AVALIAÇÃO DA QUALIDADE DO USO DE *WAVELETS* PARA RECUPERAÇÃO, CLASSIFICAÇÃO E AGRUPAMENTO DA INFORMAÇÃO TEXTUAL

Fabrício Raphael Silva Ferreira

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Geraldo Bonorino Xexéo

Rio de Janeiro
Setembro de 2011

AVALIAÇÃO DA QUALIDADE DO USO DE *WAVELETS* PARA RECUPERAÇÃO, CLASSIFICAÇÃO E AGRUPAMENTO DA INFORMAÇÃO TEXTUAL

Fabrício Raphael Silva Ferreira

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) 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.

Examinada por:

Prof. Geraldo Bonorino Xexéo, D.Sc.

Prof. Jano Moreira de Souza, D.Sc.

Prof. Jorge Lopes de Souza Leão, Dr.Ing.

RIO DE JANEIRO, RJ – BRASIL
SETEMBRO DE 2011

Ferreira, Fabrício Raphael Silva

Avaliação da qualidade do uso de *wavelets* para recuperação, classificação e agrupamento da informação textual/Fabrício Raphael Silva Ferreira. – Rio de Janeiro: UFRJ/COPPE, 2011.

XVII, 100 p.: il.; 29, 7cm.

Orientador: Geraldo Bonorino Xexéo

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2011.

Referências Bibliográficas: p. 81 – 86.

1. Wavelet. 2. Recuperação de Informação. 3. Busca e Recuperação de Informação. 4. Classificação. 5. Agrupamento. 6. Transformada Wavelet. I. Xexéo, Geraldo Bonorino. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Aos meus avós e pais pelo dom da vida, pela formação e amparo ao longo desses anos. À minha esposa Karina. À minha irmã Anna Rafaela.

Agradecimentos

Registro aqui meus sinceros agradecimentos a todos que contribuíram para que a conclusão deste trabalho pudesse ser alcançada com sucesso. Em especial, quero agradecer a:

Deus, pela permissão de viver e por todos os acontecimentos, que me proporcionaram crescimento e fortalecimento;

Meus pais, Rafael e Ivonete, pelo esforço na minha formação moral e intelectual;

Meus avós, pelo provimento de um referencial moral, bases sólidas do que eu sou;

Karina, minha esposa, pelo apoio incondicional e fundamental nesta conquista;

Minha irmã, Anna Rafaela, pela oportunidade de aprimorar a personalidade através da convivência;

Todos os meus amigos, em especial ao Vinícius e Olivério pela grande amizade construída ao longo dos últimos 3 anos;

Meu padrinho "Nonatinho" pela amizade, pelos sábios conselhos e pela experiência desinteressadamente repassada; à minha madrinha Irene, pelo suporte e incentivo sempre presentes; a todos os meus tios e primos, pela convivência.

Prof. Geraldo Bonorino Xexéo, meu orientador, pela oportunidade desta pesquisa e pelo suporte durante todo este tempo, fundamental para a conclusão deste trabalho;

Todos os meus professores que tive no decorrer da minha formação, principalmente àqueles que desempenharam seu papel não só no aspecto intelectual, mas principalmente no moral.

Todos os amigos da GPE, em especial ao meu chefe Ricardo Barros, pela oportunidade de mostrar a maneira como trabalho, e de como posso contribuir;

Amigos com os quais tive a oportunidade de estudar ou trabalhar juntos, e que com certeza contribuíram para o meu aprendizado técnico;

Sukyo Mahikari, por me mostrar que posso ser útil à Deus e à sociedade através do meu trabalho;

Amigos praticantes da Arte *Mahikari*, por me incentivarem a sempre colocar Deus em primeiro lugar, e também por terem sido os principais incentivadores para a conclusão deste trabalho;

Todos que me ajudaram na revisão e correção da escrita deste trabalho;

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte financeiro;

Todos aqueles com quem tive algum contato durante esta etapa.

Meu sincero muito obrigado a todos!

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

AVALIAÇÃO DA QUALIDADE DO USO DE *WAVELETS* PARA RECUPERAÇÃO, CLASSIFICAÇÃO E AGRUPAMENTO DA INFORMAÇÃO TEXTUAL

Fabrício Raphael Silva Ferreira

Setembro/2011

Orientador: Geraldo Bonorino Xexéo

Programa: Engenharia de Sistemas e Computação

Este trabalho avalia o emprego da transformada *wavelet* sobre os sistema de recuperação, classificação e agrupamento da informação textual, comparando-se com os modelos e algoritmos clássicos, ou bastantes difundidos em cada um desses sistemas em termos de suas eficácias. Presume-se que, a partir dos trabalhos referenciados, a utilização da *wavelet* tenha resultados tão bons ou melhores quando se reduz a quantidade da informação com o processo da transformada *wavelet* e devido a suas propriedades. Será comparada a eficácia não apenas com os modelos clássicos e/ou algoritmos difundidos, como também entre as transformadas de dois tipos de *wavelets* mais utilizadas computacionalmente: a *wavelet* de Haar e a *wavelet* de Daubechies $D4$.

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

QUALITY ASSESSMENT OF THE USE OF WAVELETS FOR RETRIEVAL, CLASSIFICATION AND CLUSTERING OF TEXTUAL INFORMATION

Fabrício Raphael Silva Ferreira

September/2011

Advisor: Geraldo Bonorino Xexéo

Department: Systems Engineering and Computer Science

This study evaluates the use of the transform wavelet on a retrieval system, classification and clustering the textual information, comparing with the classical models and traditional algorithms, or very diffused in each of these systems in terms of their effectiveness. Presumably, from the referenced work, the application of wavelet has as good as or better results when reduced the amount of information with the process of the transform wavelet and due its properties. Its effectiveness will be compared not only to the classic and/or diffused algorithms, but also among the transformed ones from two types of more computationally used wavelet: the Haar wavelet and the Daubechies $D4$ wavelet.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

C_ψ	constante de Calderón, p. 40
$L^2(\mathbb{R})$	espaço de Hilbert em $\mathbb{R}$, p. 42
Ψ	operador linear da transformada *wavelet*, p. 42
$\mathcal{F}$	operador da transformada de Fourier, p. 38
ω	frequência instantânea angular, p. 37
$\phi(t)$	função de escala, ou, *wavelet* pai, p. 44
ψ	*wavelet*, p. 36
$\psi(t)$	função da *wavelet* mãe, p. 40
$\psi_{\mathcal{E},\mathcal{D}}(t)$	função da *wavelet* filha, p. 40
e	base dos logaritmos naturais, p. 37
$f(t)$	função do sinal no domínio do tempo, p. 38
i	unidade imaginária, p. 37
t	variável independete pertencente ao domínio do tempo, p. 37
*	operador do complexo conjugado, p. 42

Lista de Abreviaturas

BRI	Busca e Recuperação de Informação, p. 1
CLEF	*Cross Language Evaluation Forum*, p. 22
COPPE	Instituto Alberto Luiz Coimbra de Pós-graduação e Pesquisa de Engenharia, p. 14
FCD	Função de Codificação de Documento, p. 14
FCT	Função de Codificação de Termos, p. 14
FT	*Fourier Transform*, p. 38
IDF	*Inverse Document Frequence*, p. 9
KNN	*K-Nearest Neighbor*, p. 24
LSI	*Latent Semantic Analysis*, p. 11
MAC	Modelo de Alta Correlação, p. 14
MRA	*Multi-Resolution Analysis*, p. 42
MSBRI	Modelo de Sinais para Busca e Recuperação de Informação, p. 14
NTCIR	*NII Test Collection for IR Systems*, p. 22
PRP	*Probability Ranking Principle*, p. 10
RI	Recuperação de Informação, p. 1, 4
RSS	*Residual Sum of Squares*, p. 28
STFT	*Short-Time Fourier Transform*, p. 38
SVD	*Singular Value Decomposition*, p. 11
SVM	*Support Vector Machine*, p. 30

TF-IDF	Produto entre o TF e o IDF, p. 10
TF	*Term Frequence*, p. 9
TREC	*Text Retrieval Conference*, p. 22
VSM	*Vector Space Model*, p. 8
WT	*Wavelet Transform*, p. 41

Capítulo 1

Introdução

1.1 Motivação

A Busca e Recuperação de Informação (BRI), ou simplesmente Recuperação de Informação (RI) do tipo texto é o ramo da computação que se dedica ao estudo do armazenamento e recuperação de documentos de texto. Diversas técnicas de RI foram e continuam sendo desenvolvidas. Algumas técnicas recentes aplicaram o uso do processamento de sinais, que até então foram usadas com eficácia em sistemas de RI Multimídia como mostram NIBLACK *et al.* [1], MURALA *et al.* [2], ao processo de indexação e recuperação de informações textuais, como a transformada de Fourier em FLESCA *et al.* [3] e a transformada *wavelet* em PARK *et al.* [4], THAICHAROEN *et al.* [5], XEXEO *et al.* [6].

Desse modo, como os referidos trabalhos sobre as *wavelets* utilizam-se da vantagem desta em ter uma performance superior ao analisar, processar e sintetizar imagens e sinais onde o método de Fourier não obtém performance aceitável, surge daí a necessidade de comprovar se o mesmo ocorre ao aplicá-las em técnicas de RI textuais. Além disso, também se faz necessário avaliar a eficácia do uso das *wavelets* em sistemas de RI textuais e compará-lo com outras técnicas aplicadas em sistemas de RI consolidados, como em sistemas de classificação e agrupamento da informação textual.

1.2 Objetivo

Esta dissertação tem dois objetivo principais. O primeiro é avaliar se a eficácia do uso das *wavelets* é comparável com outras técnicas aplicadas em sistemas de recuperação, classificação e agrupamento já consagradas. O segundo é avaliar qual tipo de função *wavelet*, e em quais configurações da sua transformada obtem-se uma eficácia superior para cada um desses sistemas.

Dessa forma, pretende-se também tornar simples e clara a aplicabilidade de *wavelets* no processamento da informação textual.

1.3 Metodologia

Para atingir o objetivo deste trabalho, as *wavelets* serão aplicadas em sistemas de recuperação, classificação e agrupamento da informação textual, de forma a comparar, em termos de eficácia, com outras ferramentas já aplicadas nesses sistemas. Dessa forma, o uso da ferramenta matemática será avaliado em três tarefas de processamento da informação textual: a recuperação de informação, a classificação e o agrupamento. Na recuperação de informação, o trabalho será realizado da seguinte forma. Primeiramente será definido um *baseline* que é o modelo do espaço vetorial. Em seguida, serão usadas as *wavelets* em RI e avaliadas combinando-as com outras maneiras possíveis para o processamento da informação nesta tarefa, para assim analisar qual a melhor combinação testada. Um primeiro componente a ser variado é a função que irá identificar unicamente cada termo da coleção, como citado nas referências há o modelo de alta correlação em que se utiliza de uma propriedade de correlação entre os termos para identificá-los. E, logicamente, o outro componente a ser variado é a própria *wavelet*, em que se podem utilizar vários tipos delas, como as *wavelets* de *Haar*, as *wavelets* de *Daubechies*, e outras a serem estudadas posteriormente. Pode-se variar também a maneira de como avaliar a similaridade entre as transformadas *wavelets* representativas de cada documento. E assim, as combinações serão comparadas com o modelo do espaço vetorial através da medida de algumas medidas, como *precision* e *recall*.

Outra tarefa a ser estudada sua viabilidade, bem como sua eficiência com o uso das *wavelets* é a classificação, onde um forte candidato a ser o *baseline* é o KNN. E por último será avaliado o uso de *wavelets* também nos algoritmos de agrupamento, podendo ser o *k-means* o *baseline*. É válido observar que embora se adote um algoritmo como *baseline*, não se impede de que outros algoritmos sejam comparados também com a aplicação de *wavelets*.

1.4 Organização da dissertação

Este trabalho está organizado em 5 capítulos. No Capítulo 1, no qual consta a introdução, foram exibidos a motivação, o objetivo, a metodologia e a organização do trabalho.

Nos Capítulos 2 e 3 é apresentada a revisão bibliográfica das tecnologias envolvidas nesta dissertação. No Capítulo 2 revisa-se ainda os problemas da recuperação, da classificação e do agrupamento da informação textual. E no Capítulo 3 procura-se

dar uma visão geral das *wavelets* e suas propriedades, e ainda evidencia-se suas aplicações e os trabalhos relacionados a suas aplicações no processamento da informação em forma de sinal, em especial informações do tipo texto.

No Capítulo 4, será feita a avaliação do uso de *wavelet* em RI, classificação e agrupamento de textos, através de experimentos e da análise de seus resultados. Serão avaliados requisitos funcionais, utilizando-se das métricas de avaliação utilizados por cada técnica em que as *wavelets* foram aplicadas. Aqui também é comentado os requisitos não funcionais como tempo e quantidade de recursos computacionais utilizados tanto para geração dos índices como para a execução final de cada técnica que empregou as *wavelets*.

No Capítulo 5, são apresentadas as contribuições e limitações desse trabalho, e trabalhos futuros.

Capítulo 2

Recuperação, Classificação e Agrupamento de Informações Textuais

Neste capítulo, serão apresentados os sistemas típicos de mineração de informação textual: recuperação, classificação e agrupamento. Serão apresentados também os modelos e técnicas utilizados e consolidados em cada um desses sistemas.

2.1 Recuperação de Informação Textual

Suponha que um indivíduo deseja buscar em vários documentos partes que contenham determinado assunto que lhe interessa. Uma vez definido o assunto, o indivíduo sugere palavras ou termos que representem o que se deseja procurar. Bom, isso é uma tarefa muito fácil e rápida quando se tem apenas um documento de pouco conteúdo. Mas, e se tivermos um conjunto de milhares ou milhões de documentos, e o indivíduo deseja procurar em todos esses textos um assunto representado por alguns termos, isso se torna uma tarefa impraticável de ser executada manualmente. Para isso, torna-se necessário criar um sistema que automatize de alguma forma a recuperação da informação desejada pelo indivíduo.

Partindo dessa necessidade, surgiu a área de estudo denominada Recuperação de Informação. A Recuperação de Informação (RI) é a busca por objetos não estruturados (geralmente documentos ou textos) que satisfazem a necessidade de uma determinada informação a partir de um universo de informações MANNING *et al.* [7]. Vale ressaltar que esta área não só se preocupa com a maneira de buscar a informação, mas também em como representar, organizar, armazenar e acessá-la visando tornar esta busca mais rápida e eficiente.

O primeiro passo para a automatização da recuperação é a maneira como os

documentos serão persistidos. Para evitar a leitura integral dos documentos a cada consulta realizada, é feito o processo de indexação destes, criando-se uma matriz de incidência termo-documento, na qual os termos são geralmente as palavras contidas nestes documentos. Entretanto, antes de compor o índice, estes termos são pré-processados pelas seguintes etapas MANNING *et al.* [7], BAEZA-YATES e RIBEIRO-NETO [8]: análise léxica ou normalização de *token*, remoção das *stop words*, *stemming*, *lemmatization*, escolha dos termos do índice (palavras-chaves) e construção de estruturas de categorização dos termos (*thesaurus*). Dentre os benefícios dessas etapas, tem-se a identificação somente de palavras, a redução da quantidade de termos que têm significados similares ou próximos e a redução do tamanho do índice.

Ao final do pré-processamento, forma-se um dicionário de termos ou vocabulário, que contém todos os termos que irão compor o índice, denominados de termo-índice. O índice deve conter de forma sucinta e concisa a informação de todas as ocorrências dos termos em cada documento contido na coleção de documentos, também conhecido como *corpus*. E cada documento é representado por um subconjunto dos termos que fazem parte do índice da coleção. Além da matriz de incidência termo-documento, outra maneira de indexar que se tornou padrão para a recuperação de informação é o índice invertido, ou arquivo invertido, que mapeia a partir de cada termo do vocabulário, suas ocorrências nos documentos. Para isso, também se faz necessário que cada documento tenha uma identificação, o identificador do documento. Quando um termo ocorre em um documento, o índice invertido mapeia o termo-índice ao identificador do documento, e cada mapeamento entre o termo e um documento é conhecido como *posting*. E assim, a lista desses mapeamentos é chamada de *postings list*, ou lista invertida. Adicionalmente, pode-se extrair desse índice algumas estatísticas, como a frequência do termo, definida pela quantidade de ocorrências de um termo em um determinado documento, ou frequência do documento, definido pelo número de documentos na coleção que contém um determinado termo.

Dessa forma, sistemas de RI geralmente utilizam termos de índice para indexar e recuperar documentos. Com isso, a recuperação baseada nesse tipo de indexação pode ser implementada eficientemente com simplicidade, pontos muito importantes a serem aplicados na representação, na organização e no armazenamento, a fim de reduzir o esforço da formulação e execução de uma dada consulta.

Uma vez estabelecido o índice, o processo de recuperação continua quando uma consulta é submetida, e então os termos da consulta são verificados com os termos existentes no índice. Isso possibilita o sistema decidir, baseado em um determinado modelo de RI, quais os documentos são relevantes e assim recuperá-los em uma ordem de acordo com o modelo empregado. Essa ordenação deve refletir a relevância

dos documentos para a consulta. A figura 2.1 ilustra de forma simples e genérica o processo de recuperação da informação.

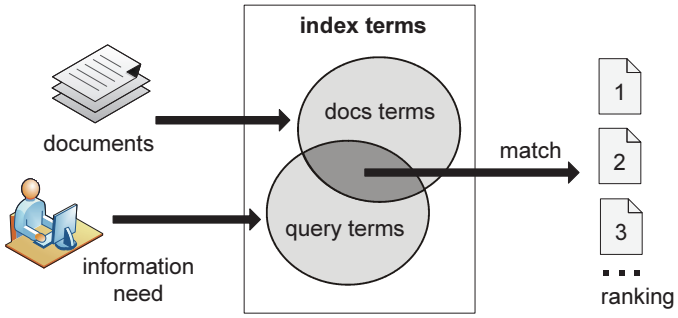


Figura 2.1: Processo de Recuperação de Informação (Fonte: BAEZA-YATES e RIBEIRO-NETO [8], Cap. 3)

Os modelos de RI são processos complexos que almejam uma ordenação dos documentos através de uma função de classificação (*ranking*) – também conhecida como função de similaridade ou função de ordenação –, que atribui pontuações (*scoring*) aos documentos de acordo com sua relevância em relação a uma determinada consulta. Como afirma BAEZA-YATES e RIBEIRO-NETO [8], um modelo de RI é definido por uma quádrupla definida em:

$$[\mathbf{D},\ \mathbf{Q},\ \mathcal{F},\ R(q_i, d_j)]$$

em que:

- $\mathbf{D}$ é o conjunto das representações lógicas dos documentos da coleção,
- $\mathbf{Q}$ é o conjunto das representações lógicas das consultas,
- $\mathcal{F}$ é a estrutura matemática (*framework*) que define a modelagem, ou a representação lógica, dos documentos e das consultas,
- $R(q_i, d_j)$ é a função de classificação ou similaridade (*ranking*) a ser aplicada às representações lógicas de um documento ($d_j \in \mathbf{D}$) e de uma dada consulta ($q_i \in \mathbf{Q}$).

A seguir, serão explanados os modelos clássicos de RI, além de outros modelos difundidos e relacionados a este presente trabalho. Os modelos clássicos são: modelo booleano, modelo probabilístico e modelo vetorial. Outros modelos que serão explanados são: o modelo de indexação semântica latente e o modelo de alta correlação XEXEO *et al.* [6]. Os modelos clássicos 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 classificação (*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 existentes entre os modelos de recuperação da informação textual, há um ponto em comum: todos consideram os documentos como uma coleção de termos. Quando um indivíduo vê um termo dentro de um documento, ele percebe a semântica envolvida com aquele termo. Mas uma máquina não consegue perceber essa semântica na sua totalidade, diferentemente do humano. Então, uma forma da máquina entender o que o termo representa para aquele documento é atribuir pesos a cada um dos termos, de forma a destacar determinados termos mediante aos outros.

Desta forma, dado uma coleção de documentos $\mathbf{D} = \{d_1, d_2, \ldots, d_m\}$, extrai-se um vocabulário $\mathbf{V} = \{k_1, k_2, \ldots, k_t\}$, e constrói-se um índice na forma de matriz de incidência termo-documento W, onde cada elemento $w_{i,j} \geqslant 0$ representa o peso do termo-índice k_i no documento d_j:

$$\begin{array}{c} \\ k_1 \\ k_2 \\ \vdots \\ k_t \end{array} \begin{array}{c} \begin{array}{cccc} d_1 & d_2 & \cdots & d_m \end{array} \\ \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,m} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{t,1} & w_{t,2} & \cdots & w_{t,m} \end{bmatrix} \end{array}$$

2.1.1 Modelo Booleano

O Modelo Booleano é um modelo simples baseado na teoria dos conjuntos e na álgebra booleana, de onde herda o formalismo, de maneira que as consultas são construídas, com uma semântica intuitiva e precisa, através de expressões booleanas e de operadores lógicos, como: AND, OR, NOT. Dada uma consulta q, esta é reescrita para sua forma normal disjunta q_{DNF}, composta por uma disjunção de componentes conjuntivos da consulta $c(q)$.

Cada documento é representado pelo modelo apenas como um conjunto de termos presentes. O peso $w_{i,j}$ atribuído a cada termo-índice é binário, 0 ou 1, indicando a presença ou ausência do termo no documento. Formalmente tem-se:

- $w_{i,j} \in \{0, 1\}$: peso associado ao par (k_i, d_j)
- $w_{i,q} \in \{0, 1\}$: peso associado ao par (k_i, q)
- q_{DNF}: forma normal disjunta da consulta q
- $c(q)$: é um componente conjuntivo da consulta q

- $c(d_j) = (w_{1,j} \wedge w_{2,j} \wedge \ldots \wedge w_{t,j})$

A função de similaridade entre um documento d_j e uma consulta q é definida em 2.1:

$$sim(d_j, q) = \begin{cases} 1 & \text{se } \exists\ c(q) \mid c(q) = c(d_j) \\ 0 & \text{caso contrário} \end{cases} \tag{2.1}$$

Assim, este modelo se limita apenas em acusar para cada documento se é, ou não, relevante para uma dada consulta, sem qualquer escala de classificação que indique o qunto cada documento é relevante para a consulta.

2.1.2 Modelo Vetorial (VSM)

O Modelo Vetorial (*Vector Space Model* - VSM) atribui pesos não binários aos termos, e assim faz uso desses pesos para calcular o grau de relevância entre uma consulta e cada documento. E então, os documentos são recuperados em ordem decrescente pelo grau de relevância. Para isso, o modelo se baseia na álgebra linear e no cálculo vetorial, de modo a representar os documentos e consulta em vetores sobre o espaço de termos. O espaço de termos é dado pelo vocabulário $\mathbf{V} = \{k_1, k_2, \ldots, k_t\}$, enquanto que as representações dos documentos e da consulta é dada por $\vec{d_j} = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$ e $\vec{q} = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$, respectivamente. E, a partir das representações vetoriais de um documento e a consulta, é possível calcular a diferença entre eles, determinando assim o grau de similaridade, ou seja, o grau de relevância para ordenar os resultados da recuperação.

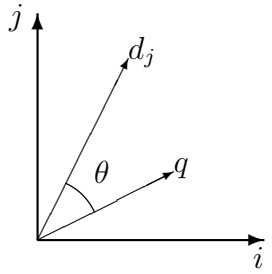


Figura 2.2: Ângulo θ formado pelos vetores $\vec{d_j}$ e $\vec{q}$

Dessa forma, para o modelo vetorial adota-se como função de similaridade o cosseno do ângulo θ formado pelos vetores $\vec{d_j}$ e $\vec{q}$ (ver Figura 2.2), definida em 2.2:

$$sim(d_j, q) = \cos(\theta) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \times |\vec{q}|} \tag{2.2}$$

onde:

- $\vec{d_j} \cdot \vec{q}$ é o produto escalar entre os vetores $\vec{d_j}$ e $\vec{q}$,
- $|\vec{d_j}| \times |\vec{q}|$ é o produto das distâncias Euclidianas dos vetores $\vec{d_j}$ e $\vec{q}$.

De outra forma, usando os elementos dos vetores (os pesos dos termos), se obtém uma definição programática em 2.3:

$$sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}} \tag{2.3}$$

E, como $w_{i,j} \geqslant 0$ e $w_{i,q} \geqslant 0$, tem-se uma similaridade no seguinte intervalo de valores: $0 \leqslant sim(d_j, q) \leqslant 1$. Mas que valores atribuir aos pesos de cada termo-índice, para que se possa chegar a calcular a similaridade? Bom, no modelo vetorial geralmente utiliza-se de informações da frequência do termo em toda coleção e em cada documento para se obter os pesos dos termos.

A quantidade de ocorrências de um termo em um documento, diante da maior quantidade de ocorrência de um termo que aparece nesse mesmo documento, é conhecida simplesmente por frequência do termo (*Term Frequence* - TF), dada pela Fórmula 2.4:

$$TF_{i,j} = \frac{f_{i,j}}{\max(f_j)} \tag{2.4}$$

onde:

- $TF_{i,j}$ é a freqüência (relativa) do termo k_i dentro do documento d_j,
- $f_{i,j}$ é a frequência absoluta do termo k_i dentro do documento d_j,
- $\max(f_j)$ é o fator normalizador da freqüência absoluta, ou seja, é a frequência máxima de um termo que aparece dentro do documento d_j.

Entretanto, essa informação não é suficiente para inferir qual a importância do termo diante de toda a coleção, assim é preciso que se conheça a quantidade de documentos nos quais um determinado termo aparece, o que é chamado de frequência de documento (df_i). Contudo, essa medida pura ainda não transmite uma idéia da escala de importância de um termo diante aos outros. Assim, utiliza-se a Fórmula 2.5, que define o inverso da frequência de documento (*Inverse Document Frequence* - IDF), para contribuir com o peso do termo. Segue a fórmula:

$$IDF_i = \log \frac{N}{df_i} \tag{2.5}$$

onde:

- IDF_i é o inverso da frequência de documento do termo k_i,
- N é o número total de documentos da coleção, ou seja, $|\mathbf{D}|$,
- df_i é a frequência de documento referente ao termo k_i.

Finalmente, pode-se definir os valores do pesos, sintetizando as informações esclarecidas anteriormente, na forma do produto entre o TF e IDF, conhecido como TF-IDF. Formalmente, tem-se 2.6:

$$w_{i,j} = TF_{i,j} \times IDF_i = \frac{f_{i,j}}{\max(f_j)} \times \log \frac{N}{df_i} \tag{2.6}$$

Já para a consulta, geralmente cada elemento do vetor de pesos que a representa é definido de forma binária de acordo como termo de índice que está presente, ou não na consulta, sendo atribuídos os valores 0 ou 1 respectivamente.

Com esse modelo, percebem-se diversos pontos que contribuem para a qualidade do resultado recuperado, como a pesagem contínua, e não binária, dos termos para os documentos. E como consequência dessa pesagem e do cálculo da similaridade do cosseno, o resultado é ordenado por grau de relevância de forma decrescente. Entretanto, mesmo que este modelo considere fatores que levem em conta a importância de cada termo para a coleção e para cada documento, o modelo vetorial ainda não contempla a correlação que possa existir entre os diferentes termos.

2.1.3 Modelo Probabilístico

O Modelo Probabilístico se baseia na teoria da probabilidade, como o ferramental matemático, para resolver o problema da recuperação da informação. Dada uma consulta, o modelo presume que exista um conjunto de resposta ideal R para a consulta, e usa das descrições desse conjunto para recuperar os documentos relevantes BAEZA-YATES e RIBEIRO-NETO [8]. As descrições desse conjunto são dadas pela consulta, que é vista como uma especificação de propriedades do conjunto ideal. Na prática, isso é feito através da recuperação – realizada de alguma maneira – de uma amostra inicial de alguns documentos, e inspecionados pelo usuário, restando apenas os que são realmente relevantes. E, com esses últimos, a descrição do conjunto de resposta ideal é melhorada pelo aperfeiçoamento da amostra, e cada vez mais com a repetição do processo. Com isso, busca-se chegar à melhor resposta que se pode obter, com base nos dados disponíveis, de acordo com o Princípio de Classificação da Probabilidade (*Probability Ranking Principle* - PRP) MANNING *et al.* [7].

Igualmente ao modelo booleano, a indexação dos termos é feita com pesos binários, ou seja, $w_{i,j} \in \{0, 1\}$. O modelo tenta estimar qual a probabilidade de um documento d_j ser relevante para uma consulta q, assumindo que essa probabilidade depende somente da representação do documento e da consulta. E faz uso de informações do conjunto de resposta ideal R, para maximizar a probabilidade da relevância.

A partir disso, sabendo que:

- R é o conjunto dos documentos relevantes para a consulta q,
- $\overline{R}$ é o complemento de R, ou seja, o conjunto de documentos não relevantes para a consulta q,
- $P(R|\vec{d_j})$ é a probabilidade de d_j ser relevante para a consulta q,
- $P(\overline{R}|\vec{d_j})$ é a probabilidade de d_j não ser relevante para a consulta q,

tem-se a função de similaridade definida em 2.7:

$$sim(d_j, q) = \frac{P(R|\vec{d_j})}{P(\overline{R}|\vec{d_j})} \tag{2.7}$$

Por *Bayes*, ainda se tem:

$$sim(d_j, q) = \frac{P(\vec{d_j}|R, q) \times P(R, q)}{P(\vec{d_j}|\overline{R}, q) \times P(\overline{R}, q)} \quad \sim \quad \frac{P(\vec{d_j}|R, q)}{P(\vec{d_j}|\overline{R}, q)} \tag{2.8}$$

onde:

- $P(\vec{d_j}|R, q)$ é a probabilidade de seleção aleatória do documento a partir do conjunto R,
- $P(R, q)$ é a probabilidade de um documento selecionado aleatoriamente, a partir da coleção, ser relevante para a consulta q,
- $P(\vec{d_j}|\overline{R}, q)$ e $P(\overline{R}, q)$ são os complementos das definições anteriores.

Esta última Fórmula (2.8) representa as chances de um documento estar entre os documentos relevantes para a consulta q. Como o modelo probabilístico age de maneira incremental, a Fórmula 2.8 é realimentada a cada rodada e refina os resultados, até chegar ao ponto em que o grau de refinamento seja considerado suficiente.

Como principal desvantagem, este modelo necessita de um bom levantamento de estimativas iniciais. Mas como vantagem, apesar do modelo empregar o mesmo peso do modelo booleano, a função de similaridade resulta em uma saída não binária, dentro de uma escala de classificação, que reflete a relevância do documento para a consulta.

2.1.4 Modelo de Indexação Semântica Latente (LSI)

O Modelo de Indexação Semântica Latente (*Latent Semantic Analysis* - LSI) é fundamentado também na álgebra linear, mas fazendo uso da Decomposição de Valor Singular (*Singular Value Decomposition* - SVD). Com esse ferramental, o Modelo

LSI objetiva relacionar a informação subjacente entre os termos ou mesmo entre os documentos, a fim de recuperar e classificar os documentos não só pelos termos do índice, mas também pela priorização da informação de correlação entre os termos. Dessa forma, documentos que compartilham conceitos semelhantes, mesmo que não contenham os mesmos termos da consulta, podem ser recuperados MANNING *et al.* [7], BAEZA-YATES e RIBEIRO-NETO [8].

Para ilustrar matematicamente esse modelo, tomemos:

- t: a quantidade total de termos do índice,
- m: o número de documentos,
- $W = [w_{i,j}]$: a matriz termo-documento $t \times m$, em que $w_{i,j}$ é o peso atribuído ao par termo-documento $[k_i, d_j]$, e este peso pode ser baseado no TF-IDF, assim como no modelo vetorial.

A matriz $W = [w_{i,j}]$ pode ser decomposta em três componentes usando a decomposição de valor singular (SVD), assim, tem-se a Fórmula 2.9:

$$W = O \cdot \Sigma \cdot P^T \tag{2.9}$$

onde:

- O é uma matriz ortogonal $t \times r$ composta por autovetores derivada de $C = W \cdot W^T$,
- P^T é uma matriz ortogonal $r \times m$ composta por autovetores derivada de $W^T \cdot W$, em que ^{T} denota ser uma matriz transposta,
- Σ é uma matriz diagonal $r \times r$ composta por valores singulares em sua diagonal principal, e estão ordenados de forma decrescente, podendo ser nulos nas últimas posições da matriz. E $r = \min(t, d)$ é conhecido como o nível (*rank*) de W.

Similarmente, para a transposta de W, tem-se a Fórmula 2.10:

$$W^T = P \cdot \Sigma \cdot O^T \tag{2.10}$$

A partir dessas equações é possível inferir, utilizando-se das operações e propriedades algébricas, novas equações que tragam informações comparativas entre termo-termo ou documento-documento. Dessa forma, multiplica-se W pela transposta W^T, e o resultado será uma matriz $t \times t$ de correlação termo-termo, que pode ser decomposta por três componentes como mostra 2.11:

$$WW^T = O \cdot \Sigma^2 \cdot O^T \tag{2.11}$$

Da mesma forma, e alterando-se a ordem da multiplicação, o resultado será uma matriz $m \times m$ de correlação documento-documento 2.12:

$$W^T W = P \cdot \Sigma^2 \cdot P^T \tag{2.12}$$

Importante salientar que as matrizes WW^T e $W^T W$ resultantes das Fórmulas 2.11 e 2.12 são matrizes diagonais simétricas.

Agora, como afirmado em relação à Σ, os valores singulares da diagonal principal estão dispostos em ordem decrescente, e os últimos valores podem ser nulos, ou pequenos o suficiente para serem considerados nulos. Dessa forma, pode-se descartar alguns dos últimos valores, e suas respectivas linhas e colunas, para os cálculos das matrizes de correlação, reduzindo a dimensão das matrizes a serem trabalhadas, MANNING *et al.* [7], BAEZA-YATES e RIBEIRO-NETO [8]. Em outras palavras, descartam-se alguns termos ou documentos, dependendo se está calculando WW^T ou $W^T W$ respectivamente, sem comprometer a significância da maior parte dos termos ou documentos.

De forma matemática, primeiramente aproxima-se Σ algumas dimensões mais abaixo, ao se descartar as últimas linhas e colunas, resultando em Σ_k, em que k ($k < r$) é a nova dimensão de Σ, e o novo nível (*rank*) de W. Uma vez que decide-se em qual nível a aproximação será feita para W, tornando W_k, aplica-se também a aproximação em O e O^T (ou P e P^T), obtendo O_k e O_k^T (ou P_k e P_k^T). Consequentemente, essas aproximações resultam na matriz de correlação aproximada $W_k W_k^T$ (ou $W_k^T W_k$). A partir disso, ilustram-se as equações enunciadas anteriormente com os componentes aproximados ao nível k em 2.13, 2.14, 2.15 e 2.16:

$$W_k = O_k \cdot \Sigma_k \cdot P_k^T \tag{2.13}$$

$$W_k^T = P_k \cdot \Sigma_k \cdot O_k^T \tag{2.14}$$

$$W_k W_k^T = O_k \cdot \Sigma_k^2 \cdot O_k^T \tag{2.15}$$

$$W_k^T W_k = P_k \cdot \Sigma_k^2 \cdot P_k^T \tag{2.16}$$

O valor de k deve ser suficiente grande para permitir a adaptação das características dos dados, como também pequeno o suficiente para descartar os detalhes não-relevantes sem comprometer a representação dos dados BAEZA-YATES e RIBEIRO-NETO [8].

Uma vez aplicada a SVD à matriz original e encontrar a matriz W_k aproximada, contendo os vetores que representam os documentos, resta ver como se representa uma dada consulta q, para se calcular a similaridade entre um documento d_j e uma consulta q. No modelo LSI a consulta é representação de uma consulta, cuja representação inicial dada por $\vec{q}$, como no modelo vetorial, é transformada usando a Fórmula 2.17:

$$\vec{q_k} = \Sigma_k^{-1} \cdot O_k^T \cdot \vec{q} \tag{2.17}$$

Finalmente, com as representações vetoriais dos documentos e da consulta dadas respectivamente por $\vec{d_{j_k}}$ e $\vec{q_k}$, é possível calcular a similaridade utilizando-se da distância do cosseno entre os vetores, assim como no modelo vetorial. Para a função de classificação, se for escolhida a distância do cosseno, tem-se 2.18:

$$sim(d_j, q) = \cos(\theta) = \frac{\vec{d_{j_k}} \cdot \vec{q_k}}{|\vec{d_{j_k}}| \times |\vec{q_k}|} = \frac{\sum_{i=1}^{k} w_{i,j} \times w_{i,q_k}}{\sqrt{\sum_{i=1}^{k} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{k} w_{i,q_k}^2}} \tag{2.18}$$

Este modelo torna-se bastante interessante no que diz respeito ao uso da correlação subjacente entre os termos, ou entre os documentos, reconhecendo representações com distribuições parecidas, mesmo com a redução de informação se aplicada na medida certa. Entretanto, no LSI a escolha da "medida certa" não é trivial. Outra desvantagem desse modelo é o elevado custo computacional quando se adiciona um novo documento na coleção, pois é preciso refazer os cálculos algébricos e aproximações para atingir uma nova estrutura semântica. Apesar disso, existem métodos para realizar essas atualizações descritas em BERRY *et al.* [9], O'BRIEN [10].

2.1.5 Modelo de Alta Correlação (MAC)

O Modelo de Alta Correlação (MAC), criado pela COPPE em XEXEO *et al.* [6], se baseia em um meta-modelo para recuperação de informação, chamado Modelo de Sinais para Busca e Recuperação de Informação (MSBRI), também criado pela COPPE em DA SILVA [11]. É considerado um meta-modelo porque define um esqueleto de um processo para BRI baseada em processamento de sinais, sem forçar a utilização de nenhuma codificação do sinal (*framework*), ou função de classificação em particular. O esquema do processo proposto pelo MSBRI é ilustrado na figura 2.3.

