Criação de um Repositório de Dados Ligados para Filtragem de *Hoax*

Adriano Rodrigues Delvoux Mattos, Jairo Francisco de Souza

Departamento de Ciência da Computação – Universidade Federal de Juiz de Fora (UFJF)

adrianodelvoux@gmail.com.br, jairo.souza@ufjf.edu.br

Abstract. *This project shows the use of linked data to provide a way to represent and consume information. This technology aims to create a knowledge base from multiple interlinked domains allowing to perform complex queries. By using linked data it is possible to create intelligent systems for consuming and analyzing semantic information. A tool was created to mark as spam messages about unmissing people on Facebook, using a database that implements the linked data principles.*

Resumo. *Este projeto exemplifica o uso de dados ligados para fornecer um mecanismo de representação e consumo de informações. Esta tecnologia visa a criação de centros de dados para vários domínios que podem interagir através de ligações entre diferentes entidades na web, tornando possível a realização de consultas detalhadas. Utilizando esta abordagem torna-se possível a criação de aplicações mais inteligentes que podem consumir e analisar estes dados. Para este projeto foi criada uma ferramenta que auxilia na identificação de mensagens falsas de pessoas desaparecidas no Facebook, utilizando uma base de dados que segue os princípios de dados ligados.*

1. Introdução

A chegada da *web* 2.0 trouxe mais dinâmica para a disponibilização de conteúdo, possibilitando que qualquer usuário publique conteúdo maneira simples, sem a necessidade de qualquer conhecimento avançado em informática. As ferramentas para criação de blogs estão se tornando cada vez mais práticas e acessíveis ao usuário leigo, de forma que as pessoas possam se preocupar mais com a qualidade da informação.

As redes sociais também se destacam entre os meios de geração de conteúdo, sendo um ambiente onde as pessoas postam informações variadas e compartilham com amigos, atingindo milhares de usuários. Junto às redes sociais o número de dispositivos conectados à internet cresceu consideravelmente. Hoje, existem *notebooks*, *smartphones*, *tablets* e até celulares mais simples com acesso a internet, prontos para que o usuário possa interagir nas redes sociais. Com todas essas tecnologias em mãos pode-se dizer que este perfil de usuário se torna o principal criador de conteúdo na *web*.

O crescimento de informações e o fato de qualquer pessoa ser capaz de postar dados sem a certeza de sua validade é um fator preocupante. Atualmente, circulam entre as redes sociais mensagens conhecidas como *hoax*, que procuram sensibilizar usuários a compartilhar informações falsas, formando correntes entre milhares de pessoas (Teixeira, 2007). Dentre as *hoax* existentes, as mensagens de pessoas desaparecidas são

cada vez mais constantes. O uso das redes sociais para encontrar pessoas é uma estratégia válida e que pode realmente ajudar. Porém, a falta de um mecanismo para eliminar estas mensagens após encontrar o indivíduo, proporciona a propagação de conteúdo falso nas redes sociais, sendo um incômodo para os usuários e principalmente para a família que passa a ser vítima de informações falsas.

Dentro deste cenário, este projeto possui como objetivo geral contribuir com a criação de um banco de dados com informações de pessoas desaparecidas, capaz de ajudar a identificar possíveis *hoaxes*, evitando que usuários sejam enganados.

Como objetivo específico o projeto propõe a implementação de uma abordagem para identificação automática de conteúdo sobre pessoas desaparecidas em um ambiente muito utilizado atualmente, as redes sociais. Automatizando este processo, uma aplicação criada poderá evitar que usuários destes serviços repassem mensagens falsas para sua rede de amigos.

