Redes neurais e sua aplicação em sistemas de recuperação de informação

Edberto Ferneda
Professor doutor do curso de ciências da informação e Documentação da Faculdade de Filosofia Ciências e Letras de Ribeirão Preto – USP
E-mail: ferneda@ffclrp.usp.br

Resumo

Redes neurais constituem um campo da ciência da computação ligado à inteligência artificial, buscando implementar modelos matemáticos que se assemelhem às estruturas neurais biológicas. Nesse sentido, apresentam capacidade de adaptar os seus parâmetros como resultado da interação com o meio externo, melhorando gradativamente o seu desempenho na solução de um determinado problema. A utilização de redes neurais em sistemas computacionais de recuperação de informação permite atribuir um caráter dinâmico a tais sistemas, dado que as representações dos documentos podem ser reavaliadas e alteradas de acordo com a especificação de relevância atribuída pelos usuários aos documentos recuperados. O presente trabalho apresenta as principais iniciativas de se aplicarem os conceitos de redes neurais aos sistemas de recuperação de informação e avalia sua aplicabilidade em grandes bases documentais, como é o caso da Web.

Palavras-chave

Redes neurais. Recuperação de informação. Sistemas adaptativos.

Neural networks and its application in information retrieval systems

Abstract

Neural networks are a field of Computer Science related to Artificial Intelligence. The field aims at implementing mathematical models that are similar to biological neural structures. It is also capable of adapting its parameters as a result of interactions with the external environment, gradually improving their performance in the solution of a particular task. By using neural networks in computer information retrieval systems one can assign a dynamic character to those systems by allowing the representation of documents to be reevaluated and modified according to specifications of relevance attributed by users to retrieved documents. This work presents the main initiatives in applying neural networks concepts to information retrieval systems and evaluates its applicability to large document databases as is the case of the Web.

Keywords

Neural networks. Information retrieval. Adaptive systems.

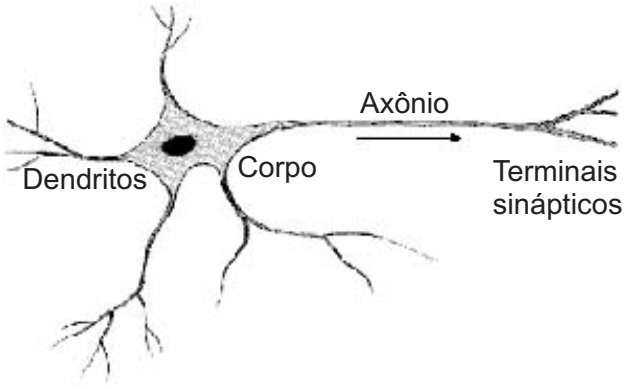
INTRODUÇÃO

Sabe-se que o cérebro é composto de bilhões de neurônios. Um neurônio é uma célula formada por três seções com funções específicas e complementares: *corpo*, *dendritos* e *axônio*. Os dendritos captam os estímulos recebidos em um determinado período de tempo e os transmitem ao corpo do neurônio, onde são processados. Quando tais estímulos atingirem determinado limite, o corpo da célula envia novo impulso que se propaga pelo axônio e é transmitido às células vizinhas por meio de sinapses. Este processo pode se repetir em várias camadas de neurônios. Como resultado, a informação de entrada é processada, podendo levar o cérebro a comandar reações físicas. A figura 1 ilustra de forma simplificada as partes de um neurônio.

A habilidade de um ser humano em realizar funções complexas e principalmente a sua capacidade de aprender advêm do processamento paralelo e distribuído da rede de neurônios do cérebro. Os neurônios do córtex, a camada externa do cérebro, são responsáveis pelo processamento cognitivo. Um novo conhecimento ou uma nova experiência pode levar a alterações estruturais no cérebro. Tais alterações são efetivadas por meio de um rearranjo das redes de neurônios, reforçando ou inibindo algumas sinapses (HAYKIN, 2001, p.32-36).

