MODELO DE SINAIS PARA BUSCA E RECUPERAÇÃO DE INFORMAÇÃO TEXTUAL

Rafael Leonardo Siqueira da Silva

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Aprovada por:

_____
Prof. Geraldo Bonorino Xexéo, D.Sc.

_____
Prof. Jano Moreira de Souza, Ph.D.

_____
Prof. Claudia Maria Garcia Medeiros de Oliveira, Ph.D.

_____
Prof. Eduardo Antônio Barros da Silva, Ph.D.

_____
Prof. Geraldo Zimbrão da Silva, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
MARÇO DE 2007

SILVA, RAFAEL LEONARDO SIQUEIRA

MSBRI: Modelo de Sinais para Busca e Recuperação de Informação [Rio de Janeiro] 2007

XI,127p., 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia de Sistemas e Computação, 2007)

Dissertação – Universidade Federal do Rio de Janeiro, COPPE

1. Busca e Recuperação de Informação
2. Processamento de Sinais
3. Transformada de Fourier
4. Transformada Wavelet
5. Multiresolução
6. Sistemas Distribuídos

I. COPPE/UFRJ II. Título (série)

Aos meus pais, Luis Carlos e Mirian,
pelo apoio e compreensão indispensáveis
para realização desse trabalho.
.

AGRADECIMENTOS

Agradeço aos meus pais, Luis Carlos Siqueira da Silva e Mirian Domingues Gonçalves Leonardo, sem os quais eu, definitivamente, não estaria aqui. Agradeço pela formação, educação e, sobretudo, pelos valores por eles passados a mim. Tenho consciência e gratidão pelo apoio e carinho que me foram dedicados e que culminaram na realização desse trabalho.

Agradeço ainda àqueles que, apesar de pontualmente, tiveram papeis decisivos em minha vida. Particularmente, à Doutora Lélia Tozatto, que há 24 anos, acreditou na criança que acabara de entrar em seu consultório. Hoje, os olhos dos quais ela vem cuidando desde então foram capazes de criar esta dissertação. Agradeço à prof. Carmem (1990), que, quando os demais me viraram o rosto, sorriu-me e deu-me força para que eu continuasse após a reprovação. Hoje, aquele que foi reprovado, vence mais uma etapa de sua vida.

Agradeço ao meu Orientador, Geraldo Bonorino Xexéo, por ter acreditado e apoiado minhas idéias. Ele se mostrou mais que um orientador, mostrou-se um amigo com o qual aprendi muito nos últimos anos. Seus conselhos e apoio foram decisivos para o direcionamento de minha vida profissional.

Agradeço a todos meus amigos que participaram comigo dessa jornada, em especial àqueles mais próximos, Amanda Varella, Vinícios Pereira, Vinicius Vonheld e José Nogueira. Tenho um especial agradecimento a Andressa Kalil, que me apoiou muito no início desse trabalho e, sem a qual, provavelmente eu teria muito mais dificuldades em meu caminho. Agradeço também a Daniela Marques Pereira, amiga, que ouviu e suportou minhas digressões durante o desenvolvimento desse trabalho.

Por fim, agradeço a Luciana do Valle Barbosa, minha namorada, pela compreensão e tolerância que teve comigo durantes estes últimos três anos que me dediquei a essa dissertação. Sem seu apoio, tenho certeza, que tudo seria mais difícil.

Em suma, agradeço a todos que de alguma forma participaram ou contribuíram para criação desse trabalho, mesmo aqueles que não foram nominalmente citados. Muito Obrigado.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

# MODELO DE SINAIS PARA BUSCA E RECUPERAÇÃO DE INFORMAÇÃO TEXTUAL

Rafael Leonardo Siqueira da Silva

Março / 2007

Orientadores: Geraldo Bonorino Xexéo
Jano Moreira de Souza

Programa: Engenharia de Sistemas e Computação

Esta dissertação propõe uma nova abordagem, baseada em técnicas de processamento de sinais, para lidar com a Busca e Recuperação de Informação (BRI) de tipo texto. Propomos a utilização da Transformada Wavelet para manipular o índice usado pelos sistemas de BRI. De fato propomos que é possível transpor um documento do domínio espacial para o domínio wavelet, bem como se faz para um sinal físico qualquer. As principais vantagens dessa nova abordagem são: a redução da dimensão do índice empregado no sistema de BRI, bem como a possibilidade de execução de consultas em diversos níveis de detalhe, devido à propriedade de multiresolução da Transformada Wavelet. Esta propriedade também pode ser usada em ambientes distribuídos como um mecanismo de distribuição da informação. Nossa proposta não se limita apenas à Transformada Wavelet, mas fornece uma estrutura básica, onde outras técnicas de processamento de sinais podem ser usadas. A escolha da Transformada Wavelet como demonstrativo dessa estrutura foi devido a suas características singulares de multiresolução, simplicidade de implementação e baixa complexidade algorítmica, O(N).

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master. of Science (M.Sc.)

SIGNAL MODEL FOR TEXTUAL INFORMATION RETRIEVAL

Rafael Leonardo Siqueira da Silva

March / 2007

Advisors: Geraldo Bonorino Xexéo
Jano Moreira de Souza

Department: Computer and Systems Engineering

This dissertation proposes a novel approach, based on signal processing techniques, to deal with text Information Retrieval (IR). We propose the usage of Wavelet Transform to manipulate the IR systems index. Indeed, we propose that is possible to transpose a document from space domain to wavelet domain, exactly as done with physical signals. The main advantages of this approach are: the reduction of IR systems index dimension, the possibility to make queries in different levels of details, because of the multiresolution property of Wavelet Transform. This property also can be used in distributed environments to disseminate information. Our proposal are not limited to Wavelet Transform, it is actually a more general framework that allow us to apply other signal processing techniques. The Wavelet Transform was chose as demonstration due to its singular features such as: multiresolution, implementation simplicity and low algorithm complexity, O(N).

ÍNDICE:

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

# 1 Introdução

A Busca e Recuperação de Informação (BRI) do tipo texto é o ramo da computação que se dedica ao estudo do armazenamento e recuperação de documentos texto. Seu principal objetivo é recuperar esses documentos de maneira a satisfazer as necessidades do usuário, as quais são expressas por meio de uma consulta (*query*). A consulta é uma representação do desejo do usuário diante do sistema, ela é composta por termos-chaves que são os elementos básicos de sistema. Um sistema de BRI é composto por duas fases: a fase de indexação, quando os índices são criados para futuras consultas; e a fase de consulta, quando o usuário é capaz de informar seus desejos (consulta) ao sistema.

Durante a fase de indexação, criamos índices que são uma representação compacta do documento. Esses índices são armazenados em uma estrutura de dados especial conhecida como índice-invertido. Essa objetiva recuperar rapidamente um documento a partir dos termos que ele contém.

Na fase de consulta, o usuário fornece ao sistema uma consulta que representa o documento que ele deseja recuperar. Essa consulta é comparada aos índices dos documentos previamente armazenados. Os documentos cujos índices sejam considerados semelhantes à consulta são recuperados e ordenados segundo sua semelhança. Durante principalmente a fase de consulta, a representação conceitual de um documento desempenha um papel fundamental. Pois essa representação possibilita que definamos uma função que meça a semelhança entre documento e a consulta.

## *1.1 Objetivo*

O objetivo dessa dissertação é propor e explorar um novo modelo conceitual para representação dos documentos dentro do sistema de BRI. Como veremos mais adiante, existem três representações clássicas de documentos que estão intimamente relacionadas à definição da estrutura matemática de suporte empregada pelo sistema de BRI. Assim nossa proposta pode ser entendida como a utilização de uma nova estrutura matemática de suporte a sistemas de BRI. As estruturas clássicas de representação de documentos são a teoria de conjuntos, a teoria das probabilidades e a álgebra linear. Elas definem respectivamente as seguintes representações para documentos: conjuntos, variável aleatória e vetor. Propomos agora adicionar mais uma estrutura matemática básica: a teoria de processamento de sinais, cuja representação para o documento é um

sinal. Por ser uma proposta muito básica ela possui vários desdobramentos que serão explorados ao longo dessa dissertação. Ao representarmos um documento como sinal, abrimos caminho para aplicação de técnicas aplicadas no campo de processamento de sinais sobre um documento. Essas técnicas possibilitarão novas interpretações e definições métodos de comparações entre consultas e documentos. Por exemplo, o uso das transformadas wavelet aplicada à representação de sinal de um documento, possibilitará o uso da propriedade de multiresolução em sistemas de BRI. Essa propriedade, como veremos, é adequada para analise de documentos em diversos níveis de detalhe e pode ser muito bem explorada em ambientes distribuídos como redes ponto-a-ponto (CHUNQIANG TANG., ZHICHEN XU et al., 2003).

## *1.2 Organização*

Ao longo dessa dissertação construiremos a base necessária para compreensão de nossa proposta. Iniciaremos com uma introdução à BRI no capítulo 2, nele apresentamos a definição formal de um sistema de BRI, bem como os modelos clássicos. No capítulo 3, apresentamos os fundamentos teóricos de processamento de sinais que servirão de base para o entendimento de nossa proposta. Apresentamos principalmente a transformada de Fourier e suas derivações, bem como a transformada Wavelet, que será ponto chave de nossa proposta, particularmente por causa da propriedade de multiresolução. O capítulo 4 apresentará os trabalhos relacionados, com especial destaque para (PARK, RAMAMOHANARAO et al., 2005) que, apesar de independente, tem muitos traços comuns aos propostos nessa dissertação. No capítulo 5 apresentamos nossa proposta, o MSBRI (Modelo de Sinais para BRI). Defenderemos que o MSBRI na verdade se trata de um meta-modelo para BRI, de forma que apresentaremos várias propostas de diferentes instâncias para ele. O capítulo 6 detalha o Modelo Wavelet, a instância do MSBRI que foi desenvolvida para fins de analise e experimentos dessa dissertação. No capítulo 7 avaliamos o Modelo Wavelet e o comparamos com o sistema de BRI comercial conhecido como Lucene (OTIS GOSPODNETIC e ERIK HATCHER, 2007). No capítulo 8 apresentamos as conclusões e os trabalhos futuros propostos para essa dissertação. O Apêndice A aprofunda nossa proposta de implementação de um algoritmo distribuído usando o MSBRI. O Apêndice B apresenta os modelos de documento usados para avaliação do MSBRI.

# 2 Introdução à BRI

Nesse capítulo faremos uma breve introdução à Busca e Recuperação de Informação (BRI). Apresentaremos a definição formal de BRI, o que possibilitará encaixar melhor nossa proposta dentro de uma estrutura teórica. Faremos uma breve distinção entre busca de informação e busca de dados. Finalmente apresentamos os três modelos clássicos de BRI, o modelo booleano, modelo probabilístico e o modelo vetorial.

## *2.1 O que é BRI*

A Busca e Recuperação de Informação (BRI) ramo da computação que se dedica à representação, armazenamento e recuperação de informação. Nesta dissertação, o termo BRI, a não ser que seja especificado diferentemente, será empregado para descrever a Busca e Recuperação de Informação de formato texto. No entanto, existem outras classes de BRI, como por exemplo, BRI em áudio (músicas) (STEPHEN DOWNIE e MICHAEL NELSON, 2000), BRI em imagens (BENG CHIN OOI, KIAN-LEE TAN et al., 1998) e BRI multimídia (CARLO MEGHINI, FABRIZIO SEBASTIANI et al., 2001).

A BRI de formato texto ganhou um grande impulso nos últimos anos devido ao crescimento da internet. Sítios especializados, como Yahoo! (SITIO DA INTERNET, 2007b), Altavista (SITIO DA INTERNET, 2007a), e mais recentemente o Google (YATES B.R. e BERTHIER R.N., 1999) ganharam projeção mundial devido suas "máquinas de busca".

A BRI não se preocupa apenas em encontrar a informação desejada, mas também em mecanismos eficientes de armazenamento. A informação deve ser registrada de maneira que aumente o desempenho de recuperação. Para tal, usamos estruturas de dados especiais conhecidas como arquivos invertidos (ou índices-invertidos) que possibilitam encontrar rapidamente os documentos que contém determinado termo. Outras estruturas secundárias também podem ser usados para casos específicos de busca, como por exemplo, o uso de arvores de sufixo e prefixo para encontrar termos derivados. Existe uma ampla gama de estruturas que podem ser usadas em sistemas de BRI, (YATES B.R. e BERTHIER R.N., 1999) tem mais referencias.

## 2.2 Busca por dados e busca por informação

Por tratar de um conteúdo do qual, a princípio, não se conhece a estrutura, a busca de informação relevante em um sistema de BRI é menos precisa do que em um sistema de busca de dados. Em um sistema de busca de dados, a informação é decomposta em elementos semânticos bem definidos, os dados.

Pare efeitos dessa dissertação, um dado é caracterizado por uma definição de tipo, um nome e um valor. Os dados são organizados logicamente em tabelas que, por sua vez, são agrupadas e gerenciadas por um Sistema Gerenciador de Banco de Dados (SGBD). Tamanha sofisticação possibilita a definição de uma álgebra, conhecida como álgebra relacional, que é capaz de distinguir as características de cada dado registrado no SGBD.

Ao efetuar uma busca em um SGBD, o usuário é capaz de fornecer uma descrição formal de todos os dados que ele deseja recuperar. Desta forma, uma busca em um SGBD é uma definição das restrições que determinados dados devem atender para serem considerados relevantes a uma consulta.

Por outro lado, a busca em um sistema de BRI é feita através de palavras-chaves que muitas vezes são ambíguas e imprecisas, portanto é improvável recuperar apenas documentos relevantes a uma consulta. Em um sistema de BRI o mais importante é a ordenação dos documentos segundo sua relevância em relação à consulta. Este princípio se baseia na suposição de que as palavras-chaves contidas na consulta podem, em certo grau, representar a necessidade do usuário, mas note que nem sempre é fácil expressar uma necessidade em termos de palavras-chaves (Veja Exemplo 1).

## 2.3 Modelos clássicos de BRI

Antes de iniciar uma discussão mais profunda sobre os modelos clássicos de busca e recuperação de informação é preciso encontrar uma formalização para o conceito de modelo. Segundo [15], um modelo de busca e recuperação de informação é uma quíntupla como mostrado na Definição (1)

$$\{D, Q, F, R(q_i, d_j)\} \tag{1}$$

Onde,

D é o conjunto composto por todas as representações dos documentos na coleção.

Q é o conjunto de todas as consultas executadas pelo usuário de um sistema de busca e recuperação de informação. Estas consultas são chamadas de queries.

F é a estrutura matemática (*framework*) usada para modelar os documentos e consultas, assim como seu relacionamento.

$R(q_i, d_j)$ é uma função de ordenação (*ranking*), ou similaridade, que associa um valor real a uma consulta $q_i \in Q$ e a uma representação $d_j \in D$. Esta função determina a ordem de relevância dos documentos em relação à consulta.

A estrutura (*framework*) deve ser escolhida de maneira que forneça um método intuitivo para modelar os documentos e consultas, assim como estabelecer uma função de classificação (*ranking*). Por exemplo, no modelo vetorial, tanto os documentos quanto as consultas são considerados como vetores, enquanto a função de ordenação (*ranking*) é dada pelo ângulo entre estes vetores.

Nas seções 2.3.1, 2.3.2 e 2.3.3 serão apresentados os modelos clássicos. Eles são chamados assim porque todos os demais modelos são extensões e/ou aperfeiçoamentos feitos sobre estes modelos. Cada um destes modelos se baseia em uma estrutura matemática (*framework*) diferente que o caracteriza, mas todos definem claramente uma função de ordenação (*ranking*) capaz de ordenar os documentos da coleção segundo sua relevância em relação a uma consulta. O modelo booleano é baseado em álgebra de conjuntos, o modelo vetorial se baseia em álgebra linear, enquanto o modelo probabilístico se baseia em teoria de probabilidades, em especial no teorema de Bayes .

Apesar das diferenças entre as estruturas e suas funções de classificação, todos os modelos de Busca e Recuperação de Informação no texto têm uma característica em comum, a saber: todos consideram os documentos como uma coleção de termos. Estes termos, quando colocados em determinada ordem dentro de um documento, dão-lhe uma semântica, um sentido aos olhos humanos. Um computador é incapaz de capturar esta semântica, mas por outro lado ele pode associar pesos a cada um dos termos de forma a valorizar determinados termos em detrimento de outros. A questão fundamental para qualquer modelo de BRI é como definir tais pesos. Uma maneira prática de fazer isso é associar um peso maior a termos mais raros, pois estes contêm maior poder de

discriminação. Por exemplo, suponha uma coleção de 1000 documentos. Se o termo t aparece em todos os documentos, então t não tem nenhum poder de distinção. O termo t não é útil como termo-índice porque não caracteriza nenhuma particularidade de qualquer documento. Por outro lado, suponha que t apareça apenas em dois documentos entre os mil da coleção. Então, t seria um ótimo termo-índice, uma vez que ele reduziria gigantescamente o espaço de busca por documentos contendo t. Para ilustrar a concepção acima, considere o Exemplo 1

Exemplo 1.

Seja a coleção de documentos:

D1 = (A, B, A, B, A)

D2 = (A, A, A, A, A)

D3 = (B, B, B, B, B)

D4 = (A, C, B, C, A)

D5 = (A, A, B, B, B)

Portanto, segundo a Definição (1), D= (D1,D2,D3,D4,D5) e T=(A,B,C), onde T é a coleção de termos presentes nos documentos. Para esta coleção o termo "A" tem pouco poder discriminatório, uma vez que ele aparece em quatro dos cinco documentos da coleção. Enquanto "C" tem grande poder discriminatório, pois ele somente aparece no documento D4. Uma consulta q=(C), por exemplo, eliminaria de imediato todos os documentos exceto D4. Portanto, um sistema de BRI deve atribuir um peso maior ao termo C do que o termo A.

### 2.3.1 Modelo booleano

O modelo booleano é baseado na teoria dos conjuntos, sendo o mais simples dentre os três modelos clássicos de busca e recuperação de informação. Sua principal característica é basear-se precisamente no formalismo da lógica booleana. Neste modelo, o peso de um termo-índice é binário, 1 ou 0, indicando respectivamente a presença ou a ausência do termo no documento. As consultas realizadas neste sistema são expressões booleanas ligadas pelos operadores AND, OR e NOT.

Exemplo 2.

Seja a coleção definida no Exemplo 1. No modelo booleano os documentos são representados pela presença ou ausência dos termos. Portanto:

D1 = (1,1,0)
D2 = (1,0,0)
D3 = (0,1,0)
D4 = (1,1,1)
D5 = (1,1,0)

Uma consulta q da forma q=(A & (B | ~C)) é mapeada na forma normal disjuntiva em $q_{fnd}$=(1,1,1) | (1,1,0) | (1,0,0), onde &, | e ~, indicam respectivamente as operações "E", "OU" e "NÃO" definidos na álgebra booleana. O diagrama de Venn da Figura 1 ilustra a visão conceitual da consulta. Neste exemplo os documentos recuperados são D1, D2, D4 e D5. Ou seja, são os documentos contidos nas interseções dos conjuntos representados no diagrama. A conjunção (1,1,0) indica a interseção entre os conjuntos A e B, como ela está presente na consulta os documentos presentes nesta interseção são considerados relevantes.

Os conjuntos A, B e C contêm os documentos que possuam o respectivo termo. Conceitualmente, cada termo dá origem a um conjunto que abriga todos os documentos no qual aquele termo aparece pelo menos uma vez.

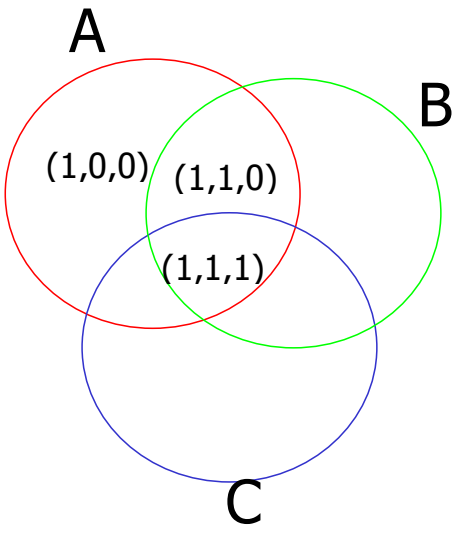



**Figura 1 diagrama de Venn da Consulta q=(A & (B | ~C))**

### 2.3.1.1 A função de similaridade

Uma parte essencial de qualquer modelo de BRI é a definição da função de similaridade que possibilita a comparação entre consultas e documentos. Segundo (YATES B.R. e BERTHIER R.N., 1999), no modelo booleano os termos-índices são todos variáveis binárias, ou seja, dado um termo t, ele está ou não presente em um documento. Uma consulta q é definida como uma expressão booleana convencional. Seja $q_{fnd}$ a forma normal disjuntiva da consulta q, e $q_{cc}$ qualquer componente conjuntivo de q. A função de similaridade entre o documento $d_j$ e a consulta q é:

$$sim(d_j, q) = \begin{cases} 1, \exists \vec{q}_{cc} \mid (\vec{q}_{cc} \in \vec{q}_{fnd}) \wedge (\forall k_i, g_i(\vec{d}_j) = g_i(\vec{q}_{cc})) \\ 0, otherwise \end{cases} \quad (2)$$

Onde,

$K_i$ é o i-ésimo termo do conjunto K=(A,B,C).

$g_i$ é uma função binária que associa um peso a um documento baseada na presença de uma componente conjuntiva.

Apesar de sua simplicidade e forte embasamento teórico, o modelo booleano sofre de uma séria limitação: a função de similaridade definida acima não possibilita

graduação no nível relevância do documento em relação à consulta. Portanto, um documento ou é relevante para consulta ou não é relevante para consulta. A graduação no nível de similaridade é introduzida pelo modelo vetorial (*VSM - Vector Space Model*) como será visto mais adiante.

### 2.3.2 Modelo probabilístico

O modelo probabilístico se baseia em uma estrutura matemática bastante conhecida, a teoria de probabilidades. Neste modelo tanto documento quanto a consulta são considerados eventos em um espaço amostral, de forma que o objetivo da função de similaridade é calcular a probabilidade de o documento $d_j$ ser relevante para consulta $q_i$. Vejamos como é definida a função de similaridade.

#### 2.3.2.1 A função de similaridade

Neste modelo os pesos dados aos termos-índice são todos binários. Uma consulta q é um subconjunto dos termos. R é o subconjunto dos documentos sabidamente relevantes e, ~R é o complemento de R. Seja $P(R \mid d_j)$ a probabilidade do documento $d_j$ ser relevante a consulta q, enquanto $P(\sim R \mid d_j)$ a probabilidade de $d_j$ não ser relevante a q. A função de similaridade sim(dj,q) entre o documento e a consulta é definida como:

$$sim(\vec{d}_j, q) = \frac{P(R \mid \vec{d}_j)}{P(\sim R \mid \vec{d}_j)} \tag{3}$$

Por Bayes temos,

$$sim(\vec{d}_j, q) = \frac{P(\vec{d}_j \mid R) \times P(R)}{P(\vec{d}_j \mid \sim R) \times P(\sim R)} \tag{4}$$

Onde $P(d_j \mid R)$ é a probabilidade de selecionarmos aleatoriamente o documento dj da coleção R de documentos relevantes. P(R) é a probabilidade de que um documento qualquer retirado aleatoriamente de toda a coleção seja relevante. $P(d_j \mid \sim R)$ e P(~R)

são os complementos da definição anterior. P(R) e P(~R) são constantes para toda a coleção e podem ser ignorados sem perda de generalidade. Então se tem:

$$sim(\vec{d}_j, q) = \frac{P(\vec{d}_j \mid R)}{P(\vec{d}_j \mid \sim R)} \tag{5}$$

A equação (5) calcula as chances de um documento está entre os documentos relevantes para a consulta q. O modelo probabilístico age de maneira incremental, a equação (5) é realimentada a cada rodada e refina os resultados, até chegar ao ponto em que o grau de refinamento seja considerado suficiente.

A principal vantagem do modelo probabilístico em relação ao modelo booleano é que a função de similaridade, apesar de baseada em pesos binários, fornece uma saída não binária que reflete a relevância do documento para determinada consulta.

### 2.3.3 Modelo vetorial

No modelo vetorial (*Vector Model Space - VSM*) (SALTON GERARD, 1989), ao contrário do que ocorre com o modelo booleano, o peso dado a um determinado termo-índice não é binário. Ao contrário, ele é uma variável real que reflete a grau de relevância do termo dentro do documento. Neste modelo, tanto o documento quanto a consulta são considerados vetores sobre o espaço de termos. Em geral, este modelo trabalha com vetores normalizados de forma que os cálculos de similaridade fiquem mais simples. A função de similaridade $sim(d, q)$ entre uma consulta e um documento é definida como o ângulo entre a representação vetorial do documento e a representação vetorial da consulta. De forma que quanto menor o ângulo entre os dois vetores mais similar é o documento à consulta. Vamos a um exemplo para tornar a idéia mais nítida.

Exemplo 3.

> Por simplicidade nesse exemplo, abribuimos pesos aos termos baseados meramente em suas freqüências absolutas dentro dos documentos. No entanto, em sistemas reais de BRI, costumamos usar

uma heurística conhecida como TF-IDF (SALTON GERARD, 1989) para atribuir pesos aos temos dentro do documento.

Seja a mesma coleção empregada no Exemplo 1, e a mesma consulta usada no Exemplo 2. No modelo vetorial, os documentos são codificados da seguinte maneira:

D1=(3,2,0)
D2=(5,0,0)
D3=(0,5,0)
D4=(2,1,2)
D5=(2,3,0)

A consulta q=(A & (B | ~C)), é representada neste modelo como q'=(1,1,0). O ângulo entre q' e cada um dos DX é apresentado abaixo.

Sim(D1,q') = 11°30'
Sim(D2,q') = 45°
Sim(D3,q') = 45°
Sim(D4,q') = 45°
Sim(D5,q') = 11°30'

Os documentos podem ser ordenados segundo sua similaridade (menor ângulo), em nosso exemplo a consulta q retornará D1, D5, D2, D3, D4. Onde D1 é o documento mais relevante e D4 o menos relevante.

#### 2.3.3.1 A função de similaridade

Formalmente o modelo vetorial define um conjunto de pesos $w_{i,j}$, não binários, associados ao par $(d_j, q_i)$. Estes pesos são fornecidos pela relevância de cada termo dentro do conjunto considerado. Assim temos que $q_i = (w_{1,i}, \cdots, w_{t,i})$ onde t=|T|, da mesma forma $d_j = (d_{1,j}, \cdots, d_{t,j})$.

Por fim, a função de similaridade é definida como o cosseno entre os vetores que representam o documento e a consulta. Desta forma temos:

$$sim(d_j, q) = \frac{\vec{d}_j \bullet \vec{q}}{\left|\vec{d}_j\right| \times \left|\vec{q}\right|} = \frac{\sum_{i=1}^{n} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w^2{}_{i,j}} \times \sqrt{\sum_{j=1}^{t} w^2{}_{i,q}}} \quad (6)$$

Note que a fórmula expressa em (6) é simplesmente o produto escalar (produto interno) entre os vetores que representam a consulta e o documento, dividido pela norma dos mesmo.

O modelo vetorial (*Vector Model Space - VSM*) por sua simplicidade e eficiência é amplamente usado em diversos sistemas de BRI. Aqui ressaltamos o SMART (BUCKELEY, SINGHAL et al., 1995)proposto e desenvolvido por Salton na década de 70 e que serve de referencia até hoje para muitos sistemas de BRI.

Uma característica importante do modelo vetorial é o fato de sua função de similaridade ter uma característica contínua, não binária. Desta forma é possível associar a um documento um grau de relevância em relação a uma consulta.

# 3 Fundamentos Teóricos

Neste capitulo apresentaremos os fundamentos teóricos que nos ajudarão a compreender os principais conceitos por trás da proposta dessa dissertação. Nossa ferramenta básica será a álgebra linear que fornece mecanismos de calculo e interpretação poderosos para compreendermos as transformadas que serão apresentadas em seguida. Nós daremos especial destaque para a transformada de Fourier, que por sua vez nos servirá de base para compreendermos a transformada Wavelet e em particular o efeito de multiresolução, em torno do qual gira a proposta fundamental deste trabalho.

## *3.1 Álgebra linear*

A álgebra linear fornece um mecanismo muito eficiente de representação e operação de vetores. Ela lida com operadores lineares e fornece uma série de teoremas que fundamentarão discussões futuras. Este capítulo não tem a intenção de esgotar o assunto, ela apenas apresenta alguns resultados que serão importantes para o tema desta dissertação. Existem muitos bons livros sobre o assunto que podem ser consultados, particularmente (BOLDRINE, 2005) e (STRANG, 2005).

### 3.1.1 Espaço vetorial

Um espaço vetorial complexo (C) ou real (R) é um conjunto de vetores **E**, associados às operações de soma e multiplicação escalar, que para qualquer **x**, **y** pertencentes a **E**, e **a**, **b** reais ou complexos, satisfazem:

(a) Comutatividade: **x**+**y** = **y**+**x**.
(b) Associatividade: (**x**+**y**)+**z** = **x**+(**y**+**z**), (**ab**)**x** = **a**(**bx**).
(c) Distributividade: **a**(**x**+**y**) = **ax**+**ay**.
(d) Identidade aditiva: Existe 0 pertencente a **E**, tal que 0+**x**=**x** para todo **x** pertencente a **E**.
(e) Inverso aditivo: para todo x pertencente a **E**, existe (-**x**) tal que **x**+(-**x**)=0.
(f) Identidade multiplicativa: Existe 1, tal que 1**x**=**x** para todo **x** pertencente a **E**

### 3.1.2 Subespaço vetorial

Um subespaço vetorial é um subconjunto **S** de um espaço vetorial **E**, que ainda é um espaço vetorial. Formalmente, S é subespaço de **E** se:

(a) Para todo **x** e **y** pertencente a **S**, **x**+**y** pertencer a **S**.
(b) Para todo **x** pertencente a **S** e **a** pertencente a **R** ou **C**, **a***x** pertencer a **S**.