Seguindo o meta-modelo, o MAC aplica uma função de codificação de termo (FCT) e uma função de codificação de documento (FCD), que permita a representação do documento em forma de um sinal de termos. De outra forma, o MAC

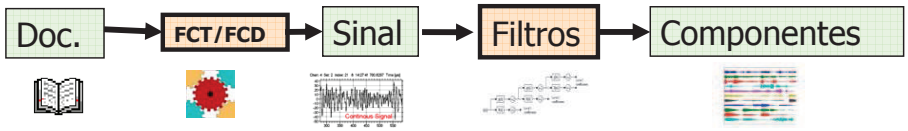


Figura 2.3: Esquema do meta-modelo MSBRI (Fonte: DA SILVA [11], Cap. 5)

representa os documentos em sinais na forma de vetores contendo uma codificação dos termos. A particularidade deste modelo está em considerar uma característica comum dos sinais, em que eles tendem a ter componentes altamente correlacionados no domínio temporal ou espacial. Isso quer dizer que um sinal pode indicar que pontos próximos geralmente apresentam atributos parecidos, e pode ainda indicar que pontos próximos, mas com atributos divergentes, acusam uma borda no sinal.

Bom, mas como transformar um texto em um sinal que tenha uma alta correlação entre pontos adjacentes? Para isso, a proposta do modelo é que cada termo deve ser associado a um índice de forma a estabelecer uma ordem parcial, em que dois termos com índices adjacentes, k_i e k_{i+1}, tenham uma correlação ρ, tal que, a correlação entre esses dois termos seja maior que a do termo de índice menor com seu antecessor. Dessa forma, a FCT é definida pela expressão 2.19:

$$k_i < k_{i+1} \quad \leftrightarrow \quad \rho(k_{i-1}, k_i) > \rho(k_i, k_{i+1}), \quad \forall i | 0 < i < t \tag{2.19}$$

em que

- k_i e k_{i+1} são dois termos com índices adjacentes,
- i é a posição do índice referente a um termo,
- ρ é a correlação entre dois termos.

A partir dessa expressão, nota-se que é necessário definir qual será o termo base k_0, e como será calculada a correlação entre dois termos. O MAC atribui ao termo base, o termo com menor IDF, já que este é o mais comumente encontrado em todos os documentos. A correlação entre dois termos é definida pelo elemento, de acordo com o posicionamento prévio entre os termos, que compõe a matriz resultante da multiplicação da matriz de incidência termo-documento por sua transposta.

Essa FCT pode ser interpretada como a "energia de ligação" entre dois termos da coleção. Uma vez definido o termo base, atribui-se à próxima posição no índice, o termo que possui a maior energia de ligação com o anterior, e assim sucessivamente, até que todos os termos sejam posicionados no índice. E a FCD definida para o MAC apenas concatena em um vetor as FCT de cada termo na ordem em que estes estão dispostos no documento. E assim, a codificação do documento define o sinal original de cada documento.

Por fim, uma alternativa para a função de similaridade do MAC é o uso da transformada de Fourier, e esta função é definida pela integral do módulo da diferença entre as transformadas de Fourier (indicada pelo símbolo $\mathcal{F}$) de um documento d_j e uma consulta q. A Fórmula 2.20 formaliza essa definição:

$$sim(d_j, q) = \sum |\mathcal{F}(FCD(d_j)) - \mathcal{F}(FCD(q))| \tag{2.20}$$

Esse modelo tem um fator complicador na comparação entre os documentos e a consulta quando se utiliza a transformada de Fourier, devido a dimensão do vetor de saída do FCD ser proporcional ao tamanho do documento. Essas limitações da transformada de Fourier serão evidenciadas no Capítulo 3, quando serão apresentadas as propriedades e aplicações de uma outra ferramenta matemática, as *wavelets*.

2.1.6 Avaliação de Sistemas de Recuperação da Informação

O desenvolvimento de novos métodos de recuperação da informação tem se demonstrado altamente empírico. Além disso, um sistema de RI deve atender não só às necessidades de um único usuário, mas sim de um grupo de usuários. Por essas questões, torna-se bastante importante uma completa e cuidadosa avaliação da eficácia e superioridade do desempenho dessas novas técnicas diante de outras em uma coleção representativa de documentos MANNING *et al.* [7].

A avaliação de sistemas de RI consiste em associar uma métrica quantitativa aos resultados produzidos pelo sistema. Tal métrica pode ser diretamente associada com a relevância dos resultados obtidos, e, normalmente, o cálculo dessa métrica se dá pela comparação entre os resultados obtidos do sistema com os resultados sugeridos pelos usuários para um mesmo conjunto de consultas BAEZA-YATES e RIBEIRO-NETO [8]. Dessa forma, a avaliação de sistemas de RI avalia o quão relevante foi o conjunto de documentos retornados pelo sistema para a necessidade de informação do usuário. Importante observar que a relevância de um documento é avaliada em relação a uma necessidade de informação, e não a uma consulta.

Segundo MANNING *et al.* [7], BAEZA-YATES e RIBEIRO-NETO [8], para mensurar a eficácia de um sistema de RI, é necessária uma coleção de teste que contenha ao menos os três pontos seguintes:

1. Uma coleção de documentos (conjunto D).

2. Um grupo de descrições de necessidades de informação para testes (conjunto Q), expressáveis como consultas.

3. Um conjunto de opiniões de relevância, geralmente uma avaliação binária classificando os documentos como relevantes ou não-relevantes para cada consulta

contida no grupo de testes. Ou seja, um conjunto de opiniões de relevância referentes a cada par consulta-documento $[q_i, d_j]$, $q_i \in Q$ e $d_j \in D$, em que os valores atribuídos a cada par pode ser 1 se o documento d_j é relevante para q_i, ou 0 caso contrário.

As opiniões de relevância devem ser produzidas por especialistas humanos. Além disso, tanto a coleção de documentos, como o grupo de necessidades de informação, devem ter um tamanho considerável para que se possa chegar a uma avaliação com resultados representativos MANNING *et al.* [7]. Para este efeito, várias bases de testes para sistemas de RI foram e continuam sendo criadas e incrementadas. Essas bases serão discutidas no tópico seguinte (Seção 2.1.7).

Duas medidas mais frequentemente utilizadas para avaliação de sistemas de RI são as métricas de precisão e abrangência (*precision* e *recall*). Elas são definidas para o simples caso em que um sistema de RI retorna um conjunto de documentos para uma consulta. A seguir são apresentados os conceitos dessas duas medidas, segundo MANNING *et al.* [7], BAEZA-YATES e RIBEIRO-NETO [8]:

- **Precisão** (*Precision*) é a fração dos documentos recuperados que são relevantes.
- **Abrangência** (*Recall*) é a fração dos documentos relevantes que foram recuperados.

A partir de uma informação q_i requisitada, e considerando os conjuntos:

- D: o conjunto de todos os documentos existentes na coleção,
- R: o conjunto de documentos relevantes para q_i,
- A: o conjunto de documentos recuperados pelo sistema de RI para q_i,
- $R \cap A$: a interseção dos conjuntos R e A, ou seja, o conjunto de documentos recuperados pelo sistema que são efetivamente relevantes.

A Figura 2.4 ilustra esses conjuntos envolvidos nos cálculos das medidas, utilizando-se do diagrama de *Venn-Euler*.

Dessa forma, seguem as definições matemáticas das medidas, dadas pelas Fórmulas 2.21 e 2.22:

$$Precision(\mathcal{P}) = \frac{|R \cap A|}{|A|} \tag{2.21}$$

$$Recall(\mathcal{R}) = \frac{|R \cap A|}{|R|} \tag{2.22}$$

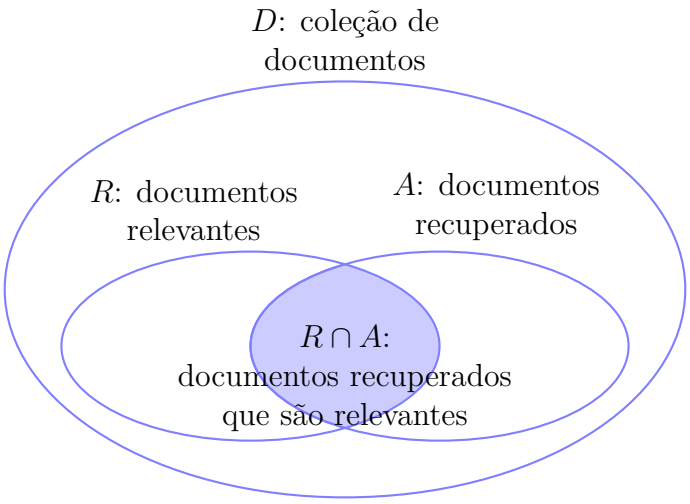


Figura 2.4: Diagrama de *Venn-Euler*

Tabela 2.1: Relação de Contingência

	Relevante	**Não-Relevante**
Recuperado	verdadeiro-positivo (vp)	falso-positivo (fp)
Não-Recuperado	falso-negativo (fn)	verdadeiro-negativo (vn)

De outra forma, essas relações podem ficar mais claras pela tabela de contingência a seguir:

Então, a partir da Tabela 2.1, as medidas ainda podem ser expressas pelas Fórmulas 2.23 e 2.24:

$$Precision(\mathcal{P}) = \frac{vp}{vp + fp} \tag{2.23}$$

$$Recall(\mathcal{R}) = \frac{vp}{vp + fn} \tag{2.24}$$

Observando isto, pode-se pensar que seria óbvio julgar sistemas de RI por uma medida de fidelidade (*accuracy*) de seus resultados, ou seja, a fração das classificações verdadeiras obtidas. De acordo com a tabela de contingência acima, essa medida seria dada por: $accuracy = (vp + vn)/(vp + fp + fn + vn)$. Apesar desta medida ser frequentemente utilizada em problemas de classificação por aprendizagem de máquina, segundo MANNING *et al.* [7] essa medida não é adequada para sistemas de recuperação da informação. O motivo se dá principalmente pela enorme distorção dos dados devido à alta porcentagem dos documentos não-relevantes da coleção, provocando uma concentração da taxa de falso-positivos, diferente das medidas de precisão e abrangência que concentram a avaliação sobre os verdadeiro-positivos.

Ainda segundo MANNING *et al.* [7], é vantajoso ter as duas medidas (precisão e abrangência), pois dependendo das cirscunstâncias uma pode ser mais importante

que a outra. Mas de maneira geral, em um bom sistema de RI, a abrangência não é decrescente em função do número de documentos recuperados, mas a precisão geralmente diminui à medida que cresce a quantidade de documentos recuperados. Dessa forma, pretende-se obter alguma quantidade de abrangência enquanto for tolerável a taxa de falso-positivos.

Em MANNING *et al.* [7], BAEZA-YATES e RIBEIRO-NETO [8] citam outra medida que combina as duas medidas citadas anteriormente, a média harmônica ponderada da precisão e abrangência, conhecida como *F-measure*. A Fórmula 2.25 define esta medida:

$$F = \frac{1}{\frac{\alpha}{\mathcal{P}} + \frac{1-\alpha}{\mathcal{R}}} = \frac{(\beta^2+1)\mathcal{P}\mathcal{R}}{\beta^2\mathcal{P}+\mathcal{R}} \quad em\ que \quad \beta^2 = \frac{1-\alpha}{\alpha} \tag{2.25}$$

onde:

- F (*F-measure*) é a média harmônica ponderada da precisão e abrangência,
- $\mathcal{P}$ é o valor da medida de precisão,
- $\mathcal{R}$ é o valor da medida de abrangência,
- α é o fator de balanceamento entre $\mathcal{P}$ e $\mathcal{R}$, em que $\alpha \in [0,1]$ e $\alpha \in \mathbb{R}$,
- β também é outro fator de balanceamento entre $\mathcal{P}$ e $\mathcal{R}$, em que $\beta \in [0,\infty]$ e $\beta \in \mathbb{R}$.

Agora, fazendo $\beta = 1$ (ou $\alpha = 1/2$), tem-se a medida reescrita na Fórmula 2.26. Geralmente, a indicação da *F-measure* com $\beta = 1$ é dada por $F_{\beta=1}$, ou simplesmente F_1.

$$F_{\beta=1} = \frac{2\mathcal{P}\mathcal{R}}{\mathcal{P}+\mathcal{R}} \tag{2.26}$$

Até o momento, foram vistas as medidas de avaliação de sistemas de RI sem levar em consideração a ordenação que possa existir entre os elementos do conjunto de documentos recuperados. Ao considerar este fato, percebe-se que os primeiros documentos recuperados de maneira ordenada pelo sistema de RI tem mais impacto para o usuário do sistema, já que na prática o usuário só examina os primeiros resultados e na ordem em que aparecem em um sistema de busca.

Ao percorrer os documentos recuperados de acordo com a ordenação em que estão dispostos, a taxa de abrangência é crescente. E, ao contrário da abrangência, a taxa de precisão, que inicalmente é alta, geralmente toma valores cada vez menores à medida em que se faz a varredura nos documentos ordenados. Ao final, para uma determinada consulta realizada, se obtém uma curva entre precisão e abrangência, como exemplificada no traçado em azul da Figura 2.5.

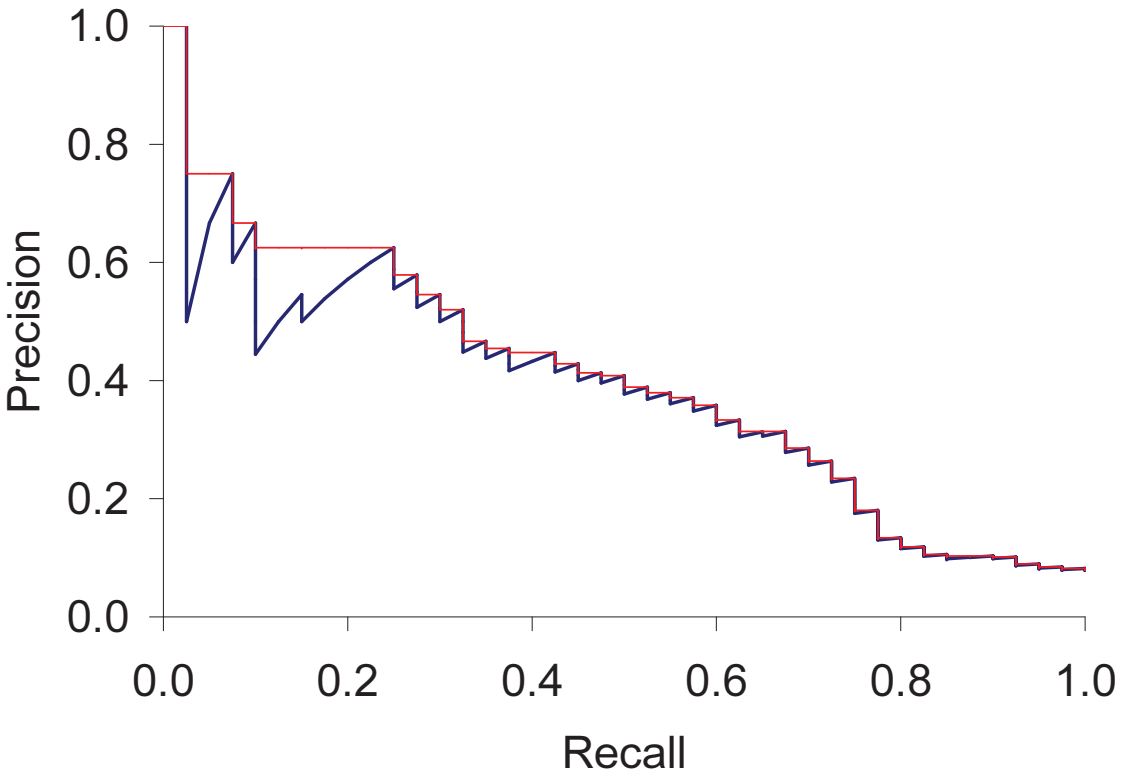


Figura 2.5: Precisão x Abrangência (Fonte: MANNING *et al.* [7], Cap. 8)

Percebe-se que, para uma única consulta realizada, o gráfico acima apresenta trêmulações na precisão entre os níveis de abrangência, devido à alternância entre documentos relevantes e não-relevantes encontrados no decorrer da varredura feita sobre os documentos recuperados a partir de uma determinada consulta. Entretanto, constuma-se remover essas tremulações pela interpolação da precisão entre os níveis de abrangência, resultando em uma precisão interpolada (traçado em vermelho da Figura 2.5) MANNING *et al.* [7].

Sabe-se que $\mathcal{P} \in \mathbb{R}$, $\mathcal{R} \in \mathbb{R}$, $\mathcal{P} \in [0,1]$ e $\mathcal{R} \in [0,1]$. E dividindo o intervalo de valores da abrangência em 11 níveis, cada nível denotado por r_j ($j \in \{0,1,2,\ldots,10\}$, correspondentes aos valores 0.0, 0.1, 0.2, ..., 1.0), define-se matematicamente a precisão interpolada para um nível de abrangência r_j na Fórmula 2.27.

$$\mathcal{P}_{interp}(r_j) = \max_{\forall r | r_j \leq r} \mathcal{P}(r) \tag{2.27}$$

Agora, tendo as precisões interpoladas de cada nível para todas as necessidades de informação existentes em uma base de testes, seria inviável avaliar a eficácia de um sistema de RI, ainda que visualmente pelos gráficos das medidas. Dessa forma, segundo MANNING *et al.* [7], BAEZA-YATES e RIBEIRO-NETO [8], torna-se necessário avaliar o esses sistemas pela média das precisões interpoladas em cada nível de abrangência. Essa média é definida por $\overline{\mathcal{P}}$ em 2.28.

$$\overline{\mathcal{P}}(r_j) = \sum_{i=1}^{|Q|} \frac{\mathcal{P}_i(r_j)}{|Q|} \tag{2.28}$$

onde:

- $\overline{P}(r_j)$ é a média da precisão interpolada no nível de abrangência r_j,
- $P_i(r_j)$ é a precisão interpolada no nível de abrangência r_j para a i-ésima necessidade de informação,
- $|Q|$ é a quantidade de necessidades de informação existentes na base de testes.

Assim, obtem-se um gráfico com uma curva decrescente, como na Figura 2.6. Gráficos como esse podem ser sobrepostos a fim de se comparar a eficácia entre sistemas de RI distintos.

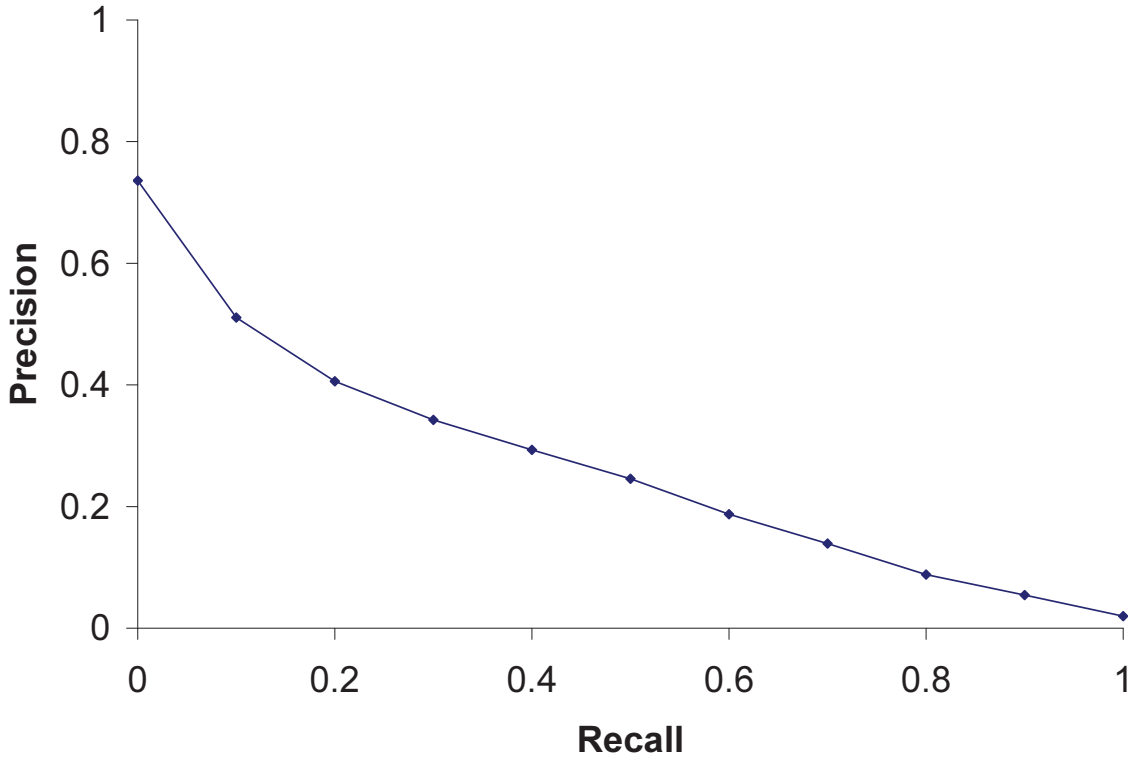


Figura 2.6: Precisão x Abrangência - Média da Precisão em 11 níveis de Abrangência (Fonte: MANNING *et al.* [7], Cap. 8)

Apesar das medidas de precisão e abrangência serem bastante utilizadas para a avaliação de sistemas de RI, elas apresentam alguns pontos que necessitam ser considerados sempre que forem utilizadas em algumas cirscunstâncias, como quando o sistema requer uma ordenação fraca dos documentos recuperados, ou quando uma medida de valor único poderia ser mais apropriada, afirma BAEZA-YATES e RIBEIRO-NETO [8]. Além disso, existem outras medidas para tal avaliação, como a *Mean Avarage Precision*, *precision at K*, *R-precision*, *Break-Even Point*,

ROC curve, *sensitivity*, *specificity*, dentre outras discutidas em MANNING *et al.* [7], BAEZA-YATES e RIBEIRO-NETO [8].

2.1.7 Bases de Testes para Avaliação de Sistemas de RI

Neste tópico serão explanadas de maneira breve algumas das bases mais utilizadas para avaliação de sistemas de RI, com uma atenção especial à TREC que será utilizada nos experimentos para os propósitos deste trabalho. Os dados e características de cada base apresentada foram extraídos de MANNING *et al.* [7], BAEZA-YATES e RIBEIRO-NETO [8]

A primeira coleção de testes para RI criada e bastante referenciada foi a *Cranfield*. Esta coleção surgiu como resultado dos primeiros experimentos em RI com um alto grau de precisão nas medidas extraídas desses estudos. Foi criada nos anos de 1950, contendo 1398 resumos de artigos e um conjunto de 225 consultas, além de um exaustivo julgamento de relavância para todos os pares de consulta e documento. Para os padrões atuais, esta base é considerada pequena, mas apropriada em primeiros experimentos de um sistema de RI.

TREC (*Text Retrieval Conference*) é uma base de teste de referência no campo de RI, e possui vários conjuntos de experimentos. Cada conjunto de experimento, ou cada coleção (como geralmente são chamados esses conjuntos) compõe-se de três partes, assim como ocorre na maioria das bases de testes de RI: os documentos (cerca de 1.89 milhões), os tópicos (450 exemplos de informações requisitadas) e um conjunto de documentos relevantes para cada tópico. A cada ano essa base de testes torna-se maior, já que a cada conferência TREC realizada mais dados são agregados.

A coleção de testes NTCIR (*NII Test Collection for IR Systems*) é composta principalmente por artigos de patentes em Inglês e Japonês. Suas coleções de testes têm tamanhos similares às coleções da TREC, e são focadas em línguas do Leste asiático, bem como em recuperação da informação entre idiomas. Outra base que é voltada para RI entre idiomas é a CLEF (*Cross Language Evaluation Forum*), mas se concentra em idiomas da Europa.

Há outras bases de tamanhos consideráveis, dentre elas tem-se a GOV2 (também referenciada como TREC-15) e INEX. Já bases de tamanhos pequenos são mais numerosas: CF (*Cystic Fibrosis*), ADI, CACM, ISI, CRAN, LISA, MED, NLM, NPL e TIME.

Para finalizar esta seção, destaca-se um ponto importante a considerar quando se avalia um sistema de RI de acordo com o tamanho da base de testes utilizada. Para bases de testes grandes é praticamente inviável avaliar todos os documentos recuperados para uma dada necessidade de informação. Então, de acordo com VOORHEES e HARMAN [12], uma alternativa é considerar somente os k (geralmente

$k = 100$) primeiros documentos recuperados pelo algoritmos de ranking. Esse método é chamado de *pooling method*, e funciona para coleções de referência com alguns milhares de documentos, como a TREC.

2.2 Classificação e Agrupamento de Informação Textual

Nesta seção serão apresentadas as técnicas de mineração de texto, de classificação e agrupamento, que serão empregadas para os fins deste trabalho. Serão discutidos seus propósitos, o embasamento teórico e seus algoritmos, trazendo apenas as questões que possam contribuir para o entendimento desta dissertação. Alguns algoritmos serão apresentados, mas o maior enfoque estará sobre o KNN (classificação) e o *K-Means* (agrupamento). Juntamente com essas informações, também serão levantadas as formas de avaliar cada uma das técnicas, bem como as bases de testes recomendadas pela literatura para se realizar as avaliações de mineração textual.

Técnicas de mineração da informação textual[1] têm como propósito geral organizar os documentos de uma coleção, como o acervo de uma biblioteca, em grupos ou classes, para uma posterior busca por documentos relacionados a um determinado assunto. Mesmo que estes documentos estejam rotulados, dependendo do tamanho do acervo, pode não ser fácil encontrar um documento que trata de um determinado assunto se eles não estiverem organizados em classes de acordo com os tópicos ou assuntos. Os nomes dados a essas classes são conhecidos como rótulos (*labels*).

2.2.1 Classificação de Informação Textual

A Classificação associa ou classifica um item a uma ou várias classes categóricas pré-definidas. Geralmente os algoritmos utilizados nessa técnica envolvem a descrição gráfica ou algébrica das características diferenciais das observações de vários documentos, além da classificação das observações em uma ou mais classes pré-determinadas. De maneira geral consiste em, a partir da observação de vários documentos de uma coleção de treinamento, em que cada documento deve ser previamente e manualmente associado a uma ou mais classes rotuladas (algumas vezes chamadas de tópicos), detectar um padrão que possa ser usado para classificar outro documento ainda desconhecido a alguma classe já rotulada MANNING *et al.* [7], BAEZA-YATES e RIBEIRO-NETO [8], HAN e KAMBER [13].

Seguindo a esta breve explanação sobre o problema da classificação de informação textual, tem-se a definição formal do mesmo MANNING *et al.* [7], BAEZA-YATES e RIBEIRO-NETO [8]:

[1] Algumas vezes genericamente referenciadas pela literatura como classificação de texto.

- $\mathbf{D} = \{d_1, d_2, \ldots, d_m\}$ é a coleção de documentos com suas respectivas representações,
- $\mathbf{C} = \{c_1, c_2, \ldots, c_l\}$ é o conjunto das l classes (categorias ou tópicos) com seus respectivos rótulos,
- $\mathcal{F} : \mathbf{D} \rightarrow \mathbf{C}$ é a função de classificação que mapeia os documentos às classes, na qual a existência do par $[d_j, c_p]$, $d_j \in \mathbf{D}$ e $c_p \in \mathbf{C}$, indica que o documento d_j é um membro da classe c_p.

E como a classificação é uma técninca de aprendizagem supervisionada MANNING *et al.* [7], BAEZA-YATES e RIBEIRO-NETO [8], ou seja, como depende de um conjunto de treinamento para que a função de classificação possa ser aplicada a partir de um padrão detectado previamente, ainda tem-se:

- $\mathbf{D_t} \subset \mathbf{D}$ é conjunto de documentos de treinamento com suas respectivas representações,
- $\mathcal{T} : \mathbf{D_t} \rightarrow \mathbf{C}$ é a função do conjunto de treinamento, em que a existência do par $[d_j, c_p]$, $d_j \in \mathbf{D_t}$ e $c_p \in \mathbf{C}$, indica que d_j foi atribuído manualmente à classe c_p.

Dessa forma, por ser uma técnica de aprendizagem supervisionada, a formação do classificador primeiramente se passa pela fase de treinamento, em que se utiliza do conjunto de treinamento formado por documentos com classes atribuídas por especialistas humanos . E, em seguida, o classificiador passa pela fase de avaliação, a qual faz uso de um conjunto de testes. Esse conjunto de testes também é composto por documentos com a classificação conhecida de acordo com a rotulação dada por um especialista humano, mas não há interseção entre o conjunto de treinamento e o de testes. A avaliação é feita em dois passos, em que primeiro utiliza-se do classificador já treinado para atribuir classes aos documentos do conjunto de testes, e em seguida essas classes atribuídas a cada documento são comparadas com as classes dadas pelo conjunto de testes. E assim, a técnica de classificação desenvolvida pode ser aperfeiçoada e novamente treinada e validada, repetindo os processos acima até se chegar a resultados que sejam satisfatórios. Só então, o classificador pode ser usado para classificar novos documentos desconhecidos MANNING *et al.* [7], BAEZA-YATES e RIBEIRO-NETO [8], HAN e KAMBER [13].

Em seguida, será visto um classificador bastante conhecido que pode ser aplicado à informação textual.

KNN

O KNN (*K-Nearest Neighbor*) é considerado um classificador sob-demanda (*on-demand*), ou classificador preguiçoso (*lazy classifier*), pois enquanto é alimentado

pelo conjunto de treinamento, este algoritmo apenas o memoriza. E a classificação propriamente dita só é feita a cada caso de teste submetido, ou seja, nenhum modelo de classificação é construído durante o treinamento, a execução inerente ao algoritmo só é feita a cada novo documento apresentado BAEZA-YATES e RIBEIRO-NETO [8], HAN e KAMBER [13].

Como o próximo nome sugere, este algoritmo se baseia nas classes dos k vizinhos mais próximos de um novo documento d_j apresentado. O KNN procura determinar os k ($k \in \mathbb{N}$) vizinhos mais próximos de um documento d_j ($d_j \notin \mathbf{D_t}$) dentro do conjunto de treinamento $\mathbf{D_t}$, e usa as classes desses vizinhos encontrados para determinar a classe de d_j BAEZA-YATES e RIBEIRO-NETO [8].

Considerando N_k um subconjunto de $\mathbf{D_t}$ ($N_k \subset \mathbf{D_t}$) contendo os k vizinhos mais próximos a d_j, determinados pela função de similaridade ($sim(d_a, d_b)$) entre dois documentos, e S_{c_p} a pontuação que d_j recebe em relação à classe c_p. O KNN é apresentado no Algoritmo 1.

Algoritmo 1: Algoritmo de Classificação KNN.

Entrada:

- $\mathbf{C}$: conjunto de classes;
- $\mathbf{D_t}$: coleção de treinamento;
- k: quantidade de vizinhos mais próximos;
- d_j: documento a ser classificado.

Saída: classe atribuída ao documento d_j

1 **início**
2 $\quad N_k \leftarrow VizinhosPróximos(\mathbf{D_t}, k, d_j)$;
3 $\quad$ **para cada** $c_p \in \mathbf{C}$ **faça**
4 $\quad\quad S(c_p) \leftarrow \sum_{d_t \in N_k, \exists [d_t, c_p]} sim(d_j, d_t)$;
5 $\quad$ **retorna** $\arg\max_{c_p} S(c_p)$;

Como a classificação presume que cada elemento classificado seja descrito por seus atributos, para o caso da classificação de documentos de texto pode-se considerar os termos como os seus atributos. Da mesma forma, pode-se adotar a representação vetorial do documento, já definida na Seção 2.1.2, como a representação de cada documento, em que os pesos são dados pelo TF-IDF (ver Fórmula 2.6). Além disso, como no KNN a similaridade deve indicar o quão próximos estão dois documentos, a distância do cosseno (definida nas Fórmulas 2.2 e 2.3) pode ser empregada como a medida de similaridade deste algoritmo. Entretanto, ao se aplicar o KNN, geralmente utiliza-se a distância Euclidiana para definir a função de similaridade MANNING *et al.* [7], BAEZA-YATES e RIBEIRO-NETO [8], HAN e KAMBER

[13]. Essa distância é definida na Fórmula 2.29.

$$sim(d_j, d_t) = \sqrt{\sum_{i=1}^{t}(w_{i,j} - w_{i,t})^2} \tag{2.29}$$

A principal desvantagem do KNN é o desempenho, já que para cada documento submetido à classificação é necessário calcular as distâncias com todos os documentos do conjunto de treinamento. Outra questão é a escolha do valor de k, que geralmente é obtido experimentalmente repetindo a execução do algoritmo sobre a coleção de testes com a variação incremental de k, normalmente iniciando com $k = 1$ HAN e KAMBER [13].

2.2.2 Agrupamento de Informação Textual

O Agrupamento, ou análise de *cluster* como preferem alguns autores, associa um item a uma ou várias classes categóricas (ou *clusters*), em que as classes são determinadas pelos dados, diversamente da classificação em que as classes são pré-definidas. Os *clusters* são definidos por meio do agrupamento de dados baseados em medidas de similaridade ou modelos probabilísticos.

A análise de *cluster* (ou agrupamento) é uma técnica que visa detectar a existência de diferentes grupos dentro de um determinado conjunto de dados e, em caso de sua existência, determinar quais são eles. Nesse tipo de análise, o procedimento inicia com o cálculo das distâncias entre os objetos estudados dentro do espaço multiplano constituído por eixos de todas as medidas realizadas (variáveis), sendo, a seguir, os objetos agrupados conforme a proximidade entre eles. Na sequência, efetuam-se os agrupamentos por proximidade geométrica, o que permite o reconhecimento dos passos de agrupamento para a correta identificação de grupos dentro do coleção de documentos estudada MANNING *et al.* [7], BAEZA-YATES e RIBEIRO-NETO [8], HAN e KAMBER [13].

Segue a definição formal do problema de agrupamento de informação textual MANNING *et al.* [7], BAEZA-YATES e RIBEIRO-NETO [8]:

- $\mathbf{D} = \{d_1, d_2, \ldots, d_m\}$ é a coleção de documentos com suas respectivas representações,
- K é quantidade de agupamentos (*clusters*) desejado,
- $\mathcal{F} : \mathbf{D} \rightarrow \{G_1, G_1, \ldots, G_K\}$ é a função de classificação que mapeia os documentos aos agrupamentos a serem encontrados, na qual a existência do par $[d_j, G_k]$, $d_j \in \mathbf{D}$ e $G_k \in \{G_1, G_2, \ldots, G_K\}$, indica que o documento d_j é um membro do agrupamento G_k, em que $G_k \subset \mathbf{D}$.

Diferente da classificação, o agrupamento é uma técnica de aprendizagem não-supervisionada em que não se faz uso de nenhuma coleção de treinamento ou de testes, ou seja, não há a participação de especialistas humanos para atribuir classes aos documentos BAEZA-YATES e RIBEIRO-NETO [8], HAN e KAMBER [13]. A seguir, será visto um algoritmo de agrupamento comumente aplicado à informação textual.

K-Means

O *K-Means* é um algoritmo baseado no conceito de centróide, ou centro de massa, e proporciona o particionamento do conjunto de m documentos em K partições, em que K é o número de agrupamentos que se pretende atingir. O centróide de um *cluster* é um documento representativo do centro de massa deste agrupamento.

Primeiramente, são escolhidos de maneira arbitrária, K documentos como os centróides iniciais de cada *cluster*, algumas vezes referenciados como sementes pela literatura, e em seguida cada um dos outros documentos são vinculados aos *clusters* que contêm o centróide mais próximo, de acordo com uma dada função de similaridade ($sim(d_a, d_b)$) entre dois documentos. Uma vez definido os *clusters* iniciais com seus documentos, os centróides são recalculados e os documentos são revinculados ao centródide mais próximo de cada documento, e assim, esse processo se repete até que nenhum centróide seja alterado MANNING *et al.* [7], BAEZA-YATES e RIBEIRO-NETO [8], HAN e KAMBER [13].

O cálculo do centróide $\vec{\mu}$ é dado pela média dos valores atribuídos às características de cada documento pertencente ao *cluster* G, de acordo com a Fórmula 2.30 que presume a representação vetorial dos documentos (ver Seção 2.1.2) MANNING *et al.* [7].

$$\vec{\mu}(G) = \frac{1}{|G|} \sum_{\vec{d} \in G} \vec{d} \tag{2.30}$$

A distância do cosseno (ver Fórmulas 2.2 e 2.3) pode ser empregada no cálculo da similaridade do *K-Means*, mas, assim como no algoritmo KNN, geralmente aplica-se à similaridade do corrente algoritmo a distância Euclidiana (ver Fórmula 2.29).

Quanto à seleção dos documentos sementes, que geralmente é feita de modo randômico, existem variações da maneira de realizar esta seleção com o objetivo de otimizar o algoritmo, como apresentadas em MANNING *et al.* [7], BAEZA-YATES e RIBEIRO-NETO [8], HAN e KAMBER [13].

Com isso, pode-se apresentar o *K-Means* formalmente no Algoritmo 2.

Algoritmo 2: Algoritmo de Agrupamento *K-Means*.

Entrada:

- K: quantidade de *clusters*;
- **D**: coleção de documentos.

