression, IB1, Bagging, Part, Ridor, ID3, LMT, SVM e Decision Tree.

WEKA se destaca dentro das ferramentas disponíveis para mineração de dados, sendo mantida na licença *General Public License*. A existência de uma grande quantidade de heurísticas, fornecidas para o tratamento de dados, também favorece essa ferramenta e ajuda a manter sua posição de destaque [10].

É importante lembrar que, os dados fornecidos para a ferramenta WEKA foram gerados a partir dos coletores, descritos na seção anterior.

Validação Cruzada

Para realizar experimentos sobre o conjunto de teste, veja seção ***Separação do Conjunto de Teste e Treino***, é possível utilizar diversas técnicas como *Holdout*, sub conjunto

[4] LibXML2 - http://xmlsoft.org
[5] LibXML2DOM - http://www.boddie.org.uk/python/libxml2dom.html
[6] WEKA - http://www.cs.waikato.ac.nz/ml/weka/

randômico, validação cruzada ou ainda *Bootstrap* [5]. A técnica escolhida foi a validação cruzada com N partições (*cross-validation k-fold*), por funcionar bem em conjuntos pequenos de dados.

Esse método de validação é interessante por utilizar todo o conjunto de dados, de forma a sempre treinar e testar em conjuntos disjuntos. Porém, não é recomendado para conjuntos de dados muito grandes, pois seu processo é demorado.

Support Vector Machine

Support Vector Machine, introduzida por Vapnik [11], é uma técnica de reconhecimento de padrões utilizada para o problema de classificação, onde uma abordagem essencialmente geométrica é adotada. Nela, um conjunto de pontos, representados em um espaço euclidiano, é separado por um hiperplano proporcionando uma classificação binária.

Para um maior entendimento de SVM recomenda-se o trabalho de [12], onde um tutorial bem formulado é apresentado. Sendo esse a principal referência de SVM utilizada neste projeto.

Experimentos

Nesta seção, são apresentados experimentos que buscam validar as linhas abordadas por este trabalho. Os atributos obtidos a partir das linhas (estudo de nós, estudo de tags e estudo de caracteres) são utilizados para gerar os modelos de classificação e testa-los.

A técnica utilizada para classificação foi a SVM, com a implementação *Sequential Minimal Optimization* (SMO). Essa implementação foi escolhida, pois é disponibilizada de forma nativa na ferramenta WEKA.

Os experimentos com o conjunto de treino, apresentado na seção ***Separação do Conjunto de Teste e Treino***, foram realizados utilizando a técnica de validação cruzada (*closs-validation*). Tal técnica foi aplicada dez vezes, sendo obtida a média dessas execuções para exibição do resultado.

Os resultados sobre o conjunto de teste, apresentados na seção ***Validação no Conjunto de Teste***, são obtidos a partir da criação de um modelo utilizando o conjunto de treino, e sua aplicação no conjunto de teste.

Experimentos do Estudo de Nós

- EN1: utiliza a quantidade total de nós, a altura máxima e a média de nós por altura;
- EN2: utiliza a quantidade de nós por altura, até a altura 49;
- EN3: utiliza a quantidade de nós por altura, até a altura 49, de maneira proporcional à quantidade de nós no documento.

Tabela 9: Validação do conjunto de treino utilizando atributos de nós com 2 classes (notícia, portal).

Informação	EN1	**EN2**	EN3
Acurácia	69.13	**71.36**	59.44
taxa de erro	31.06	**28.67**	40.53
Precision	0.64	**0.69**	0.64
Recall	0.89	**0.8**	0.47
F1	0.74	**0.73**	0.52
Area Under ROC	0.69	**0.71**	0.59
Tempo treino	0.03	**0.03**	0.03

Tabela 10: Validação do conjunto de treino utilizando atributos de nós com 3 classes (notícia, portal e outro).

