PROJETO ZIR – Estudo de Caso sobre a utilização da Metodologia DolphinSearch para Recuperação de Informações

Elias A. Zir Neto[1], Anita Maria da Rocha Fernandes[2], Benjamim Grando Moreira[3]

[1, 2, 3]Grupo de Inteligência Artificial – Universidade do Vale do Itajaí (UNIVALI)
Rua Uruguai, 458 – 88302-202 – Itajaí – SC – BRASIL

{eliaszirneto@yahoo.com.br, anita.fernandes@univali.br, nivali@gmail.com}

Abstract. *This paper presents this work a case-study of the DolphinSearch methodology utilization for Information Retrieving, which uses Artificial Neural Networks to retrieve information. This paper describes the methodology used to develop the project and to develop the Neural Network.*

Resumo. *O presente artigo apresenta um estudo de caso sobre a utilização da metodologia DolphinSearch para Recuperação de Informação, a qual utiliza Redes Neurais Artificiais para recuperar informações. Nele contêm uma descrição dos métodos utilizados durante o desenvolvimento do projeto, como a metodologia adotada, a seqüência de atividades seguidas e os métodos utilizados para o desenvolvimento do projeto.*

1. Introdução

Desde a década de quarenta o problema de armazenamento de informação, e sua posterior recuperação vem chamando a atenção. Segundo van Rijsbergen (1999) este problema é simples de se entender: tem-se uma vasta quantidade de informação para qual seu acesso rápido e preciso está ficando cada vez mais difícil. Um efeito disto é a perda de possíveis informações, que seriam relevantes, e que não serão encontradas.

A partir do momento que os computadores de alta velocidade tornaram-se disponíveis para o trabalho não numérico (cálculos) muitos acharam que eles poderiam "ler" uma coleção inteira de documentos e extrair apenas os documentos relevantes. Porém, o uso da linguagem natural no texto de um documento acarreta em problemas de entrada e armazenamento, pois o conhecimento apresentado no documento pode não estar explícito e sim de forma subentendida.

Librelotto, Ramalho e Henriques (2005) explicam que "os computadores são úteis para organização e processamento de dados, tipicamente mantendo as informações em hierarquias rígidas, enquanto a mente humana tem a habilidade especial de ligar pequenas unidades de informação de forma aleatória." Com base nesta constatação, a segunda geração da Web, cunhada como Semantic Web, apresenta um conceito de que a semântica seria fundamental, ao invés, de apenas a sintaxe para determinar uma busca. Então, se o desejo é que um computador possa entender o significado de uma expressão é necessário construir e representar um modelo que represente alguma parte do modelo mental humano.

Para tornar possível essa representação pode-se utilizar uma tecnologia para recuperação de informação chamada DolphinSearch [ROITBLAT, 2000 e 2001]. Essa metodologia utiliza os mesmos princípios de ontologia, ou seja, os documentos a serem recuperados, são resultados da extração do significado da expressão de consulta.

A equipe da DolphinSearch Inc., a fim de melhorar as pesquisas de documentos, desenvolveu uma tecnologia baseada no echolocation (eco) produzido pelo sonar dos golfinhos. Esta tecnologia imita o sonar biológico dos golfinhos, o qual identifica as características dos objetos submersos, como tamanho, estrutura, forma e composição material [ROITBLAT, 2000 e 2001].

Esta tecnologia é à base deste projeto, no qual, no momento em que o usuário faz uma consulta, o sistema cria um perfil semântico para a consulta e então compara com os perfis semânticos dos documentos. Assim, o resultado reflete a semelhança dos documentos e não apenas as palavras que o compõem. Para que essas comparações tornem-se possíveis, utilizam-se Redes Neurais Artificiais, implementando o modelo do Perceptron Multi-Camadas.

2. Desenvolvimento do protótipo ZIR

2.1. O que é a Metodologia DolphinSearch

A metodologia DolphinSearch baseia-se na tecnologia utilizada pelos golfinhos para o reconhecimento de peixes, pedras e quaisquer objetos que estejam a sua frente, ou seja, a metodologia DolphinSearch baseia-se na tecnologia de padrão de reconhecimento desenvolvida por Herbert Roitblat (2000), professor do Departamento de Psicologia da Universidade do Hawaii, que modela o echolocation dos golfinhos.

