Criação de um Repositório em Dados Ligados para Filtragem de Emails Indesejados

Adriano Rodrigues Delvoux Mattos[1], **Jairo Franscisco de Souza**[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Juiz de Fora

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

Resumo. *Este trabalho aborda sobre a importância do uso de dados ligados para a criação de aplicações mais inteligentes e complexas através de um exemplo prático que será desenvolvido. Tecnicamente, o projeto consiste em extrair informações na Web através da raspagem de dados e padronizá-las no formato RDF, criando um novo dataset para contribuir com o projeto Linking Open Data. Este projeto visa disponibilizar informações abertas para o consumo de aplicações. Os dados coletados serão utilizados como suporte para uma aplicação criada, que atuará como um analisador de e-mails filtrando conteúdos falsos em um domínio específico, neste caso sobre crianças desaparecidas. Além dos objetivos técnicos este trabalho procura amenizar os problemas enfrentados por milhares de usuários de serviços de e-mails criando uma ferramenta que irá possibilitar a estes usuários verificar a autenticidade das mensagens recebidas. Devido à diversidade de domínios em que estes e-mails estão enquadrados será necessário limitar a análise aos e-mails relacionados ao domínio de crianças desaparecidas. Desta forma, o projeto também ajudará diversas famílias que passam por este problema.*

Palavras-chave: Dados Ligados. Filtro de e-mails. Crianças desaparecidas. Dados abertos. Aplicações Web.

1. Introdução

A Web, desde seu surgimento, passou por uma série de transformações que possibilitou a criação de um ambiente cada vez mais interativo para o usuário, seja este avançado ou leigo. Porém, nem sempre ela teve esta característica.

Como citado em [Sabattini 2008], Tim Berners Lee inicialmente definiu a Web através de três tecnologias: um identificador de recursos (documentos, dados) conhecido como URL; uma forma de apresentar o conteúdo das páginas na Web e expressá-las em links, o HTML; por fim, um protocolo para gerenciamento e transferência dos dados na Web, o HTTP.

Em sua versão 1.0, a Web possuía a estruturação do conteúdo sob o domínio de desenvolvedores, visto que exigia maior conhecimento do usuário. A necessidade de criar um espaço com maior interatividade foi o próximo passo.

O surgimento de novas tecnologias e ferramentas facilitou o contato com usuários leigos, possibilitando estes contribuírem com conteúdo para a nova Web. A geração de conteúdo dinâmico e o compartilhamento de informações marcou a transição para a Web 2.0. Ferramentas como Blogs, Redes Socias, sites de compartilhamento de arquivos, vídeos, músicas passaram a ser utilizados amplamente.

Apesar do avanço da Web, um problema que existe até os dias de hoje é a falta de padronização dos dados. Até então, a Web é estruturada como uma rede de recursos ligados entre si, com mínimo de informações deste relacionamento. O futuro da Web propõe tratar este problema fornecendo mecanismos de forma que os dados possam ser manipulados abertamente por diferentes aplicações. Atualmente, vigora a Web 2.5, uma transição para a versão 3.0.

A Web Semântica é uma tecnologia que surgiu como suporte à nova Web, de forma que as informações possam ser dispostas tanto para humanos quanto para as máquinas, no caso das máquinas, teria grande importância para a criação de aplicações mais sofisticadas e inteligentes.

Como consequência de um movimento mundial para a padronização dos dados surgiu o conceito de Dados Ligados cujo objetivo é agregar informações relevantes aos recursos dispostos na Web e interligá-los, reestruturando a atual Web de documentos heterogêneos.

Para que a máquina consiga identificar uma entidade é preciso que esta esteja bem definida de acordo com um padrão. Desta forma, a Web Semântica faz uso de ontologias. Ontologia é um modelo utilizado para descrever uma série de conceitos relativos a um domínio. Atualmente existem milhares de ontologias disponíveis, para descrição de pessoas (Foaf), dados geográficos (GeoNames), dentre outras.

2. Motivação

Atualmente, grande parte dos usuários de serviços de e-mail recebem inúmeras mensagens denominadas correntes de e-mail. Estas correntes contêm informações diversas que são repassadas para conhecidos podendo atingir até milhões de pessoas. Geralmente estão relacionadas a fotos, mensagens positivas, golpes, alertas de vírus, crianças desaparecidas, dentre outros. É muito comum recerbermos correntes com histórias falsas conhecidas como *hoax*, que apresentam supostas campanhas filantrópicas, informações de falsos vírus e mensagens de apelo dramático.

