CENTRO UNIVERSITÁRIO FEEVALE

FERNANDO MERTINS

ONTOLOGIAS NA ANÁLISE DE *BLOGS*

Anteprojeto de Trabalho de Conclusão

Novo Hamburgo, abril de 2007.

FERNANDO MERTINS

fmertins@terra.com.br

ONTOLOGIAS NA ANÁLISE DE *BLOGS*

Centro Universitário Feevale
Instituto de Ciências Exatas e Tecnológicas
Curso de Ciência da Computação
Anteprojeto de Trabalho de Conclusão

Professor orientador: Rodrigo Rafael Villarreal Goulart

Novo Hamburgo, abril de 2007.

RESUMO

O principal tema do trabalho é a representação e extração de conhecimento através de ontologias. Busca-se estudar e aplicar esta tecnologia mais especificamente na análise de *blogs* da internet. Uma ontologia é um modelo de dados que define e representa um conjunto de conceitos e seus respectivos relacionamentos, geralmente dentro de um determinado domínio. Os *blogs* são uma forma padronizada e mundialmente utilizada para publicar conteúdo, podendo inclusive influenciar positiva ou negativamente a imagem de organizações. Entretanto, os mecanismos atuais de busca não consideram o significado da informação, ou seja, não conseguem, por exemplo, identificar um bloco de texto cujo contexto diz respeito a uma outra empresa concorrente do mesmo setor. Eles agregam alguns recursos específicos para pesquisa em *blogs*, porém não refinam os resultados semanticamente. Ou seja, sem um modelo de representação do conhecimento, não há como realizar buscas mais precisas e inteligentes. Sendo assim, este trabalho tem como objetivo modelar uma ontologia de domínio sobre o setor de transporte aéreo civil para que seja possível analisar o conteúdo de *blogs* que possam influenciar a reputação das organizações. Após uma revisão bibliográfica inicial sobre os principais assuntos abordados, a ontologia será modelada, populada e avaliada, sendo possível ainda a execução de experimentos.

Palavras-chave: ontologia. blog. representação do conhecimento. internet. modelagem.

SUMÁRIO

MOTIVAÇÃO

O sucesso e os benefícios da internet têm ajudado pessoas, instituições e empresas diariamente, facilitando e otimizando suas tarefas na realização dos mais variados tipos de negócios e pesquisas.

Contudo, a realização destas atividades se limita às características tecnológicas empregadas na divulgação de informações na Internet, como por exemplo a linguagem HTML[1], cujo propósito é determinar como ter acesso à informação e como apresentá-la, mas nada que considere o significado do conteúdo.

O computador entende o que é um parágrafo, corpo de texto e imagem, mas não existem meios de saber, como por exemplo, numa página de comércio eletrônico, que a foto do produto se refere ao item apresentado ao lado dele (BAUM, 2006). Nos últimos anos pesquisadores das áreas de computação e letras têm desenvolvido teorias para codificar significado ao vasto conteúdo informacional disponível na internet. Dessa forma pessoas e programas inteligentes podem realizar buscas de forma mais eficiente. O resultado dessas propostas que mais se destaca são as ontologias.

Uma ontologia, segundo Daconta (2003, p. 166), é um modelo de dados que define e representa um conjunto de conceitos (significados) e seus respectivos relacionamentos, geralmente dentro de um determinado escopo. São utilizadas por diversos tipos de sistemas e aplicações que precisam compartilhar e reaproveitar conhecimento de domínio, como um assunto da área de medicina, advocacia ou educação. As ontologias são estruturadas e funcionam através de:

- Classes (também conhecidas como Conceitos);

[1] *Hyper Text Markup Language* (Linguagem de Marcação de Hipertexto)

- Propriedades e seus respectivos valores (descrevem propriedades e características de cada classe);
- Instâncias (materialização das classes);
- Relacionamentos entre as classes;
- Restrições e regras.

Uma ontologia em conjunto com instâncias das classes constitui uma base de conhecimento. Na prática, a diferença entre a ontologia e a base de conhecimento é muito sutil (NOY, 2001).

Um exemplo de aplicação para uso de ontologias é o projeto "Os *blogs* como objeto de percepção e análise de risco à imagem das organizações" (MONTARDO, 2006), onde um dos objetivos é investigar o uso de ontologias por mecanismos de busca de *blogs* na internet.

Blog é a abreviatura comumente utilizada de *weblog* para locais da internet cujos objetivos são publicar e organizar conteúdo, geralmente de forma cronológica, como um diário. Os *blogs* são fáceis de criar e configurar e possuem uma interface agradável de navegação e podem ser utilizados por pessoas que não conheçam programação para a *web* (CIPRIANI, 2006).

