

# ExperText: Uma Ferramenta de Combinação de Múltiplos Classificadores *Naive Bayes*

Grinaldo Oliveira<sup>1</sup>, Manoel Mendonça<sup>2</sup>

<sup>1</sup> Unitech Tecnologia de Informação  
Av. ACM, 2487, CEP: 40280-630, Salvador, BA – Brasil  
grinaldo@unitech.com.br

<sup>2</sup> Universidade Salvador  
CEPERC – Centro de Pesquisas Interdepartamental em Redes de Computadores  
Rua Ponciano de Oliveira, 126, CEP: 41950-275, Salvador, BA – Brasil  
mgmn@unifacs.br

**Abstract.** Managing the ever increasing number of digital documents is a challenge to be faced in modern organizations. It is argued that text mining techniques may help to extract non-trivial information from unstructured document repositories. One of the tasks in which text mining techniques can help is the classification of documents, this task consists in classifying documents elaborated in natural language in pre-established categories. Bayesian algorithms have successfully been used to build classification models from document training samples. However, it is perceived that the accuracy of these classifiers depends on its accumulated learning. This usually demands a great mass of labeled information and, consequently, time and attention of specialists. This work presents an approach for combining multiple Bayesian classifiers. This strategy allows for distributed and incremental labeling of training sets by different specialists. The models produced for each set can then be combined into one classification model. Over the whole training set, the combined classifier has better overall performance than the individual classifiers. This paper discusses this approach and presents a tool – ExperText – that implements it.

**Keywords:** Knowledge Management, Text Mining, Learning Machine, Classification Algorithms, Theorem of Bayes.

**Resumo.** Gerenciar o crescente número de documentos digitais é um desafio a ser enfrentado pelas organizações modernas. Nesta linha, as técnicas de mineração de textos podem ajudar a extrair informações não-triviais de repositórios de documentos não estruturados. Uma destas técnicas, a classificação de documentos, consiste em classificar documentos elaborados em linguagem natural em categorias pré-estabelecidas. Algoritmos Bayesianos têm sido usados com sucesso na construção de modelos de classificação a partir de um conjunto de amostras de treinamentos. Entretanto, é percebido que a precisão destes classificadores dependem de seu conhecimento acumulado. Isto usualmente demanda uma grande massa de informação rotulada e, conseqüentemente, tempo e dedicação de especialistas. Este trabalho apresenta um método de combinação de múltiplos classificadores Bayesianos. Esta estratégia permite o uso de conjuntos de treinamentos distintos que foram rotulados por diferentes especialistas de forma distribuída e incremental. Os modelos produzidos por cada conjunto podem ser combinados em um único modelo de classificação. Em comparação a todos os modelos de treinamentos utilizados, o classificador combinado tem melhor performance que classificadores individuais. Este trabalho discute este método e apresenta a ferramenta – ExperText – que o implementa.

**Palavras-Chave:** Gestão do Conhecimento, Mineração de Textos, Aprendizado de Máquina, Algoritmos Classificadores, Teorema de Bayes.

## 1. Introdução

Organizações modernas migraram de uma era puramente industrial para uma outra baseada em conhecimento. Mais que isto, uma parcela significativa do conhecimento destas organizações está contido em meios digitais. Desta forma, ferramentas e técnicas que consigam extrair conhecimento destes meios digitais são de grande valia para estas organizações.

A criação de mapas do conhecimento, que funcionem como índice, e mostrem como encontrar a informação que se necessita, pode ajudar em muito o processo de busca por conhecimento [1]. Idéia semelhante pode ser aplicada ao mapeamento de repositórios de conhecimento explícito e não estruturado, geralmente na forma de conjuntos de documentos [2].

Entretanto, como montar mapas de documentos textuais, se quando estes são criados não pressupõem que seus autores se preocupem em rotulá-los antes de armazená-los em um sistema computacional? Por razões da própria natureza humana, uma abordagem manual seria inviável, pois, o tempo de análise de uma grande quantidade de documentos seria extremamente dispendioso em prazo e dinheiro.

O emprego de técnicas de mineração de textos pode auxiliar no mapeamento de documentos. Neste escopo, o uso de abordagens de mineração de texto para classificação de documentos é de grande interesse. Uma categoria de algoritmos de aprendizado de máquina que tem sido usada com sucesso para classificação de documentos são os classificadores estatísticos *Naive Bayes*. Eles se baseiam na aquisição de conhecimento sobre categorias distintas de documentos, a partir da representação interna de um texto em um modelo composto por um conjunto de probabilidades.

O objetivo da ferramenta apresentada neste trabalho é criar automaticamente um classificador combinando a *expertise* de múltiplos classificadores de textos baseados no teorema estatístico de Bayes. A principal contribuição deste trabalho é oferecer um mecanismo de estudo e combinação do aprendizado de vários classificadores em uma base única de conhecimento, a fim de compor um único classificador de *expertise* melhorada.

O resto deste artigo está organizado da seguinte forma. A seção 2 aborda conceitos sobre a gestão do conhecimento e os problemas associados à gerência eletrônica de documentos. A seção 3 apresenta o tema mineração de textos. A seção 4 descreve o processo de classificação de textos através de aprendizado de máquina e do algoritmo *Naive Bayes*. A seção 5 descreve a metodologia de combinação de *expertise* dos classificadores. A seção 6 descreve a ferramenta ExperText e seu modelo de implementação. E, finalmente, a seção 7 apresenta as conclusões do trabalho.

