Web Semântica para Máquinas de Busca

Eli Nogueira Milagres Júnior, Elio Lovisi Filho (orientador)

Departamento de Ciência da Computação - Faculdade de Ciência da Computação e Comunicação Social (FACICS) - Universidade Presidente Antônio Carlos - Campos Magnus - Campolide - MG

elimilagres@hotmail.com,eliolovisi@netwave.com.br

Resumo. *Este artigo tem como objetivo mostrar as características da Web Semântica, como conceitos, estruturas, ferramentas utilizadas e também mostrar como esta tecnologia pode ser utilizada em máquinas de busca, mostrando desta forma como será mais fácil fazer a busca de uma informação na Web.*

1- Introdução

A web hoje é uma biblioteca imensa que a cada segundo que passa recebe mais documentos, aumentando assim o seu tamanho. Um dos problemas da web atualmente é a recuperação da informação, isto acontece porque as mesmas não estão estruturadas e nem seguem nenhum padrão, sendo que a principal preocupação atualmente é na estética e não no sentido da informação.

Para que a recuperação da informação na web seja mais rápida e com mais qualidade é necessário que esta busca seja feita de maneira automática, ou seja, utilizando os próprios computadores. Para isto é preciso que os computadores entendam o significado das consultas e da informação não estruturada contida na web.

Com este objetivo foi criada a idéia de Web Semântica, que juntamente com as máquinas de busca muito nos ajudará a enxugar todo este “lixo eletrônico” que sempre vem junto á nossa busca, porque podemos observar na prática que geralmente estas buscas são lentas e geram respostas muito abundantes e muitas das vezes sem relevância , apesar de haver uma boa implementação das máquinas de busca . (DZIEKANIAK)

Um exemplo do que se quer chegar utilizando as técnicas de Web Semântica é o seguinte, vamos supor que um determinado usuário pretende encontrar fazendo uma busca pela Web um médico de uma determinada área da medicina, o usuário digitará em um mecanismo de busca onde através de uma máquina de inferência o próprio computador será capaz sozinho de procurar este médico dentre milhões de páginas na Web , neste caso poderá achar vários desta área, ele então irá selecionar qual deles tem um consultório mais próximo do usuário, e se ambos estiverem com suas agendas eletrônicas funcionando ele será capaz de selecionar qual o melhor horário, ou então o próprio usuário só terá o trabalho de optar por qual horário ele acha melhor.

Neste artigo poderemos conhecer as principais características da Web Semântica como conceitos, estruturas, ferramentas utilizadas e também conhecer um pouco sobre Máquinas de Busca, principalmente sobre Máquinas de Busca Semântica e também fazer um comparativo entre Máquinas de Busca Semântica e Máquinas de Busca Indexadas em HTML.

2 - Web Semântica

A Web Semântica é um projeto dirigido por Tim Berners-Lee, criador do HTML e da *World Wide Web* sob os aspectos do *World Wide Web Consortium* (W3C). O objetivo deste projeto é melhorar as potencialidades da web através da criação de ferramentas e de padrões que permitam atribuir significados claros aos conteúdos das páginas e facilitar a sua publicação e manutenção. (ARTIFICE1)

Segundo (ARAÙJO,2003) a definição mais correta para Web Semântica é :

"A Web Semântica não é uma Web Separada, mas uma extensão da atual, na qual a informação é utilizada com significado bem definido, aumentando a capacidade dos computadores em trabalharem em cooperação com as pessoas".

A principal idéia da Web Semântica é organizar melhor as informações contidas na Web, de maneira que a própria máquina possa entendê-la, facilitando assim as buscas por determinada informação na Web, porque nas buscas feitas atualmente como por exemplo

pelo Google retorna geralmente centenas de páginas que contém na maioria informações que não estão ligadas ao que procura-se, sendo que maioria utiliza-se de linguagem HTML, onde está linguagem não aceita atribuição de significado à informação, ou seja as buscas atualmente são realizadas com base em palavras chaves e as buscas semânticas são realizadas com base no significado da informação.

2.1 Características da Web Semântica

A Web Semântica é uma evolução da Web atual, não podemos continuar pensando somente em melhorar o layout de sites, precisamos pensar em organizar melhor a informação para que as buscas possam ser feitas com mais eficiência e qualidade.

