1
Introdução

1.1
Motivação

Um *esquema conceitual* é uma descrição de alto nível de como organizar os conceitos armazenados em um banco de dados. Um esquema conceitual descreve os conceitos de acordo com o modelo de dados utilizado pelo Sistema Gerenciador de Banco de Dados (SGBD). Por exemplo, em SGBDs relacionais, o esquema conceitual contém descrições de relações (Ramakrishnan e Gehrke, 2000).

Catálogos são bancos de dados descritos com esquemas conceituais simples que utilizam um *esquema de classificação* acoplado para classificar seus objetos em categorias pré-definidas. Um esquema de classificação fornece informação descritiva para a organização de objetos em grupos baseado em características que estes objetos possuem em comum. Tipos de esquemas de classificação são: palavras-chave, tesauros e taxonomias (ISO/IEC 11179-2, 2005). Alternativamente, podemos ver um esquema de classificação como uma forma de organizar o domínio de valores utilizado em um atributo enumerado. Devido a particularidades nos requisitos de projetos de software e nas preferências pessoais dos projetistas de banco de dados, até mesmo esquemas conceituais e de classificação para um mesmo domínio de aplicação possuem diferentes características e estruturas. As abordagens de alinhamento descritas nesta tese são aplicáveis à catálogos com esquemas conceituais simples e tesauros acoplados.

Com a difusão da Internet, é razoável pensarmos em arquiteturas de sistemas que mantém e utilizam dados distribuídos em larga escala. Com isso, torna-se essencial que as organizações que possuem unidades de negócio distribuídas possam manipular de forma integrada todos os dados disponíveis em suas unidades. O suporte a acesso integrado impõe desafios importantes, tais como lidar com a heterogeneidade semântica entre os esquemas das fontes de dados a serem acessadas.

A *heterogeneidade semântica* caracteriza-se pela ocorrência de conflitos semânticos entre os esquemas das fontes dados. Ela ocorre quando existe discordância a respeito do significado, interpretação ou uso pretendido, por exemplo, de um determinado atributo ou tabela de um esquema conceitual, ou um termo de um esquema de classificação.

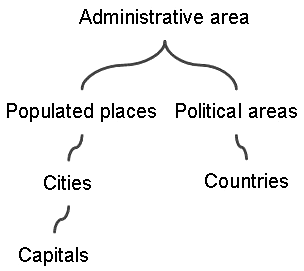
Sem dúvida, este é um dos principais problemas a serem solucionados para que fontes de dados distintas de um mesmo domínio de aplicação possam interoperar. Os nomes das entidades utilizadas em um esquema carregam uma semântica implícita, representando um conceito num determinado contexto. A interpretação desses nomes não necessariamente irá coincidir quando realizada por diferentes pessoas ou grupo, caracterizando um conflito semântico. Além desse, podem ocorrer conflitos devido a nomes iguais que representam diferentes conceitos (homônimos) ou quando nomes diferentes representam o mesmo conceito (sinônimos).

Outra situação ocorre quando dois atributos possuem a mesma semântica e apresentam heterogeneidade em suas representações. Como exemplo, temos “preço” como nome de um atributo de uma fonte de dados A e de um atributo de uma fonte de dados B. Na fonte A, o preço está representado em “dólares americanos”, enquanto na fonte B, em “reais”. Outro exemplo de heterogeneidade na representação dos atributos é identificado quando as fontes de dados utilizam esquemas de classificação diferentes. Como exemplo, considere um atributo “tipo de feição” no domínio de dados geográficos. Na fonte *A*, o tipo de feição é representado utilizando termos definidos no *ADL Feature Type Thesaurus* (ADL FTT), ilustrado na Figura 1 (a), enquanto que na fonte B é representado como uma lista de códigos, ilustrada na Figura 1 (b). Note que os termos do *GeoNames* são uma lista de códigos sem estrutura aparente, enquanto que os termos do ADL FTT são conceitos organizados em uma estrutura hierárquica. Neste contexto, o termo *Administrative area* da ADL FTT parece ter o mesmo significado que os termos ADM1, ADM2 ou ADMD do *Geonames* e o termo *Populated Places* parece ter o mesmo significado que os termos PPL, PPLA, PPLC, PPLL ou PPLX.

