Modelos Computacionais Adaptativos para Recuperação de Informação

Adaptative Computing Models for Information Retrieval

Márcio Henrique Zuchini[1] , USF, ESAMC

Resumo

A vasta quantidade de dados disponíveis na Internet requer ferramentas adequadas para efetuar a mineração de dados de forma automática, principalmente quando se considera que grande parte do conhecimento humano é expresso na forma textual. Este trabalho mostra a aplicação de dois modelos computacionais adaptativos, o Mapa Auto-Organizável e o Mapeamento Topográfico Gerativo na tarefa de recuperação de informação. Também inclui a realização de um experimento computacional, envolvendo dois conjuntos distintos de arquivos em língua portuguesa, que mostra a efetividade destas ferramentas.

Palavras-chave. mapa auto-organizável (SOM); mapeamento topográfico gerativo (GTM); recuperação de informação; descoberta de conhecimento.

Abstract

The vast amount of data available on the Internet requires suitable tools to perform data mining automatically, especially when considering that much of human knowledge is expressed in textual form. This work shows the application of two adaptive computational models, the Self-Organizing Map and the Generative Topographic Mapping, in the task of information retrieval. Also includes the results of a computational experiment, involving two distinct sets of files in Portuguese, which shows the effectiveness of these tools.

Keywords. self-organizing map (SOM); generative topographic mapping (GTM); data recovery; knowledge discovery.

1.Introdução

De acordo com Salton & McGill (1983), *recuperação de informação* está relacionada à representação, armazenamento e acesso a itens de informação. Tradicionalmente, entretanto, este termo é relacionado aos métodos de recuperação de documentos de texto contidos em conjuntos de documentos disponíveis.

Com o advento da Internet, a vasta quantidade de dados disponíveis requer ferramentas automáticas para efetuar mineração de dados, que representa uma das tarefas do processo de KDD (*Knowledge Discovery in Databases*). No caso particular de dados na forma textual, o cenário que se apresenta é altamente desafiador para qualquer iniciativa de automatização de processos de recuperação de informação, pois estão disponíveis textos de toda natureza, cuja qualidade e propósito são extremamente diversos (LAGUS, 2000). Assim, é frequente o caso de uma busca tradicional recuperar milhares de documentos – grande parte deles de valor seriamente questionável, quando se consideram os objetivos da busca – ou nenhum documento, devido a um critério muito restritivo. Sendo assim, processos de calibração de filtros eficazes para recuperação de informação relevante são altamente desejáveis.

Este trabalho mostra a aplicação de dois modelos computacionais adaptativos, o Mapa Auto-Organizável (SOM: *Self-Organizing Map*) e o Mapeamento Topográfico Gerativo (GTM: *Generative Topographic Mapping*) na tarefa de recuperação de informação.

[1]Mestre. Docente da Universidade São Francisco-USF (Itatiba-SP) e da Escola Superior de Administração Marketing e Comunicação-ESAMC (Campinas-SP).

2.Recuperação de informação aplicada a documentos textuais

Considerando que grande parte do conhecimento humano é expresso na forma textual em formato de livros, artigos, páginas da Internet etc. (doravante generalizados pelo termo *documentos*), entende-se que o termo *recuperação de informação* aplicado a documentos relaciona-se com a tarefa de satisfazer a necessidade de informação do indivíduo (LAGUS, 2000). A necessidade de informação pode ser entendida como a busca de respostas para determinadas questões ou problemas a serem resolvidos, a recuperação de um documento com particularidades específicas, a recuperação de documentos que versem sobre determinado assunto ou ainda o relacionamento entre assuntos. Embora relativamente simples em seu conceito, esta tarefa envolve questões bastante difíceis de serem resolvidas:

- como se deve armazenar o conjunto de documentos de forma a preservar e evidenciar seu conteúdo e o relacionamento entre os mesmos?
- uma vez armazenados, como recuperá-los eficientemente de forma a satisfazer a necessidade de informação de um indivíduo?

A forma de armazenamento dos documentos é crucial e intrinsecamente determina os métodos possíveis de recuperação dos mesmos. A recuperação de documentos envolve ainda critérios subjetivos, o que sugere métodos interativos. Normalmente, são consideradas duas possibilidades de encaminhamento para estas questões: (1) considerar a natureza estatística da ocorrência das palavras dentro de um documento, levando-se ou não em conta sua ordem (modelo do saco de palavras, do inglês *bag of words*); ou (2) utilizar a abordagem simbólica da linguagem natural (SCHOLTES, 1991).

A primeira abordagem reduz o documento a alguma forma estatística de representação do texto (vetores de frequência de palavras, dicionário de termos (*thesaurus*), palavras-chave etc.), o que leva ao conceito de reconhecimento de padrões. Costuma ser uma abordagem rápida, computacionalmente eficiente e pode lançar mão do conhecimento e de ferramentas estatísticas já bem fundamentadas na literatura. Entretanto, é incapaz de considerar relações simbólicas ou de efetuar inferências conceituais sobre os documentos. A segunda abordagem, ao considerar a natureza simbólica da linguagem, é teoricamente capaz de lidar com as deficiências do primeiro método, mas costuma ser computacionalmente complexa e ineficiente. Normalmente, lança-se mão de cadeias de Markov de palavras ou caracteres, mas a *memória* do método depende da ordem da cadeia de Markov e os requisitos computacionais crescem exponencialmente com o aumento da cadeia (SCHOLTES, 1991).

Os métodos clássicos de armazenamento e recuperação de documentos baseiam-se nestas duas abordagens. Embora exista ainda a forma mais tradicional de todas, a classificação manual, esta é viável apenas em conjuntos reduzidos de textos e é suscetível a aspectos subjetivos, particularmente quando se procura indexar obras que envolvam várias áreas de conhecimento simultaneamente (SALTON & MCGILL, 1983). O processo de armazenamento ou representação, também chamado indexação, busca extrair características do documento que permitam seu armazenamento de forma resumida. O processo de recuperação é booleano (considera a existência ou não de índices ou palavras-chave dentro dos documentos) ou por alguma métrica de distância envolvendo a pergunta feita e o conjunto dos índices armazenados. Todos estes métodos sofrem de problemas comuns (SCHOLTES, 1993; ZUCHINI, 2003):

- dificuldade para processar perguntas indiretas ou incompletas;
- dificuldade para manipular intenções vagas de busca (i.e., quando o usuário não conhece exatamente o tópico sobre o qual procura informação);
- ausência de habilidade de realimentação da busca em função do resultado obtido previamente;
- ausência de vínculos mais efetivos com o contexto da linguagem, pois são consideradas apenas algumas relações sequenciais entre palavras;
- dificuldade na recuperação de documentos por similaridade contextual.

