Widgets baseados em conhecimento advindo de dados referenciados e abertos na Web

Henrique Santos
Universidade de Fortaleza
Av. Washigton Soares 1321
NATI Bloco M sala 11
hensantos@gmail.com

Vasco Furtado
Universidade de Fortaleza
Av. Washigton Soares 1321
NATI Bloco M sala 11
vasco@unifor.br

Vládia Pinheiro
Universidade de Fortaleza
Av. Washigton Soares 1321
NATI Bloco M sala 11
vladiacelia@gmail.com

Caio Ferreira
Universidade de Fortaleza
Av. Washigton Soares 1321
NATI Bloco M sala 11
caioxferreira@gmail.com

José Eurico Vasconcelos
Universidade de Fortaleza
Av. Washigton Soares 1321
NATI Bloco M sala 11
Jose_eurico@yahoo.com

Guilherme Shiki
Universidade de Fortaleza
Av. Washigton Soares 1321
NATI Bloco M sala 11
gshiki@gmail.com

ABSTRACT

The use of widgets is a very popular manner to make a website customization. From widget's creation the content creator configures the website with functions that he/she consider adequate to the users. Typically, widgets relies on syndication (RSS) in which a website content is made available to other websites. Even though there is a huge popularity of this kind of widgets, they typically are constrained to what a data provider makes available. Linked Open Data (LOD) is an opportunity to cope with the today's constraint in the process of widget creation and execution. We propose a platform, called *SemWidgets* (from Semantic Widgets), for the creation and execution of piece of programs able to access and reason over LOD. With *SemWidgets*, we provide to the content producer of a website a way to describe the concept(s) that best represent the content to be explored. Widgets created from *SemWidgets* have the power to perform inferences and access external sources that constitute information that may be useful and appropriate to the context of the website.

Categories and Subject Descriptors

H.3.4 [**Information Systems**]: World Wide Web – *semantic web, knowledge-based systems.*

General Terms

Algorithms.

Keywords

Widgets, Knowledge-based Systems, Linked Open Data, RDF

Agradecimentos ao CNPQ pelo financiamento de bolsa de pesquisa contrato 306266/2008-3.

WebMedia'11: Proceedings of the 17th Brazilian Symposium on Multimedia and the Web. Short Papers.
October 3 -6, 2011, Florianópolis, SC, Brazil.
ISSN 2175-9650.
SBC - Brazilian Computer Society

1. INTRODUÇÃO

O uso de *widgets* é uma das maneiras mais populares de se realizar a customização de um sítio web. Um *widget* é um trecho fechado de código que pode ser inserido em um sítio web ou programa para realizar uma função específica. A partir da criação do *widget*, o criador de conteúdo configura o sitio web com funções que ele(a) considera adequada aos usuários.

Apesar da grande popularidade desse tipo de *widgets*, normalmente eles são restritos a apenas o que a fonte de dados disponibiliza. Por exemplo, um leitor RSS de notícias depende do conteúdo fornecido pelos sítios de notícias. Dados referenciados e abertos (*Linked Open Data* - LOD) são uma oportunidade para quebrar essa restrição no processo de criação e execução de um *widget*, pois uma consulta a uma base da LOD permite extrapolação à fonte inicial.

Este trabalho se encaixa de forma inovadora nesse contexto. Propomos uma plataforma, chamada *SemWidgets*, para a criação e execução de pedaços de programas capazes de acessar e raciocinar em cima da LOD. A principal inovação do *SemWidgets* é que ele fornece ao produtor de conteúdo de um sítio web uma maneira de caracterizar semanticamente o que ele(a) pretende exibir no corpo do *widget*. Chamamos de "caracterização semântica" o ato de descrever os conceitos que melhor representam o conteúdo a ser explorado. *Widgets* criados a partir de *SemWidgets* possuem a capacidade de executar inferências, acessar fontes externas e assim trazer informação adicional e contextualizada aos usuários do sítio web.

