10th Brazilian Congress on Computational Intelligence (CBIC'2011), November 8 to 11, 2011, Fortaleza, Ceará Brazil

TÉCNICAS BASEADAS EM REDES NEURAIS ARTIFICIAIS E LÓGICA FUZZY PARA MINERAÇÃO DE TEXTOS

Jonas Henrique Mendonça[1], Isabela Neves Drummond[2], Sandra Aparecida Sandri[1]

[1]Instituto Nacional de Pesquisas Espaciais, [2]Universidade Federal de Itajubá
jonas.henrique01@gmail.com, isadrummond@unifei.edu.br, sandra.at.lac.inpe.br@gmail.com

Abstract – There is currently a considerable amount of information available in text format on the Internet and the networks of large corporations. This information can be found in an unstructured format, difficult to handle by computer programs. It is proposed in this work, the study of a neuro-fuzzy system applied to the task of text mining. The models investigated for the clustering task of texts is a neural network fuzzy-ART and a neural network SOM. To validate this model experiments are carried out with the bases of texts *Reuters Transcribed Subset* and *SyskillWebert*. Finally, the fuzzy-ART network is compared with the neural network SOM.

Keywords – Text mining, fuzzy logic, neural networks, neuro-fuzzy, clustering

1 Introdução

Nos últimos anos, houve um enorme crescimento na quantidade de documentos e textos disponíveis em formato digital e a recente popularidade da Internet tem acelerado o crescimento. A acessibilidade a grandes coleções de documentos em formato eletrônico enfatiza a necessidade de técnicas de recuperação de informação inteligentes. É neste contexto que se insere este trabalho.

A mineração de textos (MT) é um conjunto de métodos para navegar, organizar, achar e descobrir informações em bases textuais. Pode ser vista como uma extensão da mineração de dados, focada na análise de textos [14]. O uso desta tecnologia permite recuperar informações, extrair dados, resumir documentos, descobrir padrões, associações e regras e realizar análises qualitativas e quantitativas em documentos de texto.

Este trabalho tem por objetivo o estudo de modelos baseados em redes neurais artificiais mapas auto-organizáveis (SOM) e fuzzy-ART em problemas de MT. Estes modelos foram implementados e os resultados obtidos foram comparados aos resultados apontados por um especialista humano.

O presente trabalho está organizado da seguinte maneira: a seção 2 traz uma breve explanação sobre o que é mineração de textos. As técnicas utilizadas, lógica *fuzzy* e redes neurais artificiais, são descritas nas seções 3 e 4 respectivamente. A quinta seção deste artigo mostra como o sistema de agrupamento de textos foi implementado e como os parâmetros iniciais das redes neurais artificiais foram definidos. Os resultados obtidos foram descritos na seção 6 e, ao final, a seção 7 apresenta as conclusões obtidas a partir dos experimentos realizados.

2 Mineração de Textos

Em um contexto em que grande parte dos dados corporativos encontra-se disponível em forma textual, o processo de mineração de textos surge como uma poderosa ferramenta de apoio à gestão de conhecimento.

Segundo Zanasi [14], mineração de textos é o processo de extrair, dirigido pelos dados, conhecimento não conhecido previamente através de bases de dados textuais. O processo de mineração de textos pode ser dividido em quatro etapas: identificação do problema, pré-processamento, extração do conhecimento e pós-processamento.

Este trabalho se concentra na etapa de extração do conhecimento considerando textos previamente pré-processados. Suas principais tarefas são a obtenção de regras de associação, o agrupamento e a sumarização de documentos. O agrupamento é um método de descoberta de conhecimento, utilizado para identificar relacionamentos entre objetos, facilitando a identificação de classes. No caso de textos não estruturados, o agrupamento identifica nos textos conteúdos similares, agrupando-os e gerando grupos de textos similares, sendo útil quando não se tem um dos assuntos tratados em cada texto e deseja-se separá-los por assunto [13].

