Pós-Graduação em Ciência da Computação

"SISTEMAS BASEADOS EM MAPAS AUTO-ORGANIZÁVEIS PARA ORGANIZAÇÃO AUTOMÁTICA DE DOCUMENTOS TEXTO"

Por

RENATO FERNANDES CORRÊA

TESE DE DOUTORADO

Universidade Federal
de Pernambuco
http://www.cin.ufpe.br

RECIFE, JULHO/2008

UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

RENATO FERNANDES CORRÊA

"SISTEMAS BASEADOS EM MAPAS AUTO-ORGANIZÁVEIS PARA ORGANIZAÇÃO AUTOMÁTICA DE DOCUMENTOS TEXTO"

ESTE TRABALHO FOI APRESENTADO À PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO DO CENTRO DE INFORMÁTICA DA UNIVERSIDADE FEDERAL DE PERNAMBUCO COMO REQUISITO PARCIAL PARA OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIA DA COMPUTAÇÃO.

ORIENTADORA: PROFª DRª TERESA BERNARDA LUDERMIR

RECIFE, JULHO/2008

Corrêa, Renato Fernandes

Sistemas baseados em mapas auto-organizáveis para organização automática de documentos texto / Renato Fernandes Corrêa. - Recife : O Autor, 2008.

vi, 103 p. : il., fig., tab.

Tese (doutorado) – Universidade Federal de Pernambuco. CIn. Ciência da Computação, 2008.

Inclui bibliografia e apêndice.

1. Redes neurais artificiais. I. Título.

006.3 CDD (22.ed.) MEI2008-067

Resumo

Este trabalho propõe e avalia sistemas híbridos para organização automática de documentos texto baseado em Mapas Auto-Organizáveis (do inglês *Self-Organizing Maps* - SOM). O objetivo é obter um sistema que ao combinar SOM com outros algoritmos de agrupamento seja capaz de gerar mapas de documentos de qualidade a um custo computacional baixo para grandes coleções de documentos texto.

Um mapa de documentos é resultado de pós-processamento de uma rede neural SOM treinada com os vetores representativos dos documentos de uma coleção. Um mapa de documentos é considerado de boa qualidade quando este representa bem as relações de similaridade de conteúdo entre documentos de uma coleção.

Um mapa de documentos possibilita a organização de uma coleção de documentos texto de acordo com a similaridade de conteúdo e tem aplicação na melhoria dos processos de recuperação de informação, exploração, navegação e descoberta de conhecimento sobre uma coleção.

Vários trabalhos na literatura de redes neurais têm utilizado SOM para criar mapas de documentos. Entretanto, o treinamento de redes SOM ainda é uma tarefa cara computacionalmente para grandes coleções de documentos texto. Alguns métodos propostos na literatura para construir mapas de documentos mais rapidamente reduzem drasticamente a qualidade do mapa gerado; além disso, sistemas híbridos envolvendo SOM com outros algoritmos de agrupamento têm sido pouco investigados na literatura. Estes fatos motivaram o presente trabalho.

Os resultados mostram que a combinação cuidadosa de algoritmos tradicionais de agrupamento como K-means e Leader com redes SOM é capaz de produzir sistemas híbridos bastante eficientes. Por este motivo, sistemas híbridos foram propostos, visando à construção automática de mapas de documentos com qualidade e a um custo computacional mais baixo.

Estes sistemas híbridos representam um avanço na área de sistemas de organização automática de documentos texto, bem como sistemas neurais híbridos baseados em SOM, fornecendo resultados importantes para diversas aplicações práticas no projeto de sistemas, tais como engenhos de busca, sistemas para bibliotecas digitais e sistemas para descoberta de conhecimento em texto.

Palavras-chave: Sistemas Híbridos Inteligentes, Recuperação de Informação, Redes Neurais Artificiais, Algoritmos de Agrupamento.

Abstract

This work proposes and evaluates hybrid systems for automatic text document organization based on Self-Organizing Maps (SOM). The aim is to design a system that combines SOM with other clustering algorithms, in order to generate document maps for large text document collections of good quality at a low computational cost.

The posprocessing of a neural network SOM trained with the vectors that represent documents of a collection generates a document map. Document maps of good quality are those that represent well the relations of content-based similarity between documents

A document map organizes a text document collection in accordance with the content-based similarity, and it has application in improving of the processes of information retrieval, exploration, browsing and text mining on a collection.

Several works in the literature of neural networks have used SOM to create document maps. However, the training of SOM networks is still an expensive computational task for large text document collections. Some methods considered in literature to construct document maps more quickly reduce drastically the quality of the generated map. Moreover, hybrid systems combining SOM with other clustering algorithms are not investigated enough in literature. These facts had motivated the present work.

The results show that the careful combination of traditional clustering algorithms like K-means and Leader with SOM networks is able to produce very efficient hybrid systems. For this reason, a hybrid system was proposed, in order to implement an automatic process to generate document maps of good quality at a low computational cost.

These hybrid systems represent a advance in the field of document organization systems, as well as SOM-based neural hybrid systems, by providing important results for several practical applications in design of systems as: search engines, systems for digital libraries and systems for text mining.

Keywords: Hybrid Intelligent Systems, Information Retrieval, Artificial Neural Networks, Clustering Algorithms.

Índice

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1. Motivação

Sistemas de Recuperação de Informação (SRIs) [Baeza-Yates & Ribeiro-Neto 1999] tratam essencialmente de indexação, busca e ordenação de documentos, com o objetivo de satisfazer necessidades de informação dos usuários, geralmente expressa através de consultas.

Hoje em dia, os SRIs disponibilizam a busca por meio de consultas como a principal forma de explorar uma coleção de documentos à procura de informações relevantes. As consultas são construídas pelo usuário, sendo compostas por palavras-chave e/ou expressões simples envolvendo as mesmas. O sucesso em encontrar documentos relevantes depende do casamento dos termos fornecidos pelo usuário em uma consulta, com os utilizados como índices na indexação da base de dados de documentos.

Os SRIs geralmente expressam o resultado de uma busca através de uma lista linear de links para os documentos que satisfazem a consulta, ordenados de acordo com o valor de uma métrica de relevância de cada documento em relação à consulta utilizada. Este tipo de representação de resultados pode retornar documentos ordenados de uma forma não correlata à relevância atribuída pelo usuário, bem como omite a relação de similaridade de conteúdo entre documentos, dificultando a identificação de grupo de documentos relevantes e irrelevantes [Kohonen et al. 2000].

Quando o tamanho da coleção é da ordem de milhares de documentos, formular uma consulta efetiva para uma busca é uma tarefa difícil e examinar uma lista resultante

de uma busca na qual os itens retornados são muitos e estão ordenados de forma aparentemente não significativa pode ser enfadonho para o usuário.

Assim, com o crescimento das coleções de documentos digitais, os SRIs atuais têm buscado aprimoramento diante das limitações presentes no processo de recuperação de informação, tais como a dificuldade do usuário em expressar o que ele realmente procura através de uma consulta; a forma seqüencial em que os documentos que satisfazem as buscas são apresentados; e o número excessivo de documentos retornados.

Intuitivamente, a navegação e pesquisa sobre uma coleção de documentos seriam facilitadas se os documentos fossem organizados automaticamente de acordo com a similaridade de conteúdo da seguinte maneira:

- Grupos de documentos conteriam documentos similares em conteúdo. Assim, os conteúdos dos documentos mais próximos (similares) ajudariam a entender o verdadeiro significado de cada documento individualmente e a encontrar informações relevantes ou similares, mesmo que o usuário não estivesse explicitamente procurando por estas;
- A relação de similaridade entre grupos de documentos seria expressa explicitamente, auxiliando na identificação de grupos de documentos correlacionados e permitindo ao usuário a seleção de grupos relevantes;
- Cada grupo de documentos seria descrito por palavras-chave, auxiliando na compreensão dos tópicos tratados pelo grupo de documentos sem que necessariamente o usuário tenha que ler algum documento do grupo.

Denominam-se sistemas de organização automática de documentos [Kohonen 2001] os sistemas capazes de organizar automaticamente documentos de uma coleção em grupos de acordo com a similaridade de conteúdo e que tornem explícitas as relações de proximidade entre grupos e tópicos presentes nos documentos de cada grupo. Tais sistemas são úteis na construção de SRIs mais robustos e eficientes.

A organização automática de documentos por conteúdo é uma necessidade atual que se acentuará cada vez mais com o crescimento das coleções de documentos digitais, já que o aumento do número de documentos torna cada vez mais grave o problema conhecido como sobrecarga de informação (do inglês *information overloading*) [Baeza-Yates & Ribeiro-Neto 1999]. Em outras palavras, a quantidade crescente de informação

disponível torna a tarefa de encontrar informações relevantes cada vez mais difícil e consumidora de tempo para o usuário.

A aplicação de Mapas Auto-Organizáveis (do inglês *Self-Organizing Maps* - SOM) [Kohonen 2001] tem sido bastante explorada na literatura de agrupamento de documentos (do inglês *text clustering* ou *document clustering*) visando a organização automática de documentos texto por conteúdo. A tarefa de organização automática de documentos é mais específica que a tarefa de agrupamento de documentos, pois além de gerar grupos de documentos também busca representar de forma explícita a relação de proximidade entre os grupos.

Além da capacidade de organizar vetores que representam documentos em grupos de acordo com a similaridade de conteúdo dos documentos e por permitir a obtenção de palavras-chave descritoras do conteúdo de cada grupo (no caso das redes SOM através da avaliação pós-treinamento dos pesos presentes nos vetores modelo de cada neurônio), a rede SOM se destaca entre os algoritmos de agrupamento que podem ser utilizados na tarefa de organização automática de documentos por deixar explícitas as relações entre os grupos por meio de uma projeção não-linear em um arranjo unidimensional, bidimensional ou tridimensional de neurônios.

Em experimentos realizados em [Corrêa 2002], as redes SOM mostraram-se capazes de organizar os documentos de maneira intuitiva, preservando a relação de similaridade não só entre documentos, mas também entre categorias de documentos atribuídas por especialistas. O desempenho da rede SOM na categorização de documentos é em alguns casos foi superior ao obtido por alguns algoritmos supervisionados como C4.5 e PART [Corrêa & Ludermir 2004c]. Estes resultados confirmam que redes SOM organizam intuitivamente os documentos de forma que a organização gerada é adequada para análise humana por se assemelhar à organização feita por humanos, motivando a utilização destes na construção de mapas de documentos.

Entretanto, a rede SOM apresenta uma desvantagem quando aplicada na organização automática de documentos texto de grandes coleções: o treinamento da rede é um processo computacionalmente caro quando a dimensionalidade é alta e o volume de dados é grande [Azcarraga & Yap 2001].

As alternativas relatadas na literatura para resolver este problema envolvem uso de métodos para redução da dimensionalidade, redução de volume, e aceleração do processo de treinamento das redes SOM. Entretanto, a avaliação do impacto do uso destas alternativas no desempenho do sistema tem sido pouco investigada. Alguns métodos propostos na literatura para acelerar a construção de mapas de documentos chegam a reduzir drasticamente a qualidade do mapa gerado e a aplicação destes deve ser evitada ou realizada com grande cautela. Estas investigações constituem contribuições deste trabalho.

A combinação de outros algoritmos de agrupamento com SOM formando um sistema híbrido tem sido pouco explorada. Algoritmos de agrupamento tradicionais, como Leader e K-means, são mais rápidos que a rede SOM, motivando a proposição e avaliação de sistemas híbridos compostos por tais algoritmos. Os sistemas propostos nesta tese são considerados híbridos, pois são fruto da combinação dos algoritmos K-means, Leader e SOM que são considerados algoritmos distintos por possuírem natureza e propósitos diferentes (os dois primeiros foram projetados para agrupar dados e a rede SOM para a visualização dos dados).Vários sistemas híbridos são propostos e avaliados, constituindo contribuições do presente trabalho. Não se tem até o momento conhecimento de sistemas híbridos de organização automática de documentos com as mesmas características dos sistemas propostos nesta tese.

1.2. Objetivos

O presente trabalho visa propor e avaliar sistemas híbridos baseados em SOM para construir mapas de documentos de grandes coleções de texto, buscando um compromisso entre rapidez de treinamento e qualidade do mapa gerado. A qualidade do mapa é mensurada através da capacidade de caracterizar e tornar explícitas as relações semânticas entre os documentos de uma coleção.

Além disso, este trabalho traz um resumo das abordagens de organização automática de documentos usando SOM e suas variantes. Tal estudo teve a finalidade de contribuir com um resumo sobre este tópico, pois não se tem conhecimento de trabalhos que relatem de forma abrangente a aplicação de SOM e suas variantes para organizar automaticamente documentos. Este estudo pode servir como material introdutório para

diversas linhas de pesquisa relacionadas ao uso de Mapas Auto-Organizáveis para organização automática de documentos.

Até o presente momento, não se encontra na literatura, uma combinação entre SOM e algoritmos de agrupamento como foi proposta neste trabalho, procurando explorar ao máximo as potencialidades de cada técnica, de modo que uma parte possa compensar as deficiências da outra. Dessa forma, o sistema híbrido resultante mantém um melhor compromisso entre eficácia e eficiência do que suas técnicas constituintes funcionando isoladamente.

Para desenvolver estes sistemas, é necessário definir alguns aspectos importantes na aplicação de SOM e técnicas de agrupamento na organização automática de documentos, como a forma de representar vetorialmente os documentos, a forma de reduzir a dimensionalidade dos vetores documentos, a medida de similaridade usada para medir a relação ou distância entre vetores documentos, parâmetros de treinamento, a medida de qualidade do mapa ou eficácia e a medida de eficiência na construção do mapa. Outro aspecto importante na construção dos sistemas híbridos é definir meios de combinar SOM e as técnicas de agrupamento e avaliar estas combinações.

Dessa forma, torna-se importante avaliar experimentalmente várias alternativas no treinamento de SOM e algoritmos de agrupamento, a fim de verificar quais escolhas realmente trazem vantagens na organização automática de documentos. A análise de resultados, mostrada no Capítulo 4, tem o objetivo de justificar a escolha de parâmetros e métodos, bem como avaliar os sistemas híbridos propostos.

O problema utilizado como estudo de caso nos resultados do Capítulo 4 foi a categorização de diferentes coleções de documentos: K1, Reuters-21578 e 20 Newsgroups. Estas coleções contêm documentos escritos em língua inglesa pertencentes a diferentes gêneros (páginas web, notícias e e-mails, respectivamente), sendo consideradas *benchmarks* nas áreas de categorização e agrupamento de documentos. Estas coleções são consideradas de grande porte e têm sido estudadas e utilizadas há vários anos por pesquisadores do mundo inteiro, sendo esta a principal motivação para a utilização destes conjuntos de dados a fim de testar os sistemas híbridos propostos.

Entretanto, é importante ressaltar que os sistemas híbridos propostos são customizáveis para organizar documentos escritos em outras línguas ou pertencentes a diferentes domínios.

1.3. Organização da Tese

Neste capítulo introdutório, a motivação e os objetivos deste trabalho foram apresentados.

O Capítulo 2 traz um resumo do estado da arte sobre sistemas de organização automática de documentos baseados em SOM. Inicialmente é apresentado o modelo e treinamento das redes SOM, sendo depois apresentado como esta rede neural é aplicada na organização automática de documentos. Em seguida, é apresentada a arquitetura do sistema de organização automática de documentos baseados em SOM bem como métodos de avaliação.

O Capítulo 3 trata da metodologia proposta, apresentando os detalhes do processo de construção dos sistemas híbridos. As técnicas de agrupamento K-means e Leader são apresentadas e, em seguida, a arquitetura dos sistemas híbridos é discutida. Os sistemas híbridos propostos são também especificados.

O Capítulo 4 apresenta os resultados obtidos pelos sistemas híbridos na categorização de documentos das coleções K1, Reuters-21578 e 20 Newsgroups. Inicialmente, apresentam-se explicações sobre as coleções utilizadas, tornando mais clara a composição e a escolha das mesmas. Em seguida, apresenta-se a metodologia usada nos experimentos. Por fim, são apresentados e discutidos os resultados para os sistemas definidos no Capítulo 3.

No Capítulo 5, apresentam-se as conclusões obtidas com o trabalho desenvolvido e as possibilidades de trabalhos futuros.

No Apêndice A, apresenta-se uma breve introdução sobre categorização automática de documentos.

Capítulo 2

Organização Automática de Documentos usando redes SOM

Este capítulo traz um resumo do que existe na literatura sobre sistemas de organização automática de documentos baseados em redes SOM. Inicialmente são apresentados a arquitetura e treinamento de tais redes. Em seguida, a arquitetura do sistema de organização automática de documentos baseados em SOM bem como métodos de avaliação são discutidos. Posteriormente, o estado da arte da aplicação de redes SOM na organização automática de documentos é apresentado.

2.1. Self-Organizing Maps

A rede SOM [Kohonen 2001] é um modelo de rede neural artificial que segue os paradigmas de aprendizado não supervisionado e competitivo, sendo capaz de extrair padrões de similaridade dos vetores de entrada de forma que as relações estatísticas não-lineares entre os padrões de entrada multidimensionais são convertidas em simples relações geométricas dos respectivos neurônios, que se encontram dispostos em um arranjo unidimensional, bidimensional ou tridimensional. Desta forma, a rede SOM compacta a informação preservando as mais importantes relações topológicas e/ou métricas, gerando um tipo de representação dos dados. Visualização e abstração constituem as duas principais aplicações das redes SOM [Kohonen 2001].

Durante o treinamento, um espaço de entrada de alta dimensionalidade representado por padrões de entrada é aproximado através de um conjunto finito de vetores protótipo de mesma dimensionalidade presentes nos neurônios da rede, sendo

estes neurônios organizados em arranjos geralmente unidimensionais, bidimensionais ou tridimensionais. Metaforicamente falando, a rede SOM forma uma rede elástica de neurônios que se molda na nuvem formada pelos dados de entrada. No final do treinamento, em geral, pontos do espaço de entrada próximos uns dos outros são mapeados em neurônios próximos no mapa. Assim, a rede SOM pode ser interpretada como um mapeamento que preserva a topologia do espaço de entrada em um arranjo 1-D, 2-D ou 3-D de neurônios.

O número de neurônios no mapa é que determina o nível de detalhamento na representação do espaço de entrada, a acurácia e a capacidade de generalização dos mapas auto-organizáveis. Durante o treinamento, a rede SOM se molda aos dados de entrada, e assim, o número de neurônios no mapa deve ser suficiente para representar cada uma das regiões do espaço de entrada.

Existem métodos clássicos em análise exploratória de dados e análise multivariada, que são capazes de formar projeções 2-D de uma distribuição de itens em um espaço de dimensionalidade alta, como os métodos de escalonamento multidimensional [Kohonen et al. 2000]. Segundo Kohonen [Kohonen et al. 2000], as vantagens em utilizar SOM em relação a estes métodos são as seguintes:

- Para espaços de entrada de alta densidade (isto é, contendo muitos padrões de entrada) é computacionalmente mais leve na geração da projeção;
- O mapeamento pode ser computado usando um subconjunto representativo do espaço de entrada;
- Novos itens podem ser mapeados sem recomputação de todo o mapeamento.

Do ponto de vista estatístico, a rede SOM treinada pode ser interpretada como uma projeção não-linear da função densidade de probabilidade dos dados de entrada de alta dimensionalidade no espaço unidimensional ou bidimensional.

2.1.1 Arquitetura

A estrutura básica de uma rede SOM é formada por uma camada de entrada e uma camada de saída. A camada de entrada recebe sinais de entrada e os transfere para a camada de saída. Sinais de entrada codificam um padrão de entrada e são apresentados à rede como um vetor. A camada de saída é responsável pela representação dos padrões

de entrada. Como pode ser observada na Figura 2.1, a rede SOM possui uma única camada de neurônios.

Cada neurônio na camada de saída recebe todos os sinais captados pela camada de entrada. Associado a cada neurônio há um vetor protótipo, também chamado vetor modelo (do inglês *model vector*), de mesma dimensionalidade dos padrões de entrada. O vetor protótipo contém os pesos da sinapse entre cada característica vinda da camada de entrada e o neurônio. O estado de ativação de um neurônio é o valor da distância entre o vetor protótipo e o padrão de entrada apresentado à rede.

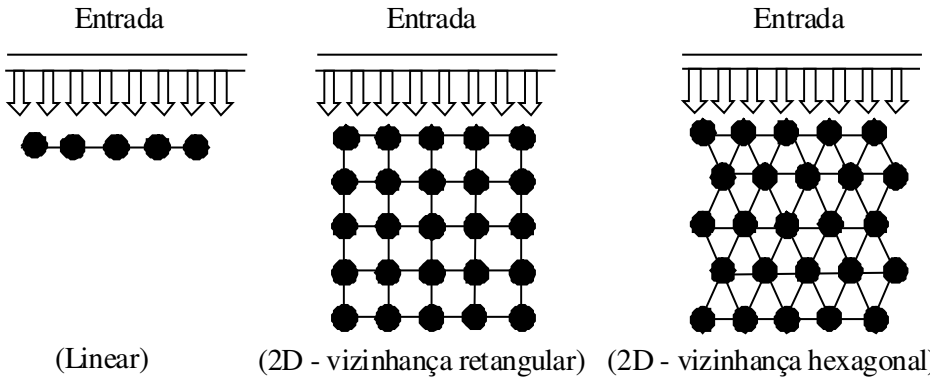

Figura 2.1 – Arquitetura da rede SOM.

A saída da rede SOM para um dado padrão de entrada é formada pelo conjunto de saídas apresentadas por todos os neurônios, isto é, pelas distâncias entre o padrão de entrada e cada um dos vetores protótipos associados aos respectivos neurônios.

Na camada de saída, os neurônios se encontram organizados regularmente em um arranjo geralmente unidimensional, bidimensional ou tridimensional. A configuração do arranjo determina o formato para a região de vizinhança de um neurônio (estabelecendo o grau do relacionamento de vizinhança entre neurônios) e atribui a cada neurônio da camada de saída coordenadas fixas no chamado espaço de saída. Na Figura 2.1, as linhas unindo os neurônios na camada de saída representam apenas quais neurônios são vizinhos imediatos de quais neurônios.

A rede SOM aproxima um espaço de entrada, normalmente representado por um elevado número de itens de dados (padrões de entrada), através de um conjunto finito de vetores protótipos. Tais vetores protótipos são vetores de características que são ajustados por um processo de aprendizado e podem ser vistos como coordenadas adaptáveis (durante o treinamento) dos neurônios no espaço de entrada, funcionando como apontadores para regiões deste mesmo espaço.

2.1.2 Treinamento

As redes SOM seguem o paradigma de aprendizado não-supervisionado. Desta forma, a única informação fornecida à rede SOM durante o treinamento está no conjunto de padrões de entrada e o ajuste dos protótipos é realizado através de um algoritmo constituído por um conjunto de regras de natureza local, sem a presença de um supervisor externo ou um esquema de punição/recompensa.

O algoritmo de aprendizado das redes SOM segue o paradigma de aprendizado competitivo em que os neurônios da camada de saída competem entre si para se tornarem ativos para um padrão de entrada. Vence o neurônio que tem o vetor protótipo mais similar ao padrão de entrada. Esta competição é chamada de o "vencedor-leva-tudo" (do inglês *winner-takes-all*). O vencedor e os neurônios vizinhos dentro de certo raio de vizinhança têm seus protótipos ajustados para se tornarem mais similares ao padrão de entrada.

Através desse processo de treinamento, no qual os neurônios competem entre si e o vencedor influencia seus vizinhos, os vetores protótipos têm seus valores adaptados e automaticamente ordenados no espaço de saída (arranjo de neurônios) de acordo com as suas similaridades mútuas.

O algoritmo de treinamento da rede SOM é aplicável a conjuntos de dados de grande volume. A complexidade computacional é $O(nmd)$, onde n é o número de vetores de entrada, m é o número de neurônios no mapa e d é a dimensionalidade dos vetores de entrada e vetores protótipos. Em muitos casos, esta complexidade é considerada $O(nm^2)$, ou seja, linearmente proporcional ao número de vetores de entrada (n) e quadraticamente proporcional ao número de neurônios no mapa (m). A memória necessária durante o treinamento basicamente é gasta no armazenamento dos vetores protótipos e vetores de entrada.

O algoritmo de treinamento da rede SOM é computacionalmente mais leve que suas variações [Kohonen et al. 2000], fator decisivo para sua utilização na criação de grandes mapas.

Existem dois tipos de algoritmos de treinamento: o algoritmo de aprendizado incremental e o algoritmo de aprendizado em lote. Embora os dois algoritmos gerem mapas diferentes a partir do treinamento de um mesmo mapa iniciado com valores

aleatórios para os protótipos, três processos estão envolvidos nos mesmos: competição, cooperação e adaptação.

1. **Competição** - Para cada padrão de entrada, os neurônios competem por mapear o padrão de entrada. O neurônio vencedor é aquele com vetor protótipo mais semelhante ao padrão de entrada.
2. **Cooperação** - O neurônio vencedor excita os neurônios dentro de um raio na sua vizinhança topológica, levando-os a cooperarem na representação do padrão de entrada.
3. **Adaptação** - Os neurônios excitados adaptam os protótipos para se tornarem mais semelhantes ao padrão de entrada, aumentando a densidade de vetores protótipos em torno daquele padrão de entrada.

O processo de competição consiste em encontrar no passo de iteração *t* o índice *c(x)* do neurônio vencedor para um dado padrão de entrada *x(t)* e depende da medida de distância utilizada.

Por exemplo, seja *c(x)* o índice do neurônio contendo o vetor protótipo $m_i(t) \in \Re^n$ que melhor aproxima o padrão de entrada $x(t) \in \Re^n$. Se é utilizada a distância euclidiana, tem-se

$$c(x) = \arg\min_i \{\|x(t) - m_i(t)\|\}. \quad (1)$$

Caso seja utilizada como métrica de distância o produto interno de vetores unitários, tem-se:

$$c(x) = \arg\max_i \{x(t) \bullet m_i(t)\}. \quad (2)$$

O processo de cooperação usa o conceito de vizinhança topológica de um neurônio para definir o campo de influência de um neurônio vencedor.

Na Figura 2.2 observam-se as formas em que a vizinhança topológica pode estar organizada no mapa unidimensional e bidimensional, bem como é ilustrada em tons de cinza escuros para claros as vizinhanças de raio 0, 1 e 2 do neurônio central do mapa. No caso bidimensional, a configuração de vizinhança hexagonal é mais utilizada por permitir uma melhor visualização do mapa [Kohonen et al. 2000].

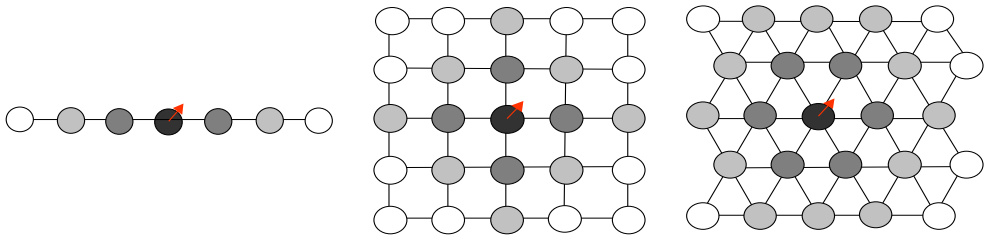

Figura 2.2 – Relação de vizinhança dos neurônios

A função vizinhança $h_{c(x),i}(t)$ é usada durante o processo de adaptação para definir a intensidade da adaptação dos vetores protótipos como função da distância de cada neurônio no arranjo ao neurônio vencedor. Esta é freqüentemente tomada como uma função gaussiana em que $0 < \alpha(t) < 1$ é a taxa de aprendizado que decresce monotonicamente com os passos do algoritmo, $r_i \in R^2$ e $r_{c(x)} \in R^2$ são respectivamente os vetores posição no arranjo do neurônio vizinho e do neurônio vencedor e $\sigma(t)$ é o tamanho da vizinhança que também decresce monotonicamente com os passos do algoritmo:

$$h_{c(x),\,i}(t) = \alpha(t)\exp\left(-\frac{\left\|r_i - r_{c(x)}\right\|^2}{2\sigma^2(t)}\right). \tag{3}$$

Na prática, por razões computacionais, a função de vizinhança é truncada quando $||r_i - r_{c(x)}||$ excede certo limite.

