O poder cognitivo das redes neurais artificiais modelo ART1 na recuperação da informação

Ethel Airton Capuano
Doutorando em ciência da informação pela Universidade de Brasília (UnB). Mestre em gestão do conhecimento e da tecnologia da Informação pela Universidade Católica de Brasília (UCB).
E-mail: eacapuano@terra.com.br

Resumo

O artigo relata um experimento de simulação computacional de um sistema de recuperação da informação composto por uma base de índices textuais de uma amostra de documentos, um *software* de rede neural artificial implementando conceitos da *Teoria da Ressonância Adaptativa*, para automação do processo de ordenação e apresentação de resultados, e um usuário humano interagindo com o sistema em processos de consulta. O objetivo do experimento foi demonstrar (i) a utilidade das redes neurais de Carpenter e Grossberg (1988) baseadas nessa teoria e (ii) o poder de resolução semântica com índices sintagmáticos da abordagem SiRILiCO proposta por Gottschalg-Duque (2005), para o qual um sintagma nominal ou proposição é uma unidade linguística constituda de sentido maior que o significado de uma palavra e menor que uma narrativa ou uma teoria. O experimento demonstrou a eficácia e a eficiência de um sistema de recuperação da informação combinando esses recursos, concluindo-se que um ambiente computacional dessa natureza terá capacidade de clusterização (agrupamento) variável *on-line* com entradas e aprendizado contínuos no modo não supervisionado, sem necessidade de treinamento em modo *batch (off-line)*, para responder a consultas de usuários em redes de computadores com desempenho promissor.

Palavras-chave

Sistema de recuperação da informação. Sintagma nominal. Semântica. Indexação sintagmática. Mineração de textos. Redes neurais artificiais. Teoria da ressonância adaptativa. Redes neurais ART. Simulação computacional. Inteligência artificial.

The cognitive power of artificial neural networks model ART1 for information retrieval

Abstract

This article reports an experiment with a computational simulation of an Information Retrieval System constituted of a textual indexing base from a sample of documents, an artificial neural network software implementing Adaptive Resonance Theory concepts for the process of ordering and presenting outputs, and a human user interacting with the system in query processing. The goal of the experiment was to demonstrate (i) the usefulness of Carpenter and Grossberg (1988) neural networks based on that theory, and (ii) the power of semantic resolution based on sintagmatic indexing of the SiRILiCO approach proposed by Gottschalg-Duque (2005), for whom a noun phrase or proposition is a linguistic unity constituted of meaning larger than a word meaning and smaller than a story telling or a theory meaning. The experiment demonstrated the effectiveness and efficiency of an Information Retrieval System joining together those resources, and the conclusion is that such computational environment will be capable of dynamic and on-line clustering with continuing inputs and learning in a non-supervised fashion, without batch training needs (off-line), to answer user queries in computer networks with promising performance.

Keywords

Information retrieval system. Noun phrase. Semantics. Syntagmatic indexing. Text mining. Artificial neural networks. Adaptive resonance theory. ART neural networks. Computational simulation. Artificial intelligence.

INTRODUÇÃO

Este artigo relata um experimento de simulação computacional de um sistema de recuperação da informação composto por uma base de índices textuais sintagmáticos de uma amostra de documentos de apresentações ocorridas nos eventos IA Summit (Encontros de Arquitetura da Informação) de 2005 a 2008, nos EUA, construída manualmente por um especialista no tema, um *software* de rede neural artificial implementando conceitos da Teoria da Ressonância Adaptativa, de Helmholz, para automação do processo de ordenação e apresentação de resultados, e um usuário humano interagindo com o sistema em processos de consulta (busca) em linguagem natural. O objetivo do experimento é demonstrar a utilidade de redes neurais artificiais baseadas nessa teoria, uma promissora tecnologia de mineração de dados, em missões de recuperação da informação textual.

Em ciência da informação o conceito de *índice* é amplo, consistindo do registro de valores de vários atributos-chave de um texto referenciado com os quais se espera identificar e recuperar, futuramente, esse mesmo texto armazenado em alguma base de conteúdos informacionais (MEADOW *et al.*, 2007). O uso de índices compostos por mais de um termo linguístico como metadados, em vez de palavras-chave, busca a redução da ambiguidade, conforme proposto por Gottschalg-Duque (2005).

O tema de pesquisa correlato, com esse tipo de composição tecnológica utilizando inteligência artificial, é abordado na literatura brasileira da ciência da informação por Ferneda (2006), que ressalta, como vantagem da utilização de redes neurais em sistemas computacionais de recuperação da informação, os aspectos dinâmicos dessas redes de reconhecimento automático de padrões (aprendizado computacional). O usuário desse modelo de sistema de recuperação da informação com redes neurais artificiais pode ajustar os parâmetros de busca (consulta) na base textual de acordo com o nível de relevância ou generalização dos conceitos pretendidos, emprestando alguma capacidade de resolução semântica ao processo de busca.

Os desafios metodológicos do experimento são também interessantes no processo de indexação devido à complexidade inerente e às implicações epistemológicas da representação da informação com índices textuais e codificação binária, problema abordado, por exemplo, por Meadow *et al.* (2007), Tsuruoka *et al.* (2007) e Yu, Wang e Lai (2008). Os resultados esperados do experimento são evidências do poder cognitivo computacional de redes neurais artificiais modelo ART1, de Carpenter e Grossberg (1988), em contextos de sistemas de recuperação de informação (SRI) com bases textuais indexadas com sintagmas nominais.

CONCEITOS

Recuperação da informação

Contemporaneamente, Gottschalg-Duque (2005, p. 11), baseado em vários outros autores, apresenta a recuperação da informação como:

> (...) o clássico problema da recuperação efetiva e eficiente de documentos pertinentes extraídos de uma grande coleção (que nos dias de hoje pode ser entendida como um armazém de informação ou uma base de dados digital) de acordo com uma necessidade de informação específica, consistindo de três processos*:* coleta*,* indexação *e* ordenação.

A coleta é um processo de identificação, avaliação e armazenamento; a indexação, um processo de categorização com uso de palavras-chave para representação de um documento; a ordenação, um processo de disponibilização de documentos aos usuários segundo critérios de representação que satisfaçam suas necessidades.

Semântica com sintagmas nominais

A construção de um sistema de recuperação da informação[1] que possa extrair, consistentemente, informações semânticas acuradas de todo tipo de texto nem sempre é possível (KONCHADY, 2006), constituindo um dos maiores problemas conceituais da recuperação da informação a ambiguidade de termos linguísticos. Meadows *et al.* (2007) observam que:

> (...) enquanto identificadores únicos são utilizados para algumas aplicações e indicadores de classificação, em outros casos há incerteza sobre valores ou significados, causando confusão ao leitor humano ou a um programa de computador, ou ambos. E uma fonte de ambiguidade é a semântica – o significado dos símbolos.

Gottschalg-Duque (2005, p. 29) apresenta uma solução metodológica para o problema propondo que se construam proposições (ou enunciados, na matriz aristotélica) para a interpretação semântica. Certamente, este autor define em seu *framework* de recuperação da informação que "uma proposição é uma unidade constituída de sentido e que é maior que o significado de uma palavra e menor que uma narrativa ou uma teoria".

Considerando o significativo poder cognitivo subjacente à proposta de Gottschalg-Duque (2005) para solução do problema semântico, o experimento executado utilizou uma base de índices de textos em linguagem natural representada por sintagmas nominais, que são partes de uma sentença constituídas de substantivos geralmente apresentadas em estruturas sintáticas como SUBSTANTIVO_PREPOSIÇÃO-SUBSTANTIVO, SUBSTANTIVO_ADJETIVO, ADJETIVO_ SUBSTANTIVO, SUBSTANTIVO_ PREPOSIÇÃO_ ARTIGO_ SUBSTANTIVO e SUBSTANTIVO_ SUBSTANTIVO. A base de índices de textos em linguagem natural elaborada para o experimento é composta, na maior parte, de estruturas sintáticas estilo SUBSTANTIVO_ SUBSTANTIVO (ou NP_NP, no padrão inglês), SUBSTANTIVO_ PREPOSIÇÃO_ARTIGO_ SUBSTANTIVO (NP_PP_ART_NP) e ADJETIVO_ SUBSTANTIVO (ADJ_NP).[2] Como exemplos de sintagmas com essas estruturas sintáticas, encontram-se, na base textual utilizada, INFORMATION STRUCTURE (NP_NP), DATABASE-INTENSIVE APPLICATION (ADJ_NP), FACETED CLASSIFICATION (ADJ_NP). Observe-se que a estrutura sintática NP_NP é predominante nos sintagmas nominais do idioma inglês, equivalente, na maioria dos casos, à estrutura NP_PP_NP do idioma português, geralmente utilizadas para expressar substantivos compostos com relação de adjetivação de um termo sobre o outro no sintagma, como em *Information Architecture* (Arquitetura da Informação) e *User Needs* (Necessidades do Usuário). Outras estruturas sintáticas do idioma inglês também aparecem nos índices textuais, ainda que menos frequentes, como ADJ_NP_NP *(Faceted Browse System)*, NP_PP_ART_NP *(Metadata for the Masses)* e NP_NP_NP *(Desktop Data Integration)*.

Como vantagens desse método de indexação de textos, tem-se que:

a) um sintagma nominal é sempre mais poderoso, no sentido de redução da ambiguidade, que um termo isolado;

b) dois ou mais sintagmas nominais correlatos em um texto podem contextualizar melhor os termos sintáticos, do ponto de vista semântico, que apenas um sintagma nominal;

c) sintagmas nominais são estruturas de texto de extração relativamente fácil, atividade que pode ser automatizada com uso de *softwares* de processamento de linguagem natural especializados em etiquetação *(tagging)* e análise sintáticas *(parsing)*, ou *softwares* de mineração de textos.