2. Dados de pessoas desaparecidas

No Brasil as Organizações Não Governamentais são as principais atuantes na busca por pessoas desaparecidas junto às famílias. As ONGs utilizam amplamente a *Web* como um meio para a divulgação de casos de desaparecimentos por atingir um grande número de pessoas. Para estes casos, as informações estão limitadas a documentos HTML simples e de fácil visualização, mas que dificultam o processamento por máquina. Uma vez que os dados não se encontram estruturados, torna-se difícil para qualquer aplicação extrair conteúdos pertinentes das páginas. Outro aspecto importante é que os sites que tratam destes assuntos agem de forma independente, cada um com sua própria base, sendo possível a ocorrência de informações duplicadas.

Como as ONGs são entidades filantrópicas, em muitos dos casos faltam recursos para manter uma equipe capaz de trabalhar com a manutenção dos dados. As informações na *Web* nem sempre estão atualizadas. Uma dificuldade do projeto está no acesso e reunião dos dados necessários. Durante o trabalho diversas buscas por informações de pessoas desaparecidas foram feitas e poucos foram os *sites* com dados atualizados disponíveis.

3. Dados Ligados

De acordo com Berners-Lee (2006) a *Web* Semântica não se resume somente em colocar dados na *web*. Sua proposta é realizar ligações entre os dados de forma que pessoas e máquinas possam reutilizá-los. Com os dados ligados entre si é possível, a partir de alguma informação, atingir outros dados relacionados.

Bizer & Heath (2009) definem Dados Ligados simplesmente como sendo uma forma de utilizar a *web* para criar ligações entre os dados de acordo com seus tipos. Como estes dados estão publicados na *web* podem-se encontrar fontes de informações em bancos de dados externos, em diferentes posições geográficas e legíveis por máquinas, uma vez que o significado dos dados é explícito.

No caso da *web* de documentos as unidades primárias são documentos HTML com links para outros documentos. Para compor a *Web* de Dados Ligados, os recursos

são representados através de um formato padrão, o RDF (W3C-RDF, 2004), que permite interligar entidades de diferentes domínios (Bizer & Heath, 2009).

Berners-Lee (2006) definiu quatro princípios que regem a publicação de dados utilizando a tecnologia de dados ligados na *web*:

1. Utilize uma URI para identificar qualquer recurso;

2. Sempre use URIs HTTP para que seja possível encontrar estes nomes na *web*;

3. Forneça os dados utilizando um formato padrão, RDF e SPARQL (W3C-SPARQL, 2008);

4. Crie ligações para outros recursos na *web* de forma que seja possível encontrar mais informações.

Estes princípios fornecem um mecanismo para publicação e conexão entre dados usando a infra-estrutura da *web* (Heath as all, 2009)**.**

Atualmente, existem vários projetos como o Wikipedia, Wikibooks, Geonames, que disponibilizam conteúdos livres na *web*. O projeto *Linking Open Data*, criado pela W3C SWEO, surgiu para incentivar a publicação de *datasets* em RDF, atribuindo ligações entre dados de diferentes fontes (W3C-SWEO, 2012).

Inicialmente o projeto contou com o apoio de pesquisadores, desenvolvedores e pequenas empresas. Com o tempo, ganhou proporções maiores e atingiu grandes organizações, principalmente devido seu caráter público, onde qualquer pessoa pode contribuir fornecendo um conjunto de dados seguindo os princípios de dados ligados (Bizer & Heath, 2009). A figura 1 mostra uma representação dos *datasets* criados e as respectivas ligações entre eles:

