1
Introdução

1.1.
Integração de dados e ambiente de mediação

Nas últimas três décadas, o problema de integrar fontes de dados heterogêneas vem sendo estudado como um dos problemas fundamentais na área de bancos de dados [32]. Recentemente, a pesquisa sobre este problema voltou a receber grande atenção, motivada principalmente pela necessidade de integrar fontes de dados na *Web* [11].

Mais precisamente, integração de fontes de dados refere-se ao problema de combinar dados que estão armazenados em diferentes fontes, fornecendo ao usuário uma visão unificada destas fontes. As consultas são então expressas em um esquema global ou esquema mediado. Um sistema de integração de fontes de dados caracteriza-se por deixar o usuário livre da necessidade de conhecer detalhes a respeito da estruturação dos dados e da forma como as consultas são processadas nas fontes de dados envolvidas.

Um tópico crucial em integração de fontes de dados trata do mapeamento entre os elementos do esquema mediado e os elementos das fontes de dados. Além disso, como o usuário utiliza o esquema mediado para formular consultas, ele deve ser definido em uma linguagem capaz de representar a semântica do domínio de interesse [11].

Uma das principais razões que torna difícil a tarefa de integrar fontes de dados em um ambiente de mediação é a necessidade de se conhecer a semântica do ambiente mediado e de cada uma das fontes envolvidas. Incluir restrições de integridade no esquema mediado contribui para um entendimento correto sobre o que a semântica das fontes de dados do ambiente de mediação tem em comum.

Na abordagem usada nesta tese, um *ambiente de mediação* consiste de (ver Figura 1):

1. Um *esquema mediado* M, que fornece uma descrição comum das fontes de dados que participam do ambiente de mediação.
2. Para cada $i \in [1,n]$, um *esquema exportado* E_i, e um *esquema importado* I_i, e um *mapeamento local* γ_i, onde E_i indica como os dados exportados pela i-ésima fonte de dados são organizados, I_i descreve como esses dados são organizados quando importados para o ambiente de mediação (por isso o nome *esquema importado)*, e γ_i define os conceitos de I_i em termos dos conceitos de E_i.
3. Um *mapeamento de mediação* γ, que define os conceitos de M em termos dos conceitos dos esquemas importados.

Em outras palavras, M define um vocabulário comum, I_i define um subconjunto de tal vocabulário que a i-ésima fonte de dados suporta, e γ_i define como interpretar tal subconjunto em termos do vocabulário de E_i. Embora não seja convencional, a noção de esquema importado justifica-se principalmente para separar a definição dos mapeamentos em dois estágios: a definição dos mapeamentos locais e a definição do mapeamento mediado.

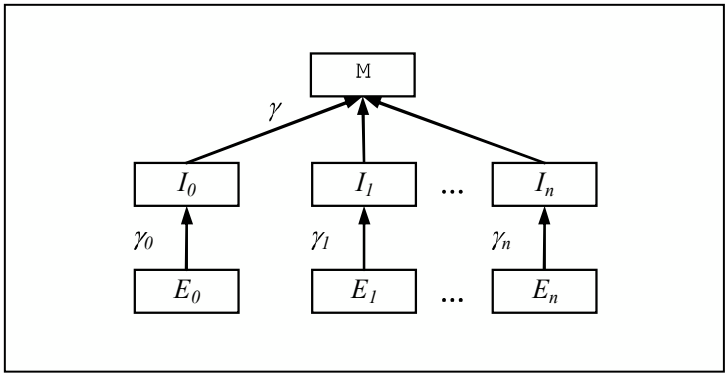

Figura 1 – Componentes de um ambiente de mediação

1.2. O problema de revisar as restrições de um esquema mediado

Esta tese trata do problema de revisar as restrições de um esquema mediado para acomodar um novo esquema exportado. De fato, inserir uma nova fonte de dados ao ambiente de mediação exige a revisão de todas as definições envolvidas sob pena de as respostas às consultas não retornarem os dados corretos. Desta forma, a modelagem das restrições é feita de forma incremental a partir da adição de cada um dos novos esquemas exportados ao esquema mediado.