[1] Utilizando-se, no caso, *recuperação* como um termo mais genérico, hiperônimo de *extração*.

[2] NP: *Noun Phrase*; ADJ: *Adjective*; ART: *Article*; PP: *Preposition*.

Mineração de textos

Os métodos e processos de descoberta de *informação* (que alguns chamam, em certos contextos, de *conhecimento*) relevantes em linguagem natural, conhecidos como *mineração de textos*, constituem variantes do que se conhece como *mineração de dados*. Chauke-Nehme (2008), como justificativa para os investimentos em pesquisa e desenvolvimento e o enorme interesse do mercado em tecnologias dessa natureza, argumenta que um importante desafio para os próximos dez anos é a preparação de um *exército de Davi*[3] para tirar vantagem do volume de informações disponíveis exponencialmente crescente no mundo em favor de uma sociedade mais sábia, justa e igualitária em termos de oportunidades para todos. Este autor observa, em particular, que a mineração de textos, que constitui um dos tipos de tecnologias úteis para o tratamento adequado dessa *inundação* de informações, tem apresentado significativa evolução na última década, desde o simples processamento de palavras na segunda metade da década de 1990 até hoje, quando o processamento de conceitos (termos ontológicos), ou mesmo a extração de conhecimento de estruturas linguísticas, tem se tornado possível.

Outros autores, como Do Prado e Ferneda (2008), definem *mineração de textos* como a aplicação de métodos e técnicas computacionais sobre dados textuais com a finalidade de encontrar informações intrinsecamente relevantes e conhecimento. Em relação às origens da mineração de textos na modernidade, Penteado e Boutin (2008) recordam o estudo de textos estruturados para mensuração da publicação científica, que surgiu e se desenvolveu a partir dos esforços de pioneiros como Solla Price, Small, van Raan, Swanson, Dou e Porter. Os autores mencionam que a mineração de textos estruturados é encontrada em campos do conhecimento tais como "bibliometria, cientometria, informetria, midiametria, museometria e webmetria", esclarecendo que nesses campos se estudam os diferentes aspectos da informação, inclusive sua qualidade, sendo a principal matéria-prima para esses estudos as palavras nos textos.

Quanto às funções que podem ser desempenhadas por sistemas de mineração de textos, Konchady (2006) observa que não há um conjunto padrão definido das funções desse tipo de tecnologia. O autor, no entanto, apresenta uma lista de soluções com uso (inclusive) de mineração de dados para problemas comuns de gestão da informação: busca *(search)*, extração de informação (busca de padrões de uso semântico), formação de *clusters*, categorização, construção de resumos (sumarização), monitoramento de informação, organização de perguntas e respostas (em aplicações de *Frequent Asked Questions*, por exemplo). O experimento relatado neste artigo pode ser classificado de modo abrangente, testando funções combinadas de busca de padrões semânticos e formação de *clusters* para ordenação de informação ao usuário de um sistema de recuperação da informação.

Simulação computacional

Experimentos com uso de técnicas de simulação computacional em pesquisas na área de ciências sociais não são comuns na literatura, caracterizando-se uma tendência histórica de produção acadêmica mais intensa com esse tipo de recurso metodológico nas denominadas c*iências exatas* e nas e*ngenharias*. Contudo, Pidd (1988) argumenta que métodos de simulação computacional têm se desenvolvido desde o início dos anos 1960, nos quais se incluem, talvez como as ferramentas mais comumente utilizadas para tanto, as utilizadas na administração. O pesquisador ou analista que utiliza simulação computacional desenvolve um modelo do sistema de interesse que imita o problema real, escreve um programa de computador que implementa esse modelo e utiliza um computador para emular o comportamento do

[3] Como metáfora relacionada ao confronto bíblico entre Davi (representando os usuários de informações) e o gigante Golias (representando o enorme volume de informações disponíveis).

sistema quando submetido a uma variedade de políticas (ou situações) operacionais (PIDD, 1988). Como resultados observados das simulações, o pesquisador ou analista terá dados para análise e condições de compor um modelo completo do problema e da melhor solução possível, comparando as implicações de adoção de uma solução ou outra em termos de desempenho.

O método que utiliza simulação pode ser considerado um método experimental de pesquisa científica à medida que testa um modelo conceitual de problema e solução, e Pidd (1988, p. 5-6) argumenta, no contexto da simulação computacional da administração, que:

> O modelo é utilizado como um veículo para experimentação, frequentemente com tentativa e erro, para demonstrar os prováveis efeitos de várias políticas. Então, aquelas que produzem os melhores resultados no modelo seriam implementadas no sistema real.

A simulação computacional consome tempo e, geralmente, é uma tarefa bastante complexa, mas apresenta também as seguintes vantagens (PIDD, 1988):

a) custo: em geral é menor que um experimento real nas organizações;

b) tempo: após a modelagem e a construção do artefato computacional, as possibilidades de emulação de situações reais com simulação são ilimitadas, podendo-se também simular períodos maiores de operações reais em questão de segundos no computador;

c) possibilidade de replicação: um experimento simulado pode ser replicado por outros pesquisadores, uma vez que as condições ambientais são modeladas e estruturadas de modo inteligível, algo difícil de ocorrer em um experimento real, pois as organizações e os ambientes reais geralmente não se repetem naturalmente de modo muito similar;

d) segurança: uma simulação geralmente não apresenta riscos para as organizações ou pessoas, o que nem sempre ocorre em experimentos reais (experiências reais malsucedidas podem acarretar danos ambientais ou prejuízos econômicos para as organizações).

Redes neurais artificiais

As redes neurais artificiais são um paradigma específico de inteligência artificial que representa uma ruptura com os conceitos tradicionais nessa área do conhecimento, estes centrados na lógica computacional e na heurística como meios de se introduzir inteligência em um artefato. Com essas redes, uma corrente de cientistas da inteligência artificial buscou, no final dos anos 1950, a construção de artefatos inteligentes com capacidade de aprendizado autônomo, sem necessidade de informação *a priori* de todos os detalhes lógicos e heurísticos que o artefato inteligente deveria conter, em determinadas circunstâncias, para executar sua missão (FREEDMAN, 1995; FREEMAN; SKAPURA, 1993).

O potencial uso de redes neurais artificiais em sistemas de recuperação da informação é apresentado, em literatura brasileira, por Ferneda (2006), comparando esse tipo de sistema a uma rede neural de três camadas: a camada de termos de busca seria a camada de entrada de dados na rede; a camada de documentos seria a saída; a camada de termos de indexação seria uma camada central. O experimento executado e relatado no presente artigo segue essa arquitetura básica de sistema; contudo, antes são apresentadas as principais características dos principais modelos de redes neurais artificiais.

Modelos e arquiteturas

Freeman e Skapura (1993) comentam que, quando a única ferramenta que se dispõe é um martelo, todos os problemas que encontramos tendem a nos parecer como se fossem pregos. De fato, o mesmo

vício ocorre entre os profissionais de tecnologia da informação: quem domina uma (e somente uma) tecnologia procura sempre implementar soluções que a utilizem, mesmo que existam outras tecnologias mais adequadas à disposição. Contribuindo para a superação desse hábito profissional, as redes neurais artificiais apresentam diversas topologias, conceitos e vários algoritmos de aprendizado, sem falar nas aplicações bastante variadas para cada tipo de rede. O desafio para o *engenheiro do conhecimento* responsável por um projeto de mineração de dados é justamente empreender a correta seleção de tecnologias e, tendo optado pelo emprego de redes neurais artificiais, a correta seleção do modelo e arquitetura da rede para solucionar o problema que se apresenta.

O primeiro ponto a ser analisado no processo de seleção se refere ao modo de aprendizado, se supervisionado, não supervisionado ou reforçado, e a decisão por um modo ou outro deve resultar do estudo cuidadoso do problema no contexto do ecossistema informacional em que ele se insere. Como regra geral, o aspecto crucial é o grau de conhecimento que se tem dos padrões a serem reconhecidos pela rede, que se relaciona ao conhecimento que se tem dos dados disponíveis para mineração. Quando os padrões de saída da rede, ou *vetores-alvo*, são previamente conhecidos, opta-se por redes com aprendizado supervisionado; neste caso, a rede opera com funções mais sensoriais do que cognitivas, pois os padrões a serem aprendidos já são conhecidos pelos usuários e o trabalho da rede se parece mais com classificação que agrupamento *(cluster)*.

As redes com aprendizado não supervisionado são especialmente úteis quando se navega em um ambiente de dados no qual o conhecimento encontra-se oculto e não se sabe os conteúdos das descobertas – a rigor, o aprendizado não supervisionado representa a essência epistemológica da mineração de dados. O experimento objeto deste artigo mostra um meio de se transpor esse construto cognitivo para mineração de textos.

Topologias

Conforme Bigus (1996), as topologias básicas de redes neurais artificiais são de alimentação progressiva *(feedforward),* limitadamente recorrentes e totalmente recorrentes (figura 1, a seguir). Os dados do fenômeno em estudo são introduzidos na camada de processamento da rede denominada "Entrada", para que a rede realize as tarefas de aprendizado para as quais foi projetada na camada "Oculta" e, por fim, mostre os resultados aos usuários nas interfaces da camada "Saída". Em redes com aprendizado supervisionado, os resultados devem mostrar quais dados de entrada se identificam com quais padrões de saída esperados.

