Anais do CELSUL 2008
ISSN ????-???? [a ser preenchido pelos editores]

Construção de Ontologias Baseadas na Wikipedia

Clarissa Castellã Xavier[1], Vera Lucia Strube de Lima[1]

[1]Faculdade de Informática – Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)

Clarissa.xavier@pucrs.br, vera.strube@pucrs.br

Abstract. *This essay brings a preliminary approach to the subject, reviewing five papers that propose techniques of ontology extraction based on Wikipedia.*

Resumo. *Este estudo traz uma abordagem preliminar ao assunto, apreciando cinco artigos científicos que propõem técnicas de extração de ontologias a partir da Wikipedia.*

Palavras-chave: ontologia; Wikipedia

1. Introdução

A palavra "ontologia" deriva do grego onto (ser) e logia (discurso escrito ou falado). Na Filosofia, a ontologia é a ciência do que é, dos tipos de estruturas dos objetos, propriedades, eventos, processos e relacionamentos em todas as áreas da realidade. Seu objetivo é fornecer sistemas de categorização para organizar a realidade.

Pesquisadores da Web e da Inteligência Artificial adaptaram o termo aos seus próprios jargões e, para eles, uma ontologia pode ser um documento ou arquivo que define formalmente as relações entre termos mais gerais e termos mais específicos.

Os recentes desenvolvimentos relacionados à Gestão do Conhecimento e à Web Semântica têm mostrado a necessidade de ontologias para descrever, de modo formal, conceitos compartilhados a respeito dos mais diferentes domínios. Para que computadores e pessoas possam trabalhar em cooperação é necessário que as informações por eles utilizadas, tenham significados bem definidos e compartilhados. Ontologias são instrumentos viabilizadores dessa cooperação. Entretanto, a construção de ontologias envolve um processo complexo e longo de aquisição de conhecimento, o que tem dificultado a utilização desse tipo de solução em mais larga escala. Em resposta a essa dificuldade, a comunidade científica vem explorando novas técnicas e abordagens que possam reduzir esse esforço.

Quando se observa o grande número de domínios para os quais não existem ontologias especificadas e compartilhadas, fica evidente a importância do estabelecimento de métodos de construção mais ágeis, tanto no momento de aquisição de conhecimento e construção inicial, quanto em momentos de atualização e ajuste das ontologias.

Verifica-se que, nas comunidades baseadas em ferramentas Wiki, em especial a Wikipedia, encontram-se um grande número de usuários habilitados a criar um expressivo domínio de representações, identificadores e definições de conceitos. A versão em inglês da Wikipedia atualmente contém mais de dois milhões de entradas. A

possibilidade da utilização de metodologias que auxiliem na geração de ontologias baseadas na Wikipedia é útil e promissora.

Este estudo propõe-se a realizar um levantamento dos trabalhos realizados nesta direção, as técnicas que estão sendo utilizadas e os resultados obtidos através da análise de cinco artigos científicos propondo técnicas de extração de ontologias a partir da Wikipedia.

2. Wikipedia

A Wikipedia é baseada no sistema MediaWiki, desenvolvido especialmente para a enciclopédia, mas que atualmente também é utilizado em outros sitios [Völkel 2006]. O software MediaWiki é escrito na linguagem de programação PHP e utiliza o banco de dados MySQL .

Atualmente, para acessar seus artigos, a Wikipedia oferece apenas duas opções: um sistema de busca de texto e uma lista de categorias de artigos, organizada hierarquicamente.

Páginas na Wikipedia são explicitamente associadas a uma ou mais categorias. As categorias devem representar tópicos principais e sua principal função é auxiliar na localização da informação. Existem dois tipos de categorias. O primeiro é utilizado na classificação de páginas em relação aos tópicos abordados, podendo ser hierarquicamente estruturados. Por exemplo, uma página pode estar na categoria "ciência" e ter subcategorias biologia ou geografia. O segundo tipo de categoria são as listas, que costumam conter *links* para instâncias de certos conceitos como, por exemplo, uma lista de países asiáticos. Além disso, os artigos possuem *links* para outros artigos, o que ao mesmo tempo facilita a navegação no conteúdo da enciclopédia, mas também representa uma forma de relação semântica entre artigos ou categorias [Chernov 2006].

