ARTIGO

Introdução a Web Semântica, Ontologia e Máquinas de Busca

Alysson Vicuña de Oliveira

***Resumo*—As estruturas das informações disponíveis na web, atualmente não estão bem definidas. Sendo assim, um programa que precise extrair dados importantes ou específicos, como uma máquina de busca, nem sempre obterão êxito, podendo recuperar informações pouco relevantes devido a grande quantidade de informações, sendo a grande maioria desestruturada. Neste contexto, a *web* semântica surge com o objetivo de introduzir uma estruturação aos dados e, para isso, utiliza-se de ferramentas tais como XML, RDF e ontologias para definição de hierarquia criando uma inter-relação de conceitos.**

***Palavras-chave*—web semântica, ontologia, máquina de busca**

I. Introdução

O computador, nos dias atuais, é ferramenta indispensável em diversas áreas do conhecimento humano, tanto para produção do conhecimento, quanto para entretenimento ou mesmo uma simples navegação na internet. Desde a disseminação do uso de computadores, a principal preocupação da área de Tecnologia de Informação tem sido prover informações para apoiar a resolução de problemas. Todos os dias milhares de novas páginas são publicadas na Internet, tornando cada vez mais difícil e demorado encontrar informações relevantes. Esta demora e dificuldade para encontrar informações úteis em tempo hábil acabam prejudicando uma grande quantidade de negócios e oportunidades, simplesmente por falta da informação na hora ou no formato certos.

No contexto específico da *Web*, este problema foi identificado e as iniciativas para tentar minimizar seus efeitos deram origem à área de pesquisa denominada *Web* Semântica (*Semantic Web*). Este artigo tem por objetivo apresentar as principais ferramentas e tecnologias que permeiam o estudo da *Web* Semântica, que está sendo reconhecida como o próximo passo evolutivo da Internet. A *Web* Semântica (WS) pode representar uma revolução na maneira de enxergamos a internet. Uma das aplicações mais evidentes da *web* semântica são as máquinas de busca e por esse motivo este artigo irá utilizá-las como motivação principal. A seção 2 irá apresentar as classificações das máquinas de busca. A seção 3 irá introduzir os principais aspectos da *web* semântica. As seções VII, VIII e IX irão apresentar respectivamente o XML, RDF e o conceito de ontologia, ferramentas que começam a ser utilizadas como referência quando se fala de *web* semântica.

O autor agradece a Escola de Tecnologia da Faculdade Projeção pelo incentivo e oportunidade de divulgação do trabalho.

I. Oliveira é professor da Escola de Tecnologia da Faculdade Projeção. Contato: alysson.vicuna@gmail.com.br.

II. Classificação de Máquinas de Busca

A internet foi criada, inicialmente com propósitos militares e, posteriormente se estendeu ao meio acadêmico. Depois disso, vem crescendo e se popularizando cada vez mais, podendo, ser acessada por qualquer cidadão em praticamente qualquer lugar em que exista um computador ligado a uma linha telefônica. Contudo esse crescimento ocorreu de forma desordenada e sem nenhum controle, pois qualquer pessoa pode publicar um documento na internet. Isso se deve à facilidade de se construir páginas, já que o HTML ainda é a principal linguagem para construção de páginas web, possuindo a característica de ser simples e possibilitar que todos os dias novos documentos sejam disponibilizados sem controle de conteúdo, facilitando assim a edição e remoção de documentos com facilidade por seus criadores.

Devido a esse grande contingente de informações disponíveis, a localização de informações relevantes ficou mais difícil e, por isso, pensou-se numa solução rápida e eficiente de acesso e localização dessas informações: as máquinas de busca. Elas utilizam como parâmetro a consulta requerida pelo usuário para vasculhar documentos *web*, processando essas informações e retornando uma lista dos documentos que apresentaram similaridade com o assunto desejado. Essas informações são classificadas como informações intrínsecas e informações extrínsecas.

Informações intrínsecas são informações contidas dentro dos documentos que estão sendo analisados pelas máquinas de busca, ou seja, o próprio texto do documento. A máquina de busca analisa a ocorrência de uma determinada palavra ou frase e sua localização no texto e classifica as páginas com maior número de ocorrência como as mais importantes para o usuário, criando um ranking de páginas tidas como relevantes. Já informações extrínsecas a uma página, são obtidas a partir dos demais documentos contidos na coleção, estrutura de *links* (*link analysis*) ou popularidade de um documento em relação a outro (*usage analysis*). Com base nestas informações podemos classificar as máquinas de busca da seguinte maneira:

- Primeira Geração: O processo de recuperação da informação consiste basicamente em utilizar as informações intrínsecas aos documentos.

- Segunda Geração: Nesta geração o processo de recuperação das informações utiliza como critério de seleção dos documentos as informações intrínsecas e extrínsecas.