3 Lógica *Fuzzy*

Segundo a teoria da lógica *fuzzy* [15], um elemento pode pertencer a um conjunto com um grau de pertinência, diferentemente da lógica clássica em que um elemento pertence totalmente ou não pertence a um determinado conjunto. Ou seja, dado um universo de discurso *X*, um subconjunto *fuzzy A* de *X* é definido por uma função de pertinência representada na Equação 1, que associa a cada elemento *x* de *X* o grau $\mu_A(x)$, compreendido entre 0 e 1, com o qual *x* pertence a *A*.

$$\mu_A(x): X \rightarrow [0,1] \quad (1)$$

Na atividade de MT, usando a teoria dos conjuntos *fuzzy*, os conjuntos que representam os documentos são compostos por duplas {termo,peso}, sendo o peso um valor difuso definido entre zero e um. Este valor indica a importância do termo, quanto mais próximo do valor um, mais relevante é o termo.

A partir da atribuição da relevância dos termos em relação ao documento, os sistemas *fuzzy* baseiam-se na similaridade, permitindo que os resultados ofereçam não apenas classificações exatas de um documento em relação a uma classe, mas também classificações parciais.

Baseado nas idéias de Oliveira [10], Loh [7] define uma fórmula, apresentada na Equação 2, para calcular similaridade considera as diferenças e as semelhanças entre os documentos, utilizando operadores *fuzzy* adequados as situações. Abaixo é dada a fórmula (Equação 2):

$$gs(X,Y) = \frac{\sum_{h=1}^{k} gi_h(a,b)}{N} \quad (2)$$

onde: gs é o grau de similaridade entre documentos X e Y; gi é o grau de igualdade entre pesos do termo h (peso a no documento X e peso b no documento Y); h é um índice para os termos comuns aos dois documentos; k é o número total de termos comuns aos dois documentos e N é o número total de termos nos dois documentos sem contagem repetida.

A partir da aplicação desta fórmula, cada vez que um termo é encontrado em ambos os documentos, um valor é acumulado. Esse valor vai definir o grau de similaridade entre os textos. O valor que deve ser acumulado é dado pelo grau de igualdade entre os pesos. Este valor é calculado pela equação 3, apresentada por Wives [13]:

$$gi(a,b) = 0{,}5 * [(a \rightarrow b) \wedge (b \rightarrow a) + (\bar{a} \rightarrow \bar{b}) \wedge (\bar{b} \rightarrow \bar{a})] \quad (3)$$

A utilização do grau de igualdade é necessária, pois, mesmo que os termos sejam iguais, eles podem ter pesos diferentes entre os documentos analisados. Estes pesos podem ser calculados pelas fórmulas de frequência de termo. Sendo assim, quando um termo aparece em ambos os documentos com pesos muito diferentes, a igualdade diminui, e com pesos semelhantes, a igualdade aumenta.

O resultado deste processo será um valor entre zero e um, como todo resultado *fuzzy*. Quanto mais próximo de zero, menos similares são os documentos e quanto mais próximo a um, mais similares.

4 Redes Neurais Artificiais (RNA)

Redes Neurais Artificiais são sistemas distribuídos altamente paralelos compostos por simples unidades de processamento que simulam o comportamento de um neurônio biológico [2], dispostas em uma ou mais camadas. Cada conexão entre dois neurônios possui um peso. Estes pesos guardam o conhecimento de uma rede neural e são usados para definir a influência de cada entrada recebida por um neurônio na sua respectiva saída. Ajustando-se os seus pesos, a rede neural assimila padrões e é capaz de fazer generalizações, isto é, produzir saídas consistentes para entradas não apresentadas anteriormente.

O neurônio é o elemento processador da rede neural. Cada neurônio gera uma saída a partir da combinação de sinas de entrada recebidos de outros neurônios, com os quais está conectado, ou a partir de sinais externos. A saída de um neurônio é, na maioria dos modelos, o resultado de uma função de ativação aplicada à soma ponderada de suas entradas.

A topologia de uma rede é descrita por um grafo de nodos (neurônios) e conexões (pesos). Ela é descrita pelo número de camadas da rede, o número de neurônios em cada camada e o tipo de conexão entre nodos.

4.1 Mapas Auto-Organizáveis (SOM)

A rede SOM [5] 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.

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. 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 (*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.

