UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE INFORMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

Geração de Regras de Extração de Dados em Páginas HTML

por

PARACELSO DE OLIVEIRA CALDAS

Dissertação submetida à avaliação como
requisito parcial para a obtenção do grau de
Mestre em Ciência da Computação

Prof. Dr. Carlos Alberto Heuser
Orientador

Porto Alegre, outubro de 2003

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Caldas, Paracelso de Oliveira

Geração de Regras de Extração de Dados em Páginas HTML / por Paracelso de Oliveira Caldas. – Porto Alegre: PPGC da UFRGS, 2003.
63 p.: il.

Dissertação (Mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR-RS, 2003. Orientador: Heuser, Prof. Dr. Carlos Alberto.

1. Dados semi-estruturados. 2. Tabelas. 3. Modelo Conceitual. 4. Regras de extração. 5. Extração de dados 6. XML I. Heuser, Carlos Alberto. II. Título.

“Só aqueles que se arriscam indo longe, têm a oportunidade de ver quão longe podem ir.”

– Thomas Elliot

Agradecimentos

A razão de nossa existência tem um, e somente um, responsável: o nosso criador. É por ele que começo os agradecimentos. Agradeço por ter-me concedido saúde, por ter-me proporcionado vontade interior para este aprendizado, por ter mantido em mim a persistência para transpor os obstáculos encontrados na empreitada; agradeço por todas as dádivas concedidas ao longo do desenvolvimento do trabalho.

Agradeço ao meu orientador Prof. Dr. Carlos Alberto Heuser que, de forma competente e organizada, teve paciência, soube compreender e realinhar minhas idéias para atingir com êxito o propósito do trabalho. Por ter propiciado sempre um ambiente agradável e otimista, tornando a missão mais leve e próspera. Por tais razões, aproveito o espaço para um “muito obrigado”.

À minha família, faço um agradecimento muito especial, porque, colocando-me no lugar de cada um, primeiramente no de minha esposa Jacira, seguida dos meus filhos Adson, Anderson, Alisson e Aline, além de minhas noras e meus netos, entendo que os privei de momentos irretornáveis.

Em meados do ano 2000, começamos a freqüentar o mestrado. Aos poucos, o envolvimento com aulas e trabalhos fora de classe vinha fortificando um convívio amigo e camarada. Assim, não poderia jamais esquecer de agradecer por tal convívio e também manter forte este relacionamento.

À UNESC e à Pós-Graduação, proporcionando-nos o mestrado e acompanhando-nos por todo o tempo os passos.

Aos meus colegas do grupo de pesquisa, orientandos do Prof. Dr. Carlos Alberto Heuser que, prontamente, atenderam-me os pedidos de material para pesquisa.

Agradeço, no momento, a todas as pessoas que contribuíram de alguma maneira para que eu chegasse com êxito à conclusão do mestrado.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Siglas e Abreviaturas

BYU	Brigham Young University
DEByE	Data Extraction by Example
GREMO	Gerador de Regras de Extração e Modelo Conceitual
HTML	Hypertext Markup Language – Linguagem de Marcação de Hipertexto
Minerva	Mapping the Internet Electronic Resources Virtual Archive
NLP	Processamento de Linguagem Natural
NoDoSE	The Northwestern Document Structure Extractor
OQL	Object Query Language
RAPIER	Robust Automated Production of Information Extraction Rules
RoadRunner	Para extração automática de dados em grandes redes locais
SoftMealy	Is a system that learns to extract data from semistructured web pages
SQL	Structured Query Language
SRV	The Sequence Rules with Validation
STALKER	Is a supervised learning algorithm for inducing extraction rules
TSIMMIS	The Stanford-IBM Manager of Multiple Information Sources
W4F	World Wide Web Wrapper Factory
Web	Rede - World Wide Web
WHISK	Is a general rule extraction system which learns regular expressions as extraction patterns
WIEN	The Wrapper Induction ENvironment
WWW	World Wide Web
XML	eXtensible Markup Language – linguagem de marcação extensível
XWRAP	eXtensible Wrapper Generation System

Resumo

Existem vários trabalhos na área de extração de dados semi-estruturados, usando diferentes técnicas. As soluções de extração disponibilizadas pelos trabalhos existentes são direcionadas para atenderem a dados de certos domínios, considerando-se domínio o conjunto de elementos pertencentes à mesma área de interesse. Dada a complexidade e a grande quantidade dos dados semi-estruturados, principalmente dos disponíveis na World Wide Web (WWW), é que existem ainda muitos domínios a serem explorados.

A maior parte das informações disponíveis em sites da Web está em páginas HTML. Muitas dessas páginas contêm dados de certos domínios (por exemplo, remédios). Em alguns casos, sites de organizações diferentes apresentam dados referentes a um mesmo domínio (por exemplo, farmácias diferentes oferecem remédios). O conhecimento de um determinado domínio, expresso em um modelo conceitual, serve para definir a estrutura de um documento.

Nesta pesquisa, são consideradas exclusivamente tabelas de páginas HTML. A razão de se trabalhar somente com tabelas está baseada no fato de que parte dos dados de páginas HTML encontra-se nelas, e, como conseqüência, elimina-se o processamento dos outros dados, concentrando-se os esforços para que sejam processadas automaticamente.

A pesquisa aborda o tratamento exclusivo de tabelas de páginas HTML na geração das regras de extração, na utilização das regras e do modelo conceitual para o reconhecimento de dados em páginas semelhantes. Para essa técnica, foi implementado o protótipo de uma ferramenta visual denominado Gerador de Regras de Extração e Modelo Conceitual (GREMO). GREMO foi desenvolvido em linguagem de programação visual Delphi 6.0.

O processo de extração ocorre em quatro etapas: identificação e análise das tabelas de informações úteis em páginas HTML; identificação de conceitos para os elementos dos modelos conceituais; geração dos modelos conceituais correspondentes à página, ou utilização de modelo conceitual existente no repositório que satisfaça a página em questão; construção das regras de extração, extração dos dados da página, geração de arquivo XML correspondente aos dados extraídos e, finalmente, realimentação do repositório.

A pesquisa apresenta as técnicas para geração e extração de dados semi-estruturados, as representações de domínio exclusivo de tabelas de páginas HTML por meio de modelo conceitual, as formas de geração e uso das regras de extração e de modelo conceitual.

Palavras-chave: dados semi-estruturados, tabelas, modelo conceitual, regras de extração, extração de dados, XML.

TITLE: "THE GENERATION OF DATA EXTRACTION RULES FOR HTML PAGES"

Abstract

There are a number of studies in the semistructured data area, which use different techniques. The extraction solutions which are available from existing studies aim at attending the data of certain domains, when one considers a domain the group of elements which belong to the same area of interest. Given the complexity and large amount of semistructured data available on the World Wide Web (WWW), there are still many domains to be explored.

The major part of the information which is available on sites on the Web is on HTML pages. Many of these pages contain data from certain domains (for example, medicine). In some cases, sites from different organizations present data referring to the same domain (for example, different pharmacies offering medicines). Knowledge about a determined domain, expressed in an conceptual model, can define the structure of a document.

This work considers exclusively tables on HTML pages. The reason for only working with tables is based on the fact that part of the data from HTML pages are on them, and, as a result, one eliminates the processing of other data, thus concentrating efforts for them to be automatically processed.

This research addresses the exclusive treatment of tables on HTML pages for the generation of extraction rules, in the use of these rules, and on the use of the conceptual model for the recognition of data on similar pages. In order to use this technique a prototype of a visual tool, called Extraction of Rules Generator and Conceptual Model (GREMO) was implemented in the Delphi 6.0 visual programming language.

The extraction process is made up of four steps: the identification and analysis of useful information on HTML pages; the identification of concepts for the elements for the conceptual models: the generation of conceptual models which correspond to the page, or the utilization of an existing conceptual model in the repository which satisfies the page in question; the construction of extraction rules, the extraction of data from the page, the generation of the XML file which corresponds to the data extracted and, finally, the realimentation of the repository.

This research presents the techniques for the generation and extraction of semistructured data, of the exclusive representation of the domain of tables on HTML pages through the conceptual model, the generation forms and the use of the extraction rules and of the conceptual model.

Keywords: semistructured data, tables, conceptual model, extraction rules, data extraction, XML.

1 Introdução

Para se processarem dados disponíveis na Web, especificamente de páginas HTML, é necessário executar um processo chamado “extração”. Por extração entende-se a retirada de parte relevante das informações de fontes de dados. Em nosso caso, as fontes de dados são páginas HTML. As informações contidas em páginas HTML são consideradas dados semi-estruturados [ABI 2000], [BAE 99], [BUN 97], [BUN 97a], pois não possuem um esquema definido. Por essa razão, precisam ser elas extraídas e modeladas para poderem ser manipuladas por técnicas e ferramentas já’utilizadas em informações estruturadas armazenadas em bancos de dados.

A maior concentração de informações disponíveis na Web encontra-se em páginas HTML. As páginas são construídas com dois tipos de informações: as de formatação e as de conteúdo. As informações de formatação servem para apresentar e definir a localização dos elementos na página e são compostas por elementos bem diversificados como: controles da linguagem, textos, figuras, sons e fotos. Por outro lado, as informações de conteúdo, neste trabalho denominadas de ***informações úteis***, são utilizadas para se identificarem ou qualificarem objetos. As informações que os usuários desejam são as informações úteis, e, para que eles possam manipulá-las, é necessário extraí-las.

No universo de páginas HTML em *sites* da Web, muitas delas contêm dados de certos domínios (conjunto de elementos pertencentes à mesma área de interesse, por exemplo, remédios) . Em alguns casos, *sites* de organizações diferentes apresentam os mesmos domínios (por exemplo, farmácias diferentes oferecem remédios). A estruturação do conhecimento de um determinado domínio, expresso em um modelo conceitual, serve para definir a estrutura de um documento.

Nesta pesquisa, são consideradas exclusivamente tabelas de páginas HTML. A razão de se trabalhar somente com elas está baseada no fato de que a maior parte dos dados de páginas HTML encontra-se em tabelas, e, como conseqüência, elimina-se o processamento dos outros dados, concentrando-se os esforços para que sejam processadas automaticamente.

As tabelas são formadas por linhas e colunas; as linhas identificam os objetos e as colunas, os atributos ou campos (por exemplo, a Figura 1.1 abaixo ilustra uma tabela de remédios) .

Nome do remédio	Validade	Preço
Biotônico Fontoura	15/2/2005	15,00
Fort Vit	12/5/2004	13,00
Tônico da vida	13/4/2006	25,00
Vitasae	11/2/2007	5,00

FIGURA 1.1 – Tabela Remédios

Modelo conceitual e ontologia são ferramentas conceituais para se representar o conhecimento sobre um domínio, ou seja, representam os objetos, os relacionamentos entre eles, as restrições de integridade, os aninhamentos, os tipos de dados e a cardinalidade.

O domínio descrito no exemplo “remédios”, acima, dá origem à representação dos dados em um modelo conceitual que pode ser visto na Figura 1.2 abaixo.

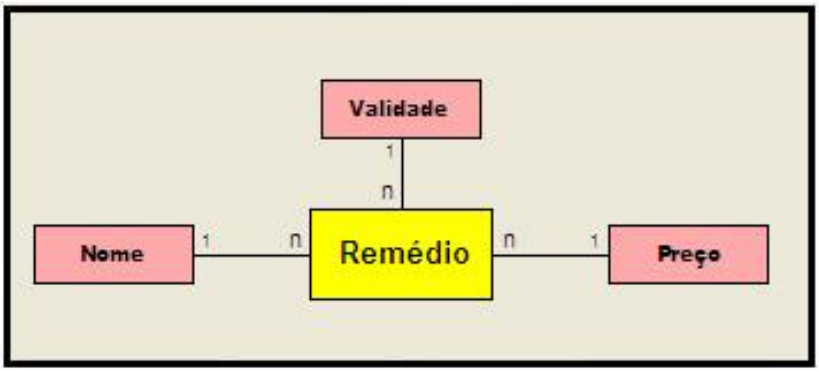

FIGURA 1.2 – Modelo Conceitual Remédios

As tabelas encontradas em páginas HTML possuem características como: quantidade de registros; quantidade de campos que compõem cada um dos registros; posição dos campos dentro dos registros; tipo de dado de cada campo; restrições de integridade e relacionamentos. Essas características podem ser usadas para constituir regras de extração as quais são armazenadas em um repositório e podem ser reutilizadas (repositório será explicado no item 3.3) .

O objetivo proposto na pesquisa é extrair informações úteis de páginas HTML referentes a um mesmo domínio, servindo, por exemplo, para se fazerem comparações de mercado (preços de remédios em farmácias diferentes). Para se tornar o processo mais eficaz, concentrando num problema específico, o foco é extrair dados representados em tabelas. Para simplificar-se a construção de regras, elas são representadas em um modelo conceitual a partir da interação do usuário e da página HTML em questão. Um dos enfoques propostos é o reuso do modelo para páginas de outros *sites* do mesmo domínio.

Esta pesquisa aborda o tratamento exclusivo de tabelas de páginas HTML na geração das regras de extração, na utilização das regras e de modelo conceitual para o reconhecimento de dados em páginas semelhantes. Para essa técnica, foi implementado, por meio deste trabalho, o protótipo de uma ferramenta visual denominada GREMO, nome que será referenciado no restante do texto.

Vários grupos de pesquisa trabalham com o problema de extração. Os trabalhos podem ser classificados[LAE 02a] em: Extração de dados com base em Ontologia [EMB98], [EMB98a], [EMB99b]; Geração de Wrappers [ASH 97], [ATZ 97], [GRU 98], [HAM 97], [KUS 97], [MUS 98], [MUS 99]; Manipulação com Linguagem Natural [COH 99], [FRE 98], [MUS 98] e Extração em páginas específicas [ABA 99], [KOR 98], [LID 99], [LIM 99].

A extração de dados semi-estruturados ganhou com essas técnicas grandes contribuições, apesar de apresentar alguns pontos fracos. A maioria deles ([ATZ 98], [ATZ 98a], [ADE 99], [ADE 98], [PAP 95], [HAM 98], [LAE 99], [LAE 99A], [LAE 00] e [SIL 01]) exige alto grau de conhecimento dos usuários para se construírem as regras de extração. Exigem também conhecimento de técnicas e ferramentas de manipulação de bancos de dados. Modificações na representação dos dados podem requerer novas definições no modelo para se possibilitar a reutilização das regras. Exceto [EMB 99, EMB 98a], os demais trabalhos não associam o resultado obtido a um modelo conceitual, o que proporcionaria maior compartilhamento de regras de extração entre aplicações. Um modelo conceitual contém informações sobre o domínio que servem para se identificar os objetos ou seja, representam-se os objetos, os relacionamentos entre eles, as restrições de integridade, os aninhamentos, os tipos de dados e a cardinalidade. Os demais trabalhos não atribuem um modelo conceitual a páginas HTML.

O processo de extração proposto nesta pesquisa se desenvolve nas seguintes etapas: 1) identificação e análise das tabelas com informações úteis em páginas HTML; 2) identificação de conceitos para os elementos dos modelos conceituais; 3) geração dos modelos conceituais correspondentes à página, ou utilização de modelo conceitual existente no repositório que satisfaça a página em questão; 4) construção das regras de extração, extração dos dados da página, geração de arquivo XML correspondente aos dados extraídos e, finalmente, realimentação do repositório.

Resumidamente, pode-se dizer que as regras e o modelo conceitual são criados e armazenados em um repositório, quando o processo trata a primeira vez uma página HTML. Regras e modelos conceituais armazenados no repositório, quando submetidos a páginas HTML que as geraram, devem necessariamente reconhecê-las automaticamente, sem nenhuma intervenção de usuário. O reconhecimento de páginas HTML semelhantes é feito por meio de regras e modelos conceituais. Por se tratarem de páginas semelhantes, as regras e os modelos conceituais devem ser alterados por usuário para se ajustarem a elas.

O trabalho está organizado como segue. Na seção 2, são apresentadas as diferenças entre a principal técnica usada por este trabalho e as demais, além de seu relacionamento com trabalhos semelhantes. Na Seção 3, é apresentada a ferramenta visual para a geração de regras de extração de dados em páginas HTML. Na Seção 4, será mostrado o funcionamento do protótipo de GREMO. Na Seção 5, far-se-ão as considerações finais e as sugestões para continuidade em trabalhos futuros.

2 Técnicas Utilizadas em Extração e Regras de Extração

As ferramentas de extração e geração de regras utilizam diferentes técnicas. Esta seção descreve as características das principais técnicas utilizadas em trabalhos semelhantes, considerando-se a classificação [LAE 2002a]. No final da seção, é apresentado um comparativo entre todas as técnicas.

2.1 Técnica baseada em HTML

As ferramentas baseadas em HTML, antes de se executar o processo de extração, transformam o documento em uma árvore, analisando-o gramaticalmente para se encontrar uma representação que reflita a hierarquia de suas *tags*. Em seguida, as regras de extração são geradas semi-automaticamente ou automaticamente e são aplicadas à árvore. Algumas ferramentas baseadas nessa técnica são W4F [SAH 2001], XWRAP [LIU 2000], e RoadRunner [CRE 2001].

2.2 Linguagem para desenvolvimento de *Wrappers*

Os *wrappers* são ferramentas de consultas e atualização em uma fonte de dados. O seu funcionamento baseia-se na especificação de quais dados serão extraídos, como extraí-los e como serão estruturados e apresentados. Uma outra característica deles é trabalhar sobre a regularidade dos dados para se encontrar aqueles que são mais relevantes. As operações realizadas por um *wrapper* são especificadas por meio de uma linguagem de extração específica.

A geração de *wrappers* teve como suas primeiras contribuições o desenvolvimento de linguagens especiais, projetadas para ajudar os usuários a construí-los. Foram propostas, como linguagens alternativas para essa tarefa, linguagens de aplicações gerais como Perl e Java. As principais ferramentas da abordagem são: Minerva[CRE 98], TSIMMIS[HAM 97a], e Web-OQL [ARO 98].

2.3 Técnica baseada em Processamento de Linguagem Natural (*NLP-based*)

O processamento de Linguagem Natural é uma técnica usada por várias ferramentas para aprender regras de extração, a fim de extrair informações

úteis em documentos. As ferramentas normalmente aplicam essas técnicas para construir relações entre frases e elementos da oração, derivando-se, assim, as regras de extração. Tais regras estão baseadas em restrições sintáticas e semânticas que ajudam a identificar a informação pertinente dentro do documento. As ferramentas baseadas em linguagem natural são normalmente mais adequadas a páginas da Web, construídas em textos livres, como listas de trabalho, anúncios de aluguel de apartamentos, anúncios de seminário, etc. Os principais trabalhos na área são: RAPIER [CAL 99], SRV [FRE 2000], e WHISK [SOD 99].

2.4 Técnicas de Indução de *Wrapper*

As ferramentas de Indução de *wrapper* geram regras de extração baseadas em delimitadores a partir do estudo de um conjunto de exemplos. A principal distinção entre essas ferramentas e as *NLP-based* é que elas não dependem de restrições lingüísticas, mas das características de formatação, que, implicitamente, delineiam a estrutura dos pedaços de dados encontrados. Isso faz tais ferramentas mais adequadas para documentos HTML do que aquelas baseadas em linguagem natural. Adotam essa abordagem: WIEN [KUS 2000], SoftMealy [HSU 98], e STALKER [MUS 2001].

2.5 Técnicas baseadas em modelo

Nessa categoria de ferramentas, encontram-se aquelas baseadas na especificação dos objetos, feitas por usuário, que servem de modelo. A especificação dos objetos compõe a estrutura do documento. A técnica procura porções de dados de páginas da Web a partir da estrutura especificada, ou seja, o processo consiste em ajustar porções de dados de páginas da Web àquela estrutura. A identificação dos objetos nas páginas é executada com algoritmos próprios das ferramentas baseadas em modelo, combinados com os algoritmos próprios das ferramentas de indução de *wrappers* com a utilização da estrutura proposta. Ferramentas que seguem essa técnica são: NoDoSE [ADE 98], DEByE [LAE 02, RIB 99] e [SIL 01].

2.6 Técnicas baseadas em ontologia

Todas as abordagens descritas anteriormente contam com as características da estrutura da apresentação dos dados dentro do documento para geração de regras ou padrões, e para realização da extração. Uma ontologia é utilizada para representar o conhecimento de um determinado domínio. Ela serve para proporcionar o compartilhamento das informações por diversas aplicações. A sua construção exige uma cuidadosa tarefa feita manualmente por um perito no domínio de ontologia. Para uma aplicação de domínio específico, uma ontologia é usada para a localização de constantes na página e a construção de objetos com ela. Quanto maior a sua representatividade, mais automatizada é a extração. Além disso, em um

mesmo domínio de aplicação, podem estar contidas páginas de muitas fontes distintas. A ferramenta principal baseada em ontologia foi desenvolvida por: Brigham Young University (BYU), Data Extraction Group [EMB 99a].

2.7 Comparativo entre a Ferramenta Baseada em Tabelas e as demais

As ferramentas estudadas nessa seção, na sua totalidade, trabalham na geração de regras e extração de dados semi-estruturados com informações da Web, e possuem técnicas próprias. Como apresentado na seção 1, páginas HTML dominam a massa de dados disponíveis na Web e muitas delas contêm tabelas.

O que precisa ser feito e não está contemplado nos demais trabalhos estudados é a concentração de esforços na análise das tabelas de páginas HTML, no sentido de se estabelecerem os seguintes processos:

- Geração de regras de extração semi-automaticamente;
- Geração de modelo conceitual automaticamente;
- Utilização de regras e de modelo conceitual em páginas semelhantes.

As técnicas e ferramentas aqui estudadas são as mais representativas, pois constituem as principais e mais recentes. As devidas comparações estão descritas como na seqüência abaixo:

a) A nossa abordagem deseja processar especificamente páginas HTML, extraindo-se delas apenas tabelas e aceitando-se tabelas aninhadas. As ferramentas baseadas em HTML (W4F, XWRAP, Road Runner) transformam todo o documento em árvore, que é analisada gramaticalmente, a fim de se encontrar uma representação que reflita a hierarquia de suas *tags*. O que esta ferramenta pretende é criar um modelo conceitual apenas das tabelas, com o domínio que as identifique.

b) As ferramentas para desenvolvimento de *wrappers* (Minerva, TSIMMIS e Web-OQL) utilizam linguagens como Perl e Java para codificar as regras de extração, ou seja, as regras são efetivamente programadas pelo usuário. A ferramenta aqui proposta não necessita de programação.

c) As ferramentas baseadas em técnicas de linguagem natural (RAPIER, SRV e WHISK) não tiram proveito de delimitadores como os que aparecem em tabelas HTML. Elas se baseiam no conteúdo das palavras que formam frases e não em delimitadores. Elas também constroem a relação entre frases e elementos da oração, filtrando *tags* de áudio, e semântica léxica, derivando-se assim as regras de extração. As suas restrições sintáticas e semânticas ajudam a identificar a informação pertinente dentro de um

documento. Por essas razões, a aplicação de Linguagem Natural não é adequada ao tipo de página considerada em nossa ferramenta.

d) Interessa-nos uma ferramenta que trabalhe com modelo conceitual e características de tabelas. Não é o caso das ferramentas de indução de *wrapper* (WIEN, SoftMealy e STALKER), que se baseiam em delimitadores para se marcar a estrutura dos pedaços de dados encontrados, determinados pelas características de formatação.

e) O modelo conceitual defendido em nossa ferramenta baseia-se em características pertinentes exclusivamente a tabelas aninhadas ou não. A diferença para os modelos utilizados nas outras ferramentas, baseadas em modelo (NoDoSE e DEByE), está na identificação de objetos de interesse, para localização de porções de páginas de dados da Web que, implicitamente, se ajustam àquele modelo. Este é formado de acordo com um cenário de primitivas de modelagem. O paradigma por meio de exemplos da ferramenta DEByE seria um modelo adequado. Entretanto, ela não é voltada para HTML, nem para tabelas e, sim, para textos com delimitadores em geral.

f) A técnica utilizada pelas ferramentas baseadas em ontologia (BYU), para localizar constantes em páginas HTML e construir objetos, constitui uma aplicação de domínio específico. Para a presente ferramenta, essa técnica não atende, tendo em vista o uso apenas de tabelas.

Resumindo, das ferramentas aqui estudadas, nenhuma delas é específica para trabalhar com tabelas. As regras de extração, geradas por nossa ferramenta, combinam idéias e técnicas de várias abordagens: ferramentas HTML; ferramentas baseadas em modelo e ferramentas baseadas em ontologia.

Nesta seção, foram abordados os aspectos relacionados ao estado da arte. Na seção 3, está descrito o processo de extração realizado por GREMO. Ele gera regras de extração e cria um modelo conceitual para aplicação em páginas HTML.

3 Ferramenta Visual para Geração de Regras de Extração e Extração de Dados em Páginas HTML

A ferramenta desenvolvida tem, como principal foco, a geração de regras e extração de dados semi-estruturados, baseada em Modelo Conceitual a partir, especificamente, de dados representados de tabelas (aninhadas ou não), reconhecidas em páginas HTML. A agregação de modelo conceitual do domínio [SIL 2001] possibilita o reconhecimento das relações semânticas dos objetos, além dos aspectos sintáticos.

A arquitetura geral, criada para essa solução, pode ser vista na Figura 3.1. A ferramenta recebe uma página HTML como entrada; a partir dela, dois processos podem ser realizados: o primeiro, a ferramenta varre o repositório em busca de um modelo conceitual e de regras que se enquadrem ou se ajustem, total ou parcialmente, aos dados das tabelas da página em questão; o segundo é decorrente de entradas a partir da interação com usuário, o qual gera, no repositório, as regras e o modelo conceitual correspondente à página. Em ambos os casos, o repositório é realimentado com as informações resultantes dos processos.

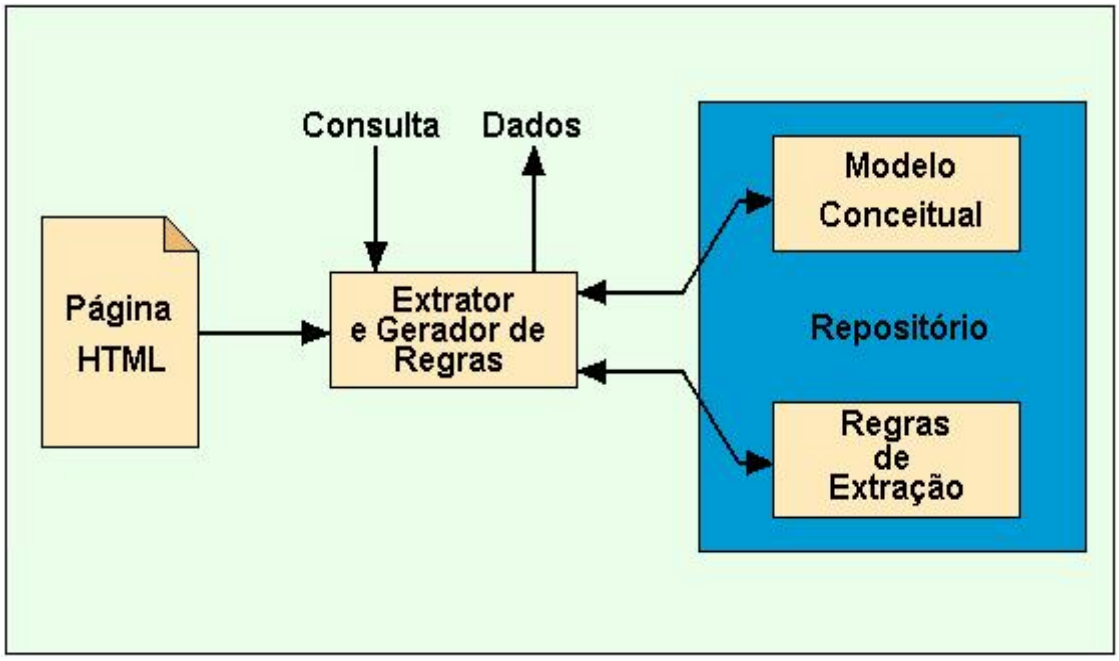

FIGURA 3.1 – Arquitetura geral da ferramenta

Todo o processo realizado pela ferramenta está descrito nessa seção. Para a experimentação da ferramenta foi desenvolvido um protótipo e, na seção 4, será mostrado um exemplo passo a passo.

FIGURA 3.2 – Página Parintins

Para se ilustrar o funcionamento da ferramenta, foram utilizadas duas páginas. A primeira, uma das diversas páginas do *site* Parintins, com informações bem diversificadas, apresentando-se textos, tabelas de controle, imagens, etc, cujos domínios são discos e livros. A segunda, uma das páginas do *site* Ouro Negro Transportes, uma página regular, contendo tabelas aninhadas, sendo seu domínio de dados endereços de matriz e filiais. As páginas são: a) Uma página de venda de livros e CD's (Figura 3.2), que será chamada "Parintins" no restante do texto; b) Uma página de Transportadora de cargas, com endereços da matriz e suas filiais (Figura 3.3), identificada no restante do texto como "Ouro Negro".

FIGURA 3.3 – Página Ouro Negro

A presente ferramenta utiliza um repositório para o armazenamento de suas informações. O funcionamento da ferramenta dá-se em duas grandes fases. A primeira, apresentada no item 3.1, descreve o processo necessário para serem geradas regras e o modelo conceitual no repositório, a partir de características de tabelas de páginas HTML. O item 3.2 descreve o processo de utilização do modelo conceitual, das regras de extração e descreve o formato de todas as informações obtidas e armazenadas no repositório. A segunda, apresentada no item 3.3, descreve o processo que permite identificar as regras e o modelo conceitual para uma tabela automaticamente a partir do repositório. Esse processo pode fazer resultar duas situações: o modelo conceitual e as regras de extração se enquadram totalmente à página e, nesse caso, não há necessidade de intervenção de usuário para a extração; caso

haja necessidade de intervenções do usuário, a ferramenta permite executá-las, reduzindo significativamente o processo.

O detalhamento das tabelas usadas pela ferramenta, armazenadas no repositório, serão descritas no item 3.3.

3.1 Criação de Regras de Extração e do Modelo Conceitual

O processo de criação de regras de extração se desenvolve ao longo de cinco etapas, que são realizadas por procedimentos manuais e automáticos. O resultado obtido com eles é armazenado no repositório como regra de extração e o modelo conceitual. A partir da carga da página HTML, concisamente, o processo de geração de regras segue as seguintes etapas:

1 – Identificação de tabelas

A identificação de tabelas em páginas HTML é realizada pela execução de três processos automáticos distintos (embutidos na ferramenta por meio de regras): primeiramente, são desconsideradas todas as informações que não pertencem a tabelas, sejam elas de formatação ou não; em segundo lugar, são desconsideradas todas as tabelas de formatação (controles da linguagem, textos, figuras, sons e fotos); em terceiro lugar, são desconsideradas as tabelas compostas por apenas um registro (essa opção é uma decisão dessa ferramenta baseada em estudo de tabelas, sendo constatada a situação em tabelas de formatação).

A execução dos processos apresenta como resultado exclusivamente tabelas com ***informações úteis***.

O passo está explicado na seção 3.1.1

2 – Determinação de conceitos

Continuando o processo, das palavras que compõem os campos dos registros das tabelas resultantes do processo anterior, algumas delas podem ser palavras pertencentes a formatos-padrão ou a palavras-chave. Em qualquer dos casos, elas identificarão conceitos, e os conceitos identificam os campos. Entende-se por conceitos, em todo este trabalho, a atribuição de nomes aos objetos léxicos e não-léxicos. A representação das informações nos documentos são os objetos léxicos. As informações não representadas diretamente nos documentos são os objetos não-léxicos, complexos ou compostos. Para exemplo de objetos léxicos e não-léxicos apresentamos, na figura 3.4, o modelo conceitual da

página “Ouro Negro”, sendo objetos léxicos: cidade, endereço, bairro, fone-fax, e-mail, gerente e tipo; e objetos não-léxicos: matriz e filial.

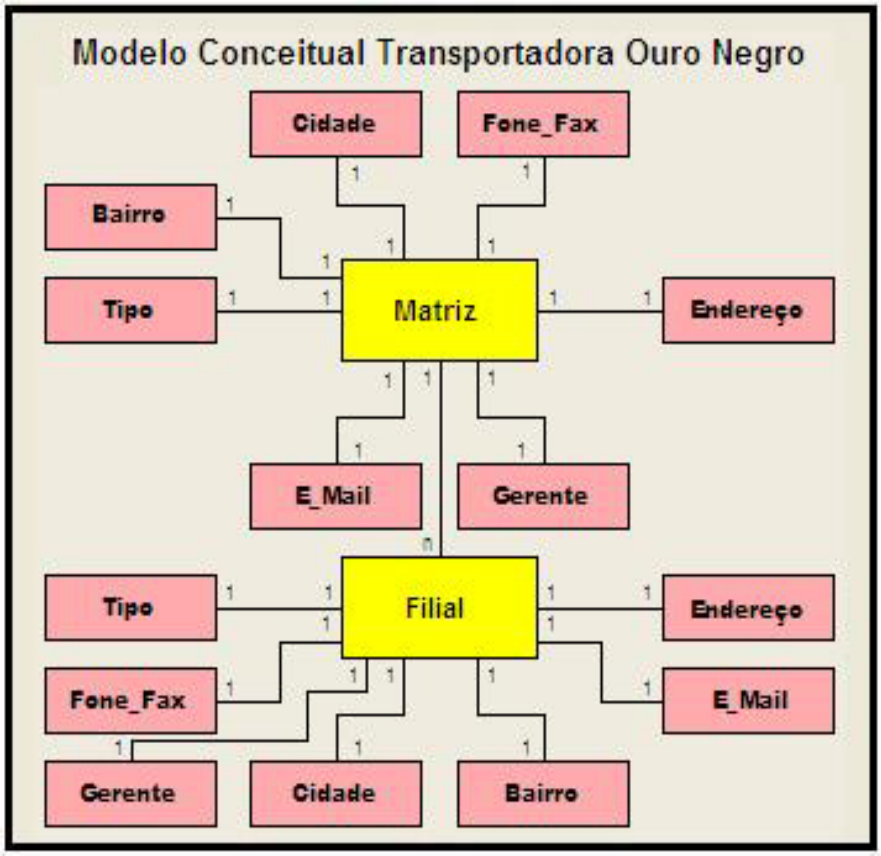

FIGURA 3.4 - Modelo Conceitual Ouro Negro

Os conceitos são obtidos de duas formas distintas:

- Primeiramente, as palavras são submetidas uma a uma à validação por algoritmos chamados de expressões regulares, que tem por objetivo reconhecê-las como formatos-padrão (CPF, CGC, PIS, Hora etc). O passo está explicado na seção 3.1.3.

- Não encontrando nenhuma palavra que se enquadre em formatos-padrão, o próximo passo é tentar reconhecer as palavras uma a uma comparando-as com as palavras-chave, contidas na tabela de palavras-chave do repositório. O passo está explicado na seção 3.1.4.

O usuário valida ou altera manualmente os conceitos gerados. A alteração manual de conceitos poderá causar o aumento do repositório, no caso em que o novo conceito atribuído não faça parte dele; por exemplo, um conceito “*endereço*” é gerado

automaticamente a partir da palavra-chave "*Rua*", porém, em função do contexto, o usuário altera para "Rodovia"; caso esse conceito não exista, então ele é inserido no repositório.

Uma tabela contém registros, cada registro contém campos, cada campo deve ser único para todos os registros. Como aos campos são automaticamente atribuídos conceitos, identificados por formatos-padrão ou palavras-chave, eles podem ser diferentes de um registro para outro. Nesse caso, os conceitos devem ser alterados, através da interação do usuário, para que todos os registros da mesma tabela sejam iguais. Por exemplo, todos os registros das filiais da transportadora Ouro Negro devem ser formados pelos campos: cidade, endereço, bairro, fone-fax, e-mail, tipo e gerente. O passo está explicado na seção 3.1.2.

3 – Construção do modelo conceitual

Agora, com todos os conceitos definidos para os campos dos registros das tabelas encontradas, os modelos conceituais são gerados automaticamente. Os modelos conceituais agregam novos elementos, os elementos não-léxicos e os relacionamentos, para os quais a sua validação ou a alteração manual é feita pelo usuário. O passo está explicado na seção 3.1.5.

4 – Extração dos dados

O processo, nessa etapa, extrai os dados da página e gera o arquivo XML, considerando-se os procedimentos acima. O passo está explicado na seção 3.1.6.

5 – Armazenamento no repositório

Finalmente, com os dados extraídos da página e, com informações geradas no modelo conceitual, o processo termina realimentando o repositório. O repositório, a partir desse momento, proporciona a extração de dados em páginas semelhantes (explicado anteriormente) àquelas já armazenadas nele. O passo está explicado na seção 3.1.7.

3.1.1 Tabelas HTML consideradas no processo

A principal característica da presente ferramenta é trabalhar com tabelas, possivelmente aninhada de páginas HTML e que contenham informações úteis. Consideram-se informações úteis todas aquelas de interesse do visitante da página HTML. Por exemplo: produtos e preços; livros com resumos e autores; notícias; índices financeiros; resultados de testes; entre outros assuntos. A obtenção de tais tabelas dá-se com a desconsideração de todas as outras informações da página, as quais são de formatação.

As informações de formatação proporcionam aos usuários das páginas HTML maior facilidade no seu entendimento, além de cativarem e motivarem pelo interesse na aquisição de suas ofertas. O objeto desse trabalho é tão somente a extração de informações úteis, portanto as tabelas de formatação precisam ser desconsideradas.

As páginas HTML são construídas, como explicado anteriormente, com informações úteis e de formatação. As informações podem estar contidas em tabelas ou não. Uma tabela pode ser reconhecida inteira ou em parte. Considera-se elemento todas as informações que não são tabelas, e parte delas que sejam informações de formatação ou registros. Para fins deste trabalho, o conteúdo de uma página é classificado em:

- Tabelas consideradas – serão consideradas, no processo, as tabelas que possuem informações úteis;

- Tabelas desconsideradas – são desconsideradas pelo processo todas as tabelas de formatação ou tabelas com apenas um registro (tabelas com um registro normalmente são tabelas de formatação e são desconsideradas automaticamente; porém, se o usuário desejar, ele pode interagir e devolvê-la ao processo, ou seja, desfazer a ação automática que a desconsiderou);

- Elementos considerados – são consideradas pelo processo informações úteis contidas nas tabelas consideradas;

- Elementos desconsiderados – são desconsideradas pelo processo todas as informações de formatação pertencentes a tabelas, todas as informações que não lhes pertencem e registros irregulares nelas. Consideram-se registros irregulares todos aqueles que possuam quantidade de campos diferentes de mais de 50% dos da tabela em questão (por exemplo, uma tabela que possua 5 registros com os campos nome, vencimento e valor, e 1 registro com os campos endereço e cep; o registro com endereço e cep é irregular).

A página Parintins foi construída com informações úteis e com informações de formatação. A seguir são diferenciadas por cores, no programa fonte HTML da página Parintins, essas informações. A identificação das cores e seus significados são mostrados na Figura 3.5 abaixo. O exemplo a seguir contém apenas parte do programa fonte HTML da página Parintins.

FIGURA 3.5 – Legenda para o programa fonte HTML Parintins

Exemplo:

```
<! DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<! -- saved from url=(0040)http://parintins.net/shopping/index.php3 -->
<HTML><HEAD><TITLE>Parintins Shopping - Cultura da Amazônia!</TITLE>
<META content="text/html; charset=iso-8859-1" http-equiv=Content-Type>
<STYLE type=text/css>A:hover {COLOR: #000000}
                              A:hover FONT {COLOR: #ff6600}
                              A:link FONT {COLOR: #000000}
                              A {TEXT-DECORATION: none}
                              A:hover {TEXT-DECORATION: underline}
                                .input {BACKGROUND: #ffffcc;
                                        COLOR: #ff0000;
                                        FONT-WEIGHT: normal}
</STYLE>
<META content="MSHTML 5.00.2614.3500" name=GENERATOR></HEAD>
<BODY  aLink=#00ff00 background=index_arquivos/bg_br.gif bgColor=#ffffff
       leftMargin=0 link=#006600 topMargin=0 vLink=#006600
       marginheight="0">
<TABLE bgColor=#00ff00 border=0 cellPadding=0 cellSpacing=0 width=780>
 <TBODY>
 <TR>
  <TD><FONT color=#000000 face="Verdana, Arial, Geneva"
      size=1><B>  <A href="http://parintins.net
      /shopping/index.php3?
```

```
        mn=info&amp;info=shopping&amp;n=2052588243&amp;f=1664221204&
        amp; d=20020217&amp;l=pt"
        >Como comprar</A>    <A href="http://parintins.net
        /shopping/index.php3?mn=info&amp;info=crypt&amp;
        n=2052588243&amp;
         f=1664221204&amp;d=20020217&amp;l=pt"
        >Sobre segurança</A></FONT></B></TD>
    <TD align=right><FONT color=#000000 face="Verdana, Arial, Geneva"
        size=1><B>   <A href="http://parintins.net/shopping
        /basket.php3?n=2052588243&amp;
        f=1664221204&amp;d=20020217&amp;l=pt"
        >Minha compra</A> </FONT></B></TD></TR>
  <TR>
    <TD><IMG height=1 src="index_arquivos/spacer.gif" width=390></TD>
    <TD><IMG height=1 src="index_arquivos/spacer.gif"
     width=390></TD></TR></TBODY></TABLE>
<TABLE bgColor=#ffffff border=0 cellPadding=0 cellSpacing=0 width=780>
  <TBODY>
  <TR>
    <TD><IMG height=4 src="index_arquivos/spacer.gif"
          width=780></TD></TR></TBODY></TABLE>
<TABLE bgColor=#ffffff border=0 cellPadding=0 cellSpacing=0 width=780>
  <TBODY>
  <TR>
  <TR vAlign=top>
    <TD align=middle vAlign=center width=140><A href="http:
        //parintins.com/"
        target=top><IMG alt=http://parintins.com border=0
        src="index_arquivos/logogray.gif"></A><BR><FONT color=#000000
        face="verdana,Arial, Geneva" size=1><B><A
        href="mailto:shopping@parintins.com"
        >shopping@parintins.com</B></A></FONT></TD>
    <TD align=middle vAlign=top><IMG height=1
```

```
        src="index_arquivos/spacer.gif"
        width=10><IMG src="index_arquivos/titulo2001.gif"><BR>
        <IMG height=1 rc="index_arquivos/spacer.gif" width=10><IMG
        src="index_arquivos/titulo2001pt.gif"><BR><B><FONT color=#000000
        face="verdana, Tahoma, Verdana, Arial, Geneva" size=1>Desde 1998
        distribuindo produtos da Amazônia para o Mundo<BR>
        <FONT color=#ff6600><A href="http://parintins.net/shopping
        /index.php3?n=2052588243&amp;
        f=1664221204&amp;d=20020217">Início</A>
        +<A href="http://parintins.net/shopping/index.php3?n=2052588243&amp;
        f=1664221204&amp;d=20020217&amp;nl=en">ENGLISH</A>
        +<A href="http://parintins.net/shopping/index.php3?
        n=2052588243&amp;
         f=1664221204&amp;d=20020217&amp;nl=es">ESPAÑOL</A>
        </FONT></B></FONT></TD>
  <TD align=middle vAlign=center width=140>
   <TABLE cellPadding=0 cellSpacing=0>
    <TBODY>
    <TR>
     <TD align=middle><IMG alt=Brasil! src="index_arquivos
     /br.gif"></TD></TR>
    <TR>
     <TD align=middle><FONT color=#000000 face="Verdana, Arial, Geneva"
     size=2><B>Amazônia é</B></FONT></TD></TR>
    <TR>
     <TD align=middle><FONT color=#000000 face="Verdana, Arial, Geneva"
     size=2><B>Brasil</B></FONT></TD></TR>
     </TBODY></TABLE></TD></TR>
 <TR>
  <TD bgColor=#ffff99 colSpan=3><IMG height=4
         src="index_arquivos/spacer.gif"width=1>
          </TD></TR></TBODY></TABLE>
         <!-- barras coloridas -->
```

O primeiro processo realizado pela ferramenta consiste em desprezar as informações de formatação, as tabelas de formatação, além de organizar as tabelas resultantes, da seguinte forma:

- Somente são tratados elementos contidos entre os elementos ***"<td>"*** e ***"</td>"***que apareçam em tabelas;

- São descartados elementos de controle, ou seja, elementos cujos conteúdos sejam: ***" "***, e ***" "*** (espaço em branco ou conjunto vazio);

- São numeradas as tabelas resultantes seqüencialmente a partir do número um, agregando endentamento também seqüencial para os aninhamentos encontrados.

O processo é totalmente automático, estando os algoritmos embutidos na ferramenta. A aplicação do processo na página Parintins retorna as tabelas listadas na Figura 3.6. A figura apresenta as seguintes situações de tabelas e registros da página:

- As tabelas com fundo azul-claro representam tabelas de formatação, não identificadas como tabelas de formatação por apresentarem conteúdos equivalentes a informações úteis.

- Os registros listados com fundo amarelo representam registros irregulares, pois são compostos por apenas um campo, também ainda não tratados.

- As tabelas e registros com fundo branco são de informações úteis.

Na figura 3.6, não aparecem as informações que foram eliminadas automaticamente, sendo elas:

- As informações de formatação que não são tabelas. Essas informações referem-se a dados do navegador e a cabeçalho da página.

- As tabelas de formatação e sem conteúdo.

Tabela	Registro	Campo	Informação útil
1	1	1	Como comprar
		2	Sobre segurança
	2	1	Minha compra
3	1	1	shopping@parintins.com
	2	1	Desde 1998 distribuindo produtos da Amazônia para o Mundo
		2	Início
		3	ENGLISH
		4	ESPAÑOL
3.1	1	1	Amazônia é
	2	1	Brasil
5	1	1	-->
		2	Use seu cartão VISA em nosso servidor seguro.
		3	Você também pode pagar com depósito bancário
		4	Fone/Fax
		5	(0xx92) 648-1800
6	1	1	Se você não encontrou um produto, faça seu pedido:
	2	1	E-mail:
7.1	1	1	Arlindo Jr
	2	1	Raízes Caboclas
	3	1	Novo CD, inclui o sucesso
		2	Só pode ser Você
		3	...
		4	R$ 20.90
	4	1	Novo CD da Banda...
		2	R$ 16.95
7.2.1.1	1	1	Garantido 2001
		2	Toadas oficiais 2001 - cd duplo
		3	R$ 23.95
7.2.2	1	1	Caprichoso: Amor e Paixão
		2	Novo cd do Caprichoso!!!
		3	R$ 17.95
7.2.3	1	1	Carlos Batata
	2	1	Novo CD, inclui o sucesso
		2	Sem Juízo
		3	R$ 16.95
7.3	1	1	Folclore Político do Amazonas
		2	Episódios hilariantes da história política regional
		3	R$ 29.90
	2	1	Garantido Antológico
		2	CD reúne toadas de todos os tempos do Bumbá Garantido, na voz de David
		3	R$ 18.95
	3	1	Os Intérpretes da Amazônia
		2	Análise de grandes obras que descreveram a Amazônia
		3	R$ 29.90
	4	1	Estatutos do Homem
		2	Edição especial trilingüe do célebre poema de Thiago de Mello.
		3	R$ 35.00
	5	1	Bar do Boi
		2	CD de aniversário de 10 anos do Bar do Boi Caprichoso,Manaus
		3	R$ 15.95
	6	1	Universo Mítico Ritual do Povo Tukano
		2	Cultura e crenças das tribos Tikunas
		3	R$ 37.90
8	1	1	Dúvidas? Sugestões?
		2	shopping@parintins.com
9	1	1	© Opera House
		2	(1998-2001)

FIGURA 3.6 – Tabelas reconhecidas

Dentre as tabelas reconhecidas pelo processo acima, algumas são de formatação e outras contêm registros irregulares. As tabelas e os registros precisam ser descartados. Para se executar esse processo, são realizados dois procedimentos: manual e automático. Os procedimentos são descritos na ordem:

- O usuário interage com a ferramenta selecionando as tabelas de formatação para descarte. O critério para indicar as tabelas é a avaliação visual da informação. Em nosso exemplo, Figura 3.6, o usuário deverá informar para descarte os números de tabelas e registros, conforme a Tabela 3.1 a seguir.

TABELA 01 – Números de tabelas para descarte

Tabela	Registro
1	1
3	2
5	1
7.1	3 e 4
8	1
9	1

- São desconsideradas, automaticamente por algoritmo embutido na ferramenta, tabelas que possuem um ou mais registros com apenas um campo em cada um, e cujo conteúdo do campo seja texto, isto é, um campo que não satisfaça os algoritmos da ferramenta, expressões regulares (nome próprio, valor, data, hora, CGC, CPF Cep e PIS), e que não contenha palavras-chave. Por opção deste trabalho, não são considerados tais registros, pois, com base nas páginas estudadas, eles não pertencem ao domínio em que estão inseridos. Podemos ver, na Tabela 3.2 abaixo, a situação encontrada na Figura 3.6.

TABELA 3.2 – Registros com um campo

Tabela	Registro
1	2
3.1	1
6	1

- São desconsiderados registros irregulares, através da interação do usuário. Como visto acima, consideram-se irregulares todos os registros que possuam quantidade de campos diferentes de mais de 50% dos da tabela em questão (por exemplo, uma tabela

que possua 5 com os campos nome, vencimento e valor; e 1 com os campos endereço e cep. O registro com endereço e cep é irregular, portanto é desconsiderado. A opção é aplicada neste trabalho com base nas páginas estudadas). Um exemplo pode ser visto na Figura 3.6, a tabela 7.2.3, registro 1.

O resultado final da aplicação das regras da etapa anterior e desta pode ser visto na figura 3.7 abaixo.

Tabela	Registro	Campo	Informação útil
7.2.1.1	1	1	Garantido 2001
		2	Toadas oficiais 2001 - cd duplo
		3	R$ 23.95
7.2.2	1	1	Caprichoso: Amor e Paixão
		2	Novo cd do Caprichoso!!!
		3	R$ 17.95
7.2.3	2	1	Novo CD, inclui o sucesso
		2	Sem Juízo
		3	R$ 16.95
7.3	1	1	Folclore Político do Amazonas
		2	Episódios hilariantes da história política regional
		3	R$ 29.90
	2	1	Garantido Antológico
		2	CD reúne toadas de todos os tempos do Bumbá ...
		3	R$ 18.95
	3	1	Os Intérpretes da Amazônia
		2	Análise de grandes obras que descreveram a Amazônia
		3	R$ 29.90
	4	1	Estatutos do Homem
		2	Edição especial trilingüe do célebre poema de Thiago ...
		3	R$ 35.00
	5	1	Bar do Boi
		2	CD de aniversário de 10 anos do Bar do Caprichoso, ...
		3	R$ 15.95
	6	1	Universo Mítico Ritual do Povo Tukano
		2	Cultura e crenças das tribos Tikunas
		3	R$ 37.90

FIGURA 3.7 – Tabelas resultantes

Resumindo, observa-se, nessa etapa, que alguns aspectos da página Parintins foram tratados pelas regras estipuladas acima e outros deverão ser tratados nas próximas:

- As tabelas de formatação e sem conteúdo foram desconsideradas;

- As tabelas 1, 3, 3.1, 5, 6, 8 e 9 não contêm informações úteis do Shopping Parintins;
- A tabela 7.1 apresenta uma estrutura irregular em seus registros e campos;
- A tabela 7.2.3 contém um registro irregular;
- As tabelas resultantes 7.2.1.1, 7.2.2, 7.2.3 e 7.3 são compostas por informações úteis e apresentam uma estrutura regular, portanto são as tabelas consideradas no processo.

3.1.2 Criação de Conceitos do modelo conceitual

Um conceito representa, no modelo conceitual, as características gerais de um objeto. Os conceitos aqui são usados para representarem as características gerais dos conteúdos dos elementos ***"<td>"***. Os conteúdos dos elementos podem ser identificados como diferentes tipos de conceitos. Existem duas formas de associar campos (conteúdos) a conceitos: palavras-chave e expressões regulares.

A ordem de busca, convencionada neste trabalho, tem a preferência por expressões regulares. Esse critério foi adotado em razão do grande número de palavras-chave em relação às expressões regulares. Isso significa que, se a expressão regular encontra um formato-padrão, o tempo envolvido no processo de busca se reduz ao número máximo de expressões regulares, inicialmente 8 (oito). Por outro lado, a maior parte dos conceitos pode ser reconhecida por meio de palavras-chave, as quais iniciam o repositório com 77 (setenta e sete), apresentando um tempo no processo de busca, significativamente menor. Por essas razões, a opção de preferência por expressões regulares. Os números de palavras-chave e de expressões regulares citados acima, são oriundos de pesquisa nas páginas estudadas para este trabalho.

3.1.3 Expressões regulares

Uma expressão regular é formada por símbolos e chaves, e ela identifica inteiramente cadeias no conteúdo, associando-as a conceitos. Por exemplo, a cadeia "88.801-440", para se representar um cep, precisa considerar todos os caracteres entre as aspas duplas. Os algoritmos que reconhecem os conceitos utilizam expressões regulares.

A ferramenta começa utilizando alguns formatos-padrão para definição de conceitos. Os formatos-padrão aqui definidos são: nomes próprios, valores monetários, datas, horas, CGC, CPF, Cep e PIS.

Os formatos-padrão são formados por regras específicas. Por exemplo: Nome próprio é um tipo especifico de conceito que é usado para nomear pessoas, empresas, nações, povoações, montes, mares, rios, etc. A composição de nomes próprios é estabelecida por regras de nossa gramática. Valor, data, hora, cgc, cpf, cep e pis também são tipos específicos e

possuem regras em suas formações. O reconhecimento desses conceitos pode ser realizado de maneira simples, utilizando-se de expressões regulares.

A seguir são apresentadas as expressões regulares utilizadas nos algoritmos para o reconhecimento dos formatos-padrão. Ilustra-se o seu uso com exemplos.

Reconhecimento de Nomes Próprios

O processo de reconhecimento de nomes próprios consiste na análise de conteúdo de elemento ***"<td>",*** desprezados todos os rótulos nele contidos. Os nomes próprios devem satisfazer as seguintes condições: ser uma única palavra ou um conjunto de palavras, todas começando por letra maiúscula, sendo as demais de cada palavra minúsculas, e ainda todas elas separadas por espaços e unidas ou não por preposições. O exemplo da página Parintins satisfaz as condições acima.

Conteúdo: **<td>Folclore Político do Amazonas</td>**

A expressão regular para se encontrar o conceito é:

[A-Z][A-Za-z]*(([A-Z\.?] | [D,d][aeiou])?([A-z][A-Za-z]*))*

Reconhecimento dos demais formatos-padrão

O reconhecimento de valores monetários dá continuidade na varredura dos conteúdos dos elementos ***"<td>"***. São identificadas como valores monetários quaisquer cadeias que estejam formatadas como moeda. São aceitos como valores monetários todas as cadeias que tenham até seis dígitos após a vírgula, que os milhares sejam ou não separados por pontos e que a quantidade de dígitos antes da vírgula varie entre um e doze.

Os demais formatos-padrão iniciais dessa ferramenta são reconhecidos, utilizando-se processos análogos, ou seja, o formato de valor monetário é um conjunto de dígitos cuja parte inteira é separada da parte fracionária por vírgula, e a parte inteira é separada aos milhares por pontos; já, num formato de data, o dia, mês e ano podem ser separados por "/", "-", ":" ou ".", assim acontecendo com os formatos hora, cgc, cpf, cep e pis.

A seguir são apresentados os formatos-padrão com as respectivas expressões regulares e seus exemplos de conteúdos.

- Valor Monetário – Conteúdo: **<td>Entrada 2.000,00 cheque</td>**

  Expressão regular: (([0-9]{3}\.){3}[0-9]{3},[0-9]{2})*

- Data – Conteúdo: **<td>Entregar em: *05/01/2001*</td>**

  Expressão regular:

  ((0[1-9] | [12][0-9] | 3[01]) / (0[1-9] | 1[012]) / [12][0-9]{3})*

- Hora – Conteúdo: **<td>Saída às *15:00:00* no portão "J"</td>**

  Expressão regular: ([0-1][0-9] | 2[0-3]) : [0-5][0-9](:[05][0-9])?)*

- CGC – Conteúdo: **<td>Empresa *98.782.456/0001-00*</td>**

  Expressão regular: ([0-9]{2}\.[0-9]{3}\.[0-9]{3} / [0-9]{4} – [0-9]{2})*

- CPF – Conteúdo: **<td>Manuel Antonio *987.782.456-00*</td>**

  Expressão regular: (([0-9]{3}\.){2}[0-9]{3}-[0-9]{2})*

- CEP – Conteúdo: **<td>Santo Amaro *97.782-000*</td>**

  Expressão regular: ([0-9]{2}\.[0-9]{3}-[0-9]{3})*

- PIS – Conteúdo: **<td>14º Salário *923.78452.66.00*</td>**

  Expressão regular: ([0-9]{3}\.[0-9]{5}\.([0-9]{2}\.?){2})*

3.1.4 Palavras-chave

Reconhecimento de Palavras-Chave

Palavras-chave são palavras que pertencem a um idioma e que expressam o sentido de um texto intuitivamente. No presente trabalho, as palavras-chave encontradas em textos de páginas HTML servem para associá-los a conceitos, sendo que os conceitos dão aos textos um sentido mais generalizado. Por exemplo, supondo-se dois textos: "Rua Antônio Teixeira, 244" e "Av. Getúlio Vargas, 100", os quais são endereços, e suas palavras-chave, "Rua" e "Av.", necessariamente o conceito para se associar a eles deve ser "Endereço". As palavras-chave permitem identificar e associar conceitos aos textos das páginas HTML a que pertencem. O exemplo a seguir ilustra o uso de palavras-chave.

Conteúdo. . . . . : <td>*Rua* Antônio Teixeira, 244</td>

Palavra-Chave. : Rua

Conceito . . . . . : Endereço

O conteúdo do exemplo acima, "Rua Antonio Teixeira, 244", contém a palavra-chave **"Rua"**. No caso, a palavra-chave "Rua" pode ser usada para se identificar o fato de que essa cadeia de caracteres representa uma instância do conceito Endereço.

O processamento automático da ferramenta está baseado em algoritmos e na procura por palavras-chave. Para a definição das palavras-chave e dos algoritmos foram analisadas e mapeadas aqui diversas páginas HTML complexas. O repositório começa com informações oriundas das análises e vai aumentando à medida que a ferramenta vai sendo usada. A seguir, são apresentadas as palavras-chave iniciais do repositório e o funcionamento dos algoritmos usados pela ferramenta com respectivos exemplos.

Os resultados obtidos podem ser vistos na Figura 3.8, mostrada abaixo.

Palavra-Chave	**Palavra-Chave**	**Palavra-Chave**
$	Entrega	Rua
Ano	Estrada	Telefone
Assunto	Fardo	Telefone:
Aut	Fax	Texto
Author	Fax:	Título
Authora	Fls	Transportador
Autor	Fornecedor	Valor
Autor:	Hora	Venda
Autora	Homepage	.
Av	Homepage:	.
Av.	Hs	.
Avenida	Idioma	
Avenida:	ISBN	
Bairro	Lançamento	
Caixas	Livro	
Carro	Logradouro	
CD	Modelo	
CDS	N:	
CEP	Name	
CEP:	Nom	
CGC	Nome	
CI	Nome ->	
Cliente	Nome Próprio	
Cliente:	Nome:	
Combustível	Obra	
Compra	Opcionais	
Cor	Pacote	
Desconto	Pç	
Edição	Pagamento	
Edifício	Páginas	
Editora	Placa	
E-Mail	Preço	
E-Mail:	Publicação	
Encadernação	R$	
Endereço	Rodovia	

FIGURA 3.8 – Palavras-chave

Analisando-se as palavras-chave obtidas nas diferentes páginas complexas, nota-se o uso intenso de sinônimos, palavras semelhantes, palavras truncadas e palavras iguais com símbolos agregados. A classificação dessas palavras e o seu agrupamento por conceitos se fazem necessários para que os modelos conceituais possam ser utilizados em páginas semelhantes. As palavras-chave obtidas nas análises realizadas nas páginas HTML deram origem a conceitos e a seus atributos mostrados na Figura 3.9.

Conceito	Tipo	TMin	TMax	Palavras-Chave	Excluir
Autor	Char	10	40	Autor:	Sim
				Author	Sim
				Aut	Sim
				Autora	Sim
				Authora	Sim
Disco	Char	10	40	Cd	Não
				Cds	Não
Endereço	Char	20	50	Av	Sim
				Av.	Sim
				Avenida	Sim
				Avenida:	Sim
				Endereço	Sim
				Estrada	Sim
				Rodovia	Sim
				Rua	Sim
Hora	Char	8	8	Hora	Sim
				Hs	Sim
Nome	Char	5	10	Nome:	Sim
				Nome ->	Sim
				Nom	Sim
				N:	Sim
				Name	Sim

FIGURA 3.9 – Conceitos associados a palavras-chave

As colunas da Figura 3.9 qualificam as informações dos conteúdos de páginas HTML como se segue:

- **Conceito** apresenta conceitos que são identificados por uma ou mais palavras-chave em conteúdos de páginas HTML.

- **Tipo** caracteriza o tipo de dado de um campo em um registro.

- **TMin** e **TMax** contêm os tamanhos mínimo e máximo em caracteres para o campo em questão.

- **Palavras-Chave** apresentam palavras-chave propriamente ditas, as quais representam um conceito ou um sinônimo dele.

- **Excluir** permite ativar duas opções para processamento sobre as informações úteis de páginas HTML: com a opção "sim" ativada, a palavra-chave é eliminada do texto da informação útil. Para exemplificar, considere-se a informação útil entre as *tags* a seguir "<td>Cliente: Luiz Otávio Fonseca</td>", a palavra-chave "Cliente:" precisa ser eliminada, tendo em vista que a representação dessa informação em XML é identificada pela *tag* <Cliente>. Com a opção "não" ativada, a palavra-chave permanece no texto da informação útil. Como exemplo, na informação útil entre as *tags* "<td>Novo cd do Caprichoso!!!</td>", pode-se ver que a palavra-chave cd não pode ser eliminada. Caso o fosse, a informação útil perderia o sentido.

Convém lembrar que a Figura 3.9 apresenta apenas parte das palavras-chave obtidas na análise.

O processo de busca de palavras-chave consiste na procura delas em informações úteis de páginas HTML. Em qual informação útil a busca será feita, descrever-se-á adiante, nas regras de extração. Localizada uma palavra-chave, um conceito é atribuído ao conteúdo, e ela é ou não excluída dele. As informações úteis resultantes do processo nas páginas Parintins e Ouro Negro ilustram a seguir o uso de palavras-chave.

Parintins

Informação útil . . : <td>Novo *cd* do Caprichoso!!!</td>

Conceito . . . . . : Disco

Excluir . . . . . : Não

Nova cadeia. . . . . : <Disco>Novo cd do Caprichoso!!!</Disco>

Ouro Negro

Informação útil . . : <td>Cliente: Luiz Otávio Fonseca</td>

Conceito . . . . . : Cliente

Excluir. . . . . : Sim

Nova cadeia . . . . . : <Cliente>Luiz Otávio Fonseca</Cliente>

A ferramenta procura no texto todas as palavras-chave contidas na tabela de palavras-chave. Caso haja mais de uma, o usuário deverá interagir e selecionar aquela que melhor expressa o conteúdo do texto. No momento que uma é selecionada, o conceito correspondente é associado ao texto e, se a propriedade "excluir" estiver ativada, a palavra-chave é retirada do texto.

Nos exemplos das páginas Parintins e Ouro Negro, os conteúdos “Novo ***cd*** do Caprichoso!!!” e ***“Cliente:*** Luiz Otávio Fonseca***”*** contêm as palavras-chave “cd” e “cliente”. Nesse caso, “cd” e “Cliente:” servem para identificar o fato de que elas representam instâncias dos conceitos ***“Disco”*** e ***“Cliente”*** .

No conteúdo **“Novo cd do Caprichoso!!!”,** se a palavra-chave “cd” for eliminada, o texto resultante será **“Novo do Caprichoso!!!”,** o qual não apresenta concordância gramatical. Portanto, a palavra-chave “cd”, nesse caso, não deve ser excluída, sendo que, para isso, a sua propriedade “excluir” precisa estar ativada em “Não”.

Já para o conteúdo ***“Cliente:*** **Luiz Otávio Fonseca*”,*** a eliminação da palavra-chave **“Cliente:”** resulta no texto **“Luiz Otávio Fonseca”,** o qual continua gramaticalmente correto. Para este caso, a propriedade “excluir” para a palavra- -chave **“Cliente:”** deve ser ativada em “Sim”, conseqüentemente eliminando-a. A razão da eliminação da palavra-chave “Cliente:”, no conteúdo, se dá pelo fato de evitar a redundância com a tag “Cliente” no arquivo de saída XML.

A definição dos conceitos referentes aos campos dos registros das tabelas encontradas, em primeira instância, é um processo automático; em um segundo momento, esses conceitos podem requerer alterações para se estabelecer concordância no domínio dos dados (como exemplo, pode-se observar o processo a seguir). Como definido anteriormente, o processo de definição desses conceitos consiste, primeiramente, na procura por formatos-padrão através de expressões regulares. Não os encontrando, a procura continua por palavras-chave ou seus sinônimos. A seguir, serão aplicados os formatos-padrão e as palavras-chave nas tabelas resultantes da página Parintins mostrada na figura 3.6. O resultado do processo pode ser visto na figura 3.10 a seguir e, mais especificamente, na coluna com fundo azul-claro.

Tabela	Registro	Campo	Informação útil	Conceito
7.2.1.1	1	1	Garantido 2001	Texto
		2	Toadas oficiais 2001 - cd duplo	CD
		3	R$ 23.95	Valor
7.2.2	1	1	Caprichoso: Amor e Paixão	Texto
		2	Novo cd do Caprichoso!!!	CD
		3	R$ 17.95	Valor
7.2.3	2	1	Novo CD, inclui o sucesso	CD
		2	Sem Juízo	Nome próprio
		3	R$ 16.95	Valor
7.3	1	1	Folclore Político do Amazonas	Nome próprio
		2	Episódios hilariantes da história política regional	Texto
		3	R$ 29.90	Valor
	2	1	Garantido Antológico	Nome próprio
		2	CD reúne toadas de todos os ...	CD
		3	R$ 18.95	Valor
	3	1	Os Intérpretes da Amazônia	Texto
		2	Análise de grandes obras que ...	Texto
		3	R$ 29.90	Valor
	4	1	Estatutos do Homem	Nome próprio
		2	Edição especial trilingüe do célebre ...	Texto
		3	R$ 35.00	Valor
	5	1	Bar do Boi	Nome próprio
		2	CD de aniversário de 10 anos do Bar ...	CD
		3	R$ 15.95	Valor
	6	1	Universo Mítico Ritual do Povo Tukano	Nome próprio
		2	Cultura e crenças das tribos Tikunas	Texto
		3	R$ 37.90	Valor

FIGURA 3.10 – Conceitos automáticos

A geração automática descrita acima gera os conceitos a partir dos padrões e das palavras-chave. Porém, esses conceitos podem ser modificados para refletirem melhor a identificação do modelo. Por exemplo: se o conceito gerado for "***Valor***", obtido a partir do texto "***2.820,56***", mas, se o contexto da página se refere a empregados, remunerações, adiantamentos, descontos de mercado ou farmácia, então esse conceito deve ser mudado de acordo com o seu uso, ou seja, o conceito dentro desse contexto poderia ser "***Salário***".

Analogamente, a geração automática poderá produzir conceitos diferentes para registros iguais em uma mesma tabela. Por exemplo: na tabela 7.3, os conceitos gerados para o primeiro registro foram: "***Nome Próprio***", "***Texto***" e "***Valor***"; para o segundo: "***Nome Próprio***", "***CD***" e "***Valor***"; terceiro: "***Texto***", "***Texto***" e "***Valor***", mas, como a tabela se refere a venda de livros e CD's, os conceitos devem ser adequados e os registros devem possuir os mesmos conceitos. Após as modificações dos conceitos, obtivemos a figura 3.11 abaixo.

Tabela	Registro	Campo	Informação útil	Conceito	Alteração
7.2.1.1	1	1	Garantido 2001	Texto	Título
		2	Toadas oficiais 2001 - cd duplo	CD	CD
		3	R$ 23.95	Valor	Preço
7.2.2	1	1	Caprichoso: Amor e Paixão	Texto	Título
		2	Novo cd do Caprichoso!!!	CD	CD
		3	R$ 17.95	Valor	Preço
7.2.3	2	1	Novo CD, inclui o sucesso	CD	Título
		2	Sem Juízo	Nome próprio	CD
		3	R$ 16.95	Valor	Preço
7.3	1	1	Folclore Político do Amazonas	Nome próprio	Título
		2	Episódios hilariantes da história ...	Texto	Livro
		3	R$ 29.90	Valor	Preço
	2	1	Garantido Antológico	Nome próprio	Título
		2	CD reune toadas de todos os ...	CD	CD
		3	R$ 18.95	Valor	Preço
	3	1	Os Intérpretes da Amazônia	Texto	Título
		2	Análise de grandes obras que ...	Texto	Livro
		3	R$ 29.90	Valor	Preço
	4	1	Estatutos do Homem	Nome próprio	Título
		2	Edição especial trilingüe do célebre ...	Texto	Livro
		3	R$ 35.00	Valor	Preço
	5	1	Bar do Boi	Nome próprio	Título
		2	CD de aniversário de 10 anos ...	CD	CD
		3	R$ 15.95	Valor	Preço
	6	1	Universo Mítico Ritual do Povo ...	Nome próprio	Título
		2	Cultura e crenças das tribos Tikunas	Texto	Livro
		3	R$ 37.90	Valor	Preço

FIGURA 3.11 – Conceitos adequados à página

3.1.5 Geração de Modelo Conceitual e Regras de Extração

Modelo Conceitual

O modelo conceitual, neste trabalho, é a forma de se representar o conhecimento sobre um domínio de dados, ou seja, representam-se os objetos, os relacionamentos entre eles, as restrições de integridade, os aninhamentos, os tipos de dados e a cardinalidade.

A representação gráfica do nosso modelo conceitual é construída, utilizando-se quatro tipos de elementos: os retângulos menores representam os conceitos referentes aos objetos léxicos; os retângulos maiores, os conceitos referentes aos objetos não-léxicos; as linhas estabelecem os relacionamentos e a cardinalidade escrita nos extremos dos relacionamentos. A Figura 3.12 abaixo apresenta os elementos gráficos com os devidos significados.

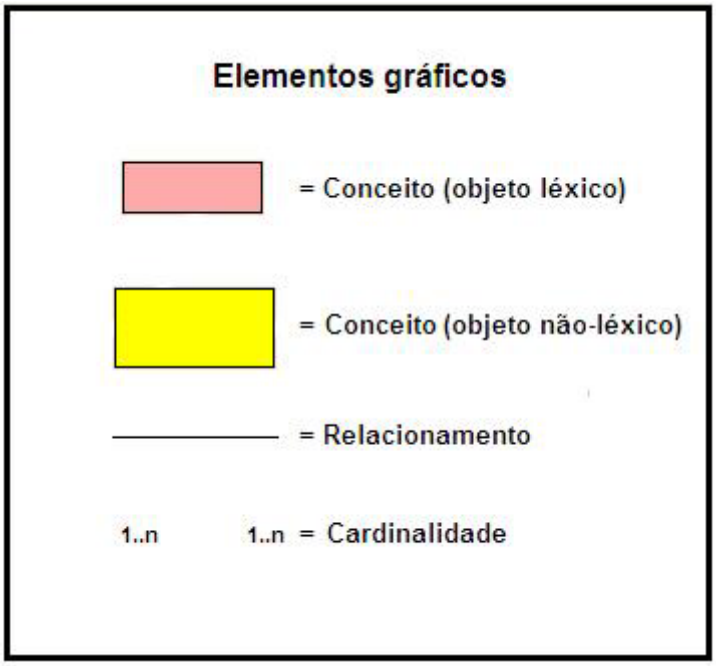

FIGURA 3.12 – Elementos gráficos

O Modelo Conceitual é gerado automaticamente a partir das definições anteriores, mas os nomes dos objetos não-léxicos do modelo o são a partir de uma tabela de nomes de objetos não-léxicos associados a conceitos de campos, tabela residente no repositório. Esse processo consiste na busca de um nome de objeto não-léxico associado aos conceitos da tabela, procedimento que se repete para tantas tabelas quantas necessárias. Novamente nos deparamos com a possibilidade da geração automática de nomes de objetos não-léxicos incompatíveis com os conceitos do modelo. Por exemplo: considerando-se um modelo com os conceitos "***Nome Próprio***", "***Texto***" e "***Preço***", e supondo-se que tenha sido gerado o nome do objeto não-léxico "***Loja***", e considerando-se ainda que, sendo o Nome Próprio "***Remédio***" e o texto "***Modo de usar***", muito provavelmente o nome do objeto não-léxico "Loja" poderia ser substituído por "***Farmácia***".

Da mesma forma que os nomes de objetos não-léxicos, a cardinalidade gerada pode não refletir a verdade, necessitando-se de alteração.

O modelo conceitual gerado originalmente e alterado conforme o contexto da página resultou nos modelos finais, apresentados nas figuras 3.13, para a página Parintins, e 3.14 para a página Ouro Negro abaixo.

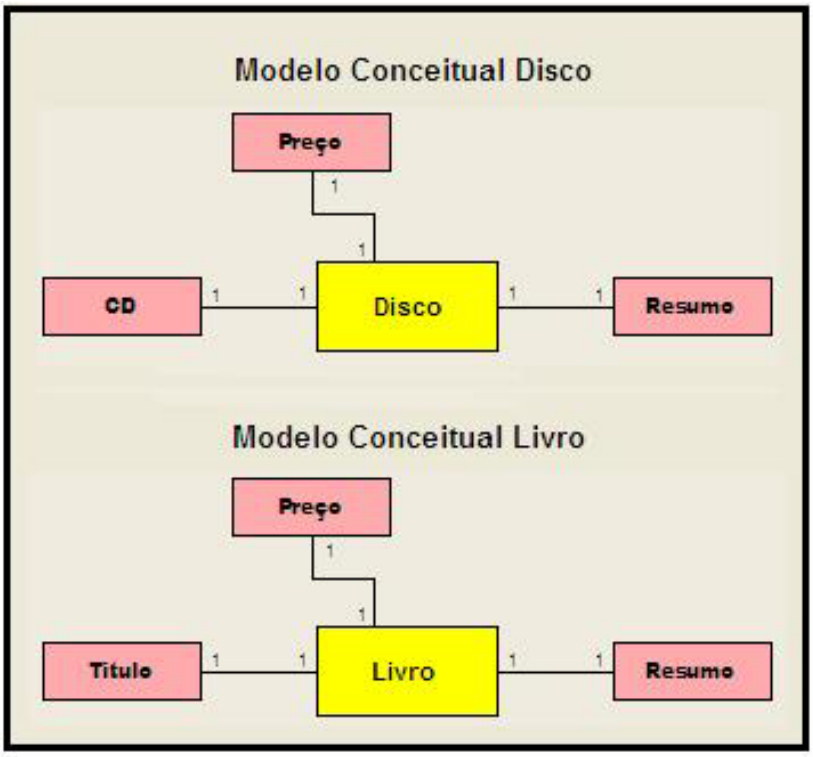

FIGURA 3.13 – Modelo Conceitual Parintins

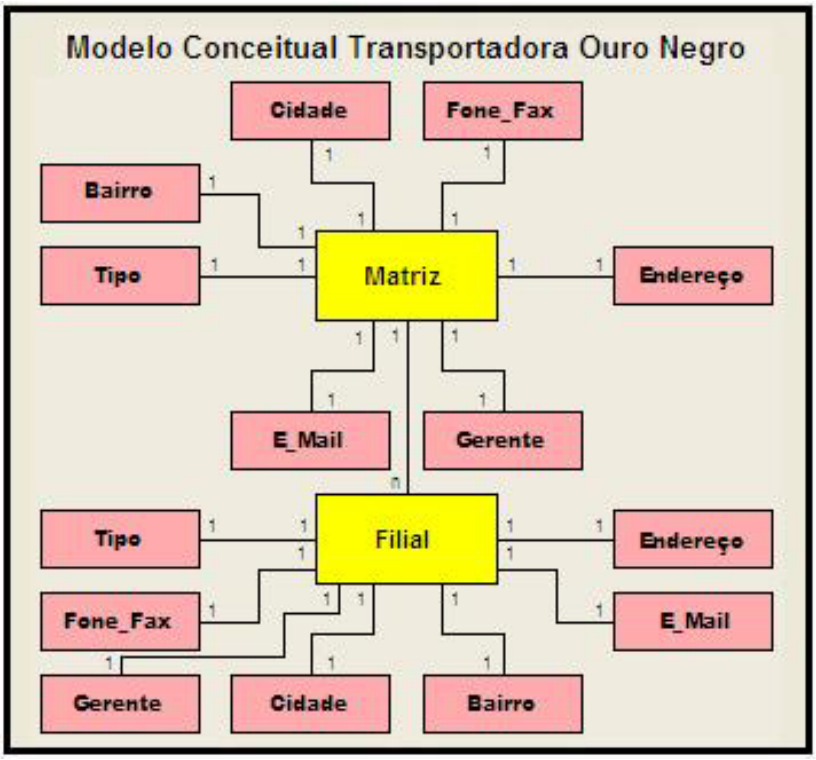

FIGURA 3.14 – Modelo Conceitual Ouro Negro

Regras de Extração

As regras de extração e a representação dos dados em um modelo conceitual, armazenadas no repositório, devem ser suficientes para realizar o mesmo processo de extração de dados que fizeram no momento de sua criação, ou seja, aqueles extraídos da página Parintins, no momento da criação das regras de extração e do modelo conceitual, devem ser os mesmos extraídos em um momento posterior, utilizando-se, sem nenhuma alteração, somente das regras e do modelo conceitual do repositório, desde que a página não tenha sido modificada..

A ferramenta baseada em tabelas, desenvolvida neste trabalho, conta com as seguintes regras de extração para reutilização na detecção de páginas semelhantes:

a) Satisfazer as condições estabelecidas nas seções 3.1.1 até 3.1.5 (geração de modelo conceitual);

   As condições descritas nas seções 3.1.1 até 3.1.5 (geração de modelo conceitual), não estão catalogadas no repositório, elas são algoritmos que são aplicados nas páginas por meio da ferramenta;

b) Os modelos conceituais que representam os objetos da página em análise devem satisfazer a modelos conceituais iguais, armazenados no repositório;

   A criação de modelo conceitual para uma página HTML considera todas as ações automáticas e manuais executadas durante o processo, armazenando-as no repositório. Por exemplo, o algoritmo de busca por palavras-chave identifica como objeto não-léxico o conceito “loja”, e, de acordo com o contexto, ele é modificado para “disco”. Em outra situação, uma expressão regular identifica uma cadeia como nome próprio e este é modificado para “Resumo”. Todas essas ações são regras que são inferidas na identificação de páginas semelhantes. A tabela 3.3 mostra as alterações de objetos durante o processo;

TABELA 3.3 – Alteração de objetos

Objetos não léxicos		**Objetos léxicos**	
Final	**Original**	**Final**	**Originais**
Disco	Loja	CD	Texto, CD, Nome próprio
		Resumo	Nome próprio
		Preço	Valor
Livro	Sem opção	Título	Nome próprio, Texto
		Resumo	Texto
		Preço	Valor

c) Eliminar todas as tabelas que não pertençam ao domínio representado no modelo conceitual aplicado à página HTML;

d) Verificar se a página em questão possui todos os objetos léxicos correspondentes na tabela representada pelo modelo conceitual;

e) Armazenar as características de todos os registros no repositório.

3.1.6 Extração de dados de Página HTML e Geração do arquivo XML

Ao final de todo o processo executado sobre a página HTML, resta ainda a extração dos dados e geração de um arquivo fonte XML.

Extração dos dados da Página HTML

Um arquivo de texto é gerado com os dados extraídos da página HTML.

Geração do arquivo XML

Os dados resultantes do processo, gerados em XML, apresentam o seguinte formato:

```
<?xml version=”1.0”?
<!doctype advent system esquema>
<disco>
        <título>Garantido 2001</título>
        <cd>Toadas oficiais 2001 - cd duplo</cd>
        <preço>23,95</preço>
</disco>
<disco>
        <título>Caprichoso: Amor e Paixão</título>
        <cd>Novo cd do Caprichoso!!!</cd>
        <preço>17,95</preço>
</disco>
<disco>
        <título>Novo CD, inclui o sucesso</título>
        <cd>Sem Juízo </cd>
        <preço>16,95</preço>
</disco>
```

```
<disco>
        <título>Garantindo Antológico</título>
        <cd>CD reúne toadas de todos os </cd>
        <preço>18,95</preço>
</disco>
<disco>
        <título>Bar do Boi</título>
        <cd>CD de aniversário de 10 anos do Bar do Boi Caprichoso,
Manaus</cd>
        <preço>15,95</preço>
</disco>
<livro>
        <título>Folclore Político do Amazonas</título>
        <resumo>Episódios hilariantes da história política regional </resumo>
        <preço>29,90</ preço>
</livro>
<livro>
        <título>Os intérpretes da Amazônia </título>
        <resumo> Análise de grandes obras que descreveram a
                Amazônia</resumo>
        <preço> 29,90</preço>
</livro>
<livro>
        <título>Estatutos do Homem</título>
        <resumo>Edição especial trilingüe do célebre poema de Thiago de Mello
                </resumo>
        <preço>29,90</ preço>
</livro>
<livro>
        <título>Universo Místico Ritual do Povo Tukano</título>
        <resumo>Cultura e crenças das Tribos Tikunas</resumo>
        <preço> 29,90</preço>
</livro>
```

3.1.7 Realimentação do Repositório

A realimentação do repositório, como foi citado no início dessa seção, é um processo automático. Para qualquer extração, todos os objetos alterados ou criados são alimentados no repositório para futuras utilizações. Em resumo, ao final da parametrização realizada ao longo dos itens 3.1.1 a 3.1.6, novas palavras-chave, conceitos, regras de extração e os modelos conceituais são criados e, obrigatoriamente, devem-se agregar ao repositório.

3.2 Utilização de Modelo Conceitual e Regras de Extração

O objetivo da utilização de modelo conceitual e regras de extração é proporcionar um alto grau de automaticidade na extração de dados semi-estruturados em páginas HTML. Na medida em que vai sendo usada, a ferramenta cria, no repositório, novas regras de extração e novos modelos conceituais. Esse acervo possibilita a reutilização de modelos conceituais e regras para páginas semelhantes, eliminando-se procedimentos manuais para geração de novas regras e novos modelos conceituais em páginas HTML. Para tanto, estão listados abaixo os passos para a utilização de modelo conceitual e regras de extração.

a) Carregar na ferramenta a página da qual se deseja realizar a extração;

b) Disparar um processo automático que seleciona, na página HTML, somente as tabelas com informações úteis, utilizando-se dos procedimentos listados na seção 3.1.1;

c) Procurar no repositório um modelo conceitual; pages

d) Verificar se o modelo conceitual se ajusta a todas ou a parte das tabelas do item b;

e) Voltar para o item c, caso o modelo não se ajuste às tabelas selecionadas no item b;

f) Eliminar as tabelas que não pertencem ao modelo conceitual encontrado;

g) Finalizar a extração.

Para exemplificar, suponha-se que tenham sido gerados, anteriormente, as regras de extração e o modelo conceitual para a página Parintins. Isso significa que existem, no repositório, as regras de extração e o modelo conceitual dessa página. Conseqüentemente, a ferramenta deve encontrá-las. Portanto, carregando-se novamente a página Parintins, a ferramenta deve encontrar as regras e o modelo conceitual existente no repositório que,

necessariamente, se enquadra a essa página. A figura 3.15 ilustra o processo, destacando, com fundo amarelo, as tabelas que devem satisfazer o primeiro modelo conceitual com suas regras e, com fundo azul, as que devem satisfazer o segundo modelo conceitual com as suas.

Tabela	Registro	Campo	Informação útil
1	1	1	Como comprar
		2	Sobre segurança
	2	1	Minha compra
3	1	1	shopping@parintins.com
	2	1	Desde 1998 distribuindo produtos da Amazônia para o Mundo
		2	Início
		3	ENGLISH
		4	ESPAÑOL
3.1	1	1	Amazônia é
	2	1	Brasil
5	1	1	-->
		2	Use seu cartão VISA em nosso servidor seguro.
		3	Você também pode pagar com depósito bancário
		4	Fone / Fax
		5	(0xx92) 648-1800
6	1	1	Se você não encontrou um produto, faça seu pedido:
	2	1	E-mail:
7.1	1	1	Arlindo Jr
	2	1	Raízes Caboclas
	3	1	Novo CD, inclui o sucesso
		2	Só pode ser Você
		3	...
		4	R$ 20.90
	4	1	Novo CD da Banda...
		2	R$ 16.95
7.2.1.1	1	1	Garantido 2001
		2	Toadas oficiais 2001 - cd duplo
		3	R$ 23.95
7.2.2	1	1	Caprichoso: Amor e Paixão
		2	Novo cd do Caprichoso!!!
		3	R$ 17.95
7.2.3	1	1	Carlos Batata
	2	1	Novo CD, inclui o sucesso
		2	Sem Juízo
		3	R$ 16.95
7.3	1	1	Folclore Político do Amazonas
		2	Episódios hilariantes da história política regional
		3	R$ 29.90
	2	1	Garantido Antológico
		2	CD reúne toadas de todos os tempos do Bumbá Garantido, na voz de David Assayag
		3	R$ 18.95
	3	1	Os Intérpretes da Amazônia
		2	Análise de grandes obras que descreveram a Amazônia
		3	R$ 29.90
	4	1	Estatutos do Homem
		2	Edição especial trilingüe do célebre poema de Thiago de Mello.
		3	R$ 35.00
	5	1	Bar do Boi
		2	CD de aniversário de 10 anos do Bar do Boi Caprichoso,Manaus
		3	R$ 15.95
	6	1	Universo Mítico Ritual do Povo Tukano
		2	Cultura e crenças das tribos Tikunas
		3	R$ 37.90
8	1	1	Dúvidas? Sugestões?
		2	shopping@parintins.com
9	1	1	© Opera House
		2	(1998-2001)

FIGURA 3.15 – Regras de Extração e Modelos Conceituais associados à Página

3.3 O Repositório

O suporte ao processamento executado por toda a ferramenta é fornecido pelo repositório por meio da manipulação de suas informações. Ele é formado por várias tabelas que pertencem a três categorias distintas: a primeira dá suporte ao processamento automático da ferramenta; a segunda contém informações do modelo conceitual; a terceira, as regras de extração.

a) Para a realização dos processos automáticos, a ferramenta conta com as tabelas descritas a seguir:

- Tabela de palavras-chave;
- Tabela de nodos;
- Tabela de conceitos.

b) Os modelos conceituais de cada página HTML são estruturados no repositório para a reutilização em páginas semelhantes:

- Tabela de modelos conceituais.

c) As regras de extração são armazenadas na seguinte tabela:

- Tabela de regras de extração.

Nessa seção, foram descritas e exemplificadas todas as etapas do processo realizado pelo GREMO, objetivando gerar regras de extração e criação de modelo conceitual para aplicação em páginas HTML. Na seção 4, será mostrado um exemplo passo a passo, utilizando-se a página Parintins para experimentação no GREMO.

4 O Protótipo da Ferramenta

A página Parintins foi escolhida para exemplificar este trabalho por se tratar de uma página HTML complexa. A execução do protótipo para essa página mostrará a interação passo a passo para a geração de regras de extração e criação do modelo conceitual.

4.1 Seleção e carga de páginas HTML

O processo inicia-se pela seleção e carga da página HTML da qual se deseja executar a extração de dados. A Figura 4.1 representa a interface e, para se executarem os procedimentos dessa etapa, são necessárias as seguintes ações: um clique no botão “seleção página HTML”, indicado por □, causando a abertura da janela “seleção página HTML □”, que permite a navegação para localização e carga da página.

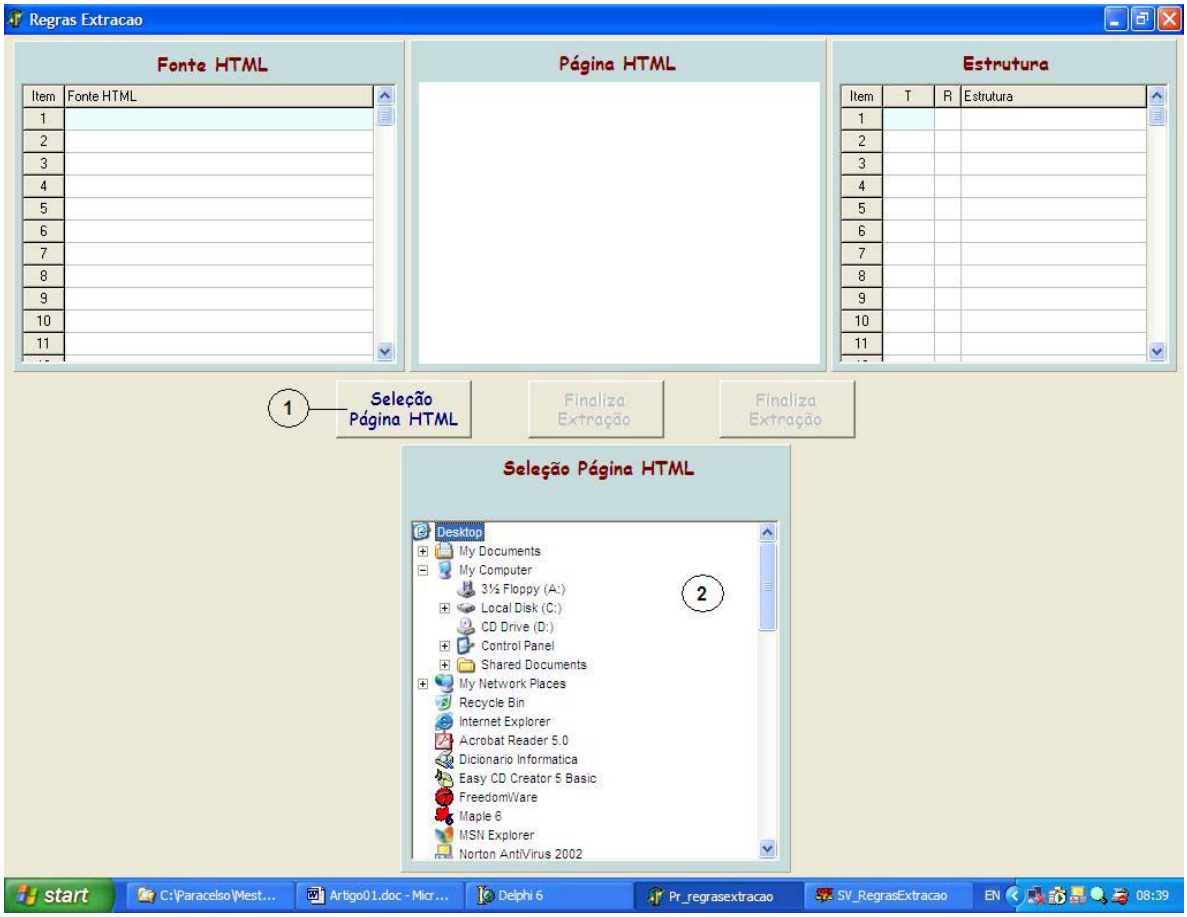

FIGURA 4.1 – Seleção e carga de página

4.2 Tabelas (aninhadas ou não) em páginas HTML

O processo continua, utilizando a Figura 4.2 para se representar a interface nessa etapa, realizando o seguinte:

- Possibilidade de uma nova seleção de página, podendo ser obtida com um clique no botão Seleção Página HTML □;
- Apresentação, na grade da janela Fonte HTML □, do arquivo-fonte da página HTML selecionada, com recursos de navegação para avaliação e identificação de tabelas;
- Apresentação da página selecionada propriamente dita, na janela página HTML □, também com recursos de navegação;
- Apresentação, na grade da janela Estrutura □, de todas as tabelas encontradas na página selecionada, sejam elas de controle ou tabelas de informações úteis (aninhadas ou não);
- Apresentação, na janela "desprezar tabelas □", dos botões que servem para esse efeito, contendo eles as mesmas cores e números das tabelas indicadas na janela Estrutura □, podendo o pressionamento de um deles desprezar a tabela correspondente;
- Apresentação do botão Avançar □, que serve para uma pré-ratificação das tabelas desprezadas pelos botões □;
- Apresentação do botão Voltar □, que serve para desfazer a pré-ratificação do botão □;
- Apresentação do botão Confirmar □, que serve para ratificar a ação do botão □, encerrando essa etapa com o fechamento da janela "desconsiderar tabelas".

FIGURA 4.2 – Seleção de tabelas

4.3 Identificação de Conceitos

A representação da interface é apresentada na Figura 4.3, e, nessa etapa, a identificação de conceitos é realizada em duas fases distintas: a) Geração automática de conceitos; b) Interação do usuário para adequação dos conceitos às tabelas.

Geração Automática de Conceitos

A geração automática de conceitos inicia com o pressionamento do botão “cria conceitos □”, na interface representada pela Figura 4.3, que dispara um processo de aplicação de regras nos registros das tabelas selecionadas, mostradas na grade da janela Estrutura □. Essas regras foram descritas no item 3.1.1.

Interação do usuário para adequação dos conceitos às tabelas

Alguns dos conceitos gerados automaticamente precisam ser modificados para expressarem melhores os objetos dentro do contexto. Para essa situação, a interface está representada pela Figura 4.3. Tais modificações podem ser realizadas como se segue:

- Para modificar um conceito, informa-se o novo nome no campo – Novo Conceito □ – e clica-se naquele que se quer modificar □. Repete-se esse procedimento tantas vezes quantas necessárias.

- Para igualar os registros das tabelas, seleciona-se um botão de rádio em "Conceitos de Tabelas Iguais □", a seguir clica-se em todos os registros □ com os mesmos conceitos daquela tabela. Repete-se esse procedimento para todos os registros que requeiram tal modificação nas tabelas selecionadas da página em questão.

- Para encerrar a etapa, confirmando todas as ações realizadas, apertar o botão "Avançar □".

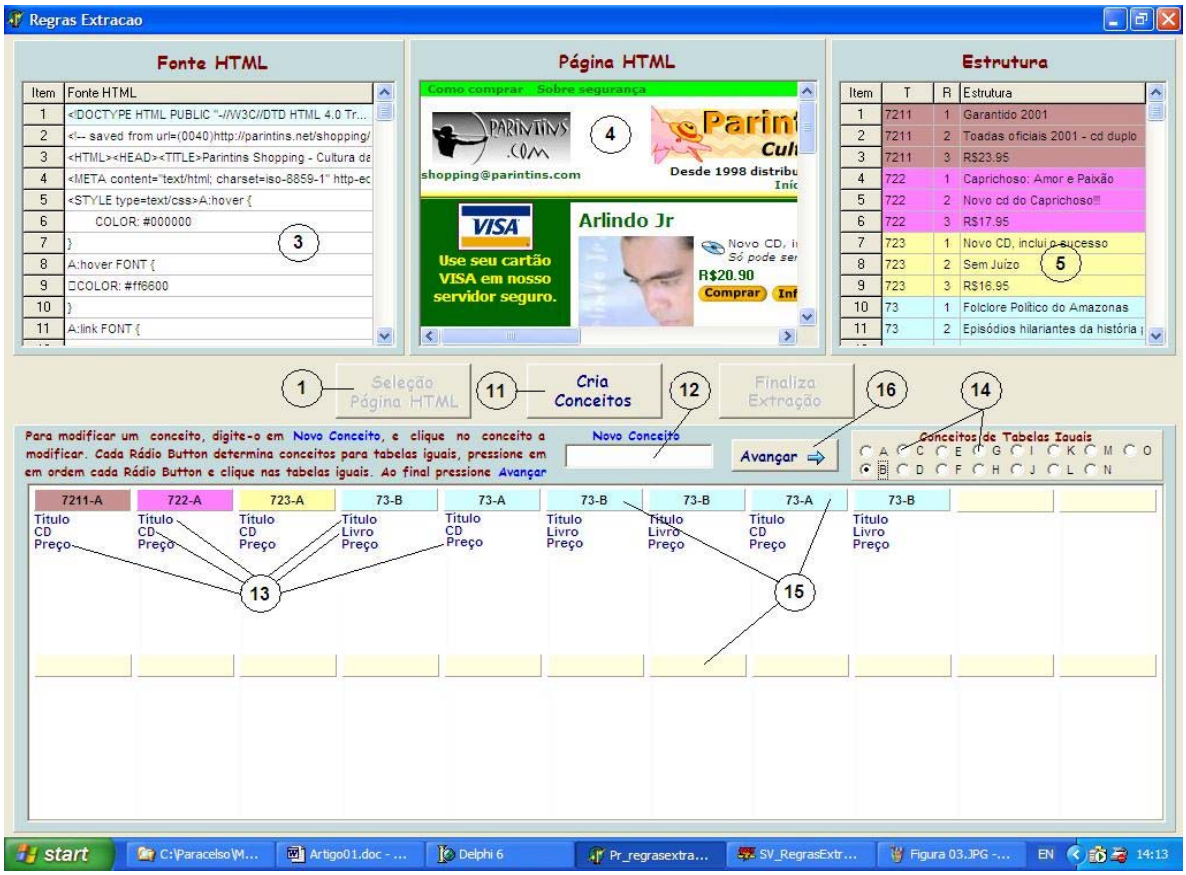

FIGURA 4.3 – Igualando Conceitos e Registros

4.4 Alteração de Modelo Conceitual

A criação de Modelo conceitual é feita automaticamente pela ferramenta e apresentada na interface representada pela Figura 4.4, na janela “Modelo Conceitual ”. Como a geração dos nomes é automática, alguns deles podem não espelhar a verdade do modelo; portanto, precisamos ter a possibilidade de alterá-los. Diante disso, para se realizar a substituição de conceitos léxicos, não-léxicos ou da cardinalidade, informa-se, na janela Modelo Conceitual , o novo conceito ou a cardinalidade na caixa de edição , e a seguir clica-se no conceito ou cardinalidade a se modificar; repete-se o procedimento tantas vezes quantas necessárias.

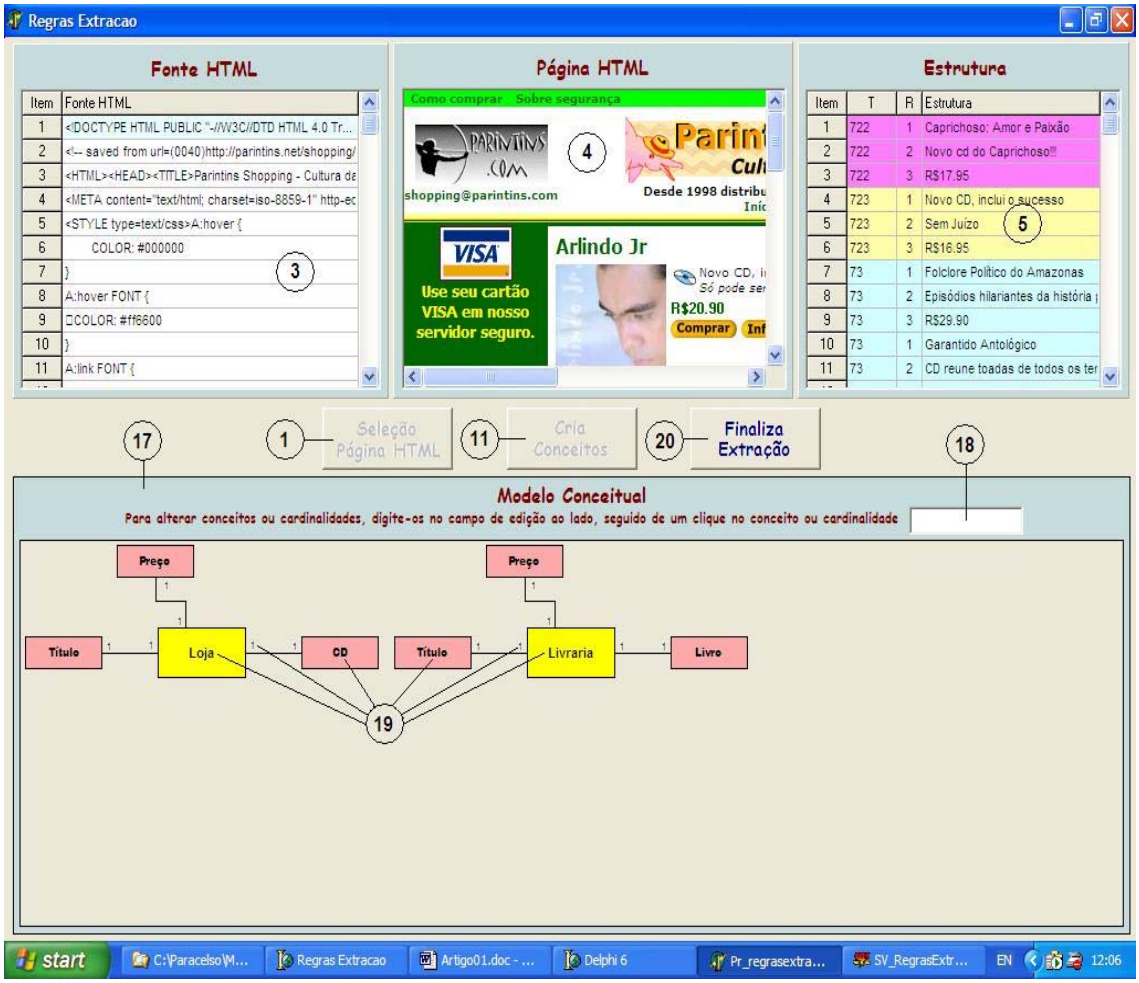

FIGURA 4.4 – Modelo Conceitual

Para encerrar essa etapa confirmando todas as ações realizadas, apertar o botão Finaliza Extração .

4.5 Finalizar a Extração

A finalização da extração compreende processos automáticos para a construção das regras de extração, a extração dos dados da página, a geração do arquivo XML e, finalmente, a realimentação do repositório, tendo-se em vista todas as ações necessárias para isso já terem sido realizadas, nas etapas anteriores, por processos automáticos ou por intervenção do usuário.

A experimentação de todo o processo de GREMO foi realizada sobre a página Parintins, sendo que os resultados obtidos atingiram totalmente as expectativas da proposta. O processo resultou em dois modelos conceituais correspondentes a duas tabelas, sendo o primeiro um modelo conceitual correspondente a discos e o segundo, a livros. A tabela de discos com 5 registros, e a de livros com 4. Foi gerado um arquivo fonte XML com o resultado da extração.

A figura 3.14 página 43 é o resultado também do experimento do processo de GREMO sobre a página Ouro Negro, para a ilustração de uma tabela aninhada.

5 Conclusões e trabalhos futuros

O desenvolvimento deste trabalho tem como principal contribuição a geração de regras de extração de dados em páginas HTML e a reutilização dessas regras para reconhecimento de páginas semelhantes. O processo de geração de regras está baseado em modelos conceituais utilizando exclusivamente tabelas (aninhadas ou não). Nessa abordagem, o foco é a busca somente por tabelas com informações úteis, sendo essa a grande diferença entre os demais métodos. O nosso trabalho utiliza alguns recursos também manipulados por outros métodos, como busca por palavras-chave, a utilização de expressões regulares para a localização de formatos-padrão, além da representação dos dados em um modelo conceitual, método também utilizado por Embley [EMB 99] e [SIL 01].

A grande vantagem de se trabalhar somente com tabelas de informações úteis prende-se ao fato de que é eliminada do método a maior parte dos dados das páginas. Portanto, causa significativa redução nas buscas, além de proporcionar uma melhor performance nos processos.

Outra facilidade de nosso método é a utilização, com ou sem adaptação, das regras de extração e de modelos conceituais existentes no repositório; ou seja, páginas HTML semelhantes podem ser reconhecidas por regras e modelos conceituais criados anteriormente, alguns sem alterações e outros com pequenas alterações. Páginas HTML semelhantes são as de qualquer organização que possuam tabelas que se ajustem a algum modelo conceitual e a regras do repositório. Isso quer dizer que as regras e o modelo conceitual não precisam se ajustar somente à página que lhes deram origem ou à evolução delas, mas sim a qualquer página HTML da rede.

O crescimento do acervo no repositório, a partir da criação de regras, de novos modelos e das adaptações daqueles já existentes, aumenta na razão direta o grau de automaticidade do método, pois maior será o número de páginas semelhantes reconhecidas.

Para a técnica proposta neste trabalho, foi desenvolvido um protótipo. Com esse protótipo foram realizados diversos experimentos, dos quais, neste trabalho, foram apresentados apenas dois, conforme demonstrado na seção 3, sendo que, em todos os casos, os resultados obtidos com a geração de regras de extração e de modelos conceituais atingiram, com sucesso, os resultados esperados.

Nosso método, por outro lado, não trata páginas HTML como se segue:

- que não possuam tabelas;
- que tenham somente tabelas de formatação;

- que possuam tabelas com registros de apenas um campo;
- que possuam tabelas com apenas um registro;
- que possuam tabelas para as quais a representação dos dados do modelo conceitual não sejam iguais; por exemplo, o método não trata uma página que tenha uma tabela com os campos *nome*, *data de fabricação* e *data de validade,* avaliada por um modelo conceitual que possua os conceitos *nome*, *data de nascimento* e *data de casamento*.

A continuidade deste trabalho pode agregar, em trabalhos futuros, ainda mais benefícios na extração de dados. Para tanto, podemos relacionar o seguinte:

- Desenvolver um estudo em tabelas com registros de apenas um campo e em tabelas com somente um registro, objetivando a possibilidade de agregação ao método;
- Desenvolver um algoritmo que reconheça a regularidade de uma tabela, retornando a igualdade em todos os seus registros. Por exemplo, em uma tabela, os campos de todas as linhas são identificados pelo cabeçalho das colunas, e a maioria das tabelas, em páginas HTML, não possui cabeçalhos de colunas. Portanto, numa tabela que tenha 10 linhas e esteja referenciando remédios, cujos campos sejam código, remédio e preço, em princípio todos os campos de todas as linhas devem ser reconhecidos como código, remédio e preço;
- Desenvolver um estudo em páginas HTML, para aumentar o número de palavras-chave, no sentido de reduzir a interação do usuário;
- Desenvolver um algoritmo para propor alternativa a conceitos similares. Por exemplo, seja, em uma tabela, um campo identificado como data de fabricação, mas também podendo ter outras identificações.

A extração de dados semi-estruturados abrange uma gama muito grande de fontes de dados, causando significativa dificuldade de uma única técnica resolver todo o problema. Com base na importância deste assunto e no interesse por ele, que muitas pesquisas se desenvolvam, produzindo técnicas que, junto com a nossa, venham somente a somar.

Referências

[ABA 99] ABASCAL, R.; SÁNCHEZ, J. A. X-tract: Structure extraction from botanical textual descriptions. In: INTERNATIONAL SYMPOSIUM ON STRING PROCESSING AND INFORMATION RETRIEVAL, 6., 1999. **Proceedings …** [S.l.: s.n.], 1999.

[ABI 2000] ABITEOUL, S.; BUNEMAN, P.; SUCIU, D. **Data on the Web**: from Relations to Semistructured Data and XML. [S.l.]: Morgan Kaufmann, 2000.

[ADE 98] ADELBERG, B. NoDoSE – A Tool for Semi-Automatically Extracting Structured and semistructured Data from text Documents. **SIGMOD Record**, New York, v.27, n.2, p. 283-294, June 1998.

[ADE 99] ADELBERG, B. **Building Robust Wrappers for Text Sources**. New York, 1999. Technical Report. Disponível em: <http://www.cs.nwu.edu/~adelberg>.Acesso em: mar.2001.

[ARO 98] AROCENA, G. O.; MENDELZON, A. O. WebOQL: restructuring documents, databases, and webs. In: INTERNATIONAL CONFERENCE ON DATA ENGINEERING, 14., 1998, Orlando. **Selected Papers**: object-oriented technology in advanced applications. New York: [s.n.], 1998. p. 24-33.

[ASH 97] ASHISH, N.; KNOBLOCK, C. Wrapper Generation for Semi-Structured Internet Sources. **SIGMOD Record**, New York, v.26, n.4, p.8-15, 1997.

[ATZ 97] ATZENI, P.; MECCA, G. Cut and Paste. In: ACM SIGART – SIGMOD – SIGART SYMPOSIUM ON PRINCIPLES OF DATABASE SYSTEMS, PODS, 16., 1997, Tucson. **Proceedings…** New York: ACM, 1997

[ATZ 98] ATZENI, P.; MECCA, G. et al. **From Database to Web-Bases**: The Araneus Experience. [S.l.: s.n.], 1998. Technical Report. Disponível em: <http://www.dia.uniroma3.it/Araneus>. Acesso em: mar.2001.

[ATZ 98a] ATZENI, P.; MECCA, G. et al. **The Araneus Web-Base Management System**. [S.l.: s.n.], 1998. Technical Report. Disponível em: <http://www.dia.uniroma3.it/Araneus>. Acesso em: mar.2001.

[BAE 99] BAEZA, R.; RIBEIRO, B. **Modern Information Retrieval**. [S.l.]: Addison Wesley, 1999. p. 257-323.

[BUN 97] BUNEMAN, P. et al. Adding Structure to Unstructured Data. In: INTERNATIONAL CONFERENCE ON DATABASE THEORY, ICDT, 1997. **Database Theory**: proceedings. Berlin: Springer-Verlag, 1997.

[BUN 97a] BUNEMAN, P. **Semistructured Data**. [S.l.: s.n.]: 1997. Technical Report.

[CAL 99] CALIFF, M. E.; MOONEY, R. J. Relational Learning of Pattern-Match Rules for Information Extraction. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, AAAI, 16., 1999. **Proceedings…** Menlo Park, CA: AAAI Press, 1999. p. 328 – 334.

[COH 99] COHEN, W. W.; SINGER, Y. A Simple, Fast, and Effective Rule Learner. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, AAAI, 16., 1999. **Proceedings…** Menlo Park, CA: AAAI Press, 1999. p. 335-342.

[CRE 98] CRESCENZI, V.; MECCA, G. Grammars have exceptions. **SIGMOD Record**, New York, v.23, n.8, p. 539-565, June 1998.

[CRE 2001] CRESCENZI, V.; MECCA, G.; MERIALDO, P. RoadRunner: Towards automatic data extraction from large Web sites. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 27., 2001. **Proceedings…** [S.l.: s.n.], 2001. p. 109-118.

[EMB 98] EMBLEY, D. W. et al. Ontology-based extraction and structuring of information from data-rich unstructured documents. In: CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, CIKM, 1998. **Proceedings…** Bethesda, Maryland: CIKM Press, 1998. p.52-59.

[EMB 98a] EMBLEY, D. W. et al. A Conceptual-modeling approach to extracting data from the WEB. In: INTERNATIONAL CONFERENCE ON CONCEPTUAL MODELING, ER, 17., 1998. **Proceedings…** [S.l.: s.n.], 1998.

[EMB 99] EMBLEY, D. W. et al. Record Boundary Discovery in Web Documents. **SIGMOD Record**, New York, v. 28, n. 2, June 1999. Trabalho apresentado na ACM SIGMOD International Conference on Management of Data, SIGMOD, 1999, Philadelphia.

[EMB 99a] EMBLEY, D. W. et al. Conceptual-model-based data extraction from multiple-record Web pages. **SIGMOD Record**, New York, v.31, n.3, p. 227-251, June 1999.

[EMB 99b] EMBLEY, D. W. et al. Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages. **SIGMOD Record**, New York, Nov. 1999.

[FRE 98] FREITAG, D. Information Extraction from HTML: Application of a General Learning Approach. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, AAAI, 15., 1998. **Proceedings…** Madison, WI: AAAI Press, 1998. p. 517-523.

[FRE 2000] FREITAG, D. Machine Learning for Information Extraction in Informal Domains. **SIGMOD Record**, New York, v.39, n.2/3, p. 169-202, July 2000.

[GRU 98] GRUSER, J. R. et al. Wrapper Generation for Web Accesible Data Sources. In: INTERNATIONAL CONFERENCE ON COOPERATIVE INFORMATION SYSTEMS, CoopIS, 13., 1998. **Proceedings…** New York, NY: 1998. p. 14-23.

[HAM 97] HAMMER, J. et al. Template-Based Wrappers in the TSIMMIS System. In: INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 1997. **Proceedings…** [S.l.: s.n.], 1997.

[HAM 97a] HAMMER, J.; MCHUGH, J.; GARCIA MOLINA, H.. semistructured data: The TSIMMIS experience. In: EAST-EUROPEAN SYMPOSIUM ON ADVANCES IN DATABASES AND INFORMATION SYSTEMS, ADBIS, 1., 1997. **Proceedings…** [S.l.: s.n.], 1997. p. 1-8.

[HAM 98] HAMMER, J.; GARCIA-MOLINA, J; ARANHA, R. et al. **Extracting Semistructured Information from the Web**. [S.l.: s.n.], 1998. Technical Report.

[HSU 98] HSU, C.-N.; DUNG, M.-T. Generating finite-state transducers for semi-structured data extraction from the Web. **Information Systems,** [S.l.], v.23, n.8, p. 521-538, 1998.

[KOR 98] KORNFELD, W.; WATTECAMPSS, J. Automatically Locating, Extracting and Analyzing Tabular Data. In: ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT ON INFORMATION RETRIEVAL, 21., 1998. **Proceedings…** [S.l.: s.n.], 1998. p. 347-348.

[KUS 97] KUSHMERICK, N.; WELD, D.; DOORENBOS, R. Wrapper Induction for Information Extraction. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, IJCAI, 15., 1997. **Proceedings…** [S.l.: s.n.], 1997. p. 729-735.

[KUS 2000] KUSHMERICK, N. Wrapper induction: Efficiency and expressiveness. **Artificial Intelligence,** Amsterdam, v.118, n. 1-2, p.15-68, 2000.

[LAE 99] LAENDER, A.; SILVA, E.; SILVA, A. DEByE – Uma Ferramenta para Extração de Dados Semi-Estruturados. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, SBBD, 1999, Florianópolis. **Anais...** Florianópolis: UFSC, 1999.

[LAE 99a] LAENDER, A.; SILVA, E.; SILVA, A. **Top-Down Extraction of Semistructured Data**. Belo Horizonte: [s.n.], 1999. Technical Report.

[LAE 2000] LAENDER, A.; SILVA, E.; SILVA, A. Extracting Semi-strucutred Data Through Examples. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, SBBD, 2000, João Pessoa. **Anais...** João Pessoa: PUCRS, 2000.

[LAE 02] LAENDER, A. H. F.; RIBEIRO-NETO, B.; DA SILVA, A. S. DEByE Data Extraction By Example. **Data and Knowledge Engineering,** [S.l.], v.40, n.2, p.121-154, 2002.

[LAE 2002a] LAENDER, A. et. Al. A Brief Survey of Web Data Extraction Toolst. **SIGMOD Record**, New York, v. 31, n. 2, June 2002.

[LID 99] LIDDLE, S. W.; CAMPBELLI, D. M.; CRAKFORD, C. Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages. In: INTERNATIONAL CONFERENCE ON INFORMATION KNOWLEDGE MANAGEMENT, CIKPM, 8., 1999. **Proceedings…** [S.l.: s.n.], 1999. p. 86- 93.

[LIM 99] LIM, S. J.; Ng, Y-K. An Automated Approach for Retrieving Hierarchical Data from HTML Tables. In: INTERNATIONAL CONFERENCE ON INFORMATION KNOWLEDGE MANAGEMENT, CIKPM, 8., 1999. **Proceedings…** [S.l.: s.n.], 1999. p. 466- 474.

[LIU 2000] LIU, L.; PU, C.; HAN, W. XWBAP: Na XML-enabled wrapper construction system for Web information sources. In: INTERNATIONAL CONFERENCE ON DATA ENGINEERING, 16., 2000, San Diego. **Selected Papers:** object-oriented technology in advanced applications. San Diego: [s.n.], 2000. p. 611-621.

[MUS 98] MUSLEA, I.; MINTON, S.; KNOBLOCK, C. Wrapper Induction for Semistructured, Web-based Information Sources. In: CONFERENCE ON AUTOMATED LEARNING AND DISCOVERY, CONALD, 1998. **Proceedings…** [S.l.: s.n.], 1998.

[MUS 99] MUSLEA, I.; MINTON, S.; KNOBLOCK, C. A Hierarchical Approach to Wrapper Induction. In: CONFERENCE ON AUTONOMOUS AGENTS, 3., 1999. **Proceedings…** Seatle, WA: ACM Press, 1999. p. 190-197.

Disponível em: <http://www.isi.edu/muslea/PS/hwi_aa99.ps.>. Acesso em: mar. 2001.

[MUS 2001] MUSLEA, I.; MINTON, S.; KNOBLOCK, C. Hierarchical wrapper induction for semistructured information sources. **SIGMOD Record,** New York, v.4, n.1-2, p. 93-114, June 2001.

[PAP 95] PAPAKONSTANTINOU, Y.; GARCIA-MOLINA, Hector; WIDOM, Jennifer. Object Exchange Across Heteregoneus Information Sources. In: INTERNATIONAL CONFERENCE ON DATA ENGINE, 1995. **Proceedings...** [S.l.: s.n.], 1995.

[RIB 99] RIBEIRO-NETO, B.; LAENDER, A. H. F.; DA SILVA, A. S. Extracting semi-structured data through examples. In: ACM CIKM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, CIKM, 1999. **Proceedings…** [S.l.: s.n.], 1999.

[SAH 2001] SAHUGUET, A.; AZAVANT, F. Building intelligent Web applications using lightweight wrappers. **SIGMOD Record,** New York, v. 36, n. 3, p. 283-316, June 2001.

[SIL 2001] SILVEIRA, I. C. da. **Extração Semântica de Dados Semi-Estruturados Através de Exemplos e Ferramentas Visuais.** 2001. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre.

[SOD 99] SODERLAND, S. Learning information extraction rules for semi-structured and free text. **SIGMOD Record,** New York, v. 34, n. 1-3, p. 233-272. June 1999.