1
Introdução

Neste capítulo é abordado o contexto no qual este trabalho se enquadra, ou seja, a integração semântica de dados através de federação de ontologias numa analogia a integração de esquemas de dados heterogêneos e distribuídos mapeados em um modelo global de um sistema de federação de dados. É apresentado inicialmente um histórico para o problema de integração de dados e em seguida é apresentado o problema de integração de dados distribuídos na Web e, a solução encontrada para a integração destes esquemas de dados representados agora em ontologias. – o *merging* de ontologias (Klein, 2001) para a criação de uma federação de ontologias.

1.1 Motivação

Vivemos em uma época em que a informação se tornou um bem valioso, a ponto de nossa sociedade estar sendo chamada de sociedade da informação. Diversas são as fontes que disponibilizam informação, como, por exemplo, rádios, televisões, jornais e mais recentemente a Internet.

Até o surgimento do computador pessoal (PC) nos anos 80, as informações armazenadas em meio digital eram quase que de uso exclusivo de universidades, grandes organizações ou de áreas estratégicas dos governos para suporte a pesquisas e a tomada de decisão. O PC, logo acompanhado pela Internet, permitiu que a computação ganhasse maior escala de uso e, assim, as informações armazenadas em meio digital começaram a aumentar e a se difundir. Os sistemas de informação que eram quase na sua totalidade manuais e se baseavam em técnicas de arquivamento e recuperação de informação, foram dando espaço aos sistemas operacionais automatizados.

Sistemas de informação foram desenvolvidos para suprir de forma mais adequada às necessidades dos usuários da informação, permitindo o acesso rápido, e garantindo a integridade e veracidade destas informações e a privacidade de acesso. Estes sistemas foram evoluindo na mesma proporção em que evoluíam as

tecnologias de computação e telecomunicação. Com o advento das redes e com o aumento da capacidade de armazenamento e processamento dos computadores pessoais, novas formas de sistemas de informação foram surgindo, permitindo no contexto empresarial uma maior integração de áreas e disponibilização de dados dispersos em vários sistemas diferentes. Os sistemas para armazenamento dos dados de forma persistente, os chamados Sistemas de Bancos de Dados, também foram se popularizando. Uma preocupação desta área de pesquisa sempre foi a integração dos dados espalhados por vários sistemas, de forma a fornecer uma informação mais completa e unificada. Integração de dados passou, então, a ser sobrenome da área de banco de dados.

Desde o surgimento dos sistemas de gerenciamento de banco de dados, o problema de integração de dados tem sido reconhecido como criticamente importante (Miller et al., 2001), e ainda demanda pesquisas em banco de dados (Sheth & Larson, 1990; Litwin et al.,1990; Silberschatz & Zdonik, 1997; Abiteboul et al., 2003; Halevy, 2003).

Várias alternativas de integração de dados foram sugeridas, diversos trabalhos foram publicados com arquiteturas que vão desde a integração manual, integração por aplicação, por uso de uma interface comum de acesso aos dados, mediador (*middleware*), sistemas de federação de bancos de dados, sistemas de *geração* de *workflows*, *data warehouses*, *web services*, e mais recentemente portais de acesso e *Peer-to-Peer* (P2P). Cada uma destas abordagens de integração de dados possui vantagens e desvantagens permitindo o acesso de forma materializada ou não (virtual) atendendo a problemas específicos.

Vários tipos de modelos foram sugeridos para representar estes dados, pois, desde a década de 70, os estudiosos da área de banco de dados vêm criando modelos conceituais, vem olhando estes dados na forma de modelos.

Para que o usuário de um sistema de informação tenha uma visão unificada dos dados provenientes de diversas fontes distribuídas e heterogêneas os dados integrados precisam ser representados usando o mesmo princípio de abstração (modelo de dados). Esta tarefa inclui detecção e resolução de conflitos que vão desde a estrutura onde os dados estão armazenados até o significado dos mesmos, os chamados conflitos semânticos.

Nas razões para a integração de dados podemos observar dois aspectos:

- Dado um conjunto de dados pré-existentes, uma visão integrada destes pode ser criada para facilitar o acesso e reuso da informação gerada por eles através de um único ponto de acesso.
- Dada certa necessidade de informação, dados provenientes de sistemas de informação complementares podem ser combinados para obter-se uma base que satisfaça a necessidade de informação de forma mais completa, uma vez que várias informações são combinadas. Existem muitas aplicações que se beneficiam deste tipo de informação combinada como, por exemplo, na área de B.I. (*Business Inteligence*) onde relatórios com informações estratégicas para suporte à decisão precisam ter uma visão integrada da informação.

O objetivo maior dos sistemas de integração de dados sempre foi o de liberar o usuário da necessidade de localizar as diversas fontes, interagir com cada uma isoladamente e combinar os dados das múltiplas fontes manualmente (Halevy, 2003).