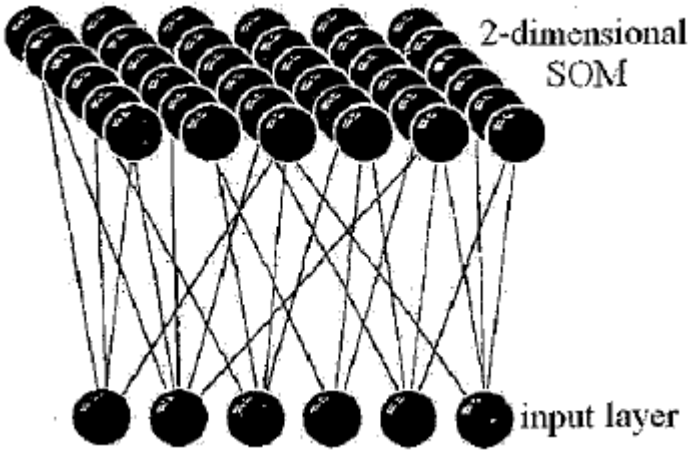

Figura 1: Arquitetura da rede neural artificial SOM

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. A Figura 1[9] mostra a arquitetura de uma rede SOM.

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.

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.

4.2 Redes *Fuzzy*-ART

A rede neural auto-organizável ART (*Adaptive Resonance Theory* – Teoria da Ressonância Adaptativa), para Fausset [3], é apropriada para aplicações de reconhecimento de padrões e classificação de dados, projetada para controlar o grau de similaridade entre padrões que são colocados em um mesmo grupo. Além disso, foi desenvolvida para solucionar o problema da estabilidade-plasticidade, ou seja, não precisa recomeçar o treinamento do ponto inicial a cada novo padrão de entrada que aparecer e ainda preserva o conhecimento adquirido.

A rede ART foi criada por Stephen Grossberg em 1976[4]. É um modelo de arquitetura de rede neural, onde os algoritmos são implementados em termos de aproximações de equações diferenciais, visando uma analogia ao modelo dos neurônios biológicos [3].

Apesar da rede ART ser uma rede não supervisionada, possui um mecanismo de controle do grau de similaridade que é função do parâmetro ρ (limiar de vigilância), cujo valor é especificado pelo usuário. Quando um novo padrão não se enquadra a qualquer grupo já existente, este mecanismo provoca a formação de um novo grupo, determinando se um novo padrão de entrada pode ser incluído em um dos agrupamentos.

A arquitetura básica das redes ART envolve duas camadas de neurônios. Uma para processar os dados de entrada e outra de saída para agrupar os dados por meio de treinamentos específicos, ligadas por meio de conexões, denominadas *feedforward* (W) e *feedback* (B). A rede também possui para cada camada unidades de controle, chamadas de C1, controlando o fluxo de dados para a camada de entrada e o C2 controlando o fluxo de dados para a camada de saída. As funções das unidades de controle resumem-se em determinar o fluxo de dados para a camada de saída e habilitar ou desabilitar neurônios da camada de saída.

A rede ART ainda possui um mecanismo de *reset*, responsável pela verificação da semelhança entre um vetor de entrada e o neurônio vencedor da fase de reconhecimento, utilizando um limiar de vigilância, determinando se o vetor de entrada pode ser incluído em um dos grupos. A Figura 2 ilustra a arquitetura da rede ART.