Na tentativa de resolver alguns destes problemas, as redes neurais artificiais são promissoras para a pesquisa, pois exibem capacidades de aprendizado, generalização e sensibilidade a alguns aspectos contextuais necessários para o cenário da recuperação de informação baseada em documentos de texto (SCHOLTES, 1993). A Seção a seguir discute brevemente alguns dos métodos mais importantes ou de reconhecido valor histórico no contexto de armazenamento e recuperação de documentos, alguns dos quais são baseados em redes neurais artificiais. Recomenda-se consultar (ZUCHINI, 2003) para uma abordagem mais extensiva sobre o assunto.

3.Métodos de armazenamento e recuperação de documentos

A comparação de desempenho de métodos de recuperação de documentos é tradicionalmente baseada em duas métricas, precisão e recuperação (SALTON & MCGILL, 1983), dadas pelas Equação 1 e Equação 2:

Equação 1. Precisão

$$Precisão = \frac{No\ de\ documentos\ relevantes\ recuperados}{No\ total\ de\ documentos\ relevantes\ recuperados}$$

Equação 2. Recuperação

$$Recuperação = \frac{No\ de\ documentos\ relevantes\ recuperados}{No\ total\ de\ documentos\ relevantes\ existentes}$$

Infelizmente, é muito difícil comparar métodos de recuperação de documentos devido à subjetividade envolvida na avaliação. A *relevância* de um documento é altamente subjetiva, pois dependente do conhecimento prévio do usuário e é sempre relativa a outros documentos recuperados. Ainda, em geral, aqueles que procuram uma informação podem não saber exatamente o que procurar devido, possivelmente, ao próprio desconhecimento sobre o assunto (SCHOLTES, 1993).

Em função do exposto acima, dificilmente pode-se esperar que um método em particular resolva todos os dilemas da área e é mais provável que modelos híbridos apresentem melhor desempenho que os modelos básicos (ZUCHINI, 2003; LAGUS *et. al.*, 1996A; LAGUS *et. al.*, 1996B; LAGUS, 2000) afirmam que a necessidade de informação está orientando o desenvolvimento de ferramentas interativas, visuais, capazes de oferecer uma visão geral do relacionamento do conjunto de documentos. Também estas ferramentas devem oferecer possibilidades de busca e navegação por entre os documentos, oferecendo níveis de detalhes que podem ser escolhidos pelo usuário (uma espécie de *zoom*). Pullwitt (2002) critica as métricas de precisão e recuperação, úteis para classificação mas não para análise de proximidade contextual entre documentos. Infelizmente, não há ainda na literatura métricas efetivas que permitam estabelecer com precisão a qualidade de métodos de recuperação de informação, embora muitos métodos já foram propostos e vêm sendo empregados na prática, cada qual com as suas potencialidades e limitações, dentre as já levantadas acima (ZUCHINI, 2003).

3.1.Modelo booleano

No modelo booleano (SALTON & MCGILL, 1983), cada documento é indexado por uma coleção de palavras extraídas do documento. Estes índices são palavras cujo valor discriminante é elevado. O valor discriminante mede a capacidade de uma palavra em identificar um documento como relevante e separá-lo de outros não relevantes numa busca. De acordo com o *princípio do menor esforço*, as pessoas tendem a usar termos repetidos em vez de criar novos termos para expressar ideias. Este mesmo princípio comprova que as palavras mais frequentes num documento (e até mesmo na fala) carregam pouco significado na expressão da ideia, sendo utilizadas como elementos de ligação numa frase (SALTON & MCGILL, 1983).

Seguindo esta abordagem, o valor discriminante depende da frequência com que a palavra ocorre, ou em um documento ou no conjunto de documentos. Assim, pode-se considerar duas medidas de frequência absoluta relacionadas entre si, como mostra a Equação 3.

Equação 3. Frequência total

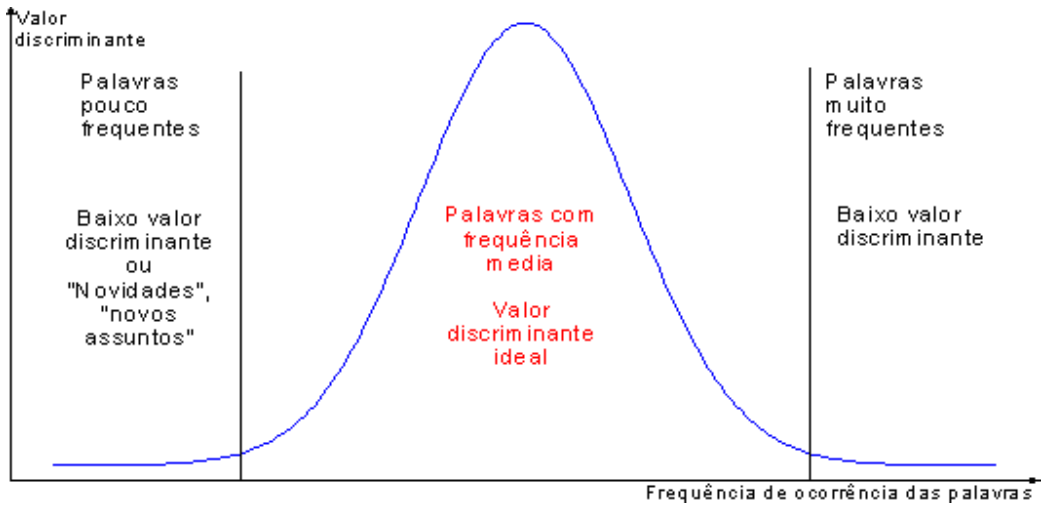
$$freqtot_k = \sum_{i=1}^{n} freq_{ik}$$

onde `n` é o número de documentos, $freq_{ik}$ é a frequência com que a palavra `k` aparece no documento `i` e $freqtot_k$ é a frequência total de ocorrência da palavra `k` no conjunto de documentos.

Assim, palavras que ocorrem com frequência muito alta ou muito baixa não são bons discriminantes e não deveriam ser consideradas como índices possíveis (LUHN, 1958).

Por outro lado, palavras com baixa frequência de ocorrência podem ser entendidas como a representação de termos novos, ou seja, termos cunhados para discutir novos assuntos (ZUCHINI, 2003). Sob este ponto de vista, descartar tais palavras teria efeito negativo sobre a qualidade do mapa, tornando-o pouco sensível a novos termos (ou termos pouco frequentes). Neste artigo considera-se esta última abordagem, ou seja, as palavras com baixa frequência de ocorrência não são descartadas, como sugere a literatura tradicional, mas consideradas como (possivelmente) relevantes. A Figura 1 ilustra esta ideia.

Figura 1. Variação do valor discriminante de um termo em relação à frequência de ocorrência deste no conjunto de documentos. Os limiares de corte são escolhas heurísticas e dependem do conjunto de documentos (ZUCHINI, 2003).