A regra de atualização dos protótipos usada durante o processo de adaptação depende da métrica de distância utilizada e também se o algoritmo é iterativo ou em lote, sendo especificada mais adiante.

No caso de um conjunto de dados discretos e uma função de vizinhança fixa h_{cj}, a função de erro da rede SOM é

$$E = \sum_{i=1}^{N}\sum_{j=1}^{M} h_{cj}\left\|x_i - m_j\right\|^2 . \tag{4}$$

em que N é o número de padrões para treinamento e M é o número de neurônios no mapa. A função de vizinhança h_{cj} é avaliada para o neurônio j sendo centrada no neurônio c que é o neurônio vencedor para o vetor x_i. Se o valor da função de vizinhança é igual a um para o neurônio vencedor e zero caso contrário, a rede SOM se reduz ao algoritmo K-means [Kohonen et al. 2000]. Se não for assim, conforme mostra a Equação (7) adiante, os vetores protótipos não são os centróides dos seus respectivos

conjuntos Voronoi, mas são as médias locais de todos os vetores no conjunto de dados ponderados pelo valor da função vizinhança.

A. Algoritmo de aprendizado incremental

A rede SOM é originalmente treinada de forma iterativa. Em cada passo do treinamento, um vetor x é randomicamente escolhido do conjunto de dados de entrada. As distâncias entre x e todos os vetores modelo são computadas (originalmente distância euclidiana). O neurônio vencedor é aquele cujo vetor protótipo mais se aproxima do vetor entrada. Então, os vetores protótipos do neurônio vencedor e os neurônios topologicamente na vizinhança são atualizados, movidos para a região do espaço próximo do vetor entrada.

O algoritmo de treinamento original da rede SOM é um processo de regressão recursiva [Kohonen et al. 2000]. Regressão de um conjunto ordenado de vetores protótipos $m_i \in R^n$ no espaço de vetores observação $x \in R^n$ pode ser feito recursivamente por

$$m_i(t+1) = m_i(t) + h_{c(x),\,i}(t)[x(t) - m_i(t)]\,, \tag{5}$$

em que t é o passo da regressão, sendo a regressão realizada a cada apresentação de um item (padrão) de x, denotado por $x(t)$. O multiplicador escalar $h_{c(x),i}(t)$ é a função vizinhança centrada no neurônio vencedor cujo índice é $c(x)$.

B. Algoritmo de aprendizado em lote

Buscando-se acelerar o treinamento da rede SOM, o princípio *batch-map* [Kohonen 2001] mostra-se muito efetivo. Assumindo que a convergência para um estado ordenado é verdadeira, espera-se que os valores de $m_i(t+1)$ e $m_i(t)$ para $t \to \infty$ sejam iguais; em outras palavras, no estado estacionário tem-se que

$$\forall i,\, E_t\{h_{c(x),\,i}(t)[x(t) - m_i^*(t)]\} = 0\,, \tag{6}$$

em que $E_t\{.\}$ é o valor esperado em t.

Por simplicidade, considere que a função de vizinhança mantenha-se invariante no tempo ao menos nas últimas iterações. No caso especial em que se tem um número finito de padrões de treinamento $x(t)$, baseando-se em (6) tem-se

$$m_i^* = \frac{\sum_t h_{c(x),\ i}\, x(t)}{\sum_t h_{c(x),\ i}}. \qquad (7)$$

Esta não é uma solução explícita para m_i^* porque o índice $c(x)$ ainda depende de $x(t)$ e todos os m_i^*. Entretanto, (6) pode ser resolvido iterativamente.

O algoritmo *batch-map* SOM foi formulado baseando-se em dois passos computacionais: quantização vetorial (do inglês *vector quantization*) e o suavizamento (do inglês *smoothing*) dos valores numéricos sobre a grid 2-D. O algoritmo segue abaixo.

"Dado um conjunto finito $\{x(t)\}$ de exemplos. Seja V_i o conjunto de todos os $x(t)$ que tem m_i^* como o vetor protótipo mais próximo – o chamado conjunto Voronoi de i. O número de exemplos $x(t)$ em V_i é chamado n_i.

1. Inicie cada m_i^* aleatoriamente.
2. Para cada neurônio encontre o seu conjunto Voronoi.
3. Compute um passo de quantização vetorial, em que $\bar{x}_i$ é o vetor médio de $x(t)$ sobre V_i:

$$\forall i, \bar{x}_i = \frac{\sum_{x(t)\in V_i} x(t)}{n_i} \qquad (8)$$

4. Suavizamento - Tome o novo vetor protótipo para cada neurônio como sendo a média ponderada dos vetores protótipos dos neurônios da vizinhança (inclusive). A ponderação é feita em relação ao tamanho do conjunto de cada neurônio e sua distância em relação ao neurônio corrente:

$$m_i^* = \frac{\sum_j n_j h_{ji} \bar{x}_j}{\sum_j n_j h_{ji}} \qquad (9)$$

5. Até que o conjunto de vetores $\{m_i^*\}$ seja estacionário vá para o passo 3."

Outra vantagem em utilizar o princípio *batch-map* é que a taxa de aprendizado não é utilizada. O único parâmetro ajustável do algoritmo é o raio da vizinhança.

C. Fases do treinamento e inicialização dos protótipos

O treinamento da rede SOM ocorre em duas fases: fase de ordenação e fase de convergência (ou ajuste fino).

Durante a fase de ordenação, ocorre a ordenação topológica dos vetores de protótipos iniciados aleatoriamente, criando um mapeamento grosseiro dos padrões de entrada. Esta fase dura em torno de 1000 iterações do algoritmo iterativo [Kohonen 2001] [Haykin 1999]. O valor inicial da taxa de aprendizado é escolhido do intervalo [1; 0,1], sendo os extremos do intervalo sugeridos por [Kohonen 2001] e [Haykin 1999] respectivamente. A taxa de aprendizado é então gradualmente reduzida até um valor próximo a 0,02 [Kohonen 2001], mas superior a 0,01 [Haykin 1999]. O tamanho da vizinhança é inicialmente próximo a metade do diâmetro do mapa e decresce até um valor próximo a um [Kohonen 2001] [Haykin 1999].

A fase de convergência faz um ajuste fino no mapa, melhorando o mapeamento realizado na fase anterior. Esta fase requer aproximadamente 10 vezes mais iterações que a fase anterior [Haykin 1999]. Para mapas já próximos do estado final e treinados com grande volume de dados, cinco iterações do algoritmo em lote foram utilizadas na fase de convergência em [Kohonen et al. 2000]. Como regra geral, o número total de iterações (iterações da fase de ordenamento mais iterações da fase de convergência) deve ser de 150 [Kohonen et al. 2000] a 500 [Kohonen 2001] vezes o número de neurônios na rede para garantir acurácia estatística. A taxa de aprendizado deve ser iniciada com um valor da ordem de ou menor do que 0,02 [Kohonen 2001] ou 0,01 [Haykin 1999] e deve decrescer sem, no entanto, chegar ao valor zero [Haykin 1999]. O raio de vizinhança deve cobrir os vizinhos imediatos de um neurônio (raio um) podendo chegar a cobrir somente o neurônio vencedor (raio menor do que um) [Kohonen 2001] [Haykin 1999]. Usualmente, durante a fase de convergência, o tamanho da vizinhança é deixado fixo [Kohonen 2001].

Ao iniciar o treinamento, os vetores peso têm seus valores escolhidos aleatoriamente dentro de algum intervalo. Escolhendo valores de magnitudes pequenas, próximas a zero, nenhuma ordem a priori é imposta ao mapa [Haykin 1999]. Outro modo de iniciar a rede é selecionar, de forma aleatória, vetores do conjunto de entrada.

A rede SOM pode ser iniciada com um estado já ordenado e grosseiramente de acordo com a função densidade de probabilidade do espaço de entrada. Se isto é feito o

processo de aprendizado converge rapidamente, podendo o raio de vizinhança inicial ser da ordem do valor final desejado e o valor inicial da taxa de aprendizado mais próximo de 0,2 ou 0,1 na fase de ordenação [Kohonen 2001].

A rede pode também ser iniciada tomando vetores distribuídos regularmente de uma projeção linear dos dados de entrada, denominada iniciação linear [Kohonen 2001]. Inicialmente são determinados os dois primeiros componentes principais da matriz de autocorrelação. Um arranjo retangular (com formato retangular ou hexagonal) é definido ao longo do subespaço expandido pelos dois componentes principais, com o centro coincidindo com a média dos padrões de entrada. Se for desejado um espaçamento uniforme dos vetores protótipos, o número relativo de posições na horizontal e vertical deve ser proporcional aos respectivos autovalores dos componentes principais. Utilizando este tipo de iniciação, o treinamento da rede SOM pode começar já na fase de convergência [Kohonen 2001].

2.2. Arquitetura do Sistema

Denomina-se sistema de organização automática de documentos, o sistema capaz de receber uma coleção de documentos texto e organizar os documentos desta coleção em uma estrutura em que são definidos grupos e a relação entre os grupos baseados no conceito de similaridade de conteúdo dos documentos. A estrutura gerada de forma automática tem aplicação na construção de sistemas de recuperação de informação sobre aquela coleção.

A Figura 2.3 mostra os processos realizados por um sistema de organização automática de documentos baseado em rede SOM.

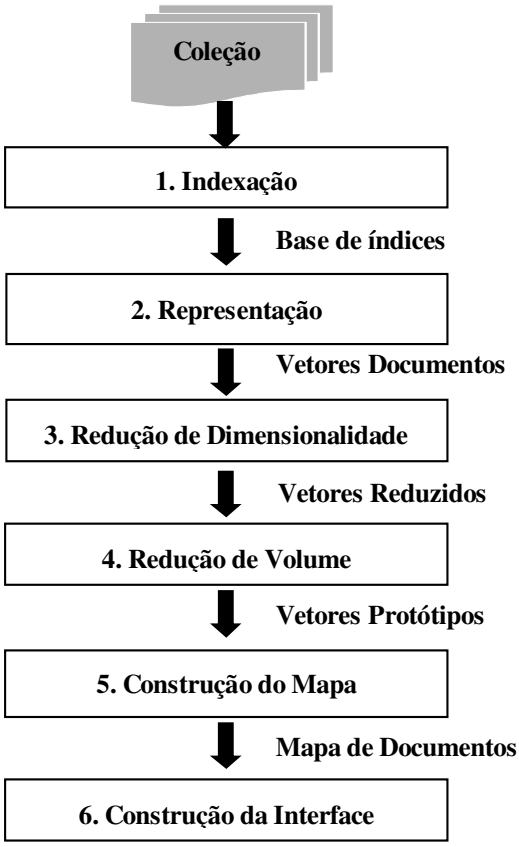

Figura 2.3 – Passos envolvidos na construção de um mapa de documentos.

As seções que se seguem detalham cada uma destas etapas.

2.2.1 Indexação

O processo de indexação consiste na obtenção de termos ou palavras-chave que melhor representam o conteúdo de cada um dos documentos da coleção. Estas palavras-chave são relacionadas em uma lista, chamada base de índices, que contém um apontador para os documentos que contém uma dada palavra-chave e pode também informar a freqüência de ocorrência daquela palavra-chave em cada documento. É através desta lista de índices que o documento poderá ser acessado pelo processo de recuperação de informação.

A indexação envolve comumente as seguintes etapas: análise léxica, eliminação de palavras irrelevantes, remoção de afixos das palavras. As próximas subseções fornecem maiores detalhes sobre cada uma destas etapas.

A. Análise Léxica do Texto

O objetivo da análise léxica do texto é eliminar ou tratar marcadores, dígitos e sinais de pontuação, isolar as palavras e efetuar a conversão das letras de maiúscula para minúscula, com a finalidade de montar a lista de palavras a serem consideradas como termos de indexação.

B. Eliminação de Palavras Irrelevantes

Esta fase é o primeiro passo na redução da dimensionalidade da lista de termos. A fase de eliminação de palavras irrelevantes consiste na exclusão de palavras que possuem nenhum ou pouco valor semântico, que ocorrem com alta freqüência nos documentos da coleção e não constituem elementos discriminativos. Em geral, estas palavras irrelevantes (do inglês *stopwords*) são filtradas dos documentos, consistindo, por exemplo, em artigos, preposições, conjunções e alguns advérbios, verbos e adjetivos. Uma das principais vantagens desta eliminação de palavras é a redução do tamanho do documento, em [Salton & McGill 1983] já era reportado que a eliminação destes termos muito freqüentes promoveria uma redução de 40 a 50% dos textos dos documentos a serem analisados. Adicionalmente, pode-se optar pela exclusão de *stopwords* específicas do domínio em questão. Por exemplo, para páginas da WWW, as palavras "home" e "html" aparecem com bastante freqüência em quase todo tipo de assunto.

Apesar de existirem diversas listas de palavras irrelevantes para a língua inglesa (listas conhecidas individualmente como *stoplist*), elas são muito parecidas e reúnem em torno de 250 a 540 termos. Um exemplo destas listas para língua inglesa é aquela sugerida pelo Laboratório de Recuperação de Informações da Universidade de Massachusetts em Amherst, relatadas no trabalho de [Lewis 1992a] e contém 266 termos.

C. Remoção dos afixos das palavras

A eliminação de afixos (sufixos e prefixos) das palavras tem como objetivo eliminar as várias palavras com a mesma raiz que possuem significados parecidos, facilitando o "casamento" entre os termos representativos presentes no índice e aqueles presentes nos documentos. Palavras no plural, alguns advérbios e flexões verbais são os exemplos mais comuns de variações sintáticas de uma palavra que podem ser tratadas

nesta fase do pré-processamento. Por exemplo, seria necessário um único radical "*connect*" para representar os termos "*connect*, *connected*, *connecting*, *connection*, *connections*". Desta forma, a cobertura na resposta a uma consulta que utilizasse um destes termos e passasse por um processo de remoção de afixos, poderia ser bastante significativa porque a resposta ao usuário conteria todos os textos daquela coleção, cujo radical "*connect*" participasse de sua chave de acesso. Por outro lado, a melhoria da cobertura pode implicar no decréscimo da precisão.

Além da vantagem de melhorar o desempenho do sistema com a redução de variantes de um mesmo radical para um mesmo "conceito", o processo de eliminação de afixos ou radicalização (do inglês *stemming*) também contribui para a redução do número de termos significativos no documento.

A construção de algoritmos de eliminação de afixos é fruto da análise morfológica cuidadosa do idioma em questão. Para a língua inglesa, o algoritmo de extração de radicais de Martin Porter [Porter 1980] é o mais conhecido e aceito pela comunidade científica. O algoritmo de extração de radicais de Porter realiza um processo de remoção de letras do final de palavras da língua Inglesa, que possuem mesma variação morfológica e de flexão, para isso, é utilizado um conjunto de regras.

2.2.2 Representação dos Documentos

Para que redes SOM e algoritmos de agrupamento possam ser utilizados na construção de mapas de documentos, estes documentos precisam ser representados por um vetor de características numéricas. A idéia fundamental para alcançar tal representação vetorial de documentos é que a semântica de um documento pode ser expressa por um conjunto de termos de indexação nele presentes. Os termos de indexação consistem em palavras isoladas ou grupos de palavras relacionadas, presentes no texto dos documentos. Assim, os documentos são representados por vetores, em que os índices correspondem aos termos utilizados e os valores em cada posição representam a importância do termo dentro do documento. Os vetores formados são chamados vetores documentos.

Para a tarefa de representação dos documentos e determinação do peso associado a cada termo em um documento, geralmente são utilizados dois modelos: o modelo booleano e o modelo espaço vetorial [Baeza-Yates & Ribeiro-Neto 1999].

O modelo booleano consiste em representar os documentos por vetores binários indicando a presença ou ausência de termos. O modelo booleano avalia somente a presença ou ausência do termo de indexação no documento, portanto, os pesos atribuídos a esses termos são binários, isto é {0,1}. Assim, o conteúdo dos documentos é especificado por uma conjunção de ocorrência de termos. As maiores vantagens do modelo booleano são sua simplicidade e por exigir menos espaço de armazenamento. A maior desvantagem é a de não oferecer uma ordem de relevância dos termos dentro dos documentos.

O modelo de representação de documentos mais utilizado e bem sucedido é o modelo espaço vetorial (do inglês *vector space* [Kohonen et al. 2000]). Segundo o modelo espaço vetorial, o documento é representado por um vetor real em que cada componente corresponde à freqüência da ocorrência de um termo no documento, ou seja, o documento é representado por um histograma de ocorrência dos termos. Desta forma, um histograma descreve a coleção de termos utilizados num documento bem como a freqüência de ocorrência de cada termo dentro do documento. A vantagem do modelo espaço vetorial é considerar a quantidade de vezes que os termos aparecem em cada documento, o que reflete a importância de cada termo no texto. As desvantagens do modelo espaço vetorial são o custo computacional do cálculo dos pesos e o espaço de armazenamento requisitado.

Uma extensão do modelo espaço vetorial consiste em ponderar o histograma com pesos que correspondem à importância discriminatória de cada palavra. Os pesos podem ser computados por vários métodos bem especificados na literatura de recuperação de informação [Sebastiani 2002], como o método *Inverse Document Frequency* (IDF) que atribui como peso a uma palavra um valor inversamente proporcional ao número de documentos em que ela ocorre. A representação dos documentos baseada no modelo espaço vetorial ponderado com IDF é denominada *tfidf* [Salton & McGill 1983], atualmente esta é a representação mais utilizada na literatura.

A grande desvantagem em se utilizar vetores para representar os documentos é a grande dimensionalidade dos vetores que pode chegar a ser igual ao tamanho do vocabulário da coleção, sendo necessário muitas vezes limitar o tamanho do vocabulário para impedir a dimensionalidade de crescer muito. Na indexação de grandes coleções de

textos em linguagem natural, necessita-se de um grande vocabulário, algo em torno de 50.000 palavras [Kohonen et al. 2000].

2.2.3 Redução de Dimensionalidade

Para uma coleção com documentos extensos, ou com grande número de documentos, a dimensionalidade dos vetores documentos poderá ser intratável pelos algoritmos de aprendizado (da ordem de dezenas de milhares).

A dimensionalidade dos vetores documento pode ser reduzida utilizando-se métodos de seleção de características ou métodos de extração de características.

Na primeira abordagem, as características selecionadas são um subconjunto das características originais. Esta redução pode ser realizada através da seleção daquelas melhores, de acordo com algum critério de representatividade do conteúdo dos documentos ou de categorias de documentos.

Na segunda abordagem, as características são obtidas pela combinação ou transformação das originais, através da síntese de um conjunto n' de novas características, a partir do conjunto de características N' original, de modo a maximizar a eficácia do sistema. Os métodos de extração de características têm como principal objetivo a criação de características artificiais que não sofram dos problemas de polissemia[1] e sinonímia[2]. Uma técnica bastante utilizada para este fim é a Semântica Latente (*Latent Semantic Indexing*) [Deerwester et al. 1990].

A abordagem da seleção de características possui vantagens significativas, tais como sua simplicidade computacional e a interpretabilidade direta do conjunto de características resultante. Entretanto, algumas de suas desvantagens como não reduzir informação redundante (termos correlacionados) e a exclusão de termos individuais não significativos que poderiam ter poder discriminativo em combinação com outros, abrem espaço para pesquisas com técnicas de extração de características.

Há quatro maneiras principais de se reduzir a dimensionalidade do vetor documento, sem perder essencialmente o poder de discriminação entre documentos:

---

[1] Termo lingüístico que possui vários sentidos, e.g., a palavra manga.

[2] Várias palavras que possuem o mesmo significado semântico, e.g., as palavras pinga e aguardente.

1) O uso de métodos de seleção de características – os métodos de seleção de características visam selecionar do conjunto de características ou dimensões originais um subconjunto menor e mais significativo segundo uma métrica de importância. Entre estes métodos o Chi-square e o método de Ganho de Informação são citados como mais efetivos na categorização dos documentos [Yang & Pedersen 1997], sendo ambos de caráter supervisionado. Um método não supervisionado utilizado com freqüência nos trabalhos é o baseado em freqüência de documentos (do inglês *document frequency*), este método consiste em descartar palavras ou termos que tem uma freqüência muito alta ou muito baixa no conjunto de documentos da coleção.
2) *Latent Semantic Indexing* (LSI) [Deerwester et al. 1990] – trata-se de um método de extração de características que consiste em organizar uma matriz em que cada coluna corresponde ao vetor representativo de um documento da coleção e então decompor o espaço expandido pelos vetores coluna em um conjunto ordenado de fatores pelo método SVD (do inglês *singular-value decomposition*). A decomposição tem a propriedade de que os últimos fatores têm influência mínima sobre a matriz. Os fatores que menos influenciam podem ser descartados, diminuindo a dimensionalidade. Este método tem uma complexidade alta, porém produz melhores resultados na recuperação de informação.
3) Agrupar termos em categorias semânticas e trabalhar com histograma de categorias – É um método de extração de características que consiste basicamente num método de agrupamento aplicado sobre alguma representação dos termos. No sistema WEBSOM1 [Kaski et al. 1998], a redução de dimensionalidade foi realizada através do agrupamento das palavras utilizando a rede SOM. O mapa de categorias de palavras foi treinado com os vetores formados pela concatenação de três vetores representativos de palavras adjacentes sob uma janela deslizante ao percorrer o texto dos documentos, sendo cada palavra no vocabulário representada por um único vetor aleatório. Os documentos são representados por vetores histogramas no mapa de categorias de palavras, sendo estes últimos utilizados para treinar a rede SOM. Este método foi abandonado por ser menos efetivo que a projeção aleatória, utilizada no WEBSOM2 [Kohonen et al. 2000].

4) Reduzir a dimensionalidade dos vetores documento por um método de projeção aleatória denominado RM (do inglês *Random Mapping*) [Kaski 1997] – este método de extração de características é utilizado no sistema WEBSOM2 [Kohonen et al. 2000] por ser computacionalmente mais leve que os métodos apresentados em 2) e 3) e por permitir ao sistema uma acurácia na categorização de documentos próxima a obtida com LSI [Kohonen et al. 2000]. Consiste na obtenção de um vetor projetado x_i de dimensão m, pela multiplicação do vetor n_i, de dimensão n, por uma matriz de projeção R_{mxn} ($m << n$), na qual os valores presentes nos vetores coluna são assumidos normalmente distribuídos e gerados aleatoriamente, sendo normalizados para o tamanho unitário. Uma versão do método de projeção aleatória chamada SRM (do inglês, *Sparse Random Mapping*) [Kohonen et al. 2000], que consiste na construção de uma matriz esparsa de projeção apenas contendo zeros e uns foi estabelecida como padrão no método WEBSOM2.

Independentemente do método a ser utilizado, encontrar um subconjunto de termos ou um conjunto de novas variáveis extraídas que represente as características essenciais dos documentos é um problema difícil.

2.2.4 Redução de Volume

A utilização de redução de volume na organização automática de documentos foi proposta por [Azcarraga & Yap 2001]. Consiste de um procedimento de quantização vetorial, onde um conjunto de padrões de treinamento é representado por um conjunto menor de vetores chamados protótipos. Algoritmos de agrupamento podem ser utilizados nesta etapa para obtenção dos vetores protótipos de forma não supervisionada.

Em [Azcarraga & Yap 2001] foi proposto o primeiro sistema híbrido de organização automática de documentos baseados em SOM. Nesse trabalho, vetores obtidos como saída de um algoritmo de geração de protótipos foram utilizados para treinar uma rede SOM. Entretanto, neste trabalho apenas foi proposto e avaliado um algoritmo supervisionado de aprendizado incremental para geração de protótipos. Além disso, não foi reportada uma análise do tempo total de treinamento (geração dos protótipos e treinamento da rede SOM) e na avaliação da qualidade do mapa obtido não foi reportada medida de avaliação padrão.

Até o início do presente trabalho, esta linha de pesquisa se encontrava abandonada.

2.2.5 Construção do Mapa de Documentos

Este passo corresponde ao treinamento de uma rede SOM com os vetores documentos obtidos nas fases anteriores.

Diretrizes para treinar as redes SOM tornam-se necessárias e vários são os parâmetros a serem especificados, por exemplo: as dimensões do mapa, número de épocas de treinamento, tamanho da vizinhança, taxa de aprendizado, além dos vetores protótipos iniciais.

O treinamento da rede SOM pode ser feito em um estágio ou múltiplos estágios [Kohonen et al. 2000]. O treinamento feito em um estágio consiste em treinar um mapa iniciado aleatoriamente até este chegar a um estado estacionário, passando pelas fases de ordenação e convergência uma única vez, este é modo tradicional de realizar o treinamento de um mapa SOM. O treinamento em múltiplos estágios consiste em treinar um mapa pequeno em um estágio e posteriormente realizar vários estágios de estimação e refinamento de um mapa maior baseado no estado estacionário de um menor, passando assim por tantas fases de convergência quantas forem os estágios de estimação de mapas maiores.

A grande maioria dos trabalhos utiliza o treinamento em um estágio, sendo também comum o uso de uma organização hierárquica de mapas treinados independentemente em detrimento do uso de um único grande mapa.

Várias alterações no algoritmo de treinamento da rede SOM foram propostas a fim de acelerar a construção de grandes mapas em [Kohonen et al. 2000]. Torna-se necessário uma avaliação mais sistemática da qualidade da solução obtida.

2.2.6 Construção da Interface com o Usuário

Após o treinamento com vetores representativos dos documentos, a rede SOM passa por um processo de rotulação no qual são atribuídos a cada neurônio palavras-chave e/ou a categoria da maioria dos documentos do conjunto de treinamento nela mapeados (esta última opção é possível caso os documentos estejam manualmente categorizados).

O mapa pode ser visualizado através de um gráfico de suporte bidimensional, no qual se encontram projetados os vetores representativos de cada cluster. Este gráfico pode ser apresentado ao usuário como uma imagem interligada a sub-gráficos ou a páginas HTML descrevendo o conteúdo de neurônios, de forma a permitir: (i) a exploração dos neurônios e dos documentos dentro de cada neurônio; (ii) visualizar as regiões em que há documentos que satisfazem uma busca dirigida por palavras-chave ou por endereçamento de conteúdo.

Para mapas auto-organizáveis, a imagem pode ser obtida através da matriz de distâncias entre os vetores protótipos dos neurônios do mapa [Ultsch & Siemon 1990] [Kraaijveld et al. 1995] bem como através da densidade de documentos em cada neurônio [Kohonen et al. 2000]. Mapeando as distâncias ou densidades em uma escala de cores, a imagem passa a expressar o nível de proximidade de conteúdo ou distribuição dos documentos em cada região do mapa respectivamente. Há trabalhos que determinam áreas nos mapas contendo neurônios de conteúdo similar por meio do agrupamento dos que possuem palavras-chave em comum, sendo as palavras-chave obtidas através do processo de rotulação dos neurônios [Lin et al. 1991] [Roussinov & Chen 1998].

A imagem geralmente é enriquecida com palavras automaticamente selecionadas que caracterizam os documentos em cada região ou neurônio no mapa, este processo é chamado rotulação. As palavras-chave ficam distribuídas sobre o mapa e servem como pontos de referência durante a navegação, fornecem informação sobre tópicos discutidos pelos documentos na respectiva área, auxiliam a encontrar informações interessantes e na coletividade podem servir como sumário da coleção de documentos.

As interfaces reportadas na literatura são bastante intuitivas e facilitam a exploração iterativa de uma coleção de documentos e podem ser construídas para qualquer tipo de coleção de documentos textuais.

Em [Lagus & Kaski 1999] foi proposto um método para escolha e posicionamento dos rótulos em regiões. Os demais trabalhos na literatura propuseram apenas métodos para determinar rótulos a fim de caracterizar o conteúdo dos neurônios, por exemplo o método LABELSOM [Rauber 1999].

