Uso de Redes Neurais Artificiais para processos de Recuperação de Informação

Nathalie Portugal Vargas

Sumário

- Introdução
- Trabalhos Relacionados
  - Recuperação da Informação com redes ART1
  - Mineração de Dados com Redes SOM
  - RNA na extração da Informação
  - Filtragem de Informação com Redes Hopfield
- Considerações Finais

Introdução

- A **internet** é o meio mais utilizado para **extrair informações** úteis. Estas informações pelo geral estão em forma de artigos, relatórios, notícias, etc.

- O crescimento da informação vem motivando o **desenvolvimento de novas técnicas de extração** para melhorar os existentes modelos e algoritmos de Recuperação de Informação.

- Os atuais sistemas de **Recuperação de Informação** (RI) têm sido desenvolvidos para **facilitar a pesquisa e extração de dados**. Baseados em palavras chave.

Introdução

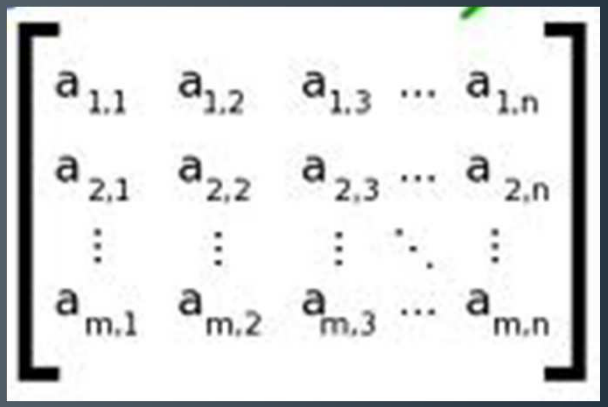
- O principal **objetivo** de um sistema de RI é **recuperar informação** (contida nos documentos) que possa ser útil ou **relevante para o usuário.**

- Tal informação é normalmente chamada de **necessidade de informação** do usuário.

- Caracterizar a necessidade de informação do usuário **não é** uma tarefa **simples**
  - "Encontre todos os documentos contendo informações sobre a doença Neoplasma Benigno de forma que: (1) O paciente com a doença possua idade inferior a 50 anos e (2) seja diabético."

Introdução

- A informação pode ser tratada desde dois pontos de vista:
  - Estatístico
    - Abordagem rápida.
    - Computacionalmente Eficiente.
    - Não considera relações simbólicas.
    - Não efetua inferências
  - Simbólico
    - Abordagem custosa.
    - Computacionalmente Ineficiente.
    - Normalmente utilizado com Cadeias Markov



Recuperação da Informação com Redes ART1 (Capuano,2000)

- Rede ART
  - Apresenta um **padrão** → Se faz **ressoar** com os protótipos das categorias conhecidas pela rede → Se o padrão **entra em ressonância** com alguma classe então **é associado** → O **centro de cluster é deslocado ligeiramente** para se adaptar melhor ao novo padrão.

  - Capa de saída **estática** = **saturação** pois não pode criar uma nova classe para o padrão apresentado
  - Capa de saída **dinâmica** = criar-se-á uma **nova classe** para dito padrão e isto não afetará às classes já existentes.

Recuperação da Informação com Redes ART1 (Capuano,2000)

- Experimento de simulação computacional de um **sistema de RI** composto por uma **base de índices textuais** de uma amostra de documentos, um **software** de Rede Neural Artificial implementando conceitos da **ART**.

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						

Recuperação da Informação com Redes ART1 (Capuano,2000)