Escolhidas as palavras que comporão o índice, um vetor é associado a cada documento, onde cada dimensão corresponde a um índice, contendo, por exemplo, "1" e "0" conforme o índice esteja presente ou ausente no documento. Uma operação de busca consiste na formulação de uma expressão booleana que é aplicada sobre o conjunto de índices. Embora seja um modelo muito usado por sua simplicidade, há vários inconvenientes com esse método (ZUCHINI, 2003):

- é difícil realizar uma busca adequada (isto é, que recupere documentos relevantes) especialmente quando o usuário não domina o conjunto de palavras-chave (índices) do assunto em questão;
- não há forma de obter resultados aproximados, isto é, a busca ou é bem sucedida ou é mal sucedida, sendo comum a recuperação de milhares ou nenhum documento;
- não havendo um critério de aproximação, não há como classificar os documentos recuperados segundo sua relevância.

3.2.Modelo de espaço vetorial

O modelo de espaço vetorial (SALTON & MCGILL, 1983) é uma variação do modelo booleano. Diferentemente deste, onde apenas a frequência absoluta de ocorrência de uma palavra é considerada,

o modelo de espaço vetorial busca privilegiar palavras que ocorrem de forma concentrada em alguns textos (mesmo que a frequência absoluta destas palavras seja elevada em relação ao conjunto de documentos).

Neste modelo, cada documento é representado por um vetor em que cada dimensão corresponde à frequência relativa de ocorrência de uma determinada palavra dentro deste mesmo documento (diferentemente do modelo booleano, onde apenas a presença ou ausência da palavra é considerada). Agora, o valor discriminante de uma palavra é considerado proporcional à frequência relativa de ocorrência da palavra no documento e inversamente proporcional ao número de documentos do conjunto em que esta aparece. Assim, palavras menos frequentes e concentradas em alguns documentos são boas candidatas para identificar um texto em particular e isto pode ser expresso pela Equação 4.

Equação 4. *Term Frequency*

$$TF_{ik} = \frac{freq_{ik}}{|d_i|}$$

onde TF_{ik} (TF: *Term Frequency*) é a frequência do termo k no documento i e $|d_i|$ é o número de palavras presentes no documento i.

Já aquelas palavras que aparecem em muitos textos de maneira mais ou menos uniforme têm menor valor discriminante e uma possível expressão para este conceito é dada na Equação 5.

Equação 5. *Inverse Document Frequency*

$$IDF_k = \log_2\left(\frac{n}{freqdoc_k}\right) + 1$$

onde IDF_k (IDF: *Inverse Document Frequency*) é o inverso da frequência de ocorrência do termo k em relação ao total de documentos, n, e $freqdoc_k$ é o número de documentos nos quais a palavra k é encontrada pelo menos uma vez.

Uma equação para ponderação de cada palavra no vetor representante dos documentos é sugerida por Salton & McGill (1983) sendo conhecida como TF-IDF (*Term Frequency – Inverse Document Frequency*), dada na Equação 6.

Equação 6. *Term Frequency – Inverse Document Frequency*

$$w_{ik} = TK_{ik} \times IDF_i$$

onde w_{ik} é o valor discriminante da palavra k no documento i.

A busca é executada calculando-se a *distância* entre o vetor representando o texto de busca e os vetores representantes dos documentos, recuperando os mais próximos (dentro de um intervalo dado) ordenadamente. Uma vantagem do modelo é a possibilidade de aplicação direta de algoritmos baseados em métricas de distância vetorial (normalmente euclidianas). Porém, a dimensão dos vetores representantes torna esta abordagem impraticável para conjuntos reais de textos, dada a grande quantidade de palavras envolvidas (ZUCHINI, 2003).

3.3.Outros modelos e variações

Os modelos variantes para recuperação de informação buscam incluir mais informação semântica e de contexto na codificação dos documentos. Isto implica na tentativa de criar modelos que possam incorporar, ao máximo, informações da área de processamento de linguagem natural.

SCHOLTES (1993; 1991) apresenta propostas baseadas em redes neurais artificiais que buscam integrar as informações léxicas, sintáticas e semânticas para aumentar a performance de sistemas de recuperação de informação.

Miikkulainen (1999; 1997) sugere o uso de processamento sub-simbólico da linguagem natural através de SOMs hierárquicos com alguns traços de recorrência. A abordagem parte do princípio que o conhecimento é, de alguma forma, armazenado sob a forma de roteiros (*scripts*). Roteiros são, assim, esquemas estereotípicos de sequências de eventos. Seres humanos possuem, segundo o autor, centenas ou milhares de roteiros.

BOLEY at. al. (1999) aplicam diversos algoritmos de agrupamento por particionamento em documentos obtidos diretamente da Internet. Os métodos testados e propostos não utilizam técnicas de redes neurais artificiais.

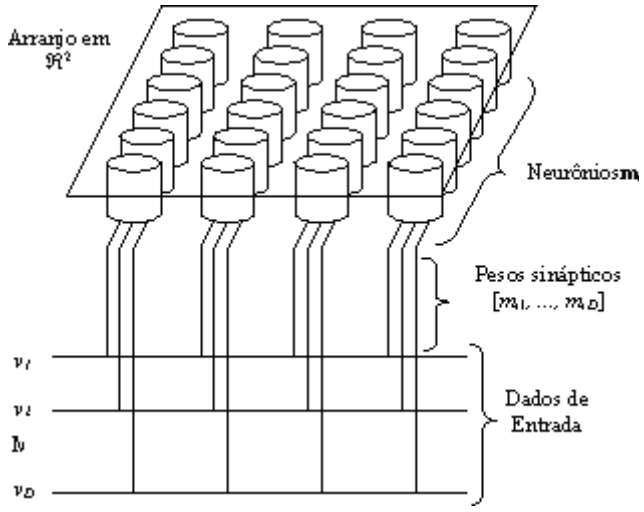
Finalmente, VISA at. al. (2000) propõem uma hierarquia multinível de SOMs para tentar codificar a informação de contexto através de mapas de palavras (SOM semântico), mapas de sentenças e mapas de parágrafos.

4.Os modelos SOM e GTM

O *Mapa Auto-Organizável de Kohonen* (SOM) (KOHONEN, 1982; 1997) é um tipo de rede neural artificial baseada em aprendizado competitivo e não supervisionado, sendo capaz de mapear um conjunto de dados, de um espaço de entrada contido em R^D, em um conjunto finito de neurônios organizados em um arranjo normalmente bidimensional.

A ideia fundamental do SOM é a de que neurônios próximos entre si no arranjo representem dados próximos entre si no espaço de dados, segundo a ideia de preservação da topologia dos dados (ZUCHINI, 2003). Representar um dado aqui significa ter um vetor de pesos que seja mais próximo do dado que qualquer outro vetor de pesos da rede neural. Com isso, a topologia dos dados no espaço original acabará sendo preservada, dentro do possível, pelo arranjo de neurônios em um espaço de menor dimensão. Essencialmente, cada neurônio i é representado por um vetor de pesos sinápticos $m_i = [m_{i1}, \ldots, m_{iD}]^T$ em R^D e todos os neurônios são conectados ao sinal de entrada ou dado recebido como na Figura 2.

Figura 2. Todos os neurônios do arranjo, representados por vetores de pesos sinápticos $m_i = [m_{i1}, \ldots, m_{iD}]$, $i = 1, \ldots, 24$, recebem o mesmo dado de entrada.