### 3.1.3 Base de um espaço vetorial

Uma base para um espaço vetorial **E**, é um conjunto de vetores **B**, para todo **x** pertencente a **E**, **x** pode ser expresso como combinação linear dos vetores de **B**. Formalmente:

$$Base(E) = \left\{ \sum_{i=1}^{n} a_i x_i \,\middle|\, a_i \in C \vee R, x_i \in B \right\} \tag{7}$$

Os elementos de **B** formam uma base para o espaço **E**. Se os elementos de **B** forem todos unitários, então é dito que **B** é uma base ortonormal de **E**.

### 3.1.4 Produto escalar (interno)

O produto escalar, ou produto interno, sobre um espaço vetorial E é uma função < . , . >, definida em $E^2$ com as seguintes propriedades.

(a) < **x** + **y** , **z** > = < **x** , **z** > + < **y** , **z** >.
(b) < **x** , **ay** > = **a***< **x** , **y** >, onde *denota o complexo conjugado.
(c) < **x** , **y** >* = < **y**, **x** >, onde * denota o complexo conjugado.
(d) < **x**, **x** > = 0 <=> **x** = 0.

As propriedades (a) e (b) implicam em linearidade, ou seja, o produto escalar é um operador linear. Ao definir a propriedade (c) foi usado o complexo conjugado (*), com a propriedade de que se **a** pertence a R então **a***=**a**, se a pertence a C, então se **a**=**b**+**i**.**c**, onde i é o elemento imaginário, então **a*** = b-**i.c**.

### 3.1.5 Norma

A noção de norma é fundamental, ela é uma abstração do conceito de tamanho. A norma de um vetor é denotada por || . || e ela pode ser definida em função do produto escalar da seguinte maneira:

$$\|x\| = \sqrt{\langle x, x \rangle} \tag{8}$$

Desta forma a distância entre dois vetores é definida como a norma de sua diferença, ou ||x-y||.

### 3.1.6 Ortogonalidade

O conceito de produto escalar também pode ser usado para definir ortogonalidade entre dois vetores. Dois vetores x e y são ortogonais, se e somente se:

$$\langle x, y \rangle = 0 \tag{9}$$

### 3.1.7 Espaço de Hilbert

Um espaço vetorial completo equipado com o produto escalar < . , . > para todos os seus vetores é um espaço de Hilbert (MARTIN VETTERLI e JELENA KOVACEVIC, 1995).

### 3.1.8 Espaço das seqüências de somas quadráticas

Nesta dissertação nosso interesse será quase que exclusivamente em seqüências do tipo x[n] cuja soma dos quadrados seja finita, ou em outras palavras, cuja energia seja finita (em física, a soma ou a integral do quadrado de uma função quase sempre corresponde à energia). Portanto, x[n] é um elemento de um espaço de Hilbert $l^2(Z)$, onde n é um elemento de Z. Neste espaço definimos o produto escalar

$$\langle x, y \rangle = \sum_{n=-\infty}^{\infty} x[n] y[n] \tag{10}$$

E a norma,

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{\sum_{n \in Z} (x[n])^2} \tag{11}$$

Portanto, $l^2(Z)$ é o espaço das seqüências $\|x\| < \infty$. Este espaço é obviamente de seqüências finitas cuja base do espaço vetorial é definida como δ[n-k], função delta de Dirac.

As operações feitas no espaço de seqüências podem ser estendidas para o caso contínuo. Neste caso, ao invés do vetor x[n], operamos sobre funções complexas f(t). Mas as definições anteriores continuam válidas com pequenos ajustes. Portanto, é possível se mostrar que uma função f(t) definida em R pertence ao espaço de Hilbert $L^2(Z)$, se $\|f(t)\|^2$ é integrável e finito. Ou seja:

$$\sqrt{\int_{t \in T} \|f(t)\|^2 \, dt} < \infty \tag{12}$$

Neste caso, o produto escalar entre duas funções é definido como

$$\langle f, g \rangle = \int_{t \in R} f(t) g(t) dt \tag{13}$$

Enquanto a norma é

$$\|f\| = \langle f, f \rangle^{1/2} = \sqrt{\int_{t \in R} |f(t)|^2 dt} \tag{14}$$

Como podemos perceber a álgebra linear fornece mecanismos de interpretação para os vetores tanto para o caso discreto quanto para o caso contínuo, esta caracteristica a torna uma ferramenta importante para o entendimento das transformadas.

### 3.1.9 Bases ortogonais e melhor aproximação

Como na álgebra linear, as bases ortogonais desempenham um papel muito importante no espaço de Hilbert, pois elas possibilitam a representação não redundante de um vetor (função ou sinal).

Dado um vetor y no espaço de Hilbert E, sempre é possível projetá-lo sobre o subespaço S e formar um novo vetor $y_p$, por meio da operação (15).

$$\mathrm{y_p} = \sum_i \langle \mathrm{x_i}, \mathrm{y} \rangle x_i \tag{15}$$

A Figura 2 apresenta uma projeção ortogonal sobre um subespaço vetorial. Por simplicidade aqui é adotado um vetor y em $\mathrm{R}^3$ e sua projeção $y_p$ em $R^2 = \{x_1, x_2\}$.

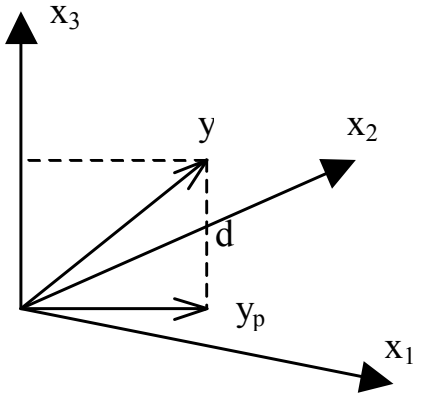



**Figura 2**

Esta ilustração mostra que $d = y - y_p$, note que d é ortogonal a $\{x_1, x_2\}$, e em particular d é ortogonal a $y_p$, como sabemos (Teorema de Pitágoras, no caso particular) que

$$\|y\|^2 = \|y_p\|^2 + \|d\|^2 \tag{16}$$

Note que esta aproximação é ótima no sentido de mínimos quadrados, uma vez que d representa o erro quadrático da aproximação de y por $y_p$. Ou seja,

$$\min(\|y - x\|) \tag{17}$$

Onde x pertence ao subespaço S e pode ser gerado por

$$x = \sum_i a_i x_i \tag{18}$$

Onde $a_i$ é a projeção de y sobre o respectivo eixo $x_i$, matematicamente:

$$a_i = \langle x_i, y \rangle \tag{19}$$

Desta forma podemos definir recursivamente projeções em subespaços vetoriais, de maneira que um vetor $y_n$ em $R^n$, seja projetado em vetores $y_{n-m}$ em $R^{n-m}$. Note que esta projeção vale para qualquer espaço vetorial definido sobre bases ortogonais, não só para $R^n$ como apresentado no na Figura 2.

## 3.2 Processamento de sinais

O processamento de sinais é o campo da engenharia que lida com a análise e síntese de sinais. Ela pode ser dividida em dois grandes subcampos complementares, o processamento de sinais analógicos e o processamento de sinais digitais. Entendemos por um sinal qualquer elemento do espaço de Hilbert como apresentado na seção 3.1.7. Portanto, do ponto de vista matemático, um sinal é qualquer função que obedeça a equação (12).

Historicamente o processamento de sinais lida com ondas de natureza física, como ondas eletromagnéticas ou ondas mecânicas. Portanto, é natural que haja uma interpretação física coerente com a definição matemática apresentada anteriormente.

A equação (12) quando usada para representar um sinal físico, como uma onda mecânica, por exemplo, é interpretada como a energia total da onda. Portanto, outra maneira de interpretar um sinal é qualquer função de onda que contenha energia finita.

No restante deste capítulo trataremos das técnicas mais importantes empregadas em processamento de sinais. Na seção 3.2.4 serão apresentadas as técnicas de amostragem usadas para converter um sinal analógico em digital e vice-versa. Tais técnicas servirão de fundamento para dedução das propriedades das equações do processamento de sinais digitais. Na seção 3.3 será apresentada a transformada de Fourier, uma técnica matemática que possibilita transição do domínio temporal para o domínio de freqüência de um sinal. Em seguida, será apresentada a transformada de Fourier com janelamento, ou *Short Time Fourier Transform*, e nesta seção será introduzida a noção de resolução. Na seção 3.4 apresentamos a transformada wavelet como uma evolução da transformada de Fourier com janelamento. Na seção 3.4.4.1 é fundamentada a multiresolução na análise de sinais, assim terminamos os fundamentos de processamento de sinais necessários para a compreensão da proposta desta dissertação.

### 3.2.1 Processamento de sinais analógicos

Sinais analógicos são aqueles contínuos no tempo. Portanto, uma representação matemática adequada é através de uma função f(t), onde t é uma variável real. As

técnicas de análise e síntese de sinais analógicos são de grande valia para a engenharia, elas possibilitam o estudo das estruturas que compõe o sinal. Mas para os fins desta dissertação estas técnicas são úteis apenas para derivarmos as propriedades do caso discreto, que é nosso verdadeiro interesse.

### 3.2.2 Processamento de sinais digitais

Os sinais digitais têm uma natureza discreta. Um sinal analógico pode ser convertido em um sinal digital através de técnicas de amostragem que garantem a posterior reconstrução do sinal. Após a segunda guerra mundial, com a advento dos computadores digitais, o processamento de sinais digitais ganhou muita força, pois os sinais digitais são mais facilmente armazenados e processados nos modernos computadores.

Hoje as técnicas de processamento de sinais digitais não se limitam apenas a sinais físicos, mas são aplicadas em muitas outras áreas, como estatística, computação gráfica, teoria da informação.

Para fins desta dissertação o processamento de sinais digitais é de maior valia devido à natureza discreta das funções que serão empregadas, mas o processamento de sinais analógicos ainda se faz necessário pois possibilita a derivação de diversas propriedades presentes nos sinais digitais.

### 3.2.3 O que é um sinal

Antes de iniciar uma discussão mais profunda é necessário definir precisamente a que nos referimos exatamente quando mencionamos o termo sinal. Portanto, é preciso que se defina formalmente o que é um sinal. Para efeito deste trabalho, adotaremos a definição fornecida por (MARTIN VETTERLI e JELENA KOVACEVIC, 1995).

Do ponto de vista matemático um sinal é qualquer função matemática que pertença ao espaço de Hilbert, ou seja, que possua produto interno definido (finito). Fisicamente, o produto interno de uma função é interpretado como a energia desta função. Portanto, ao se dizer que uma função tem produto interno definido (finito), na verdade está se dizendo que ela possui uma quantidade de energia definida (finita). Esta idéia é expressa matematicamente por (20).

$$\langle f(t), f(t) \rangle^{1/2} = \sqrt{\int_{t \in T} |f(t)|^2 \, dt} < \infty \tag{20}$$

Onde,

f(t) é uma função complexa; t pertence ao conjunto dos números real e <,> é o operador de produto interno.

Em particular, neste trabalho lidaremos não com funções contínuas em t, mas com funções discretas, ou seja, com parâmetro t pertencente aos números inteiros, como mostrado em (21).

$$\langle f[n], f[n] \rangle^{1/2} = \sqrt{\sum_{n \in Z} (f[n])^2} < \infty \tag{21}$$

Onde,

f[n] é uma função complexa discreta e n é o parâmetro inteiro, novamente <,> é o operador de produto interno.

Note que a condição (21) não é uma restrição muito forte e qualquer função contínua, limitada e restrita em um intervalo finito a respeita. Portanto, matematicamente, a definição de sinal é bastante ampla. É devido a esta amplitude que o campo de processamento de sinais tem diversas aplicações nos mais diferentes contextos.

Apresentamos aqui duas definições matemáticas para o conceito de sinal, uma contínua e outra discreta. Este não é um fato isolado. Na verdade, todo o campo de processamento de sinais pode ser dividido em duas grandes subcampos: o processamento de sinais analógicos e o processamento de sinais digitais. Ao contrário do que possam parecer, estes dois campos são complementares. Em sistemas reais eles agem juntos para alcançar melhor qualidade na análise de sinais físicos.

### 3.2.4 A amostragem (*sampling*)

A conversão de um sinal analógico (contínuo) para um sinal digital (discreta) acontece por um processo de amostragem (sampling). A amostragem é uma técnica que permite que um sinal contínuo seja representado apenas por alguns (finitos) pontos, de forma que posteriormente possa ele possa ser reconstruído.

Para que a reconstrução do sinal seja possível é necessário que um número mínimo de pontos do sinal seja amostrado. Este número mínimo é baseado na freqüência máxima do sinal e é conhecido como freqüência de amostragem de Nyquist, ou simplesmente freqüência de Nyquist. A equação (22) apresenta a relação entre a freqüência de amostragem e a freqüência máxima do sinal.

$$\omega_s > 2\omega_0 \tag{22}$$

Onde,

$\omega_s$ é a freqüência de amostragem de Nyquist

$\omega_0$ é a freqüência máxima do sinal

Em palavras podemos dizer que a freqüência de amostragem tem que ser, no mínimo, duas vezes maior que a freqüência máxima de um sinal. De outra forma, para cada período do sinal original temos que ter dois pontos de amostragem.

Apesar de simples a equação (22) é um resultado fundamental e poderoso, pois ela define o critério de amostragem de um sinal. Se tal critério não for respeitado, então a reconstrução não pode ser feita.

#### 3.2.4.1 A função δ

A amostragem de um sinal é feita com o auxilio da função impulso δ (delta) de Dirac, como definida em (23)

$$\begin{aligned} &\int_{-\infty}^{\infty} \delta(t)dt = 1 \\ &\delta(t) = 0, t \neq 0 \end{aligned} \tag{23}$$

A função δ(t) é definida como zero para todo t, exceto na origem, onde tem uma área unitária. Geometricamente a função δ(t) é concebida como um pulso estreito, com largura 1/ε. No limite $\varepsilon \to 0$ e a altura h tende a infinito.

O produto de um sinal (função) f(t) por δ(t) será igual ao valor do sinal na origem (f(0)), uma vez que δ(t) é sempre zero para qualquer valor de t≠0. Podemos usar uma variável

auxiliar T, de tal forma que δ(t) seja descolado da origem, possibilitando assim a amostragem de um ponto T qualquer do sinal original.

$$\int_{-\infty}^{\infty} f(t)\delta(t-T)dt = f(T) \qquad (24)$$

A equação (24) fornece um mecanismo matemático que possibilita a amostragem de qualquer ponto de um sinal. Este mecanismo envolve multiplicar o sinal f(t) pela função impulso δ(t) e, em seguida, integrar este produto para todo t.

Exemplo 4

A Figura 3 está dividida em três partes, na parte superior é apresentada uma senoide cuja freqüência é de aproximadamente 1,5Hz. Na parte central é mostrada a representação discreta desta mesma função. Mas note que a freqüência de amostragem é de apenas 1 Hz, ou seja, menor do que os 3 Hz previstos pela freqüência de amostragem de Nyquist. A terceira e última da figura mostra a reconstrução do sinal. Note como o sinal obtido não é idêntico ao original. A este efeito de distorção por falha no processo de amostragem denominamos aliasing.

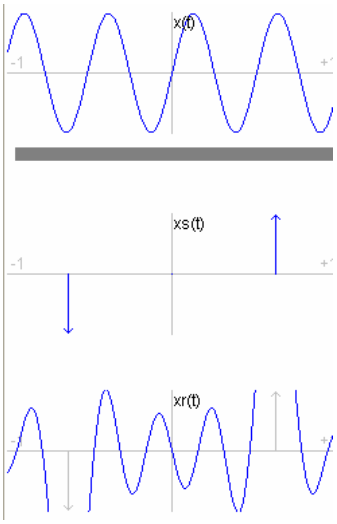



**Figura 3**

Está fora do escopo desta dissertação apresentar ou discutir o efeito de *aliasing*, mas (DINIZ, 2006) apresenta explicações completas e uma ampla discussão sobre o assunto.

### 3.2.5 O que é uma transformada

Uma transformada é um operador linear que promove a mudança de domínio de um sinal (função). O domínio de um sinal (função) refere-se ao domínio de sua variável independente. Particularmente nesta dissertação quando nos referimos a uma transformada estamos, na verdade, tratando sobre uma transformada integral. Qualquer transformada integral tem forma como definida em (25).

$$T(f)(\omega) = \left\langle f(t), K^*(\omega,t) \right\rangle = \int_{t1}^{t2} K(\omega,t) f(t) dt \qquad (25)$$

Onde,

K(ω,t) é chamada de função núcleo (*kernel*)

f(t) é a função (sinal) original

T é o operador linear que representa a transformada

<.,.> é o operador de produto escalar

Note que a integral do lado direito da igualdade (25) é função de t, enquanto K é função de ω e t. Portanto, o resultado desta integral ele será função de ω. A este efeito chamamos de mudança do domínio t para o domínio ω. A função núcleo, K, determina as características da transformada. A interpretação dada à mudança de domínio depende de K.

Para alguns tipos especiais de K, é possível se definir $K^{-1}$, a transformada de inversa. Esta transformada possibilita o retorno ao domínio original. A equação (26) o esquema genérico de uma transformada integral inversa.

$$f(t) = \left\langle T(f)(\omega), K^{'}(\omega,t) \right\rangle = \int_{\omega 1}^{\omega 2} K^{'}(\omega,t) T(f)(\omega) d\omega \qquad (26)$$

Onde,

K é a função núcleo

$K'$ é o inverso de K

f é a função original no domínio temporal

T é o operador linear que representa a transformada

<,> é o operador de produto escalar

A motivação para criação das transformadas integrais foi a possibilidade de mudança de domínio. Muitas operações complexas em um domínio são simplificadas em outros. Neste sentido uma transformada integral tem o mesmo efeito que o operador logarítmico, que possibilita operações como produto, torne-se simples adições quando no "domínio" logarítmico.

### 3.2.6 A ortogonalidade da base

Quando K é ortogonal, ou seja, quando o produto escalar de K por K' é igual a função delta de dirac, dizemos que K forma uma base ortogonal para um determinado domínio. Matematicamente (27) apresenta esta concepção.

$$\left\langle K(\omega_1,t), K'(\omega_2,t)\right\rangle = \int_{t1}^{t2} K(\omega_1,t)K'(\omega_2,t)d\omega = \delta(\omega_1-\omega_2) \qquad (27)$$

Todas as propriedades de bases ortogonais apresentadas na seção 3.1.6 são válidas aqui também. Este é um resultado muito importante, porque ele possibilita que interpretemos (26) como a projeção de f(t) em K(ω,t). Em outras palavras, um transforma é apenas uma mudança de eixo ortogonal, é uma nova representação diferente para uma mesma função, no mesmo sentido de uma projeção ortogonal.

A precursora de todas as transformadas foi a expansão de funções por séries de Fourier que possibilita a representação de uma função em termos de senoides de freqüências diferentes. A série de Fourier veio, mais tarde, a dar origem à transformada de Fourier, que hoje é a transformada mais empregada. Na seção 3.3 será apresentada mais detalhadamente a transformada de Fourier.

## 3.3 A transformada de Fourier

A transformada de Fourier é uma transforma integral cujo núcleo é como definido em (28), esta relação é conhecida como Fórmula de Euler em homenagem ao grande matemático alemão de mesmo nome.

$$e^{it\omega} = \cos(\omega t) + isen(\omega t) \tag{28}$$

Onde,

e é a base dos logaritmos naturais,

ω é a freqüência das senoides em radianos, e

t é uma variável independente geralmente interpretada como o tempo.

Uma interpretação geométrica útil para a equação (28) é que ela representa uma circulo unitário, descrito em coordenadas polares, no plano imaginário. Esta interpretação possibilita o entendimento de diversas propriedades da transformada de Fourier que estão fora do escopo desta dissertação.

A equação (29) apresenta a forma matemática da transformada de Fourier. Note que por ser integrada para todo t, seu resultado é função de ω.

$$\left\langle f(t), e^{it\omega_0} \right\rangle = F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-it\omega_0} dt$$
$$t \in \Re \tag{29}$$

Onde,

F é a representação do operador

$e^{-it\omega}$ é a fórmula de Euler

A equação (30) apresenta a transformada inversa de Fourier, com ela é possível reconstruir a função original, ou seja, voltar do domínio de freqüência para o domínio do tempo.

$$f(t) = \int_{-\infty}^{\infty} F(\omega) e^{it\omega_0} d\omega \qquad (30)$$
$$\omega \in \Re$$

Onde,

$e^{it\omega}$ é o complexo conjugado de $e^{-it\omega}$

### 3.3.1 Série de Fourier

A transformada de Fourier é derivada da série de Fourier, que foi a precursora de todas as transformadas. Joseph Fourier (1768 - 1830) foi um matemático e físico francês que iniciou desenvolveu um mecanismo para resolver equações de calor baseado em séries matemáticas. Hoje esta série é conhecida como série de Fourier.

Formalmente Fourier estabeleceu o seguinte: qualquer função contínua e periódica em R pode ser representada como a soma de senos e cossenos de diferentes freqüências. A Figura 4 apresenta o esquema gráfico desta idéia. A função dente de serra da figura pode ser obtida a partir da soma de diferentes senos, de forma que quanto mais senos forem somados, mais precisa fica a nova representação.

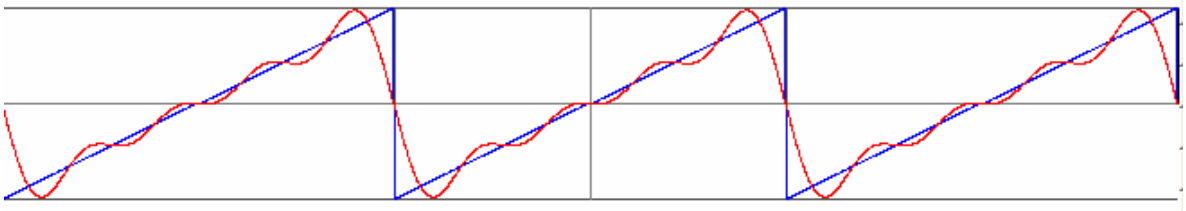


**Figura 4 Função dente de serra aproximada por uma série de Fourier**

Matematicamente a série de Fourier é representada pela equação (31)

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos(\omega_n t) + b_n sen(\omega_n t)) \qquad (31)$$

Onde,

f é uma função periódica com período T=t2-t1,

$$\omega_n = n\frac{2\Pi}{T},$$

$$a_n = \frac{2}{T}\int_{t_1}^{t_2} f(t)\cos(\omega_n t)dt,$$

$$b_n = \frac{2}{T}\int_{t_1}^{t_2} f(t)sen(\omega_n t)dt$$

Os coeficientes $a_n$ e $b_n$ são respectivamente conhecidos como coeficientes de Fourier pares e impares da função f. Eles determinam o quão forte é a influência de determinada freqüência ω na composição da função f.

Outra forma de expressar expansão por serie de Fourier é utilizando a função (28), de maneira que obtemos (32)

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{iwt} \tag{32}$$

Onde,

$$c_n = \frac{1}{T}\int_{t1}^{t2} f(t)e^{-iw_n t}$$

A principal vantagem da representação exponencial da série de Fourier é sua elegância. Além de ser mais compacta e simples, ela reduz os coeficientes $a_n$ e $b_n$ em apenas $c_n$. Novamente este coeficiente representa a contribuição de uma senóide de freqüências ωn para a formação da função f(t). Podemos organizar estes coeficientes em um plano cartesiano de forma a obter o que é conhecido como espetro discreto de freqüências da função f(t).

Exemplo 5.

O espectro discreto de freqüências de uma função (sinal) f(t) consiste em dispor em um plano cartesiano todos os coeficientes, $c_n$, da série de Fourier associados com o seu valor. O eixo x do plano deve conter todos os coeficientes listados na ordem que aparecem na expansão da série. No eixo y, cada coeficiente deve está associado com o seu valor.

Desta forma obtemos uma espécie de histograma de freqüências e suas respectivas contribuições para a formação da função original.

A Figura 5 apresenta uma função de onda quadrada aproximada por uma série de Fourier. Esta aproximação foi feita a partir da soma de três coeficientes da série.

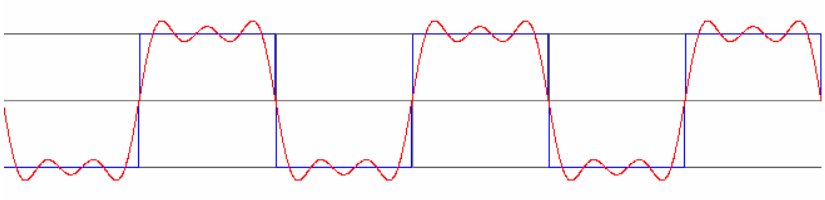


**Figura 5 Uma onda Quadrada aproximada por uma soma de senóides.**

A Figura 6 apresenta as três componentes que formam a aproximação supracitada. A componente de mais baixa freqüência representa a tendência da curva original. As demais componentes acrescentam os detalhes à componente inicial.

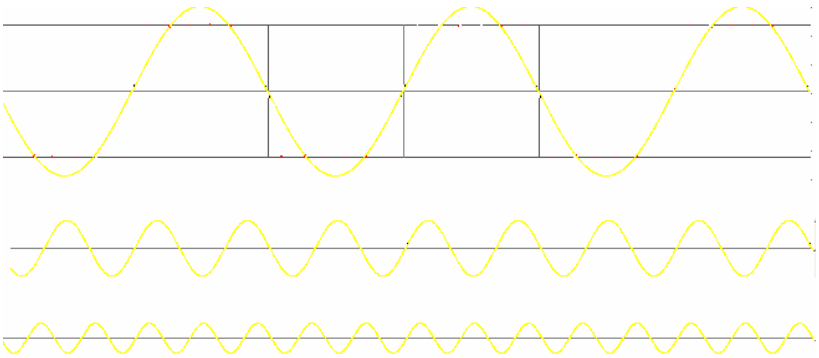


**Figura 6 Componentes de freqüência qua aproximam uma onda quadrada**

A aproximação por série de Fourier é uma soma ponderada de cada uma das componentes. O peso associado a cada componente é dado por $c_n$ como mostrado em (32). Podemos colocar estes pesos ordenados em um plano cartesiano como mostrado na Figura 7.

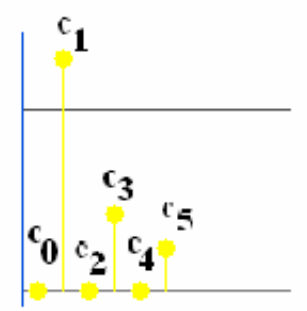



**Figura 7 Intensidade das componentes de Freqüência de uma onda quadrada**

Note que na verdade a aproximação possui seis coeficientes dos quais três são iguais a zero e, portanto não contribuem em nada para a aproximação. Os coeficientes diferentes de zero foram apresentados na Figura 6. O coeficiente $c_1$ é o de maior magnitude e, portanto o que mais contribuem para a aproximação, seguido de $c_3$ e $c_5$.

O Exemplo 5 apresentou o espectrograma discreto de freqüências de uma função. A transformada de Fourier pode ser pensada como uma extensão contínua deste espectrograma. (LATHI, 1974) apresenta uma dedução da transformada de Fourier a partir da série de Fourier. Basicamente conseguimos este resultado supondo que o período da função é infinito e fazendo-se o limite para ω contínuo. Portanto, a transformada de Fourier de uma função pode ser interpretada como um histograma contínuo de freqüências.

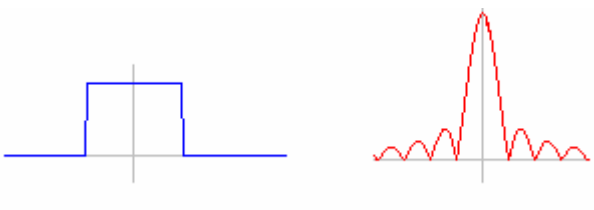


**Figura 8 Uma onda quadrada e sua Transformada de Fourier**

Na Figura 8, a direita podemos ver uma função quadrada no domínio temporal, mais a esquerda está a transformada de Fourier desta função. Ou seja, o espectrograma contínuo de freqüências da função original.

### 3.3.2 Transformada com janelamento (*Short Time Fourier Transform - STFT*)

A transformada de Fourier herda das séries de Fourier a premissa de que a função (sinal) que esta sendo operada é uma função periódica. Infelizmente para a maior parte dos sinais práticos esta premissa não é verdadeira. Isso causa alguns efeitos indesejáveis, o mais comum deles é conhecido como efeito de borda. O efeito de borda é causado pela variação abrupta de freqüência em uma função (sinal) no domínio temporal. Esta variação possui componentes de freqüência alta que interferem em toda a transformada no domínio de freqüência.

#### 3.3.2.1 Efeitos de borda

Seja f(t) uma função (sinal) estacionária como mostrado na equação (33).

$$f(t) = \cos(4*2*\Pi*t) + \cos(2*2*\Pi*t) + 2*\cos(1*2*\Pi*t) + \cos(3*2*\Pi*t) \quad (33)$$

A Figura 9 apresenta a função (sinal) f(t) e sua transformada.