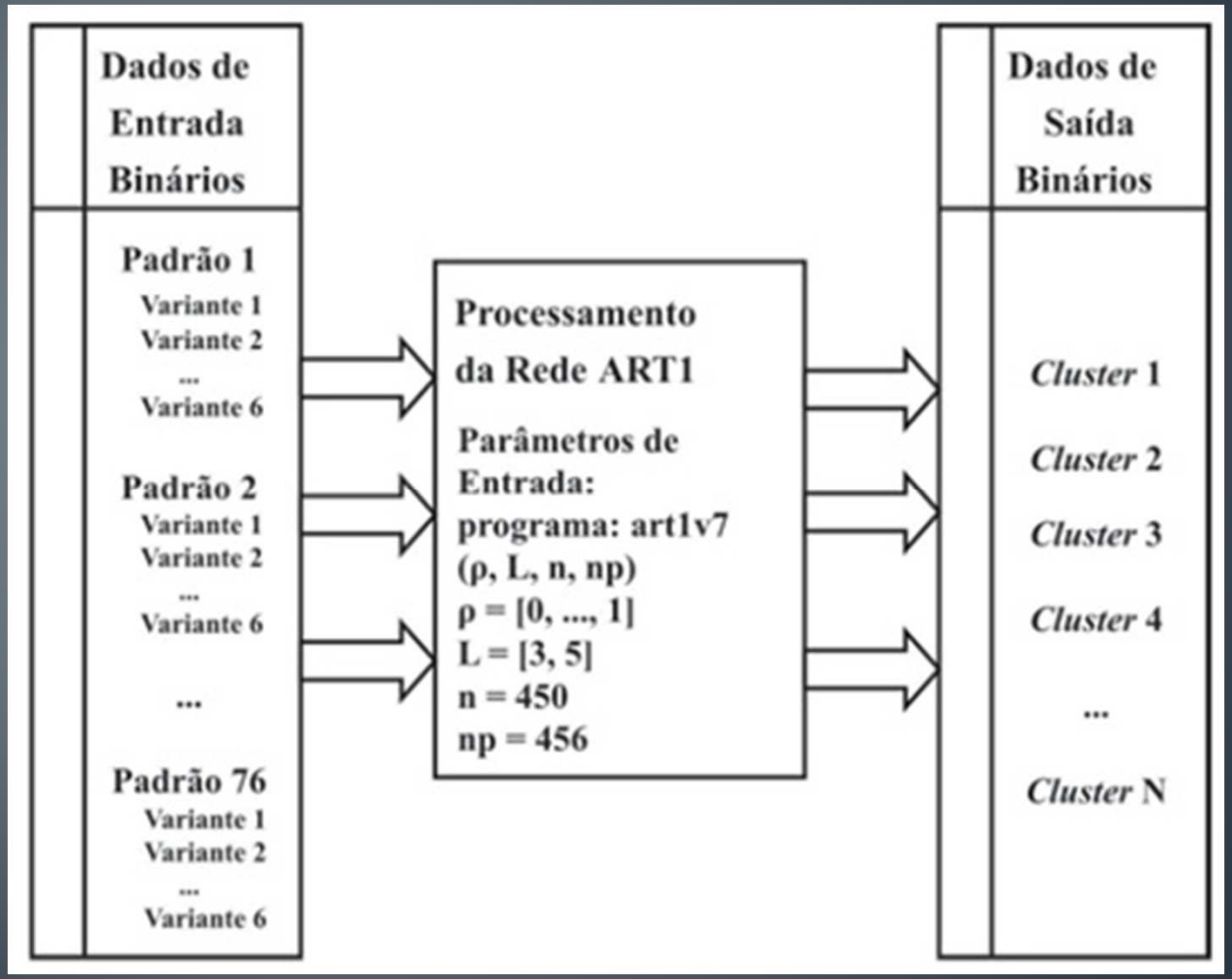
	A	B	C	D	E	F	G	H	I	J	...
	00000	00001	00010	00011	00100	00101	00110	00111	01000	01001	...
	1	2	3	4	5	6	7	8	9	10	...
1	Social classification, freetagging, metadata for the masses.		10010	01110	00010	01000	00000	01011	11010	00010	...
2	Social classification, metadata for the masses, freetagging.		10010	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	...

Recuperação da Informação com Redes ART1 (Capuano,2000)

- Redes ART1 operam com vetores dimensionais de posições fixas.
  - Quantidade Padrões Testados = 456. Sintagmas Nominais = 3.

  - Usuário informa inicialmente os parâmetros de pesquisa.
    - CAMPO 1: Rich Internet Application
    - CAMPO 2: Information Findability
    - CAMPO 3: User Needs
  - Consulta estendida
    - 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;

Recuperação da Informação com Redes ART1 (Capuano,2000)



Recuperação da Informação com Redes ART1 (Capuano,2000)

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.

Mineração de Dados com Redes SOM (Zuchini, 2003)

- Representar um **conjunto de palavras** por vectores de forma que seu **significado semântico** seja de alguma forma, **capturado pelo mapa neural.**
  - Símbolos semanticamente próximos sejam mapeados topograficamente próximos.
- Assume-se que a **palavra** em si **não carrega seu significado,** mas este **depende** principalmente do **contexto** em que ela está inserida.
- **Vetor composto** pela concatenação de outros dois, um representando respectivamente o **símbolo** em si (a palavra) e outro, o **conjunto de atributos** (contexto) associado ao símbolo.

Mineração de Dados com Redes SOM (Zuchini, 2003)

- Os símbolos constantes são representados pela concatenação de seus atributos e de um código para cada símbolo.
- Ao vetor de atributos é concatenado um vetor para a codificação do símbolo.