Os dados (representados por seus vetores v_n) são então apresentados ao SOM estocasticamente e o arranjo é adaptado para representar os dados da melhor forma possível. A ideia de aprendizado

competitivo diz que o neurônio mais próximo de um item de dado (BMU: *Best Matching Unit*) deve ser adaptado para melhor representar o sinal de entrada, *movendo-se* na direção do dado. No SOM, não apenas o neurônio que ganhou a competição é adaptado mas também seus vizinhos, estabelecendo uma interação local entre os neurônios que, ao longo do aprendizado, promove a organização geral do mapa (KOHONEN, 1997; ZUCHINI, 2003). A Figura 3 representa um função h_{ci} sobre um mapa bidimensional cuja projeção pode ser vista na Figura 4. Quanto mais próximo um neurônio encontra-se do BMU, isto é, quanto menor a distância $||r_c - r_i||$, maior é a adaptação aplicada ao neurônio. O neurônio com maior adaptação é, obviamente, o BMU.

Figura 3. Representação da função h_{ci} sobre um mapa bidimensional.

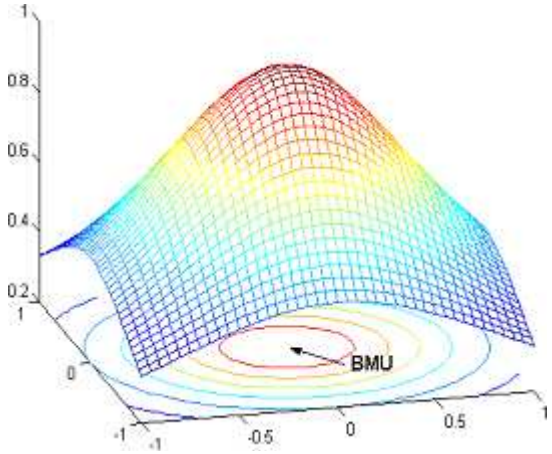
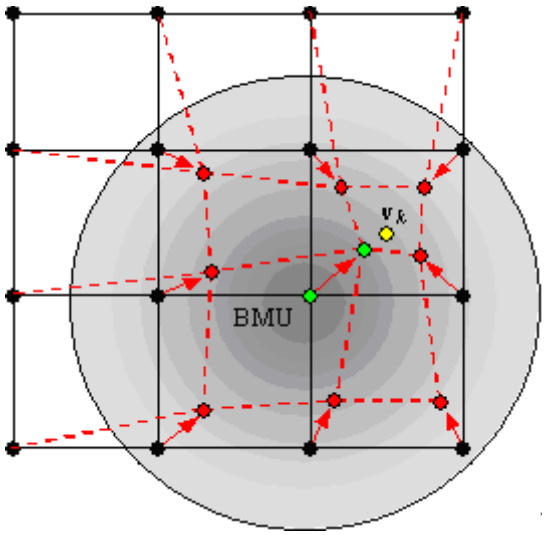
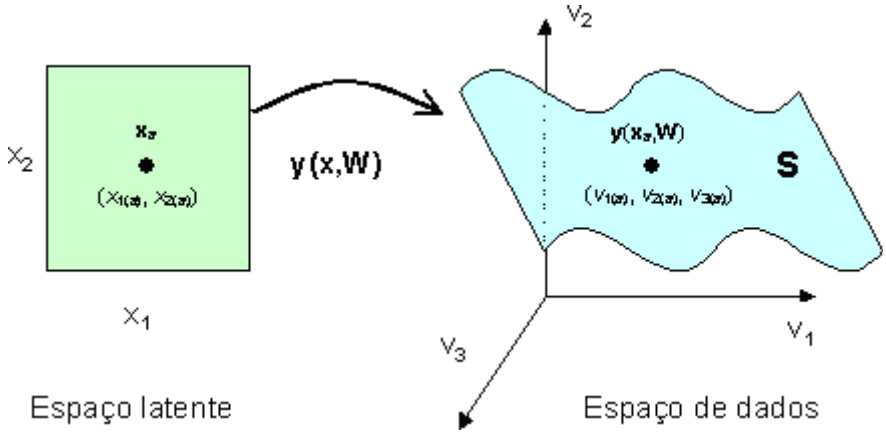


Figura 4. Projeção da função h_{ci} sobre um mapa bidimensional.



O modelo GTM (BISHOP, SVENSÉN & WILLIAMS, 1998) é um modelo que executa um mapeamento paramétrico não linear de um espaço `L`-dimensional de variáveis (chamadas latentes) para um espaço `D`-dimensional de dados de entrada onde, normalmente, `L<D`. Este mapeamento define um subespaço `S` (contido no espaço de entrada) que representa o espaço de variáveis latentes segundo a transformação `y(x,W)`, a qual mapeia pontos `x` do espaço latente para pontos `v` no espaço de dados, como ilustrado na Figura 5 para o caso em que o espaço latente reside em R^2 `(L=2)` e o espaço de dados, em R^3 `(D=3)`.

Figura 5. ideia geral do mapeamento de variáveis latentes.



Na Figura 5 cada ponto do espaço latente (`X`-espaço, à esquerda) é levado ao espaço de dados (`V`-espaço, à direita) através de um mapeamento paramétrico não linear `y(x,W)`, o qual define um subespaço `S` contido no espaço de dados. Cada ponto pertencente a `S` é resultante da aplicação de `y(x,W)` sobre um ponto pertencente ao `X`-espaço. Assim, a transformação `y(x,W)` leva um ponto x_a residente no espaço latente e definido pelas suas coordenadas $(x_{1(a)}, x_{2(a)})$, para um ponto $y(x_a,W)$, pertencente ao espaço `S` e definido por suas coordenadas $(v_{1(a)}, v_{2(a)}, v_{3(a)})$ no espaço de dados.

A hipótese feita pelo modelo GTM é a de que o comportamento do conjunto de dados no espaço `D`-dimensional pode de fato ser expresso por um conjunto menor de atributos[2] (as variáveis latentes) através de um mapeamento paramétrico não linear `y(x,W)`. Uma aproximação para esse raciocínio é imaginar que, embora a dimensão do conjunto de entrada possa ser elevada, muitas das variáveis são correlacionadas entre si, resultando num conjunto potencialmente mais simples que pode representar o comportamento dos dados no espaço original (BISHOP, SVENSÉN & WILLIAMS, 1998). Os modelos baseados nesta ideia são chamados modelos de variáveis latentes (BARTHOLOMEW, 1987).

O mapeamento realizado pelo GTM utiliza-se normalmente de um modelo de probabilidade baseado em mistura de gaussianas, o qual é adaptado pelo algoritmo EM (*Expectation-Maximization*) (DEMPSTER, LAIRD & RUBIN, 1987; BISHOP, SVENSÉN & WILLIAMS, 1996, 1998; SVENSÉN, 1998).