Um grande problema é identificar a veracidade destes e-mails. No caso de crianças desaparecidas, existem muitos casos de spams que atrapalham as verdadeiras mensagens que poderiam ajudar uma família. Também existe a possibilidade destes e-mails circularem entre milhares de pessoas sendo que a criança já tenha sido encontrada. Mesmo existindo muitos sites que possuem inúmeras informações destas crianças os usuários não acessam estas fontes para verificar, e acabam ignorando ou repassando as mensagens.

A tecnologia de Dados Ligados está sendo utilizada amplamente para padronização e publicação de dados na Web. Este conjunto de informações de crianças espalhadas na internet poderia ser aproveitado de uma forma melhor se estivessem padronizados utilizando dados ligados, evitando assim que se propaguem informações desnecessárias que atrapalham o verdadeiro objetivo destes e-mails, que é ajudar.

3. Justificativa

O problema de se identificar automaticamente a veracidade de e-mails é a falta de uma base de dados padronizada que possa ser utilizada pelo próprio computador para buscar informações relevantes. Como os dados estão em diversos sites, é necessário realizar um

trabalho de Mineração de Dados para obter informações. Logo após, podemos padronizar e disponibilizar utilizando Dados Ligados. Tais informações serão distribuídas de forma legível ao usuário ou em um formato que poderia ser processado pelo computador.

Com uma base pronta, seria interessante criar uma aplicação que automatize o processo de busca em serviços de e-mail. Os navegadores mais conhecidos atualmente possuem suporte a plugins, que são programas que complementam os navegadores oferecendo alguma funcionalidade extra. Uma opção seria utilizar este recurso para criar um plugin e integrar a algum serviço de e-mail de forma que o usuário possa clicar em um botão e automaticamente realizar uma busca de termos presentes no e-mail, verificando a autenticidade da mensagem.

Oferecendo esta facilidade ao usuário torna-se mais fácil identificar o que realmente deve ser repassado na Web, evitando que os spams circulem infinitamente e as famílias que realmente precisam de ajuda possam receber melhor atenção.

4. Objetivos

4.1. Objetivos Gerais

Este projeto possui como objetivo contribuir com um novo dataset referente a um domínio específico e a partir destes dados evitar que usuários sejam enganados por spams perigosos.

4.2. Objetivos Específicos

Este projeto propõe a implementação de uma abordagem para identificação automática de spams. Automatizando este processo a aplicação criada poderá evitar que usuários de serviços de e-mail sejam vítimas de golpes. Outro objetivo importante, que está relacionada no contexto de Dados Ligados, é a disponibilização de informações de crianças desaparecidas em formato aberto e padronizado, possibilitando ainda que outras aplicações possam consumir e manipular tais informações.

5. Fundamentação Teórica

De acordo com [Berners-Lee 2006] a Web Semântica não se preocupa somente em colocar dados na Web. Sua proposta é realizar ligações entre os dados de forma que máquinas e pessoas possam aproveitar melhor os dados espalhados na Web. Com Dados Ligados a interação com os dados torna-se mais inteligente, onde uma informação se relaciona com outra gerando uma rede de dados ligados entre si.

[Bizer et al. 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 podemos ter fontes de informações em bancos de dados externos, em diferentes posições geográficas, e legível por máquinas, uma vez que temos o significado dos dados explícito.

[Heath 2009] realiza um paralelo entre a atual Web de Documentos e a Web de Dados Ligados dizendo que o uso de dados Ligados encoraja a reutilização de informações, reduz a quantidade de informações redundantes e amplia o relacionamento entre os dados.

Com os dados organizados de acordo com as práticas vistas até então é inevitável o surgimento de aplicações que utilizem estas informações. Surge o conceito de Aplicações

com Dados Ligados que segundo [Hausenblas 2011], pode ser entendido de duas formas. Primeiramente, seria uma aplicação utilizando dados ligados em um domínio específico, seja na Engenharia, na Estatística, na Biologia, entre outros. Também podemos entender como sendo a construção de aplicações Web baseadas em dados ligados.

As Aplicações Web baseadas em Dados Ligados basicamente consomem e manipulam as informações dispostas na Web. [Hausenblas 2011] cita que atualmente a grande maioria das aplicações somente consomem os dados. Apesar da estrutura, em que os dados ligados são baseados, ser conhecida esta é uma área que ainda deve ser muito estudada para que estes dados passem a ser utilizados não só para leitura.

Um problema de se criar estas aplicações é o número mínimo de dados ligados disponibilizados na Web. Muitos são os trabalhos que coletam dados na Web para criar uma base com dados padronizados. [dos Santos and de Souza 2010] comenta sobre a dificuldade de uma aplicação fazer uso de dados ligados devido a falta de informações no formato RDF, padrão para descrever estes dados.