Para implementarmos semanticamente é necessário que tenhamos um bom conhecimento de algumas tecnologias como Dublin Core, RDF (*Resource Description Framework*) e XML(*Extensible Markup Language*), porque a união destas três ferramentas é que nos possibilitará o desenvolvimento de bases de dados e plataformas semânticas que visam a organização do conhecimento possibilitando ao homem e a máquina poderem trabalhar em conjunto.

Ao falarmos de Web Semântica devemos lembrar que milhões de páginas que estão publicadas na Web atualmente não utilizam RDF e como veremos à frente, é que, utilizar RDF pode ser complicado para desenvolvedores médios.

De acordo com (PARREIRAS) na futura Web, ou na Web Semântica, o conjunto de recursos e links são identificados por URI's, mas podem ser "tipados". Isto consiste em atribuir um tipo à relação entre dois recursos. Forma-se o conceito chamado de triplas (Um recurso, uma propriedade e um valor). Neste contexto, uma relação entre dois recursos possui uma propriedade que permite atribuir significado à ligação. No exemplo dos sites de filme, pode-se atribuir a propriedade "está entre os dez melhores" ao link no primeiro site e a propriedade "está entre os dez piores" ao link no segundo site. A diferença entre os modelos é que agora o conhecimento está formalizado de uma maneira estruturada. As linguagens que permitem a criação do conhecimento estruturado são RDF e XML.

3 - Representação do Conhecimento

A representação do conhecimento é feita através do RDF e seguem geralmente 3 condições:

- **Interoperabilidade Estrutural** – Permite que os dados sejam representados de forma distinta, permitindo especificar tipos e valores para cada forma de representação.
- **Interoperabilidade Sintática** - Constitui-se de regras precisas que permitem o intercambio de dados na web.
- **Interoperabilidade Semântica** - Possibilita a compreensão e associação entre dados.

Mais adiante conheceremos o RDF e o XML que serão ferramentas que nos possibilitaram aplicar esta representação.

4 - Metadados

Metadados são dados capazes de descrever outros dados, ou seja explicar o que se trata um determinado arquivo. A grande aplicabilidade dos metadados está na Web Semântica, onde com a utilização dos metadados será possível fazer a recuperação da informação.

O que acontece na realidade é que todo documento deve ser catalogado. Este catálogo consiste de vários registros eletrônicos onde em cada registro existe uma descrição do documento, ou seja não há necessidade do leitor ler ou ouvir toda obra para que ele possa entender o que tem em determinada obra.

4.1 – Padrões de Metadados

Seria impossível utilizarmos um padrão de metadados para todas as áreas do conhecimento humano, por isto existem vários tipos como por exemplo GILS (***Government Information Locator Service***) que é utilizada para descrever informações governamentais, FGDC (***Federal***

Data Geografic Committee) que é utilizado para informações geoespaciais e muitos outros. Já o que nos interessa é o Dublin Core que é um padrão de fácil utilização que veremos mais informações dele logo a seguir.

4.2 Dublin Core – DC

O padrão *Dublin Metadata Core Element*, é um padrão de metadados que possui uma lista de quinze elementos descritores, sendo um dos padrões mais utilizados na Web, inclusive é o padrão recomendado pela W3C. Possui pares(nome atributo / valor atributo) como estrutura e sua utilização não possui maior complexidade.

Os descritores fornecidos pelo DC são genéricos e simples e não cobrem todas as necessidades de descrição dos recursos, mas porém para sanar esta ausência são desenvolvidos elementos extras que podem ser denominados de qualifiers. O DCMI (*Dublin Core Metadata Iniciative*) recebe sugestões de qualifiers que são analisados e se aprovados podem fazer parte do conjunto de descritores às aplicações

Abaixo na Figura A segue um exemplo de descrição da página inicial do site www.fernando.parreiras.nom.br, utilizando o padrão de descrição de metadados Dublin Core. Estão descritas informações sobre o autor, direitos autorais e data de criação da página: (PARREIRAS)

```
<html xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:con="http://www.w3.org/2000/10/swap/pim/contact#">
  <head>
    <title>Site pessoal de Fernando Parreiras</title>
    <meta property="dc:creator">
      <meta property="con:fullName">Fernando Parreiras</meta>
    </meta>
    <meta property="dc:rights">&copy; 2004 </meta>
    <meta property="dc:subject">vida do autor,
          currículo, publicações</meta>
    <meta property="dc:date">2004-11-06T08:49:37+00:00</meta>
  </head>
```

Figura A (Exemplo de Dublin Core)

5 – Ferramentas na Web Semântica

Como já foi abordado anteriormente, para que a Web Semântica se torne uma realidade é necessário que algumas ferramentas trabalhem em conjunto, como é o caso do RDF e do XML que abordaremos logo a seguir.