Sabe-se que qualquer expressão vocal pode representar frases que expressem a mesma idéia, por exemplo, as frases "O menino deslizou no hall" e "O jovem homem escorregou ao longo do corredor", possuem o mesmo significado, porém as únicas palavras que elas têm em comum são duas ocorrências da palavra "o". Atualmente, os sistemas de procura de texto livre tradicionais não reconheceriam a semelhança das duas orações do exemplo acima, mas, se pudesse-se extrair o significado de vários termos, então poder-se-ia recuperar o texto apropriado, não importando a forma utilizada para sua escrita [ROITBLAT, 2000].

O DolphinSearch explora regras semelhantes baseadas em perfis semânticos, ao invés da freqüência da ocorrência da palavra no documento. Quando um usuário submete uma solicitação (busca), o sistema cria um perfil semântico para cada solicitação e então compara com o perfil semântico de cada documento. O resultado da busca reflete a semelhança significante dos documentos, não apenas o número das palavras que o sobrepõem. Em uma procura, podem-se encontrar documentos relevantes sem que contenha a palavra (expressão) utilizada para busca [ROITBLAT, 2000].

2.1. Metodologia

Para o desenvolvimento do protótipo, primeiramente, fez-se a preparação dos dados a serem coletados. Foram escolhidas cinco palavras do dicionário português que tivessem duplo sentido para comporem o dicionário do protótipo. Manga, mangueira, macaco, tomada e vaso, foram as palavras escolhidas para este protótipo. Após a seleção das palavras, procurou-se, em média, cinco pequenos textos para cada significado de cada

palavra que seriam utilizadas para treinamento da Rede Neural Artificial. Selecionados estes pequenos textos, leu-se todos, identificando as palavras do dicionário e suas palavras anteriores e posteriores.

Depois, criou-se uma matriz com os verbos de ligação da língua portuguesa e outra com alguns pronomes. Também foi criada posteriormente uma tabela com as palavras anteriores, palavras do dicionário e palavras posteriores encontradas nos textos. Após a preparação dos dados, passou-se para a fase de treinamento da Rede Neural Artificial. Para, posteriormente, passar para a fase de modelagem do sistema e implementação e testes do protótipo.

2.2. Modelagem da Rede Neural Artificial

A modelagem da Rede Neural Artificial divide-se em preparar os dados para construção das matrizes utilizadas para treinamento da rede; treinar a rede e testes.

2.2.1. Preparação dos dados

Para o desenvolvimento do protótipo foram selecionadas cinco palavras do dicionário português que têm duplo sentido para comporem o dicionário do protótipo. As palavras escolhidas foram:

- Manga: pode ser uma fruta ou uma parte da camisa;
- Mangueira: pode ser uma árvore frutífera ou uma mangueira de água, por exemplo, mangueira de jardim, bombeiro;
- Macaco: pode ser um animal ou um objeto utilizado para trocar pneus em veículos;
- Vaso: pode ser um vaso sangüíneo ou um vaso de flor; e
- Tomada: pode ser uma tomada elétrica ou então, uma conquista, por exemplo, os índios tomaram (conquistaram) o forte.

Após a seleção das palavras foram selecionados, em média, cinco parágrafos para cada significado de cada palavra, totalizando quarenta e três parágrafos utilizados para o treinamento da Rede Neural Artificial.

Com as palavras selecionadas e seus parágrafos definidos, partiu-se para a criação manual das matrizes. Criou-se, então, uma matriz, de sessenta e uma linhas por duas colunas (61 x 2), a qual contém vários pronomes da língua portuguesa. Posteriormente, montou-se, uma matriz com duzentas e noventa e oito linhas por duas colunas (298 x 2), contendo os verbos de ligação, que são: ser, estar, parecer, ficar e continuar.

Depois, criou-se uma matriz com os significados das palavras e seus respectivos identificadores, conforme a Tabela 1. Em seguida, foram lidos todos os parágrafos e analisados um a um. O texto era lido. Encontrava-se a(s) palavra(s) principal, por exemplo, manga. Verificava-se qual a palavra anterior e a posterior a palavra principal. Caso a palavra fosse um pronome (o, a, de, da, apenas, etc.) voltava-se uma palavra, até que fosse um adjetivo ou um verbo.