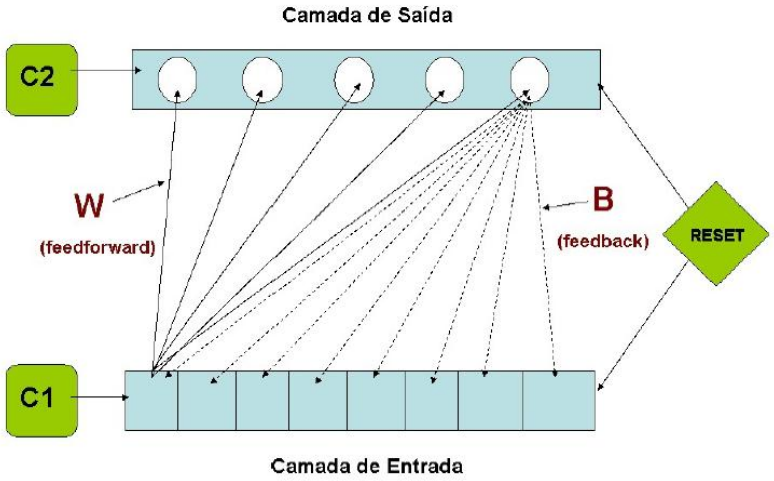

Figura 2 – Arquitetura da rede fuzzy-ART

O aprendizado na rede ART é competitivo. A camada de entrada recebe o padrão e calcula sua ativação *Tj* (Equação 4) para cada neurônio *J* da camada de saída. O neurônio *J*, com maior atividade da camada de saída, torna-se um candidato para codificar o padrão de entrada. Nesse ponto, os outros neurônios tornam-se inativos e a camada de entrada combina a informação entre o padrão de entrada e o neurônio candidato. O neurônio torna-se vencedor e aprende o vetor de entrada (Equação 5), dependendo da similaridade entre o vetor de entrada e o de pesos. Essa decisão é tomada pela unidade de *reset*, que compara sinais provenientes da camada de entrada, verificando a similaridade entre o vetor de entrada e o de pesos do neurônio candidato. Se a similaridade entre o vetor de entrada e o vetor de pesos for menor que o limiar ρ, o neurônio candidato é marcado como inibido e um novo candidato é escolhido. Tal sequência é mantida até encontrar um neurônio capaz de representar o padrão ou até que todos os neurônios da camada de saída estejam inibidos. Nesse caso, a rede cria um novo neurônio para armazenar o padrão ou informa que o padrão não pode ser representado.

$$T_j = \frac{|I \cap W_j|}{\propto + |W_j|} \tag{4}$$

$$W_j^{novo} = txAprendizado * \left(I \cap W_j^{velho}\right) + \left((1 - txAprendizado) * W_j^{velho}\right) \tag{5}$$

Os neurônios da camada de saída podem estar em três estados diferentes: ativo, inativo e inibido. O neurônio encontra-se no estado ativo quando se torna candidato a aprender o padrão de entrada; inativo, quando perde a competição para ser um candidato; e inibido, quando o neurônio foi previamente escolhido como candidato para o padrão de entrada, porém não mostrou similaridade suficiente para aprender o padrão. No estado inibido, o neurônio não poderá competir novamente para aprender o padrão corrente. A condição de parada é a quantidade de ciclos de treinamento para a rede.

O sinal de *reset* é calculado de acordo com a similaridade entre o vetor de entrada e o vetor de pesos do neurônio candidato. Tal sinal define se o candidato será aprovado ou não. Se o candidato for aprovado, o vetor de pesos do neurônio vencedor é adaptado, combinando-se ao vetor de entrada para produzir outro vetor de pesos. O treinamento é encerrado quando a condição de parada é encontrada, ou seja, até que se atinja o número de iterações de treinamento pré-definido.

As redes *Fuzzy*-ART são modelos formalmente parecidos com as redes ART, porém realiza duas operações adicionais; uma relacionada à aprendizagem e outra ao pré-processamento das entradas. Uma rede *Fuzzy*-ART gera agrupamentos de vetores de características *fuzzy*. Mais especificamente, segundo Silva [11], cada componente do vetor de entrada i é um valor de pertinência da função membro de uma determinada característica difusa, indicando o quanto esta característica está presente na amostra.

As computações dos operadores nebulosos E e OU são implementados por meio das funções de mínimo (∧) e máximo (∨), respectivamente, conforme demonstrado na Tabela 1.

Tabela 1 - Analogia entre as redes ART e Fuzzy-ART

	ART (BINÁRIO)	FUZZY-ART
Escolha de Categoria	$T_j = \frac{\lvert I \cap W_j \rvert}{\alpha + \lvert W_j \rvert}$	$T_j = \frac{\lvert I \wedge W_j \rvert}{\alpha + \lvert W_j \rvert}$
Critério de Similaridade	$\frac{\lvert I \cap W_j \rvert}{\lvert I \rvert} \geq \rho$	$\frac{\lvert I \wedge W_j \rvert}{\lvert I \rvert} \geq \rho$
Aprendizado Rápido	$W_j^{novo} = I \cap W_j^{velho}$	$W_j^{novo} = I \wedge W_j^{velho}$

onde I são os vetores de entrada, W_j são os vetores de pesos adaptativos e $\rho \in [0,1]$ é o parâmetro de vigilância.