5.1 - RDF (*Resource Description Framework*)

È uma linguagem utilizada para representar a informação na Internet através de esquemas semânticos. O RDF é um sistema de auxilio para desenvolvimento de metadados cuja finalidade é promover a interoperabilidade entre as aplicações que compartilham informações que sejam entendidas por sistemas na Web (ZANETE2002?). Metadados representados em RDF são usados para dar significados aos recursos da Web Semântica por permitir que este sejam manipulados e compreendidos por máquinas.

Os arquivos RDF têm três componentes básicos que são o recurso, a propriedade e a indicação, o que torna a linguagem altamente escalável.
Recurso: Qualquer coisa que pode conter um URI, incluindo as páginas da web, assim como elementos de um documento XML.
Propriedade: Um recurso que tenha um determinado nome e possa ser utilizado como uma propriedade.
Indicação: consiste na combinação de um recurso, de uma propriedade, e de um valor.

Abaixo na figura B mostra um arquivo RDF que descreve quem é o criador do recurso www.fernando.parreiras.nom.br: (PARREIRAS)

```
<?xml version="1.0"?>
   <rdf:RDF
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:dc="http://purl.oclc.org/DC#">
   <rdf:Description about=" http://www.fernando.parreiras.nom.br"/>
     <dc:Creator>Fernando Parreiras</dc:Creator>
   </rdf:Description>
   </rdf:RDF>
```

Figura B (Exemplo de RDF)

5.2 - XML (*Extensible Markup Language*)

È uma linguagem de marcação de dados que permite aos usuários criarem tags personalizadas sobre os documentos criados, diferentemente da linguagem HTML que possui tags fixas, impedindo a criação de novos tipos de descritores.

Oferece aos seus usuários a descrição de dados estruturados facilitando declarações precisas do conteúdo de documentos e mais ainda, facilitando a recuperação destes documentos via Web. A linguagem XML supre as deficiências do HTML, permitindo a criação de marcações pelo próprio usuário, e desta forma proporcionando uma maior descrição dos recursos em termos de metadados. A formatação do documento é feita separada de sua estrutura resolvendo assim um dos grandes problemas do HTML.

O XML é a representação textual do dado. O componente básico do XML é element, isto é o texto é limitado entre os delimitadores tags < > ..... </>(incluindo os próprios delimitadores). Tal como pessoa, nome e e-mail . È possível associar atributo a elementos. Um atributo é definido como um par (nome , valor).

Uma linguagem XML deve respeitar duas restrições: tags devem estar corretamente alinhadas e atributos devem ser únicos. Um documento organizado corretamente ficará fácil de representá-lo via XML na Web e de fazer a recuperação da informação.

Abaixo temos a figura C que possui um exemplo que irá facilitar o entendimento.

Esse exemplo demonstra a síntaxe flexível do XML sendo usada para descrever uma receita de pão. (WIKIPEDIA)

```
<?xml version="1.0" encoding="UTF-8"?>
<Receita nome="pão" tempo_de_preparo="5 minutos" tempo_de_cozimento="3
horas">
  <título>Pão simples</título>
  <ingrediente quantidade="3" unidade="xícaras">Farinha</ingrediente>
  <ingrediente quantidade="7" unidade="gramas">Fermento</ingrediente>
  <ingrediente quantidade="1.5" unidade="xícaras"
estado="morna">Água</ingrediente>
  <ingrediente quantidade="1" unidade="colheres de chá">Sal</ingrediente>
  <Instruções>
    <passo>Misture todos os ingredientes, e dissolva bem.</passo>
    <passo>Cubra com um pano e deixe por uma hora em um local morno.</passo>
    <passo>Misture novamente, coloque numa bandeja e asse num forno.</passo>
  </Instruções>
</Receita> Onde temos na 1ª linha
<Receita nome="pão" tempo_de_preparo ="5 minutos" tempo_de_cozimento="3
horas">
```

"Receita" é o nome principal para o seu documento. Note que a semelhança entre XML e HTML é grande, na 1ª linha abrimos a tag Receita e na última linha fechamos a mesma, como em HTML e assim se estende por todo o exemplo.