Tabela 1. Significado das palavras

Identificador	Palavra	Significado
1	manga	fruta
2	manga	roupa
3	macaco	animal
4	macaco	objeto carro
5	tomada	elétrica
6	tomada	conquista
7	vaso	sangüíneo
8	vaso	de flor
9	mangueira	de jardim
10	mangueira	árvore

O próximo passo foi a criação de uma matriz de cento e cinqüenta linhas por duas colunas (150 x 2), a qual se tem a palavra (anterior ou posterior) e seu respectivo identificador. Com essa matriz foi possível a criação da matriz de entrada da Rede Neural, que contém noventa e nove linhas por quatro colunas (99 x 4). Juntamente com ela pôde-se criar a matriz de saída ideal da rede, contendo noventa e nove linhas por uma coluna (99 x 1).

Nesta fase, foram seguidos os seguintes passos:

1. É solicitado que se digite uma palavra para treinamento da rede;
2. É feita uma consulta os documentos do banco de dados buscando expressões que tenham a palavra (expressão) digitada;
3. A seguir, gera-se o código da palavra anterior a palavra principal da expressão, o código da palavra principal e o código da palavra posterior a palavra principal, através de um dicionário de palavras;
4. Cria-se uma matriz de 3 (três) colunas por n linhas (3 x n), onde as três colunas representam os códigos das palavras anterior, principal e posterior e as linhas representam a quantidade de vezes que a palavra principal foi encontrada nos documentos do banco de dados;
5. A seguir, é gerada uma matriz de pesos, atribuídos aleatoriamente entre 0 (zero) e 1 (um);
6. É feita a multiplicação das matrizes (entrada x peso);
7. Gera-se o somatório de cada linha da matriz resultante da multiplicação;
8. Cria-se uma matriz de 1 (uma) coluna por n linhas, que especifica a ocorrência da palavra principal da expressão em cada conjunto;
9. A seguir, é aplicada a função de ativação na matriz criada no passo 8. A função de ativação dependerá da quantidade de significados que a palavra em estudo tenha. Por exemplo, a palavra "manga", teria dois significados estabelecidos no dicionário (fruta e parte de uma peça de vestuário). Logo poderia se usar uma função linear. Estas funções serão pré-estabelecidas conforme a quantidade de significados que uma palavra possa ter, para ser utilizada na função de ativação;

10. Gera-se a saída da Rede Neural;
11. Os valores gerados na saída da Rede Neural são comparados com os resultados esperados;
12. Caso o erro seja maior que 5%, aplica-se o algoritmo de backpropagation (retro propagação); e
13. Caso o erro seja menor que 5%, armazenam-se os pesos gerados.

A Figura 1 ilustra os passos descritos na forma de um fluxograma.

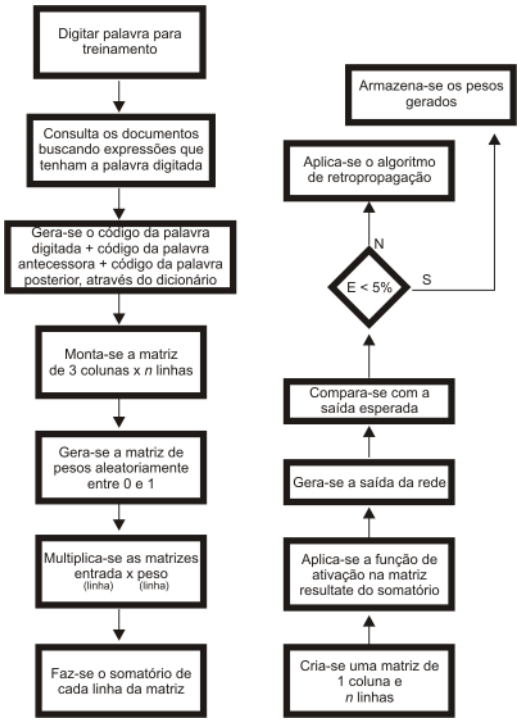

Figura 1. Fluxograma da fase de treinamento da Rede Neural Artificial do protótipo.

2.2.2. Fase de testes

Utilizando o programa para calcular a retro propagação, Figura 10, foi separado um conjunto de quinze pares de entrada para realizar alguns testes, que mostraram que a convergência da Rede Neural Artificial estava correta.

Para fazer os testes, foram seguidos os seguintes passos:

1. Carregar o arquivo de teste;
2. Criar um vetor para armazenar cada padrão de teste;
3. Laço para testar cada padrão:
    a) Armazenar um padrão no vetor;
    b) Inserir o bias no padrão de teste;
    c) Carregar os pesos da entrada até a camada escondida, obtidos no treinamento;

d) Carregar os pesos da camada escondida até a saída, já obtidos no treinamento;

e) Multiplicar o padrão pelo resultado da etapa "c";

f) Aplicar uma função matemática ao resultado da etapa "e";

g) Inserir bias ao resultado da etapa "g" pelos pesos carregados na etapa "d";

h) Aplicar uma função matemática ao resultado da etapa "h".