Saída: Um conjunto de K *clusters*.

```
1  início
2      (s⃗_1, s⃗_2, ..., s⃗_K) ← SelecionarSementes(K, D);
3      para k ← 1 até K faça
4          μ⃗_k ← s⃗_k;
5      repita
6          para k ← 1 até K faça
7              G_k ← ∅;
8          para j ← 1 até |D| faça
9              p ← arg max_p' sim(μ⃗_p', d⃗_j);
10             G_p ← G_p ∪ {d⃗_j};
11         para k ← 1 até K faça
12             μ⃗_k ← (1/|G_k|) Σ_{d⃗∈G_k} d⃗;
13     até que não haja alteração em {μ⃗_1, ..., μ⃗_K};
14     retorna {μ⃗_1, ..., μ⃗_K};
```

Para encontrar o melhor valor para K no *K-Means*, utiliza-se uma medida que representa o quão bem os membros de um cluster são representados pelo seu centróide, conhecida como soma dos quadrados de resíduos ou RSS (*Residual Sum of Squares*)) dada pela Fórmula 2.31. Fazendo uso dessa medida, procura-se minimizá-la configurando o valor de K, como afirmam MANNING *et al.* [7], HAN e KAMBER [13].

$$RSS = \sum_{k=1}^{K} \sum_{\vec{d} \in G_k} |\vec{d} - \vec{\mu}(G_k)|^2 \tag{2.31}$$

Dessa forma, o *K-Means* é um algoritmo que despende a maior parte do tempo calculando as distâncias, e dependendo do conjunto de documentos, pode ser muito custoso encontrar o melhor valor para K.

2.2.3 Outros Algoritmos de Mineração de Informação Textual

Nesta seção, serão apresentados em linhas gerais outros algoritmos de mineração de dados que podem ser aplicados à informação textual. Lista-se: Classificação Bayesiana Simplificada, Agrupamento Hierárquico, Classificador SVM e Redes Neu-

rais. Existem muitos outros algoritmos para mineração de texto, cada um com uma abordagem diferente ou com variações nos procedimentos adotados.

Classificação Bayesiana Simplificada

Este classificador se baseia na teoria da probabilidade e no teorema de Bayes. Em classificadores probabilísticos, atribui-se à existência de cada par documento-classe $[d_j, c_p]$ uma probabilidade $P(c_p|\vec{d_j})$, dada pela Fórmula 2.32:

$$P(c_p|\vec{d_j}) = \frac{P(c_p) \times P(\vec{d_j}|c_p)}{P(\vec{d_j})} \tag{2.32}$$

em que

- $P(\vec{d_j})$ é a probabilidade do documento $\vec{d_j}$ ser selecionado arbitrariamente,
- $P(c_p)$ é a probabilidade de um documento da classe c_p ser selecionado arbitrariamente,

Uma vez realizado o treinamento com este classificador, levantando a distribuição de probabilidade do conjunto de treinamento, essa probabilidade é aplicada ao novo documento para cada classe, e a classificação é escolhida pela classe que recebeu o maior valor. Há muitas variações deste algoritmo como BAEZA-YATES e RIBEIRO-NETO [8] explica.

Agrupamento Hierárquico

O agrupamento hierárquico tem como objetivo criar uma hierarquia de *clusters* pela decomposição de um *cluster* grande em menores, ou, pela aglomeração de *clusters* pré-definidos formando *clusters* maiores.

Em BAEZA-YATES e RIBEIRO-NETO [8] é apresentado um algoritmo genérico do agrupamento hierárquico, cuja a idéia é a partir de M documentos, formar M *clusters* no passo inicial, cada qual contendo apenas um documento. Em seguida, agrega-se os dois *clusters* mais próximos de acordo com a similaridade entre eles, para formar um *cluster*, e assim esse procedimento ocorre sucessivamente até que se obtenha um único *cluster*.

A distância entre dois *clusters* depende dos valores das similaridades ou distâncias entre os documentos ($sim(d_j, d_l)$) que compões os *clusters*. E, baseado nessa premissa, de acordo com BAEZA-YATES e RIBEIRO-NETO [8] costuma-se usar um dos três seguintes métodos para se calcular a distância entre dois *clusters*, dada por $sim(G_p, G_r)$:

- *Single-Link*

$$sim(G_p, G_r) = \min_{\forall d_j \in G_p, d_l \in G_r} sim(d_j, d_l) \tag{2.33}$$

- *Complete-Link*

$$sim(G_p, G_r) = \max_{\forall d_j \in G_p, d_l \in G_r} sim(d_j, d_l) \tag{2.34}$$

- *Average-Link*

$$sim(G_p, G_r) = \frac{1}{M_p + M_r} \sum_{d_j \in G_p} \sum_{d_l \in G_r} sim(d_j, d_l) \tag{2.35}$$

Classificador SVM

O classificador SVM (*Support Vector Machine*) é um método que emprega o espaço vetorial para problemas de classificação binária. A partir de uma coleção de documentos representados em um espaço t-dimensional, o método constrói um superfície (ou hiperplano) de decisão nesse espaço que melhor separa os documentos em duas classes. E, dessa forma, quando um novo documento é submetido ao classificador, este posiciona o documento no hiperplano de acordo com sua representação vetorial, possibilitando a decisão binária de acordo com o lado da superficie em que foi posicionado.

Este método é definido de maneira que maximize soma das distâncias dos dois documentos mais próximos do hiperplano, sendo que estes dois documentos não pertencem à mesma classe. A soma destas distâncias é definida como margem. Dessa forma, tem-se as seguintes definições:

- $\mathcal{H}_w$ um hiperplano que separa os documentos nas classes c_a e c_b,
- m_a a distância do documento pertencente à classe c_a mais próximo a $\mathcal{H}_w$,
- m_b a distância do documento pertencente à classe c_b mais próximo a $\mathcal{H}_w$,
- $m_a + m + b$ a margem m do SVM, tal que o hiperplano de decisão $\mathcal{H}_w$ maximiza o valor de m.

Mesmo sendo um método de classificação binária, ele pode ser empregado para problemas com múltiplas classes, através da redução de um problema como este para uma classificação binária. Isso pode ser feito criando múltiplos problemas de classificação binária, cada um relativo a uma classe. Assim, para classificar um novo documento, aplica-se a classificação deste documento para cada classe. Em seguida, o resultado das margens de cada classe é comparado com o resultado de todas as outras, e finalmente o documento é atribuído à classe que apresentou os maiores valores de margem nas comparações.

Redes Neurais

O campo das redes neurais foi originalmente aberto por psicólogos e neurobiólogos, que procuraram desenvolver e testar métodos computacionais análogos aos neurônios. Como afirma HAN e KAMBER [13], uma rede neural é um conjunto de unidades de entrada e saída conectados em que cada conexão tem um peso associado. Durante a fase de aprendizagem da rede, ajusta-se os pesos de modo a serem capazes de prever corretamente o rótulo da classe de cada documento submetido.

As redes neurais têm várias propriedades que tornam seu uso popular em agrupamento, devido a basicamente três motivos, segundo HAN e KAMBER [13]:

1. as redes neurais são compostas por uma arquitetura inerente ao processamento paralelo e distribuído;

2. as redes neurais aprendem ajustando os pesos as suas conexões, de modo a melhor ajustar os dados, e isto permite prototipar os padrões e tornar-se apropriado para extrair atributos para os vários agrupamentos.

3. as redes neurais processam vetores numéricos e exigem padrões de objeto que possam ser representado por apenas características quantitativas. Assim, se o dado original não for numérico, é preciso transformá-lo em uma representação com atributos quantitativos.

A abordagem das redes neurais para o agrupamento procura representar cada *cluster* como um exemplar, o qual tem a função de protótipo do *cluster* e não é necessário que seja a instância de um documento. Assim, novos documentos submetidos podem ser atribuídos ao *cluster* cujo exemplar seja o mais similar, baseado em alguma medida de distância.

As redes neurais requerem um longo tempo de treinamento e, portanto, são mais adequadas para aplicações onde isso é viável. Elas exigem uma série de parâmetros (como a topologia da rede) cuja melhor forma de determinar é empiricamente. E, além disso, as redes neurais geralmente são criticadas por sua fraca interpretabilidade, devido aos seus significados que, além de "ocultos", são de difícil percepção e entendimento por um humano. Apesar dessas desvantagens, de acordo com HAN e KAMBER [13] as redes neurais tem uma elevada tolerância a ruídos presentes nos dados, assim como possuem uma elevada capacidade de classificar padrões para os quais a rede não foi treinada.

Há muitos tipos diferentes de redes neurais e algoritmos de rede neural. O algoritmo mais popular da rede neural é o *backpropagation*. E os *self-organizing feature maps* são um dos mais populares métodos de redes neurais para análise de *cluster*.

2.2.4 Avaliação de Algoritmos de Classificação e Agrupamento de Informação Textual

Análogamente à recuperação da informação, os algoritmos desenvolvidos para tarefas de classificação e agrupamento necessitam de uma avaliação de sua eficácia e desempenho. Para isso, muitas métricas foram desenvolvidas, e nesta seção serão apresentadas as formas de avaliação mais difundidas na literatura.

Antes de prosseguir com a apresentação dessas métricas, ressalta-se que, de modo geral, por uma questão instrínseca às duas técnicas abordadas, as métricas que são utilizadas para avaliar a classificação, também podem ser aplicadas ao agrupamento com poucas adaptações, partindo do pressuposto que se tenha uma base de testes para tal efeito, já categorizada ou classificada por especialistas humanos.

Avaliação para Classificação

Para se apresentar qualquer medida para avaliação de algoritmos de classificação de texto faz-se necessário expor primeiramente os conjuntos e as variáveis envolvidas na avaliação e como se relacionam em uma tabela de contigência. Desssa forma, sendo:

- $\mathbf{D}$ a coleção de documentos,
- $\mathbf{D_t}$ a coleção dos documentos já classificados por especialistas humanos (tanto os documentos de treinamento como os de testes),
- $\mathbf{C} = \{c_1, c_2, \ldots, c_l\}$ é o conjunto de todas as l classes,
- $\mathcal{T} : \mathbf{D_t} \rightarrow \mathbf{C}$ a função do conjunto de documentos já classificados por especialistas humanos,
- $\mathcal{F} : \mathbf{D} \rightarrow \mathbf{C}$ a função de classificação de texto.

Ao submeter a coleção $\mathbf{D_t}$ ao classificador, chega-se à seguinte tabela de contingência 2.2:

Tabela 2.2: Relação de Contingência para Classificação

	$\mathcal{T}(d_j) = c_p$	$\mathcal{T}(d_j) \neq c_p$
$\mathcal{F}(d_j) = c_p$	verdadeiro-positivo (vp)	falso-positivo (fp)
$\mathcal{F}(d_j) \neq c_p$	falso-negativo (fn)	verdadeiro-negativo (vn)

Por meio dessa tabela pode-se calcular duas das métricas que são comumente usadas para avaliar classificadores, a exatidão (*accuracy*) e o erro (*error*), relativas a uma dada classe c_p, que estão definidas pelas Fórmulas 2.36 e 2.37, respectivamente.

$$Acc(c_p) = \frac{vp + vn}{vp + fp + fn + vn} \tag{2.36}$$

$$Err(c_p) = \frac{fp + fn}{vp + fp + fn + vn} \tag{2.37}$$

E ainda, observa-se que $Acc(c_p) + Err(c_p) = 1$. Em outras palavras, a exatidão é a taxa de documentos classificados corretamente pelo classificador, enquanto o erro é a taxa de documentos classificados de maneira errada pelo classificador HAN e KAMBER [13].

Apesar destas medidas serem comumente usadas, elas podem apresentar desvantagens em determinadas situações, pois evidenciam taxas que não são representativas da realidade, como afirma BAEZA-YATES e RIBEIRO-NETO [8].

Outras duas medidas bastante empregadas nessas avaliações são variações das mencionadas para recuperação de informação, a precisão (*precision*) e a abrangência (*recall*). E estão respectivamente definidas nas Fórmulas 2.38 e 2.39 com base na tabela de contingência para determinada classe c_p.

$$Precision \quad \mathcal{P}(c_p) = \frac{vp}{vp + fp} \tag{2.38}$$

$$Recall \quad \mathcal{R}(c_p) = \frac{vp}{vp + fn} \tag{2.39}$$

Assim, interpreta-se que a precisão é a fração de todos os documentos atribuídos à classe c_p pelo classificador que realmente pertencem à classe c_p, e que a abrangência é a fração de todos os documentos que pertencem à classe c_p que foram corretamente atribuídos a essa classe pelo classificador BAEZA-YATES e RIBEIRO-NETO [8]. E como essas medidas são aplicadas a cada classe existente na coleção, dependendo da quantidade de classes e da quantidade de documentos, pode-se chegar a uma enorme quantitade de valores para essas medidas, a serem tratados e interpretados. A partir disso, surge uma outra medida que combina a precisão e abrangência, consequentemente variação de uma medida também existente para RI, a *F-measure*. Para uma única classe essa medida é dada pela Fórmula 2.40:

$$F_1(c_p) = \frac{2\mathcal{P}(c_p)\mathcal{R}(c_p)}{\mathcal{P}(c_p) + \mathcal{R}(c_p)} \tag{2.40}$$

Mas para derivar a uma única medida de *F-measure* a partir das medidas individuais de cada classe, é preciso aplicar alguma função de média a elas. Em MANNING *et al.* [7], BAEZA-YATES e RIBEIRO-NETO [8] são mencionadas duas medidas derivadas para todas as classes: a macro-média F_1 ($macF_1$) e a micro-média F_1 ($micF_1$). Essas medidas estão definidas nas Fórmulas 2.41 e 2.42.

$$macF_1 = \frac{\sum_{p=1}^{|\mathbf{C}|} F_1(c_p)}{|\mathbf{C}|} \tag{2.41}$$

$$micF_1 = \frac{2\mathcal{P}\mathcal{R}}{\mathcal{P}+\mathcal{R}} \tag{2.42}$$

Para o cálculo de $micF_1$, a obtenção de $\mathcal{P}$ e $\mathcal{R}$ são dadas pelas Fórmulas 2.43 e 2.44, respectivamente.

$$\mathcal{P} = \frac{\sum_{c_p \in \mathbf{C}} vp}{\sum_{c_p \in \mathbf{C}} (vp + fp)} \tag{2.43}$$

$$\mathcal{R} = \frac{\sum_{c_p \in \mathbf{C}} vp}{\sum_{c_p \in \mathbf{C}} (vp + fn)} \tag{2.44}$$

A micro-média F_1 trata cada documento com a mesma importância. Já a macro-média F_1 trata cada categoria com a mesma importância, e ainda percebe o quão hábil é o classificador quando se tem muitas classes. Além disso, BAEZA-YATES e RIBEIRO-NETO [8] afirma que quando há distorção na distribuição das classes, ambas as medidas devem ser consideradas.

Avaliação para Agrupamento

Todas as métricas citadas acima para a classificação também podem ser aplicadas ao agrupamento de texto. Entretanto, é necessário uma adaptação que antecede ao cálculo da medida, que consiste em associar um *cluster* à classe cujo rótulo é mais frequente no *cluster* a ser atribuído.

Em MANNING *et al.* [7] é levantada uma simples e transparente medida de avaliação, a pureza (*purity*). Esta medida representa o quanto são verdadeiros os *clusters* é em relação às categorias que foram atribuídos. A pureza é definida pela Fórmula 2.45:

$$purity(\{G_1, G_1, \ldots, G_K\}, \mathbf{C}) = \frac{1}{|\mathbf{D}|} \sum_{k=1}^{K} \max_j |G_k \cap c_j| \tag{2.45}$$

onde

- $\mathbf{D}$ a coleção dos documentos submetidos ao agrupamento,
- $\mathbf{C} = \{c_1, c_2, \ldots, c_l\}$ é o conjunto de todos as l classes,
- $\{G_1, G_1, \ldots, G_K\}$ é o conjunto dos K *clusters* formados pelo agrupamento.

Partindo dessa medida, MANNING *et al.* [7] ainda explana sobre outras medidas conhecidas, como a *normalized mutual information* e a entropia.

2.2.5 Bases de Testes para Avaliação de Algoritmos de Mineração de Textos

Para se realizar a avaliação dos algoritmos de mineração de textos, além dos mecanismos de avaliação, fazem-se necessárias as bases de testes para tais avaliações. Nesta seção serão citadas algumas coleções populares, cujas informações foram extraídas de MANNING *et al.* [7], BAEZA-YATES e RIBEIRO-NETO [8], incluindo a Reuters-21578 que será utilizada neste trabalho.

A coleção de testes conhecida como Reuters-21578 é a coleção de referência mais largamente usada. Sendo a mesma constituída de artigos de notícias da Reuters referentes ao ano de 1987, os quais são classificados em diversas categorias relacionadas à economia. Quanto aos número, a base contem cerca de 9.603 documentos para treinamento e 3.299 para testes, com 90 categorias que ocorrem em ambos. As categorias são atribuídas desde 1,88% até 20,96% do conjunto de treinamento, enquanto que no conjunto de testes essas taxas variam de 1,7% a 32,95%.

A agência de notícias Reuters também criou outras coleções mais gigantes, conhecidas como RCV1 e a RCV2, sendo esta última uma versão corrigida e modificada da RCV1. Estas coleções contém aproximadamente 800.000 documentos, organizados em 103 categorias, e assim, espera-se que, com o tempo, esta coleção substitua a Reuters-21578.

Uma outra coleção conhecida é a OSHUMED, que é um subconjunto da Medline (outra coleção de testes), e contem documentos médicos, englobando 23 classes.

Há também a coleção 20 *NewsGroups* que está entre as 3 mais usadas, contendo aproximadamente 20.000 mensagens postadas por usuário web do *newsgroups*. Como o nome indica, essas mensagens são distribuídas em 20 diferentes grupos de notícias, que são as próprias categorias.

Por fim, cita-se outras coleções de testes: WebKB, ADM-DL e ODP.

Capítulo 3

Wavelets: Conceito, Propriedades e Aplicações

Como foi revisado no Capítulo 2, existem vários modelos desenvolvidos para a Recuperação de Informação do tipo texto, alguns considerados clássicos, a saber: o Modelo Booleano, o Modelo Vetorial e o Modelo Probabilístico. Estes modelos têm por objetivo representar os documentos de algum modo que possam ser indexados e posteriormente recuperados sob determinados critérios.

Assim, estes e outros modelos foram desenvolvidos baseando-se em diferentes ferramentas matemáticas. Além disso, ao final da Seção 2.1.5 visualizou-se de maneira breve que uma ferramenta matemática possibilitou a representação de textos em sinais, a transformada de Fourier. Essa ferramenta ainda permitiu fazer uso das propriedades dos sinais em sistemas de RI do tipo texto, uma vez que já são utilizadas nos sistemas de RI multimídia.

Neste capítulo, iremos apresentar os conceitos e propriedades de uma outra ferramenta matemática que possibilitou a representação e o tratamento de texto como sinal: as *wavelets*. Além disso, serão apresentadas diversas aplicações das *wavelets*, das quais se detalharão trabalhos de aplicações em sistemas de RI multimídia e textual.

3.1 O que são *Wavelets*?

Wavelet é uma forma de onda de duração efetiva limitada que tem um valor médio de zero, geralmente expressa na literatura por uma função ψ. Um exemplo ilustrativo de *wavelet* está na figura 3.9. Definindo de maneira prática, *wavelets* são ondas pequenas com determinadas propriedades que as tornam adequadas a servirem de base para decomposição de outras funções, assim como senos e cossenos servem de base para decomposições (análise) de Fourier. De outra forma, *wavelet* é uma

ferramenta matemática para analisar, processar e sintetizar sinais onde o método de Fourier não obtém uma representação aceitável do sinal.

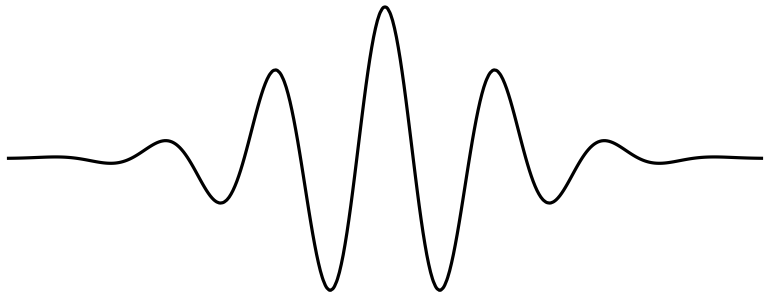


Figura 3.1: *Wavelet* ψ

3.1.1 Um Breve Histórico: da Análise de Fourier à Análise *Wavelet*

Antes de mais nada, é preciso compreender o contexto do surgimento da teoria *wavelet*, bem como suas motivações. Antes da descoberta das *wavelets*, FOURIER [14] mostrou que uma dada função $f(x)$ pode ser representada (ou aproximada) por uma combinação linear de componentes senoidais, cada um com um dado coeficiente. O conjunto $\{w_n(t) = e^{int} | n \in \mathbb{Z}, t \in \mathbb{R}\}$ de funções ortogonais, de período 2π, forma a base para a análise de Fourier. Tornou-se então bastante comum o emprego desse ferramental em vários tipos de funções, inclusive de sinais. Assim, de outra forma, Fourier mostrou que pode-se representar uma função como uma soma de infinitos senos e cossenos, cada um com um certo coeficiente. Em termos de função de um sinal no decorrer do tempo, quer dizer que a análise de Fourier é usada para representar o sinal por funções de seno e cosseno, cada um de acordo com uma dada frequência.

Dito isto, a transformada de Fourier é uma transformada integral cujo núcleo é definido pela famosa Fórmula de Euler em 3.1.

$$e^{i\omega t} = \cos(\omega t) + i\,\mathrm{sen}(\omega t) \tag{3.1}$$

Em que:

- e é a base dos logaritmos naturais,
- i é a unidade imaginária ($i = \sqrt{-1}$),
- ω pode ser interpretada como a frequência instantânea angular das senoides,
- t é uma variável independente, que pode ser interpretada como o tempo.

Em função disso, a Transformada de Fourier (FT - *Fourier Transform*) de um sinal, definida pelo operador $\mathcal{F}$, é dada pela Fórmula 3.2. O sinal é dado pela função $f(t)$.

$$\mathcal{F}(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}\mathrm{d}t \tag{3.2}$$

Em outras palavras, a transformada de Fourier decompõe o sinal no domínio da frequência. E, para se obter novamente o sinal no domínio temporal, aplica-se a transformada inversa de Fourier (síntese de Fourier) definida na Fórmula 3.3.

$$f(t) = \mathcal{F}^{-1}(\mathcal{F}(\omega)) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{i\omega t}\mathrm{d}\omega \tag{3.3}$$

A transformada de Fourier é perfeitamente aplicável em sinais estacionários, que são sinais cuja função é periódica, como a função de seno ou cosseno ilustrados na Figura 3.2.

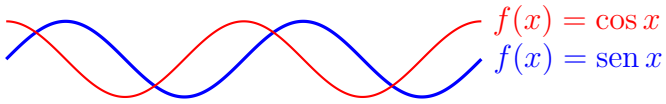


Figura 3.2: Seno e Cosseno

Para exemplificar a representação de um sinal na forma de uma onda quadrada em componentes senoidais, a Figura 3.3 mostra sucessivas aproximações pelas somas de senóides.

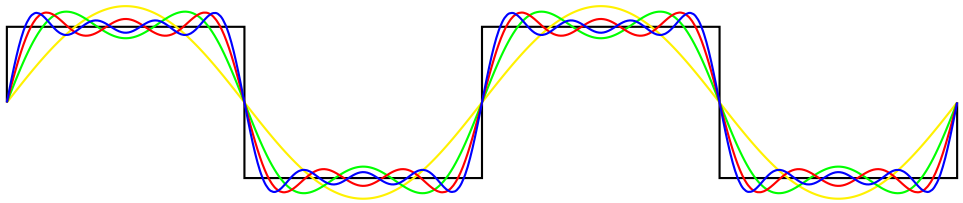
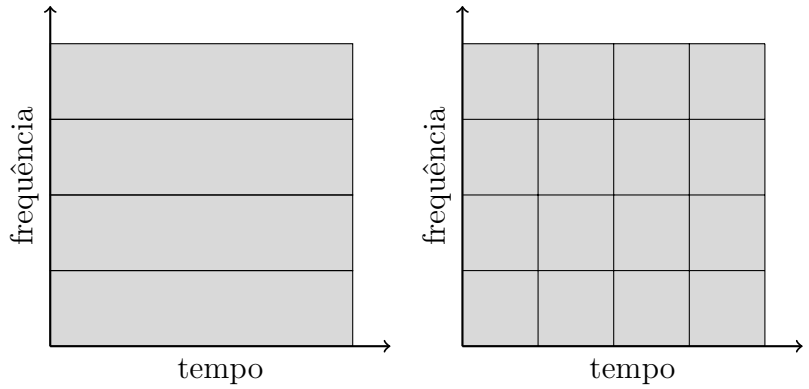


Figura 3.3: Uma onda quadrada com sucessivas aproximações pelas somas de senóides.

Por outro lado, a aplicação da análise de Fourier apresenta limitações devido ao comportamento dos sinais práticos, que em sua maioria apresentam um comportamento não-estacionário, ou seja, sinais que apresentam variações abruptas na sua frequência no decorrer do tempo. Essa limitação se dá pelo fato dessas variações abruptas, características dos sinais de função não periódica, possuirem componentes de alta frequência que interferem em toda a transformada no domínio da frequência, causando perda de informação temporal. Assim, ao se utilizar a análise de Fourier, muitas vezes recorre-se à transformada de Fourier com Janelamento (STFT - *Short-Time Fourier Transform*).

Proposta por GABOR [15], a STFT não trata o sinal como um todo, mas o separa em intervalos (ou janelas) termporais uniformes suficientemente pequenos para que as características de sinal sejam consideradas estacionárias. Dessa maneira, para uma análise localizada do sinal, o STFT é satisfatório. Entretando, ainda com o janelamento, a transformada de Fourier apresenta algumas questões não triviais. A primeira é que cada janela analisada pode apresentar um efeito de borda que é causado pelas variações que podem ocorrer nas proximidades dos limites do intervalo. Em segundo lugar, essa análise é totalmente dependente da escolha das janelas, pois, a partir do momento em que se fixa os intervalos temporais para análise, fixa-se também a resolução em uma dada frequência. E, em terceiro lugar, as funções trigonométricas possuem energia finita, ou seja, estas funções compreendem todo intervalo $-\infty$ e ∞.

Para ilustrar o tratamento tanto da FT, como da STFT sobre um sinal, a Figura 3.4 mostra a relação entre a resolução temporal e a resolução da frequência.



(a) FT (b) STFT

Figura 3.4: Relação entre a resolução temporal e da frequência da FT e da STFT.

Foi diante dessas limitações ao fazer uso da STFT para resolver problemas geofísicos que Jean Morlet percebeu a necessidade de desenvolver uma função matemática base, ψ, que deveria possuir energia finita (intervalo finito, ou seja, com início e fim) e ainda pudesse ser capaz de dilatar ou comprimir esta função, pois dessa forma seria eliminado o problema da janela fixa da STFT. Assim, Jean Morlet recorreu a um trabalho de HAAR [16], que continha a primeira família das *wavelets* sem ainda formular o conceito de *wavelets*, e juntamente com Alex Grossmann consolidaram esse conceito (referenciado primeiramente pelo termo francês *ondelettes*) em seus trabalhos GROSSMANN e MORLET [17], GOUPILLAUD *et al.* [18]. E foi nessa busca que algumas propriedades das *wavelets* foram se consolidando, como a condição de serem na forma de pequenas ondas, em que, mesmo decaindo para zero rapidamente, fosse possível aplicar translações (ou deslocamentos) à função base de modo que cobrisse todo o eixo dos reais. E, como também, fosse possível aplicar

dilatações em sua frequência para se atingir a função base, assim como já ocorre na transformada de Fourier. Essas propriedades serão discutidas na próxima seção, e em seguida será apresentada formalmente a transformada *wavelet*.

3.2 Propriedades das *Wavelets*

Formalmente, as *wavelets* são ondas cuja função ($\psi(t)$) apresentam as seguintes características:

1. a área total sob a curva da função é zero:

$$\int_{-\infty}^{\infty} \psi(t)\mathrm{d}t = 0 \tag{3.4}$$

2. a energia da função é finita:

$$\int_{-\infty}^{\infty} |\psi(t)|^2\mathrm{d}t < \infty \tag{3.5}$$

Além disso, uma família de funções *wavelets* podem ser definidas com base na *wavelet* mãe $\psi(t)$ dessa família. Segue abaixo, na Fórmula 3.6, a definição das *wavelets* filhas ($\psi_{\mathcal{E},\mathcal{D}}(t)$) que compõem uma família.

$$\psi_{\mathcal{E},\mathcal{D}}(t) = \frac{1}{\sqrt{\mathcal{E}}}\psi\left(\frac{t-\mathcal{D}}{\mathcal{E}}\right), \quad \mathcal{E},\mathcal{D} \in \mathbb{R},\ \mathcal{E} > 0 \tag{3.6}$$

Para tal efeito, uma *wavelet* mãe ψ, além de atender às condições 3.4 e 3.5, também deve atender aos seguintes critérios:

1. a função é absolutamente integrável:

$$\int_{-\infty}^{\infty} |\psi(t)|\mathrm{d}t < \infty \tag{3.7}$$

2. condição de admissibilidade, em que C_ψ é a constante de Calderón, $\mathcal{F}_\psi(\omega)$ é a transformada de Fourier de ψ, e ω é a frequência instantânea:

$$C_\psi = \int_{-\infty}^{\infty} \frac{|\mathcal{F}_\psi(\omega)|^2}{\omega}\mathrm{d}\omega < \infty \tag{3.8}$$

3.2.1 Escala e Deslocamento

Escala e deslocamento são duas propriedades para a análise *wavelet* evidenciadas pela fórmula 3.6, em que $\mathcal{E}$ é o parâmetro de escala e $\mathcal{D}$ é o parâmetro de descolamento da *wavelet*. Dessa forma, a análise *wavelet* se diferencia da análise de Fourier, pois enquanto esta decompõe o sinal em componentes caracterizados pela

frequência, as componentes da análise *wavelet* são caracterizadas pelos coeficientes de escala ($\mathcal{E}$) e deslocamento ($\mathcal{D}$). E, apesar desses componentes não serem caracterizados diretamente pela frequência ω, o parâmetro de escala $\mathcal{E}$ está relacionado a ela, de tal forma, que sinais decompostos em *wavelets* de menor escala apresentam componentes em frequências mais altas, enquanto os componentes de uma escala maior são de frequências mais baixas. Já o parâmetro de deslocamento $\mathcal{D}$ afeta o posicionamento da *wavelet* sobre o sinal em análise.

GROSSMANN e MORLET [17], GOUPILLAUD *et al.* [18] introduziu esse ajuste da escala visando garantir a isometria, ou seja, a conservação de energia de acordo com a Fórmula 3.9.

$$||\psi(t)||^2 = ||\psi_{\mathcal{E},\mathcal{D}}(t)||^2 \tag{3.9}$$

Assim, com o parâmetro de escala, pode-se comprimir uma *wavelet* tornando $\mathcal{E} < 1$, como também expandí-la tornando $\mathcal{E} > 1$. Enquanto o parâmetro de deslocamento pode deslocar para a *wavelet* para a esquerda, se $\mathcal{D} < 0$, e para a direita, se $\mathcal{D} > 0$. Para ilustar o efeito de $\mathcal{E}$ e $\mathcal{D}$ sobre uma *wavelet*, a Figura 3.5 exibe uma *wavelet* mãe juntamente com algumas de suas variações em escala e deslocamento.

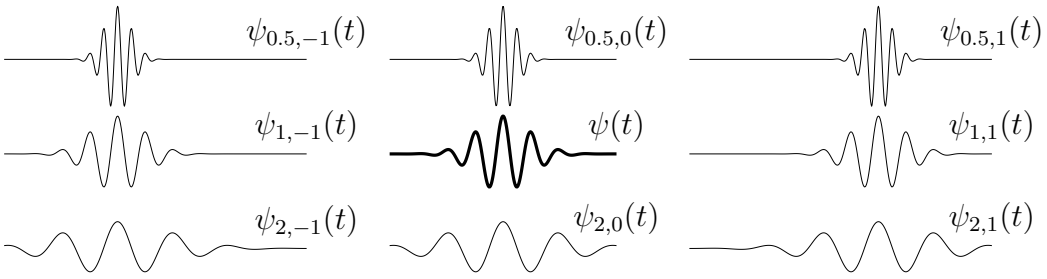


Figura 3.5: Exemplo de uma *wavelet* mãe ψ e algumas de suas variações $\psi_{\mathcal{E},\mathcal{D}}$ em escala ($\mathcal{E}$) e deslocamento ($\mathcal{D}$).

3.3 A Transformada *Wavelet*

Uma vez que o conceito de *wavelets* foi explanado, bem como suas principais propriedades, esta breve seção irá apresentar a definição da transformada *wavelet*, e complementar sua interpretação já iniciada em na seção anterior.

A Transformada *Wavelet* (WT - *Wavelet Transform*) decompõe o sinal em bases (ou núcleos) caracterizadas pelos coeficientes de escala e deslocamento, como já mencionado na Seção 3.2.1 referindo-se à expressão análise *wavelet*. Formalmente, a transformada *wavelet*, definida pelo operador Ψ, é dada pela Fórmula 3.10. De forma prática, a transformada *wavelet* defini-se pela soma sobre todo o domínio temporal do sinal multiplicado por versões escalonadas e deslocadas da função *wavelet* mãe,

ou seja, define-se pelo produto escalar entre a função do sinal no domínio temporal, $f(t)$, com a função *wavelet* ψ mãe em diferentes escalas e deslocamentos.

$$\Psi_{f(t)}(\mathcal{E},\mathcal{D}) = \langle \psi_{\mathcal{E},\mathcal{D}}(t), f(t) \rangle = \int_{-\infty}^{\infty} f(t)\psi^*_{\mathcal{E},\mathcal{D}}\mathrm{d}t \tag{3.10}$$

Em que:

- $\psi(t)$ é a função *wavelet* mãe,
- Ψ é o operador linear da transformada *wavelet*,
- $\mathcal{E}$ é o coeficiente de escala, sendo $\mathcal{E} \in \mathbb{R}^+$,
- $\mathcal{D}$ é o coeficiente de deslocamento, sendo $\mathcal{D} \in \mathbb{R}$,
- $f(t)$ é a função do sinal no domínio temporal,
- * indica o complexo conjugado.

Para evidenciar o tratamento que a WT realiza sobre o sinal, em contraste com o tratamento realizado pela STFT, a Figura 3.6 mostra a relação entre a resolução temporal e a resolução da frequência nas duas transformadas.

Assim como na transformada de Fourier, é possível resconstruir do sinal original através da transformada *wavelet* inversa a partir da WS, dada pela Fórmula 3.11.

$$f(t) = \Psi^{-1}_{f(t)}(\mathcal{E},\mathcal{D}) = \frac{1}{C_\psi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{\mathcal{E}^2}\Psi_{f(t)}(\mathcal{E},\mathcal{D})\psi_{\mathcal{E},\mathcal{D}}\mathrm{d}\mathcal{E}\mathrm{d}\mathcal{D} \tag{3.11}$$

3.4 Análise de Multi-Resolução

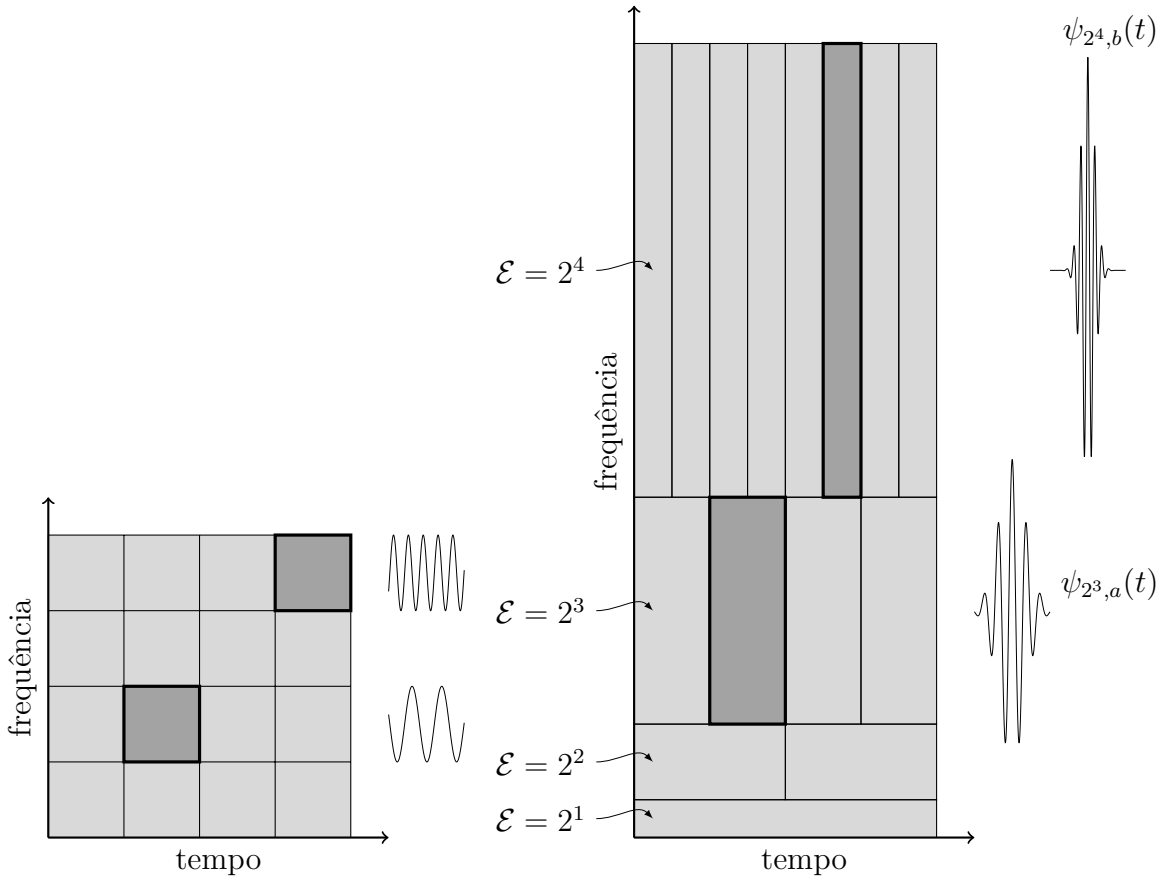
A teoria da Análise de Multi-Resolução (MRA - *Multi-Resolution Analysis*) foi introduzida por MALLAT [19], com o intuito de analisar o sinal em várias escalas de resolução, através de uma função de escala $\phi(t)$ e de *wavelets* $\psi_{\mathcal{E},\mathcal{D}}(t)$.

