Proceedings of V Congresso Brasileiro de Redes Neurais - VI Escola de Redes Neurais
pp. 517-522, april 2-5, 2001 - Rio de Janeiro - RJ - Brazil

Uma Proposta para Categorização de Textos por uma Rede Neural

Claudia Brandelero Rizzi[1], João Francisco Valiati[2], Paulo Martins Engel[2]
[1]Instituto de Informática da UFRGS e CCET da UNIOESTE
[2]Instituto de Informática da UFRGS
E-mails : claudiab@inf.ufrgs.br, jvaliati@inf.ufrgs.br, engel@inf.ufrgs.br

Abstract

The objective of this work is to present the results of two experiments in which a Multilayer Perceptron neural network, trained with the Backpropagation algorithm was used for categorization of texts in English (the categorization of texts is the classification of texts with regard to a group of one or more existent categories). The approach of organization of the input data adopted for this purpose is presented. It proved to be efficient in the training process, implying in a reduction of up to 1/3 the necessary time for its conclusion. The obtained results are satisfactory, since that in the first experiment, accomplished with the "IA Collection", 72% of recall and 74% of precision were reached, and in the second one, accomplished with a sub-collection of Reuters-21578, 79% of recall and precision were obtained.

1. Introdução

Há muitos registros, na história da humanidade, sobre o esforço do ser humano na tentativa de relatar e registrar sua vida, suas descobertas, seus pensamentos. Desenhos e símbolos foram sendo aprimorados e com o desenvolvimento da linguagem, culminaram com o advento da escrita, um meio bastante apropriado para efetuar e divulgar esses relatos. Como conseqüência, pequenas e grandes bibliotecas se formaram ao longo dos tempos, e ali foram sendo depositados textos de diferentes temas, povos, línguas e costumes.

Visando facilitar o acesso a estes textos, passou-se a guardá-los seguindo certos critérios. Nas bibliotecas, os manuscritos começaram a ser organizados conforme a temática de que tratavam. Para isso, pessoas os liam e definiam onde seriam colocados, já considerando como poderiam, mais facilmente, serem reavidos.

Se por um lado, a necessidade de guardar e reaver informação ainda é semelhante àquela do passado, por outro, as técnicas e metodologias utilizadas para realizá-las, e em particular as computacionais, têm que estar adaptadas a este novo tempo, em que o volume de informações é muito grande e cresce exponencialmente. Além disso, com o advento da Internet, atuando como um "repositório distribuído" de informações, demandando sistemas com características particulares, essa temática de manter e recuperar informações têm sido objeto de pesquisas. Essas pesquisas visam basicamente aprimorar estas técnicas e permitir que respostas cada vez mais rápidas e precisas sejam oferecidas aos usuários.

Dentre essas pesquisas, estão aquelas que incluem a realização de experimentos na tentativa de contribuir para o aprimoramento e desenvolvimento de técnicas aplicadas a área dos Sistemas de Recuperação de Informações (SRI). É neste contexto que se insere a presente pesquisa.

O objetivo deste trabalho é, a partir da realização de dois experimentos feitos por [5], analisar resultados do desempenho de uma rede neural artificial no processo de categorização de informações textuais. Este texto foi assim estruturado : as seções 2 e 3 introduzem os temas da categorização de textos e das redes neurais utilizadas nos experimentos respectivamente. A seção 4 apresenta os elementos utilizados nos experimentos, desde as bases de dados até o método de avaliação de resultados. As seções 5 e 6 relatam a sistemática de organização dos dados de entrada e sua implicação na aceleração da convergência. As seções 7 e 8 mostram os resultados e a 9 as conclusões e perspectivas para trabalhos futuros.

2. Categorização de Textos

Categorizar textos é a atividade de classificá-los com respeito a um conjunto de uma ou mais categorias pré-existentes. Sua aplicação mais comum é a indexação de documentos para os SRIs [1], mas também são utilizados na categorização de mensagens e notícias, de resumos de publicações, na filtragem e sumarização de textos e outras [2], além de auxiliar profissionais da área a desempenharem estas funções.