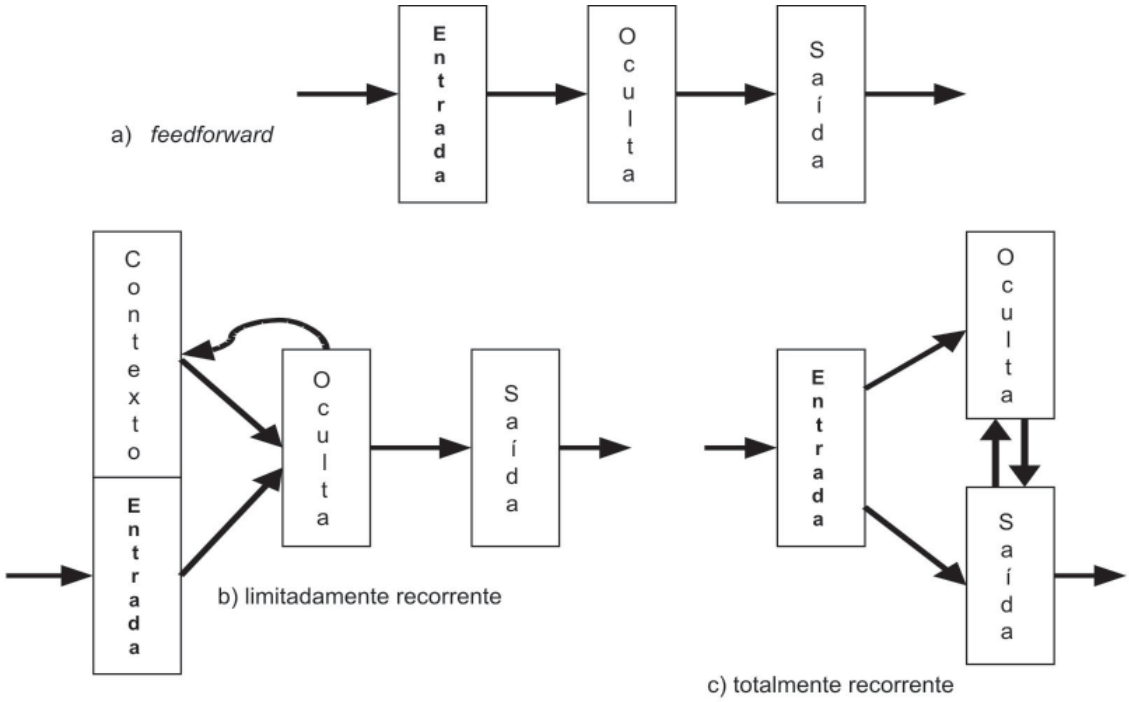
Como no conceito mais primitivo de caixa-preta de modelagem de sistemas, o modelo básico de três camadas é o modelo-padrão nas redes neurais, moldando sua topologia de modo explícito ou implícito. Embora o modelo de três camadas clássico possa apresentar variantes, o que diferencia um modelo topológico de rede de outro se relaciona com o método de aprendizagem, pois este é o trabalho das camadas intermediárias (ocultas).

Redes *feedforward*

As topologias de alimentação para frente *(feedforward)* são as mais comuns, sendo utilizadas em situações em que se introduzem todos os dados para solução do problema, na camada de processamento de "Entrada", de uma só vez. Nesse tipo de topologia, os dados fluem através das camadas de processamento na rede em um só sentido (daí sua denominação *alimentação para frente*), sem reavaliações recursivas do processo de aprendizado, e a resposta é baseada somente no conjunto corrente de dados de entrada (BIGUS, 1996).

A figura 1a mostra os blocos de construção de uma rede neural artificial com alimentação para frente típica, onde os dados entram na rede através das unidades de entrada posicionadas à esquerda, sendo submetidas a essas unidades como "valores de ativação". O trabalho da camada intermediária, em

FIGURA 1
Topologias de redes neurais



suma, será encontrar, mediante comparações iterativas entre entradas e saídas, o melhor conjunto de pesos (números reais) que, multiplicados pelos valores dos dados de entrada, possam identificar semelhanças deste produto com os valores dos dados esperados na camada de "Saída". A natureza computacional e heurística de uma rede neural, portanto, não é definida caso a caso, situação por situação, como nos algoritmos de inteligência artificial tradicional, mas sim como um modelo cognitivo padrão de mais alto nível – alvo da crítica dos cientistas da corrente tradicional, porque essa vantagem cognitiva torna-se uma desvantagem lógica na medida em que não se pode explicitar e formalizar, com precisão, os mecanismos de aprendizado do artefato.

Cada unidade de processamento combina todos os sinais de entrada ponderados com um valor limiar, e a soma dos sinais de entrada no "neurônio" é submetida a uma função de ativação, como no modelo de ativação elétrica dos neurônios biológicos das redes neurais naturais, que determina a saída atual da unidade da camada de processamento de saída que, por sua vez, torna-se entrada para outra camada. A função matemática de ativação mais utilizada é a sigmóide, que converte um valor de entrada em uma saída com amplitude entre 0,0 e 1,0 (é uma função de normalização, portanto). Os valores dos dados que saem das unidades de entrada são modulados pelos pesos das conexões da rede e amplificados, se o respectivo peso da conexão for positivo e maior que 1,0 (um), ou reduzidos, se o peso da conexão se encontra entre os valores 0,0 (zero) e 1,0 (um). O processo de aprendizado da rede consiste, portanto, no ajuste dos pesos de modo que as classificações dos padrões de entrada sejam executadas corretamente.

As funções sigmóides têm a propriedade de apresentar resultados (variáveis dependentes) muito próximos a partir de certo limiar do valor dos dados de entrada (variáveis independentes), razão de seu uso em redes neurais. Os produtos matemáticos resultantes desses cálculos são utilizados como filtros para classificação dos dados de entrada segundo os interesses dos usuários

expressos na camada de dados de saída. O processo de alimentação dos dados para frente, ou para a camada seguinte, é executado com base nos "sinais de ativação" resultantes do processamento com os pesos, sendo o sinal positivo (geralmente, + 1) representativo da "aprovação" do sistema de pesos utilizado e o sinal negativo de sua "reprovação" (com sinal – 1). Ou seja, dados "aprovados" são classificados em determinado padrão de saída esperado e dados "reprovados" são desclassificados nesse padrão, embora possam ser "aprovados" em outro padrão de saída na rede.

Redes limitadamente recorrentes

Em uma rede neural artificial com topologia limitadamente recorrente (figura 1b), a sequência das entradas é importante e espera-se que a rede armazene, de algum modo, um registro das entradas anteriores (de "contexto") e os fatores com os dados correntes para produzir uma resposta. Nesse tipo de rede, informações sobre entradas passadas são reintroduzidas e misturadas com as entradas presentes através das conexões recorrentes ou de retroalimentação nas unidades das camadas ocultas ou de saída. Dessa forma, a rede contém uma memória das entradas passadas a partir das ativações, incrementando o processo de aprendizado.

Redes limitadamente recorrentes estabelecem um compromisso de equilíbrio entre a simplicidade da proposta *feedforward* e o poder de separação lógico-matemática não linear de redes totalmente recorrentes e permitem a utilização do algoritmo de treinamento de retropropagação (*backpropagation)*. Com esse tipo de rede neural, pode-se operar, portanto, com um nível de complexidade maior na classificação dos dados de entrada que as redes de alimentação para frente.

Redes totalmente recorrentes

Redes neurais com topologia totalmente recorrente propiciam conexões de duas vias entre todas as unidades processadoras (figura 1c). Os dados de entrada fluem da primeira camada para todas as demais camadas adjacentes e circulam para frente e para trás até que a ativação das unidades se estabilize. As ativações das camadas ocultas e de saída são, então, recomputadas até que toda a rede se estabilize; nesse ponto, os valores de saída podem ser lidos das unidades processadoras da camada de saída.

Essas redes são sistemas muito complexos e dinâmicos e exibem comportamento de sistemas caóticos (da Teoria do Caos). Apesar de seu alto poder de simulação de sistemas e soluções complexas, o sistema revela alta instabilidade e, muitas vezes, não converge para uma única solução ou, quando converge, consome muito tempo com as iterações recorrentes entre os neurônios. As mesmas características que trazem desvantagens apresentam também vantagens, e essas redes, por isso, são úteis para solução de problemas muito complexos que envolvem funções de otimização.

Paradigmas e algoritmos de aprendizado

Em termos de modelos completos de redes neurais artificiais, assim definidos com base na topologia de conexões e no paradigma de aprendizado, as redes *backpropagation* (com retropropagação) e *Kohonen Feature Map* (Mapa de Características de Kohonen) são as mais populares para mineração de dados (BIGUS, 1996). Redes supervisionadas com aprendizado por *retropropagação* utilizam topologia de alimentação progressiva, sendo responsáveis pelo retorno do interesse pelas redes neurais na década de 1980, após as frustrações da década de 1960.

Essas redes tornaram-se populares graças ao trabalho de Rumelhart, Hinton e Williams e superaram as críticas dos pesquisadores da inteligência artificial tradicional, como Minsky e Papert, do Massachusetts Institute of Technology. As principais vantagens das redes com retropropagação são a simplicidade do paradigma de aprendizado e seu ecletismo em termos de aplicações concretas, representando o ideal da inteligência artificial com base na natureza. Como desvantagens, tem-se o tempo de resposta

frequentemente moroso e sua inaplicabilidade de aprendizado em tempo real devido ao conhecido dilema entre a estabilidade necessária para convergência da solução com um conjunto de padrões de treinamento e a plasticidade indispensável para que a rede possa aprender novos padrões sem *esquecer* os antigos.

Outros paradigmas de redes neurais artificiais são apresentados em Freeman (1993). Redes neurais com aprendizado baseado na Teoria da Ressonância Adaptativa (*Adaptive Resonance Theory*, ou ART) são consideradas uma família de redes recorrentes, com variantes supervisionadas e não supervisionadas, úteis na mineração de dados para identificação de *clusters*. O dilema estabilidade-plasticidade tem motivado pesquisas e desenvolvimentos de modelos de redes neurais artificiais que imitam de forma mais plausível o aprendizado cumulativo, contínuo e metamórfico do ser humano. O paradigma de aprendizado das redes ART, de Grossberg (1988), representa uma solução para o problema.

