Construtor de Ontologias Baseado no WordNet

KARLA DE SOUZA TORRES [1]
JOSÉ LUIS BRAGA [2]
[1] UFV – Universidade Federal de Viçosa
Alojamento Novo, ap. 534
Campus Universitário
CEP: 35761-000
ktorres@dpi.ufv.br
[2] UFV – Universidade Federal de Viçosa
DPI – Departamento de Informática
Campus Universitário
CEP 36571-000
zeluis@mail.ufv.br

Resumo: O objetivo deste projeto é implementar um novo sistema baseado no WordNet de modo que suporte inserção de novos conceitos e termos, fazendo com que o sistema possa ser usado em aplicações que não sejam exclusivamente de processamento da língua inglesa. Finalizada esta fase, o sistema ficará disponível para download e para uso via Internet a partir de um sitio FTP.

O sistema WordNet em sua versão para plataforma Windows funciona com os arquivos originais, projetados para plataforma UNIX e portados para uso no Windows com a mesma estrutura. Nessa configuração, só é possivel a inserção de novos termos, conceitos ou mesmo ontologias se os arquivos forem todos gerados novamente, usando um aplicativo específico, a partir dos arquivos originais. Com isso, fica limitado o uso do WordNet em novas aplicações.

Este trabalho apresenta uma solução para esse problema, migrando o sistema de arquivos para um banco de dados relacional, com uma interface de consulta e atualização baseada em princípios de cooperatividade.

Palavras Chaves: WordNet, ontologias, tesauro.

1 Introdução

A definição de conceitos em áreas distintas da ciência é o primeiro passo para que cientistas possam se fazer entender sem risco de ambigüidades, permitindo assim a troca de idéias e o avanço científico. A cada novo avanço nas pesquisas, novos conceitos e termos são criados que, com algum tempo de maturação, passam a fazer parte do vocabulário e da semântica de cada área.

Tem sido cada vez mais complexo o problema de construir o dicionário dos termos e conceitos de cada área. O crescimento é grande, e o número de grupos ao redor do mundo lidando com o mesmo assunto tende a crescer na medida em que crescem os recursos disponíveis para pesquisa e desenvolvimento. Para se manter atualizado, um cientista deve ler um número crescente de artigos científicos, jornais e revistas especializados.

Uma ferramenta intelectual muito utilizada desde os primórdios da ciência são as denominadas **ontologias**[10,14]. Podem ser definidas como depósitos de conceitos e termos em áreas de conhecimento humano, mostrando não apenas o(s) significado(s) de cada termo ou conceito, mas principalmente as relações existentes entre os conceitos. Na grande maioria das vezes, o conhecimento avança mais no sentido das relações do que na caracterização do termo propriamente dito. O desejo de avançar por essa linha é expressado fisicamente pelas obras de referência conhecidas como os **tesauro**, que são um tipo mais simples de ontologia, mostrando um número pequeno de relações entre os conceitos.

1.1 O Sistema WordNet

Dentre os tesauro de uso geral disponíveis na comunidade acadêmica, está o WordNet. Desenvolvido pelo grupo de Ciências Cognitivas da Princeton University, o WordNet [5, 11] é um sistema público, que supre essa necessidade de estabelecimento do significado preciso, ou semântica, para conceitos simbolizados. Uma caracterização resumida, extraída do sítio do sistema na Internet[11], é:

"WordNet® é um sistema de referência léxica on-line cujo projeto foi inspirado em teorias psicolinguísticas atuais sobre a memória léxica humana. Nomes, verbos, adjetivos e advérbios da língua inglesa são organizados em conjuntos de sinônimos, cada conjunto representando um conceito ou categoria léxica diferente. Os conjuntos são então relacionados por relações estruturais. "

WordNet é portanto um tesauro disponível por enquanto apenas para a língua inglesa, contendo substantivos, verbos, adjetivos e advérbios. A classe dos substantivos, sem dúvida a mais numerosa, é organizada em classes de substantivos sinônimos. Por volta de outubro de 1998, o sistema continha mais de 91600 conjuntos de sinônimos. Esses conjuntos são relacionados entre si por relações especiais, do tipo "classe mais geral" ou generalização, "classe mais específica" ou especialização, "parte-de", "tem-partes" e as hierarquias "é-um" . Essas relações, juntamente com os conjuntos de sinônimos, formam um grafo de relacionamentos que é usado para definir a semântica de cada conceito, e as relações entre conceitos distintos.