5 Sistema Implementado

Com o objetivo de se obter um enriquecimento prático acerca da metodologia de mineração de textos estudada, foi implementado um sistema de agrupamento de textos contendo modelos baseados nas redes neurais *fuzzy*-ART e SOM.

Ao fim do processo de treinamento é exibida para o usuário uma tabela mostrando as categorias criadas e os documentos que as compõem como mostrado na Figura 3.

Treino1

Categoria 1	Categoria 2	Categoria 3	Categoria 4
textos_Maid/Bands/11	textos_Maid/Bands/18	textos_Maid/Bands/16	textos_Maid/BioMedic...
textos_Maid/Bands/20	textos_Maid/Bands/19	textos_Maid/Bands/36	textos_Maid/BioMedic...
textos_Maid/Bands/3		textos_Maid/Bands/43	textos_Maid/BioMedic...
textos_Maid/Bands/37		textos_Maid/Bands/50	textos_Maid/BioMedic...
textos_Maid/Bands/39		textos_Maid/Bands/52	textos_Maid/BioMedic...
textos_Maid/Bands/49		textos_Maid/Bands/57	textos_Maid/BioMedic...
textos_Maid/Bands/7		textos_Maid/Bands/in...	textos_Maid/BioMedic...
		textos_Maid/BioMedic...	textos_Maid/BioMedic...
		textos_Maid/BioMedic...	textos_Maid/Goats/1
		textos_Maid/BioMedic...	textos_Maid/Goats/10
		textos_Maid/BioMedic...	textos_Maid/Goats/22
		textos_Maid/BioMedic...	textos_Maid/Goats/27
		textos_Maid/BioMedic...	textos_Maid/Goats/30
		textos_Maid/BioMedic...	textos_Maid/Goats/32
		textos_Maid/BioMedic...	textos_Maid/Goats/36

Figura 3 – Modo como os resultados são exibidos ao usuário

5.1 Treinamento rede SOM

O treinamento da rede SOM emprega vizinhança quadrangular e a função gaussiana. O número de iterações (quantidade de vezes que cada documento é apresentado a RNA) é definido pelo usuário. O número inicial de neurônios é igual ao número de documentos que serão apresentados a RNA. Os demais parâmetros são gerados aleatoriamente.

Para efetuar o treinamento, documentos são selecionados aleatoriamente e apresentados a RNA até que se atinja o número de iterações escolhido pelo usuário. Assim como o número de iterações, a porcentagem de documentos selecionados para treinamento e teste são determinados pelo usuário do sistema.

Ao fim da fase de treinamento, os neurônios que compõem a rede SOM são rearranjados utilizando-se as métricas de similaridade *fuzzy* apresentadas nas equações 2 e 3. Tal procedimento garante que neurônios com características muito próximas sejam mantidos como representantes de categorias diferentes.

5.2 Treinamento rede *fuzzy*-ART

Os documentos selecionados para treinamento são apresentados a RNA até que seja atingida a condição de parada. A criação de novas categorias é limitada pelo número de documentos que compõem a base de dados conforme a Equação 6.

$$\max Categorias = 5 * \sqrt{(numerodedocumentos)} \quad (6)$$

O limiar de vigilância ρ presente no mecanismo de *reset* da RNA é escolhido pelo usuário de acordo com sua necessidade de especialização ou generalização dos dados apresentados.

Durante a etapa de treinamento da RNA são armazenadas as taxas de similaridade entre os neurônios que formas a camada de saída da RNA e os documentos apresentados. Ao fim do treinamento, calcula-se a média aritmética e o desvio padrão destas medidas. Calcula-se então, a similaridade difusa definida pelas equações 2 e 3 entre todas as categorias criadas.

Aquelas categorias cujos neurônios representantes possuem um grau de similaridade superior à similaridade média entre as categorias acrescida do desvio padrão são eliminadas. Os neurônios restantes correspondem às categorias que representam a base de documentos apresentadas à RNA.