## 2. Gestão do Conhecimento

O conceito de gestão do conhecimento pode ser entendido como o conjunto de processos que direcionam a criação, utilização e disseminação do conhecimento na organização, de forma a atingir seu objetivo de negócio. Desta forma, é dito que a administração do conhecimento não é mais do que o gerenciamento do fluxo da

informação certa para as pessoas que precisam dela a fim de que possam agir com rapidez [3]. Nesta lógica, nos últimos 10 anos, as tarefas relacionadas ao gerenciamento baseado no conteúdo de documentos ou recuperação de informação, como é mais comumente conhecido, ganharam importância significativa em virtude da crescente disponibilidade de documentos na forma digital e sua necessidade de acesso de forma mais flexível possível [4].

Um documento pode ser percebido como um objeto que contém elos e regras que o associam a outros componentes informacionais. Neste caso, o documento está deixando de ser apenas uma entidade física, e sim uma entidade lógica e dinâmica que faz parte dos ativos de uma empresa [5].

Neste novo paradigma, esta entidade não é necessariamente algo que consigamos carregar em nossas mãos. Entretanto, como no mundo real, quando as pessoas encontram rapidamente e facilmente os documentos de que precisam, elas são capazes de investir seu tempo em trabalho efetivo, ao invés de gastá-lo tentando localizá-los sem sucesso [6].

Com o rápido progresso das tecnologias de rede e computadores, tornou-se fácil coletar e armazenar uma grande quantidade de textos, não estruturados ou semi-estruturados, como páginas armazenadas na grande rede, arquivos com linguagem de marcação HTML/XML, mensagens de correio eletrônico e arquivos em formato texto [7]. Estes tipos de dados, por não possuir uma estruturação bem definida de seu conteúdo, dificultam sua localização e rápido entendimento.

### **3. Mineração de Textos**

A mineração de textos, também conhecida como mineração de informação documental, mineração de dados textuais, ou descoberta de bancos de dados textuais é uma tecnologia emergente para análise de grandes coleções de documentos não estruturados para os propósitos de extração de padrões ou conhecimentos interessantes e não triviais [8].

Alguns problemas típicos envolvendo a busca por conhecimento em arquivos com conteúdo puramente lingüísticos têm sido resolvidos com a mineração de textos:

- **Identificação de idiomas:** A identificação de idiomas é indicada para descobrir a língua em que o texto foi escrito, ou percentual de participação, no caso do emprego de mais de uma língua.
- **Extração e seleção de características:** Indicado para reconhecer itens significativos do vocabulário empregado no texto. Entre exemplos de características reconhecidas, podem ser citados nomes de pessoas, organizações ou lugares, abreviações, datas, valores em moeda corrente e outros tipos de itens qualificados.
- **Aglomeración:** É uma técnica que divide uma coleção de documentos em grupos. Os documentos de cada grupo são homogêneos entre si. A aglomeración divide uma população com base na auto-similaridade entre os dados.
- **Sumarização:** Consiste em identificar segmentos relevantes de um texto e compô-los a fim de produzir os sumários correspondentes.

- Visualização: Consiste em descrever conjuntos complexos de dados em cenas visuais de fácil interpretação. As propriedades ou características de grandes itens textuais podem ser visualizadas através de gráficos de várias dimensões.
- Categorização de textos: Consiste em examinar os atributos de um determinado documento e, baseado nos valores destes atributos, associar seu conteúdo a uma determinada categoria.

Este artigo se foca na categorização de texto, em particular no processo de criar e combinar diferentes classificadores derivados a partir de diferentes conjuntos de documentos para um mesmo conjunto de categorias.

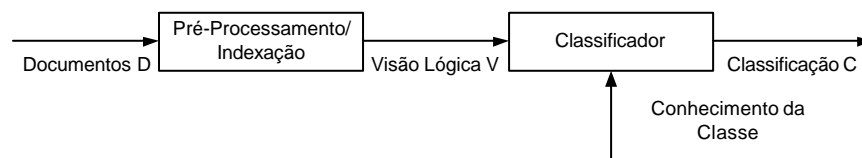
#### 4. O Processo de Categorização de Textos

A atividade de classificação de textos consiste no agrupamento de documentos elaborados em linguagem natural em diversas categorias ou classes [4][9][10]. Até o final da década de 1980, o processo de automação desta atividade consistia em manualmente definir um conjunto de regras, que representava o conhecimento de especialistas, para classificar documentos em uma categoria específica.

Esta abordagem mudou a partir da década de 1990, com a introdução de algoritmos de aprendizado de máquina para classificação de textos [4]. O objetivo destas técnicas é ensinar classificadores, a partir de exemplos que assimilem automaticamente características intrínsecas de cada categoria.

Matematicamente, a categorização de textos pode ser vista como a busca pela função  $f: D \times C = \{0,1\}$  que seja próxima à função ótima de classificação  $g: D \times C = \{0,1\}$ , onde  $D = \{d1, d2, d3, \dots, dj\}$  é um conjunto de documentos e  $C = \{c1, c2, c3, \dots, ci\}$  é um conjunto de classes pré-definidas.

