Processamento Semântico de Consultas para Sistemas de Biodiversidade

Bruno S. C. M. Vilar[1], **Claudia M. Bauzer Medeiros**[1]

[1]Instituto de Computação – Universidade Estadual de Campinas (UNICAMP)
Caixa Postal 6.176 – 13.084-971 – Campinas – SP – Brasil

bruno.vilar@students.ic.unicamp.br,cmbm@ic.unicamp.br

Nível: Mestrado
Ano de ingresso: 2007-2
Previsão de término: 2009-1
Etapa já concluída: Defesa da proposta
Palavras-chave: Processamento de Consultas, Ontologias, Biodiversidade, Serviços Web

Resumo. *Sistemas de informação de biodiversidade lidam com um conjunto heterogêneo de informações providas por grupos de pesquisa, como espécies estudadas, estruturação das informações e locais de estudo. Esta heterogeneidade de dados, usuários e procedimentos dificulta o reuso e o compartilhamento de informações. O objetivo deste trabalho é melhorar o processo de consulta às informações em sistemas de biodiversidade. Para tanto, é proposto um módulo que pré-processa uma consulta de usuário (cientista) agregando informações, provenientes de ontologias, para desambigüizar a consulta. O trabalho pressupõe que os dados a serem consultados estão distribuídos em repositórios na Web, os quais são mantidos por grupos de cientistas e têm seus conteúdos acessíveis por serviços Web.*

1. Introdução e Motivação

Estudos em biodiversidade se baseiam em diversos tipos de modelos para definir fatores como riqueza de espécies, abundância, endemismo, distribuição e diferentes outras variáveis [Jr 2007], correlacionadas a dados geográficos. Conforme explica Bisby [Bisby 2000], três fatores contribuem para que a ciência da biodiversidade exija estudos globais: a distribuição geográfica dos envolvidos; a interdependência global de eventos que faz com que uma região seja afetada por fenômenos que ocorrem em outras regiões; a necessidade que há de sintetizar numerosas observações e estudos feitos por observadores, equipes e instituições locais.

Os Sistemas de Informação de Biodiversidade estão neste contexto. Estes sistemas permitem a condução dos processos de avaliação, predição, planejamento e tomada de decisão sobre biodiversidade [Xu et al. 2000]. Para isso, podem ter que englobar coleções heterogêneas de informações de grupos de pesquisa, os quais atuam com metodologias e visões diferentes.

Um dos desafios encontrados no desenvolvimento de Sistemas de Informação de Biodiversidade é fornecer aos usuários a capacidade de consultar as diferentes fontes de informação sem lidar com os aspectos técnicos e específicos das formas pelas quais elas são estruturadas e representadas. O objetivo deste trabalho é melhorar o processo de consulta às informações em sistemas de biodiversidade. Para tanto, é proposto um módulo de processamento de consultas que pré-processa uma consulta de usuário, desambigüizando termos e agregando informações.

A base para o aperfeiçoamento das consultas reside em utilizar ontologias para enriquecimento semântico. Uma ontologia define um conjunto de primitivas de representação com as quais se pode modelar um domínio do conhecimento ou de discurso [Gruber 2008]. Entre as informações que podem ser usadas estão sinônimos, especialização e generalização de termos, seus relacionamentos, entre outros.

O trabalho pressupõe que os dados a serem consultados estão distribuídos em repositórios na *Web*. Cada repositório é mantido por um grupo de cientistas e seu conteúdo é disponibilizado por serviços *Web*. As ontologias a serem usadas são disponibilizadas pelo Aondê, um serviço *Web* desenvolvido no IC-UNICAMP [Daltio and Medeiros 2008], que será estendido e adaptado para atender aos requisitos desta proposta.

Esta proposta está inserida no WeBIOS, um Sistema de Informação de Biodiversidade desenvolvido pelo IC em conjunto com o Instituto de Biologia. Seu objetivo é fornecer apoio a cientistas e pesquisadores da área da Biologia para que estes possam realizar consultas exploratórias multimodais sobre fontes de dados heterogêneas a respeito de biodiversidade.

2. Fundamentação teórica

Técnicas de processamento de consulta são utilizadas como forma de ajustar uma consulta às fontes de informação ou de aperfeiçoar algumas de suas características, seja por uma semântica melhor definida ou por uma sintaxe que beneficie sua execução. Na literatura podem ser encontradas diferentes técnicas, dentre as quais: reescrita [Godfrey and Gryz 1996], expansão [Andreou 2005], substituição [Jones et al. 2006] e relaxamento [Lian et al. 2007, Bosc et al. 2006].

