1
Introdução

O volume de informações que gerenciamos tem crescido a cada dia. Realizamos milhões de pesquisas na Web, postamos mensagens em blogs, enviamos mensagens de correio eletrônico, gerenciamos fotos, músicas e vídeos em nossas comunidades virtuais, colaboramos para a solução de problemas em fóruns de discussão, conversamos através de mensagens instantâneas; e a lista de atividades que desempenhamos na Web não termina por aí. É cada vez mais evidente que novas ferramentas precisam ser criadas para nos auxiliar no processo de criar e recuperar nossas informações pessoais. A Web Semântica (Berners-Lee et al., 2001), tenta prover uma infra-estrutura para que no futuro tenhamos aplicações mais autônomas e que facilitem a realização de nossas tarefas. Em seu cerne, ela fornece um framework para comunicação de dados entre aplicações, permitindo que máquinas e pessoas colaborem na solução de tarefas, o que na prática amplia nossa capacidade de gerenciar nossas informações pessoais.

Atualmente, uma tarefa que permeia todas as atividades que desenvolvemos na Web é recuperar informação. Tradicionalmente, na pesquisa por informação, utilizamos busca por palavra-chave, navegação ou browsing (Carmel et al., 1992). No entanto, nenhum destes mecanismos é suficiente por si só, pois o processamento do resultado da consulta é sempre demandado do usuário, o que quase sempre gera uma elevada carga mental ao ser humano. Até mesmo ferramentas que possuem uma capacidade de processamento adicional, permitindo ao usuário expressar a sua consulta utilizando meta-dados (Yee et al., 2004), linguagem natural (Lin et al., 2003), ou até mesmo especificando o seu contexto (Lawrence et al., 2000) demandam algum filtro ou pós-processamento dos resultados por parte dos usuários. É evidente que recuperar informação na Web não é uma atividade isolada, mas sim um processo que pode incorporar busca, navegação e *browsing*, além de envolver o conhecimento do usuário sobre o domínio pesquisado.

O objetivo da Web Semântica é dar semântica aos dados, o que implica em estruturá-los (Berners-Lee et al). De fato, parte dos dados estarão expressos em RDF (Resource Description Framework)[1] e grande parte de acordo com alguma ontologia[2]. Esse método geral de modelar informação cria um cenário em que os usuários evoluirão de uma Web com dados sem estrutura, para uma Web semi-estruturada. Um desafio, portanto, será permitir o usuário (homem ou máquina) tirar proveito da estrutura inerente do modelo RDF para ampliar as possibilidades de exploração da informação.

1.1. Ontologia na Web Semântica

Para a Web Semântica, ontologia é um modelo de dados que representa um conjunto de conceitos, e seus relacionamentos, dentro de um domínio de conhecimento (Zhang, J., 2007). Em suma, ontologias são vocabulários abertos que nos permitem descrever objetos do nosso universo de conhecimento e definir relação entre eles. Na concepção de Breitman (2005, p. 7): "Ontologias são especificações formais e explícitas de conceitualizações compartilhadas. Ontologias são modelos conceituais que capturam e explicitam o vocabulário utilizado nas aplicações semânticas. Servem como base para garantir uma comunicação livre de ambigüidades. Ontologia será a língua franca da Web Semântica."

Ontologias podem ser descritas fazendo uso de linguagens tais como: RDFS (Resource Description Framework Schema)[3] e OWL (Web Ontology Language)[4]. Tais linguagens possuem elementos básicos que podem ser utilizados na construção de ontologias sobre um domínio arbitrário. Por exemplo, o elemento *rdfs:Class* que nos permite declarar recursos como uma classe de recursos.

[1] http://www.w3.org/RDF/

[2] Na Web Semântica, ontologia é um modelo de dados que representa um conjunto de conceitos, e seus relacionamentos, dentro de um domínio de conhecimento (Zhang, J., 2007).

[3] http://www.w3.org/TR/rdf-schema/

[4] http://www.w3.org/TR/owl-features/

1.2. RDF (Resource Description Framework)

RDF é um modelo de dados utilizado para descrever dados na Web Semântica. Basicamente, o RDF estrutura a informação em triplas na forma: Sujeito-Predicado-Objeto. Cada sujeito denota um recurso na Web e os predicados denotam a relação entre dois recursos (chamada *Object Property*), ou entre um sujeito e um valor (chamada *Datatype Property*).