A Figura 1 ilustra o processo de classificação de um texto. Inicialmente existe um processo de conversão do documento para uma visão lógica que possa ser compreensível para o algoritmo classificador. Em seguida, a partir do conhecimento da classe, obtido via aprendizado de máquina, a categorização do documento submetido é realizada.



**Fig 1.** Processo de classificação de um documento

#### 4.1 Indexação do Documento

Um documento, em seu formato original, não pode ser diretamente interpretado por um algoritmo classificador. Devido a este fato, um procedimento de indexação, usado mapear um documento para uma representação compacta de seu conteúdo, é necessária para permitir a uniformização do processo de classificação de documentos.

A idéia neste caso é coletar um conjunto de termos oriundos do documento, já formatados a uma forma padrão através de processos de análise léxica, conversão de caracteres, remoção de *stopwords*, normalizações morfológicas, reduções de dimensionalidade e outros métodos aplicáveis.

#### 4.2 Classificação do Documento

Os algoritmos classificadores de documentos utilizam processos indutivos. Nesta linha, um classificador para uma categoria  $c_i$  é construído observando as características de um conjunto de documentos, previamente rotulados sob  $c_i$  por um especialista no domínio. Esta é uma abordagem de aprendizado supervisionado, onde um novo documento é classificado de acordo com as características aprendidas por um classificador construído e treinado a partir de dados rotulados [11].

Para o problema de classificação de documentos apresentado neste trabalho, o classificador *Naive Bayes* é construído utilizando dados de treinamento para estimar a probabilidade de um documento pertencer a uma classe. O teorema de Bayes, mostrado no Quadro 1, é utilizado para estimar estas probabilidades [12].

$$P(C = c_i | \vec{x}) = \frac{P(\vec{x} | C = c_i) \times P(C = c_i)}{P(\vec{x})}, \text{ onde:}$$

$\vec{x}$  representa um vetor de termos e  $c_i$  representa uma classe.

**Quadro 1.** Fórmula de Bayes aplicada à classificação de documentos

Para fins de simplificação, uma vez que o cálculo de  $P(\vec{x} | C = c_i)$  é freqüentemente impraticável computacionalmente, o classificador *Naive Bayes* assume que as características dos termos  $\{x_1, \dots, x_n\}$  são independentes, dado a categoria variável  $C$ . Isto simplifica em muito a equação. Como  $P(\vec{x})$  é um denominador comum, pode-se ignorá-lo da equação. Assim, a fórmula fica simplificada conforme ilustrado no Quadro 2.

$$P(C = c_i | \vec{x}) = \prod_i P(x_i | C = c_i) \times P(C = c_i)$$

**Quadro 2.** Fórmula *Naive Bayes*

Apesar da suposição de independência condicional não ser inteiramente verdadeiro, o algoritmo de *Naive Bayes* é bastante efetivo [13].

A probabilidade de cada classe pode ser facilmente encontrada levando em consideração a quantidade de documentos assimilados à classe, dividido pelo conjunto total de documentos utilizados no treinamento do classificador. No caso específico da estimação de probabilidades de cada termo, foi utilizada na ferramenta, a fórmula expressa no Quadro 3 [14].

$$P(w_k | v_j) = \left( \frac{n_k + 1}{n + |\text{Vocabulário}|} \right), \text{ onde:}$$

$P(w_k | v_j)$  representa a probabilidade da verossimilhança da evidência de termo  $w_k$  dado a hipótese da classe  $v_j$ .

$n_k$  representa a quantidade de vezes que o termo  $w_k$  aparece no conjunto de treinamento designado para a classe  $v_j$ .

$n$  representa o total de termos coletados no conjunto de treinamento para a classe  $v_j$ .

**Vocabulário** representa o total de termos encontrados nos dados de treinamento de todas as classes.

**Quadro 3.** Fórmula de estimação de probabilidade para um termo de documento

## 5. Combinação de Classificadores

Um conjunto  $h^*$  de  $L$  classificadores individuais  $\{h_1, h_2, \dots, h_L\}$  podem ter suas predições combinadas a fim de determinar melhor o rótulo de um determinado documento [16]. Na sua forma mais simples, para um problema de  $k$  classes  $\{C_1, C_2, \dots, C_K\}$ , a combinação é efetuada utilizando o voto majoritário. Neste tipo de combinação, a classe prevista com maior frequência por todos os classificadores é a classe prevista pelo classificador final  $h^*$ .

É interessante perceber que, quando uma estratégia deste tipo é definida, o resultado final do conjunto tende a ser muito melhor do que o de um classificador base isolado. No caso de um grupo formado por  $n$  classificadores independentes, e com uma taxa de erro individual  $p$ , a probabilidade de  $x$  classificadores errarem em conjunto, na afirmação de uma hipótese final, obedece a uma distribuição binomial.

Por exemplo, se for admitido um conjunto de 21 classificadores bases, com uma probabilidade de erro individual de 0.3, uma hipótese falsa seria gerada se sua maioria errasse (mais de 11 classificadores). Neste caso, o erro calculado seria de aproximadamente 0.026, um valor muito menor em comparação ao erro individual de um único classificador.

Há, entretanto, alguns efeitos colaterais indesejados nesta abordagem. O processo de aprendizado individual dos classificadores do conjunto, muitas vezes, demanda tempo e memória [17][18].

Além disso, a operação do conjunto de classificadores exige uma avaliação individual de hipóteses de cada classificador base, que pode exigir um esforço computacional pequeno, quando avaliado isoladamente, mas substancial quando visto como empenho conjugado.