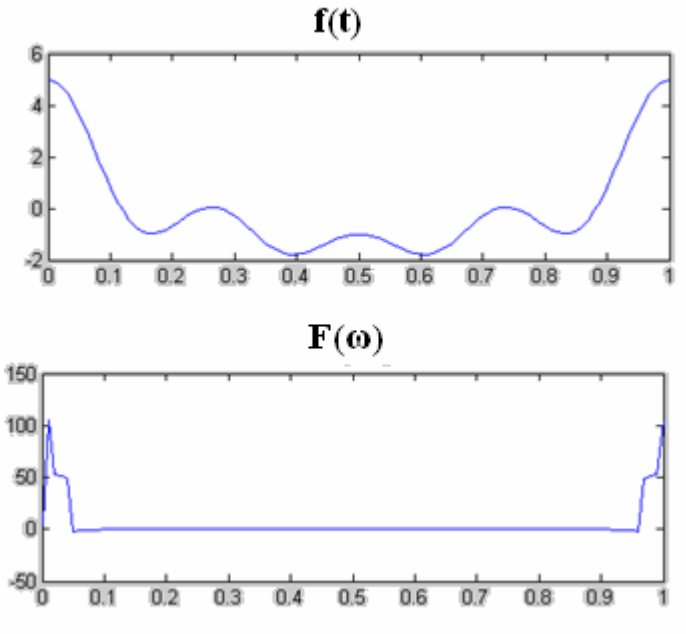



**Figura 9 Acima: sinal estacionário. Abaixo: a respectiva transformada de Fourier.**

Em um sinal estacionário as freqüências estão distribuídas por todo o tempo. Por exemplo, o cosseno de 4 Hz está somado para todo valor de t, como o cosseno de 3 Hz.

Portanto, não faz sentido tentar descobrir em que parte do sinal a freqüência de 4 Hz tem mais influência.

Veja agora como se comporta um sinal não estacionário com as mesmas freqüências presentes na equação (28). A equação (34) o comportamento do sinal, cada freqüência está confinada dentro de um intervalo de tempo determinado.

$$f(t) = \begin{cases} \cos(4 * 2 * \Pi * t), 0 < t < 1 \\ \cos(2 * 2 * \Pi * t), 1 < t < 2 \\ \cos(2 * \Pi * t), 2 < t < 4 \\ \cos(3 * 2 * \Pi * t), 4 < t < 5 \end{cases} \quad (34)$$

A Figura 10 apresenta o gráfico do sinal (34), nele se pode ver claramente que cada freqüência está presente em um intervalo de tempo bem definido. Entre zero e um o sinal tem freqüência de 4 Hz; entre um e dois tem 2 Hz; entre dois e quatro tem 1 Hz; e, por fim, entre quatro e cinto tem 3 Hz.

A transformada de Fourier desse sinal é apresentada logo abaixo na mesma figura. Note que apesar de diferente da primeira transformada mostrada na Figura 9, essencialmente elas são iguais. Pois ambas apresentam a contribuição de cada freqüência para formação do sinal original. No entanto, na Figura 10 os efeitos de borda alteram o comportamento da transformada criando componentes de alta freqüência nos pontos de transição do sinal. Estes componentes de certa forma interferem na transformada, mas ainda assim percebemos picos de freqüência semelhantes em ambas as figuras.

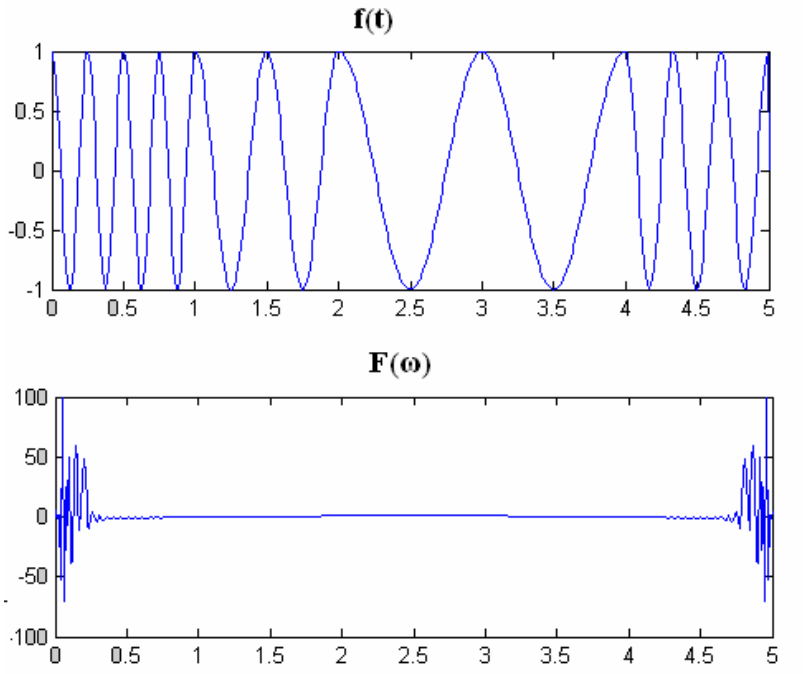



**Figura 10 Acima: sinal não estacionário; Abaixo: a respectiva transformada de Fourier**

A diferença básica entre os dois sinais, no domínio temporal, apresentados na Figura 9 e Figura 10, não são as componentes de freqüência que os compõem, mas sim a distribuição destas componentes. No sinal estacionário, cada cosseno contribui para todo t na formação do sinal. Já no sinal não estacionário cada cosseno contribui apenas em determinados intervalos para a formação do sinal, não se misturando aos demais.

Contudo a informação sobre a distribuição das freqüências dentro do sinal não é destacada na transformada de Fourier. Com base apenas na transformada, não se consegue distinguir em que intervalo cada componente tem mais influência. Sabemos apenas a influência total de cada componente.

Para vencer esta limitação criou-se a transformada de Fourier com Janelamento (*Short Time Fourier Transform* - STFT). A idéia básica deste tipo de transformada é aplicar a transformada de Fourier em um sinal não estacionário apenas para os valores de t onde este sinal é aproximadamente estacionário. Por exemplo, podemos dividir o sinal apresentado em (34) em cinco sinais estacionários. O primeiro no intervalo [0,1[; o segundo em [1,2[; o terceiro em [2,3[; o quarto em [3,4[ ; e, por fim, o quinto entre [4,5]. Para cada um destes intervalos o sinal não estacionário é claramente estacionário. Isso significa que se aplicarmos a transformada de Fourier independentemente em cada um destes intervalos obteremos um resultado sem efeitos de borda. Mais que isso, saberemos qual a freqüência dominante em cada um destes intervalos. O resultado final

desta operação pode ser agrupado em um gráfico tridimensional conhecido como espectrograma de freqüências do sinal. Nele se pode ver a distribuição de freqüências do sinal em cada intervalo de tempo selecionado e determinar qual a freqüência dominante para dado intervalo.

Formalmente, a equação (35) apresenta equação da transformada STFT. Basicamente ela continua sendo a transformada original, mas o núcleo da transformada é multiplicado por w, uma função janela, com suporte compacto[1]. Por ter um suporte compacto a função janela é definida como diferente de zero apenas para um intervalo de tempo curto. Para todos os demais valores de t ela é igual a zero.

$$F(\tau,\omega) = STFT(f(t)) = \int_{-\infty}^{\infty} f(t)w(t-\tau)e^{-i\omega t}dt \tag{35}$$

Onde,

w é uma função janela com suporte compacto.

A escolha do tamanho e formato da janela são pontos chaves da STFT. Trivialmente, podemos escolher uma janela retangular de largura unitária. Tal janela pode funcionar para equações simples como a apresentada em (34), mas falharia na maior parte das funções de caráter pratico. Uma janela de forma quadrada pode até mesmo provocar mais efeitos de borda uma vez que interrope um sinal abruptamente. Normalmente, escolhemos uma janela gaussiana (normal) para atenuar o efeito de borda.

#### 3.3.2.2 A largura da janela

A escolha da largura da janela também é um ponto delicado da STFT, caso escolhamos uma janela muito longa perdemos resolução temporal, ou seja, ficamos incapazes de dizer em que parte da janela cada freqüência aparece. Neste caso, ela pode conter muitas freqüências que poderiam ser separadas por uma janela menor. No limite

[1] Suporte compacto é uma região finita, geralmente pequena, do domínio de uma função na qual essa função é diferente de zero. Por exemplo, funções temporais no domínio do tempo que possuem suporte compacto são limitadas a um intervalo de tempo de finito.

podemos escolher o tamanho da janela do tamanho do próprio sinal, então voltaremos ao caso trivial da transformada de Fourier.

Uma janela curta demais, por outro lado, também causa inconvenientes. Pois as componentes de baixa freqüência são perdidas, uma vez que a janela não abrange um período de uma componente. Portanto, para janelas curtas demais temos baixa resolução de freqüência. Em outras palavras capturamos apenas as componentes de alta freqüência ignorando as freqüências mais baixas.

Gostaríamos de escolher uma janela que capturasse simultaneamente as componentes de altas e baixas freqüências, assim como em que intervalo de tempo elas ocorrem. Mas isso não é possível devido ao princípio de incerteza de Heisenberg (LATHI, 1974).Um vez escolhida a largura da janela, esta deve ser mantida para todo t. Portanto, devemos encontrar um ponde de equilíbrio entre a resolução temporal e a resolução de freqüência. Infelizmente não é nada trivial encontrar tal ponto a priori, sem conhecer o sinal que será analisado.

### 3.3.3 Teorema de Parseval

Agora que entendemos o funcionamento das transformadas, apresentaremos um resultado muito importante para o restante dessa dissertação: o Teorema de Parseval. Esse teorema basicamente estabelece que a energia (informação) de um sinal (função) no domínio temporal é igual a energia (informação) desse mesmo sinal, quando transladado para o domínio de freqüência. Matematicamente temos:

$$\|x(t)\| = \langle x(t), x(t) \rangle = \langle F(x(t)), F(x(t)) \rangle = \|F(x(t))\| \qquad (36)$$

Onde,

$x(t)$ é um sinal com variável independente t,

$F(x(t))$ é a Transformada de Fourier do sinal x(t)

$\|\bullet\|$ é o operador de módulo de um sinal

$\langle \bullet, \bullet \rangle$ é o operador de produto escalar

Em física e engenharia, o teorema de Parseval é interpretado como conseqüência direta da conservação de energia, que garante que os dois sinais, antes e depois da transformada sejam equivalentes. Para ciência da computação, a informação contida na função antes e depois da transformação permanece a mesma. Portanto, podemos

considerar a transformada de Fourier apenas uma nova forma de representar um mesmo sinal.

O teorema de Parseval pode ser generalizado e aplicado a qualquer transformação de eixos ortogonais em um espaço de Hilbert, H. Essa demonstração pode ser encontrada em (MARTIN VETTERLI e JELENA KOVACEVIC, 1995). Esse é um resultado poderoso pois ele garante que o teorema de Parseval pode ser aplicado a transformada wavelet.

## 3.4 A transformada wavelet

Do ponto de vista estritamente matemático, a transformada Wavelet é uma transformada integral cujo núcleo é uma função wavelet, ψ. A função wavelet, também conhecida como wavelet mãe (mother wavelet), deve obedecer alguns critérios de admissibilidade para que ela seja considerada um núcleo válido da transformada. Estes critérios serão detalhados mais adiante (seção 3.4.1).

As wavelets funcionam como base na decomposição de outras funções pertencentes a $L^2(R)$, espaço das funções finitamente mensuráveis (seção 3.1.7). Elas podem ser redundantes ou ortogonais. Nosso maior interesse está nas wavelets ortogonais, pois estas possibilitam a reconstrução da função (sinal) original. A equação (37) apresenta a fórmula matemática da transformada wavelet contínua

$$\langle \psi, f(t) \rangle = \Psi_g[f(t)](a,b) = |a|^{-1/2} \int f(t)\psi^*(\frac{t-b}{a})dt = \Psi_g f(a,b) = F(a,b) \qquad (37)$$

Onde,

ψ é a função wavelet mãe,

Ψ é o operador linear que representa a transformada,

*a* é o coeficiente de escala,

b é o coeficiente de deslocamento.

f(t) é a função (sinal) no domínio temporal,

* é o operador de complexo conjugado.

Novamente, como no caso de Fourier, a álgebra linear nos fornece uma estrutura matemática adequada para o estudo da transformada. A equação (37) é o produto escalar

entre a função original no domínio temporal com a função wavelet mãe em diferentes escalas e deslocamentos (coeficientes **a** e **b**). O resultado da operação é função de **a** e **b**. O coeficiente de escala **a** age sobre a freqüência da wavelet mãe, enquanto o coeficiente de deslocamento b age sobre seu posicionamento.

### 3.4.1 Critério de admissibilidade

O termo wavelet vem do inglês e significa "onda pequena" ou onduleta (como originalmente do francês), porque é exatamente esta característica que o núcleo da transformada deve possuir para ser uma função admissível. Em outras palavras, o núcleo da transformada wavelet deve ser uma "onda pequena". Quando nos referimos à onda significa que a função wavelet mãe deve oscilar. E quando dizemos que ela deve ser pequena, significa que ela não deve se perpetuar por todo o tempo. Matematicamente este é critério de admissibilidade é alcançado caso a função $\psi$, wavelet mãe, obedeça a relação (38).

$$\int \frac{\left|F_{\psi}(\omega)\right|^2}{\left|\omega\right|} d\omega < \infty \tag{38}$$

Onde,

$F_{\psi}(\omega)$ é a transformada de Fourier de $\psi$, a função wavelet mãe.

$\omega$ é a freqüência instantânea

Como mostrado em (TAO LI, QI LI et al., 2005), (38) é suficiente (mas não necessário)[2] para garantir que a função $\psi$ seja uma "onda pequena". Foge do escopo dessa dissertação aprofundar-se nesse tema, mas uma análise rápida pode nos ajudar compreender melhor o critério de admissibilidade. Primeiro note que (38) implica em (39).

$$\left|F_{\psi}(\omega)\right|^2\Big|_{\omega=0} = 0 \Rightarrow \int \psi(t)dt = 0 \tag{39}$$

Onde,

$F_{\psi}(\omega)$ é a transformada de Fourier de $\psi(t)$

[2] Existem wavelet mães que não precisam necessariamente seguir (38), elas estão fora do escopo dessa dissertação, mas podem ser encontradas em (MARTIN VETTERLI e JELENA KOVACEVIC, 1995)

$\psi(t)$ é a núcleo da transformada wavelet, a wavelet mãe.

A relação (39) é uma implicação. A conclusão dessa implicação (segunda parte da relação) nos indica que a wavelet mãe oscila (área sob a curva é igual a zero), e é, portanto uma onda. A primeira parte da implicação (premissa) é conseqüência direta de (38) e informa que a função wavelet mãe decai rapidamente com o tempo, portanto é pequena.

Como podemos perceber, o critério de admissibilidade não é muito rigoroso. Portanto, existe mais de uma função que o atende. Desta forma podemos falar em famílias de funções wavelets que atendem o critério de admissibilidade. Na próxima seção veremos algumas das principais famílias de funções wavelet.

### 3.4.2 Algumas funções wavelet

Existem muitas funções wavelet que podem ser usadas como núcleo da transformada. As Figura 11, Figura 12, Figura 13 e Figura 14 apresentam as principais funções wavelets empregadas. Note que todas possuem as mesmas características: têm suporte compacto e possuem média igual a zero.

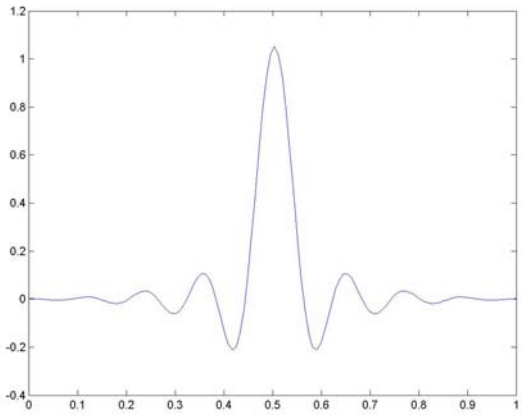


**Figura 11 Chapéu Mexicano**

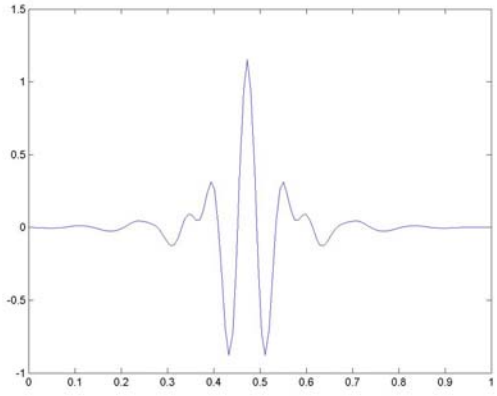


**Figura 12 Meyer**

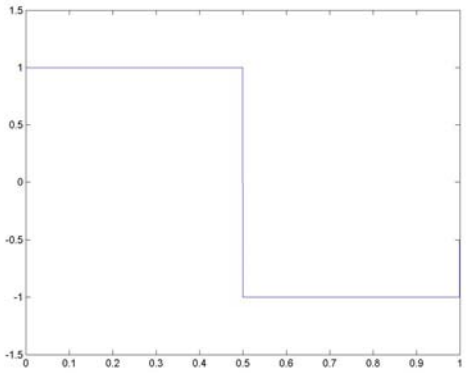


**Figura 13 Haar Wavelet**

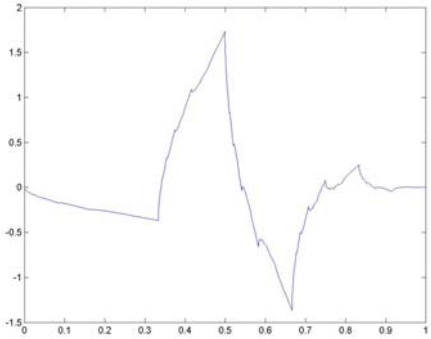


**Figura 14 Daubechies Wavelet**

A Figura 11 apresenta a wavelet mãe conhecida como chapéu mexicano por causa de seu formato que lembra um *sombrero*. Ao contrário do desejado nessa dissertação, ela não é ortogonal e as propriedades decorrentes do Teorema de Parseval não se aplicam nesse caso.A equação (40) mostra função explicita da wavelet chapéu mexicano.

$$\psi(t) = \left( \frac{2}{\sqrt{3}} \pi^{-1/4} \right) \left(1 - t^2\right) e^{-t^2/2} \qquad (40)$$

Na Figura 12 apresentamos a wavelet de Meyer, essa wavelet tem a seguinte peculiaridade: ela é biortogonal. A princípio essa wavelet serviria aos fins dessa dissertação, no entanto duas características impedem que isso aconteça: Primeiro, ela

não possui um suporte compacto sobre o eixo real, ou seja, ela existe para todo o tempo, apesar de tender a zero. Segundo, ela não tem uma representação discreta simples, o que dificulta seu desenvolvimento via software. Ao contrário do que ocorre normalmente, a descrição da wavelet mãe de meyer é descrita normalmente no domínio da freqüência ω, como podemos ver na equação (41).

$$\hat{\psi}(\omega)=\begin{cases} 2\pi^{-1/2}e^{i\omega/2}sen\left(\frac{\pi}{2}v\left(\frac{3}{2\pi}|\omega|-1\right)\right) & se \quad \frac{2\pi}{3}\le|\omega|\le\frac{4\pi}{3} \\ 2\pi^{-1/2}e^{i\omega/2}sen\left(\frac{\pi}{2}v\left(\frac{3}{4\pi}|\omega|-1\right)\right) & se \quad \frac{4\pi}{3}\le|\omega|\le\frac{8\pi}{3} \\ 0 & se \quad |\omega|\notin\left[\frac{2\pi}{3},\frac{8\pi}{3}\right] \end{cases} \tag{41}$$

Onde,

$$v(a)=a^4(35-84a+70a^2-20a^3) \quad a\in[0,1]$$

Destacamos a Figura 13, conhecida como Haar Wavelet (MALLAT S.G., 1989). Ela é a mais simples de todas as wavelets que atendem a condição de admissibilidade. Pode-se mostrar que a wavelet de Haar é apenas um caso particular das wavelets de Daubechies (DAUBECHIES, 1992; DAUBECHIES, 1990) com apenas um momento nulo. A equação (42) mostra a função explicita da wavelet mãe de Haar (psi). A wavelet de Haar tem um suporte compacto bem definido, além de um simples desenvolvimento via software, o que a torna ideal para os fins dessa dissertação.

$$\psi(t)=\begin{cases} 1 & 0\le t<0,5 \\ -1 & 0,5\le t\le 1 \\ 0 & \text{caso contrário} \end{cases} \tag{42}$$

A família de wavelets conhecida como Daubechies foi criada pela matemática de mesmo nome na década de 80 (DAUBECHIES, 1990). Baseada no nos trabalhos de Mallat (MALLAT S.G., 1989), ela mostrou a relação entre as teorias wavelets desenvolvidas até então com o processamento de sinais, abrindo portas para um grande

número de estudos sobre o emprego e utilização das wavelets nos mais diversos campos (TAO LI, QI LI et al., 2005). A Figura 14, mostra a wavelet mãe de Haar (psi) de segunda ordem. Infelizmente essa função não possui uma função explicita (A wavelet mãe de Daubechies de segunda ordem é definida em função da função de transferência do filtro que ela representa).

### 3.4.3 A escala e o deslocamento

A função wavelet mãe, quando usada como núcleo da transformada wavelet, deve sofrer duas operações básicas que lhe permitem fazer a análise multi-resolucional de uma função (sinal). Estas operações são conhecidas como escalamento e deslocamento e são regidas pelos coeficientes a e b da equação (37). Seja (43) a equação de uma função wavelet mãe qualquer. Por meio dos parâmetros a e b podemos controlar, respectivamente, a freqüência e o posicionamento (suporte) da função.

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \tag{43}$$

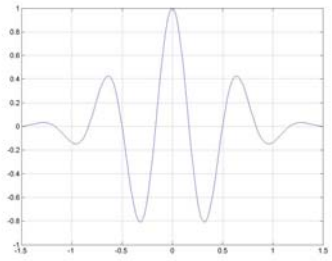
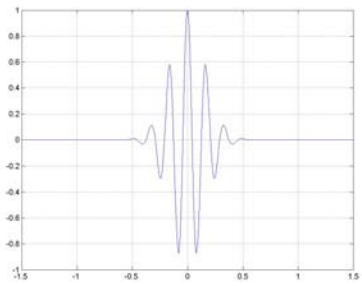
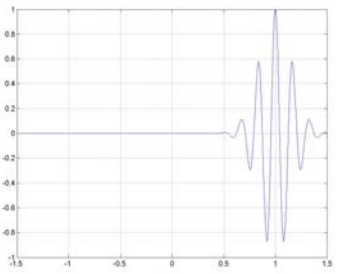
Onde,

$\psi$ é a wavelet mãe,

a o coeficiente de esclada e,

b o coeficiente de deslocamento.

O coeficiente $\frac{1}{\sqrt{a}}$ visa manter a área sob a função (energia do sinal) constante após o escalamento. A Figura 15 apresenta uma wavelet mãe de Morlet, cuja forma é $\psi(t) = e^{-\alpha t^2} e^{i\omega t}$. A Figura 16 apresenta a mesma wavelet, mas com os parâmetros a=0,5, o que proporciona o dobro da freqüência original. Por fim, a Figura 17 apresenta a mesma wavelet anterior, mas deslocada sobre seu suporte devido as alterações no parâmetro b.

| | | |
|---|---|---|
|  **Figura 15 wavelet mãe** |  **Figura 16 wavelet mãe com a freqüência dobrada** |  **Figura 17 wavelet mãe com a freqüência dobrada e deslocada a direita** |

Note que o parâmetro de escala é inversamente proporcional ao número de componentes de alta freqüência da wavelet mãe, portanto para valores muito grandes de a obtemos freqüências mais baixas, enquanto valores muito pequenos obtêm freqüências mais altas.

Agora podemos entender melhor como a análise de funções empregando wavelet funciona: para cada valor de a, resolvemos a integral da equação (37) para todos os valores de b. Sabemos que a integral entre as duas funções é apenas o produto escalar entre elas, ou em outras palavras, o produto escalar fornece o "ângulo" entre as duas funções em um espaço de dimensão infinita. Este ângulo serve, em certo sentido, como uma medida de semelhança entre as duas funções: a função wavelet (deslocada e escalada) e a função original. Ao final da operação obtemos a representação da função original em termos de suas "componentes wavelet". O Exemplo 6 mostra como esta operação acontece.

Exemplo 6.

A Figura 18 apresenta a função wavelet mãe e uma função qualquer no mesmo gráfico. O produto escalar entre estas duas funções fornece a medida de similaridade entre ambas. Como a função wavelet tem suporte compacto, para maior parte dos pontos, o produto escalar será igual a zero. A Figura 19 mostra esta mesma wavelet mãe deslocada para direita. Novamente o produto escalar pode ser feito para comparar a segunda parte da função analisada com a wavelet mãe. Este procedimento é feito para todos os deslocamentos possíveis (todos os

valores de *b*).

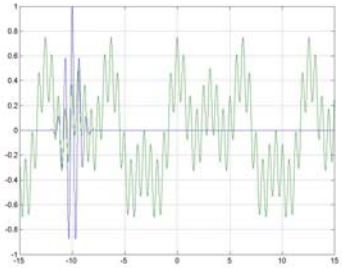


**Figura 18 Wavelet comparada com uma função qualquer**

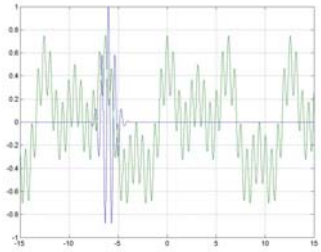


**Figura 19 Wavelet deslocada e comparada**

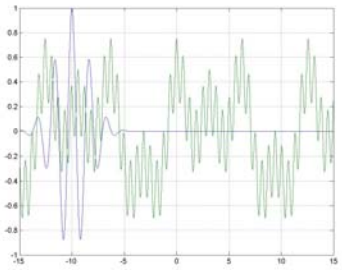


**Figura 20 Wavelet escalada e comparada com uma função**

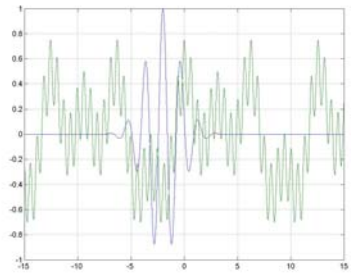


**Figura 21 Wavelet escalada e deslocada comparada com uma função.**

Em seguida devemos alterar a escala da função wavelet mãe por meio do parâmetro **a**. Na Figura 20 aumentamos o valor de **a**, diminuindo a freqüência da wavelet mãe. O produto escalar novamente fornece a similaridade entre as duas funções. Finalmente a Figura 21 apresenta a função wavelet escalada e deslocada. Este procedimento deve ser repedido para todas as escalas e deslocamentos possíveis. Ao final do procedimento teremos o resultado da transformada wavelet, que é a representação da função (combinação linear) original em termos da wavelet mãe em diferentes escalas e deslocamentos.

### 3.4.4 Transformada de Fourier com janelamento e transformada wavelet

Tanto a transformada de Fourier com janelamento (STFT – *Short Time Fourier Transform*), quanto a transformada wavelet são mecanismos de análise tempo-freqüência. Na STFT os "átomos" da análise são senoides limitadas por uma janela,

geralmente uma gaussiana. Na transformada wavelet a função wavelet mãe age como um "átomo". A característica comum entre as duas transformadas está no fato de ambas empregarem "átomos" como suporte compacto tanto no domínio temporal quanto no domínio de freqüência.

Na STFT o "tamanho" da janela gaussiana empregada é constante durante toda a operação da transformada. Já na transformada wavelet o tamanho do "átomo" (wavelet mãe) varia conforme o parâmetro de escala empregado. Esta é a base para a análise de multiresolução. A Figura 22 e a Figura 23 ilustram o comportamento das janelas para a transformada STFT e a transformada wavelet.

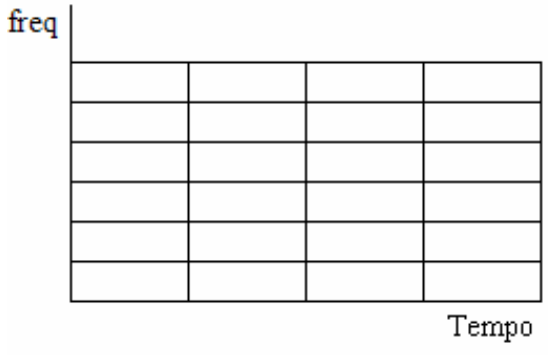



**Figura 22 Grade da STFT**

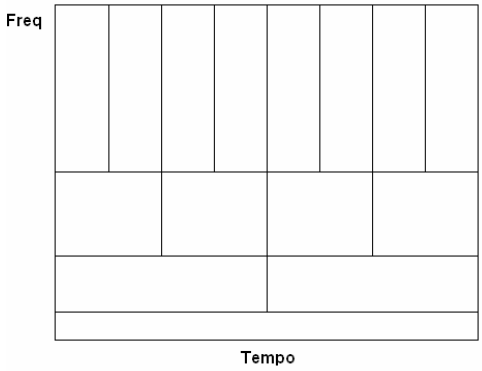



**Figura 23 Grade da Transformada Wavelet**

Nas duas figuras, a base do retângulo representa a largura da janela empregada, e a altura representa a largura desta mesma janela no domínio de freqüência. A transformada de Fourier com janelamento usa uma janela constante para análise de todo o sinal independentemente da freqüência analisada. Este fato se reflete nos retângulos de base e altura iguais da Figura 22. Já a transformada wavelet, por meio da operação de escalamento, altera e utiliza funções wavelet mães em diferentes escalas para analisar o sinal de forma que as freqüências mais altas são analisadas por escalas grandes (pequenas janelas temporais), e as freqüências mais baixas são analisadas por escalas pequenas (grandes janelas temporais). Este fato se reflete nos retângulos de áreas iguais, mas bases diferentes apresentados na Figura 23.