Tudo, seja um sujeito, predicado ou objeto, é identificado por uma URI[5] (Uniform Resource Identifier). Uma URI é uma seqüência de caracteres utilizada para identificar um recurso na Internet, tal como: <http://www.w3.org/> ou <http://purl.org/dc/elements/1.1/title>.

Descrever dados RDF utilizando URIs pode ser algo pouco prático e ilegível, devido aos tamanhos das URIs. Por conveniência, podemos usar abreviações (QName[6] – *XML qualified name*) para as URIs quando criarmos arquivos RDF. Uma abreviação é formada por um prefixo e o nome local (*localname*[7]) da URI. Por exemplo, rdfs:class pode ser uma abreviação para a URI: http://www.w3.org/2000/01/rdf-schema#class. O prefixo é sempre associado a um espaço de nome (*namespace*[8]). No nosso exemplo, rdfs estaria associado ao espaço de nome: http://www.w3.org/2000/01/rdf-schema#. O nome local da URI, no nosso exemplo, seria o *string* "class". Alguns prefixos comumente utilizados na literatura são:

- prefixo **rdf:**, *namespace* URI: http://www.w3.org/1999/02/22-rdf-syntax-ns#
- prefixo **rdfs:**, *namespace* URI: http://www.w3.org/2000/01/rdf-schema#
- prefix **dc:**, *namespace* URI: http://purl.org/dc/elements/1.1/
- prefixo **owl:**, *namespace* URI: http://www.w3.org/2002/07/owl#
- prefixo **ex:**, *namespace* URI: http://www.example.org/ (ou http://www.example.com/)

[5] http://labs.apache.org/webarch/uri/rfc/rfc3986.html

[6] http://www.w3.org/TR/REC-xml-names/#ns-qualnames

[7] http://www.w3.org/TR/REC-xml-names/#ns-qualnames

[8] http://www.w3.org/TR/REC-xml-names/#ns-qualnames

- prefixo **xsd:**, *namespace* URI: http://www.w3.org/2001/XMLSchema#

Existem algumas notações para descrevermos dados RDF: Notation3[9] ou N3, N-Triples[10] ou NT, Trix[11], Turtle[12], TriG[13] e RDF/XML[14]. Adotaremos no decorrer dessa dissertação o uso da notação N3 pois é uma notação de fácil leitura em comparação as demais. Abaixo seguem alguns exemplos de uso dessa notação.

Como vimos, em RDF, a informação é especificada através de uma coleção de declarações, cada uma com um sujeito, predicado e objeto - e nada mais. Em N3, você pode escrever uma tripla RDF da seguinte forma:

```
<#joao> <#conhece> <#maria> .
```

Quadro 1 – Exemplo da notação N3.

Para descrevermos diversas sentenças sobre um mesmo sujeito fazemos o seguinte:

```
<#joao> <#conhece> <#maira> .
<#joao> <#idade> 24 .
```

Quadro 2 – Exemplo de um recurso com duas propriedades, em N3.

Existem dois atalhos para quando você tem várias declarações sobre um mesmo sujeito: um ponto e vírgula (";") introduz outra propriedade ao mesmo sujeito, e uma vírgula introduz um outro objeto com o mesmo sujeito e predicado.

[9] http://www.w3.org/DesignIssues/Notation3.html

[10] http://www.w3.org/2001/sw/RDFCore/ntriples/

[11] http://sw.nokia.com/trix/TriX.html

[12] http://www.dajobe.org/2004/01/turtle/

[13] http://www4.wiwiss.fu-berlin.de/bizer/TriG/

[14] http://www.w3.org/TR/rdf-syntax-grammar/

```
<#joao> <#pai>  <#ana>, <#luiz>, <#manuel> ;
    <#idade>    24 ;
    <#olhos> "castanhos" .
```

Quadro 3 – Exemplo do uso de , e ; da notação N3.

No exemplo a seguir utilizaremos a diretiva *@PREFIX* para definirmos prefixos. O exemplo a seguir mostra como a informação de que *Brasília é a capital de Brasil* pode ser representada usando esta estratégia:

```
@prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix ex: <http://www.exemplo.com> .
ex:Pais rdfs:type rdfs:Class .
ex:Cidade rdfs:type rdfs:Class .
ex:Brasilia rdfs:type ex:Cidade .
ex:Brasil rdfs:type ex:Pais .
ex:Brasilia ex:capital ex:Brasil .
```

Quadro 4 - Exemplo de RDF na sintaxe N3

Existem outras formas de representação dos dados RDF na notação N3, no entanto não utilizaremos nenhuma abordagem diferente das exemplificadas até aqui.