6 Experimentos e Resultados

Os experimentos foram realizados utilizando-se as bases de textos *Reuters Transcribed Subset* [6] e *Syskill Webert* [11] . A base de textos *SyskillWebert* é formada por páginas *Web* agrupadas por um especialista humano em 4 classes (Bands, BioMedical, Goats e Sheep) de acordo com seu conteúdo. Estas classes são compostas de 61, 136, 74 e 71 textos, respectivamente, totalizando 342 documentos. A base de textos *Reuters Transcribed Subset* contém 10 categorias formadas por 20 textos cada. Os textos que compõem esta base foram retirados de notícias transmitidas pela agência de notícias *Reuters*. Ambas as bases de textos utilizadas são escritas em língua inglesa. O critério para seleção das bases de textos para este trabalho foi a utilização destes documentos em outros sistemas de agrupamento possibilitando, portanto, uma avaliação comparativa.

Os textos foram pré-processados com a ferramenta *Pretext* e transformados em uma matriz atributo-valor. Um dos problemas encontrados na mineração de textos é a dimensionalidade dos atributos de um corpus, ou seja, a relação entre o número de documentos da coleção, a quantidade de termos que aparece no total da coleção e a quantidade de termos que aparece em cada documento, que pode resultar numa matriz esparsa. Para resolver este problema, foram aplicados os cortes de Luhn [8] especificando as freqüências mínimas e máximas dos termos da coleção. Um ponto importante a ser ressaltado é a validade do método de amostragem utilizado. Como o objetivo deste trabalho é avaliar o comportamento da implementação e não a qualidade do pré-processamento de textos, o pré-processamento realizado, isto é, a medida utilizada *tfidf*, bem como os cortes de Luhn são muito simples e não utilizam informações adicionais dos conjuntos de teste. Caso for realizado um pré-processamento utilizando medidas mais sofisticadas para determinar o valor de cada termo nos documentos, bem como uma redução mais apurada da dimensão dos atributos, então, cada um dos conjuntos de treinamento e teste deveriam ser pré-processados independentemente [1].

Os testes realizados consistem em comparar o número de categorias geradas pela RNA e o número de categorias definidas por um especialista humano para cada base de textos variando-se a taxa de aprendizado e o limiar de vigilância da RNA. Para todos os testes realizados o número de iterações de treinamento foi fixado em 5. Este valor foi firmado através de experimentos realizados durante a implementação do sistema.

Os experimentos foram realizados variando-se o grau de similaridade utilizado para reagrupar os neurônios ao fim do treinamento da RNA SOM. Os mesmos valores foram utilizados como limiar de vigilância da RNA *fuzzy*-ART. Foram efetuados 108 experimentos utilizando a RNA *fuzzy*-ART, 54 utilizando a base de textos *Reuters Transcribed Subset* e outros 54 experimentos utilizando a base de textos *SyskillWebert*. Os resultados obtidos são apresentados nas Tabelas 2 e 3.

Tabela 2 – Agrupamentos formados para a base de textos *Reuters Transcribed Subset* utilizando a RNA Fuzzy-ART

Reset	Porcentagem de documentos para treinamento					
	30%	40%	50%	60%	70%	80%
0,2	7	**10**	13	14	**10**	13
0,25	13	**9**	**11**	14	**11**	13
0,3	**10**	**10**	**9**	**10**	**9**	15
0,35	8	**11**	**11**	**11**	**10**	**11**
0,4	**9**	**10**	**10**	**10**	12	**11**
0,45	13	**9**	**11**	**11**	**11**	12
0,5	**10**	12	13	13	12	13
0,55	13	**10**	16	**10**	**10**	15
0,6	13	13	**10**	**11**	15	12

Tabela 3 – Agrupamentos formados para s base de textos *SyskillWebert* utilizando a RNA Fuzzy-ART

Reset	Porcentagem de documentos para treinamento					
	30%	40%	50%	60%	70%	80%
0,2	**4**	**4**	6	6	6	**5**
0,25	**4**	**3**	**4**	**4**	**4**	**4**
0,3	**5**	**4**	**5**	6	**5**	**5**
0,35	**4**	**4**	**4**	**5**	**5**	**4**
0,4	6	**4**	**4**	**4**	**4**	**4**
0,45	**5**	**4**	**4**	**4**	**5**	**5**
0,5	**4**	**4**	48	6	**3**	**3**
0,55	57	**4**	**4**	**3**	**5**	37
0,6	51	43	52	**5**	**4**	**4**

Assim como os dados mostrados nas Tabelas 2 e 3, as Tabelas 4 e 5 trazem os resultados obtidos quando as mesmas bases de dados foram apresentadas à rede SOM sendo que, para esta RNA, foram efetuados 36 experimentos para a base de textos *Reuters Transcribed Subset* e outros 36 experimentos para a base de textos *SyskillWebert*, totalizando 72 experimentos.