Entre os objetivos visados por essas técnicas estão:

- Aumento de desempenho: menor tempo de execução ou menos recursos utilizados;
- Integração de dados: consulta a bases de dados heterogêneas;
- Variação de resultados: consultas diferentes com mesmo significado;
- Aumento de precisão: fortalecer os critérios aplicados.

O **processamento de consulta** consiste em reformular uma consulta de usuário de tal modo que a consulta resultante forneça resultados significativos adicionais que correspondam à intenção do usuário. Entre as formas utilizadas para aperfeiçoar as consultas está a utilização de conhecimento semanticamente representado através de ontologias. Essa técnica visa reformular uma consulta em outra mais eficiente, semanticamente equivalente [Necib and Freytag 2004].

A **reescrita de consulta** é uma das técnicas que se beneficiam do conhecimento semanticamente representado. As técnicas de reescrita que têm sido propostas exploram *caches* semânticos de consulta, visões materializadas e conhecimento semântico sobre o domínio da base de dados para otimizar a avaliação da consulta [Godfrey and Gryz 1996]. Semelhante à reescrita, a técnica de **substituição de consulta** tem por objetivo substituir a consulta original do usuário por outra similar. A nova consulta se mantém fortemente relacionada à original, contendo termos intimamente relacionados aos originais [Jones et al. 2006].

A **expansão de consulta** é o processo de aumentar a consulta do usuário com termos adicionais, com propósito de melhorar os resultados obtidos [Andreou 2005]. A adoção desta técnica pode levar a problemas como *query drift* [Andreou 2005, Jones et al. 2006], *outweighting* [Andreou 2005] e custo alto de processamento [Jones et al. 2006]. O *query drift* consiste tornar uma consulta distante do interesse original do usuário. Considerado um tipo de *query drift*, o *outweighting* é caracterizado por ter os termos de expansão relacionados a termos específicos de uma consulta e não da consulta como um todo [Andreou 2005].

O **relaxamento de consultas** consiste em generalizar uma consulta falha em uma bem sucedida, por meio da remoção de algumas sub-consultas da consulta original [Lian et al. 2007]. Em um sentido mais amplo, o objetivo pode ser a expansão do escopo de uma consulta pelo relaxamento de restrições envolvidas [Bosc et al. 2006]. O problema desta técnica está na redução da especificidade da consulta, que resulta no atendimento incompleto do objetivo do usuário [Jones et al. 2006].

3. Caracterização da contribuição

O módulo de processamento de consultas a ser criado na dissertação deve permitir aos biólogos consultar informações de repositórios variados a fim de encontrar os registros procurados, detectar relações que não estão diretamente representadas e correlacionar informações. Nesse processo, é preciso facilitar a especificação da consulta para que as características de cada repositório não limitem a abrangência do resultado e os dados coletados não sejam subutilizados. A dissertação deve aproveitar a infra-estrutura do WeBIOS e os serviços do Aondê para dar suporte às consultas dos pesquisadores. A Figura 1 apresenta uma primeira arquitetura para o processamento de consultas.

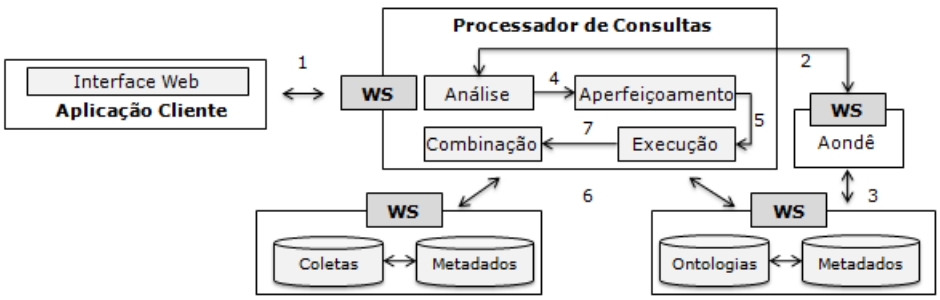

Figura 1. Arquitetura do Módulo de Processamento de Consulta

A arquitetura prevê que as consultas recebidas pela aplicação cliente (1) serão analisadas, em um processo de decomposição de consulta, desambigüização dos termos por meio

das ontologias, usando o Aondê (2,3), e identificação de modificações pertinentes. Após essa análise as consultas serão modificadas (4) e executadas (5) por meio dos serviços *Web* que dão acesso às informações dos repositórios (6). Os resultados obtidos serão combinados (7) e retornados para a aplicação cliente. Para recuperar ontologias relacionadas às consultas serão usados os metadados, referentes aos domínios representados, e processos de análise de similaridade entre os conceitos das ontologias e os termos das consultas e das bases de dados.

3.1. Exemplo: Uso de subclasses (hipônimo)

Considere a consulta em linguagem natural: Retornar insetos da ordem *lepidoptera* que tenham antenas clavadas.