3. Trabalhos Analisados

Foram estudados em detalhe cinco trabalhos propondo extração de ontologias da Wikipedia. A escolha desses trabalhos foi determinada por sua publicação recente em eventos relevantes e, principalmente, por descreverem com clareza propostas consideradas implementáveis no contexto de um projeto de um ano de duração no âmbito de um curso de mestrado.

3.1. Semantic Wikipedia

O objetivo do trabalho Semantic Wikipedia [Völkel 2006] é implementar uma extensão para o MediaWiki que permita tornar partes importantes do conhecimento contido na Wikipedia processáveis automaticamente, com o menor esforço possível. A técnica descrita introduz *Links Tipados* e *Atributos Tipados*.

A meta do *Link Tipado* é explicitar a relação do *link* com o artigo em que este se encontra, através de uma pequena extensão sintática, permitindo que os usuários expressem a relação entre duas páginas, ou entre seus respectivos assuntos. Ao invés de se descrever, em um artigo, um *link* para *England*, por exemplo, no texto "London is the Capital of England", como [[England]], o trabalho propõe que seja aplicada a seguinte

extensão: [[capital of:: England]] na descrição do *link*, descrevendo uma relação "Capital of" entre "London" e "England".

Na versão corrente da Wikipedia, os valores de atributos geralmente são apresentados como texto simples, não havendo marcação para este tipo de informação, em contraste com os *links* que já possuem um marcador especial.

Ao invés de simplesmente digitar o texto "*The total resident population of London was estimated 7,241,328*", Völkel e seus co-autores propõem uma extensão introduzindo *Atributos Tipados*, ou seja, rótulos junto a um valor, como por exemplo: [[population := 7,421,328]] na descrição anterior. O exemplo a seguir, obtido de [Völkel 2006], apresenta o código fonte de um artigo sobre Londres da Wikipedia em inglês, semanticamente estendido com o uso da linguagem de marcação proposta:

'''London''' is the capital city of [[capital of::England]] and of the [[is capital of::United Kingdom]]. As of [[2005]], the total resident population of London was estimated [[population:=7,421,328]]. Greater London covers an area of [[area:=609 square miles]]. [[Category:City]]

Este trabalho é interessante por apresentar detalhadamente a arquitetura proposta e por disponibilizar na internet o código fonte do algoritmo apresentado[1].

3.2. Extração a partir de Wikipedia Templates - What Have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content Semantic Wikipedia

Auer e Lehmann, em [Auer 2007], apresentam um método para extrair conteúdo semântico estruturado, a partir de instâncias de modelos Wiki. O trabalho sugere meios de pesquisar o grande volume de informações extraídas da Wikipedia em inglês, analisa a qualidade do conteúdo extraído e propõe estratégias para melhorar a qualidade destas informações com poucas alterações nos sistemas Wiki utilizados atualmente. Para isso ele propõe uma combinação simples de sistemas Wiki já existentes e paradigmas da representação Web Semântica.

A ferramenta MediaWiki possui modelos que servem para incluir conteúdo predefinido ou apresentar conteúdo. Por exemplo, os *Infoboxes* são um tipo especial de modelos que geraram caixas formatadas para determinados conteúdos em artigos que descrevam instâncias de um determinado tipo. O MediaWiki permite que os autores de artigos representem a informação estruturadamente através de uma notação atributo-valor, que é processada dentro da página Wiki por um modelo associado. O artigo propõe métodos para: separar a informação que é realmente importante para a avaliação; extrair esta informação dos modelos nos textos Wiki e convertê-la em o RDF; buscar e apresentar esta informação.