Desta forma, ao invés de combinar as hipóteses de um conjunto de especialistas sobre um determinado assunto, por que não somar a *expertise* individual de cada um para compor um único especialista?

Esta estratégia possui algumas peculiaridades interessantes a observar:

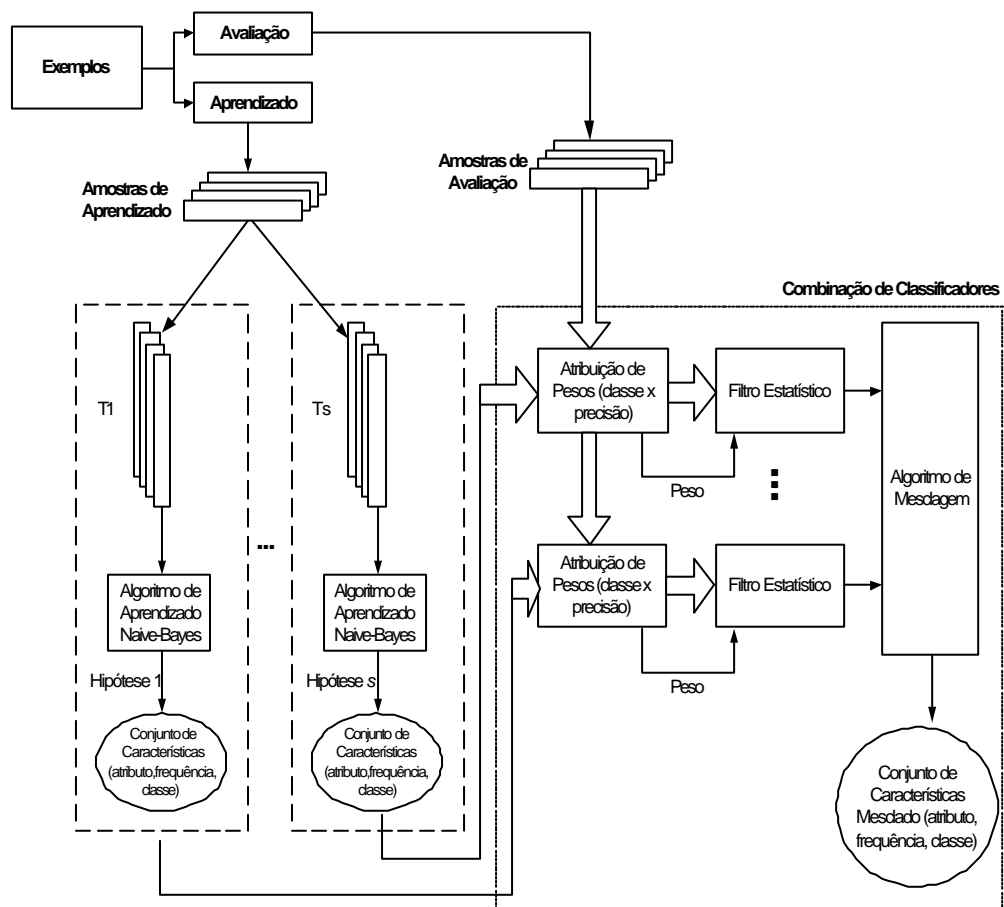
1. Os especialistas podem ser treinados isoladamente, em momentos diferentes, sem necessidade de consumo de recursos de um único ambiente. Além disso, o histórico de aprendizado de cada especialista contribui para uma diversidade de suposições interessantes no processo de combinação.
2. Os especialistas contribuem enviando sua *expertise* a fim de serem combinadas em um único especialista. Assim, o que irá ser demandado em um processo de avaliação de hipóteses final será apenas o esforço do super especialista, e não o esforço individual de cada especialista.
3. No processo de combinação é necessária uma avaliação de cada *expertise* para evitar a presença de maus especialistas que possam denegrir ou diminuir a *expertise* final. Além disso, os conhecimentos que possam ser supérfluos ao alvo de estudo das hipóteses devem ser descartados para não gerar perturbações na *expertise* final.
4. A *expertise* final pode ser devolvida a cada especialista para uma melhor avaliação de hipóteses no ambiente ao qual está inserido.

Esta estratégia foi empregada em um sistema, denominado XRULER, que induzia um agrupamento de regras lógicas em cada algoritmo de aprendizado de um conjunto e compunha um classificador final com base em regras pré-selecionadas de cada classificador [16].

A ferramenta ExperText, proposto neste trabalho, realiza um procedimento semelhante, a partir do aprendizado de máquina baseado no algoritmo de classificação *Naive Bayes*.

## 5.1 Método de Combinação da Ferramenta ExpertText

A Figura 2 ilustra a arquitetura utilizada pela ferramenta ExpertText na combinação de *expertise* em classificadores *Naive Bayes* distintos. A idéia é coletar, a partir de um conjunto de documentos denominados exemplos, dois conjuntos distintos. Um deles é destinado ao treinamento dos classificadores (*aprendizado*) e o outro é destinado à avaliação da *expertise* de cada classificador base (*avaliação*).