A consulta pode ser representada em uma consulta SQL:

”SELECT * FROM colecao WHERE tipo=’inseto’ AND ordem = ’lepidoptera’ AND antenas = ’clavada’ ”

Suponha que a Tabela 1 ilustre uma tabela do banco de dados a ser utilizado. Se todos os atributos correspondessem aos critérios da consulta, esta poderia ser aplicada diretamente. Porém, como a tabela não possui atributos correspondentes aos critérios especificados, como ’ordem’, é preciso encontrar termos alternativos que satisfaçam à consulta original.

Id	Tipo	Família	Antena	Coloração
1	inseto	hesperiidae	clavada	colorida
2	inseto	lepidotrichidae	filiforme	monótona
3	inseto	cerambycidae	genículo-clavada	colorida

Tabela 1. Tabela de banco de dados sem correspondência total com a consulta.

No caso, como o conceito taxonômico ordem não está representado diretamente, pode-se obter uma consulta alternativa a partir de uma ontologia como a representada na Figura 2.

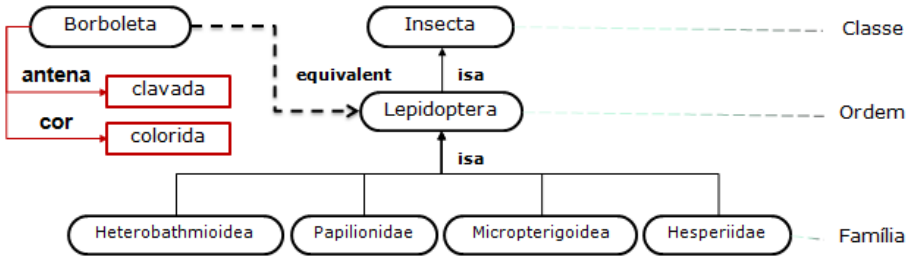

Figura 2. Ontologia que descreve parte da classe *Insecta*

Por meio da relação de herança entre os conceitos, é possível reconhecer que *hesperiidae*, *papilionidae*, *micropterigoidea* e *heterobathmioidea* são sub-conceitos de *lepidoptera*. A consulta então pode ser reformulada da seguinte forma:

”SELECT * FROM colecao WHERE tipo=’inseto’ AND familia in (’hesperiidae’, ’papilionidae’, ’micropterigoidea’, ’heterobathmioidea’) AND antenas = ’clavada’ ”

O resultado é a utilização de um conjunto de famílias, pertencentes ao campo ordem especificado na consulta, do qual participam os mesmos indivíduos que estariam presentes em na ordem *lepidoptera*. Tais indivíduos, portanto, satisfazem ao critério especificado na consulta.

O exemplo utilizou as informações referentes às subclasses para chegar a definições diferentes de um mesmo conceito. Além deste exemplo, as ontologias têm um conjunto de recursos que pode ser empregado para chegar a consultas equivalentes, tais como:

- Conceito (Idêntico, Equivalente, Subclasse, Superclasse, Instância, Interseção e União);
- Relacionamento/Propriedade (Idêntico, Equivalente, Transitivo, Simétrico e Parte/todo);
- Axiomas.

4. Trabalhos relacionados

Na literatura foram encontrados vários trabalhos que lidam com o processamento de consulta. Para atingirem seus objetivos, os trabalhos utilizam diferentes recursos e técnicas, como ontologias de domínio, métodos probabilísticos e mapeamento de esquemas. A Tabela 2 apresenta um quadro comparativo em que cada publicação representa um tipo de processo.

Autor	**Método**	**Ontologias**	**Técnica**	**Aplicação**
Andreou (2005)	Expansão de Consulta	WordNET - Ontologias léxica para avaliar a similaridade semântica e desambiguização	Híbrido - Ontologias e método probabilístico (*Pseudo-Relevance Feedback: Local Context Analysis*)	*Web*
Xiao (2006)	Reescrita de Consulta (global e local)	Mapeamento e integração de BD heterogêneos - Ontologias (RDF)	Integração dos dados	BD heterogêneos (RDF e XML)
Jones et al. (2006)	Substituição de Consulta	-	*Log* das sessões dos usuários: identificar reformulações e palavras correlacionadas	*Web* - Buscas patrocinadas e propagandas
Lian et al. (2007)	Relaxamento de Consulta	Ontologia: descrição para a descoberta de serviços *Web*.	Divide ontologia em visões e classifica conceitos por sua densidade de informações.	Descoberta de serviços *Web*
Necib e Freytag (2004)	Processamento de Consulta (reescrita)	Ontologia para descrever o contexto do BD e as relações entre ambos.	Regras de dedução, regras de generalização e hierarquias conceituais	BD relacional único

Tabela 2. Trabalhos com Processamento de Consulta.