O funcionamento básico de um Sistema de Categorização de Textos (SCT) envolve as fases de indexação e resposta. Na primeira fase, os termos de um documento são analisados considerando a existência de categorias pré-definidas, que normalmente são utilizadas por um longo período na existência do SCT. A fase de resposta informa ao usuário se aquele documento processado pertence a nenhuma, uma ou mais daquelas categorias.

A maior dificuldade na categorização de textos, está na grande quantidade de termos que precisam ser processados; tecnicamente conhecida como alta dimensionalidade do espaço de características. Este espaço se constitui de termos únicos ou compostos que são extraídos ou adaptados (por exemplo via thesaurus) a partir dos textos dos documentos processados. O fato é

que podem existir centenas ou milhares destes termos, até mesmo para uma pequena coleção de textos.

Visando otimizar o desempenho computacional, é importante reduzir ao máximo este espaço de características sem sacrificar a identificação dos conteúdos dos textos, pois isso significa alterar o desempenho do processo de categorização. Esta redução pode ser feita considerando que muitos desses termos, como artigos, preposições, advérbios (chamados de *stopwords*), são irrelevantes para a categorização [3] e podem ser suprimidos.

É fator determinante para o bom desempenho de um SCT a definição precisa dos termos que melhor representam os conteúdos dos textos e as categorias consideradas. Esta definição é feita a partir da verificação automática dos assuntos ali tratados, que são formulados através de uma linguagem. Ela envolve análises das frases e/ou dos termos que o texto contém.

Se esta análise é feita aplicando-se métodos semânticos e sintáticos, então é utilizada a abordagem lingüística. Se nesta análise são aplicados métodos que incluem seleção e contagem de termos nos textos, a abordagem é estatística. Neste último caso, o termo é o meio de acesso ao texto, é a maneira pela qual os textos são identificados e diferenciados. Este trabalho está inserido no conjunto daqueles que utilizam a abordagem estatística na análise automática de textos.

Uma vez definidos os termos que representam cada categoria considerada, e consequentemente o domínio a ser tratado, o processo de categorização já pode ser iniciado. Neste trabalho, este processo é realizado por uma rede neural Perceptron Multicamadas.

3 Perceptron Multicamadas

A rede Perceptron Multicamadas (PM) é constituída por uma topologia de múltiplos neurônios divididos em 3 camadas, interconectados para frente. A camada de entrada é responsável pela distribuição dos valores de entrada para os neurônios da camada escondida. A camada escondida, através de um mapeamento intermediário do problema, gera vetores linearmente separáveis, que servem de entrada para a camada de saída, que realiza a rotulação das categorias.

Esta rotulação das categorias ocorre a partir da aprendizagem da rede, decorrente de seu treinamento, onde exemplos e suas respectivas categorias corretas são apresentados. Assim, através de aproximações de funções, ela é capaz de generalizar o conhecimento aprendido, e padrões similares são classificados com o mesmo rótulo [13], efetivando a categorização.

A rede utilizada nos experimentos foi treinada através do algoritmo de retropropagação de erros, o Backpropagation, tendo como função de ativação a tangente hiperbólica. O Backpropagation é um dos algoritmos mais conhecidos e utilizados por pesquisadores que se valem das redes neurais [14], e é encontrado facilmente na literatura da área, como por exemplo em [15] e [16].

4. Ferramentas e Métodos Utilizados nos Experimentos

Visando efetuar a categorização de textos em ambas as bases de dados utilizadas, nesta seção são descritas as ferramentas e métodos empregados. A figura 1 mostra a seqüência com que foram utilizados.

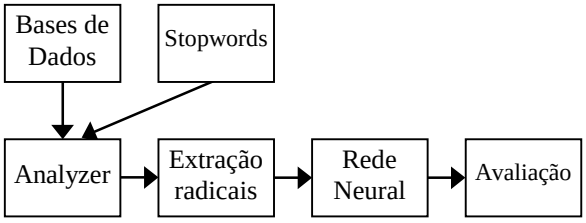

Figura 1 - Seqüência de Processos

As bases de dados utilizadas nos experimentos, foram a Coleção Reuters-21578 [7] e a Coleção IA [5], ambas em língua inglesa.