Esse sistema está sendo muito usado como apoio a processamento semântico de informação, por exemplo em interfaces que utilizem linguagem natural para permitir a comunicação homem-máquina. Foi desenvolvido sob ambiente UNIX, e seu banco de dados, muito complexo e volumoso, tem a estrutura dos arquivos desse sistema. Recentemente foi liberada uma versão utilizável sob Windows/DOS, que manteve a estrutura de todos os arquivos e sistema, tendo apenas uma interface de uso sob Windows.

A base de dados do sistema é fixa, e para incluir nela novos termos relativos, por exemplo, a novas áreas de aplicação, é necessário que os autores do sistema original gerem todos os arquivos novamente, a partir dos arquivos originais, incluindo os novos termos. Ou seja, em termos práticos, é impossível usar o sistema em aplicações que não sejam exclusivamente de processamento da língua inglesa. Isso limita seu uso numa grande variedade de situações em que seria útil e factível utilizá-lo.

O uso do sistema atual se dá via Internet[11] ou via a interface da Figura 1, para ambiente Windows:

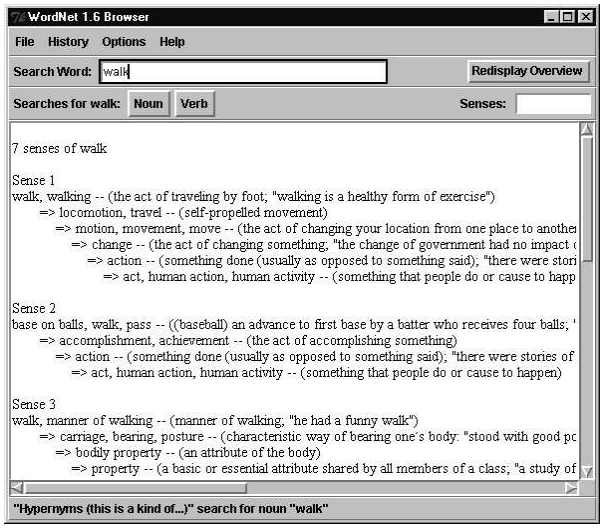

Figura 1: Hierarquia para o substantivo "walk"

Na Figura 1 está um exemplo em que se busca o significado e a estrutura da relação "é um tipo de" do substantivo "walk". Nesta relação o sistema WordNet

mostra a hierarquia da palavra como um nome. O programa também mostra todos os significados que o nome possui e para cada um deles existe uma frase explicativa entre parênteses.

O WordNet permite que o usuário conheça diversos tipos de relações entre o termo procurado e outros termos diversos e a hierarquia semântica da palavra é apenas uma destas relações.

2 Caracterização do Problema

Como citado anteriormente, a versão original do WordNet foi feita em ambiente UNIX. Recentemente, o grupo de Ciências Cognitivas da Princeton University produziu uma versão, sem modificações estruturais na versão original, para ser usada em ambiente Windows ou via interface HTML na web[11]. Nos dois casos, o sistema só funciona para consultas, não sendo permitida a inserção de novos termos e conceitos.

Essa limitação na inserção dificulta o uso do sistema em aplicações práticas. Para que um cientista use o sistema em sua área de conhecimento, inserindo nele os conceitos necessários, ele tem que usar um utilitário especial, o Grinder, juntamente com os arquivos de dados originais do sistema, o que até o momento só pode ser feito pelo grupo criador do sistema, em Princeton.