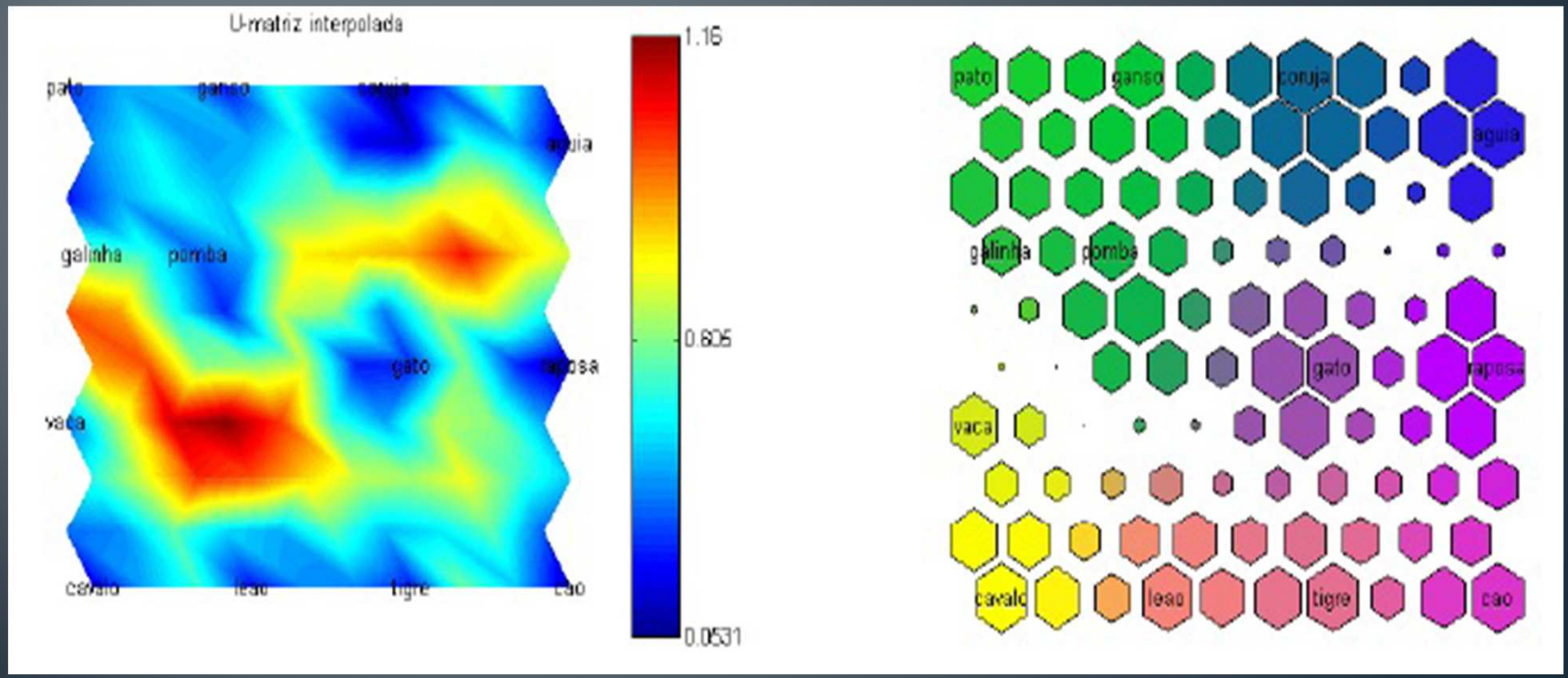
		pomba	galinha	pato	ganso	coruja	falcão	águia	raposa	cão	lobo	gato	tigre	leão	cavalo	zebra	vaca
é	pequeno	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0
	médio	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
	grande	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
tem	2 pernas	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
	4 pernas	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
	pelos	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
	cascos	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
	crina/juba	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0
	penas	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
gostam de	caçar	0	0	0	0	1	1	1	1	0	1	1	1	1	0	0	0
	correr	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0
	voar	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0
	nadar	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0

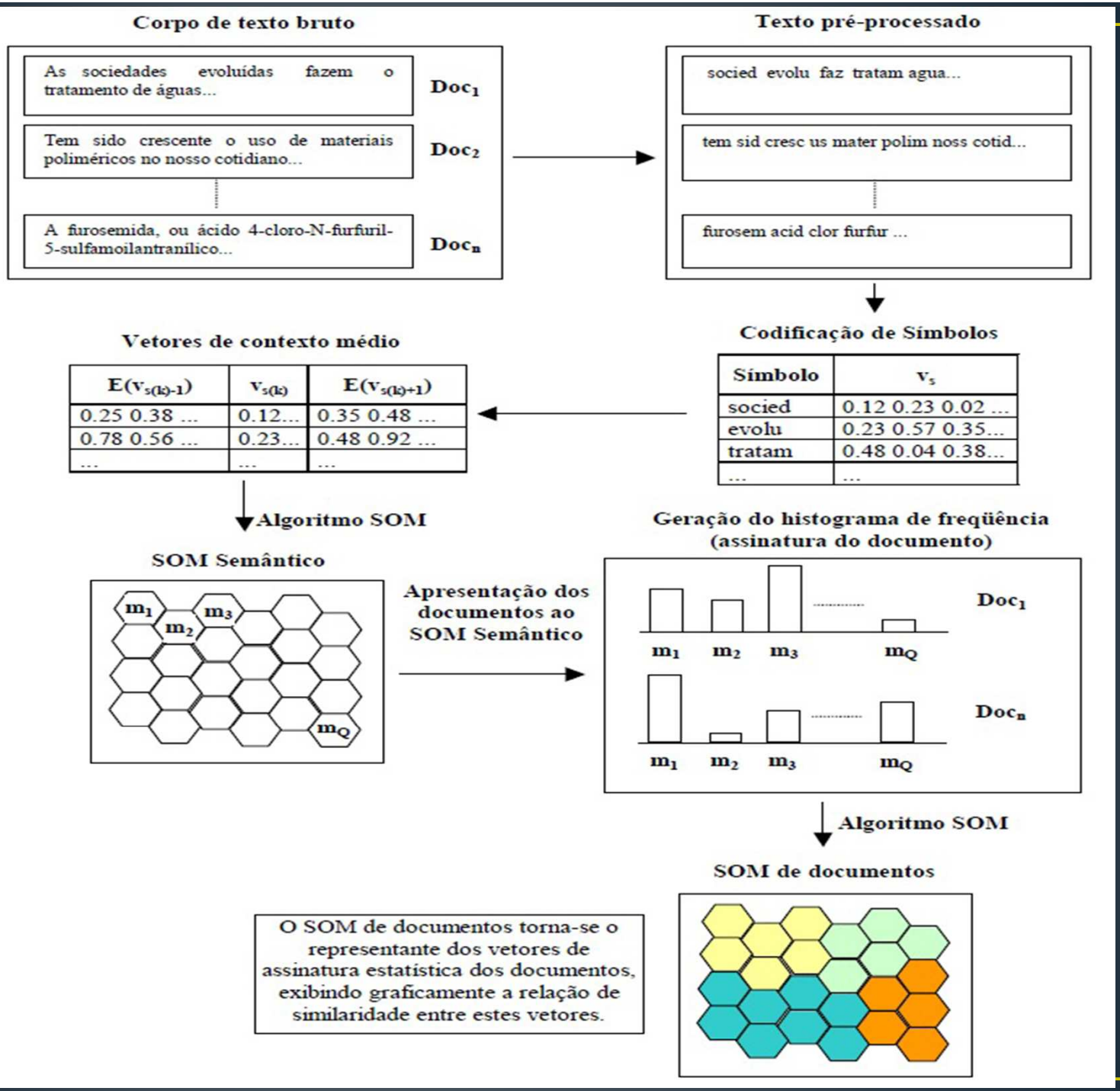
Mineração de Dados com Redes SOM (Zuchini, 2003)