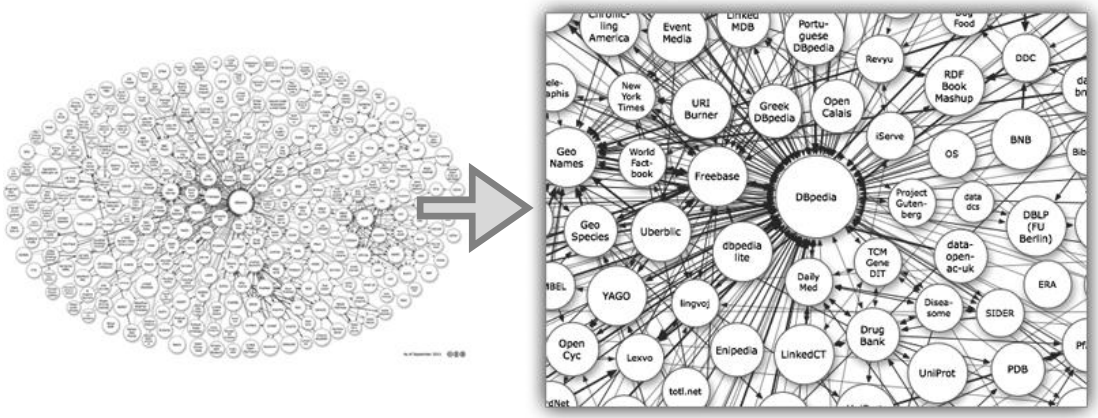

Figura 1. Diagrama do projeto Linking Open Data.

O grafo representado na figura 1 apresenta uma centena de *datasets*, porém somente dois contém informações em português. O DBPedia[1] surgiu como um projeto cujo objetivo é mapear os dados da Wikipédia e oferecê-los no formato de dados ligados. Este trabalho contou com a ajuda de vários países inclusive do Brasil, em que o Departamento de Ciência da Computação da Universidade Federal de Juiz de Fora

[1] http://pt.dbpedia.org/

contribuiu para o mapeamento dos dados em português. Outro projeto de destaque é o trabalho que surgiu também no Departamento de Ciência da Computação da UFJF que criou um novo *dataset* para o projeto *Linking Open Data* com informações de políticos brasileiros, o Ligado nos Políticos[2].

4. Criação de um *dataset* de pessoas desaparecidas

Para formar um *dataset* de pessoas desaparecidas foi necessário inicialmente recolher dados na *web*. Como não existem informações em formatos abertos utilizaram-se os dados disponíveis em sites de ONGs. O método utilizado escolhido para obter os dados é chamado de raspagem de dados. Através deste mecanismo, um *script* conhecido como *web crawler* acessa o HTML das páginas e retira o conteúdo de acordo com as especificações do desenvolvedor.

Os dados obtidos foram armazenados em um banco relacional e mapeados para um formato aberto e disponível para consultas. Objetivando facilitar e otimizar a formação do *dataset* utilizou-se uma tecnologia chamada Virtuoso, um software criado para atuar como um servidor universal que comporte as mais variadas tecnologias, incluindo servidor *web*, servidor de arquivos, banco de dados e armazenamento de XML nativo, disponíveis em uma única ferramenta. É possível trabalhar com o gerenciamento de bancos de dados relacionais e de dados no formato RDF e XML, como mencionado anteriormente possui um servidor *web* além de um servidor de Dados Ligados e de aplicações *web*. Outra característica importante é a possibilidade de trabalhar com Serviços *Web*, SOAP ou REST (OpenLink-Documentation, 2008).

Criou-se também um site para exibir os dados no formato HTML e RDF, http://desaparecidos.ice.ufjf.br. Os dados das pessoas desaparecidas ficam acessíveis através de um identificador, gerado de acordo com os princípios de dados ligados, uma URI HTTP. Toda URI deve ser dereferenciável, ou seja, o cliente pode utilizar um protocolo HTTP e obter informações relacionadas ao recurso identificado pela URI. A informação retornada depende do cliente que fez a requisição, para seres humanos uma representação HTML seria mais apropriada, enquanto para aplicações que consomem dados, o uso do padrão RDF é indicado (Heath & Bizer, 2011).

Para a representação RDF foram reaproveitadas algumas ontologias como *foaf*[3] para descrição de pessoas, *geonames*[4] para dados de localização, e *dbpprop*[5] para descrever características físicas. Propriedades importantes para descrever pessoas desaparecidas como data de localização e a situação do indivíduo, foram criadas originando uma nova ontologia nomeada como *des*.