5.Uso de SOM e GTM em recuperação de informação

Ambos os modelos descritos possuem propriedades muito interessantes quando se trata de mapear dados de espaços multidimensionais para um espaço de menor dimensão, passível de análise. Esta característica os torna úteis na recuperação de informação. Várias características da língua portuguesa a tornam um experimento bastante distinto nas tarefas de recuperação de informação, destacando-se:

- elevado número de vocábulos existentes;
- elevado número de sinônimos entre vocábulos;
- elevado número de flexões verbais (comum em línguas latinas);
- diversas possibilidades de construção sintática para a expressão de ideias;
- elevado número de partículas textuais com flexões em gênero, número e grau (como artigos, preposições e advérbios); e

[2]A mesma hipótese é assumida por modelos como *Factor Analysis* (BARTHOLOMEW, 1987) e *Probabilistic PCA* (TIPPING & BISHOP, 1997) com a diferença de que o GTM executa um mapeamento não linear.

- finalmente, grande número de exceções a praticamente todas as regras.

Experimentos realizados com o SOM e o GTM sugerem que a língua portuguesa demanda um tratamento mais cuidadoso para que sejam obtidos resultados aproveitáveis (ZUCHINI, 2003). Especialmente, optou-se pela radicalização[3] das palavras, dada a grande variedade de flexões dos vocábulos da língua portuguesa. O experimento aqui apresentado busca responder uma pergunta: "se o conjunto de documentos possui poucos temas bastante distintos (isto é, com poucas palavras representativas comuns entre si), mesmo um número pequeno de textos (um conjunto estatisticamente pequeno) pode ser classificado conforme seus assuntos?". Uma segunda pergunta possível seria: "se o conjunto de documentos possui muitos temas sem uma distinção expressiva em termos de contexto, há formas para melhor evidenciar a separação dos conjuntos?".

O conjunto de textos "Esporte e Culinária" (EC) possui um total de 52 textos, sendo 25 sobre esporte de competição de carros (*Stock Car*) e 27 tratando de receitas culinárias, num total de 12187 palavras (média de 230 palavras por texto). O conjunto EC possui uma separação de contexto bastante clara (considerando análise manual) e foi utilizado para testar a primeira hipótese. Para um tratamento aprofundado e resultados sobre a segunda pergunta, consulte (ZUCHINI, 2003).

5.1.Mapas semânticos

Ritter & Kohonen (1989) demonstraram num experimento prático que o SOM é capaz de representar graficamente a relação entre valores simbólicos (palavras, no caso) através de uma codificação apropriada do contexto em que se encontram.

A proposta é representar um conjunto de palavras por vetores de forma que seu significado semântico seja, de alguma forma, capturado pelo mapa neural e que, portanto, símbolos *semanticamente próximos* sejam mapeados *topograficamente próximos*. Palavras são particularmente difíceis de serem representadas em forma vetorial. Para tanto, assume-se que a palavra em si não carrega significado intrínseco, mas este depende principalmente do *contexto em que ela está inserida.*

Cada símbolo (palavra) foi codificado através de um vetor real. Para operar com textos livres, entretanto, é necessário carregar o contexto do símbolo. A estratégia utilizada é considerar como atributos va de um símbolo vs a média de todos os símbolos sucessores e predecessores de vs, o que é chamado de contexto médio. A expressão que representa esta ideia é denotada pela Equação 7.

Equação 7. Contexto Médio

$$v_k = \begin{bmatrix} E\left[v_{s(k)-1}\right] \\ \varepsilon v_{s(k)} \\ E\left[v_{s(k)+1}\right] \end{bmatrix}$$

onde $E\{v_{s(k)-1}\}$ é a média (vetorial) de todos os símbolos que precedem a palavra v_k e $E\{v_{s(k)+1}\}$ é a média de todos os símbolos que sucedem a palavra v_k no corpo de texto.

O objetivo final é obter um SOM que organize os documentos de texto conforme sua proximidade contextual. Para tanto, o processo tem as seguintes fases:

1. É executado um pré-processamento no conjunto de documentos, onde se eliminam palavras com valor discriminante baixo (veja Figura 1), sendo também realizada a *radicalização* (*stemming*) das palavras;

2. para cada palavra restante é gerado um vetor $v_{s(k)}$ que representa a palavra k;

3. treina-se um SOM semântico com os vetores de contexto médio obtidos pela codificação de todos os símbolos relevantes do corpo de texto, conforme Equação 7;

[3]Radicalização (*stemming*) é o processo de remoção de sufixos e flexões para obter o radical da palavra, reduzindo assim o número de palavras de um conjunto textual.

4. apresenta-se o texto de cada documento, palavra por palavra, dentro da janela de contexto considerada (no caso, uma palavra de contexto à esquerda e à direita do símbolo) ao SOM semântico já treinado. Deste, obtém-se um *histograma* do documento em relação à frequência com que os neurônios são excitados, considerando a resposta de cada BMU como uma dimensão do vetor. Este vetor é uma espécie de *assinatura estatística* do documento e as respostas dos neurônios BMU são acumuladas à medida que o texto é apresentado ao SOM semântico;

5. os histogramas de documentos, gerados a partir do SOM semântico, são usados para treinar o SOM de documentos, que agora pode representar as relações entre os documentos do conjunto.

5.2.Experimento com SOM e GTM

O conjunto EC foi manipulado segundo o procedimento descrito na Seção 5.1. O conjunto inicial continha 52 textos e 12187 palavras, transformadas as letras em minúsculas e sinais de pontuação e algarismos tendo sido removidos. Após isso, foram removidas, também, 5286 palavras de uso comum e que não agregam informação ao contexto. Estas palavras são artigos (“a, as, os, algum, ...”), preposições (“ante, até, após, ...”), conjunções (“e, ou, porque, quando, onde, ...”) e alguns verbos, incluindo suas flexões (“ser, estar, ter, fazer”). Estes verbos foram escolhidos previamente, antes de qualquer manipulação do conjunto de documentos. Das 6901 palavras restantes, uma análise mostrou haver um total de 2108 palavras diferentes, incluindo as flexões de gênero, número e grau ainda presentes.