Código dos atributos $(v_a)^T$												Código do símbolo $(v_s)^T$																Símbolo
1	0	0	1	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	pomba
1	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	galinha
1	0	0	1	0	0	0	0	1	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	pato
1	0	0	1	0	0	0	0	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	ganso
1	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	coruja
1	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	falcão
0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	águia
0	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	raposa
0	1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	cão
0	1	0	0	1	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	lobo
1	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	gato
0	0	1	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	tigre
0	0	1	0	1	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	leão
0	0	1	0	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	cavalo
0	0	1	0	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	zebra
0	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	vaca

Mineração de Dados com Redes SOM (Zuchini, 2003)

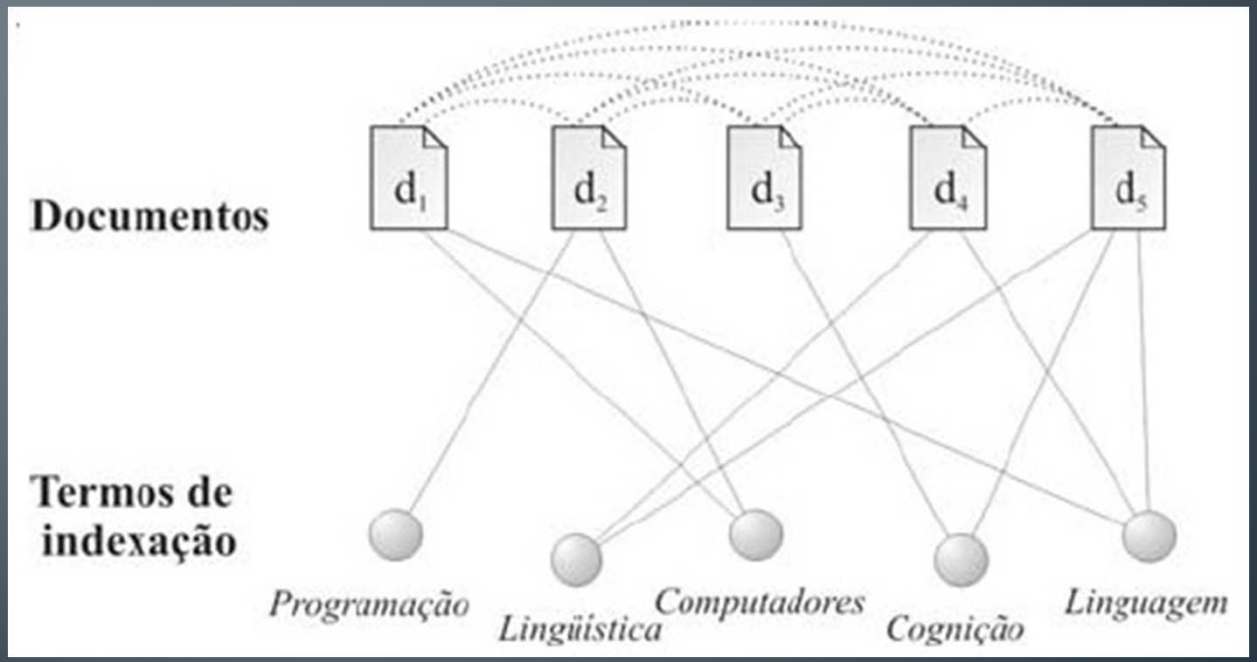
- Objetivo final é obter um SOM que organize os documentos de texto conforme sua proximidade contextual