**Fig 2.** Arquitetura da ferramenta ExpertText para a combinação de classificadores

O conjunto de treinamento é dividido em  $S$  conjuntos de documentos seleccionados aleatoriamente  $\{T1, T2, \dots, TS\}$ , que, uma vez inseridos no treinamento *Naive Bayes*, geram as hipóteses  $\{H1, H2, \dots, HS\}$  em cada classificador respectivo. Estas hipóteses



são obtidas a partir de uma função alvo  $y=f(x)$  que, no caso do aprendizado baseado no teorema de Bayes, se resume à fórmula ilustrada no Quadro 4 [4][15].

$$\arg \max P(\text{classe} | a_1..a_n) = \arg \max \prod_i P(a_i | \text{classe}) \times P(\text{classe}), \text{ onde:}$$

$a_1..a_n$  representa um conjunto de características obtidos do conjunto de treinamento e cada classe possui um conjunto de características específico ao classificado.

#### Quadro 4. Fórmula de cálculo da hipótese máxima *a posteriori*.

O passo inicial no procedimento de combinação é a junção da *expertise* de cada classificador, cuja habilidade individual está diretamente relacionado ao seu conjunto de características, e que foi fruto de seu processo de aprendizado. É recolhido paulatinamente cada conjunto de características e submetido a um teste de categorização. Isto é necessário para possibilitar a verificação de sua precisão. Neste caso, é empregada, para esta verificação, um conjunto de documentos pré-rotulados, extraído de exemplos e denominado conjunto de avaliação.

É gerada, para cada documento contido em avaliação, uma hipótese de classe a partir do conjunto de características do classificador base avaliado. Ao final do processo de classificação, a eficiência do conjunto de características é aferida, via fórmula de cálculo de precisão, ilustrado no Quadro 5.

$$\text{precisão} = \frac{a}{(a+b)}, \text{ onde:}$$

$a$  é número de documentos rotulados a uma classe  $c_i$  e corretamente classificados como pertencentes a esta classe.

$b$  é o número de documentos rotulados como não pertencentes a uma classe  $c_i$ , porém incorretamente classificados como pertencentes a esta classe. São também chamados de falso-positivos.

#### Quadro 5. Fórmula de cálculo de precisão.

A precisão calculada irá influenciar no peso associado aos termos, contidos no conjunto de características do classificador avaliado. A idéia é associar um peso por classe analisada, conforme o nível de precisão obtido na classificação dos documentos em avaliação, e atribuir seu valor aos termos selecionados durante o processo de filtro estatístico.

Esta estratégia em atribuir pesos, conforme a precisão calculada, atua como um mecanismo de compensação de crenças, que visa valorizar os classificadores que possuam características que melhor avaliem determinadas classes, em detrimento a

outros que possuam termos, em seu conjunto de características, que venham contribuir com julgamentos equivocados.

No caso da ferramenta ExperText, pode-se definir os pesos de duas maneiras distintas: a primeira, atribuindo pesos a três faixas distintas de precisão, e a outra, definindo um peso proporcional à frequência aferida. Neste último caso, é definida uma faixa de pesos, cujo valor inicial e final são utilizados dentro de um cálculo proporcional.

O filtro estatístico se baseia em critérios de redução de dimensionalidade fundamentados em critérios de frequência de termos, frequência inversa de documento, *TFIDF*, probabilidade, informação mútua e ganho de informação.

Ao final da avaliação de um classificador base, os termos que sobreviverem ao filtro estatístico terão um determinado peso aplicado e serão mesclados, por classe, aos demais termos avaliados em cada classificador base.

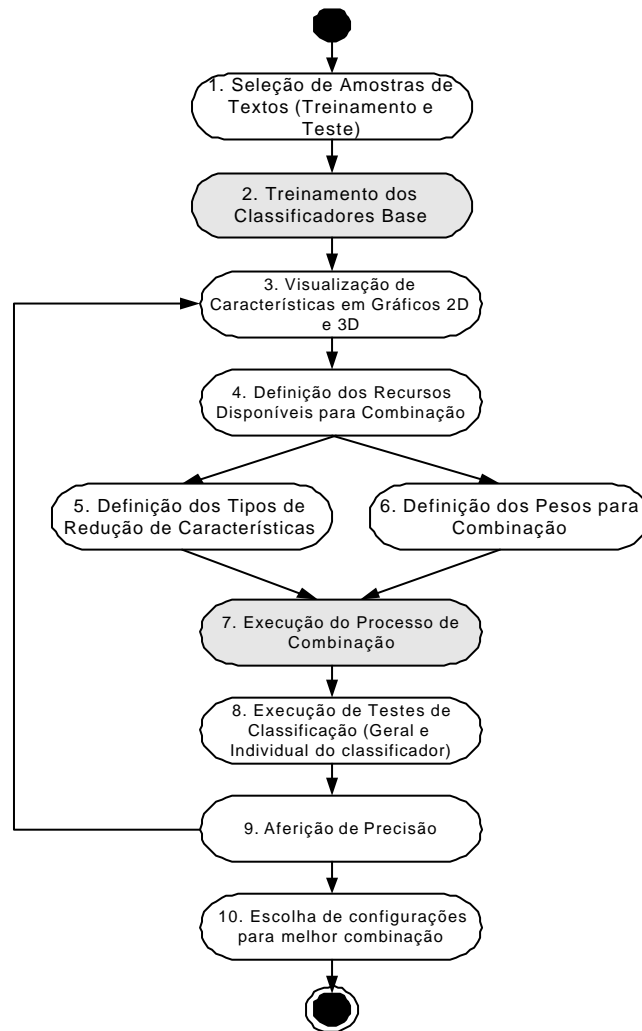
## 6. Ferramenta ExperText

A ferramenta ExperText foi construída com o objetivo de fornecer uma forma de acompanhar todo o processo de combinação de *expertise* de classificadores de textos baseados no teorema de *Bayes*, desde a seleção de conjuntos de documentos para treino e testes, passando pelo treinamento específico de cada algoritmo e sua posterior combinação. Nesta oportunidade, a observação das características de cada classificador pode ser realizada em gráficos cruzados, com duas ou três dimensões, para uma melhor definição da estratégia de redução de atributos e configuração de pesos em cada *expertise* armazenada.