Figura C (Exemplo de XML)

6 – Máquinas de Busca

O conteúdo da Web vem crescendo muito rápido e de maneira desordenada. Devido à facilidade na criação e postagens de páginas, hoje qualquer usuário consegue colocar e retirar quando quiser qualquer tipo página sem seguir nenhum tipo de padrão, o que dificulta e muito fazer uma busca para encontrar determinado assunto.

Devido a isto foram criadas as máquinas de busca para minimizar este trabalho que na verdade tem a função de processar as informações contidas nos documentos durante o processo de recuperação da informação e comparar com a consulta requerida pelo usuário, e retornar uma lista de documentos que possuem maior similaridade com o tema desejado. Estas informações são classificadas como intrínsecas e extrínsecas.

Segundo (CAMPOS) as informações intrínsecas são aquelas que estão dentro do documento que está sendo analisado pela máquina de busca como, por exemplo, uma

palavra em um determinado texto. Já as informações extrínsecas são obtidas através de documentos contidos na coleção, estrutura de links ou popularidade de um documento com relação a outro.

Mas atualmente chegamos a um volume tão grande de informação que na maneira que as máquinas de buscas trabalham atualmente que é utilizando palavras chaves para fazerem as buscas está se tornando cansativo também fazer este tipo de busca, por isto mostraremos aqui os tipos de máquinas de buscas existentes e qual delas devemos utilizar na Web Semântica.

6.1 - Classificação das Máquinas de Busca

As máquinas de buscas são classificadas de acordo com a maneira que é recuperada a informação, abaixo elas estão classificadas:

- **Primeira Geração** – Durante o processo de recuperação da informação utiliza-se de informações intrínsecas aos documentos.
- **Segunda Geração** - Utiliza-se também de informações extrínsecas.
- **Terceira Geração** - A recuperação da informação é feita com base na semântica das informações, ou seja os dados são recuperados com base em informações estruturadas semanticamente.

6.2 – Máquinas de Buscas Semânticas

São as máquinas classificadas como de terceira geração, que são as que estão ligadas ao tema do nosso artigo.

As máquinas de busca Semânticas são baseadas na recuperação da informação e para que isto aconteça é necessário que as informações estejam postadas na Web de maneira estruturada e semanticamente correta e além disso é necessário que o próprio usuário deixe bem claro o que ele quer buscar, para que a busca seja bem realizada.

Segundo (BAX) ,atualmente existem máquinas que com base nos dados regionais, ou seja de onde veio a consulta, nos dados semânticos , que são se os nomes da consulta

são nome de cidades, de pessoas, e com base em dados de determinação do idioma, se é português inglês, com base neste dados eles levantam qual foi a consulta do usuário e fazem uma busca pela Web utilizando estes dados acima, isto já é uma maneira de tentar entregar ao usuário uma reposta de maior qualidade. Como exemplo de máquinas que evoluem nesta direção podemos citar a Wisenut (WISENUT) e a Teoma (TEOMA) onde, pode-se observar que ambas as máquinas como trabalham com base na semântica das informações nos trazem resultados piores que as máquinas de primeira e segunda geração, porque a web atual não está preparada ainda para receber as máquinas de busca semânticas.

Para deixar mais claro como é feito esta busca iremos citar um exemplo onde uma determinada senhora de nome Juliana pretende viajar de Belo Horizonte para Campinas no dia 15 julho no horário das 10;00 em um avião na classe de luxo, como se trata de uma viagem de negócios de muita importância ela quer comprar a passagem aérea com antecedência para não correr o risco de perder o compromisso.

Neste caso teremos duas situações:

A primeira que está ilustrada pela figura 1 onde a senhora Juliana entra na Internet para procurar qual empresa aérea faz esta linha, neste caso ela digita em uma Máquina de Busca que ela pretende comprar uma passagem com as características citadas anteriormente, a Máquina de Busca retornará para ela alguns sites onde foram encontradas estas palavras chaves, mas ela terá que abrir cada página e ver qual das páginas disponibilize que ela compre a determinada passagem, e mesmo assim ela terá que entrar em várias páginas e ir anotando as características de cada empresa, como preço da passagem, se tem exatamente o horário que ela quer, para somente depois decidir e realizar a compra da passagem.