TEORIA DA RESSONÂNCIA ADAPTATIVA

A Teoria da Ressonância Adaptativa é um paradigma de rede neural desenvolvido no Centro de Sistemas Adaptativos da Universidade de Boston (EUA) por Carpenter e Grossberg, tendo como principal característica sua similaridade com os processos de aprendizado humano. Esse paradigma se contrapõe ao tradicional, baseado na lógica de primeira ordem e na heurísica, porque busca na própria natureza os processos de aprendizado, entendendo que os seres vivos sobrevivem porque aprendem a se adaptar continuamente ao ambiente mutante. Os artigos seminais desses pesquisadores pioneiros, que lançaram os fundamentos epistemológicos desse tipo de rede neural, foram publicados originalmente em 1988 (GROSSBERG, 1988; CARPENTER; GROSSBERG, 1988) e 1991 (CARPENTER; GROSSBERG; ROSEN, 1991), a partir dos quais se desenvolveu uma série de variantes do modelo original.

Com um trabalho de mais de quatro décadas em teorias sobre o funcionamento do cérebro e inteligência artificial, Carpenter e Grossberg construíram, gradualmente, um conjunto de modelos de redes neurais que parecem explicar alguns dos mais desconcertantes problemas do pensamento e permitem simular sistemas de inteligência artificial que "veem, se movem e aprendem" da maneira humana (FREEDMAN, 1995). A principal diferença entre o trabalho desses pesquisadores e a retropropagação ou outros tipos de redes neurais é que ele eliminou todas as condições simplificadoras que lhes permitiam funcionar. Com efeito, a maioria das redes neurais artificiais funciona somente com informações estacionárias e controle externo, ou seja, podem manejar somente um conjunto de padrões invariável, introduzido na rede lentamente e supervisionado por um treinador, do mesmo modo que se executa um algoritmo computacional.

Grossberg insiste que uma rede neural artificial verdadeiramente semelhante à humana deveria ser "autônoma, de aprendizagem rápida e adaptável"; isso significa uma rede capaz de aprender rapidamente como organizar e manejar um mundo pleno de surpresas (FREEDMAN, 1995, p. 106). As redes ART, portanto, são inspiradas na inteligência artificial baseada na natureza, que é pensada em termos de três princípios elementares (FREEDMAN, 1995, p. 27-95):

I. a melhor maneira de se compreender como funciona a inteligência humana é estudar, antes, como funciona um modelo de inteligência mais elementar, como a inteligência animal;

II. a inteligência pode ser emergente, uma propriedade da interação complexa de elementos mais simples;

III. a inteligência é demasiadamente complexa para que possa ser projetada a partir do zero.

Essa característica de autoaprendizagem das redes neurais modelo ART pode ser considerada como um quarto princípio de aproximação à inteligência

artificial baseada na natureza: não se pode inserir a inteligência em um sistema, mas sim desenvolvê-la mediante interação desse sistema com o mundo circundante (FREEDMAN, 1995, p. 107).

EXPERIMENTO DE SIMULAÇÃO COMPUTACIONAL

Objetivo e critérios de avaliação

O experimento de recuperação da informação textual a partir de índices sintagmáticos com uso de redes neurais modelo ART teve como principal objetivo testar a utilidade desse tipo de tecnologia combinada com os conceitos desenvolvidos por Gottschalg-Duque (2005) no processamento de sintagmas nominais para associação semântica de termos. Ou seja, testar o *framework* conceitual de recuperação da informação sintagmática baseado na arquitetura de rede neural denominada ART1 desenvolvida por Carpenter e Grossberg (1988). O conceito de sintagma nominal é explorado, no experimento, pela sua promissora capacidade de superação, ao menos parcial, do conhecido problema semântico inerente aos sistemas tradicionais de recuperação da informação com uso de técnicas estatísticas e sintáticas de processamento de termos da linguagem natural.

O critério de avaliação do experimento adotado é o da viabilidade funcional do *framework* no atual contexto da recuperação da informação, subdividindo-se, do ponto de vista técnico, em (i) revocação e precisão (KENT *et al.*, 1955), (ii) eficiência computacional e (iii) usabilidade. Define-se *revocação* como uma métrica de conjuntos que mostra a relação entre o número de documentos (ou registros) relevantes recuperados pelo sistema e o número de documentos relevantes existentes na base ou repositório de informação; *precisão* como a relação entre o número de documentos relevantes recuperados e o número total de documentos recuperados. A eficiência computacional se refere, no caso, ao tempo de resposta do Sistema de Recuperação da Informação, ou seja, ao tempo de processamento dos dados de entrada no sistema até a apresentação dos resultados ao usuário. A usabilidade é um conceito da ciência da computação que se refere à qualidade do *software* do ponto de vista do usuário; no caso, avalia-se como tal sistema poderia ser utilizado pelo usuário sem que o mesmo tenha necessidade de se envolver com a complexidade das redes neurais artificiais.

Esses três itens de avaliação que compõem o critério de viabilidade adotado são discutidos ao longo deste capítulo, mostrando-se também, no capítulo das conclusões, algumas questões em aberto para futuros experimentos.

Motivação

Conforme experiências anteriores de Chauke-Nehme (1996), Serrano-Gotarredona, Linares-Barranco e Andreou (1998), Capuano (2001), Capuano e Chauke-Nehme (2002) e outros, as redes neurais ART1, um modelo variante da *família* de redes ART, têm se mostrado eficientes no reconhecimento de padrões binários e flexíveis para ajustes de granularidade na formação de *clusters*. Contudo, sua capacidade de resposta não tem sido testada com entradas textuais, ou seja, em contextos de mineração de textos, com dados representativos de termos da linguagem natural, despertando curiosidade sobre seus resultados nesse cenário funcional. E talvez uma de suas principais vantagens sobre outras tecnologias de recuperação da informação, mesmo outros modelos de redes neurais artificiais, seja a capacidade de clusterização *on-line* com entradas contínuas no modo não supervisionado, recursos que a habilitam, naturalmente, para implementação em sistemas de recuperação da informação para usuários em redes de computadores *on-line* operando em tempo real.

Os experimentos apresentados por Capuano (2001) e Capuano e Chauke-Nehme (2002), por exemplo, mostraram que o parâmetro de vigilância ρ das redes ART1 podem ser ajustados a cada sessão (processo de consulta computacional) de modo que a rede produza *clusters* mais ou menos agrupados de acordo com a necessidade do usuário, demonstrando uma capacidade de clusterização que pode ser utilizada para o agrupamento de termos mais específicos ou

mais genéricos, assim como para a conexão semântica (hierárquica ou facetada) de um grupo de termos a outros com variados graus de similaridade. Essa capacidade de clusterização variável é obtida de uma rede ART1 com relativa facilidade de manuseio, configurando-se e reconfigurando-se o parâmetro ρ como um dado de entrada no processamento computacional. Com isso, ART1 se mostra também uma tecnologia de maior usabilidade que outros modelos de redes neurais artificiais, na qual o próprio usuário final de um serviço de recuperação da informação poderia ajustar os parâmetros de granularidade de sua consulta à base textual por meio de uma interface de *software*.

Metodologia

Codificação dos padrões textuais

Como mencionado anteriormente, os padrões de entrada em uma rede neural ART1 são vetores com características dimensionais representadas de forma binária (com *zeros* e *uns*). Consequentemente, o desafio metodológico inicialmente encontrado no experimento se deveu à necessidade de representação binária da informação dos sintagmas nominais de indexação responsáveis pela expressividade semântica dos textos referenciados na base.

Os textos utilizados no experimento constituem uma amostra dos resumos das apresentações dos encontros *IA Summit* (nos EUA) de 2005 a 2008. O número de sintagmas nominais utilizados para indexação de cada texto dessa base é três, pois com esse número pode-se obter significativo poder de resolução semântica no processo de recuperação da informação sem pressionar demais a infraestrutura computacional, como observado na simulação executada. Com isso em mente e sopesando-se os prós e contras de vários modelos de codificação baseados no padrão binário, escolheu-se um padrão simples de correlação entre sinais gráficos do idioma inglês e números binários, decisão que teve como princípio norteador também a garantia de não colisão (ou ambiguidade) de códigos. Considerando que os índices sintagmáticos construídos para representação dos textos contêm apenas as 26 letras do alfabeto inglês mais o espaço em branco (entre termos), o traço, a apóstrofe e os sinais de interrogação e exclamação, o código completo precisa expressar 31 instâncias de representações diferentes de sinais gráficos idiomáticos. Optou-se, em função desses requisitos, por uma representação com 5 (cinco) *bits* por símbolo idiomático da linguagem natural (idioma inglês), que poderia, em um extremo, representar mais um símbolo que os 31 previstos. Ou seja, com cinco *bits* pode-se representar (mapear) até 32 símbolos diferentes (a figura 2, que mostra uma parte da planilha de codificação binária elaborada, ilustra como essa codificação foi implementada).