Portanto, ainda não é possível ao profissional que deseja ter sua própria ontologia com termos que são relativos à sua área, usar o sistema livremente para fazer experimentação na definição dos novos conceitos, entender sua relação com os demais conceitos, etc. Esse usuário qualificado e especialista em uma área de conhecimento não é atualmente atendido pelo sistema nessa sua necessidade de uso.

O que se visualiza como situação ideal de uso do sistema pode ser exemplificado pela Figura 2.

O ideal é que usuários geograficamente distantes possam usar o sistema de forma independente, criando suas ontologias localmente, que seriam a seu tempo incluídas na ontologia geral, centralizada em servidores especializados. Repare-se que a seta dupla no esquema da Figura 2 significa que o usuário(cliente) pode tanto consultar a base de dados principal como fazer inserção na mesma. Isto quer dizer que toda inserção feita em qualquer banco de dados local, será em algum momento espelhada também no banco de dados principal, de forma que qualquer outro cliente possa visualizá-la. Esse problema não tem solução simples, pois a consistência da base local tem que ser verificada com relação à base centralizada, para que não haja conflitos ou inconsistências entre conceitos, o que minaria a confiança dos usuários do sistema.

Para que qualquer inserção de dados possa ser feita, o usuário deverá obter uma autorização junto aos responsáveis pelo sistema. A partir daí, o termo a ser inserido sofrerá uma rigorosa avaliação que vai determinar se este é um conceito já existente ou um conceito inútil. A criação de normas e regras para a inserção de novos termos, bem como os critérios usados para a aprovação de um novo usuário do sistema é o objetivo para a continuidade deste projeto.

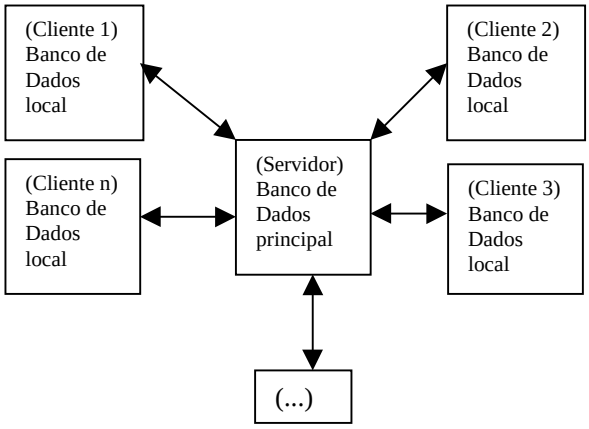

Figura 2: WordNet em arquitetura cliente-servidor

2.1 Sistemas Disponíveis

Como já dito, o problema do sistema original consiste no fato de que a base do sistema é fixa e incluir nela novos termos relativos a novas áreas de aplicação é uma tarefa complicada. Em termos práticos, é impossível usar o sistema em aplicações que não sejam exclusivamente de processamento da língua inglesa. Isso, claro, limita seu uso numa grande variedade de situações em que seria útil e factível utilizá-lo.

Existem algumas soluções disponíveis na literatura como o Loom[10] que é um software que fornece ferramentas para a aquisição de conhecimento construído com técnicas de Inteligência Artificial. Como extraído da própria página do sistema, *"Loom é uma linguagem e um ambiente para a construção de aplicações inteligentes"*.

Outro exemplo é o sistema Sophia[9], que é um servidor de conhecimento para ser usado via web. Este sistema foi construído a partir de um sistema comercial de banco de dados relacional. O sistema é simples e não dispendioso, e fornece ainda uma avançada funcionalidade. Sophia é acessível a usuários via Web ou via aplicações cliente através de um API.

