Universidade do Porto

Faculdade de Engenharia

FEUP

# Semantic Web

Ana Isabel Pinto Correia
Sérgio Sobral Nunes

Armazenamento e Recuperação da Informação
**Mestrado em Gestão de Informação, Junho 2002**

# Introdução

A Internet, dos dias de hoje, é um meio de comunicação poderoso onde se pode encontrar uma quantidade abundante de informação sobre qualquer assunto que se pretenda pesquisar. Esta rede global (World Wide Web) foi idealizada para a comunicação humana o que se revela um obstáculo quando se tenta evoluir para uma rede em que haja maior cooperação com agentes computacionais.

O W3C (World Web Consortium) tem-se empenhado em maximizar o potencial da rede global tornando a Internet uma ferramenta acessível a todos. Esta organização visa criar um conjunto de regras e protocolos que devem ser seguidos por todos que participam, como publicadores de informação, na construção do grande repositório global de informação que é a Internet.

Um recente projecto da W3C é a criação da *Web Semântica* (**WS**). A WS surge como uma extensão à rede global actual na qual é possível atribuir conteúdo às páginas para que os computadores sejam capazes de captar o significado da informação e de a relacionar. Ou seja, ser possível a um agente computacional fazer inferência sobre os dados equiparando-se a um humano.

Tim Berners Lee, James Handler e Ora Lassila focaram o seu estudo no futuro desta nova rede com o objectivo de melhorar e facilitar a comunicação entre o homem e a máquina.

# Semantic web

A WS é apresentada como uma rede que visa permitir expressar a informação num formato que as máquinas possam processar. A ideia consiste em desenvolver tecnologias que permitam às máquinas tirar mais partido da Web, com o objectivo de a tornar mais útil para os humanos.

A recuperação de documentos relevantes torna-se uma tarefa pouco eficaz nos dias de hoje. Embora os motores de busca estejam cada vez mais poderosos, o que recuperam a partir de alguma palavras-chave são várias respostas que são necessárias analisar para verificar se são adequadas ao contexto que esperávamos. Uma das possibilidades para optimizar esta tarefa seria automatizar a pesquisa onde agentes computacionais fossem recuperar, apenas, a informação que fosse interessante. A WS possibilita que a automatização da pesquisa como foi referida não seja inatingível.

A *Web Semântica* permite a estruturação da informação e a atribuição de significado para que seja possível a um computador processar a informação contida em diferentes páginas Web captando o seu conteúdo, interpretando a sua lógica e relacionando-as entre si. Deste modo, definindo linguagens para a expressão de conteúdos, representação e interpretação dos dados o computador pode “compreender” a informação facilitando a sua pesquisa, a sua integração e a sua utilização.

Na base deste conceito estão 2 tecnologias em fase avançada de desenvolvimento: a XML (eXtended Markup Language) para estruturação de dados, e a RDF (Resource Description Framework) para exprimir significado. Com recurso a sistemas periciais e um mecanismo de ontologias, os agentes podem "perceber" e processar a informação existente.

## XML

XML é uma linguagem estrutural que foi desenvolvida partindo dos princípios básicos do HTML(HyperText Markup Language). A sua estrutura pretende melhorar a gestão da informação para satisfazer os novos requisitos do crescimento da Internet. No XML, os programadores criam a sua própria linguagem de marcas (“tags”) para organizar os conteúdos que pretendem disponibilizar.

## RDF

A tecnologia RDF permite equivaler um significado aos dados estruturados em XML. A cada marca (“tag”) é associado o seu conceito real. Essa equivalência é realizada por uma estrutura de três elementos: o sujeito, o predicado e o objecto. Esta tripla de elementos pode ser comparada com a forma de organizar uma frase em linguagem natural. A informação referida é armazenada também em XML sendo o conteúdo de cada elemento

da tripla um URI (Uniform Resource Identifier). Um URI é um apontador para outro recurso disponibilizado na Internet e que contém o significado de cada um desses elementos.

Estes recursos externos referenciados pelos URI são a fonte onde se deve obter o significado real da marca XML utilizada. Este método é usado em detrimento da pesquisa num repositório de conceitos e definições para as marcas. O que permite ao autor do documento em XML ser o único responsável pelo o significado das marcas que ele próprio cria.

O uso destas tecnologias coloca sobre o controlo do publicador da informação o conceito que pretende equivaler à sua informação. A associação entre informação e seu significado é realizada de uma forma completamente livre, sem rigidez de regras e sem quaisquer limites. Cada triplo de URI associado a uma marca é considerado como sendo único. Esta informação é invisível ao utilizador que está a ver a página, mas permite ao indexador automático armazená-la de modo a refinar as suas pesquisas.