A busca por um modelo computacional que simule o funcionamento das células do cérebro data dos anos 40,

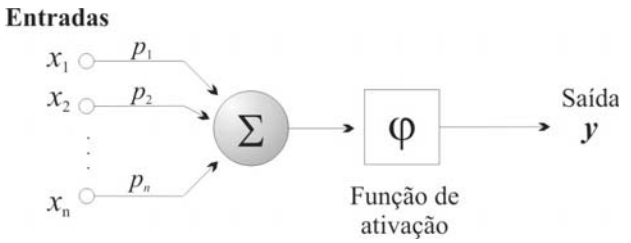
FIGURA 1
Representação simplificada de um neurônio



com o trabalho de McCulloch e Pitts (1943). O entusiasmo pela pesquisa neste campo cresceu durante os anos 50 e 60. Nesse período, Rosenblatt (1958) propôs um método inovador de aprendizagem para as redes neurais artificiais denominado *percepton*. Até 1969, muitos trabalhos foram realizados utilizando o *percepton* como modelo. No final dos anos 60, Minsky e Pappert (1969) publicam um livro no qual apresentam importantes limitações do *perceptron*. As dificuldades metodológicas e tecnológicas, juntamente com os ataques extremamente pessimistas de Papert e Minsky, fizeram com que as pesquisas arrefecessem nos anos seguintes. Durante os anos 70, a pesquisa contava apenas com um número ínfimo de cientistas. Porém, durante os anos 80, o entusiasmo ressurge graças a avanços metodológicos importantes e ao aumento dos recursos computacionais disponíveis.

O modelo de neurônio artificial da figura 2 é uma simplificação do modelo apresentado por Haykin (2001, p. 36).

FIGURA 2
Modelo matemático de um neurônio



Este modelo é composto por três elementos básicos:

• um conjunto de *n* conexões de entrada (x_1, x_2, ..., x_n), caracterizadas por pesos (p_1, p_2, ..., p_n);

• um somador (Σ) para acumular os sinais de entrada;

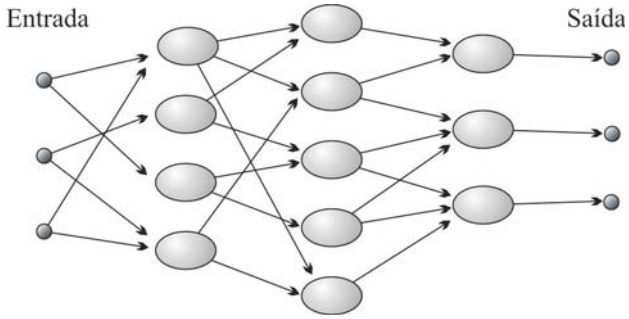
• uma função de ativação (φ) que limita o intervalo permissível de amplitude do sinal de saída (*y*) a um valor fixo.

O comportamento das conexões entre os neurônios é simulado por meio de seus pesos. Os valores de tais pesos podem ser negativos ou positivos, dependendo de as conexões serem inibitórias ou excitatórias. O efeito de um sinal proveniente de um outro neurônio é determinado pela multiplicação do valor (intensidade) do sinal recebido pelo peso da conexão correspondente ($x_i \times p_i$). É efetuada a soma dos valores $x_i \times p_i$ de todas as conexões, e o valor resultante é enviado para a *função de ativação*, que define a saída (**y**) do neurônio.

Combinando diversos neurônios, forma-se uma rede neural artificial. As redes neurais artificiais são modelos que buscam simular o processamento de informação do cérebro humano. São compostas por unidades de processamentos simples, os neurônios, que se unem por meio de conexões sinápticas.

De uma forma simplificada, uma rede neural artificial pode ser vista como um grafo onde os nós são os neurônios e as ligações fazem a função das sinapses, como exemplificado na figura 3.

FIGURA 3
Representação simplificada de uma rede neural artificial



As redes neurais artificiais se diferenciam pela sua arquitetura e pela forma como os pesos associados às conexões são ajustados durante o processo de aprendizado. A arquitetura de uma rede neural restringe o tipo de problema no qual a rede poderá ser utilizada, e é definida pelo número de camadas (*camada única* ou *múltiplas camadas*), pelo número de nós em cada camada, pelo tipo de conexão entre os nós (*feedforward* ou *feedback*) e por sua topologia (HAYKIN, 2001, p. 46-49).