Outro sistema, mais antigo e que de fato deu origem a vários outros, é o Ontolíngua[14]. Este é um laboratório de sistemas de conhecimento que fornece um grande número de serviços através da Web que qualquer usuário pode acessar usando um browser. Alguns destes serviços são simplesmente informações fornecidas por

documentos estáticos na Web. Outros já são documentos dinâmicos ou serviços tais como edição interativa de uma base conhecimento.

Entre outros sistemas que foram investigados, pode ainda ser citado o The Brain[12] que tem uma boa proposta de estrutura e interface para um browser de conhecimento. Este site tenta trazer para as pessoas em geral, o conhecimento e o aumento de produtividade fornecendo a indivíduos e organizações um ambiente para visualização de informações muito rico e de uso intuitivo. O The Brain dá às pessoas a habilidade para organizar informação dentro de uma estrutura que converte relações em uma estrutura visual de entendimento fácil.

O sistema Inxight[13] fornece outra sugestão de interface de browser hiperbólico. Este também fornece uma versão demo de um programa que gera o browser para um banco de dados qualquer ou uma página da Web. É muito utilizado por "portais" na Internet ou sites de grandes empresas visando a facilitar a interação com o usuário.

3 A Solução Visualizada

Está sendo projetado e implementado um novo sistema de organização dos dados e uma interface de consulta e atualização pela Internet. Cada usuário do sistema será capaz de fazer consultas na base de dados já existente, de maneira que o resultado destas consultas o guiem para a solução do seu problema. Além disto o usuário também será capaz de criar sua própria ontologia, inserindo novos termos que lhe sejam úteis. A inserção na base de dados geral do sistema, que estará localizada em um servidor na UFV será somente permitida a pessoas autorizadas, e cada novo termo a ser inserido deverá ser rigorosamente avaliado de acordo com a consistência semântica e de suas relações com os demais termos. Uma vez que um termo é inserido na base de dados, ele não mais poderá ser retirado, pelo fato de que o sistema se baseia em uma forma de aquisição de conhecimento e este jamais é perdido. Exatamente por causa disto a avaliação da consistência do termo deve ser feita antes de este ser inserido no banco de dados do sistema.

Os arquivos do WordNet possuem uma estrutura fixa, e são divididos de acordo com o tipo dos termos. Estes podem ser nomes, adjetivos, verbos e advérbios. Para cada uma destas classes existe um conjunto específico de arquivos, estruturados de forma a facilitar a consulta de dados. Para os nomes por exemplo, existe um arquivo de índices onde é feita a busca do termo desejado. Após ser localizado são extraídos deste arquivo os synsets(conjunto de sinônimos) que são relativos ao termo. Com esta informação, o programa do WordNet navega por um outro arquivo contendo diversos synsets e relações. De acordo com esta estrutura, a inserção de um termo que fosse um nome, por exemplo, deveria ser inserido em todos os arquivos e dentro deles estabelecidas todas as relações. No entanto trabalhar com um sistema de arquivos possui a desvantagem de que uma atualização deve reorganizar todos os demais dados que estão contidos no mesmo. Isto é trabalhoso e demorado, e por isso os autores originais do sistema só o fazem usando um aplicativo especial.

Visando a acabar com este problema, foi realizado um trabalho de reengenharia reversa sobre os arquivos do WordNet. O resultado desta fase foi um esquema ER que descreve a estrutura conceitual do sistema. Esse esquema está apresentado na Figura 3, a seguir.

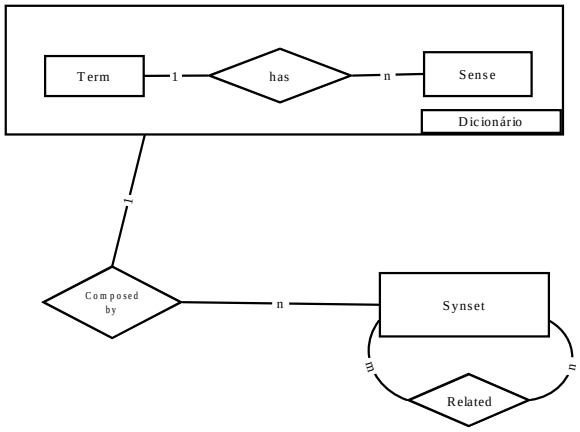

Figura 3: Modelagem E-R do sistema para conversão dos dados.

