Classificando Automaticamente Documentos Digitais no *Site* de Notícias do UOL

Elias Oliveira, Patrick Marques Ciarelli,
Marcos Hercules Santos e Bruno Oliveira da Costa
Departamento de Ciências da Informação
Universidade Federal do Espírito Santo
Campus de Goiabeiras, Av. Fernando Ferrari, s/n,
Cx Postal 5011, 29060-970 – Vitória, ES.
http://www.inf.ufes.br/∼elias
elias@inf.ufes.br

Resumo

O crescente volume de documentos tem trazido preocupações metodológicas entre os profissionais da área de Ciências da Informação. Se por um lado temos o difícil problema da escolha acertada de documentos contendo a informação desejada pelo usuário/cliente, de outro lado temos o árduo trabalho da pré-organização destes mesmos documentos para posterior recuperação. Acrescenta-se a esse contexto a falta de pessoal em que, em geral, vivem as unidades de informação neste país. Este trabalho apresenta um modelo de representação algébrica de documentos textuais, o qual pode ser uma alternativa metodológica para o problema de classificação de documentos. Utilizamos como forma de comparação de nosso processo automático, documentos já classificados por especialistas em *site* de notícias UOL. Os resultados se mostram promissores indicando que tal metodologia poderia ser utilizada na organização de documentos em uma biblioteca digital.
Palavras-chave: Classificação automática, Modelo vetorial, Recuperação da informação, Biblioteca Digital.

1 Introdução

O volume de informação codificada disponível ao público, de maneira geral, vem crescendo vertiginosamente desde a iniciativa da imprensa de Gutemberg (CHARTIER, 1998). Hoje, o fato de termos maior acesso à diversas informações via a grande rede Internet e a facilidade de publicarmos o que quisermos nesta rede, vem inundando-nos de *informação* de uma forma jamais vista na história da humanidade (TEIXEIRA; SCHIEL, 1997).

Por outro lado, o excedente informacional produzido nestes últimos anos, em particular na Internet, trouxe junto consigo uma nova dificuldade aos usuários da informação eletrônica (MARCONDES; SAYÃO, 2002). Em conseqüência disso, vemos que cada vez mais torna-se crítico o problema de identificação da informação especificamente relevante para um usuário alvo. Isso nos leva ao caos organizacional provocado por essa enxurrada de documentos disponíveis na rede e, ainda, a falta de ferramental apropriado para o tratamento dessa informação. Essa carência temos evidenciado nos atuais sistemas de busca que ainda produzem uma alta revocação e baixa precisão na informação recuperada.

Neste contexto apresenta-se um dos grandes desafios aos profissionais da informação de hoje (CUNHA, 2005): lidar de forma produtiva com a informação dispersa na Internet. Não temos como ignorar este grande repositório de informação que é a Internet, mas não podemos deixar somente por conta do usuário o árduo trabalho de garimpar pedras preciosas, por ele almejadas, neste moderno repositório digital. Mesmo nesta nova estrutura do mundo moderno, devemos nos preocupar em fornecer *a cada livro seu leitor* (GIGANTE, 1995), como nos diz a terceira lei fundamental de Ranganathan (1996), ou reformulando esta lei para os novos meios eletrônicos: *a cada porção de informação o seu consumidor*.

Assim, este artigo trata da apresentação de uma metodologia que vem sendo utilizada para lidar automaticamente com uma grande massa de documentos no que diz respeito a indexação destes, utilizando-se da extração dos termos relevantes do documento. A partir desta metodologia, utilizaremos um modelo vetorial de representação dos documentos para avaliarmos similaridades entre os mesmos. Com isso produziremos classes de documentos segundo seus enfoques temáticos. Compararemos estes resultados, produzidos de forma automática, com aqueles gerados pelo especialista humano para avaliarmos a eficácia e eficiência desta metodologia automática.

Este artigo está organizado da seguinte forma: Na Seção 2 fazemos uma breve revisão da literatura relacionada com o trabalho desenvolvido aqui. Apresentamos alguns modelos para representação abstrata de documentos para manipulação automática. Nossos experimentos são apresentados na Seção 3. Nossa conclusão é apresentada na Seção 4, onde também lançamos algumas idéias para futuros trabalhos.

2 Lidando com Documentos Digitais

Muitas iniciativas têm surgido nos últimos anos no sentido de disponibilizar uma larga quantidade de materiais bibliográficos. Mais recentemente tivemos, também, iniciativa como a *Google Book Search* (`http://books.google. com`) com o projeto de digitalizar o acervo de várias bibliotecas de universidades Norte-Americanas, incluindo algumas no Brasil. Indo em sentido semelhante, já a algum tempo importantes editores de jornais científicos vêm disponibilizando seus acervos em meio digital.

A parte esses projetos milionários, podemos constatar o crescente número de bibliotecas digitais de dissertação e teses que estão sendo implantadas recentemente

(CUNHA; MCCARTHY, 2006) no Brasil. Entretanto, para realmente tirarmos proveito desse imenso acervo digital que está sendo formado aqui e no mundo, será necessário que processemos, de forma mais *inteligente* (POLTRONIERI; OLIVEIRA, 2005) as muitas páginas de esforço intelectual que estão sendo disponibilizadas e, também muitas outras que estão à caminho.

