7 Conclusão

Aplicações web que obtêm dados de diferentes (*Mashups*) estão ganhando importância na internet. Um importante requisito dessas aplicações é a habilidade de assegurar a qualidade dos dados, provendo uma solução viável para o problema de identificação de duplicatas.

Não há uma solução única que atenda a todos os domínios. Mesmo com uma estratégia geral como guia, ainda é preciso entender como diferenciar os objetos deste domínio para escolher técnicas e parâmetros adequados.

7.1 Contribuições

A maioria das soluções existentes utilizam algoritmos de aprendizado de máquina, que são dependentes de uma base de treinamento pré-existentes. Soluções deste tipo não são adequadas quando utilizamos fontes de dados da Deep Web. Nesta dissertação propomos uma solução para o enriquecimento de bases de dados através de consultas em fontes de dados na Deep Web. Esta solução pode ser decomposta em duas contribuições principais. A primeira é uma estratégia para a busca (rastreamento) de informações na Deep Web orientado a duplicatas, cujo objetivo é obter resultados precisos de consultas construídas a partir de um conjunto conhecido de objetos e seus atributos. A segunda contribuição, é uma estratégia para a detecção de duplicatas em resultados de consultas realizadas sobre fontes de dados da Deep Web.

Nossa abordagem não requer uma base de treinamento previamente definida, e utiliza um classificador baseado no Vector Space Model (VSM) em combinação com funções de cálculo de similaridade para prover uma solução viável. Em contrapartida, nossa solução requer um esforço adicional para escrever parsers de dados, filtros e wrappers para normalizar dados minerados da web para um formato consistente com nosso esquema global.

De modo a validar a solução proposta, construímos um framework que implementa as estratégias de busca e incorporação de dados. Ilustramos a utilização do framework através da instanciação do mesmo em uma aplicação de comércio eletrônico voltada para o domínio de vinhos. Os detalhes da implementação e as lições aprendidas no processo foram descritas cuidadosamente, para que possam ser úteis aqueles que pretendem replicar os experimentos aqui realizados. Os resultados, quando comparados com outras técnicas, são promissores.

PUC-Rio - Certificação Digital Nº 0821369/CA

7.2 Trabalhos Futuros

Dados os resultados obtidos com a estratégia proposta, é possível, como trabalho futuro, instanciar e investigar o framework utilizando outros domínios e investigar o comportamento do framework com um conjunto maior de teste. Espelhando-se em trabalhos de referência como o de Su [36], pode-se experimentar incluir um mecanismo de auto-calibração do classificador.

Alguns experimentos superficiais foram realizados utilizando fontes de dados do tipo OWL com o domínio de livros. Os primeiros resultados demonstraram que utilizar a mesma solução que foi dada ao domínio dos vinhos para este domínio não foi adequado. É necessária uma investigação acerca deste domínio para a criação de uma nova instância do framework.

PUC-Rio - Certificação Digital Nº 0821369/CA