De maneira formal, a análise de multi-resolução no espaço de funções $L^2(\mathbb{R})$, consiste em uma sequência crescente de subespaços fechados $\{V \subset L^2(\mathbb{R}), j \in \mathbb{Z}\}$ que aproximam $L^2(\mathbb{R})$, de maneira a satisfazer as seguintes relações, enunciadas em MORETTIN [20], DE OLIVEIRA [21]:

$$V_j \subset V_{j+1}, \forall j \in \mathbb{Z} \tag{3.12}$$

$$\overline{\bigcup_{j\in\mathbb{Z}} V_j} = L^2(\mathbb{R}) \tag{3.13}$$

$$\bigcap_{j\in\mathbb{Z}} V_j = 0 \tag{3.14}$$



(a) STFT (b) WT

Figura 3.6: Relação entre a resolução temporal e da frequência da STFT e da WT.

$$f(t) \in V_j \Leftrightarrow f(2t) \in V_{j+1}, \forall j \in \mathbb{Z} \tag{3.15}$$

$$\exists \phi(t) \in L^2(\mathbb{R}) \mid \{\phi_{j,k}(t), k \in \mathbb{Z}\} \text{ é uma base ortonormal de } V_j \tag{3.16}$$

$$V_{j+1} = V_j \oplus W_j \quad | \quad W_j \perp V_j \tag{3.17}$$

O que se pretende por trás dessas relações é obter aproximações da função de sinal $f(t)$ em vários níveis de resolução j, ou seja, em várias escalas $\mathcal{E} = 2^{-j}$, em que o valor de cada escala é definido pelo inverso de potência de dois correspontente a um nível j. Além disso, cada subespaço V_j é constituído por funções aproximantes, sendo que a melhor aproximação é obtida considerando-se a projeção ortogonal de $f(t)$ sobre cada V_j. E, como afirma MORETTIN [20], o fato de $V_j \subset V_{j+1}$, indicado em 3.12, significa que ao passar do nível de resolução j para $j + 1$ (diminuição na escala de 2^{-j} para $2^{-(j+1)}$) ganha-se informação, ou seja, adiciona-se detalhes característicos do sinal original. Assim, quando se aumenta a resolução, fazendo

$j \to \infty$, a função aproximada converge para a função original do sinal, obtendo 3.13.

Em contrapartida, quando se aproxima $f(t)$ a níveis de resolução cada vez menores (aumenta-se a escala), reduz-se a informação. Em outras palavras, fazendo $j \to -\infty$ a aproximação de $f(t)$ converge para a função nula e tem-se 3.14. Já em 3.15, indica-se que o espaço V_j é obtido de V_{j+1} escalonando-se (comprimindo-se ou expandindo-se) funções aproximadas pela razão dos respectivos níveis de resolução.

A relação 3.16 refere-se à função de escala que foi apenas mencionada no início desta seção. A função de escala, também conhecida conhecida por *wavelet* pai, dada por $\phi(t)$, é definida e provada sua existência em MALLAT [19]. Algumas vezes, ela é referenciada como *wavelet* pai, pois é capaz de gerar outras *wavelets*. Caso se adicione o sinal de uma função *wavelet* ao sinal de uma função de escala, obtém-se uma nova função de escala. Essa função é a solução da Fórmula 3.18, e gera uma família ortonormal em $L^2(\mathbb{R})$ dada pela Fórmula 3.19.

$$\phi(t) = \sqrt{2} \sum_k l_k \phi(2t - k) \quad | \quad k \in \mathbb{Z} \tag{3.18}$$

$$\phi_{j,k}(t) = \sqrt{2^j} \phi(2^j t - k) \quad | \quad j, k \in \mathbb{Z} \tag{3.19}$$

Assim, a *wavelet* mãe pode ser obtida pela *wavelet* pai, ou seja, a partir função de escala ϕ, de acordo com 3.20. As Fórmulas 3.18 e 3.20, conhecidas como funções de dilatação, são as duas equações centrais da MRA, pois permitem a representação de funções *wavelets* em uma resolução j pelas funções de escala.

$$\psi(t) = \sqrt{2} \sum_k h_k \phi(2t - k) \quad | \quad k \in \mathbb{Z} \tag{3.20}$$

Por esse contexto, geralmente toma-se os valores especiais para $\mathcal{E} = 2^{-j}$ e $\mathcal{D} = k\mathcal{E}$ em 3.6, resultando na função *wavelet* reescrita na Fórmula 3.21.

$$\psi_{j,k}(t) = \sqrt{2^j} \psi(2^j t - k) \quad | \quad j, k \in \mathbb{Z} \ \wedge \ j = \log_{1/2} \mathcal{E}, \ k = \frac{\mathcal{D}}{\mathcal{E}} \tag{3.21}$$

Os fatores l_k e h_k são os coeficientes de filtros passa-baixo (*low-pass*, captam as oscilações de baixa frequência) e passa-alto (*high-pass*, captam as oscilações de alta frequência), respectivamente, e são usados para calcular a transformada *wavelet* discreta (passível de implementação). Esses coeficiente são dados por 3.22 e 3.23.

$$l_k = \sqrt{2} \int_{-\infty}^{\infty} \phi(t) \phi(2t - k) \mathrm{d}t \tag{3.22}$$

$$h_k = (-1)^k l_{1-k} = \sqrt{2} \int_{-\infty}^{\infty} \psi(t) \phi(2t - k) \mathrm{d}t \tag{3.23}$$

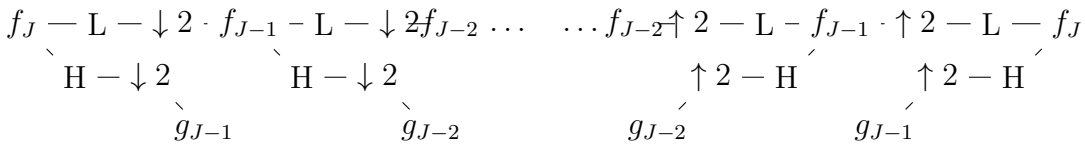
E sobre a última relação 3.17, ao representar W_j como o complemento ortogonal de V_j em V_{j+1}, indica que W_j é o espaço que representa a informação "descartada" (os detalhes) ao se reduzir de V_{j+1} para V_j. E que a soma direta dos espaços V_j em W_j reconstói o espaço de uma resolução acima. Assim, enquanto que W_j contém os detalhes da função do sinal na resolução de nível j, V_j detém a função $f_j(t)$ representativa do comportamento geral do sinal aproximado na mesma resolução.

Dessa forma, sendo uma função $f(t) \in L^2(\mathbb{R})$, para cada nível de resolução $j \in \mathbb{Z}$ existe uma função $f_j(t) \in V_j$ que melhor aproxima-se de $f(t)$ dentro do espaço V_j. E, também, para cada resolução $j \in \mathbb{Z}$ existe uma função $g_j(t)$ que corresponde aos detalhes do sinal dentro do espaço W_j. Essas funções são obtidas aplicando-se sucessivamente, para cada resolução, os filtros passa-baixa e passa-alta sobre o sinal aproximado da resolução imediatamente acima, resultando nas funções $f_j(t)$ e $g_j(t)$.

Com isso, e a partir de 3.17, para uma dada resolução j_0, é possível atingir uma melhor aproximação, dada por $f_J(t)$, do sinal original, dado pela função $f(t)$, através das somas da função $f_{j_0}(t)$ com todas as funções $g_j(t)$ em que $j \geq j_0$, como indica a Fórmula 3.24. E como são resultados dos filtros passa-baixa e passa-alta, as funções $f_j(t)$ e $g_j(t)$ são combinações lineares das funções dadas por 3.19 e 3.21, respectivamente. MALLAT [19], DAUBECHIES *et al.* [22] mostraram como realizar a decomposição *wavelet* do sinal em diferentes resoluções, como também a reconstrução do mesmo, computacionalmente.

$$f(t) \simeq f_J(t) = f_{j_0}(t) + \sum_{\forall j \geq j_0} g_j(t) \tag{3.24}$$

Para uma fácil visualização dos aspectos evidenciados da análise do sinal em *wavelets* em diferentes resoluções, a Figura 3.7 ilustra o processo da aplicação dos filtros a partir do sinal original e em cada nível resolução, e sua reconstrução. Os filtros passa alta e passa-baixa estão indicados por L e H, respectivamente. E, se há $\downarrow n$, o nível da resolução está sendo diminuído pelo fator n, enquanto que $\uparrow n$ indica que a resolução é aumentada pelo fator n.



(a) Decomposição (b) Reconstrução

Figura 3.7

A Figura 3.8 mostra um sinal, e suas transformadas *wavelets* $\Psi_{f(t)}(\in^{-|}, \mathcal{D})$ em alguns de seus níveis de resolução, cujas escalas compreendem intervalo dado por $2^{-7} \leq 2^{-j} \leq 2^{-3}$, e curva ao final é a aproximação do sinal dada pelas frequências

baixas correspondentes às escalas maiores que 2^3, ou aos níveis de resolução menores que 3.

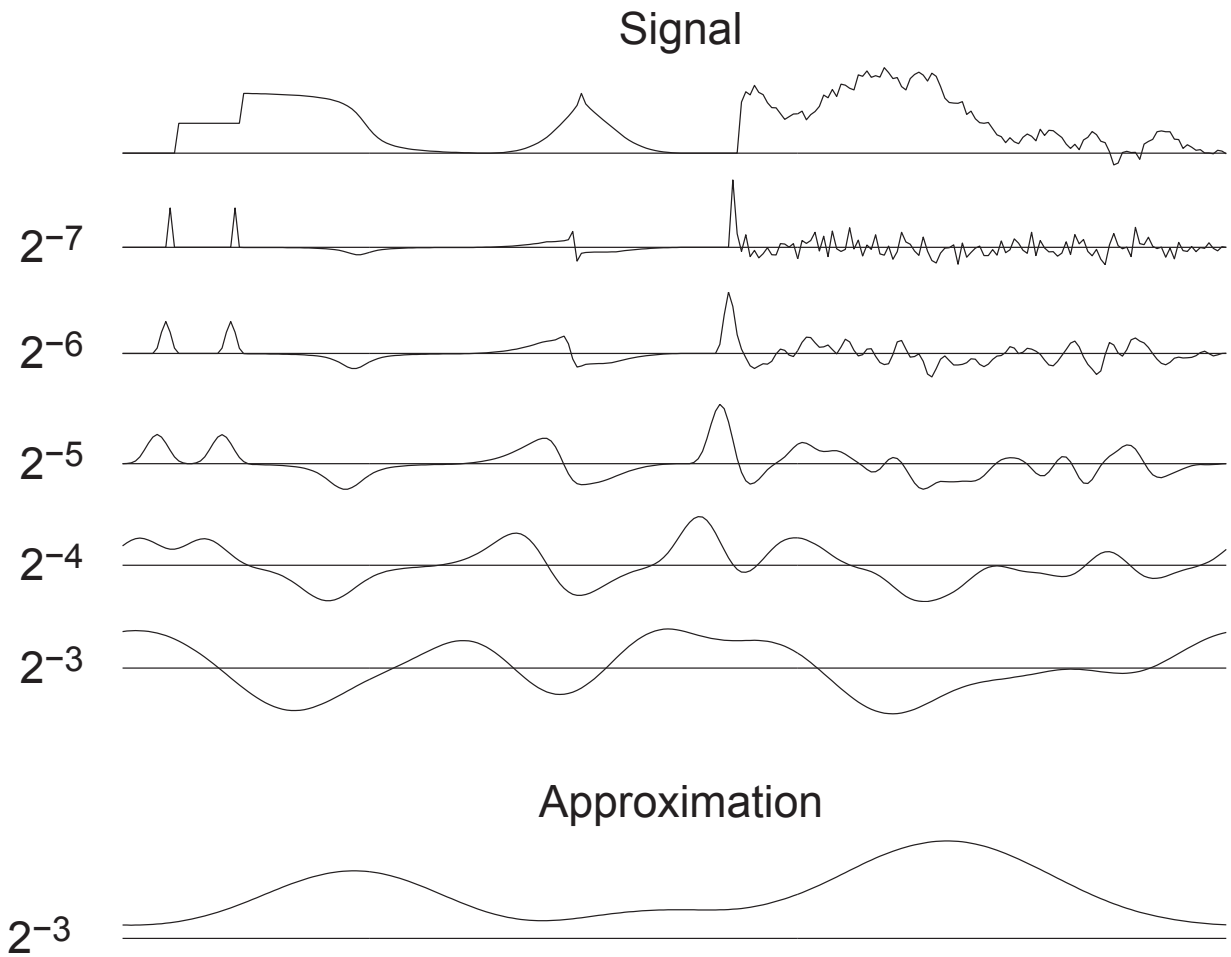
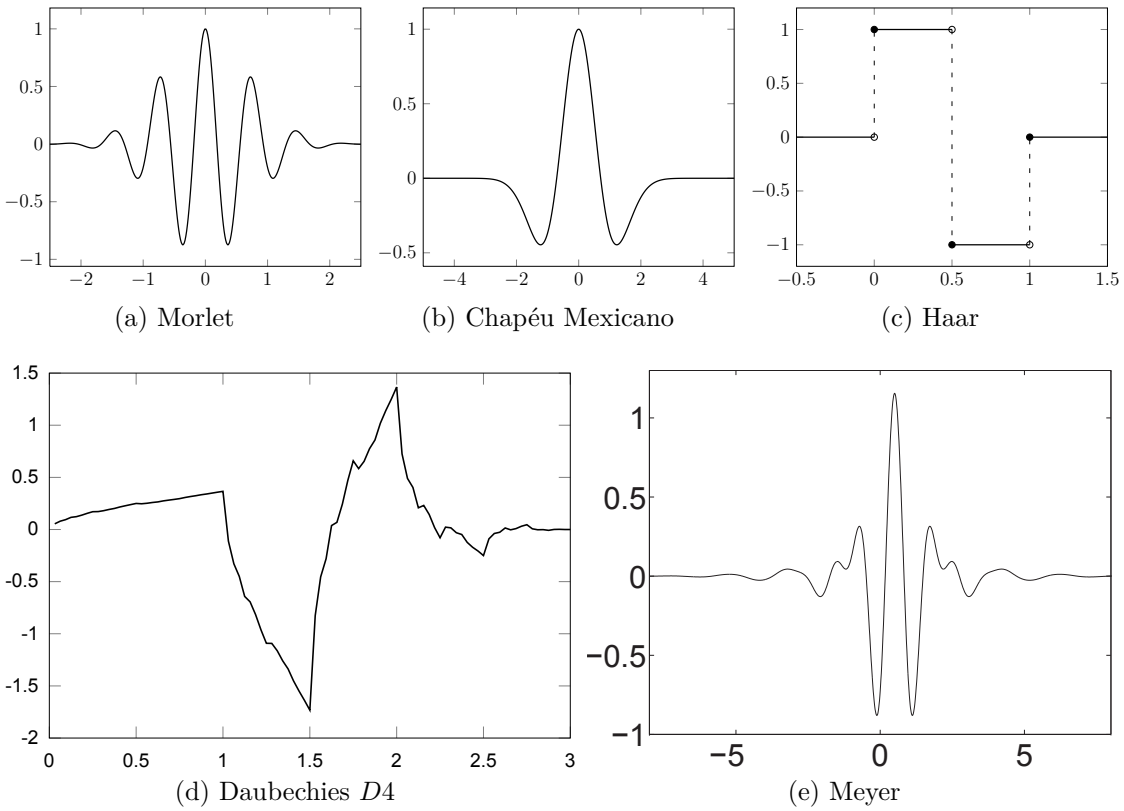


Figura 3.8: Um sinal seguido por suas transformadas *wavelets* $\Psi_{f(t)}(\in^{-|}, \mathcal{D})$ nas escalas $2^{-7} \leq 2^{-j} \leq 2^{-3}$. E a última curva é a aproximação do sinal dada pelas frequências baixas correspondentes às escalas maiores que 2^3. (Fonte: MALLAT [23], Cap. 5)

3.5 Algumas Funções *Wavelets*

Várias famílias de funções *wavelets* foram sendo contruídas com os mais diversos propósitos principalmente nas diversas áreas de engenharias que tratam do processamento de sinais. A primeira ilustração de uma *wavelet* deste capítulo (ver Figura 3.9) é um exemplo da *wavelet* de Morlet, desenvolvida por GROSSMANN e MORLET [17]. Abaixo, seguem algumas ilustrações das *wavelets* mais conhecidas.

Nesta seção, duas dessas *wavelets*, a *wavelet* de Haar e de Daubechies, serão detalhadas brevemente, e assim, proporcionar um rápido entendimento de suas implementações e sua aplicabilidade neste trabalho.



(a) Morlet (b) Chapéu Mexicano (c) Haar

(d) Daubechies $D4$ (e) Meyer

Figura 3.9: Exemplos de *Wavelet*

3.5.1 *Wavelet* de Haar

A família das *wavelets* de Haar foi formulada por HAAR [16]. A função da *wavelet* mãe é dada dela Fórmula 3.25.

$$\psi(t) = \begin{cases} 1, & \text{se } t \in [0, \frac{1}{2}[\\ -1, & \text{se } t \in [\frac{1}{2}, 1[\\ 0, & \text{caso contrário} \end{cases} \tag{3.25}$$

Como percebe-se em 3.9c, essa função representa um pulso quadrado. Essa família é a mais simples das *wavelets* que atendem à condição de admissibilidade, além de sua implementação computacional também ser simples. E foi a partir dessas condições que se optou pelo seu uso neste presente trabalho.

3.5.2 *Wavelet* de Daubechies

Em DAUBECHIES [24, 25], mostra-se que a *wavelet* de Haar é um caso particular de outro tipo de *wavelet*: a *wavelet* de Daubechies. Essas *wavelets* foram criadas por DAUBECHIES [25], e através de trabalhos como DAUBECHIES *et al.* [22], DAUBECHIES [26, 27], baseados em MALLAT [19], sua aplicação nas mais diferentes áreas de estudo foi impulsionada, pois evidenciou a relação entre a teoria *wavelet*

até então desenvolvida com o processamento de sinais.

Esse tipo de *wavelet* está dividido em casos particulares, de acordo com sua ordem (ou momento) N, em que cada caso é indicado por DN. A *wavelet* de Haar pode ser vista como um caso particular com ordem $N1$, sendo indicada por $D1$, enquato que na figura 3.9d é ilustrado o caso particular das *wavelets* de Daubechies $D4$. Esse último caso será empregado neste trabalho, pois possui uma implementação computacional factível, embora mais lenta que a implementação da *wavelet* de Haar.

3.6 Aplicações de *wavelets*

Hoje, as *wavelets* são aplicadas em diversos campos. Lista-se a maioria destes campos:

- Geologia sísmica (predição de terremotos e maremotos);
- Visão computacional e humana;
- Computação gráfica;
- Medicina e biomedicina (análise de sinais biomédicos, distinção celular, modelos para trato auditivo, análise de sequências de DNA, neurofisiologia, detecção de curtos eventos patológicos);
- Telecomunicações;
- Detecção de rupturas e bordas;
- Análise de sinais acústicos;
- Modelagem de sistemas lineares;
- Modelagem geométrica;
- Óptica e eletromagnetismo;
- Descontaminação de sinais (*denoising*);
- Processamento de imagens (codificação de imagens, compressão de imagens);
- Busca e recuperação de informação; ...

Agora, serão apresentados, de maneira breve, algumas destas áreas em que as *wavelets* são aplicadas com o processamento de sinais. A descontaminação de sinais é o processo de remover os ruídos de um sinal. Como não há fronteiras de distinção entre o sinal e o ruído, este problema se torna bastante difícil. Entretanto, as

wavelets tem se mostrado ferramentas importantes no combate e remoção de ruído. Donoho propôs uma das técnicas mais importantes de redução de ruído com base em decomposição via *wavelets*, em DONOHO [28, 29].

Quanto à codificação de imagens, usa-se para simplesmente proporcionar uma representação alternativa da imagem no domínio *wavelets*. Assim, ao invés de operar com os *pixels*, opera-se com os coeficientes *wavelets*. Em STOLLNITZ *et al.* [30] foi mostrado como alterar a representação em pixels para o domínio *wavelets*, através de um mecanismo de decomposição padrão de uma imagem, em que a transformada *wavelet* é aplicada sucessivamente nas linhas e depois sobre as colunas da imagem.

Muitas técnicas de compressão de imagens vem usando transformadas, como a transformada de cosseno discreta STOLLNITZ *et al.* [30]. E agora, a transformada *wavelet* também está sendo aplicada, como mostrado em MALLAT [19], DAUBECHIES *et al.* [22], MANDUCA [31], MULCAHY [32]. Vale ressaltar a relevância das *wavelets* como mecanismo de compressão ao ser inclusa no padrão internacional JPEG2000 CHRISTOPOULOS *et al.* [33], como também no padrão do FBI (*Federal Bureal of Investigations*) para o armazenamento de impressões digitais BRADLEY *et al.* [34].

Outros trabalhos referentes à aplicações das *wavelets* relacinadas a este tópico podem ser visto em DE OLIVEIRA e DE SOUZA [35].

3.6.1 *Wavelets* em processamento de sinais aplicados à recuperação, classificação e agrupamento da informação

No campo de RI, as *wavelets* começaram a avançar primeiramente sobre o processamento de informação multimídia, como áudio, imagens e vídeos, a exemplo de NIBLACK *et al.* [1]. O reconhecimento automático de impressões digitais também pode ser implementado com o auxílio de *wavelets*. TICO *et al.* [36] propõe combinar a imagem adquirida com imagens de um banco de dados contendo impressões digitais, baseando-se em atributos extraídos diretamente do domínio *wavelet*.

Outros trabalhos relacionados a esta seção podem ser vistos em CHANG e KUO [37], LI *et al.* [38], AL-DUBAEE e AHMAD [39], AL-DUBAEE *et al.* [40], MONTOYA ZEGARRA *et al.* [41], SUTAGUNDAR *et al.* [42]. Todos estes trabalhos fazem uso da análise de multiresolução para aproximar ou reduzir o sinal original (áudio, imagem, vídeo ou texto) devido a algum propósito de reconhecimento do padrão da informação contida no sinal.

3.6.2 *Wavelets* em processamento de texto

Mais recentemente, algumas pesquisas apontaram a viabilidade e as vantagens do uso da ferramenta *wavelet* para extrair informações relevantes de documentos tex-

tuais. Para exemplificar, *TOPIC ISLAND* MILLER *et al.* [43] aplica técnicas de processamento de sinais empregadas em textos através da transformada *wavelet* para resumir e visualizar textos.

Outro trabalho bastante relevante é o de Park em PARK *et al.* [4], que emprega a transformada *wavelet* para a realização de RI em documentos de texto. Este trabalho utiliza-se da transformada *wavelet* para indexar partições de cada documento da coleção de documentos. E, para isso, estabelece um conceito de sinal do termo (*term signal*), que nada mais é que um sinal representativo da distribuição do termo dentro da partição em questão. Entretanto, esse trabalho limita-se na escolha da quantidade de partições e no tamanho de cada partição, pois estes fatores influenciam nos sinais gerados, e assim comprometem a recuperação da informação desejada.

Em AL-DUBAEE e AHMAD [39], AL-DUBAEE *et al.* [40] também se faz uso das wavelets para a recuperação e a clusterização textual. No primeiro artigo é feito uma avaliação do impacto de perda de informação com uso de transformadas de diferentes *wavelets* mãe sobre as mesmas consultas em línguas diferentes. Enquanto que o segundo artigo faz uso da transformada *wavelet* de Haar para propôr uma nova abordagem de agrupamento dos resultados de uma consulta na *web*.

3.6.3 Uso de *wavelets* no Modelo de Alta Correlação (MAC)

XEXEO *et al.* [6], DA SILVA [11] também empregam a transformada *wavelet* para a realização de RI e classificação em documentos de texto. Mas não particiona os documentos da coleção a fim de indexá-los. Cada indexação é feita de maneira única e integral sobre cada documento da coleção. Dessa forma não se atém à escolha de valores de configurações que podem comprometer a recuperação e a classificação da informação, como em PARK *et al.* [4]. E, para que se possa realizar a indexação da maneira descrita, este trabalho propõe o Modelo de Alta Correlação (MAC), já apresentado na Seção 2.1.5, em que os sinais dos termos são organizados de forma que os parecidos fiquem em posições adjacentes, tornando o comportamento dos sinais de cada documento, como um todo, similares aos sinais de imagens. Assim, essa concentração de informação pode ser mais bem aproveitada pela transformada *wavelet*, assim como feita quando se aplica essa transformada na compactação ou indexação de imagens.

De outra forma, pode-se entender que o MAC adiciona informação para formar o sinal a ser tratado pela transformada *wavelet*, em que essa informação adicional é dada pela ordenação dos termos de acordo com a correlação entre eles. Com isso, a intrigante propriedade de multi-resolução da transformada *wavelet* MALLAT [19] pode ser empregada para reduzir a resolução do sinal, de maneira a conservar a semântica original do documento.

Dessa forma, outra alternativa para a função de similaridade do MAC, inicialmente definida na Fórmula 2.20, é o emprego da transformada *wavelet* dada pela Fórmula 3.26, em que $dist(\cdot,\cdot)$ é a distância entre dois vetores, podendo ser empregada a distância do cosseno (ver Fórmula 2.2) ou a distância Euclidiana (ver Fórmula 2.29).

$$sim(d,q) = dist(\Psi_{FCD(d)}(\mathcal{E},\mathcal{D}), \Psi_{FCD(q)}(\mathcal{E},\mathcal{D})) \tag{3.26}$$

Perceba-se que a resolução do sinal a ser usado no cálculo da similaridade pode ser configurada previamente na indexação de cada documento como sinal. E assim, para cada resolução j tem-se 2^j elementos discretos que compõe o vetor de saída da transformada. A quantidade de elementos da sinal do documento em sua resolução máxima é dada pela potência de dois imediatamente maior ou igual à quantidade de termos utilizados para representar o sinal original dos documentos. Desse modo, para que o sinal original se torne um sinal tratável pela transformada *wavelet* em diferentes resoluções, são adicionados zeros no final do sinal até que o tamanho seja uma potência de dois.

Capítulo 4

Avaliação Experimental

Este capítulo, primeiramente, apresenta os objetivos dos experimentos, e em seguida a metodologia experimental empregada, explicando as ferramentas, tecnologias envolvidas e como essas tecnologias foram aplicadas. Será detalhada também a organização dos experimentos e como foram executados. Além disso, para cada experimento, os resultados serão evidenciados bem como a análise feita sobre os mesmos.

4.1 Objetivos dos Experimentos

Em seções anteriores foram referenciados trabalhos que aplicam as *wavelets* em sistemas de RI Multimídia e mostram, comparavelmente, que o seu uso é bastante eficiente diante de outras técnicas típicas desses sistemas. Também foi mostrado que a aplicação das *wavelets* começou a atingir a área de processamento textual. No entanto, ainda se faz necessário comprovar que estas aplicações são realmente eficientes, comparando-a com as técnicas de processamento textual.

Partindo disso, cada experimento realizado objetiva permitir a avaliação do quão eficaz é o uso da transformada *wavelet*, para suas diferentes configurações de resolução, sobre as técnicas de processamento de texto revisadas no Capítulo 2, comparando-se com o algoritmo clássico ou mais comumente utilizado em cada uma dessas técnicas, nas quais também poderão variar algumas de suas configurações na medida do possível. Esse algoritmo, na sua forma original, escolhido será o *baseline* para efeitos de comparação e avaliação.

Em especial, será avaliado qual a taxa de queda (ou de possível aumento) de eficácia da aplicação da transformada *wavelet* sobre as técnicas a cada redução do nível de resolução da transformada sobre o sinal do documento. Ou seja, pretende-se averiguar qual a perda nos processos de recuperação, classificação ou agrupamento da informação, quando se reduz a quantidade de informação tratada pelo algoritmo, redução esta que é proporcionada pela análise de multiresolução da teoria *wavelet*.

4.2 Ferramental e Bases de Dados

Antes de prosseguir com a metodologia, é necessário descrever quais foram as ferramentas, tecnologias e bases de dados para testes empregados nos experimentos.

4.2.1 Ferramentas e Tecnologias

Todos os processos de experimentos foram implementados utilizando a linguagem Java. Para a recuperação da informação foi utilizada uma biblioteca de software livre e de código aberto de RI desenvolvida pela Apache, o Lucene na versão 2.9.4 (Luc [44]). Esta ferramenta implementa os algoritmos de alguns dos modelos mais empregados em RI, como os clássicos modelo booleano (ver Seção 2.1.1) e modelo vetorial (ver Seção 2.1.2). Para a técnica de RI, este trabalho irá empregar como *baseline* o algoritmo do modelo vetorial. Para aplicação da transformada *wavelet*, estendeu-se dessa biblioteca classes envolvidas no algoritmos do modelo vetorial, para que ao invés de utilizar o vetor de termos com os seus valores de TF-IDF, utiliza-se o sinal processado pela transformada em determinada resolução para a representar o documento. Foi necessário não só a extensão dessas classes, também foram feitas adaptações destas para permitir tal efeito.

Para a classificação, como também para o agrupamento, utilizou-se a Weka (*Waikato Environment for Knowledge Analisys*) na versão 3.6.5, um software de mineração de dados que também é livre e de código aberto desenvolvido pela Universidade de Waikato (Wek [45]). Entretanto, essa biblioteca não foi extendida; na verdade, o ambiente de experimento desta ferramenta foi utilizado para configurar, executar e extrair os resultados de cada algoritmo. Para isso, foram gerados arquivos que pudessem ser reconhecidos pela Weka (.arff), contendo representações vetoriais dos documentos tanto na forma original de acordo com a *baseline* – que geralmente é definido pelo vetor de termos com seus valores de TF-IDF –, como nas formas das diferentes resoluções da transformada *wavelet*.

A Weka implementa diversos algoritmos tanto de classificação como de agrupamento. Além disso, proporciona diversas medidas de avaliação para a classificação, indo muito além das discutidas na Seção 2.2.4, por outro lado, quando se refere ao agrupamento, o ambiente de experimento deixa muito a desejar. Mas ainda considera-se que foi o suficiente para extrair medidas que proporcionassem uma análise comparativa do uso de wavelets na técnica de agrupamento. O *baseline* adotado para a classificação é o algoritmo KNN (ver Seção 2.2.1). E o *baseline* escolhido para o agrupamento é o *K-Means* (ver Seção 2.2.2).

Pode-se visualizar rapidamente na Tabela 4.1 a relação de escolha do *baseline* para cara técnica empregada nos testes. Vale ressaltar que, mesmo para cada *baseline*, diferentes execuções foram feitas variando-se a configuração do algoritmo

quando possível, da mesma forma essa variação na configuração do algoritmo foi feita ao se processar a transformada do sinal em cada resolução.

Tabela 4.1: Relação de escolha do algoritmo *baseline* para cada técnica processamento textual.

Técnica	**Algoritmo *Baseline***
Recuperação da Informação	Modelo Vetorial
Classificação da Informação	KNN
Agrupamento da Informação	*K-Means*

4.2.2 Bases de Dados para Testes e seu Pré-Processamento

Todas os documentos das bases de dados utilizados no experimento deste presente trabalho foram pré-processados aplicando as fases enunciadas na Seção 2.1. Assim, para cada coleção foi montado um vocabulário de termos, além de uma matriz de incidência termo-documento.

As coleções de testes tem como origem duas fontes já discutidas nas Seções 2.1.7 e 2.2.5. Para os experimentos em RI, as coleções utilizadas foram extraídas da base CF (*Cystic Fibrosis*) e TREC (TIPSTER). Esta base contém várias coleções diferentes, e são exibidas na Tabela 4.2, juntamente com algumas de suas características, como a quantidade de termos e documentos, tamanho da matriz de incidência termo-documento, e o tamanho em MB que os documentos pré-processados da coleção ocupam na memória secundária. Além disso, a base CF, contém 50 consultas com os julgamentos de especialistas, enquanto que a base TREC possue aproximadamente 300 consultas com julgamentos para os documentos.

Tabela 4.2: Coleções que compõe a base de testes CF e TREC (TIPSTER) para os experimentos em IR.

Nº	**DataSet**	**Documentos**	**Termos**	**TxD**	**Tamanho (MB)**
1	CF	1.225	8.957	10.972.325	4,78
2	TREC-AP	32.887	86.191	2.834.563.417	162
3	TREC-DOE	2.232	11.479	25.621.128	8,71
4	TREC-FR	1.981	54.354	107.675.274	82,08
5	TREC-SJM	5.296	33.571	177.792.016	27,4
6	TREC-WSJ	18.599	67.893	1.262.741.907	119
7	TREC-ZIFF	17.036	80.168	1.365.742.048	131

Entretanto, por limitações de memória, nem todas as coleções da base de testes TREC puderam compor os experimentos. Inclusive, esse foi um motivador de esforço deste trabalho, buscando consumir o mínimo de memória primária, sem diminuir consideravelmente o tempo de execução dos experimentos. Dessa forma, quatro

coleções são empregadas nos experimentos: CF, TREC-DOE, TREC-FR e TREC-SJM.

Com relação à classificação e ao agrupamento, a coleção de documentos utilizada é a Reuters-21578. Seus atributos são apresentados na Tabela 4.3. A Reuters-21578 contém aproximadamente 118 classes, e seus documentos estão classificados dentre estas classes de tal forma que apenas 10 dessas classes concentram mais de 50% dos documentos da coleção. Além disso, para os experimentos deste trabalho, foi escolhido a porcentagem de 70% dos documentos para formar o conjunto de testes, enquanto os outros 30% formam o conjunto de treinamento.

Tabela 4.3: Coleção que compõe a base de testes Reuters-21578 para os experimentos de Classificação e Agrupamento da Informação.

Nº	**DataSet**	**Documentos**	**Termos**	**TxD**	**Tamanho (MB)**
1	Reuters-21578	21.578	32.644	704.392.232	85

4.3 Metodologia e Organização dos Experimentos

O primeiro ponto a se frisar na metodologia, além do estabelecimento de um algoritmo *baseline* para cada técnica, é quanto à forma padrão de entrada para cada *baseline*, que consiste no vetor de termos com seus valores de TF-IDF, ordenados de acordo com o grau de correlação entre os termos como definido no modelo MAC, (ver Seção 2.1.5). E, da mesma forma como o MAC refere-se ao vetor de termos ordenados desta forma, este será referenciado daqui para frente como **sinal original** do documento. Enquanto que os sinais processados pela transformada em uma dada resolução, será referenciado como **transformada do sinal na resolução** j, em que j é o nível da resolução aplicado ao cálculo da transformada.

A metodologia da avaliação experimental da aplicação da teoria *wavelet* nas técnicas apresentadas neste trabalho é dada pela seguinte ordem em cada coleção de testes:

1. Para cada técnica de processamento textual (recuperação, classificação e clusterização) seleciona-se um algoritmo *baseline*;

2. Para cada algoritmo *baseline* definido são executados um ou mais experimentos, dependendo das variações das configurações desse algoritmo;

3. Para cada experimento, aplica-se o algoritmo *baseline* sobre a coleção de testes, juntamente com a representação padrão do documento, ou seja, o sinal original de cada documento. Então, a partir dos resultados desta execução, calcula-se os valores para cada medida de avaliação deste experimento;

4. Calcula-se o nível máximo de resolução $j = J$ permitida pela transformada *wavelet* dado o sinal original dos documentos da coleção;

5. Para cada experimento, varia-se a função *wavelet* mãe, podendo ser a *wavelet* de Haar, ou a de Daubechies $D4$;

6. Para cada função *wavelet* mãe, varia-se a resolução, a partir da máxima $j = J$, reduzindo até mínima ($j = 0$);

7. Para cada nível de resolução j, aplica-se também o algoritmo *baseline* sobre a coleção de testes, utilizando-se agora, como a representação do documento, a transformada do sinal no nível de resolução atual. Da mesma forma, para cada nível resolução e a partir dos resultados da execução em um desses níveis, calcula-se os valores para cada medida de avaliação em cada nível de resolução.

A partir do processo sequencial acima, verifica-se que ainda é necessário definir melhor a etapas 2, 3 e 6. Quanto à etapa 2, necessita-se definir se há alguma configuração pertinente a ser variada para cada algoritmo *baseline*, e se houver, definir também quais variações serão feitas sobre essa configuração. Para as etapas 3 e 6, precisa-se definir quais medidas de avaliação serão calculadas para cada técnica experimentada.