Informação	EN1	**EN2**	EN3
Acurácia	51.45	**56.12**	51.57
taxa de erro	48.37	**43.81**	48.28
Precision	0.5	**0.56**	0.5
Recall	0.51	**0.62**	0.39
F1	0.5	**0.58**	0.42
Area Under ROC	0.66	**0.71**	0.62
Tempo treino	0.07	**0.08**	0.08

Nesse teste é possível reparar que analisar os dados de maneira proporcional não adiciona qualidade ao modelo de classificação. No entanto, a idéia de observar os nós por altura é válida, e traz melhoras para os resultados.

Experimentos do Estudo de Tags

- ET1: utiliza a quantidade de tags escolhidas como importantes, exibidos na tabela 5;
- ET2: resultados obtidos a partir da quantidade de tags escolhidas como importantes por altura;
- ET3: resultados obtidos a partir da quantidade de tags escolhidas como importantes de forma proporcional à quantidade total de tags no documento;
- ET4: resultados obtidos a partir da quantidade de tags escolhidas como importantes de forma proporcional à quantidade total de tags no documento por altura.

Tabela 11: Validação do conjunto de treino utilizando atributos de tags com 2 classes (notícia, portal).

Informação	ET1	**ET2**	ET3	ET4
Acurácia	71.36	**74.68**	59.5	69.58
taxa de erro	28.67	**25.39**	40.4	30.48
Precision	0.68	**0.78**	0.6	0.93
Recall	0.8	**0.7**	0.58	0.43
F1	0.73	**0.73**	0.58	0.57
Area Under ROC	0.71	**0.74**	0.6	0.69
Tempo treino	0.02	**0.05**	0.02	0.04

Tabela 12: Validação do conjunto de treino utilizando atributos de tags com 3 classes (notícia, portal e outro).

Informação	ET1	**ET2**	ET3	ET4
Acurácia	52.64	**58.76**	45.75	53.74
taxa de erro	47.25	**41.24**	54.21	46.05
Precision	0.58	**0.7**	0.48	0.82
Recall	0.68	**0.55**	0.36	0.39
F1	0.61	**0.6**	0.4	0.51
Area Under ROC	0.73	**0.73**	0.64	0.67
Tempo treino	0.07	**0.13**	0.07	0.12

Observado os resultados acima, é possível notar que, novamente, a informação proporcional não adiciona qualidade ao modelo, enquanto que a informação por altura aumenta a qualidade da classificação.

Experimentos do Estudo de Caracteres

- EC1: resultados gerados a partir do total de caracteres existentes no documento, total de nós existentes no documento e a média de caracteres por nós;

- EC2: resultados gerados a partir do total de caracteres existentes no documento por altura;
- EC3: resultados gerados a partir do total de caracteres existentes no documento, de maneira proporcional à quantidade total de caracteres no documento, por altura;
- EC4: resultados gerados a partir da distância entre nós com quantidade de caracteres maior que 250. Foram geradas ainda 3 informações que podem ser obtidas a partir da distância dos nós que tâm mais que 250 caracteres. O primeiro é obtido a partir da divisão do desvio padrão dos ids pela quantidade de ids de nós que têm mais que 250 caracteres, esse é chamado de "Valor1". O segundo é a um valor binário para aqueles que tem ``Valor1" maior que um parâmetro, o parâmetro que proporciona uma melhor separação de classes para o conjunto de treino, ou seja, o utilizado é 30. Finalmente, o último, também, é um valor binário para os documentos que têm desvio padrão dos ids maior que um parâmetro, o escolhido para o experimento por proporcionar a melhor separação das classes do conjunto de treino é 10.

Tabela 13: Validação do conjunto de treino utilizando atributos de caracteres com 2 classes (notícia, portal).

Informação	EC1	EC2	EC3	**E4**
Acurácia	69.45	65.18	64.22	**81.51**
taxa de erro	30.67	34.92	35.82	**18.36**
Precision	0.64	0.71	0.67	**0.78**
Recall	0.92	0.55	0.6	**0.92**
F1	0.75	0.6	0.62	**0.84**
Area Under ROC	0.69	0.65	0.64	**0.82**
Tempo treino	0.02	0.02	0.02	**0.02**

Tabela 14: Validação do conjunto de treino utilizando atributos de caracteres com 3 classes (notícia, portal e outro).