O conjunto foi então radicalizado, sendo obtidas 1316 palavras, ou seja, uma redução de aproximadamente 37,6% na quantidade de símbolos. O algoritmo de radicalização utilizado é uma versão adaptada por (ZUCHINI, 2003), sendo superior à versão para português do algoritmo de Porter (ORENGO & HUYCK, 2001). O algoritmo opera em 8 estágios, buscando reduzir, por ordem: (1) a forma plural, (2) transformar a forma feminina para a forma masculina, (3) redução de advérbios pela exclusão do sufixo “-mente”, (4) redução de grau (diminutivo, aumentativo e superlativo), (5) redução do sufixo de substantivos (por exemplo, “contagem -> cont”), (6) redução do sufixo de verbos e flexões para sua raiz, (7) redução da vogal final de palavras como “menino -> menin” e, finalmente, (8) remoção de acentos. Embora com resultados positivos na redução do número de palavras do corpo de texto, o algoritmo apresenta incorreções:

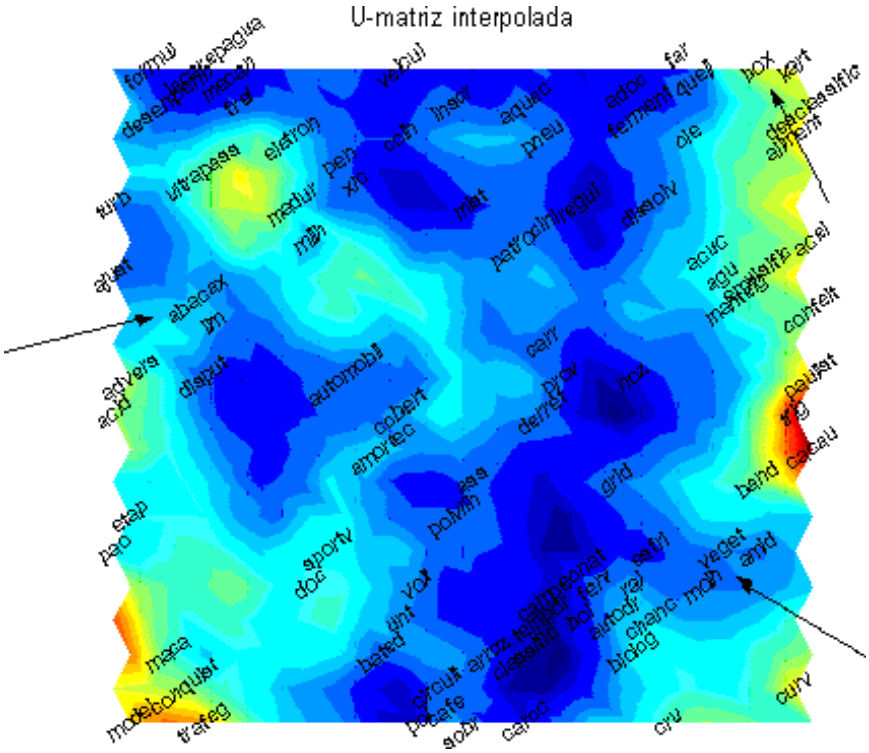
- as formas “alta”, “alto” e “alterado” foram reduzidas para a forma única “alt”.

- “alimentos” e “alimentícios” são palavras semanticamente próximas, mas foram reduzidas para as formas “aliment” e “alimentici”.

O primeiro erro é chamado sobre-radicalização (*overstemming*) e prejudica o índice de precisão na recuperação (Equação 1), pois coloca sob o mesmo símbolo palavras semanticamente distintas. Já o segundo erro, a sub-radicalização (understemming), prejudica o índice de recuperação de documentos (Equação 2) por não entender como semanticamente relacionadas palavras que o são.

Segundo o processo descrito, foi treinado o SOM semântico (Figura 6). Optou-se por um mapa de 20×20 neurônios em um arranjo hexagonal com função de vizinhança gaussiana, o que corresponde a uma média de 3,29 símbolos por neurônio, bem abaixo do máximo sugerido de 10 símbolos por neurônio (KOHONEN, 1998).

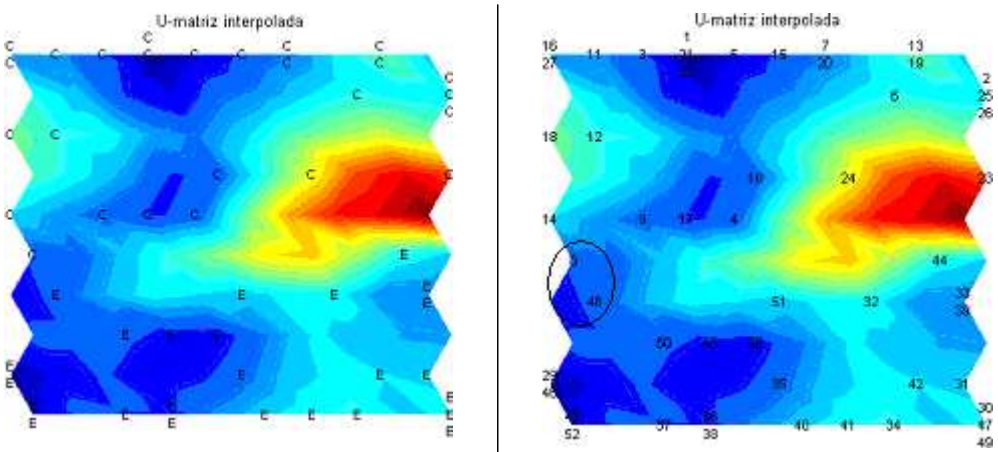
Figura 6. Matriz-U do SOM semântico de 20 × 20 neurônios com a sobreposição de algumas palavras escolhidas manualmente do corpo de texto.



As setas apontam algumas palavras que foram mapeadas proximamente entre si, indicando que o mapa agrupou as palavras semanticamente próximas baseadas em seu contexto.

Após a geração do SOM semântico, o conjunto de texto foi apresentado ao mapa semântico e os histogramas de palavras de cada documento foram obtidos, gerando uma matriz composta por 52 vetores de dimensão 400, representando os mesmos. Estes vetores foram utilizados para treinar um SOM de documentos de 10 × 10 neurônios. A Figura 7 apresenta a matriz-U do SOM com os documentos com os rótulos apresentados.

Figura 7. SOM de documentos com rótulos de documentos.

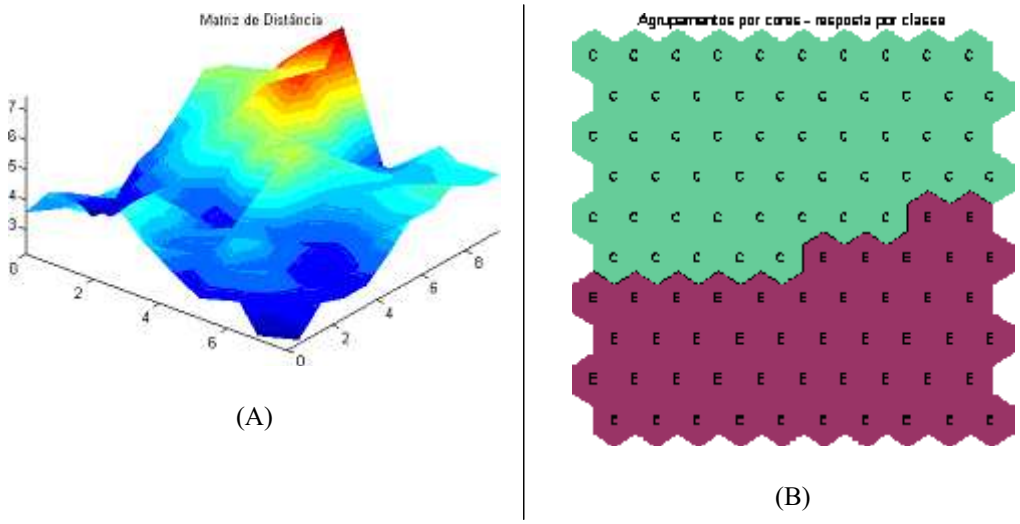


(A) (B)

Os rótulos foram previamente escolhidos com “E” e “C” para representar textos relativos ao esporte ou culinária, respectivamente. Em (A) vê-se os rótulos e em (B) os números identificando cada documento, praticamente separados em dois conjuntos ocupando as metades superior e inferior dos mapas. Os dois itens destacados em (B) sugerem uma proximidade contextual inexistente de fato.