Tabela 4 – Nº de categorias obtidas pela rede SOM para a base de textos *Reuters Transcribed Subset*

Grau de Similaridade Fuzzy	Nº de categoria / % de documentos utilizados para treinamento			
	40%	50%	60%	70%
0,1	12	**10**	8	**10**
0,2	8	12	7	**9**
0,3	**9**	**11**	**9**	**9**
0,4	8	8	**11**	**9**
0,5	**9**	**10**	**10**	**11**
0,6	**10**	8	**9**	7
0,7	**9**	**10**	**9**	**10**
0,8	**9**	**11**	8	**9**
0,9	**11**	8	**11**	**11**

Tabela 5 – Nº de categorias criadas pela rede SOM para a base de textos *SyskillWebert*

Grau de Similaridade Fuzzy	Nº de categoria / % de documentos utilizados para treinamento			
	40%	50%	60%	70%
0,1	6	**4**	2	**4**
0,2	2	7	1	**3**
0,3	**3**	**5**	**3**	**3**
0,4	2	**4**	**5**	**3**
0,5	**3**	**5**	**4**	**5**
0,6	**4**	2	**3**	1
0,7	**3**	**4**	**3**	**4**
0,8	**3**	**5**	2	**3**
0,9	**5**	2	**5**	**5**

Comparando-se as Tabelas 2, 3, 4 e 5 verifica-se que ambas as redes apresentaram um comportamento semelhante, visto que a rede *fuzzy* ART apresentou uma taxa de acerto igual a 67%, enquanto a rede SOM acertou 70% dos experimentos realizados. Para as duas bases de textos, a taxa de acerto é obtida tendo como base o número de categorias apontadas por um especialista humano, sendo que este valor pode variar em uma unidade para mais ou para menos.

Mas, quando comparamos os resultados obtidos nestes trabalho com o número de categorias apontado por um especialista humano (10 categorias para a base de textos *Reuters Transcribed Subset* e 4 categorias para a base de textos *SyskillWebert*), a rede *fuzzy*-ART obtive sucesso em 42 dos 108 casos de teste realizados (somando as duas bases de textos) e a rede SOM acertou 14 dos 72 testes realizados, ou seja, 39 e 20% de acerto, respectivamente.

Outro importante fato a ser descrito foi como o uso da similaridade difusa reduziu a vulnerabilidade da rede *fuzzy*-ART à variação de seus parâmetros de entrada. A Tabela 6 mostra a reação desta RNA à variação do limiar de vigilância sem a etapa de pós-processamento do treinamento proposta.

Tabela 6 – Número de categorias geradas pela rede *fuzzy*-ART sem etapa *fuzzy* de pós-processamento

Reset	Porcentagem de documentos para treinamento					
	30%	40%	50%	60%	70%	80%
0,2	**5**	**4**	6	**5**	**3**	**5**
0,25	**4**	**3**	**4**	**4**	**4**	**4**
0,3	**5**	**4**	**5**	6	**5**	**5**
0,35	**4**	**4**	**4**	6	**5**	**4**
0,4	6	**4**	**4**	**4**	**4**	**4**
0,45	10	**4**	11	**4**	**5**	**5**
0,5	**4**	15	48	6	**3**	17
0,55	60	20	37	**3**	**5**	37
0,6	67	43	52	**5**	12	91

Observando os resultados exibidos na Tabela 6 e comparando-os aos mostrados na Tabela 3 pode-se observar que a rede mostrou-se bem menos sensível as variações impostas sobre o limiar de vigilância . Pode-se ver que à medida que o limiar de vigilância aumenta o número de categorias geradas também aumenta porém, de maneira menos abrupta.