No coração do processo de caracterização semântica está o uso de Processamento de Linguagem Natural (PLN) para suportar interativamente a elicitação e expressão dos conceitos a serem consultados na LOD. Esse processo utiliza uma base de conceitos de senso comum e um raciocinador. Utilizamos a base InferenceNet [6] com conceitos em inglês e português, gerados a partir da ConceptNet [5], mas construindo sua própria representação baseada nas chamadas teorias inferencialistas [2]. Em suma, essas teorias precisam que o conteúdo de um conceito seja expresso pelo conjunto de relações inferenciais (pré-condições e pós-condições) com outros conceitos, permitindo que

o significado de uma sentença seja inferido. O raciocinador SIA [6] explora os conceitos representados na base de senso comum. Para utilizarmos o poder da LOD, geramos uma versão da InferenceNet em RDF que nos permitiu conectá-la a WordNet [4], Yago [7], DBPedia [1], e outras.

Demonstraremos como um widget semântico criado por SemWidgets explora a base de conceitos da Wikipédia.

2. FONTES DE CONHECIMENTO

Nosso trabalho utiliza essencilamente duas fontes de conhecimento: dados abertos referenciados e uma base de senso comum.

2.1 Dados Referenciados Abertos

Dados referenciados abertos (*Linked Open Data* - LOD) se referem a uma série de práticas e padrões adotados para a publicação de informação na web para que máquinas sejam capazes de processá-las eficientemente. Mais ainda, LOD visa utilizar a web para ligar dados que não estavam referenciados anteriormente, ou ainda utilizar a web para diminuir a dificuldade de referenciar dados através da interferência de especialistas.

Tim Berners-Lee [3] definiu quatro máximas que são caracterizadas como as melhores práticas para publicação na LOD: (i) Uso de URIs para identificar coisas, (ii) Atribuição de endereços *http* a uma URI para permitir que ela seja facilmente recuperada na web; (iii) Uso de RDF como padrão para expressão informação; (iv) Uso de dados ligados através de URIs permitindo a busca por novos dados.

Muitas iniciativas de descrever conteúdo livre na web através de ontologias em RDF surgiram, como DBPedia e Freebase (http://www.freebase.com). Entretanto, foram as recentes iniciativas de grandes conglomerados da mídia, tais como *The New York Times* e *BBC* que impulsionaram a área.

2.2 Bases de senso comum

SemWidgets se baseia em um processo interativo de caracterização semântica que utiliza uma base conceitual bilíngue e LOD, tendo o Yago e a DBpedia como principais conexões. Um dos recursos da base conceitual é que expressa conhecimento semântico inferencialista e conhecimento de senso comum – InferenceNet [6]. A base InferenceNet expressa conteúdo semântico através de uma rede que conecta um conceito a vários outros através de diversas relações semânticas de senso comum e que são inferenciais por natureza (pré-condições e pós-condições no uso de conceitos). Formalmente, essa base é representada em um grafo direcionado de conceitos e relações $G_c(C,R_c)$. Cada relação inferencial $r_{cj} \in R_c$ é representada por uma tupla (*nomeRelação*, c_i, c_k, tipo), onde *nomeRelação* é o nome de uma relação semântica no InferenceNet (CapazDe, PropriedadeDe, EfeitoDe etc.), c_i e c_k são conceitos de linguagem natural, c_i, $c_k \in C$, e tipo = "Pre" or "Pos" (pré-condição ou pós-condição para uso do conceito c_i). Os conceitos da InferenceNet e seu conteúdo inferencial foram convertidos para RDF e ligados ao Yago pelo nome do conceito em inglês (propriedade "hasEnglishName") que, por sua vez, é ligado a DBpedia. Abaixo mostramos um trecho do RDF que descreve a ligação entre o conceito político ("http://inferencenet.org/rdf/conceito41738") no InferenceNet com o conceito Yago relacionado "http://www.mpii.de/yago/resource/wordnet_politician_109772277".

O conteúdo expresso na InferenceNet e suas conexões com a LOD permitem que sejam realizadas inferências diversas e ricas, visto que se consegue aqui uma associação entre uma base conceitual taxonômica com uma base de senso comum e inferencialista.