É interessante notar que a matriz-U praticamente separou os textos em dois conjuntos. Entretanto, esta percepção não é óbvia a menos que se recorra aos rótulos pré-definidos. A Figura 8 apresenta a matriz-U e uma classificação por cores tomando cada neurônio e observando a que classe corresponde o vetor de documento mais próximo por ele representado. O resultado torna-se evidente, mas não é razoável admitir a existência de duas classes com base nestas informações apenas. Embora seja inegável que os contextos foram separados eficientemente, a necessidade do mapa em representar os itens pode sugerir proximidade de contexto onde ela, de fato, não existe. Este fato pode ser observado exatamente na fronteira que separa os textos na Figura 7-B e sugere precaução na interpretação dos resultados do SOM de documentos.

Figura 8. Representação da superfície da matriz-U do SOM de documentos.

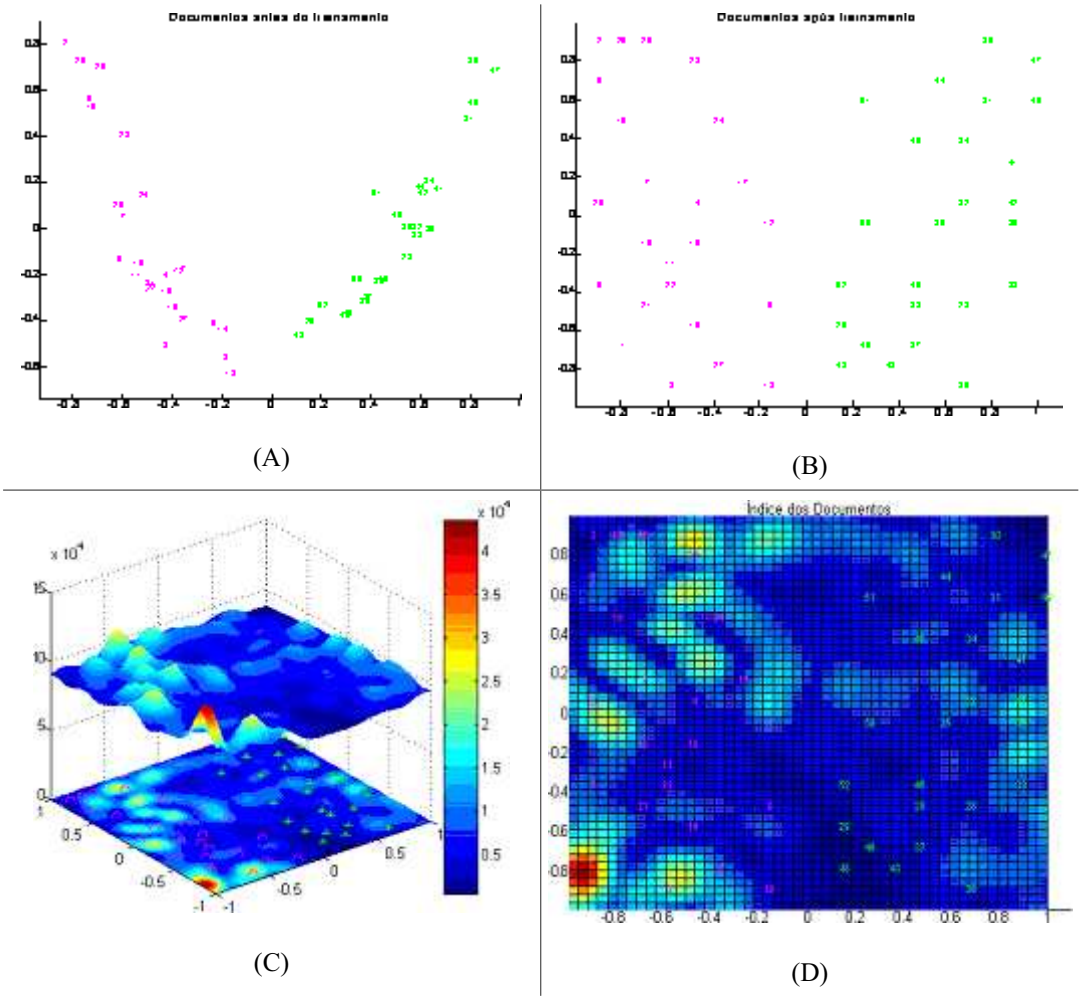


(A) (B)

Embora a resposta neural por classe confirme a separação dos contextos, a matriz-U não é capaz, sozinha, de sugerir esta separação com clareza.

Numa abordagem híbrida (ZUCHINI, 2003), os vetores gerados a partir do SOM semântico foram usados para se adaptar um modelo GTM de 20 × 20 pontos latentes e 12 × 12 funções-base com espalhamento 0,8 em 8 ciclos. O modelo escolhido foi aquele com maior logaritmo da verossimilhança dentre 10 testes onde foram variados diversos parâmetros. O modelo GTM possui uma convergência muito rápida e a separação entre os objetos pode ser percebida claramente mesmo antes do modelo ser adaptado, como pode ser verificado na Figura 9. Por outro lado, assim como no SOM de documentos, não é óbvia a separação dos documentos em dois conjuntos, embora seja inegável que houve separação dos contextos.

Figura 9. Projeção da média *a posteriori* da distribuição dos dados (documentos) sobre o espaço latente antes (A) e depois (B) do treinamento do modelo GTM.



(A) (B)

(C) (D)

Em (C) e (D) os fatores de ampliação, sobre os quais foi projetada a média *a posteriori* dos documentos. A esquerda de todas as figuras encontram-se os documentos de culinária (“C”, em vermelho) e à direita, os de esportes (“E”, em verde).

6.Conclusões

Percebe-se que, em ambas as ferramentas, os resultados foram coerentes e indicam que a primeira hipótese é verdadeira: mesmo um conjunto estatisticamente pequeno de textos, mas com temas bastante distintos, pode ser separado, em termos de contexto, sem grandes dificuldades. Se a informação prévia das classes for disponível, ambas as ferramentas, SOM e GTM, são bastante efetivas na classificação de novos textos.