7 Conclusão

Este artigo apresentou o estudo de técnicas utilizando redes neurais artificiais e lógica *fuzzy* com o objetivo de gerar agrupamento de textos. Comparando os resultados gerados pela rede *fuzzy*-ART e os resultados obtidos pelas redes SOM percebe-se que apesar de a taxa de acerto média ser muito próxima, 67% nas redes *fuzzy*-ART e 70% na rede de SOM, quando leva-se em conta somente os experimentos cujos resultados foram idênticos aos apontados por um especialista humano, a rede *fuzzy*-ART mostrou uma taxa de acerto superior a apresentada pela rede SOM.

Os resultados obtidos mostraram a viabilidade de aplicação das técnicas de redes neurais e lógica *fuzzy* na fase de agrupamento dentro do processo de mineração de texto. Trata-se de uma abordagem promissora.

Cabe ressaltar que durante o desenvolvimento deste trabalho observou-se que uma das etapas mais complexas do agrupamento de textos é o pré-processamento. Para efetuar esta etapa é necessário um grande conhecimento a respeito da base de textos, o que torna o agrupamento de textos um sistema "semi-supervisionado".

Como trabalho futuro é interessante a investigação dos modelos de agrupamento de neurônios aqui propostos aplicados a mineração de dados e agrupamento de dados em geral. Apesar da particularidade do dado aqui tratado, os modelos podem se adaptar e solucionar outros problemas.

Referências Bibliográficas

[1] Batista, G.E.A.P.A. Pré-processamento de dados em Aprendizado de Máquina Supervisionado. Tese de Doutorado, ICMC-USP, 2003.

[2] Braga, A. P., Ludermir, T. B., Carvalho, A. C. P. L. F. Redes Neurais Artificiais: Teoria e aplicações. Rio de Janeiro: LTC - Livros Técnicos e Científicos Editora S.A, 2000.

[3] Fausset, L.V. Fundamental of Neural Networks Architectures, Algorithms and Applications. New Jersey: Prentice Hall International, 1994.

[4] Grossberg, 1976. Adaptive pattern classification and universal recoding, 1: Parallel development and coding of neural feature detectors. *Biological Cybernetics*, 23:187-202.

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

[6] Lewis, D. D. HTTP://www.davidddlewis.com/resources/testcollections/reuters21587, 2006.

[7] Loh, S. Abordagem Baseada em Conceitos para Descoberta de Conhecimento em textos. PhD thesis, Universidade Federal do Rio Grande do Sul, Instituto de Informática, 2001.

[8] Luhn, H. P. The automatic creation of literature abstracts. IBM Journal os Research and Development, 2(2):159–165, 1958.

[9] MARTINS, Weber; NALINI, Laura Eugênio Guimarães; TSUKAHARA, Fernando Pirkel. Context-sensitive multidimensional ranking: an alternative technique to data complexity. **Rev. Psicol., Organ. Trab.**, Florianópolis, v. 6, n. 1, jun. 2006 . Disponível em <http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1984-66572006000100010&lng=pt&nrm=iso>.

[10] Oliveira, H. M. Seleção de Entes Complexos utilizando lógica difusa. Porto Alegre, Pontifícia Universidade Católica do Rio Grande do Sul, 1996.

[11] Pazzani, M. J.; Muramatsu, J.; Billsus, D.(1996) Syskill Webert: Identifying Interesting Web Sites. In: AAAI/IAAI, VOL. 1, 1996. Anais. [S.l.: s.n.], p.54–61, 1996.

[12] Silva, N. C. Utilização de operadores genéticos para otimizar classificadores neurais não-supervisionados de imagens. Brasília. 200p. Tese de Doutorado em Geociências - Universidade de Brasília, 2002.

[13] Wives, L. K., Técnicas de Descoberta de Conhecimento em Textos Aplicada à Inteligência Competitiva. Programa de pós-graduação em computação (PRGC) Instituto de Informática, Universidade Federal do Rio Grande do Sul – UFRGS, 2002.

[14] Zanasi, A. . *Discovering Data Mining*. *Prentice Hall*, 1997.

[15] Zadeh, L. A. Fuzzy Sets. *Information Control*, 8:338--353, New York,1965.