Informação	EC1	EC2	EC3	**EC4**
Acurácia	47.45	50.43	51.79	**56.87**
taxa de erro	52.53	49.61	47.98	**43.32**
Precision	0.42	0.58	0.54	**0.6**
Recall	0.93	0.34	0.43	**0.88**
F1	0.58	0.42	0.46	**0.7**
Area Under ROC	0.64	0.61	0.6	**0.79**
Tempo treino	0.07	0.08	0.08	**0.06**

É interessante notar que, no teste com duas classes a informação proporcional não melhora o resultado, enquanto que no teste com as três classes a acurácia da classificação sofre uma pequena melhora.

As evoluções no estudo mostram qualidade quando analisamos os resultados das três classes, enquanto que o estudo inicial (EC1) apresenta resultados melhores para duas classes. No entanto, a classificação sofre uma melhora significativa quando é realizada a análise sobre os dados para que os atributos gerem o máximo de separação no conjunto de treinamento, como apresentado no E4.

Experimentos com Todos os Estudos

Para finalizar, utilizamos os melhores atributos de cada estudo (EN2, ET2, EC4) para realizar um último teste sobre o conjunto de treino. Repare que a união dos melhores atributos, das três linhas de estudos, gera resultados com aproximadamente 5% a mais de acurácia do que os testes com os atributos dos estudos isolados. Na tabela 15 são apresentados dois resultados. O resultado TT1 é o teste com duas classes (notícia e portal de notícias) e o resultado TT2 é o teste com as três classes.

Novamente, é possível notar como a classe outro reduz a qualidade da classificação. Esse fato é motivado pela falta de padrão dentro dessa classe, o que prejudica a criação do modelo de classificação.

Tabela 15: Validação do conjunto de treino utilizando os melhores atributos de todos os estudos.

Informação	TT1	TT2
Acurácia	83.99	64.92
taxa de erro	15.98	34.97
Precision	0.88	0.78
Recall	0.81	0.71
F0	0.83	0.73
Area Under ROC	0.84	0.84
Tempo treino	0.05	0.12

Validação no Conjunto de Teste

Na tabela 16 são apresentados os resultados, apenas na classe notícia e portal de notícia, utilizando os coletores que apresentaram os melhores resultados nos experimentos no conjunto de teste. Em seguida, na tabela 17 o experimento é repetido, sendo adicionada a classe outro.

- TF1 - melhor experimento de nós (EN2)
- TF2 - melhor experimento de tags (ET2)
- TF3 - melhor experimento de caracteres (EC4)
- TF4 - todos os melhores juntos (EN2, ET2 e EC4)

Tabela 16: Validação no conjunto de teste utilizando duas classes

Informação	TF1	TF2	TF3	TF4
Acurácia	65	67.5	70	**75**
taxa de erro	35	32.5	30	**25**
Precision	0.63	0.73	0.67	**0.81**
Recall	0.75	0.55	0.8	**0.65**
F0	0.68	0.63	0.73	**0.72**
Area Under ROC	0.65	0.68	0.7	**0.75**
Tempo treino	0.07	0.1	0.03	**0.16**

Tabela 17: Validação no conjunto de teste utilizando três classes

Informação	TF1	TF2	TF3	TF4
Acurácia	40	48.33	**55**	45
taxa de erro	60	51.67	**45**	55
Precision	0.37	0.53	**0.5**	0.48
Recall	0.5	0.45	**0.8**	0.55
F0	0.43	0.49	**0.62**	0.51
Area Under ROC	0.59	0.64	**0.69**	0.7
Tempo treino	0.1	0.16	**0.07**	0.19

Conclusão e Trabalhos Futuros

Este trabalho apresenta resultados, utilizando SVM, para a classificação das páginas de duas classes funcionais, partindo da hipótese de que é possível classificar uma página da Web baseando-se apenas em informações estruturais. Com isso, é razoável afirmar que a utilização de elementos estruturais para a classificação é viável, chegando à acurácia de 75% para a classificação em duas classes (notícia e portal de notícias).