Uma metodologia para realizar busca por palavras-chave em mapas de documentos é introduzida e avaliada em [Lagus 2002]. Pesquisas por palavras-chave e pesquisas baseadas em conteúdo são essenciais na implementação de sistemas de recuperação de documentos.

2.3. Avaliação do Sistema

As medidas de eficácia na categorização de documentos permitem uma avaliação menos subjetiva da qualidade dos mapas de documentos gerados. A definição e cálculo destas medidas de desempenho encontram-se no Apêndice A. A acurácia ou erro de classificação na categorização de documentos foi inicialmente utilizada como índice de validação externo para mapas SOM em estudos de caso do projeto WEBSOM [Kohonen et al. 2000] e em [Strehl et al. 2000]. Mais recentemente, a F-measure tem sido utilizada para avaliar a qualidade de mapas de documentos [Ontrup & Ritter 2006] [Freeman & Yin 2004].

Poucos trabalhos fazem uma análise da eficiência de seus sistemas, reportando a complexidade computacional e o tempo total de treinamento.

2.4. Estado da Arte

A capacidade da rede SOM em organizar os padrões de entrada de forma que padrões similares encontrem-se mapeados em neurônios próximos, a torna muito útil na organização de grandes coleções de dados em geral, incluindo coleções de documentos.

Baseando-se na hipótese de que características textuais elementares dos documentos que abordam tópicos similares são estatisticamente similares, a rede SOM tem sido aplicada na tarefa de organização automática e classificação de documentos.

Nesta seção, são dispostos em ordem cronológica os trabalhos, projetos e sistemas presentes na literatura, focando as metodologias e contribuições na construção de sistemas de organização automática de documentos texto utilizando redes neurais SOM. A ordem cronológica retrata aproximadamente a evolução das pesquisas na área, podendo ser dividida nas seguintes fases: primeiros trabalhos, grandes projetos, diversificação e convergência.

Cada uma dessas fases e respectivos trabalhos são descritos nas próximas subseções.

2.4.1 Primeiros Trabalhos (1991 – 1995)

Nesta fase, estão descritos os primeiros trabalhos utilizando redes SOM para organização de coleções de documentos. A preocupação maior destes trabalhos foi mostrar a viabilidade e utilidade da aplicação de redes SOM para a organização e posterior exploração de coleções de documentos. Os trabalhos envolviam coleções pequenas (da ordem de uma centena de documentos), mapas bidimensionais de dimensões 10x14 ou 10x10, os documentos eram representados por vetores binários indicando a presença ou ausência de certos termos ou palavras nos documentos e o algoritmo utilizado era o algoritmo padrão de treinamento da rede SOM com distância euclidiana.

A primeira tentativa de utilizar redes SOM para a tarefa de recuperação de informação foi descrita em [Lin et al. 1991] inspirado no trabalho de [Doley 1961], onde são lançadas diretrizes para construção de um mapa semântico de documentos, uma interface entre um pesquisador e uma coleção de documentos. Na Figura 2.4, é exibido o mapa de documentos obtido neste trabalho. Neste artigo os autores utilizaram redes SOM para classificar 140 documentos do *LISA database*. Os documentos foram caracterizados por 25 termos de indexação retirados dos respectivos títulos. Estes termos de indexação foram escolhidos após eliminação das palavras irrelevantes, redução das palavras restantes aos respectivos radicais e posterior eliminação dos termos mais freqüentes e dos que ocorreram menos que três vezes. Os documentos foram representados por vetores binários. O mapa possuía 10x14 neurônios e foi treinado por 2.500 ciclos. Os documentos foram mapeados em regiões conceituais que foram formadas e rotuladas em um arranjo bidimensional. A rotulação do mapa foi realizada em dois estágios: primeiramente foram definidas as áreas conceituais, isto foi feito comparando cada neurônio com todos os vetores unitários contendo somente uma palavra e atribuindo a cada um a palavra que o vetor representa, as áreas foram definidas por neurônios vizinhos que possuíam características comuns; as áreas foram então rotuladas comparando cada vetor unitário a todas os neurônios e rotulando o neurônio vencedor com a palavra correspondente ao vetor unitário, as palavras foram agrupadas se mapeadas na mesma área (palavras freqüentemente co-ocorrentes). Pela dimensão da coleção e o número de termos de indexação este experimento é pouco realístico.

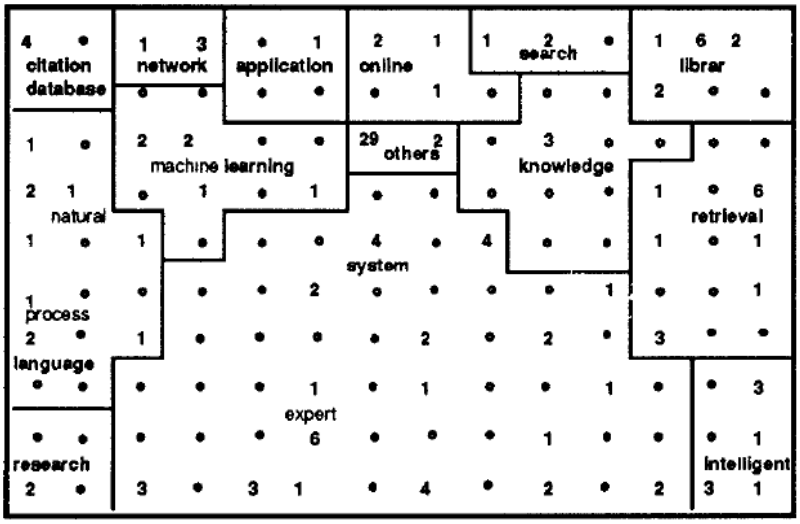

Figura 2.4 – Mapa de Documentos extraído de [Lin et al. 1991]**.**

Seguindo a linha de pesquisa de [Lin et al. 1991], Merkl e Tjoa [Merkl & Tjoa 1994] sugeriram o uso de SOM para organizar bibliotecas de softwares. A coleção utilizada consistiu da especificação textual de 36 comandos do sistema operacional MSDOS. Os documentos foram representados por vetores binários com 39 posições, correspondendo às 39 palavras selecionadas excluindo-se o nome dos comandos. Foi utilizado um mapa de 10x10 neurônios treinado por 9.000 iterações. O mapa foi rotulado atribuindo o nome de cada comando ao neurônio que melhor o representa. Experimentos foram realizados para analisar a capacidade de recuperação de informação gerada pelo mapa.

Em [Merkl 1995a] e [Merkl 1995b] a rede SOM é utilizada para organizar uma coleção de documentos consistindo da descrição textual de 80 classes da biblioteca NHI Class Library (NHICL). Os documentos foram representados por vetores binários contendo a informação sobre a presença ou não de 489 palavras-chave. O mapa possuía 10x10 neurônios e foi treinado por 8.000 iterações.

Na área de processamento de informação legal, a rede SOM foi utilizada em [Merkl et al. 1994] [Schweighofer et al. 1995] para análise exploratória de conceitos judiciais no espaço de documentos.

2.4.2 Grandes Projetos (1996 – 2000)

Nesta fase, grandes projetos foram lançados visando à criação de sistemas de organização automática de documentos utilizando redes SOM que gerassem uma interface de navegação intuitiva e fossem escaláveis. Inicialmente nesta fase os trabalhos organizavam coleções da ordem de centenas de documentos, passando a utilizar coleções da ordem de dezenas de milhares de documentos e no final chegando a organizar coleções da ordem de centena de milhares e milhões de documentos. Os maiores mapas possuíam da ordem de milhares a milhões de neurônios. Novas maneiras de representar os documentos foram propostas, o produto interno de vetores unitários passou a ser utilizada fazendo contraponto à distância euclidiana e várias otimizações do algoritmo de treinamento da rede SOM foram propostas e utilizadas. Passou a ser explorado também nesta fase o conceito de hierarquia de mapas SOM para organizar coleções.

São quatro os grandes projetos lançados nesta fase, sendo cada um descrito a seguir.

A. Arizona Digital Library

Desenvolvido pelo grupo de pesquisa do laboratório de inteligência artificial da Universidade de Arizona, conhecido como *Arizona AI Lab*[3], este projeto adotou e revisou SOM para classificação de documentos texto e visualização [Chen et al. 1996] [Orwig et al. 1997] [Roussinov & Chen 1998]. O objetivo geral deste projeto era construir ferramentas para a construção de bibliotecas digitais [Schatz & Chen 1996] [Schatz et al. 1996].

Adaptando a idéia de Ichiki et al. [Ichiki et al. 1991], que propuseram uma arquitetura multicamadas de mapas SOM para classificação, em [Chen et al. 1996] foi proposto a rede MSOM (*multi-layered* SOM). A utilização de MSOM permitiu a categorização de coleções de média escala, isto é da ordem de dezenas de milhares de documentos. Em [Chen et al. 1996] 10.000 páginas da web relacionadas com entretenimento foram retiradas da porção *entertainment* da hierarquia de diretórios do

[3] http://ai.bpa.arizona.edu/research/dl/index.htm

Yahoo[4] e categorizadas (classificadas) de acordo com o conteúdo das mesmas utilizando uma hierarquia de redes SOM. Este foi o primeiro trabalho a utilizar hierarquia de redes SOM para organização de coleções de documentos. Os documentos foram representados por vetores binários. Inicialmente um mapa é construído e recursivamente, para cada região no mapa com mais de 100 homepages, outro mapa é construído num nível mais baixo. Os mapas-folhas desta árvore agrupavam no máximo 100 homepages. Ao todo foram quatro níveis. Todos os mapas tinham dimensão 20x10 neurônios. Detalhes sobre o treinamento das redes não foram reportados. A rotulação dos neurônios foi realizada submetendo vetores com um único termo não nulo para a rede treinada e atribuindo ao neurônio vencedor o nome do termo. Neurônios na vizinhança que contém o mesmo nome (termo) formam então uma região. Regiões que são similares (conceitualmente) aparecem próximas umas das outras. A hierarquia de mapas pode ser vista no endereço eletrônico: http://ai.bpa.arizona.edu/ent/et-map.html . Foram realizados experimentos visando a avaliação por usuários da utilização do mapa como interface de navegação e busca de homepages.

Orwig et al. [Orwig et al. 1997] descreveram a aplicação da rede SOM para o problema de classificação da saída do *eletronic brainstorming* (EBS) e avaliação dos resultados. EBS é uma das ferramentas mais produtivas no sistema de cooperação eletrônica chamado *GroupSystems*. O maior passo na resolução de problemas por meio de grupos de discussão envolve a classificação da saída do EBS em uma lista de conceitos, tópicos ou assuntos que pode ser mais bem analisada pelo grupo. A sobrecarga de informação e a demanda cognitiva para processar uma grande quantidade de texto tornam este passo problemático. Comparando a saída produzida por uma rede SOM com a produzida por uma rede Hopfield e por especialistas, usando o mesmo conjunto de dados, foi encontrado que a rede SOM teve melhor desempenho que a rede Hopfield, obtendo desempenho igual ao dos especialistas em representar a associação de termos na saída do EBS, em outras palavras, as categorias geradas pela rede SOM foram comparáveis àquelas geradas por humanos. Adicionalmente, a cobertura obtida pela rede SOM foi equivalente à obtida pelos especialistas, porém a precisão foi menor. Os 202 comentários do EBS foram representados por vetores binários de 190

[4] http://www.yahoo.com

características, correspondendo aos 190 termos mais freqüentes na coleção. O mapa construído possuía 20x10 neurônios com vizinhança hexagonal e foi treinado em duas fases: na fase inicial foram utilizadas 1000 iterações, uma taxa de aprendizado de 0,05 e um raio de vizinhança compreendendo 10 neurônios; na fase de refinamento foram utilizadas 10.000 iterações, uma taxa de aprendizado de 0,02 e um raio de vizinhança de três neurônios. A rotulação do mapa foi realizada atribuindo a cada neurônio o termo (vetor unitário contendo um único termo) apropriado e os agrupando em regiões e depois rotulando estas regiões.

Em [Roussinov & Chen 1998] é apresentado um sistema escalável de categorização e classificação textual baseado na rede SOM, o SSOM (*Scalable Self-organizing Map*), que pode ser utilizado para geração automática de tesauro (do latim *thesaurus)*. Tesauro é uma lista estruturada de termos associados que descrevem relações entre conceitos, podendo estar organizada por classe de assuntos, de forma hierárquica ou não e ser gerada utilizando diversos métodos. Nesta pesquisa, este algoritmo é utilizado com o objetivo de extrair uma taxonomia hierárquica de grupos de documentos bem como conceitos (categorias) descobertos neles. Os documentos são representados por vetores binários. Depois da rede ter sido treinada, regiões do mapa representam conceitos ou categorias. Um rótulo é criado para um neurônio atribuindo o termo que corresponde à coordenada de maior valor no respectivo vetor protótipo, o chamado termo vencedor. Regiões vizinhas tendo os mesmos termos vencedores são agrupadas para produzir regiões e o termo vencedor é designado como categoria para toda a região. Os documentos apresentados são categorizados quando colocados em áreas no mapa próximas aos conceitos que representam. Documentos pertencentes à mesma categoria são recursivamente usados para produzir mapas menores que correspondem a um nível mais baixo na hierarquia de conceitos. Esta hierarquia é compreendida como um tesauro.

O SSOM foi aplicado nas coleções utilizadas em [Chen et al. 1996] e [Orwig et al. 1997] variando-se a dimensionalidade no intervalo de 25 a 400 termos, bem como na coleção COMPENDEX. O objetivo destes experimentos foi medir a escalabilidade do algoritmo SSOM. A coleção COMPENDEX consiste em 247.721 resumos de artigos de campos relacionados à engenharia elétrica, ciência da computação, sistemas de informação, etc. Depois da indexação da coleção COMPENDEX, 160.000 termos

únicos foram extraídos, não contando palavras irrelevantes (presentes na *stoplist*) ou palavras que aparecem menos de três vezes. Destes termos, os 10.000 mais freqüentes foram escolhidos como características. Um mapa de 30x30 neurônios foi treinado por 4.500.000 ciclos de treinamento, de forma que cada documento foi apresentado 15 vezes. O processo de treinamento durou 15,7 CPU horas numa máquina DEC Alpha 3000/600 200 MHz com 128 MB de RAM. Utilizando o algoritmo de treinamento da rede SOM esta tarefa duraria 1000 vezes mais tempo (aproximadamente 625 CPU dias).

O mapa gerado na organização das 10.000 homepages relacionadas a entretenimento é mostrado na Figura 2.5. Cada região representa um importante tópico relacionado com entretenimento e o número de homepages classificadas em cada região é também demonstrado. No endereço http://ai.bpa.arizona.edu/ent/ pode-se utilizar a versão iterativa do mapa, onde cada região pode ser clicada para visualizar um mapa SOMmais detalhado representando sub-tópicos classificados na região.

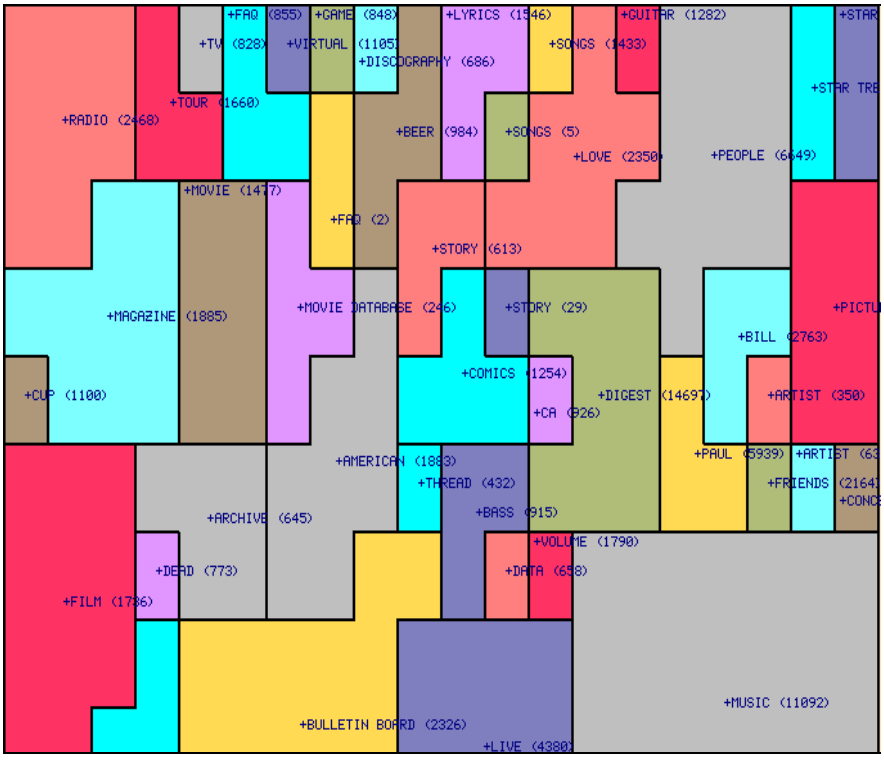

Figura 2.5 – Mapa de Documentos obtido de [Roussinov & Chen 1998].

B. WEBSOM

WEBSOM foi um projeto desenvolvido no *Neural Network Research Centre da Helsinki University of Technology* liderado por T. Kohonen, com o objetivo de

automaticamente ordenar, ou organizar, coleções arbitrárias de documentos textos não estruturados possibilitando a fácil navegação e exploração destas coleções [Honkela et al. 1996a]. O WEBSOM utiliza a rede SOM para automaticamente organizar os documentos num mapa de neurônios. O site do projeto se encontra no endereço eletrônico: http://websom.hut.fi/websom/. A organização dos documentos foi realizada segundo dois métodos:

1. Dois níveis de análise: no primeiro, as categorias de palavras são extraídas por uma rede SOM, baseada na informação estatística dos curtos contextos em que as palavras aparecem no texto dos documentos, este mapa de categorias de palavras é denominado "mapa semântico"; no segundo, cada documento é codificado pelo histograma de categorias formado sobre o mapa semântico, sendo organizados por outra rede SOM denominada "mapa de documentos". Este método foi denominado WEBSOM1;
2. Um nível de análise: os vetores documentos (histogramas ponderados de palavras) são projetados aleatoriamente para um espaço de dimensionalidade reduzida e são agrupados por uma rede SOM. Este método foi denominado WEBSOM2.

O maior mapa de documentos construído usando o WEBSOM1[Kohonen 1998] consistiu de 104.040 neurônios. A dimensão dos vetores protótipos foi 315. O mapa semântico possuía 13.432 neurônios e os histogramas de categorias gerados a partir deste foram randomicamente projetados para 315 dimensões. A coleção de documentos utilizada consistia de 1.124.134 mensagens retiradas de 80 diferentes *Usenet newsgroups*. O vocabulário final consistiu de 63.773 palavras. A computação dos dois mapas durou cerca de um mês de computação. A acurácia de classificação dos documentos em um dos 80 grupos foi cerca de 80%. As dimensões do mapa de documentos foram determinadas de forma que, em média, 10 a 15 mensagens fossem mapeadas em um neurônio no mapa. Detalhes sobre o treinamento dos mapas não foram reportados. O método WEBSOM2 consistiu na utilização do algoritmo *batch-map* [Kohonen et al. 2000] para treinamento da SOM, bem como na adoção de medidas para diminuir a complexidade da construção de grandes mapas. O mapa semântico foi abandonado após a realização de experimentos que identificaram que a utilização da

projeção randômica dos histogramas de palavras gerava um melhor desempenho na classificação de documentos.

O maior mapa de documentos construído utilizando o método WEBSOM2 [Kohonen et al. 2000] consistiu de 1.002.240 neurônios e foi utilizado para organizar 6.840.568 resumos de patentes escritos em língua inglesa. A coleção foi subdivida em 21 categorias, correspondendo a subseções do sistema de categorização de patentes (agricultura, transporte, química, construção, engenhos, eletricidade, etc.). Os documentos da coleção foram pré-processados transformando as palavras em seus radicais utilizando um algoritmo de radicalização (do inglês *stemmer)* [Koskenniemi 1983] e eliminando aquelas entre as 1335 consideradas irrelevantes ou que apareceram menos que 50 vezes na coleção. O vocabulário final consistiu de 43.222 termos. Os vetores documentos foram randomicamente projetados para a dimensão 500. A construção do mapa durou cerca de seis semanas em um computador SGI O2000 com seis processadores e exigiu 800 Mb de memória RAM. O mapa obteve uma acurácia de 64% na categorização dos documentos. Os parâmetros necessários durante o treinamento do mapa não foram reportados.

Foi criado o servidor WEBSOM [Honkela et al. 1996b], que demonstra a capacidade de categorizar milhares de mensagens de newsgroups. Uma interface gráfica bidimensional intuitiva foi criada para navegação no mapa SOM. Uma metodologia de rotulação automática de neurônios e regiões do mapa foi desenvolvida [Lagus & Kaski 1999]. A interface de visualização do mapa foi construída baseada na WWW, permitindo que documentos em interessantes áreas do mapa pudessem ser visualizados. A navegação pode também ser estendida para tópicos relacionados, que aparecem em áreas próximas no mapa. A técnica usada para ilustrar o mapa foi projetar os vetores protótipos num espaço de cor de forma que a neurônios similares do mapa são atribuídas cores similares [Varfis 1993] [Kaski et al. 1999]. A imagem do mapa foi enriquecida utilizando-se de um método automático para selecionar palavras-chave. As palavras-chave ficam distribuídas sobre o mapa e servem como pontos de referência durante a navegação, fornecendo informação sobre tópicos discutidos pelos documentos na respectiva área, auxiliando a encontrar informações interessantes e na coletividade podem servir como sumário da coleção de documentos.

A Figura 2.6 é a imagem do mapa criado a partir de mensagens do *newsgroup comp.ai.neural-nets* disponibilizado no site do projeto WEBSOM. As áreas mais claras representam áreas de grande densidade de mensagens.

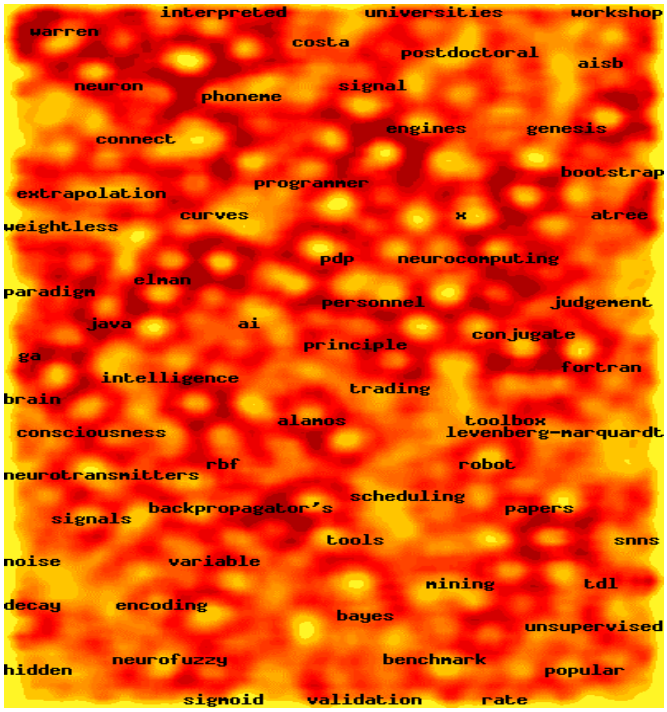

Figura 2.6 – Mapa de mensagens do newsgroup comp.ai.neural-nets disponível no servidor WEBSOM.

Em [Lagus 2002] um método para realizar busca por palavras-chave em mapas de documentos foi proposto e avaliado, tendo sido utilizada a coleção CISI, um *benchmark* na área de recuperação de informação. Os resultados obtidos foram animadores, já que o método proposto teve melhor desempenho médio que métodos consagrados da área.

O método WEBSOM foi relatado de forma mais abrangente em [Lagus et al. 2004]. No entanto, nenhuma novidade foi apresentada.

C. SOMLib

Uma série de experimentos usando HFM (do inglês *Hierachical Feature Maps*) [Miikkulainen 1990] para organização de documentos foi reportado. HFM consiste de uma hierarquia de redes SOM treinadas de forma independente. Esse modelo de rede neural permite agrupamento hierárquico de documentos por contar com uma arquitetura predefinida de redes SOM dispostas de forma hierárquica. As unidades da hierarquia de

mapas SOM são treinadas de forma independente, sendo que para cada unidade no mapa num nível da hierarquia uma rede SOM é treinada no nível inferior. Os benefícios são a redução dramática no tempo de treinamento, e uma representação intuitiva da similaridade de documentos. A maior desvantagem é a necessidade de definir a arquitetura hierárquica da rede antes do treinamento o que requer certo conhecimento das características particulares da coleção de documentos a ser organizada.

Em [Merkl 1997] e [Merkl 1998] a HFM foi utilizada para organizar classes da biblioteca NHI Class Library (NHICL). Em [Merkl & Schweighofer 1997] foi organizada uma coleção contendo os 100 mais importantes tratados internacionais. Em [Merkl & Rauber 1998] os formulário da *CIA world factbook* relativos a descrição de cada país no mundo foram organizados.

Em [Merkl & Rauber 1998], trabalho divulgado paralelamente aos da criação do projeto SOMLib, a rede SOM foi usada para organizar os formulários da *CIA world factbook* (edição de 1990) sobre os países do mundo. Foram ao todo 245 documentos representados por 959 termos. Foram eliminadas as palavras que apareciam em mais de 196 documentos e menos de 15 documentos. Um mapa de 10x10 neurônios foi criado e rotulado com o nome dos países. Neste trabalho a rede SOM foi comparada com a hierarquia de mapas gerada pela rede HFM, sendo esta última mais rapidamente construída e capaz de revelar maiores detalhes sobre regiões específicas do espaço de entrada.

O projeto SOMLib[5] [Rauber & Merkl 1999] foi criado com o objetivo de construir um sistema para bibliotecas digitais que usa como núcleo redes neurais SOM para representar coleções de documentos e processar buscas sobre as mesmas. Em [Rauber & Merkl 1998] uma coleção distribuída de documentos foi organizada através de hierarquia de mapas SOM onde cada mapa foi automaticamente rotulado. Neste trabalho, seis mapas treinados de forma independente foram integrados por meio de um mapa de 10x15 neurônios. Cada um destes seis mapas representa entre 53 e 87 artigos da coleção *TIME Magazine*, o mapa de integração representa toda a coleção de 420 artigos. A dimensão dos seis mapas variou de 42 a 70 neurônios e a dimensionalidade dos vetores protótipos variou de 1.255 a 2019 termos. A dimensionalidade do mapa que

[5] http://www.ifs.tuwien.ac.at/~andi/somlib/

faz a integração foi de 3303 termos correspondendo à conjunção das características dos vetores protótipos dos seis mapas, sendo treinado sobre os vetores protótipos destes mapas. Detalhes sobre o treinamento das redes não foram reportados. O método utilizado na rotulação dos mapas foi denominado LabelSOM [Rauber 1999] e permite automaticamente descrever as categorias de documentos extraindo as características aprendidas pela rede SOM. Este método visa auxiliar o usuário a entender a coleção de documentos representada pelo mapa. Ele foi construído sobre a observação que os pesos presentes nos vetores protótipos da rede SOM treinada, servem como protótipos de um conjunto de sinais de entrada, isto é, exibem as características dos documentos mapeados num neurônio particular. Assim, foi assumido que aquelas características compartilhadas pela maioria dos documentos mapeados em um neurônio, servem como descritores para o respectivo neurônio. O método LabelSOM rotula cada neurônio com as características altamente similares para todos os padrões de entrada mapeados no mesmo, ou seja, para cada neurônio é realizada uma seleção de características tendo como base o vetor erro de quantização que armazena o erro de quantização para cada característica em um neurônio. O vetor erro de quantização é calculado através do somatório das distâncias euclidianas do vetor peso a cada um dos padrões de entrada mapeados em um neurônio particular. Características que exibem o menor erro de quantização são escolhidas como as mais prováveis candidatas para rotular o respectivo neurônio.