Podemos ver em [Bize et al. 2008] que existe um projeto que visa melhorar a diponibilização destes dados na Web denominado *Linking Open Data (LOD)*. Ele tem o objetivo de disponibilizar os dados em diversas bases de dados livremente formando uma nuvem de dados ligados, denominada LOD *Cloud*.

6. Metodologia

Para que o objetivo do trabalho seja alcançado será necessário seguir um conjunto de etapas descritas a seguir:

- Avaliação das ontologias que servirão como suporte para a representação dos dados no padrão RDF.
- Criação do modelo de dados a ser utilizado para o armazenamento das informações.
- Construção de um dataset para a disponibilização dos dados no formato RDF a partir de uma URI HTTP única.
- Construção de um plugin que atue junto a um serviço de e-mail e acesse o dataset criado para a análise do conteúdo dos e-mails do usuário.
- Criação de um procedimento para que o usuário possa avaliar o funcionamento da ferramenta a fim de estabelecer um controle de qualidade da aplicação e dos dados.

7. Estrutura do Trabalho

1. Introdução
    - 1.1 Motivação
    - 1.2 Justificativa
    - 1.3 Objetivos
        - 1.31 Objetivos Gerais
        - 1.32 Objetivos Específicos
2. Dados Ligados
    - 2.1 Conceitos básicos
        - 2.11 Web Semântica

        2.12 Ontologias
        2.13 Padrão RDF
    2.2 Definição
    2.3 Linking Open Data
3. Aplicações com dados ligados
    3.1 DBPedia
    3.2 Aplicações que Consomem Dados Ligados
4. Projeto: Criação de um Dataset de Crianças Desaparecidas e de uma Aplicação para Consumo de Dados
    4.1 Dataset de Crianças Desaparecidas
        4.11 Visão Geral
        4.12 Fonte de Dados
        4.13 Web Crawler
        4.14 Estrutura do Banco de Dados
        4.15 Modelo RDF para Disponibilização dos Dados
    4.2 Aplicação
        4.21 Visão Geral
        4.22 Realizando Busca sobre os Dados no formato RDF
        4.23 Criando um plugin para navegadores
        4.24 Analisando o conteúdo dos e-mails
        4.25 Controle de Qualidade da Ferramenta
5. Conclusão
    5.1 Considerações Finais
    5.2 Limitações
    5.3 Trabalhos Futuros

8. Cronograma

9. Resultados Esperados

Espera-se que este trabalho contribua com mais um dataset de dados sobre o domínio de crianças desaparecidas que possa ser utilizado abertamente por outras aplicações Web. O desenvolvimento detalhado de uma aplicação que utiliza dados ligados deverá incentivar o consumo destes dados contribuindo para o surgimento de novas aplicações inteligentes e de trabalhos futuros.

10. Aceite do Orientador

Local/Data:	Juiz de Fora, 06 de Abril de 2011.
Parecer do Orientador:	
Assinatura do Aluno	
Assinatura do Orientador	

CRONOGRAMA	Agosto				Setembro				Outubro				Novembro			
Semanas	1ª	2ª	3ª	4ª	1ª	2ª	3ª	4ª	1ª	2ª	3ª	4ª	1ª	2ª	3ª	4ª
Definir as ontologias que serão utilizadas	X															
Definir fontes de dados	X															
Criar estrutura do banco de dados		X														
Coletar informações			X	X												
Criação do modelo RDF					X	X										
Avaliação da base de dados							X									
Construção do plugin								X	X	X						
Realização de testes											X					
Procedimento de controle de qualidade												X				
Elaboração da Monografia	X	X			X	X			X	X	X	X	X			
Revisão do Orientador														X		
Alterações solicitadas														X	X	
Entrega para a Banca de Avaliação															X	
Defesa do TCC																X

Figura 1. Cronograma do projeto

Referências

Berners-Lee, T. (2006). Linked data. disponível em: http://www.w3.org/designissues/linkeddata.html.

Bize, C., Heath, T., Idehen, K., and Berners-Lee, T. (2008). Linked data on the web. *Workshop at the 17th International World Wide Web Conference Beijing, China, April 22, 2008*.

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data - the story so far.

dos Santos, A. M. and de Souza, J. F. (2010). Manipulação de dados abertos para construção de novas aplicações. *I Workshop de Trabalhos de Graduação e Pós-Graduação - DCC/UFJF*.

Hausenblas, M. (2011). Linked data application. *DERI - Digital Enterprise Research Institute*.

Heath, T. (2009). An introduction to linked data. *Platform Division. Talis Information Ltd. Austin, Texas*.

Sabattini, R. (2008). Internet e web: passado, presente e futuro. *Conip 2008*.