Desta forma, o processo de combinação assemelha-se um pouco a um procedimento de mineração visual de dados, devido ao seu caráter exploratório. No caso da ferramenta ExperText, os passos para se descobrir a melhor forma de mesclar as características enviadas por cada classificador base e que compõem sua *expertise*, está ilustrado na Figura 3.

Existe um conjunto de passos a serem seguidos, a fim de que todo o processo seja realizado da melhor forma possível. É interessante ressaltar que os passos 2 e 7, treinamento dos classificadores base e execução do processo de combinação, respectivamente, são realizados pela própria ferramenta, enquanto que as demais atividades dependem exclusivamente do usuário e de seu conhecimento sobre o tema.

A fim de realizar uma avaliação de sua aplicabilidade, foram realizados dois estudos de casos com a ferramenta ExperText em abordagens distintas e sua eficácia foi comprovada observando-se um aumento na abrangência média e acurácia no processo de classificação quando utilizada a *expertise* combinada. O estudo de caso no contexto acadêmico envolveu a utilização de uma base de artigos amplamente utilizada por pesquisadores, nas áreas de recuperação de informação, classificação de textos e processamento em linguagem natural, denominada Reuters-21578. No contexto empresarial, a ferramenta foi aplicada na classificação de correspondência eletrônica mantida pela área de tecnologia da Secretaria da Fazenda do Estado da Bahia com outras áreas técnicas da instituição.



**Fig 3.** Diagramas de atividades válidas para o processo de combinação

### 6.1 Arquitetura da Ferramenta

Tendo em vista, a simplificação de sua construção, a arquitetura da ferramenta ExperText foi definida levando em consideração uma modularização do aplicativo em quatro partes inter-relacionadas: o módulo de entrada de documentos, o módulo de configuração de pesos e filtros, o módulo de treinamento, classificação e combinação de classificadores e o módulo de visualização de dados. Estes módulos estão associados às seguintes funções básicas da ferramenta.

- Cadastro das referências de documentos com fins de treinamento e teste;
- Criação de uma interface para entrada das opções de filtro e atribuição de pesos;
- Treinamento e classificação de textos juntamente com a combinação de classificadores;
- Representação e visualização de características e resultados do processo de classificação de textos.

## 6.2 Avaliação de Resultados

A eficácia do processo de combinação sugerida neste trabalho foi comprovada avaliando as métricas de abrangência média e acurácia de cada classificador base. Neste caso, um conjunto de documentos, previamente rotulados e não utilizados nas amostras de treinamento e avaliação sugeridos na Figura 2, foi submetido a cada classificador base, empregando sua própria *expertise* adquirida e uma outra obtida a partir do processo de combinação.

Compreendendo os conceitos, a abrangência é a métrica que define o percentual verificado entre o número de textos corretamente associados a uma classe e o número de textos que realmente pertencem à classe. O valor médio é calculado dividindo-se todos os valores de abrangência avaliados por classe, pelo número de classes utilizadas pelo classificador.

A acurácia define o percentual relativo à quantidade total de documentos corretamente rotulados pelo classificador sobre o total de documentos utilizados.

Estas duas métricas são calculadas conforme as fórmulas ilustradas no Quadro 6.

$$\text{abrangência} = \frac{a}{(a + c)}, \text{acurácia} = \frac{(a + d)}{(a + b + c + d)} \text{ onde:}$$

**a** é número de documentos rotulados a uma classe  $c_i$  e corretamente classificados como pertencentes a esta classe.

**b** é o número de documentos rotulados como não pertencentes a uma classe  $c_i$ , porém incorretamente classificados como pertencentes a esta classe. São também chamados de falso-positivos.

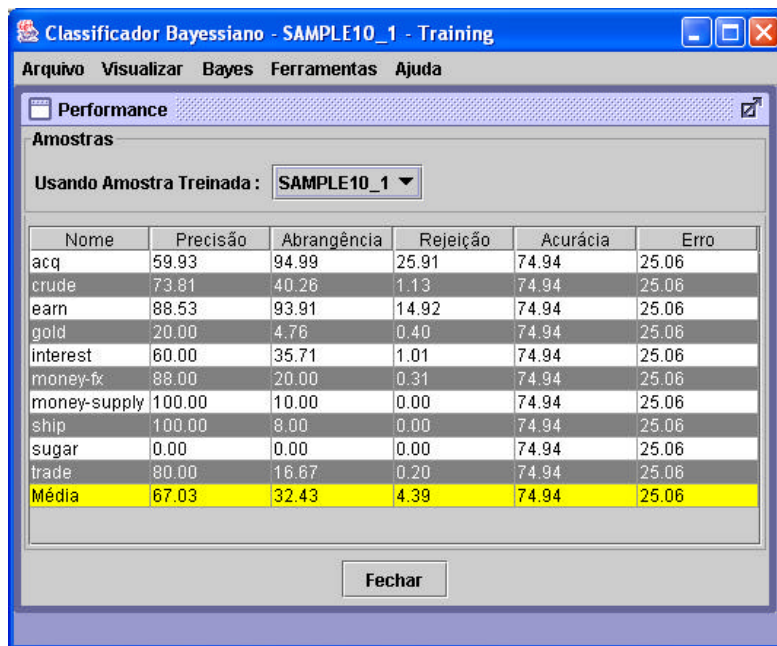
**c** é o número de documentos rotulados como pertencente a uma classe  $c_i$ , porém incorretamente classificados como não pertencentes a esta classe. Também chamados de falso-negativos.