Um dos princípios dos dados ligados requer que as informações estejam conectadas entre si. No intuito de satisfazer esta condição foi necessário estabelecer ligações entre os dados de desaparecidos com outras bases existentes. Para este caso utilizaram-se dados de localização e a fonte em que a informação foi retirada.

---

[2] http://ligadonospoliticos.com.br

[3] http://www.foaf-project.org/

[4] http://www.geonames.org/

[5] http://dbpedia.org/

A figura 2 ilustra de forma simplificada os processos envolvidos desde a coleta de informações até a geração dos dados em formato aberto.

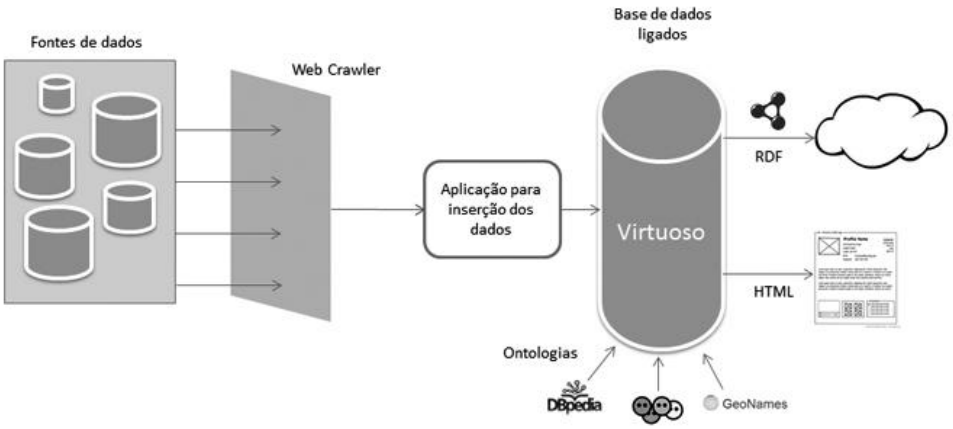

Figura 2. Esquema do processo de criação da base de dados ligados

5. Aplicação para redes sociais

Com um repositório de informações abertas a disposição torna-se mais fácil a criação de aplicações. O problema identificado é a propagação de *hoax* nas redes sociais e a falta de praticidade em descobrir a validade destas mensagens. Para este cenário uma aplicação poderia contribuir reduzindo os compartilhamentos desnecessários.

Sugeriu-se para este projeto o desenvolvimento de uma aplicação *web mobile* capaz de acessar dados do mural de usuários no Facebook e, através de uma interface simples, retornar a situação de um indivíduo e informações mais detalhadas.

O Facebook disponibiliza uma API para que os desenvolvedores criem aplicações e obtenham dados de usuários da rede social. Ao acessar as atualizações do mural é possível utilizar técnicas para identificar possíveis mensagens de pessoas desaparecidas. Neste caso, optou-se por utilizar um *array* de *tokens*. A escolha dos melhores *tokens* foi realizada com o apoio de uma técnica chamada s*temming*, que propõe um conjunto de etapas para chegar a um termo comum, eliminando os fatores que geram variações. Um exemplo é a palavra “desaparecido”, que pode aparecer em diversas postagens sofrendo variações. Aplicando o algoritmo *stemming,* o *token* para busca passa a ser somente “desaparec”.

Ao identificar uma postagem como sendo de um desaparecido, a aplicação verifica a ocorrência desta pessoa na base. O Virtuoso oferece uma interface REST, acessível através do protocolo HTTP, onde é possível enviar uma consulta SPARQL e ter como resultado um XML com o resultado.