O processo manual de organização documental pode ser feito por profissionais da informação, como bibliotecários, ou por especialistas da área de conhecimento do corpus (FUJITA, 2003). Entretanto este processo é lento e requer a presença constante de um especialista, esse nem sempre disponível. Packer (1998) aponta o elevado tempo gasto para a extração de elementos da estrutura de um documento para a construção dos metadados na publicação de uma revista eletrônica.

Além disso, mesmo utilizando uma equipe de profissionais qualificados e uma política de indexação consistente para a organização documental, a subjetividade desse processo pode levar à situações em que um mesmo documento poderá ser representado de diferentes formas (FERNEDA; PINHEIRO, 2005). Em conseqüência destes inconvenientes, a alternativa do uso de uma metodologia automatizada pode auxiliar o profissional da informação a realizar o tratamento técnico documental trazendo, dessa forma, várias vantagens, como por exemplo poupar do indexador o trabalho de realizar uma leitura exaustiva dos documentos para a escolha de descritores dos mesmos (DZIEKANIAK; KIRINUS, 2004, pag. 32). Diante disso precisamos repensar o fazer tradicional de organização bibliográfico para que possamos dar conta de acompanhar o crescimento dessa massa documental.

Nas próximas seções introduziremos o assunto do tratamento automático de texto. Para isso começaremos com o processo de indexação. Os modelos que iremos apresentar ainda estão longe de reproduzirem o especialista humano quando fazendo a mesma tarefa, porém o que desejamos é alcançar um resultado com qualidade aceitável em um tempo bem inferior àquele quando tendo um humano na realização da mesma tarefa.

2.1 Indexação Automática

A indexação é uma etapa importante do tratamento técnico documental para facilitar a recuperação da informação (PIEDADE, 1977). Esta etapa consiste em extrair termos de um documento que melhor represente seu conteúdo. Há décadas os profissionais da informação vêm desempenhando essa atividade. Porém, com a explosão documental surge a necessidade destes profissionais utilizarem métodos mais automatizados para a indexação (LANCASTER, 2003). Soma-se ao alto desempenho do processo automático a redução da subjetividade nos processos manuaias de indexação (MAMFRIM, 1991, p. 191). Indexação automática é, segundo Robredo *apud* (SILVA; FUJITA, 2004), qualquer procedimento que permita identificar e selecionar os termos que representem o conteúdo dos documentos, sem a intervenção direta do documentarista. Como no processo manual, os métodos automáticos de indexação consistem também em extrair os termos que se encontram em certa posição de um documento, como por exemplo no título ou no resumo (LANCASTER, 2003). Um outro método alternativo

de indexação consiste em se escolher os termos de indexação através da contagem de palavras que ocorram com uma determinada freqüência, em um documento como todo.

A indexação automática baseada na freqüência de termos surgiu na década de 50 (LANCASTER, 2003). Contudo, não são quaisquer palavras que servem como termo de indexação. O sistema automático utiliza-se de uma lista de palavras proibidas, as quais possuem pouco significado semântico. Tais palavras, portanto, não serão consideradas como termos de indexação. Às palavras relevantes para a indexação devemos encontrar pesos apropriados para distinguir umas das outras no contexto em estudo. Buscar os melhores pesos para tais termos não é uma tarefa trivial, entretanto com ajuda de modelos Matemáticos e técnicas de Inteligência Artificial poderemos obter bons resultados, como veremos a seguir neste trabalho.

Na seção seguinte iremos apresentar uma metodologia de representação algébrica de documentos. Nesta metodologia, os documentos são representados de forma vetorial baseados na freqüência de ocorrência de seus termos. Como conseqüência desta representação seremos capazes de lidar com uma base de dados de documentos com instrumentos vindos da Matemática e Estatística.

2.2 Alguns Modelos de Representação de Documentos

Em virtude da grande massa documental existente no mundo contemporâneo, urge utilizarmos alguma forma abstrata para representação destes documentos para então tratarmos. A literatura (BAEZA-YATES; RIBIERO-NETO, 1998) é rica em apresentar modelos de representação de documentos textuais. Entre muitos outros modelos de representação podemos citar as Redes Neurais Artificiais (HAYKIN, 1998), os processos estatísticos Bayesianos (PEARL, 1988), a técnica *Latent Semantic Indexing* (LSI) (BERRY, 2003; BERRY; DUMAIS; O'BRIEN, 1995), entre outras.

A maioria dos métodos utilizados, em particular o escolhido para os experimentos nesta pesquisa, fazem uso da comparação lexical entre as palavras existentes no índice dos documentos para a realização do processo de classificação dos documentos ali representados. Isto acontece por ser ainda muito custosos, do ponto de vista computational, técnicas como as de extração automática de ontologia formal e análise conceitual destes documentos como as apontatas por Alvarenga (2001), ou mesmo da extração dos sintagmas como propõem outros autores (KURAMOTO, 2002).

Neste trabalho estaremos adotando o modelo vetorial de representação de documentos textuais. Escolhemos este modelo pela simplicidade de implementação e por atender bem aos propósitos ilustrativos deste trabalho.