**d** é o número de documentos rotulados como não pertencentes a uma classe  $c_i$  e corretamente classificados como não pertencentes a esta classe.

**Quadro 6.** Fórmula de cálculo da abrangência e acurácia de um classificador.

### 6.3 Visualização de Performance

O procedimento de classificação tem sua performance calculada através de um processo de batimento de dados entre uma tabela de dados que armazena o resultado de uma classificação e outra que guarda os dados originais de classe de cada documento. Desta forma, os parâmetros de desempenho, tais como precisão, abrangência, rejeição, acurácia e erro, são visualizados em formato tabular, em uma interface visual. A Figura 4 ilustra uma tela disponível na ferramenta para este propósito.



The screenshot shows a window titled 'Classificador Bayesiano - SAMPLE10\_1 - Training'. The menu bar includes 'Arquivo', 'Visualizar', 'Bayes', 'Ferramentas', and 'Ajuda'. The 'Performance' tab is active, showing a table of performance metrics for various samples. The table has columns for 'Nome', 'Precisão', 'Abrangência', 'Rejeição', 'Acurácia', and 'Erro'. The 'Média' row is highlighted in yellow.

Nome	Precisão	Abrangência	Rejeição	Acurácia	Erro
acq	59.93	94.99	25.91	74.94	25.06
crude	73.81	40.26	1.13	74.94	25.06
earn	88.53	93.91	14.92	74.94	25.06
gold	20.00	4.76	0.40	74.94	25.06
interest	60.00	35.71	1.01	74.94	25.06
money-fx	88.00	20.00	0.31	74.94	25.06
money-supply	100.00	10.00	0.00	74.94	25.06
ship	100.00	8.00	0.00	74.94	25.06
sugar	0.00	0.00	0.00	74.94	25.06
trade	80.00	16.67	0.20	74.94	25.06
Média	67.03	32.43	4.39	74.94	25.06

Fig 4. Tela de visualização de performance da ferramenta ExperText

### 6.4 Visualização de Gráficos

O algoritmo de visualização de dados utilizado pela ferramenta ExperText é baseado no pacote *jmathplot* (código fonte aberto e livre) encontrado no site SOURCEFORGE.NET. O pacote *jmathplot* é composto por um conjunto de classes e interfaces que fazem a representação de uma coleção de dados em uma estrutura visual de duas ou três dimensões.

A Figura 5 exibe um exemplo de gráfico em três dimensões possível de ser obtida através do módulo da ferramenta para visualização das características de um documento. Através deste módulo é possível cruzar todos os valores associados ao conjunto de termos que define a *expertise* de um classificador, e cuja semântica são

justamente as opções de filtragem oferecidas pelo módulo de configuração de filtros e pesos, antes de serem aplicadas aos classificadores em cada amostra de documentos.