3. CARACTERIZAÇÃO SEMÂNTICA DE UM WIDGET

3.1 Visão Geral

Para criar um *widget* semântico capaz de consultar a LOD, o produtor de conteúdo precisa expressar um conceito ou uma entidade específica que o *widget* utilizará para procurar notícias, vídeos, fotos ou artigos relacionados. Exemplos desses conceitos são políticos, times de futebol, locais onde o lixo está acumulado, ou entidades específicas como pessoas (ex: Obama), instituições (ex: Stanford, CNN), lugares (ex: Rio de Janeiro) etc. Para que *SemWidgets* seja capaz de realizar inferências e recuperar informações relacionadas sobre um conceito ou uma determinada instância, seu conteúdo semântico deve ser definido. Em suma, o produtor de conteúdo, ao informar uma expressão linguística que descreve um conceito ou uma instância de um conceito, é guiado em um processo cujo objetivo é associar conceitos oriundos da base *InferenceNet* a essa expressão linguística. Com isso, *SemWidgets* começa a ter conhecimento sobre as relações inferenciais de senso comum oriundas da base *InferenceNet* que definem o valor semântico do conceito e é capaz de realizar inferências sobre a rede de dados ligados.

A Figura 1 mostra parte do processo de caracterização semântica de um *widget*. Nela vemos como *SemWidgets* exibe, em um grafo, as relações inferenciais semânticas de um possível conceito a ser associado à expressão linguística. No exemplo, o produtor de conteúdo informa a expressão – "Político" – e o sistema procura na InferenceNet por um conceito relacionado, exibindo suas principais relações semânticas com outros conceitos (ex: "político" tem a propriedade de "corrupto"), através de um grafo conceitual. O usuário então é questionado se o grafo se refere ao conceito encontrado. Se sim, o usuário pode adicionar novas relações e remover uma ou mais que não deseje para o contexto do *widget* a ser criado.

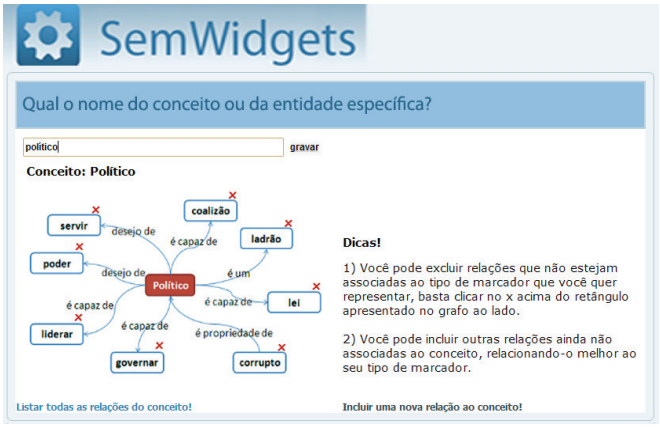

Figura 1. Grafo de conceitos e interação entre *SemWidgets* e o produtor de conteúdo do site.

Quando a expressão linguística se refere a uma instância de um conceito (ex: Pelé – instância de jogador de futebol, ou Barack Obama – instância de político), *SemWidgets* procura a DBpedia para descobrir as classes Yago que melhor representam a instância. O conceito mais específico é recuperado a partir da navegação no grafo completo de conceitos para uma instância específica. Primeiro, os conceitos diretamente ligados às instâncias são recuperados. *SemWidgets* calcula então o número de saltos de cada conceito do grafo ao conceito "entity" (a raiz da

ontologia Yago) e os classifica. O conceito que melhor representa a instância é aquele que possui o maior número de saltos e que está associado ao *InferenceNet*. A intuição por trás dessa estratégia de escolha de conceito se baseia no fato de que classes específicas possuem informação mais precisa (ou menos superficial) sobre a instância.

No fim desse processo interativo de caracterização semântica, conceitos da *InferenceNet*, expressados através de relacionamentos inferenciais de senso comum, são associados à expressão linguística. Ao fazer isso, permitimos que *SemWidgets* recupere conteúdo relacionado (notícias, artigos, vídeos, fotos, perfis de redes sociais, etc.), enquanto o site estiver sendo usado.

3.2 Heurísticas para busca de conteúdo na LOD

As inferências geradas podem ser de dois tipos:

1. Inferências baseadas em instâncias e nas relações inferenciais do conceito relacionado a essa instância, expressa na *InferenceNet*.
2. Inferências baseadas no conceito genérico e suas relações inferenciais expressas na *InfereceNet*.