2.2.1 Representação Vetorial de Documentos

No modelo por nós adotado neste trabalho, o vetorial, os documentos são representados por vetores no espaço R^n (BAEZA-YATES; RIBIERO-NETO, 1998). n representa o número de termos-palavras nos documentos considerados. Cada documento é considerado portanto um vetor de termos. Formalizando o que foi dito acima, consideremos um conjunto de documentos $D = \{d_1, d_2, \ldots, d_j, \ldots, d_n\}$, onde d_i é um dos

elementos deste conjunto. O documento d_i será representado portanto por um vetor de pesos $d_i = [w_1, w_2, \ldots, w_k, w_{k+1}, w_{k+2}, \ldots, w_n]$, sendo que k é o número de todos termos $\{t_1, t_2, \ldots, t_k\}$ distintos que aparecem no documento d_i. Os demais termos $\{t_{k+1}, t_{k+2}, \ldots, t_n\}$, associados aos pesos $[\ldots, w_{k+1}, w_{k+2}, \ldots, w_n]$, são termos que aparecem em outros documentos. Portanto, $\{t_1, t_2, \ldots, t_k, t_{k+1}, t_{k+2}, \ldots, t_n\}$ são todos os termos do vetor do documento d_i e a freqüência dos termos $t_{k+1} = t_{k+2} = \ldots t_n = 0$ neste vetor. Assim, podemos concluir que um termo (palavra no documento) pode aparecer em mais de um documento. Portanto, a cada termo será atribuído um peso w_i. Este peso será relativo a ocorrência do termo t_i, tanto no documento onde ele aparece em relação aos demais termos deste mesmo documento, como também quanto ao número de documentos do conjunto em que o termo aparece. Através disso ponderamos a importância deste termo no conjunto de documentos onde o mesmo aparece. Uma das propostas de ponderação desta importância apresentada na literatura (BAEZA-YATES; RIBIERO-NETO, 1998) é dado pela função $idf_i = log\frac{N}{n_i}$, onde idf_i (*inverse document frequency*) é o valor desta ponderação para o termo t_i, N é o total de documentos no conjunto D e n_i o número de documentos em que o termo t_i aparece. Com esta função queremos tornar sensível o fato de que se um termo aparece em todos os documentos, esta função assumirá valor próximo de zero.

Tabela 1: Representação vetorial de um documento.

Índice i	Peso w_i	Termo t_i
		d_1
1	**3**	**campeonato**
2	1	brasileiro
3	1	próximo
4	1	fim
5	1	foi
6	1	prejudicado
7	1	desorganização
8	**2**	**times**
9	1	famosos
10	1	poderão
11	1	rebaixados
12	1	entrando
13	1	justiça
14	1	pedir
15	1	anulação

Para dar uma ilustração do formalizado acima, vejamos este exemplo dos procedimentos de construção do vetor representativo do documento dado a seguir. Considere que tenhamos a seguinte notícia na área de esporte: – d_1: *O campeonato brasileiro está próximo ao fim. Tal campeonato foi muito prejudicado pela desorganização e times famosos poderão ser rebaixados. Alguns times estão entrando na Justiça para*

pedir a anulação do campeonato.

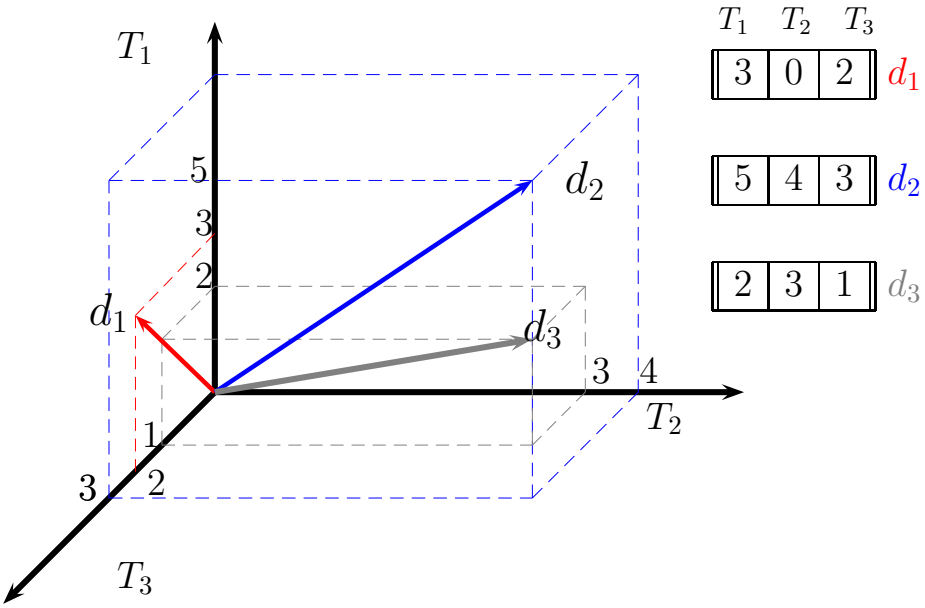
Primeiramente devemos excluir as palavras sem muito significado: os artigos e preposições, por exemplo. São as *stop words* (BAEZA-YATES; RIBIERO-NETO, 1998). Ficamos com a seguinte lista de palavras apresentada na Tabela 2.2.1 quando analizarmos o documento d_1. Para facilitação do entendimento, neste exemplo estaremos considerando a influência dos $idf = 1$ para todos pesos dos termos. Outra estratégia que estaremos adotando neste trabalho será a de utilizarmos na representação vetorial do documento apenas as palavras que tiverem peso maior que 50% do termo de maior peso. No caso da Tabela 2.2.1 o termo de maior peso é a palavra *campeonato*, com peso 3. Assim somente utilizaremos aquelas palavras com peso igual ou superior a $3/2 = 1,5$. Com isso ficamos somente com *campeonato* e *times* para a representação vetorial deste documento.