- Terceira Geração: Recuperam a informação com base na semântica (sentido) das informações contidas nos documentos, ou seja, recuperam dados com base em informações estruturadas semanticamente.

As máquinas da terceira geração introduziram novas necessidades na *web* (internet), como por exemplo, ter conhecimento semântico do conteúdo do documento, ou seja saber do que se trata uma imagem inserida em uma página da *web*, obrigando uma estruturação dos documentos disponibilizados. A proposta da resolução dos problemas na busca de informações relevantes dentro de um contexto desejado é trazida pela *web* semântica, juntamente com outros mecanismos funcionais.

III. WEB SEMÂNTICA (*WEB SEMANTIC*)

Segundo o idealizador da *web* semântica (WS) Tim Berners-Lee[1], *Web* Semântica é uma extensão da *Web* atual na qual a informação possui significado bem definido, permitindo assim que computadores e pessoas trabalhem melhor, possibilitando que haja cooperação entre eles. Ele também foi o responsável pela criação de conceitos importantíssimos para o sucesso da internet, tais como a WWW, URLs, http e o próprio HTML. Atualmente ele trabalha liderando um grupo de pesquisadores no *World Wide Web Consortium* ou W3C, com o objetivo de melhorar, estender e padronizar os sistemas *web*.

A *Web* Semântica surge como uma possível solução para a estruturação dos dados na *Web*, permitindo a criação de um contexto no qual a informação possa ter significado para humanos e para máquinas, que se encarregarão de levar a informação relevante para o usuário. Segundo Berners-Lee [1] o principal desafio da *Web* Semântica é criar uma linguagem que consiga expressar o significado e ao mesmo tempo estabelecer regras para processar esse significado de forma a inferir novos dados e regras. As regras para o processamento do significado devem permitir que outros sistemas inteligentes possam interagir.

Berners-Lee[1], cita um exemplo do que a *Web* Semântica será capaz de fazer: o usuário realizará uma pesquisa na Internet para encontrar um médico de uma determinada área da Medicina, estabelecendo algumas restrições, tais como "o médico deve ter consultório no mesmo bairro onde moro e deve estar ligado à comunidade acadêmica". Quando a busca for solicitada um agente de pesquisa navegará pela rede e encontrará algumas possibilidades. De maneira inteligente e autônoma, o agente deverá comparar a agenda do usuário com a agenda do médico e oferecer opções de horários para consulta. O usuário só terá o trabalho de escolher o horário que melhor lhe convier.

Na proposta de desenvolvimento da *Web* Semântica é sugerida uma arquitetura de 3 camadas:

- Camada Esquema: responsável por estruturar os dados e definir seu significado, para que possa elaborar um "raciocínio lógico". Essa camada é o primeiro passo em direção a *Web* Semântica;

- Camada Ontologia: responsável por definir as relações entre os dados. Essa camada é responsável pela formação do entendimento comum e compartilhado de um domínio;

- Camada Lógica: responsável pela definição de mecanismos de inferência sobre os dados, sendo composta por um conjunto de regras de inferência que os agentes poderão utilizar para relacionar e processar informações.

A Fig. 1 apresenta uma arquitetura proposta por Berners-Lee para a *Web* Semântica, composta de três camadas:

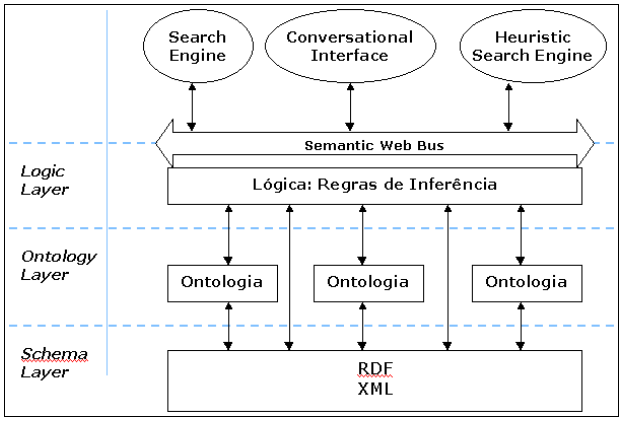

Fig. 1. Arquitetura da Web Semântica.

IV. CAMADA DE ESQUEMA (*SCHEMA LAYER*)

A camada esquema provê uma forma de definir os dados do documento e o significado associado a esses dados. Trata também da estruturação e disposição dos dados de forma que os programas que rodam na *web* possam fazer inferência a partir dos mesmos.

Para que haja a representação do conhecimento são necessárias três condições:

- Interoperabilidade Estrutural: Permite que os dados sejam representados de forma distinta, permitindo especificar tipos e possíveis valores para cada forma de representação;

- Interoperabilidade Sintática: Constitui-se de regras precisas que permitem o intercâmbio de dados na *Web*;

- Interoperabilidade Semântica: Possibilita a compreensão e associação entre os dados.

