6
Conclusões e trabalhos futuros

Esta tese investiga, documenta e experimenta abordagens baseadas em instâncias para alinhamento de esquemas de classificação (tesauros) e esquemas conceituais. Os processos de alinhamento apresentados em cada uma das abordagens apresentadas nesta tese podem ser aplicados para resolver as questões relacionadas à mediação de consultas e ao procedimento de transformação de dados.

As abordagens discutidas nesta tese são classificadas em *a priori* e adaptativas. Portanto, podem ser aplicadas tanto a problemas onde existe a necessidade dos mapeamentos serem descobertos *a priori* quanto em ocasiões onde os mapeamentos podem ser descobertos de forma incremental.

Para alinhar tesauros, uma alternativa bastante utilizada são os algoritmos de similaridade sintática entre os conceitos. Porém, esta abordagem pode ser inútil, por exemplo, quando um dos tesauros utiliza palavras para designar seus conceitos e o outro adota siglas. Uma alternativa ao uso de conceitos seria o uso das definições dos conceitos. Esta abordagem varre os textos das definições em busca de palavras similares entre eles. Assim, quando detectada a similaridade entre as definições, os conceitos serão alinhados. Porém, a maioria das definições ocupam, no máximo, três linhas, caracterizando-se textos muito pequenos para os algoritmos de mineração de texto. Uma fragilidade desta abordagem é a análise da estrutura das frases, pois conceitos semanticamente inversos podem ser alinhados por possuírem a maioria das palavras em comum. Neste contexto, a solução para alinhamento de tesauros adotada nesta tese mostrou-se uma alternativa eficaz para alinhamentos baseados em técnicas puramente sintáticas. Nesta tese, foi apresentado o uso de instâncias equivalentes para o alinhamento *a priori* dos termos dos tesauros. As instâncias são previamente coletadas e analisadas, gerando evidências para os alinhamentos.

A tese descreve ainda uma variante adaptativa para esta abordagem para alinhamento de tesauros, em que a descoberta dos alinhamentos é feita de forma incremental. A abordagem adaptativa para alinhamento de tesauros usa

as instâncias retornadas nas respostas de consultas de usuários como fonte para levantar as evidências para os alinhamentos.

Em seguida, a tese aplica estratégias baseadas em instâncias para alinhamento de esquemas conceituais. Assim, surgiu a abordagem *a priori* para alinhamento de esquemas conceituais, baseada no trabalho de Wang et al. (2004). Esta abordagem baseia-se na definição de um esquema global para o domínio das fontes que serão integradas e na coleta de instâncias globais para servirem como iscas para descobrir os alinhamentos. As instâncias globais são submetidas como palavras chave das consultas às fontes a serem integradas, recuperando, assim, instâncias equivalentes das fontes a serem integradas. Estas instâncias equivalentes servem como evidências dos alinhamentos dos elementos dos esquemas das fontes com o esquema global previamente definido.

Diferente da abordagem proposta por Wang et al. (2004), que usa fontes disponíveis na Web através de formulários HTML, a abordagem proposta nesta tese usa fontes disponíveis via Web Services. A abordagem proposta poupa todo o trabalho de extração dos elementos dos esquemas e das respostas das consultas a partir de informações em HTML.

Porém, o processo de criação do esquema global e de levantamento das instâncias de referência pode ser trabalhoso. Por isso, foi proposta a abordagem adaptativa para alinhamento de esquemas conceituais. Nesta abordagem, não há a necessidade de criação do esquema global e definição das instâncias globais. A abordagem adaptativa usa as instâncias equivalentes retornadas de consultas dos usuários como evidências para os alinhamentos dos elementos dos esquemas.

Como prova de conceito, as abordagens apresentadas foram validadas através de experimentos utilizando fontes de dados geográficos disponíveis na Web. Os experimentos mostraram bons índices de cobertura e precisão das abordagens. Outros experimentos utilizando o domínio de livros podem ser vistos em (Brauner et al., 2008) e (Gazola 2008). Além disso, Gazola (2008) apresenta uma implementação da arquitetura para abordagem de alinhamento adaptativo de tesauros e uma implementação da arquitetura para abordagem de alinhamento adaptativo de esquemas.

Seguem-se as principais vantagens das abordagens propostas nesta tese.