Os *blogs* armazenam uma enorme quantidade de conhecimento relacionado a diversos assuntos, desde diários pessoais até opiniões sobre empresas e questões polêmicas. Toda essa informação pode ser utilizada por grandes instituições para extrair conteúdo relevante e analisá-lo, a fim de melhorar a tomada de decisões, planejamentos estratégicos, campanhas de marketing entre outros benefícios.

Dois sistemas de busca de *blogs* mundialmente utilizados, o *Google Blogs* e o *Technorati* agregam alguns mecanismos especiais para pesquisa exclusiva em *blogs*, porém ambos não refinam os resultados encontrados com base em informações de semântica (CARVALHO et al., 2006).

Por exemplo, ao pesquisar pelo nome de uma empresa, não é possível diferenciar o conteúdo que realmente se refere a esta empresa do conteúdo que apenas cita o nome dela, mas tem outro foco principal. Ou seja, os mecanismos de busca não conseguem avaliar o contexto dos *posts*.

Para aprimorar os mecanismos atuais de busca, é necessária a utilização de outras tecnologias além da HTML, que caracterizam a Web Semântica. Estas tecnologias são a XML

(*eXtensible Markup Language*), a RDF (*Resource Description Framework*) e a OWL (*Web Ontology Language*). Conforme Baum (2006, p.43), "estes padrões e descritores permitem aos desenvolvedores adicionar camadas de significado nos documentos da *Web*, criando uma estrutura que define como a informação é conectada e como seus relacionamentos são expressos".

Mas como descobrir se um determinado *blog* faz referência ao assunto ou tema que se está procurando? Os mecanismos atuais de busca fazem uma pesquisa sempre relacionada ao(s) termo(s) utilizado(s) pelo usuário, ou seja, são encontradas todas as ocorrências da expressão informada. Nesta sistemática faltam recursos para especificar a semântica da informação — o significado e o contexto ao qual ela está inserida.

Sem um modelo de representação do conhecimento, não há como realizar buscas mais precisas e inteligentes. "Se os documentos da *Web* fossem definidos semanticamente, buscas e consultas simples poderiam mais facilmente identificar conteúdo relacionado" (BAUM, 2006, p. 43).

Este estudo propõe a utilização de uma ontologia de domínio para modelar as informações do contexto das empresas de aviação civil (companhias aéreas), identificando seus principais objetos, atributos e relacionamentos. A partir disso, poderá ser implementado um mecanismo mais sofisticado de busca, que realize inferência sobre a ontologia.

Acredita-se que os esforços de pesquisa e desenvolvimento deste trabalho trarão benefícios inicialmente para as companhias aéreas, permitindo que o conteúdo de *blogs* possa ser analisado e interpretado de forma mais completa. Posteriormente, novas ontologias poderão ser modeladas para outros domínios, como por exemplo, empresas de telecomunicações.

Os *blogs* são umas das principais formas de publicação de conteúdo na internet, espera-se contribuir com novas possibilidades de busca e extração do conhecimento armazenado neles.

OBJETIVOS

Objetivo geral

Modelar uma ontologia de domínio sobre o setor de transporte aéreo civil para que seja possível analisar o conteúdo de *blogs* que possam influenciar a reputação das organizações.

Objetivos específicos

- Estudar duas metodologias para construção de ontologias;
- Pesquisar e avaliar até três exemplos de aplicações científicas e comerciais que utilizam ontologias, com foco, se possível, no tema do objetivo geral;
- Pesquisar e avaliar até três ferramentas para modelagem de ontologias;
- Desenvolver uma ontologia para tema do trabalho e populá-la manualmente com instâncias, criando uma base de conhecimento;
- Estabelecer um método para avaliar a ontologia proposta;
- Avaliar a ontologia que foi modelada.

METODOLOGIA

O desenvolvimento inicial do trabalho será realizado através da leitura e revisão bibliográfica dos principais assuntos abordados; são eles: ontologias, sistemas de representação de conhecimento e a estrutura de funcionamento dos *blogs*.

1) Estudar duas metodologias diferentes para a criação manual de ontologias e adotada aquela que se concluir ser a mais adequada para o projeto.

2) Pesquisar e analisar até três exemplos de sistemas em utilização que façam uso de ontologias, verificando propostas similares.

3) Pesquisar e avaliar até três ferramentas (e/ou linguagens) para modelagem de ontologias, entre elas com certeza a ferramenta *Protégé*, por já ter sido objetivo de estudos iniciais anteriores.