A Figura 2 ilustra a heterogeneidade semântica entre esquemas conceituais *S* e *T*. Para exemplificar foram utilizados apenas os elementos similares *S.Client* e *T.Customer*, que armazenam informações de consumidores de duas lojas *online* distintas. A Figura 2 (a) ilustra os alinhamentos corretos entre os atributos dos esquemas, entre eles: *id* ↔ *cNumber*, *first* ↔ *cName*, *last* ↔ *cNumber*, *address* ↔ *cEmail*, *home* ↔ *cAddress* e *phone* ↔ *cPhone*. Note

que os elementos *first* e *last* do esquema *S* representam o primeiro nome e o sobrenome do consumidor, respectivamente. Ambos mapeiam para o mesmo atributo *cName* do esquema *T*, que representa o nome completo do consumidor. Em contrapartida, a Figura 2 (b) ilustra um exemplo de alinhamento utilizando pistas sintática para criar as correspondências. Nesta abordagem, observam-se os alinhamentos *address* ↔ *cAddress* e *phone* ↔ *cPhone*, porém com um erro semântico pois o atributo *address* do esquema *S* representa o endereço eletrônico, enquanto o atributo *cAddress* do esquema *T* representa o endereço residencial do cliente.

Para lidar com o problema de heterogeneidade semântica, o correto *alinhamento de esquemas* torna-se uma questão fundamental em diversas aplicações de bancos de dados, tais como mediação de consultas (discutido na seção 2.1.1) e transformação de dados no contexto de *data warehousing* (discutido na seção 2.1.2).



(a)

Fragmento *ADL Feature Type Thesaurus* [1]

ADM1
ADM2
ADMD
PPL
PPLA
PPLC
PPLL
PPLX

(b)

Fragmento *GeoNames Feature Codes* [2]

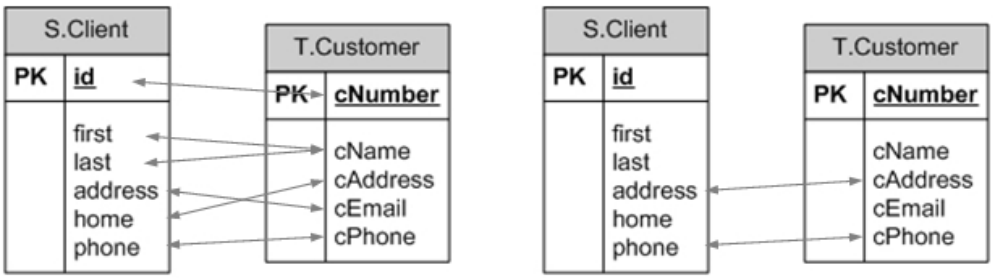
Figura 1 – Exemplo de heterogeneidade semântica em tesauros.

O alinhamento de esquemas é considerado uma operação fundamental na manipulação de esquemas de bancos de dados. Dados dois esquemas como entrada, o processo de alinhamento produz um mapeamento entre os elementos desses esquemas que correspondem semanticamente um ao outro (Rahm & Bernstein, 2001). Então, alinhar um esquema *S* com um esquema *T* significa encontrar um mapeamento μ dos conceitos de *S* nos conceitos de *T* de maneira que, se $t = \mu(s)$, então *s* e *t* possuem o mesmo significado (Figura 3). A operação de alinhamento pode ser baseado somente nos esquemas, utilizando os elementos do esquema como recursos básicos de entrada, ou pode ser baseado

[1] http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/ver070302/

[2] http://www.geonames.org/export/codes.html

em instâncias, utilizando instâncias das fontes de dados como recursos básicos de entrada. Opcionalmente, a operação de alinhamento pode contar com o auxílio de recursos adicionais como *thesauri*, ontologias ou um corpus com exemplos de mapeamentos anteriormente encontrados.



(a) Alinhamento correto (b) Alinhamento utilizando pistas sintáticas

Figura 2 – Exemplo de heterogeneidade semântica em esquemas conceituais.

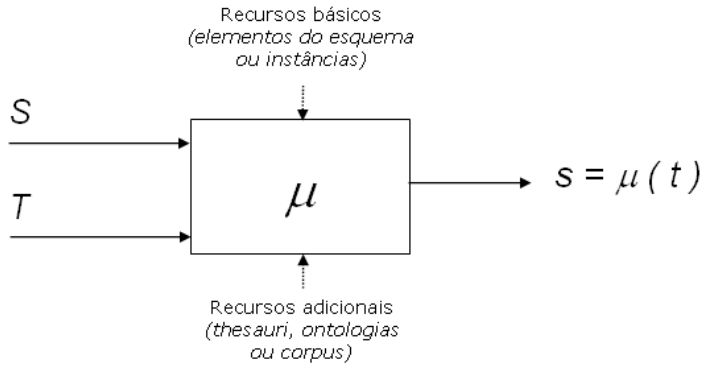


Figura 3 – A operação de alinhamento.