Agora considere outros dois documentos que depois do procedimento acima teriam os seguintes termos representativos:

1. d_2: peso 5 para o termo **campeonato**, 4 para **brasileiro** e 3 para **times**;
2. d_3: peso 2 para o termo **campeonato**, 3 para **brasileiro** e 1 para **times**;

Através deste exemplo ilustrativo criado e a representação descrita acima é possível agora visualizar os três documentos de forma gráfica. Na forma gráfica podemos ver a relação de *distância* que existe entre os documentos quando olhamos o ângulo que um vetor tem com o outro. Este conceito de *distância* será muito utilizado mais adiante neste trabalho.

Figura 1: Representação gráfica de três documentos de acordo com o modelo vetorial.



Na Figura 2.2.1 apresentamos a representação vetorial, de forma gráfica, de três documentos ilustrativos desta metodologia. Os termos $T_1 =$ *campeonato*, $T_2 =$ *brasileiro* e $T_3 =$ *times* representam os termos que aparecem nos documentos d_1, d_2 e d_3

representados em vermelho, azul e cinza, respectivamente. No gráfico, o peso dado ao termo T_1 no documento d_2 foi 5, enquanto em d_3 foi 2, o que significa que este termo tem uma importância maior para o segundo documento em relação ao terceiro. Notamos que o termo T_2 não ocorre em d_1, por isso está com valor nulo na segunda posição do vetor representativo deste documento.

Esta forma de representar um documento nos mostra que enquanto nós seres humanos *pensamos*, as máquinas *fazem contas*. Portanto, o que está por trás de um modelo como esse é o fato de transformar o processo de indexação e classificação em um processo de *contagem* para que o computador possa nos auxiliar a tratar grandes volumes de documentos.

Desta forma, considaremos a pequena base ilustrativa $D = \{d_1, d_2, d_3\}$ de documentos. O que queremos agora é saber precisamente quão similar é um documento do outro. O que desejamos é calcular o valor de $sim(d_i, d_j)$ entre quaisquer dois documentos da base. Uma vez tendo a representação vetorial dos documentos da base, como já feito acima, a conta que agora devemos fazer é a seguinte (BAEZA-YATES; RIBIERO-NETO, 1998):

$$sim(d_i, d_j) = \frac{\mathbf{d_i} \bullet \mathbf{d_j}}{\mid \mathbf{d_i} \mid \times \mid \mathbf{d_j} \mid} = \tag{1}$$

$$= \frac{\sum_{k=1}^{n} w_k^i \times w_k^j}{\sqrt{\sum_{k=1}^{n} \{w_k^i\}^2} \times \sqrt{\sum_{k=1}^{n} \{w_k^j\}^2}} = cos(\theta) \tag{2}$$

Onde, $\mid \mathbf{d_i} \mid$ é o módulo do vetor d_i. $cos(\theta)$ é o cosseno do ângulo entre os vetores que representam os dois documentos d_i e d_j. O valor do cosseno de um ângulo varia em um intervalo de 0 à 1. Esse fato nos dará uma interpretação de distância entre os documentos, onde 0 significará o mais alto grau de dissimilaridade e 1 de completa similaridade. Já o valor w_k^i indica o peso referente ao termo t_k, no documento d_i, como descrito anteriormente.

Vamos exemplificar utilizando os três documentos ilustrativos acima. Para os documentos d_1 e d_2, a conta é a seguinte:

$$sim(d_1, d_2) = \frac{3 \times 5 + 0 \times 4 + 2 \times 3}{\sqrt{3^2 + 0^2 + 2^2} \times \sqrt{5^2 + 4^2 + 3^2}} = \frac{21}{25.49} = 0.82 = cos(\theta_{1,2})$$

$$sim(d_1, d_3) = \frac{3 \times 2 + 0 \times 3 + 2 \times 1}{\sqrt{3^2 + 0^2 + 2^2} \times \sqrt{2^2 + 3^2 + 1^2}} = \frac{8}{13.49} = 0.59 = cos(\theta_{1,3})$$

$$sim(d_2, d_3) = \frac{5 \times 2 + 4 \times 3 + 3 \times 1}{\sqrt{5^2 + 4^2 + 3^2} \times \sqrt{2^2 + 3^2 + 1^2}} = \frac{25}{24.49} = 0.94 = cos(\theta_{2,3})$$

As contas realizadas acima nos indicam que os documentos d_2 e d_3 têm o mais alto grau de similaridade entre os três documentos, 0.94. Note que intuitivamente podemos visualizar este resultado no gráfico da Figura 2.2.1.

O exemplo acima foi criado de forma a ilustrar as partes importantes do modelo que estamos abordando, por isso escolhemos situações em que apenas três termos foram utilizados. Na próxima seção estaremos trabalhando com documentos de mais de 600 termos, o que não nos permitirá a representação gráfica destes documentos.