Vários itens de pesquisa futura podem ser citados:

- desenvolvimento de modelos GTM hierárquicos, com ampliação e exploração automáticas de regiões de dados com grande densidade de pontos;
- desenvolver regras para remoção de palavras com baixo valor discriminante, utilizando técnicas adaptativas, como lógica nebulosa ou algoritmos genéticos, de forma que as regras sejam dependentes do conjunto de textos;
- desenvolvimento de um modelo híbrido de processamento de linguagem natural e redes neurais artificiais, de forma a ampliar a quantidade de informação disponível *a priori* na mineração de textos;

- possibilidade de uso de gaussianas com diferentes variâncias no modelo GTM, de forma a flexibilizar a modelagem dos dados;
- utilização de mapas SOM N-dimensionais com aplicação de algoritmos de segmentação e rotulação automáticos (por exemplo, SL-SOM) na construção de mapas semânticos; e
- desenvolvimento de um algoritmo de radicalização adaptativo ao *corpus* de texto.

7.Referências bibliográficas

BARTHOLOMEW, D. J. *Latent Variable Models and Factor Analisys*. Charles Griffin and Co. Ltd, London, 1987.

BISHOP, Christopher M.; SVENSÉN, Markus; WILLIAMS, Christopher K.I. *EM Optimization of Latent-Variable Models*. In: Touretzky, D.S.; Mozer, M.C.; Hasselmo, M.E. (editors), Advances in Neural Information Processing Systems 8, The MIT Press, Cambridge, MA, pg. 465-471, 1996. Disponível em http://www.ncrg.aston.ac.uk/Papers/postscript/NCRG_96_011.ps.Z. Acesso em 15/09/1999.

_____. *GTM: The Generative Topographic Mapping*. Technical Report NCRG/96/015, Aston University, UK. Published in: Neural Computation 10, pg. 215-234, 1998. Disponível em http://www.ncrg.aston.ac.uk/Papers/postscript/NCRG_96_015.ps.Z. Acesso em 15/09/1999.

BOLEY, Daniel; GINI, Maria; GROSS, Robert; HAN, Eui-Hong; HASTINGS, Kyle; KARYIPIS, George; KUMAR, Vipin; MOBASHER, Bamshad; MOORE, Jerome. *Partitioning-based clustering for Web document categorization*. Decision Support Systems, nº 27, pg. 329-341, 1999.

DEMPSTER, A.P.; LAIRD, N.M.; RUBIN, D.B. *Maximum likelihood for incomplete data via the EM algorithm*. Journal of the Royal Statistical Society, Series B, n.º 39, pg. 1-38, 1977.

KOHONEN, Teuvo. *Self-Organized Formation of Topologically Correct Feature Maps*. Biological Cybernetics 43, pg. 59-69, 1982.

_____. *Self-Organizing Maps*. Series in Information Sciences, vol. 30, 2nd edition. Springer-Verlag, Heidelberg, 1997.

_____. *Self-Organization of Very Large Document Collection: State of the Art*. In: Niklasson, L.; Bodem, M.; Ziemke, T. (editors). Proceedings of the 8th International Conference on Artificial Neural Networks, vol. 1, Springer, London, pg. 65-74, 1998. Disponível em http://websom.hut.fi/websom/doc/ps/kohonen98.ps.gz. Acesso em 04/12/1998.

LAGUS, Krista. *Text Mining with the SOM*. Acta Polytechnica Scandinavica, Mathematics and Computing Series n.º 110. Dr. Tech Thesis, Helsinki University of Technology, Finland, 2000.

LAGUS, Krista; HONKELA, Timo; KASKI, Samuel; KOHONEN, Teuvo. *Self-Organizing Maps of Document Collections: A New Approach to Interactive Exploration*. In: Simoudis, E., Han, J., Fayyad, U. (editors), Proceedings of the Second International Conference on Knowledge Discovery & Data Mining, AAAI Press, Menlo Park, California, pg. 238-243, 1996. Disponível em http://websom.hut.fi/websom/doc/ps/lagus96kdd.ps.gz. Acesso em 04/12/1998.

LAGUS, Krista; KASKI, Samuel; HONKELA, Timo; KOHONEN, Teuvo. *Browsing Digital Libraries with the Aid of Self-Organizing Maps*. In: Proceedings of the Fifth International World Wide Web Conference (WWW5), Paris, France, pg. 71-79, 1996. Disponível em http://websom.hut.fi/websom/doc/ps/lagus96.ps.gz. Acesso em 04/12/1998.

LUHN, H. P. *The Automatic Creation of Literature Abstracts*. IBM Journal of Research and Development, vol. 2, nº 2, pg. 159-165, 1958.

MIIKKULAINEN, Risto. *Script Recognition With Hierarchical Feature Maps*. Connection Science 2, pg. 83-101, 1990. Disponível em http://www.cs.utexas.edu/users/nn/pages/publications/abstracts.html#miikkulainen.script-recognition.ps.Z. Acesso em 22/09/1999.

_____. *Natural Language Processing With Subsymbolic Neural Networks*. In: A. Browne (editor), Neural Network Perspectives on Cognition and Adaptive Robotics. Institute of Physics Publishing, 1997. Disponível em http://www.cs.utexas.edu/users/nn/pages/publications/abstracts.html#miikkulainen.perspectives.ps.Z. Acesso em 22/09/1999.

ORENGO, Viviane Moreira; HUYCK, Christian. *A Stemming algorithm for the Portuguese Language*. In: Proceedings of SPIRE'2001 Symposium on String Processing and Information Retrieval, Laguna de San Raphael, Chile, November 2001.

PULLWITT, Daniel. *Integrating Contextual Information to Enhance SOM-Based Text Document Clustering*. Neural Networks 15, Special Issue, pg. 1099-1106, 2002.

RITTER, Helge; KOHONEN, T. *Self-Organizing Semantic Maps*. Biological Cybernetics 61, pg. 241-254, 1989.

SALTON, G.; MCGILL, M.J. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, NY, 1983.

SCHOLTES, Johannes C. *Neural Nets and Their Relevance for Information Retrieval*. CL-1991-02, ITLI Prepublication Series for Computational Linguistics, Institute for Logic, Language and Computation (ILLC), University of Amsterdam, 1991.

_____. *Neural Networks in Natural Language Processing and Information Retrieval*. PhD Thesis, Institute for Logic, Language and Computation (ILLC), University of Amsterdam, 1993.

SVENSÉN, Johan Fredrik Markus. *GTM: The Generative Topographic Mapping*. PhD Thesis, Aston University, April 1998. Disponível em http://neural-server.aston.ac.uk/GTM/thesis.html. Acesso em 24/09/1998.

TIPPING, Michael E.; BISHOP, Christopher M. *Probabilistic Principal Component Analysers*. Technical Report, Neural Computing Research Group, Aston University, 1997.

VISA, Ari; TOVONEN, Jarmo; RUOKONEN, Piia; VANHARANTA, Hannu; BACK, Barbro. *Knowledge Discovery from Text Documents Based on Paragraph Maps*. In: Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS'33), Maui, Hawaii, January 4, CDROM, 2000.

ZUCHINI, Márcio H. *Aplicações de Mapas Auto-Organizáveis em Mineração de Dados e Recuperação de Informação*. Dissertação de mestrado. Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação. Campinas, SP: 2003.