Para a técnica de RI, cujo *baseline* é o modelo vetorial, não percebeu-se nenhuma configuração pertinente a ser variada para melhor comparar e avaliar o uso de *wavelet*. Embora não haja esse tipo de variação, optou-se por calcular as seguintes medidas para cada experimento: Precisão, Abrangência e *F-measure* (ver Seção 2.1.6).

Em relação à técnica de classificação da informação, com o KNN como o algoritmo *baseline*, verificou-se rapidamente a possibilidade de variar a quantidade de vizinhos mais próximos dado pelo valor de k, e se mostrou perfeitamente factível pelo ambiente de experimentos da Weka. Assim, foram executados experimentos com o KNN para o seguinte conjunto de valores $k \in \{1, 3, 5, 15, 30, 45, 60, 75, 90\}$. Já, as medidas utilizadas nesta técnica, são as mesmas escolhidas para RI, entretanto, para a classficiação calcula-se uma medida única para a Precisão, Abrangência e *F-measure* (ver Seção 2.2.4) em cada configuração do experimento, ou seja, uma valor de cada medida para cada valor de k).

E por último, com técnica de agrupamento da informação, para a qual foi escolhido o *K-Means* como o *baseline*, também percebeu-se que se pode variar a quantidade de *clusters* a serem formados pelo algoritmo, em que essa quantidade é dada por um valor K. O objetivo seria realizar essa variação dos valores $K \in \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Entretanto, a execução para cada configuração pretendida não se tornou tão trivial no ambiente de experimentação da Weka, pois

foram encontrados obstáculos operacionais, como a sobrecarga da memória e o processo incrivelmente lento. Tais fatores são intrísecos à natureza do experimento de agrupamento, e à enorme quantidade de dados que necessitam ser processados em conjunto de acordo com a implementação interna da Weka. Dessa forma serão exibidos os resultados obtidos até o momento da escrita deste trabalho. Pretendia-se também, calcular a medida da pureza (ver Seção 2.2.4) para este experimento. Mas, novamente, por uma limitação do ambiente de experimento da weka para o agrupamento, só é retornado uma medida conhecida como *log-likelihood* (MANNING *et al.* [7], HAN e KAMBER [13]). A pureza pode ser obtida na Weka, mas não pelo ambiente de experimentos automatizados, e sim pelo ambiente de exploração, cujo objetivo são tarefa pontuais. Ou seja, é um ambiente *stand-alone*, o que levaria ainda mais tempo e mais recursos computacionais para se obter todos os resultados desejados.

Além de tudo isso, após se calcular todas essas medidas de avaliação, tanto para a execução dos algoritmos com o sinal original, como para a transformada em cada nível de resolução, pretende-se ainda obter uma relação entre a redução do nível de resolução, com a possível perda ou ganho de eficácia com o uso das *wavelets*.

4.4 Resultados

Nesta seção serão apresentados e discutidos os resultados das medidas de avaliação obtidos em cada experimento. E ao final desta, será feita uma análise conclusiva destes resultados, avaliando o quão o uso da transformada *wavelet* contribui para cada uma das técnicas às quais foi aplicada.

Todos os valores numéricos dos resultados obtidos e utilizados para a representação gráfica dos mesmos, podem ser consultados no Apêndice A.

4.4.1 Experimentos em Recuperação da Informação

A seguir, os tópicos referem-se aos resultados obtidos na utilização da técninca de recuperação da informação, com seu *baseline* e as variações na resolução da transformada *wavelet*, sobre cada umas das três coleções de dados: CF, TREC-DOE e TREC-FR.

Coleção: CF

Em uma primeira observação breve dos gráficos a seguir (Figuras 4.1 e 4.2), percebe-se que o uso da transformada na resolução máxima teve praticamente o mesmo efeito do *baseline*, seja com a *wavelet* de Haar, ou com a *wavelet* de Daubechies $D4$. Além

disso, os valores para a duas *wavelet* mãe são exatamente iguais em sua resolução máxima.

E, observando mais detalhamente cada resolução na Figura 4.1 principalmente nos primeiros níveis de abrangência, nota-se que a precisão cai a medida em que se reduz o nível da resolução. Mas essa queda não é constante, como também não é proporcional à quantidade de informação reduzida ("perdida"), já que a cada descida no nível de resolução, reduz-se a quantidade de informação utilizada para se calcular a similaridade pela metade. Isso também vale para as duas *wavelets*, entretanto, a partir da resolução de nível 11, para Daubechies $D4$, e de nível 12, para Haar, a precisão começa a cair de forma mais acentuada. Isso permite perceber uma leve melhora na eficácia de Daubechies $D4$ em relação a Haar para este experimento, já que esta última começa a declinar acentuadamente mais cedo.

Como essa coleção é considerada pequena, não é plausível ter alguma conclusão consistente nesse primeiro instante. Assim, será verificado se esse comportamento também é válido para coleções mais robustas, que fazem parte da base TREC.

Coleção: TREC-DOE

Neste experimento, a primeira percepção do gráfico nas Figuras 4.3 e 4.4, um pouco diferente do anterior (ver Figura 4.1), é que a precisão da transformada *wavelet* na sua resolução máxima, apesar de tentar aproximar, não alcançou a precisão do baseline. E da mesma forma ao experimento anterior, os valores dos resultados para as transformadas nos dois tipos de *wavelet* mãe na resolução máxima foram idênticos.

Apesar disso, os primeiros níveis de abrangência no gráfico da Figura 4.3 também mostram que a queda da precisão não é proporcional à redução da quantidade de informação dada uma resolução. E neste experimento, a queda acentuada se dá aproximadamente partir dos níveis 12 e 10 das *wavelets* de Haar e de Daubechies $D4$, respectivamente. Assim, novamente a *wavelets* de Daubechies $D4$ apresenta uma melhora na eficácia com relação a Haar.

Coleção: TREC-FR

Este experimento com a técinca de IR foi realizado sobre uma base maior ainda que as duas anteriores. A quantidade de informação em cada sinal original é quase cinco vezes maior que a informação de cada sinal original do experimento anterior. Pode-se observar nas Figuras 4.5 e 4.6, que o comportamento nas resoluções máximas das transformadas é meio termo das duas observações anteriores, em que a precisão resultante das transformadas *wavelets* só alcança a precisão do algoritmo *baseline* nos últimos níveis de abrangência. E, mais uma vez, essas precisões, dada a resolução

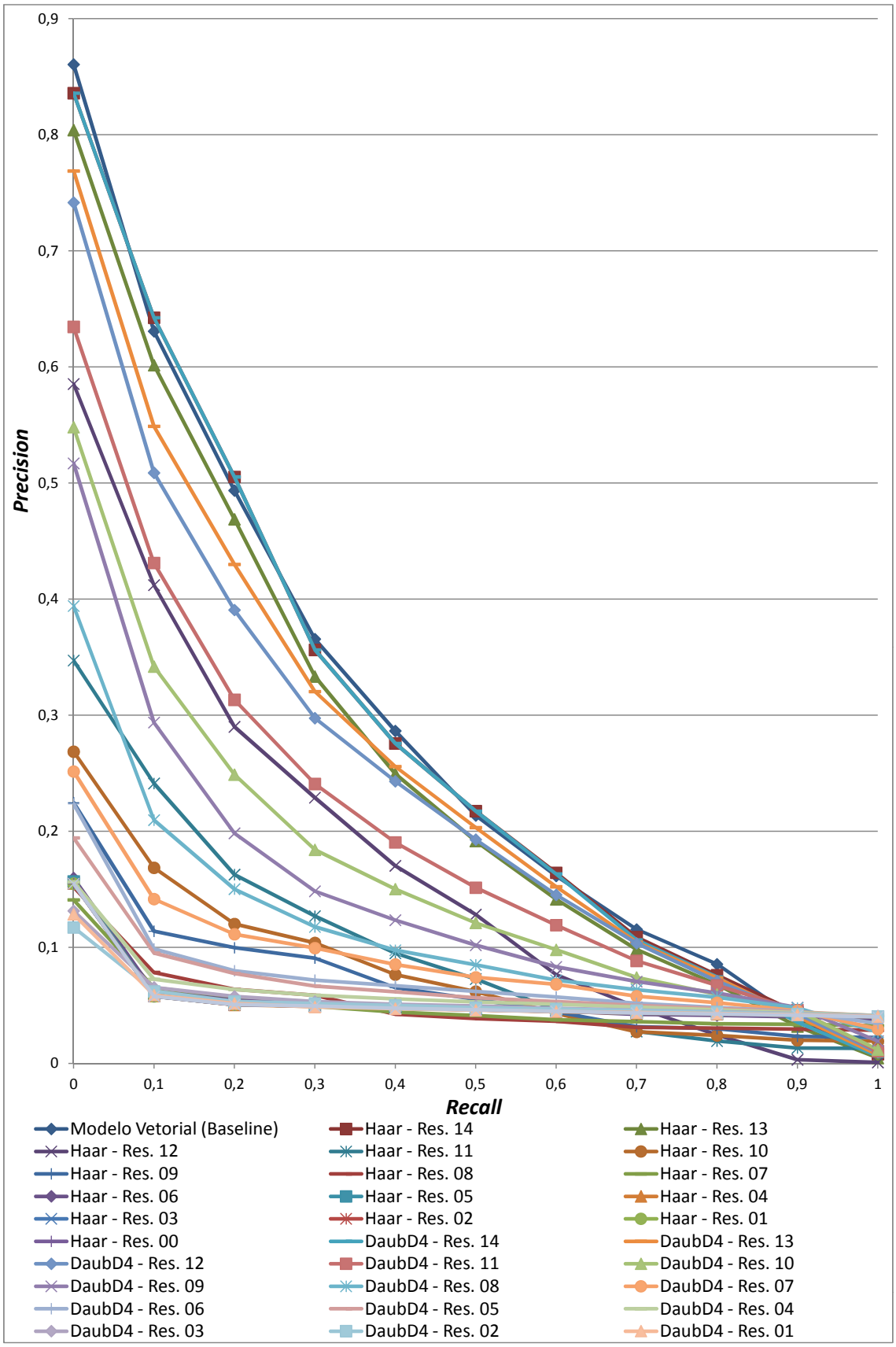


Figura 4.1: Resultados de RI (Precisão x Abrangência) sobre a coleção de documentos CF. (ver Tabela A.1)

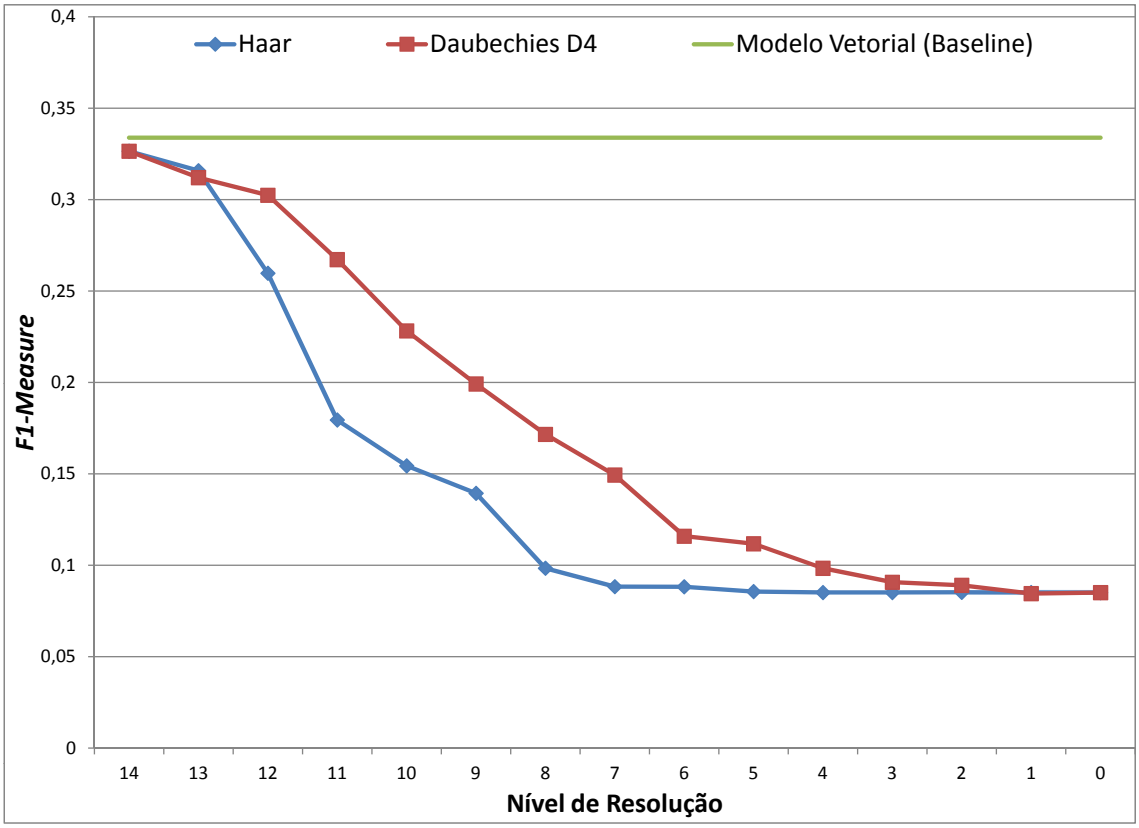


Figura 4.2: Resultados de RI (*F1-Measure* x Abrangência) sobre a coleção de documentos CF. (ver Tabela A.1)

máxima, são idênticas para as transformadas das duas *wavelets* mãe.

Da mesma forma, como nos experimentos anteriores, nota-se uma queda da precisão que não é proporcional à redução da informação em cada nível de resolução da transformada. Mas essa queda se acentua nos níveis 14 e 12 de resolução para as *wavelets* de Haar e de Daubechies $D4$. Com isso, reforça a melhora da eficácia quando se utiliza a *wavelet* de Daubechies $D4$ em relação a Haar. Salientenado que, para este experimento, a queda entre os níveis 15 e 14 da *wavelet* de Haar foi muito mais acentuada ao se comparar com os esperimentos anteriores.

Coleção: TREC-SJM

Este é último experimento na técinca de IR, no qual a base é ainda maior no que se refere à quantidade de documentos. Nas Figuras 4.7 e 4.8, que o comportamento nas resoluções máximas das transformadas é similar a todos os anteriores.

Aqui, vale ressaltar uma pequena diferença, em que a transformada da *wavelet* de Haar tem uma melhor eficácia na primeira redução do nível de resolução, mas logo em seguida apresenta comportamento semelhando aos experimentos anteriores. E, apesar de que a *wavelet* de Daubechies $D4$ te uma superioridade em relação à de Haar, nos níveis seguites de resolução ela declina acentuadamente.

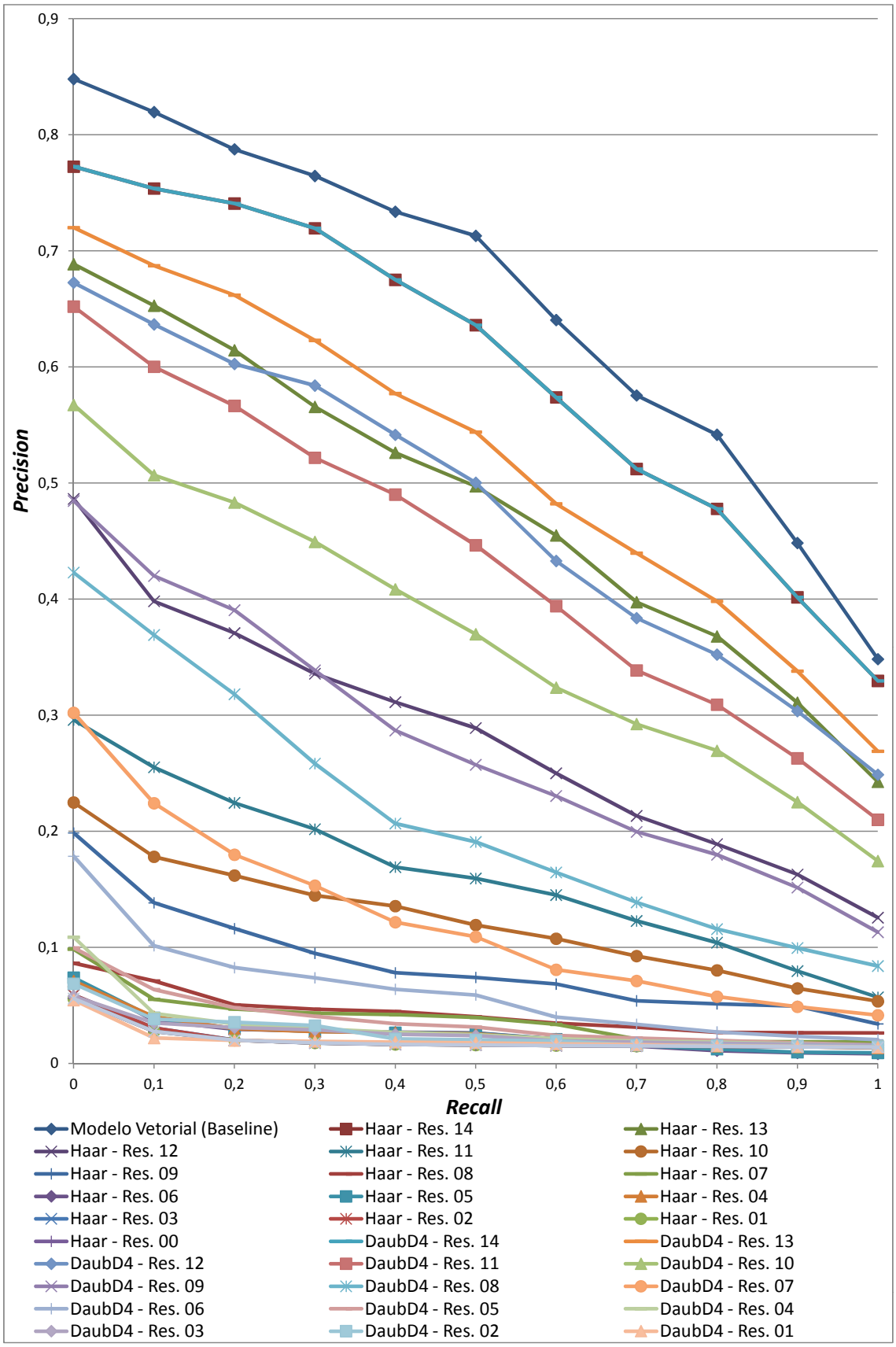


Figura 4.3: Resultados de RI (Precisão x Abrangência) sobre a coleção de documentos TREC-DOE. (ver Tabelas A.3 e A.4)

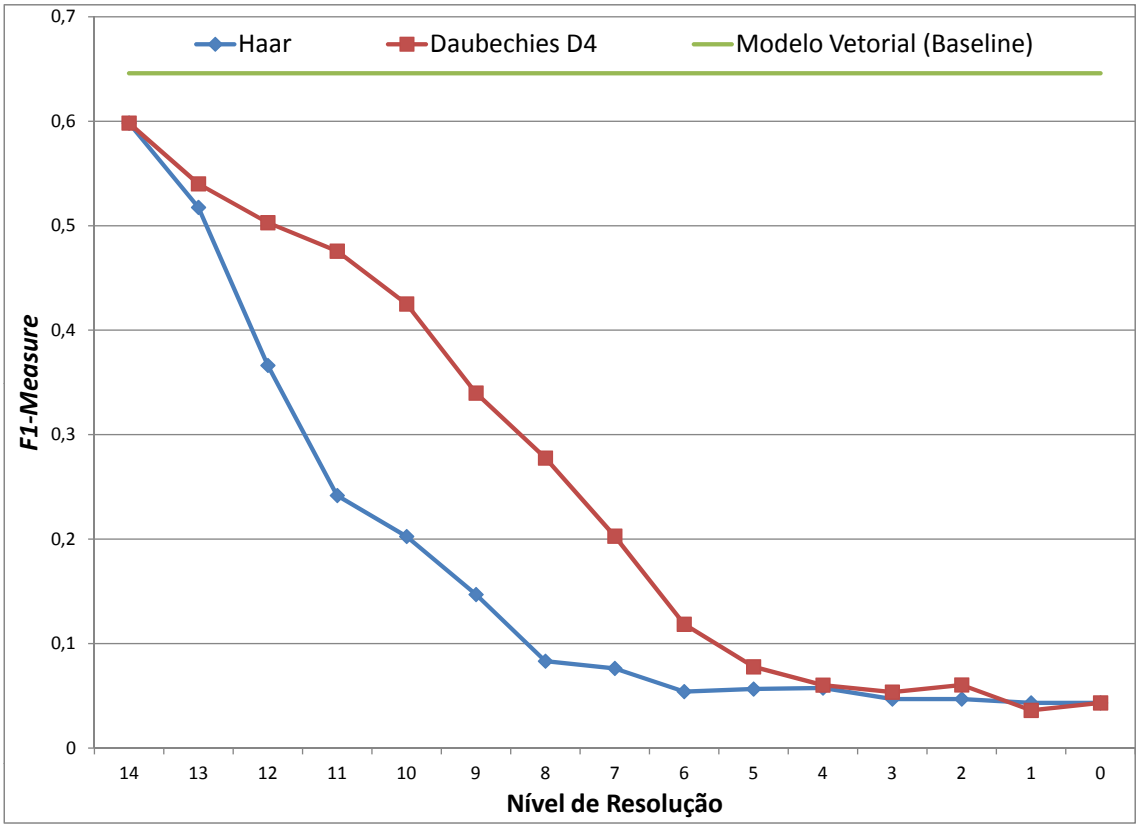


Figura 4.4: Resultados de RI (*F1-Measure* x Abrangência) sobre a coleção de documentos TREC-DOE. (ver Tabelas A.3 e A.4)

4.4.2 Experimentos em Classificação da Informação

Os experimentos para se avaliar o uso da transformada *wavelet* em técnicas de classificação da informação foram feitos somente em uma coleção de dados, a Reuters-21578, que se caracteriza por ser bastante robusta e largamente utilizadas em experimentos desta espécie.

Coleção: Reuters-21578

Para esta base, variou-se o número (k) de vizinhos mais próximos, e para cada k executou-se um experimento, do qual foi extraído um único valor de cada medida de avaliação. E assim, cada um dos gráficos nas Figuras 4.9, 4.10, 4.11 e 4.12 contém a representações gráfica dos resultados de cada experimento para um dado valor de $k \in \{1, 3, 5, 15, 30, 45, 60, 75, 90\}$.

Um primeiro fato muito interessante que pode ser observado é que tanto para o algoritmo *baseline*, como para a utilização das transformadas das duas *wavelets* mãe na resolução máxima, tem-se valores exatamente iguais em todas as medidas de avaliação. E ainda, estes valores estão muito abaixo dos valores de outras resoluções menores, de forma que o efeito aproximando dos níveis de resoluções quase nulos.

Os valores da medida de avaliação da precisão estão representados na Figura

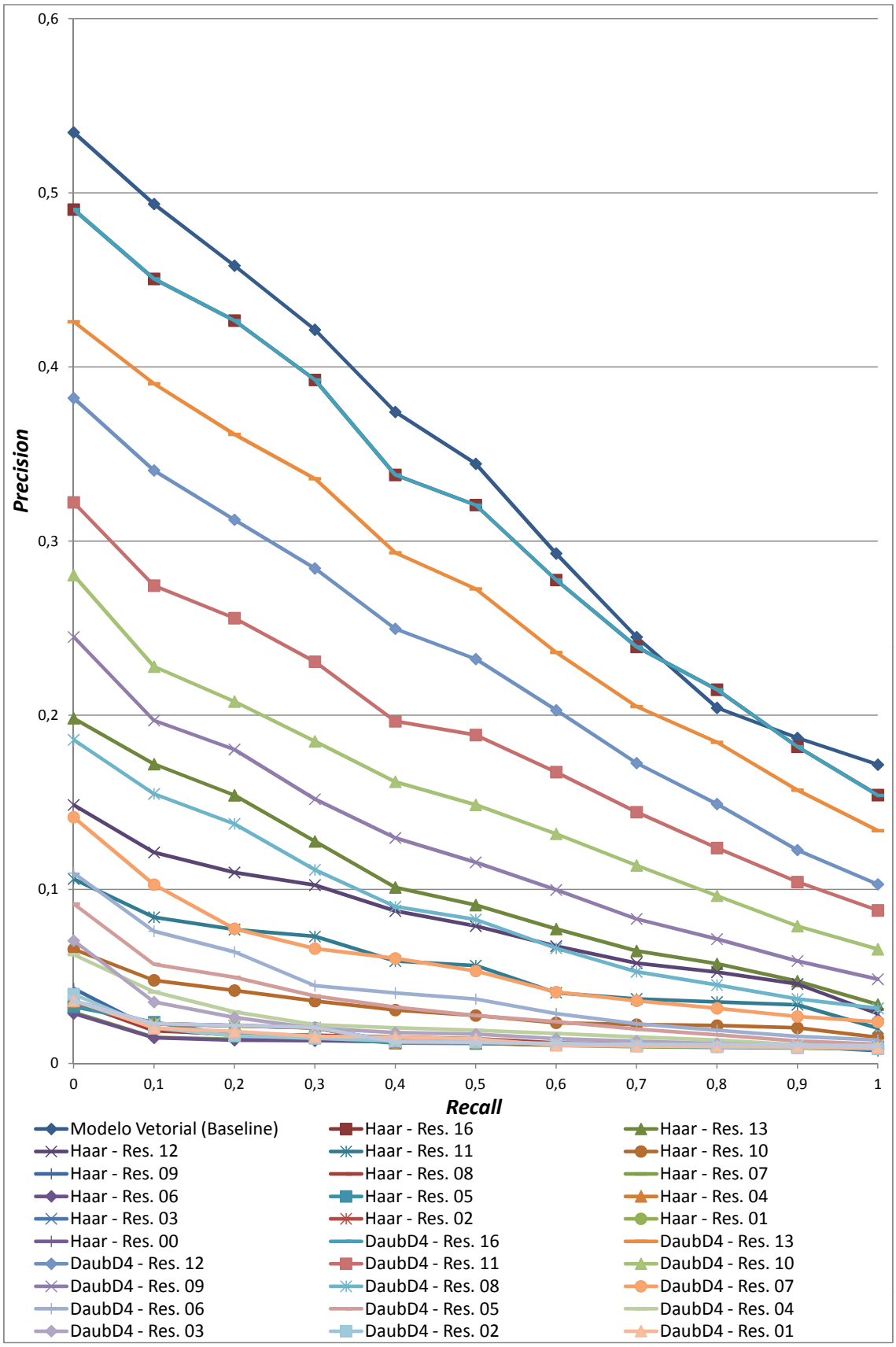


Figura 4.5: Resultados de RI (Precisão x Abrangência) sobre a coleção de documentos TREC-FR. (ver Tabelas A.5 e A.6)

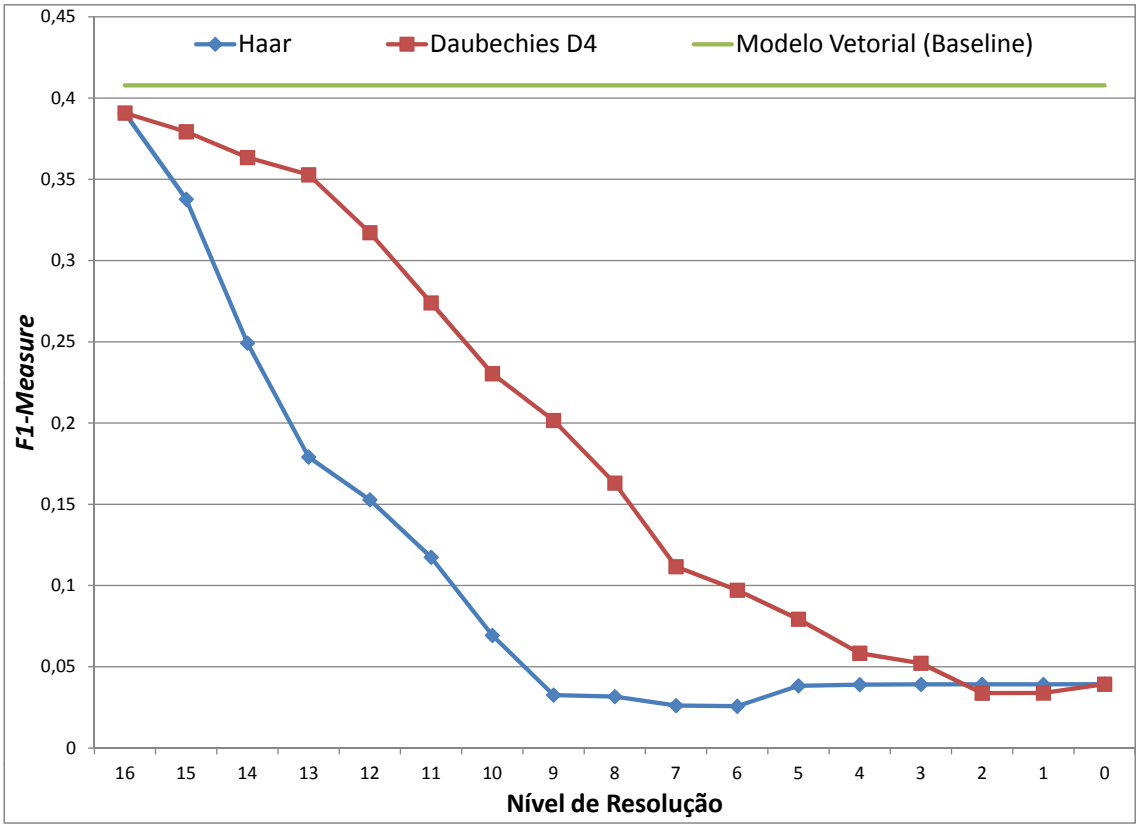


Figura 4.6: Resultados de RI (*F1-Measure* x Abrangência) sobre a coleção de documentos TREC-FR. (ver Tabelas A.5 e A.6)

4.9, e a primeira impressão refere-se à não uniformidade no comportamento entre cada nível de resolução. Ou seja, a redução da resolução não implica obrigatoriamente na redução da precisão, podendo esta aumentar ou diminuir bruscamente. Observa-se que os melhores resultados foram com a resolução 11 para as duas transformadas, Haar e Daubechies $D4$. As curvas das dos resultados das transformadas de Daubechies $D4$ nas resoluções 11 e 12 apresentam um comportamento, no mínimo, intrigante. Essas curvas, mesmo sofrendo uma queda inicial, começam a assumir valores cada vez maiores quando k aumenta, atingindo valores máximos para $k > 15$, e quando $k = 90$ a precisão dessas duas resoluções se iguala.

No gráfico da Figura 4.10, que apresenta os valores da medida de abrangência, mostra grande desigualdade para os primeiros valores de k para algumas resoluções das transformadas. E, a partir do valor de $k = 45$, essas medidas sofrem poucas variações em cada nível de resolução.

E em relação às medidas *F1-measure* e *accuracy*, cujos resultados estão representados respectivamente nas Figuras 4.11 e 4.12, apresentam comportamento bastante similar ao da medida de precisão, em que também evidenciam a não uniformidade do mesmo entre cada nível de resolução.

Dessa forma, observa-se pelos gráficos das Figuras 4.9, 4.10, 4.11 e 4.12, que a

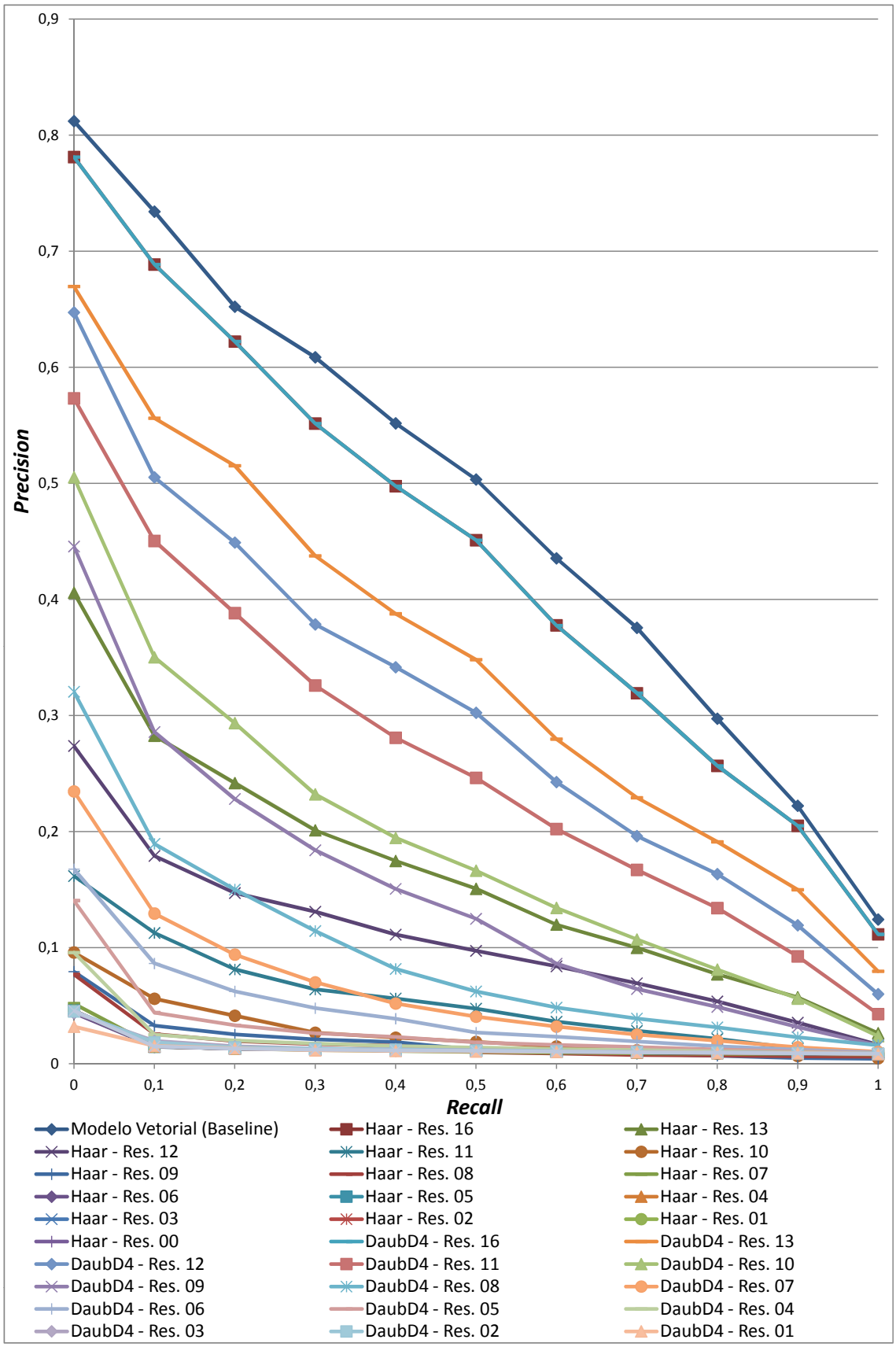


Figura 4.7: Resultados de RI (Precisão x Abrangência) sobre a coleção de documentos TREC-SJM. (ver Tabelas A.7 e A.8)

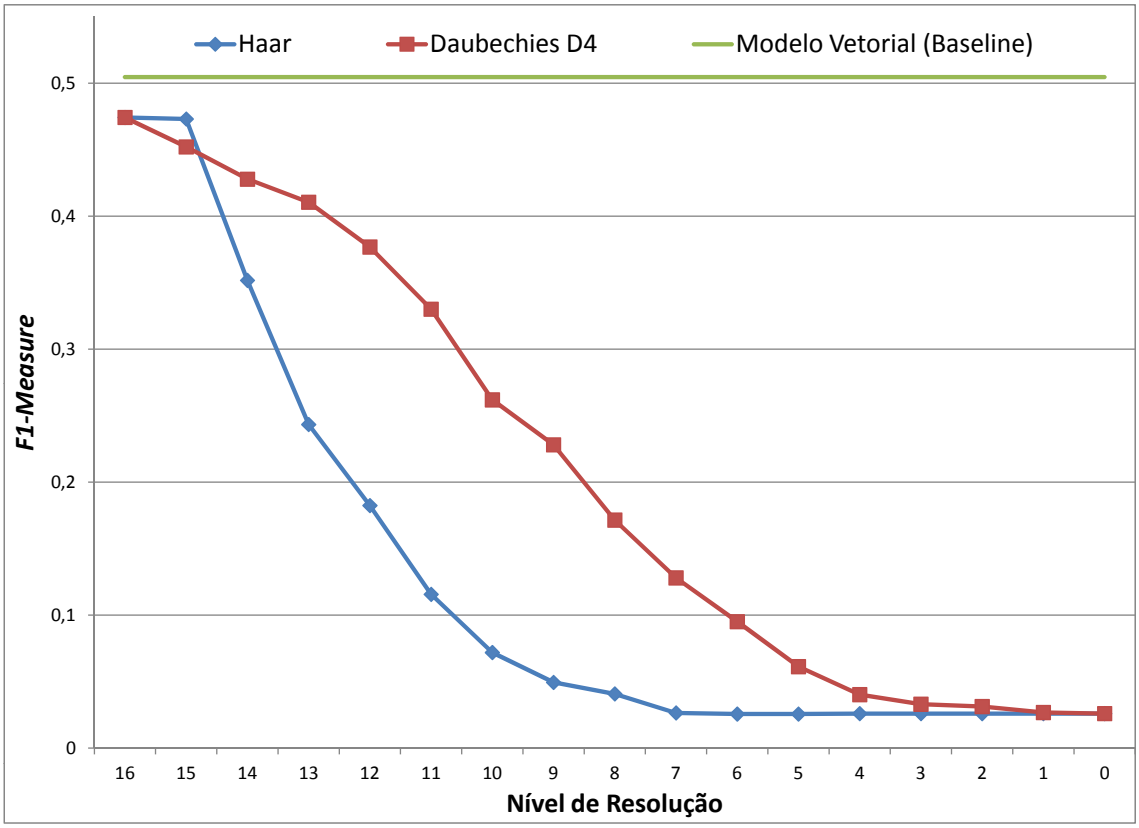


Figura 4.8: Resultados de RI (*F1-Measure* x Abrangência) sobre a coleção de documentos TREC-SJM. (ver Tabelas A.7 e A.8)

transformada de Haar na resolução 11 demonstrou-se com uma eficácia superior ao *baseline* ou a qualquer outra transformada para valores de $k \leq 15$. Já para $k > 15$, foi a transformada de Daubechies $D4$ no nível de resolução 11 que teve uma eficácia superior, seguido imediatamente da transformada de Daubechies $D4$ na resolução 12.