### 3.4.4.1 A multiresolução

A multiresolução (GRAPS A., 1995)é a propriedade da transformada wavelet que permite que a função (sinal) analisada seja decomposta em diversos graus de

detalhe. Como visto na seção anterior, o "átomo" da análise wavelet é a função wavelet mãe que através do escalamento toma diferentes freqüências. O escalamento da função wavelet mãe equivale à mudança de largura da janela de uma transformada STFT. Nesta seção veremos com mais detalhes como ocorre a multiresolução.

A propriedade de multiresolução está associada à função de escala φ (phi). A função de escala é definida de tal forma que a partir dela podemos definir toda uma família de funções, como mostrado em (44)

$$f(t) = \sum_{b} c_k \varphi(t-b) \tag{44}$$

Onde,

f(t) é uma função gerada a partir de φ

φ é a função de escala

b é o índice de deslocamento

Note que, na equação (44), f(t) é uma representante da família de funções que podem ser representadas a partir da combinação linear de φ(t-b). Para cada conjunto distinto de coeficientes $c_k$, é definida uma função f(t) diferente. Chamemos ao conjunto de todas as funções que podem ser representadas por φ(t-b) de $V_0$.

Seja (45) a função de escala com deslocamento b

$$\varphi_b(t) = \varphi(t-b) \tag{45}$$

Então,

$$\varphi_{a,b}(t) = 2^{a/2}\varphi(2^a t - b) \tag{46}$$

Onde,

a é o coeficiente de escala e

b é o coeficiente de deslocamento.

A equação (46) é a função de escala em diferentes resoluções. Note que uma função representada por φ(t) pode ser representada, com maior resolução (nível de

detalhes) por $\varphi_{1,0}(t)$, mas o oposto não é verdade. Se chamarmos o conjunto de funções que pode ser representada por $\varphi_{1,0}(t)$ de $V_1$, então a relação (47), é verdade:

$$V_0 \subset V_1 \subset \cdots \subset V_n \tag{47}$$

A relação (47) deve ser interpretada da seguinte maneira: uma função representada por $\varphi_{0,b}(t)$, também pode ser representada por $\varphi_{1,b}(t)$, que por sua vez pode ser representada por $\varphi_{2,b}(t)$, e assim sucessivamente, com resoluçõesão cada vez mais altas (nível de detalhe mais alto). Esta é a base para multiresolução, mas como a função φ(t) está associada às wavelets?

Seja $\varphi_f^{(a)}(t)$ a representação da função f(t) por meio de $\varphi_{a,b}(t)$, como representado em (48).

$$f(t) = \varphi_f^{(a)}(t) = \sum_a \sum_b c_{a,b}\varphi_{a,b}(t) \tag{48}$$

É possível se provar a seguinte relação (W.M.SHAW, J.B.WOOD et al., 1993b),

$$\varphi_f^{(a)}(t) - \varphi_f^{(a-1)}(t) = c_{a-1,b}\psi_{a-1,b}(t) \tag{49}$$

Onde,

a é o coeficiente de escala,

b é o coeficiente de deslocamento,

c é o coeficiente da função de escala,

$\varphi_f^{(a)}(t)$ é a representação de f(t) como definido em (48),

ψ é a função wavelet mãe como definida em (43).

A equação (49) mostra como a função wavelet está associada a função de escala φ. De certa forma ela nasce da diferença entre duas representações em níveis de resolução diferentes da função f(t). Se representarmos como $W_a$ o espaço gerado pela função wavelet mãe com resolução **a**, então a relação (50) nos fornece uma maneira de

representar qualquer função pertencente a $V_a$ em função de funções com menor resolução.

$$V_a = V_{a-1} \oplus W_{a-1} \tag{50}$$

A relação (50) pode ser expandida para

$$V_a = V_k \oplus W_k \oplus W_{k+1} \oplus \cdots \oplus W_{a-1} \tag{51}$$

Onde,

k e a pertencem a N e

k < a

Em outras palavras, uma função contida em $V_a$, ou seja, que pode ser perfeitamente representada pela função de escala com resolução **a**, é decomposta na soma de sua representação na resolução **k** (mais baixa) com seguidas somas de funções wavelet com resoluções entre **k** e **a**.

### 3.4.5 A implementação da transformada wavelet

O resultado da relação (51), obtido na seção 3.4.4.1, é muito importante porque ele nos fornece uma maneira de construir a transformada wavelet através de filtros digitais empregados no campo de processamento de sinais. Filtros, no jargão de processamento de sinais, são operações combinações lineares da amostra do sinal.

Existem basicamente dois tipos de filtros, conhecidos como filtros passa-baixa e filtros passa-alta. Os filtros passa-baixa são funções que recebem de entrada um sinal e retornam como saída o mesmo sinal filtrado de todas as suas freqüências acima de um determinado limite. Em outras palavras, este tipo de filtro permite somente que determinadas freqüências abaixo do limite definido passem por ele. Analogamente, filtros passa-alta permitem apenas que freqüências acima de determinado limite passem por ele.

Existem ainda, os filtros passa-faixa que podem ser obtidos pela combinação de um filtro passa-alta com um filtro passa-baixa.

No domínio do tempo, um filtro passa-baixa suaviza o sinal de entrada, mantendo a tendência geral do sinal e removendo as imperfeições. Um filtro passa-alta, permite que apenas os detalhes de um sinal passem, removendo sua tendência global.

Através da utilização de banco de filtros é possível se implementar a transformada apresentada na equação (51). Um banco de filtros pode ser um arranjo em forma de árvore de filtros passa-alta e passa-baixa ligados como mostrado na Figura 24

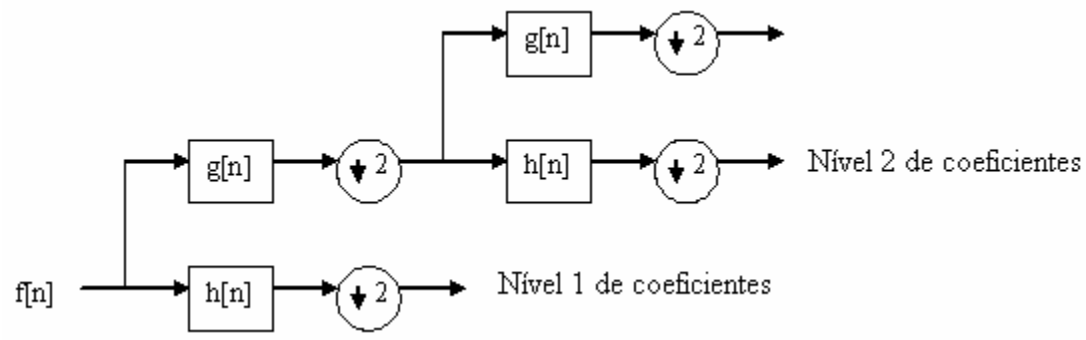



**Figura 24 Banco de filtros**

Na Figura 24 h[n] representa um filtro passa-alta e g[n] representa um filtro passa-baixa. O número 2 dentro do circulo representa uma decimação (*downsampling*) a cada duas amostras do sinal original.

Esquematicamente o banco de filtros pode ser visto em diversos níveis. Cada nível é composto por um filtro passa-alta e um filtro passa-baixa. A saída do filtro passa-baixa realimenta o próximo nível do banco. A saída do filtro passa-alta é o coeficiente da função wavelet mãe em cada nível de resolução. A saída do filtro passa-baixa é o coeficiente da função de escala em cada nível.

### 3.4.5.1 A transformada de Haar

A transformada wavelet de Haar, ou simplesmente transformada de Haar, é a mais simples de todas as transformadas wavelet. Ela foi descoberta pelo matemático de mesmo nome em 1909. Seu trabalho só foi reconhecido mais tarde com as extensões feitas por Daubechies (DAUBECHIES, 1990), que construiu uma família de funções wavelet com diferentes momentos nulos, das quais Haar é um tipo.

A função wavelet de Haar é uma onda quadrada como mostrada na Figura 13. Sua equação é apresentada em (52).

$$\psi(t) = \begin{cases} 1 & 0 \le t < \frac{1}{2} \\ -1 & \frac{1}{2} \le t < 1 \end{cases} \quad (52)$$

A função wavelet de Haar é obtida a partir da função de escala de Haar, que é simplesmente uma reta no intervalo [0,1] como mostra a equação (53)

$$\phi(t) = \begin{cases} 1 & 0 \le t < 1 \\ 0 & c.c. \end{cases} \quad (53)$$

Note que por ser uma reta, a função de escala, que age no domínio temporal representa (aproxima) qualquer função que tenha intervalos constantes em sua composição. Devido a sua simplicidade, a Wavelet de Haar é muito fácil de ser implementada por software e este foi um dos motivos pelo qual escolhemos usá-la nesta dissertação. Além disso, as wavelets de Haar formam uma base ortogonal para o espaço $L^2$, o espaço de funções que possuem métrica finita. O teorema de Parseval (36) garante que, se as bases são ortogonais, então a representação da função em qualquer um dos espaços é equivalente. Em outras palavras, não há perda de informação durante o processo da transformada.Por isso, ao aplicarmos uma transformada como essa a um documento em um sistema de BRI, não temos perda durante o processo de indexação. Portanto, a transformada wavelet de Haar é uma escolha adequada para os fins desta dissertação. O Exemplo 7 mostra como funciona a transformada wavelet de Haar.

Exemplo 7.

A transformada wavelet de Haar não normalizada pode ser construída como mostrada na figura Figura 25.

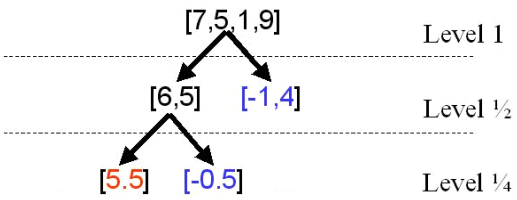



**Figura 25 Árvore de decomposição Wavelet**

A função original é [7,5,1,9]. Ao passarmos está função pelo banco de filtros como mostrado na Figura 24, obtemos duas saídas. A saída

do filtro passa-alta e a saída do filtro passa-baixa. A primeira representa os detalhes do sinal e está no ramo da direita da árvore. Este ramo é obtido a partir da diferença, duas a duas, das dimensões do vetor que representa a função original. Ou seja, $[-1,4]=[\frac{5-7}{2},\frac{9-1}{2}]$. Esta saída representa os coeficientes da função wavelet pertencente ao nível de resolução mais alta, ou seja, $W_1$.

A segunda saída (relativa ao filtro passa-baixa) é colocada a esquerda da árvore. Ela é obtida a partir da média, duas a duas, das dimensões do vetor original, ou seja, $[6,5]=[\frac{7+5}{2},\frac{9+1}{2}]$. Esta saída é um elemento de $V_1$ e realimenta o processo para gerar o próximo nível de resolução. O processo se repete até alcançarmos a representação em menor resolução possível. No caso de nosso exemplo, isso ocorre no próximo nível, portanto vejamos como fica: o vetor que representa a função original no nível de resolução mais baixo é [6,5]. Ele deve sofrer o mesmo procedimento do vetor original ([7,5,1,9]), ou seja, deve passar pelos filtros passa-alta e passa baixa. Os respectivos resultados são $[-0.5]$ e $[5.5]$. Neste ponto o processo pára e o resultado da transformada é a concatenação dos coeficientes da função de escala seguido dos coeficientes da função wavelet, ou seja, $[0.5,5.5,-1,4]$.

A transformada inversa pode ser obtida facilmente, como esquematizada na árvore da Figura 26.

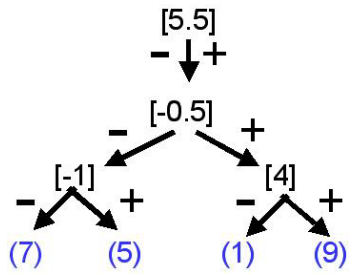



**Figura 26 Árvore de reconstrução Wavelet de Haar**

Iniciamos pelo coeficiente da função de escala, somamos e

subtraímos o primeiro coeficiente da função wavelet. O resultado da subtração é posto a esquerda da arvore e a soma a direita. Repetimos o processo com o próximo nível, somando e subtraindo os próximos coeficientes da função wavelet e assim sucessivamente. Ao final do processo, quando não houver mais coeficientes para serem contabilizados, obtemos a função (sinal) original.

# 4 Trabalhos relacionados

Nas seções 3.3 e 3.4 foram apresentados os fundamentos de processamento de sinais necessários para a compreensão da proposta desta dissertação. A utilização de transformadas não é uma novidade no campo de BRI. Neste capítulo, apresentamos alguns exemplos de como as técnicas apresentadas anteriormente são empregadas, assim como trabalhos relacionados.

## *4.1 Sistemas de BRI multimídia*

Os sistemas de BRI Multimídia (BRIM) há muito já empregam técnicas de processamento de sinais para a análise de objetos complexos como, por exemplo, músicas, vídeos e imagens. O MULTOS, Office Server, (YATES B.R. e BERTHIER R.N., 1999) e QBIC (NIBLACK et al.) são exemplos de sistemas BRI Multimídia.

Os sistemas de BRI multimídia geralmente possuem algum tipo de estrutura, ainda que simples, para representar as diferentes partes de um documento. Por isso em geral estes sistemas estão associados a sistemas de banco de dados o que de certa forma os diferenciam dos sistemas de BRI tradicionais.

Músicas, vídeos e imagens são naturalmente codificados como sinais e tiram, , portanto proveito das técnicas já mencionadas. O GEMINI, *GEneric Multimedia object INdexIng*, (FALOUTSOS) é uma abordagem abstrata para indexação e busca de objetos complexos. Esta abordagem é definida da seguinte forma:

- Seja uma coleção de N objetos: $O_1, O_2, \ldots, O_N$
- A distância ou similaridade entre dois objetos é dada por $D(O_i, O_j)$
- A consulta Q deve ser especificada com um fator de tolerância ε
- F é um conjunto de características (*features)*, que são valores numéricos que de fornecem alguma informação parcial importante que caracterize o objeto.

A execução de uma consulta, segundo a abordagem GEMINI, ocorre da seguinte forma: primeiro um teste rápido, mas impreciso, é feito na coleção de N objetos com base em suas características *(features)*. A idéia é a seguinte: objetos com características *(features)* semelhantes às da consulta são candidatos à resposta. Este processo de

filtragem inicial reduz o número de objetos que devem ser analisados. Na segunda etapa, os objetos restantes são comparados com a consulta usando-se a função de similaridade D(). Todos os objetos que estiverem dentro da tolerância ε são recuperados, enquanto que os demais são ignorados.

Exemplo 8.

Seja um sistema de BRI multimídia para busca de dados em séries temporais. Por exemplo, a variação da taxa de inflação mensal. Para usa série temporal como esta, uma possível função de similaridade é a distância euclidiana como em (54).

$$D(I,Q) = \sqrt{\sum (I-Q)^2} \qquad (54)$$

Onde,
D é a função de distância/similaridade entre dois objetos.
I é a série temporal do índice de inflação
Q é a serie temporal de consulta

Esta operação é bastante custosa do ponto de vista computacional, pois envolve subtrações, multiplicações e radiciações. Segundo a abordagem GEMINI primeiro devemos analisar as características *(features)* da série temporal para realizar uma filtragem inicial, limitando o número de comparações que serão feitas.
Uma característica (*feature)* possível para uma série temporal é a média. Duas séries com médias iguais são provavelmente parecidas, mas se as médias forem muito diferentes as séries sem dúvidas são diferentes. Se quisermos refinar a filtragem, podemos, além da média, usar o desvio padrão ou momentos mais altos. Outros candidatos fortes a característica (*feature)* de uma série temporal, seriam os primeiros coeficientes da transformada de Fourier da série.

A extração de *características (features)* é uma etapa importante em sistemas de busca e recuperação de informação multimídia, pois através delas podemos realizar uma

filtragem inicial no nosso conjunto de dados e limitar bastante o número de operações feitas para recuperar os documentos. Como visto no Exemplo 8, técnicas de processamento de sinais são aplicadas muitas vezes para obter as *características (features)* de documentos multimídia como músicas, vídeos e imagens.

Mais recentemente, alguns pesquisadores têm apontado para as vantagens de usar tais técnicas para extrair informações relevantes de documentos do tipo texto. As iniciativas ainda são isoladas, mas apontam para uma tendência que vem crescendo nos últimos anos. Podemos citar alguns exemplos de técnicas de processamento de sinais empregados em texto como TOPIC ISLAND (MILLER et al), que utiliza transformadas wavelet para resumo e visualização de texto. Ou os trabalhos de (FLESCA, MANCO et al., 2005), (TAO LI, QI LI et al., 2005), (PARK, RAMAMOHANARAO et al., 2005).

## 4.2 Topic Island

Apesar de não ser um software de BRI, Topic Island é uma das primeiras iniciativas a utilizar a transformada wavelet em documentos texto. Seu principal objetivo é encontrar uma representação em forma de vales e picos para os tópicos (temas) de um documento texto em diferentes níveis de resolução. O software funciona da seguinte maneira: primeiro o documento texto é codificado na forma de um sinal respeitando a narrativa original, ou seja, a ordem dos termos dentro do documento é mantida; depois, a transformada wavelet é empregada para analisar, em diversos níveis de detalhe, a narrativa do texto e determinar qual o tema abordado. Os temas mais relevantes são apresentados ao usuário como picos separados por vales que representam a mudança de narrativa do texto. Desta forma, livros inteiros podem ser resumidos e apresentados de maneira gráfica possibilitando que o usuário tenha um resumo dos temas abordados.

## 4.3 BRI de documentos semi-estruturados

Flesca, et al (FLESCA, MANCO et al., 2005) propõem uma abordagem nova para BRI em documentos semi-estruturados, particularmente documentos XML. No entanto, sua abordagem não é baseada em conteúdo, mas sim na estrutura dos documentos. Esta abordagem emprega a transformada de Fourier para descobrir similaridades entre documentos XML muito longos. A idéia é a seguinte: um documento XML é caracterizado por rótulos (*tags*) iniciais e finais que definem sua

estrutura. Esses rótulos são definidas em um documento especial conhecido como *XMLSchema*. Uma instância bem formada de um documento XML é qualquer documento que respeite a hierarquia de rótulos estabelecidas no *XMLSchema.* Dada uma instância, codificamos os rótulos presentes de maneira a associar um código numérico único para cada uma. A maneira mais trivial de associar este código aos rótulos é empregar um inteiro que indica o nível de indentação de cada rótulo em relação a raiz do documento. Se considerarmos o documento XML como uma árvore, estaríamos associando um inteiro para cada nível da árvore a partir da raiz.

Os documentos são representados como uma série temporal onde cada rótulo é disposta no eixo X conforme aparece no documento. Dois documentos com o mesmo *XMLSchema* são considerados iguais quando possuem a mesma representação de série temporal.

A transforma de Fourier é usada para transpor á serie temporal que representa o documento XML para o domínio de freqüência. O autor defende que a análise de freqüência é mais propícia para detectar similaridades entre documentos XML do que os tradicionais algoritmos empregados para a comparação de documentos semi-estruturados de ordem quadrática.

## 4.4 Métodos empregando wavelets

Talvez o trabalho mais parecido com a proposta desta dissertação seja o de (PARK, RAMAMOHANARAO et al., 2005), onde a transformada wavelet é empregada para a realização de BRI em documentos texto. Neste trabalho o autor propõe a utilização das transformadas wavelets para a indexação de documentos texto da seguinte maneira: primeiro todos os documentos de uma coleção são partidos em N partições. Dessa forma todos os documentos passam a ter a mesma dimensão (partições). Cada partição do documento é analisada individualmente como se fosse um documento independente. A transformada wavelet de cada documento é a concatenação das transformadas de cada pedaço do documento.

Park estabelece o conceito de termo sinal, que nada mais é que um sinal que representa a distribuição do termo dentro de partição em questão. Assim, ao aplicar a transformada wavelet em cada uma das partições ele obtém a representação no domínio wavelet das partições do documento. (PARK, RAMAMOHANARAO et al., 2005) exploram a magnitude do sinal que representa a partição do documento para retirar

informação sobre a freqüência e a posição de um termo, respectivamente, dento de uma partição.

Assim, a proposta de (PARK, RAMAMOHANARAO et al., 2005) na verdade é a utilização da transformada wavelet para analisar a freqüência e a posição dos termos de uma consulta em diferentes resoluções.

Vejamos com um pouco mais de detalhe a proposta de (PARK, RAMAMOHANARAO et al., 2005). Considere a seguinte consulta feira ao sistema de BRI "amor de mãe". O sistema ignorará a preposição "de" por se tratar de um termo com baixo IDF para um texto escrito e português. Portanto a consulta seria "amor mãe". A Figura 27 mostra como dois documentos são representados na proposta de Park. Todos documentos têm três partições; cada termo é representado por uma linha horizontal e a pequena linha vertical indica sua posição dentro do documento. Por exemplo, para a primeira partição do Documento 1, o termo amor aparece 3 vezes enquanto o termo mãe aparece duas vezes. Pela posição dos termos podemos perceber que o termo "amor mãe" aparece duas vezes nessa primeira partição. Já na segunda partição do Documento 2, tanto o termo amor e mãe aparecem ambos duas vezes, mas não formam "amor mãe".

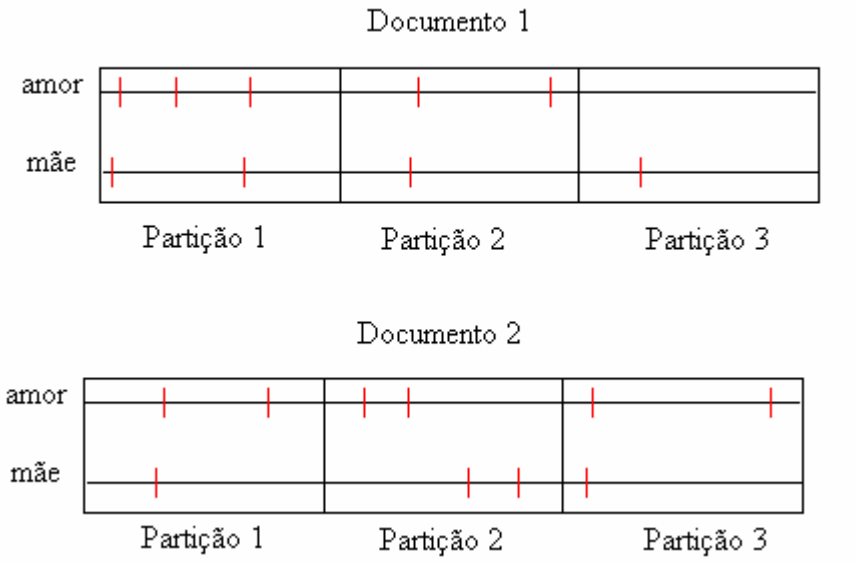



**Figura 27 Distribuição de termos dentro de dois documentos como proposto por Park et al.**

A proposta de (PARK, RAMAMOHANARAO et al., 2005) é utilizar a transformada wavelet de cada partição para representar o documento em diferentes resoluções e utilizar a magnitude e a fase .para determinar a freqüência e posição de cada termo. Assim, ele é capaz de determinar com diferentes graus de resolução qual a proximidade dos termos que figuram na consulta. Como por exemplo, na consulta "amor de mãe".

Essa proposta apesar de utilizar a transformada wavelet para fins de BRI, difere bastante de nossa proposta como será visto no capítulo 5. Primeiro, nossa proposta se trata de um meta-modelo, segundo porque nossa proposta opera diferente sobre os documentos, não considerando sua posição dentro do documento para cada documento individualmente, mas o fazendo de maneira global baseado na semântica de cada termo. Além disso, a proposta de (PARK, RAMAMOHANARAO et al., 2005) pode se tornar muito custosa se o número de partições de um documento for muito grande. Enquanto por outro lado, o tamanho variável de cada partição pode afetar significativamente o resultado das comparações entre documentos.

# 5 Proposta de modelo para BRI

Nesse capítulo nós iniciamos a defesa detalhada de nossa proposta para a utilização de uma nova estrutura matemática para sustentar o modelo de Busca e Recuperação de Informação (BRI) de documentos texto. Na seção 2.3 definimos um sistema de busca e recuperação de informação como uma tupla {T,Q,D,F,S}, onde T é o conjunto de termos; Q o conjunto de consultas; D o conjunto de documentos, F a estrutura matemática de suporte ao sistema; S a função de similaridade que compara as consultas com os documentos. Agora propomos a utilização das técnicas matemáticas empregadas no campo de processamento de sinais como estrutura matemática do modelo de um sistema de BRI. Convencionamos chamar esse modelo de Modelo de Sinais para Busca e Recuperação de Informação ou MSBRI.

Como estamos propondo redefinir a estrutura matemática F, devemos redefinir todos os demais elementos da quintupla acima. Portanto, vejamos como cada um desses elementos é definido.

**O Documento**

Enquanto no modelo vetorial os documentos são vetores no espaço de termos; e no modelo booleano os documentos são conjuntos; nossa proposta estabelece que os documentos são amostras discretas de um sinal físico imaginário (fictício), respeitando o princípio de amostragem de Nyquist. Desta fora abrimos caminho para a aplicação das técnicas apresentadas nas seções 3.3 e 3.4.

Nosso documento, portanto, é uma série espacial (sinal) discreta. Podemos representá-lo de diversas maneiras distintas, como veremos mais adiante. A forma mais simples de representar um documento é associando um índice inteiro a cada termo que aparece dentro dele. Assim, o eixo das abscissas representa a posição dentro do documento e as ordenadas representam o índice do termo em questão.

Existem muitas outras formas de se representar o documento, cada uma dessas formas dará origem a uma instância do MSBRI. Nessa dissertação exploraremos algumas propostas de representação de documento nas seções 5.2, 5.3, 5.4, 5.5 e capítulo 6.

**A consulta**

A consulta é considerada um documento e, portanto tem a mesma interpretação. Na prática, esperamos que a consulta em média seja pequena, mas o modelo não estabelece nenhum limite para ela. Pelo contrário, quanto maior a consulta, melhor nós poderemos estabelecer a similaridade entre ela e os documentos.

**A função de similaridade**

A função de similaridade será definida em cada uma das instâncias do MSBRI de maneira a não limitar sua usabilidade. Diversas métricas serão propostas, as mais simples são apenas a diferença ponto a ponto da curva que representa o documento e a curva que representa a consulta. As métricas mais complexas são o produto escalar entre a consulta e o documento no domínio wavelet. Mas de qualquer forma, a função de similaridade sempre tem o mesmo objetivo: fornecer uma maneira de quantificarmos o quanto um documento está próximo de uma consulta.

**Os termos**

Em um sistema de BRI de documentos texto os termos são as menores frações que podem ser manipuladas. No modelo vetorial os termos são considerados dimensões do espaço vetorial; no modelo booleano eles são elementos dos conjuntos. Em nossa proposta, a interpretação dos termos depende do tipo de codificação que estamos empregando. Esta é a principal diferença entre o MSBRI e os demais modelos clássicos: nosso modelo é um meta-modelo, pois sua interpretação depende da codificação empregada para representar os termos. Definimos duas funções básicas com este intuito: a Função de Codificação de Termos e a Função de Codificação de Documentos.

**A Função de Codificação de Termos**

A Função de Codificação de Termos (FCT) é o elemento fundamental de nossa proposta. Ela objetiva fornecer uma identificação única a cada termo distinto da coleção de termos T.

A idéia básica por trás da FCT é codificar os termos do documento de maneira que estes possam formar um sinal com características de um sinal físico e aproveitem, dessa maneira, as técnicas de processamento de sinais empregadas. Por exemplo, tipicamente um sinal de uma imagem, por mais complexa que seja, apresenta uma alta correlação entre pontos adjacentes. Uma possível função de codificação pode utilizar a correlação entre os termos dos documentos para estabelecer seu índice. Outro exemplo seria no caso de sinais estacionários que se formam a partir da soma de senóides de

freqüência constante em todo o domínio. Neste caso, uma função de codificação possível poderia associar uma freqüência a um determinado termo.

Portanto, a FCT possibilita que atribuamos uma interpretação aos termos do documento e consequentemente ao próprio documento.

**A Função de Codificação de Documento**

A Função de Codificação de Documento (FCD) objetiva fornecer uma representação única para cada documento distinto da coleção de documentos D. A representação do documento deve ser coerente com os critérios de Nyquist e a condição (21) , apresentada na seção 3.2.3, para que o documento possa ser considerado um sinal. Portanto, a FCD recebe como entrada o documento propriamente dito, e sua saída é a representação, enquanto sinal, desse documento. A relação entre a FCT e a FCD é a seguinte: A FCD deve usar a FCT para criar a representação do documento como um sinal.

## *5.1 O MSBRI é um meta-modelo*

A conseqüência direta da definição da FCD é que o MSBRI pode ser considerado um meta-modelo, uma vez que ele define uma estrutura para BRI baseada em processamento de sinais, mas não nos força a utilizar nenhuma codificação em particular. Dessa forma, ficamos livres para definir a codificação mais adequada para nossos fins. No decorrer desta dissertação apresentaremos algumas codificações possíveis para um documento e, discutiremos particularmente a FCD empregada para efeito dos experimentos desta dissertação.

A Figura 28 mostra esquematicamente como o MSBRI funciona. Primeiro o documento deve passar por um processamento para que obtenhamos sua representação enquanto sinal. Este processamento é feito pela FCT e FCD. Uma vez obtido o sinal que representa o documento devemos aplicar as técnicas de processamento de sinais, apresentadas nas seções 3.3 e 3.4, para extrair as componentes deste sinal. Essas componentes são usadas como índices do sistema de BRI e podem ser empregadas para fazer comparações entre os documentos e as consultas.

As consultas, como já referido anteriormente, devem passar pelo mesmo processo a fim de serem comparados com os documentos.