O algoritmo de extração semântica dos modelos opera em cinco estágios:

1. Selecionar todas as páginas da Wikipedia contendo modelos;

3. Extrair e selecionar modelos significativos, ou seja, aqueles com alta probabilidade de conter informação estruturada, baseado na seguinte heurística:

[1] http://sourceforge.net/projects/semediawiki

são descartados modelos com apenas um ou dois atributos e modelos raramente utilizados (este conceito de “raramente” permanece em aberto para os autores);

3. Realizar o *parsing* de cada modelo e gerar as triplas apropriadas onde cada atributo do modelo corresponde ao predicado da tripla e o valor do atributo é convertido no seu objeto.

4. Pós-processar os objetos para gerar referências URI adequadas ou valores literais;

5. Determinar os membros de uma classe para a página da Wikipedia sendo processada.

Este trabalho caracteriza-se por ter como base para a extração semântica de dados os modelos pré-definidos da enciclopédia, atendo-se a implementação atual da Wikipédia, sem adição de extensões.

3.3. Performing Cross-Language Retrieval with Wikipedia

Schönhofen, Benczur, Biro e Csalogany propõem a extração de uma ontologia básica da Wikipedia em língua inglesa como o primeiro passo na construção de um sistema de tradução da língua inglesa para o búlgaro [Schönhofen 2007].

O objetivo é converter o corpo regular da enciclopédia, da marcação Wiki para XML, omitindo imagens, tabelas de conteúdo, notas de rodapé e outros elementos complementares não diretamente relacionados com o conteúdo dos textos. Páginas de categorias são excluídas, redirecionamentos são considerados títulos adicionais dos artigos para os quais eles apontam. Páginas de desambiguação, que listam as diferentes interpretações possíveis do mesmo termo, são quebradas em fragmentos menores com o mesmo título, para que o algoritmo diferencie os artigos referenciados.

Em seguida, títulos de artigos de redirecionamento são adicionados como títulos adicionais dos artigos alvo, e referências entre artigos são coletadas do corpo dos artigos. Aplica-se, aos artigos da Wikipedia, o mesmo pré-processamento aplicado aos termos em inglês no dicionário, a fim de tornar a futura correspondência entre os termos da consulta e da Wikipedia tão precisa quanto possível. Após, são removidos os caracteres especiais e por fim se faz o balanceamento da árvore de decisão montada. O algoritmo computa o número de ocorrências de bigramas, ou seja, as ocorrências da palavra *w* seguida pela palavra *v*.

Este trabalho mostrou-se interessante pela simplicidade do algoritmo descrito.

3.4. Extração de Relações Semânticas entre Categorias

Chernov, Iofciu, Nejdl e Zhou [Chernov 2006] propõem a extração de informação semântica da Wikipedia através da análise dos *links* entre as categorias de artigos da enciclopédia e apresentam resultados experimentais através da análise da Wikipedia em língua inglesa.

O artigo propõe a construção automática de um esquema de banco de dados, que enfatiza os relacionamentos significativos entre categorias, visto que categorias altamente conectadas representam relações semânticas fortes. Por exemplo: a categoria “*Country*” possui *links* para a categoria “*Capital*”, inferindo-se que deve haver uma

relação "*Country to Capital*" entre duas instâncias destas categorias. Entretanto, se há poucos *links* entre duas categorias como "*Actor*" e "*Capital*", conclui-se que não há relação semântica regular "*Actor to Capital*".

