1 Introdução

Estudos recentes demonstram que uma fração significativa dos recursos disponíveis através da Web não podem ser alcançados através de links [4]. Páginas geradas dinamicamente são exemplos típicos destes recursos. Este fato impossibilita que mecanismos de busca tradicionais como Google, Yahoo e Bing possam indexar os conteúdos destas páginas e, conseqüentemente, apresentá-las entre seus resultados. Esses recursos ganharam o nome de *Hidden Web* ou *Deep Web*. Estima-se que menos de 40% das páginas Web são acessíveis através de mecanismos de busca [17]. Estima-se, ainda, que estes dados apresentam grande qualidade, são bem estruturados e crescem em uma proporção bem maior que a Web.

Os sites de e-commerce são os maiores produtores de conteúdo da *Hidden Web.* Segundo He, 94% destes sites que possuem conteúdo dinâmico, e fazem acesso a um banco de dados com apenas três links de navegação [15]. As páginas geradas dinamicamente por este tipo de sites são normalmente obtidas como respostas consultas realizadas através de formulários de pesquisa. Podemos, desta forma, encarar a *Deep Web* como o maior banco de dados (*Very Large Data Base* - VLDB) existente.

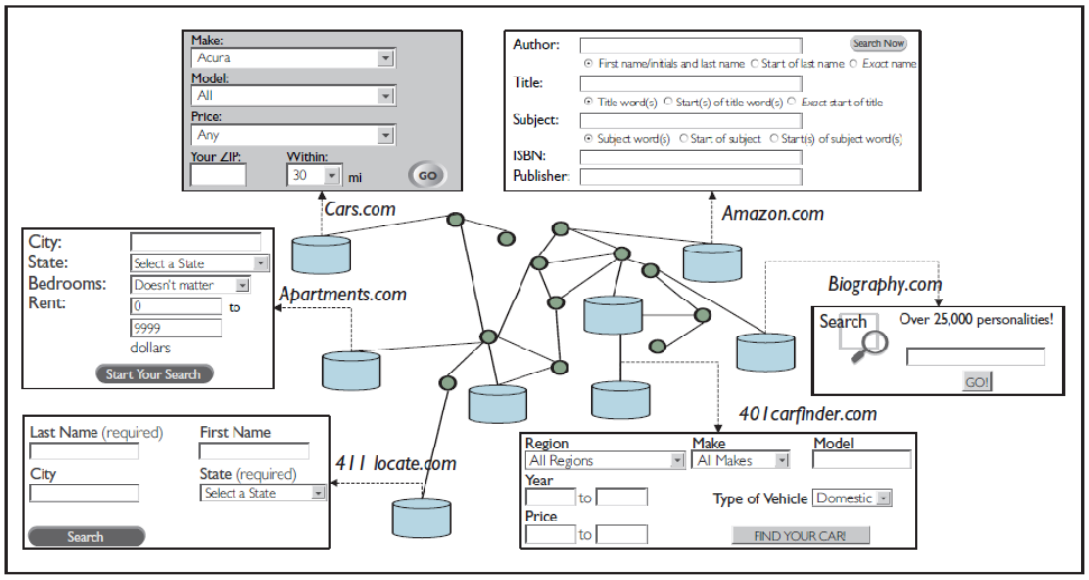

Figura 1 - Vista conceitual da Deep Web [17]

1.1 Objetivo

O objetivo desta dissertação é explorar a possibilidade de se utilizar dados da Deep Web para enriquecer bases de e-commerce existentes. Nossa intenção é propor um framework capaz de (1) identificar fontes de dados relevantes na Deep Web, (2) construir consultas automáticas para estas fontes de dados, (3) filtrar os resultados obtidos, (4) identificar a correspondência entre os resultados e as instâncias da base original (computar similaridade) e, finalmente, (5) enriquecer a base original com novas informações.

Aplicações geradas a partir do framework proposto podem servir a muitos propósitos, e.g., construção de novos sites, construção de mecanismos para a consulta e comparação de atributos, e aplicativos de *mashup* de dados. O foco deste trabalho, porém, é sua utilização como um mecanismo para enriquecer informações incompletas ou faltantes.

São dois desafios deste trabalho. Em primeiro lugar temos que construir mecanismos de rastreamento (*crawlers*) capazes de buscar informações na Deep Web. Rastreadores do tipo tradicional foram concebidos para funcionamento na Web de Superfície, e são programados para a coleta de informações contidas em páginas Web estáticas, que podem ser atingidas através da navegação por links. No caso da Deep Web, é preciso que estes rastreadores sejam capazes de submeter consultas através da interface Web de bancos de dados relacionais. A maior parte da literatura da área [28] trata de casos onde se deseja obter a totalidade dos dados, através do menor numero de consultas possível (*recall*). Desta forma, a maior parte dos artigos trata de questões como o limite do número de respostas imposto por determinados sites, a identificação de atributos relevantes de forma a reduzir o número de consultas necessárias, sobreposição de respostas, entre outras [3, 9, 10, 37].