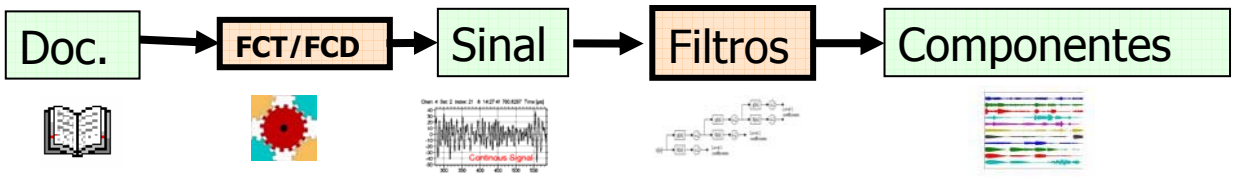



**Figura 28 Esquema do meta-modelo proposto**

Nas seções 5.2, 5.3, 5.4 e 5.5 , nós examinaremos algumas propostas para a FCT e FCD. Discutiremos qual o efeito e aplicabilidade de tais funções. Discutiremos a aplicação de filtros sobre a saída da FCD, particularmente as transformadas de Fourier e Wavelet. Finalmente, discutiremos detalhadamente a FCT/FCD implementadas para fins desta dissertação, capítulo (0), bem como a utilização de banco de filtros para obtenção da transformada wavelet.

## *5.2 Modelo Trivial (MT)*

O Modelo Trivial (MT) é a instância do MSBRI que faz uso de FCT e FCD triviais, muito simples e intuitivas. Nesse modelo, cada termo recebe um código único e seqüencial que o identifica. A representação de um documento é fornecida simplesmente pela disposição espacial dos termos na ordem em que eles aparecem no documento. Note que apesar da simplicidade, a FCD cria uma representação para o documento que respeita o critério estabelecido em (21), e portanto do ponto de vista de processamento de sinais, é considerado um sinal.

### 5.2.1 A Função de Codificação de Termo (FCT)

A função de Codificação de Termo (FCT) do MT é apresentada na equação (55). Nela cada termo é associado com um índice inteiro positivo. Esse índice é incrementado assim que um novo termo aparece, enquanto que a coleção de documentos é varrida linearmente.

$$FCT(t) \in N$$
$$FCT(t_i) = \begin{cases} 1 & \text{se } |T| = 0 \\ i & \text{se } t_i \in T \\ j=|T|+1 & T \cup \{t_j\} \end{cases} \qquad (55)$$

Onde,

FCT é a função de codificação de termos

T é o conjunto de termos do sistema

### 5.2.2 A Função de Codificação de Documentos (FCD)

A função de codificação de documento (FCD) do MT é igualmente simples: ela apenas concatena os índices dos termos que aparecem dentro do documento na mesma ordem em que eles aparecem. Portanto, o resultado da FCD é uma série espacial que representa o documento. A equação (56) mostra a representação paramétrica da FCD do MT.

$$FCD(d) = (s, FCT(t_s)) \tag{56}$$

Onde,

d é um documento com s termos

s é a s-ésima posição do documento

$t_s$ é o s-ésimo termo do documento

O feito final da aplicação da FCD sobre um documento real pode ser visto na Figura 29. Nele, o documento 01 da coleção Reuters (DAVID D.LEWIS, YIMING YANG et al., 2004) foi processado segundo o modelo trivial. O documento original tem cerca de 230 termos, cada qual com um índice diferente que varia de 1 a 200.

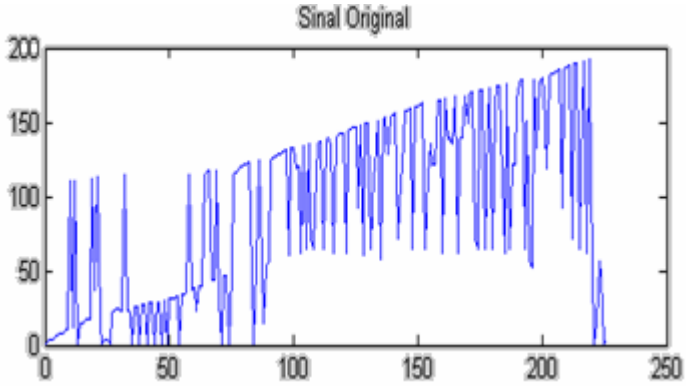



**Figura 29 Efeito da FCD do MT sobre documento Reuters**

Após definirmos a FCD, devemos escolher qual o tipo de filtro que será empregado para analisar o sinal obtido. No MT escolhemos um filtro que implementa a transformada de Fourier. Assim, o índice de nosso sistema será não a representação espacial do documento como mostrado na Figura 29, mas sim sua transformada de Fourier como mostrado na Figura 30.

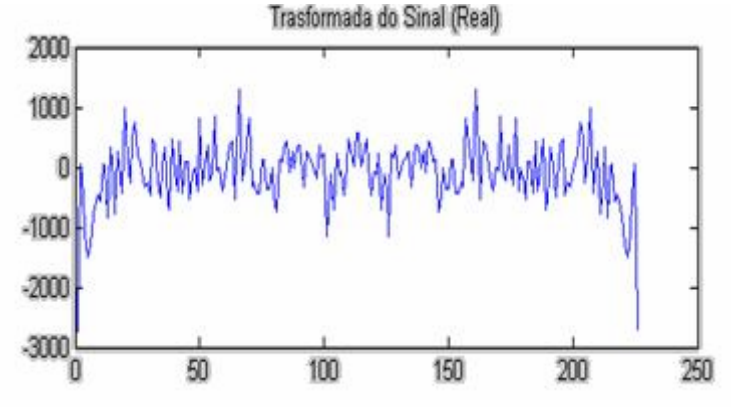



**Figura 30 Transformada de Fourier (parte real) da FCD(d)**

Cada documento deve ter sua transformada calculada e armazenada. Essas transformadas, segundo o teorema de Parseval (Seção 3.3.3), são representações equivalentes do documento, mas no domínio de freqüência. Portanto, há uma preservação da informação. Inclusive, podemos utilizar a transformada inversa para se obter a representação original do documento. A coleção de transformadas forma o índice do MT.

### 5.2.3 A consulta

A consulta no MT é considerada um documento como outro qualquer, mas de tamanho reduzido. Desta forma todas as operações efetuadas sobre um documento comum devem ser feitas sobre a consulta também.

### 5.2.4 A função de similaridade

A função de similaridade para o modelo trivial não poderia ser mais simples: ela é simplesmente a integral do módulo da diferença entre as transformadas de Fourier da FCD de um documento e da consulta. A equação (57) mostra como podemos operar os documentos e consulta a fim de compará-los.

$$sim(q,d) = \sum |\, FFT(FCD(d) - FCD(q))\,| \qquad (57)$$

Onde,

q é uma consulta

d é um documento

#### 5.2.4.1 Interpretação geométrica

A interpretação geométrica do FCD do MT nos ajuda a compreender melhor como este modelo pode ser empregado. A saída da FCD fornece uma curva que tem

uma forma; é razoável supor que documentos parecidos tenham formas parecidas. Os coeficientes de menor freqüência da Transformada de Fourier da saída da FCD fornecem a tendência da curva que representa o documento original. Os coeficientes de maior freqüência representam os detalhes desta mesma curva. Sendo assim, ao utilizarmos a equação (57) para comprar o documento com a consulta, na verdade estamos comparando duas formas.

### 5.2.5 Análise do MT

O Modelo Trivial sofre de algumas limitações: primeiro o tamanho da saída da FCD empregada neste modelo é proporcional ao número de termos do documento de entrada. Portanto, para documentos pequenos a dimensão da saída da FCD é pequena e para documentos grandes a dimensão da saída é grande. Do ponto de vista de processamento de sinais, se assumimos que a representação de sinal do documento seguiu o número de amostras mínimo para a futura reconstrução do hipotético sinal original, então a freqüência máxima de cada documento varia em função do número de termos do mesmo. Isso gera um complicador para efeitos da comparação entre documentos de tamanhos diferentes, uma vez que a função de classificação espera que as transformadas tenham dimensões iguais.

Para resolver este problema é preciso mudar ligeiramente a definição da função de codificação de documentos de forma que esta forneça uma saída de dimensão fixa independente do documento de entrada. Uma solução possível é apresentada por (PARK, RAMAMOHANARAO et al., 2005), onde o documento é "quebrado" em um número N fixo de pedaços e, a transformada de cada pedaço é feita independentemente.

Outro limite do Modelo Trivial (MT) é que ele associa aleatoriamente os índices aos termos, de maneira que a curva que representa um documento possui uma forte componente aleatória. Isso dificulta a concentração de informação no domínio de freqüência e a interpretação da transformada.

## *5.3 Modelo de Alta Correlação (MAC)*

O Modelo de Alta Correlação (MAC) é a instancia do MSBRI que funciona como uma extensão do Modelo Trivial (MT) proposto na seção 5.2. A principal contribuição desse modelo é fornecer uma Função de Codificação de Termos (FCT) mais elaborada que a empregada no MT.

A FCT do Modelo Trivial associa um número natural único a cada termo do documento. Este número é associado a cada termo segundo a ordem que ele aparece na coleção de documentos. Já o Modelo de Alta Correlação (MAC) propõe o seguinte: uma característica comum dos sinais que são normalmente analisados com o conjunto de técnicas matemáticas propostas nessa dissertação é que eles tendem a ter componentes altamente correlacionadas no domínio temporal (ou espacial). Por exemplo, em uma imagem, dois pontos (*pixels*) adjacentes tendem a ter aproximadamente a mesma intensidade. Variações repentinas na intensidade indicam alguma anomalia, em geral uma borda. A Transformada de Fourier aplicada a sinais como estes, tende a ter mais informação concentrada nos coeficientes de menor freqüência. Esta concentração de informação é aproveitada pelos algoritmos de compactação com perda. Ela possibilita que ignoremos os coeficientes de mais alta freqüência que representam os detalhes da imagem. O Exemplo 9 ilustra esse fenômeno.

Exemplo 9.

A Figura 31 apresenta o resultado visual da aplicação da transformada de Fourier sobre uma imagem periódica. A metade de cima apresenta duas figuras listradas verticalmente e horizontalmente. Abaixo de cada figura são apresentadas as suas respectivas transformadas de Fourier. Por se tratarem de figuras de cossenos com freqüência constante, a transformada de Fourier destas figuras são impulsos. Estes impulsos são os pequenos pontos brilhantes das imagens da metade de baixo da figura.

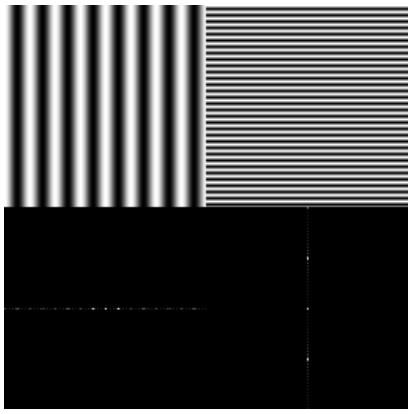


**Figura 31 Imagem de dois cossenos de freqüência constante (Fonte: (SITIO DA INTERNET, 2007c))**

Neste exemplo, fica claro como a transformada tende a concentrar informação nos coeficiente de baixa freqüência (centrais). Toda a imagem da metade de cima da Figura 31 é sintetizada pelos pequenos pontos brilhantes da metade de baixo. Apesar de os demais pontos existirem, eles têm baixa intensidade e seu valor é praticamente zero (preto na imagem).

A Figura 32 ilustra também o efeito da concentração de informação (energia) obtido pela transformada de Fourier. A metade de cima apresenta fotos de uma modelo. A foto mais a esquerda é a original e a foto a direita é a mesma foto sem os componentes de baixa freqüência. As metades de baixo da figura ?? apresenta a transformada de Fourier de cada foto respectivamente acima.

**Figura 32 Exemplo de concentração de informação obtido pela transformada de Fourier (Fonte: –image fourier)**

Note que na transformada da direita (abaixo), o ponto central mais brilhante foi removido. Como este ponto concentra a maior parte da informação da imagem, a modelo foi praticamente apagada. Note, no entanto, que os detalhes (bordas) permanecem na imagem.

O MAC tenta reproduzir essa condição de alta correlação entre pontos adjacentes através de sua FCT. A idéia é a seguinte: a cada termo deve ser associado um índice de forma a estabelecer uma ordem parcial. Assim, dois termos com índices adjacentes i e i+1 têm correlação ρ tal que a correlação entre esses dois termos é maior do que a do termo de índice menor com seu antecessor. A equação (58) representa

matematicamente essa idéia. Note que a equação (58) não nos mostra explicitamente como calcular $\rho$, nem mesmo como estabelecer a ordem proposta. No capitulo 6 veremos que o Modelo Wavelet também emprega essa mesma equação, e em momento oportuno mostraremos como ela pode ser calculada. Por hora, basta que compreendamos a idéia subjacente a equação (58).

$$t_i < t_{i+1} \leftrightarrow \rho(t_{i-1}, t_i) > \rho(t_i, t_{i+j}),\ \forall j > 0 \qquad (58)$$

Onde,

$t_i$ e $t_{i+1}$ são dois termos com índices adjacentes

$j$ é um índice natural qualquer

ρ é a correlação entre dois termos

Note que a definição recursiva da correlação da equação (59) nos leva a definição de um termo base $t_0$, em relação a qual o coeficiente de correlação inicial é calculado. O MAC define este termo base como o termo de menor IDF entre todos os termos. Portanto, o termo base é o termo que mais comumente aparece nos documentos (um termo de baixa entropia).

### 5.3.1 Interpretação da FCT

A Função de Codificação de Termos pode ser interpretada como a "energia de ligação " entre dois termos da coleção. Associamos o índice zero (0) ao termo com menor IDF, pois este termo é o mais comumente encontrado em todos os documentos, e de certa forma o termo que tem menor energia absoluta. Daí por diante, associamos os índices consecutivos aos termos que têm maior "energia de ligação" (correlação) com o termo de base consecutivamente.

### 5.3.2 A FCD do MAC

A Função de Codificação de Documento do Modelo de Alta Correlação é similar à empregada pelo Modelo Trivial, mas ao invés de os termos terem índices aleatórios, agora nós usamos a FCT definida anteriormente. Desta forma, a série espacial que representa o documento tende a ter menos componentes de alta freqüência, portanto a informação do efeito das transformadas tende a concentrar mais informação (energia) nos coeficientes de baixa freqüência, que representam a tendência da curva.

### 5.3.3 A função de similaridade

Assim como a FCD, a função de similaridade do MAC também é a mesma empregada no MT, ou seja, é a subtração dos coeficientes da transformada de Fourier da saída da FCD do MAC. Todas as observações feitas para o MT são válidas aqui também.

### 5.3.4 Análise do MAC

Por usar as mesmas FCDs empregadas no MT, o MAC sofre de problemas e limitações semelhantes: a dimensão do vetor de saída da FCD é proporcional ao tamanho do documento. Portanto, a comparação entre consulta e documento ganha um fator complicador, como descrito em 5.2.5. O MAC também privilegia a forma da curva de saída da FCD, e em geral a consulta de sistemas de BRI é muito curta para ser uma boa aproximação da forma da curva. (FLESCA, MANCO et al., 2005) mostra como sistemas de BRI podem explorar esta característica para uma comparação rápida de arquivos semi-estruturados. Ele apresenta o uso da transformada de Fourier para comparação de *schemas* de arquivos XML. Portanto, FCD's como as usadas no MAC e no MT são mais apropriadas para comparar formas do que propriamente a semântica dos documentos.

## *5.4 Modelo de Sinal Estacionário (MSE)*

O modelo de Sinal Estacionário é uma instância do MSBRI que tem uma abordagem diferente para a codificação de documentos como sinais. A proposta desse modelo é codificar os documentos de maneira que eles fiquem mais favoráveis à aplicação das transformadas de Fourier. Para isso, usamos sinais estacionários como os mostrados na seção 3.3.2.1 e relembrados abaixo.

### 5.4.1 Sinais Estacionários

Sinais estacionários são aqueles que apresentam os mesmos componentes de freqüência durante toda sua duração. Para sinais temporais, significa que para qualquer intervalo de tempo considerado, desde que respeitando o princípio de amostragem de Nyquist (22), obteremos os mesmos componentes de freqüência ao aplicarmos uma transformada de Fourier.

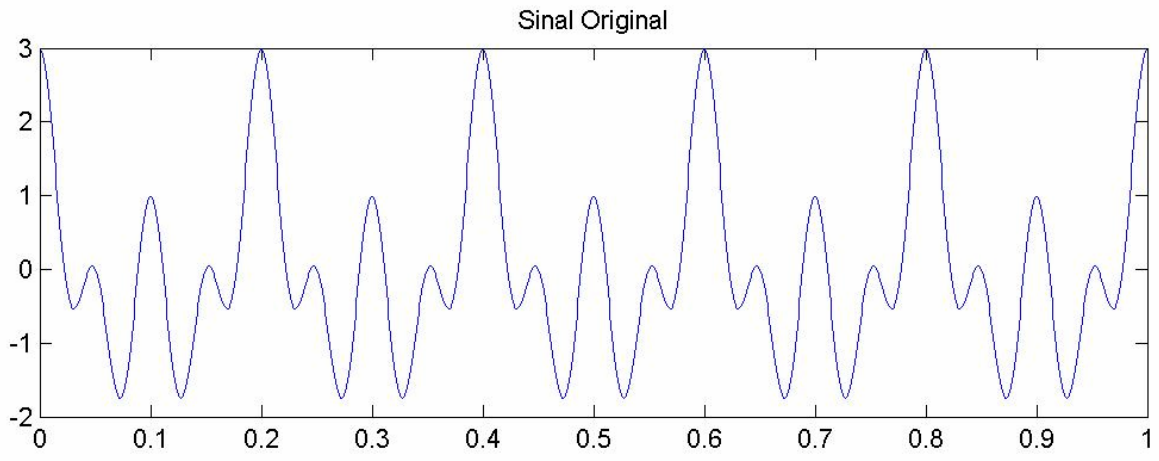



**Figura 33 Sinal Estacionário com freqüências 5, 10 e 20 Hertz**

. A Figura 33 mostra um exemplo de sinal estacionário. Ela é formada pela soma de três cossenos cujas freqüências são 5, 10 e 20 hertz. A equação (59) apresenta a forma matemática desse sinal . Perceba como para todo valor de t, todos os cossenos de diferentes freqüências são somados.

$$f(t) = \cos(5*2\pi t) + \cos(10*2\pi t) + \cos(20*2\pi t) \qquad (59)$$

A transformada de Fourier de um sinal estacionário possui picos que indicam a presença das freqüências que o compõe, ou de outra forma, a transformada de Fourier de um sinal estacionário é um histograma de freqüências de um sinal.

### 5.4.2 A Função de Codificação de Termos (FCT)

A função de codificação de termos desse modelo é bastante simples: ela apenas atribui um número inteiro a cada termo diferente da coleção, exatamente como feito no Modelo Trivial.

### 5.4.3 A Função de Codificação de Documento (FCD)

A função de codificação de documento do Modelo de Sinal Estacionário (MSE) propõe a codificação dos documentos como um sinal estacionário cujas freqüências são fornecidas pela FCT citada acima, como mostrado em (60).

$$FCD(d) = \sum_{s=0}^{N-1} \cos(FCT(d(s)) * 2 * \pi * t) \qquad (60)$$

Onde,

FCT é a função de codificação de termos

d(s) é o termo da posição s do documento d

### 5.4.4 A consulta

A consulta, assim como o documento, também deve ser codificada como um sinal estacionário cujas freqüências são fornecidas pelos termos que nela aparecem.

### 5.4.5 A função de similaridade

A função de similaridade é trivialmente a mesma que a usada no Modelo Trivial. Portanto, a função de classificação nada mais é que a diferença entre as transformadas de Fourier dos sinais que representam os documentos e as consultas.

### 5.4.6 Análise do MSE

O MSE é trivialmente equivalente ao modelo vetorial, uma vez que a transformada de Fourier dos sinais fornece o histograma de palavras que representa o documento e a consulta. A diferença desse modelo é que ele possibilita que outras técnicas que não a transformada de Fourier sejam aplicadas à apresentação do documento no domínio de freqüência. Por exemplo, poderíamos usar uma transformada de Fourier com Janelamento (*Short Time Fourier Transform)* para analisar o documento com diferentes janelas. Mas para isso seria interessante que o documento não fosse um sinal estacionário, fornecendo assim a informação da localização da freqüência ao longo de um sinal. Como veremos, essa é a idéia básica do próximo modelo proposto: o Modelo de Sinais Não Estacionário.

## *5.5 Modelo de Sinal Não Estacionário (MSNE)*

O modelo de sinal não estacionário (MSNE) é uma instância do MSBRI que propõe a codificação dos documentos como sinais não estacionários de forma a indicar o posicionamento de um termo dentro do documento. Vejamos como ele funciona.

### 5.5.1 Sinais não estacionários

Os sinais não estacionários são aqueles cujas componentes de freqüência diferem ao longo do sinal. Se estivermos tratando de um sinal no tempo, por exemplo, as componentes de freqüências desse sinal serão diferentes para o intervalo de tempo considerado. A grande maioria dos sinais com alguma função pratica é não estacionário, mas como vimos na seção 3.3.2, a transformada de Fourier não é muito adequada para tratar este tipo de sinal. Nessa mesma seção foi apresentada uma variante da transformada de Fourier conhecida como Transformada de Fourier com Janelamento

que possibilita a análise de sinais não estacionários por meio da aplicação de uma janela gaussiana de determinada largura.

A Figura 34 apresenta um sinal não estacionário cujas freqüências são 5, 10 e 20 hertz. Note que ao contrário do sinal estacionário, as freqüências aparecem sozinhas em determinados intervalos de tempo (segundos).

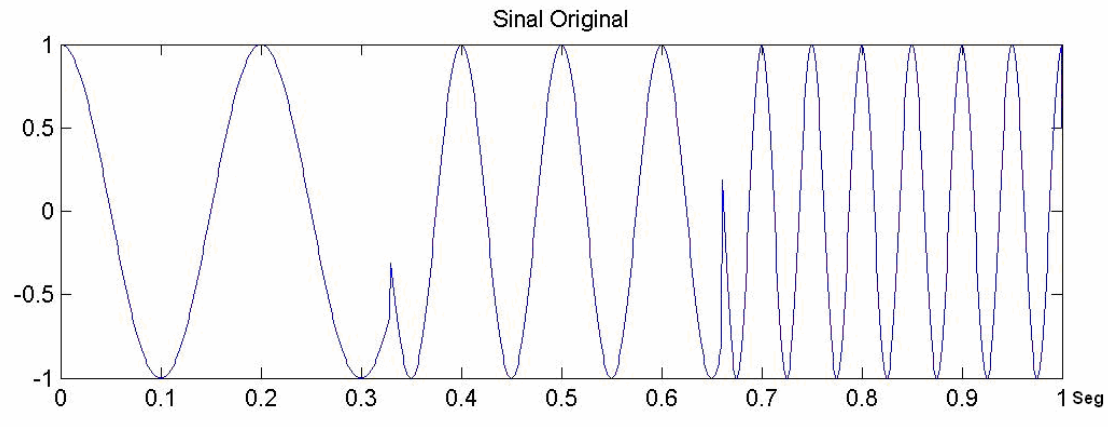



**Figura 34 Sinal Não Estacionário**

No intervalo entre 0 e 0,33 o sinal possui 5 hertz. No intervalo entre 0,33 e 0,66 o sinal apresenta 10 hertz e no intervalo entre 0,66 e 1 o sinal tem 20 hertz.

A equação (60) apresenta a fórmula que originou o sinal acima apresentado. Note que as freqüências aparecem em intervalos de tempo distintos.

$$f(t) = \begin{cases} \cos(5 * 2\pi t),\ 0 \le t<0{,}66 \\ \cos(10 * 2\pi t),\ 0{,}66 \le t<0{,}99 \\ \cos(20 * 2\pi t),\ 0{,}99 \le t<1 \end{cases} \qquad (61)$$

A transformada de Fourier do sinal apresentado na Figura 34 é apresentado abaixo. Note que apesar das irregularidades causadas pelos efeitos de borda, ela apresenta picos que indicam a presença mais forte das componentes de freqüência mostrados em (60).

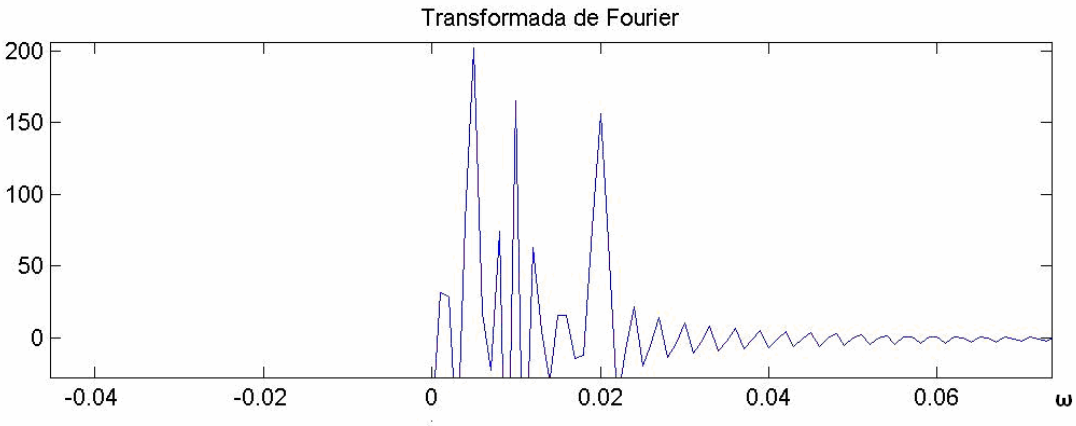



**Figura 35 Transformada de Fourier de um sinal não estacionário.**

A transformada de Fourier com Janelamento (*Short Time Fourier Transform* ) aplicada a este mesmo sinal pode ser ajustada para capturar os intervalos estacionários desse sinal com diferentes resoluções de freqüência. Em último caso, quando a janela escolhida contiver todo o sinal, o resultado da transformada será como o da Figura 35.

### 5.5.2 A Proposta do MSNE

A proposta do modelo de sinal não estacionário é a utilização da Transformada de Fourier com Janelamento (STFT) no lugar da transformada de Fourier na análise de sinais não estacionários, da mesma maneira que o MSE a usa para sinais estacionários.

### 5.5.3 A Função de Codificação de Termos

Da mesma maneira que o MSE, este modelo usa uma Função de Codificação de Termos (FCT) proposta no modelo trivial (MT), ou seja, atribuímos um inteiro seqüencial a cada novo termo que aparece na coleção de documentos à medida que eles vão aparecendo.

### 5.5.4 A Função de Codificação de Documento

O diferencial desse modelo é a Função de Codificação de Documentos (FCD). Atribuímos a cada termo do documento uma freqüência baseada na FCT. ,No entanto ao invés de somarmos cossenos para todo intervalo [a,b] dividimos o domínio em |T|, o número total de termos, os intervalos iguais e atribuímos uma freqüência específica a cada intervalo baseado na ordem espacial dos termos dentro do documento. A equação (62) apresenta matematicamente está idéia.

$$FCD(d) = \sum_{s=0}^{n-1} \cos(FCT(d(s)) * 2\pi t) * w(t,s) \quad (62)$$

Onde,

w(t,s) é uma janela quadrada no intervalo t=[T/s,T/(s+1)]

s é a s-ésima posição do documento d

T é o valor máximo de t considerado.

Basicamente, (62) cria um sinal estacionário como o apresentado na Figura 34, onde para cada intervalo de tamanho **T/(n-1)** é apresentada uma freqüência referente ao termo que ocupa a posição **s** do documento **d**. Para este tipo de sinal, a transformada de Fourier fornece o histograma de termos do documento como no MSE, mas possibilita que usemos técnicas mais apuradas como a STFT para analisá-lo.

### 5.5.5 A função de similaridade

Como no Modelo Trivial, a função de similaridade é a diferença entre as transformadas de Fourier aplicada ao sinal que representa a consulta e o documento.

### 5.5.6 A análise

A principal característica do Modelo de Sinais Não Estacionário é que, dada sua FCD, ele possibilita que usemos técnicas como STFT para fazer uma análise sobre o documento em diversos níveis de resolução. Potencialmente podemos fazer uma análise de contexto, encontrando termos segundo a ordem em que aparecem dentro do documento.

# 6 Modelo corrente (Modelo Wavelet)

O Modelo Wavelet merece um capítulo a parte por se tratar da proposta central dessa dissertação. Esse modelo, como veremos, é a única instância do MSBRI que foi totalmente implementada e avaliada. No capítulo 5 apresentamos a proposta do MSBRI como um meta-modelo de Busca e Recuperação de Informação. Demos exemplos de instâncias desse modelo e destacamos a importância da FCT e FCD. Agora entraremos em detalhes sobre o Modelo Wavelet.

## *6.1 Modelo_Wavelet*

O modelo Wavelet (MW) é resultado da contribuição de todos os modelos anteriormente apresentados. A idéia básica desse modelo é a utilização da Transformada Wavelet Discreta (Discrete Wavelet Transform, DWT) para analisar documentos codificados como sinais. Diferentemente da Transformada de Fourier (Fourier Transform, FT) e da Transformada de Fourier com Janelamento (Short Time Fourier Transform, STFT), a Transformada Wavelet possui a propriedade de multiresolução que pretendemos explorar neste modelo.

A especificação do modelo Wavelet é mais detalhada do que os modelos anteriores por dois motivos: Primeiro porque esse foi o único modelo completamente implementado e analisado para fins dessa dissertação. Segundo, porque esse modelo foi o único avaliado e comparado com um sistema de busca e recuperação de informação comercial, o LUCENE (OTIS GOSPODNETIC e ERIK HATCHER, 2007). Como na descrição dos modelos anteriores, passamos agora a descrição de sua estrutura, partindo da Função de Codificação de Termos, FCT, até a função de classificação. Ao longo de nossa descrição serão apresentados exemplos que ilustram as idéias aqui propostas e facilitam o entendimento do modelo por parte do leitor.