Uma das propriedades mais importantes de uma rede neural artificial é a capacidade de aprender por intermédio de exemplos e fazer inferências sobre o que aprendeu, melhorando gradativamente o seu desempenho. As redes neurais utilizam um *algoritmo de aprendizagem* cuja tarefa é ajustar os pesos de suas conexões (BRAGA; CARVALHO; LUDEMIR, 2000, cap. 2).

Existem duas formas básicas de aprendizado de redes neurais: *aprendizado supervisionado* e *aprendizado não-supervisionado*. No aprendizado supervisionado, um agente externo (professor) apresenta à rede neural alguns conjuntos de padrões de entrada e seus correspondentes padrões de saída. Portanto, é necessário ter um

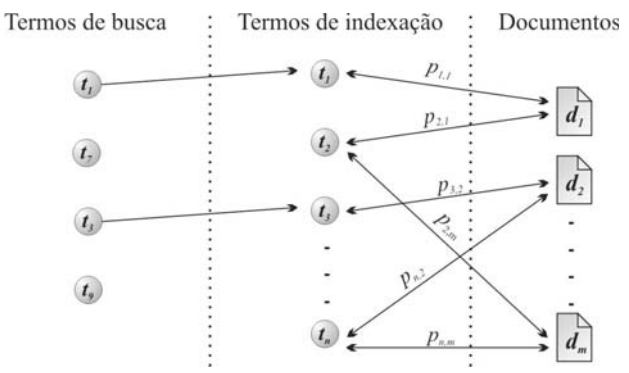
conhecimento prévio do comportamento que se deseja ou se espera da rede. Para cada entrada, o professor indica explicitamente se a resposta calculada é boa ou ruim. A resposta fornecida pela rede neural é comparada à resposta esperada. O erro verificado é informado à rede para que sejam feitos ajustes a fim de melhorar suas futuras respostas.

Na aprendizagem não supervisionada, ou aprendizado auto-supervisionado, não existe um agente externo para acompanhar o processo de aprendizado. Neste tipo de aprendizagem, somente os padrões de entrada estão disponíveis para a rede neural. A rede processa as entradas e, detectando suas regularidades, tenta progressivamente estabelecer representações internas para codificar características e classificá-las automaticamente. Este tipo de aprendizado só é possível quando existe redundância nos dados de entrada, para que se consiga encontrar padrões em tais dados.

REDES NEURAIS NA RECUPERAÇÃO DE INFORMAÇÃO

De uma forma simplificada, a recuperação de informação lida com documentos, termos de indexação e as expressões de buscas dos usuários. Pode-se dizer que, em um sistema de recuperação de informação, de um lado estão as expressões de busca, do outro lado estão os documentos e no centro estão os termos de indexação. Essa estrutura pode ser vista como uma rede neural de três camadas: a camada de termos de busca seria a camada de entrada da rede neural, a camada de documentos seria a saída, e a camada de termos de indexação seria uma camada central. A figura 4 mostra um exemplo genérico da aplicação das redes neurais na recuperação de informação.

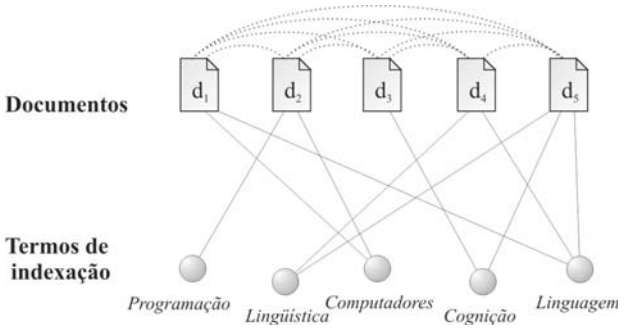
FIGURA 4
Representação de rede neural aplicada à recuperação de informação