Fim do laço.

Neste protótipo a função matemática utilizada foi a linear.

Ainda nesta fase, foram selecionados dois documentos para cada palavra utilizada no protótipo, resultando em dez pares de entrada. Com esses pares, foram feitos testes de consistência da Rede Neural Artificial.

2.3. Modelagem do Protótipo

A próxima fase no desenvolvimento do protótipo foi à modelagem do mesmo. Foi modelado um diagrama de Use Case do sistema, onde representa a interação do usuário com a ferramenta. Também foi modelado um diagrama de seqüência do sistema para compreensão dos passos executados pelo protótipo.

2.4. Desenvolvimento do Sistema

O protótipo terá uma interface com o usuário, como mostra a Figura 2, onde o usuário digitará a expressão para a busca. Após, pressionado o botão de "OK" o protótipo transforma a palavra anterior, a principal e a posterior da expressão em um identificador numérico em relação às tabelas de palavras pré-definidas. Posteriormente, o sistema faz a classificação da palavra principal para gerar uma expressão para consulta na Web, que será retornada para o usuário.

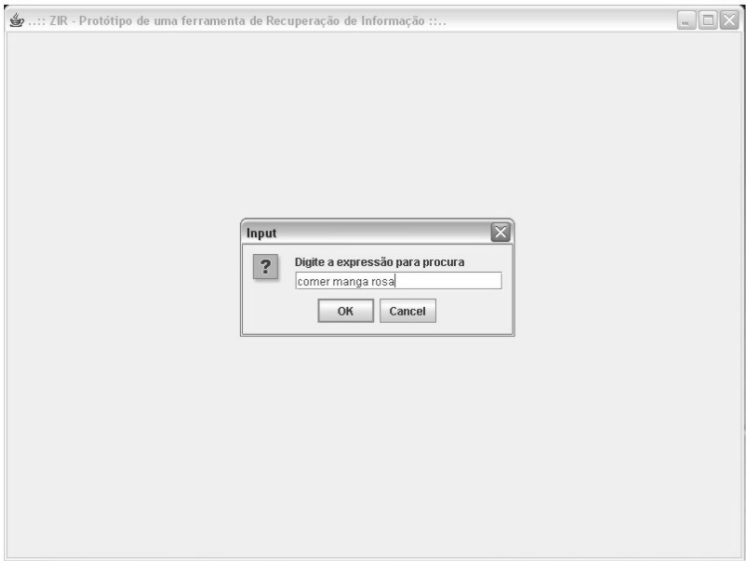

Figura 2. Tela onde o usuário digitará a expressão para procura.

Os resultados da pesquisa serão retornados para o usuário em outra tela, onde conterá um identificador, o nome do documento, as primeiras frases do documento (síntese), seu tipo (doc, ppt, html), seu tamanho e o endereço na Web. A Figura 3 mostra um protótipo da tela de resultados.

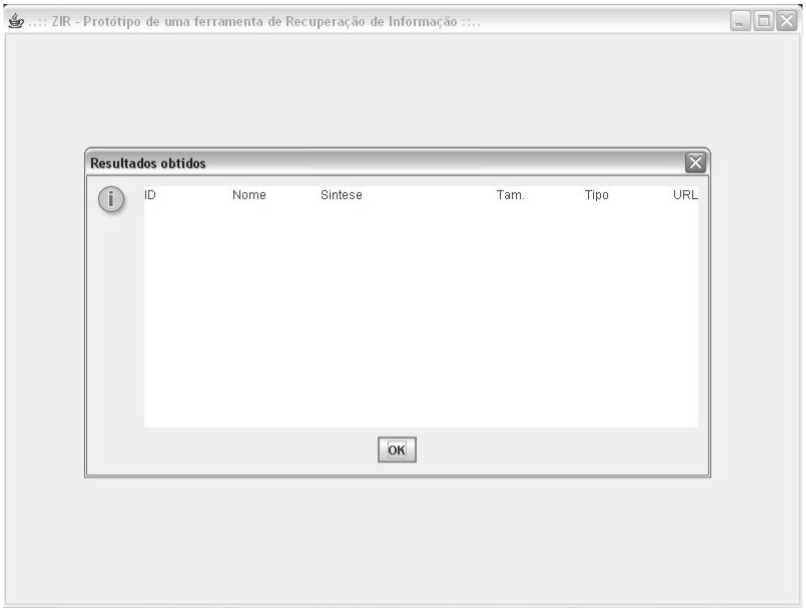

Figura 3. Tela de resultados obtidos na procura.