### 6.1.1 A Função de Codificação de Termos

A função de codificação de termos empregada pelo Modelo Wavelet é a mesma empregada pelo Modelo de Alta Correlação (MAC). O objetivo principal desta função é manter os termos que tenham uma semântica próxima com índices próximos. Para isso partimos do princípio que a semântica de um termo pode ser capturada estatisticamente. Por exemplo, se em uma coleção de documentos os termos “rede” e “computador”,

aparecem sempre juntos. Podemos considerá-los sinônimos estatísticos. Uma vez que seria equivalente pesquisa por rede ou computador, para a coleção em questão. De uma maneira mais geral, termos que tenham sua distribuição entre os documentos iguais podem ser considerados de sinônimos estatísticos. Este princípio não é exclusividade do modelo proposto, ele já foi empregado com muito sucesso na Analise de Semântica Latente (DUMAIS S.T., DEERWESTER S. et al., 1990).

Na seção 5.3, quando apresentamos o MAC, definimos a FCT de forma abstrata, apenas dissemos que a correção entre dois termos com índices adjacentes deveria ser máxima. Agora é o momento de detalhar esse conceito.

Primeiro partimos da definição do peso dos termos dentro do documento, para isso utilizamos a freqüência de termos (TF, *Term Frequency*) e inverso da freqüência de documento (IDF, *Inverse Document Frequency*). Basicamente, TF é uma medida de importância local de um termo, quanto mais freqüente o termo dentro de um documento, mais importante ele é considerado. Enquanto IDF é uma medida de importância global, quando menos documentos contiverem o termo em questão, mais importante ele é, pois seu fator de discriminação é mais alto.

A equação (6) mostra como podemos obter o fator TF a partir de uma simples contagem dos termos de um documento.

$$TF_{ij} = \frac{f_{ij}}{\max(f_j)} \tag{63}$$

Onde,

$TF_{ij}$ é a freqüência relativa do termo i dentro do documento j

$f_{ij}$ é a freqüência absoluta do termo i dentro do documento j

$\max(f_j)$ é o fator normalizador da freqüência absoluta, ou seja, é a freqüência máxima de um termo que aparece dentro do documento j.

A equação (64) mostra como podemos calcular o valor de IDF a partir da contagem do número de documentos que contém determinado termo.

$$IDF_i = \log_{10}(\frac{N}{n_i}) \tag{64}$$

Onde,

N é o numero total de documentos da coleção, ou seja, N=|D|

$n_i$ é o número de documentos que contém o termo i

O peso (importância) de um termo dentro de um documento é dado pela equação (65), que é o produto entre TF e IDF.

$$w_{ij} = IDF_i * TF_{ij} \tag{65}$$

Onde,

IDF é como definido em (64) e,

TF é como definido em (6).

Podemos, exatamente como no modelo vetorial (VSM), utilizar a matriz termo-documento para organizar os pesos de todos os termos em relação a todos os documentos. A partir dessa matriz podemos obter a matriz termo-termo que indica a 'energia de ligação' entre os termos da coleção. A equação (66) mostra como podemos obter esta matriz.

$$M_{tt} = \overline{M}_{td} \bullet \overline{M}^{t}_{td} = \begin{bmatrix} \overline{w}_{11} & \cdots & \overline{w}_{N1} \\ \vdots & \ddots & \vdots \\ \overline{w}_{1t} & \cdots & \overline{w}_{Nt} \end{bmatrix} \bullet \begin{bmatrix} \overline{w}_{11} & \cdots & \overline{w}_{t1} \\ \vdots & \ddots & \vdots \\ \overline{w}_{1N} & \cdots & \overline{w}_{tN} \end{bmatrix} \tag{66}$$

Onde,

$M_{tt}$ é a matriz termo-termo,

$\overline{M}_{td}$ é a matriz termo-documento ¨com as linhas normalizadas quadraticamente, e

$\overline{w}_{ij}$ ¨são os pesos normalizados quadraticamente.

Um elemento aij da matriz $M_{tt}$ representa a 'energia de ligação' entre o termo i e j. Vamos analisar mais cuidadosamente esta operação: o produto das duas matrizes é equivalente ao produto escalar dos vetores normalizados que representam as distribuições dos termos nos documentos. Portanto, o que estamos efetivamente fazendo

ao efetuarmos a operação representada pela Equação (66), é, na verdade, a comparação entre as distribuições dos termos. Portanto, termos com distribuições 'próximas' têm coeficientes mais elevados. Trivialmente, o resultado dessa operação é uma matriz simétrica com elementos da diagonal igual a 1.

Uma vez de posse da matriz $M_{tt}$, nós podemos definir uma ordem parcial para os termos da seguinte maneira: Primeiro, é necessário estabelecer qual será o termo inicial. Em nossa implementação escolhemos o termo de menor IDF que é o termo mais comum entre todos os documentos. A partir dele, caminhamos para o termo com maior correlação e continuamos este processo até que todos os termos tenham sido visitados uma única vez. Em caso de empate das correlações, trivialmente caminhamos para o termo de maior IDF. Ao final do processo, cada termo terá um índice que será usado como valor da FCT.

Exemplo 10.

Neste exemplo mostramos como a FCT do MW funciona. A partir da coleção de documentos, calculamos os pesos dos termos e em seguida montamos a matriz termo-documento. Por fim, calculamos a matriz de 'energia' entre os termos. Esta matriz nos fornece o resultado da FCT. Vejamos concretamente como proceder.

Seja a coleção de documentos abaixo:

| |
|---|
| D1 = (A, B, B, C, C) |
| D2 = (A, A, B, B) |
| D3 = (B, C, B, C) |
| D4 = (C, D) |

O primeiro passo é construir a matriz termo-documento, mas para isso precisamos calcular o peso de cada termo dentro de cada documento. Portanto, precisamos calcular os parâmetros TF e IDF, como mostrado nas equações (6) e (64) respectivamente. As seguintes tabelas nos ajudam nos cálculos.

A Tabela 1 apresenta a freqüência de documentos que contém cada termo ($n_i$).

**Tabela 1 Freqüência de documentos que contém os respectivos termos.**

| Termo | $n_i$ |
|---|---|
| A | 2 |
| B | 3 |
| C | 3 |
| D | 1 |

**Tabela 2 Valores de IDF para cada termo.**

| Termo | IDF = $\log_{10}(N/n_i)$ |
|---|---|
| A | 0,30 |
| B | 0,12 |
| C | 0,12 |
| D | 0,60 |

A matriz termo-documento apresentada na equação (67) é resultado da aplicação de (65) sobre a coleção de documentos considerada neste exemplo. Cada elemento $a_{ij}$ da matriz $M_{td}$ indica o peso (energia) de um termo dentro do respectivo documento. Note que inicialmente os termos da matriz estão ordenados segundo uma ordem qualquer, no caso, a ordem lexicográfica.

$$M_{td} = \begin{matrix} \\ A \\ B \\ C \\ D \end{matrix} \overbrace{\begin{bmatrix} 0,15 & 0,30 & 0,00 & 0,00 \\ 0,12 & 0,12 & 0,12 & 0,00 \\ 0,12 & 0,00 & 0,12 & 0,12 \\ 0,00 & 0,00 & 0,00 & 0,60 \end{bmatrix}}^{Documentos} \tag{67}$$

A partir de (67) calculamos o valor de $M_{tt}$ como mostrado em (69). Note, no entanto, que primeiro normalizamos os vetores do espaço

linha da matriz $M_{td}$, de forma que o produto por sua transposta nos forneça o cosseno do ângulo entre as distribuições de termos. A matriz apresentada em (68) é a versão normalizada de (67).

$$\widehat{M}_{td} = \begin{bmatrix} 0,45 & 0,89 & 0,00 & 0,00 \\ 0,58 & 0,58 & 0,58 & 0,00 \\ 0,58 & 0,00 & 0,00 & 0,58 \\ 0,00 & 0,00 & 0,00 & 1,00 \end{bmatrix} \quad (68)$$

A partir da matriz normalizada, (68), obtemos resultado apresentado em (69). Esta operação é análoga ao cálculo de uma matriz de correlação. E sua interpretação é uma medida de similaridade entre as distribuições dos termos entre os documentos segundo seus pesos definidos pela equação (65).

$$M_{tt} = \begin{bmatrix} 0,45 & 0,89 & 0,00 & 0,00 \\ 0,58 & 0,58 & 0,58 & 0,00 \\ 0,58 & 0,00 & 0,00 & 0,58 \\ 0,00 & 0,00 & 0,00 & 1,00 \end{bmatrix} \bullet \begin{bmatrix} 0,45 & 0,58 & 0,58 & 0,00 \\ 0,89 & 0,58 & 0,00 & 0,00 \\ 0,00 & 0,58 & 0,00 & 0,00 \\ 0,00 & 0,00 & 0,58 & 1,00 \end{bmatrix} = \begin{bmatrix} 1,00 & 0,77 & 0,26 & 0,00 \\ 0,77 & 1,00 & 0,67 & 0,00 \\ 0,26 & 0,67 & 1,00 & 0,58 \\ 0,00 & 0,00 & 0,58 & 1,00 \end{bmatrix} \quad (69)$$

A partir de $M_{tt}$ é possível se estabelecer a ordem parcial entre os termos. Iniciamos pelo termo de menor IDF, no caso B, que equivale à segunda linha da matriz. De B transitamos para A, pois este é o termo com maior energia de ligação (0,5). Em seguida vamos para C, mas note que deveríamos voltar à B, mas como este já fora visitado, nós passamos para o próximo termo. Por fim, chegamos a D. A ordem final estabelecida é, portanto, (B, A, C, D). Cada termo deve receber um código inteiro segundo sua posição, e este é o resultado da FCT do MW. A Tabela 3 resume este resultado.

**Tabela 3 Resultado da FCT.**

| Termo | FCT |
|---|---|
| A | 2 |
| B | 1 |

| | |
|---|---|
| C | 3 |
| D | 4 |

### 6.1.2 Função de Codificação de Documentos (FCD)

A função de codificação de documento empregada no Modelo Wavelet (MW) pode ser entendida como o histograma de termos, ordenados segundo seus respectivos índices fornecidos pela FCT, transladado para o domínio wavelet. Vejamos mais detalhadamente como funciona esta FCD.

Primeiro precisamos reordenar as linhas da matriz $M_{td}$ segundo a ordem estabelecida pela FCT, como mostrado em (70).

$$M_{\sin} = \begin{matrix} \\ T_1 \\ \vdots \\ T_t \end{matrix} \overbrace{\begin{bmatrix} w_{11} & \cdots & w_{1N} \\ \vdots & \ddots & \vdots \\ w_{t1} & \cdots & w_{tN} \end{bmatrix}}^{Documentos} \tag{70}$$

Onde,

$M_{\sin}$ é a matriz sinal,

$T_i$ é o termo cujo FCT é igual a i e,

$w_{ij}$ é o peso do termo i no documento j como mostrado em (65)

A matriz sinal, $M_{\sin}$, nos fornece a representação enquanto sinal de todos os documentos da coleção. Individualmente, cada elemento do espaço coluna dessa matriz representa um documento e pode ser interpretada como um histograma de termos ordenados.

Finalmente, devemos aplicar a transformada wavelet sobre os vetores do espaço coluna de $M_{\text{sin}}$ para obter uma matriz equivalente no domínio wavelet, que chamaremos de matriz wavelet, $M_{wav}$, como mostrado em (71).

$$M_{wav} = \overbrace{\left[ W[\vec{d}_1] \quad \cdots \quad W[\vec{d}_N] \right]}^{Documentos} \tag{71}$$

Onde,

$M_{wav}$ é a representação de $M_{\text{sin}}$ no domínio wavelet,

W[] é o operador da transformada wavelet e,

$\vec{d}_i$ é o histograma de termos ordenados que representa um documento em $M_{\text{sin}}$.

Cada um dos vetores $W[\vec{d}_i]$ é resultado da FCD aplicada a um documento especifico. Note que uma vez no domínio wavelet, os documentos herdam a propriedade de multiresolução da transformada, que agora pode ser explorada pelo modelo. Na próxima seção veremos como interpretar e tirar proveito dessa propriedade.

Exemplo 11.

Continuando o Exemplo 10, agora podemos montar a matriz sinal que dará origem a matriz wavelet. Vejamos como proceder:

A ordem estabelecida pela FCT foi a seguinte: (B,C,A,D). Portanto a matriz sinal é simplesmente a matriz termo-documento com algumas linhas permutadas como mostrado em (72).

$$M_{\text{sin}} = \begin{matrix} \\ B \\ A \\ C \\ D \end{matrix} \overbrace{\begin{bmatrix} 0,12 & 0,12 & 0,12 & 0,00 \\ 0,15 & 0,30 & 0,00 & 0,00 \\ 0,12 & 0,00 & 0,12 & 0,12 \\ 0,00 & 0,00 & 0,00 & 0,60 \end{bmatrix}}^{Documentos} \tag{72}$$

A partir de $M_{\text{sin}}$ podemos calcular $M_{\text{wav}}$, que é a matriz equivalente no

domínio wavelet como mostrado em (73).

$$M_{wav} = \begin{bmatrix} 0,19 & 0,21 & 0,12 & 0,36 \\ 0,07 & 0,21 & 0,00 & -0,36 \\ -0,02 & -0,12 & 0,08 & 0,00 \\ 0,08 & 0,00 & 0,08 & 0,34 \end{bmatrix} \quad (73)$$

### 6.1.3 A Multiresolução sobre os documentos

A forma (*shape*) do histograma de termos ordenados reflete a semântica do documento. Por exemplo, vamos supor que para uma coleção em particular o termo "amor" tenha índice igual a 10, é esperado que os termos que tenham distribuições próximas a de "amor" tenham índices próximos. É provável que o termo "mãe" receba o índice 11, por se tratar de um termo que está fortemente ligado a "amor". Outros possíveis índices seriam, (pai, 12), (paixão,13) e assim sucessivamente. Como a diversidade de termos de toda a coleção é, geralmente, muito maior do que os termos contidos em um documento, o histograma de termos se torna uma função muito esparsa, ou seja, ela é igual a zero sobre quase todo o domínio. Os pontos diferentes de zero refletem os termos que aparecem dentro de um documento. Em um histograma comum, onde os termos estão distribuídos em uma ordem qualquer, os pontos diferentes de zero não refletem nenhuma tendência. São apenas 'picos semânticos' dentro do documento. Já no caso do histograma usado como FCD do MW (histograma de termos ordenados), há uma tendência de aglutinação de termos correlatos, uma vez que eles estão ordenados segundo alguma semântica. Esta característica possibilita tirarmos proveito da propriedade de multiresolução da transformada wavelet.

O 'efeito wavelet' pode ser entendido a partir da propriedade de multiresolução. O Modelo Wavelet (MW) propõe a utilização deste efeito para criar índices em diferentes graus de detalhes. O índice de um sistema de Busca e Recuperação de Informação é representado matematicamente pela matriz termo-documento, $M_{td}$, que relaciona cada termo com seus respectivos documentos. Em nossa proposta, $M_{wav}$ é a matriz equivalente a $M_{td}$, mas ao contrário desta última, as linhas da primeira não são termos, mas sim coeficientes da função wavelet mãe. Por isso a interpretação semântica não é direta. Vejamos como interpretar os documentos no domínio wavelet.

Primeiro considere apenas um elemento do espaço coluna de $M_{wav}$, este elemento representa um documento. Por razões de simplicidade vamos trabalhar com a versão não normalizada da transformada wavelet. Neste caso, os cálculos se tornam muito simples como mostrado pela árvore do Exemplo 7.

Ao reduzirmos a resolução de uma função, obtemos uma versão mais suave da mesma, isso porque filtramos as componentes de freqüência mais alta. No caso discreto, ao reduzirmos a resolução de um pulso, obtemos dois novos pulsos de intensidade igual à metade do anterior, este efeito de certa forma é equivalente suavização obtida no caso contínuo (veja exemplo).

Exemplo 12.

Neste exemplo mostramos como é o comportamento da transformada wavelet para pulsos e mostramos como podemos nos beneficiar do efeito da multiresolução para efeitos da Busca e Recuperação da Informação.

Para o mesmo conjunto de termos considerados nos exemplos anteriores, seja um novo documento hipotético cuja representação em $M_{sin}$ seja $\vec{d} = [0,00 \quad 1,00 \quad 0,00 \quad 0,00]^T$, em outras palavras um documento que contenha apenas o termo A. A transformada wavelet desse vetor é mostrada em (74)

| | |
|---|---|
| $W[\vec{d}] = [\frac{1}{4} \quad \frac{-1}{4} \quad \frac{1}{2} \quad 0]^T$ | (74) |

Onde,
W[] é o operador wavelet

Note que a interpretação de termos é perdida quando estamos no domínio wavelet, mas como a transformada Wavelet é um operador linear equivalente a uma rotação (preserva distância), ela garante que estas representações são equivalentes. Ou seja, através da transformada inversa podemos obter o vetor (sinal) original. Vamos reduzir a resolução de nosso vetor no domínio wavelet e observar o

que ocorre no domino espacial (domínio do documento).

No MW a redução de um nível de resolução equivale a poda pela metade do vetor no domínio wavelet. O vetor apresentado em (75), mostra o vetor (74) podado pela metade.

| | |
|---|---|
| $W[\vec{d}] = [\frac{1}{4} \quad \frac{-1}{4}]^T$ | (75) |

Ao aplicarmos a transformada inversa para obter o vetor original, alcançamos o vetor mostrado em (76). De volta ao domínio espacial (domínio do documento), podemos interpretar cada um das dimensões do vetor como um termo diferente.

| | |
|---|---|
| $\vec{d} = [\frac{1}{2} \quad \frac{1}{2} \quad 0 \quad 0]^T$ | (76) |

Agora podemos perceber o efeito da redução de resolução, obtivemos um novo documento cujos termos são A e B. De fato se estes dois termos estão fortemente relacionados (e estão porque têm índices próximos), uma busca por A feita em uma resolução menor, pode acabar por encontrar B. Este pode ser exatamente o desejo do usuário. Uma vez, que A e B estão fortemente relacionados.

Portanto a redução de resolução significa a sobreposição de termos relacionados que podem indicar a presença de um conceito único, os termos “mãe”, “amor”, “pai” e “paixão” para uma coleção particular podem estar tão relacionados que redução de resolução pode ser bastante benéfica se aplicada na dose correta. É claro que a função de codificação de termos (FCT) não reconhece a verdadeira semântica por trás dos termos, mas sim supõe que termos com distribuições parecidas são parecidos, o que do ponto de vista estatístico é correto, mas pode apresentar aberrações semânticas. Mas na prática, estudos com LSI mostram que este efeito é em geral benéfico se aplicado na medida certa.

Por fim, por se tratar de um meta-modelo, o MW não precisa necessariamente utilizar formulas estatísticas para determinara a associação entre termos, é possível, com

a ajuda de thesaurus, que obtenhamos mais termos correlatos do que empregando meramente estatística.

### 6.1.4 A poda

A propriedade de multiresolução da transformada wavelet é obtida a partir da operação de poda dos últimos coeficientes da transformada. Em nossa implementação a poda da matriz $M_{wav}$ deve ser feita sempre se reduzindo a metade o número de linhas da matriz original. Isso equivale à redução de um grau de resolução no domínio do documento. Essa perda de detalhes pode ser feita em diversos níveis, podendo ser prejudicial ou beneficiar a qualidade da resposta do sistema de BRI. A princípio esta afirmação pode parecer paradoxal, mas se analisarmos melhor o efeito da poda da matriz wavelet entenderemos este fenômeno.

A redução do número de linhas da matriz $M_{wav}$, segundo o ponto de vista do campo de Processamento de Sinais, age como uma filtragem das altas freqüências. No domínio do documento (domínio espacial[3]) esta filtragem age como uma suavização do sinal (Veja Exemplo 12). As altas freqüências podem ser meramente ruídos e, portanto a filtragem é benéfica ao sistema de BRI. Por outro lado, se exagerarmos na filtragem, nós podemos obter uma perda de informação significativa e, portanto piorar a qualidade do sistema de BRI. O grau de filtragem necessário depende exclusivamente das características da coleção que está sendo indexada.

O número de níveis ou graus de resolução que uma coleção possui é proporcional ao número de termos como mostrado em (77).

$$R_{\max} = 1 + \log_2(|T|) \qquad (77)$$

Onde,

$R_{\max}$ é o nível máximo de resolução atingido pela um documento da coleção.

|T| é o número de termos da coleção.

[3] Tomamos a liberdade de comparar o domínio do documento com o domínio espacial por analogia para facilitar o entendimento de leitor. Note no entanto, que rigorosamente não o domínio do documento não é verdadeiramente um domínio espacial, apesar da analogia.

Perceba que a base 2 do logaritmo é devido a poda ser feita sempre se reduzindo o número de termos da matriz $M_{wav}$ pela metade.

### 6.1.5 A Função de Similaridade

Igualmente ao modelo vetorial (VSM), assumimos que o histograma de termos ordenados, que representa um documento é um vetor. Fazemos os mesmo para a consulta e utilizamos o cosseno do ângulo entre os vetores que os representam como métrica para classificação dos documentos segundo seu grau de similaridade com a consulta. No entanto, precisamos fazer uma pequena adaptação na equação original para embutir a propriedade de multiresolução. A equação (78) apresenta a função de similaridade.

$$\cos(\alpha) = sim(q,d) = \left\langle \overline{W_n[q]}, \overline{W_n[d_i]} \right\rangle = \overline{W_n[q]} \bullet \overline{W_n[d_i]} \qquad (78)$$

Onde,

< . , .> é o operador de produtor escalar

$\overline{W_n[.]}$ é o resultado normalizado da transformada wavelet (de um documento ou consulta) na resolução n.

q é a consulta realizada

di é o i-ésimo documento da coleção

α é o ângulo entre os vetores q e d

A equação (78) nos mostra como devemos calcular a similaridade entre um documento e uma consulta. Note que $W_n[.]$ é uma coluna da matriz $M_{wav}$ na resolução n, ou seja, com uma poda equivalente ao grau de resolução estabelecido.

Exemplo 13

Nesse exemplo, mostramos como uma consulta pode ser feita em diversos níveis de resolução. Vamos aproveitar a matriz $M_{wav}$ usada no Exemplo 12, reproduzida abaixo. Como sempre as colunas dessa matriz

representam um documento de nossa coleção.

| | |
|---|---|
| $M_{wav} = \begin{bmatrix} 0,19 & 0,21 & 0,12 & 0,36 \\ 0,07 & 0,21 & 0,00 & -0,36 \\ -0,02 & -0,12 & 0,08 & 0,00 \\ 0,08 & 0,00 & 0,08 & 0,34 \end{bmatrix}$ | (79) |

A matriz $M_{wav}$ é apresentada em (79) tem $\log_2(4)+1=3$ níveis de resolução possível. No nível mais alto, temos uma equivalência ao modelo vetorial (VSM) nas demais resoluções temos uma perda de gradativa de detalhes do documento original.

Seja uma consulta pelo termo A, ou seja, no domínio do documento esta consulta seria representada pelo seguinte vetor q=(0, 1, 0, 0). Devemos aplicar a FCD a este vetor a fim de obtermos sua representação no domínio wavelet. A matriz (80) mostra o resultado dessa operação

| | |
|---|---|
| $W[\vec{q}] = [0,50 \quad 0,50 \quad -0,71 \quad 0,00]$ | (80) |

Onde,

W[.] é a transformada wavelet do vetor consulta.

Agora, como mostrado na equação (78), devemos calcular o cosseno do ângulo entre os vetores que representam a consulta e os documentos no domínio wavelet em diferentes níveis de resolução. Por medida de simplicidade vamos normalizar as matrizes apresentadas em (79) e (80) de forma que seu produto nos forneça diretamente o cosseno do ângulo. Em (81) mostramos o resultado dessa operação, a matriz resultante do produto é uma medida de similaridade. Quando mais próxima de 1, maior a similaridade entre o documento e a consulta.

$$\overbrace{\begin{bmatrix} 0,50 \\ 0,50 \\ -0,71 \\ 0,00 \end{bmatrix}^{T}}^{Consulta} \bullet \overbrace{\begin{bmatrix} 0,86 & 0,65 & 0,71 & 0,59 \\ 0,33 & 0,65 & 0,00 & -0,59 \\ -0,09 & -0,39 & 0,50 & 0,00 \\ 0,37 & 0,00 & 0,50 & -0,55 \end{bmatrix}}^{Documentos} = \overbrace{[0,66 \quad 0,93 \quad 0,00 \quad 0,00]}^{Resultados} \qquad (81)$$

Note que o resultado, como esperado, é equivalente ao modelo VSM. Analisando a matriz de resultados concluímos que o documento $d_2$ é o mais similar à consulta, se voltarmos ao domínio do documento, nós chegaremos a essa mesma conclusão uma vez que o peso do termo A nesse documento é mais alto, seguido do peso do termo A no documento $d_1$. Por fim, nem o documento $d_3$ nem o documento $d_4$ possuem o termo A e, portanto suas similaridades são zero.

Faremos o mesmo procedimento mostrado acima, mas dessa vez com a resolução reduzida, ou seja, com as matrizes podadas pela metade. O resultado da operação é apresentado em (82)

$$\overbrace{\begin{bmatrix} 0,71 \\ 0,71 \end{bmatrix}^{T}}^{Consulta} \bullet \overbrace{\begin{bmatrix} 0,93 & 0,71 & 1,00 & 0,71 \\ 0,36 & 0,71 & 0,00 & -0,71 \end{bmatrix}}^{Documentos} = \overbrace{[0,91 \quad 1,00 \quad 0,71 \quad 0,00]}^{Resultados} \qquad (82)$$

Note que devido a redução de resolução passamos a ter um falso positivo, d3, que de fato não possui termo A, mas acaba por ser encontrado durante o procedimento de recuperação devido a sobreposição de conceitos que ocorre quando reduzimos a resolução.

Por fim, vamos para o último grau de resolução possível, como mostrado em (83).

$$\overbrace{[1]^{T}}^{Consulta} \bullet \overbrace{[1 \quad 1 \quad 1 \quad 1]}^{Documentos} = \overbrace{[1 \quad 1 \quad 1 \quad 1]}^{Resultados} \qquad (83)$$

Neste último caso, os vetores se degeneram no caso trivial, onde todos

os documentos têm similaridades iguais a 1.

Podemos compilar todos estes resultados em outra matriz, $M_{res}$, que nos dá uma visão geral da comparação em todos os níveis de resolução. $M_{res}$ é apresentada em (84).

$$M_{res} = \begin{bmatrix} 0,66 & 0,93 & 0,00 & 0,00 \\ 0,91 & 1,00 & 0,71 & 0,00 \\ 1,00 & 1,00 & 1,00 & 1,00 \end{bmatrix} \begin{matrix} \text{Resolução 3} \\ \text{Resolução 2} \\ \text{Resolução 1} \end{matrix} \quad (84)$$

Cada linha de $M_{res}$ é o resultado de uma comparação feita em um nível de resolução. Esta matriz possibilita que analisemos a evolução das comparações segundo o nível de resolução. Possibilita que, uma vez conhecido o resultado correto, comparar o número de falso-positivos que aparecem com a redução de resolução.

### 6.1.6 Analise do Modelo

O Modelo Wavelet (MW) é resultado da experiência com os demais modelos que foram apresentados. O MW supera várias das limitações dos modelos anteriores. Primeiro a função de codificação de termos não é meramente um inteiro aleatório, mas sim procura relacionar a semântica dos termos com os índices atribuídos. Como mostraremos nos experimentos esta função melhora sensivelmente a qualidade da recuperação dos documentos.

Segundo, a função de codificação de documentos usada possui um tamanho fixo para todo e qualquer documento analisado, isso simplifica em muito o funcionamento do sistema.

Terceiro, a utilização da transformada wavelet no lugar da transformada de Fourier, abre novas possibilidades de aplicação para o sistema. Por causa da propriedade de multiresolução, podemos balancear a precisão e o tamanho do índice analisado. Isso

possibilita a criação de novas aplicações como, por exemplo, o roteamento de consultas na em redes ponto a ponto ou a consulta em diversos níveis de resolução (detalhe).

# 7 Avaliação do Modelo Wavelet

Este capítulo descreve os experimentos realizados para avaliar o modelo proposto. Como citado na seção anterior, o Modelo Wavelet foi o único que foi efetivamente implementado e avaliado e por isso, este capítulo trata apenas desse modelo. Esta avaliação visa mostrar que, como esperado, a propriedade de multiresolução se manifesta no processo de busca e recuperação da informação. Além disso, com os experimentos realizados será possível traçar exatamente o ponto onde a perda de resolução do índice começa a causar sensíveis alterações na qualidade da busca.

## *7.1 Objetivos da avaliação*

A avaliação do Modelo Wavelet tem dois objetivos principais: o primeiro, e mais essencial, é mostrar como, conforme a teoria, a propriedade de multiresolução aparece e influencia na qualidade da operação de busca; o segundo, é mostrar o papel importante da função de codificação de termos para o Modelo Wavelet.

## *7.2 Metodologia*

A metodologia empregada para avaliar o Modelo Wavelet é a clássica medida de cobertura versus precisão (*recall versus precision*). Este metodologia foi escolhida em detrimento de outras, por mostrar o comportamento global do sistema em todos os níveis de cobertura possíveis.

### 7.2.1 Descrição da metodologia

A cobertura e a precisão são dois conceitos relacionados que, quando usados em conjunto, fornecem um gráfico do comportamento médio do sistema de busca e recuperação de informação. Vejamos a definição formal desses dois conceitos, segundo (YATES B.R. e BERTHIER R.N., 1999).