O esquema da Figura 3 demonstra o relacionamento entre três entidades do WordNet: a entidade representativa dos termos(Term), a entidade representativa dos sentidos(Sense) e a entidade representativa dos sinônimos(Synset). Estas entidades serão mapeadas em tabelas no Modelo Relacional de Banco de Dados[1]. Os requisitos originais do WordNet são atendidos perfeitamente com este esquema, pois ele foi baseado na estrutura original do sistema. Com esta mesma estrutura, é possível ter a função de inserção bastando para isto usar o esquema de tabelas e não de arquivos. O passo seguinte foi mapear o esquema ER obtido, em tabelas do modelo relacional de dados, completando assim a reengenharia do sistema WordNet. Na nova estrutura, não houve perda de informação com relação à estrutura original. Por exemplo, a consulta do termo "walk" feita no WordNet(Figura 1) poderá ser feita da mesma forma no novo sistema, sendo que a diferença é que a consulta pode ser estruturada de forma

a facilitar a interação com o usuário e pode conter mais dados de resposta que a pesquisa original.

A estrutura relacional, de forma como está proposta, facilitará a inserção de novos termos e conceitos na estrutura do WordNet, o que é o objetivo mais forte do projeto aqui descrito.

O esquema relacional[1] obtido a partir do esquema ER ficou da seguinte forma:

***Nomes*(lemma, pos, sense_cnt, p_cnt, pointers, tagsense)**
***NomeSenses*(lemma*, Sense*)**
***NomeSynsets*(Synset_offset, ss_type, w_cnt, words, p_cnt, gloss)**
***NomesRelacoes*(Syn_From*, Tipo_Relacao, Syn_To*, Tipo)**

No esquema, os termos sublinhados são as chaves primárias e o símbolo "*" representa uma chave estrangeira. Os demais termos são os atributos de cada tabela.

Terminada esta fase de reengenharia, foram implementados programas (*drivers*) em Java para fazer a migração dos arquivos.

4 Detalhes de Implementação

Os passos metodológicos para obter a implementação do sistema podem ser relacionados abaixo:

- Estudo e entendimento da linguagem Java, do ambiente JBuilder3[3] e do WordNet[5,11];
- Implementação de um *applet* como protótipo de interface de consulta;
- Modelagem E-R de toda a estrutura do sistema. (engenharia reversa)[6, 7];
- Projeto relacional(reengenharia)[1, 6];
- Projeto e implementação de programas em Java para a conversão dos dados;
- Implementação de um *applet* como protótipo de interface de consulta sob a nova estrutura das informações em banco de dados relacional.

Após ter sido feita a conversão dos dados foi necessário projetar um *applet* como protótipo de interface de consulta sobre a nova estrutura de organização das informações, ou seja, um *applet* que navega sobre as tabelas do banco de dados gerado e monta as informações necessárias à busca. A conversão dos dados e a implementação deste *applet* de consulta dão início ao novo sistema criado que virá a permitir a inserção de novos termos aumentando assim a base de dados já criada e expandindo a área de abrangência do sistema, não ficando restrito somente ao domínio da língua inglesa. Este novo sistema que está sendo criado recebeu o nome de ATENA – Construtor de Ontologias baseado no WordNet.

Quando digita-se na interface do programa a palavra "abacus" e dica-se no botão "Hierarquia" o programa irá exibir os dois significados da palavra dizendo que esta é um substantivo. E para cada significado, é exibido na tela também toda a hierarquia do tipo "É um subconjunto de:". Além disto, para cada um dos significados da palavra procurada, o Dicionário Eletrônico dá uma frase de exemplo que define esta palavra dentro de cada significado. Na Figura 4, mostra-se o resultado obtido com a busca da palavra "abacus" no *applet*.