Conhecendo-se a amostra de sintagmas de indexação produzidos na fase anterior ao experimento, observou-se que seria necessária uma *string* (sequência de caracteres) padrão de no mínimo 30 símbolos léxicos de cinco *bits* cada, totalizando 150 *bits* por sintagma nominal. Com isso, um dos mais extensos sintagmas de indexação produzidos, *information-processing devices*, poderia ter seus 30 símbolos léxicos codificados. Observe-se, no entanto, que em sintagmas mais curtos em *bits* significativos o preenchimento da *string* de 150 *bits*, na parte restante, dá-se com séries de *bits* codificadas como 11010, que corresponde a um espaço em branco. O termo *information architecture*, por exemplo, é codificado nesse esquema como:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
01000	01101	00101	01110	10001	01100	00000	10011	01000	01110	01101	11010	00000	10001	00010	00111
I	**N**	**F**	**O**	**R**	**M**	**A**	**T**	**I**	**O**	**N**		**A**	**R**	**C**	**H**

17	18	19	20	21	22	23	24	25	26	27	28	29	30
01000	10011	00100	00010	10011	10100	10001	00100	11010	11010	11010	11010	11010	11010
I	**T**	**E**	**C**	**T**	**U**	**R**	**E**						

O tipo de codificação adotado deve contar com auxílio de programação computacional para sua execução, pois do contrário tem-se uma atividade extremamente morosa e penosa pela frente, também sujeita a erros. Utilizou-se, no entanto, apenas uma planilha eletrônica para a preparação de séries de 76 diferentes padrões de entrada (*strings* de 450 *bits* representando cada vetor de três sintagmas nominais codificado) para realização do experimento, com o leiaute da figura 2, sendo 75 padrões referentes aos índices da base textual testada mais um padrão referente ao argumento de pesquisa do usuário introduzida no sistema.

Considerando que redes ART1 operam com vetores dimensionais de posições fixas, nos quais os padrões posicionais de características de cada vetor são utilizados para clusterização, o número total de padrões testados em cada sessão computacional com a rede ART1 é 456, pois cada vetor-padrão de entrada, constituído de três sintagmas nominais, necessita ser apresentado de seis formas diferentes para que nenhuma possibilidade de comparação vetorial com os padrões existentes da base textual possa ser ignorado no teste (6 x 76 = 456). O uso desse modelo de apresentação dos dados de entrada na rede é necessário para se evitar que a ordem de inserção dos três sintagmas de busca do usuário no sistema interfira na eficácia do processo de recuperação da informação. As seis formas de apresentação dos padrões de entrada resultam do cálculo fatorial do número dos arranjos combinatórios possíveis no caso, ou seja: $A^n = n! = 3! = 6$. Os 450 *bits* de cada sintagma nominal (índice textual) resultam da seguinte multiplicação: 5 *bits*/símbolo x 30 símbolos/sintagma x 3 sintagmas/vetor = 450 *bits*/vetor.

Sistema computacional utilizado

O experimento foi realizado utilizando-se o algoritmo e o código-fonte de *software* de redes ART1 publicados por Serrano-Gotarredona,

FIGURA 2
Codificação binária de índices sintagmáticos

	A	B	C	D	E	F	G	H	I	J	...
	00000	00001	00010	00011	00100	00101	00110	00111	01000	01001	...
	1	2	3	4	5	6	7	8	9	10	...
1	Social classification, freetagging, metadata for the masses.		10010	01110	00010	01000	00000	01011	11010	00010	...
2	Social classification, metadata for the masses, freetagging.		10010	01110	00010	01000	00000	01011	11010	00010	...
3	Freetagging, social classification, metadata for the masses.		00101	10001	00100	00100	10011	00000	00110	00110	...
4	Freetagging, metadata for the masses, social classification.		00101	10001	00100	00100	10011	00000	00110	00110	...
5	Metadata for the masses, social classification, freetagging.		01100	00100	10011	00000	00011	00000	10011	00000	...
6	Metadata for the masses, freetagging, social classification.		01100	00100	10011	00000	00011	00000	10011	00000	...
7	Information architecture, content management rules, content modeling.		01000	01101	00101	01110	10001	01100	00000	10011	...
8	Information architecture, content modeling, content management rules.		01000	01101	00101	01110	10001	01100	00000	10011	...
9	Content management rules, information architecture, content modeling.		00010	01110	01101	10011	00100	01101	10011	11010	...
10	Content management rules, content modeling, information architecture.		00010	01110	01101	10011	00100	01101	10011	11010	...
11	Content modeling, information architecture, content management rules.		00010	01110	01101	10011	00100	01101	10011	11010	...
12	Content modeling, content management rules, information architecture.		00010	01110	01101	10011	00100	01101	10011	11010	...

Linares-Barranco e Andreou (1998), no modo de codificação complementar (vetores de entrada de dimensão dobrada), com algumas alterações implementadas por Capuano (2001) para adequação de interfaces de entrada e saída de dados. Esse tipo de configuração dos padrões de entrada é necessário para se evitar inadequada proliferação de *clusters*, como explicado por Carpenter, Grossberg e Rosen (1991) e Carpenter *et al.* (1992).

A linguagem de programação adotada foi MATLAB, uma linguagem de alto nível muito utilizada nas engenharias para operações com matrizes, construída sobre bibliotecas de linguagens de programação de nível mais baixo. O microcomputador utilizado para processamento da rede é um *notebook* HP Compaq nx9005, com microprocessador *mobile* AMD Athlon XP2200+ de 1,79 GHz, 1,0 GB de memória RAM, 40 GB de disco rígido SATA e Sistema Operacional Microsoft Windows XP Pro, versão 2002 (*Service Pack* 2).

Organização dos dados e processamento

Os dados (padrões sintagmáticos) utilizados no experimento foram organizados de modo a representar várias situações de um sistema de recuperação da informação no atendimento de consultas de usuários. Simulou-se uma série de seis consultas de usuários que teriam acesso ao sistema por meio de uma tela de computador, quando o mesmo precisaria informar ao sistema apenas os argumentos de busca constituídos por três sintagmas nominais por consulta, que sugerem o conteúdo pesquisado, com alguma correlação semântica entre si. Como exemplo dessas simulações, um usuário informaria, na tela de entrada dos parâmetros de pesquisa, os sintagmas seguintes (três campos de dados textuais):

[CAMPO 1]: **Rich Internet Application**
[CAMPO 2]: **Information Findability**
[CAMPO 3]: **User Needs**

Os sintagmas nominais simulados como dados das seis consultas dos usuários ao sistema são apresentados a seguir:

- CONSULTA 1: **Semantic Web**; **Information Retrieval System**; **Digital Library**;
- CONSULTA 2: **Web Design**; **Vignette**; **Information Architects**;
- CONSULTA 3: **Rich Internet Application**; **Information Findability**; **User Needs**;
- CONSULTA 4: **Web Design**; **User Needs**; **Faceted Classification**;
- CONSULTA 5: **Findability**; **Search Engine**; **Google**;
- CONSULTA 6: **Information Findability**; **Information Architects**; **User Needs.**

Conforme se pode comparar na lista de índices sintagmáticos que compuseram a base de informações de referência do sistema, algumas das consultas (como a 1 e a 5) não deveriam produzir *clusters* com muito poder semântico em razão da pouca similaridade com os conteúdos da base, enquanto as demais (2, 3 e 4) poderiam se agrupar em *clusters* mais interessantes (poderosos) do ponto de vista semântico, com mais de um sintagma coincidente entre a consulta e a base de dados do sistema. Essa variação dos padrões de consulta foi idealizada propositalmente para se testar os efeitos na rede ART1 do Sistema de Recuperação da Informação simulado. Com a consulta 6, no entanto, testou-se a capacidade de a rede reconhecer um argumento de consulta totalmente coincidente com um índice existente na base textual.

Como recursos configuráveis de uma rede ART1, os parâmetros de precisão da busca também podem ser informados pelo usuário, que poderá optar por uma busca mais genérica ou mais restrita (específica), algo que se implementou no experimento simulado. Caso opte inicialmente por uma busca mais genérica, o usuário deverá configurar o parâmetro de vigilância (ρ) do sistema para um valor adequado, geralmente entre 0,60 e 0,75 (CAPUANO; CHAUKE-NEHME, 2002);

caso opte, de imediato, por uma busca mais específica, tentando maior precisão, o usuário deverá informar um valor mais alto e próximo de 1,0 para o parâmetro de vigilância (entre 0,75 e 0,85). O parâmetro L da rede neural deverá ser configurado pelos próprios administradores do aplicativo, pois sua melhor opção somente poderá ser avaliada com testes na base de dados, algo pouco recomendável para um ambiente computacional *on-line*.

As dimensões do vetor de dados binários (n) e o número de padrões de entrada (np) a serem submetidos para processamento são parâmetros do próprio sistema e deverão ser configurados pelo administrador do sistema de recuperação da informação. O parâmetro *np* se relaciona com o próprio tamanho da base textual, sem limite teórico de quantidade de índices armazenados.

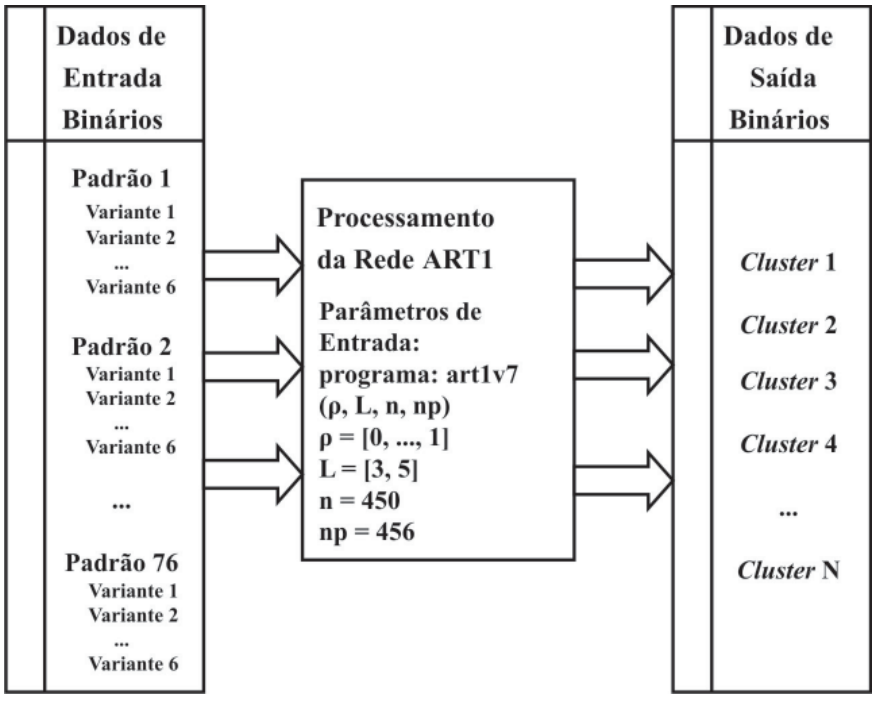
O esquema lógico de entradas, processamento e saídas desse sistema simulado com redes ART1 é esboçado na figura 3.