A Coleção Reuters-21578 contém textos de notícias divulgadas no período entre 1987 e 1991. Esta coleção é composta por 21.578 textos cronologicamente organizados em 5 grupos (*topics, places, people, orgs,* e *exchanges*) e suas respectivas sub-categorias. Esses textos estão em formato SGML (Standard Generalized Markup Language), com diversos marcadores (tags) indicando, início e fim do texto, título, autores, data da publicação, entre outras informações.

A Coleção Inteligência Artificial, chamada "Coleção IA", contém 280 registros separados em 13 sub-categorias. Estes registros são formados por títulos, resumos e palavras-chave, extraídos de publicações da área da IA, ou de áreas que a envolvesse.

A lista de 266 *stopwords* utilizada, foi aquela elaborada pelo Laboratório de Recuperação de Informações da Universidade de Massachusetts em Amherst, obtida no trabalho de [1], e relacionadas em [5]. Esta lista é formada por termos que se repetem muitas vezes nos textos e são pouco significativos [17]. A eliminação desses termos promove uma redução de 40 a 50% dos textos a serem analisados [18].

A ferramenta *Analyzer*, implementada a partir do sistema *Eurekha* [4], foi utilizada para processar os textos, na primeira etapa dos trabalhos. Com ela foi realizada a eliminação de *stopwords* e a seleção das características representantes de cada uma das categorias. Dentre os vários métodos existentes para seleção de características, tais como a Freqüência Relativa [18], Ganho de Informação e Informação Mútua [19], Qui-Quadrado e Semântica Latente [11], o método Escore de Relevância [11], [12] foi o escolhido para ser usado neste trabalho. Ele foi aplicado tanto para o conjunto de textos destinados à aprendizagem quanto aos destinados aos testes da rede. O método Escore de Relevância calcula a relevância relativa de um determinado termo (r_t) pela expressão:

$$r_t = log \frac{A/B + 1/6}{\overline{A}/\overline{B} + 1/6} \quad (1)$$

Onde A é o número de textos da categoria que contém o termo, e B o número total de textos da categoria. $\overline{A}$ e $\overline{B}$ são os mesmos, porém para a presença do termo em outras categorias. Este método proposto por [11], foi adaptado dos estudos de [18] sobre atribuição de pesos a termos. Tanto [18] quanto [11] obtiveram bons resultados com sua aplicação.

O algoritmo utilizado para conversão de termos em seus respectivos radicais (*stemming*), foi o de Martin Porter. Este algoritmo, chamado de algoritmo de Porter, é o mais conhecido e aceito pela comunidade da área [8], [9]. Ele realiza um processo de remoção de letras do final de palavras, que possuem mesma variação morfológica e de flexão. Em [6] há uma descrição completa do algoritmo, bem como suas regras e etapas de processamento.

A rede neural, implementada em linguagem C no ambiente Linux, é treinada com um conjunto de vetores contendo termos representantes de cada categoria a ser aprendida. A forma com que os dados de entrada lhe são apresentados é descrita na próxima seção. Ao término do processo de treinamento, o conjunto de textos que compõem o conjunto de testes são apresentados à rede. Ela o processa e emite como saída, uma lista de valores reais, indicando o "grau de pertinência" daquele vetor, portanto daquele texto que ele representa, em relação às categorias possíveis.

O maior valor real apresentado (que varia entre –1 e 1) necessariamente deve atingir um valor mínimo de 0,5 para que a categorização seja considerada como efetuada com sucesso. Além disso, esse valor indica que o documento pertence à categoria que lhe é correspondente. Assim, a opção pelo limiar 0,5 permite que se tenha no mínimo 75% de segurança para aceitação do rótulo de categoria, atribuído pela rede.