4.4.3 Experimentos em Agrupamento da Informação

Estes experimentos avaliam o uso da transformada *wavelet* em técnicas de agrupamento da informação, e também foram executados somente sobre a robusta coleção de dados Reuters-21578.

Coleção: Reuters-21578

Para esta base, pretende-se variar o número (K) de clusters formados pelo algoritmo, e para cada K executar um experimento, e assim, extrar um único valor de cada medida de avaliação.

O experimento iria variar o número de clusters de acordo com conjunto de valores de $K \in \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Entretanto, até o momento da escrita deste trabalho, foi possível executar o os experimentos para $k = 10$ e para os seguintes valores de

resolução $j \in \{15, 13, 12, 11, 10, 9, 8, 7, 6\}$ da transformada *wavelet* de Haar, devido a limitações computacionais já esclarecidas na Seção 4.3.

Os gráficos da Figura 4.13 mostram os resultados obtidos nesse experimento dentro das condições previamente explicadas. O gráfico da Figura 4.13a tenta expor todas as medidas obtidas, mas por uma questão de escala do gráfico, não é possível visualizar todos detalhes. Por esse motivo, foi exposto dois outros gráficos nas Figuras 4.13b e 4.13c, afim de expor mais detalhes dos valores de *log-likelihood* em todas as resoluções experimentadas.

Percebe-se claramente que há uma grande semelhança nos valores do *baseline*, que utiliza o sinal original, e da transformada na sua resolução máxima ($j = 15$), e particularmente são valores negativos. Por outro lado, pode-se perceber que há uma enorme diferença entre estes valores e os valores das resoluções seguintes ($j \leq 13$), como também há alguma diferença considerável entre os resultados da transformada na resolução de nível 13 e os de níveis consecutivos (ou seja, entre $j = 13$ e $j \leq 12$).

4.5 Análise dos Resultados

Agora serão discutidos os resultados apresentados até o momento, evidenciando a comparação entre os resultados do *baseline* e os resultados das transformadas *wavelets* nas diferentes resoluções para cada técnica experimentada, como também, entre as *wavelets* mãe utilizadas nestes trabalho.

Primeiramente, foi experimentada a técnica de recuperação da informação utilizando-se como *baseline* o modelo vetorial. E para todas as bases, observou-se que este *baseline* teve o comportamento aproximado pelas transformadas das duas *wavelets* mãe, Haar e Daubechies $D4$, na resolução máxima. Enquanto que nos níveis de resolução imediatamente seguintes, houve pouca perda na sua eficácia, e assim pode-se afirmar que foi possível ter uma quantidade de resultados ainda relevantes, mesmo reduzindo a quantidade de informação no cálculo da similaridade pela metade, ou em até 1/4. Já para outros outros níveis, a perda na eficácia se acentuou gradativamente até o nível 0.

Agora, ao se comparar por observação visual dos gráficos de precisão e abrangência os resultados entre os dois tipos de *wavelets*, nota-se uma considerável superioridade do uso das *wavelets* de Daubechies $D4$ em relação à *wavelet* de Haar. Mas essa observação visual ainda é confusa. Por isso, representa-se graficamente os valores da medida *F1-Measure* nas Figuras 4.2, 4.4 e 4.6, em função da variação do nível de resolução de cada *wavelet* mãe. Com isso, pode-se comparar com maior clareza as eficácias das duas *wavelets*, e com o *baseline* para as três coleções de dados em questão. Isso, permite uma comparação clara e precisa entre os objetos de estudo.

Nessa figura, ainda se nota nos primeiros níveis de resolução uma certa "de-

mora" para que a *F1-measure* da transformada *wavelet* de Daubechies $D4$ decline mais acentuadamente, em especial na Figura 4.6 que se refere ao experimento da coleção que contém a maior quantidade de informação. Dessa forma, percebe-se que além da *wavelet* de Daubechies $D4$ superar a *wavelet* de Haar, quando aplicadas à recuperação da informação textual, ainda se aproxima da eficácia do modelo vetorial mesmo com a redução em aproximadamente 50%, em 25% ou mesmo em 12,5% da quantidade de informação inicial. Isso, com certeza, é uma grande vantagem das *wavelets* para se tratar da recuperação da informação, em relação aos modelos clássicos de RI.

Com relação à experimentação sobre a técnica de classificação da informação textual, percebeu-se uma eficácia muito superior ao algoritmo *baseline*, e esta percepção está quantificada e representada graficamente nas Figuras 4.14, 4.15, 4.16 e 4.17. Estes gráficos representam a área sob cada medida de avaliação (Precisão, Abrangência, *F1-Measure* e *Accuracy*), em função da variação do nível de resolução da transformada *wavelet*. Nos gráficos estão resumidos os valores dos resultados para as duas *wavelts* mãe, Haar e Daubechies $D4$, além de uma linha de referência correspondente ao *baseline*. E, da mesma forma, permite-se comparar claramente, e com precisão, as eficácias entre as *wavelets*, como com o algoritmo *baseline* da classificação.

Os gráficos da precisão (4.14), *F1-measure* (4.16) e *accuracy* (4.17), além de evidenciarem que o uso de *wavelet* obteve resultados melhores que os algoritmos tradicionais de classificação da informação, embora na resolução máxima tenha resultado em valores exatamente iguais ao KNN, mostra também que para as 2 resoluções mais altas e menores que a máxima, a *wavelet* de Haar teve resultados melhores que a de Daubechies $D4$. Enquanto que esta, parecia estar um nível atrasado em relação à de Haar para esses níveis de resolução.

Descendo mais níveis ainda, a *wavelet* de Haar chega ao seu pico, mas logo em seguida declina profundamente, tornando-se inferior ao algoritmo *baseline* a partir da sexta resolução. Ou seja, a utilização da transformada dessa *wavelet* só tem uma eficácia inferior ao KNN, na forma tradicional, quando a resolução implica em aproximadamente 0,19% da informação original. Realmente é incrível esse resultado. E pode-se ter resultados ainda mais surpreendentes quando se usa a *wavelet* de Daubechies $D4$, que teve seu pico na resolução 11, ultrapassando a eficácia de qualquer nível de resolução da transfromada de Haar. E a partir dessa resolução, apenas se distância mais da curva da *wavelet* de Haar, tendo valores inatingíveis até resoluções quase nulas. Assim, a transformada *wavelet* de Daubechies, aplicada à classificação da informação, só se torna inferior ao KNN tradicional quando utilizada com uma resolução correspondente a aproximadamente 0,006% da informação original.

Entretanto, quando se refere à medida de abrangência (Figura 4.15), nota-se que

o valor máximo é definido pelo *baseline*, enquanto que as cuvas da *wavelet* de Haar e de Daubechies $D4$ variam, de modo a igualar-se ao *baseline* na resolução máxima e mínima. A curva da *wavelet* de Haar já se iguala para $k \leq 6$, e a de Daubacheis $D4$ iguala-se enquanto $k \geq 13$. Interessante que enquanto a curva relativa à Haar apresenta uma queda inicial e logo depois se recupera, a curva relativa à Daubachies apesar de declinar pouco no ínício, logo entra em uma queda significativa até os últimos níveis de resolução.

Já em relação ao experimento com a técnica de agrupamento da informação, ainda não se pode tirar muitas conclusões sobre a eficácia da aplicação da transformada *wavelet* sobre essa técnina. Entretanto, a Figura 4.13 sugere que o uso das *wavelets* de Haar com as suas transformadas em níveis de resolução intermediários, podem obter valores consideravelmente superiores ao algoritmo *baseline*, o K-Means na sua forma tradicional.

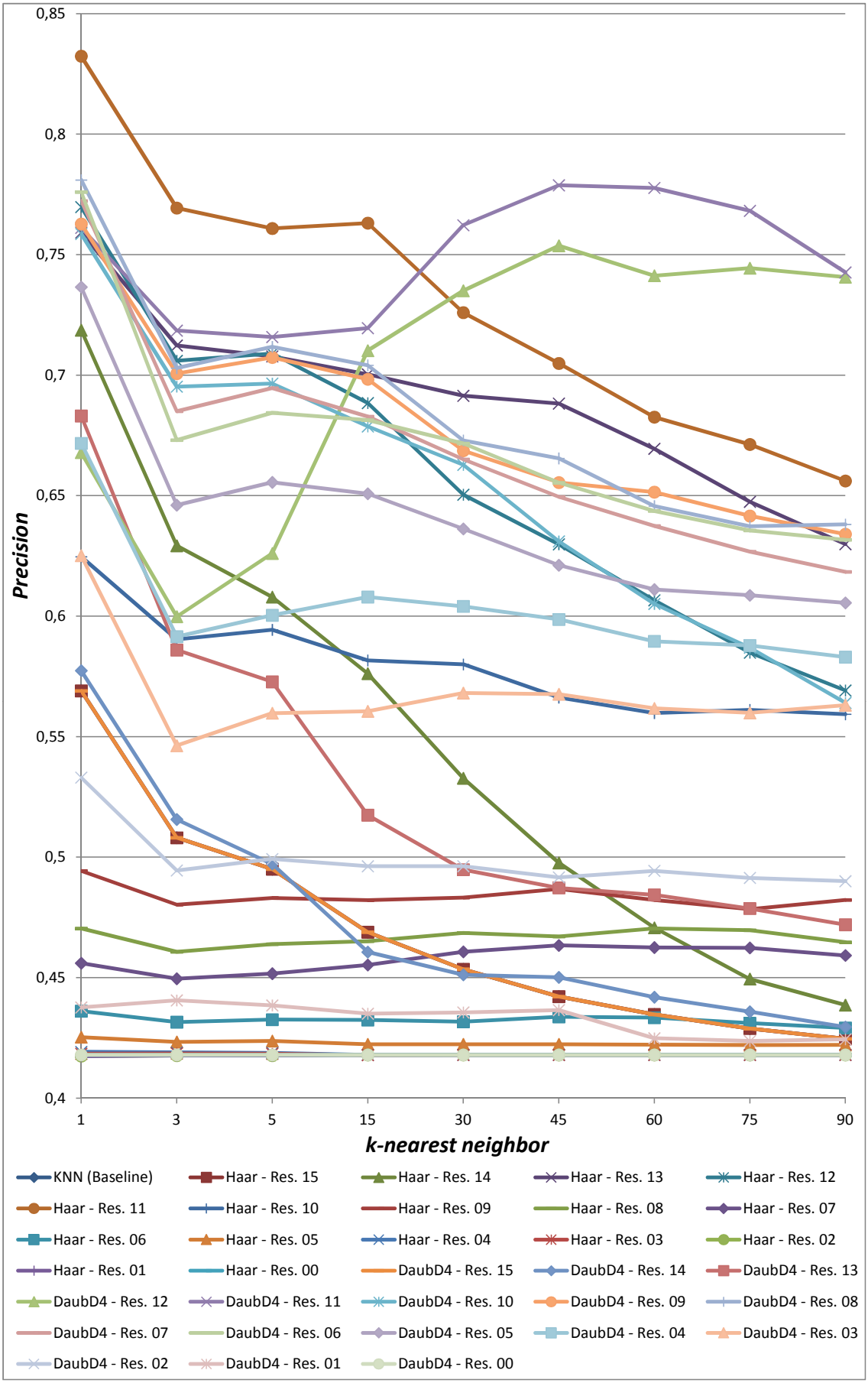


Figura 4.9: Resultados da medida de Precisão na Classificação da Informação sobre a coleção de documentos Reuters-21578. (ver Tabela A.9)

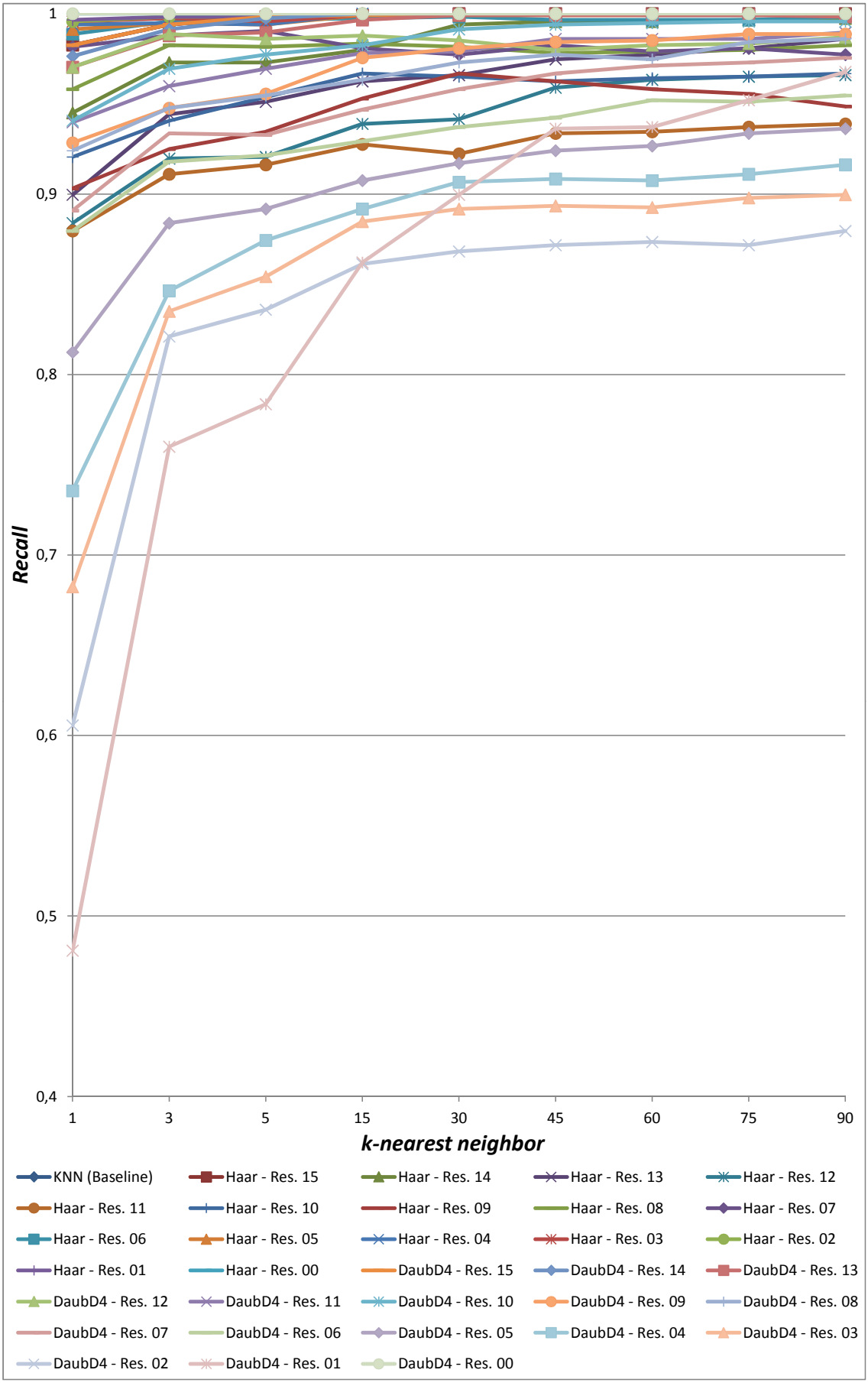


Figura 4.10: Resultados da medida de Abrangência na Classificação da Informação sobre a coleção de documentos Reuters-21578. (ver Tabela A.10)

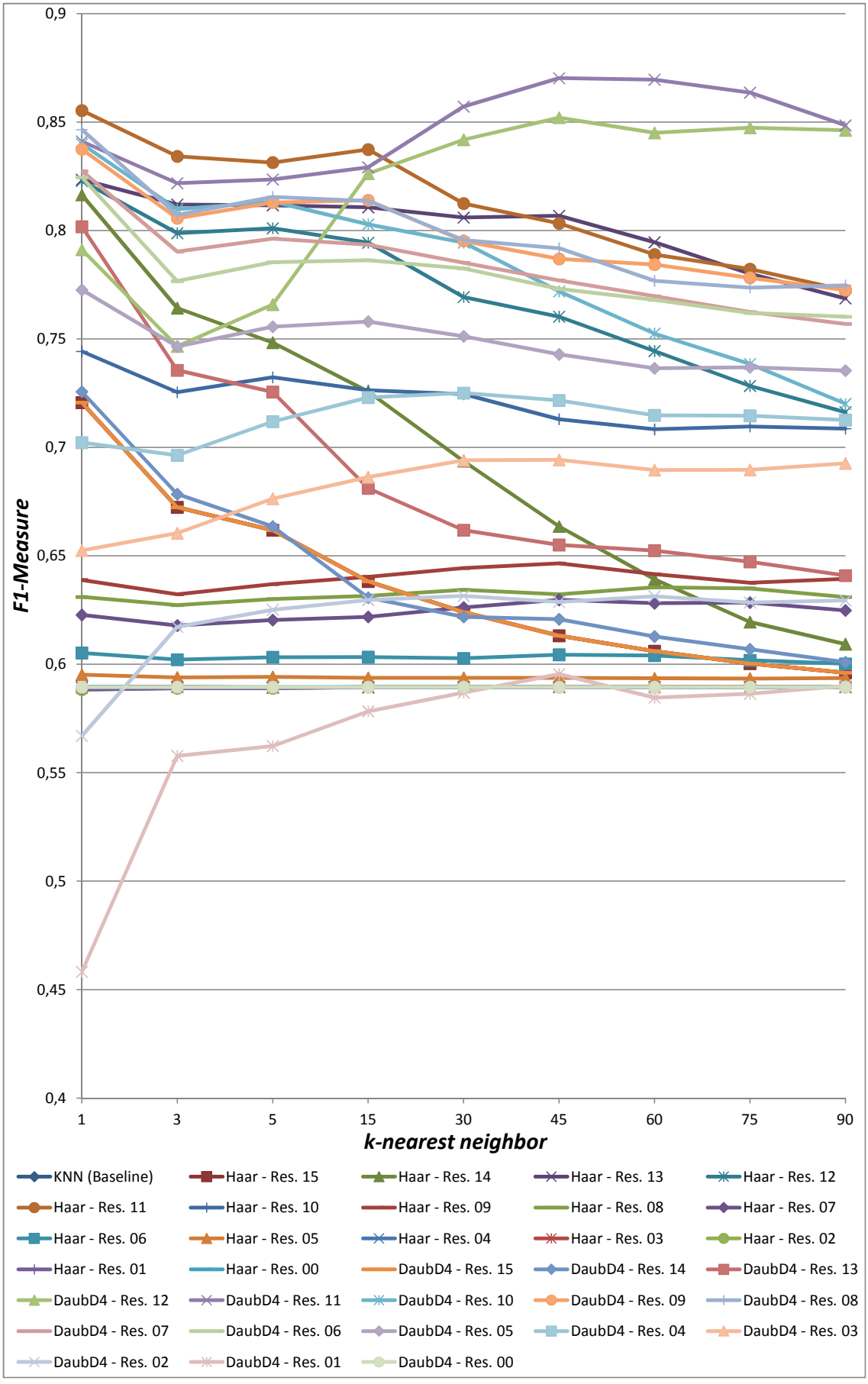


Figura 4.11: Resultados da *F1-Measure* de Precisão na Classificação da Informação sobre a coleção de documentos Reuters-21578. (ver Tabela A.11)

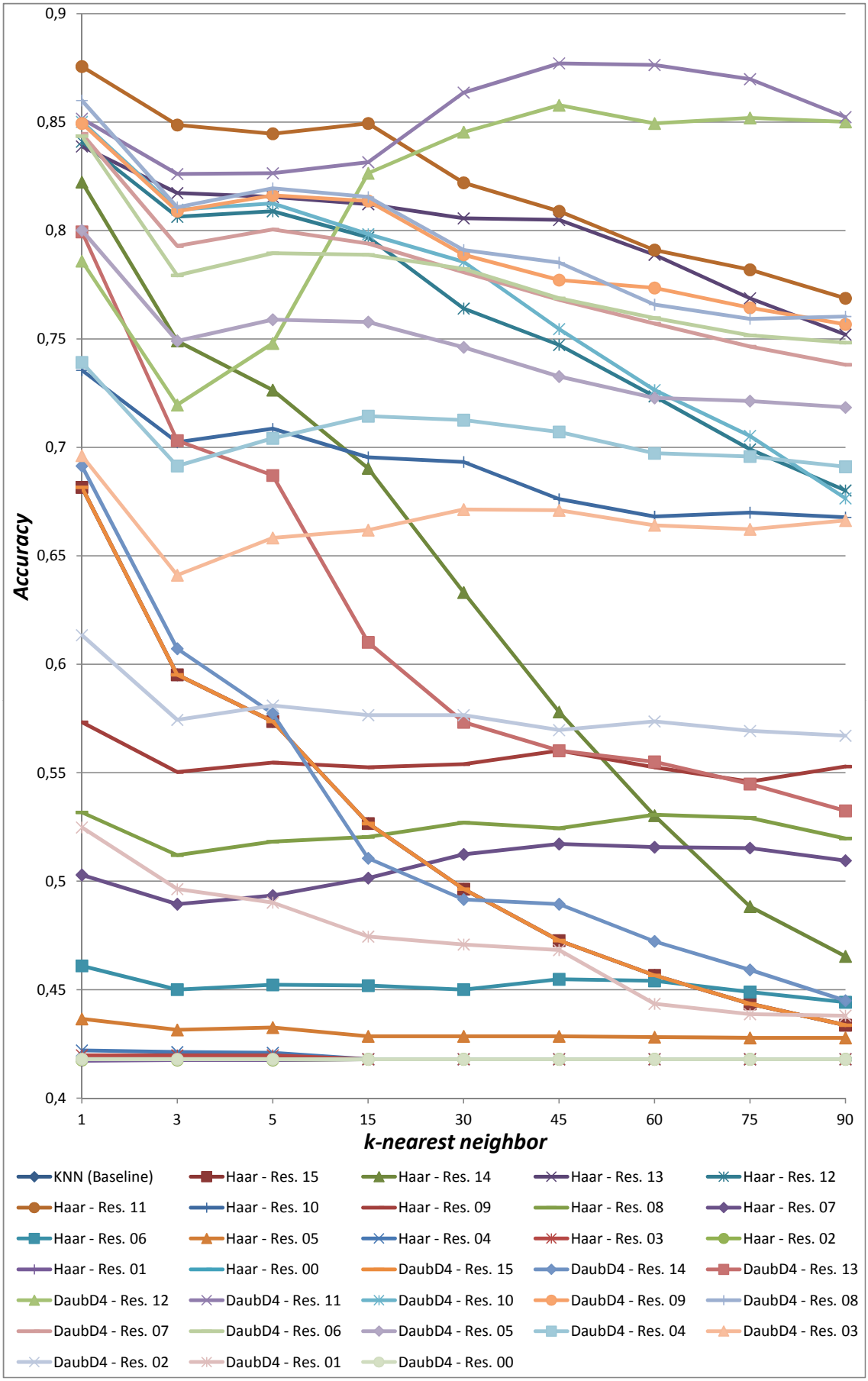
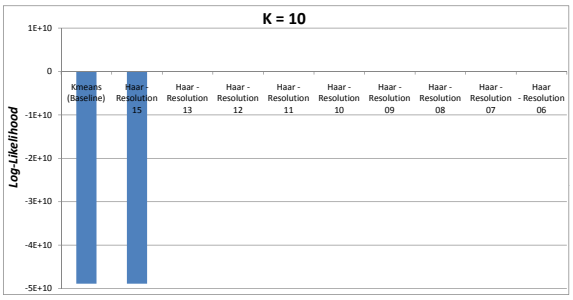
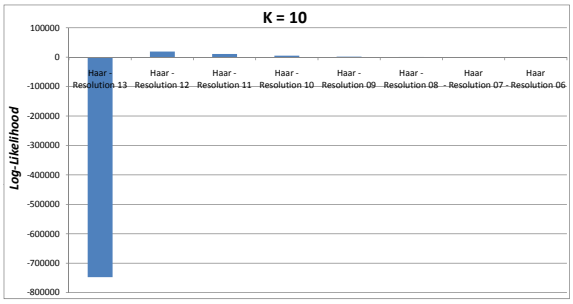


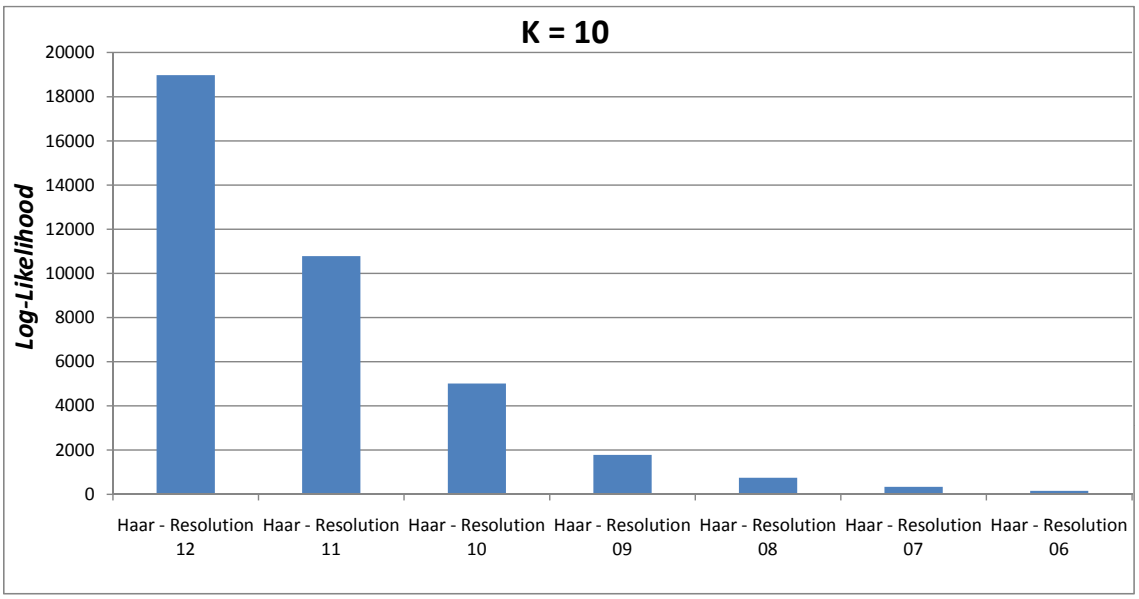
Figura 4.12: Resultados da medida de *Accuracy* na Classificação da Informação sobre a coleção de documentos Reuters-21578. (ver Tabela A.12)



(a) Log-Likelihood ($K = 10$)



(b) Log-Likelihood ($K = 10$ e $j \leq 13$)



(c) Log-Likelihood ($K = 10$ e $j \leq 12$)

Figura 4.13: Resultados das medidas de avaliação de Agrupamento da Informação sobre a coleção de documentos Reuters-21578. (ver Tabela A.13)

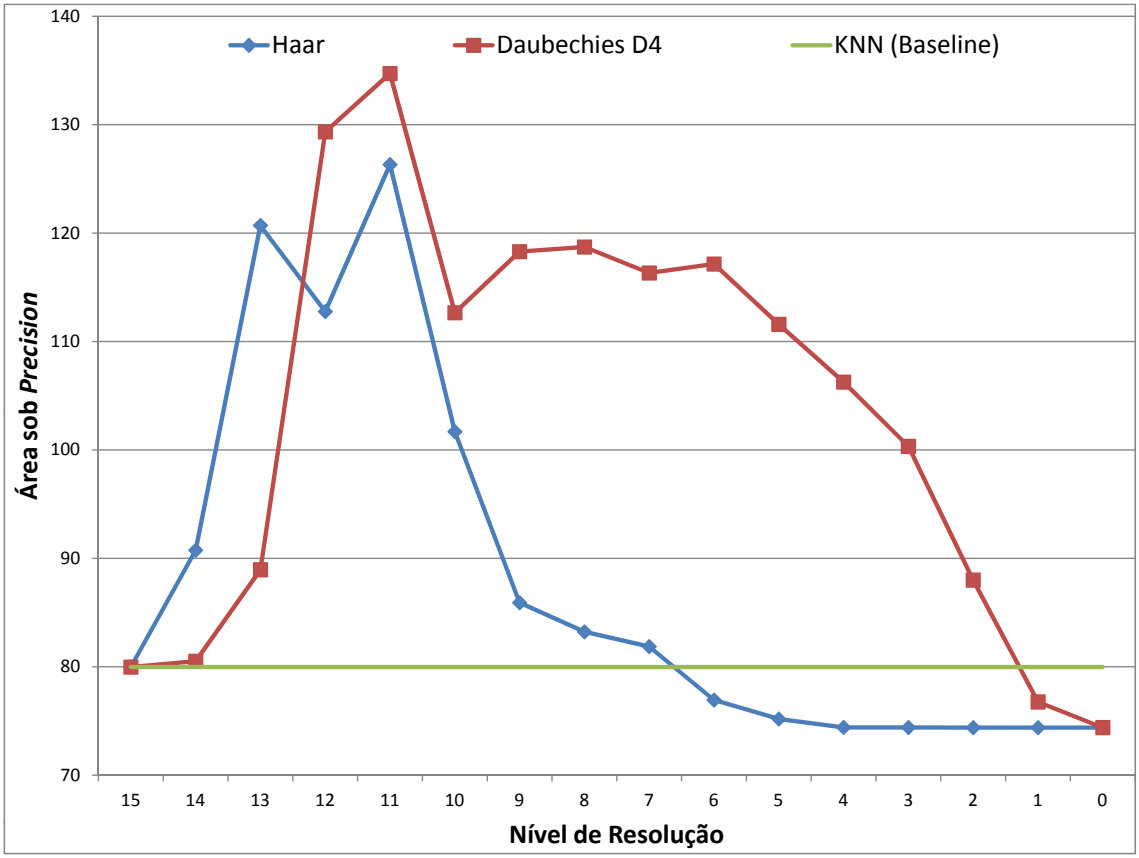


Figura 4.14: Comparação entre as áreas da medida de Precisão das transfomadas *wavelet* de Haar e Daubechies $D4$, obtidos a partir dos experimentos de classificação sobre a base Reuters-21578, em função da variação do nível de resolução. (ver Tabela A.9)

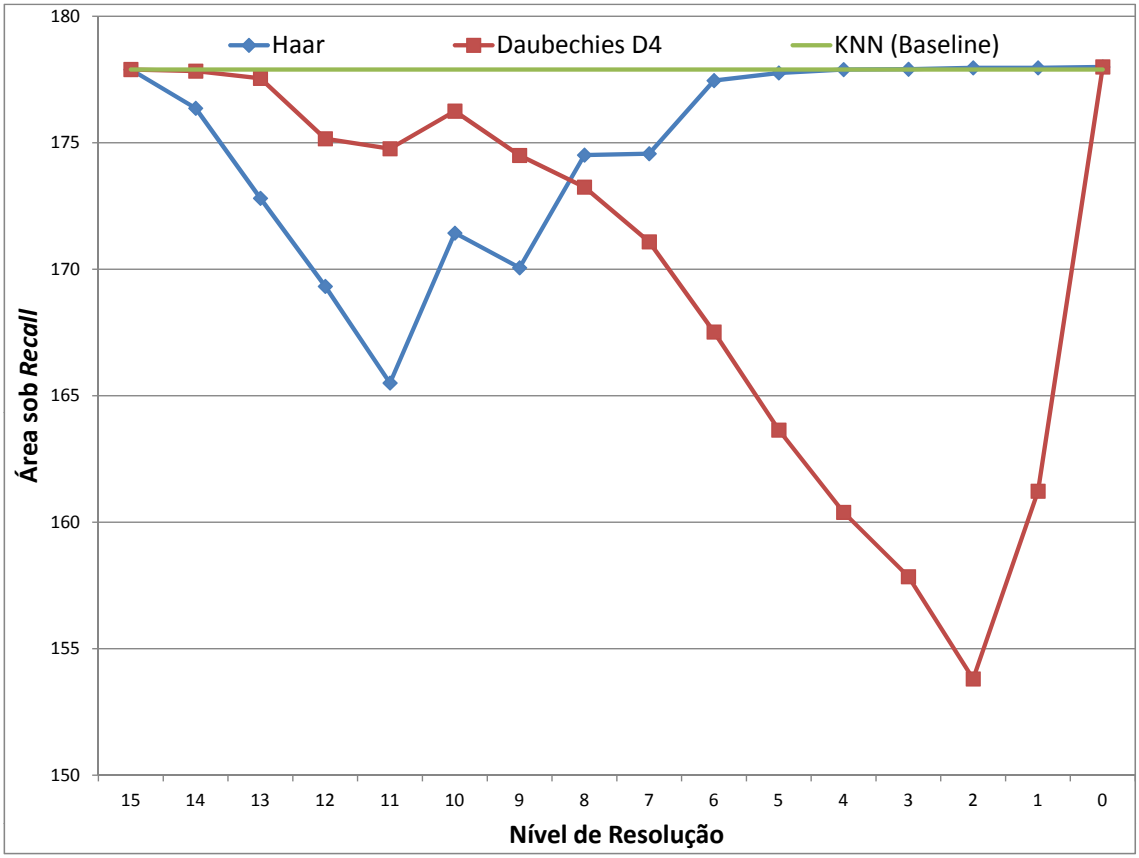


Figura 4.15: Comparação entre as áreas da medida de Abrangência das transfomadas *wavelet* de Haar e Daubechies $D4$, obtidos a partir dos experimentos de classificação sobre a base Reuters-21578, em função da variação do nível de resolução. (ver Tabela A.10)

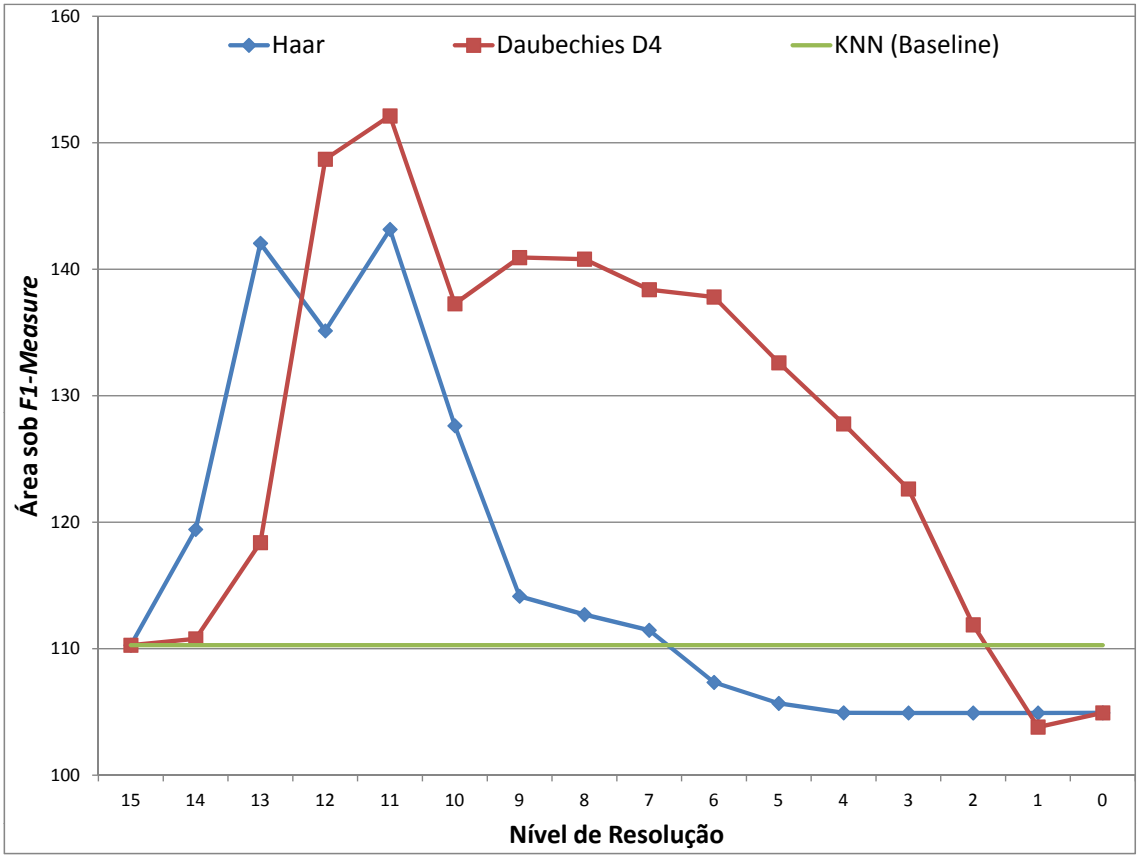


Figura 4.16: Comparação entre as áreas da medida de *F1-Measure* das transfomadas *wavelet* de Haar e Daubechies $D4$, obtidos a partir dos experimentos de classificação sobre a base Reuters-21578, em função da variação do nível de resolução. (ver Tabelas A.11)

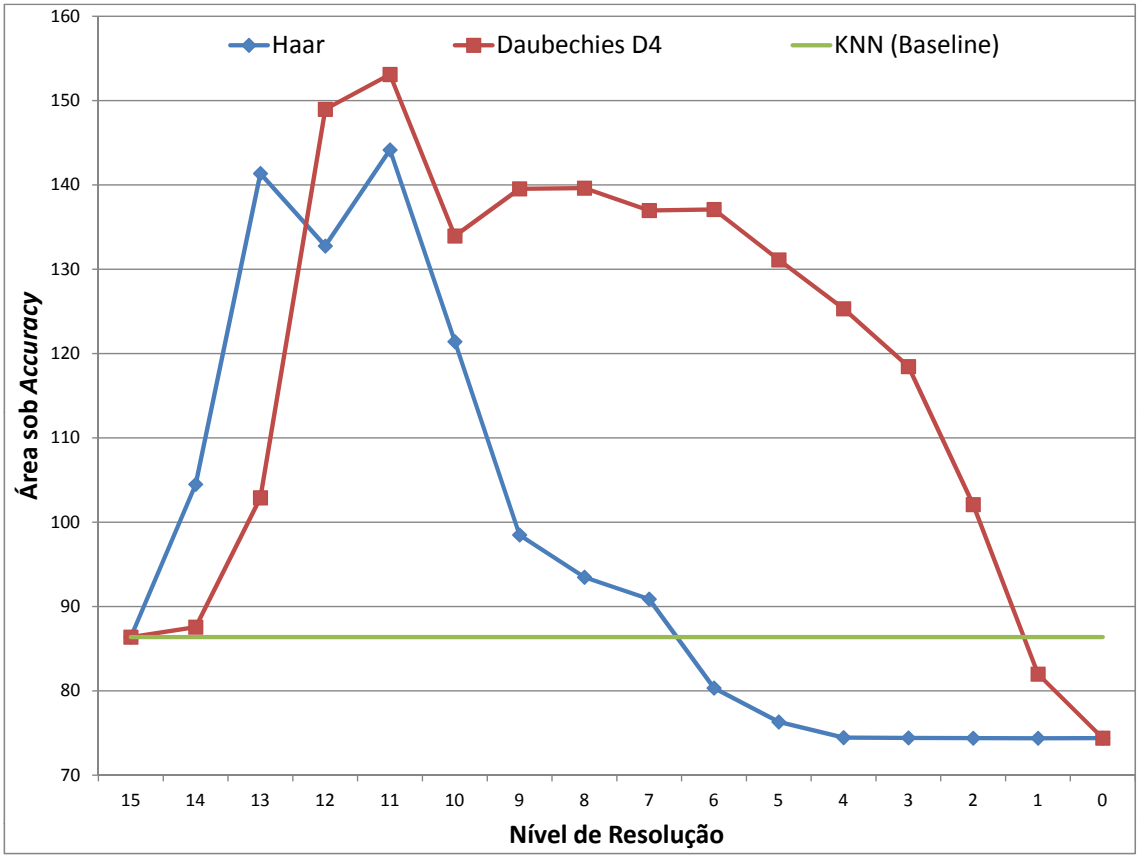


Figura 4.17: Comparação entre as áreas da medida de *Accuracy* das transfomadas *wavelet* de Haar e Daubechies $D4$, obtidos a partir dos experimentos de classificação sobre a base Reuters-21578, em função da variação do nível de resolução. (ver Tabelas A.12)

Capítulo 5

Conclusão e Trabalhos Futuros

Finalmente, este capítulo apresentará uma breve síntese da avaliação da utilização da transformada *wavelet* em recuperação, classficiação e agrupamento da informação textual e contribuições deste trabalho. Em seguida, serão levantadas algumas vertentes de possibilidades para a continuidade deste trabalho.