A hierarquia de ambos os significados termina em "entity, something" que são os termos mais genéricos na classe de substantivos.

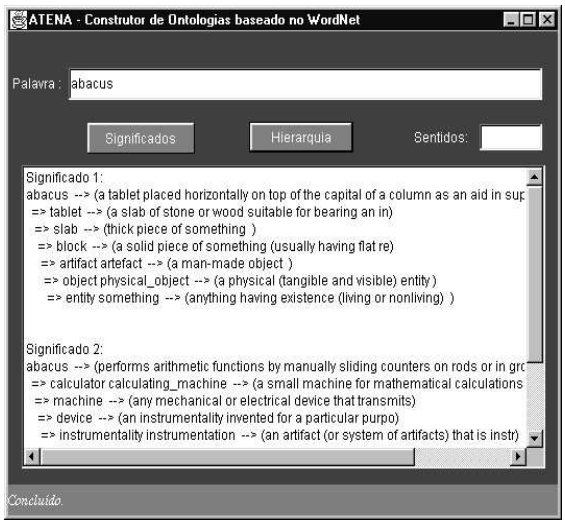

Figura 4: A hierarquia do termo "abacus" no *applet* que navega sobre o banco de dados gerado**.**

Outro exemplo de uso pode ser obtido ao digitar-se a palavra "frog" no *applet* ATENA e clicar-se no botão "Hierarquia". O resultado pode ser visto na Figura 5. Para os três sentidos que o termo "frog" apresenta é mostrada a hierarquia de cada um.

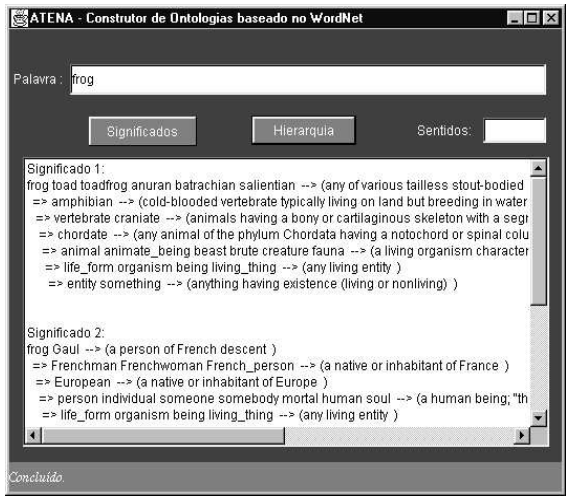

Figura 5: A hierarquia para o termo "frog".

Apesar de o *applet* ATENA já ter seu sistema de busca sobre o banco de dados do WordNet bem estruturado, a interface do programa final não será deste tipo, pois além de ter que permitir a inserção de novos termos por pessoas autorizadas, o sistema deverá ter uma interface mais fácil de manipular por usuários de todos os níveis de experiência. A idéia principal é que esta interface seja construída dentro do conceito de "interface ativa" ou "interface cooperativa"[8] que auxiliam o usuário no momento da busca de informações.

Com o objetivo de aperfeiçoar a interface de consulta e inserção do sistema que está sendo criado, está sendo desenvolvido um outro projeto de pesquisa integrado a este, que vai buscar a melhor solução de interação com os usuários deste sistema.

Após ter convertida toda a base de informações do WordNet para o sistema de banco de dados relacional, o próximo passo é transferir essa base de informações para um banco de dados de melhor desempenho, que suporte o sistema cliente-servidor visto na Figura 2. O banco de dados escolhido foi o SQL Server 7 que possui todos os componentes necessários à nova fase do projeto. A partir deste ponto, será projetado uma componente para inserção de novos termos por pessoas autorizadas. Neste momento surge a necessidade da definição de normas para a inserção, de forma a evitar que a consistência dos dados seja ameaçada e que não sejam inseridos dados considerados como dados insatisfatórios.