Com o advento da Web, que revolucionou os meios de comunicação e a forma de se obter informação rápida e em tempo real, estas necessidades de integração cresceram. Se por um lado a facilidade de acesso à informação aumentou consideravelmente, por outro lado a publicação desta informação se deu de forma desestruturada, ou seja, qualquer pessoa pode publicar informações na Web de forma bastante simples. A informação está disponível e é de fácil acesso de praticamente qualquer lugar do mundo, mas nem sempre os resultados de buscas por informação correspondem àquilo que o usuário esperava encontrar ou são suficientemente confiáveis. Podemos considerar que a Web interliga uma quantidade gigantesca de informações dispersas em inúmeras fontes de dados heterogêneas. Com o uso de banco de dados na Web, surgiu a necessidade de integrá-los. Para tanto, pesquisas têm sido realizadas com o objetivo de incorporar mecanismos tradicionais da área de banco de dados, de forma a possibilitar a integração de fontes heterogêneas e distribuídas.

Dada a heterogeneidade dos dados distribuídos e publicados na Web sem seus metadados associados, busca-se definir uma infra-estrutura capaz de possibilitar a comunicação entre computadores na Web que trate deste problema

da heterogeneidade semântica. Surge, então, a extensão da Web atual - a Web Semântica.

A idéia, segundo (Berners-Lee et al., 2001), é fazer com que o conteúdo da Web, seja acessível por máquinas usando marcação semântica, diferente do tratamento sintático normalmente feito usando XML. Mesmo que o conteúdo da Web seja processável por máquina, ainda pode existir o problema de integração semântica, pois pessoas usam termos iguais para representar diferentes conceitos e um mesmo conceito pode ser representado por diferentes termos (inclusive usando línguas diferentes). Traduções e mapeamentos são feitos entre diferentes comunidades para compartilhar conhecimento. Desta forma, a tecnologia relacionada com o desenvolvimento e aplicação de ontologias, exercem um papel central junto ao problema de heterogeneidade semântica. Segundo (Bradshaw et al., 2004) e (Wache et al., 2001), uma ontologia é usada por um agente, uma aplicação ou outro recurso de informação para definir os termos utilizados e os seus significados. Se esta informação estiver disponível, a semântica pode ser compartilhada utilizando-se os mesmos conjuntos de conceitos representados pelos mesmos conjuntos de termos em qualquer situação, alcançando-se desta forma uma alta fidelidade semântica.

Qualquer modelo conceitual de banco de dados pode ser visto como uma estrutura lógica de primeira ordem, portanto um modelo semântico lógico. Por outro lado existem estruturas lógicas (de primeira ordem) que não podem ser vistas como modelos conceituais de bancos de dados mesmo quando a forma esteja representando um pedaço do mundo e o conhecimento sobre ele. Enquanto a aproximação lógica tem sido usada com sucesso na direção de representação do conhecimento, a aproximação baseada em modelo de dados parece ser mais restritiva.

Com o uso de ontologias para representação do conhecimento pode-se ter uma visão unificada de dados e conhecimento. Se alguém usa uma ontologia para descrever um modelo de dados ou uma base de conhecimento então, em ambos os casos, as estruturas lógicas dos modelos são quase idênticas. De qualquer forma, a existência de uma maneira uniforme para representar ambos os mundos é possível. Este trabalho parte desta visão, de representação uniforme, mas em direção oposta. Tecnologias bem conhecidas de bancos de dados são aplicadas para representar conhecimento.

Federação de bancos de dados é um conceito bem conhecido e uma tecnologia que tem sido usada partindo da necessidade de integração de diferentes bases de conhecimento cobrindo diferentes conteúdos do mundo dos dados. Por outro lado, a integração de ontologias tem sido provida pelo conceito usual de *merging,* alinhamento e outros relacionados. A forma para integração de ontologias é muito geral e pode levar em conta somente a estrutura das ontologias considerando uma teoria em *Description Logic* (Baader et al., 2003).

A proposta deste trabalho é fazer uso destes conceitos provenientes de banco de dados para integração de dados, como o de federação de bancos de dados, para estruturar a integração de ontologias numa federação de ontologias. Cada conceito da federação de bancos de dados vai ajudar a estruturar a federação de ontologias. Esquemas de exportação, local e global vão ter um papel fundamental na construção da federação de ontologias. A nossa intenção é promover a integração semântica de dados com uma metodologia logicamente fundamentada - uma aproximação para integração de ontologias de uma forma federada.

1.2 Organização da dissertação

Esta dissertação está organizada da seguinte forma. O capítulo 2 ressalta o problema da interoperabilidade entre os sistemas de informação. São destacados os diversos tipos de heterogeneidades que podem aparecer na tentativa de promover esta interoperabilidade. Ontologias são apresentadas como alternativa de solução e é feita uma análise das diversas situações que podem aparecer na integração de ontologias. No final do capítulo a abordagem de federação de dados é apresentada como proposta de integração dos dados.

O Capítulo 3 introduz a metodologia utilizada para viabilizar a interoperabilidade entre dados através do *merging* de ontologias. No final do capítulo, a arquitetura de federação de ontologias é apresentada, baseada na arquitetura de federação de dados que será utilizada para o estudo de caso apresentado no capítulo 4.

O capítulo 4 detalha o estudo de caso realizado, ou seja, o protótipo desenvolvido como prova de conceito da metodologia e arquitetura para a federação de ontologias apresentada no capítulo 3. São mostradas as ontologias

que serão integradas e o resultado da integração no processo de *merging*. A consistência da solução é testada.

Finalmente no capítulo 5 são apresentadas as conclusões e trabalhos futuros.