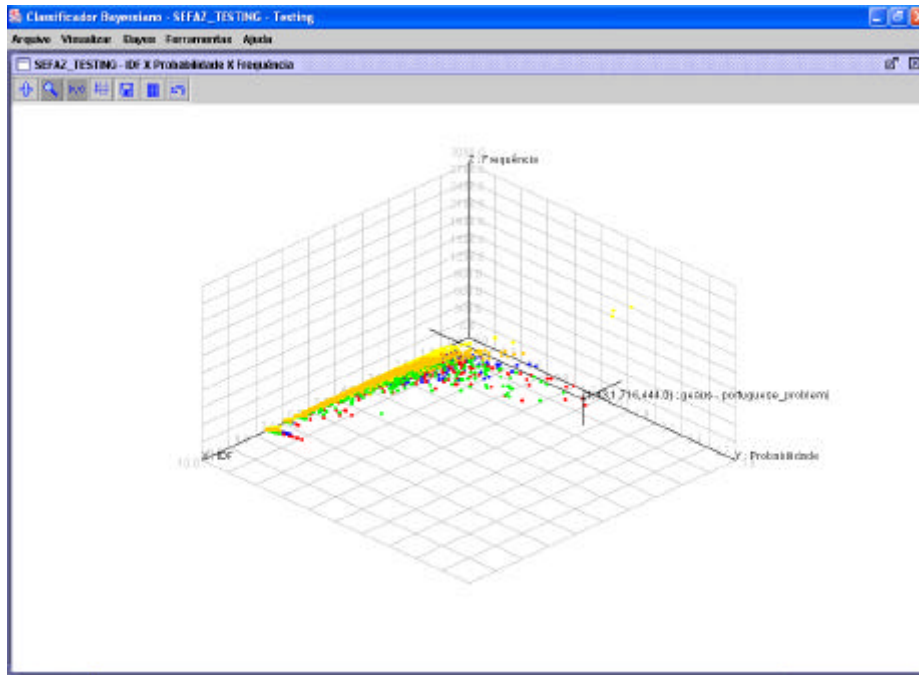


Fig 5. Exemplo de gráfico em três dimensões exibido pela ferramenta

## 7. Conclusões

Este artigo apresentou uma proposta de combinação de *expertise* de classificadores utilizando algoritmos estatísticos e fórmula *Naive Bayes*. Esta abordagem englobou a definição e implementação de uma ferramenta que permitiu materializar a metodologia sugerida para o processo de combinação. A ferramenta realiza análises visuais no espaço de características associadas aos documentos de uma determinada classe e avaliações de métricas que podem justificar a eficácia de todo o processo.

Os trabalhos futuros relacionados à evolução da ferramenta ExpertText são:

- Inclusão de uma funcionalidade de simulação de combinação mais abrangente que possa diminuir o tempo de análise dos critérios de atribuição de pesos e filtros.
- Integração da ferramenta ExpertText com uma ferramenta de mineração visual, a fim de permitir a exploração dinâmica do espaço de características de um classificador, facilitando o julgamento dos critérios de combinação de *expertise*.

- Integração da ferramenta ExperText a servidores de arquivos a fim de permitir uma automação no processo de categorização de documentos em um ambiente empresarial.
- Utilização da metodologia de combinação e demais objetos que compõem a ferramenta ExperText em uma abordagem orientada a agentes com o objetivo de coletar *expertise* de maneira remota e descentralizada sobre a categorização de textos em tópicos específicos, como lixo eletrônico. A idéia neste caso é criar uma ferramenta poderosa na detecção de SPAM.

## Referências

- [1] STEWART, T. Capital Intelectual. 2 ed, RJ: Campus, 1998.
- [2] DAVENPORT, T., PRUSAK, L. Conhecimento Empresarial: Como as Empresas Gerenciam o seu Capital Intelectual. 7 ed, RJ: Campus, 1998.
- [3] GATES, B. A Empresa na Velocidade do Pensamento. São Paulo, Companhia das Letras, 1998.
- [4] SEBASTIANI, F.; DEBOLE. F. An Analysis of the Relative Hardness of Reuters-21578 Disponível em: <http://Subsets.faure.isti.cnr.it/~fabrizio/Publications/JASIST04.pdf>. Acesso em: julho de 2004.
- [5] DAVIDSON. K; FRAPPAOLO, C. Document Power: The New Management Paradigm. 1999. Disponível em: <http://www.delphigroup.com>. Acesso em: Maio de 2004.
- [6] CARVALHO, R. Aplicações de softwares de gestão do conhecimento: tipologia e usos. Belo Horizonte: Escola de Ciência da Informação da UFMG, 2000. 144p.(Dissertação, Mestrado em Informação Gerencial e Tecnológica).
- [7] ARIMURA, H; ABE, J.; FUJINO, R.; SAKAMOTO, H.; SHIMOZONO, S.; ARIKAWA, S. Text Data Mininig: Discovery of Important Keywords in the Cyberspace. In Proc. IEEE. Kyoto ICDL, 2001.
- [8] VISA. A. Technology of Text Mining. Proceedings of Machine Learning and Data Mining in Pattern Recognition, Second International Workshop, MLDM 2001, Leipzig, Alemanha, 2001
- [9] YANG, Y.; PEDERSEN, J. O. A Comparative Study on Feature Selection in Text Categorization. In: International Conference On Machine Learning (ICML), 1997.
- [10] JOACHIMS, T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. Proceedings of International Conference on Machine Learning (ICML), 1997.
- [11] MARTINS, J. Classificação de páginas na internet. Trabalho de Conclusão (Mestrado). Instituto de Ciências Matemáticas e de Computação. USP. São Carlos, 2003.
- [12] DUMAIS, S. T., PLATT, J., HECKERMAN, D., AND SAHAMI,M. 1998. Inductive learning algorithms and representations for text categorization. In Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management (Bethesda, MD, 1998).

- [13] DOMINGOS, P. e PAZZANI, M. J. On the optimality of the simple Bayesian classifier under zero-one loss. *Mach. Learn.* 1997.
- [14] MITCHEL, T. *Machine Learning*. McGraw-Hill, 1997.
- [15] PARDO, T.A.S. e NUNES, M.G.V. *Aprendizado Bayesiano Aplicado ao Processamento de Línguas Naturais*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação - ICMC, Universidade de São Paulo, 2002.
- [16] BARANAUSKAS, J. A. *Extração automática de conhecimento por múltiplos indutores*. Tese de Doutorado, ICMC-USP. 2001.
- [17] DIETTERICH, T. *Machine Learning Research: Four Current Directions*. *AI Magazine*, 18(4), 1997.
- [18] KOTSIANTIS, S. B.; PINTELAS, P. E. *An Online Ensemble Of Classifiers*, The Fourth International Workshop on Pattern Recognition in Information Systems – PRIS-2004, In conjunction with 6th International Conference on Enterprise Information Systems, Porto - Portugal 14-17, Abril de 2004.