A partir dos experimentos, pode-se concluir que a aplicabilidade das *wavelet* foi não só possível para sistemas de recuperação e classificação de texto, como também para o agrupamento da informação textual. Além disso, com as análises desses resultados, constatou-se que para sistemas de recuperação da informação, o uso da transformada *wavelet* não ultrapassa a eficácia do modelo clássico desses sistemas. Entretanto, mesmo reduzindo exponencialmente a informação pelo uso da transformada, ainda se obtém um eficácia próxima ao modelo vetorial de RI. Quanto aos sistemas de classificação da informação, a análise dos resultados indicou um ganho surpreendente com a aplicação da transformada *wavelet* nesses sistemas, quando se comparado como KNN, algoritmo clássico para tais sistemas. Esse ganho se deu de duas formas, primeiro no que se refere à superioridade da classificação em si, obtendo resultados melhores que o KNN, e segundo por isso ter sido atingido reduzindo exponencialmente a informação utilizada, através da transformada.

Já para o agrupamento da informação, ainda não pôde se concluir sobre a eficácia da aplicabilidade da transformada *wavelet* nesse sistema, embora a análise tenha sugerido que possa ter um comportamento semelhando ao que ocorreu na classificação. Assim, pretende continuar com os experimentos para se obter resultados suficientes que permitam uma conclusão sobre a eficácia da transformada *wavelet* em sistemas de agrupamento da informação textual.

E ainda, nos dois primeiros sistema citados, a transformada da *wavelet* mãe de Daubechies $D4$ teve uma eficácia superior à transformada *wavelet* de Haar, ficando evidente de maneira intrigante na classificação.

Com isso, pode-se afirmar que a transformada *wavelet* além de ser perfeitamente aplicável a essas técnincas, possui propriedades que tornam vantajoso seu uso, como

a propriedade da multiresolução que permite reduzir a quantidade de informação, mantendo a semântica de forma aproximada da informação original. Essa propriedade, com certeza, foi a principal responsável dos resultados obtidos.

Quanto à continuidade deste trabalho, pode-se citar primeiramente a aplicação da implementação da transformada de outras *wavelets* mãe, afim de verificar se ocorre algum caso que supere a eficácia da transformada de Daubechies $D4$ nesses sistemas.

Outro ponto observado, e que deve ser averiguado em mais detalhes, é que o grau de esparsividade do sinal original, bem como dos vetores de saída das transformada nas diferentes resoluções, a serem utilizados no cálculo de similaridade, tenha alguma relação com a variação do nível de resolução juntamente com a eficácia obtida nesse nível. Isso se refere quando se aplicou a transformada na técninca de classificação.

Durante a execução dos experimentos foi detectado uma demora na indexação, devido ao cálculo das transformadas. Então, na adição de novos documentos no índice, isso pode tomar grande custo computacional, mas isso pode ser melhorado ou, de forma similar ao que aplica-se no modelo LSI, pode adotar-se métodos de atualização do índice, como o *folding-in* e o *SVD-Updating* citados no Capítulo 2.

Outra possibilidade de trabalho, é avaliação da transformaa *wavelet* comparando-se com outros *baselines* para cada um dos sistemas em questão. E ainda, pode-se utilizar outras coleções de dados, como as coleções da TREC que ainda não foram experimentadas neste trabalho, como também em coleções de testes de mineração de dados maiores que a Reuters-21578, como por ex. a Reuters-RCV2.

Referências Bibliográficas

[1] NIBLACK, C. W., BARBER, R., EQUITZ, W., et al. "QBIC project: querying images by content, using color, texture, and shape". In: *Retrieval Methods and Systems for Image Databases II*, v. 1908, pp. 173–187, San Jose, CA, USA, Feb 1993. SPIE. doi: 10.1117/12.143648.

[2] MURALA, S., GONDE, A., MAHESHWARI, R. "Color and Texture Features for Image Indexing and Retrieval". In: *Advance Computing Conference, 2009. IACC 2009. IEEE International*, pp. 1411–1416, Mar 2009. ISBN: 978-1-4244-2927-1. doi: 10.1109/IADCC.2009.4809223.

[3] FLESCA, S., MANCO, G., MASCIARI, E., et al. "Fast Detection of XML Structural Similarity", *Knowledge and Data Engineering, IEEE Transactions on*, v. 17, n. 2, pp. 160–175, Feb 2005. ISSN: 1041-4347. doi: 10.1109/TKDE.2005.27. Student Member-Pugliese, Andrea.

[4] PARK, L. A. F., RAMAMOHANARAO, K., PALANISWAMI, M. "A novel document retrieval method using the discrete wavelet transform", *ACM Trans. Inf. Syst.*, v. 23, n. 3, pp. 267–298, 2005. ISSN: 1046-8188. doi: 10.1145/1080343.1080345. Disponível em: <`http://doi.acm.org/10.1145/1080343.1080345`>.

[5] THAICHAROEN, S., ALTMAN, T., CIOS, K. J. "Structure-Based Document Model with Discrete Wavelet Transforms and Its Application to Document Classification". In: Roddick, J. F., Li, J., Christen, P., et al. (Eds.), *Seventh Australasian Data Mining Conference, 2008. AusDM'2008*, v. 87, *CRPIT*, pp. 209–217, Glenelg, South Australia, 2008. ACS.

[6] XEXEO, G., DE SOUZA, J., CASTRO, P., et al. "Using Wavelets to Classify Documents". In: *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08.*, v. 1, pp. 272–278, Los Alamitos, CA, USA, Dec 2008. IEEE Computer Society. ISBN: 978-0-7695-3496-1. doi: 10.1109/WIIAT.2008.221. Disponível em: <`http://doi.ieeecomputersociety.org/10.1109/WIIAT.2008.221`>.

[7] MANNING, C. D., RAGHAVAN, P., SCHTZE, H. *Introduction to Information Retrieval*. New York, NY, USA, Cambridge University Press, 2008. ISBN: 978-0-5218-6571-5. Disponível em: <http://nlp.stanford.edu/IR-book/>.

[8] BAEZA-YATES, R. A., RIBEIRO-NETO, B. A. *Modern Information Retrieval: The Concepts and Technology behind Search*. Addison-Wesley. 2 ed. Harlow, England, Pearson Higher Education Ltd., 2011. ISBN: 978-0-321-41691-9.

[9] BERRY, M. W., DUMAIS, S., O'BRIEN, G., et al. "Using Linear Algebra for Intelligent Information Retrieval", *SIAM Review*, v. 37, pp. 573–595, 1995.

[10] O'BRIEN, G. W. "Information Management Tools for Updating an SVD-Encoded Indexing Scheme". 1994.

[11] DA SILVA, R. L. S. *Modelo de Sinais para Busca e Recuperação de Informação Textual*. Dissertação de Mestrado, Universidade Federal do Rio de Janeiro, COPPE, PESC, Mar 2007. Disponível em: <http://www.cos.ufrj.br/uploadfiles/1177328414.pdf>.

[12] VOORHEES, E. M., HARMAN, D. "Overview of the sixth text REtrieval conference (TREC-6)", *Inf. Process. Manage.*, v. 36, pp. 3–35, Jan 2000. ISSN: 0306-4573. doi: 10.1016/S0306-4573(99)00043-6. Disponível em: <http://portal.acm.org/citation.cfm?id=342528.342531>.

[13] HAN, J., KAMBER, M. *Data mining: concepts and techniques*. The Morgan Kaufmann series in data management systems. Elsevier, 2006. ISBN: 978-1-5586-0901-3.

[14] FOURIER, J. B. J. "Théorie analytique de la chaleur". 1822. Disponível em: <http://books.google.com/books?id=TDQJAAAAIAAJ>.

[15] GABOR, D. "Theory of Communication", *J. IEE*, v. 93, n. III, pp. 429–457, Nov 1946.

[16] HAAR, A. "Zur Theorie der orthogonalen Funktionensysteme", *Mathematische Annalen*, v. 69, pp. 331–371, 1910. ISSN: 0025-5831. doi: 10.1007/BF01456326. Disponível em: <http://dx.doi.org/10.1007/BF01456326>.

[17] GROSSMANN, A., MORLET, J. "Decomposition of Hardy functions into square integrable wavelets of constant shape", *SIAM J. of Math. Anal.*, v. 15, pp. 723–736, 1984.

[18] GOUPILLAUD, P., GROSSMANN, A., MORLET, J. "Cycle-Octave and related transforms in seismic signal analysis", *Geoexploration*, v. 23, pp. 85–102, 1984.

[19] MALLAT, S. "A theory for multiresolution signal decomposition: the wavelet representation", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 11, n. 7, pp. 674–693, Jul 1989. ISSN: 0162-8828. doi: 10.1109/34.192463. Disponível em: <http://doi.ieeecomputersociety.org/10.1109/34.192463>.

[20] MORETTIN, P. A. *Ondas e Ondaletas: Da Análise de Fourier à Análise de Ondaletas*. EdUSP - Editora da Universidade de São Paulo, 1999. ISBN: 978-85-314-0509-9.

[21] DE OLIVEIRA, H. M. *Análise de Sinais para Engenheiros: Uma Abordagem via Wavelets*. Sociedade Brasileira de Telecomunicações - SBrT/Brasport. Brasport, 2007. ISBN: 978-85-7452-283-8.

[22] DAUBECHIES, I., ANTONINI, M., BARLAUD, M., et al. "Image coding using wavelet transform", *Image Processing, IEEE Transactions on*, v. 1, n. 2, pp. 205–220, Apr 1992. ISSN: 1057-7149. doi: 10.1109/83.136597. Disponível em: <http://dx.doi.org/10.1109/83.136597>.

[23] MALLAT, S. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Wavelet Analysis and Its Applications Series. Academic Press, 2008. ISBN: 978-0-1237-4370-1.

[24] DAUBECHIES, I. "Orthonormal bases of compactly supported wavelets", *Communications on Pure and Applied Mathematics*, v. 41, n. 7, pp. 909–996, Oct 1988. doi: 10.1002/cpa.3160410705. Disponível em: <http://dx.doi.org/10.1002/cpa.3160410705>.

[25] DAUBECHIES, I. "The wavelet transform, time-frequency localization and signal analysis", *Information Theory, IEEE Transactions on*, v. 36, n. 5, pp. 961–1005, Sep 1990. ISSN: 0018-9448. doi: 10.1109/18.57199. Disponível em: <http://dx.doi.org/10.1109/18.57199>.

[26] DAUBECHIES, I. "Where do wavelets come from? A personal point of view", *Proceedings of the IEEE*, v. 84, n. 4, pp. 510–513, Apr 1996. ISSN: 0018-

9219. doi: 10.1109/5.488696. Disponível em: <http://dx.doi.org/10.1109/5.488696>.

[27] DAUBECHIES, I. "The wavelet transform, time-frequency localization and signal analysis", *IEEE Transactions on Information Theory*, v. 36, n. 5, pp. 961 –1005, Sep 1990. ISSN: 0018-9448. doi: 10.1109/18.57199. Disponível em: <http://dx.doi.org/10.1109/18.57199>.

[28] DONOHO, D. "Progress in wavelet analysis and WVD: a ten minute tour". In: Meyer, Y., Roques, S. (Eds.), *Progress in wavelet analysis and applications*, Frontières, pp. 109–128, 1993.

[29] DONOHO, D. "De-noising by soft-thresholding", *Information Theory, IEEE Transactions on*, v. 41, n. 3, pp. 613–627, May 1995. ISSN: 0018-9448. doi: 10.1109/18.382009.

[30] STOLLNITZ, E. J., DEROSE, T. D., SALESIN, D. H. "Wavelets for Computer Graphics: A Primer, Part 1", *IEEE Computer Graphics and Applications*, v. 15, n. 3, pp. 76–84, 1995. ISSN: 0272-1716. doi: 10.1109/38.376616.

[31] MANDUCA, A. "Compressing Images with Wavelets/Subband Coding". In: *IEEE Engineering in Medicine and Biology*, v. 14, pp. 639–646, Sep/Oct 1995.

[32] MULCAHY, C. "Image compression using the Haar wavelet transform", *Spelman Science and Math Journal*, 1997.

[33] CHRISTOPOULOS, C., SKODRAS, A., EBRAHIMI, T. "The JPEG2000 still image coding system: an overview", *Consumer Electronics, IEEE Transactions on*, v. 46, n. 4, pp. 1103–1127, Nov 2000. ISSN: 0098-3063. doi: 10.1109/30.920468.

[34] BRADLEY, J., BRISLAWN, C., HOPPER, T. "The FBI Wavelet/Scalar Quantization Standard for Gray-Scale Fingerprint Image Compression", *SPIE: Visual Information Processing II*, v. 1961, n. 4, pp. 293–304, 1993.

[35] DE OLIVEIRA, H. M., DE SOUZA, D. F. "Wavelet Analysis as an Information Processing Technique", *VI International Telecommunications Symposium, 2006. ITS'2006.*, pp. 7–12, Sep 2006. ISSN: 0306-4573. doi: 10.1109/ITS.2006.4433232. Disponível em: <http://dx.doi.org/10.1109/ITS.2006.4433232>.

[36] TICO, M., KUOSMANEN, P., SAARINEN, J. "Wavelet domain features for fingerprint recognition", *Electronics Letters*, v. 37, n. 1, pp. 21–22, Jan 2001. ISSN: 0013-5194. doi: 10.1049/el:20010022.

[37] CHANG, T., KUO, C.-C. "Texture analysis and classification with tree-structured wavelet transform", *Image Processing, IEEE Transactions on*, v. 2, n. 4, pp. 429 –441, oct 1993. ISSN: 1057-7149. doi: 10.1109/83.242353. Disponível em: <http://dx.doi.org/10.1109/83.242353>.

[38] LI, T., LI, Q., ZHU, S., et al. "A survey on wavelet applications in data mining", *SIGKDD Explor. Newsl.*, v. 4, n. 2, pp. 49–68, 2002. ISSN: 1931-0145. doi: 10.1145/772862.772870. Disponível em: <http://doi.acm.org/10.1145/772862.772870>.

[39] AL-DUBAEE, S., AHMAD, N. "New Direction of Applied Wavelet Transform in Multilingual Web Information Retrieval". In: *Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08.*, v. 4, pp. 198–202, Los Alamitos, CA, USA, Oct 2008. IEEE Computer Society. ISBN: 978-0-7695-3305-6. doi: 10.1109/FSKD.2008.551. Disponível em: <http://doi.ieeecomputersociety.org/10.1109/FSKD.2008.551>.

[40] AL-DUBAEE, S. A. R. S., AHMAD, N., ABDULLA, H. D., et al. "A New Search Result Clustering Using Haar Wavelet Transform". In: *Proceedings of the 2009 International Conference on Future Computer and Communication*, pp. 653–657, Washington, DC, USA, 2009. IEEE Computer Society. ISBN: 978-0-7695-3591-3. doi: 10.1109/ICFCC.2009.142. Disponível em: <http://dl.acm.org/citation.cfm?id=1584345.1585112>.

[41] MONTOYA ZEGARRA, J. A., LEITE, N. J., DA SILVA TORRES, R. "Wavelet-based fingerprint image retrieval", *J. Comput. Appl. Math.*, v. 227, pp. 294–307, May 2009. ISSN: 0377-0427. doi: 10.1016/j.cam.2008.03.017. Disponível em: <http://dl.acm.org/citation.cfm?id=1518333.1518571>.

[42] SUTAGUNDAR, A., MANVI, S., BHARAMGOUDAR, S. "Wavelet Based Image Indexing and Retrieval". In: *First International Conference on Emerging Trends in Engineering and Technology, 2008. ICETET'08.*, v. 0, pp. 52–55, Los Alamitos, CA, USA, Jul 2008. IEEE Computer Society. ISBN: 978-0-7695-3267-7. doi: 10.1109/ICETET.2008.129. Disponível em: <http://doi.ieeecomputersociety.org/10.1109/ICETET.2008.129>.

[43] MILLER, N. E., WONG, P. C., BREWSTER, M., et al. "TOPIC ISLANDS^{TM}- A Wavelet-Based Text Visualization System". In: *Proceedings IEEE Visualization, 1998. VIS'98.*, v. 0, pp. 189–196, Los Alamitos, CA, USA, Oct 1998. IEEE Computer Society. ISBN: 0-8186-9176-X. doi: 10.1109/VISUAL.1998.745302. Disponível em: <http://doi.ieeecomputersociety.org/10.1109/VISUAL.1998.745302>.

[44] "Apache Lucene". 2011. Disponível em: <http://lucene.apache.org/>.

[45] "Weka 3: Data Mining Software in Java". 2011. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/index.html>.

[46] MALA, T., GEETHA, T. V., KUMAR, S. "Fundamental Papers in Wavelet Theory". v. 4099, *Lecture Notes in Computer Science*, cap. Topical and Temporal Visualization Using Wavelets, pp. 839–843, Heidelberg, Berlin, Springer, Jul 2006. ISBN: 978-3-5403-6667-6. doi: 10.1007/11801603. Disponível em: <http://dx.dor.org/10.1007/11801603>.

[47] MITCHELL, T. M. *Machine learning*. McGraw Hill series in computer science. McGraw-Hill, 1997. ISBN: 978-0-0704-2807-2.

[48] MARSLAND, S. *Machine Learning: An Algorithmic Introduction*. New Jersey, USA, CRC Press, 2009. ISBN: 978-1-4200-6718-7.

[49] GREENGRASS, E. "Information Retrieval: A Survey". 2000. Disponível em: <http://www.cs.umbc.edu/cadip/readings/IR.report.120600.book.pdf>.

[50] BERRY, M. W., CASTELLANOS, M. "Text Mining: Clustering, Classification, and Retrieval". Springer Preface, 2007. Disponível em: <http://www.inma.ucl.ac.be/~blondel/publications/08-textmining.pdf>. Second Edition.

Apêndice A

Valores Numéricos dos Resultados

Este apêndice contém todos os valores numéricos dos resultados encontrados em todos os experimentos realizados para os fins deste presente trabalho. Esses valores estão representados graficamente nas figuras localizadas no Capítulo 4.

Tabela A.1: Valores numéricos dos resultados da Precisão e Abrangência com a técnica de RI sobre a coleção de documentos CF. (Representação gráfica na Figura 4.1)

Recall	**0**	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**	**1**
Modelo Vetorial (*Baseline*)	0,86059	0,63077	0,49359	0,36558	0,28651	0,21364	0,16160	0,11556	0,08557	0,03759	0,00841
Haar - Res. 14	0,83587	0,64251	0,50531	0,35629	0,27577	0,21742	0,16432	0,10888	0,07616	0,04404	0,00808
Haar - Res. 13	0,80406	0,60155	0,46879	0,33341	0,24953	0,19163	0,14144	0,09832	0,06794	0,03193	0,00470
Haar - Res. 12	0,58535	0,41211	0,29003	0,22902	0,17031	0,12828	0,07659	0,04849	0,02457	0,00325	0,00094
Haar - Res. 11	0,34714	0,24129	0,16274	0,12695	0,09499	0,07275	0,04586	0,02762	0,01937	0,01338	0,01296
Haar - Res. 10	0,26860	0,16860	0,12018	0,10394	0,07658	0,06199	0,04309	0,02726	0,02420	0,02002	0,01924
Haar - Res. 09	0,22471	0,11397	0,09975	0,09077	0,06543	0,05540	0,04449	0,03169	0,02999	0,02341	0,02244
Haar - Res. 08	0,15190	0,07847	0,06392	0,05882	0,04254	0,03876	0,03642	0,03113	0,03040	0,02963	0,02872
Haar - Res. 07	0,14096	0,06545	0,05665	0,04922	0,04432	0,04136	0,03766	0,03608	0,03428	0,03376	0,03251
Haar - Res. 06	0,15976	0,06194	0,05446	0,05171	0,04864	0,04644	0,04527	0,04231	0,04144	0,04022	0,03784
Haar - Res. 05	0,15692	0,05894	0,05110	0,04964	0,04794	0,04682	0,04495	0,04374	0,04301	0,04181	0,04049
Haar - Res. 04	0,15511	0,05819	0,05059	0,04914	0,04761	0,04622	0,04469	0,04326	0,04254	0,04150	0,04035
Haar - Res. 03	0,15496	0,05803	0,05077	0,04958	0,04754	0,04627	0,04491	0,04338	0,04251	0,04144	0,04022
Haar - Res. 02	0,15493	0,05790	0,05077	0,04963	0,04767	0,04631	0,04499	0,04346	0,04277	0,04156	0,04019
Haar - Res. 01	0,15492	0,05787	0,05072	0,04959	0,04762	0,04626	0,04497	0,04342	0,04273	0,04151	0,04017
Haar - Res. 00	0,15489	0,05784	0,05064	0,04953	0,04755	0,04620	0,04493	0,04339	0,04271	0,04157	0,04014
DaubD4 - Res. 14	0,83587	0,64251	0,50531	0,35629	0,27577	0,21742	0,16330	0,10659	0,07453	0,03559	0,00618
DaubD4 - Res. 13	0,76876	0,54895	0,42992	0,32037	0,25580	0,20326	0,15229	0,10453	0,07374	0,04046	0,00763
DaubD4 - Res. 12	0,74161	0,50880	0,39065	0,29751	0,24308	0,19291	0,14529	0,10353	0,07102	0,04273	0,00961
DaubD4 - Res. 11	0,63457	0,43118	0,31331	0,24082	0,19049	0,15146	0,11918	0,08847	0,06739	0,04595	0,01072
DaubD4 - Res. 10	0,54812	0,34194	0,24882	0,18414	0,15028	0,12114	0,09802	0,07409	0,05965	0,04533	0,01199
DaubD4 - Res. 09	0,51696	0,29379	0,19831	0,14840	0,12347	0,10189	0,08326	0,07057	0,06073	0,04824	0,01933
DaubD4 - Res. 08	0,39411	0,20968	0,15026	0,11771	0,09759	0,08498	0,07172	0,06362	0,05703	0,04788	0,02695
DaubD4 - Res. 07	0,25152	0,14165	0,11139	0,09943	0,08538	0,07423	0,06822	0,05803	0,05237	0,04542	0,02973
DaubD4 - Res. 06	0,22366	0,09864	0,07992	0,07181	0,06701	0,06200	0,05730	0,05187	0,04823	0,04480	0,03446
DaubD4 - Res. 05	0,19435	0,09512	0,07754	0,06667	0,06179	0,05670	0,05338	0,05009	0,04738	0,04439	0,04120
DaubD4 - Res. 04	0,15642	0,07257	0,06377	0,05880	0,05564	0,05305	0,05104	0,04860	0,04626	0,04374	0,04127
DaubD4 - Res. 03	0,13151	0,06551	0,05785	0,05343	0,05108	0,04951	0,04765	0,04593	0,04468	0,04299	0,04094
DaubD4 - Res. 02	0,11719	0,06217	0,05344	0,05178	0,05009	0,04805	0,04693	0,04566	0,04429	0,04260	0,04061
DaubD4 - Res. 01	0,12855	0,05902	0,05167	0,04876	0,04724	0,04598	0,04485	0,04394	0,04294	0,04196	0,04053
DaubD4 - Res. 00	0,15489	0,05784	0,05064	0,04953	0,04755	0,04620	0,04493	0,04339	0,04271	0,04157	0,04014

Tabela A.2: Valores numéricos dos resultados da *F1-Measure* com a técnica de RI sobre a coleção de documentos CF. (Representação gráfica na Figura 4.2)

	F1-Measure
Modelo Vetorial (*Baseline*)	0,33387
Haar - Res. 14	0,32647
Haar - Res. 13	0,31582
Haar - Res. 12	0,25975
Haar - Res. 11	0,17946
Haar - Res. 10	0,15439
Haar - Res. 09	0,13937
Haar - Res. 08	0,09836
Haar - Res. 07	0,08829
Haar - Res. 06	0,08821
Haar - Res. 05	0,08562
Haar - Res. 04	0,08509
Haar - Res. 03	0,08510
Haar - Res. 02	0,08519
Haar - Res. 01	0,08511
Haar - Res. 00	0,08502
DaubD4 - Res. 14	0,32647
DaubD4 - Res. 13	0,31205
DaubD4 - Res. 12	0,30239
DaubD4 - Res. 11	0,26717
DaubD4 - Res. 10	0,22821
DaubD4 - Res. 09	0,19915
DaubD4 - Res. 08	0,17160
DaubD4 - Res. 07	0,14936
DaubD4 - Res. 06	0,11588
DaubD4 - Res. 05	0,11175
DaubD4 - Res. 04	0,09833
DaubD4 - Res. 03	0,09071
DaubD4 - Res. 02	0,08903
DaubD4 - Res. 01	0,08450
DaubD4 - Res. 00	0,08502

Tabela A.3: Valores numéricos dos resultados da Precisão e Abrangência com a técnica de RI sobre a coleção de documentos TREC-DOE. (Representação gráfica na Figura 4.3)

Recall	**0**	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**	**1**
Modelo Vetorial (*Baseline*)	0,84803	0,81952	0,78750	0,76451	0,73378	0,71297	0,64043	0,57550	0,54166	0,44838	0,34834
Haar - Res. 14	0,77253	0,75374	0,74086	0,71947	0,67509	0,63604	0,57383	0,51217	0,47779	0,40161	0,32963
Haar - Res. 13	0,68855	0,65286	0,61440	0,56567	0,52617	0,49722	0,45500	0,39743	0,36793	0,31081	0,24270
Haar - Res. 12	0,48638	0,39821	0,37075	0,33577	0,31135	0,28901	0,25005	0,21334	0,18901	0,16283	0,12567
Haar - Res. 11	0,29599	0,25524	0,22456	0,20186	0,16919	0,15941	0,14513	0,12284	0,10411	0,07965	0,05705
Haar - Res. 10	0,22489	0,17811	0,16196	0,14475	0,13561	0,11930	0,10751	0,09258	0,08026	0,06472	0,05360
Haar - Res. 09	0,19861	0,13845	0,11622	0,09490	0,07816	0,07418	0,06843	0,05401	0,05142	0,04936	0,03415
Haar - Res. 08	0,08640	0,07120	0,05056	0,04671	0,04482	0,04028	0,03467	0,03132	0,02683	0,02645	0,02633
Haar - Res. 07	0,09856	0,05528	0,04697	0,04337	0,04215	0,03971	0,03380	0,02109	0,01888	0,01869	0,01840
Haar - Res. 06	0,07320	0,03704	0,02954	0,02805	0,02693	0,02630	0,02109	0,01501	0,01108	0,00922	0,00833
Haar - Res. 05	0,07392	0,03940	0,03009	0,02758	0,02657	0,02596	0,02099	0,01545	0,01238	0,00974	0,00921
Haar - Res. 04	0,07108	0,04033	0,03003	0,02747	0,02646	0,02550	0,02080	0,01551	0,01499	0,01443	0,01342
Haar - Res. 03	0,05904	0,03062	0,02016	0,01766	0,01647	0,01593	0,01551	0,01504	0,01473	0,01424	0,01396
Haar - Res. 02	0,05909	0,03065	0,02011	0,01776	0,01647	0,01593	0,01553	0,01504	0,01474	0,01424	0,01397
Haar - Res. 01	0,05603	0,02759	0,02002	0,01776	0,01646	0,01592	0,01552	0,01504	0,01474	0,01423	0,01398
Haar - Res. 00	0,05603	0,02759	0,02002	0,01776	0,01647	0,01593	0,01553	0,01505	0,01474	0,01423	0,01399
DaubD4 - Res. 14	0,77253	0,75374	0,74086	0,71947	0,67509	0,63604	0,57383	0,51217	0,47779	0,40161	0,32963
DaubD4 - Res. 13	0,72011	0,68732	0,66192	0,62293	0,57714	0,54399	0,48213	0,43966	0,39812	0,33805	0,26900
DaubD4 - Res. 12	0,67267	0,63669	0,60265	0,58400	0,54158	0,50029	0,43292	0,38363	0,35225	0,30358	0,24882
DaubD4 - Res. 11	0,65204	0,60018	0,56650	0,52177	0,49011	0,44641	0,39399	0,33860	0,30909	0,26288	0,20999
DaubD4 - Res. 10	0,56729	0,50677	0,48323	0,44940	0,40847	0,36979	0,32385	0,29244	0,26943	0,22516	0,17441
DaubD4 - Res. 09	0,48463	0,42001	0,39053	0,33869	0,28690	0,25735	0,23048	0,19957	0,17986	0,15160	0,11313
DaubD4 - Res. 08	0,42292	0,36916	0,31812	0,25837	0,20675	0,19087	0,16464	0,13884	0,11579	0,09955	0,08397
DaubD4 - Res. 07	0,30211	0,22423	0,17984	0,15325	0,12171	0,10913	0,08078	0,07111	0,05759	0,04885	0,04158
DaubD4 - Res. 06	0,17859	0,10148	0,08262	0,07385	0,06395	0,05899	0,04008	0,03376	0,02718	0,02339	0,02056
DaubD4 - Res. 05	0,09958	0,06367	0,04813	0,04063	0,03426	0,03145	0,02411	0,02196	0,01994	0,01771	0,01588
DaubD4 - Res. 04	0,10880	0,04312	0,03408	0,02989	0,02687	0,02485	0,02140	0,01929	0,01843	0,01691	0,01554
DaubD4 - Res. 03	0,05800	0,03497	0,03088	0,02896	0,02521	0,02397	0,01961	0,01836	0,01743	0,01663	0,01575
DaubD4 - Res. 02	0,06863	0,03824	0,03559	0,03270	0,02130	0,02032	0,01944	0,01589	0,01550	0,01496	0,01448
DaubD4 - Res. 01	0,05469	0,02199	0,01986	0,01909	0,01838	0,01782	0,01695	0,01621	0,01520	0,01459	0,01400
DaubD4 - Res. 00	0,05603	0,02759	0,02002	0,01776	0,01647	0,01593	0,01553	0,01505	0,01474	0,01423	0,01399

Tabela A.4: Valores numéricos dos resultados da *F1-Measure* com a técnica de RI sobre a coleção de documentos TREC-DOE. (Representação gráfica na Figura 4.4)

	F1-Measure
Modelo Vetorial (*Baseline*)	0,64596
Haar - Res. 14	0,59827
Haar - Res. 13	0,51754
Haar - Res. 12	0,36629
Haar - Res. 11	0,24175
Haar - Res. 10	0,20255
Haar - Res. 09	0,14701
Haar - Res. 08	0,08318
Haar - Res. 07	0,07626
Haar - Res. 06	0,05406
Haar - Res. 05	0,05653
Haar - Res. 04	0,05748
Haar - Res. 03	0,04688
Haar - Res. 02	0,04692
Haar - Res. 01	0,04325
Haar - Res. 00	0,04325
DaubD4 - Res. 14	0,59827
DaubD4 - Res. 13	0,54009
DaubD4 - Res. 12	0,50295
DaubD4 - Res. 11	0,47565
DaubD4 - Res. 10	0,42515
DaubD4 - Res. 09	0,33980
DaubD4 - Res. 08	0,27763
DaubD4 - Res. 07	0,20287
DaubD4 - Res. 06	0,11852
DaubD4 - Res. 05	0,07780
DaubD4 - Res. 04	0,06026
DaubD4 - Res. 03	0,05350
DaubD4 - Res. 02	0,06043
DaubD4 - Res. 01	0,03613
DaubD4 - Res. 00	0,04325

Tabela A.5: Valores numéricos dos resultados da Precisão e Abrangência com a técnica de RI sobre a coleção de documentos TREC-FR. (Representação gráfica na Figura 4.5)