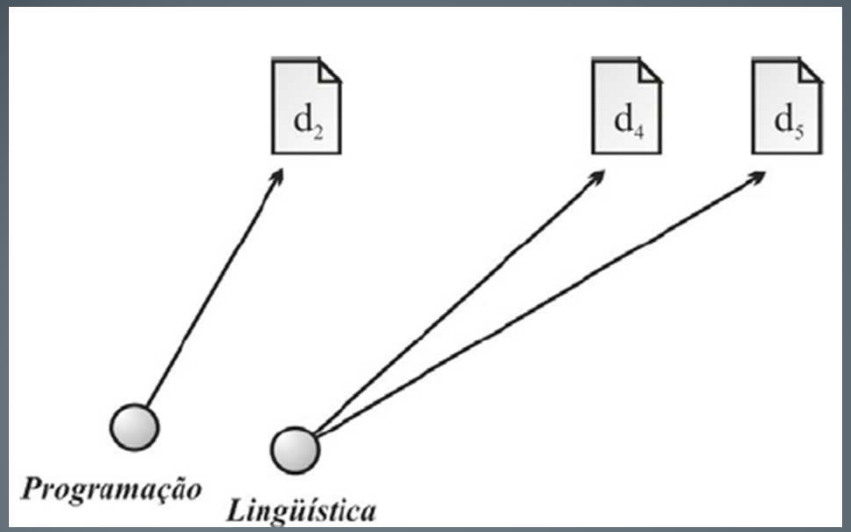

Corpo de texto bruto
As sociedades evoluídas fazem o tratamento de águas...
Doc1
Tem sido crescente o uso de materiais poliméricos no nosso cotidiano...
Doc2
A furosemida, ou ácido 4-cloro-N-furfuril-5-sulfamoilantranílico...
Docn
Texto pré-processado
socied evolu faz tratam agua...
tem sid cresc us mater polim noss cotid...
furosem acid clor furfur ...
Codificação de Símbolos
Símbolo
vs
socied
0.12 0.23 0.02 ...
evolu
0.23 0.57 0.35...
tratam
0.48 0.04 0.38...
Vetores de contexto médio
E(vs(k)-1)
vs(k)
E(vs(k)+1)
0.25 0.38 ...
0.12...
0.35 0.48 ...
0.78 0.56 ...
0.23...
0.48 0.92 ...
Algoritmo SOM
SOM Semântico
m1
m3
m2
mQ
Apresentação dos documentos ao SOM Semântico
Geração do histograma de freqüência (assinatura do documento)
Doc1
Docn
m1
m2
m3
mQ
Algoritmo SOM
SOM de documentos
O SOM de documentos torna-se o representante dos vetores de assinatura estatística dos documentos, exibindo graficamente a relação de similaridade entre estes vetores.

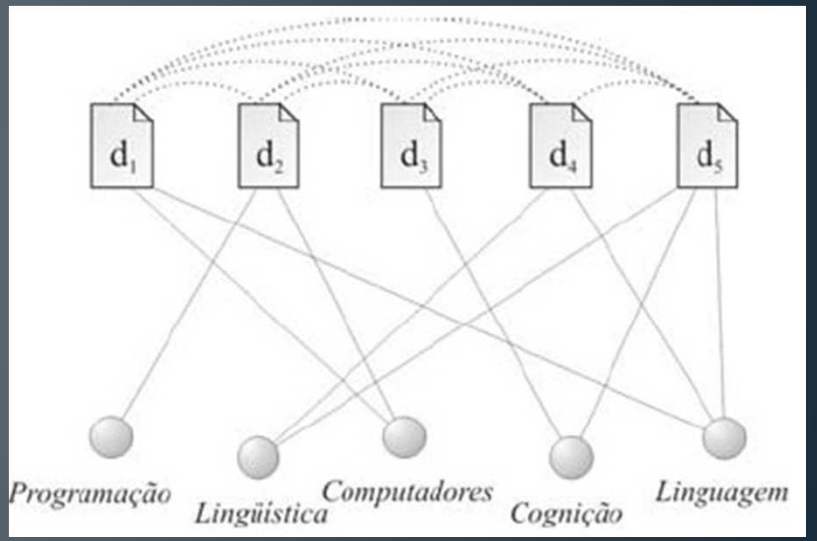
Redes Neurais na extração da Informação (Ferneda, 2006)

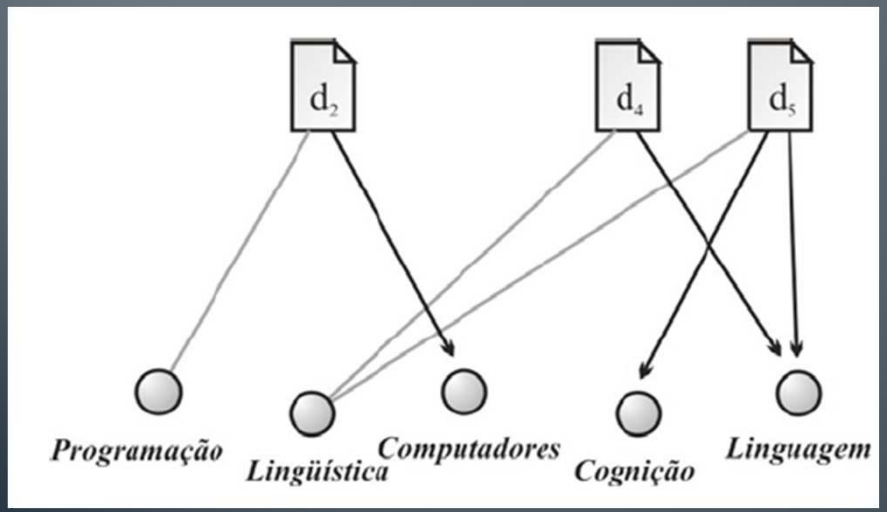
- Mozer (1984), utilizou uma arquitetura de Rede Neural para processos de RI, **sem capacidade de aprendizado.**
- Ás **ligações** entre os documentos são **inibitórias**, isto é, um documento, quando ativado, reduz o nível de ativação dos demais documentos