Para ilustrar este problema, considere o seguinte exemplo bastante simples. O ambiente refere-se a uma loja virtual que está mediando acesso a livrarias *on-line*. Todas as livrarias que participam do ambiente de mediação exigem que todos os livros possuam pelo menos um autor. Logo, as restrições definidas no esquema mediado devem comportar esta característica. Agora, assuma que uma nova livraria é adicionada e que esta permite livros sem um autor definido. A partir deste fato, as restrições do esquema mediado devem acomodar tal modificação, ou seja, as restrições do esquema mediado devem permitir livros com zero ou mais nomes de autores para comportar dados retornados por qualquer livraria.

Em geral, modificar um esquema mediado M para acomodar um novo esquema exportado E_0 é uma tarefa difícil que envolve uma série de problemas. Primeiramente, deve-se ajustar as classes e propriedades de M para refletir as classes e propriedades de E_0. Em seguida, deve-se revisar o mapeamento de mediação e os mapeamentos locais. Por fim, deve-se modificar as restrições de M para acomodar as restrições de E_0. Note ainda que tanto os mapeamentos locais quanto o mapeamento de mediação devem continuar corretos mesmo face às novas alterações.

Em mais detalhe, considere que o esquema mediado M possua um vocabulário MV e restrições MC. Suponha que o novo esquema exportado E_0 possua um conjunto de restrições EC_0.

O processo de adicionar ao ambiente de mediação o novo esquema exportado E_0 pode ser quebrado em três etapas: *revisão de conceito*, *revisão de mapeamento* e *revisão de restrições*.

1. *Etapa de revisão de conceitos.*

   Cria o vocabulário revisado MV_r do esquema mediado, através da inclusão (quando necessário) em MV de classes e propriedades originalmente definidas em E_0, e define o esquema importado I_0 para E_0.

2. *Etapa de revisão dos mapeamentos*

   Cria o mapeamento mediado revisado γ_r, e define o mapeamento local γ_0 entre I_0 e E_0.

3. *Etapa de revisão das restrições*

   Cria o conjunto de restrições revisado MC_r de M através da definição do conjunto IC_0 de restrições de I_0, e aplica um conjunto mínimo de modificações a MC para acomodar IC_0. Esta etapa é sub-dividida em dois outros passos:

   a. *Passo da tradução das restrições:*

      Traduz o conjunto de restrições de modo que γ_0 mapeie estados de E_0 que satisfazem EC_0 em estados de I_0 que satisfazem IC_0. Intuitivamente, como resultado deste passo, a semântica de E_0 é expressa em termos de I_0.

   b. *Passo de mudança mínima de restrições*:

      Aplica um conjunto mínimo de mudanças ao conjunto de restrições de M para acomodar IC_0 de modo que todos os mapeamentos de esquema continuem corretos. O que significa, intuitivamente, harmonizar a semântica de E_0 com a semântica de todos os esquemas exportados adicionados previamente ao ambiente de mediação, capturados nas restrições de M. A questão central neste passo é como definir precisamente o que significa aplicar um conjunto mínimo de mudanças ao conjunto de restrições, e como garantir que o mapeamento continue correto.

Para resolver as Etapas 1 e 2, será usada a abordagem já proposta em [10, 30, 31]. A tese concentra-se nas Etapas 3(a) e 3(b).

1.3. Contribuições

As contribuições do Capítulo 3 dizem respeito à Etapa 3(a), ou seja, ao problema da definição do conjunto de restrições de um esquema importado a partir do conjunto de restrições do esquema exportado e do mapeamento entre os dois esquemas. Este problema e outros problemas correlatos são reduzidos diretamente ao problema de subsunção em Lógica de Descrição (LD). Neste capítulo, através da escolha cuidadosa de um dialeto de LD, mostra-se como estender o procedimento de decisão de tableau tradicional para o problema de subsunção, sem fazer uso de reduções. Mostra-se ainda como modificar o procedimento de subsunção estrutural de [21, 34] para acomodar as classes de restrições consideradas, sem alterar a complexidade do algoritmo. Tanto o procedimento do tableau quanto o procedimento de subsunção estrutural são de significância prática para o problema endereçado na Etapa 3(a). Também são relevantes para a construção de otimizadores de consulta, no processo de eliminação de subconsultas redundantes. Os resultados deste capítulo estão registrados em [29].