Para a geração de ambos os tipos de inferência, *SemWidgets* define as seguintes heurísticas:

a) Escolha de relações inferenciais do InferenceNet (*nomeRelação*, c_1, c_2, Pre/Pos), onde c_1 é o conceito associado à expressão linguística e c_2 é uma frase verbal, frase nominal ou frase adjetiva, que contém pelo menos um nome ou um adjetivo ligado a um recurso do Yago e DBpedia. Por exemplo, a relação inferencial (*CapazDe*, "político", "propor lei", "Pre") é escolhida, pois a expressão linguística "Barack Obama" foi associada ao conceito "político" e o conceito c_2 é do tipo frase verbal, com o nome "lei" ligado via RDF ao recurso do Yago e DBpedia. Para o caso onde c_2 é o tipo frase adjetiva, o nome relacionado ao adjetivo é também procurado no Yago e DBpedia. Por exemplo, para o relacionamento (*PropriedadeDe*, "político", "corrupto", "Pre"), o adjetivo "corrupto" e o nome "corruption" são pesquisados no Yago e DBpedia.

b) Recuperação da URI na DBpedia que representa a instância do conceito (ex: "Barack Obama") e a URI do conceito associado à expressão linguística (através da ligação entre o Yago e o InferenceNet). Note que essa estratégia é genérica e válida para todas as instâncias e conceitos associados à expressão linguística.

c) Inferências, no Yago e DBpedia, em subclasses do conceito associado com a expressão linguística (ex: "político") para recuperar a URI dos subconceitos (ex: senador, prefeito, legislador, etc.). Essa inferência é realizada quando não é possível recuperar conteúdo utilizando a URI da estratégia (b). As subclasses são recuperadas por raciocínio transitivo (rdfs:subPropertyOf e rdfs:subClassOf) na DBpedia e Yago.

d) A recuperação de conteúdo por tipo funciona como se segue. Primeiro, definimos alguns arquivos (datasets) como provedores de um certo tipo de conteúdo. *The New York Times*, por exemplo, fornece notícias, a *BBC Artist Info API* fornece informação sobre artistas, a *Open Library Book* fornece informações sobre livros, *flickr™ wrapper* fornece fotos, e assim sucessivamente. Localizamos na DBpedia, que funciona como um *hub* de conceitos, um artigo sobre o conceito e procuramos por conteúdo adicional através da propriedade "owl:sameAs". Tal estratégia não é completa, pois não garante que existe informação adicional sobre um conceito em outros arquivos (os quais pré-definimos como provedores de conteúdo). Para mitigar essa limitação, procuramos por propriedades que se tornaram um padrão na LOD para a descrição de conteúdo de um determinado tipo. A propriedade *foaf:depiction*, por exemplo, tornou-se informalmente padrão para ligação de imagens. Nós então procuramos através dos arquivos por essa propriedade para encontrar imagens sobre um assunto.

O produtor de conteúdo interage com *SemWidgets* informando os conceitos a serem procurados na LOD. O trecho de código do *widget* é criado com geração automática de consultas SPARQL na LOD. O produtor de conteúdo copia e cola o código em seu sítio web. O *widget* explora a LOD para os usuários finais.

4. EXEMPLO DE FUNCIONAMENTO

Para exemplificar como o conhecimento representado em um *widget* gerado pelo *SemWidgets* funciona, vamos assumir a existência de um mapa colaborativo chamado "Política". Num mapa colaborativo, informações advindas de vários usuários são mapeadas em um serviço de mapas digital como *Google Maps* (maps.google.com). Mapas colaborativos podem ser customizados com *widgets* que aportam informações adicionais ao contexto do mapa. No exemplo em questão, vislumbramos um produtor de conteúdo que criou um mapa colaborativo para mapear políticos de uma determinada região. Nesse mapa, o produtor de conteúdo, utilizando *SemWidgets*, cria um *widget* a partir da expressão linguística "Político", o qual foi associado ao conceito "político" do *InferenceNet* que, por sua vez, é ligado ao Yago e DBpedia, como descrito na sessão anterior.