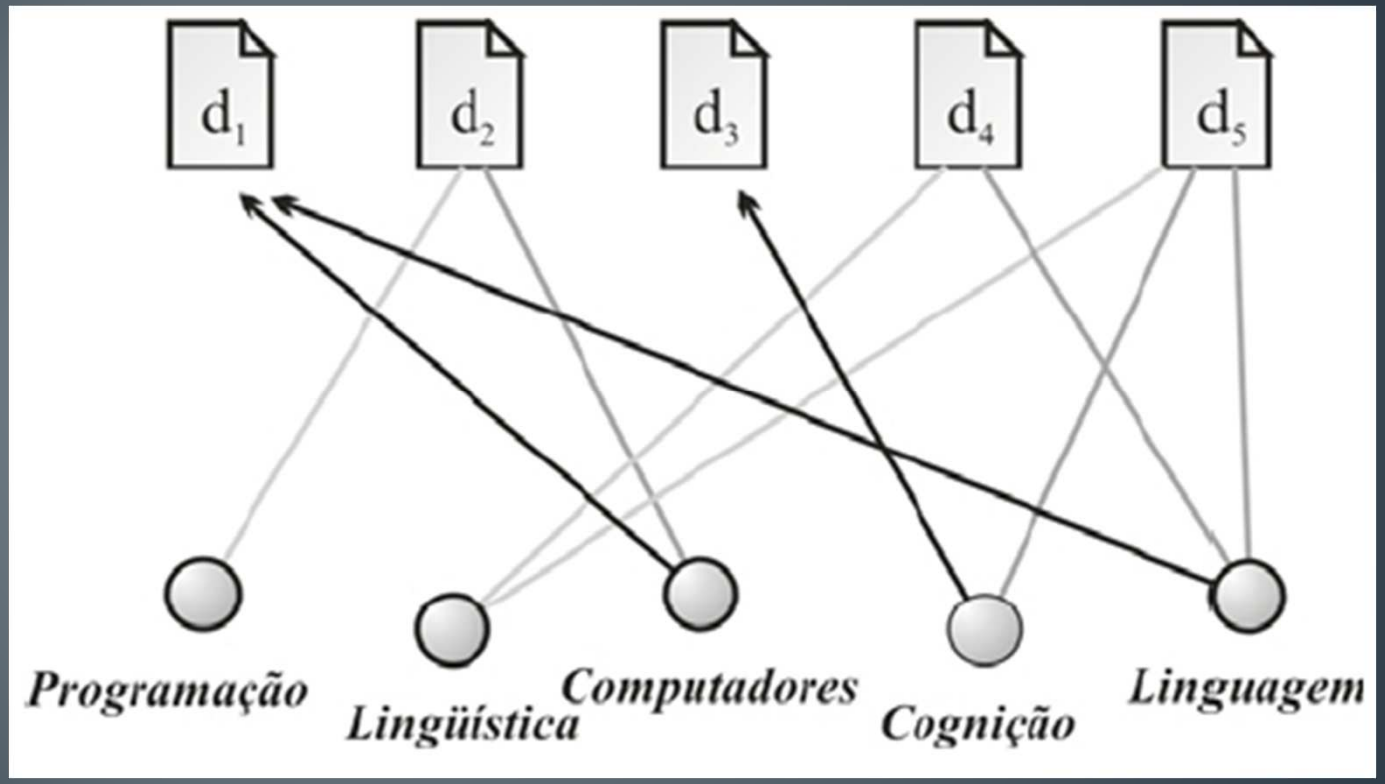
Redes Neurais na extração da Informação (Ferneda, 2006)