D. CDS Astronomy bibliographical Map

Este sistema começou a ser desenvolvido em 1996 por integrantes do observatório astronômico de Estrasburgo na França (*The centre de Donnés astronomique de Strasbourg - CDS*) [Lesteven et al. 1996] [Poinçot et al. 1998]. O objetivo era criar uma ferramenta baseada em redes SOM para organizar documentos texto da área de astronomia, a fim de facilitar a exploração e busca na coleção de documentos.

No trabalho [Poinçot et al. 2000] são especificados com detalhes a criação da interface de navegação do mapa de documentos, bem como é realizada uma análise do uso destes mapas na tarefa de recuperação de informação. A representação dos documentos utilizada consiste de vetores binários ponderados com *idf* (o inverso da

freqüência das palavras nos documentos da coleção) por permitir maior espalhamento dos documentos pelos neurônios no mapa que a representação binária.

Em [Lesteven et al. 2001], alguns avanços ao sistema são reportados, tais como a indexação completa dos documentos com uso de *stoplist*, uso do algoritmo de radicalização de Porter [Porter 1980], eliminação das palavras menos freqüentes e armazenamento da associação entre palavras (duas ou três); o uso de mapas secundários; e um processo de rotulagem automática. Foi utilizada uma hierarquia de mapas SOM para organizar os documentos. Inicialmente 9.450 artigos do periódico *Astronomy and Astrophysics* foram organizados. Os documentos foram indexados utilizando 269 palavras-chave que apareceram no mínimo em cinco diferentes artigos. O mapa principal tinha dimensões 15x15 e foi treinado por 50 épocas. O mapa secundário tem dimensões 5x5 e é criado caso um neurônio tenha muitos documentos nele mapeados, este mapa é treinado com os documentos associados com o neurônio e seus neurônios vizinhos no mapa principal. A interface do sistema foi gerada colorindo os mapas de acordo com a densidade de documentos em cada neurônio. Um neurônio pode ser selecionado clicando na figura e então os documentos nele mapeados são listados juntamente com a quantidade deles e as palavras que os descrevem, ou então o mapa secundário é mostrado. A rotulação automática consiste em obter a palavra mais freqüente em cada neurônio. Entretanto, somente neurônios de maior densidade são rotulados para evitar sobreposição de rótulos.

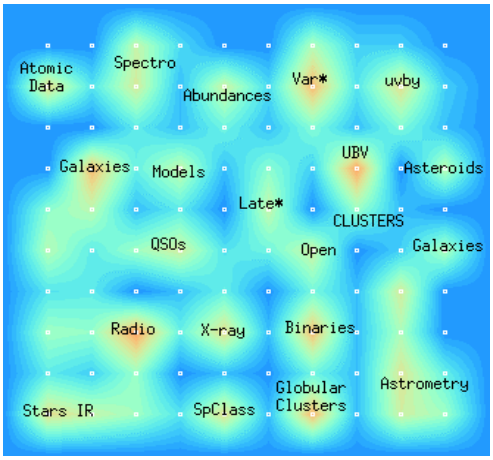

Figura 2.7 – Mapa de documentos extraído do VizieR[6].

[6] http://vizier.u-strasbg.fr/viz-bin/VizieR#Qkmap

Na Figura 2.7 é apresentada a interface de um mapa disponibilizado on-line para navegação através de catálogos. Cada ponto marca uma área no mapa; cores denotam a densidade ou a tendência de agrupamento dos documentos; áreas em escuro têm menor densidade. Ao clicar em uma área no mapa uma lista de catálogos de documentos que se encontram mapeados naquela área é demonstrado.

E. Outros trabalhos

Uma série de experimentos foram realizados em [Strehl et al. 2000] para medir o impacto das medidas de similaridade (euclidiana, co-seno, *jaccard* estendida e correlação de pearson) sobre a qualidade dos agrupamentos gerados em conjunção as técnicas *K-means* generalizado e particionamento de grafos com pesos. As redes SOM e técnicas de agrupamento por particionamento de hipergrafos foram também avaliadas. Foram usadas duas coleções de páginas Web retiradas da hierarquia de diretórios do Yahoo[7], uma sobre indústrias [Craven et al. 1998] e outra sobre notícias [Boley et al. 1999]. Os documentos (páginas Web) foram representados por vetores reais contendo a freqüência normalizada das palavras ou termos. A base de indústrias contém 966 documentos distribuídos em 10 categorias (setores industriais), a dimensionalidade dos vetores foi de 2896 palavras e, para todos os métodos, o número de grupos a serem gerados foi igual a 20. A base de notícias contém 2340 documentos classificados em 20 categorias, os documentos foram representados por 2903 termos distintos e foram agrupados em 40 grupos por todos os métodos. Cada grupo gerado foi rotulado com a categoria mais freqüente nos documentos nele contido. O desempenho das técnicas foi medida utilizando a métrica de informação mútua (do inglês *mutual information*) que mede o grau de dependência entre o agrupamento e a categorização. A rede SOM e os demais algoritmos que utilizaram a medida euclidiana, inclusive *K-means* , tiveram os piores resultados. O uso das métricas co-seno e *jaccard* estendida levou a resultados mais próximos (similares) ao do comportamento humano na categorização. O número de neurônios nas redes SOM foi igual ao número de grupos a serem gerados. Os mapas foram treinados por 5.000 ciclos ou 10 minutos (o que for atingido primeiro). A normalização utilizada e o raio inicial não foram especificados.

[7] http://www.yahoo.com

2.4.3 Diversificação (2001 – 2005)

Esta fase se caracteriza pela diversificação das linhas de pesquisa na construção de sistemas de organização automática de documentos usando a rede SOM. São testadas abordagens híbridas e a utilização de variantes de SOM na construção dos sistemas.

A maior dificuldade do uso de SOM é o longo tempo de treinamento requerido para construção do mapa para coleções onde o volume é grande e a dimensionalidade é alta. Em [Azcarraga & Yap 2001] foi proposta uma metodologia, chamada LiGHtSOM, que permite a redução drástica da dimensionalidade e volume de uma coleção e a construção de um sistema híbrido de arquivamento de documentos. O sistema híbrido utiliza seleção de características, projeção aleatória, um método de aprendizado incremental baseado em protótipos para realizar redução de volume e uma rede SOM. A idéia consiste em comprimir a base de dados numa taxa que permita rapidez no treinamento, rotulação e arquivamento, sem perder muito da informação original do conteúdo necessário para um arquivamento de qualidade. Entretanto, Azcarraga e Yap não reportaram uma análise do tempo total de treinamento do sistema híbrido e não realizaram uma avaliação cuidadosa da qualidade do mapa de documentos gerado pelo sistema. O sistema híbrido proposto por estes autores é um sistema semi-supervisionado por utilizar informação das categorias de cada documento para realizar a redução de volume.

GHSOM (do inglês *growing hierarchical SOM*) [Rauber et al. 2002] consiste de uma associação hierárquica de mapas, sendo estes mapas construídos por meio de uma variante da rede SOM que mantém uma grade regular dos neurônios enquanto cresce incrementalmente. Na construção dos mapas é utilizado um algoritmo similar ao GG (do inglês *growing grid*) [Fritzke 1995] para crescimento do mapa. A rede GHSOM foi aplicada a várias coleções de documentos, como documentos retornados de um engenho de busca [Rauber & Bina 2000], o *CIA world fact book* [Merkl & Rauber 2000] e artigos de notícias [Rauber et al. 2002]. O método LABELSOM [Rauber 1999] é usado para atribuir rótulos significativos a cada neurônio. Os mapas-filhos são iniciados usando os neurônios pais [Dittenbach et al. 2002]. A rede GHSOM é um método automático e eficiente ao gerar uma hierarquia de mapas. Entretanto, este método é sensível a parâmetros que determinam o crescimento dos mapas e a profundidade da hierarquia.

A Figura 2.8 mostra o exemplo de uma topologia construída através da rede GHSOM.

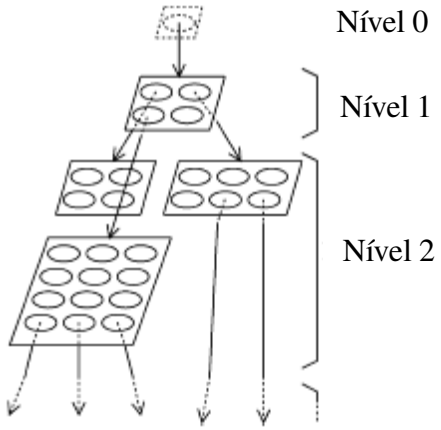

Figura 2.8 – Exemplo de topologia do GHSOM, adaptado de [Rauber et al. 2002].

A rede GHSOM é capaz de identificar o número necessário de neurônios em cada mapa durante o processo de aprendizado não supervisionado. A estrutura da hierarquia é determinada dinamicamente para retratar a estrutura dos dados de entrada. A construção da hierarquia se inicia com o cálculo do vetor protótipo de um neurônio (nível zero), o vetor protótipo será a média dos vetores documentos da coleção. No nível inferior (nível um) é realizado o treinamento de um pequeno mapa que será responsável por oferecer uma visão geral do conteúdo da coleção. Níveis subseqüentes de mapas são adicionados quando necessário de acordo com uma função do erro de quantização do mapa no nível atual e do nível anterior. Os mapas em níveis mais baixos apresentam maiores detalhes da região do espaço de entrada mapeado por um neurônio no nível superior, mostrando uma subdivisão mais fina dos tópicos presentes na coleção.

O método ATTS (do inglês *Adaptive Topological Tree Structure*) [Freeman & Yin 2004], também denominado TOC (do inglês *Topological Organization of Content*) [Freeman & Yin 2005a] ou *Treeview SOMs* [Freeman & Yin 2005b], utiliza um conjunto hierarquicamente organizado de redes SOM construtivas unidimensionais independentemente expandidas, chamadas individualmente de GC (do inglês *growing chain*). O processo da construção da árvore é realizado treinando cada cadeia individualmente a partir da cadeia raiz. Uma vez que cada cadeia foi treinada, cada neurônio é testado para determinar se os documentos mapeados no mesmo devem ser

agrupados em uma cadeia filha. O tamanho de cada cadeia é determinado independentemente através de um processo de validação usando a medida de validação baseada em entropia BIC (do inglês *Bayesian Information Criterion*). ATTS incorpora métodos para lidar com vetores esparsos no cálculo do produto interno rápido, na pesquisa pelo vencedor e atualização dos protótipos. A regra de atualização dos protótipos é a mesma do algoritmo de treinamento da rede SOM utilizando produto interno [Kohonen 2001]. A fim de reduzir o ruído dos vetores protótipos esparsos, os pesos menores que um limiar (0,00001) recebem o valor zero. As cadeias crescem até um número máximo de neurônios determinado a priori. A inserção de neurônios é realizada de forma adaptativa nas cadeias. Um método de interpolação ou extrapolação é usado para iniciar o vetor protótipo do neurônio a ser inserido próximo ao neurônio de maior atividade (maior número de vitórias desde o último reinício do treinamento). A inserção é realizada após convergência (estabilização da similaridade média). As cadeias-filhas têm vocabulário menor comparado com a respectiva cadeia-pai a fim de permitir agrupamentos mais especializados e a redução significativa da complexidade computacional. Os protótipos das cadeias-filhas são iniciados baseados nos protótipos dos neurônios na cadeia-pai para acelerar a convergência. Todos os neurônios na árvore são rotulados com termos representativos levando em conta a freqüência dos termos e pesos dos neurônios. Os documentos são indexados usando uma janela deslizante de tamanho máximo igual a quatro, sendo os termos formados por uma, duas, três e quatro palavras consecutivas na mesma sentença, enquanto palavras comuns em uma *stoplist* são ignoradas. Termos menos significativos são descartados, isto é, termos que ocorrem em poucos documentos e aqueles ocorrendo em mais que uma porcentagem do total de documentos. É usado o modelo espaço vetorial (do inglês *vector space*) para representar os documentos sendo os termos ponderados pelo esquema *tfidf* combinado com uma função que atribui mais peso a termos contendo múltiplas palavras. Os vetores protótipos são normalizados. Em [Freeman & Yin 2004] o método ATTS foi usado para organizar subcoleções da Reuters-21578 e comparado com SOM e *bisecting K-means* em termos de F-measure. ATTS obteve melhor desempenho, entretanto as diferenças ficaram abaixo de 0,1.

A Figura 2.9 mostra exemplos de uma topologia e um mapa de documentos obtido com o ATTS respectivamente.

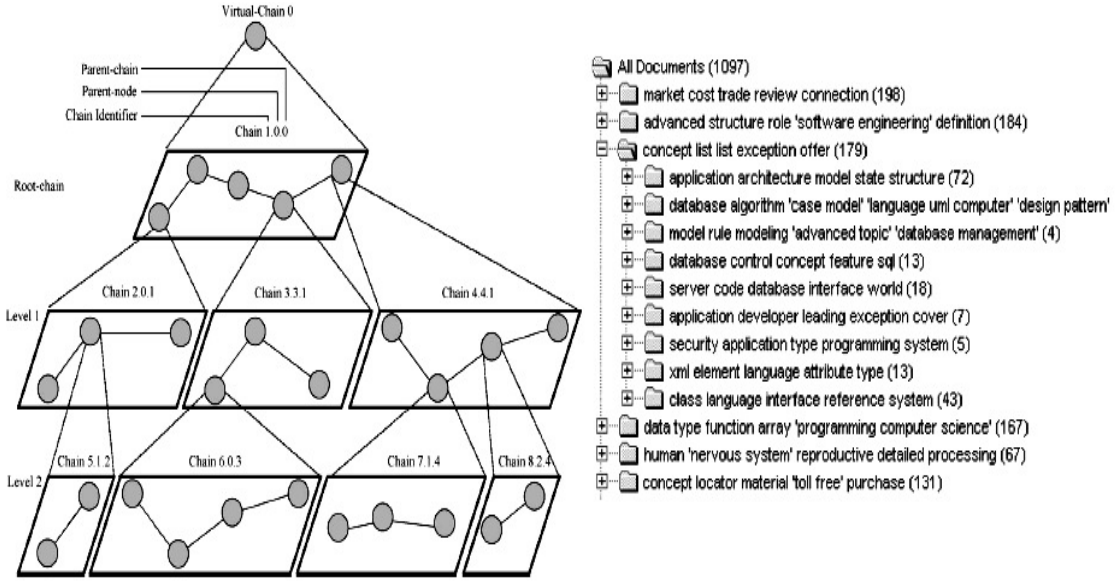

Figura 2.9 – Exemplos de topologia e visualização de um mapa de documentos, extraídos de [Freeman & Yin 2004].

2.4.4 Consolidação (2006 – 2008)

Esta fase está sendo marcada pela consolidação das diferentes linhas de pesquisa definidas nas fases anteriores e a busca pela comparação entre métodos a fim de obter sistemas que geram mapas de documentos de boa qualidade a um custo computacional mais baixo.

Em [Ontrup & Ritter 2006] foi proposto uma rede neural chamada SOM hiperbólico que cresce hierarquicamente (do inglês *Hierarchically Growing Hyperbolic Self-Organizing Map* - H2SOM). Esta rede é uma extensão da rede SOM hiperbólico (HSOM) visando resolver o problema de escalonamento das redes quando o número de neurônios é grande. HSOM e H2SOM são variantes de SOM que utilizam neurônios no espaço de saída hiperbólico ao invés do espaço euclidiano. O espaço hiperbólico facilita a visualização do mapa no espaço bidimensional. H2SOM permite organização hierárquica dos dados, crescimento adaptativo para uma granularidade requerida, bom escalonamento e suavização, bem como navegação sobre o mapa usando o Poincaré Disk. H2SOM e SOM foram aplicados sobre o subconjunto R20 da coleção Reuters-21578. Os documentos foram representados utilizando o modelo espaço vetorial com esquema de ponderação *tfidf*, sendo os vetores normalizados. Os vetores eram compostos por 5903 termos após pré-processamento, redução ao radical e remoção de palavras irrelevantes. O complemento de um da medida co-seno foi utilizado para medir

a dissimilaridade entre documentos. Entretanto, o algoritmo de treinamento reportado utiliza a distância euclidiana. Na comparação entre SOM e H2SOM, foram reportados tempo de treinamento, erro de quantização, três medidas do grau de preservação da topologia e F-measure. H2SOM mostrou melhor desempenho que SOM. Cada neurônio da H2SOM foi rotulado como os termos que tem valores mais altos no vetor protótipo dos neurônios (termos descritivos) e todo mapa foi visualizado por meio do Poincaré Disk. A Figura 2.10 mostra a topologia da H2SOM e um exemplo de mapa de documentos da coleção Reuters-21578.

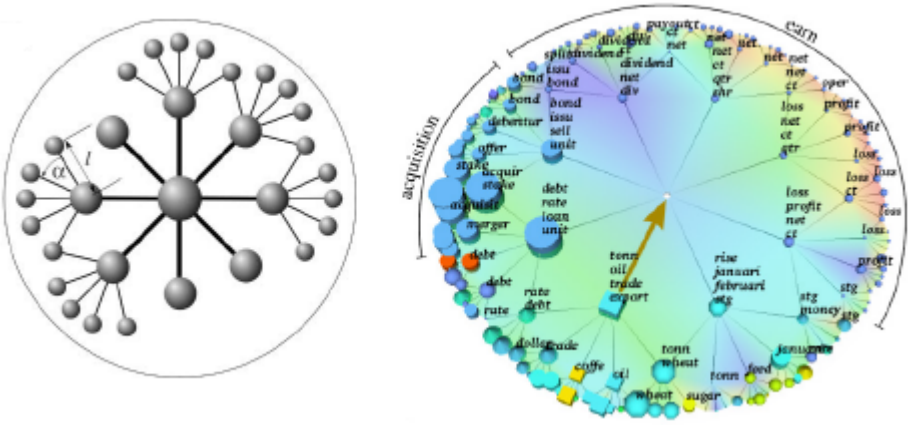

Figura 2.10 – Topologia da H2SOM e mapa de documentos extraídos de [Ontrup & Ritter 2006].

Usando o Poincaré Disk, regiões de interesse podem ser vistas trazendo-as ao foco por meio do ajuste da fóvea. O grande tempo de treinamento reportado para a rede SOM indica o uso de uma implementação padrão de SOM, sem uso de estruturas de dados para manipulação de vetores esparsos.

Estão inclusos nesta fase os trabalhos publicados pelo autor desta tese, sendo os mesmos apresentados e discutidos no Capítulo 4.

2.5. Conclusão

O primeiro trabalho visando a utilização das redes SOM em recuperação de informação foi [Lin et al. 1991]. Após este seguiram vários outros trabalhos e projetos foram criados visando utilizar SOM na organização de documentos. Dentre os projetos, os de maior impacto na área de organização automática de documentos baseado em redes SOM são WEBSOM [Kohonen et al. 2000], SOMLib [Rauber & Merkl 1998] e Arizona Digital Library [Roussinov & Chen 1998].

A ordem cronológica retrata aproximadamente a evolução das pesquisas na área. Inicialmente trabalhos utilizam coleções pequenas e tentam estabelecer ou definir uma avaliação sistemática dos resultados, passando por uma fase onde o objetivo era a construção de mapas de documentos de grandes coleções e outra fase marcada pela diversificação das linhas de pesquisa, até chegar à fase atual onde se observa uma tendência em focar mais na proposta e avaliação sistemática de sistemas, buscando a construção de mapas de documentos de melhor qualidade ou eficácia e construídos mais eficientemente.

A utilização de hibridismo na construção do sistema de recuperação de informação, ou mais precisamente, do sistema de organização automática de documentos, é uma possibilidade promissora que requer estudos mais intensos e experimentos mais elaborados.

Capítulo 3

Sistemas Híbridos Baseados em SOM para Organização Automática de Documentos

Atualmente, tem merecido crescente atenção, na área de Inteligência Artificial, o desenvolvimento de sistemas híbridos, que resultam da combinação de duas ou mais técnicas distintas para resolver um dado problema [Goonatilake & Khebbal 1995]. A motivação para tais sistemas está no fato de que as diversas técnicas existentes de Inteligência Artificial podem ser adequadas para determinados casos, mas podem apresentar deficiências para resolver outros tipos de problemas. Estas limitações estimulam o estudo dos sistemas híbridos, os quais procuram combinar as vantagens de duas ou mais técnicas, com o intuito de superar as desvantagens que cada uma apresenta individualmente na resolução do problema de interesse.

As principais contribuições a serem efetivadas nesta tese consistem na construção e avaliação de sistemas híbridos baseados em SOM para organização automática de documentos. Basicamente, os sistemas híbridos propostos combinam algoritmos de agrupamento tradicionais (K-means [McQueen 1967] e Leader [Hartigan 1975]) com redes SOM visando obter sistemas mais eficientes na organização automática de documentos, mantendo um nível aceitável de eficácia ou qualidade do mapa de documentos. K-means, Leader e SOM são considerados algoritmos distintos já que possuem natureza e propósitos diferentes (os dois primeiros foram projetados para agrupar dados e a rede SOM para a visualização dos dados).

Neste capítulo serão abordados os sistemas híbridos propostos, tais sistemas seguem a arquitetura especificada no Capítulo 2. O hibridismo ocorre nas etapas de redução de dimensionalidade e redução de volume. Algoritmos de agrupamento são introduzidos nestas etapas a fim de gerar sistemas híbridos de organização automática de documentos capazes de gerar mapas de documentos de boa qualidade a um custo computacional baixo.

Segundo a classificação dos sistemas híbridos proposta em [Goonatilake & Khebbal 1995], os sistemas propostos nesta tese são do tipo "híbridos intercomunicativos", pois os algoritmos de agrupamento são utilizados para superar a limitação de SOM quanto à visualização de grande volume de dados de alta dimensionalidade, resolvendo de forma independente as subtarefas de redução de dimensionalidade e redução de volume.

3.1. Algoritmos de Agrupamento

Nesta seção serão descritos os algoritmos de agrupamento utilizados na especificação dos sistemas híbridos de organização automática de documentos baseados em SOM.

K-means e Leader foram os algoritmos escolhidos por se tratar de algoritmos simples e bem conhecidos e por terem complexidade linear em relação ao tamanho do conjunto de treinamento.

3.1.1 Algoritmo K-means

O algoritmo K-means [McQueen 1967] é um algoritmo iterativo que agrupa vetores em k grupos a fim de minimizar uma função critério de dissimilaridade. Pode ser implementado para operar em lote (K-means em lote) ou iterativamente (K-means iterativo). O algoritmo K-means originalmente utiliza a distância euclidiana entre vetores, assim minimizando o critério do erro quadrático mínimo. Neste trabalho, uma variação de K-means utilizando co-seno como medida de similaridade entre vetores foi investigada, isto porque o co-seno demonstra capturar melhor a similaridade de conteúdo entre documentos representados por vetores que à distância euclidiana, fato relatado em [Strehl et al. 2000] e confirmado na literatura e nos experimentos apresentados no Capítulo 4. A variante do K-means usando co-seno minimiza a soma do complemento de um do co-seno.

No algoritmo K-means cada agrupamento é representado pelo seu centro, isto é a média aritmética dos padrões de entrada mapeados no mesmo. Os centros inicialmente recebem valores de k padrões de entrada aleatoriamente selecionados. Cada padrão de entrada é então mapeado para o agrupamento cujo centro está mais próximo ou é mais similar. A subseqüente recomputação da média para cada agrupamento e re-mapeamento dos padrões de entrada são realizados até convergir para um mapeamento fixo ou terminar o número máximo de iterações.

O algoritmo consiste dos seguintes passos:

(1) Escolha k centros de agrupamentos escolhidos aleatoriamente entre os padrões de entrada ou entre os pontos contidos no hipervolume contendo o cojunto de padrões.

(2) Atribua cada padrão ao centro de agrupamento mais próximo.

(3) Calcule o centro dos agrupamentos usando a atual pertinência aos agrupamentos.

(4) Se o critério de convergência não for alcançado, vá para o passo 2. Típicos critérios de convergência são: nenhuma (ou mínima) reatribuição de padrões a novos centros, ou mínimo decréscimo no erro quadrático.

A complexidade do algoritmo é *O(ndk)*, onde *n* é o número de padrões no conjunto de treinamento, *d* é a dimensão dos vetores de entrada e *k* é o número desejado de agrupamentos.

K-means é o algoritmo de agrupamento mais popular. As razões por trás desta popularidade são as seguintes: (i) é um algoritmo fácil de apresentar; (ii) tem complexidade linear ao tamanho do conjunto de treinamento; (iii) não é sensível a ordem em que os padrões de entrada são apresentados.

As maiores desvantagens deste algoritmo são a elevada sensibilidade a inicialização dos centros ou partição inicial, a possibilidade de convergência para um mínimo local da função critério se a partição inicial não é propriamente escolhida e no melhor caso a produção de somente agrupamentos de formato geométrico bem definido, por exemplo, hiperesféricos ao usar a distância euclidiana e hipercones ao usar co-seno.

3.1.2 Algoritmo Leader

O algoritmo Leader [Hartigan 1975] é um algoritmo de agrupamento incremental muito rápido, sendo o mais simples de todos [Jain et al. 1999]. Este

algoritmo ficou popular após sua implementação através do modelo de rede neural ART1 [Carpenter & Grossberg 1988]. Entretanto, este algoritmo é mais vantajoso que ART1 por trabalhar com vetores reais e por ser mais simples e eficiente.

Leader requer uma passada pelos dados para mapear cada padrão de entrada em um agrupamento. Associado a cada agrupamento existe o protótipo líder, que é um padrão contra o qual novos padrões serão comparados para determinar se o novo padrão pertence ou não ao agrupamento em questão.

Essencialmente, o Leader inicia com nenhum protótipo e adiciona um novo protótipo quando nenhum dos protótipos existentes é similar o suficiente ao padrão de entrada. O novo protótipo adicionado é uma cópia do padrão de entrada que é chamado líder do agrupamento. O co-seno do ângulo entre o padrão de entrada e cada protótipo é usado como medida de similaridade. Um limiar de influência, cujo valor se encontra no intervalo de zero a um, é um parâmetro do algoritmo que determina o quão similar o protótipo deve ser para ser considerado "próximo o suficiente". Nos casos onde existe um protótipo suficientemente próximo ao padrão de entrada corrente, este último é mapeado naquele agrupamento.

O algoritmo Leader consiste nos seguintes passos:

(1) Atribua o primeiro padrão para um agrupamento e o torne líder deste agrupamento.

(2) Considere o próximo padrão. Ou atribua o mesmo a um agrupamento existente ou atribua-o a um novo agrupamento e o torne líder do novo agrupamento criado. Essa atribuição é feita baseando-se no limiar de influência e a distância entre o padrão e cada um dos líderes de agrupamento, sendo o padrão comparado aos líderes na ordem de inclusão dos mesmos.

(3) Repita o passo 2 até que todos os padrões estejam agrupados.

O primeiro padrão apresentado será sempre o líder do primeiro agrupamento. Se o segundo padrão é próximo o suficiente ao líder (como determinado pelo limiar de influência), o segundo padrão é mapeado no primeiro agrupamento; caso contrário, o segundo padrão se tornará o líder do segundo agrupamento. O próximo padrão será comparado como o líder do primeiro agrupamento, se for próximo o suficiente será mapeado naquele agrupamento, se não será comparado ao líder do próximo agrupamento. O padrão será mapeado em um agrupamento ou, depois de ter sido

comparado com todos os líderes e não mapeado em nenhum, se tornará líder de um novo agrupamento. O próximo padrão segue o mesmo processo e assim se prossegue até que cada padrão esteja mapeando em algum agrupamento.

A desvantagem deste algoritmo é que ele é sensível a ordem de apresentação dos padrões. Por exemplo, o primeiro padrão apresentado sempre será líder de um agrupamento, e uma vez determinado o protótipo líder de um agrupamento, esse não sofre adaptações. Outra desvantagem é que grupamentos criados primeiro tendem a ser mais volumosos, pois cada padrão é mapeado no primeiro agrupamento em que o líder é próximo o suficiente.