As contribuições do Capítulo 4 referem-se à Etapa 3(b), ou seja, ao problema de aplicar um conjunto mínimo de mudanças no conjunto de restrições do esquema mediado para acomodar as restrições de um novo esquema importado de modo que todos os mapeamentos de esquema continuem corretos. Os resultados deste capítulo estão registrados em [15, 17].

O Capítulo 4 apresenta, também, uma nova abordagem para o problema de decidir implicação lógica e de computar o *ínfimo* (*greatest lower bound*) de dois conjuntos de restrições. Em particular, lidar com restrições de cardinalidade é uma tarefa complexa e esta tese também traz contribuições para superar os problemas técnicos pertinentes a esta questão. O procedimento de decisão proposto é baseado no algoritmo de satisfatibilidade para fórmulas booleanas na forma normal conjuntiva, com no máximo dois literais por cláusula, descrito em [2]. O procedimento para calcular o ínfimo de dois conjuntos de restrições é uma conseqüência direta do procedimento de decisão e explora, essencialmente, a estrutura de um conjunto de restrições capturadas como um grafo. Estes resultados também são encontrados em [16, 17].

1.4. Trabalhos relacionados

Pesquisas em construção de esquemas mediados se concentram em técnicas de alinhamento de vocabulário, na definição dos mapeamentos de esquema, e em processamento de consultas. De fato, a maioria ignora a questão da revisão das restrições.

Técnicas de alinhamento são úteis para o processo de revisar o vocabulário do esquema mediado, um ponto que não será o foco desta tese. Euzenat e Shvaiko [20] apresentam uma abrangente pesquisa sobre alinhamento de ontologias. Rahm e Bernstein [42] fazem um estudo sobre alinhamento de esquemas, e Bernstein e Melnik [5] descrevem os requisitos para sistemas de gerenciamento de modelos que suportam o processo de alinhamento. Köpcke e Rahma [28] fazem uma análise comparativa entre onze *frameworks* para alinhamento de entidades.

Técnicas de alinhamento de esquemas podem ser classificadas em: sintáticas, semânticas ou híbridas [14]. Melnik et al. [38] e Madhavan et al. [37] apresentam técnicas sintáticas baseadas em modelagem de esquemas como grafos. Bilke e Naumann [8] propõem uma técnica semântica baseada em uma análise de instâncias duplicadas. Brauner et al. [9] adotam esta estratégia para alinhar tesauros. Wang et al. [45] descrevem uma técnica semântica baseada em sondagem dos bancos de dados envolvidos.

Partindo desta classificação, Qi e Linga [41] apresentam algoritmos para resolver as discrepâncias esquemáticas através da transformação de metadados em valores de atributos de tipos de entidade, mantendo as informações e restrições do esquema original. Zhaoa e Ramb [46] propõem um procedimento iterativo para detecção de alinhamentos tanto em nível de esquema quanto em nível de instância de fontes de dados heterogêneas.

Lonsdale et. al. [35] e Simperla [44] apresentam uma estratégia proveitosa para superar problemas de integração através do reuso de esquemas e ontologias. O uso de modelos para auxiliar o intercâmbio de esquemas, como proposto em Papott and Torlone [40], é uma estratégia similar que também pode ser usada para driblar as dificuldades de integração.

Para o mapeamento entre o esquema externo e o esquema mediado, duas abordagens básicas vêm sendo usadas [32]. A primeira chamada de *global-as-view (GAV)*, requer que o esquema mediado seja expresso em termos das fontes de dados. Mais precisamente, uma visão sobre as fontes de dados é

associada a cada elemento do esquema global, de forma que o significado do elemento é especificado em termos dos dados armazenados nas fontes de dados. Isto significa que adicionar uma nova fonte de dados pode impactar nos mapeamentos definidos previamente, que possivelmente deverão ser atualizados. Vários projetos adotam a abordagem GAV, tais como: TSIMMIS [22], IBIS [11] e INFOMIX [33].

