UNIVERSIDADE FEDERAL DE PERNAMBUCO
GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
CENTRO DE INFORMÁTICA – UFPE

Proposta de Trabalho de Graduação

ESTUDO DE FUNÇÕES DE SIMILARIDADE SEMÂNTICA DE TERMOS APLICADAS A UM DOMÍNIO

Aluno: Daniel Ferreira da Silva (dfs3@cin.ufpe.br)
Orientadora: Ana Carolina Salgado (acs@cin.ufpe.br)
Co-orientadora: Rosalie Barreto Belian (rbb@cin.ufpe.br)

Outubro de 2007

Conteúdo

Contexto

O crescente surgimento de informações a cada dia na web e em qualquer meio de armazenamento compartilhado tem feito surgir a necessidade de sofisticados mecanismos de busca a essas informações. Apenas a busca por palavras-chave, por exemplo, já não tem sido satisfatória em alguns casos. Além disso, a grande heterogeneidade das bases de dados disponíveis atualmente faz surgir também a necessidade de sistemas que integrem tais bases num sistema unificado para facilitar posteriores consultas com maior precisão e eficiência.

Neste contexto surgiu o termo "web semântica" referindo toda a potencialidade do processamento inteligente da informação disponível na Web. A Web Semântica consiste numa web com toda sua informação organizada de forma que não somente seres humanos possam entendê-la, mas principalmente máquinas [2]. Através da utilização das ferramentas e técnicas voltadas para a web semântica, o processo de extração dos dados da web obtém um melhor resultado, já que os mesmos já se apresentam estruturados. Sistemas de integração de informações na Web compõem o cenário da Web semântica [1] constituindo um dos pré-requisitos para a completa interoperabilidade entre aplicações desta área. Neste sentido, conceitos da Web semântica têm sido assimilados no desenvolvimento de sistemas de integração de informações na Web.

Muito se tem feito para que a integração dos dados dessas bases heterogêneas seja feita da forma mais automática possível, sem intervenção humana. Porém, um dos principais problemas dos sistemas de integração de dados diz respeito à integração semântica dos esquemas das fontes de dados e uniformizá-los, é um processo reconhecidamente difícil de ser realizado automaticamente.

Proposta

Conceitos como ontologias, metadados, contextos e similaridade semântica exercem uma importante influência no entendimento e no desenvolvimento do resultado final num processo de integração de dados em bases heterogêneas. Ontologias são especificações explicitas de uma conceitualização em algum domínio [4]. Já os metadados [5] são comumente definidos como “dados sobre dados”, e podem descrever significado, conteúdo, organização ou objetivos de algum conjunto de dados. Os contextos “contém metadados relacionados ao seu significado, propriedades (tais como fonte, qualidade e precisão), e organização” [3, 7]. Contextos são considerados ferramentas eficazes no tratamento da heterogeneidade da informação [7]. Por fim, temos o conceito fundamental e amplamente usado, [6] o de similaridade semântica. Através do cálculo de similaridade entre termos baseado em ontologias, é possível descrever que termos devem ser usados dentro do processo de integração de dados.

Este trabalho se propõe a pesquisar, estudar e comparar as várias funções existentes que calculem a similaridade semântica entre termos através de consultas em ontologias, de forma a minimizar a interferência humana no processo de integração de bases heterogêneas e nas correspondências semânticas de entidades e atributos. A função também deverá consultar ontologias baseadas em contexto, para que o resultado seja mais preciso e confiável.

Dentre todas as funções estudadas, será escolhida e analisada a função que melhor atender as necessidades de cálculo de similaridade semântica, e será utilizada no processo de geração de um esquema de mediação apresentando aspectos semânticos pertinentes à integração de informações na web identificados no processo de especificação do sistema Integra. O Integra é um sistema para integração de informações distribuídas em fontes de dados na Web [9] e possui uma arquitetura baseada em mediação que adota a abordagem GAV (Global as View) [8] na definição de mapeamentos entre o esquema de mediação e os esquemas das fontes de dados. Os testes serão realizados utilizando base de dados clínicas e o vocabulário utilizado será obtido através da UMLS [10].

Estrutura do Trabalho

Introdução	Esta será a parte inicial do documento onde constará a contextualização, motivação e objetivos do trabalho que será desenvolvido.
Análise das técnicas existentes	Capítulo que mostrará as principais técnicas desenvolvidas para a construção de funções de similaridade.
Testes comparativos das funções de similaridade encontradas	Esta parte mostrará os testes comparativos entre as funções encontradas para escolha da técnica mais apropriada para o trabalho.
Implementação da função na Ontologia UMLS e experimentos	Este capítulo mostrará os resultados obtidos da função escolhida dentro da Ontologia UMLS utilizada no Integra.
Conclusões	Capítulo da monografia que apresentará os objetivos alcançados e idéias para trabalhos futuros.
Referências Bibliográficas	Lista de materiais usados como embasamento para a realização da monografia.

Cronograma

Atividades	**Outubro**				**Novembro**				**Dezembro**				**Janeiro**			
Revisão bibliográfica																
Pesquisa de técnicas aplicáveis																
Realização de testes comparativos																
Implementação da função de similaridade																
Testes e refinamentos																
Escrita do trabalho de graduação																
Preparação da versão final																
Defesa																

Referências Bibliográficas

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web, Scientific America184, no. 5, pp. 34-43, 2001.
2. Elis, Diego. A Web Semântica. Disponível em < http://www.tableless.com.br/a-web-semantica/>. Acesso em 01 outubro 2007.
3. Goh, C., Madnik, S., Siegel, M.: Semantic Interoperability through Context Interchange: Representing and Reasoning about Data Conflicts in Heterogeneous and Autonomous Systems. Sloan School of Management, MIT, http://citeseer.ist.psu.edu/191060.html, 1997.
4. Gruber, T.: A Translation Approach to Portable Ontologies. Knowledge Acquisition, V.5, n.2, p.199-200, 1993.
5. Kashyap, V., Sheth, A.: Semantic Heterogeneity in Global Information Systems: The Role of Metadata, Context and Ontologies. Chapter in Cooperative Information Systems: Current Trends and Directions, M. Papazoglou and G. Schlageter Editors, 1996.
6. Lin, Dekang.: An Information-Theoretic Definition of Similarity. Departament of Computer Science, University of Manitoba, Canada, 2000.
7. Wache, H., Stuckenschmidt, H.: Practical Context Transformation for Information System Interoperability. In Proceedings of the 3rd International Conference on Modeling and Using Context (CONTEXT'01), Lecture Notes in AI, Springer Verlag, 2001.
8. Levy, A.: Logic-Based Techniques in Data Integration. In: J. Minker, editor Logic-based Artificial Intelligence, Kluwer Publishers, 2000.
9. Lóscio, B.: Managing the Evolution of XML-based Mediation Queries. PHD Thesis, Federal University of Pernambuco, Brazil, 2003.
10. UMLS: Unified Medical Language System. Disponível em: <http://www.nlm.nih.gov/research/umls/about_umls.html>. Acesso em: 04 outubro 2007.

Datas e Assinaturas

Recife, 09 de outubro de 2007

Daniel Ferreira da Silva

Aluno

Ana Carolina Salgado

Orientadora

Rosalie Barreto Belian

Co-orientadora