O algoritmo Leader modificado [Hartigan 1975] suprime a desvantagem dos primeiros agrupamentos serem os mais volumosos ao comparar o padrão a todos os líderes e mapeando-o para aquele que tem a menor distância e é próximo o suficiente de acordo com o limiar. Se nenhum líder for próximo o suficiente, o padrão é adicionado como novo líder. Desta forma, cada agrupamento tem a mesma chance de ter novos padrões mapeados nele, não é dada vantagem aos agrupamentos criados primeiro. Mesmo com esta mudança este algoritmo é ainda sensível à ordem de apresentação dos padrões e é mais caro computacionalmente que o algoritmo Leader.

Em ambas as versões do algoritmo, o usuário deve fornecer o valor do limiar de influência e valores diferentes podem gerar resultados muito diferentes. Um valor muito pequeno pode resultar em cada padrão sendo mapeado em um agrupamento. Um valor muito grande pode resultar em um único agrupamento mapeando todos os padrões. Para encontrar agrupamentos naturais inerentes aos dados, o limiar deve ser maior que a distância típica dentro de cada agrupamento e menor que a distância típica entre agrupamentos.

Para um dado número máximo de agrupamentos desejados k, a complexidade dos algoritmos é limitada superiormente por $O(ndk)$, onde n é o número de padrões no conjunto de treinamento, d é a dimensionalidade dos vetores padrão de entrada. O tempo de treinamento de ambas as versões é menor que o do algoritmo K-means.

3.2. Arquitetura dos Sistemas Híbridos Propostos

Os sistemas híbridos propostos seguem a arquitetura descrita no Capítulo 2 (ver Figura 2.3). Tais sistemas são fruto da inserção de algoritmos de agrupamento nas

etapas de redução de dimensionalidade e redução de volume. Em outras palavras, a proposta desta tese pode ser sintetizada como a combinação de técnicas de redução de dimensionalidade e redução de volume utilizando algoritmos de agrupamento com as redes SOM, visando à construção de sistemas híbridos de organização automática de documentos.

Estes sistemas realizam todas ou algumas das tarefas de indexação, representação de documentos, redução de dimensionalidade, redução de volume e construção do mapa de documentos, sendo que pelo menos uma ou ambas as etapas de redução de dimensionalidade e redução de volume estão presentes. A seguir, descreve-se a configuração de cada etapa.

3.2.1 Indexação

Esta etapa é obrigatória num sistema de organização automática de documentos. Os documentos podem ser indexados usando palavras isoladas ou palavras compostas como termos. Esta etapa é a única que depende de customização para uma determinada língua e domínio, pois envolve o tratamento de palavras, números, caracteres especiais e sinais de pontuação, elaboração e uso de arquivo *stoplist* contendo palavras a serem ignoradas, redução das palavras ao radical usando um algoritmo de radicalização. As etapas posteriores independem da configuração interna desta etapa desde que sejam armazenadas em arquivo as informações da freqüência dos termos em cada documento da coleção.

3.2.2 Representação dos Documentos

Realizada com base nas informações capturadas pela etapa de indexação, a representação dos documentos é uma etapa obrigatória para um sistema de organização automática de documentos que use algoritmos de aprendizagem de máquina. Os documentos são representados por vetores-documentos utilizando para isto ou modelo espaço vetorial ou modelo booleano, podendo a importância do termo em cada documento ser computada como uma função qualquer desde que o valor cresça proporcional à importância do termo no documento.

3.2.3 Redução de Dimensionalidade por Mapeamento Semântico

A etapa de redução de dimensionalidade é o primeiro ponto de inserção de hibridismo no sistema de organização automática de documentos baseado em Redes SOM. É proposta a redução de dimensionalidade dos vetores documento por Mapeamento Semântico (MS) [Corrêa & Ludermir 2004a] [Corrêa & Ludermir 2004b] [Corrêa & Ludermir 2006a]. MS é um método baseado no agrupamento semântico de termos, onde cada grupo de termos corresponde a uma única dimensão num espaço reduzido. Teoricamente, este método é mais leve que o LSI e mais representativo da semântica dos documentos que o método RM (do inglês *Random Mapping*) [Ritter & Kohonen 1989] [Kohonen et al. 2000]. Inicialmente foi proposta a utilização de SOM para a tarefa de agrupamento de termos e depois o uso de K-means e Leader. Assim, MS é utilizado para reduzir a dimensionalidade dos vetores documentos gerando vetores no espaço reduzido que preservam o conteúdo semântico dos documentos.

O método MS foi elaborado como uma especialização do método RM, produzindo um subconjunto do conjunto de possíveis matrizes de projeção geradas por RM. Este método consiste dos passos listados na Figura 3.1.

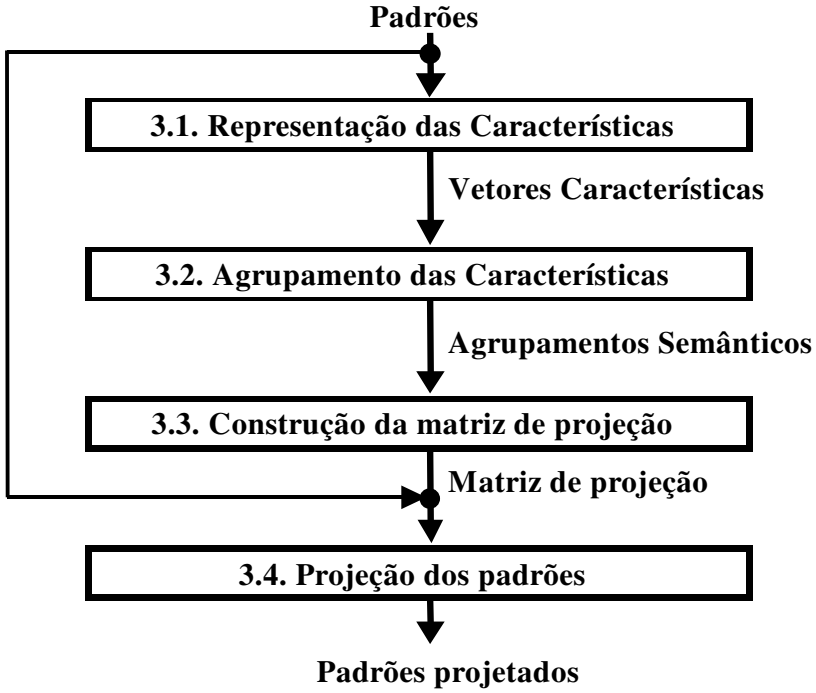

Figura 3.1 – Mapeamento Semântico.

O método MS consiste em construir uma matriz de projeção em que a semântica das características originais, capturada a partir dos dados, é utilizada na construção de características extraídas num espaço reduzido.

Inicialmente, dada uma matriz P de padrões por características, cada característica ou dimensão original deve ser representada por um vetor, de forma que a similaridade entre vetores aproxime a correlação ou proximidade semântica das características originais. Uma boa representação vetorial das características originais pode ser construída tomando os respectivos vetores colunas de uma sub-matriz de P com N padrões aleatoriamente ou heuristicamente selecionados. O subconjunto de padrões deve ser tal que a maioria dos vetores características tenham norma não-nula. No contexto de organização automática de documentos os padrões são vetores documentos e as características são os termos. Assim, cada termo é representado por um vetor contendo a freqüência do termo nos documentos do conjunto de treinamento. Deste modo, a semântica ou significado de um termo será deduzido analisando o contexto onde este é aplicado (o conjunto de documentos onde ocorre) e termos co-ocorrentes são geralmente semanticamente relacionados.

No segundo passo, agrupamentos semânticos são gerados através da utilização de um algoritmo de agrupamento treinado com os vetores características. Como vetores similares indicam características co-ocorrentes, agrupamentos de características co-ocorrentes são formados. Na organização automática de documentos, tais agrupamentos correspondem tipicamente a tópicos ou assuntos presentes nos documentos e provavelmente contém termos semanticamente relacionados. O conceito de agrupamento é então correlacionado com o conceito de característica extraída num espaço reduzido. O número de agrupamentos deve ser igual ao número de características extraídas desejadas. O algoritmo de agrupamento utilizado deve ter complexidade computacional linear para permitir que o método MS seja utilizado em grandes coleções. MS foi originalmente proposto em [Corrêa & Ludermir 2004a] e [Corrêa & Ludermir 2004b] usando a rede SOM como algoritmo de agrupamento, mas depois estendido para utilizar algoritmos de agrupamento mais rápidos como K-means e Leader [Corrêa & Ludermir 2007] [Corrêa & Ludermir 2008a].

A construção da matriz de projeção M, que é o terceiro passo, é realizada mapeando cada característica original t_j à k agrupamentos que melhor a representam (k

centros mais próximos do vetor característica) e depois construindo a matriz com a estrutura mostrada na Figura 3.2 da seguinte forma: Seja n o número de características originais e d o número de características extraídas, a matriz de projeção M deve ser construída com d linhas e n colunas, com m_{ij} igual a um se a característica original t_j foi mapeada no agrupamento semântico de índice c_i, e zero para o caso contrário. O número k de agrupamentos em que cada característica é mapeada é fixo e determina o número de uns em cada coluna de M.

A posição dos uns nas colunas de M indica em que característica extraída cada característica original irá participar. Assim, enquanto no método RM a posição dos uns em cada coluna da matriz de projeção é determinado aleatoriamente, no método MS esta é determinada de acordo com os agrupamentos semânticos onde cada característica original é mapeada.

$$\begin{array}{c} \quad t_1 \quad t_2 \quad \cdots \quad t_n \leftarrow \textbf{Termos} \\ \begin{array}{c} c_1 \\ c_2 \\ \vdots \\ c_d \end{array} \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1n} \\ m_{21} & m_{22} & \cdots & m_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ m_{d1} & m_{d2} & \cdots & m_{dn} \end{bmatrix} \\ \uparrow \\ \textbf{Agrupamentos Semânticos} \end{array}$$

Figura 3.2 – Estrutura da Matriz de Projeção.

O conjunto de matrizes de projeção geradas por MS é um subconjunto das matrizes geradas por RM, assim MS também preserva aproximadamente a similaridade mútua entre vetores de dados após a projeção para o espaço reduzido de características como provado em [Kaski 1997].

Finalmente, o mapeamento ou projeção dos vetores n-dimensionais para o espaço de dimensionalidade reduzida d é feito multiplicando a matriz M por cada vetor x_z, gerando o respectivo vetor reduzido y_z, segundo a Equação

$$y_z = M\, x_z, \forall\, z\,. \tag{10}$$

Depois da projeção, os vetores reduzidos podem ser opcionalmente normalizados para vetores unitários.

Os parâmetros do método MS são: um algoritmo de agrupamento (juntamente com seus respectivos parâmetros de treinamento); o número de dimensões reduzidas desejadas; e o número de uns em cada coluna da matriz de projeção.

A complexidade computacional do método MS é *O(ndN)* que é basicamente a complexidade do algoritmo de agrupamento em gerar *d* grupos (número de características extraídas) de *n* vetores características originais com *N* dimensões (número de documentos no conjunto de treinamento). Esta complexidade é menor do que a complexidade do método PCA (do inglês *Principal Component Analysis*) e ainda linear ao número de características no espaço original com o método RM [Corrêa & Ludermir 2006a].

As características extraídas por MS são, analítica e experimentalmente [Corrêa & Ludermir 2004a] [Corrêa & Ludermir 2007], mais representativas do conteúdo dos documentos, além de melhor interpretáveis que aquelas geradas por RM, permitindo a geração de mapas de documentos de melhor qualidade.

3.2.4 Redução de Volume por Algoritmos de Agrupamento

A redução de volume é um procedimento de quantização vetorial e consiste em representar o conjunto de padrões de treinamento por um conjunto menor de vetores representativos. Para tanto, os vetores documentos (originais ou reduzidos) são agrupados utilizando-se de algoritmos de agrupamento e o vetor representando cada agrupamento é tomado como protótipo, isto é, como vetor representativo dos vetores documentos mapeados no agrupamento. Os vetores protótipos são então utilizados para treinar o mapa SOM, sendo o mapeamento dos vetores documentos para os neurônios da rede realizado ao término do treinamento utilizando a mesma estratégia de mapeamento do algoritmo de agrupamento.

Os parâmetros da etapa de redução de volume consistem do algoritmo de agrupamento (juntamente com seus respectivos parâmetros de treinamento) e o número de protótipos a ser gerado.

Por razões de escalabilidade, o algoritmo de agrupamento usado na geração de protótipos deve ter complexidade computacional linear ao tamanho da coleção de documentos. Exemplos de algoritmos de agrupamento rápidos com esta característica são K-means e Leader, sendo estes os algoritmos propostos para realização desta tarefa.

3.2.5 Treinamento do Mapa de Documentos

A construção do mapa de documentos consiste em treinar a rede SOM com os vetores reduzidos obtidos da etapa de redução de dimensionalidade ou com os vetores protótipos obtidos da etapa de redução de volume, trata-se de uma etapa obrigatória para um sistema de organização automática de documentos baseados em SOM.

No caso de uso de redução de volume, um esquema de ponderação pode ser utilizado durante o treinamento. O esquema de ponderação consiste em atribuir um peso a cada protótipo igual ao número de vetores documentos que este representa.

O treinamento da rede SOM pode ser realizado em um estágio ou múltiplos estágios, bem como uma hierarquia de mapas pode ser construída ao invés de um único grande mapa.

3.2.6 Construção da Interface com o Usuário

A interface de um sistema de organização automática de documentos tem como objetivo fornecer ao usuário a capacidade de explorar ou navegar sobre o mapa de documentos, visualizando os documentos presentes em cada neurônio. Neurônios também devem ser rotulados com palavras-chave que descrevem o conteúdo dos documentos nele presentes. Além disso, podem também ser adicionadas ao sistema as funcionalidades de realizar buscas por palavras-chave e buscas baseadas em conteúdo sobre o mapa. Basicamente, a construção da interface exige a geração de páginas HTML descrevendo o conteúdo de cada neurônio (termos que funcionem como palavras-chave, número de documentos mapeados, link para documentos mapeados, etc.), interligadas a uma ou mais figuras ilustrativas dos mapas de documentos.

Embora seja imprescindível na construção de um sistema de organização automática de documentos funcional, será deixada de lado a implementação da interface de navegação sobre as coleções, já que a avaliação do sistema será feita após a etapa de construção do mapa de documentos e o maior objetivo deste trabalho é propor sistemas híbridos que permitam a construção de mapas de documentos de qualidade a um custo computacional mais baixo.

3.3. Sistemas Híbridos Propostos

Seguindo a arquitetura proposta na seção anterior, sistemas híbridos de organização automática de documentos baseados em SOM podem ser elaborados e avaliados.

O objetivo da proposição destes sistemas é avaliar qual a melhor arquitetura do sistema híbrido, ou seja, a melhor combinação do uso de Mapeamento Semântico para redução de dimensionalidade e algoritmos de agrupamento para redução de volume a fim de obter um sistema híbrido que permita a construção de mapas de documentos de qualidade com menor custo computacional possível.

A seguir, na especificação da arquitetura dos sistemas híbridos, os "*" codificam a posição onde será especificado o algoritmo de agrupamento que será utilizado no Mapeamento Semântico ("MS-*") e na redução de volume ("RV-*") respectivamente, o "+" indica que uma etapa sucede a outra, e que a etapa à esquerda fornece entrada para a etapa à direita.

Os primeiros sistemas híbridos a serem avaliados recebem o codinome SH1 e possuem arquitetura "MS-*+SOM". Tal arquitetura consiste em utilizar o Mapeamento Semântico para redução de dimensionalidade, não utilizar a redução de volume e treinar a rede SOM sobre os vetores documentos de dimensionalidade reduzida. O principal objetivo em avaliar este tipo de sistema é tornar claro o impacto da etapa de redução de dimensionalidade por Mapeamento Semântico no desempenho do sistema híbrido.

A complexidade do sistema SH1 é $O(nNd)+O(Ndm)$, onde N é o número de documentos no conjunto de treinamento, n é a dimensionalidade original dos vetores documentos, $\boldsymbol{d}$ é o número de dimensões dos vetores reduzidos, e m é o número de neurônios desejados no mapa SOM. Os dois termos na equação da complexidade computacional são respectivamente a complexidade da redução de dimensionalidade e treinamento do mapa SOM. O quanto este sistema híbrido é mais rápido que o sistema SOM tradicional , que tem complexidade $O(Nnm)$, é dado por

$$v = \frac{1}{d.\left(\frac{1}{m}+\frac{1}{n}\right)}. \tag{11}$$

Os próximos sistemas híbridos a serem avaliados recebem o codinome SH2 e possuem arquitetura "RV-*+SOM", consistindo em suprimir a redução de

dimensionalidade por Mapeamento Semântico, utilizando somente a redução de volume por algoritmos de agrupamento e treinamento da rede SOM sobre os vetores protótipos. O objetivo ao avaliar este tipo de sistema é mensurar o impacto da redução de volume no desempenho do sistema híbrido.

A complexidade do sistema SH2 é $O(Nnk)+O(knm)$, onde N é o número de documentos no conjunto de treinamento, n é a dimensionalidade original dos vetores documentos, k é o número de protótipos, e m é o número de neurônios desejados no mapa SOM. Os dois termos na equação da complexidade computacional são respectivamente a complexidade da redução de volume e treinamento do mapa SOM. O quanto este sistema híbrido é mais rápido que o sistema SOM tradicional, que tem complexidade $O(Nnm)$, é dado por

$$v = \frac{1}{k.\left(\frac{1}{m}+\frac{1}{N}\right)}. \tag{12}$$

Finalmente, os últimos sistemas híbridos a serem avaliados recebem o codinome SH3 e possuem a arquitetura "MS-*+RV-*+SOM", consistindo da combinação de redução de dimensionalidade por Mapeamento Semântico, redução de volume por algoritmos de agrupamento e treinamento da rede SOM sobre os vetores protótipos de dimensionalidade reduzida. Ao avaliar estes sistemas busca-se mensurar o impacto conjunto da redução de dimensionalidade por Mapeamento Semântico e redução de volume no desempenho do sistema híbrido.

A complexidade do sistema SH3 é $O(ndN)+O(Ndm)+O(m^2d)$, onde n é a dimensionalidade original dos vetores documentos, d é a dimensionalidade reduzida, N é o número de documentos no conjunto de treinamento e m é o número de neurônios desejados no mapa SOM. Os três termos na equação da complexidade computacional são, respectivamente, a complexidade do MS, redução de volume e treinamento do mapa SOM. O sistema híbrido é aproximadamente m/d ordens de magnitude mais rápido que o sistema SOM tradicional, que tem complexidade $O(Nnm)$. O quanto este sistema híbrido é mais rápido que o sistema SOM tradicional é dado por

$$v = \frac{n.m}{d.\left(n+k.\left(1+\frac{m}{N}\right)\right)}. \tag{13}$$

A análise da complexidade dos sistemas híbridos e cálculo do ganho de velocidade em relação ao sistema SOM tradicional permite a determinação de limites e valores para o número de protótipos e/ou número de dimensões reduzidas a fim de capacitar os sistemas híbridos a serem mais eficientes que o sistema baseado na rede SOM. Por exemplo, para o sistema SH2 o número de protótipos deve ser menor ou igual ao número de nodos no mapa SOM.

3.4. Conclusão

Neste capítulo especificou-se a arquitetura dos sistemas híbridos propostos. Basicamente os sistemas híbridos são gerados através do uso de algoritmos de agrupamento nas etapas de redução de dimensionalidade e redução de volume presentes na arquitetura dos sistemas de organização automática de documentos discutida no Capítulo 2.

Foi proposto o método Mapeamento Semântico para redução de dimensionalidade. Trata-se de um método capaz de gerar matrizes de projeção utilizando a informação de grupos semânticos de termos gerados por algoritmos de agrupamento dado um conjunto de vetores representativos da semântica dos termos.

A redução de volume é realizada tomando-se os centróides ou protótipos representativos dos grupos de vetores documentos gerados por algoritmos de agrupamento.

A metodologia e os resultados dos experimentos de avaliação dos sistemas propostos são apresentados no próximo Capítulo.

Capítulo 4

Experimentos e Resultados

Neste capítulo são apresentados os resultados obtidos na avaliação dos sistemas híbridos propostos nesta tese.

Inicialmente, as coleções utilizadas são descritas, tornando mais clara a composição e a escolha das mesmas. Em seguida, apresenta-se a metodologia usada nos experimentos e por fim, são apresentados e discutidos os resultados.

Parte dos resultados reportados neste capítulo foram apresentados em publicações anteriores e são suficientes para indicar que os sistemas híbridos propostos são relevantes para a área de organização automática de documentos, tendo também grande influência na implementação final dos sistemas híbridos.

No final do capítulo são apresentados experimentos que envolvem a comparação de todos os sistemas híbridos propostos na categorização de três coleções de documentos.

4.1. Problema e Bases de Dados

Uma maneira de medir o quanto um sistema de organização automática de documentos gera uma organização intuitiva de uma coleção de documentos para o ser humano é medir a capacidade deste sistema em separar as categorias de documentos manualmente rotulados por especialistas humanos. Levando em conta que dentro das categorias encontram-se documentos de conteúdo similares ou que tratam do mesmo tema, os agrupamentos formados pela rede SOM ou qualquer algoritmo de agrupamento devem maximizar a separabilidade dos documentos de categorias distantes semanticamente.

Assim, utilizou-se de medidas de eficácia da área de categorização de documentos para medir a qualidade dos mapas gerados nos experimentos.

O uso de medidas de categorização de documentos gera como pré-requisito aos experimentos o emprego de coleções de documentos categorizados manualmente.

Foram escolhidas para os experimentos três coleções, sendo a primeira considerada um *benchmark* da área de agrupamento de documentos e as duas últimas consideradas como *benchmarks* da área de categorização de documentos, a saber: K1, Reuters-21578 e 20 Newsgroups.

Todas estas coleções contêm da ordem de milhares de documentos, sendo consideradas de médio a grande porte. Estas coleções foram escolhidas por se tratarem de *benchmarks*, por possuírem natureza e características diversas, bem como pelo fato de que o processamento das mesmas não exige grande poder computacional. A seguir, é descrita cada uma destas coleções.

4.1.1 Coleção K1

Consiste de uma coleção de 2340 páginas da Web em língua inglesa classificadas em 20 categorias extraídas do diretório de notícias do Yahoo[8] envolvendo assuntos como saúde, negócios, esportes, política, tecnologia e entretenimento.

Esta base foi usada inicialmente em [Boley et al. 1999] para testar a escalabilidade de algoritmos de agrupamento. Posteriormente ela foi utilizada [Strehl et al. 2000], para medir o impacto de medidas de similaridade no agrupamento de páginas Web.

A coleção se encontra disponível na internet no endereço eletrônico: ftp://ftp.cs.umn.edu/dept/users/boley/PDDPdata/K1 .

As 20 categorias de notícias presentes na coleção são listadas na Tabela 4.1.

A coleção é distribuída já pré-processada (após as fases de indexação e representação dos documentos) como uma matriz esparsa de 21.839 termos por 2.340 documentos contendo a freqüência de ocorrência de cada termo nos documentos, bem como um arquivo contendo a listagem dos termos. Nesta matriz apenas 0,68% das entradas são não nulas [Boley et al. 1999]. Os termos correspondem a palavras reduzidas aos radicais utilizando o algoritmo de radicalização de Porter [Porter 1980].

[8] http://www.yahoo.com

Seguindo a metodologia reportada em [Strehl et al. 2000], os termos cuja freqüência média por documento foi menor que 0,01 ou maior que 0,10 foram eliminados da matriz por serem considerados insignificantes ou muito genéricos respectivamente; a dimensionalidade foi então reduzida para 2.903 termos.

A coleção foi particionada aleatoriamente dividindo os documentos pertencentes a cada categoria em metade para o conjunto de treinamento e a outra metade para o conjunto de teste. Cada conjunto ficou com 1.770 documentos. Observando a Tabela 4.1 vê-se que o número de documentos por categoria varia muito. As categorias majoritárias (categorias que possuem os maiores números de documentos e somadas correspondem a mais de 50% dos documentos da coleção) são Healt, Entertainment film, Entertainment people e Entertainment television. As categorias minoritárias (categorias que possuem os menores números de documentos e somadas correspondem a menos de 10% dos documentos da coleção) são Entertainment, Entertainment multimedia, Entertainment stage, Entertainment media e Entertainment art, Entertainment cable e Entertainment variety.

Tabela 4.1 – K1: Distribuição de documentos por categoria.

Categoria	Código	Conjunto de Treinamento	Conjunto de Teste	Total
Bussiness	B	71	71	142
Entertainment	E	4	5	9
Entertainment art	Ea	12	12	24
Entertainment cable	Ec	22	22	44
Entertainment culture	Ecu	37	37	74
Entertainment film	Ef	139	139	278
Entertainment industry	Ei	35	35	70
Entertainment media	Em	10	11	21
Entertainment multimedia	Emm	7	7	14
Entertainment music	Emu	62	63	125
Entertainment online	Eo	33	32	65
Entertainment people	Ep	124	124	248
Entertainment review	Er	79	79	158
Entertainment stage	Es	9	9	18
Entertainment television	Et	94	93	187
Entertainment variety	Ev	27	27	54
Healt	H	247	247	494
Politics	P	57	57	114
Sports	S	71	70	141
Technology	T	30	30	60
Total		1170	1170	2340

Esta coleção apresenta as seguintes características: (i) cada documento pertence a uma categoria, (ii) as categorias não são balanceadas, isto é, algumas contêm poucas unidades enquanto outra possui centenas de documentos classificados nas mesmas, (iii) existem categorias altamente correlacionadas (por exemplo, as categorias Entertainment) bem como categorias bem distintas (por exemplo, Sports e Politics).

4.1.2 Coleção Reuters-21578

Trata-se de um *benchmark* da área de categorização de documentos. Consiste de 21.578 notícias provenientes da agência Reuters newswire e divulgadas em 1987, sendo classificadas de acordo com 135 categorias temáticas tratando principalmente de temas envolvendo negócios e economia.

A coleção Reuters [Lewis 1997] foi originalmente criada por funcionários do grupo *Reuters Ltd*. e *Carniege Group Inc*, que classificaram manualmente os artigos organizando-os em diversas categorias predefinidas.

Por volta de 1990, a coleção foi disponibilizada à comunidade científica. Nesta época a coleção era chamada Reuters-22173, pois era constituída de 22.173 documentos. Com o passar dos anos a coleção passou a ser utilizada por diversos pesquisadores. Observando a possibilidade de poder comparar os resultados, a coleção tornou-se padrão entre os pesquisadores que poderiam então validar seus estudos e algoritmos.

Em 1996, Steve Finch e David D. Lewis realizaram uma série de alterações na coleção, refinando e reorganizando-a. Com isso, alguns documentos foram excluídos e anomalias corrigidas. A partir de então a coleção passou a chamar-se Reuters-21578, correspondendo à nova quantidade de documentos.

A coleção pode ser livremente distribuída e utilizada para fins de pesquisa e estudo, desde que a publicação que a utilize indique seu nome (“Reuters-21578, Distribution 1.0”) e o local onde pode ser encontrada (http://www.research.att.com/~lewis).

A coleção Reuters-21578 é distribuída em 22 arquivos de dados no formato SGML (do inglês *Standard Generalized Markup Language*), cada qual contendo 1000 artigos (com exceção o último que possui 578 artigos). Além disso, seis arquivos

descrevendo as categorias utilizadas para indexar os artigos também estão inclusos da coleção.

Os artigos se encontram enquadrados manualmente em 5 grupos distintos de categorias predefinidas: Topics, Exchange, Orgs, People e Places. As categorias no conjunto Topics são relacionadas com assuntos de interesse econômico, por exemplo: Gold (ouro), Coconut (coco), Money-Supply (empréstimos). Os conjuntos de categorias Exchanges, Orgs, People e Places correspondem aos diferentes tipos de entidades que são foco no artigo: instituições financeiras, organizações, pessoas e lugares, respectivamente. Pode-se citar como exemplo destes conjuntos as categorias Nasdaq (Exchanges), Gatt (Orgs), Perez-de-Cuellar (People) e Australia (Places).