3 Pondo à Prova o Modelo Apresentado

Esta seção está dividida em duas partes. Na primeira, Seção 3.1, mostramos como o modelo escolhido neste trabalho pode ser ajustado com documentos *corretamente classifiados*. A expressão *corretamente classifiados* se refere ao que o(s) especialista(s), ou grupo social local de indivíduos, concordam com a diferenciação/similaridade entre documentos que servirão de parâmetro para o modelo. É com base nesta escolha inicial que nosso modelo fará as futuras escolhas, agora sim de forma automática. Na Seção 3.2, fazemos a validação do modelo introduzindo novos documentos para serem testados de acordo com o modelo de classificação automática.

3.1 Calibrando o Modelo

Os experimentos realizados neste trabalho tiveram como objetivo principal a exemplificação das metodologias algébricas de indexação e de representação de documentos textuais, como mais uma ferramenta para o profissional da informação. Além disso, nosso sub-objetivo vai no sentido de mostrar que este conjunto de técnicas pode ser utilizado para classificar documentos de forma automática (ou semi-automática em certas circunstâncias em que a máquina não conseguir *ter certeza*) e, em conseqüência disso, muito mais rápido do que faria um ser humano. Entendemos que, em muitas situações do dia-a-dia a máquina não será capaz de superar o especialista humano. Porém, também entendemos que o especialista está muitas das vezes assorberbado de pequenas tarefas que, nos dias de hoje, a máquina poderia realizar mais rápido e com um bom nível de qualidade. Advogamos que agrupamento de documentos textuais, de interesse de um usuário particular, ou mesmo para outros fins (SANTOS; COSTA; OLIVEIRA, 2005), seja uma destas atividades.

Utilizamos o repositório de notícias RSS do UOL para realização de nossos experimentos. A escolha deste repositório, assim como outros similares, deveu-se ao fato de caracterizar-se como uma boa fonte de documentos publicamente disponível e já classificados por especialistas humanos. Desta forma, poderemos comparar os resultados da classificação de documentos produzidos em nossos experimentos com os existentes no repositório. Deste repositório extraímos, manualmente e ao acaso, cinco documentos de notícias de cada um dos seguintes assuntos: *cinema*, *economia* e *esporte*.

Como os textos, por vezes são longos, apenas indicamos aqui os *hiperlinks* onde os mesmos poderão ser encontrados.

1. Na área de economia:

```
eco1: http://noticias.uol.com.br/ultnot/economia/2005/11/04/ult35u44044.jhtm
eco2: http://noticias.uol.com.br/ultnot/economia/2005/11/04/ult1767u53812.jhtm
eco3: http://noticias.uol.com.br/ultnot/economia/2005/11/04/ult1767u53813.jhtm
eco4: http://noticias.uol.com.br/economia/ultnot/efe/2005/11/04/ult1767u53802.jhtm
eco5: http://noticias.uol.com.br/economia/ultnot/afp/2005/11/04/ult35u44037.jhtm
```

2. Na área de esportes:

```
esp1: http://noticias.uol.com.br/ultnot/esporte/2005/11/05/ult1777u36742.jhtm
esp2:http://www.gazetaesportiva.net/ge_noticias/newsarch/ch_119/noticia.php?wt=
uolnot&p=bndpZC0zODk5MDQtbm51bS0g
esp3: http://noticias.uol.com.br/ultnot/esporte/2005/11/05/ult1777u36727.jhtm
esp4: http://noticias.uol.com.br/ultnot/esporte/2005/11/04/ult1777u36710.jhtm
esp5: http://noticias.uol.com.br/ultnot/esporte/2005/11/04/ult1777u36707.jhtm
```

3. E, por último, na área de cinema:

```
cin1: http://cinema.uol.com.br/ultnot/2005/11/04/ult32u12544.jhtm
cin2: http://www1.folha.uol.com.br/fsp/ilustrad/fq0411200531.htm
cin3: http://www1.folha.uol.com.br/fsp/ilustrad/fq3110200520.htm
cin4: http://www1.folha.uol.com.br/fsp/ilustrad/fq3010200518.htm
cin5: http://cinema.uol.com.br/ultnot/2005/10/16/ult831u1924.jhtm
```

Os títulos eco1, eco2, eco3, eco4 e eco5 são os documentos da área econômica. Já os da área esportiva são esp1, esp2, esp3, esp4 e esp5 e os da área de cinema como cin1, cin2, cin3, cin4 e cin5, respectivamente.

Os algoritmos para extração dos termos de indexação dos documentos foram todos implementados na linguagem de programação Java. Para a indexação desconsideramos as palavras sem muito significado, como por exemplos: artigos e preposições; conhecidas na literatura como *stop words* (BAEZA-YATES; RIBIERO-NETO, 1998).

Após a indexação dos documentos geramos, para cada uma das áreas acima, um documento artificial contendo somente os termos com freqüência superior a 50% em relação ao termo de maior freqüência no documento no qual ambos aparecem. Cada um destes documentos artificiais são dinâmicos, ou seja, sempre que um novo documento vier a ser agrupado em uma dada classe seus termos serão considerados para, possivelmente, comporem os termos já existentes no documento artificial daquela classe. Dessa forma, buscamos acompanhar a linguagem correntemente utilizada em cada área, naquele tempo, uma vez que consideramos a linguagem como um sistema vivo e, portanto, dinânmico.