Nos experimentos conduzidos para testar este método de filtragem, os autores extraem o conjunto principal de páginas que possuem um tópico comum (no caso, *Country*). Para estas páginas, eles extraem todas as categorias a que elas pertencem, e também duas listas de categorias, uma para as páginas com *links* que levam a *Countries (inlinks)* e outra para páginas referidas em *Countries (outlinks)*. Foram selecionados três conjuntos de páginas, denominados *Countries, Inset e Outset*. O principal critério de avaliação utilizado foi a qualidade das relações semânticas extraídas. Para isso, utilizaram a medida *Semantic Connection Strenght (SCS)*. Na avaliação, foi dada a seguinte instrução aos avaliadores: "A categoria A está fortemente relacionada à categoria B (valor 2) se você acredita que todas as páginas na categoria A possuem conceitualmente pelo menos uma ligação semântica para B; A e B estão relacionadas com valor 1, se acredita que 50% das páginas de A possuem ligação semântica com B; caso contrário, A e B são pouco relacionadas (valor 0)".

Os resultados experimentais mostraram um alto nível de desacordo entre avaliadores (por vezes 40%), indicando que *SCS* é uma medida muito subjetiva e deve ser melhorada no futuro. Também observou-se que, para uma determinada categoria, *inlinks* possuem um desempenho superior a *outlinks*. Isso pode ser um sinal da importância dos *inlinks* ou uma evidência de uma propriedade especial da categoria *Countries*. Além disso, o experimento verificou que uma medida normalizada da *Connection Strenght* é melhor para a extração de relações semânticas entre as categorias.

Este trabalho caracteriza-se por uma proposta interessante, que infelizmente ainda não alcançou os resultados desejados, o que levou os autores a concluir que o número de experimentos ainda não é suficiente para levar a uma avaliação convincente.

3.5. Harvesting Wiki Consensus - Using Wikipedia Entries as Ontology Elements

Hepp, Bachlechner e Siorpaer [Hepp 2006] propõem o reuso do vasto material de entradas da Wikipedia como componentes de uma ontologia. Seu objetivo é mostrar que a tecnologia Wiki pode ser facilmente utilizada como um ambiente de desenvolvimento de ontologias, reduzindo barreiras para a participação dos usuários na criação e manutenção de ontologias leves; apresentando uma análise quantitativa de Wikipedia corrente no momento da elaboração do trabalho e suas propriedades; provando que as URIs das entradas da Wikipedia são confiáveis como conceitos de uma ontologia; e finalmente, demonstrando como entradas disponíveis na Wikipedia podem ser utilizadas como elementos de uma ontologia.

Sua estratégia é:

1. Desenvolver uma solução técnica para o uso das entradas da Wikipedia como elementos de uma ontologia em RDF;
2. Retirar uma amostra aleatória de n artigos (n = 100) de uma parte da versão em inglês da Wikipedia;

3. Analisar se o conceito representado pela URI, no momento da adição da entrada, ainda é coerente com a descrição mais atual recuperável na respectiva URI, ou seja, se anotações feitas utilizando o URI, no passado permanecem corretas, apesar do fato de as entradas Wiki poderem ser facilmente modificadas pela comunidade. Em especial, analisar a quantidade de páginas de desambiguação, que são inseridas quando um mesmo termo refere-se a conceitos distintos em inúmeros contextos.

4. Analisar quantitativamente propriedades como a média da idade das entradas e a quantidade de mudanças por hora, provando que, uma vez que o trabalho assume que amostras aleatórias trazem estimativas confiáveis para o total da população (ou seja, o conteúdo total da Wikipedia), a abordagem retorna dados precisos sobre a adequação dos conteúdos Wikipedia como conceitos.

O estudo verificou que apenas 3% da amostra tornaram-se páginas de desambiguação durante o período avaliado. Além disso, 94 de 100 entradas permaneceram estáveis e podem ser utilizadas como fontes de dados sem maiores problemas. Uma entrada apresentou uma pequena alteração e 5 mudanças significativas.

Em relação às propriedades de distribuição das modificações por URI, em média houveram 9,5 alterações durante o período avaliado. Ou seja, 50% das entradas foram alteradas 9,5 vezes ou menos no período. Em relação ao período de sua existência, 50% foram alteradas 1,2 vezes por mês ou menos.