Seja R o conjunto de documentos sabidamente relevantes para uma consulta q; $D_r$ o conjunto de documentos recuperados pelo sistema de BRI; e $D_c$ a interseção entre R e $D_r$, ou seja, o conjunto de documentos recuperados pelo sistema de BRI que se encontra entre os documentos sabidamente relevantes para coleção. Dessa forma, podemos definir cobertura como mostrado em (85).

$$\text{cobertura} = \frac{|D_c|}{|R|} \tag{85}$$

Onde,

|.| é o operador de cardinalidade do conjunto.

$D_r$ é o conjunto de documentos recuperados pelo sistema

R é o conjunto de documentos sabidamente relevantes para consulta q

Em palavras, cobertura é a fração de documentos recuperados que são relevantes. Idealmente, um sistema de BRI sempre deve alcançar um nível de cobertura 1 (100%), ou seja, ele deve encontrar todos os documentos relevantes.

De maneira similar, podemos definir precisão, como mostrado em (86)

$$\text{precisão}=\frac{|D_c|}{|D_r|} \tag{86}$$

Onde,

|.| é o operador de cardinalidade

$D_c$ é o conjunto de documentos relevantes recuperados pelo sistema de BRI

R é o conjunto de documentos sabidamente relevantes para consulta q.

Em palavras, precisão é a fração dos documentos relevantes para consulta que foram recuperados pelo sistema de BRI. Idealmente, a precisão deve ser sempre 1 (100%), para todos os níveis de cobertura considerados.

A relação entre cobertura e precisão fica mais explicita quando as organizamos na forma do diagrama de Venn apresentado na Figura 36. Nessa figura podemos ver que existe uma interseção entre o conjunto R, o conjunto dos documentos sabidamente relevantes para consulta q, e o conjunto $D_r$, o conjunto de documentos recuperados pelo sistema de BRI. Idealmente esta interseção deveria ser total e completa, ou seja, os conjuntos R e $D_r$ deveriam ser iguais. Nesse caso, teríamos uma cobertura e precisão máximos.


**Figura 36 Diagrama de Venn que demonstra a relação entre cobertura e precisão.**

A avaliação de um sistema de BRI é feita da seguinte forma: para cada documento, na ordem que ele foi recuperado, são calculadas a cobertura e a precisão, como mostrado em (85) e (86), respectivamente. Note que como $D_c$ é um valor fixo para cada documento recuperado, a precisão varia seu valor e a cobertura permanece inalterada até que um documento relevante seja recuperado. Portanto, para uma mesma cobertura, temos diferentes precisões. Para efeitos do experimento, usamos como indicador do grau de precisão, o valor máximo atingido para uma determinada cobertura. O resultado final desse procedimento é um gráfico de cobertura versus precisão como mostrado Exemplo 14.

Cada consulta tem seu próprio gráfico de cobertura versus precisão, portanto para que obtenhamos uma estatística válida é preciso sintetizar um conjunto razoavelmente grande de gráficos em um único gráfico que reflita o comportamento médio do sistema. Este gráfico é obtido a partir da padronização de 11 níveis (ou graus) de cobertura que refletem a porcentagem de documentos recuperados e relevantes. Estes graus de cobertura são numerados de 0 a 10. Caso um determinado grau de cobertura não seja atingido, os resultados são interpolados de maneira a repetir o último resultado obtido.

Exemplo 14. [4]

Considere um sistema de BRI cujo conjunto de documentos relevantes para uma consulta q seja igual a R=($d_3$, $d_5$, $d_9$, $d_{25}$, $d_{39}$, $d_{44}$, $d_{56}$, $d_{71}$, $d_{89}$, $d_{123}$). Suponha que a resposta do sistema a esta mesma consulta seja como a mostrada pela Tabela 4, onde a última coluna indica quais dos documentos recuperados se encontra no conjunto do documentos relevantes.

[4] Exemplo retirado de (Yates e Neto), páginas 76 e 77

**Tabela 4 Documentos recuperados pelo sistema de BRI hipotético. A última coluna indica os documentos que foram recuperados e são relevantes.**

| Classificação | Documento | Relevância |
|---|---|---|
| 1 | $d_{123}$ | X |
| 2 | $d_{84}$ | |
| 3 | $d_{56}$ | X |
| 4 | $d_{6}$ | |
| 5 | $d_{8}$ | |
| 6 | $d_{9}$ | X |
| 7 | $d_{511}$ | |
| 8 | $d_{129}$ | |
| 9 | $d_{187}$ | |
| 10 | $d_{25}$ | X |
| 11 | $d_{38}$ | |
| 12 | $d_{48}$ | |
| 13 | $d_{250}$ | |
| 14 | $d_{113}$ | |
| 15 | $d_{3}$ | X |

Vamos examinar a classificação dos documentos iniciando a partir do primeiro: o primeiro documento recuperado é relevante, portanto temos uma precisão de 100% (um documento relevante em um conjunto de um documento recuperado). A cobertura também é de 10% (um documento recuperado em um conjunto de 10 documentos relevantes); o segundo documento recuperado não é relevante, portanto temos uma redução de precisão (50%, ou seja, um documento relevante em dois recuperados), mas permanecemos no mesmo grau de cobertura. Como estamos interessados apenas no valor máximo de cobertura em cada grau de

relevância, ignoramos este resultado; o terceiro documento recuperado, que também é relevante apresenta uma precisão de 66% (dois documentos relevantes em três recuperados), e uma cobertura de 20% (dois documentos relevantes recuperados em um conjunto de dez documentos relevantes). Se continuarmos este procedimento, podemos criar um gráfico de cobertura versus precisão que mostra o comportamento do sistema de BRI para a consulta em particular que foi realizada.

## 7.3 Coleções de teste

As coleções de teste são formadas por três conjuntos: um conjunto de documentos, um conjunto de consultas e um conjunto de respostas que associa os documentos às consultas. O conjunto de respostas é criado por especialistas humanos que avaliam as consultas e associam os documentos relevantes a elas. Portanto, ao usarmos uma coleção de teste já sabemos de antemão quais são os documentos relevantes para uma consulta. Dessa forma, podemos calcular facilmente a cobertura e a precisão como definidas anteriormente.

Para os fins dessa dissertação, usamos a coleção conhecida como *Cystic Fibrosis Collection* (CF) (W.M.SHAW, J.B.WOOD et al., 1993a), que é composta por 1239 documentos e 100 consultas. Originalmente a CF é composta pelos seguintes campos:

- Número da Consulta;
- Autor;
- Título;
- Fonte;
- Assuntos majoritários;
- Assuntos minoritários;
- Resumo;
- Referência;
- Citações.

No entanto, durante nossos experimentos, somente os campos assuntos majoritários, assuntos minoritários e resumo foram concatenados e empregados. Os demais campos foram descartados.

## 7.4 O ambiente

Os experimentos foram realizados em um microcomputador com processador AMD 2400 de 2 GHz, com 1 GB de memória RAM. No momento do ensaio, apenas 250 MB estavam ocupados, restando assim 750 MB disponíveis para o experimento.

O sistema operacional utilizado foi o Microsoft Windows XP Service Pack 2.0, a linguagem de programação empregada foi JAVA versão 1.5.

## 7.5 Experimento 1

O experimento 1 consistiu na prévia indexação da coleção CF e na aplicação da Função de Codificação de Documento como descrita na seção 5.1. Uma vez de posse do índice, iniciamos a segunda etapa do experimento: a busca por documentos baseado em consultas pré-estabelecidas cujos resultados esperados já são conhecidos. Dessa forma, podemos calcular os pontos de cobertura versus precisão, originando um gráfico com 11 níveis de cobertura.

Por se tratar de um índice com múltiplos níveis de detalhe (resoluções), a segunda etapa do experimento se repete diversas vezes, cada uma em um nível de resolução diferente. Iniciamos o experimento com a resolução máxima e vamos reduzindo um grau a cada etapa. O número total de etapas depende do tamanho original do índice (veja equação (77)). Ao final do experimento, obtemos N gráficos de cobertura versus precisão, cada um referente a um grau de resolução diferente.

### 7.5.1 Resultados obtidos

Como podemos ver no gráfico da Figura 37, a propriedade de multiresolução se materializa no resultado de maneira a reduzir a precisão conforme diminuímos o grau de resolução. Nas resoluções mais altas, temos um resultado compatível com o modelo vetorial. Nas demais resoluções, obtemos uma redução do tamanho físico do índice ao custo de perda da resolução. Para coleção CF em particular, foram realizados 100 consultas, cada uma em um dos 15 níveis de resolução, totalizando 1500 consultas.

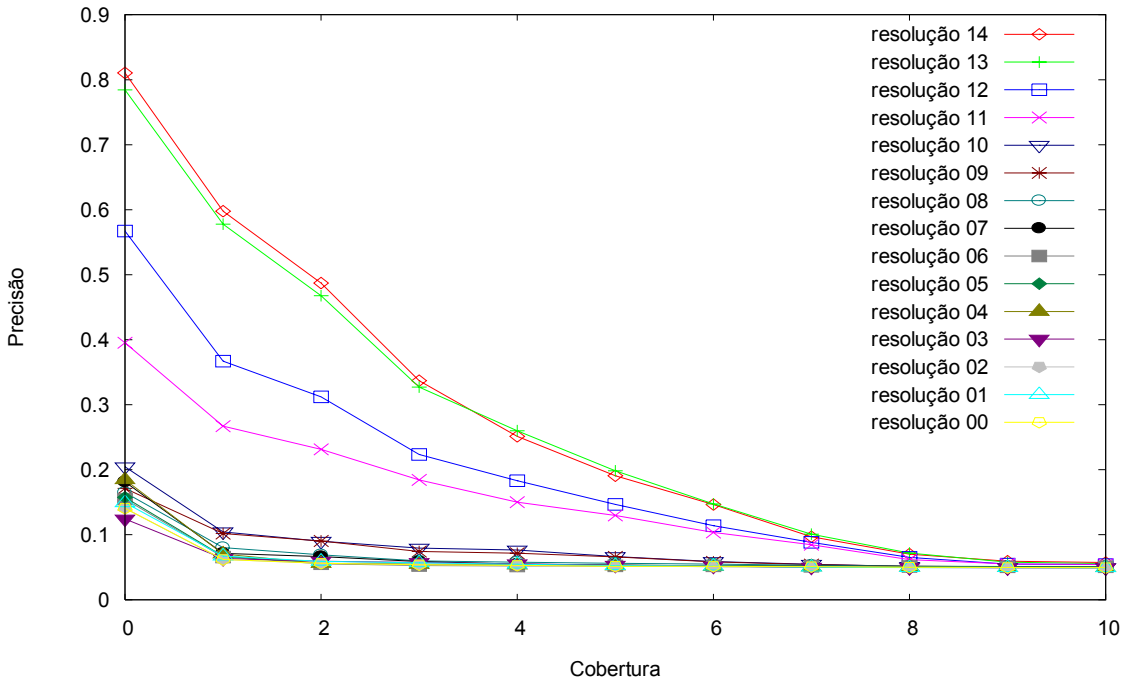



**Figura 37 Gráfico de cobertura versus precisão em diferentes níveis de resolução.**

A Tabela 5 apresenta os dados, com duas casas decimais de precisão, que foram usados para marcar o gráfico acima. Nela podemos observar como, numericamente, a partir do nível de resolução 10 as precisões sofrem uma brusca queda e a diferença entre duas precisões para o mesmo grau de cobertura em níveis de precisão diferentes deixam de ser tão significativas.

**Tabela 5 Precisões médias de 100 consultas realizadas em cada nível de resolução totalizando 1500 consultas.**

| Cobert. | Nível 14 | Nível 13 | Nível 12 | Nível 11 | Nível 10 | Nível 09 | Nível 08 | Nível 07 | Nível 06 | Nível 05 | Nível 04 | Nível 03 | Nível 02 | Nível 01 | Nível 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0,81 | 0,78 | 0,57 | 0,40 | 0,20 | 0,17 | 0,16 | 0,18 | 0,15 | 0,16 | 0,18 | 0,12 | 0,14 | 0,15 | 0,14 |
| 1 | 0,60 | 0,58 | 0,37 | 0,27 | 0,10 | 0,10 | 0,08 | 0,07 | 0,07 | 0,07 | 0,07 | 0,06 | 0,06 | 0,07 | 0,06 |
| 2 | 0,49 | 0,47 | 0,31 | 0,23 | 0,09 | 0,09 | 0,07 | 0,07 | 0,06 | 0,06 | 0,06 | 0,06 | 0,06 | 0,06 | 0,06 |
| 3 | 0,34 | 0,33 | 0,22 | 0,18 | 0,08 | 0,07 | 0,06 | 0,06 | 0,05 | 0,05 | 0,05 | 0,06 | 0,05 | 0,06 | 0,05 |
| 4 | 0,25 | 0,26 | 0,18 | 0,15 | 0,08 | 0,07 | 0,06 | 0,05 | 0,05 | 0,05 | 0,05 | 0,06 | 0,05 | 0,05 | 0,05 |
| 5 | 0,19 | 0,20 | 0,15 | 0,13 | 0,07 | 0,07 | 0,06 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 |
| 6 | 0,15 | 0,15 | 0,11 | 0,10 | 0,06 | 0,06 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 |
| 7 | 0,10 | 0,10 | 0,09 | 0,08 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 |
| 8 | 0,07 | 0,07 | 0,06 | 0,06 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 |
| 9 | 0,06 | 0,06 | 0,05 | 0,06 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 |
| 10 | 0,06 | 0,06 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 |

Podemos ainda observar claramente pela Tabela 5, o comportamento não monótono da precisão em relação à resolução. Por exemplo, para o grau de cobertura 0 (zero), vemos uma ligeira melhoria na precisão entre os níveis de resolução 5 e 4. Este

efeito é causado pela não linearidade do produto escalar que é utilizado como métrica de comparação entre os documentos.

### 7.5.2 Análise dos resultados

Como sabemos, o 'efeito wavelet' promove uma perda de detalhamento no índice usado na busca de documentos. Esta perda acaba por se refletir na precisão do sistema de BRI. A Tabela 6 mostra como esta perda é sentida pelo sistema. Nesta tabela medimos, em porcentagem, o quanto a precisão para um determinado nível de cobertura caiu em relação ao nível de resolução anterior. Trivialmente, o nível de resolução 14 é comparado consigo mesmo e, portanto não há perda de precisão. Os demais níveis sempre são comparados com o anterior e a perda de precisão é apresentada na mesma coluna. Desta forma, para cobertura 0, o nível de resolução 13 perdeu 3,70% de precisão em relação ao nível de resolução anterior (nível 14). A última linha da Tabela 6 apresenta a perda de precisão média de cada nível. Nesta última linha temos uma síntese do comportamento da perda da precisão em todos os níveis de cobertura.

**Tabela 6 Perdas relativas causas pela redução de resolução do indice do Modelo Wavelet.**

| Cobertura | Nível 14 | Nível 13 | Nível 12 | Nível 11 | Nível 10 | Nível 09 | Nível 08 | Nível 07 | Nível 06 | Nível 05 | Nível 04 | Nível 03 | Nível 02 | Nível 01 | Nível 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0,00% | 3,70% | 26,92% | 29,82% | 50,00% | 15,00% | 5,88% | -12,50% | 16,67% | -6,67% | -12,50% | 33,33% | -16,67% | -7,14% | 6,67% |
| 1 | 0,00% | 3,33% | 36,21% | 27,03% | 62,96% | 0,00% | 20,00% | 12,50% | 0,00% | 0,00% | 0,00% | 14,29% | 0,00% | -16,67% | 14,29% |
| 2 | 0,00% | 4,08% | 34,04% | 25,81% | 60,87% | 0,00% | 22,22% | 0,00% | 14,29% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% |
| 3 | 0,00% | 2,94% | 33,33% | 18,18% | 55,56% | 12,50% | 14,29% | 0,00% | 16,67% | 0,00% | 0,00% | -20,00% | 16,67% | -20,00% | 16,67% |
| 4 | 0,00% | -4,00% | 30,77% | 16,67% | 46,67% | 12,50% | 14,29% | 16,67% | 0,00% | 0,00% | 0,00% | -20,00% | 16,67% | 0,00% | 0,00% |
| 5 | 0,00% | -5,26% | 25,00% | 13,33% | 46,15% | 0,00% | 14,29% | 16,67% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% |
| 6 | 0,00% | 0,00% | 26,67% | 9,09% | 40,00% | 0,00% | 16,67% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% |
| 7 | 0,00% | 0,00% | 10,00% | 11,11% | 37,50% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% |
| 8 | 0,00% | 0,00% | 14,29% | 0,00% | 16,67% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% |
| 9 | 0,00% | 0,00% | 16,67% | -20,00% | 16,67% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% |
| 10 | 0,00% | 0,00% | 16,67% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% |
| Média | 0,00% | 0,44% | 24,60% | 11,91% | 39,37% | 3,64% | 9,78% | 3,03% | 4,33% | -0,61% | -1,14% | 0,69% | 1,52% | -3,98% | 3,42% |

O nível de resolução 10 é onde ocorre a maior redução relativa de precisão, nele, para uma cobertura 0, temos 50% de perda de perda de precisão. Que é o pior resultado obtido por todo o sistema. A média de perda de precisão também é maior no nível de resolução 10. Portanto, para a coleção CF em particular, o nível 10 de resolução é o nível que tem maior impacto sobre a precisão.

A partir do nível de resolução 5 o número de zeros da tabela aumenta acentuadamente, portanto a partir deste ponto há poucas mudanças no comportamento médio do sistema. Ou seja, a perda de resolução passa a afetar pouco o a precisão das consultas. De fato, como podemos observar no gráfico da Figura 37 a partir do nível de resolução 10 há pouca variação no comportamento médio do sistema.

## *7.6 Experimento 2*

Este experimento visa comparar os resultados obtidos com o Modelo Wavelet com outros sistemas comerciais. Utilizamos como parâmetro de comparação a biblioteca Lucene (OTIS GOSPODNETIC e ERIK HATCHER, 2007), amplamente usada pela comunidade dedicada à busca e recuperação de informação. Da mesma maneira como no experimento anterior utilizamos um gráfico de cobertura versus precisão para comparar o Modelo Wavelet com os resultados obtidos pela biblioteca Lucene. Para esse experimento foram realizadas 100 consultas nos cinco níveis mais altos de resolução e mais 100 consultas realizadas com a biblioteca Lucene, totalizando 600 consultas.

### 7.6.1 Resultados obtidos

Como podemos constatar no gráfico da Figura 38, o Modelo Wavelet com resoluções mais altas foi superior aos resultados obtidos pela biblioteca Lucene em todos os níveis de cobertura. Este resultado, de certa forma é inesperado, uma vez que no mais alto nível de resolução o Modelo Wavelet é perfeitamente equivalente ao modelo vetorial (VSM).

A partir do nível de resolução 12, percebemos que a biblioteca Lucene passa a ter resultados melhoras que o Modelo Wavelet. Mas em contrapartida estamos lidando com um índice cujo número de ‘termos’ é quatro vezes menor.

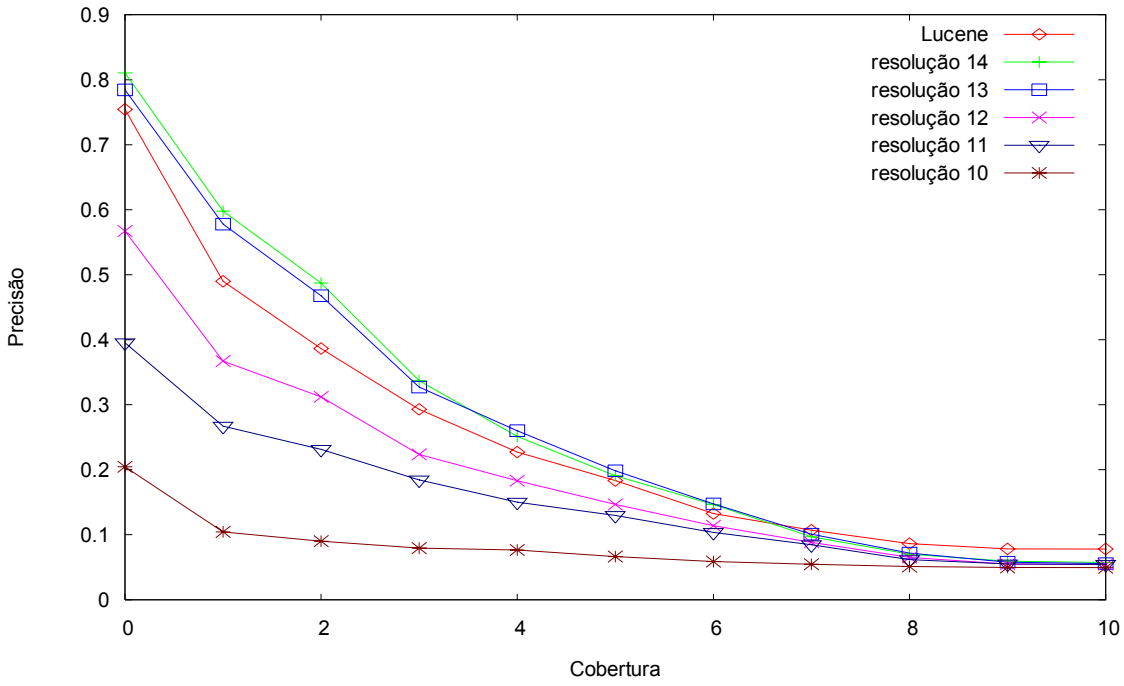



**Figura 38 Gráfico de Cobertura versus Precisão comparando Lucene e Modelo Wavelet**

A Tabela 7 mostra os resultados numéricos que são apresentados no gráfico acima. Podemos

**Tabela 7 Comparação entre o Modelo Wavelet e a bibliteca lucene.**

| Cobertura | Lucene | Nível 14 | Nível 13 | Nível 12 | Nível 11 | Nível 10 |
|---|---|---|---|---|---|---|
| 0 | 0.75 | 0.81 | 0.78 | 0.57 | 0.40 | 0.20 |
| 1 | 0.49 | 0.60 | 0.58 | 0.37 | 0.27 | 0.10 |
| 2 | 0.39 | 0.49 | 0.47 | 0.31 | 0.23 | 0.09 |
| 3 | 0.29 | 0.34 | 0.33 | 0.22 | 0.18 | 0.08 |
| 4 | 0.23 | 0.25 | 0.26 | 0.18 | 0.15 | 0.08 |
| 5 | 0.18 | 0.19 | 0.20 | 0.15 | 0.13 | 0.07 |
| 6 | 0.13 | 0.15 | 0.15 | 0.11 | 0.10 | 0.06 |
| 7 | 0.11 | 0.10 | 0.10 | 0.09 | 0.08 | 0.05 |
| 8 | 0.09 | 0.07 | 0.07 | 0.06 | 0.06 | 0.05 |
| 9 | 0.08 | 0.06 | 0.06 | 0.05 | 0.06 | 0.05 |
| 10 | 0.08 | 0.06 | 0.06 | 0.05 | 0.05 | 0.05 |

## *7.7 Experimento 3*

Este experimento visa destacar o importante papel da Função de Codificação de Documentos para o Modelo Wavelet. Ele consiste na execução das mesmas consultas realizadas no experimento 1, mas desta vez usando a mesma Função de Codificação de Documentos empregada no Modelo Trivial, ou seja, associamos um número inteiro a cada novo termo da coleção à medida que eles aparecem. Dessa forma, perdemos o potencial do que convencionamos chamar de efeito Wavelet, que é a propriedade da

transformada wavelet de sobrepor conceitos relacionados. Para esse experimento foram realizadas 100 consultas em cada um dos 15 níveis de resolução, totalizando 1500 consultas.

### 7.7.1 Resultados obtidos

O gráfico da Figura 39 apresenta os resultados obtidos nos diferentes níveis de resolução para o experimento realizado com a FCD trivial. Podemos comparar este gráfico com o apresentado pela Figura 37 e notar uma sensível perda na precisão das consultas em todos os níveis de cobertura. Este resultado ratifica a importância a FCD para o modelo e mostra também como o comportamento do sistema pode variar conforme mudamos a configuração da FCD.

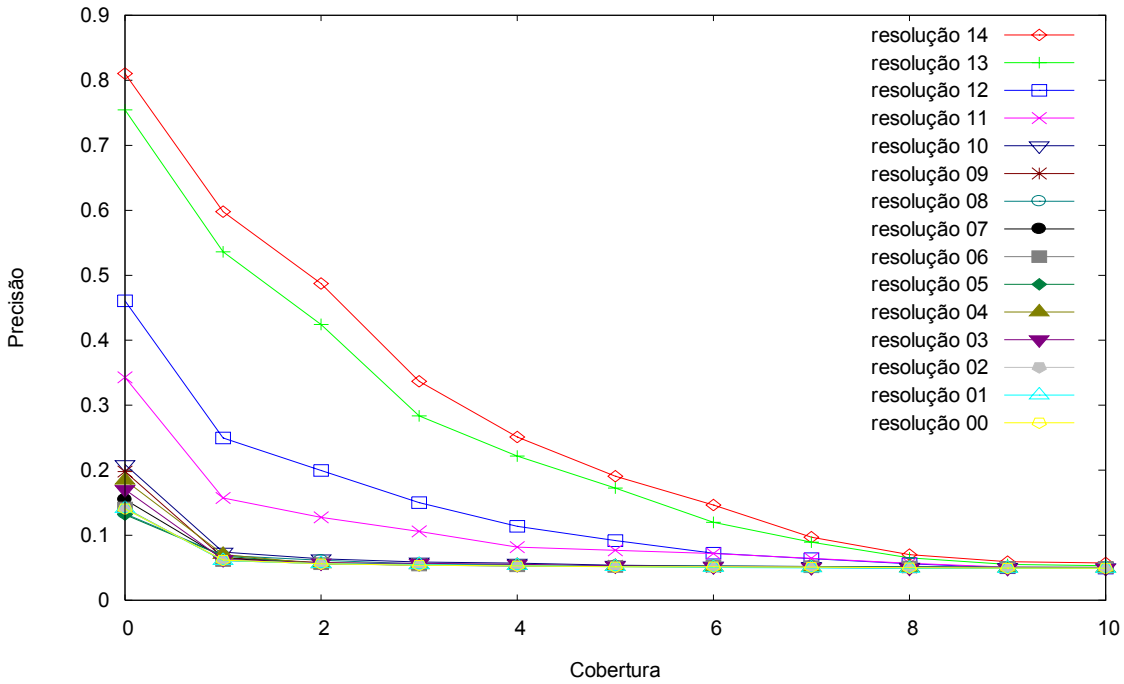



**Figura 39 Experimento realizado com FCD trivial**

Como esperávamos, a perda de resolução é causada pela falta de uma FCD que se aproveite da propriedade de multiresolução da transformada wavelet. Esta perda é mais sensivelmente percebida no gráfico no nível de resolução 12 (linha com quadrado), onde, para o primeiro nível de cobertura, temos uma precisão abaixo de 50%, enquanto no gráfico da Figura 37 nós atingimos quase 60%. Esta diferença fica mais explicita quando comparamos numericamente a Tabela 5 e a Tabela 8. Ambas as tabelas apresentam as precisões em 11 níveis de cobertura em diferentes níveis de resolução, mas a última usa a FCD apresentado no modelo trivial.

**Tabela 8 Cobertura versus Precisão com FCD trivial**

| Cobertura | Nível 14 | Nível 13 | Nível 12 | Nível 11 | Nível 10 | Nível 9 | Nível 8 | Nível 7 | Nível 6 | Nível 5 | Nível 4 | Nível 3 | Nível 2 | Nível 1 | Nível 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0,81 | 0,75 | 0,46 | 0,34 | 0,21 | 0,20 | 0,13 | 0,15 | 0,14 | 0,13 | 0,18 | 0,17 | 0,14 | 0,14 | 0,14 |
| 1 | 0,60 | 0,54 | 0,25 | 0,16 | 0,07 | 0,07 | 0,07 | 0,06 | 0,06 | 0,07 | 0,07 | 0,06 | 0,06 | 0,06 | 0,06 |
| 2 | 0,49 | 0,42 | 0,20 | 0,13 | 0,06 | 0,06 | 0,06 | 0,06 | 0,06 | 0,06 | 0,06 | 0,06 | 0,06 | 0,06 | 0,06 |
| 3 | 0,34 | 0,28 | 0,15 | 0,11 | 0,06 | 0,06 | 0,06 | 0,05 | 0,05 | 0,05 | 0,05 | 0,06 | 0,05 | 0,05 | 0,05 |
| 4 | 0,25 | 0,22 | 0,11 | 0,08 | 0,06 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,06 | 0,05 | 0,05 | 0,05 |
| 5 | 0,19 | 0,17 | 0,09 | 0,08 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 |
| 6 | 0,15 | 0,12 | 0,07 | 0,07 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 |
| 7 | 0,10 | 0,09 | 0,06 | 0,06 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 |
| 8 | 0,07 | 0,07 | 0,06 | 0,06 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 |
| 9 | 0,06 | 0,06 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 |
| 10 | 0,06 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 |

Excetuando-se o nível de resolução 14, que é comum a ambas as tabelas, podemos perceber nitidamente a perda de precisão, principalmente nos primeiros níveis de cobertura.

A Tabela 9 apresenta a perda de resolução relativa de precisão que tivemos ao usarmos a FCD trivial. A perda mais sensível foi no nível de resolução 11, onde perdemos 26,85% de precisão, mas em, praticamente, todos os níveis de resolução obtivemos alguma perda. Novamente, uma análise mais cuidadosa dos números, mostra que para algumas resoluções como a de nível 3, nós tivemos uma leve melhoria relativa da precisão média do sistema. Nós atribuímos este comportamento a natureza não linear da função de classificação empregada, somada a degeneração do vetor original que representa tanto o documento quanto a consulta. Nos níveis de resolução mais baixos, o grau de sobreposição de conceitos devido ao 'efeito wavelet' é tão grande que qualquer análise se torna inócua.