A distribuição de categorias em cada um dos grupos, bem como uma idéia de como bem representadas estão estas categorias pode ser vistos na Tabela 4.2.

Tabela 4.2 – Reuters-21578: Distribuição de categorias por grupo de categorias.

Conjunto de Categorias	Número de Categorias	Número de Categorias com mais de uma ocorrência	Número de Categorias com mais de 20 ocorrências
EXCHANGES	39	32	7
ORGS	56	32	9
PEOPLE	267	114	15
PLACES	175	147	60
TOPICS	135	120	57

Dos cinco conjuntos de categorias existentes (Topics, Exchange, Orgs, People e Places), o mais utilizado, nos experimentos de categorização de documentos é o conjunto Topics. Este conjunto abrange assuntos de interesse econômico, num total de 135 assuntos ou categorias. As categorias deste conjunto são descritas nos arquivos ALL-TOPICS-STRINGS.LC.TXT e CAT-DESCRIPTIONS_120396.TXT que acompanham a coleção.

A coleção Reuters, por ter sido utilizada em vários experimentos, já possui partições elaboradas visando à definição dos conjuntos de treinamento e teste. As subdivisões mais conhecidas e trabalhadas da coleção *Reuters* são três: *Lewis Split*, *Apté Split* e *Hayes Split*. Os detalhes de como cada uma destas subdivisões foram obtidas são descritos no arquivo README.TXT que acompanha a coleção.

Nos experimentos apresentados neste trabalho, utilizou-se do subconjunto R90 [Debole & Sebastiani 2005] desta coleção com partionamento *ModApté* para definir os documentos nos conjuntos de treinamento e teste. Os pesquisadores da área de categorização de documentos têm adotado este subconjunto e este particionamento como padrão [Debole & Sebastiani 2005]. O subconjunto R90 contém somente documentos categorizados em no mínimo uma das categorias que contém ao menos um exemplo no conjunto de treinamento e um exemplo no conjunto de teste, que são ao todo 90 categorias.

O conjunto de treinamento possui 7770 documentos e o conjunto de teste possui 3019 documentos. A distribuição das categorias nos conjuntos de treinamento e teste é demonstrada na Tabela 4.3.

Tabela 4.3 – Reuters-21578: Distribuição de documentos por categoria.

Categoria	Conjunto de Treinamento	Conjunto de Teste	Total
acq	1650	719	2369
alum	35	23	58
barley	37	14	51
bop	75	30	105
carcass	50	18	68
castor-oil	1	1	2
cocoa	55	18	73
coconut	4	2	6
coconut-oil	4	3	7
coffee	111	28	139
copper	47	18	65
copra-cake	2	1	3
corn	181	56	237
cotton	39	20	59
cotton-oil	1	2	3
cpi	69	28	97
cpu	3	1	4
crude	389	189	578
dfl	2	1	3
dlr	131	44	175
dmk	10	4	14
earn	2877	1087	3964
fuel	13	10	23
gas	37	17	54
gnp	101	35	136
gold	94	30	124
grain	433	149	582

groundnut-oil	1	1	2
heat	14	5	19
hog	16	6	22
housing	16	4	20
income	9	7	16
instal-debt	5	1	6
interest	347	131	478
ipi	41	12	53
iron-steel	40	14	54
jet	4	1	5
jobs	46	21	67
l-cattle	6	2	8
lead	15	14	29
lei	12	3	15
lin-oil	1	1	2
livestock	75	24	99
lumber	10	6	16
meal-feed	30	19	49
money-fx	538	179	717
money-supply	140	34	174
naphtha	2	4	6
nat-gas	75	30	105
nickel	8	1	9
nkr	1	2	3
nzdlr	2	2	4
oat	8	6	14
oilseed	124	47	171
orange	16	11	27
palladium	2	1	3
palmkernel	2	1	3
palm-oil	30	10	40
pet-chem	20	12	32
platinum	5	7	12
potato	3	3	6
propane	3	3	6
rand	2	1	3
rape-oil	5	3	8
rapeseed	18	9	27
reserves	55	18	73
retail	23	2	25
rice	35	24	59
rubber	37	12	49
rye	1	1	2
ship	197	89	286
silver	21	8	29
sorghum	24	10	34
soybean	78	33	111
soy-meal	13	13	26

strategic-metal	16	11	27
sugar	126	36	162
sun-meal	1	1	2
sun-oil	5	2	7
sunseed	11	5	16
tea	9	4	13
tin	18	12	30
trade	369	117	486
veg-oil	87	37	124
wheat	212	71	283
wpi	19	10	29
yen	45	14	59
zinc	21	13	34
Total	7770	3019	10789

Esta coleção apresenta as seguintes características: (i) cada documento pode pertencer a nenhuma, uma, ou mais de uma categoria, (ii) algumas categorias tem poucos documentos classificados sobre a mesma, enquanto outras têm milhares, (iii) existem várias relações semânticas entre as categorias.

4.1.3 Coleção 20 Newsgroups

A coleção 20 Newsgroups é considerada como um *benchmark* na área de categorização de documentos e agrupamento de documentos. Consiste de aproximadamente 20.000 mensagens de e-mail capturadas de 20 categorias extraídas do Usenet newsgroups. As mensagens de e-mail lidam com os seguintes tópicos: computador, vendas, religião, política, ciência e recreação.

Nos experimentos foi utilizado o particionamento padrão "por data" (do inglês "By Date"), onde os documentos são ordenados por data e os primeiros 60% são utilizados para o conjunto treinamento e os 40% remanescentes para o conjunto de teste.

No pré-processamento desta coleção, alguns cuidados devem ser tomados com a não inclusão de duplicatas originadas de *cross-posts*, a não inclusão de cabeçalhos de identificação do newsgroup (como por exemplo, *Xref*, *Newsgroups*, *Path*, *Followup-To* e *Date*) e a remoção de *PGP keys*.

Foi utilizada uma distribuição desta coleção já pré-processada e disponibilizada em: http://web.ist.utl.pt/~acardoso/datasets/. Mais precisamente foi utilizado os arquivos 20ng-train-stemmed e 20ng-test-stemmed. Cada um destes arquivos contém o conteúdo dos documentos do conjunto de treinamento e do conjunto de teste respectivamente. Em

cada arquivo, cada documento ocupa uma linha, onde a primeira palavra denota a categoria do documento seguida de tabulação e da seqüência de termos que compõem o documento na ordem de apresentação dos mesmos separados por espaço.

O conjunto de treinamento possui 11.293 documentos e o conjunto de teste possui 7.528 documentos. A distribuição das categorias nos conjuntos de treinamento e teste é demonstrada na Tabela 4.4.

Tabela 4.4 – 20 Newsgroups: Distribuição de documentos por categoria.

Categoria	Conjunto de Treinamento	Conjunto de Teste	Total
alt.atheism	480	319	799
comp.graphics	584	389	973
comp.os.ms-windows.misc	572	394	966
comp.sys.ibm.pc.hardware	590	392	982
comp.sys.mac.hardware	578	385	963
comp.windows.x	593	392	985
misc.forsale	585	390	975
rec.autos	594	395	989
rec.motorcycles	598	398	996
rec.sport.baseball	597	397	994
rec.sport.hockey	600	399	999
sci.crypt	595	396	991
sci.electronics	591	393	984
sci.med	594	396	990
sci.space	593	394	987
soc.religion.christian	598	398	996
talk.politics.guns	545	364	909
talk.politics.mideast	564	376	940
talk.politics.misc	465	310	775
talk.religion.misc	377	251	628
Total	11293	7528	18821

Esta coleção apresenta as seguintes características: (i) cada documento pertence a uma categoria, (ii) as categorias são balanceadas, isto é, contém aproximadamente o mesmo número de documentos classificados nas mesmas, (iii) existem categorias altamente correlacionadas (por exemplo, comp.sys.ibm.pc.hardware e

comp.sys.mac.hardware) bem como categorias bem distintas (por exemplo, misc.forsale e soc.religion.christian).

4.2. Metodologia dos Experimentos

Encontra-se especificada nesta seção a metodologia comum a todos os experimentos no tocante à avaliação, implementação e metodologia de uso dos sistemas híbridos. Particularidades da metodologia adotada na avaliação dos sistemas híbridos são especificadas nas respectivas seções de resultado.

Todos os experimentos foram realizados utilizando o ambiente de programação MATLAB[9] e recursos do SOM Toolbox[10].

4.2.1 Avaliação dos Sistemas

Os sistemas híbridos serão avaliados em termos de eficiência e eficácia na categorização dos documentos [Sebastiani 2002].

A eficácia diz respeito à qualidade do mapa de documentos obtido. As seguintes medidas percentuais foram escolhidas para medir a eficácia dos sistemas híbridos: erro de classificação na categorização ou a métrica recíproca chamada acurácia; micro e macro F1 na categorização.Maiores detalhes sobre estas métricas podem ser vistos no Apêndice A. Estas medidas são típicas da área de categorização de documentos e são utilizadas aqui por serem uma maneira de avaliar sistematicamente se a forma com que os documentos são organizados no mapa assemelha-se à forma com que estes seriam organizados por seres humanos.

As medidas de eficácia serão avaliadas nos conjuntos de treinamento e teste. O uso do conjunto de teste tem como objetivo medir a generalização do mapa de documentos e manter a compatibilidade de comparação com trabalhos desenvolvidos na área de categorização de documentos, de onde foram extraídas as coleções a serem utilizadas nos experimentos.

A eficiência diz respeito ao tempo total gasto em segundos da etapa de redução de dimensionalidade até a finalização da etapa de construção do mapa de documentos. Para o sistema tradicional baseado na rede SOM (sistema SOM tradicional), o tempo

[9] http://www.mathworks.com

[10] http://www.cis.hut.fi/projects/somtoolbox/

total corresponde ao tempo de treinamento do mapa. Para os sistemas híbridos, o tempo total corresponde ao somatório do tempo gasto nas etapas de redução de dimensionalidade, redução de volume e construção do mapa de documentos.

A realização de testes de hipóteses para a comparação dos resultados obtidos pelos sistemas tornou-se necessário, já que apenas a comparação dos valores absolutos das médias gerais obtidas para cada métrica pode levar a conclusões erradas. Sendo assim, torna-se necessário verificar se estas médias são significativamente distintas do ponto de vista estatístico e a realização de testes de hipóteses permite fazer esta investigação, utiliza-se o t-teste [Wilcox 2001] para fazer análise dos resultados. Uma avaliação desta natureza é extremamente importante para verificar o nível de melhoria obtida com os sistemas propostos.

Os sistemas híbridos têm seu desempenho comparado com o sistema SOM tradicional sem redução de dimensionalidade ou volume, ou sistemas SOM utilizando métodos alternativos para a realização destas tarefas.

4.2.2 Implementação e metodologia de uso dos sistemas

São descritos a seguir aspectos de implementação e uso de cada etapa dos sistemas híbridos propostos.

A. Indexação e Representação dos documentos

Um programa escrito na linguagem JAVA[11] foi elaborado para fazer a indexação dos documentos. No tratamento do conteúdo dos documentos os caracteres especiais (tabulação, nova linha, retorno de carro, pontuação, etc.) e números foram substituídos por espaço, múltiplos espaços foram substituídos por espaço simples e todos os caracteres foram transformados em caracteres minúsculos. A indexação foi realizada tomando palavras isoladas como termos de indexação. Palavras com menos de 3 caracteres ou presentes numa lista de *stopwords* foram removidas e as restantes foram reduzidas à forma base utilizando o algoritmo de radicalização de Porter[12] [Porter 1980]. Documentos vazios foram eliminados.

[11] http://java.sun.com/

[12] http://www.tartarus.org/~martin/PorterStemmer/

Duas listas de palavras *stopwords* foram obtidas, uma elaborada pelo Laboratório de Recuperação de Informações da Universidade de Massachusetts em Amherst [Lewis 1992a] e obtida no trabalho [Lewis 1992b] contendo 292 palavras, e outra proveniente do projeto SMART[13] consistindo de 524 palavras.

Na indexação das coleções Reuters-21578 e 20 Newsgroups, uma lista padrão de *stopwords* foi usada para remover palavras irrelevantes e palavras remanescentes foram reduzidas ao radical usando o algoritmo de Porter [Porter 1980]. Os termos de indexação foram reduzidos eliminando termos genéricos e não informativos, sendo esses detectados por meio da freqüência (termos pouco freqüentes e termos muito freqüentes). A dimensão final dos vetores documento foi 5.180 termos e 8.165 termos para Reuters-21578 e 20 Newsgroups respectivamente.

A coleção K1 já se encontrava pré-processada, consistindo em uma matriz de 2340 documentos por 2903 termos contendo a freqüência de ocorrência de cada termo em cada documento.

Para as coleções K1, Reuters-21578 e 20 Newsgroups, os vetores documentos foram montados usando o modelo booleano e o modelo espaço vetorial com as representações *tf* e *tfidf*. A representação *tfidf* geralmente gera melhores resultados e, por isto, é a representação mais utilizada em trabalhos recentes da literatura. Os vetores documentos são geralmente normalizados.

B. Algoritmos de agrupamento

Os algoritmos K-means em lote e Leader modificado foram utilizados para agrupar vetores. Por simplicidade, estes algoritmos serão referenciados como K-means e Leader respectivamente.

O algoritmo K-means é Iniciado com vetores selecionados aleatoriamente. As condições de finalização do treinamento são: melhora no erro de quantização médio abaixo de um limiar (0,1%), nenhuma mudança no mapeamento dos vetores aos centros, e o número de épocas ser atingido (20 épocas).

A versão do algoritmo Leader implementado permite o estabelecimento de um número máximo de agrupamentos.

[13] ftp://ftp.cs.cornell.edu/pub/smart/english.stop

Os algoritmos de agrupamento foram utilizados nas etapas de redução de dimensionalidade e redução de volume.

C. Redução de Dimensionalidade

O método de Mapeamento Semântico (MS) foi implementado para receber como parâmetros o algoritmo de agrupamento, o número de dimensões reduzidas e o número de uns em cada coluna da matriz de projeção. Além dos parâmetros, o método MS recebe como entrada os vetores documentos do conjunto de treinamento.

Ao serem utilizados no método MS, os algoritmos de agrupamento geram tantos grupos quanto forem às dimensões reduzidas. As dimensionalidades reduzidas testadas foram: 100, 200, 300, 400 e 500.

D. Redução de Volume

Ao serem utilizados na redução de volume, os algoritmos de agrupamento geram tantos grupos quanto forem os de neurônios no mapa, isto é o número de protótipos gerados é igual ao número de neurônios desejados no mapa. Com esta escolha, se deseja avaliar o impacto da redução de volume nos sistemas SH2 e SH3 quando se busca maximizar a eficácia em detrimento da diminuição da eficiência dos mesmos.

Na especificação de alguns sistemas SH2, um "W." antecede o nome do algoritmo de agrupamento na descrição do sistema, isto indica o uso de ponderação dos protótipos durante o treinamento da rede SOM, no caso contrário o sistema não fez uso de ponderação dos protótipos. No uso de ponderação, o peso atribuído a cada protótipo é igual ao número de vetores documentos que cada um representa.

E. Construção do Mapa

O algoritmo usado para treinar os mapas foi o *batch* SOM por ser rápido e ter menos parâmetros ajustáveis.

Nos experimentos, foi utilizado o treinamento da rede SOM em um estágio por facilitar a implementação e análise dos resultados experimentais.

A topologia dos mapas de documentos bem como parâmetros do algoritmo de treinamento foram estabelecidos após experimentos pré-liminares.

A topologia foi definida como mapas de formato plano, vizinhança hexagonal, função de vizinhança como sendo a gaussiana truncada. A métrica utilizada nos

primeiros experimentos foi distância euclidiana e depois o co-seno se tornou padrão por levar a melhores resultados.

Os parâmetros de treinamento foram definidos como: 10 épocas na fase de ordenação e 20 épocas na fase de convergência, tamanho da vizinhança linearmente decrescente nas fases de ordenação (valor inicial igual à metade da maior dimensão do mapa mais um e valor final um) e convergência (valor inicial e final iguais a um). As condições de finalização do treinamento são: melhora no erro de quantização médio abaixo de um limiar (0,01%), nenhuma mudança no mapeamento dos vetores aos neurônios, o número de épocas ser atingido.

O número de neurônios no mapa SOM foi estabelecido de acordo com a heurística proposta no projeto WEBSOM [Kohonen et al. 2000], isto é, aproximadamente um décimo do número de documentos na coleção. Assim, para a coleção K1 o mapa tem dimensões 12x10 (120 neurônios) e para as coleções Reuters-21578 e 20 Newsgroups o mapa tem dimensões 30 x 30 (900 neurônios).

Os sistemas híbridos propostos são comparados com o sistema SOM tradicional, assim, para efeitos de comparação de desempenho, cinco mapas tiveram seus pesos iniciados com valores aleatórios retirados uniformemente do intervalo [0; 1] sendo utilizados para treinamento no sistema híbrido e no sistema SOM.

4.3. Resultados Publicados

Os resultados reportados nesta seção foram apresentados em publicações anteriores, sendo o motivo maior de apresentá-los o de respaldar a afirmação de que os sistemas híbridos propostos são relevantes para a área de organização automática de documentos, bem como fornecer argumentos que influenciaram no desenvolvimento dos sistemas híbridos propostos.

4.3.1 Experimentos com SH1 (Mapeamento Semântico + SOM)

Sistemas de arquitetura "MS-*+SOM" foram propostos em [Corrêa & Ludermir 2004a] e [Corrêa & Ludermir 2004b], sendo posteriormente avaliados e estendidos em [Corrêa & Ludermir 2006a] e [Corrêa & Ludermir 2007] respectivamente.

Em [Corrêa & Ludermir 2004a] e [Corrêa & Ludermir 2004b] o método Mapeamento Semântico (MS) foi proposto. O sistema híbrido "MS-SOM+SOM" foi

avaliado na categorização da coleção K1 utilizando a representação booleana na criação dos vetores documentos.

O artigo [Corrêa & Ludermir 2006a] foi uma extensão destes primeiros trabalhos, testando o sistema híbrido "MS-SOM+SOM" sobre a coleção K1 utilizando a representação *tfidf* na criação dos vetores documentos.

Nestes três trabalhos, MS foi comparado com SRM (do inglês *Sparse Random Mapping*) e PCA (do inglês *Principal Component Analysis*) em quatro dimensões de projeção: 100, 200, 300 e 400. Nestes trabalhos, o algoritmo de treinamento da rede SOM utilizava a métrica euclidiana e os vetores reduzidos foram normalizados para tamanho unitário. A Tabela 4.5 mostra os melhores resultados experimentais obtidos para cada método de redução de dimensionalidade nestes trabalhos.

Tabela 4.5 – Melhores resultados gerados por método de redução de dimensionalidade para a coleção K1.

Método	Representação	Dimensão	Treinamento Erro de classificação	Teste Erro de classificação
PCA	booleana	200	25,93 ± 0,82	37,91 ± 1,08
PCA	*tfidf*	100	28,19 ± 1,75	34,39 ± 1,79
MS-SOM	booleana	400	31,46 ± 0,96	38,46 ± 1,43
MS-SOM	*tfidf*	400	33,66 ± 1,18	41,15 ± 1,51
SRM	booleana	400	43,69 ± 1,72	51,72 ± 2,13
SRM	*tfidf*	400	52,75 ± 2,14	64,1 ± 3,15

Na Tabela 4.5 as médias e desvios padrões foram calculados sobre 150 execuções para os métodos MS e SRM (combinação de 30 matrizes de projeção geradas e treinamento de cinco mapas SOM Iniciados aleatoriamente) e cinco execuções para o método PCA (combinação de uma matriz de projeção gerada e treinamento de cinco mapas SOM iniciados aleatoriamente).

Em [Corrêa & Ludermir 2006a], são apresentadas as seguintes conclusões após análise dos resultados experimentais dos três trabalhos: (i) MS e SRM obtiveram melhores resultados ao utilizar 2 uns em cada coluna da matriz de projeção, independentemente do tipo de representação de documentos ser *tfidf* ou booleana; (ii) a projeção de vetores booleanos permitiu um melhor desempenho dos sistemas que a projeção dos vetores *tfidf*; (iii) A representação *tfidf* dos documentos a serem projetados faz com que os métodos MS e SRM fiquem menos sensíveis ao número de uns

utilizados na matriz de projeção que a representação booleana; (iv) nos experimentos realizados, o desempenho do sistema híbrido usando MS foi superior ao desempenho do sistema SOM tradicional com uso de SRM e próxima ao desempenho do sistema SOM tradicional com uso de PCA.

Em [Corrêa & Ludermir 2007] o método MS foi estendido para utilizar outros algoritmos de agrupamento além de SOM, sendo testados os algoritmos K-means e Leader. O desempenho de MS foi novamente comparado com o desempenho de SRM e PCA na categorização da coleção K1. Assim, os sistemas híbridos "MS-SOM+SOM", "MS-K-means+SOM" e "MS-Leader+SOM" foram avaliados na categorização da coleção K1. Foi utilizada a representação *tfidf* dos documentos e diferentemente dos trabalhos anteriores, a rede SOM utilizou a medida co-seno, o que possibilitou grande redução no erro de classificação.

As figuras a seguir ilustram graficamente os resultados do experimento apresentado em [Corrêa & Ludermir 2007]. Na Figura 4.1 é mostrado o desempenho para os métodos MS, PCA e SRM. Na Figura 4.2 é mostrado um zoom da parte inferior do gráfico da Figura 4.1, mostrando em mais detalhes o desempenho de MS e PCA.

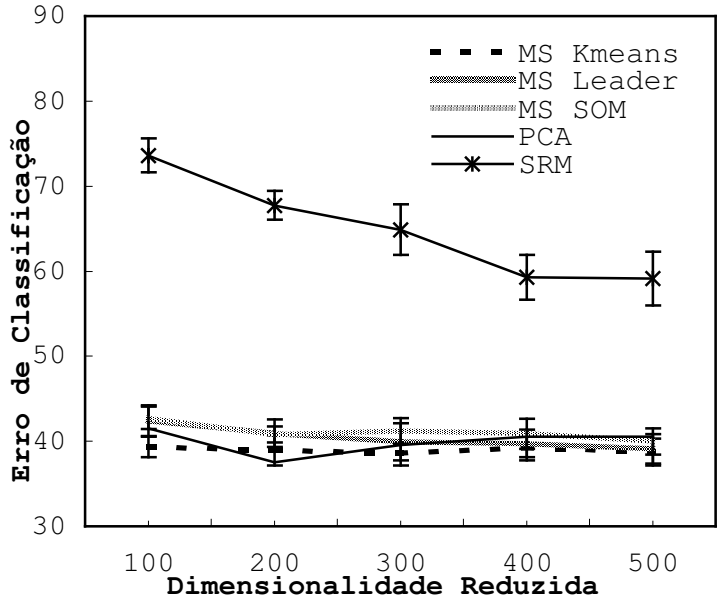

Figura 4.1 – Redução de dimensionalidade para a coleção K1.

Na Figura 4.1 observa-se que o erro de classificação obtido utilizando MS ficou bem próximo ao obtido utilizando PCA e foi muito menor ao erro obtido pelo método

SRM. O sistema híbrido "MS-K-means+SOM" teve melhor desempenho, seguido de "MS-Leader+SOM" e "MS-SOM+SOM".

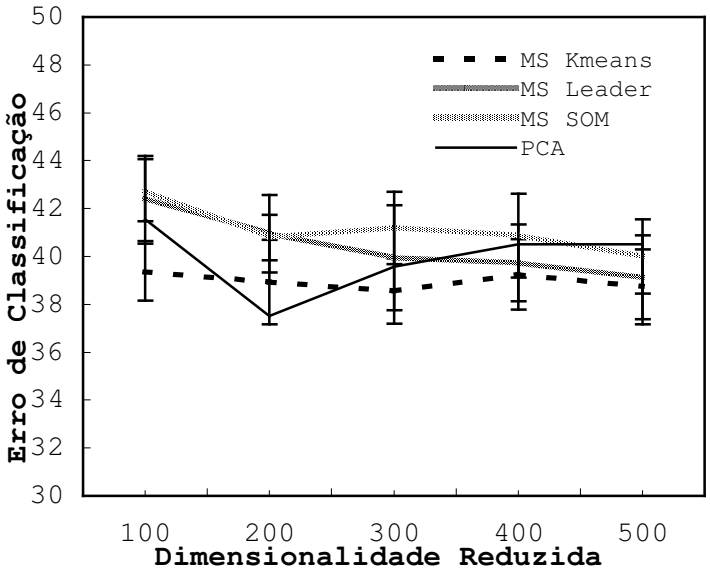

Figura 4.2 – Redução de dimensionalidade por MS e PCA para K1.

Na Figura 4.2 fica mais evidente o quanto o método MS tem desempenho próximo ao do método PCA, sendo que em algumas dimensões reduzidas chega a ter melhor desempenho.

A Tabela 4.6 mostra os resultados numéricos obtidos nos experimentos reportados neste trabalho e que deram origem aos gráficos apresentados. As médias e desvio padrões foram calculados sobre 15 execuções dos métodos MS e SRM (15 matrizes de projeção foram geradas por cada método). Um único mapa SOM Iniciado aleatoriamente foi utilizado em todos os sistemas.

Estes resultados mostram o quanto MS é um método eficiente e eficaz na redução de dimensionalidade de vetores documentos, tendo desempenho melhor ou próximo ao PCA a um custo computacional cerca de duas ordens de magnitude menor. Entretanto, resultados com outras coleções de documentos são necessários para se caracterizar melhor o desempenho de MS em relação ao método PCA.

Tabela 4.6 – Resultados gerados pelos métodos de redução de dimensionalidade para a coleção K1.

Método	Dimensão	Treinamento Erro de classificação	Teste Erro de classificação	Tempo Total
PCA	100	32,99 ± 0,00	41,54 ± 0,00	1006,00 ± 0,00
MS-K-means	100	32,08 ± 1,18	39,35 ± 1,19	10,66 ± 0,95
MS-Leader	100	35,58 ± 1,38	42,43 ± 1,78	8,21 ± 0,71
MS-SOM	100	35,10 ± 0,89	42,76 ± 1,30	15,33 ± 0,97
SRM	100	60,79 ± 1,25	73,62 ± 2,02	6,52 ± 0,57
PCA	200	33,25 ± 0,00	37,52 ± 0,00	1008,00 ± 0,00
MS-K-means	200	31,7 ± 1,31	38,93 ± 1,76	13,62 ± 0,92
MS-Leader	200	34,15 ± 1,26	40,95 ± 1,62	10,91 ± 0,98
MS-SOM	200	34,19 ± 1,01	40,79 ± 0,96	24,34 ± 1,44
SRM	200	56,77 ± 1,55	67,74 ± 1,70	8,28 ± 0,75
PCA	300	34,27 ± 0,00	39,57 ± 0,00	1009,00 ± 0,00
MS-K-means	300	31,98 ± 1,15	38,55 ± 1,37	16,42 ± 1,25
MS-Leader	300	33,58 ± 1,48	39,95 ± 2,20	12,49 ± 1,16
MS-SOM	300	34,61 ± 1,21	41,20 ± 1,51	31,73 ± 2,17
SRM	300	54,05 ± 2,09	64,88 ± 2,98	9,75 ± 0,78
PCA	400	33,50 ± 0,00	40,51 ± 0,00	1010,00 ± 0,00
MS-K-means	400	31,97 ± 1,11	39,25 ± 1,48	20,06 ± 1,41
MS-Leader	400	32,80 ± 1,54	39,74 ± 1,60	14,67 ± 1,00
MS-SOM	400	34,03 ± 1,13	40,88 ± 1,75	40,84 ± 2,39
SRM	400	50,56 ± 1,77	59,3 ± 2,63	10,89 ± 0,71
PCA	500	33,50 ± 0,00	40,51 ± 0,00	1012,00 ± 0,00
MS-K-means	500	31,73 ± 1,24	38,73 ± 1,56	22,31 ± 1,93
MS-Leader	500	32,16 ± 1,24	39,13 ± 1,74	16,88 ± 1,50
MS-SOM	500	33,64 ± 1,34	40,00 ± 1,55	50,97 ± 5,20
SRM	500	50,31 ± 2,57	59,13 ± 3,14	12,04 ± 0,85

4.3.2 Experimentos com SH2 (Redução de Volume + SOM)

Em [Corrêa & Ludermir 2006b] é proposta a redução de volume utilizando o algoritmo de agrupamento K-means. No papel de gerador de protótipos, K-means é comparado com a versão não-supervisionada do método proposto em [Azcarraga & Yap 2001] referenciado aqui como método AY. O método AY foi abandonado por ser um método computacionalmente mais caro que o K-means (apesar de ter a mesma complexidade computacional) e por ter desempenho inferior ao mesmo.