O mapa colaborativo é assim customizado com a inclusão de um *widget* gerado por *SemWidget*, que será responsável por trazer informações de diversas fontes como Wikipédia, New York Times, BBC, etc. relacionadas com políticos. Além de informações sobre políticos e instâncias deles conforme as heurísticas apresentadas anteriormente, o *widget* consulta a base *InferenceNet* e encontra relações de senso comum relacionadas ao conceito político. Relações como a existente entre políticos, corrupção e escândalos são exemplos disso.

A Figura 2 mostra os resultados de uma consulta feita pelo *widget*. Podemos ver links para notícias bem como artigos sobre políticos de uma determinada região do mapa (no caso, o Estado de Illinois, US). Nesse exemplo, a requisição foi feita ao *The Article Search API* do *New York Times e à DBpedia*, enviando: (i) URI do conceito "político" na DBpedia, recuperado através da ligação entre InferenceNet e Yago, e (ii) URI dos conceitos "lei" , escândalo" e "corrupção" (relacionados à "político"), os quais foram recuperados através da ligação entre *InferenceNet* e Yago.

A Figura 2 mostra ainda links representando as relações de senso comum que foram identificadas. Caso o usuário decida por navegar nos mesmos, outras consultas são realizadas, desta feita filtrando informações relacionadas às relações como matérias jornalísticas sobre leis de políticos, casos de corrupção ou escândalos quaisquer. Note que os assuntos sugeridos por *SemWidgets* ao usuário são definidos a partir do conceito *InferenceNet* associado à expressão linguística. No exemplo mostrado na figura, os possíveis assuntos foram criados pelo conteúdo inferencial do conceito "político", expresso no *InferenceNet*: "um político é capaz de propor leis" e "um político tem a propriedade de corrupto". Esse conhecimento de senso comum é expresso na InferenceNet através de relacionamentos (CapazDe, "político", "propor lei", "Pre") e (PropriedadeDe, "político", "corrupto", "Pre"), respectivamente. Além disso, os conceitos "político", "lei" e "corrupção" são associados – via LOD – aos recursos Yago e DBpedia.