Já na outra situação que está ilustrada pela figura 2 a Juliana simplesmente irá entrar na Internet e digitar as características da passagem como já foi citado acima, em uma Máquina de Busca, neste caso os dados da passagem serão a requisição que a Juliana estará passando para que a Máquina de Inferência receba está requisição e vá até as empresas aéreas e ache qual delas tem as características da requisição que a Juliana passou,

neste caso a Máquina de Inferência retornará para a Juliana algumas opções, onde a Juliana terá somente o trabalho de escolher qual das opções ela acha melhor, ou seja, com o uso da Web Semântica e das Máquinas de Busca Semântica poderemos fazer com que a maior parte do trabalho seja realizado pela própria máquina.

7 – Comparativo entre Máquinas de Buscas Semânticas e Máquinas de Buscas Indexadas em HTML.

Como citamos o exemplo acima da Senhora Juliana, iremos agora fazer um comparativo como mostra as figuras 1 e 2 que estão logo a seguir.

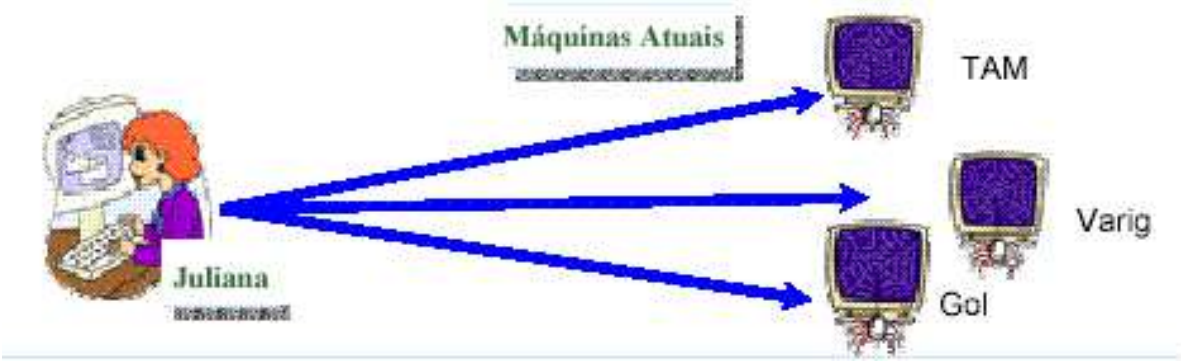

Figura 1(Máquinas Indexadas em HTML)

Figura 2(Máquinas Semânticas)

Podemos ver que na Web Semântica o que irá acontecer é o seguinte, quando a Juliana digitar sua requisição em uma Máquina de Busca, que neste caso será uma máquina de terceira geração, está máquina irá procurar pela web, que neste caso estará

semanticamente correta e estruturada, ou seja estará seguindo o Padrão de Metadados DC e utilizando RDF e XML , ao fazer está procura ela poderá achar ou não a informação desejada, caso ache retornará para a Juliana as opções que tem disponíveis, e a Juliana terá somente o trabalho de optar por qual das descrições ela irá querer.

Vimos por estas figuras que as buscas semânticas serão muito mais eficientes que as buscas indexadas em HTML, mas porém a falta de informações semânticas no serviço de informação nos impedirá de migrar para as máquinas de busca semânticas, ou seja para haver esta migração será necessário nos lembrarmos dos seguintes tópicos:

- O usuário deve digitar corretamente os dados para a busca;
- A Web deve estar com suas informações postadas de maneira estruturada e semanticamente corretas;
- As máquinas de buscas devem ser de terceira geração, que fazem a recuperação da informação com base na semântica dos dados;
- Além de serem de terceira geração estas máquinas devem estar funcionando totalmente corretas

8 – Considerações Finais

Com este artigo é possível observar que tudo na vida evolui, como o carro, as máquinas, os estudos, as pesquisas, assim também a web precisa evoluir e a evolução chama-se Web Semântica.

Podemos observar que encontraremos várias dificuldades para implantar a Web Semântica, por existir um universo muito grande de páginas, de websites, ou seja, onde cada uma delas segue um modelo próprio sem padronização, ou melhor muitas delas utilizam linguagem HTML que nos impossibilita de fazer implantarmos a Web Semântica devido a suas restrições.

Uma das grandes dificuldades que vemos hoje na implantação da Web Semântica é que as milhões de Páginas publicadas na web não utilizam RDF e como já vimos o RDF é complicado de ser utilizado por desenvolvedores médios.