As Linguagens utilizadas para atender esses requisitos são XML e RDF, pois permitem expressar os dados para definir regras de raciocínio. A XML e a RDF que serão descritos mais adiante neste documento, nas seções VII e VIII, respectivamente.

V. A CAMADA ONTOLOGIA (*ONTOLOGY LAYER*)

Duas bases de dados podem utilizar terminologias diferentes para referir-se à mesma informação, resultando em divergências para um mesmo conjunto semântico de dados. Pode ocorrer também de uma mesma terminologia estar sendo utilizada com significados diferentes, por aplicações distintas. Para tratar esses conflitos, existe a camada de ontologia que define mecanismos capazes de estabelecer um padrão entre as páginas de *web*. As ontologias serão tratadas com mais detalhes na seção IX.

VI. A CAMADA LÓGICA (*LOGIC LAYER*)

É na camada lógica que são possíveis os relacionamentos de informação e as inferências de conhecimento da Web Semântica. As regras de inferência fornecem aos agentes (programas) poder de raciocinar sobre os termos e seus significados, que foram definidos na camada esquema e de raciocinar a respeito dos relacionamentos entre os conceitos segundo a sua definição na camada ontologia.

Os agentes são sistemas computacionais autônomos que travam diálogos, negociam e coordenam transferência de informações para atingir os objetivos do seu criador ([2]). De acordo com [3], os agentes possuem algumas características como autonomia (funcionam sem intervenção humana), reatividade (percebem o ambiente tomam as decisões), têm comportamento colaborativo, possuem objetivos, são flexíveis, sociáveis e têm a capacidade de aprender.

A WS possuirá vários agentes interagindo entre si, compreendendo, trocando ontologias, adquirindo novas capacidades racionais quando adquirirem novas ontologias, formando cadeias que facilitam a comunicação e a ação humana.

VII. XML (*EXTENSIBLE MARKUP LANGUAGE*)

O surgimento da XML, em 1996, revolucionou a Web e as formas como as aplicações trocam e representam dados. Mas afinal de contas o que vem a ser XML? Uma linguagem de programação? Ou um substituto do HTML? Nem uma coisa e nem outra. De acordo com [4], XML nada mais é do que uma linguagem de representação de dados cujo foco é a semântica dos dados representados e não sua forma de apresentação.

XML é uma linguagem de marcação extensível (*eXtensible Markup Language*), derivada do SGML, de onde também veio o HTML, por isso a confusão. Contudo ao contrário do HTML, que possui *tags* limitadas, em XML as *tags* não são pré-definidas. Elas podem ser definidas de acordo com o significado do dado que se quer representar.

No exemplo simples ilustrado na Fig. 2, é mostrada uma implementação em XML. Perceba que em XML existem *tags* específicas para objetos como monitor, modelo. Portanto é atribuído um significado bem definido de certas unidades na página criada. Tais unidades podem ser então manipuladas por aplicações que conhecem seus significados. É um primeiro passo em direção à *Web* Semântica.

Fig. 2. Exemplo Flexibilidade do XML.

```
<microcomputador>
      <modelo>Core 2 Quad</modelo>
      <ram>8 GB</ram>
      <monitor>LCD 17 Polegadas</monitor>
      <teclado> Microsoft Wireless</teclado>
</microcomputador>
```

VIII. RDF (*RESOURCE DESCRIPTION FRAMEWORK*)

O RDF é um modelo de dados para objetos (recursos) e relações entre eles; ela provê uma semântica simples para o modelo, o qual pode ser representado em sintaxe XML.

Frequentemente chamada de "linguagem", RDF é essencialmente um modelo de dados. Seu bloco de construção básico é uma tripla objeto-atributo-valor, chamada de *statement* (declaração). O significado da tripla <objeto, atributo, valor> é de que o objeto X tem o valor Y para certo atributo Z.

O exemplo abaixo ilustra a escrita RDF utilizando a sintaxe XML:

```
<disciplina nome = "Engenharia de Software">
   <professor>José</professor>
</disciplina>
```

A sentença precedente sobre Jose é um *statement*. É claro que um modelo de dados abstrato precisa de uma sintaxe concreta para ser representado e transmitido, e, nesse sentido, o RDF tem sido usado sobre XML. Como resultado, ele herda os benefícios associados com XML. Contudo, é importante entender que outras representações sintáticas de RDF, não baseadas em XML, são também possíveis; a sintaxe baseada em XML não é um componente necessário do modelo RDF [1].