Resultados

O experimento simulado de um sistema de recuperação da informação baseado em redes neurais artificiais modelo ART1, com dados de entrada binários representando índices textuais compostos por sintagmas nominais, produziu resultados interessantes para a ampliação do conhecimento sobre esse tipo de tecnologia e seu uso potencial para desenvolvimento de soluções para o problema da semântica na recuperação da informação. As tabelas 1 e 2, a seguir, mostram os resultados do processamento das seis consultas de usuários simuladas e seu significado em termos de utilidade no contexto. Os parâmetros ou argumentos de consulta (busca) do usuário aparecem no bloco de colunas à esquerda nas tabelas intitulado *Parâmetros de Consulta do Usuário Simulado,* e os resultados da busca, no bloco da direita.

O usuário deve introduzir, nas interfaces do sistema simulado, os sintagmas em linguagem natural (idioma inglês, no caso) que representam os conteúdos de interesse na busca textual e o

FIGURA 3
Esquema lógico do SRI simulado



parâmetro de vigilância da rede ART1 ρ, que regula o nível de generalização das respostas do sistema. Esse parâmetro necessitaria ser traduzido para um termo mais próximo do senso comum, para que um usuário leigo nesse tipo de tecnologia possa utilizar, conscientemente, o recurso de *calibragem* do sistema – por exemplo, CAMPO 1: **Consulta Ampla** e CAMPO 2: **Consulta Restrita**, onde o primeiro nível de generalização da consulta implicaria a parametrização de ρ com valores [0,6; 0,65; 0,70] e o segundo nível com valores [0,75; 0,80; 0,85].

O bloco *Resultados* mostra que o sistema devolve (i) um dado denominado *npa*, que corresponde ao número de *clusters* formados pela rede ART1, (ii) os números de identificação sequencial dos *clusters* formados que contenham, ainda que parcialmente, os sintagmas de busca introduzidos pelo usuário, (iii) os índices de textos da base (ou *componentes*) que foram considerados similares no respectivo *cluster* enumerado e (iv) os textos, em linguagem natural, dos próprios índices de textos recuperados pelo sistema, para que o usuário possa avaliar sua utilidade no contexto. Certamente, desses dados de saída, apenas os índices em linguagem natural seriam mostrados ao usuário na tela do computador.

Análise da Consulta 1

Como se pode observar na tabela 1, a seguir, os sintagmas nominais introduzidos pelo usuário simulado na primeira consulta são os seguintes: **Semantic Web**; **Information Retrieval System**; **Digital Library**. Com o parâmetro de vigilância (ou de generalização) da sessão de busca ρ = 0,60, o usuário obteria como resultado um texto tendo como índice sintagmático[4] **Wireframing**; **Customer Expectation**; **Web Development**. Esse texto se encontraria indexado na base textual com o número *449*. O único fragmento de sintagma que indica alguma similaridade entre o argumento de busca e o resultado, nesta sessão, é o termo **Web** (destacado na coluna *Índices Sintagmáticos Recuperados* da tabela 1, a seguir). Os conteúdos semânticos dos demais sintagmas, ou fragmentos (termos) de sintagmas, teriam de ser avaliados pelo usuário para concluir sobre sua utilidade.

Quando a rede ART1 executa a sessão de clusterização, nessa mesma consulta, com o parâmetro de configuração ρ = 0,65, os resultados são ampliados em números de textos recuperados, obtendo o usuário, como resposta do sistema, três textos recuperados com os seguintes índices sintagmáticos: **IA Practices**; **Information Architects**; **Local IA Groups** (número *426* na base textual), **Wireframe**; **Information Architect**; **IA Community** (número *427*) e **Consistency and Indexing**; **Information Retrieval**; **Information Architecture** (número *436*). Como destacado na tabela 1, o termo **Information** e o sintagma nominal **Information Retrieval** é que caracterizaram a similaridade sintática vetorial entre os argumentos de consulta e os índices textuais de referência produzidos pela rede nesta sessão.

Observa-se, com o uso do parâmetro de vigilância ρ = 0,65, um aumento do poder semântico dos índices textuais referenciados resultantes da busca, pois *Information Retrieval* apresenta maior poder de esclarecimento semântico do texto referenciado do que simplesmente *Information* ou *Web*. Com ρ = 0,70 a busca não melhorou em termos de resultados, retornando apenas o termo **Information**; por algum motivo, a rede ART1, ao tentar especializar um pouco mais a clusterização, apresentou um resultado teoricamente pior que com ρ = 0,65. Como explicação para o caso, pode-se especular que nesta sessão os índices semanticamente mais bem obtidos na sessão anterior foram agregados a outros *clusters* formados pela rede, concluindo-se que talvez o processamento de várias sessões computacionais, cada uma com um valor de ρ diferente, ampliando-se o conjunto dos sintagmas nominais resultantes mediante a composição dos resultados das várias sessões, seja a solução mais adequada para se evitarem problemas dessa natureza.

[4] O sintagma apresentado na tabela é a variante original que serviu como índice do texto na base e não, necessariamente, o sintagma variante agregado ao *cluster* pela rede ART1. Como em ambos os casos refere-se ao mesmo índice sintagmático, não há problema em se apresentarem os resultados dessa forma.

TABELA 1
Informações recuperadas nas consultas 1, 2 e 3

Parâmetros de Consulta do Usuário Simulado			Resultados		
Nº	**Sintagmas de Busca**	**ρ**	**npa**	***Clusters* e Componentes**	**Índices Sintagmáticos Recuperados**
1	Semantic Web; Information Retrieval System; Digital Library.	0,60	144	*Cluster* 141: [449]	Wireframing; Customer Expectation; **Web** Development.
		0,65	173	*Cluster* 163: [426, 427]	426: IA Practices; **Information** Architects; Local IA Groups. 427: Wireframe; **Information** Architect; IA Community.
				Cluster 166: [436]	Consistency and Indexing; **Information Retrieval**; **Information** Architecture.
		0,70	192	*Cluster* 180: [426, 432]	426: IA Practices; **Information** Architects; Local IA Groups. 432: Wireframe; **Information** Architect; IA Community.
2	Web Design; Vignette; Information Architects.	0,70	190	*Cluster* 80: [177, 197]	177: **Information Architecture**; Creative **Design; Design** Insights. 197: **Information Architects;** Skills; Experience.
				Cluster 86: [193]	**Information Architects;** Skills; Experience.
				Cluster 172: [401]	Digital Library; Museum; Collections.
				Cluster 180: [426, 432]	426: IA Practices; **Information Architects**; Local IA Groups. 432: Wireframe; **Information Architect**; IA Community.
				Cluster 184: [437]	Consistency and Indexing; **Information** Retrieval; **Information Architecture**.
		0,75	220	*Cluster* 39: [77, 83]	77: **Web Design**; Shared References; **Information Architects.** 83: **Web Design**; User Needs; Product **Information**.
				Cluster 41: [80, 96]	80: **Web Design**; User Needs; Product **Information**; 96: **Information** Findability; **Information Architects**; User Needs.
				Cluster 99: [195, 224]	195: **Information Architects**; Skills, Experience. 224: Enterprise **Websites**, Intranets, **Information Architects**.
				Cluster 208: [424, 429]	424: IA Practices; **Information Architects**; Local IA Groups. 429: Wireframe; **Information Architect**; IA Community.
		0,80	265	*Cluster* 120: [194, 227]	194: **Information Architects**; Skills, Experience. 227: Enterprise **Websites**, Intranets, **Information Architects**.
3	Rich Internet Application; Information Findability; User Needs.	0,65	173	*Cluster* 163: [426, 427]	426: IA Practices; **Information** Architects; Local IA Groups. 427: Wireframe; **Information** Architect; IA Community.
				Cluster 166: [436]	Consistency and Indexing; **Information** Retrieval; **Information** Architecture.
				Cluster 171: [449]	Wireframing; Customer Expectation; Web Development.
		0,80	267	*Cluster* 97: [154]	**Rich Internet Application**, stateless GUI; real-time activity.

Com valores de ρ mais altos, o sistema não produziu nenhum resultado, pois a rede ART1 não encontrou padrões similares para clusterização de padrões de entrada e padrões da base textual nesses níveis de especialização maiores. Decerto, quando submetida a uma base de índices maior, com milhares ou dezenas de milhares de índices, uma rede ART1 poderia encontrar mais padrões similares aos de entrada, com menor probabilidade de resultado nulo em uma consulta.

Análise da Consulta 2

Com os sintagmas nominais de busca **Web Design**, **Vignette** e **Information Architects,** o usuário simulado comandou a segunda consulta ao sistema experimental. Os três sintagmas nominais introduzidos no sistema nesta consulta apresentam maior similaridade com índices sintagmáticos da base, esperando-se, portanto, resultados melhores que no caso da consulta anterior.