A idéia por trás da criação destes documentos artificiais veio de uma técnica muito conhecida na Estatística como *Análise Discriminante* de dados (JOHNSON; WICHERN, 1992, cap. 11). Ou seja, estamos dizendo que os termos existentes em cada um destes documentos artificiais são termos que *discriminam*, ou separam, os documentos da classe relacionada ao documento de outras. No modelo por nós adotado neste trabalho, o vetorial, os documentos são representados por vetores, como descrito na Seção 2.2.1.

Para sabermos quão similar um documento será do documento discriminante, nós utilizaremos um procedimento que consiste em se calcular o *produto vetorial* entre dois vetores (veja Equações (1) e (2), na Seção 2.2.1).

Com esta metodologia, transformamos o procedimento de análise de documentos em um procedimento de cálculo. Portanto, o espaço de busca por documentos similares se torna um sub-espaço do R^n, onde estaremos interessados em encontrar vetores que mais se assemelhem a um dado vetor, que no nosso experimento será o vetor representativo da classe, o documento discriminante.

Tabela 2: Cálculo de similaridade entre os documentos analizados e os discriminates das classes – parte I.

	Classes de Documentos		
	cinema	economia	esporte
cin1	0.499	0.0	0.0
cin2	0.417	0.0	0.039
cin3	0.408	0.0	0.0
cin4	0.512	0.0	0.0
cin5	0.399	0.053	0.0
eco1	0.0	0.415	0.0
eco2	0.0	0.626	0.0
eco3	0.0	0.357	0.0
eco4	0.0	0.409	0.0
eco5	0.0	0.643	0.0
esp1	0.0	0.0	0.419
esp2	0.0	0.0	0.418
esp3	0.03	0.0	0.467
esp4	0.0	0.01	0.552
esp5	0.0	0.0	0.370

Os resultados obtidos com estes experimentos estão apresentados na Tabela 2. Nesta tabela, as colunas *cinema*, *economia* e *esporte* representam os documentos discriminantes citados acima. As linhas da tabela representam os documentos utilizados para este experimento. Assim, podemos ver que os documentos se agrupam com mais alto grau de similaridade em torno dos documentos discriminantes de suas respectivas classes. Por outro lado, o grau de similaridade deste com respeito à outras classes é bem mais baixo, quando não é nulo. Por exemplo, o documento *cin5* tem uma similaridade de 0.399 com o documento *cinema* enquanto, por outro lado, tem uma similaridade de 0.053 com a classe de economia. Uma similaridade bem baixa como podemos ver. Um outro exemplo é o documento *eco2* que tem uma similaridade de 0.626 com o documento discriminante de economia, *economia*, e zero com as demais

classes.

Em dados não apresentados na tabela mencionada acima, pudemos constatar que o documento *eco4* obteve uma alta similaridade com *eco1*, 0.418, maior do que o valor apresentado em relação ao documento discriminante de economia. O que pudemos perceber analisando os dois documentos é que *eco1* e *eco4* falam sobre o mesmo assunto: bolsa de valores.

Um outro exemplo curioso foi com respeito ao documento *esp2*. Este documento apresenta similaridade zero em relação à todos os outros documentos utilizados como exemplos, inclusive alguns da classe de esporte. Em nosso entendimento, isso foi possível dado a grande variedade de esportes e modalidades dos mesmos. Portanto, ao analisar a notícia existente neste documento, *esp2*, descobrimos que o assunto se tratava de *handebol*, enquanto os documentos *esp1*, *esp4* e *esp5* relatam futebol e *esp3* motovelocidade. Porém, isso não nos trouxe nenhuma dificuldade em classificá-lo corretamente como sendo de esporte, com um alto grau de similaridade de 0.418 como mostra a Tabela 2.

Para considerarmos um documento como pertencente à uma determinada classe, adotamos um ponto de corte *pc*. Desta forma, bastará calcularmos a similaridade do novo documento em relação aos documentos discriminantes, se a similaridade deste documento for menor que este *pc*, significará que este documento pode, ou não, pertencer à classe do documento discriminante. Se este dado documento estiver abaixo do valor de *pc* de todos as outras classes, pode-se adotar a alternativa de se deixar a cargo do especialista humano a decisão de escolher a que classe esse documento melhor se enquadraria. O valor *pc* é calculado através do procedimento descrito a seguir.

Considere a média m_c, onde c representa a classe sendo avaliada, de similaridade dos documentos *corretamente* classificados em uma classe. Por exemplo, no caso apresentado na Tabela 2, nós temos cinco documentos corretamente classificados em economia. A média de similaridades destes documentos é portanto calculada da seguinte forma:

$$m_{economia} = \frac{0.415 + 0.626 + 0.357 + 0.409 + 0.643}{5} = 0.49$$

Agora temos que adotar um limite inferior de similaridade que representará nosso ponto de corte *pc*. Para isso calculamos o desvio padrão através da fórmula:

$$dp = \left(\frac{(d_1 - m_c)^2 + (d_2 - m_c)^2 + \ldots + (d_n - m_c)^2}{n} \right)^{1/2}$$

Finalmente, o ponto de corte é calculado da seguinte forma:

$$pc = m_c - dp = 0.370$$

Para o exemplo mostrado na Tabela 2, temos na Tabela 3 os respectivos valores de ponto de corte para cada uma das classes.