A segunda abordagem, chamada de *local-as-view (LAV)*, requer que o esquema mediado seja especificado independentemente das fontes de dados. As fontes de dados são definidas uma a uma como visões sobre o esquema mediado [25]. O que significa que para adicionar uma nova fonte de dados é necessário apenas adicionar uma nova declaração ao mapeamento mediado. Obviamente este enfoque melhora a conservação e extensibilidade dos sistemas [6]. Agora [36], StyX [1] e Picsel [24] são exemplos de sistemas que usam a abordagem LAV.

Mapeamentos podem também ser classificados de acordo com sua precisão em *consistentes, completos* e *exatos* [11, 32]. Seja V uma visão associada a um elemento E do esquema mediado. Na abordagem GAV, V é dita *parcial* quando todos os dados fornecidos por V satisfazem E, mas podem existir dados adicionais satisfazendo E que V não provê. A visão V é *completa* uma vez que nem todos os dados fornecidos por V precisam satisfazer E, mas todos os dados que satisfazem E são fornecidos por V. Finalmente V é dita *exata*, quando todos os dados fornecidos por V satisfazem E, e todos os que satisfazem E são fornecidos por V [11].

Rull et al. [43] apresentam uma abordagem para validar mapeamentos de esquema que permite ao projetista do mapeamento perguntar se eles possuem certas propriedades desejáveis.

O enfoque que será usado para definir o ambiente de mediação usado neste trabalho é semelhante à visão consistente. No entanto, as restrições devem ser incluídas no esquema mediado para capturar a semântica que as fontes de dados têm em comum, ao contrário da maioria das propostas baseadas no conceito de visão exata, que assumem que o esquema mediado não possui restrições, como observado em [32].

Calì et al. [11] argumentam que as restrições do esquema mediado devem ser levadas em consideração durante o processamento da consulta e que a linguagem de definição do esquema deve incorporar mecanismos flexíveis e poderosos para restrições de integridade. Os autores também afirmam que,

quando um esquema mediado possui restrições, a semântica do sistema de integração de dados é mais bem descrita em termos de um conjunto de bancos de dados, e que o processamento de consultas deve ser baseado na noção de consultas a bancos de dados incompletos.

Calvanese et al. [13] introduzem um *framework* de Lógica de Descrição, similar ao que é usado nesta tese, para tratar os problemas de integração de esquemas e resposta a consultas. Atzeni et al. [3] estudam o problema de re-escrever um esquema de um modelo para outro, mas não tocam no problema mais complexo: o de gerar um novo conjunto de restrições que generaliza um par de conjuntos de restrições de esquemas diferentes, que é tratado aqui. Hick and Hainauta [27] mostram como as mudanças de requisitos são propagadas para esquemas de bancos de dados, aos dados e aos programas através de uma estratégia geral.

Hartmanna et al. [26] aplicam técnicas de lógica proposicional para oferecer suporte a decisão para especificar dependências booleanas e multivaloradas.

Outro aspecto também será examinado nesta tese, é o problema de subsunção em Lógica de Descrição (LD) que se refere à questão de decidir se uma descrição de conceito sempre denota um subconjunto do conjunto denotado por uma outra descrição do conceito. O problema de subsunção é decidível para dialetos expressivos de LD, mas geralmente pertence às classes de complexidade dos problemas difíceis [4], especialmente na presença de axiomas (ou restrições) [19]. Para certos dialetos de LD, existem procedimentos de decisão polinomial para o problema da subsunção que exploram a estrutura das descrições de conceito e que são, por esta razão, chamados de *procedimentos de subsunção estrutural* [21]. No entanto, tais procedimentos não levam em consideração axiomas. Além disso, as reduções propostas para codificar os axiomas fazem uso dos dialetos para os quais o problema subsunção é difícil [19].

1.5. Organização da tese

O Capítulo 2 introduz o formalismo de Lógica de Descrição necessário para a tese, além de definir formalmente o ambiente de mediação utilizado. O Capítulo 3 aborda o problema da definição do conjunto de restrições de um esquema importado a partir do conjunto de restrições do esquema exportado e

do mapeamento entre os dois esquemas, bem como problemas correlatos. O Capítulo 4 mostra, passo a passo, como as restrições de um ambiente de mediação são construídas, endereça o problema de revisão mínima das restrições e o cálculo do ínfimo de dois conjuntos de restrições. Finalmente, o Capítulo 5 contém as conclusões e sugestões para trabalhos futuros.