Esta forma de armazenar o significado e a estrutura do documento permite constituir uma rede de informação relacionada, onde será possível aos homens e às máquinas aplicar regras de inferência para criar deduções a partir dos significados descritos nos URIs de cada tripla de informação.

## Ontologias

Depois de atingir este nível de organização, estruturação e significado para a informação que é disponibilizada na Web, surge um novo problema: como relacionar os significados que na realidade são semelhantes e podiam ser agrupados num só. A capacidade de descobrir estes agrupamentos iria permitir criar uma rede mais ligada entre si. Existem diversos conceitos que significam a mesma coisa, mas que podem ter sido explicados de forma ligeiramente diferente, ou mesmo numa língua diferente.

A W3C e os responsáveis pelo projecto da *Web Semântica* avançaram com a utilização de ontologias para permitir que toda a informação possa ser processada e relacionada, se possível, estabelecendo relações entre os conceitos encontrados.

Ontologia é a “parte da metafísica que estuda o ser em si, as suas propriedades e os modos por que se manifesta”. Este significado foi adaptado pelos investigadores de inteligência artificial e da Web para o seu próprio meio e, para eles, uma ontologia, é apenas um ficheiro ou documento que define relações entre os conceitos. Estes ficheiros têm uma organização mais rígida e são escritos numa linguagem com limitações de sintaxe para esquematizar as relações. Adicionalmente as ontologias fornecem um conjunto de regras de inferência, oferecendo ainda mais poder às relações entre os conceitos.

Essas regras de inferência permitem aos agentes computacionais realizar deduções lógicas expressamente definidas nas regras. Sendo estabelecida a regra que a cidade do Porto pertence a Portugal, quando for definida a regra que a FEUP está situada no Porto seria imediatamente descoberto que o seu país era Portugal. A utilização destas regras,

permite deduções que, não sendo directamente do entendimento do agente computacional, podem fazer sentido e ser bastante úteis aos humanos.

Com a definição destas regras de inferência, a informação disponibilizada na Web ganha um significado muito mais detalhado, pois para cada termo correctamente identificado é possível obter o seu conceito real e outras informações que lhe possam ser relacionadas. Este detalhe vem eliminar diversos problemas de indefinição e confusão de termos utilizados na Web. Problemas que são bastante evidentes nos nossos dias na simples tentativa de recuperação de informação num indexador de páginas pelo uso do seu motor de pesquisa.

O uso de ontologias oferece aos motores de busca a possibilidade de refinarem as suas pesquisas, adaptando as suas bases de dados e o funcionamento dos seus agentes, a estas novas tecnologias.

## Agentes

O grande “poder” da WS será atingido quando forem criados os programas que serão capazes de recolher, tratar e armazenar a informação publicada com definições de semântica.

Actualmente existem já diversos agentes espalhados pela Internet, pequenos programas que percorrem a rede em busca do relacionamento de qualquer tipo de informação. A forma mais comum de agentes são os utilizados pelos indexadores de páginas que realizam acções de manutenção ou expansão da sua base de dados: percorrem todos os endereços que existem em cada página em busca de páginas ainda não catalogadas e verificam alterações realizadas nas já catalogadas.

No caso da WS, espera-se que os agentes realizem estas acções catalogando da melhor forma a informação, utilizando a sua definição de estruturas, os conceitos de cada estrutura, as suas relações e as regras de inferência entre conceitos.

Seguindo as linhas deste projecto da *Web Semântica*, que pretende promover as sinergias e a ligação de informação, os agentes, além das capacidades já referidas, iriam também comunicar entre si partilhando a informação quando esta tiver o seu significado já definido. Esta partilha de informação entre os agentes computacionais vem reforçar a ideia de que a Web pode ter uma linguagem única.

Para assegurar a veracidade e a fiabilidade da troca de informação entre os agentes, surge a utilização de uma tecnologia bastante desenvolvida e actualmente usada, principalmente nos serviços de banca e no comércio electrónico, os certificado digitais. Como poderia um agente assegurar-se que o outro pertencia a uma entidade responsável e credível, e não a alguma entidade ou pessoa mal intencionada que pretendia gerar confusão na forma como funciona a Internet?

Em todos os serviços disponibilizados actualmente, é vulgar surgirem técnicas de exploração dos pontos fracos de algumas tecnologias com o simples intuito de perturbar o seu funcionamento. A forma de garantir a responsabilização pelas partilhas de informação

é conseguida através da certificação digital, um processo simples que consiste numa assinatura e numa referência para uma fonte tida como de confiança. Os agentes devem categorizar a troca de realizada com outros agentes depois de verificar as fiabilidade da fonte.