Note que estes pontos de corte conseguem decidir que, *cin2* com similaridade 0.417 com a classe de cinema, pertence a esta classe e não a classe de esporte, com uma

Tabela 3: Cálculo dos valores de ponto de corte para cada uma das classes consideradas nos experimentos.

Ponto de Corte para as Classes			
	cinema	economia	esporte
pc:	0.399	0.370	0.384

similaridade de 0.039, já que o ponto de corte para esporte exigiria que o documento tivesse um grau de similaridade maior que 0.384. Neste sentido é interessante é notar que o documento *esp5* estaria fora da classe de esporte por ter um grau de similaridade com o documento discriminador da classe inferior ao ponto de corte para esta classe, de apenas 0.370. Este seria o caso onde o especialista humano deverá tomar a decisão de escolher a que classe esse documento melhor se enquadraria. Todavia, este especialista humano tem agora uma *pré-análise* deste documento em que, de acordo com esta *pré-análise* o documento teria mais chances de pertencer à classe de esportes e não as outras, as quais este documento não tem nenhuma aparente afinidade (ver Tabela 2).

Nessa metodologia, quanto maior o número de documentos representativos de cada classe melhor será o processo decisório para os novos documentos. Isso é devido ao fato de que os documentos já classificados corretamente servirão de base, no tocante a variabilidade de seus termos, para os cálculos feitos acima. Portanto, como já dissemos anteriormente, a cada novo documento que é classificado em uma determinada classe, este novo documento *ensina* ao modelo *novas lições*, através da introdução de novos termos ao documento discriminante da classe.

Para validar o processo descrito acima, na próxima seção escolheremos outros três documentos e avaliar se a técnica apresentada consegue distinguí-los em uma das três classes apresentadas acima.

3.2 Classificando Novos Documentos

Uma vez tendo gerado uma base de dados com documentos classficados corretamente, podemos agora utilizar o modelo/sistema para tentarmos classificar automaticamente outros documentos. Desta forma, escolhemos outros três documentos, dentre as três classes, para mostrarmos como se daria o processo como um todo.

Os documentos escolhidos foram:

```
cin6: http://noticias.uol.com.br/ultnot/efe/2005/01/30/ult1817u2706.jhtm
eco6:http://noticias.uol.com.br/economia/ultnot/efe/2006/04/20/ult1767u65477.jhtm
esp6:http://espnbrasil.uol.com.br/scripts/noticia/artigo.asp?idArtigo=38669
```

O cálculo de similaridade foi suficiente para determinar a classe para dois dos três documentos acima selecionados. Os documentos nas áreas de economia e esportes,

Tabela 4: Cálculo de similaridade entre os documentos analizados e os discriminates das classes – parte II.

	Classes de Documentos		
	cinema	economia	esporte
cin6	0.296	0.000	0.000
eco6	0.000	0.575	0.000
esp6	0.120	0.000	0.541

eco6 e *esp6*, respectivamente, têm seus valores de similaridades acima do ponto de corte determinado na tabela 3. Todavia, vemos que o modelo não foi capaz de identificar, com alto grau de precisão a classe para o mesmo. O documento *cin6* tem um grau de similaridade com a classe cinema de 0.296, quando o ponto de corte para a classe de cinema é de 0.399. Este é o momento onde, como já apontamos em outro caso anterior, a interferência humana se faz necessária.

Mesmo quando não conseguimos com grande grau de certeza apontar uma classe para um determinado documento, o modelo que apresentamos aqui indicará qual das classes tal documento terá maior afinidade. Assim, o especialista humano terá uma sugestão a mais para sua tomada de decisão. Quando este especialista decidir colocar o documento *cin6* associado a classe cinema, ele estará fazendo com que o modelo *aprenda*. Isto se dá pelo fato de que uma nova contagem deverá ser realizada com os termos existentes entre os documentos da classe e, em decorrência disso, o ponto de corte *pc* será alterado, dando assim uma dinamicidade ao modelo.

4 Conclusão

Diante do crescimento vertiginoso de repositórios de informação no Brasil e também no mundo. O problema que surge daí é no *como* recuperarmos de forma mais inteligente a informação necessária para o nosso usuário/cliente. Os métodos tradicionais de tratamento da informação não são mais compatíveis com repositórios do tamanho da Internet. Portanto, para novos problemas devemos buscar novas soluções.

Este artigo discute a representação abstrata de documentos. A representação vetorial escolhida neste trabalho é tal que, nos permite representar graficamente um documento e visualizá-lo, quando em até três dimensões. Desta representação extrai-se os termos que servirão de índices para tais documentos.

Os documentos sendo representatos através de vetores, nos permite utilizar o cálculo do ângulo entre vetores como medida de similaridade entre quaisquer dois documentos. Com isso obtemos uma forma, automática, de agrupamento destes documentos em classes de semelhança.

Para testar o modelo apresentado neste trabalho, escolhemos um conjunto de documentos já previamente classificado pelo especialista humano. Com isso submetemos os documentos ao modelo de indexação e, posteriormente, a classificação. Os resultados nos mostraram que o modelo trouxe, de forma automática, a mesma classificação dada pelo especialista humano. Entendemos que mais testes precisarão ser realizados, entretanto, os experimentos nos mostrou da possibilidade de utilização desta ferramenta para auxílio ao especiliasta de classificação.