Os termos de busca (t_1, t_7, t_3, t_9) iniciam o processo de inferência mediante a ativação dos respectivos termos de indexação. Alguns termos da expressão de busca podem não fazer parte do conjunto de termos de indexação, como é o caso do termo t_7 e t_9. Nesse caso, esses termos não ativarão nenhum termo de indexação e, portanto, não serão considerados durante o processo ativação da rede. Os termos de indexação ativados pelos de termos de busca enviam sinais para os documentos. Estes sinais serão multiplicados pelos pesos de cada ligação ($p_{1,1}, p_{1,2}, ..., p_{n,m}$). Os documentos ativados enviam sinais que são conduzidos de volta aos termos de indexação. Ao receberem estes estímulos, os termos de indexação enviam novos sinais aos documentos, repetindo o processo. Os sinais tornam-se mais fracos a cada iteração, e o processo de propagação eventualmente pára. O resultado final de uma busca será um conjunto dos documentos que foram ativados, cada qual com um nível ativação que pode ser interpretado como o grau de relevância do documento em relação à busca do usuário. Entre os documentos resultantes, podem aparecer alguns que não estão diretamente relacionados aos termos utilizados na expressão de busca, mas que foram inferidos durante a pesquisa e possuem certo grau de relacionamento com a necessidade de informação do usuário. A ativação do termo de indexação t_1, por exemplo, ativou a conexão com o documento d_1. O documento d_1, por sua vez, ativou o termo t_2, que não fazia parte do conjunto de termos de busca. O termo t_2 poderá ativar o documento d_m, que, dependendo do seu grau de ativação, pode vir a fazer parte do conjunto de documentos recuperados.

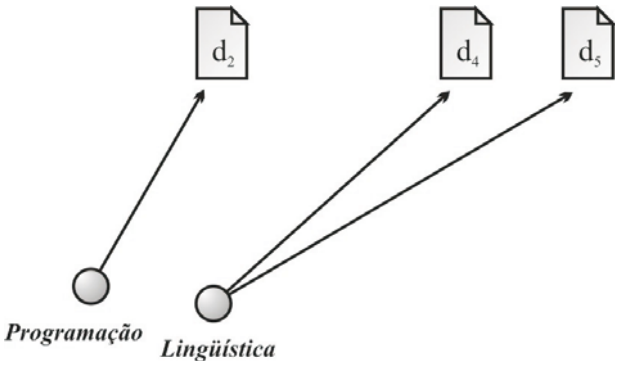
Mozer (1984) foi o pioneiro na utilização de técnicas de redes neurais na recuperação de informação. Ele utilizou uma arquitetura bastante simples que não empregava uma das principais características das redes neurais, que é a capacidade de aprender. A figura 5 mostra um exemplo apresentado por Ford (1991, p.108), que utiliza a arquitetura de rede neural idealizada por Mozer.

FIGURA 5
Exemplo de rede neural utilizando arquitetura de Mozer

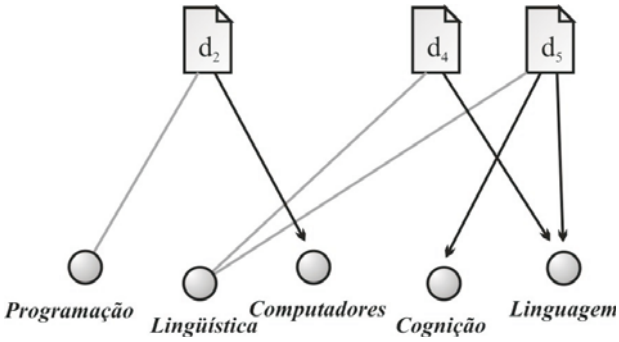


As linhas contínuas representam ligações *excitatórias* entre os termos de indexação e os documentos. As linhas pontilhadas, que ligam pares de documentos, representam ligações *inibitórias*, isto é, ligações que reduzem a força de associação entre os nós. Os termos de indexação ativam os documentos que são indexados por eles e vice-versa. Um documento, ao ser ativado, reduz o nível de ativação dos demais documentos.