O método de avaliação dos resultados adotado foi aquele proposto por David Lewis [10]. Ele sugere um modelo em que uma “tabela de contingência”, é inicialmente preenchida. A organização desta tabela parte do princípio de que um sistema realiza *n* decisões binárias, sendo que cada uma delas tem ou não, exatamente uma resposta correta. A partir desta tabela, podem ser extraídas importantes medidas de eficácia para sistemas de categorização de textos como *Abrangência*, *Precisão*, *Falha*, *Acurácia* e *Erro* e que são representadas pelas colunas desta tabela.

Feitas essas considerações sobre as ferramentas e métodos utilizados, a próxima seção relata como os dados de entrada da rede neural foram organizados.

5. Sistemática de Organização das Entradas

Selecionadas as características que melhor representam cada uma das categorias das duas coleções, elas foram convertidas em seus respectivos radicais que, seqüencialmente numerados em ordem crescente, foram chamados de *vetor base*.

Um *vetor base* de tamanho c_n representa, em cada uma de suas posições, uma das características selecionadas. Quando um vetor *de treinamento* ou *de teste* é construído, este *vetor base* é consultado para que sejam obtidas as posições correspondentes que representam as características desejadas.

Assim, tanto o *vetor base* quanto os vetores *de treinamento* e *de teste*, são formados por *n* posições. Estas posições já definem o número de neurônios da camada de entrada da rede. Para o caso do experimento com a Coleção IA, *n* é igual a 325; para a Coleção Reuters, *n* é igual a 120.

A organização dos vetores foi estruturada de maneira que as primeiras *x* posições representam as *x* características mais importantes para a categoria 1; as próximas *x* posições do vetor representavam as *x* características mais importantes da categoria 2 e assim sucessivamente. A figura 2 apresenta o formato de um vetor da Coleção IA, que compreende um total de 13 categorias e 325 termos. Os elementos *+1* e *0*, inseridos no vetor, indicam a presença ou ausência do termo correspondente indicado pelos números seqüenciais marcados abaixo do vetor, e fornecidos pelo *vetor base*.

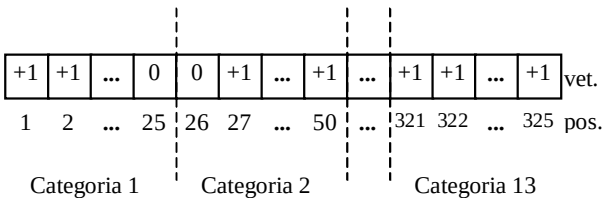

Figura 2 - Formato de um Vetor da Coleção IA

Assim, por exemplo, na primeira parte do vetor que representa a categoria 1, o elemento +1 na primeira e segunda posições, indicam que os radicais dos dois primeiros termos do *vetor base* foram encontrados no corpo do texto que está sendo codificado. Os números *1* e *2* representam ao mesmo tempo a posição e o código do termo. O radical do 25º termo representado pela posição *25*, não foi encontrado no corpo deste texto. Por este motivo, o elemento *0* está naquela posição. Este processo é válido para as demais categorias, elementos e posições, e é o mesmo para a coleção Reuters.

Concluída a etapa de construção dos vetores *de treinamento* e *de testes*, tomando como referência o *vetor base* da Coleção, a rede foi treinada e testada.

6. Aceleração da Convergência

A aplicação da sistemática de organização dos dados, descrita na seção anterior, promoveu uma significativa aceleração de convergência, durante o processo de treinamento.

Testes foram realizados em ambas as coleções, aplicando-se esta sistemática, e também outra em que as

características foram organizadas em ordem alfabética. Os pesos aleatórios iniciais da rede, gerados uma única vez, foram utilizados em todos os testes.

A tabela 1 mostra a taxa de aprendizado e a quantidade de épocas necessárias para treinamento da rede, com e sem a adoção desta sistemática, em 3 testes com uma sub-coleção da Reuters-21578. Ilustrada pela figura 3, ela mostra o paralelo da quantidade de épocas necessárias para alcançar a taxa de erro de 0,001 nesses 3 testes, variando-se apenas a taxa de aprendizado.

