5
Conclusões e trabalhos futuros

Nesta dissertação, abordamos o problema de alinhamento de catálogos heterogêneos, uma operação fundamental para lidar com aplicações que requerem interoperabilidade ou integração de dados, como é o caso de mediadores. Para endereçar o problema, discutimos estratégias baseadas em instâncias para realizar o alinhamento dos esquemas e tesauros de catálogos.

Descrevemos e implementamos uma técnica para a geração automática de esquemas globais a partir de um conjunto de catálogos heterogêneos, bastante adequada para uso em mediadores. Na implementação do protótipo apresentado, consideramos o uso de instâncias resultantes de consultas feitas por usuários para realizar o alinhamento dos esquemas das diversas fontes e produzir um esquema mediado adequado ao domínio de aplicação em questão.

Uma publicação inicial dessa idéia foi feita em Brauner et al. (2008). A abordagem assume que, além dos catálogos pertencerem a um mesmo domínio, é necessário que os mesmos possuam uma parcela de objetos em comum. Se, por exemplo, nos testes apresentados no capítulo 4, considerássemos catálogos de objetos geográficos de regiões diferentes ou livrarias especializadas em assuntos distintos, a técnica utilizada teria sua efetividade reduzida. Heterogeneidade semântica no nível de instâncias e a aleatoriedade da seleção do catálogo e das instâncias para sondagem também pode afetar a qualidade dos alinhamentos descobertos. Quanto maior o número de catálogos registrados no mediador, menos propensa a aplicação estará a esses tipos de interferência.

Apresentamos o *CatalogMatcher*, uma infra-estrutura de software para alinhamento de catálogos heterogêneos. A infra-estrutura contém componentes que implementam estratégias de alinhamento de catálogos heterogêneos utilizando abordagens baseadas em instâncias. Para validação do *CatalogMatcher*, implementamos uma aplicação de mediação de consultas a catálogos, aplicando-a para testes em cenários no domínio de objetos geográficos e de livrarias virtuais.

A seguir, enumeramos alguns pontos que podem ser considerados no desenvolvimento de trabalhos futuros:

- **Consideração do problema da heterogeneidade estrutural**. Por utilizarmos catálogos neste trabalho, nos abstraímos do problema da heterogeneidade estrutural, que traz a necessidade de técnicas mais elaboradas para lidar com estruturas mais complexas.
- **Consideração de alinhamentos complexos**. As técnicas de alinhamento poderiam ser estendidas para incorporar alinhamentos complexos (i.e.,1:n, m:1 e m:n), em vez de apenas alinhamentos simples (i.e., 1:1).
- **Inclusão de taxas de alinhamento de esquemas**. Diferentemente da técnica que utilizamos para alinhamento de tesauros, não estabelecemos taxas de alinhamento entre os atributos alinhados de dois esquemas. Essa informação poderia ser incorporada, utilizando-se algumas métricas para estimar a confiança dos alinhamentos encontrados. Por exemplo, quanto mais instâncias são utilizadas para a descoberta de alinhamentos, maior poderia ser a taxa de alinhamento dos atributos.
- **Inclusão de técnicas de validação**. O *CatalogMatcher* poderia executar algum processo com o objetivo de validar os alinhamentos descobertos num determinado instante.
- **Consideração de alinhamento de tesauros utilizando palavras-chave**. No protótipo apresentado em Gazola et al. (2007), além do alinhamento de termos de tesauros por consultas baseadas em classificação de objetos, também implementamos o alinhamento dos termos dos tesauros considerando instâncias retornadas por consultas por palavras-chave. No entanto, existem problemas na abordagem desenvolvida, pois é necessário que o número de instâncias utilizadas tenha algum fator de impacto na fórmula para o cálculo das taxas.
- **Consideração de algoritmos para detecção de duplicatas**. Atualmente, existe a necessidade de que sejam manualmente especificados os atributos que identificam universalmente os objetos armazenados. A técnica de detecção de duplicatas desenvolvida por Bilke & Naumann (2005), por exemplo, poderia ser implementada e incorporada à infra-estrutura para possibilitar a descoberta automática dessa informação.

Apesar dos bons resultados obtidos com técnicas baseadas em instâncias, a tarefa de alinhar esquemas e tesauros é uma tarefa inerentemente difícil de ser automatizada por completo, pois boa parte da semântica exata dos dados só é conhecida de maneira completa pelos projetistas do esquema, não sendo capturada no esquema em si (Madhavan et al., 2005). Um dos motivos para que isso aconteça está na falta de expressividade dos modelos de dados existentes, agravada por projetos deficientes e com documentação escassa (Madhavan et al., 2005). Avanços nas pesquisas com ontologias podem melhorar esse cenário (Necib & Freytag, 2005), além da combinação de diversas abordagens (Rahm & Bernstein, 2001).