Uso de instâncias nos alinhamentos

O uso de instâncias para descoberta dos alinhamentos é uma contribuição significativa desta tese. Diversas técnicas de alinhamento utilizavam, e muitas ainda utilizam, similaridades sintáticas para alinhar esquemas de classificação (tesauros) e esquemas conceituais. Porém, como os esquemas são criados por diferentes grupos e organizações, até mesmo esquemas num mesmo domínio de aplicação podem possuir conceitos sintaticamente similares mas semanticamente diferentes, criando falsos alinhamentos.

Alinhamento automático de esquemas

Mediadores que implementam uma das abordagens de alinhamento apresentadas nesta tese podem ser utilizados em domínios que possuam tesauros muito grandes, ou esquemas exportados com grande número de atributos, onde seria inviável ou extremamente trabalhosa a definição dos mapeamentos de forma manual.

Evita retrabalho em domínios variáveis

A característica adaptativa das abordagens apresentadas nesta tese demonstrou ser importante quando aliada à tarefa de alinhamento de tesauros e esquemas conceituais em domínios variáveis. Mediadores adaptativos podem ser utilizados em domínios que variam com muita freqüência os tesauros ou os esquemas conceituais das suas fontes de dados. Desta forma, não há a necessidade de retrabalho do engenheiro de integração em redefinir os mapeamentos a cada alteração nos esquemas.

Mediadores como consumidores de Web Services

Atualmente, a adoção de Web Services por empresas e organizações que desejam disponibilizar dados na Web é crescente. Portanto, as abordagens apresentadas nesta tese viabilizam que os mediadores atuem como consumidores de Web Services que forneçam uma interface de consulta a fontes de dados.

O uso de mediadores como consumidores de Web services torna mais fácil a captura automática dos elementos pertencentes aos esquemas conceituais das fontes a serem integradas. Na abordagem de alinhamento de esquemas conceituais o mediador se comunica com o banco de dados via troca de mensagens entre Web Services. Neste contexto, a sintaxe correta das mensagens contendo os elementos utilizados na consulta e os elementos do

esquema de exportação é definida através da descrição do serviço, geralmente em XML.

O uso de Web Services facilita também o cadastramento de novas fontes no mediador. Por utilizar Web Services, as arquiteturas propostas nesta tese podem ser implementadas provendo serviços de registro simples contendo apenas o endereço do serviço e a operação a ser utilizada. Desta forma, novas fontes podem ser automaticamente cadastradas no mediador sem a necessidade de um engenheiro de integração definir todos os elementos dos esquemas *a priori*. Devido à instabilidade que a Web pode proporcionar, com Web Services ora em funcionamento, ora não, este torna-se um requisito essencial para novas aplicações na Web.

Abaixo são apresentados alguns tópicos elegíveis para trabalhos futuros tanto para extensão das abordagens apresentadas nesta tese quanto para elucidar os tópicos em aberto viabilizando dar continuidade às pesquisas envolvendo alinhamento de esquemas utilizando instâncias.

Inclusão de diversas fontes de dados

As abordagens propostas nesta tese foram testadas utilizando-se apenas duas fontes de dados. Como trabalho futuro sugere-se a inclusão de mais fontes de dados para analisar a performance dos processos de alinhamento. Desta forma, será possível verificar e sugerir possíveis extensões, tanto a nível arquitetural quanto com relação aos próprios processos de alinhamento.

Transformação de valores dos atributos

Para aumentar a precisão dos alinhamentos, sugere-se a realização de transformações de valores dos atributos. A transformação pode ser feita por meio de algoritmos de lematização, remoção de stop-words e por procedimentos de padronização. Um algoritmo de lematização (*stemming*) realiza a remoção de sufixos e prefixos das palavras, reduzindo a palavra a um radical (*stem*). A remoção de Stop-words elimina palavras consideradas irrelevantes, tais como artigos, pronomes, interjeições, advérbios, preposições, etc. Essas palavras, normalmente, são eliminadas porque não traduzem a essência do termo e, por isso têm baixo valor semântico.