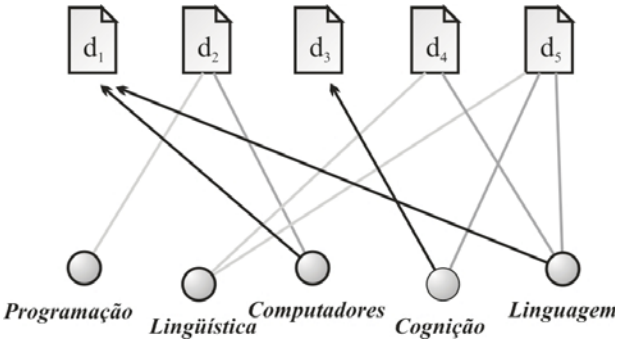
Utilizando uma expressão de busca que contém os termos "programação" e "lingüística", por exemplo, a rede neural da figura 5 apresentará a seguinte seqüência de ativação:



1. Inicialmente serão ativados os nós correspondentes aos termos de busca ("*programação*" e "*lingüística*"). O termo "*programação*" irá ativar o documento $\mathbf{d}_2$, e o termo "*lingüística*" ativará os documentos $\mathbf{d}_4$ e $\mathbf{d}_5$:



2. O documento $\mathbf{d}_2$ ativará todos os termos de indexação usados para indexá-lo: "*programação*" e "*computadores*". Assim, o termo "*programação*" é reforçado, e o termo "*computadores*" é ativado pela primeira vez. Os documentos $\mathbf{d}_4$ e $\mathbf{d}_5$ ativarão o termo "*linguagem*" e reforçarão a ativação do termo "*lingüística*". O documento $\mathbf{d}_5$ ativará também o termo "*cognição*":



3. O termo "*computadores*" ativará os documentos indexados por ele. Assim, o documento $\mathbf{d}_2$ é reforçado, e o documento $\mathbf{d}_1$ é ativado. O termo "*linguagem*" reforçará a ativação dos documentos $\mathbf{d}_4$ e $\mathbf{d}_5$ e também o documento $\mathbf{d}_1$. O termo "*cognição*" ativará o documento $\mathbf{d}_3$.

Este processo se propagará até a estabilização da rede neural, quando cessam as ativações entre seus nós. O nível de ativação de cada documento representará a sua relevância em relação à expressão de busca. No exemplo, os documentos $\mathbf{d}_2$, $\mathbf{d}_4$ e $\mathbf{d}_5$, que foram ativados diretamente pelos termos de busca, terão um nível de ativação maior do que o documento $\mathbf{d}_3$, que é indexado por apenas um termo ("*cognição*") e que foi indiretamente ativado durante a busca.

Para que sejam apresentados resultados satisfatórios, os parâmetros da rede neural (pesos das conexões, funções de ativação etc.) devem ser configurados de forma precisa. Porém, o sistema pode compensar algumas inconsistências na indexação e até possíveis imprecisões nas expressões de busca dos usuários. Mozer (1984) enfatiza que a grande vantagem deste modelo é a habilidade em produzir resultados não esperados, recuperando documentos que não possuem nenhum termo em comum com a expressão de busca, mas, mesmo assim, podem vir a ser relevantes para o usuário. No exemplo apresentado, em resposta à expressão de busca contendo os termos "*programação*" e "*lingüística*". O documento $\mathbf{d}_1$, que é indexado pelos termos "*computadores*" e "*linguagem*", obteve também certo nível de ativação (FORD, 1991, p.109).

Na arquitetura proposta por Mozer, as ligações entre os documentos são inibitórias, isto é, um documento, quando ativado, reduz o nível de ativação dos demais documentos. Isso causa uma competição entre os documentos, fazendo com que apenas os documentos mais ativados durante o processo de busca sejam efetivamente recuperados, reduzindo assim o número de documentos resultantes.

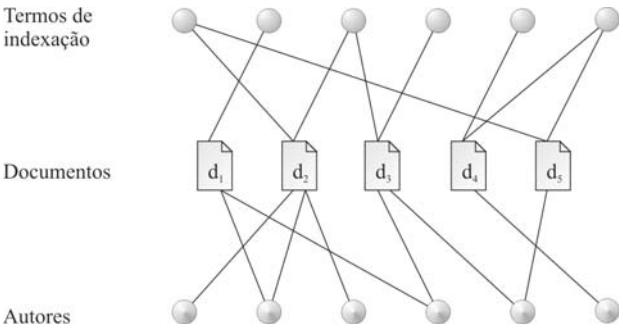
Ao final do processo de pesquisa, o grau de ativação de cada documento pode ser utilizado como critério de ordenamento dos itens resultantes. Os documentos com maior nível de ativação são geralmente aqueles que possuem todos os termos utilizados na expressão de busca, seguidos dos documentos que possuem somente alguns dos termos de busca e dos que foram apenas inferidos durante o processo de pesquisa.