4) Modelar uma ontologia, específica de uma empresa de transporte aéreo civil, ou seja, identificando e modelando os principais conceitos, regras e relacionamentos deste domínio, utilizando-se a metodologia definida no segundo item. Para esta etapa serão coletadas diversas informações sobre o setor aéreo no Brasil, principalmente em *blogs*. Também está prevista uma entrevista com um usuário especialista do setor, para complementar o trabalho de coleta de informações.

Além da criação da ontologia, a mesma será populada com as informações coletadas, para formar uma base de conhecimento.

5) Pesquisar e/ou desenvolver um método para avaliar a ontologia. Nesta etapa pretende-se conseguir implementar algum tipo de algoritmo ou protótipo de software, que possibilite a execução de testes e experiências utilizando-se a base de conhecimento — por exemplo, realizar inferências. Serão utilizadas preferencialmente ferramentas e tecnologias de código aberto (software livre), possivelmente as desenvolvidas pelos alunos de TC Leandro

Fussiger (*NLPSearch - Framework* para integração de Sistemas de PLN[2] e *API's* de Mecanismos de Busca) e Diego Cirino Kern (Sistema de inferência baseado em ontologia).

6) Executar experimentos para avaliação da ontologia. Esta etapa dependerá diretamente da etapa anterior; estima-se realizar testes e experiências da implementação, ou seja, colocar tudo em prática e funcionamento para no fim realizar as conclusões finais.

A última etapa será entrega de todo o trabalho realizado e apresentação à banca avaliadora.

[2] Processamento de Linguagem Natural

CRONOGRAMA

Trabalho de Conclusão I

Etapa	Meses																			
	Mar.				Abr.				Mai.				Jun.				Jul.			
Desenvolvimento do Anteprojeto.																				
Estudo de duas metodologias para construção de ontologias.																				
Pesquisar e avaliar até três exemplos de aplicações científicas e comerciais.																				
Pesquisar e avaliar até três ferramentas para modelagem de ontologias.																				
Elaboração do TC I.																				

Trabalho de Conclusão II

Etapa	Meses																			
	Ago.				Set.				Out.				Nov.				Dez.			
Desenvolver a ontologia e populá-la manualmente com instâncias.																				
Estabelecer um método para avaliar a ontologia proposta.																				
Avaliar a ontologia que foi modelada.																				
Elaboração do TC II.																				
Apresentação do TC para a banca.																				

BIBLIOGRAFIA

BAUM, David. **Semantic Break Through**. Oracle Magazine, California, V. 20, n. 3, p. 42-46, mai./jun. 2006.

BECKER, Júnior. **Ontologia Terminológica para apoio a ferramentas de Recuperação de informações e de Text Mining**. 2006. Projeto de Diplomação (Bacharelado em Ciências da Computação) – Instituto de Ciências Exatas e Tecnológicas (ICET), Centro Universitário FEEVALE, Novo Hamburgo.

BREITMAN, Karin. **Web Semântica**. São Paulo: LTC, 2005. 212p.

CARVALHO, Cíntia et al. **Monitoramento da imagem das organizações e as ferramentas de busca de blogs**. Prisma.Com, n. 3, p. 420-447, out. 2006. Disponível em < http://prisma.cetac.up.pt/artigospdf/23_sandra_portella_montardo_cintia_carvalho_prisma.pdf >. Acesso em: 07 abr. 2007.

CIPRIANI, Fabio. **Blog Corporativo**. São Paulo: Novatec, 2006. 208p.

CRISTANI, Matteo; CUEL, Roberta. **A Comprehensive Guideline for Building a Domain Ontology from Scratch**. In: I-KNOW '04, 2004. Proceedings… Austria. p. 205-212. Disponível em <http://i-know.know-center.tugraz.at/previous/iknow04/papers/cristani.pdf>

DACONTA, Michael C.; OBRST, Leo J.; SMITH, Kevin T. **The Semantic Web:** A Guide to the Future of XML, Web Services, and Knowledge Management. Wiley Publishing: 2003. 281p.

MONTARDO, S. P. ; CARVALHO, C. ; GOULART, R. ; ROSA, H. . **Blogs e gerenciamento da imagem das organizações**: análise de ferramentas de busca de blogs. In: XXIX Intercom, 2006, Brasília, DF. Anais do XXIX Congresso Brasileiro de Ciências da Comunicação, 2006.

NOY, Natalya F.; MCGUINNESS, Debora L. **Ontology Development 101:** A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, Março 2001. Disponível em: <http://www-ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html>. Acesso em: 07 abr. 2007.

The Protégé Ontology Editor and Knowledge Acquisition System. Disponível em: <http://protege.stanford.edu/>. Acesso em: 07 abr. 2007.

World Wide Web Constortium. **OWL Web Ontology Language Overview**. Disponível em: <http://www.w3.org/TR/owl-features/>. Acesso em: 08 abr. 2007.