Taxa de aprendizado	Nº. de Épocas c/sistemática	Nº. de Épocas s/sistemática
0,009	196	892
0,008	221	923
0,007	253	970

Tabela 1 - Taxa de Aprendizado e Nº de Épocas

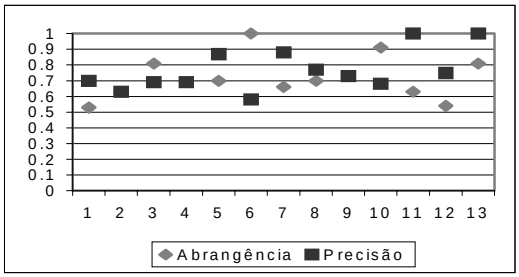

Figura 3 - Gráfico Aceleração de Convergência

7. Resultados - Coleção IA

A rede que processou a coleção IA foi constituída por 325 neurônios na camada de entrada, representando as 25 características de cada uma das 13 categorias. Na camada de saída foram alocados 13 neurônios que representavam estas 13 categorias possíveis. Na camada escondida foram utilizados 65 neurônios, número obtido após a realização de diversos testes.

Foram utilizados 10 vetores de treinamento para cada categoria e 150 vetores para teste. Utilizou-se a taxa de aprendizagem de 0,003, demandando 874 épocas para que o processo fosse concluído.

A rede informa os resultados do reconhecimento em forma de 13 valores reais. O *reconhecimento* ocorre somente quando pelo um dos valores é maior ou igual a 0,5. O *não reconhecimento* ocorre quando nenhum dos valores apresentados por ela atinge este limiar.

A tabela 2 mostra a contingência da Coleção IA. De forma idêntica para a tabela 4, a coluna "*a*" mostra a quantidade de documentos que foram corretamente atribuídos a cada categoria. A coluna "*b*" mostra a quantidade de documentos incorretamente atribuídos a elas. A coluna "*c*" mostra a quantidade de documentos incorretamente rejeitados para as categorias. A coluna "*d*" os documentos corretamente rejeitados para elas.

Contingência – Coleção IA				
Nome da Categoria	**a**	**b**	**c**	**d**
C01: Knowledge R.Reasoning	7	3	6	134
C02: Machine Learning	7	4	4	135
C03: Logic Programming	9	4	2	135
C04: Expert Systems	9	4	4	133
C05: Natural Language	7	1	3	139
C06: Multi-Agents Systems	10	7	0	140
C07: Intelligent Tutoring	8	1	4	137
C08: Genetic Algorithms	7	2	3	138
C09: Neural Networks	11	4	4	131
C10: Fuzzy Logic	11	5	1	133
C11: Comp.Vision /Robotics	7	0	4	139
C12: Digital Signal Processing	6	2	5	137
C13: Sound /Music Computing	9	0	2	139
Totais	108	37	42	
Abrangência/ Precisão	**72%**	**74%**		

Tabela 2 - Contingência da Coleção IA

A Abrangência e Precisão gerais obtidas para esta coleção, foram 72% e 74% respectivamente. A partir da tabela 2, a tabela 3 foi elaborada, e informa, para cada uma das 13 categorias, cinco medidas de eficácia (abrangência, precisão, falha, acurácia e erro).

Cat.	a/(a+c) **abr.**	a/(a+b) **prec.**	b/(b+d) **fal.**	(a+d)/n **acur.**	(b+c)/n **err.**
C01	0,53	0,70	0,021	0,94	0,060
C02	0,63	0,63	0,028	0,94	0,053
C03	0,81	0,69	0,028	0,96	0,040
C04	0,69	0,69	0,029	0,94	0,053
C05	0,70	0,87	0,007	0,97	0,026
C06	1,00	0,58	0,047	1,00	0,046
C07	0,66	0,88	0,007	0,96	0,033
C08	0,70	0,77	0,014	0,96	0,033
C09	0,73	0,73	0,029	0,94	0,053
C10	0,91	0,68	0,036	0,96	0,040
C11	0,63	1,00	0,000	0,97	0,026
C12	0,54	0,75	0,014	0,95	0,046
C13	0,81	1,00	0,000	0,98	0,013