Bein e Smolensky (1988) implementaram e testaram esse modelo de rede neural proposta por Mozer utilizando 12.990 documentos e 6.832 termos de indexação. Eles avaliaram os resultados apresentados como satisfatórios e sugeriram novos testes utilizando bases de dados maiores e com características diversas.

Como foi observado anteriormente, o modelo proposto por Mozer não utiliza uma das características mais fortes das redes neurais: a habilidade de aprender por meio da alteração dos pesos associados às ligações entre os nós. Um sistema mais recente que explora tal habilidade das redes neurais é o sistema AIR.

Desenvolvido por Belew (1989), o sistema Adaptative Information Retrieval (AIR) utiliza uma arquitetura de rede neural composta de três camadas que representam os termos de indexação, os documentos e os seus autores. As ligações são feitas entre os documentos e seus autores e entre documentos e seus termos de indexação, como apresentado na figura 6.

FIGURA 6
Arquitetura de rede neural do sistema AIR



Uma busca pode ser expressa não apenas pela ativação dos termos de indexação, mas por qualquer tipo de nó (autor documento ou termo de indexação), ou por alguma combinação deles. Durante a pesquisa, é feita a ativação dos nós da rede e, quando o sistema se estabiliza, os nós e as ligações que foram inferidos são apresentados ao usuário. O sistema AIR fornece uma interface apropriada para que o usuário possa atribuir um grau de relevância para cada um dos itens recuperados. Este *feedback* é utilizado na aprendizagem da rede neural, que modifica os pesos associados às conexões entre seus nós, buscando adaptar a rede às necessidades de informação dos usuários.

Portanto, o sistema AIR (BELEW, 1989) permite uma participação ativa de seus usuários por meio da atribuição de relevância aos resultados obtidos em suas buscas. Essa interação é responsável por mudanças estruturais na rede neural artificial utilizada pelo sistema e pode ser vista como um processo contínuo de aprendizagem e adaptação do sistema aos interesses de seus usuários, resultando presumivelmente em melhoria progressiva de seu desempenho.

Essa adaptabilidade, porém, permite inferir que este tipo de aplicação das redes neurais só é possível em ambientes nos quais os usuários possuam interesses comuns, para que seja possível ao sistema convergir para um desempenho ótimo. Assim, em um ambiente tipicamente heterogêneo como a Web, esta forma de aplicação das redes neurais só seria viável em sistemas com domínio bem específico e restrito, como, por exemplo, uma biblioteca digital especializada ou em sistemas de recuperação de informação ligados a grupos de pesquisa em determinadas áreas do conhecimento.

Portanto, as idéias apresentadas por Mozer (1984) e por Belew (1989) sofrem algumas restrições quando inseridas no ambiente web. No entanto, existem outras formas de aplicação desenvolvidas especificamente para o ambiente web no qual tais restrições inexistem, como será exemplificado na próxima seção.

REDES NEURAIS NO AMBIENTE WEB

A recuperação de informação no complexo ambiente da Internet é relativamente facilitada pelos mecanismos de busca (*search engines*), que coletam e indexam uma parte da imensa quantidade de páginas disponíveis na Web. Para facilitar a seleção dos itens recuperados, a maioria dos mecanismos de busca realiza um ordenamento dos resultados, utilizando algum algoritmo que tenta predizer a relevância de cada item para a necessidade de informação do usuário. As primeiras referências são presumivelmente mais relevantes do que as últimas.

Cada mecanismo de busca utiliza seu próprio algoritmo para a coleta e indexação de páginas. Como decorrência desta diversidade, para uma mesma expressão de busca, os resultados apresentados pelos diferentes mecanismos podem variar consideravelmente. Pode-se supor, então,

que a combinação de vários mecanismos de busca pode aumentar a área de cobertura da Web e, conseqüentemente, permitir obter resultados mais completos do que um mecanismo de busca tomado isoladamente. Esta combinação de vários mecanismos de busca é denominada *metabuscador* (*metasearch engine*). Um metabuscador obtém os resultados de diferentes mecanismos de busca e, após retirar as referências repetidas, apresenta as páginas ordenadas e em uma interface adequada.