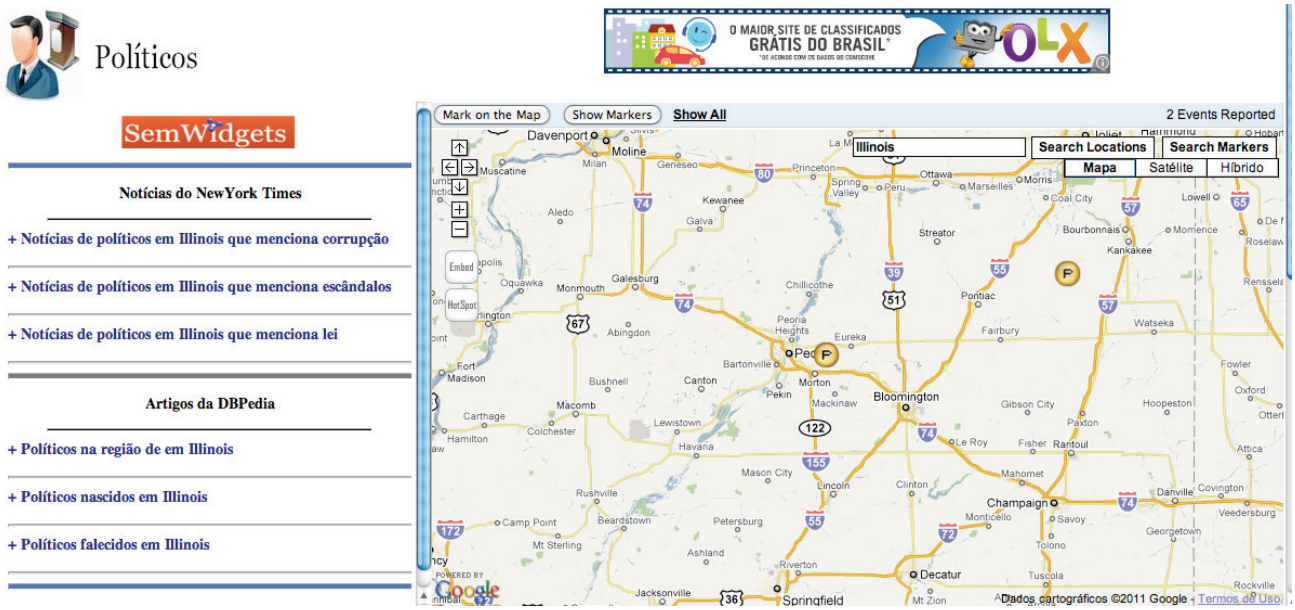

Figura 2. Exemplo de como as informações advindas da LOD são estruturadas no *widget* semântico gerado a partir de *SemWidget*

Ainda na Figura 2 vê-se uma outra faceta da estruturação da informação que é possível a partir do conhecimento da semântica do conceito e que foi implementada especificamente para *widgets* a serem usados em mapas.

Trata-se da classificação dos artigos da DBpedia a partir de referencias geográficas do local que o mapa está apontando. Além disso as informações referentes ao local de nascimento, morte, de trabalho bem como outras que fazem referencia à base geonames[1] permitem ao widget semântico automaticamente classificar os artigos por essas características.

5. CONCLUSÃO

Alguns projetos como Wolfram Alpha (www.wolframalpha.com) têm como objetivo fornecer uma "máquina inteligência computacional" e permite igualmente a geração de *widgets* de conhecimento. Entretanto, utiliza uma alternativa de exploração proprietária de bases de conhecimento que poderia eventualmente dar suporte ou complementar nossa abordagem.

Nós descrevemos *SemWidgets*, uma plataforma para criação de *widgets* que procura conteúdo da LOD. *SemWidgets* permite que um usuário que queira customizar um sítio web o faça com a descrição do conceito ao qual ele(a) deseja recuperar informações relacionadas na LOD. Mais ainda, conceitos relacionados oriundos de uma base de senso comum também são recuperados. Com isso, os *widgets* gerados pelo *SemWidgets* possuem a capacidade de realizar inferências, acessar fontes externas e estruturar informações de forma que sejam informativas e de fácil acesso, constituindo-se assim mais apropriadas ao contexto do site.

A caracterização semântica potencializa a exploração e classificação automática de informações dentro de um contexto específico (no caso, em um determinado sitio web). A aplicação das heurísticas de caracterização semântica na plataforma de geração de *widgets* mostrou-se ainda útil para suprir a falta de plataformas para o desenvolvimento de aplicações, preferencialmente por usuários não especialistas, onde dados são caracterizados semanticamente e posteriormente explorados por essas aplicações. Ao fazer isso, fornecemos uma ferramenta para customização de sítio web com capacidade de inferência.

[1] http://www.geonames.org/ontology/

REFERENCIAS BIBLIOGRÁFICAS

[1] Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, C.; and Ives, Z. 2007. DBpedia: A Nucleus for a Web of Open Data. Proceedings of the 6th International Semantic Web Conference (ISWC2007).

[2] Brandom, R.B. 2000. Articulating Reasons: An Introduction to Inferentialism. Cambridge, MA, Harvard University Press.

[3] Berners-Lee, T. 2006. Linked Data - Design Issues. http://www.w3.org/DesignIssues/LinkedData.html.

[4] Fellbaum, C. 1998. (Ed.): WordNet: An electronic lexical database. MIT Press.

[5] Liu, H.; and Singh, P. 2004. Commonsense Reasoning in and over Natural Language. In: Knowledge-Based Intelligent Information and Engineering. LNCS (3215/2004). Springer, Heidelberg.

[6] Pinheiro, V.; Pequeno, T.; Furtado, V.; and Franco, W. 2010. InferenceNet.Br: Expression of Inferentialist Semantic Content of the Portuguese Language. In: T.A.S. Pardo et al. (eds.): PROPOR 2010, LNAI 6001(90-99). Springer, Heidelberg.

[7] Suchanek, F.M.; Kasneci, G.; and Weikum, G. 2007. Yago: a core of semantic knowledge, Proceedings of the 16th international conference on World Wide Web, May 08-12, Banff, Alberta, Canada.