Recall	**0**	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**	**1**
Modelo Vetorial (*Baseline*)	0,53470	0,49359	0,45821	0,42136	0,37418	0,34438	0,29292	0,24495	0,20423	0,18698	0,17161
Haar - Res. 16	0,49043	0,45062	0,42667	0,39258	0,33800	0,32072	0,27770	0,23930	0,21471	0,18192	0,15419
Haar - Res. 15	0,42072	0,38833	0,35935	0,32090	0,27691	0,25501	0,23278	0,20297	0,17046	0,14316	0,11753
Haar - Res. 14	0,29484	0,25760	0,23485	0,21060	0,18098	0,15954	0,14314	0,12775	0,10408	0,08958	0,06919
Haar - Res. 13	0,19827	0,17194	0,15404	0,12763	0,10115	0,09103	0,07725	0,06466	0,05727	0,04736	0,03388
Haar - Res. 12	0,14851	0,12120	0,10972	0,10248	0,08759	0,07895	0,06739	0,05766	0,05251	0,04566	0,02850
Haar - Res. 11	0,10599	0,08403	0,07693	0,07297	0,05882	0,05625	0,04067	0,03713	0,03535	0,03374	0,02026
Haar - Res. 10	0,06562	0,04781	0,04195	0,03586	0,03070	0,02760	0,02342	0,02234	0,02179	0,02047	0,01500
Haar - Res. 09	0,04302	0,01947	0,01626	0,01453	0,01374	0,01330	0,01122	0,01062	0,01020	0,00984	0,00734
Haar - Res. 08	0,03542	0,01881	0,01688	0,01625	0,01523	0,01422	0,01188	0,01121	0,01104	0,01009	0,00862
Haar - Res. 07	0,02939	0,01504	0,01396	0,01329	0,01263	0,01178	0,01045	0,00966	0,00924	0,00893	0,00831
Haar - Res. 06	0,02873	0,01476	0,01341	0,01319	0,01252	0,01183	0,01068	0,01036	0,00958	0,00911	0,00850
Haar - Res. 05	0,03263	0,02369	0,01639	0,01505	0,01169	0,01154	0,01079	0,01041	0,01013	0,00989	0,00880
Haar - Res. 04	0,03549	0,02270	0,02163	0,02041	0,01193	0,01178	0,01114	0,01051	0,01024	0,00996	0,00935
Haar - Res. 03	0,03572	0,02283	0,02173	0,02055	0,01214	0,01191	0,01126	0,01063	0,01041	0,00999	0,00963
Haar - Res. 02	0,03573	0,02282	0,02173	0,02054	0,01213	0,01191	0,01127	0,01064	0,01042	0,00999	0,00976
Haar - Res. 01	0,03572	0,02282	0,02172	0,02054	0,01214	0,01191	0,01127	0,01065	0,01043	0,01000	0,00977
Haar - Res. 00	0,03591	0,02291	0,02181	0,02072	0,01238	0,01221	0,01150	0,01099	0,01083	0,01057	0,01025
DaubD4 - Res. 16	0,49043	0,45062	0,42667	0,39258	0,33800	0,32072	0,27770	0,23930	0,21471	0,18192	0,15400
DaubD4 - Res. 15	0,46329	0,42597	0,39934	0,37237	0,32809	0,30549	0,26359	0,22331	0,19381	0,15681	0,13274
DaubD4 - Res. 14	0,44419	0,39842	0,37400	0,34979	0,30769	0,28546	0,24924	0,21287	0,18950	0,15440	0,13040
DaubD4 - Res. 13	0,42602	0,39040	0,36135	0,33587	0,29331	0,27255	0,23612	0,20499	0,18445	0,15701	0,13374
DaubD4 - Res. 12	0,38217	0,34060	0,31227	0,28436	0,24967	0,23224	0,20290	0,17254	0,14905	0,12260	0,10294
DaubD4 - Res. 11	0,32226	0,27445	0,25578	0,23077	0,19652	0,18866	0,16734	0,14437	0,12377	0,10427	0,08797
DaubD4 - Res. 10	0,28048	0,22793	0,20791	0,18501	0,16179	0,14856	0,13192	0,11376	0,09635	0,07893	0,06568
DaubD4 - Res. 09	0,24496	0,19702	0,18035	0,15182	0,12958	0,11554	0,09967	0,08305	0,07146	0,05890	0,04844
DaubD4 - Res. 08	0,18589	0,15485	0,13764	0,11125	0,09011	0,08271	0,06619	0,05270	0,04509	0,03711	0,03172
DaubD4 - Res. 07	0,14145	0,10273	0,07741	0,06600	0,06040	0,05318	0,04107	0,03602	0,03173	0,02704	0,02392
DaubD4 - Res. 06	0,10939	0,07599	0,06416	0,04468	0,04051	0,03695	0,02864	0,02283	0,01907	0,01562	0,01349
DaubD4 - Res. 05	0,09168	0,05700	0,04946	0,03880	0,03238	0,02758	0,02407	0,01975	0,01642	0,01295	0,01109
DaubD4 - Res. 04	0,06259	0,04120	0,02963	0,02232	0,02057	0,01912	0,01725	0,01529	0,01352	0,01074	0,00946
DaubD4 - Res. 03	0,07	0,04	0,03	0,02	0,02	0,02	0,01	0,01	0,01	0,01	0,01
DaubD4 - Res. 02	0,04	0,02	0,02	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01
DaubD4 - Res. 01	0,04	0,02	0,02	0,02	0,02	0,01	0,01	0,01	0,01	0,01	0,01
DaubD4 - Res. 00	0,04	0,02	0,02	0,02	0,01	0,01	0,01	0,01	0,01	0,01	0,01

Tabela A.6: Valores numéricos dos resultados da *F1-Measure* com a técnica de RI sobre a coleção de documentos TREC-FR. (Representação gráfica na Figura 4.6)

	F1-Measure
Modelo Vetorial (*Baseline*)	0,40785
Haar - Res. 16	0,39078
Haar - Res. 15	0,33776
Haar - Res. 14	0,24921
Haar - Res. 13	0,17908
Haar - Res. 12	0,15277
Haar - Res. 11	0,11739
Haar - Res. 10	0,06935
Haar - Res. 09	0,03259
Haar - Res. 08	0,03166
Haar - Res. 07	0,02615
Haar - Res. 06	0,02572
Haar - Res. 05	0,03831
Haar - Res. 04	0,03904
Haar - Res. 03	0,03920
Haar - Res. 02	0,03920
Haar - Res. 01	0,03918
Haar - Res. 00	0,03933
DaubD4 - Res. 16	0,39078
DaubD4 - Res. 15	0,37926
DaubD4 - Res. 14	0,36343
DaubD4 - Res. 13	0,35279
DaubD4 - Res. 12	0,31716
DaubD4 - Res. 11	0,27395
DaubD4 - Res. 10	0,23039
DaubD4 - Res. 09	0,20161
DaubD4 - Res. 08	0,16306
DaubD4 - Res. 07	0,11162
DaubD4 - Res. 06	0,09715
DaubD4 - Res. 05	0,07931
DaubD4 - Res. 04	0,05836
DaubD4 - Res. 03	0,05
DaubD4 - Res. 02	0,03
DaubD4 - Res. 01	0,03
DaubD4 - Res. 00	0,04

Tabela A.7: Valores numéricos dos resultados da Precisão e Abrangência com a técnica de RI sobre a coleção de documentos TREC-SJM. (Representação gráfica na Figura 4.7)

Recall	**0**	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**	**1**
Modelo Vetorial (*Baseline*)	0,81211	0,73415	0,65228	0,60874	0,55183	0,50339	0,43554	0,37569	0,29736	0,22228	0,12439
Haar - Res. 16	0,78121	0,68865	0,62223	0,55162	0,49768	0,45110	0,37771	0,31932	0,25668	0,20510	0,11160
Haar - Res. 15	0,78650	0,68286	0,62004	0,54832	0,49115	0,44902	0,37465	0,31760	0,25350	0,20359	0,11107
Haar - Res. 14	0,58872	0,48212	0,41297	0,34753	0,30860	0,27130	0,22564	0,18406	0,14519	0,10728	0,05222
Haar - Res. 13	0,40594	0,28273	0,24185	0,20119	0,17487	0,15075	0,11995	0,09995	0,07713	0,05730	0,02653
Haar - Res. 12	0,27385	0,17913	0,14712	0,13105	0,11128	0,09736	0,08400	0,06947	0,05372	0,03547	0,01649
Haar - Res. 11	0,16178	0,11278	0,08133	0,06405	0,05631	0,04756	0,03628	0,02843	0,02153	0,01314	0,00775
Haar - Res. 10	0,09594	0,05606	0,04144	0,02679	0,02252	0,01884	0,01493	0,01158	0,00881	0,00672	0,00477
Haar - Res. 09	0,07939	0,03281	0,02524	0,02103	0,01881	0,01192	0,00974	0,00864	0,00695	0,00497	0,00433
Haar - Res. 08	0,07704	0,02557	0,01955	0,01699	0,01564	0,00981	0,00889	0,00751	0,00699	0,00645	0,00615
Haar - Res. 07	0,05187	0,01523	0,01303	0,01189	0,01102	0,01022	0,00955	0,00906	0,00869	0,00844	0,00808
Haar - Res. 06	0,04377	0,01469	0,01320	0,01236	0,01146	0,01092	0,01028	0,00964	0,00913	0,00868	0,00843
Haar - Res. 05	0,04592	0,01469	0,01318	0,01235	0,01157	0,01117	0,01056	0,00990	0,00940	0,00892	0,00860
Haar - Res. 04	0,04595	0,01487	0,01318	0,01232	0,01158	0,01117	0,01057	0,00990	0,00940	0,00896	0,00862
Haar - Res. 03	0,04564	0,01489	0,01306	0,01227	0,01153	0,01117	0,01054	0,00989	0,00947	0,00903	0,00862
Haar - Res. 02	0,04564	0,01489	0,01306	0,01226	0,01152	0,01117	0,01054	0,00988	0,00947	0,00903	0,00862
Haar - Res. 01	0,04564	0,01489	0,01306	0,01226	0,01152	0,01117	0,01054	0,00991	0,00947	0,00903	0,00862
Haar - Res. 00	0,04573	0,01490	0,01303	0,01224	0,01151	0,01116	0,01054	0,00992	0,00948	0,00903	0,00863
DaubD4 - Res. 16	0,78121	0,68865	0,62223	0,55162	0,49768	0,45110	0,37764	0,31932	0,25668	0,20510	0,11159
DaubD4 - Res. 15	0,74829	0,64476	0,58820	0,51647	0,45577	0,41270	0,34128	0,28058	0,23608	0,18119	0,10330
DaubD4 - Res. 14	0,71001	0,59209	0,53042	0,46983	0,41534	0,37395	0,30942	0,25843	0,20839	0,15923	0,08965
DaubD4 - Res. 13	0,66963	0,55615	0,51525	0,43758	0,38770	0,34822	0,27980	0,22920	0,19127	0,15000	0,07972
DaubD4 - Res. 12	0,64731	0,50531	0,44905	0,37863	0,34168	0,30241	0,24276	0,19619	0,16347	0,11937	0,06006
DaubD4 - Res. 11	0,57337	0,45043	0,38832	0,32591	0,28084	0,24638	0,20223	0,16713	0,13414	0,09257	0,04284
DaubD4 - Res. 10	0,50530	0,35033	0,29354	0,23218	0,19466	0,16639	0,13440	0,10715	0,08119	0,05620	0,02383
DaubD4 - Res. 09	0,44569	0,28588	0,22818	0,18398	0,15071	0,12494	0,08631	0,06451	0,04906	0,03177	0,01576
DaubD4 - Res. 08	0,32059	0,18949	0,15012	0,11446	0,08176	0,06233	0,04845	0,03903	0,03144	0,02292	0,01599
DaubD4 - Res. 07	0,23469	0,12950	0,09413	0,07021	0,05198	0,04060	0,03210	0,02514	0,01987	0,01431	0,01030
DaubD4 - Res. 06	0,16770	0,08645	0,06236	0,04798	0,03885	0,02688	0,02334	0,01928	0,01517	0,01224	0,00969
DaubD4 - Res. 05	0,14067	0,04413	0,03325	0,02631	0,02322	0,01848	0,01623	0,01454	0,01227	0,01073	0,00896
DaubD4 - Res. 04	0,09582	0,02517	0,02006	0,01750	0,01533	0,01401	0,01265	0,01182	0,01076	0,00987	0,00885
DaubD4 - Res. 03	0,04	0,02	0,02	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01
DaubD4 - Res. 02	0,05	0,02	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01
DaubD4 - Res. 01	0,03	0,02	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01
DaubD4 - Res. 00	0,05	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01

Tabela A.8: Valores numéricos dos resultados da *F1-Measure* com a técnica de RI sobre a coleção de documentos TREC-SJM. (Representação gráfica na Figura 4.8)

Modelo Vetorial (*Baseline*)	0,50471
Haar - Res. 16	0,47429
Haar - Res. 15	0,47314
Haar - Res. 14	0,35174
Haar - Res. 13	0,24335
Haar - Res. 12	0,18242
Haar - Res. 11	0,11564
Haar - Res. 10	0,07184
Haar - Res. 09	0,04941
Haar - Res. 08	0,04073
Haar - Res. 07	0,02643
Haar - Res. 06	0,02562
Haar - Res. 05	0,02562
Haar - Res. 04	0,02589
Haar - Res. 03	0,02592
Haar - Res. 02	0,02592
Haar - Res. 01	0,02592
Haar - Res. 00	0,02594
DaubD4 - Res. 16	0,47429
DaubD4 - Res. 15	0,45217
DaubD4 - Res. 14	0,42788
DaubD4 - Res. 13	0,41053
DaubD4 - Res. 12	0,37688
DaubD4 - Res. 11	0,33010
DaubD4 - Res. 10	0,26188
DaubD4 - Res. 09	0,22808
DaubD4 - Res. 08	0,17151
DaubD4 - Res. 07	0,12801
DaubD4 - Res. 06	0,09508
DaubD4 - Res. 05	0,06124
DaubD4 - Res. 04	0,04022
DaubD4 - Res. 03	0,03
DaubD4 - Res. 02	0,03
DaubD4 - Res. 01	0,03
DaubD4 - Res. 00	0,03

Tabela A.9: Valores numéricos dos resultados da medida de avaliação de Precisão com a técnica de Classificação da Informação sobre a coleção de documentos Reuters-21578. (Representação gráfica na Figuras 4.9 e 4.14)

K	**1**	**3**	**5**	**15**	**30**	**45**	**60**	**75**	**90**	**Área**
KNN (*Baseline*)	0,5689742	0,5080285	0,4950238	0,4688780	0,4535022	0,4421296	0,4347496	0,4288922	0,4246017	79,9795323
Haar - Res. 15	0,5689742	0,5080285	0,4950238	0,4688780	0,4535022	0,4421296	0,4347496	0,4288922	0,4246017	79,9795323
Haar - Res. 14	0,7186463	0,6292325	0,6079607	0,5761929	0,5327409	0,4976014	0,4707096	0,4494294	0,4386032	90,7480590
Haar - Res. 13	0,7592047	0,7123107	0,7077922	0,7003175	0,6914428	0,6882317	0,6694561	0,6474654	0,6298774	120,7151366
Haar - Res. 12	0,7697568	0,7059612	0,7090054	0,6884197	0,6503918	0,6297994	0,6065934	0,5848757	0,5691517	112,7690014
Haar - Res. 11	0,8323699	0,7693441	0,7608696	0,7631012	0,7259615	0,7048748	0,6826004	0,6712500	0,6560976	126,3321478
Haar - Res. 10	0,6246300	0,5903614	0,5943448	0,5816273	0,5799685	0,5662218	0,5597771	0,5611365	0,5593135	101,6863438
Haar - Res. 09	0,4942693	0,4802900	0,4830853	0,4821192	0,4832098	0,4867608	0,4822134	0,4783748	0,4822538	85,9102740
Haar - Res. 08	0,4704370	0,4607201	0,4639175	0,4651259	0,4685548	0,4670833	0,4704403	0,4696780	0,4647132	83,2222987
Haar - Res. 07	0,4560195	0,4495631	0,4516514	0,4552450	0,4607158	0,4633745	0,4624897	0,4623612	0,4592046	81,8675364
Haar - Res. 06	0,4361047	0,4315431	0,4325758	0,4324631	0,4316981	0,4337258	0,4333966	0,4311061	0,4291073	76,9352748
Haar - Res. 05	0,4253089	0,4233197	0,4237351	0,4223534	0,4223534	0,4223534	0,4221976	0,4220420	0,4220994	75,1874365
Haar - Res. 04	0,4192124	0,4189636	0,4187500	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	74,4101764
Haar - Res. 03	0,4184085	0,4185280	0,4184085	0,4178832	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	74,4012309
Haar - Res. 02	0,4174899	0,4176707	0,4176707	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	74,3886080
Haar - Res. 01	0,4173977	0,4176707	0,4176707	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	74,3884234
Haar - Res. 00	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	74,3938731
DaubD4 - Res. 15	0,5689742	0,5080285	0,4950238	0,4688780	0,4535022	0,4421296	0,4347496	0,4288922	0,4246017	79,9795323
DaubD4 - Res. 14	0,5773994	0,5156605	0,4971727	0,4606109	0,4511811	0,4501178	0,4419143	0,4358584	0,4295352	80,5139640
DaubD4 - Res. 13	0,6830467	0,5859213	0,5727273	0,5174445	0,4948142	0,4872340	0,4843486	0,4786789	0,4719472	88,9500975
DaubD4 - Res. 12	0,6678679	0,5997882	0,6260388	0,7101631	0,7350260	0,7536913	0,7412772	0,7443857	0,7406680	129,3428577
DaubD4 - Res. 11	0,7611307	0,7184847	0,7158505	0,7195122	0,7622283	0,7787733	0,7777013	0,7681849	0,7426326	134,7203324
DaubD4 - Res. 10	0,7586207	0,6952441	0,6965174	0,6787221	0,6627771	0,6310249	0,6050955	0,5869342	0,5640138	112,6596400
DaubD4 - Res. 09	0,7627240	0,7006452	0,7073643	0,6983136	0,6686496	0,6554329	0,6514714	0,6415629	0,6340235	118,2980959
DaubD4 - Res. 08	0,7809735	0,7029126	0,7117762	0,7040816	0,6729028	0,6654783	0,6456647	0,6372881	0,6380576	118,7278371
DaubD4 - Res. 07	0,7723147	0,6850192	0,6946069	0,6828194	0,6650515	0,6494725	0,6374570	0,6267566	0,6183628	116,3280434
DaubD4 - Res. 06	0,7759815	0,6730646	0,6843811	0,6813820	0,6716698	0,6553398	0,6436578	0,6355685	0,6316397	117,1530171
DaubD4 - Res. 05	0,7365506	0,6460459	0,6555484	0,6508135	0,6361985	0,6211144	0,6110472	0,6086462	0,6055305	111,5873497
DaubD4 - Res. 04	0,6717131	0,5914634	0,6003595	0,6079714	0,6040698	0,5986199	0,5895692	0,5878378	0,5830094	106,2609215
DaubD4 - Res. 03	0,6250999	0,5462329	0,5597484	0,5605307	0,5680934	0,5676275	0,5617792	0,5598477	0,5630803	100,3320172
DaubD4 - Res. 02	0,5330261	0,4944824	0,4992183	0,4962293	0,4962594	0,4916339	0,4943210	0,4913920	0,4900340	87,9990308
DaubD4 - Res. 01	0,4376489	0,4405665	0,4384766	0,4350506	0,4355725	0,4365338	0,4248418	0,4236893	0,4244164	76,7609138
DaubD4 - Res. 00	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	74,3938731

Tabela A.10: Valores numéricos dos resultados da medida de avaliação de Abrangência com a técnica de Classificação da Informação sobre a coleção de documentos Reuters-21578. (Representação gráfica nas Figuras 4.10 e 4.15)

K	**1**	**3**	**5**	**15**	**30**	**45**	**60**	**75**	**90**	**Área**
KNN (*Baseline*)	0,9825480	0,9938918	0,9982548	0,9991274	1,0000000	1,0000000	1,0000000	1,0000000	1,0000000	177,8979058
Haar - Res. 15	0,9825480	0,9938918	0,9982548	0,9991274	1,0000000	1,0000000	1,0000000	1,0000000	1,0000000	177,8979058
Haar - Res. 14	0,9450262	0,9729494	0,9729494	0,9799302	0,9938918	0,9956370	0,9956370	0,9965096	0,9973822	176,3664921
Haar - Res. 13	0,8996510	0,9441536	0,9511344	0,9624782	0,9659686	0,9746946	0,9773124	0,9808028	0,9860384	172,8054101
Haar - Res. 12	0,8839442	0,9197208	0,9205934	0,9389180	0,9415358	0,9589878	0,9633508	0,9650960	0,9659686	169,3254799
Haar - Res. 11	0,8795812	0,9109948	0,9162304	0,9275742	0,9223386	0,9336824	0,9345550	0,9371728	0,9389180	165,5034904
Haar - Res. 10	0,9205934	0,9406632	0,9537522	0,9668412	0,9650960	0,9624782	0,9642234	0,9650960	0,9668412	171,4293194
Haar - Res. 09	0,9031414	0,9249564	0,9345550	0,9528796	0,9668412	0,9624782	0,9581152	0,9554974	0,9485166	170,0584642
Haar - Res. 08	0,9581152	0,9825480	0,9816754	0,9834206	0,9816754	0,9781850	0,9790576	0,9799302	0,9825480	174,5157068
Haar - Res. 07	0,9816754	0,9877836	0,9904014	0,9808028	0,9773124	0,9825480	0,9790576	0,9808028	0,9773124	174,5706806
Haar - Res. 06	0,9886562	0,9956370	0,9965096	0,9973822	0,9982548	0,9965096	0,9965096	0,9965096	0,9982548	177,4598604
Haar - Res. 05	0,9912740	0,9947644	0,9938918	0,9991274	0,9991274	0,9991274	0,9991274	0,9991274	1,0000000	177,7617801
Haar - Res. 04	0,9938918	0,9947644	0,9938918	1,0000000	1,0000000	1,0000000	1,0000000	1,0000000	1,0000000	177,8935428
Haar - Res. 03	0,9956370	0,9973822	0,9956370	0,9991274	1,0000000	1,0000000	1,0000000	1,0000000	1,0000000	177,9066318
Haar - Res. 02	0,9956370	0,9982548	0,9982548	1,0000000	1,0000000	1,0000000	1,0000000	1,0000000	1,0000000	177,9633508
Haar - Res. 01	0,9965096	0,9982548	0,9982548	1,0000000	1,0000000	1,0000000	1,0000000	1,0000000	1,0000000	177,9650960
Haar - Res. 00	1,0000000	1,0000000	1,0000000	1,0000000	1,0000000	1,0000000	1,0000000	1,0000000	1,0000000	178,0000000
DaubD4 - Res. 15	0,9825480	0,9938918	0,9982548	0,9991274	1,0000000	1,0000000	1,0000000	1,0000000	1,0000000	177,8979058
DaubD4 - Res. 14	0,9764398	0,9912740	0,9973822	1,0000000	1,0000000	1,0000000	0,9991274	0,9991274	1,0000000	177,8342059
DaubD4 - Res. 13	0,9703316	0,9877836	0,9895288	0,9965096	0,9991274	0,9991274	0,9991274	0,9991274	0,9982548	177,5479930
DaubD4 - Res. 12	0,9703316	0,9886562	0,9860384	0,9877836	0,9851658	0,9799302	0,9825480	0,9834206	0,9869110	175,1579407
DaubD4 - Res. 11	0,9397906	0,9598604	0,9694590	0,9781850	0,9790576	0,9860384	0,9860384	0,9860384	0,9895288	174,7652705
DaubD4 - Res. 10	0,9406632	0,9694590	0,9773124	0,9825480	0,9912740	0,9938918	0,9947644	0,9956370	0,9956370	176,2521815
DaubD4 - Res. 09	0,9284468	0,9476440	0,9554974	0,9755672	0,9808028	0,9842932	0,9851658	0,9886562	0,9886562	174,5000000
DaubD4 - Res. 08	0,9240838	0,9476440	0,9546248	0,9633508	0,9729494	0,9773124	0,9746946	0,9842932	0,9860384	173,2460733
DaubD4 - Res. 07	0,8909250	0,9336824	0,9328098	0,9467714	0,9581152	0,9668412	0,9712042	0,9729494	0,9755672	171,0863874
DaubD4 - Res. 06	0,8795812	0,9179756	0,9214660	0,9293194	0,9371728	0,9424084	0,9520070	0,9511344	0,9546248	167,5226876
DaubD4 - Res. 05	0,8123909	0,8839442	0,8917976	0,9075044	0,9171030	0,9240838	0,9267016	0,9336824	0,9363002	163,6413613
DaubD4 - Res. 04	0,7356021	0,8464223	0,8743456	0,8917976	0,9066318	0,9083770	0,9075044	0,9109948	0,9162304	160,3926702
DaubD4 - Res. 03	0,6823735	0,8350785	0,8542757	0,8848168	0,8917976	0,8935428	0,8926702	0,8979058	0,8996510	157,8490401
DaubD4 - Res. 02	0,6055846	0,8211169	0,8359511	0,8612565	0,8682373	0,8717277	0,8734729	0,8717277	0,8795812	153,8071553
DaubD4 - Res. 01	0,4808028	0,7600349	0,7835951	0,8621291	0,8996510	0,9363002	0,9371728	0,9520070	0,9677138	161,2277487
DaubD4 - Res. 00	1,0000000	1,0000000	1,0000000	1,0000000	1,0000000	1,0000000	1,0000000	1,0000000	1,0000000	178,0000000

Tabela A.11: Valores numéricos dos resultados da medida de avaliação de *F1-Measure* com a técnica de Classificação da Informação sobre a coleção de documentos Reuters-21578. (Representação gráfica nas Figuras 4.11 e 4.16)

K	**1**	**3**	**5**	**15**	**30**	**45**	**60**	**75**	**90**	**Área**
KNN (*Baseline*)	0,7206400	0,6723731	0,6618455	0,6382386	0,6240131	0,6131621	0,6060286	0,6003143	0,5960988	110,2759068
Haar - Res. 15	0,7206400	0,6723731	0,6618455	0,6382386	0,6240131	0,6131621	0,6060286	0,6003143	0,5960988	110,2759068
Haar - Res. 14	0,8164342	0,7642221	0,7483221	0,7256866	0,6936663	0,6635650	0,6392157	0,6194738	0,6092751	119,4285360
Haar - Res. 13	0,8234824	0,8120075	0,8116158	0,8107313	0,8059701	0,8067895	0,7946080	0,7800139	0,7687075	142,0547253
Haar - Res. 12	0,8229082	0,7987874	0,8010630	0,7943891	0,7693405	0,7602906	0,7444370	0,7283503	0,7162731	135,1300945
Haar - Res. 11	0,8553246	0,8341990	0,8313539	0,8373375	0,8124520	0,8033033	0,7889503	0,7822287	0,7724336	143,1516614
Haar - Res. 10	0,7442681	0,7254374	0,7323283	0,7263192	0,7245332	0,7129929	0,7083333	0,7096567	0,7086665	127,6316890
Haar - Res. 09	0,6388889	0,6322696	0,6369313	0,6402814	0,6443734	0,6465416	0,6415425	0,6375546	0,6394118	114,1486067
Haar - Res. 08	0,6310345	0,6272981	0,6300756	0,6315495	0,6343389	0,6322617	0,6355140	0,6350014	0,6309891	112,6992213
Haar - Res. 07	0,6227512	0,6179039	0,6203881	0,6218534	0,6262231	0,6297539	0,6282195	0,6284596	0,6248257	111,4601776
Haar - Res. 06	0,6052350	0,6021108	0,6032752	0,6033254	0,6027397	0,6043927	0,6040730	0,6018445	0,6002099	107,3359979
Haar - Res. 05	0,5952319	0,5939047	0,5941575	0,5937257	0,5937257	0,5937257	0,5935718	0,5934180	0,5936286	105,6767792
Haar - Res. 04	0,5896971	0,5896043	0,5892395	0,5895062	0,5895062	0,5895062	0,5895062	0,5895062	0,5895062	104,9296736
Haar - Res. 03	0,5892073	0,5896312	0,5892073	0,5892949	0,5895062	0,5895062	0,5895062	0,5895062	0,5895062	104,9231332
Haar - Res. 02	0,5882960	0,5889318	0,5889318	0,5895062	0,5895062	0,5895062	0,5895062	0,5895062	0,5895062	104,9204882
Haar - Res. 01	0,5883565	0,5889318	0,5889318	0,5895062	0,5895062	0,5895062	0,5895062	0,5895062	0,5895062	104,9206093
Haar - Res. 00	0,5895062	0,5895062	0,5895062	0,5895062	0,5895062	0,5895062	0,5895062	0,5895062	0,5895062	104,9320988
DaubD4 - Res. 15	0,7206400	0,6723731	0,6618455	0,6382386	0,6240131	0,6131621	0,6060286	0,6003143	0,5960988	110,2759068
DaubD4 - Res. 14	0,7256809	0,6784115	0,6635704	0,6307100	0,6218123	0,6208017	0,6127910	0,6069441	0,6009439	110,7802323
DaubD4 - Res. 13	0,8017304	0,7355426	0,7255278	0,6811810	0,6618497	0,6550343	0,6524217	0,6472583	0,6408964	118,3918569
DaubD4 - Res. 12	0,7911775	0,7466227	0,7658421	0,8262774	0,8419090	0,8520486	0,8450281	0,8473684	0,8462402	148,7001124
DaubD4 - Res. 11	0,8410777	0,8218155	0,8235730	0,8291420	0,8571429	0,8702349	0,8695652	0,8635843	0,8484848	152,1239330
DaubD4 - Res. 10	0,8398909	0,8097668	0,8133624	0,8028521	0,7944056	0,7719417	0,7524752	0,7385113	0,7201010	137,2720299
DaubD4 - Res. 09	0,8374656	0,8056380	0,8129176	0,8139789	0,7951892	0,7868852	0,7843001	0,7781593	0,7725878	140,9268014
DaubD4 - Res. 08	0,8465228	0,8071349	0,8155050	0,8135593	0,7955762	0,7917992	0,7767733	0,7736626	0,7747686	140,8024947
DaubD4 - Res. 07	0,8273906	0,7902511	0,7962756	0,7934186	0,7851269	0,7769986	0,7697095	0,7623932	0,7569397	138,3875020
DaubD4 - Res. 06	0,8245399	0,7766704	0,7854221	0,7862680	0,7825137	0,7730852	0,7680394	0,7619713	0,7602502	137,8095669
DaubD4 - Res. 05	0,7726141	0,7464996	0,7556377	0,7580175	0,7512509	0,7428972	0,7364771	0,7369146	0,7354352	132,6070401
DaubD4 - Res. 04	0,7022074	0,6963388	0,7119005	0,7230279	0,7250523	0,7216638	0,7147766	0,7145791	0,7125891	127,7792654
DaubD4 - Res. 03	0,6524823	0,6604555	0,6763385	0,6862944	0,6940577	0,6942373	0,6895854	0,6896783	0,6926436	122,6366252
DaubD4 - Res. 02	0,5669935	0,6172516	0,6251223	0,6296651	0,6315455	0,6286973	0,6313466	0,6284995	0,6294099	111,8899053
DaubD4 - Res. 01	0,4582121	0,5577970	0,5623043	0,5782850	0,5869627	0,5954495	0,5846489	0,5864015	0,5900505	103,8070262
DaubD4 - Res. 00	0,5895062	0,5895062	0,5895062	0,5895062	0,5895062	0,5895062	0,5895062	0,5895062	0,5895062	104,9320988

Tabela A.12: Valores numéricos dos resultados da medida de avaliação de *Accuracy* com a técnica de Classificação da Informação sobre a coleção de documentos Reuters-21578. (Representação gráfica nas Figuras 4.12 e 4.17)

K	**1**	**3**	**5**	**15**	**30**	**45**	**60**	**75**	**90**	**Área**
KNN (*Baseline*)	0,6816193	0,5951860	0,5736689	0,5266229	0,4963530	0,4726477	0,4566010	0,4434719	0,4336251	86,3701678
Haar - Res. 15	0,6816193	0,5951860	0,5736689	0,5266229	0,4963530	0,4726477	0,4566010	0,4434719	0,4336251	86,3701678
Haar - Res. 14	0,8223924	0,7490883	0,7264770	0,6903720	0,6331145	0,5780452	0,5302699	0,4883297	0,4653538	104,4912473
Haar - Res. 13	0,8388038	0,8172867	0,8154632	0,8121809	0,8056163	0,8048869	0,7888403	0,7687819	0,7520058	141,3606856
Haar - Res. 12	0,8409920	0,8063457	0,8088986	0,7968636	0,7640408	0,7472648	0,7235594	0,6991247	0,6801605	132,7578410
Haar - Res. 11	0,8756382	0,8486506	0,8446390	0,8493800	0,8220277	0,8088986	0,7910284	0,7819110	0,7687819	144,1637491
Haar - Res. 10	0,7355945	0,7024070	0,7086069	0,6954778	0,6932896	0,6761488	0,6681255	0,6699489	0,6677608	121,4128373
Haar - Res. 09	0,5733042	0,5503282	0,5547046	0,5525164	0,5539752	0,5601751	0,5525164	0,5459519	0,5528811	98,4890591
Haar - Res. 08	0,5317287	0,5120350	0,5182349	0,5204230	0,5269876	0,5244347	0,5306346	0,5291758	0,5196937	93,4733771
Haar - Res. 07	0,5029176	0,4894238	0,4934354	0,5014588	0,5123997	0,5171408	0,5156820	0,5153173	0,5094821	90,8796499
Haar - Res. 06	0,4609774	0,4500365	0,4522247	0,4518600	0,4500365	0,4547775	0,4540481	0,4489424	0,4442013	80,3424508
Haar - Res. 05	0,4365427	0,4314369	0,4325310	0,4285193	0,4285193	0,4285193	0,4281546	0,4277899	0,4277899	76,3085339
Haar - Res. 04	0,4219548	0,4212254	0,4208607	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	74,4500365
Haar - Res. 03	0,4197666	0,4197666	0,4197666	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	74,4266958
Haar - Res. 02	0,4175784	0,4175784	0,4175784	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	74,3873085
Haar - Res. 01	0,4172137	0,4175784	0,4175784	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	74,3865791
Haar - Res. 00	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	74,3938731
DaubD4 - Res. 15	0,6816193	0,5951860	0,5736689	0,5266229	0,4963530	0,4726477	0,4566010	0,4434719	0,4336251	86,3701678
DaubD4 - Res. 14	0,6914661	0,6072210	0,5773158	0,5105762	0,4916120	0,4894238	0,4722830	0,4591539	0,4449307	87,5521517
DaubD4 - Res. 13	0,7994165	0,7031364	0,6870897	0,6101386	0,5733042	0,5601751	0,5550693	0,5448578	0,5324581	102,8989788
DaubD4 - Res. 12	0,7859227	0,7195478	0,7479942	0,8264041	0,8453683	0,8577681	0,8493800	0,8519329	0,8501094	148,9711889
DaubD4 - Res. 11	0,8515682	0,8260394	0,8264041	0,8315098	0,8636032	0,8770970	0,8763676	0,8698031	0,8522976	153,1024799
DaubD4 - Res. 10	0,8501094	0,8096280	0,8125456	0,7983224	0,7855580	0,7545587	0,7264770	0,7053246	0,6765135	133,9525894
DaubD4 - Res. 09	0,8493800	0,8088986	0,8161926	0,8136397	0,7888403	0,7771699	0,7735230	0,7644055	0,7567469	139,5390226
DaubD4 - Res. 08	0,8599562	0,8107221	0,8194748	0,8154632	0,7910284	0,7851933	0,7658643	0,7592998	0,7603939	139,6305616
DaubD4 - Res. 07	0,8446390	0,7928519	0,8005106	0,7939460	0,7808169	0,7680525	0,7571116	0,7465354	0,7381473	136,9631656
DaubD4 - Res. 06	0,8435449	0,7793581	0,7895697	0,7888403	0,7822757	0,7687819	0,7596645	0,7516411	0,7483589	137,0966448
DaubD4 - Res. 05	0,8001459	0,7490883	0,7589351	0,7578410	0,7461707	0,7326769	0,7228301	0,7213713	0,7184537	131,1181619
DaubD4 - Res. 04	0,7392414	0,6914661	0,7042305	0,7144420	0,7126185	0,7071481	0,6973012	0,6958425	0,6911014	125,3099927
DaubD4 - Res. 03	0,6962071	0,6411379	0,6582786	0,6619256	0,6714077	0,6710430	0,6641138	0,6622903	0,6663020	118,4646244
DaubD4 - Res. 02	0,6134209	0,5743982	0,5809628	0,5765864	0,5765864	0,5696572	0,5736689	0,5692925	0,5671043	102,0933625
DaubD4 - Res. 01	0,5247994	0,4963530	0,4901532	0,4744712	0,4708242	0,4682713	0,4434719	0,4387309	0,4380015	81,9876003
DaubD4 - Res. 00	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	0,4179431	74,3938731

Tabela A.13: Valores numéricos dos resultados da medida de avaliação de *Log-Likelihood* com a técnica de Agrupamento da Informação sobre a coleção de documentos Reuters-21578. (Representação gráfica na Figura 4.13

K	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Área**
Kmeans (*Baseline*)	-	-	-	-	-	-	-	-	-48878508927	-
Haar - Res. 15	-	-	-	-	-	-	-	-	-48878507328	-
Haar - Res. 14	-	-	-	-	-	-	-	-	-	-
Haar - Res. 13	-	-	-	-	-	-	-	-	-747301,2343000	-
Haar - Res. 12	-	-	-	-	-	-	-	-	18977,8034600	-
Haar - Res. 11	-	-	-	-	-	-	-	-	10780,2956400	-
Haar - Res. 10	-	-	-	-	-	-	-	-	5011,0849090	-
Haar - Res. 09	-	-	-	-	-	-	-	-	1782,5220900	-
Haar - Res. 08	-	-	-	-	-	-	-	-	744,0863100	-
Haar - Res. 07	-	-	-	-	-	-	-	-	334,0714307	-
Haar - Res. 06	-	-	-	-	-	-	-	-	154,5221151	-
Haar - Res. 05	-	-	-	-	-	-	-	-	-	-
Haar - Res. 04	-	-	-	-	-	-	-	-	-	-
Haar - Res. 03	-	-	-	-	-	-	-	-	-	-
Haar - Res. 02	-	-	-	-	-	-	-	-	-	-
Haar - Res. 01	-	-	-	-	-	-	-	-	-	-
Haar - Res. 00	-	-	-	-	-	-	-	-	-	-
DaubD4 - Res. 15	-	-	-	-	-	-	-	-	-	-
DaubD4 - Res. 14	-	-	-	-	-	-	-	-	-	-
DaubD4 - Res. 13	-	-	-	-	-	-	-	-	-	-
DaubD4 - Res. 12	-	-	-	-	-	-	-	-	-	-
DaubD4 - Res. 11	-	-	-	-	-	-	-	-	-	-
DaubD4 - Res. 10	-	-	-	-	-	-	-	-	-	-
DaubD4 - Res. 09	-	-	-	-	-	-	-	-	-	-
DaubD4 - Res. 08	-	-	-	-	-	-	-	-	-	-
DaubD4 - Res. 07	-	-	-	-	-	-	-	-	-	-
DaubD4 - Res. 06	-	-	-	-	-	-	-	-	-	-
DaubD4 - Res. 05	-	-	-	-	-	-	-	-	-	-
DaubD4 - Res. 04	-	-	-	-	-	-	-	-	-	-
DaubD4 - Res. 03	-	-	-	-	-	-	-	-	-	-
DaubD4 - Res. 02	-	-	-	-	-	-	-	-	-	-
DaubD4 - Res. 01	-	-	-	-	-	-	-	-	-	-
DaubD4 - Res. 00	-	-	-	-	-	-	-	-	-	-