Se forem multiplicados o número de modificações por mês de existência (2,9) com o total de entradas da Wikipedia no momento da pesquisa (850.000 em novembro de 2005), revela-se uma média de 3.465.000 alterações em entradas em inglês por mês, apontando para uma atividade intensa dos colaboradores.

Este trabalho é interessante por apresentar resultados objetivos que comprovam a viabilidade da metodologia proposta e a viabilidade do uso do conteúdo da Wikipedia na extração de ontologias.

4. Considerações

Os trabalhos analisados propõem a extração de ontologias de alto nível baseadas no corpus completo da versão em língua inglesa da enciclopédia. A finalidade desta tarefa é tornar a maior fonte de dados editados colaborativamente da atualidade processável de modo totalmente automático.

A linguagem utilizada na descrição das ontologias foi, em todos trabalhos analisados, RDF. Ela foi escolhida principalmente por ser escrita em um formato simples (XML) e gozar de flexibilidade na criação de ontologias, além de ser o padrão atual de linguagem para a construção de ontologias para a Web Semântica.

Apenas o trabalho de Völkel *et al*. [Völkel 2006] propõe uma extensão da Wikipedia, ou seja, uma modificação no modo de edição da enciclopédia através da inserção de dois novos tipos de rótulos: *Links Tipados* e *Atributos Tipados*. Esta proposta implica a alteração do software de edição MediaWiki e alinha a Wikipedia com o projeto Web Semântica, visando tornar a enciclopédia uma base mais robusta para a busca de informações. Os outros trabalhos apresentam uma proposta menos

ambiciosa, dada a dificuldade da implantação prática de uma alteração no MediaWiki, limitando-se a propor algoritmos baseados na ferramenta de edição já existente.

Os principais desafios às propostas foram a usabilidade, visto que pelos princípios Wiki, qualquer um pode ser capaz de editar a enciclopédia, e a escalabilidade, dado o crescimento constante da base de dados.

Os próximos passos da pesquisa que está sendo conduzida são: avaliação com critérios formais de algoritmos que propõem a extração de ontologias a partir da Wikipedia, visando com isso capturar qual o melhor caminho nesta tarefa dentro do que está sendo proposto na literatura e assim propor uma ferramenta que extraia ontologias de domínio da Wikipedia em língua portuguesa.

5. Referencias e Citações

AUER, Sören; LEHMANN, Jens. **What Have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content**. In Franconi et al. (eds), Proceedings of European Semantic Web Conference (ESWC'07), LNCS 4519, pp. 503-517, Springer, 2007.

BREITMAN, Karin Koogan. **Web semântica: a internet do futuro**. Rio de Janeiro: LTC, 2005.

CHERNOV, Sergey; IOFCIU, Tereza; NEJDL, Wolgang; ZHOU, Xuan. **Extracting Semantic Relationships between Wikipedia Categories**. In: First Workshop on Semantic Wikis: From Wiki to Semantic[SemWiki2006]. Proceedings. Budva, Montenrego, Jun, 2006.

HEPP, Martin; BACHLECHNER, Daniel; SIORPAER, Katharina. **Harvesting Wiki Consensus - Using Wikipedia Entries as Ontology Elements**. In: First Workshop SemWiki2006 - From Wiki to Semantics co-located with the 3rd Annual European Semantic Web Conference (ESWC). Proceedings. Budva, Montenrego, Jun, 2006.

SCHÖNHOFEN, Peter; BENCZUR, András; BIRO, István; CSALOGANY, Károly. **Performing Cross-Language Retrieval with Wikipedia**. In: Working Notes for the CLEF 2007. Workshop. Budapeste, Hungria, Set, 2007.

VÖLKEL, Max; KRÖTZSCH, Markus; VRANDECIC, Denny; HALLER, Heiko; STUDER, Rudi. **Semantic wikipedia**. In: Proceedings of the 15th international conference on World Wide Web, Edinburgo, Escócia, Mai, 2006).