Esta fase também inclui o projeto de interfaces de interação com o usuário, tanto a interface tida como "stand-alone", que é a interface que rodará localmente na máquina do usuário, quanto a interface a ser acessada via Internet.

5 Conclusões

Após ter-se feito toda esta discussão sobre o WordNet e o que vem a ser o novo método de implementação baseado no mesmo, pode-se concluir que o novo sistema trará uma enorme melhoria no que diz respeito à estrutura atual. Esta melhoria vem do fato de que a nova versão permitirá a inserção de novos termos em sua estrutura, o que amplia sobremaneira as possibilidades de uso do WordNet por pesquisadores das mais diversas áreas.

Apesar de ter sido feita a conversão dos arquivos do WordNet para uma estrutura de registros, vislumbra-se a possibilidade da superioridade de uma solução usando Banco de Dados Orientado a Objetos, ao invés de um BD Relacional. O estudo sobre esta possiblidade já está sendo feito e busca-se recursos para a conversão dos arquivos atuais para um gerenciador de objetos.

Após a conversão de dados, os próximos passos da pesquisa consistem na estruturação de formas de consulta e de inserção utilizando diversos recursos para a melhor visualização dos dados e facilidades no uso do sistema por parte do usuário. É importante salientar também que o programa estará disponível para diversos usuários através de um sitio ftp localizado na UFV, de forma que todos os usuários interessados poderão utilizar o sistema. Mais um aspecto que deve ficar claro também é que os usuários autorizados estarão sempre ajudando no crescimento da ontologia do sistema, inserindo e relacionando novos termos.

6 Referências Bibliográficas

1. KORTH, Henry F.; SILBERSCHATZ, Abrahan; *Sistema de Banco de Dados*. 2ª edição, São Paulo, Brasil; MAKRON Books, 1993.
2. FURLAN, José Davi; *Modelagem de Objetos através da UML –The Unified Language*; São Paulo, Brasil; MAKRON Books, 1998.
3. JENSEN, Cary; STONE Blake; ANDERSON Loy; *JBuilder Essentials;* 1ª edição, Berkeley, California, EUA; Osborne, 1998.
4. ECKEL, Bruce; *Thinking in Java*; 1ª edição, New Jersey, EUA: Prentice Hall PTR, 1998
5. FELLBAUM, Christiane*; WordNet: an eletronic lexical database* ; 1ª edição; Cambridge, Massachusetts, EUA; Christiane Fellbaum, 1998.
6. STAA, Arndt von; *Programação Modular: desenvolvendo programas complexos de forma organizada e segura;* Rio de Janeiro, RJ; Ed. Campus, 2000.
7. MYNATT, B.T. *Software Engineering With Student Project Guidance*, Englewood Cliffs, NJ, Prentice Hall International Editions, 1990.
8. BRAGA, J. L., LAENDER, A.H.F., Ramos, C.V.A. Knowledge-based approach to cooperative relational database quering. International Journal of Pattern Recognition and Artificial Intelligence 14(1):73-90, Feb. 2000.
9. *Knowledge-Base Server*. Stanford University . Url: http://sophia.stanford.edu/
10. Artificial Intelligence Research Group. *LOOM Project Home Page*. University of Southern California, 1999. Url: http://www.isi.edu/isd/LOOM/LOOM-HOME.html
11. Cognitive Science Laboratory. *WordNet - a Lexical Database for English*. Princteon University, 1998. Url: http://cogsci.princeton.edu/~wn/
12. FUCHS, Peter. *The Brain*. USA, 1998. Url: http://www.thebrain.com
13. LEE, Robert P. *Inxight Software*. USA. Url : http://www.inxight.com
14. Stanford Knowledge Systems Laboratory. *Stanford KSL Network Services*. Stanford, 1993. Url: http://ontolingua.stanford.edu