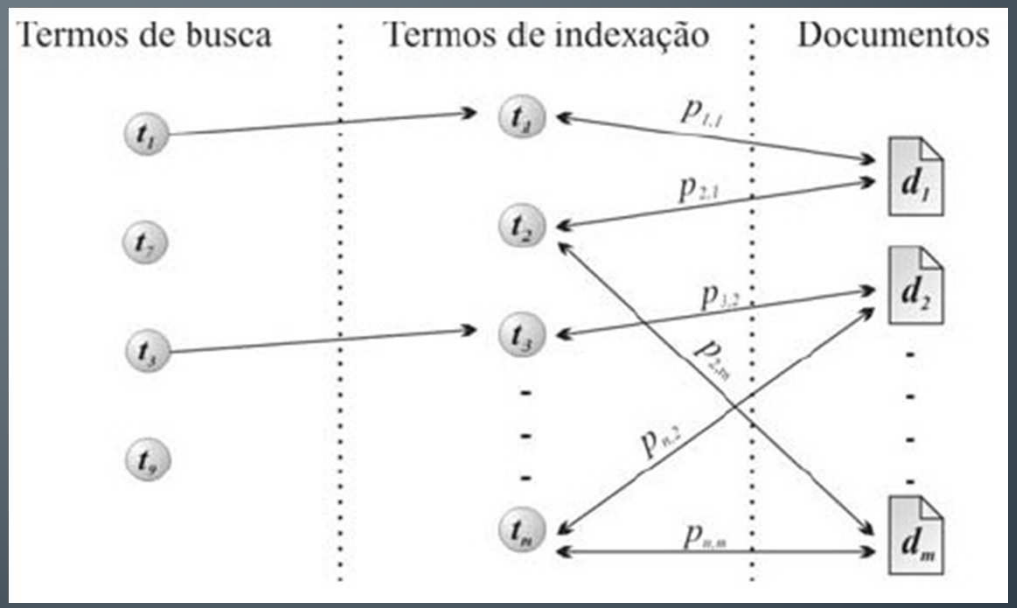


Redes Neurais na extração da Informação (Ferneda, 2006)



Redes Neurais na extração da Informação (Ferneda, 2006)

- Pode-se dizer que nesses sistemas se tem por um lado às expressões de busca e pelo outro os documentos, e no meio encontra-se os termos de indexação



Redes Neurais na extração da Informação (Ferneda, 2006)

- O resultado final será um **conjunto dos documentos** que foram **ativados**, cada um com um **nível de ativação** que pode ser interpretado como o **grau de relevância do documento** em relação à busca do usuário.

- Entre esses documentos resultantes, pode ocorrer que alguns **documentos não estejam diretamente relacionados** aos termos utilizados na expressão de busca, mas que possuem **certo grau de relacionamento com os termos de busca** dados pelo usuário

Filtragem de Informação utilizando Hopfield (Mandelli, 2010)

- Atribui-se uma importância maior nas necessidades dos usuários, sendo que a informação pode variar de usuário a usuário.
- Esses sistemas, tem que preencher três requisitos:
  - **Especialização** → Seleciona **documentos relevantes** e vai descartando os outros;
  - **Adaptação** → Considera a adaptação de **filtragem de informação** (interesses dos usuários são sujeitos a mudanças durante o tempo);
  - **Exploração** → Sistemas devem ser capazes de **explorar novos domínios**, a fim de encontrar algo novo, e potencialmente interessante para cada usuário

Filtragem de Informação utilizando Hopfield (Mandelli, 2010)

- É utilizado um agente pessoal que vasculha fontes de informação na web para recuperar documentos de acordo com os interesses dos usuários → Hopfilter.
- Etapas:
  - Usuário → **selecionar** manualmente uma **coleção de documentos** sobre um **assunto** de interesse
  - Mecanismo de Indexação Automática → Gera **termos que representem** melhor o **conteúdo** da coleção
  - Funções Estatísticas → **calcular a co-ocorrencia** dos termos na coleção, **gerando** uma **matriz de similaridade** (Rede de Conhecimento ou um Espaço de Conceitos)

Filtragem de Informação utilizando Hopfield (Mandelli, 2010)

- Um **documento** a ser filtrado é **usado** como um **padrão de entrada** para a rede.

- Quando a rede consegue **convergir**, a **quantidade de neurônios ativos** indica se o **documento é compatível** com os conceitos armazenados na memória da rede ou não. Determinando assim se o documento é relevante para o usuário.

- Método **Limitação** na **ativação inicial** dos neurônios da rede de Hopfield, já que eles são inicializados com 0 ou com 1.