Neste trabalho, foram estabelecidos requisitos para que os sistemas híbridos que utilizam a redução de volume possam constituir alternativas ao sistema SOM tradicional: (i) o algoritmo de agrupamento deve possuir a complexidade computacional menor ou igual a K-means, bem como não ser computacionalmente mais caro que o K-means (como é o caso do método AY); (ii) o limite superior para o número de protótipos gerados é o número de neurônios desejados no mapa de documentos, isto para se garantir que o tempo de treinamento seja menor do que o sistema SOM tradicional.

A Tabela 4.7 exibe os resultados experimentais obtidos neste trabalho na categorização da coleção Reuters-21578. As médias e desvios padrões foram obtidos em 10 execuções da etapa de redução de volume. Os documentos foram representados usando o modelo espaço vetorial sem ponderação (esquema de representação *tf*). O algoritmo de treinamento da rede SOM usou a medida co-seno. O "W." na descrição do sistema significa o uso de ponderação dos protótipos igual ao número de vetores documentos que cada um representa no treinamento da rede SOM. O número de protótipos gerados foi igual ao número de neurônios desejados no mapa.

Tabela 4.7 – Desempenho dos sistemas apresentados em [Corrêa & Ludermir 2006b] para a coleção Reuters-21578.

Sistema	Acurácia	Micro F1	Macro F1	Tempo Total
SOM	0,7867 ± 0,0000	0,7260 ± 0,0000	0,1656 ± 0,0000	590 ± 00
RV-W.K-means + SOM	0,7703 ± 0,0085	0,7148 ± 0,0078	0,1699 ± 0,0114	500 ± 18
RV-K-means + SOM	0,7636 ± 0,0131	0,7114 ± 0,0070	0,1707 ± 0,0171	514 ± 27
RV-W.AY + SOM	0,7545 ± 0,0115	0,7065 ± 0,0072	0,1925 ± 0,0056	928 ± 39
RV-AY + SOM	0,7490 ± 0,0090	0,7025 ± 0,0056	0,1933 ± 0,0097	922 ± 34

Baseando-se na comparação dos resultados apresentados neste trabalho, utilizando t-teste, conclui-se que: (i) os sistemas híbridos geram mapas de documentos com desempenho próximo ao sistema SOM tradicional; (ii) K-means tem melhor desempenho na geração de protótipos que o método AY, é mais eficiente (mais rápido) e permite que o sistema híbrido seja mais eficaz em termos de acurácia e micro F1; (iii)

A ponderação dos protótipos tende a melhorar a representação das classes majoritárias da coleção em detrimento das minoritárias, embora o uso da ponderação não traga ganhos significativos no desempenho.

O trabalho [Corrêa & Ludermir 2008b], comparou sistemas híbridos utilizado os algoritmos K-means e Leader na redução de volume. O algoritmo AY não foi mais utilizado nos experimentos subseqüentes por ser mais caro computacionalmente que o K-means. Os sistemas híbridos foram comparados com o sistema SOM tradicional na categorização das coleções Reuters-21578 e 20 Newsgroups. Os documentos foram representados utilizando o esquema *tfidf* normalizado. O algoritmo de treinamento da rede SOM utilizou a medida co-seno.

A Tabela 4.8 e a Tabela 4.9 mostram respectivamente os resultados para as coleções Reuters-21578 e 20 Newsgroups, sendo as médias e desvios obtidos da execução da etapa de redução de volume 10 vezes. O número de protótipos gerado foi igual ao número de neurônios desejados no mapa.

Tabela 4.8 – Desempenho dos sistemas apresentados em [Corrêa & Ludermir 2008b] para a coleção Reuters-21578.

Sistema	Acurácia	Micro F1	Macro F1	Tempo Total
SOM	0,8278 ± 0,0000	0,7390 ± 0,0000	0,2158 ± 0,0000	141 ± 00
RV-W.K-means + SOM	0,8021 ± 0,0108	0,7163 ± 0,0097	0,2039 ± 0,0184	116 ± 11
RV-K-means + SOM	0,8057 ± 0,0107	0,7193 ± 0,0096	0,2106 ± 0,0155	127 ± 13
RV-W.Leader + SOM	0,7865 ± 0,0091	0,7022 ± 0,0081	0,1797 ± 0,0145	84 ± 10
RV-Leader + SOM	0,7851 ± 0,0103	0,7009 ± 0,0092	0,1772 ± 0,0132	82 ± 08

Analisando os resultados apresentados na Tabela 4.8, percebe-se que o uso de K-means na etapa de redução de volume permite ao sistema híbrido um melhor desempenho que o uso do algoritmo Leader. O desempenho do sistema híbrido utilizando K-means foi aproximadamente 2% inferior ao do sistema SOM tradicional. Para os sistemas híbridos, o tempo total de treinamento é significantemente menor que do sistema SOM tradicional. O sistema híbrido utilizando K-means é aproximadamente

10 a 18% mais rápido que o sistema SOM tradicional. O sistema híbrido utilizando Leader é aproximadamente 41 a 42% mais rápido que o sistema SOM tradicional.

Analisando os resultados apresentados na Tabela 4.9, percebe-se que o uso de K-means na etapa de redução de volume permite ao sistema híbrido um melhor desempenho que o uso do algoritmo Leader. O desempenho do sistema híbrido utilizando K-means foi aproximadamente 2% inferior ao sistema SOM tradicional. Para os sistemas híbridos, o tempo total de treinamento é significantemente menor que o do sistema SOM tradicional. O sistema híbrido utilizando K-means é aproximadamente 23 a 28% mais rápido que o sistema SOM tradicional. O sistema híbrido utilizando Leader é aproximadamente 51 a 53% mais rápido que o sistema SOM tradicional.

Tabela 4.9 – Desempenho dos sistemas apresentados em [Corrêa & Ludermir 2008b] para a coleção 20 Newsgroups.

Sistema	Acurácia	Micro F1	Macro F1	Tempo Total
SOM	0,6823 ± 0,0000	0,6823 ± 0,0000	0,6738 ± 0,0000	281 ± 00
RV-W.K-means + SOM	0,6657 ± 0,0088	0,6657 ± 0,0088	0,6584 ± 0,0094	202 ± 17
RV-K-means + SOM	0,6735 ± 0,0088	0,6735 ± 0,0088	0,6651 ± 0,0091	217 ± 31
RV-W.Leader + SOM	0,5809 ± 0,0184	0,5809 ± 0,0184	0,5714 ± 0,0183	131 ± 13
RV-Leader + SOM	0,5782 ± 0,0130	0,5782 ± 0,0130	0,5679 ± 0,0134	136 ± 06

Baseando-se nos resultados apresentados, as seguintes conclusões foram tomadas: (i) os sistemas híbridos "RV-Kmeans+SOM" e "RV-Leader+SOM" geraram mapas de documentos com desempenho próximo ao sistema SOM tradicional, (ii) K-means é um melhor gerador de protótipos que Leader quando eficácia é o principal objetivo, (iii) Leader é um melhor gerador de protótipos que K-means quando menor tempo de treinamento ou eficiência é preferido, (iv) esquema de ponderação de protótipos usado não melhora o desempenho nem eficiência dos sistemas híbridos.

4.3.3 Experimentos com SH3 (MS+RV+SOM)

Em [Corrêa & Ludermir 2008a] um sistema híbrido "MS-K-means+RV-K-means+SOM" é proposto e comparado com um sistema SOM tradicional na

categorização de documentos da coleção Reuters-21578. Os documentos foram representados utilizando o esquema *tf*. O algoritmo de treinamento da rede SOM utilizou a medida co-seno.

O sistema híbrido utilizou todo conjunto de treinamento na etapa de redução de dimensionalidade e o número de protótipos gerados no sistema híbrido por RV foi o número de neurônios no mapa, assim este sistema híbrido possui o maior custo computacional possível para um sistema deste tipo. O sistema híbrido foi testado nas dimensões reduzidas 100, 200, 300, 400 e 500, sendo realizadas 10 execuções de MS seguido de RV para cada dimensão.

A Tabela 4.10 exibe os resultados obtidos no experimento. O sistema híbrido teve eficácia inferior ao sistema SOM tradicional, porém as diferenças foram inferiores a 2%. No quesito eficiência, o sistema híbrido foi aproximadamente duas vezes mais rápido no treinamento que o sistema SOM tradicional para as dimensões reduzidas 100, 200 e 300.

Tabela 4.10 – Desempenho dos sistemas apresentados em [Corrêa & Ludermir 2008a] para a coleção Reuters-21578.

Sistema	Erro de classificação	Micro F1	Macro F1	Tempo Total
SOM	21,76 ± 0,00	0,7494 ± 0,0000	0,1716 ± 0,0000	249 ± 00
SH100	23,71 ± 0,84	0,7186 ± 0,0054	0,1259 ± 0,0060	95 ± 08
SH200	23,59 ± 0,93	0,7244 ± 0,0044	0,1337 ± 0,0107	113 ± 08
SH300	23,20 ± 0,86	0,7236 ± 0,0050	0,1336 ± 0,0064	126 ± 11
SH400	22,44 ± 0,93	0,7260 ± 0,0050	0,1372 ± 0,0077	138 ± 09
SH500	22,79 ± 0,93	0,7242 ± 0,0057	0,1367 ± 0,0083	150 ± 13

4.4. Resultados Recentes

Nesta seção são apresentados experimentos que envolvem a comparação dos sistemas híbridos propostos na categorização das três coleções de documentos.

Nestes experimentos, utilizou-se o número de dimensões reduzidas igual a 100 para todos os sistemas que realizam redução de dimensionalidade. Esta escolha favorece a eficiência dos sistemas híbridos em detrimento da eficácia. Apesar disto, alguns sistemas híbridos tiveram a eficiência ligeiramente pior que o sistema SOM tradicional para a coleção K1, isto se explica pelo fato de que o número de dimensões reduzidas e o número de protótipos estão próximos do limite máximo permitido para esta coleção.

As tabelas que se seguem mostram os resultados dos sistemas na categorização das coleções K1, Reuters-21578 e 20 Newsgroups respectivamente. Os sistemas estão ordenados em ordem decrescente de acurácia média. As médias e desvio padrões foram calculados sobre 10 execuções dos sistemas.

Na Tabela 4.11 pode-se observar que o sistema SOM tradicional foi o mais eficaz para a coleção K1, gerando melhor acurácia e micro e macro F1.

Tabela 4.11 – Desempenho dos sistemas para a coleção K1.

Sistema	Acurácia		MicroF1		Macro F1		Tempo total	
	Média	Desvio	Média	Desvio	Média	Desvio	Média	Desvio
SOM	0,6726	0,0000	0,6726	0,0000	0,4420	0,0000	4,23	0,00
MS-K-means + SOM	0,6578	0,0129	0,6578	0,0129	0,4243	0,0204	6,07	0,48
MS-K-means+ RV-K-means+ SOM	0,6293	0,0178	0,6293	0,0178	0,3949	0,0182	5,18	0,37
MS-Leader + SOM	0,6161	0,0218	0,6161	0,0218	0,3796	0,1135	3,87	0,70
RV-W.K-means + SOM	0,6063	0,0218	0,6063	0,0218	0,3779	0,0186	4,99	0,32
RV-K-means + SOM	0,5982	0,0167	0,5982	0,0167	0,3634	0,0315	5,13	0,56
MS-Leader+ RV-K-means+SOM	0,5901	0,0076	0,5901	0,0076	0,3610	0,0204	2,85	0,14
RV-W.Leader + SOM	0,5082	0,0235	0,5082	0,0235	0,3060	0,0230	3,34	0,19
RV-Leader + SOM	0,4886	0,0246	0,4886	0,0246	0,2847	0,0233	3,57	0,41
SRM+SOM	0,3010	0,0210	0,3010	0,0210	0,1687	0,0158	2,59	0,20
SRM+ RV-Kmeans+ SOM	0,2721	0,0154	0,2721	0,0154	0,1458	0,0147	1,60	0,10

A Tabela 4.11 mostra que os sistemas que realizaram MS foram os mais eficazes entre os sistemas híbridos propostos, sendo seguidos pelos sistemas que realizaram redução de volume usando o algoritmo K-means. O uso do algoritmo K-means proporcionou aos sistemas híbridos melhor eficácia que o algoritmo Leader ao ser utilizado na redução de dimensionalidade e/ou volume. Sistemas híbridos utilizando simultaneamente redução de dimensionalidade e volume tiveram o desempenho intermediário em relação aos sistemas utilizando somente uma das duas etapas. Dos seis sistemas híbridos com acurácia 10% inferiores à do sistema SOM tradicional, somente

dois tiveram tempo inferior ou igual a este sistema, fato já esperado devido à escolha do número de dimensões reduzidas e número de protótipos muito próximos aos limites máximos permitidos para esta coleção, levando o custo computacional dos sistemas híbridos serem muito próximos do sistema SOM tradicional. Percebe-se que os sistemas híbridos utilizando SRM tiveram a pior eficácia.

Na Tabela 4.12 pode-se observar que o sistema SOM tradicional foi o mais eficaz para a coleção Reuters-21578, gerando melhor acurácia e micro e macro F1. Sistemas que realizaram redução de volume foram os mais eficazes entre os sistemas híbridos propostos, sendo seguidos pelos sistemas que realizaram MS.

Tabela 4.12 – Desempenho dos sistemas para a coleção Reuters-21578.

Sistema	Acurácia		MicroF1		MacroF1		Tempo total	
	Média	Desvio	Média	Desvio	Média	Desvio	Média	Desvio
SOM	0,8288	0,0000	0,7399	0,0000	0,2159	0,0000	141,95	0,00
RV-K-means + SOM	0,8057	0,0107	0,7193	0,0096	0,2106	0,0155	127,10	12,96
RV-W.K-means + SOM	0,8021	0,0108	0,7161	0,0097	0,2039	0,0184	115,80	11,05
RV-W.Leader + SOM	0,7865	0,0091	0,7022	0,0081	0,1797	0,0145	85,50	13,60
RV-Leader + SOM	0,7851	0,0103	0,7009	0,0092	0,1772	0,0132	81,80	8,31
MS-K-means + SOM	0,7751	0,0081	0,6920	0,0072	0,1654	0,0138	98,85	4,80
MS-K-means+ RV-K-means+SOM	0,7636	0,0087	0,6817	0,0078	0,1523	0,0098	63,88	4,99
MS-Leader + SOM	0,7581	0,0102	0,6769	0,0091	0,1486	0,0290	105,86	15,17
MS-Leader+ RV-K-means +SOM	0,7469	0,0104	0,6669	0,0093	0,1406	0,0131	62,85	2,67
SRM+SOM	0,6202	0,0073	0,5537	0,0066	0,1097	0,0071	132,11	4,93
SRM+ RV-Kmeans+SOM	0,5891	0,0177	0,5260	0,0158	0,0950	0,0079	72,92	1,45

A Tabela 4.12 mostra que o uso do algoritmo K-means proporcionou aos sistemas híbridos melhor eficácia que o algoritmo Leader ao ser utilizado na redução de dimensionalidade e/ou volume. Sistemas híbridos utilizando simultaneamente redução de dimensionalidade e volume tiveram o desempenho inferior aos sistemas utilizando

somente uma das duas etapas. Ao contrário do ocorrido com a coleção K1, os sistemas híbridos foram mais eficientes que o sistema SOM tradicional. Percebe-se que os sistemas híbridos utilizando SRM tiveram a pior eficácia.

Na Tabela 4.13 pode-se observar que o sistema SOM tradicional foi o mais eficaz para a coleção 20 Newsgroups, gerando melhor acurácia e micro e macro F1. Sistemas que realizaram redução de volume foram os mais eficazes entre os sistemas híbridos propostos, sendo seguidos pelos sistemas que realizaram MS usando o algoritmo K-means.

Tabela 4.13 – Desempenho dos sistemas para a coleção 20 Newsgroups.

Sistema	Acurácia		MicroF1		MacroF1		Tempo total	
	Média	Desvio	Média	Desvio	Média	Desvio	Média	Desvio
SOM	0,6823	0,0000	0,6823	0,0000	0,6738	0,0000	323,50	0,00
RV-K-means + SOM	0,6735	0,0088	0,6735	0,0088	0,6651	0,0091	217,00	31,11
RV-W.K-means + SOM	0,6657	0,0088	0,6657	0,0088	0,6584	0,0094	202,10	16,50
MS-K-means + SOM	0,6349	0,0129	0,6349	0,0129	0,6238	0,0134	172,28	13,78
MS-K-means+ RV-K-means+SOM	0,6200	0,0107	0,6200	0,0107	0,6087	0,0117	96,84	3,43
RV-W.Leader + SOM	0,5809	0,0184	0,5809	0,0184	0,5714	0,0183	131,30	13,12
RV-Leader + SOM	0,5782	0,0130	0,5782	0,0130	0,5679	0,0134	136,40	6,31
MS-Leader + SOM	0,5501	0,0247	0,5501	0,0247	0,5393	0,1844	179,44	32,58
MS-Leader + RV-K-means +SOM	0,5432	0,0137	0,5432	0,0137	0,5324	0,0159	98,17	2,58
SRM+SOM	0,2014	0,0124	0,2014	0,0124	0,1973	0,0112	199,21	5,22
SRM+ RV-Kmeans+SOM	0,1652	0,0102	0,1652	0,0102	0,1594	0,0095	97,65	2,74

A Tabela 4.13 mostra que o uso do algoritmo K-means proporcionou aos sistemas híbridos melhor eficácia que o algoritmo Leader ao ser utilizado na redução de dimensionalidade e/ou volume. Sistemas híbridos utilizando simultaneamente redução de dimensionalidade e volume tiveram o desempenho inferior aos sistemas utilizando somente uma das duas etapas. Ao contrário do ocorrido com a coleção K1, os sistemas

híbridos foram mais eficientes que o sistema SOM tradicional. Percebe-se que os sistemas híbridos utilizando SRM tiveram a pior eficácia.

Sumarizando os resultados apresentados nesta seção, pode-se afirmar que os sistemas híbridos envolvendo MS e redução de volume por K-means oferecem um bom compromisso entre eficácia e eficiência.

Na Tabela 4.14 é mostrado o desempenho de alguns algoritmos supervisionados na categorização das coleções de documentos testadas. Pode-se observar que em alguns casos os sistemas híbridos e SOM tradicional, mesmo tendo natureza não supervisionada, tem diferença no desempenho inferior a 10% dos obtidos por algoritmos como MLP, SVM e KNN. Isto indica que os sistemas testados nesta tese são sistemas não supervisionados de grande importância para a área de organização automática de documentos.

Tabela 4.14 – Desempenho de algoritmos de classificação supervisionados.

Coleção	Algoritmo	Micro F1	Macro F1	Fonte
K1	MLP	0,7222	-	[Corrêa 2002]
Reuters-21578	SVM	0,8600	0,4000	[Debole & Sebastiani 2005]
Reuters-21578	KNN	0,7350	0,4900	[Debole & Sebastiani 2005]
20 Newsgroups	SVM	0,8278	-	[Cachopo & Oliveira 2007]
20 Newsgroups	KNN	0,7593	-	[Cachopo & Oliveira 2007]

4.5. Conclusão

Os resultados mostram que combinando a rede SOM com algoritmos de agrupamento é possível gerar sistemas híbridos bastante eficientes e eficazes para organização automática de documentos.

Algoritmos de agrupamento podem ser utilizados com sucesso nas fases de redução de dimensionalidade e redução de volume, gerando sistemas híbridos que geram mapas de qualidade a um custo computacional mais baixo que sistemas SOM tradicionais, desde que o número de dimensões reduzidas e o número de protótipos sejam estabelecidos levando em consideração a complexidade computacional dos sistemas.

Como será visto no Capítulo 5, os resultados obtidos originam perspectivas interessantes para trabalhos futuros.

Capítulo 5

Conclusões e Trabalhos Futuros

5.1. Contribuições

Considerando o que foi apresentado, a principal conclusão que pode ser extraída do presente trabalho é que os sistemas híbridos propostos no Capítulo 3 se mostram promissores e devem apresentar resultados satisfatórios quando forem aplicados para outras coleções, pois combinam as principais vantagens de algoritmos de agrupamento e SOM.

Algoritmos de agrupamento podem ser utilizados na redução de dimensionalidade dos vetores documentos, gerando vetores reduzidos que preservam as relações semânticas entre o conteúdo dos documentos. Com este objetivo, foi proposto a método de Mapeamento Semântico, que utiliza algoritmos de agrupamento para gerar grupos de termos semanticamente relacionados e utiliza os grupos formados na construção da matriz de projeção.

Algoritmos de agrupamento também podem ser utilizados na redução de volume de grandes coleções, gerando protótipos que vão representar grupos de vetores próximos no espaço de vetores documentos. A utilização dos protótipos, ao invés de todo o conjunto de vetores documentos, permite uma redução do tempo de treinamento da rede SOM bem como uma redução na quantidade de memória necessária para a realização deste treinamento.

Entretanto, na configuração das etapas de redução de dimensionalidade e redução de volume alguns cuidados devem ser tomados quanto à escolha do algoritmo de agrupamento bem como a escolha do número de agrupamentos a serem gerados para que o sistema híbrido não seja mais caro que o sistema SOM tradicional equivalente. O

estudo e especificação de como os sistemas híbridos devem ser compostos constituem contribuições deste trabalho.

A utilização da redução de dimensionalidade pelo método de Mapeamento Semântico e a redução de volume por algoritmos de agrupamento na composição de sistemas híbridos de organização automática de documentos baseados em SOM, permitem a estes sistemas uma eficácia próxima ao que seria obtido por um sistema SOM tradicional, com uma eficiência maior que este último sistema.

Como contribuições deste trabalho, destacamos também as publicações científicas, que permitiram uma divulgação dos resultados da pesquisa, bem como validação e refinamento das contribuições propostas, tendo respaldo da comunidade científica. Foram publicados os seguintes trabalhos: [Corrêa & Ludermir 2006a]; [Corrêa & Ludermir 2006b]; [Corrêa & Ludermir 2007]; [Corrêa & Ludermir 2008a]; e [Corrêa & Ludermir 2008b].

É importante ressaltar que os sistemas híbridos de organização automática de documentos propostos podem ser facilmente customizados para outras coleções de documentos contendo documentos escritos em outros idiomas e de outros domínios, bastando ajustar as configurações da etapa de indexação dos documentos.

5.2. Trabalhos Futuros

Este trabalho origina diversas perspectivas futuras, entre as quais se destaca o ajuste dos sistemas híbridos propostos, principalmente na especificação de métodos para construir a matriz de projeção e geração de protótipos que favoreçam a qualidade dos mapas de documentos gerados e/ou a eficiência dos sistemas híbridos. Esta linha de trabalhos futuros inclui: a determinação de limites inferiores para o número de documentos a ser utilizado no Mapeamento Semântico e do número de protótipos a ser gerado pela etapa da redução de volume, a fim de garantir uma boa qualidade mantendo o custo computacional baixo; considerar outros algoritmos de agrupamentos, bem como considerar outras formas de hibridismo de SOM com os algoritmos de agrupamento. Considerar o uso de estratégias de aceleração de algoritmos de agrupamento de dados como *single pass clustering* [Alex et al. 2007] e computação hierárquica em blocos semelhante à realizada por GHSOM [Rauber et al. 2002], para permitir a construção eficiente da matriz de projeção no Mapeamento Semântico e a geração de protótipos.

Quanto à construção do mapa de documentos, a fim de obter melhor acurácia dos mapas de documentos gerados, destacam-se as seguintes linhas de trabalhos futuros: o estudo de métodos para o ajuste automático dos parâmetros da rede SOM explorando a sintonia dos parâmetros à distribuição dos dados de treinamento; explorar o uso de métodos de remoção de *outliers* do conjunto de treinamento; o uso de variantes de SOM, como por exemplo, o uso de PLSOM (Parameterless SOM) que não necessita de ajuste de parâmetros, bem como outros algoritmos de preservação de topologia como o Neural Gas [Fritzke et al. 1995]; e propor e avaliar métodos para construção de hierarquias de mapas de documentos, bem como métodos para a avaliação da qualidade das hierarquias formadas.

Outros trabalhos futuros incluem: a especificação e avaliação de métodos para construção de interfaces de navegação sobre mapas de documentos; propor e avaliar métodos para realizar buscas por palavras-chave e baseadas em conteúdo; além de especificar métodos para construção de portais web que utilizam mapa de documentos para permitir ao usuário navegar e realizar buscas sobre coleções de documentos presentes numa biblioteca digital.

Quanto à aplicabilidade dos sistemas híbridos propostos, outros trabalhos futuros vislumbrados são: investigar a aplicabilidade do Mapeamento Semântico e a redução de volume por algoritmos de agrupamento no treinamento de outros algoritmos de aprendizado de máquina em tarefas de agrupamento e classificação de documentos; escrever artigos contendo resultados de experimentos que utilizem outras coleções de documentos de grande volume consideradas *benchmarks* da área de categorização de documentos. O objetivo de usar outras coleções de documentos é testar se os sistemas propostos também geram desempenhos satisfatórios em outros domínios e em coleções maiores. Inicialmente têm-se as coleções RCV1[14] e OHSUMED[15] como candidatas a serem utilizadas nos experimentos.

[14] http://www.daviddlewis.com/resources/testcollections/rcv1/

[15] http://trec.nist.gov/data/t9_filtering.html

Apêndice A

Introdução à Categorização de Documentos

O objetivo deste apêndice é caracterizar a tarefa de categorização de documentos, deixando claro sua definição, aplicação, tipos e medidas típicas de avaliação de desempenho.

A categorização de documentos é a classificação de documentos em um conjunto de uma ou mais categorias [Sebastiani 2002].

As categorias em que os documentos são classificados são predefinidas, tipicamente pelo projetista ou mantenedor do sistema de categorização. Os usuários finais geralmente não são envolvidos no processo de definir as categorias. Dependendo do conjunto de categorias, pode haver sobreposição entre diferentes categorias, ou seja, um documento pode pertencer a mais de uma categoria.

Durante o processo de categorização, os documentos são atribuídos a nenhuma, uma ou várias categorias. Os documentos processados são então armazenados no banco de dados junto como a lista de categorias atribuídas a cada um deles.

Uma vez que os documentos estejam categorizados, o usuário pode identificar um conjunto de categorias que podem conter documentos relevantes para suas necessidades e ignorar documentos em categorias que são provavelmente irrelevantes. Dessa maneira, o espaço de informação que o usuário tem para pesquisar é bastante reduzido, o que acelera o processo de encontrar informação relevante.

Portanto, uma típica aplicação de sistemas de categorização é limitar o espaço de busca para sistemas de recuperação de informação. Além de especificar uma consulta, o

usuário pode limitar o escopo da busca especificando um conjunto de categorias a serem pesquisadas.

Outras aplicações de categorização de documentos correspondem à filtragem de mensagens e notícias, categorização de resumos de publicações para sistemas específicos e sumarização de textos.

Os métodos automáticos de categorização de textos são procedimentos que envolvem a utilização de algoritmos de aprendizagem de máquina e que efetivamente classificam documentos com respeito a um conjunto de nenhuma, uma ou mais categorias existentes baseado no conteúdo dos mesmos.

