6 Trabalhos relacionados

Existem trabalhos na literatura que lidam com diversas questões envolvidas no desenvolvimento deste trabalho. Neste capítulo discutimos diferenças e similaridades à luz dos resultados apresentados nos capítulos anteriores.

6.1 Rastreadores para Deep Web

Os trabalhos existentes na literatura [3, 30, 32, 37] demonstram aspectos envolvidos na construção de rastreadores para Deep Web. Estes trabalhos relacionados têm o objetivo de aumentar a cobertura de indexação de páginas por mecanismos de busca. Em nosso trabalho há a preocupação de direcionar o rastreador (*crawler*) para obtenção eficiente de duplicatas, montando consultas capazes de retornar apenas informações que possam ser adicionadas ao nosso banco de dados, já que o processo de incorporação de informações, descrito no Capítulo 3, é custoso computacionalmente.

6.2 Resolução de Entidades

Relacionando os trabalhos apresentados no Capítulo 2 com o framework apresentado, temos algumas diferenças e similaridades, a ver:

O framework UDD, publicado durante o desenvolvimento deste trabalho, é o primeiro, segundo os autores, a tentar realizar o casamento de entidades de forma online na Deep Web [36]. Este é o mesmo propósito deste.

O UDD de Su, assim como nosso trabalho, não apresenta um mecanismo de extração de dados das páginas. Cabe a quem instanciar o framework codificar um mecanismo de extração de informações das páginas.

As principais diferenças entre MARLIN [6], UDD [36] e o método proposto neste trabalho são:

1. Tanto os frameworks MARLIN quanto UDD utilizam classificadores SVM para classificação. Em nossos testes, foi verificado que a utilização de classificadores mais simples pode ser tão eficiente quanto estes para determinados domínios. Dado que nos problemas que objetivamos tratar a

quantidade de dados é pequena, a utilização de um classificador deste tipo pode vir a introduzir ruídos na classificação.

2. O framework MARLIN foi concebido para cenários de integração de dados off-line, enquanto os frameworks UDD e o framework proposto neste trabalho foram desenvolvidos para operar em um cenário on-line. Por conta disso, as entidades utilizadas nestes últimos são sempre objetos com múltiplos campos do tipo cadeia de caracteres, ao contrário do primeiro que pode ter campos com outros tipos de estruturas como números ou referências a outros objetos.

6.3 Comparação de resultados

É possível obter indícios sobre a qualidade da solução proposta comparando os resultados com o obtidos pelo UDD, um framework que propõe resolver o problema de detecção de duplicatas online [36]. Os experimentos foram realizados com diversos domínios, no entanto, Su [36] compara seu framework com outros utilizando o domínio dos livros. O UDD teve a melhor performance, com 92,4% de precisão, 91,5% de r*ecall* e 91,9% de *F-Measure*. É claro que muito pouco pode ser dito ao compararmos resultados de domínios diferentes, mas a intuição diz que o domínio de livros deve ser menos problemático que o de vinhos, já que o ISBN normalmente provê uma identificação única e não há tanta variação nos títulos e nomes de autores como há nas etiquetas dos vinhos.

Na proposta de dissertação de doutorado de Bilenko [7], o autor conduziu um experimento com o domínio dos vinhos, de modo a testar algumas técnicas de clusterização do MARLIN. Os seus resultados mostram que nos melhores casos é possível alcançar valores próximos de 90% de F-Measure usando seu framework. Não foram dados detalhes sobre precisão e recall. Na comparação com essas duas abordagens, é possível constatar que os resultados obtidos estão na mesma faixa, o que nos faz acreditar que a solução proposta é aceitável. A tabela 11 resume os valores de precisão, *recall* e *F-Measure* obtidas nos frameworks.

Framework	Precisão	Recall	F-Measure
Abordagem proposta	88.89%	90.0%	89.44%
UDD [34]	92.40%	91.50%	91.90%
MARLIN [7]	N/D	N/D	~90%

Tabela 11 - Comparação de performance