Assim como o modelo relacional é uma representação de dados em tabelas, o modelo RDF é uma representação dos dados na forma de grafo. A mesma informação do exemplo anterior poderia ser representada em um grafo como a seguir:

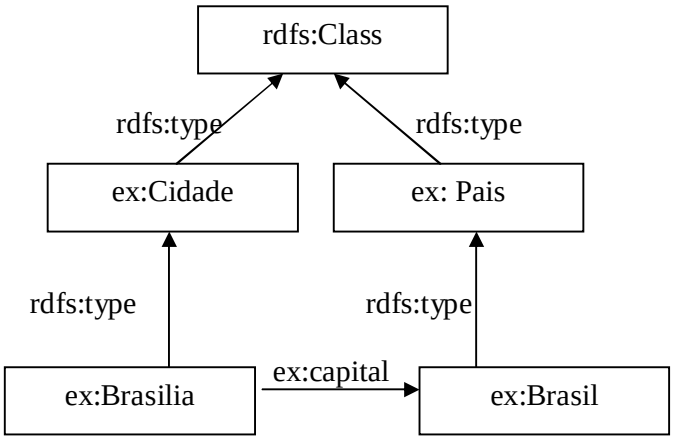

Figura 1 - Representação em grafo do RDF.

1.3. Explorando a Web Semântica

Uma área de pesquisa denominada busca exploratória (Marchionini G, 2006) tem tentado elaborar soluções que suportem exploração da informação. Busca exploratória é aplicável em situações em que a tarefa do usuário e o ambiente de busca possuem elementos complexos e requerem uma constante interpretação do usuário durante o processo de exploração. Por exemplo, como apoiar a tarefa de busca do usuário quando ele não é familiarizado com o domínio que pesquisa ou não tem conhecimento suficiente do domínio para formular uma consulta; como apoiar a navegação em espaços de informações complexos, ou quando a navegação, busca e *browsing* não são suficientes. Exemplificando o último caso, tente pesquisar na Web pela lista de telefones celulares e máquinas fotográficas que possuam o mesmo cartão de memória. Os mecanismos de busca atuais não são capazes de processar essa consulta em linguagem natural; e a natureza da navegação e *browsing* na Web nos forçaria a levar horas para obter o resultado esperado.

Marchionini G. (2006) fez uma distinção entre busca exploratória, busca simples e busca para recuperação. Segundo ele busca exploratória é baseada não somente em uma simples busca, mas também em investigação e aprendizado. Ele argumenta que busca investigativa e busca para aprendizagem requerem maior iteração humana no processo de exploração da informação do que uma simples

busca, pois são processos que suportam tarefas que demandam a capacidade cognitiva e interpretativa do usuário.

Esses tipos de tarefas são comumente encontradas no processo de exploração de informação de bases RDF, tais como DBPedia[15], onde os usuários necessitam identificar classes e propriedades das instâncias e do esquema, no intuito de compreenderem conceitos, adquirirem conhecimento e aprenderem sobre o domínio. Por exemplo, descreveremos a seguir um cenário que envolve tais características. Suponha que um usuário esteja pesquisando, no domínio do DBPedia, por informações sobre músicos italianos do século XVIII. Inicialmente o usuário tentaria obter todos os músicos italianos do século XVIII. No entanto, para formular tal consulta, ele necessitaria compreender a relação entre as instâncias que procura e as possíveis classes e propriedades existentes neste domínio, tais como Músico, Pessoa, Era, Nacionalidade, etc. Num segundo momento, o usuário poderia descobrir outras informações relevantes sobre os músicos, por exemplo, a lista de instrumentos musicais tocados por tais músicos. Note que neste cenário, o usuário necessitaria absorver informações sobre o domínio, tanto para ser capaz de formular as perguntas certas quanto para enriquecer seu conhecimento sobre o universo.