Tabela 3 - Eficácia por Categoria - Coleção IA

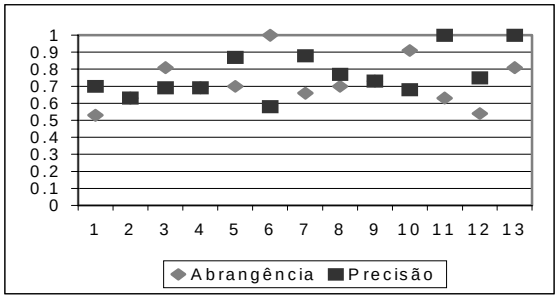

Figura 4 - Abrangência e Precisão - Coleção IA

O gráfico que representa a abrangência e precisão alcançadas em cada uma das treze categorias é ilustrado pela figura 4. Pela tabela 3, pode-se observar que a acurácia obtida neste experimento foi significativa, atingindo percentuais iguais ou superiores a 94%. A relação falha e erro alcançadas, também foram satisfatórias.

8. Resultados - Coleção Reuters-21578

Os textos extraídos da Coleção Reuters-21578 foram aqueles cujo tema estava relacionado com metais, mais especificamente, com alumínio, cobre, ouro, ferro, prata e zinco. Ao todo, 6 categorias e 312 textos foram selecionadas para a realização do experimento. Este conjunto foi chamado de "Sub-Coleção Metais".

A rede que processou esta sub-coleção foi organizada com 120 neurônios na camada de entrada, representando as 20 características mais importantes das 6 categorias. Na camada de saída 6 neurônios representavam essas categorias. Na escondida foram colocados 45 neurônios (definidos depois de testes).

Dos 312 textos disponíveis, foram utilizados 107 para treinamento (além de vetores heurísticos) e 205 para testes. A taxa de aprendizagem foi de 0,009, demandando 129 épocas para a conclusão do processo. O limiar de 0,5 foi utilizado neste experimento. A tabela 4 mostra a Contingência da Sub-Coleção Metais.

Contingência – Sub-Coleção Metais				
Categoria	**a**	**b**	**c**	**d**
C1: Alumínio	23	3	10	169
C2: Cobre	40	6	4	155
C3: Ouro	57	14	18	116
C4: Ferro	28	8	5	164
C5: Prata	4	5	6	190
C6: Zinco	10	7	0	188
Totais	162	43	43	
Abrangência/Precisão	**79%**	**79%**		

Tabela 4 - Contingência da Sub-Coleção Metais

Cat.	a/(a+c) **abr.**	a/(a+b) **prec.**	b/(b+d) **fal.**	(a+d)/n **acur.**	(b+c)/n **err.**
C1	0,69	0,88	0,017	0,93	0,063
C2	0,90	0,86	0,037	0,95	0,048
C3	0,76	0,80	0,107	0,84	0,156
C4	0,84	0,77	0,046	0,93	0,063
C5	0,40	0,44	0,025	0,94	0,053
C6	1,00	0,58	0,035	0,96	0,034

Tabela 5 - Eficácia por Categoria - Sub-Coleção Metais

A abrangência e precisão gerais obtidas para esta sub-coleção, foram ambas de 79%. A partir da tabela 4, foi elaborada a tabela 5 que informa outras medidas de eficácia para as 6 categorias consideradas. A figura 5 mostra o gráfico dos índices de abrangência e precisão obtidos em cada uma das categorias.