Para reduzir o conteúdo a ser processado utilizou-se uma técnica conhecida como s*top words*. Comum entre os mecanismos de buscas esta técnica propõe a remoção de palavras não relevantes em uma pesquisa, com o objetivo de simplificar a consulta e reduzir o tempo de resposta (Rouse, 2005). As *stop words* são artigos, preposições, pronomes, entre outras. A lista de *stop words* utilizada para o projeto encontra-se acessível em http://snowball.tartarus.org/algorithms/portuguese/stop.txt.

A sequência de termos resultantes, agora simplificada, é analisada para retirar possíveis nomes próprios. Como os nomes são encontrados entre os textos iniciando, na maioria das vezes, com a letra maiúscula, criou-se um algoritmo que fizesse esta verificação e guardasse tais nomes. Por fim, os nomes encontrados são utilizados na consulta realizada na base criada.

Caso a pessoa seja encontrada, o sistema exibe o status do indivíduo para o usuário. Se ela não existir, o usuário pode acessar o site do projeto e contribuir com informações que ajudem a encontrá-la, permitindo assim ampliar o banco de dados sem que as informações fiquem atreladas somente aos dados de ONGs. A figura 3 mostra a aplicação em execução:

Figura 3. Aplicação *Web mobile*

6. Conclusão

Este projeto possui como resultado a inclusão de um novo *dataset* com informações de pessoas desaparecidas seguindo os princípios de dados ligados, a fim de contribuir com o projeto *Linking Open Data*. Esta nuvem de dados permite que desenvolvedores coletem e manipulem dados semânticos para serem utilizados em aplicações diversas. O formato padronizado e livre associado ao crescente número de domínios que estão sendo incluídos favorece o surgimento de aplicações cada vez mais inteligentes e capazes de trabalhar com informações interligadas na *web*.

Além da criação desta base, a proposta também incluiu o desenvolvimento de uma aplicação capaz de exemplificar o uso destes dados em um caso prático, a identificação de pessoas desaparecidas.

Além de representar um caráter social, ajudando famílias que possuem parentes desaparecidos, a aplicação também ajudará a reduzir o número de compartilhamento de *hoax* nas redes sociais.

Referências

Teixeira, R. C. (2007) **Boatos (Hoax)**. Disponível em: http://informatica.terra.com.br/virusecia/spam/interna/0,,OI198466-EI2403,00.html [Online; acessado em 01-10-2012]

Berners-Lee (2006) T. **Linked data**. Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html> [Online; acessado em 12-agosto-2012].

Heath, T.; Hepp, M.; Bizer, C. (2009) **Linked data - the story so far**. Disponível em: <http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf> [Online; acessado em 31-junho-2012].

W3C-SWEO (2012) **W3C SWEO Community Project**. Disponível em: <http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

LOD (2012) **Linked Data - Connect Distributed Data across the *Web***. Disponível em: <http://linkeddata.org> [Online; acessado em 12-agosto-2012]

W3C-RDF (2004) **Resource Description Framework (RDF)**. Disponível em: <http://www.w3.org/RDF/> [Online; acessado em 15-agosto-2012]

W3C-SPARQL (2008) **SPARQL Query Language for RDF**. Disponível em: <http://www.w3.org/TR/rdf-sparql-query/> [Online; acessado em 15-agosto-2012]

OpenLink-Documentation (2008) **OpenLink Virtuoso Universal Server: Documentation**, OpenLink Software Documentation Team. Disponível em: http://docs.openlinksw.com/pdf/virtdocs.pdf [Online; acessado em 15-agosto-2012]

Heath, T.; Bizer, Y. (2011) **Linked Data: Evolving the Web into a Global Data Space (1st edition)**. Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool. Disponível em: http://linkeddatabook.com/editions/1.0/ [Online; acessado em 15-agosto-2012]

Orengo, V. M.; Huyck, C. (2001) **A Stemming Algorithm for the Portuguese Language**

Rouse, M. (2005) **Definition: Stop word**. Disponível em: < http://searchsoa.techtarget.com/definition/stop-word> [Online; acessado em 02-outubro-2012]