A rede ART1 com ρ = 0,70 produziu como resultados os seguintes índices sintagmáticos de textos: **Information Architecture; Creative Design**; **Design Insights** (177), **Information Architects**; **Skills; Experience** (197), **Digital Library**; **Museum**; **Collections** (401), **IA Practices**; **Information Architects**; **Local IA Groups** (426), **Wireframe**; **Information Architect**; **IA Community** (432), **Consistency and Indexing**; **Information Retrieval**; **Information Architecture** (437). O sintagma nominal *Information Architects* revelou-se predominante como padrão de entrada, encontrando vários índices similares em vários *clusters* formados pela rede. Os resultados apresentados pelo sistema simulado possibilitam ao usuário decidir se o conteúdo semântico que mais se aproxima de suas necessidades se encontra em um ou outro texto referenciado, ou se vários deles lhe serão úteis. Exemplos: (i) o usuário poderá decidir pela recuperação do texto com índice composto pelos sintagmas *Information Architects, Skills, Experience,* se o seu contexto de interesse for o que relaciona os profissionais Arquitetos da Informação às habilidades *(skills)* e experiências dessa ocupação; ou (ii) decidir pela recuperação dos textos com índices *IA Practices, Information Architects, Local IA Groups* e *Wireframe, Information Architect, IA Community,* caso o interesse seja mais pelo estudo de aspectos sociotécnicos da profissão de Arquiteto da Informação.

Os demais textos referenciados também poderão ser úteis, caso o usuário entenda que textos indexados com, inclusive, o sintagma *Information Architecture* apresentem possíveis conteúdos de seu interesse no contexto, tais como *Creative Design* e *Design Insights*, que aparecem no primeiro texto recuperado (mas que não contêm o sintagma *Information Architects* como índice).

Com esse tipo de consulta, demonstra-se que redes ART1 também podem reconhecer, em uma base de dados textual referenciada, padrões próximos aos padrões de entrada, conferindo-lhe alguns recursos (ainda que limitados) de processamento de linguagem natural conhecidos como *lematização* (redução de um termo flexionado a um termo mais primitivo).

Os resultados dessa consulta com parâmetros ρ = 0,75 e ρ = 0,80 se mostraram melhores ainda, quando o sistema recuperou referências semanticamente mais completas de textos da base. Com ρ = 0,75 o sistema recuperou os textos indexados com os sintagmas **Web Design**, **Shared References**, **Information Architects** (77), **Web Design**, **User Needs**, **Product Information** (83), **Information Findability**, **Information Architects**, **User Needs** (96), **Information Architects**, **Skills**, **Experience** (195), **Enterprise Websites**, **Intranets**, **Information Architects** (224), **IA Practices**, **Information Architects**, **Local IA Groups** (424) e **Wireframe**, **Information Architect**, **IA Community** (429). Ou seja, além de ampliar o conteúdo semântico dos índices de textos recuperados apresentando itens com mais de um sintagma ou termo identificado com a consulta, a rede ART1 formou um conjunto de índices novos que incluiu também

os obtidos na consulta na sessão anterior com ρ = 0,70, demonstrando seu poder de clusterização com vários níveis de generalização taxonômica. A sessão de consulta com ρ = 0,80 especializou os conteúdos recuperados, concentrando-se em apenas dois textos apresentados em sessões anteriores.

O texto recuperado que melhor deverá atender, teoricamente, às necessidades do usuário simulado, conforme os resultados desta consulta, é o de índice número 77, coincidindo com o argumento da consulta em dois dos três sintagmas do modelo de indexação: *Web Design* e *Information Architects*.

Análise das consultas 3, 4, 5 e 6

Os resultados das consultas 3, 4, 5 e 6 (tabelas 1 e 2) apresentaram-se como nas consultas anteriores, demonstrando a capacidade de redes neurais artificiais ART1 na recuperação de padrões textuais indexados e codificados com padrão de representação binária. Observou-se que os vetores mais completos em termos de símbolos significantes (padrões binários que representam termos e não espaços em branco) produziram *clusters* a partir de valores de ρ mais baixos (0,65 nos casos das consultas 3, 4 e 6), ao passo que o vetor mais curto em termos de *bits* significantes (consulta 5) produziu um *cluster* útil somente com valor de ρ bem mais alto (0,80, no caso). Como consequência dessa inferência e com os resultados das sessões de busca na consulta 4, por exemplo, resta evidente que, com valores de ρ mais baixos, redes ART1 podem reconhecer termos (ou fragmentos) de sintagmas e com valores mais altos podem reconhecer sintagmas completos, como nos índices *Yahoo,* **User Needs**, **Faceted** *Browse System* e **Web Design**, **User Needs**, *Product Information.*

A rede também reconhece, com precisão, padrões de índices completos como o da consulta 6, na qual propositalmente se realizou uma consulta de índice existente na base textual *in totum*, ou seja, um índice com os três sintagmas idênticos à busca.

Revocação e precisão

Define-se *revocação* como uma métrica de conjuntos que mostra a relação entre o número de documentos (ou registros) relevantes recuperados e o número de documentos relevantes existentes na base ou repositório de informação. E *precisão* como a relação entre o número de documentos relevantes recuperados e o número total de documentos recuperados. Os resultados do experimento anotados segundo essas duas métricas são apresentados na tabela 3, a seguir.

É importante observar-se que em buscas com sintagmas nas quais a rede reconhece padrões por aproximação na forma vetorial, como é o caso de ART1, a revocação é bem maior que a precisão, algo que, do ponto de vista de uma solução para busca textual semântica, é natural. Observe-se, conforme os dados da tabela 3, que quanto maior o parâmetro de vigilância ρ, melhor o resultado em termos de precisão e menor o resultado da revocação, tornando a busca mais específica.

O critério de cálculo de revocação e precisão adotado no experimento assume que apenas os índices recuperados da base textual com dois ou mais sintagmas nominais coincidentes com os sintagmas da consulta serão considerados, em termos de relevância (variável *N* na tabela 3), para uso nas fórmulas matemáticas de cálculo de revocação e precisão (*R* corresponde ao número de documentos recuperados pela rede):

Revocação: $| \mathbf{N} \cap \mathbf{R} | / | \mathbf{N} |$ e
Precisão: $| \mathbf{N} \cap \mathbf{R} | / | \mathbf{R} |$

As consultas 3-B e 4-B, na tabela 3, referem-se aos mesmos sintagmas das consultas 3-A e 4-A, mas com os espaços em branco para completar o tamanho do vetor (30 símbolos de 5 *bits* cada, ou 150 bits), tanto no padrão de entrada como no respectivo índice da base textual, preenchidos com os símbolos léxicos, mapeados para o código binário, dos mesmos sintagmas. O sintagma *Web design*, de apenas 10 símbolos significantes

TABELA 2

Informações recuperadas nas consultas 4, 5 e 6

Parâmetros de Consulta do Usuário Simulado			Resultados		
Nº	**Sintagmas de Busca**	**ρ**	**npa**	***Clusters* e Componentes**	**Índices Sintagmáticos Recuperados**
4	Web Design; User Needs; Faceted Classification.	0,65	172	*Cluster* 171: [449]	Wireframing; Customer Expectation; **Web** Development.
		0,75	221	*Cluster* 15: [27]	CMS Watch; **Web** Content Management; WCM Space.
				Cluster 100: [199]	**Faceted Classification;** Content Analysis; Metadata Schema.
		0,80	265	*Cluster* 125: [203]	**Faceted Classification;** Content Analysis; Metadata Schema.
				Cluster 137: [222]	Yahoo; **User Needs**; **Faceted** Browse System.
		0,85	319	*Cluster* 62: [83]	**Web Design**; **User Needs**; Product Information.
5	Findability; Search Engine; Google.	0,70	190	*Cluster* 172: [401]	Digital Library; Museum; Collections.
		0,75	221	*Cluster* 172: [342]	Designers; Design Messages; Communication Theory.
		0,80	265	*Cluster* 258: [443]	**Findability**; Ubiquitous Computing; User Experience.
6	Information Findability; Information Architects; User Needs.	0,60	141	*Cluster* 27: [76]	Web Design; Shared References; **Information Architects**.
				Cluster 30: [84, 92, 123, 163]	84: Web Design; **User Needs**; Product **Information**. 92: **Information Findability**; **Information Architects**; **User Needs**. 123: **User** Interfaces; **Information** Systems; **Information Architects**. 163: **Information** Platform; Mobile Devices; **Information Architects**.
				Cluster 34: [195]	**Information Architects**; Skills; Experience.
				Cluster 72: [224]	Enterprise Websites; Intranets; **Information Architects**.
		0,70	189	*Cluster* 36: [78, 91]	78: Web Design; Shared References; **Information Architects**. 91: **Information Findability**; **Information Architects**; **User Needs**.
				Cluster 39: [84]	Web Design; **User Needs**; Product **Information**.
				Cluster 98: [225]	Enterprise Websites; Intranets; **Information Architects**.
		0,80	263	*Cluster* 47: [70, 94, 228]	70: **Information** Science; **Information** Architecture; Content Genres. 94: **Information Findability**; **Information Architects**; **User Needs**. 228: Enterprise Websites; Intranets; **Information Architects**.
				Cluster 49: [75]	Web Design; Shared References; **Information Architects**.
				Cluster 59: [91, 124]	91: **Information Findability**; **Information Architects**; **User Needs**. 124: **User** Interfaces; **Information** Systems; **Information Architects**.
				Cluster 60: [95]	**Information Findability**; **Information Architects**; **User Needs**.

(contando com o espaço em branco entre os termos *Web* e *design*), codificado para o código binário utilizado seria (o símbolo ¢ se refere a espaço em branco) **Web¢design¢¢¢¢¢¢¢¢¢¢¢¢¢¢¢¢¢¢¢¢¢¢**.

O recurso para aumento da precisão empregado para melhoria do modelo de representação da informação vetorial codificaria esse mesmo sintagma como **Web¢design¢Web¢design¢ Web¢desi**.

Observam-se, nos resultados do experimento, percentuais de revocação altos, alcançando até 100%, com percentuais de precisão baixos, de 10% a 20%. É de se esperar que, com uma base de índices em escala de produção (com dezenas ou centenas de milhares de textos indexados) e uma codificação completa, sem espaços em branco, em todos os vetores dos sintagmas de consulta e da base, os resultados apresentem precisão maior, pois, nesse cenário, (i) existirão mais índices com sintagmas similares e (ii) os vetores serão mais diferenciados entre si, reduzindo-se a probabilidade de clusterizações com base nos espaços em branco dos vetores binários (espúrias, do ponto de vista da significância).