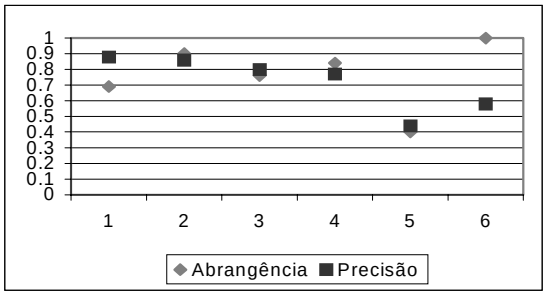

Figura 5 - Abrangência e Precisão Sub-Coleção Metais

Também neste experimento os índices de acurácia obtidos foram interessantes, tendo em vista que, com excessão da categoria 3 (ouro), todas as demais alcançaram índices iguais ou superiores a 93%.

Há algumas semelhanças entre este experimento e aquele feito por [12], em que um pacote de software desenvolvido na Carnegie Mellon University foi utilizado para categorizar textos da Coleção Reuters, relacionados com metais. O trabalho de [12] originou-se a partir dos experimentos feitos por [11], onde uma rede neural foi treinada para cada uma das categorias consideradas. Assim, a pesquisa de [12] visava determinar se uma única rede neural de 3 camadas, treinada com o algoritmo Backpropagation poderia ser utilizada para categorizar documentos de várias categorias.

Mesmo considerando as diferenças entre aqueles dois experimentos e este, tais como base de dados utilizada (naqueles Reuters-22173 que deu origem a Reuters-21578 utilizada neste), quantidade de textos para treinamento e teste, e número de categorias consideradas, a título de simples comparação de desempenho, a tabela 6 mostra os resultados obtidos em termos de abrangência e precisão.

Experimento	**Abrangência**	**Precisão**
Descrito em [11]	67%	75%
Descrito em [12]	78%	79%
Este	79%	79%

Tabela 6 - Abrangência e Precisão em Experimentos Semelhantes

9. Conclusões e Trabalhos Futuros

Este trabalho apresentou dois experimentos cujas aplicações referem-se à categorização de textos, nos quais foram utilizadas duas redes neurais PM, treinadas com o algoritmo Backpropagation. Foram apresentados também os resultados obtidos pela adoção de uma sistemática de organização das entradas da rede.

O primeiro experimento, realizado com a Coleção IA, categorizou 13 sub-áreas da Inteligência Artificial.

Os resultados obtidos foram considerados satisfatórios tendo em vista o alcance de 72% de abrangência e 74% de precisão. O segundo experimento, realizado com uma sub-coleção da Reuters-21578, categorizou 6 áreas e alcançou 79% de abrangência e precisão. Além desses resultados também as medidas de falha, erro e acurácia em ambos os casos atingiram percentuais significativos.

Uma aplicação prática da categorização de textos, a título de ilustração, por exemplo, é o caso de um sistema de roteamento (routing) em que 2 das 6 áreas da sub-coleção metais fossem selecionadas e atribuídas a um funcionário de uma empresa, responsável por elas. **E**le passaria a receber somente as informações daquelas categorias, com um índice de acerto de 79%. Isso pode significar uma grande redução na quantidade de informações que esse funcionário recebe. Mesmo com um índice de erro de 21%, as informações recebidas tornam-se mais específicas, facilitando sua leitura e análise, contribuindo para evitar o excesso de informação. Se ao contrário, não houvesse um sistema como esse que considerasse as categorias de interesse, muito provavelmente este mesmo funcionário não conseguiria triar e analisar as informações importantes de sua responsabilidade.

Em uma análise mais geral sobre os experimentos, os resultados mostram que a escolha da rede Perceptron Multicamadas bem como do algoritmo Backprogation foi satisfatória para a aplicação de categorização de textos. A adoção da sistemática de organização dos dados de entrada, mostrou-se eficiente na aceleração do processo de aprendizagem da rede. Com sua aplicação, a convergência foi na média, 3,5 vezes mais rápida em se comparando o mesmo processo, sem utilizá-la.