Tivemos a oportunidade de vermos também que de acordo com um estudo conduzido pela *Georgia Institute of Technology*, 85% dos clientes em potencial usam máquinas de busca para encontrar soluções e produtos na web, ou seja ai está quanto é importante para nós da área de computação investirmos neste campo de trabalho, porque se tivermos mecanismos que realizem estas buscas qualidade, poderemos aumentar cada vez mais o comércio eletrônico, aumentando assim cada vez mais esta fatia de mercado que é de tanta importância para nós.

Vejo também que a Web Semântica depende, enormemente, da criação de padrões de nomenclatura para a descrição de recursos na Web. Caso contrário, o problema persistirá com criadores de conteúdo atribuindo descritores diferentes para o mesmo significado.

9 - Trabalhos Futuros

Abaixo deixamos duas idéias como trabalhos futuros que são muito interessantes:

- Fazer um estudo de como migrar de máquinas de busca Indexadas em HTML para máquinas de busca Semântica.
- Fazer um estudo de como é qualidade da informação retornada pelas Máquinas de Terceira Geração existentes.

10 - Referências Bibliográficas

QUEIROZ, Rubens - O futuro da Internet. Disponível em: http://www.comciencia.br/reportagens/internet/net08.htm - Acessado em 22 de fevereiro de 2006

SILVA, George Henrique – Construção de Agentes inteligentes para Web Semântica. Disponível em: http://www.linux.ime.usp.br/~ghsilva/websemantica/ - acessado em 19 de fevereiro de 2006

DZIEKANIAK, Gisele Vasconcelos - Semantic Web. Disponível em: http://www.encontros-bibli.ufsc.br/Edicao_18/2_Web_Semantica.pdf - acessado em 15 de fevereiro de 2006

CAMPOS, Fernando - Como os Mecanismos de Busca Indexam Páginas HTML – http://www.bax.com.br/research/publications/ComoMaquinasBuscaIndexamPaginasWeb.pdf - acessado em 22 de fevereiro de 2006

ARTIFICE1, - O que é Web Semântica ? Disponível em: http://www.artifice.web.pt/blog.html#pag=00/web-design/02-11-2004-web-semantica.html - acessado em 03 de junho de 2006.

OLIVEIRA, Fábio Braga - Problemas com a Web Semântica – Disponível em : http://www.linux.ime.usp.br/~fbdo/websemantica/node6.html - Acessado em 04 de junho de 2006.

TABLELESS, Conceitos sobre a Web Semântica – Disponível em: http://www.tableless.com.br/aprenda/a-web-semantica/ - Acessado em 02 de junho de 2006.

PARREIRAS, Fernando - Web semântica: a solução para o caos da Internet – Disponível em: http://www.fernando.parreiras.nom.br/content/view/36/42/ - Acessado em 04 de junho de 2006.

METADADOS, – O que são Metadados ? . Disponível em: http://www.linhadecodigo.com.br/artigos.asp?id_ac=298 – acessado em 06 de junho de 2006.

BAX, Marcelo Peixoto - Projeto Indexa “Ferramentas de auxilio à divulgação de informação na Web”. Disponível em : http://cuba.paradigma.com.br/paradigma/artigos/artigos_04.pdf - acessado em 08 de maio de 2005.

ARAÚJO, Moises - Educação a Distância e a Web Semântica: Modelagem Ontológica de Materiais e Objetos de Aprendizagem para a Plataforma COL. – Escola Politécnica Pontifica de São Paulo maio de 2003.

WIKIPEDIA, – Conhecendo o XML - Disponível em: http://pt.wikipedia.org/wiki/XML - Acessado em 12 de Junho de 2006.

ZANETE, N.H – Introdução ao RDF - [2002?] – Disponível em: www.faccar.com.br/zanett/zanete.RDF.htm. Acessado em 12 de junho de 2006.

WISENUT, - Máquina de Busca - Disponível em *www.wisenut.com* *Acessado em 12 de junho de 2006.*

TEOMA, - Máquina de Busca – Disponível em www.teoma.com - Acessado em 12 de junho de 2006.

Web Semântica para Máquinas de Busca.

Autor: ELi Nogueira Milagres Júnior

Orientador : Élio Lovisi Filho

Barbacena , junho de 2006

Web Semântica para Máquinas de Busca.

Autor: ELi Nogueira Milagres Júnior

Projeto de pesquisa apresentado como pré-requisito da disciplina trabalho de graduação II.

Barbacena , junho de 2006