1
Introdução

1.1.
Motivação

A maior parte dos bancos de dados existentes é projetada de maneira independente e, portanto, é geralmente implementada utilizando diferentes esquemas conceituais, criando um contexto de heterogeneidade. Não obstante, quando um conjunto de bancos de dados se refere a um mesmo domínio, eventualmente, surge a necessidade de integrá-los em um mesmo banco, ou de intermediar o acesso ao conjunto de bancos de forma transparente. Essa necessidade aumentou substancialmente com a massiva disponibilização e descentralização de dados promovida pela Web.

Sheth & Larson (1990) destacam três tipos comuns de heterogeneidade: estrutural, sintática e semântica. A heterogeneidade estrutural acontece quando os dados são organizados segundo esquemas conceituais diferentes. A heterogeneidade sintática ocorre quando sintaxes diferentes são atribuídas a conceitos correspondentes. Por fim, a heterogeneidade semântica considera as diferentes interpretações para os dados. É no tratamento desta última que está a maior dificuldade, pois geralmente o significado preciso dos dados não está explícito.

Algumas soluções têm sido propostas pela comunidade de banco de dados para endereçar os problemas relacionados à heterogeneidade de dados. Num cenário ideal, os projetistas de bancos de dados deveriam empregar padrões ao elaborarem esquemas conceituais, resolvendo de maneira a priori o problema (Casanova et al., 2007). No entanto, como mencionado anteriormente, cada projetista geralmente trabalha independentemente, forçando a consideração de abordagens a posteriori para o problema.

A abordagem de alinhamento de esquemas visa identificar e tratar os relacionamentos entre os diversos esquemas (Rahm & Bernstein, 2001). Esse procedimento geralmente é feito por especialistas de domínio, mas tende a ser um trabalho muito tedioso e propenso a erros. Brauner et al. (2006a) empregam uma abordagem complementar, visando o alinhamento de tesauros em atributos

com domínio enumerável, por meio da comparação de instâncias comuns retornadas a partir de consultas a diversas fontes de dados. Zhou et al. (2007) apresentam uma técnica de relaxação de consultas que possibilita buscas efetivas em esquemas maleáveis, os quais são esquemas que intencionalmente contêm definições heterogêneas e sobrepostas de estruturas de dados que podem ser enriquecidas e consultadas a qualquer instante. Como uma última abordagem para o problema da heterogeneidade, citamos o emprego de ontologias, pois estas fornecem um vocabulário comum compartilhado de um determinado domínio de aplicação e, com isso, se mostram bastante promissoras para lidar com o problema de integração de dados (Necib & Freytag, 2005).

Neste trabalho, consideramos uma combinação das abordagens de alinhamento de esquemas e tesauros para tratar o problema da heterogeneidade semântica entre bancos de dados. Mais especificamente, esta dissertação apresenta o *CatalogMatcher*, uma infra-estrutura de software para alinhamento de catálogos heterogêneos. Um catálogo armazena dados sobre um conjunto de objetos de um determinado domínio, tipicamente classificados por algum tipo de taxonomia ou tesauro. O *CatalogMatcher* contém componentes que implementam estratégias de alinhamento de catálogos heterogêneos utilizando abordagens baseadas em instâncias.

1.2. Trabalhos relacionados

Inúmeras propostas para endereçar o problema de alinhamento de esquemas têm sido propostas. Rahm & Bernstein (2001) apresentam um resumo de várias das abordagens existentes, classificando-as em diversos tipos, como sintática vs. semântica, nível de elemento vs. nível de estrutura, etc.

Bilke & Naumann (2005) descrevem um algoritmo baseado em instâncias para alinhamento de esquemas por meio do uso de duplicatas. A técnica consiste na detecção de instâncias idênticas (duplicatas) entre dois bancos de dados com esquemas não alinhados e na aplicação de algoritmos de similaridade sobre essas instâncias para a realização do alinhamento.

Wang et al. (2004) apresentam uma técnica de alinhamento de esquemas baseada em sondagem por consultas em domínio específico para bancos de dados Web. A abordagem depende de intervenção humana para a seleção de um conjunto de instâncias de exemplo, com as quais é feita a sondagem. Além

disso, os autores propõem um modelo de esquemas combinado para descrever os vários esquemas associados a um banco de dados Web. Também utilizando técnicas baseadas em sondagem por consultas, apresentamos em Brauner et al. (2008) uma abordagem para construção de esquemas mediados através do pós-processamento de consultas de usuários.

Madhavan et al. (2005) propõem o uso de um corpo (*corpus*) de esquemas e alinhamentos como informações extras para fornecer maiores evidências sobre os esquemas a serem alinhados, aumentando a chance de descoberta de alinhamentos. Além disso, o método possibilita o aprendizado de estatísticas sobre os elementos sendo alinhados, permitindo que restrições possam ser inferidas e, consequentemente, a precisão dos resultados obtidos possa ser melhorada.

Com o objetivo de oferecer uma plataforma para alinhamento de esquemas, Do & Rahm (2002) desenvolveram o COMA, um sistema para a combinação flexível de abordagens de alinhamento de esquemas. Os autores provêem uma biblioteca com um grande número de algoritmos individuais e híbridos para alinhamento de esquemas, oferecendo um bom suporte para combinação de resultados de alinhamento. Dessa forma, o usuário pode, então, selecionar as abordagens que melhor atendem ao seu problema. Em Gazola et al. (2007) apresentamos um protótipo de mediador para *gazetteers*, o qual suporta consultas por palavra-chave e por termos de tesauros. Bernstein & Melnik (2007) consideram o uso de um motor de gerenciamento de modelos cujo objetivo é suportar diversos tipos de operações, como alinhamento de esquemas, composição de mapeamentos, intercalamento de esquemas, tradução de esquemas para diferentes modelos de dados, e etc.

1.3. Organização do trabalho

Este trabalho está organizado da seguinte forma. O capítulo 2 introduz os principais conceitos utilizados ao longo do texto. O capítulo 3 apresenta o *CatalogMatcher*, uma infra-estrutura de software para alinhamento de catálogos heterogêneos. Descreve ainda as principais estratégias de alinhamento implementadas, além da arquitetura da infra-estrutura. O capítulo 4 apresenta um mediador na Web que utiliza o *CatalogMatcher* para intermediar consultas a catálogos heterogêneos, relatando ainda alguns cenários de uso do mediador com catálogos de objetos geográficos e catálogos de livrarias virtuais. Por fim, o

capítulo 5 traz algumas conclusões deste trabalho e indica possibilidades de trabalhos futuros.