**Tabela 9 Perda relativa de precisão obtida com a utilização da FCD Trivial**

| Cobertura | Nível 14 | Nível 13 | Nível 12 | Nível 11 | Nível 10 | Nível 09 | Nível 08 | Nível 07 | Nível 06 | Nível 05 | Nível 04 | Nível 03 | Nível 02 | Nível 01 | Nível 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0,00% | 3,27% | 19,18% | 14,28% | -4,36% | -16,38% | 16,83% | 13,95% | 5,26% | 17,76% | -2,54% | -41,98% | -0,52% | 5,90% | -0,53% |
| 1 | 0,00% | 7,55% | 32,52% | 41,73% | 26,01% | 34,61% | 13,88% | 7,50% | 12,90% | 2,95% | -0,49% | -4,39% | -2,99% | 12,85% | -3,06% |
| 2 | 0,00% | 9,70% | 35,62% | 44,57% | 29,27% | 35,60% | 12,16% | 20,03% | 6,31% | 6,22% | 7,17% | 4,42% | 6,58% | 6,06% | 6,57% |
| 3 | 0,00% | 14,05% | 31,72% | 41,19% | 26,60% | 21,17% | 7,06% | 9,14% | -8,19% | -7,80% | -8,63% | 5,37% | -7,75% | 8,70% | -7,65% |
| 4 | 0,00% | 14,69% | 36,84% | 45,52% | 28,83% | 23,09% | 9,66% | -7,09% | -5,66% | -5,16% | -6,82% | 7,73% | -5,56% | -6,74% | -5,48% |
| 5 | 0,00% | 13,67% | 38,90% | 40,96% | 23,21% | 24,58% | 11,75% | -5,04% | -3,89% | -4,48% | -3,65% | -5,88% | -3,13% | -3,13% | -2,85% |
| 6 | 0,00% | 20,16% | 34,27% | 28,22% | 12,49% | 14,04% | -5,06% | -3,52% | -4,65% | -4,25% | -1,66% | -3,15% | -2,68% | -0,61% | -2,77% |
| 7 | 0,00% | 10,68% | 29,08% | 19,30% | -3,57% | -2,52% | -3,86% | -2,02% | -3,07% | -3,41% | -0,73% | -0,69% | -1,71% | 0,84% | -1,58% |
| 8 | 0,00% | 6,88% | 6,45% | 5,13% | -1,75% | -1,65% | -2,09% | -1,43% | -2,03% | -2,05% | 0,46% | 1,11% | -0,01% | 2,06% | 0,01% |
| 9 | 0,00% | 8,31% | 0,12% | 15,16% | -0,54% | -0,62% | -1,21% | -0,88% | -1,20% | -0,84% | 1,14% | 1,57% | 1,03% | 1,99% | 1,00% |
| 10 | 0,00% | 11,46% | 1,27% | -0,76% | 0,01% | -0,24% | -0,68% | -0,73% | -1,05% | -0,65% | 1,33% | 1,80% | 1,38% | 2,18% | 1,42% |
| **Média** | **0,00%** | **10,95%** | **24,18%** | **26,85%** | **12,38%** | **11,97%** | **5,31%** | **2,72%** | **-0,48%** | **-0,15%** | **-1,31%** | **-3,10%** | **-1,40%** | **2,74%** | **-1,36%** |

# 8 Conclusão

Esta dissertação teve como objetivo propor uma nova abordagem para sistemas de Busca e Recuperação de Informação (BRI) de documentos texto. A definição formal de um sistema de BRI prevê a existência de uma estrutura matemática que fornece suporte teórico às operações realizadas. Essa estrutura matemática desempenha um papel fundamental em qualquer sistema de BRI, tanto que podemos usá-la para caracterizar o próprio sistema. As estruturas matemáticas mais usadas em sistemas de BRI são: a teoria dos conjuntos, as probabilidades e a álgebra linear. Esta última origina o, assim chamado, modelo vetorial (VSM – *Vector Space Model*), onde tanto os documentos quanto as consultas são considerados vetores no espaço dos termos. Neste âmbito, nossa proposta foi utilizar a estrutura matemática empregada no campo de Processamento de Sinais para operar documentos textos. Assim como a álgebra linear, a estrutura matemática de processamento de sinais é fortemente consolidada e amplamente aplicada em outras áreas da engenharia. A inovação trazida nessa dissertação é a aplicação dessa estrutura já consagrada em um campo onde ela ainda é pouco explorada: A Busca e Recuperação de Informação (BRI) de documentos do tipo texto.

Nas seções iniciais apresentamos os três principais modelos de BRI empregados na atualidade, o modelo Booleano, sustentado pela teoria dos conjuntos; o modelo probabilístico, sustentado pela teoria das probabilidades e o teorema de Bayes; e o Modelo Vetorial, sustentado pela Álgebra Linear.

Nas seções seguintes, prosseguimos com os fundamentos matemáticos que sustentam o campo de processamento de sinais. Para terminar a parte introdutória dessa dissertação, explicamos sucintamente os fundamentos de processamento de sinais empregados ao longo desse trabalho, explicamos como as transformadas funcionam e, em especial, detalhamos o funcionamento da transformada de Fourier, que é talvez a ferramenta mais empregada no campo de processamento de sinais. Esta transformada serviu de base para a atingirmos nosso verdadeiro objetivo, que foi a explicação de como se comporta a transformada wavelet. Em particular, descrevemos como age a multiresolução, que é um dos pontos chaves de nossa proposta.

Uma vez entendidos os fundamentos de processamento de sinais, explicamos sua relação com a Busca e Recuperação de Informação (BRI). Mostramos, em particular,

que as técnicas propostas nessa dissertação não são de todo desconhecidas no campo de BRI, já sendo aplicadas com sucesso em BRI multimídia, por exemplo. Nesta mesma seção aproveitamos para citar trabalhos relacionados que contribuíram para a formação das idéias expostas nessa dissertação, com destaque particular para os trabalhos de (PARK, RAMAMOHANARAO et al., 2005) e (FLESCA, MANCO et al., 2005). Por fim, definimos o que chamamos de *MSBRI*, ou modelo de Sinais para BRI, que se trata de um meta-modelo de BRI baseado na estrutura matemática de processamento de sinais anteriormente exposta.

O *MSBRI* define duas funções básicas, chamadas de Função de Codificação de Termo (FCT) e Função de Codificação de Documento (FCD). A primeira tem por objetivo associar um identificador único a cada termo de uma coleção de documentos, já a segunda, usa a primeira para representar o documento como um sinal que posteriormente é analisado com as técnicas de processamento de sinais adequadas. A FCT e FCD desempenham um papel muito importante em nosso modelo, pois, a cada especificação diferente dessas funções, o sistema passa a se comportar como se fosse outro. Portanto, nosso modelo é na verdade um meta-modelo, uma estrutura que possibilita a instanciação de outros modelos que são independentes uns dos outros. Assim, começamos a explorar diversos possíveis modelos que poderiam ser derivados da estrutura básica do *MSBRI*. O primeiro modelo foi chamado de Modelo Trivial (MT) por causa de sua simplicidade. Em seguida apresentamos o Modelo de Alta Correlação (MAC), onde estabelecemos uma nova FCT que associa, a cada termo, um índice inteiro dependente de algumas características estatísticas desse termo. Em seguida, apresentamos o modelo de sinal estacionário, onde, pela primeira vez, explicitamente tentamos atribuir alguma semântica à transformada da representação do documento. A evolução natural desse último modelo é sua extensão a sinais não estacionários. Nesse modelo pela primeira vez propomos a utilização das propriedades de multiresolução para análise de documentos texto. Finalmente chegamos à proposta do Modelo Wavelet. Nesse modelo, propomos a utilização da propriedade de multiresolução sobre a representação de documentos textos. Propomos ainda que a FCD fosse modelada de tal forma que tirasse proveito dessa propriedade. Basicamente, usamos a FCT usada no MAC para atribuir um índice aos termos e criamos uma FCD trivial que representa o documento como um histograma de termos ordenados por sua FCT. Neste momento passamos a analisar mais cuidadosamente o que chamamos de ‘efeito wavelet’, que é a

sobreposição de conceitos que ocorre devido à perda de resolução (detalhes). O Modelo Wavelet foi o único totalmente implementado. Portanto, ele pôde ser avaliado e comparado com a biblioteca Lucene, amplamente usada pela comunidade de BRI. Os resultados da avaliação foram apresentados na seção de experimentos, tanto na forma tabular como gráfica, uma profunda análise foi feita sobre os dados coletados e uma avaliação foi materializada.

Uma vez avaliado, propusemos algumas formas de se tirar proveito do Modelo Wavelet, sugerimos que sistemas BRI distribuída podem usar o Modelo Wavelet para determinar a rota das consultas e evitar o problema conhecido como inundação (*flood*). Outra forma de se tirar proveito do Modelo Wavelet é na comparação de documentos em diversos níveis de detalhes.

Em suma, nossa proposta difere das demais porque exploramos não a transformada de Fourier, mas sim as modernas transformadas wavelet que vêm causando grande impacto e mudanças no campo de processamento de sinais. Apenas 4.2 e 4.4 empregam esta técnica, mas apenas 4.4 o faz com objetivos de BRI. No entanto, nossa proposta como será mostrado é independente e difere sob vários aspectos do trabalho de (PARK, RAMAMOHANARAO et al., 2005): primeiro, porque exploramos a propriedade de multiresolução da transforma wavelet, que é deixado de lado no trabalho de Laurence; segundo, porque propomos um novo modelo de BRI.e não nos limitando apenas a uma forma de calcular a similaridade entre consultas e documentos; terceiro, nossos estudos foram direcionados para ambientes distribuídos e mostramos como as transforma wavelet são apropriadas para BRI distribuída.

## 8.1 Trabalhos Futuros

Acreditamos que um dos principais méritos dessa dissertação é a junção de dois campos de conhecimentos correlatos antes separados: a Busca e Recuperação de Documentos Texto e o Processamento Digital de Sinais. Acreditamos, ainda que as publicações advindas desse trabalho chamarão mais atenção da comunidade cientifica para esta relação e nos ajudará no desenvolvimento de novas técnicas de Busca e Recuperação de Informação.

Esta dissertação marca apenas o início de um grande trabalho em torno do novo modelo proposto, vários aspectos dignos de análise ficaram de fora dessa dissertação por fugirem do escopo do trabalho, mas merecem ser citados como trabalhos futuros.

O primeiro trabalho relevante que pode ser destacado é a adequação do Modelo Wavelet para ambientes distribuídos. Os diferentes níveis de resoluções podem ser usados, por exemplo, para evitar o problema de inundação (*flood*) em ambientes de busca ponto-a-ponto (CHUNQIANG TANG, SANDHYA DWARKADAS et al., 2004; CHUNQIANG TANG., ZHICHEN XU et al., 2003). Ou ainda, ele pode ser usado para se fazer buscas hierárquicas em ambientes onde o tamanho do índice seja muito grande para ser processado de uma só vez (Veja Apêndice para mais detalhes).

Outros caminhos podem ser seguidos com a intenção de aperfeiçoar o modelo. Por exemplo, cabe uma comparação mais profunda de Modelo Wavelet com sistemas de LSI (DUMAIS S.T., DEERWESTER S. et al., 1990), pois de certa forma, o que convencionamos chamar de efeito wavelet, de fato, tem semelhanças com o trabalho de Dumais.

Por fim, do ponto de vista teórico, podemos explorar infinitas cominações de Funções de Codificação de Termos (FCT's) e Funções de Codificação de Documento (FCD's). No entanto, algumas combinações se mostram mais promissoras que outras. Sugerimos como trabalho futuro a investigação de outras funções wavelet mãe para analise multiresolucional, como por exemplo, Daubechies de ordens mais elevadas. Para tal, uma integração com algum software de calculo numérico pode contribuir bastante.

# 9 Referências Bibliográficas


BENG CHIN OOI, KIAN-LEE TAN, TAT SENG CHUA, et al, 1998, "Fast image retrieval using color-spatial information", *The VLDB Journal - The International Journal on Very Large Data Bases*, pp. 115-128

BUCKELEY, C., SINGHAL, A., MITRA, M., et al, 1995, "New Retrieval Approches using SMART"

CARLO MEGHINI, FABRIZIO SEBASTIANI, UMBERTO STRACCIA, 2001, "A model of multimedia information retrieval", *Journal of the ACM (JACM)*, pp. 909-970

CHUNQIANG TANG, SANDHYA DWARKADAS, ZHICHEN XU, 2004, "**On Scaling Latent Semantic Indexing for Large Peer-to-Peer Systems**".www.acm.org

CHUNQIANG TANG., ZHICHEN XU, SANDHYA DWARKADAS, 2003, "**Peer-to-Peer Information Retrieval Using Self-Organizing Semantic Overlay Networks**".www.acm.org

DAUBECHIES, I., 1992, "Ten Lectures on Wavelets"

DAUBECHIES, I., 1990, "The wavelet transform time-frequency localization and signal analysis", v. 36, pp. 961-1004

DAVID D.LEWIS, YIMING YANG, TONY G.ROSE, et al, 2004, "RCV1: A New Benchmark Collection for Text Categorization Research", *The Journal of Machine Learning Research*

DINIZ, 2006, *Digital signal Processing, a Computer Based Approach*. Second Edition, Mc Graw Hill

DUMAIS S.T., DEERWESTER S., FURMAS G.W., et al, 1990, "Indexing by Latent Semantic Analysis", *Journal of the American Society of Information Science*

FLESCA, S., MANCO, G., MASCIARI, E. P. L., et al, 2005, "Fast Detection of XML Structural Similarity", *Transactions on knowledge and data Engineering*, v. 17

GRAPS A., 1995, "An Introduction to wavelets", pp. 50-61.http://www.amara.com/IEEEwave/IEEEwavelet.html

LATHI, B. P., 1974, *Signal, systems and Controls*, Intext

MALLAT S.G., 1989, "A theory for multiresolution analysis: the wavelet representation", pp. 674-693

MARTIN VETTERLI,JELENA KOVACEVIC, 1995, *Wavelets and Subband Coding*. 1, Rio de Janeiro, Prencite-Hall

OTIS GOSPODNETIC,ERIK HATCHER, 2007, *Lucene in Action*, Manning

PARK, L. A. F., RAMAMOHANARAO, K., PALANISWAMI MARIMUTHU, 2005, "A novel Document Retrieval Method Using the Discrete Wavelet Transform", *Transactions on Information systems*, v. 23, pp. 267-298

SALTON GERARD, 1989, *Automatic Text Processing*, addison-wesley

SITIO DA INTERNET, 2007a, "altavista", Accessed in 01/02/2007a.

SITIO DA INTERNET, 2007b, "yahoo", Accessed in 01/02/2007b.

SITIO DA INTERNET, 2007c, "INTRODUCTION TO FOURIER TRANSFORMS FOR IMAGE PROCESSING", Accessed in 11/11/2006c.

STEPHEN DOWNIE,MICHAEL NELSON, 2000, "Evaluation of a simple and effective music information retrieval method", pp. 73-80

TAO LI, QI LI, SHENGHUO ZHU, et al, 2005, "A survey on Wavelets Application on Datamining"

W.M.SHAW, J.B.WOOD, R.E.WOOD, et al, 1993b, "THe Cistic Fibrosis database: Content and Research Opportunities"

W.M.SHAW, J.B.WOOD, R.E.WOOD, et al, 1993a, "The Cystic Fibrosis database: Content and Research Opportunities"

YATES B.R.,BERTHIER R.N., 1999, *Modern Information Retrieval*. 1, ACM Press

MILLER, N E, WONG, P. C., BREWSTER, M., FOOTE, H. TOPIC ISLAND - a wavelet-based text visualization system. 1998, Proceeding of Conference on Visualization 1998. IEEE Computer Society Press, Los Alamitos, C.A., pp. 189-196.

BOLDRINE, J. L. Algebra Linear, Ed. Harbra. 2º edição. 2005. pp 300

STRANG, G. Linear algebra and its Application. 4° edição 2005, Hardcover. ISBN: 0-15-551005-3. pp 496.

FALOUTSOS, C. RANGANATHAN, M., MANOLOPOULOS Y. “Fast Subsequence Matching in Time-Series Databases”, Proceedings. of SIGMOD, ACM, May 1994

NIBLACK, W., et al. 1993, "The QBIC Project: Querying Images by Content using Color, Texture, and Shape". Proceedings on Storage and Retrieval form Image and Video Databse, SPIE.

# Apêndice A. Ambientes distribuídos

A próxima etapa natural de evolução do Modelo Wavelet será a sua implementeação e avaliação do um ambiente distribuído de forma a aproveitar os benefícios da multiresolução em tal ambiente

Seja uma rede ponto a ponto, com os seguintes nós: A, B, C, D, E. Suponha que a rede forme um grafo conexo, como mostrado na Figura 40.

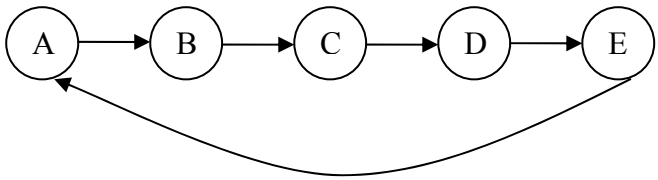



**Figura 40 Rede ponto-a-ponto formando um grafo conexo.**

A matriz de adjacência desse grafo é apresentada em (87), nela o algarismo 1 indica a presença de ligação entre dois nós e 0 indica a ausência. Trivialmente, todo nó é adjacente a si próprio.

$$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \overbrace{\begin{array}{ccccc} A & B & C & D & E \\ \hline 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{array}} \quad (87)$$

Por último, suponha que os documentos D1 a D10 estejam distribuídos entre os nós da rede como mostrado na Tabela 10.

**Tabela 10 Distribuição dos documentos**

| Nó | Lista de documentos |
|---|---|
| A | D1, D2 |
| B | D3, D4 |
| C | D5, D6 |
| D | D7, D8 |
| E | D9, D10 |

A idéia básica da utilização do *MSBRI* em ambientes distribuídos é a seguinte: cada nó deve indexar independentemente seus documentos e disponibilizar para seus vizinhos uma representação em menor resolução de seu índice. Os vizinhos, por sua vez, fazem o mesmo com seus próprios índices e com os índices de menor resolução recebidos de seus vizinhos. Caso um nó receba mais de um índice de um mesmo vizinho, ele deve ficar com o índice de maior resolução. O algoritmo abaixo mostra como a distribuição dos índices pode ser feita.

```
1.    DistribuirIndice()
2.    {
3.      indiceReduzido = meuIndice.reduzirResolucao();
4.
5.      ParaCada vizinho em listaDeVizinhos faça
6.      Inicio
7.           Enviar(vizinho, indiceReduzido)
8.           ParaCada índice em listaIndiceVizinhos faça
9.           Inicio
10               Enviar(vizinho,índice.reduzirResolucao())
11           Fim
12      Fim
13    }
```

Onde,

`listaDeVizinhos` é uma lista com uma referencia para os nós vizinhos. Esta lista pode ser obtida com ajuda da matriz de adjacência apresentada em (87).

`meuIndice` é uma variável local que guarda o índice dos documentos indexados pelo nó em questão.

A função `reduzirResolucao()` retorna a versão reduzida de um nível de resolução do índice em questão.

`listaIndiceVizinhos` é uma variável local que guarda os índices anteriormente recebidos dos vizinhos do nó em questão.

Com o algoritmo descrito acima, um nó da rede e capaz de distribuir seus índices por toda a rede em resoluções cada vez mais baixas. Note que o laço mais interno do algoritmo (linhas ii até iv) reduz sucessivamente os índices anteriormente recebidos por um nó. Dessa forma a propagação de informação é sempre reduzida a cada salto na rede.

Ao final da propagação, cada nó terá uma biblioteca de índices em diferentes resoluções. Quando maior a distância entre dois nós A e B, menor a resolução do índice que cada um tem em relação ao outro. A noção de distância entre dois nós adotadas aqui é o número de saltos que os separam. Outras métricas podem ser usadas, o fator importante é a redução gradual de resolução.

A Tabela 11 apresenta a resolução da biblioteca de índice de todos os nós da rede apresentada em (87). Nela estamos supondo que resolução máxima de um índice é 4. Por exemplo, na primeira linha da tabela encontramos a resolução da biblioteca de índices do nó 'A'. Vemos que ele possui, para os documentos locais, um índice de resolução total (4). Para os documentos vindos do nó 'B', o nó 'A' possui uma resolução de nível 3; Para os documentos vindos de 'C', 'A' possui uma resolução de nível 2; e assim sucessivamente.

**Tabela 11 Resolução da biblioteca de índices.**

| Nó | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 4 | 3 | 2 | 1 | 0 |
| B | 0 | 4 | 3 | 2 | 1 |
| C | 1 | 0 | 4 | 3 | 2 |
| D | 2 | 1 | 0 | 4 | 3 |
| E | 3 | 2 | 1 | 0 | 4 |

Portanto, cada nó possui algum tipo de informação de seus vizinhos, mesmo que em resoluções muito baixas. Portanto, é possível que tiremos proveito dessa informação para encaminhar uma consulta ao próximo nó da rede. Vejamos como seria um algoritmo de busca em uma rede ponto-a-ponto como está.

```
1.   Busca(No no, Consulta q)
2.   {
3.      Se jaVisitado(no) retorna
4.         max = obterResolucaoMaxima(no.biblioteca)
5.      ParaCada Doc d em no.biblioteca faça
6.      Iniciar
7.           Se resolução(d) == max faça
8.           Iniciar
9.                Se similar(d,q,max)
10               Iniciar
```

```
11                      resultado.adicionar(d)
12                 Terminar
13            Terminar
14            Senão Se similar(d,q,resolução(d))
15            Iniciar
16                 novoNo = origem(d)
17                 Busca(novoNo,q)
18            Terminar
19
20    }
```

**Figura 41 Proposta de Algoritmo de busca distribuído**

Onde,

`max` é uma variável que contém a resolução máxima do documento analisado.

`Resolução()` é uma função que retorna um inteiro correspondente a resolução corrente do documento dentro da biblioteca de índices contida no nó em questão. Esta função pode usar a equação (78) para calcular a resolução.

`jaVisitado()` é uma função que testa se o nó em questão já fora visitado.

`similar()` é uma função que calcula a similaridade entre a consulta e o documento em questão.

`resultado` é uma variável global que guarda ordenadamente todos os resultados obtidos ordenados por similaridade.

Note que na linha 17 há uma chamada recursiva do algoritmo que possibilita que cada nó da rede seja visitado até que toda a rede seja varrida. A linha 14 desempenha um papel fundamental, pois ela decide a partir da função similar() se deve ou não visitar o próximo nó da rede. Esta decisão pode eliminar o efeito de inundação muito comum em redes ponto a ponto, desde que a função similar seja calibrada de maneira adequada.

# Apêndice B. Modelo de documentos

Para realização de qualquer experimento de Busca e Recuperação de Informação do tipo texto, usamos uma coleção de testes que é composta por um conjunto de documento, um conjunto de consultas e o resultado esperada de cada consulta.

Um sistema de BRI geralmente é confrontado contra uma coleção de testes e parte do pré-suposto que a avaliação das consultas feita nessas coleções é perfeita. Em geral essa avaliação é feita por serem humanos e, portanto, os sistemas de BRI partem do pré-suposto que a avaliação humana é a melhor possível e tentam sempre chegar o

mais próximo dela. Essa suposição é passível de críticas que vão além do escopo dessa dissertação. Apesar disso, nós também partimos dessa mesma suposição para fins de nossa avaliação.

Nessa seção nos limitamos apenas a apresentar o modelo de documento que foi empregado em nossos experimentos. Ele faz parte da coleção conhecida como Cystic Fibrosis (CF). A Figura 42 apresenta um exemplo desse documento, nele podemos destacar os seguintes campos PN, um numero serial que indica o ano (1978) E o número do artigo (1); RN, que é um número serial único para todos os anos. AN; que é um número de acesso do banco de dados da MedLine, instituição responsável pela coleção. AU, nome dos autores do artigo. TI, título do artigo. SO, fonte de onde o artigo foi removido. MJ, assunto principal, é um resumo dos principais assuntos citados no artigo. MN é um resumo dos assuntos secundários citados no artigo. AB/EX é um resumo do artigo. RF é a lista de referências usadas dentro do artigo. CT é a lista de citações recebidas pelo artigo em questão.

PN 78001

RN 00782

AN 78231664

AU Mei-Liu-H.

TI Reactive neuroaxonal dystrophy in children. Clinical pathological correlation.

SO Acta-Neuropathol (Berl). 1978 Jun 30. 42(3). P 237-41.

MJ BRAIN-DISEASES: pa.

MN AUTOPSY. BILE-DUCTS: ab. BRAIN-DISEASES: ci, et. BRAIN-STEM: pa.

CHILD. CYSTIC-FIBROSIS: co. HEART-FAILURE-CONGESTIVE: co. HUMAN.

KIDNEY-FAILURE-CHRONIC: co. NUTRITION-DISORDERS: co.

SPINAL-CORD: pa. TIME-FACTORS. VINCRISTINE: ae.

AB The aim of the present study was to determine the frequency and etioloical factors of neuroaxonal dystrophy (N.D.) in brainstem and spinal cord of children with various non-neurological diseases. The materials used in this study consisted of 266 consecutive autopsies from 1974-1976 and an additional 13 cases from previous years. By far, the most common cause of N.D. in children was chemotherapeutic drugs, particularly vincristine, used for the treatment of malignant tumors. Approximately 90% of these children developed N.D. and the changes ranged from mild to severe. Approximately two-thirds of children with cystic fibrosis, congestive heart failure and chronic renal failure developed N.D. and the changes were mostly mild. Only one-third of patients with congenital biliary atresia and malnutrition developed N.D. and the changes were always mild. The frequency and severity of N.D. increased in patients who had prolonged clinical courses; no N.D. was seen in patients who died from acute cases such as trauma, acute infection, intoxication, acute renal failure or prematurity. This type of N.D. may be considered as reactive to a wide variety of injurious factors such as drug toxicity, malnutrition, chronic hypoxia and chronic renal failure, either alone or in combination. In contrast to previous reports of a low incidence of N.D. in children, there has been a sharp increase in recent years due to the advent of chemotherapy.

RF 001 BERARD-BADIER M ACTA NEUROPATH (BERL) 28 261 974

002 BORISY GG J CELL BIOL 34 525 967

003 BRANNON W ACTA NEUROPATH (BERL) 9 1 967

004 CHOU SM ACTA NEUROPATH (BERL) 3 428 964

005 FUJISAWA K ACTA NEUROPATH (BERL) 8 255 967

006 JELLINGER K GERMAN MEDICAL MONTHLY BD 13 341 968

007$ LAMPERT P J NEUROPATH EXP NEUROL 23 60 964

008 LIU HM ACTA NEUROPATH (BERL) 37 207 977

009 LIU HM NEUROLOGY (MINN) 24 547 974

010 LUI HM ACTA NEUROPATH (BERL) 27 201 974

011 MARTIN JJ EUR NEUROL 8 239 972

012 PENTSCHEW A ACTA NEUROPATH (BERL) 1 313 962

013 SEITELBERGER F ACTA NEUROPATH (BERL) SUPPL 5 17 971

014 SEITELBERGER F PROC CONG NEUROPATH 1ST 3 323 952

015 SUNG JH J NEUROPATH EXP NEUROL 25 341 966

016 SUNG JH J NEUROPATH EXP NEUROL 25 341 966

017 WISNIEWSKI H J CELL BIOL 38 224 968

CT 1 AICARDI J BRAIN 102 727 979

2 GRAFE MR J NEUROPATHOL EXP NEUROL 39 555 980

3 SUNG JH J NEUROPATHOL EXP NEUROL 39 584 980

4 SUNG JH J NEUROPATHOL EXP NEUROL 40 37 981

5 TOMIWA K J NEUROL 229 267 983

6 ROBAIN O J NEUROL NEUROSURG PSYCHIAT 47 65 984

7 WISNIEWSKI KE ACTA NEUROPATH (BERL) 66 68 985

**Figura 42 Modelo de documento da coleção Cystic Fibrosis**

A Figura 43 apresenta um trecho documento cquery, que faz parte da coleção de teste Cystic Fibrosis. Esse documento contém uma lista consultas seguidos de suas respectivas respostas e avaliações feitas por um quatro especialistas. Nessa figura apresentamos apenas uma consulta, mas o documento completo contém um total de 100 consultas.

Nesse documento podemos destacar os seguintes campos: QN é um número serial que indica o número da consulta. QU indica o corpo da consulta. NR indica o número de documentos relevantes à consulta em particular. RD é uma lista contendo o número do documento relevante que deve ser recuperada pelo sistema de BRI, seguido de um número que fornece a avaliação.

A avaliação da coleção é feita por quatro especialistas, cada um pode dar uma nota entre zero e dois. Zero (0) indica que o especialista não considera o documento relevante para consulta; um (1) indica que o especialista considera o documento relevante para a consulta e; dois (2) indica que o especialista considera o documento muito relevante para a consulta.

Por exemplo, na Figura 43 o documentos 139 é considerado relevante pelo primeiro especialista, e muito relevante pelos outros 3, portanto suas nota é 1222.

```
QN 00001
QU What are the effects of calcium on the physical properties of mucus
   from CF patients?
NR 00034
RD  139 1222  151 2211  166 0001  311 0001  370 1010  392 0001  439 0001
   440 0011  441 2122  454 0100  461 1121  502 0002  503 1000  505 0001
   520 1010  522 0001  526 1011  527 1001  533 2222  593 1000  619 0001
   737 0100  742 0100  789 0001  827 0010  835 0001  861 0001  875 1121
   891 0001  921 1010  922 1010 1175 0100 1185 0001 1222 0001
```

**Figura 43 Documento contendo as consultas e respectivas respostas esperadas**