Considerando que na Web Semântica os dados são semi-estruturados e expressos em RDF, nosso objetivo nesta dissertação é propor um modelo que suporte a busca exploratória em uma Web de dados RDF. Atualmente, a Web de dados já se tornou uma realidade. Por exemplo, projetos tais como LOD (*Linking Open Data Project*)[16] incluem uma variedade de conjuntos de dados publicados em RDF: DBpedia, Geonames, US Census, EuroStat, MusicBrainz, BBC Programmes, Flickr, DBLP, PubMed, UniProt, FOAF, SIOC, OpenCyc, UMBEL e Yago. A disponibilidade desses e outros conjuntos de dados levantam diversas oportunidades e desafios, e um deles é a exploração dessa extensa base de dados que está sendo criada.

Nosso modelo de exploração define um conjunto de operações e uma interface visual. Tal interface permite a um usuário, com mínimo conhecimento

[15] Casos de uso do DBpedia - http://wiki.dbpedia.org/UseCases?v=sxq

[16] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

do modelo RDF, explorar uma base RDF, sem conhecimento prévio do domínio.

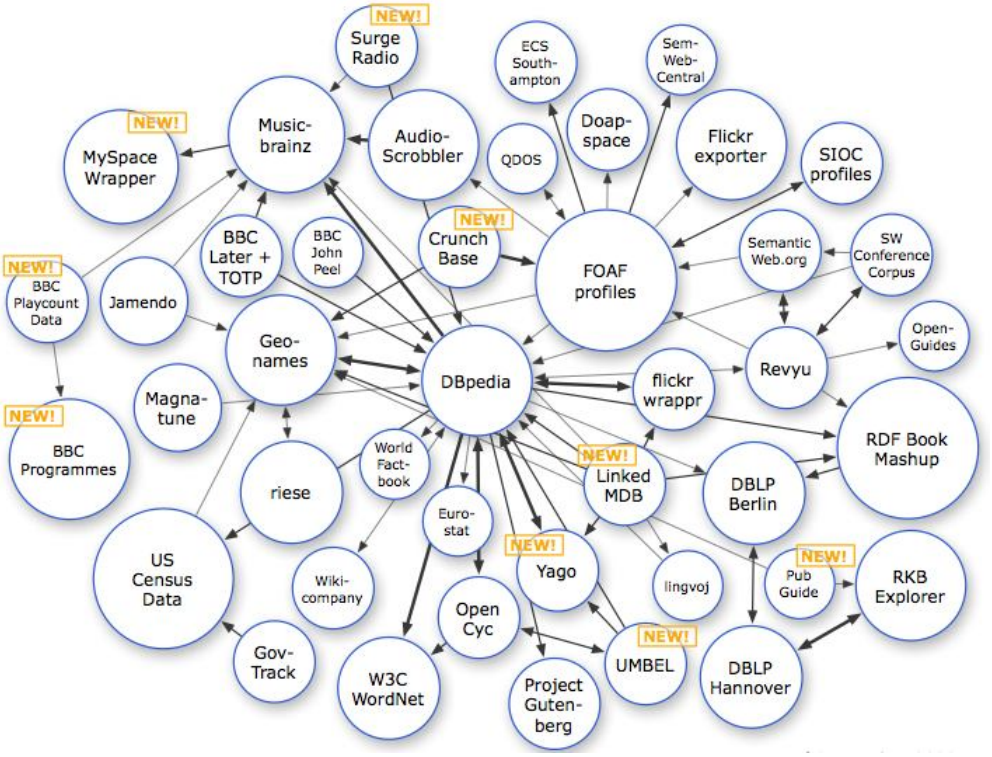

Figura 2 – Nuvem exemplificando a interconexão das bases RDF do projeto LOD. (nuvem de setembro de 2008). Fonte: http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

1.4. Organização da dissertação

Os capítulos a seguir estão organizados da seguinte forma. No capítulo 2 faremos uma revisão dos principais modelos de exploração existentes para Web semântica, e apresentaremos nosso modelo para dados RDF. No capítulo 3 abordaremos uma solução de navegação facetada para dados RDF. No capítulo 4 descreveremos a interface do usuário e apresentaremos um exemplo de uso da ferramenta que desenvolvemos. No capítulo 5 descrevemos detalhes da arquitetura de implementação da ferramenta de exploração proposta. No capítulo 6 apresentaremos os resultados de um experimento feito sobre essa ferramenta. E por fim, no capítulo 7 faremos uma conclusão.