Porém, o método de avaliação dos resultados utilizado, que é binário, não mostrou-se adequado para os casos em que mais de um valor real superior a 0,5 é emitido pela rede. Isto significa que a rede detectou que o texto sendo categorizado pertence simultaneamente, a mais de uma das categorias pré-existentes. Ao adotar este método, foi preciso nesses casos, considerar o maior valor emitido pela rede como o representante da categoria correta, descartando-se os demais. Porém, em se tratando de textos cujos temas realmente podem ser relevantes ao mesmo tempo para mais de uma categoria, parece mais apropriado informar esta característica ao usuário, e deixar que ele resolva como lidar com esta questão. Sendo assim, seria mais interessante utilizar um método de avaliação de resultados não binário, que tratasse esses casos de maneira mais apropriada.

A partir dos resultados obtidos neste trabalho, algumas questões de pesquisa foram elencadas e deram origem a necessidade de realizar outros experimentos.

Pretende-se de imediato: a) definir os termos a serem utilizados nos experimentos a partir de dois parâmetros que são 1) a quantidade máxima de termos por categoria e 2) um escore mínimo obtido pelos termos, conforme sugerido pelo método Escore de Relevância. Além disso, aqueles termos que obtiverem um escore maior do que um limite estabelecido, serão representados por duas posições no *vetor base*, dada a sua importância; e b) a extração de radicais será efetuada em toda a coleção, antes da aplicação do método Escore de Relevância.

Referências

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

[2] P. Geutner; U. Bodenhausen; A.Waibel. *Neural Networks for Text Categorization.* Disponível em : http://werner.ira.uka.de/~pgeutner/publications.html .

[3] I. Moulinier; G. Raskinins; J. Ganascia. *Text Categorization: A Symbolic Approach.* Disponível em : http://www-poleia.lip6.fr/~moulinie/.

[4] L. Wives. *Um estudo sobre Agrupamento de Documentos Textuais em Processamento de Informações não Estruturadas Usando Técnicas de Clustering.* Porto Alegre: PPGC da UFRGS, 1999. Dissertação de Mestrado.

[5] C. B. Rizzi. *Categorização de Textos por Rede Neural - Estudo de Caso.* Porto Alegre: PPGC da UFRGS, 2000. Dissertação de Mestrado ainda não publicada.

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

[7] D. D. Lewis. *Reuters-21578 Text Categorization Test Collection*. AT&T Labs Research. Disponível em : http://www.research.att.com/~lewis .

[8] B. Croft; J. Xu. *Corpus-Based Stemming using Co-occurrence of Word Variants*. Disponível em : http://www.acm.org/pubs/citations/journals/tois/1998-16-1/p61-xu .

[9] G. Kowalski. *Information Retrieval Systems : Theory and Implementation*. Boston: Kluwer Academic Publishers, 1997.

[10] D. D. Lewis. *Evaluating Text Categorization.* Disponível em http://www.research.att.com/~lewis/.

[11] E. Wiener; J. Pedersen; A. Weigend. *A Neural Network Approach to Topic Spotting*. Disponível em : http://cora.jprc.com/Information_Retrieval/Retrieval/index.html.

[12] B. Norman; E. Yu; E. Liddy. *A Learning Subject Field Coder*. Disponível em http://www.npac.syr.edu/REU/reu96/project_reu.html .

[13] P.M. Engel. *Redes Neurais*. Porto Alegre : PPGC da UFRGS, 1999. Notas de aula.

[14] Z. Kovács. *Redes Neurais Artificiais*. São Paulo: Collegium Cognitio. 2ª Edição, 1996.

[15] J. A. Freeman; D.M. Skapura. *Neural Networks, Algorithms, Applications, and Programming Techniques.* Massachusetts: Addison-Wesley, 1991.

[16] S. Haykin. *Neural Netwoks - A Comprehensive Foundation*. New York: Macmillan College Publishing Company, 1994.

[17] C. Rijsbergen. Information Retrieval. Disponível em : http://130.209.240.1/mira/cv/vanRijsbergen.html.

[18] G. Salston. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.

[19] Y. Yang; J. O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. Disponível em : http://www.cs.cmu.edu/~yiming/publications.html.