A Figura 4 ilustra um fluxograma do sistema para melhor compreensão do funcionamento do protótipo.

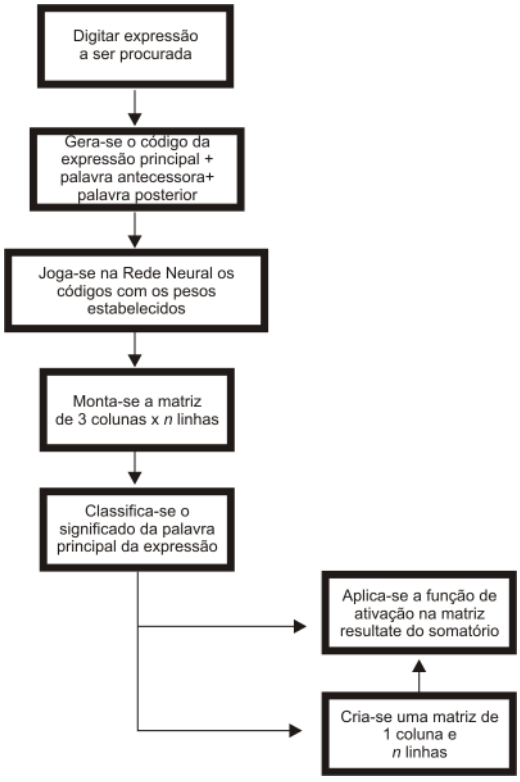

Figura 4. Fluxograma do sistema.

3. Conclusão

A idéia inicial para este trabalho era a de se construir uma ferramenta de recuperação de informação voltada para a língua portuguesa e que utilizasse as técnicas da metodologia do DolphinSearch para fazer a recuperação. Porém, ao terminar a primeira fase (TCC I), percebeu-se que o projeto seria muito extenso e não haveria tempo hábil para sua conclusão, por isso, optou-se por desenvolver um estudo de caso da utilização da metodologia DolphinSearch para recuperação de informações.

Para tanto, determinaram-se os requisitos para o desenvolvimento da ferramenta de Recuperação de Informação. Foram definidas quais seriam as variáveis de entrada, a variável de saída, o erro quadrático utilizado para o treinamento da Rede Neural Artificial. Foram criadas as matrizes de pronomes, verbos de ligação. Também foram definidas as palavras utilizadas no projeto, que, após vários testes percebeu-se que, talvez, o número de palavras utilizadas tenha sido muito pequeno, pois o universo de palavras da língua portuguesa é muito vasto, isto é, o tamanho do conjunto de treinamento, tanto em relação às palavras selecionadas quanto em relação à quantidade de documentos tenha sido muito pequeno. O ideal seria escolher por volta de umas mil palavras da língua portuguesa que tenham duplo sentido, e, pelo menos, uns trezentos documentos relacionados com cada significado das palavras.

Como, infelizmente, a ferramenta não funcionou completamente, pois não estava classificando os significados corretamente das palavras, e por isso, a posterior recuperação não estava adequada não foi possível à execução de testes e validações da ferramenta. Com isso, sugere-se para trabalhos futuros a alteração da função de ativação para, por exemplo, uma função tangente ou outra função matemática mais adequada para a questão. Sugere-se também a utilização de outro algoritmo de aprendizagem, pois, o algoritmo de backpropagation pode não ser o mais apropriado para palavras da língua portuguesa.

4. Referências Bibliográficas

Librelotto, G. R.; Ramalho, J. C.; Henriques, P. R. (2005) "Representação de Conhecimento da Semantic Web". In: Anais do XXV Congresso da Sociedade Brasileira de Computação. Cap. 1, págs. 1210 à 1224.

Roitblat, Herbert L. (2000) "DolphinSearch: Proprietary Information", http://www.dolphinsearch.com, Novembro.

Roitblat, Herbert L. (2001) "DolphinSearch: Scientific Background", http://www.dolphinsearch.com, Novembro.

van Rijsbergen, C. J. (1999) "Information Retrieval". University of Glasgow, Scotland, 2.ed.