Estes métodos podem realizar categorização:

1. Binária ou graduada – dependendo de como será expressa a relação de pertinência entre documentos e categorias;
2. Simples ou múltipla – dependendo da existência ou não de interseção entre categorias.

A categorização binária é mais comum em pesquisas sobre categorização. Na categorização binária cada documento é classificado como pertence (1) ou não (0) a cada uma das categorias. A categorização graduada ocorre quando cada documento recebe o grau de pertinência em relação a cada uma das categorias.

A categorização múltipla ocorre quando cada documento pode ser associado a mais de uma classe. Na categorização simples um documento só pode ser atribuído a uma categoria. Tendo sido associado um grau de pertinência a cada uma das classes (categorização graduada), pode-se obter a categorização múltipla atribuindo os documentos às classes em que a probabilidade associada ultrapassa algum limiar pré-estabelecido, ou a categorização simples atribuindo a cada documento a classe de maior probabilidade associada.

As principais medidas de avaliação do desempenho dos sistemas de categorização automática são a precisão (P) e a cobertura (C) [Baeza-Yates & Ribeiro-Neto 1999] calculadas para cada uma das categorias consideradas:

$$C = \frac{\text{número de documentos corretamente atribuídos}}{\text{número total de documentos da categoria}} \qquad \textbf{(14)}$$

$$P = \frac{\text{número de documentos corretamente atribuídos}}{\text{número total de documentos atribuídos}} \quad \textbf{(15)}$$

A cobertura refere-se à porção de documentos pertencentes à categoria *c* que foram atribuídos a ela. A cobertura é máxima quando todos os documentos pertencentes a *c* são atribuídos a ela pelo classificador. A precisão refere-se à porção dos documentos atribuídos a *c* pelo classificador que realmente pertencem a ela. A precisão é máxima quando são atribuídos a *c* somente documentos pertencentes a esta categoria. A Figura A.1 ilustra graficamente a obtenção destas medidas.

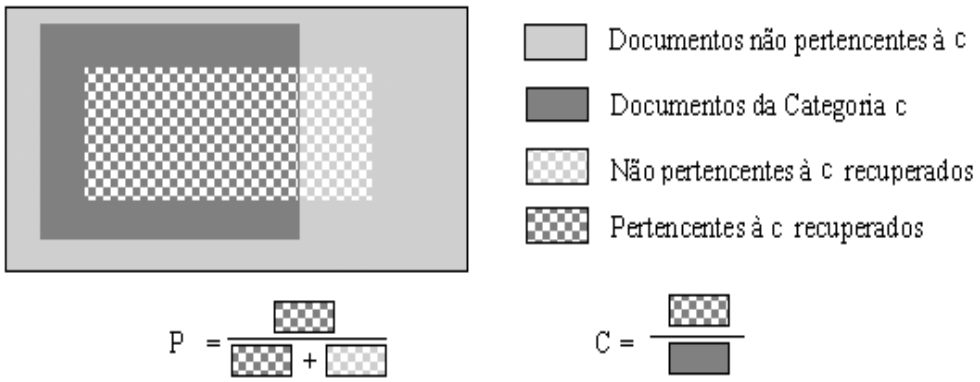

Figura A.1 - Medidas de eficácia para Sistemas de Categorização.

No contexto específico de métodos de avaliação de resultados de sistemas de categorização de textos, sugere-se na literatura a adoção de uma metodologia simples e direta baseada na construção de uma "tabela de contingência", ilustrada pela Tabela A.1.

Tabela A.1 - Tabela de Contingência

	Sim é correto	Não é correto	
Decide Sim	*a*	*b*	*a + b*
Decide Não	*c*	*d*	*c + d*
	a + c	*b + d*	*a + b + c + d = n*

A Tabela de contingência é inicialmente preenchida para cada categoria. A organização desta tabela parte do princípio de que um sistema de categorização realiza *n* decisões binárias, cada uma das quais tem exatamente uma resposta correta. Nesta tabela, a célula "*a*" representa a quantidade de documentos corretamente atribuídos a uma categoria em particular. A célula "*b*" representa a quantidade de documentos

incorretamente atribuídos a esta categoria. A célula "*c*" representa a quantidade de documentos incorretamente rejeitados para esta categoria, a célula "*d*" os corretamente rejeitados para ela e "*n*" representa o número total de documentos da coleção.

A partir desta tabela, podem ser extraídas importantes medidas de eficácia para sistemas de categorização de documentos como o grau de Falha (*fallout*), grau de acerto ou acurácia (do inglês *accuracy*) e Erro, além das já descritas precisão e cobertura. Estas medidas são expressas da seguinte forma:

$$\begin{aligned} \text{Cobertura} &= \frac{a}{a + c} \\ \text{Precisão} &= \frac{a}{a + b} \\ \text{Falha} &= \frac{b}{b + d} \\ \text{Acurácia} &= \frac{a+d}{n} \\ \text{Erro} &= \frac{b+c}{n} \end{aligned} \qquad \textbf{(16)}$$

Para o caso de sistemas de categorização com várias categorias, dois são os métodos de se avaliar o desempenho médio do sistema: por macro-média (*macroaveraging)* e por micro-média (*microaveraging)*. O método da macro-média consiste na avaliação de cada matriz de contingência de cada categoria separadamente seguida pelo cálculo de uma média global das medidas obtidas para cada categoria. O método da micro-média calcula o desempenho do sistema por meio da criação de uma matriz de contingência global cujas células sejam a soma das células correspondentes em cada matriz por categoria, calculando-se as medidas a partir dessa matriz global. A diferença entre estes dois métodos é que por micro-média são atribuídos pesos iguais para todos os documentos, enquanto que por macro-média são atribuídos pesos iguais para todas as categorias [Yang 1999]. O método mais utilizado nos trabalhos presentes na literatura é o calculo de medidas de avaliação por micro-média.

Como exemplo, a precisão e cobertura gerais por micro-média para um sistema de categorização que considera várias categorias simultaneamente sobre um conjunto de documentos, são assim calculadas: a precisão é o total de documentos corretamente atribuídos às categorias divido pelo total de documentos que foram atribuídos (correta ou incorretamente) às categorias; a cobertura é dado pelo total de documentos

corretamente atribuídos às categorias, dividido pelo total de documentos de fato pertencentes às categorias do conjunto. A precisão e cobertura gerais por macro-média consistem na respectiva média dos valores destas medidas obtidas para cada categoria.

O exame das medidas de precisão e cobertura separadamente pode levar a uma má avaliação do sistema, pois em geral, ao se aumentar a precisão de um sistema, diminui-se sua cobertura. Portanto, há a necessidade de se investigar outras formas de avaliar o sistema de modo a obter a configuração mais adequada.

As medidas do ponto de equilíbrio (do inglês *breakeven point*) e a medida do F-Measure combinam os valores de precisão e cobertura de modo a se obter o desempenho geral do sistema. O ponto de equilíbrio já foi bastante utilizado em sistemas de categorização: através do traçado dos vários pares de precisão e cobertura obtidos, pode-se obter por interpolação o ponto de equilíbrio, isto é, o ponto em que a precisão e a cobertura se igualam. A medida do F-Measure permite um balanceamento entre os valores de precisão e cobertura através da expressão

$$F = \frac{(\beta^2 + 1) \cdot P \cdot C}{\beta^2 \cdot (P + C)}, \qquad \textbf{(17)}$$

onde β é o parâmetro que permite a atribuição de diferentes pesos para as medidas de precisão (P) e cobertura (C), sendo 1 o valor geralmente adotado e neste caso se refere a esta métrica como F1. O valor de F é maximizado quando a precisão e a cobertura são iguais ou muito próximas, de modo que nesta situação, por definição, o valor do F-Measure é o próprio valor da precisão ou da cobertura, que por sua vez, é o ponto de equilíbrio do sistema.

A micro-média e macro-média de F1, ou simplesmente micro F1 e macro F1, são bastante utilizadas como medidas de desempenho médio de sistemas de categorização de documentos.

A acurácia e o erro de classificação são medidas típicas de avaliação de sistemas que realizam classificação de padrões. A acurácia corresponde ao número de documentos classificados corretamente dividido pelo número total de documentos classificados. O erro de classificação consiste do número de documentos classificados incorretamente dividido pelo número total de documentos classificados. No caso de categorização múltipla, se considera que um documento foi classificado incorretamente se ele não foi enquadrado em nenhuma das categorias a ele atribuídas.

Referências Bibliográficas

[Alex et al. 2007] Alex, N., Hammer, B. & Klawon, F. Single pass clustering for large data sets. *Proceedings of 6th International Workshop on Self-Organizing Maps* (*WSOM*), 2007, ISBN: 978-3-00-022473-7.

[Azcarraga & Yap 2001] Azcarraga, A. & Yap, T. SOM-Based Methodology for Building Large Text Archives. *In Proceedings of DASFAA01*, 2001, pp. 66-73.

[Baeza-Yates & Ribeiro-Neto 1999] Baeza-Yates, R. & Ribeiro-Neto, B. *Modern Information Retrieval*. Addison-Wesley. 1999.

[Boley et al. 1999] Boley, D., Gini, M., Gross, R., Han, E., Hastings, K., Karypis, G., Kumar, V., Mobasher, B., & Moore, J. 1999. Partitioning-based clustering for web document categorization. *Decision Support Systems*, v.27, 1999, pp. 329-341.

[Cachopo & Oliveira 2007] Cachopo, A. C. and Oliveira, A. L. Semi-supervised single-label text categorization using centroid-based classifiers, *Proceedings of 2007 ACM symposium on Applied computi*ng (*SAC' 2007*), ACM Press New York, NY, USA, 2007, pp. 844-851.

[Carpenter & Grossberg 1988] Carpenter, G. A. & Grossberg, S. The ART of adaptive pattern recognition by a self-organizing neural network. *IEEE Computer*, v. 21, 1988, pp. 77-88.

[Chen et al. 1996] H. Chen, C. Schuffels, & R. Orwig. Internet categorization and search: a machine learning approach. *Journal of Visual Communications and Image Representation*, v. 7, n. 1, 1996, pp. 88-102.

[Corrêa & Ludermir 2004a] Corrêa, R. F. & Ludermir, T. B. Dimensionality Reduction by Semantic Mapping. *Proceedings of VIII Brazilian Neural Networks Symposium (SBRN 2004)*, v. 1, 2004.

[Corrêa & Ludermir 2004b] Corrêa, R. F. & Ludermir, T. B. Dimensionality Reduction by Semantic Mapping in Text Categorization. *Proceedings of 11th International Conference on Neural Information Processing (ICONIP 2004), Lectures Notes in Computer Science*, v. 3316, Heidelberg (Alemanha): Springer Verlag, 2004, pp. 1032-1037.

[Corrêa & Ludermir 2004c] Corrêa, R. F. & Ludermir, T. B. Web Documents Categorization using Neural Networks. *Proceedings of 11th International Conference on Neural Information Processing (ICONIP 2004), Lectures Notes in Computer Science*, v. 3316, Heidelberg (Alemanha): Springer Verlag, 2004, pp. 758-763.

[Corrêa & Ludermir 2006a] Corrêa, R. F. & Ludermir, T. B. Improving Self-Organization of Document Collections by Semantic Mapping. *Neurocomputing (Amsterdam)*, v.70, 2006, pp. 62-69.

[Corrêa & Ludermir 2006b] Corrêa, R. F. & Ludermir, T. B. A Hybrid SOM-Based Document Organization System. *Proceedings of IX Brazilian Neural Networks Symposium (SBRN 2006)*, Los Alamitos: IEEE Computer Society, 2006, pp.16-23.

[Corrêa & Ludermir 2007] Corrêa, R. F. & Ludermir, T. B. Dimensionality Reduction of very large document collections by Semantic Mapping. *Proceedings of Workshop on Self-Organizing Maps (WSOM 2007)*, 2007, ISBN: 978-3-00-022473-7, Disponível em: <http://biecoll.ub.uni-bielefeld.de/volltexte/2007/133>. Acesso em: 26 mar. 2008.

[Corrêa & Ludermir 2008a] Corrêa, R. F. & Ludermir, T. B. Semantic Mapping and K-means applied to Hybrid SOM-Based Document Organization System Construction. *Proceedings of the 23rd Annual ACM Symposium on Applied Computing (SAC 2008)*, ACM, 2008, pp.1111-1115.

[Corrêa & Ludermir 2008b] Corrêa, R. F. & Ludermir, T. B. A Quickly Trainable Hybrid SOM-Based Document Organization System. *To appear in Neurocomputing, 2008*.

[Corrêa 2002] Corrêa, R. F. *Categorização de Documentos utilizando Redes Neurais: Análise comparativa com técnicas não-conexionistas*. Dissertação de Mestrado, Centro de Informática, Universidade Federal de Pernambuco, Recife-PE, 2002.

[Craven et al. 1998] Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., Slattery, S. Learning to extract symbolic knowledge from the world wide web. *In: AAAI-98*, 1998, pp. 509-516.

[Debole & Sebastiani 2005] Debole, F., Sebastiani, F. "An analysis of the relative hardness of Reuters-21578 subsets", *Journal of the American Society for Information Science and technology*, v. 56, n.6, 2005, pp. 584–596.

[Deerwester et al. 1990] S. Deerwester, S. T. Dumais, G. W. Furnas, & T. K. Landauer. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, v.41, 1990, pp. 391-407.

[Dittenbach et al. 2002] Dittenbach, M., Rauber, A., Merkl, D. Uncovering hierarchical structure in data using the growing hierarchical self-organizing map. *Neurocomputing*, v. 48, 2002, pp.199-216.

[Doley 1961] Doyle, L. B. Semantic road maps for Literature searchers. *Journal of ACM*, v.8, 1961, pp. 553-578.

[Freeman & Yin 2004] Freeman, R. T. & Yin, H. Adaptive topological tree structure for document organisation and visualization. *Neural Networks,* v. 17, 2004, pp. 1255-1271.

[Freeman & Yin 2005a] Freeman, R. T. & Yin, H. Web content management by self-organization. *IEEE Transactions on Neural Networks*, v. 16, 2005, pp.1256-1268.

[Freeman & Yin 2005b] Freeman, R. T. & Yin, H. Tree view self-organisation of web content. *Neurocomputing*, v. 63, 2005, pp. 415-446.

[Fritzke 1995] Fritzke, B. Growing grid – a self-organizing network with constant neighbourhood range and adaptation strength. *Neural Processing Letters,* v. 2, n. 5, 1995, pp.9-13.

[Fritzke et al. 1995] Fritzke, B. and Tesauro, G. and Touretzky, D.S. and Leen, T.K. A growing neural gas network learns topologies. *Advances in Neural Information Processing Systems*, MIT Press, 1995,pp.625-632.

[Goonatilake & Khebbal 1995] Goonatilake, S. & Khebbal, S. *Intelligent Hybrid Systems*, Wiley, London, 1995.

[Hartigan 1975] Hartigan, J. *A. Clustering Algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 1975.

[Haykin 1999] Haykin, S. *Neural Networks: a compreensive foundation*. 2nd ed. Prentice Hall, 1999.

[Honkela et al. 1996a] Honkela, T., Kaski, S., Lagus, K., Kohonen, T. Exploration of full-text databases with self-organizing maps. *Proceedings of IEEE International Conference on Artificial Neural Networks (ICNN'96)*, v. I. Piscataway, NJ: IEEE Service Center, 1996, pp. 56-61.

[Honkela et al. 1996b] Honkela, T., Kaski, S., Lagus, K., Kohonen, T. Newsgroup exploration with WEBSOM method and browsing interface. In: *Report A32, Helsinki University of Technology*, MD, Jan. 1996.

[Ichiki et al. 1991] H. Ichiki, M. Hagiwara, N. Nakagawa. Self-organizing multi-layer semantic maps. *Proceedings of International Conference on Neural Networks*, 1991, pp. 357-360.

[Jain et al. 1999] Jain, A. K., Murty, M. N., Flynn, P. J. “Data clustering: a review”, *ACM Computing Surveys*, v.31, n.3, 1999, pp. 264 - 323.

[Kaski 1997] Kaski, S. *Dimensionality reduction by random mapping: Fast similarity computation for clustering.* D. Sc. Thesis, Helsinki Univ. Technol., Finland, Mar. 1997.

[Kaski et al. 1998] Kaski, S., Honkela T., Lagus, K., Kohonen, T. *Websom – Self-organizing maps of document collections. Neurocomputing*, v. 21, 1998, pp. 101-117.

[Kaski et al. 1999] Kaski, S., Venna, J., Kohonen, T. Coloring that reveals high-dimensional structures in data. *Proceedings of 6th International Conference on Neural Inform. Process.*, 1999, pp. 794-734.

[Kohonen 1998] Kohonen, T. Self-organization of very large document collections: State of the art. *Proceedings of the 8th International Conference on Artificial Neural Networks (ICANN 98)*, v.1, 1998, pp. 65-74.

[Kohonen 2001] Kohonen, T. *Self-Organizing Maps*. 3rd extended edition. Berlim, Germany: Springer, 2001.

[Kohonen et al. 2000] Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., Saarela, A. Self Organization of a Massive Document Collection. *IEEE Transaction on Neural Networks*, v. 11, n. 3, May 2000, pp. 574-585.

[Koskenniemi 1983] Koskenniemi, K. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. Ph.D. thesis, University of Helsinki, Department of General Linguistics, 1983.

[Kraaijveld et al. 1995] Kraaijveld, M. A., Mao, J., Jain, A. K. A nonlinear projection method based on Kohonen's topology preserving maps. *IEEE Transactions on Neural Networks*, v. 6, May 1995, pp. 548-559.

[Lagus & Kaski 1999] Lagus, K., Kaski, S. *Keyword selection method for characterizing text document maps. In Proceedings of ICANN99, Ninth International Conference on Artificial Neural Networks*, v.1, 1999, pp. 371-376.

[Lagus 2002] Lagus, K. Text retrieval using self-organized document maps. *Neural Processing Letters*. v. 15, n. 1, February 2002, pp. 21-29.

[Lagus et al. 2004] Lagus, K., Kaski, S. & Kohonen, T. Mining massive document collections by the WEBSOM method. *Information Sciences*, v. 163, n. 1-3, 2004, pp. 135-156.

[Lesteven et al. 1996] Lesteven, S., Poinçot, P., Murtagh, Neural Networks and Information Extraction in Astronomical Information Retrieval, *Vistas in Astronomy,* v. 40, n. 3, 1996, pp.395-400.

[Lesteven et al. 2001] Lesteven, S., Poinçot, P., Murtagh, F. Visual Exploration of Astronomical Documents. *Astronomical Data Analysis Software and Systems X, ASP Conference Proceedings Series*, v. 238, 2001, pp. 78-81.

[Lewis 1992a] Lewis, D. D. *Representation and Learning in Information Retrieval*. Thesis of Doctor of Philosophy. Massachusetts: Department of Computer and Information Science, University of Massachusetts, 1992.

[Lewis 1992b] Lewis, D. D. Text Representation for Intelligent Text Retrieval: A Classification-Oriented View. In: Jacobs, Paul S., ed. *Text-Based Intelligent Systems*. Hillsdale, NJ: Lawrence Erlbaum, 1992, c. 9, pp.179-197.

[Lewis 1997] Lewis, D. D. *Reuters-21578 Text Categorization Test Colletion*. 1997. Disponível em: <http://www.daviddlewis.com/resources/testcollections/reuters21578/>. Acesso em: 26 mar. 2008.

[Lin et al. 1991] Lin, X., Soergel, D., Marchionini, G. A self-organizing semantic map for information retrieval. *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, 1991, pp. 262-269.

[McQueen 1967] McQueen, J. Some methods for classification and analysis of multivariate observations. *In Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967*, pp. 281-297.

[Merkl 1997] Merkl, D. Exploration of text collections with hierarchical feature maps, *Proc Int'l ACM SIGIR Conf R&D Information Retrieval*, July 27-31, 1997. Philadelphia, PA, 1997, pp. 186-195.

[Merkl 1998] Merkl, D. Text classification with self-organizing maps: Some lessons learned, *Neurocomputing*, v. 21, 1998, pp. 1-3.

[Merkl & Rauber 1998]. Merkl, D. & Rauber, A. CIA's view of the world and what neural networks learn from it: A comparison of geographical document space

representation metaphors. *Proceedings 9th International Conference on Database and Expert Systems Applications (DEXA98)*, 1998, pp.816-825.

[Merkl & Rauber 2000] Merkl, D. & Rauber, A. Document classification with unsupervised artificial neural networks. In F. Crestani, & G. Pasi (Eds.), *Soft computing in information retrieval*. Wurzburg, Wien: Physica-Verlag, 2000, pp.102-121.

[Merkl & Schweighofer 1997] Merkl, D. & Schweighofer, E. Exploration of Legal Text Corpora with Hierarchical Neural Networks: A Guided Tour in Public International Law, *Proceedings of the 6th Int'l Conference on Artificial Intelligence and Law (ICAIL'97)*, June 30 - July 3, 1997, Melbourne, Australia. Disponível em: <http://www.ifs.tuwien.ac.at/~dieter/LoP.html>. Acesso em 14 abr. 2008.

[Merkl & Tjoa 1994]. Merkl, D. & Tjoa, A. M. The representation of semantic similarity between documents by using maps: application of an artificial neural network to organize software libraries. *Proceedings of the General Assembly Conference and Congress of the International Federation for Information and Documentation*, 1994.

[Merkl 1995a] Merkl, D. A connectionist view on document classification. *Proceedings of the Australasian Database Conference (ADC'95)*, 1995, pp. 153-161.

[Merkl 1995b] Merkl, D. Content-based software classification by self-organization. *Proceedings of the IEEE International Conference on Neural Networks (ICNN'95)*, v.2, 1995, pp. 1086-1091.

[Merkl et al. 1994] Merkl, D., Schweighofer, E., W. Winiwarter. CONCAT: Connotation analysis of thesauri based on the interpretation of context meaning. *Proceedings of the International Conference on Database and Expert Systems Applications (DEXA'94)*, Sept 7-9, 1994. Athens, Greece, 1994, pp. 329-338.

[Miikkulainen 1990] Miikkulainen, R. Script recognition with hierarchical feature maps, *Connection Science*, v. 2, 1990, pp. 83-101.

[Ontrup & Ritter 2006] Ontrup, J. & Ritter, H. Large-scale data exploration with the hierarchically growing hyperbolic SOM. *Neural Networks, Advances in Self Organising Maps - WSOM'05,* v. 19, 2006, pp. 751-761.

[Orwig et al. 1997] Orwig, R., Chen, H., Nunamaker, J. F. A graphical, self-organizing approach to classifying electronic meeting output. *Journal of the American Society for Information Science*, v. 48, n. 2, February 1997, pp.157-170.

[Poinçot et al. 1998] Poinçot, P., Lesteven, S., Murtagh, F. A spatial user interface to the astronomical literature. *Astronomy and Astrophysics Supplement Series*, v. 130, 1998, pp. 183-191.

[Poinçot et al. 2000] Poinçot, P., Lesteven, S., Murtagh, F. Maps of Information Spaces: Assessments from Astronomy. *Journal of the American Society for Information Science,* v. 51, 2000, pp. 1081-1089.

[Porter 1980] Porter, M. An Algorithm for suffix stripping. *Program*. v.14, n.3, 1980, pp.130-137.

[Rauber et al. 2002] Rauber, A., Merkl, D. & Dittenbach, M. The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data. IEEE Transactions on Neural Networks, v.13, 2002, pp. 1331-1341.

[Rauber & Bina 2000] Rauber, A. & Bina, H."'Andreas, Rauber'? Conference pages are over there, German documents on the lower left …": an "old-fashioned" approach to Web search results visualization. *Proceedings of the 11th international workshop on database and expert systems applications*, 2000, pp.615-619.

[Rauber & Merkl 1998] Rauber, A. & Merkl, D. Creating an order in distributed digital libraries by integrating independent self-organizing maps. *Proceedings of International Conference on Artificial Neural Networks (ICANN'98)*, 1998.

[Rauber & Merkl 1999] Rauber, A. & Merkl, D. SOMLib: A digital library system based on neural networks. *Proceedings of fouth ACM International Conference on Digital Libraries*, 1999, pp.240-241.

[Rauber 1999] Rauber, A. LabelSOM: On the labeling of self-organizing maps. *Proceedings of International Joint Conference on Neural Networks (IJCNN'99)*, v. 2, 1999, pp.3524-3532.

[Ritter & Kohonen 1989] Ritter, H., & Kohonen, T. Self-organizing semantic maps, *Biological Cybernetics*, v. 61, 1989, pp.241-254.

[Roussinov & Chen 1998] Roussinov, Dmitri & Chen, Hsinchun. A Scalable Self-organizing Map Algorithm for Textual Classification: A Neural Network Approach to Thesaurus Generation. *Communication and Cognition in Artificial Intelligence Journal (CC-AI)*, v. 15, n. 1-2, 1998, pp. 81-111.

[Salton & McGill 1983] Salton, G. & McGill, M. J. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.

[Schatz & Chen 1996] Schatz, B. R. & Chen, H. Building large-scale digital libraries. *IEEE COMPUTER*, v. 29, n. 5, May 1996, pp. 22-27.

[Schatz et al. 1996] Schatz, B. R., Mischo, B., Cole, T., Hardin, J. A. Bishop, & H. Chen. Federating repositories of scientific literature. *IEEE COMPUTER*, v. 29, n. 5, May 1996, pp. 28-36.

[Schweighofer et al. 1995] Schweighofer, E., Winiwarter, W., & Merkl, D. Information filtering: The computation of similarities in large corpora of legal text. *Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL'95)*, 1995, pp.119-126.

[Sebastiani 2002] Sebastiani, F. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR),* v. 34, n. 1, March 2002, pp.1-47.

[Strehl et al. 2000] Strehl, A., Ghosh, J., Mooney, R. Impact of Similarity Measures on Web-page Clustering. *Proceedings of the 17th National Conference on Artificial Intelligence: Workshop of Artificial Intelligence for Web Search (AAAI 2000)*, 2000, pp. 58-64.

[Ultsch & Siemon 1990] Ultsch, A., Siemon, H. P. Kohonen's self organizing feature maps for exploratory data analysis. *In Proceedings of International Neural Network Conference, Dordrecht*, The Netherlands, 1990, pp. 305-308.

[Varfis 1993] Varfis, A. On the use of two traditional statistical techniques to improve the readability of Kohonen Maps. *Proceedings of NATO ASI Workshop Statistics Neural Networks*, 1993.

[Wilcox 2001] Wilcox, R. R. *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy*, Springer-Verlag, New York, 2001.

[Yang & Pedersen 1997] Y. Yang & J. P. Pedersen. A comparative study on feature selection in text categorization. *Proceedings of Fourteenth International Conference on Machine Learning (ICML'97)*, 1997, pp.412-420.

[Yang 1999] Yang, Y. An Evaluation of Statistical Approaches to Text Categorization. *Journal of Information Retrieval*, v. 1, n. 1-2, 1999, pp. 67-88.

Tese de Doutorado apresentada por **Renato Fernandes Corrêa** a Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, sob o título **"Sistemas Baseados em Mapas Auto-Organizáveis para Organização Automática de Documentos Texto"** orientada pela Profa. Teresa Bernarda Ludermir e aprovada pela Banca Examinadora formada pelos professores:

Prof. Francisco de Assis Tenório de Carvalho
Centro de Informática / UFPE

Prof. George Darmiton da Cunha Cavalcante
Centro de Informática / UFPE

Prof. Ricardo Bastos Cavalcante Prudêncio
Departamento de Ciência da Informação / UFPE

Prof. Guilherme de Alencar Barreto
Departamento de Engenharia de Teleinformática / UFC

Prof. Adrião Duarte Dória Neto
Departamento de Engenharia Elétrica / UFRN

Visto e permitida a impressão.
Recife, 7 de julho de 2008.

Prof. FRANCISCO DE ASSIS TENÓRIO DE CARVALHO
Coordenador da Pós-Graduação em Ciência da Computação do
Centro de Informática da Universidade Federal de Pernambuco.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo