paper:82

Definição e Avaliação de uma Abordagem para Extração e Rotulação de Conteúdo Obtido da *DeepWeb*

Augusto Ferreira de Souza, Ronaldo dos Santos Mello

Departamento de Informática e Estatística–Centro Tecnológico – Universidade Federal de Santa Catarina (INE/CTC/UFSC) Caixa Postal 476– Florianópolis, SC - Brasil.

augustofs@gmail.com, r.mello@ufsc.br

Abstract. *This paper presents an approach for the extraction and labeling of data presented in Deep Web databases. Such a data are extracted from a set of HTML pages generated as the result of a query posed on the hidden database through a Web form. Data labeling (and persistence) aims at providing further structured queries over this hidden content. Preliminary experiments had demonstrated that the proposed approach is promising, if compared with baselines. Other contributions are a joint-process for simultaneous data extraction and labeling, an automatic approach with the support of a knowledge base, and a labeling process of extracted records with content self-filling support for attributes with missing values.*

Resumo. *Este artigo apresenta uma solução para a extração e rotulação de dados contidos em bancos de dados na Deep Web. Esses dados são extraídos de um conjunto de páginas HTML gerado como resultado de uma consulta submetida ao banco de dados através de um formulário Web. A rotulação (e conseqüente persistência) destes dados viabiliza futuras consultas estruturadas sobre este conteúdo escondido. Uma avaliação preliminar demonstrou a eficácia da abordagem proposta em relação a baselines. Outros diferenciais deste trabalho são a realização simultânea de um processo de extração e de rotulação de dados, uma abordagem automática com suporte de uma base de conhecimento e um processo de rotulação de registros extraídos com suporte ao auto-preenchimento de atributos com valores ausentes.*

1. Introdução

O aumento do volume de dados disponíveis na *Deep Web* [Halevy et al. 2009] faz com que também aumente o interesse no acesso a essas informações por parte dos usuários. A *Deep Web* representa, dentre outras fontes de dados, bancos de dados disponíveis na *Web* cuja estrutura e conteúdo tornam-se visíveis (ou parcialmente visíveis) apenas quando apresentados em páginas dinâmicas criadas a partir do resultado de uma consulta definida sobre um formulário *Web* [Bergman, 2001]. O formulário *Web* é a principal interface de busca para estes bancos de dados ditos "escondidos" (BDEs).

Para se ter acesso às informações de um BDE são necessárias soluções para a sua descoberta na Web, submissão automática de consultas através dos seus formulários, assim como técnicas de extração e rotulação dos dados exibidos nas páginas de resultado, viabilizando o posterior consumo humano. A atividade de extração, em particular, é complexa devido à existência de uma grande variedade de *Web sites* com padrões diferenciados para a exibição do conteúdo de BDEs, bem como a

existência de muitas informações irrelevantes (menus, anúncios, etc.) que dificultam o reconhecimento do que é realmente relevante [Hong, 2010][Oro et al. 2011]. Já as abordagens de rotulação de dados geralmente utilizam um dicionário para ajudar na comparação dos dados para uma correta rotulação, porém não são capazes de detectar novas informações e assim construir registros com conteúdo mais completo [Zhao et al. 2008][Silva et al. 2011].

Assim sendo, o objetivo deste artigo é detalhar uma abordagem chamada *DeepEL* (*Deep Web Extraction and Labeling Process*) que realiza de forma automática a extração e rotulação de conteúdo obtido da *Deep Web*. O processo de rotulação é capaz de detectar informações que não estão disponíveis nos registros extraídos. Esta detecção é possível com o suporte de uma base de conhecimento (BC) construída para alguns domínios da *Deep Web*. A intenção da *DeepEL* é a construção sistemática de um BD sobre a *Deep Web* que possa servir de base para diversos serviços, como por exemplo, sistemas de busca na *Deep Web* e a construção de catálogos de BDEs. As principais contribuições deste artigo são:

- A definição de um processo de extração de conteúdo existente em BDEs e que não estão necessariamente presentes em rótulos e valores dos seus formulários de busca;
- Experimentos preliminares que indicam que o processo de extração é promissor, se comparado com *baselines* conhecidos na literatura;
- A definição de um processo de rotulação cujo benefício é uma melhoria de qualidade para futuras buscas estruturadas sobre dados de BDEs, uma vez que o conhecimento sobre os mesmos fica enriquecido com a complementação do conteúdo dos seus atributos.

Este artigo está organizado conforme segue. A seção 2 sumariza os principais trabalhos relacionados à extração e rotulação de dados na *Web*. A seção 3 detalha a abordagem *DeepEL*. A seção 4 apresenta os experimentos preliminares e a seção 5 é dedicada às conclusões e trabalhos futuros.

2. Trabalhos Relacionados

2.1 Extração de Dados

Soluções para extração de dados na *Web* definem algoritmos capazes de identificar e recuperar informações presentes em páginas *Web* [Kaiser and Miksch 2005]. Nenhum dos trabalhos relacionados ao tratamento desta problemática inclui um processo de rotulação dos dados extraídos, nem um processo automatizado quando há o suporte de uma ontologia ou BC [Embley et al. 1998] [Muslea et al. 2001] [Liu et al. 2003] [Phan et al.2005]. Essas limitações motivaram o desenvolvimento da *DeepEL*.

Dentre as soluções existentes, este trabalho considerou dois *baselines* durante a realização de experimentos: *Road Runner* [Muslea et al. 2001] e MDR [Liu et al. 2003]. Estes métodos são comumente referenciados na literatura [Hong, 2010][Oro et al. 2011]. *Road Runner* é uma abordagem automática que compara conteúdos HTML a fim de definir uma expressão regular para um conjunto de páginas de amostra. Depois de resolver incompatibilidades, uma expressão comum é utilizada para extrair registros de

dados de outras páginas na *Web*. O método MDR (*Mining Data Records*) gera uma árvore DOM de uma página HTML e a percorre de maneira *bottom-up* verificando a similaridade entre nodos adjacentes. A partir desta análise de similaridade, ele determina regiões de dados que serão consideradas para fins de extração.

2.2 Rotulação de Dados

Com relação à rotulação de dados, o trabalho de [Zhao et al. 2008] emprega uma técnica de aprendizado de máquina para rotulação de texto. Ele utiliza tabelas com dados estruturados como amostra para treinamento. Durante a fase de treinamento, a rotulação de conteúdos textuais é definida a partir de tabelas de referência. Assumindo que as sequências de texto a serem segmentadas vêm em lotes, elas devem estar em conformidade com a ordem dos atributos definidos no treinamento, para serem devidamente identificadas e rotuladas.

O trabalho de [Silva et al. 2011] introduz a abordagem JUDIE. Ela recebe como entrada um texto contendo registros de dados. Após, ocorre a segmentação dos dados e uma rotulação de valores potenciais, realizados através da comparação com dados relacionados ao domínio mantidos em uma BC. Após a segmentação e a primeira rotulação, um algoritmo baseado em um modelo de posicionamento e sequenciamento de dados rotula-os novamente, confirmando ou efetuando possíveis correções nos rótulos inicialmente rotulados.

Limitações desses trabalhos que motivaram o desenvolvimento da *DeepEL* são:

- No trabalho de [Zhao et al. 2008], o esquema assume uma ordem fixa conforme a amostra, prejudicando a rotulação dos dados que não seguem este padrão;
- Nenhuma abordagem realiza a complementação de informações que, apesar de pertencerem ao domínio, não estão presentes na fonte de dados a ser rotulada.

3. DeepEL

A abordagem *DeepEL* está centrada em dois processos: extração e rotulação de dados da *Deep Web*. A Figura 1 apresenta a arquitetura da abordagem e os componentes considerados. O componente externo *Motor* (1) encapsula o processo de descoberta de BDEs na *Web*, preenchimento de formulários *Web* e submissão de consultas, bem como a geração de páginas HTML com os resultados obtidos. Esse conjunto de páginas (2) é a entrada considerada pela *DeepEL*.

A *DeepEL* utiliza uma BC (4) como auxílio nos processos de extração e rotulação de dados. Ela foi especificada no formato XML e, nesta primeira versão, mantém conhecimento relativo a dois domínios populares na *Deep Web*: *Automóveis* e *Livros*. A modelagem da BC considerou a coleta dos principais atributos de cada domínio, conforme descrito no *corpus* do repositório colaborativo de dados *Freebase*[1], bem como o enriquecimento semântico deste conhecimento através da coleta (e posterior cadastro na BC) de termos sinônimos presentes na base de dados léxica *WordNet*[2]. Apesar da estrutura complexa no formato XML, a BC tende a ser reduzida,

[1] *http://www.freebase.com/*
[2] *http://wordnet.princeton.edu/*

pois mantém apenas atributos mandatários do domínio e exemplos de valores de atributos relevantes. Outras BCs disponíveis são muito mais volumosas, requerendo a indexação de muitos dados. Maiores detalhes sobre a modelagem da BC são omitidos devido a restrições de espaço. A Figura 2(b) apresenta um fragmento da BC com conteúdos de rótulos de atributos presentes no domínio de *Automóveis*, inclusive com sinônimos para alguns rótulos enriquecidos a partir do *WordNet*.

Os demais componentes da arquitetura são comentados à medida que os processos de Extração (3) e Rotulação (6) são detalhados, conforme segue.

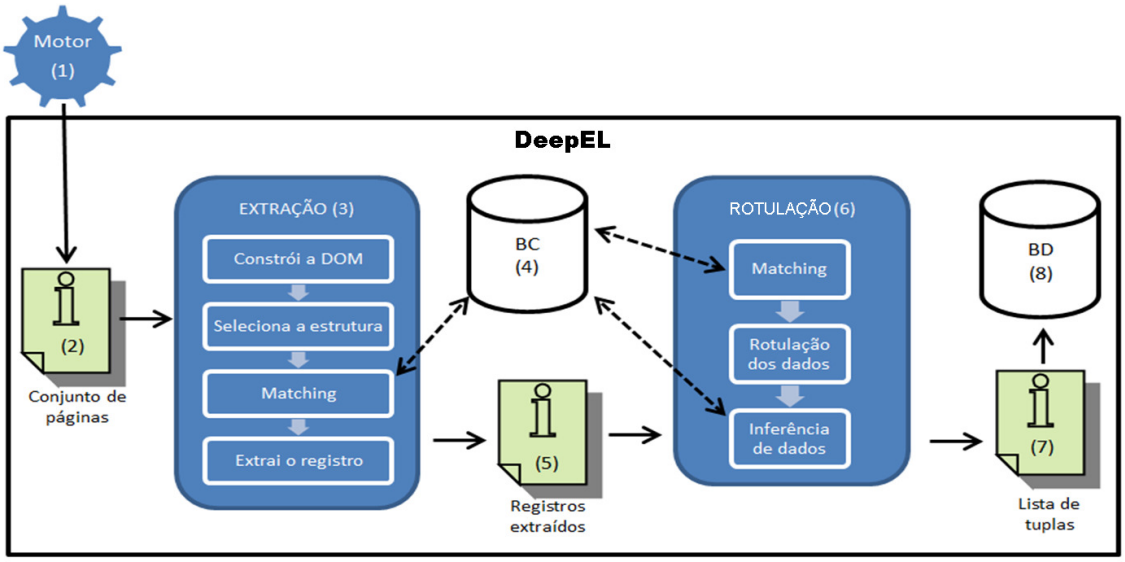

Figura 1. Arquitetura da Abordagem DeepEL.

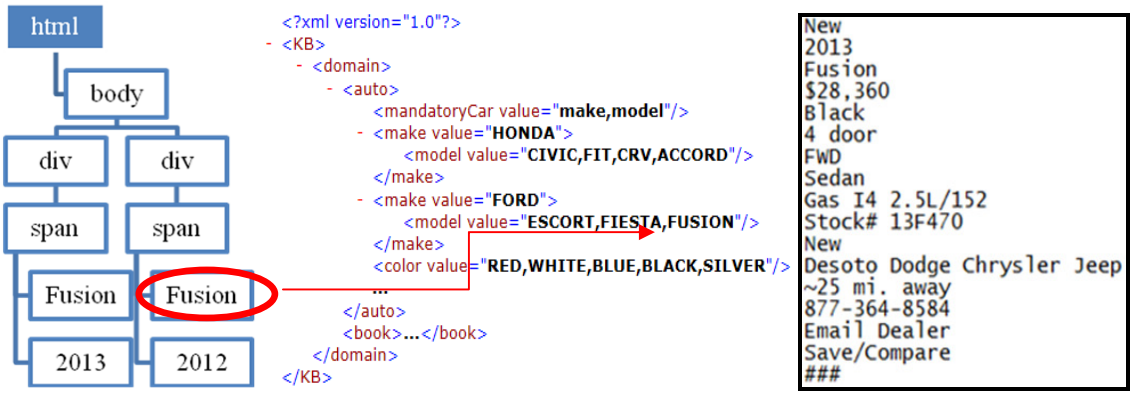

Figura 2. Página HTML no formato DOM (a). Exemplo parcial da BC no formato XML (b). Exemplo de arquivo com os registros extraídos (c).

3.1 Processo de Extração

O processo de extração adotado pela *DeepEL* é baseado nas Heurísticas 1 a 3, que visam melhorar a qualidade da recuperação dos registros de dados de BDEs presentes nas páginas HTML. Estas heurísticas são definidas a seguir.

Definição 1 (Heurística da Estrutura Irrelevante - HEI). As estruturas HTML definidas pelas *tags* "*script*", "*select*" e "*option*" são descartadas, pois provavelmente são funções ou campos de formulários sem conteúdo relevante.

Definição 2 (Heurística de Atributo Mandatário - HAM). Um atributo mandatário A_m é um atributo significativo de um determinado domínio, servindo para caracterizar o domínio de um registro que possua um valor de A_m.

Exemplo. O atributo *make* é um atributo mandatário no domínio de *Automóveis*, pois é uma propriedade inerente a qualquer registro de automóvel.

Definição 3 (Heurística da Estrutura Relevante - HER). A estrutura de representação de dados que mais se repetir na página HTML é onde provavelmente estão localizados os registros relevantes advindos do BDE.

A heurística HER é a mais importante para o bom funcionamento do extrator, pois é ela que decide o que é considerado relevante. Ela foi definida com base no trabalho de [Hong, 2010], porém, a *DeepEL* tem como diferencial o suporte da BC para confirmar se a estrutura mais frequente contém realmente os dados desejados. Com a heurística HEI analisa-se de forma mais minuciosa a estrutura HTML e desconsidera-se *tags* que não possuem conteúdo útil.

O Algoritmo 1 detalha o processo de extração. Ele recebe como entrada uma página HTML e considera o suporte da BC. Na linha 5, a página é instanciada no modelo DOM e na linha 6 é aplicada a HEI que remove estruturas irrelevantes. Na linha 7 é realizada uma busca por atributos mandatórios de um determinado domínio disponíveis na BC (HAM).

Algoritmo 1: Método de Extração.

```
1:  Entrada: páginaHTML;
2:  Início
3:    CaminhosCandidatos←{ }
4:    CaminhosRelevantes←{ }
5:    PDOM ← DOM_Parser(pagina)
6:    SDOM ← HEI(PDOM )
7:    para cada valor de HAM(v)∈ BC faça
8:      para cada termo t de um nodo folha n ∈ SDOM faça
9:        se matching(v, t) então
10:         CaminhosCandidatos←CaminhosCandidatos + Caminho (n)
11:       fim se
12:     fim para
13:   fim para
14:   CaminhosRelevantes← HER (CaminhosCandidatos)
15:   para ∀ termo t de um nodo folha n ∈ CaminhosRelevantes faça
16:     extrai(t)
17:   fim para
18: Fim
```

As Figuras 2a e 2b exemplificam o processo de *matching* entre o valor de um atributo mandatório da BC e um termo presente nas folhas de um nodo DOM da página HTML (linha 9). A função de similaridade utilizada aqui foi a *Jaro-Winkler* [Winkler,

1990], por ter obtido melhor desempenho nos experimentos preliminares. Vale ressaltar que, mesmo a BC sendo extensa, essa comparação por similaridade é realizada somente com o conteúdo das *tags* correspondentes a atributos mandatários dos domínios presentes na BC, como é o caso da *tag model*, que é um atributo mandatário no domínio de *Automóveis*. Após a identificação de tais registros, o caminho da *tag* raiz que representa o registro de dado até os nodos folha é considerado um *caminho candidato* (linha 10). Na sequência, na linha 14, é aplicada a HER sobre os caminhos candidatos para selecionar apenas os caminhos relevantes.

Por fim, na linha 16, ocorre a extração dos conteúdos que se encontram nos caminhos relevantes, gerando um arquivo de saída com os registros extraídos (5), conforme ilustrado na Figura 2c. Cada linha do arquivo é o valor de um campo do registro extraído de um nodo. Cada registro, por sua vez,é delimitado por "###".

3.2 Processo de Rotulação

O processo de rotulação adota uma abordagem de análise e caracterização de segmentos de texto, utilizando também a BC como apoio. Ele recebe como entrada o arquivo com os registros extraídos de uma página HTML e analisa cada um deles, campo a campo, para fins de construção de uma tupla a ser rotulada. Para tanto, o conteúdo de cada campo é comparado com os conteúdos da BC. O Algoritmo 2 detalha este processo. Inicialmente, na linha, 3 é considerada uma heurística para detecção do domínio ao qual pertencem os registros a serem rotulados. Ela é definida a seguir.

Definição 4 (Heurística de Detecção de Domínio - HDD). Uma detecção de um domínio *D* ocorre quando conteúdos de um registro extraído casam, por similaridade, somente com valores de atributos mandatários de *D* presentes na BC.

Algoritmo 2: Método de Rotulação.

```
1:  Entrada: conjunto de registros extraídos RegEx;
2:  Início
3:    d ← HDD(RegEx)
4:    se d ≠ nulo então
5:      para cada atributo com tag t e valor v ∈ BC(d) faça
6:        para cada registro r ∈ RegEx faça
7:          para cada campo c ∈ r faça
8:            se matching(v, c) então
9:              rotula(c,t,d)
10:           Senão
11:             se detectaPadrao(c) então
12:               rotula(c,t,d)
13:             fim se
14:           fim se
15:         fim para
16:         para cada atributo t da tupla rotulada tp |valor (t,tp) = nulo faça
17:           itemDetectado ← HDC (t)
18:           se itemDetectado ≠ nulo então
19:             rotula(itemDetectado,t,d)
20:           fim se; fim para; fim para; fim para; fim se
21: Fim
```

A HDD se baseia na hipótese considerada pela heurística HAM. Uma vez que os atributos mandatários são específicos de cada domínio, eles possibilitam a detecção deste domínio. Se um domínio *D* é detectado (linha 4), cada registro é lido e seus campos comparados com os termos da BC pertencentes a *D* (linhas 5 a 7). As similaridades entre termos da BC e os campos dos registros são identificadas (linha 8) utilizando-se, mais uma vez, a função *Jaro-Winkler*. Quando uma correspondência é identificada, o termo e o seu significado (*tag*) são armazenados na tabela do BD destino (8) correspondente ao domínio *D* (linha 9).

O processo de rotulação considera o reconhecimento de alguns padrões de valores para determinados atributos de domínios (linha 11). Esse reconhecimento é realizado através de expressões regulares e funções específicas. Para o domínio de *Automóveis*, por exemplo, padrões descobertos são ano, preço e quilometragem. Ainda, um diferencial deste processo é a detecção de conteúdos para atributos nulos de registros a serem rotulados. Isto ocorre através da identificação das correspondências entre campos de registros e amostras de dados da BC para determinados domínios, como é o caso do modelo *Fusion* para a marca *Ford* mostrada na Figura 2b. Neste caso, se o modelo é conhecido e a marca não, é possível determinar o valor da marca caso esta dependência de valores esteja presente na BC, conforme rege a heurística HDC definida a seguir. Desta forma, enriquece-se o conteúdo deste registro no ato da rotulação.

Definição 5 (Heurística de Detecção de Conteúdo - HDC). Uma detecção de conteúdo de atributo ocorre quando existe uma dependência de valor entre atributos X e Y, sendo X o atributo determinante e Y o atributo determinado na hierarquia da BC. Neste caso, se o valor de Y é rotulado, o valor de X também o será.

Esta heurística é considerada na linha 17 após a verificação da existência de algum valor de atributo nulo na tupla construída para fins de rotulação (linha 16). Caso ocorra a detecção de conteúdo, o mesmo é rotulado (linha 19). Por fim, as tuplas criadas (7) são armazenadas em um BD relacional (8). A Figura 3 mostra exemplos, gerados pelo processo de rotulação para o domínio de *Automóveis*, de registros e campos devidamente preenchidos e prontos para serem persistidos como tuplas no BD relacional. Detalhes sobre o projeto deste BD são omitidos por restrições de espaço.

Nr:1	make: FORD	model: FUSION	door: 4	price: 28.360	year: 2013	color: BLACK	
Nr:2	make: FORD	model: FUSION	door: 2	price: 22.888	year: 2012	color: GRAY	mileage: 10.054
...	...	...	...	...	...	...	...

Figura 3. Exemplos de registros rotulados para o domínio de *Automóveis*.

4. Experimentos

As amostras de dados utilizadas nos experimentos foram obtidas a partir do repositório de páginas *Web* mantido pelo sistema de busca *Deep Peep* [Barbosa et. al 2010]. *Deep Peep* é uma máquina de busca para formulários *Web* pertencentes a alguns domínios da *Deep Web*. Os dois domínios escolhidos para os experimentos são os que possuem maior volume de dados indexados por esta máquina de busca. A partir disso, realizou-se o acesso a 18 páginas desses domínios e a submissão de consultas aos formulários presentes nelas. As consultas geraram 60 páginas de resultado, que foram consideradas como entrada para a *DeepEL*. Essas páginas geraram 1376 registros no domínio de

Automóveis e 352 registros no domínio de *Livros*. Uma pequena amostra destes dados foi utilizada também na população da BC.

Os experimentos avaliaram a qualidade da extração e da rotulação de dados nos dois domínios. Primeiramente, realizaram-se experimentos somente para a etapa de extração. Nesta etapa, os resultados obtidos foram comparados com os resultados gerados pelos *baselines* MDR e *Road Runner*. A justificativa para a comparação da *DeepEL* com essas duas abordagens é que, apesar delas não considerarem aspectos semânticos em seus métodos, como o suporte de uma BC, elas se enquadram dentre as poucas que realizam o processo de extração de forma automática, assim como a *DeepEL*, e possuem códigos-fonte disponíveis para *download*.

O desempenho da *DeepEL* foi também avaliado através de experimentos para a etapa de rotulação utilizando os registros recuperados durante a etapa de extração. Cabe salientar que um trabalho anterior definiu informalmente o processo de rotulação e fez uma avaliação apenas dos dois *baselines* [Souza e Mello, 2013]. Este artigo propõe um processo de extração para a *DeepEL*, o compara com os *baselines* e ainda define formalmente os dois processos em termos de heurísticas consideradas.

A Tabela 1 apresenta os resultados do processo de extração da *DeepEL* em comparação com os *baselines*. A tabela apresenta os resultados em termos de precisão (P), revocação (R) e medida F (F), bem como o tempo de processamento necessário para a extração dos registros. Utilizou-se *make* e *model* como atributos mandatórios para o domínio de *Automóveis*, bem como *publisher* e *author* para o domínio de *Livros*.

Tabela 1. Resultados do processo de extração de dados.

Abordagem	Domínio	P	R	F	Tempo (s)
Road Runner	Automóveis	0,94	0,95	0,95	18,10
MDR		0,92	0,95	0,93	13,56
DeepEL (Extração)		0,94	**0,96**	0,95	**10,09**
Road Runner	Livros	0,84	0,91	0,88	4,95
MDR		0,81	0,90	0,85	3,62
DeepEL (Extração)		0,87	0,90	**0,89**	**2,88**

Nota-se que a etapa de extração da *DeepEL* apresenta resultados próximos à precisão e à revocação dos *baselines*. Obteve-se a melhor revocação para o domínio de *Automóveis*, ou seja, o melhor percentual de acertos em termos de recuperação de registros desejados presentes nas páginas de resultado, pois este domínio apresenta propriedades mais homogêneas, facilitando assim a identificação dos registros desejados. Outro ponto positivo da *DeepEL* foi a melhor medida F para o domínio de *Livros*. Além disso, ela realizou a extração com rapidez (aproximadamente 20% mais veloz que o *baseline* mais rápido), mesmo considerando o acesso a uma BC que, por ser enxuta, pode ser manipulada totalmente em memória.

A Tabela 2 apresenta o desempenho da *DeepEL* em termos de rotulação, ou seja, os valores de precisão, revocação e medida F considerando o que é rotulado em relação à quantidade de registros extraídos disponíveis nos arquivos. Os bons resultados (medida F superior à 90%) se justificam pelo fato de grande parte dos registros rotulados possuírem algum tipo de informação presente na BC, como por exemplo, ocorrências de modelos de carros para o domínio de *Automóveis*. Cabe ressaltar que, mesmo com a BC contendo uma pequena amostra de valores de alguns atributos

mandatórios dos domínios, isso já torna possível a rotulação correta de um conjunto de dados muito mais volumoso, pois basta haver um casamento parcial de conteúdo dos dados da amostra com o registro extraído para ocorrer a rotulação.

Tabela 2. Avaliação da rotulação dos conteúdos extraídos pela *DeepEL*.

Domínio	Precisão	Revocação	Medida F
Automóveis	0,94	0,96	0,95
Livros	0,90	0,93	0,91

Por fim, a Tabela 3 apresenta a quantidade de valores de campos que foram complementados para o conjunto de registros considerado nos experimentos. Observa-se aqui que, aplicando o mecanismo de detecção de conteúdo, foi possível um ganho de qualidade na rotulação de registros de mais de 10%. Este percentual não foi elevado pois grande parte dos registros a serem rotulados possuía valores para todos os campos previstos no domínio. Mesmo assim, é uma contribuição da abordagem *DeepEL*.

Tabela 3. Ganho com o enriquecimento de informação durante a rotulação.

Domínio	#Registros	Valores de Campos Detectados	Ganho em Termos de Registros Complementados
Automóveis	1376	150	10,90 %
Livros	352	25	7,10 %

5. Conclusão

Diferente do estado da arte sobre extração e rotulação de dados da *Deep Web*, que não trata essas atividades como um processo único e possui limitações em termos da extração distinta de metadados e valores, além da falta de complementação de informações, este artigo define e avalia uma abordagem, denominada *DeepEL*, para extração de registros de dados estruturados de páginas de resultados de consultas submetidas a BDEs da *DeepWeb*, bem como a rotulação e complementação de conteúdo destes registros. A *DeepEL* é uma contribuição para a problemática de tornar visível, estruturado e contextualizado o conteúdo presente em BDEs, contando ainda com o suporte de uma BC projetada para manter metadados e amostras de valores mais significativos dos principais domínios da *DeepWeb*. A BC é utilizada na comparação e inferência de conteúdo para registros a serem extraídos.

Os resultados obtidos através de uma avaliação experimental indicaram que a *DeepEL* executa no mínimo com qualidade similar a dos principais métodos de extração automáticos, porém com um melhor desempenho de tempo e de qualidade de rotulação nos domínios considerados. Além disso, ela proporcionou um ganho de até 10% no enriquecimento do conteúdo dos registros rotulados. Estima-se que este ganho seja maior para um volume maior de registros em domínios cujo conteúdo seja mais homogêneo, como é o caso do domínio de *Automóveis*, um dos mais extensos da *Deep Web*. Deseja-se executar novos experimentos com um maior volume de dados extraídos de páginas *Web* para validar esta hipótese.

Alguns trabalhos futuros estão em planejamento, entre eles: *(i)* avaliar o desempenho da *DeepEL* em outros domínios da *DeepWeb*; *(ii)* permitir a extensão da BC a partir de novos conteúdos extraídos, visando ampliar o seu conhecimento; *(iii)* comparar o desempenho da *DeepEL* com outros métodos de rotulação, caso se obtenha acesso às suas implementações. Por fim, cabe salientar que a *DeepEL* não está livre de

problemas de ambigüidade durante o processo de rotulação. Entretanto, os exemplos na BC e seu posterior enriquecimento, assim como uma análise de posicionamento de conteúdos dos registros, conforme utilizado por [Silva et al. 2011], podem minimizar esta questão, que também é alvo de investigações futuras.

Referências

Barbosa, L., Nguyen, H., Nguyen, T., Pinnamaneni, R., and Freire, J. (2010). Creating and Exploring Web Form Repositories. In: ACM SIGMOD Int. Conf. on Management of Data, p.1175-1178.

Bergman, M. K. (2001). White Paper: The Deep Web: Surfacing Hidden Value. Journal of Electronic Publishing, v.7, n.1.

Embley, D. W., Campbell, D. M. and Smith, R. D. (1998). Ontology-Based Extraction and Structuring of Information from Data-Rich Unstructured Documents. In: Int. Conf. on Information and Knowledge Management, p. 52-59.

Halevy, A., Madhavan, J., Afanasiev, L. and Antova, L. (2009). Harnessing the Deep Web: Present and Future. In: Int. Conf. on Innovative Data Systems Research.

Hong, Jun; He, Zhongtian and Bell, David A. (2010). An Evidential Approach to Query Interface Matching on the Deep Web. Information Systems, v.35, n.2, p.140-148.

Kaiser, K. and Miksch, S. (2005). Information Extraction. A Survey. Technical Report,Vienna University of Technology.

Liu, B., Grossman, R. and Zhai, Y. (2003). Mining Data Records in Web Pages. In: ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining, p.601-606.

Muslea, I., Minton, S. and Knoblock, C. (2001). Hierarchical Wrapper Induction for Semistructured Information Sources. Autonomous Agents and Multi-Agent Systems Archive, v.4, n.1, p.93-114.

Oro, E. and Ruffolo, M. (2011). SILA: a Spatial Instance Learning Approach for Deep Web Pages. In: Int. Conf. on Information and Knowledge Management, p.2329-2332.

Phan, X., Horiguchi, S. and Ho, T. (2005). Automated Data Extraction from the Web with Conditional Models. Int. Journal of Business Intelligence and Data Mining, v.1, n.2, p.194-209.

Silva, A. S., Cortez, E., Oliveira, D., Moura, E. S. and Laender, A. H. F. (2011). Joint Unsupervised Structure Discovery and Information Extraction. In: ACM SIGMOD Int. Conf. on Management of Data, p.12-16.

Souza, A. F. and Mello, R. S. (2013). DeepEC: An Approach for Deep Web Content Extraction and Cataloguing. In: European Conf. on Information Systems.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In: Section on Survey Research Methods, American Statistical Association, p.354-359.

Zhao, C., Mahmud, J. and Ramakrishnan, I. V. (2008). Exploiting Structured Reference Data for Unsupervised Text Segmentation with Conditional Random Fields. In: Int. Conf. on Data Mining, p.420-431.