Neste contexto, outro exemplo é a padronização nos valores dos dados. Suponha um atributo de latitude da fonte A utilizando um Sistema de Referência Espacial (*Spatial Reference System* - SRS), por exemplo representado em

graus, e outro atributo latitude da fonte B utilizando um SRS diferente, representado por um valor decimal. Neste contexto, seria praticamente impossível identificar as duplicatas em virtude dos diferentes formatos. Porém este problema poderia ser contornado utilizando um procedimento de transformação de forma a padronizar ambos os valores. Outro exemplo são os formatos das datas que podem ser: DD/MM/AAAA, DD/MM/AA, MM/DD/AAAA, etc. Neste contexto também instâncias com nome “Arroio Pelotas” e “Pelotas, Arroio” podem ser alinhadas se passarem por um processo de transformação.

Filtragem de atributos

Algumas análises realizadas durante os experimentos podem ser previamente descartadas se forem corretamente filtradas. Um exemplo disso é comparando os tipos de dados e criando restrições a respeito sobre quais pares de atributos devem ser comparados e quais pares devem ser descartados. Datas, por exemplo, devem ser comparadas apenas com atributos do tipo *string* ou *date*. Desta forma, o tempo consumido pelo algoritmo de comparação seria reduzido devido à exclusão de pares de atributos que possivelmente gerariam zero co-ocorrências.

Identificação automática de duplicatas

As abordagens de alinhamento de tesauros propostas nesta tese, consideram que é possível identificarmos duplicatas através da identificação prévia de chaves. Porém, como trabalho futuro, sugere-se a implementação de uma técnica para identificação automática de duplicadas partindo, por exemplo, da análise das tuplas completas e não apenas atributos isolados.

Uso de corpus

Suponha que utilizando as abordagens apresentadas nesta tese, foram descobertos mapeamentos entre uma fonte de dados A e uma fonte de dados B. Se uma nova fonte de dados C for incluída, os mapeamentos podem ser realizados apenas entre B e C, e esta informação pode ser utilizada para inferir alguns mapeamentos entre A e C por transitividade. Assim, apenas precisarão ser executados os alinhamentos entre A e C dos atributos de A que não tinham mapeamentos com os atributos de B.

Geração de esquema conceitual global

Com base nos mapeamentos descobertos pelas abordagens de alinhamento de esquemas conceituais apresentadas nesta tese é possível gerar um *esquema conceitual global*. Em (Brauner et al., 2008) são ilustrados alguns exemplos simples de esquemas globais intuitivamente derivados dos mapeamentos obtidos através dos experimentos.

Além disso, o esquema conceitual global derivado a partir dos mapeamentos gerados pela abordagem *adaptativa* pode ser considerado *extensível*. Isto porque ele poderá ser acrescido de novos atributos a medida que novos mapeamentos são descobertos.

Considerar a co-ocorrência de valores de dois ou mais atributos

Alguns valores de atributos podem ser avaliados em conjunto no caso da base de dados não possuir um identificador de domínio. Suponha a instância que representa a cidade do “Rio de Janeiro”. Se analisados os valores de dois atributos em conjunto, por exemplo os atributos nome e tipo, com os valores “Rio de Janeiro” e “Populated place”, respectivamente, seria possível distingui-la da instância que representa o Estado do Rio de Janeiro, cujos atributos nome e tipo recebem os valores “Rio de Janeiro” e “Administrative Area”, respectivamente. Desta forma, seria possível contar a co-ocorrência não apenas dos valores dos atributos isoladamente, mas sim, a co-ocorrência da instância como um todo.

Heterogeneidade estrutural

As abordagens discutidas nesta tese abordam o problema de heterogeneidade semântica. O problema de heterogeneidade estrutural é outra questão de extrema relevância que pode ser atacada pela abordagem de alinhamento baseadas em instâncias. Sugere-se que, com base nos mapeamentos descobertos pelas abordagens apresentadas nesta tese, os mapeamentos entre tributos de esquemas conceituais e entre conceitos de tesauros possam ser generalizados ou especializados para gerar mapeamentos estruturais. A mesma idéia pode ser aplicada à esquemas complexos, com relacionamentos entre tabelas ao invés de usar uma tabela simples.

Alinhamento de operações de Web Services

Utilizando a abordagem de alinhamento de esquemas para esquemas de exportação de Web Services, pode ser possível inferir o alinhamento das operações. Por exemplo, se todos os atributos de entrada e saída da operação *a*

do Web Service *A* casam com os atributos de entrada e saída da operação *b* do Web Service *B*, há indícios que *a* alinha com *b*, o que significa que, na ausência de *A*, os serviços que consomem a operação *a* podem substituí-la pela operação *b*.