O estudo de nós, tags e caracteres, podem ser aprofundados, sendo assim, deixados para trabalhos futuros, já que apresentam grande potencial para a extração de novos atributos. Em

especial, o estudo da distribuição de caracteres apresenta características fortes e provavelmente o seu aprofundamento nessa linha chegará a resultados mais precisos.

Durante a evolução deste trabalho, problemas com a codificação (*encoding*) dos documentos e a má formação do documento HTML foram encontrados. Problemas como esses resultam na impossibilidade de realizar o *parser* do documento HTML e, conseqüentemente, gerar a árvore DOM. A capacidade de contorná-los, sem resolve-los por definitivo foi aplicada, pois tais problemas são extensivamente encontrados por trabalhos na área, e soluções definitivas não são fornecidas.

Referências

1 - Netcraft Web Server Survey . 2008. Disponivel em: <http://news.netcraft.com/archives/web_server_survey.html>. Acessado em: mai. 2008

2 - Elgersma, E.; Rijke, M. Learning to Recognize Blogs: A Preliminary Exploration. EACL 2006 Workshop on New Text-Wikis and Blogs and Other Dynamic Text Sources . Abril. 2006. Disponível em: <http://www.sics.se/jussi/newtext/>. Acesso em: set. 2007

3 - Amitay, E.; Carmel, D.; Darlow, A.; Lempel R.; Soffer, A. The Connectivity Sonar: Detecting Site Functionality by Structural Patterns. 14th ACM Conference on Hypertext and Hypermedia . Agosto. 2003. Disponível em: <http://einat.webir.org/Hypertext_2003_p38-amitay.pdf>. Acesso em: jan. 2008.

4 - Mitchell, T. M. Machine Learning. . New York: McGraw-Hill. 1997.

5 - Tian, Y.; Huang, T.; Gao W.; Cheng, J.; Kang, P. Two-Phase Web Site Classification Based on Hidden Markov Tree Models. IEEE/WIC International Conference on Web Intelligence (WI'03) . 2003. Disponivel em: <http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1241198>. Acessado em: nov. 2007

6 - Eissen, S. M.; Stein, B. Two-Phase Web Site Classification Based on Hidden Markov Tree Models. 27th German Conference on Artificial Intelligence, Berlin. . 2004. Disponivel em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1241198>. Acessado em: mai. 2008

7 - Gyöngyi, Z.; Garcia-Molina, H. Link spam alliances 31st International Conference on Very Large Data Bases (VLDB), trondheim, Norway . 2005. Disponivel em: <http://portal.acm.org/citation.cfm?id=1083654>. Acessado em: mai. 2008

8 - Lindemann, C.; Littig, L. Coarse-grained Classification of Web Sites by Their Structural Properties. 8th Int. Workshop on Web Information and Data Management, Arlington, VA . 2006. Disponível em: <http://doi.acm.org/10.1145/1183550.1183559>. Acessado em: out. 2007

9 - Hégaret, P.; Whitmer, R.; Wood, L. Document Object Model (DOM) Disponível em: <http://www.w3.org/DOM/> Acessado em: abr. 2008.

10 - Silva, M. P. S. Mineração de Dados - Conceitos, Aplicações e Experimentos com Weka IV Escola Regional de Informática RJ/ES . novembro. 2004. Disponivel em: <http://www.sbc.org.br/bibliotecadigital/download.php?paper=35>. Acessado em: mai. 2008

11 - Boser, B.; Guyon, I.; Vapnik, V. A training algorithm for optimal margin classifiers Annual Workshop on Computational Learning Theory. 1992. Disponível em: <http://portal.acm.org/citation.cfm?id=130401>. Acessado em: mai. 2008

12 - Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery . 1998. Disponível em: <http://citeseer.ist.psu.edu/burges98tutorial.html>. Acessado em: jan. 2008