Já existem algumas capacidades semelhantes às que se pretendem dos agentes. Existem serviços na Web que são capazes de atribuir informação semântica a alguns dados. Estes apenas podem ser utilizados pelos agentes caso exista uma definição correcta do que cada função disponibilizada representa e como deve ser utilizada. Os agentes e esses serviços podem publicitar e partilhar as suas definições de funções permitindo a outros agentes utilizar as associações semânticas com a certeza do tipo de equivalência de informação que fornecem. Alguns exemplos mais comuns de serviços deste tipo são os dicionários, os tradutores de línguas, conversores de medidas e moedas.

# O futuro da Semantic Web

Neste artigo, os autores descrevem um cenário futurista onde agentes autónomos utilizam a *Web Semântica* para desempenhar tarefas a pedido dos utilizadores.

O objectivo da WS é tirar todo o potencial da Web, adicionando-lhe informação que pode ser entendida e processada por máquinas, automatizando serviços que permitem expandir largamente as suas actuais capacidades. A automatização dos serviços irá melhorar a forma de auxiliar os humanos a alcançarem os seus objectivos, "percebendo" melhor os conteúdos da Web e fornecendo a pesquisa, filtragem e categorização da informação de um modo muito mais preciso. Este processo irá conduzir a um caminho onde a Web é um sistema mais cognitivo e que disponibilizará mais serviços especializados de equiparação lógica da informação.

As ideias e conceitos avançados neste artigo parecem-nos muito promissoras e viáveis a curto prazo, pelo menos no que diz respeito a informações simples (por exemplo: cotações, contactos, etc). Por outro lado, se tentarmos imaginar a forma como serão tratadas as informações mais complexas, encontraremos problemas que poderiam levar a uma má estruturação das bases do conhecimento. Para o conceito da WS ser bem sucedido, será ainda necessário muito trabalho de pesquisa e de esquematização na tentativa de representar o conhecimento. Estas técnicas já foram bastante aprofundadas na área da inteligência artificial e continuam a existir divergências sobre a melhor forma de modelizar o conhecimento.

As tecnologias referidas para representar o conhecimento são inovadoras e necessitam ainda de serem reconhecidas por todos os intervenientes na Web. Apesar do XML ser alvo de um consenso e apoio generalizado, o mesmo não acontece com as outras tecnologias apresentadas, nomeadamente a RDF. Achamos que o sucesso desta arquitectura está muito dependente da aceitação destas ideias por parte da indústria. Veja-se, a título de exemplo, o caso dos *browsers* e das guerras travadas entre a Microsoft e a Netscape, que ainda dificultam muito a generalização das normas definidas pelo W3C. Se compararmos os intuitos originais de Berners-Lee ao criar a linguagem HTML com o que na realidade aconteceu, temos uma boa percepção do poder dos grandes fabricantes no que respeita à definição de normas.

Outro aspecto que nos parece importante para a difusão e popularização destas ideias é a necessidade em especificar mais em detalhe a aplicação prática destas noções. Estes incentivos podiam ser conseguidos, por exemplo, disponibilizando alguns casos práticos já em funcionamento.

Uma das maiores dificuldades para o sucesso da WS será incentivar os publicadores da informação a estruturarem-na segundo as normas definidas pela W3C. A aplicação de novos conceitos irá gerar custos de reformulação da sua presença na Internet. A curto prazo não existirá o retorno em relação ao investimento o que influencia na aceitação de projectos para a aplicação destas novas tecnologias.

# Bibliografia

Berners-Lee T., Hendler J., Lassila O. - The Semantic Web. *Scientific American* (Maio 2001).

Berners-Lee T. –Semantic web road map
http://www.w3.org/DesignIssues/Semantic.html

Brickley, D. – Semantic web history: nodes and arcs 1989 – 1999
http://www.w3.org/1999/11/11-WWWProposal/

Patel-Scheinder P., Fensel, D. - Layering the Semantic Web: Problems and Directions (2001)
http://www.cs.vu.nl/~dieter/ftp/paper/layering.pdf

semanticweb.org - Markup Languages and Ontologies
http://www.semanticweb.org/knowmarkup.html

Swartz, A. - The Semantic Web In Breadth
http://logicerror.com/semanticWeb-long

The Semantic Web Community Portal
http://www.semanticweb.org/ Junho 2002

w3.org – Semantic Web
http://www.w3.org/2001/sw/