Esperamos em breve estarmos avaliando esta mesma ferramenta em uma comparação com a classificação manual de dissertações e teses em nossa biblioteca digital.

Referências

ALVARENGA, L. A Teoria do Conceito Revisitada em Conexão com Ontologias e Metadados no Contexto das Bibliotecas Tradicionais e Digitais. *DataGramaZero – Revista de Ciência da Informação*, v. 2, n. 6, 2001. Disponível em: <http://www.dgzero.org/dez01/F_I_art.htm>.

BAEZA-YATES, R.; RIBIERO-NETO, B. *Modern Information Retrieval*. 1. ed. New York: Addison-Wesley, 1998.

BERRY, M. W. *Survey of Text Mining: Clustering, Classification, and Retrieval*. New York: Springer-Verlag, 2003.

BERRY, M. W.; DUMAIS, S. T.; O'BRIEN, G. W. Using Linear Algebra for Intelligent Information Retrieval. *SIAM Review*, v. 37, n. 4, p. 537–595, 1995.

CHARTIER, R. *A Aventura do Livro – do Leitor ao Navegador – Conversações com Jean Lebrun & Roger Chartier*. São Paulo: Ed. da UNESP, 1998.

CUNHA, M. B. A Biblioteca em Tempos de Internet. Janeiro 2005. Disponível em: <http://gnomo.fe.up.pt/∼ci02005/blog/Newsletter-A-Informacao.pdf>.

CUNHA, M. B.; MCCARTHY, C. Estado Atual das Bibliotecas Digitais no Brasil. In: MARCONDES, C. H. et al. (Ed.). *Bibliotecas Digitais: Saberes e Práticas*. 2. ed. Salvador/Brasília: UFBA/IBICT, 2006. cap. 2, p. 25–54.

DZIEKANIAK, G. V.; KIRINUS, J. B. WEB Semântica. *Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação*, v. 2, n. 18, p. 20–40, 2004. Disponível em: <www.encontros-bibli.ufsc.br/Edicao_18/2_Web_Semantica.pdf>.

FERNEDA, E.; PINHEIRO, C. Rrepresentação Dinâmica de Documentos em Bibliotecas Digitais. São Paulo, Novembro 2005.

FUJITA, M. S. L. A Identificação de Conceitos no Processo de Análise de Assunto para Indexação. *Revista Digital de Biblioteconomia e Ciência da Informação*, v. 1, n. 1, 2003. Disponível em: <http://eprints.rclis.org/archive/00003723/>.

GIGANTE, M. C. Os Sistemas de Classificação Bibliográfica como Interface Biblioteca/Usuário. *Ciência da Informação*, v. 25, n. 2, 1995.

HAYKIN, S. *Neural Networks – A Comprehensive Foundation*. [S.l.]: Pearson Education, 1998.

JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Anaysis*. New Jersey: Prentice Hall, 1992.

KURAMOTO, H. Sintagmas Nominais: uma Nova Proposta para a Recuperação de Informação. *DataGramaZero – Revista de Ciência da Informação*, v. 3, n. 1, 2002. Disponível em: <http://www.dgz.org.br/fev02/F_I_art.htm>.

LANCASTER, F. W. *Indexação e Resumos: Teoria e Prática*. 2. ed. Illinois: University of Illinois, 2003.

MAMFRIM, F. P. B. Representação de Conteúdo via Indexação Automática em Textos Integrais em Língua Portuguesa. *Ciência da Informação*, v. 20, n. 2, p. 191–203, 1991.

MARCONDES, C. H.; SAYÃO, L. F. Documentos Digitais e Novas Formas de Cooperação entre Sistemas de Informação em C&T. *Ciência da Informação*, Brasília, v. 37, n. 3, p. 42–54, 2002.

PACKER, A. L. SciELO: uma Metodologia para Publicação Eletrônica. *Ciência da Informação*, v. 27, n. 2, 1998.

PEARL, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988.

PIEDADE, M. A. R. *Introdução á Teoria da Classificação*. 2. ed. Rio de Janeiro: Interciência, 1977.

POLTRONIERI, A.; OLIVEIRA, E. Finding Related Articles by a Bibliometric Approach. In: *9º International Congress on Medical Librarianship*. Salvador: [s.n.], 2005.

RANGANATHAN, S. R. *Five Laws of Library Science*. 1. ed. [S.l.]: Stosius Inc/Advent Books Division, 1996.

SANTOS, M. N. dos; COSTA, B. O. da; OLIVEIRA, E. Utilizando Comparações Ponderadas em Classificação Automática de Documentos. In: *III Simpósio Internacional de Bibliotecas Digitais*. São Paulo: [s.n.], 2005.

SILVA, M. R. da; FUJITA, M. S. L. A Prática de Indexação: Análise da Evolução e Tendências Teóricas e Metodológica. *TransInformação*, v. 0, n. 0, p. 133–161, 2004.

TEIXEIRA, C. M.; SCHIEL, U. A Internet e seu Impacto nos Processos de Recuperação da Informação. *Ciência da Informação*, v. 26, n. 1, 1997.