Os metabuscadores ampliam consideravelmente a abrangência das buscas na Web. Porém, como se apóiam nos recursos oferecidos por um conjunto de mecanismos de busca, os metabuscadores herdam deles todas as suas limitações.

Com o objetivo de melhorar, na Web, a precisão das buscas, Shu e Kak (1999) implementaram um metabuscador que se apóia em quatro mecanismos de busca: Yahoo, Excite, Infoseek e WebCrawler. Após a execução de uma busca, os resultados de cada um desses mecanismos são ordenados utilizando um algoritmo de classificação baseado em uma rede neural. Este algoritmo é o principal componente do metabuscador denominado "Anvish". Para o treinamento da rede neural, o Anvish utiliza as duas primeiras páginas do resultado de cada mecanismo de busca como exemplo de respostas relevantes. As duas últimas páginas são apresentadas à rede neural como exemplo de respostas não relevantes. Uma vez terminado este processo de aprendizagem, o Anvish apresenta as referências em ordem decrescente de relevância, baseando-se no que foi aprendido.

Resultados experimentais mostraram que o Anvish apresenta um desempenho significativamente superior à maioria dos metabuscadores que utilizam processos estatísticos (SHU; KAK, 1999).

CONCLUSÃO

A utilização das redes neurais artificiais em sistemas de recuperação de informação permite atribuir a tais sistemas um caráter dinâmico. Esta dinamicidade pode ser implementada mediante a participação ativa dos usuários em um processo contínuo de representação dos documentos do *corpus* ou mediante a aprendizagem de certas condições específicas no contexto de uma única busca, como é o caso do metabuscador Anvish.

Assim como as redes neurais, diversas outras idéias e conceitos desenvolvidos pela ciência da computação podem ser utilizados no tratamento e recuperação da informação. Porém, é desejável que essas idéias sejam implementadas e avaliadas levando-se em conta os avanços teóricos e metodológicos já realizados pelos processos documentários no âmbito da ciência da informação.

Os métodos e técnicas desenvolvidos pela ciência da computação devem ser continuamente estudados e até absorvidos pela ciência da informação. Porém, o profissional da informação deve ter sempre em mente que a ciência da informação não poderá ser desenvolvida no vazio cultural de um sistema de raciocínio algorítmico.

Artigo submetido em 14/06/2006 e aceito em 10/07/2006.

REFERÊNCIAS

BEIN, J.; SMOLENSKY, P. *Application of the interactive activation model to document retrieval*. Colorado: University of Colorado at Boulder, Department of Computer Science, 1988. (Technical Report CU-CS-405-88).

BELEW, R. K. Adaptive information retrieval. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 12. , 1989, Cambridge. *Proceedings*... Cambridge: ACM, 1989. p.11-20.

BRAGA, A. P.; CARVALHO, A. C. P. L. F.; LUDEMIR, T. B. *Redes neurais artificiais*: teoria e aplicações. Rio de Janeiro: LTC, 2000.

FORD, N. *Expert systems and artificial intelligence*: an information manager´s guide. London: Library Association Publishing, 1991.

HAYKIN, S. *Redes neurais*: princípios e prática. Porto Alegre: Bookman, 2001.

MCCULLOCH, W. S.; PITTS, W. H. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, n. 5, p.115-133, 1943.

MINSKY, M. L.; PAPPERT, S. *Perceptron*: an introduction to computational geometry. Cambridge: MIT Press, 1969.

MOZER, M.C. *Inductive information retrieval using parallel distributed computation*. San Diego: University of California, 1984. (ICS Technical Report 8406).

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and retrieval in the brain. *Psychological Review*, v. 65, p. 386-408, 1958.

SHU, B.; KAK, S. A neural network-based intelligent metasearch engine. *Information Sciences*, v. 120, n. 1-4, p. 1-11, 1999.