Note que, como desejamos encontrar potencial informação adicional sobre registros da nossa base de dados, não estamos interessados em obter um grande número de respostas, mas sim respostas mais precisas. A situação ideal é encontrar uma combinação de atributos (não nulos) do registro de nossa base que retorne o mesmo objeto na base que está sendo pesquisada.

O segundo desafio, fundamental a qualquer sistema capaz de integrar dados de diversas fontes, é reconhecer se uma determinada entidade presente em uma fonte corresponde a uma entidade no mundo real, ou seja, ser capaz de identificar duplicatas.

Este é um problema tem atraído a atenção de pesquisadores de diferentes áreas, incluindo Bancos de Dados, Data Mining, Inteligência Artificial e Processamento de Linguagens Naturais. Na literatura é apresentado sob várias denominações, entre outras: “Record Linkage”, “Merge/Purge”, “Reference Conciliation”, “Entity Resolution” ou “Duplicate Detection” [5, 8, 14, 20, 21, 29, 36, 40].

Muitas soluções já foram propostas para tratar este problema [19]. A maioria, no entanto, funcionam muito bem para dados estruturados e em casos onde é possível aplicar técnicas de aprendizado de máquina a uma base de dados conhecida. Nosso foco, no entanto, é utilizar recursos encontrados na *Deep Web*. Neste caso, soluções codificadas e/ou treinamento *offline* não são adequados, primeiro porque o conjunto de dados a ser analisado não está disponível e é difícil de obter. Segundo, e mais importante, mesmo que haja um conjunto representativo de dados etiquetado para treinamento, as regras aprendidas podem não adequadas a todo o conjunto de dados ainda desconhecido.

Neste trabalho propomos uma estratégia que não requer uma base para treinamento. Como nossa proposta utiliza uma base de dados, a qual deseja-se enriquecer de dados, utilizamos funções de cálculo de similaridade em conjunto com um classificador para responder de forma online se duas entidades representam a mesma.

1.2 Contribuições

A contribuição deste trabalho é apresentar 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:

1. Uma estratégia para a busca (rastreamento) de informações na Deep Web orientada a duplicatas, cujo objetivo é obter resultados precisos de consultas construídas a partir de um conjunto conhecido de objetos e seus atributos.

2. Uma estratégia para a detecção de duplicatas em resultados de consultas realizadas sobre fontes de dados da Deep Web, independente de bases de treinamento especificadas a *priori*. Apresentamos uma solução baseada em classificadores e funções de cálculo de similaridade.

Uma contribuição adicional deste trabalho é fornecer uma descrição detalhada da instanciação do framework proposto para o domínio dos vinhos, que podem ajudar na processo de instanciação deste framework em contextos análogos no futuro.

Esta dissertação está estruturada como se segue: no próximo capítulo, apresentamos os conceitos básicos que sustentam a solução apresentada neste trabalho; no Capítulo 3, é demonstrada a estratégia utilizada neste trabalho; o framework proposto na dissertação é apresentado no Capítulo 4; o Capítulo 5 apresenta a implementação de uma instância do framework para o domínio dos vinhos; no Capítulo 6 temos a apresentação de trabalho relacionados; os resultados desde projeto encontram-se no Capítulo , juntamente com a conclusão do trabalho e, finalmente, no Capítulo 8 temos a bibliografia estudada.

1.3 Resumo

Neste capítulo apresentamos a motivação básica e principais contribuições deste trabalho. Nosso objetivo é explorar a possibilidade de utilização da Deep Web em um processo que batizamos de enriquecimento de dados. Este processo consiste na obtenção de mais informações, e.g., texto e imagens, para complementar a informação de uma base de dados existente. Este processo pode ser decomposto em duas partes: busca e incorporação. No passo de busca é necessário construir rastreadores capazes de buscar informações contidas na Deep Web. A peculiaridade do nosso caso é a necessidade de se encontrar na bases de dados da Deep Web os objetos mais próximos daqueles pertencentes à base de dados original (precisão). A segunda parte do processo é identificação de duplicatas, i.e., identificar se o(s) objeto(s) recuperado(s) e aquele da base original são, de fato, o mesmo objeto real. No caso da Deep Web, o grande desafio é a impraticalidade de se utilizar um conjunto de dados para o treinamento, comuns à maior parte das estratégias de aprendizado de máquina utilizadas na identificação de duplicatas.