Atualmente, a tarefa de alinhamento de esquemas é feita manualmente, eventualmente suportada por alguma ferramenta com uma interface gráfica que facilita o trabalho do projetista. Porém, o processo de especificar os mapeamentos entre esquemas é tedioso e dispendioso. Além disso, o nível de esforço é linear com relação ao número de mapeamentos a serem realizados, o que pode se tornar um fator limitante quando o número de fontes de dados a integrar é substancial, como é o caso de aplicações na Web. Bernstein & Melnik (2007) indicam que os departamentos de TI gastam em torno de 40% do tempo de trabalho alinhando esquemas de bancos de dados.

Em (Madhavan et al., 2007) são expostos alguns desafios impostos pelo problema de heterogeneidade na escala da Web. Grande parte do conteúdo da Web está armazenado em bancos de dados e é disponibilizado via formulários

HTML utilizando esquemas de diferentes estruturas e semântica. É estimado que estes formulários podem representar 2.5% do conteúdo total da Web. Atualmente, a existência desta vasta coleção de dados estruturados e heterogêneos impõe um dos maiores desafios para as ferramentas de busca.

Por estes motivos, o alinhamento automático ou mesmo semi-automático de esquemas é uma área de pesquisa que ainda requer amadurecimento das técnicas, de forma a facilitar e acelerar o trabalho de integração de bancos de dados. De fato, existem abordagens de alinhamento de esquemas que usam estratégias semi-automáticas utilizando comparações sintáticas entre os elementos dos esquemas. Entretanto, estas abordagens são suscetíveis a erros devido à falta de formalização da semântica embutida nos domínios dos esquemas envolvidos, não resolvendo o problema de heterogeneidade semântica. Nesta tese, são investigados métodos semânticos para alinhamento de esquemas.

1.2 Solução proposta

Esta tese investiga técnicas de alinhamento para esquemas de classificação, em particular tesauros, e esquemas conceituais simples utilizando instâncias. As abordagens apresentadas são classificadas em dois tipos: *adaptativas* e *a priori*. Neste contexto, *adaptativa* refere-se a descoberta e adaptação dos mapeamentos de forma incremental. Já *a priori* refere-se à necessidade de descoberta dos mapeamentos antes de permitir o acesso às fontes.

As abordagens para alinhamento de tesauros utilizam instâncias equivalentes como evidências dos mapeamentos entre os termos dos tesauros de fontes distintas em um mesmo domínio de aplicação. Para isso, deve ser possível detectar instâncias equivalentes utilizando valores de atributos que sirvam como identificadores únicos dos objetos. Nestas abordagens, o problema de heterogeneidade dos esquemas conceituais das fontes utilizadas deve ter sido resolvido, manualmente ou através de uma técnica de alinhamento de esquemas conceituais. Nesta tese, o alinhamento de esquemas conceituais também é discutido.

As abordagens para alinhamento de esquemas conceituais propostas nesta tese utilizam uma técnica que consiste em submeter consultas às fontes de dados utilizando valores de atributos de um conjunto de instâncias. De posse

dos resultados destas consultas, são computados os valores co-ocorrentes que computam evidências para os mapeamentos dos atributos dos esquemas conceituais a serem integrados.

As abordagens adaptativas apresentadas nesta tese foram criadas para se aproveitar da interação do usuário com o mediador de forma a evitar a coleta prévia de coleções de instâncias equivalentes (abordagem *a priori* de alinhamento de tesauros) e a definição prévia de esquemas globais e instâncias globais (abordagem *a priori* de alinhamento de esquemas). As abordagens adaptativas utilizam instâncias retornadas nas respostas de consultas de usuários como fonte para levantar as evidências para os alinhamentos.

Por fim, as abordagens propostas nesta tese foram validadas através de experimentos. Todos os experimentos, tanto das abordagens de alinhamento de tesauros quanto das abordagens de alinhamento de esquemas conceituais, adaptativas e *a priori*, utilizaram fontes de dados geográficos disponíveis na Web.

Os resultados sobre alinhamentos *a priori* de tesauri são apresentados em Brauner et al. (2007a), e a abordagem adaptativa em Brauner et al. (2006). Os resultados sobre alinhamentos *a priori* de esquemas conceituais são apresentados em Brauner et al. (2007b), e a abordagem adaptativa em Brauner et al. (2008).

1.3 Organização da tese

O Capítulo 2 introduz os conceitos básicos que motivam o desenvolvimento desta tese. Os Capítulos 3 e 4 introduzem as abordagens de alinhamento de tesauros e esquemas, respectivamente, bem como os testes realizados e validação das abordagens. O Capítulo 5 resume algumas técnicas existentes para integração que inspiraram a execução deste trabalho ou que possuem algum tipo de similaridade com as abordagens introduzidas nesta tese. Finalmente, o Capítulo 6 contém as conclusões e sugestões para trabalhos futuros.