CONCLUSÕES E QUESTÕES EM ABERTO

O experimento simulado demonstrou a capacidade de aprendizado de redes neurais artificiais modelo ART1 no reconhecimento de padrões de dados binários referentes a bases de textos indexados com sintagmas nominais em um contexto de implementação de sistemas de recuperação da informação baseados em computadores. Com esse experimento, demonstrou-se também a utilidade de simulações computacionais na ciência da informação, uma metodologia de pesquisa que tem origem nas engenharias e na administração, mas que pode ser também eficiente em contextos de uso das ciências sociais.

Com esse tipo de rede neural artificial, amplia-se o leque de possibilidades de integração e sinergia entre tecnologias de mineração de textos,

TABELA 3
Revocação e precisão com ART1

Consultas do Usuário Simulado		Dados Observados		*Cálculos*	
Nº	**Sintagmas de Consulta**	**N**	**R**	**Revocação**	***Precisão***
1	Semantic Web; Information Retrieval System; Digital Library.	0	4	n. d.[5]	*0*
2	Web Design; Vignette; Information Architects.	1	10	100%	*10%*
3-A	Rich Internet Application; Information Findability; User Needs.	1	5	0	*0*
4-A	Web Design; User Needs; Faceted Classification.	2	5	50%	*20%*
5	Findability; Search Engine; Google.	0	3	n. d.	*0*
6	Information Findability; Information Architects; User Needs.	1	9	100%	*11%*
3-B	Rich Internet Application; Information Findability; User Needs.	1	1	100%	*100%*
4-B	*Web Design; User Needs; Faceted Classification.*	*1*	*3*	*100%*	*33%*

[*] Resultado matematicamente não definido (divisão de zero por zero).

recuperação da informação e processamento da linguagem natural. As redes ART1, fundamentadas na *Teoria da Ressonância Adaptativa*, apresentam propriedades que são características funcionais úteis na construção de um sistema de recuperação da informação, tais como aprendizado não supervisionado, controle do grau de generalização dos conteúdos dos *clusters*, autoescalabilidade, autoestabilização com pequeno número de iterações computacionais, aprendizado *on-line*, captura de eventos raros, acesso direto a padrões de entrada similares, acesso direto a subconjuntos e padrões de subconjuntos e enviesamento da rede para formação de novas categorias. O aprendizado não supervisionado é essencial nesse tipo de implementação e talvez seja sua propriedade mais importante, dispensando a necessidade de qualquer interação do usuário com o sistema que não seja a inclusão dos argumentos (sintagmas nominais) de busca em linguagem natural e do nível de generalização de conteúdos pretendido. Outra propriedade bastante útil em contextos de sistemas de recuperação da informação é a capacidade de aprendizado *on-line* das redes ART1, podendo criar novos padrões ou classificar padrões de entrada em padrões conhecidos (aprendidos anteriormente) sem necessidade de nova etapa de aprendizado prévio com processamento *off-line (batch)*.

Outra característica de redes ART1 observada em Capuano (2001) e confirmada no experimento apresentado neste artigo é o reduzido número de iterações de seu algoritmo computacional (geralmente em torno de duas iterações), que propicia um tempo de resposta relativamente reduzido (no pior caso, com valores de ρ mais altos, a rede consumiu apenas 45 segundos no processamento dos dados). Os experimentos anteriores com redes ART1 sugerem que esse tipo de tecnologia de processamento paralelo é bastante escalável, sendo mais sensível, em termos de desempenho computacional, a operações de clusterização mais específicas (com valores de ρ mais próximos de 1,0) do que ao aumento do número de registros (vetores) na base de dados.

Embora não constitua um item de prova de conceito de uma grande implementação de sistema de recuperação da informação, os tempos de respostas parecem menores que os tempos de resposta observados em redes neurais com outras arquiteturas, especialmente os modelos mais populares conhecidos como *backpropagation*. Com a contínua redução histórica dos custos de *hardware* computacional previstos na Lei de Moore, é de se esperar que sistemas de recuperação da informação com esse tipo de tecnologia de rede neural se tornem economicamente viáveis e possam ser implementados corporativamente, no modo de aprendizado *on-line*, possibilitando o acesso de milhares de usuários simultâneos.

Os possíveis esquemas de codificação binária, embora eficazes, podem se tornar mais eficientes à medida que os padrões vetoriais possam ser representados de modo mais discriminatório, sem lacunas que necessitem de preenchimento com fragmentos de *strings* de *bits* comuns (como os espaços em branco após o término da parte significativa dos padrões binários, por exemplo).

O modelo de representação da informação adotado no experimento revelou-se adequado para atendimento dos requisitos enunciados em (MEADOW *et al.*, 2007, p. 64): (i) capacidade de discriminação entre padrões, (ii) capacidade de identificação de similaridades entre padrões, (iii) acurácia e (iv) capacidade de redução de ambiguidades. Entretanto, algumas questões permanecem em aberto para estudos e pesquisas em torno de soluções mais otimizadas para o problema da recuperação da informação textual com agregação de conteúdos semânticos. A mineração de textos, com metodologias e tecnologias *par excellence* nesse tipo de missão computacional, envolve modelos de representação da linguagem natural que necessitam de aperfeiçoamentos, como sugerido na análise crítica do presente experimento.

Artigo submetido em 07/12/2008 e aceito em 08/04/2009.

REFERÊNCIAS

BIGUS, Joseph P. *Data mining with neural networks*: solving business problems from application development to decision support. [São Paulo]: McGraw-Hill, 1996.

CAPUANO, Ethel Airton. *Mineração de dados com redes neurais não-supervisionadas modelo* ART: parâmetros e aplicações. 2001. Dissertação (Mestrado) - Universidade Católica de Brasília, Brasília, 2001.

_____; CHAUKE-NEHME, C. Exploring the parameter space of unsupervised ART neural networks for data mining. In: INTERNATIONAL CONFERENCE ON DATA MINING, 3., 2002. *Proceedings…* Southampton: WIT Press, 2002. p. 461-472.

CARPENTER, G. A. et al. Fuzzy-ARTMAP: a neural network architecture for incremental supervised learning of analog muldimensional maps. *IEEE Transactions on Neural Networks*, n. 3, p. 698-713, 1992.

_____; GROSSBERG, S.; ROSEN, D. B. Fuzzy ART: fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, n. 4, p. 759-771, 1991.

_____. The ART of adaptive pattern recognition by a self-organizing neural network. *IEEE*, Mar. 1988.

CHAUKE-NEHME, Cláudio. Foreword. In: PRADO, Hércules Antonio do; FERNEDA, Edilson (Org.). *Emerging technologies of text mining*: techniques and applications. Hershey: Information Science Reference, 2008. p. 11.

_____. *Modelo neurocomputacional de elementos centralizadores com interações laterais aplicado ao reconhecimento de padrões temporais*. 1996. Tese (Doutorado)- Universidade Federal do Rio de Janeiro, Rio de Janeiro, 1996.

FERNEDA, Edberto. Redes neurais e sua aplicação em sistemas de recuperação de informação. *Ciência da Informação*, Brasília, v. 35, n. 1, p. 25-30, jan./abr. 2006.

FREEDMAN, David H. *Los hacedores de cérebros*: cómo los científicos están perfeccionando las computadoras, creando um rival del cerebro humano. [S. l.]: Andres Bello, 1995.

FREEMAN, James A.; SKAPURA, David M. *Redes neuronales*: algoritmos, aplicaciones y técnicas de programación. Wilmington, USA: Addison-Wesley Iberoamericana, 1993.

GOTTSCHALG-DUQUE, Cláudio. *SiRILiCO*: uma proposta para um sistema de recuperação de informação baseado em teorias da linguística computacional e ontologia. 2005. Tese (Doutorado)- Escola de Ciência da Informação da Universidade Federal de Minas Gerais, Belo Horizonte, 2005.

GROSSBERG, Stephen. Non-linear neural networks: principles, mechanisms, and architectures. *Neural Networks*, v. 1, 1988.

KENT, A. et al. Machine literature searching VII: operational criteria for designing information retrieval systems. *American Documentation*, v. 6, n. 2, p. 93-101, 1955.

KONCHADY, Manu. *Text mining application programming*. Boston: Charles River Media, 2006.

MEADOW, Charles T. et al. *Text information retrieval systems*. 3rd ed. [S. l.]: Elsevier, 2007.

PENTEADO, Roberto; BOUTIN, Eric. Creating strategic information for organizations with structured text. In: PRADO, Hércules Antonio do; FERNEDA, Edilson (Org.). *Emerging technologies of text mining*: techniques and applications. Hershey: Information Science Reference, 2008. p. 34-53.

PIDD, Michael. *Computer simulation in management science*. 2nd ed. [S. l.]: John Wiley & Sons, 1988.

PRADO, Hércules Antonio do; FERNEDA, Edilson (Org.). *Emerging technologies of text mining*: techniques and applications. Hershey: Information Science Reference, 2008.

SERRANO-GOTARREDONA, Teresa; LINARES-BARRANCO, Bernabé; ANDREOU, Andreas G. *Adaptive resonance theory microchips*: circuit design techniques. Boston: Kluwer Academic, 1998.

TSURUOKA, Yoshimasa. et al. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics Advance Access*, p. 1-6, 12 ago. 2007.

YU, Lean; WANG, Shouyang; LAI, Kin Keung. A multi-agent neural network system for web text mining. In: PRADO, Hércules Antonio do; FERNEDA, Edilson (Org.). *Emerging technologies of text mining*: techniques and applications. Hershey: Information Science Reference, 2008. p. 162-183.