IX. ONTOLOGIAS (*ONTOLOGYS*)

Uma ontologia define termos para que um agente de software consiga extrair o máximo de informação possível de um documento. Ela fornece um entendimento comum e compartilhado de um domínio, que pode ser comunicado através de pessoas e sistemas de aplicação, tornando-se fator chave para o desenvolvimento da *Web* Semântica [5].

A ontologia tem um papel crucial no sentido de permitir o acesso, a inter-operação e a comunicação baseados em conteúdo, fornecendo à *Web* um nível de serviço qualitativamente novo, que consideramos na *Web* Semântica, pois permitem expressar regras possibilitando a um programa deduzir significados da informação guardados no documento, ou seja, permitem manipular os termos de uma maneira mais útil e eficiente.

Ela une em rede incríveis porções do conhecimento humano, complementando-as com capacidade de processamento de máquina.

Segundo [6], a utilização de ontologias permite lidar com conceitos, representando-os formalmente, e de se livrar de problemas inerentes ao vocabulário da linguagem natural tais como homonímia (nomes iguais), sinônimos, metonímia etc.

De acordo com [7], facilidades de documentação, manutenção e confiabilidade também são características importantes das ontologias assim como as propriedades compartilhamento e filtragem. [8], afirmam que "ontologia permite acesso inteligente aos documentos na *Web* e infere ou deduz o conhecimento implícito das regras e fatos declarados explicitamente na ontologia".

X. Considerações Finais

Durante a elaboração deste artigo, procurou-se mostrar de forma clara e especifica as inúmeras facilidades que a *Web* Semântica trará para as pessoas que utilizam computador, ajudando-as a obterem informação de qualidade, em meio aos inúmeros documentos existentes na *web*.

Com o auxilio das ontologias e *Web* Semântica, quando um usuário efetuar uma pesquisa utilizando uma maquina de busca, esta irá retornar ao requerente, apenas resultados relevantes ao contexto desejado, evitando que os usuários venham a sofrer prejuízos no tocante a qualidade de negócios e oportunidades.

Mas para que as facilidades da WS estejam disponíveis a todos os usuários, faz se necessário a construção de ferramentas que preparem o conteúdo das paginas de forma semanticamente estruturadas. Deseja-se ainda a automação da arquitetura da *web* Semântica, que atualmente segue uma proposta utilizando XML para estruturação dos dados, modelos de RDF para a representação semântica e os relacionamentos são gerados utilizando ontologias, manipulando manualmente diferentes arquivos.

Espera-se ainda que com o desenvolvimento de ferramentas especializadas, seja possível trabalhar apenas com as ontologias, sendo as outras duas camadas do modelo criadas automaticamente.

Referências

[1] BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web – a new form of the Web content that is meaningful to computer will unleash a revolvution of new possibilities. Scientific American, May 17, 2001. Disponível em: <http://www.sciam.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>. Acesso em: 20 abr. 2006.

[2] BONIFÁCIO, A. S.; HEUSER, C. A. Ontologias e consultas semânticas: uma aplicação ao caso Lattes. Disponível em: < http://www.uel.br/pessoal/ailton/Trabalhos/Disserta%E7ao%20de%20Mestrado-Ailton-Final.pdf >. Acesso em: 05 jun. 2011.

[3] DUARTE, O. C. M. B.; FURTADO JUNIOR, M. B. Tutorial XML. Disponível em: <http://www.gta.ufrj.br/grad/00_1/miguel/>. Acesso em: 18 abr. 2006.

[4] BRAGANHOLO, Vanessa P.: Gerenciamento de Dados XML. Revista Computação Brasil, ano VII- n°21-Março/Abril e Maio de 2006. Sociedade Brasileira de Computação.

[5] JASPER, R.; USCHOLD, M. A framework for understanding and classifying ontology applications. Disponível em: <http://sern.ucalgary.ca/KSI/KAW/KAW99/papers/Uschold2/final-ont-apnfmk.pdf>. Acesso em: 23 abr. 2006.

[6] MORAIS, Erikson Freitas de and SOARES, Marcelo Borghetti: Web Semântica para Máquinas de Busca. Universidade Federal de Minas Gerais.

[7] CUNHA, Luiz. M. Silva. Web semântica: Estudo Preliminar,. Campinas: Embrapa, 2002. 16 p.

[8] STAAB, S.; MAEDCHE, A. Knowledge portals – ontologies at work. Disponível em: <http://www.aifb.uni-karlsruhe.de/WBS/Publ/2001/KP-OaW_sstama_2001.pdf>. Acesso em: 24 abr. 2006.

Alysson Vicuña de Oliveira Graduado em Ciência da Computação pela Universidade de Rio Verde – FESURV, Especialista em Desenvolvimento de Software para Web pela Faculdade Cathedral. Professor da Escola de Tecnologia da Faculdade Projeção. Analista de Sistemas Pleno e atualmente consultor OEI do Ministério da Educação - MEC.