Xiao [Xiao 2006] emprega a Reescrita de Consulta dentro do conceito de mediadores, para que uma consulta global a múltiplas fontes de dados seja reescrita em várias consultas locais através de mapeamentos. Desta forma, obtém-se a transparência no processo de consulta a bases de dados múltiplas. O trabalho de Jones et al. [Jones et al. 2006] deriva as consultas a partir das sessões dos usuários e adota recursos como a mudança ortográfica, substituição de sinônimo, generalização e especialização para que uma nova consulta dê lugar à do usuário. Dessa forma, a consulta do usuário é relacionada a um conjunto de termos de buscas.

No trabalho de Andreou [Andreou 2005], as técnicas e recursos utilizados para a expansão de consultas incluem ontologias e métodos probabilísticos. O trabalho de Necib e Freytag [Necib and Freytag 2004] emprega uma ontologia de domínio para criar regras de derivação das consultas para a reescrita.

Com o conceito de relaxamento de consultas, o trabalho de Lian et al. [Lian et al. 2007] realiza generalizações e busca conceitos que contêm outros, como um estado que contém uma cidade, para encontrar serviços *Web* que se aproximem das necessidades do usuário.

5. Resultados Esperados e Aplicabilidade das Contribuições

Este trabalho se destina a sistemas de informação utilizados por diferentes perfis de usuário e que necessitem facilitar o processo de recuperação das informações. Tal necessidade surge das diferentes concepções que um conhecimento pode ter e que leva a diferentes representações.

O trabalho está na fase de fundamentação teórica, com o estudo dos conceitos envolvidos e da arquitetura do WeBIOS, incluindo o Aondê. Para a sua validação está prevista a utilização de informações do Instituto de Biologia da UNICAMP e das ontologias criadas no trabalho de [Daltio 2007] para a criação de um estudo de caso real.

As principais contribuições esperadas são: (1) levantamento de técnicas de processamento de consulta que envolvam enriquecimento ontológico; (2) proposta e implementação de técnicas para o processamento de consultas com informações de múltiplas ontologias; (3) desenvolvimento de um módulo que processe consultas e facilite a recuperação das informações; (4) aperfeiçoamento dos recursos do Aondê.

Agradecimentos: Apoio financeiro recebido do CNPq e da *Microsoft Research* financiadora do projeto WeBIOS.

Referências

Andreou, A. (2005). Ontologies and query expansion. Master's thesis, University of Edinburgh.

Bisby, F. A. (2000). The Quiet Revolution: Biodiversity Informatics and the Internet. *Science*, 289(5488):2309–2312.

Bosc, P., HadjAli, A., and Pivert, O. (2006). Relaxation paradigm in a flexible querying context. In Larsen, H. L., Pasi, G., Arroyo, D. O., Andreasen, T., and Christiansen, H., editors, *FQAS*, volume 4027 of *Lecture Notes in Computer Science*, pages 39–50. Springer.

Daltio, J. (2007). Aondê: Um serviço web de ontologias para interoperabilidade em sistemas de biodiversidade (aondê: An ontology web service for interoperability across biodiversity information systems). Master's thesis, Instituto de Computação - Unicamp.

Daltio, J. and Medeiros, C. B. (2008). Aondê: An ontology web service for interoperability across biodiversity applications. *Information Systems*. Accepted for publication.

Godfrey, P. and Gryz, J. (1996). A framework for intensional query optimization. In *DDLP*, pages 57–68.

Gruber, T. R. (2008). Ontology. In Liu, L. and Özsu, M. T., editors, *Encyclopedia of Database Systems*. Springer, Berlin, Heidelberg. to appear.

Jones, R., Rey, B., Madani, O., and Greiner, W. (2006). Generating query substitutions. In Carr, L., Roure, D. D., Iyengar, A., Goble, C. A., and Dahlin, M., editors, *WWW*, pages 387–396. ACM.

Jr, L. C. G. (2007). An architecture for querying biodiversity repositories on the web. Master's thesis, Instituto de Computa??o - Unicamp.

Lian, L., Ma, J., Lei, J., Song, L., and Zhang, D. (2007). Query relaxing based on ontology and users' behavior in service discovery. In *Fourth International Conference on Fuzzy Systems and Knowledge Discovery*.

Necib, C. B. and Freytag, J. C. (2004). Using ontologies for database query reformulation. In *ADBIS (Local Proceedings)*.

Xiao, H. (2006). *Query Processing for Heterogeneous Data Integration Using Ontologies*. PhD thesis, University of Illinois.

Xu, H., Wang, D., and Sun, X. (March 2000). Biodiversity clearing-house mechanism in china: present status and future needs. *Biodiversity and Conservation*, 9.