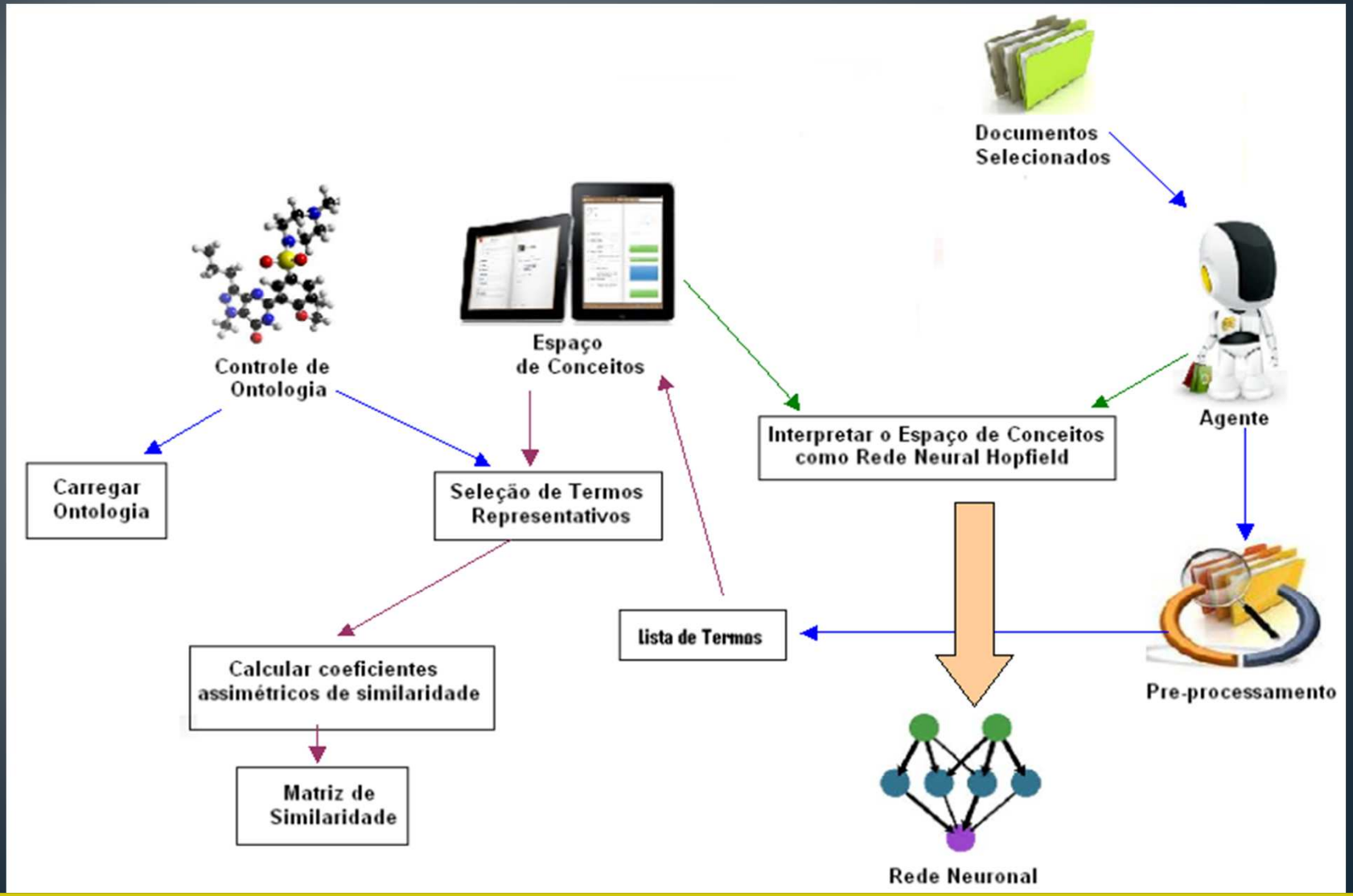
Filtragem de Informação utilizando Hopfield (Mandelli, 2010)

- Este problema pode-se solucionar utilizando uma ontologia.
- A ontologia fornece relações entre um termo no espaço de conceitos, possibilitando assim a determinação de uma relação entre um termo com outro documento.
- Assim é possível inicializar um neurônio não apenas com valor 0 ou 1, mas com qualquer valor no intervalo real [0,1]

Filtragem de Informação utilizando Hopfield (Mandelli, 2010)

- Espaço de Conceitos Gerado representa Rede Neural
  - Neurônios → Termos.
  - Pesos Sinápticos → Coeficientes de Relacionamento.
- É preciso escolher um espaço de conceitos específico.
- Palavras que não for encontradas no dicionário são levadas a uma tabela para correção de possíveis erros ortográficos, e palavras que não faz parte do dicionário podem ser incluidas.

Filtragem de Informação utilizando Hopfield (Mandelli, 2010)



Considerações Finais

- Neste trabalho foram abordadas algumas **técnicas de Redes Neurais** que **ajudam** no processo de **Recuperação de Informação**, cada uma dessas técnicas aborda de forma diferente a extração de dados.

- Pode-se observar que já desde os inicios das pesquisas com Redes Neurais, começou-se trabalhar com processos de Recuperação de Informação. No entanto, se tem que considerar que o **nível das tecnologias empregadas** foi **evoluindo** cada vez mais, ate o ponto das pesquisas atuais experimentar com **métodos híbridos** e novos conceitos que permitem melhorar os resultados explorados antigamente.

- Uma das **tendências** atuais nos processos de extração de dados, é a **utilização de conceitos semânticos**, que dão um sentido para a informação.

Referências Bibliográficas

- **Capuano, E. A (2000)**. O poder cognitivo das redes neurais artificiais modelo art1 na recuperação da informação. In Ci. Inf. Brasília
- **Daniel Mandelli Martins, J. M. A. C. (2010)**. Filtragem de informação na web usando rede neural de hopfield e ontologia. Anais do XV Encontro de Iniciação Científica da PUC - Campinas.
- **Ferneda, E. (2006)**. Redes neurais e sua aplicação em sistemas de recuperação de informação.
- **Mozer, M. (1984)**. Inductive information retrieval using parallel distributed computation. Technical Report 8406, University of California, San Diego - USA

Referências Bibliográficas

- **Scholtes, J. (1991).** Neural nets and their relevance for information retrieval. CL-1991-02, ITLI Prepublication Series for Computational Linguistics, Institute for Logic, Language and Computation (ILLC).

- **Zuchini, M. H. (2003).** Aplicações de mapas auto-organizáveis em mineração de dados e recuperação de informação. Dissertação de mestrado, Universidade Estadual de Campinas.

Uso de Redes Neurais Artificiais para processos de Recuperação de Informação

Nathalie Portugal Vargas