MICHELLE DE LOURDES PIMENTA

AVALIAÇÃO DA USABILIDADE DE *SITE*S DE BUSCA QUE UTILIZAM TÉCNICAS DE EXTRAÇÃO DE DADOS SEMI-ESTRUTURADOS

Monografia apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do Curso de Ciência da Computação, para obtenção do título de Bacharel.

Orientadora
Profa. Olinda Nogueira Paes Cardoso

LAVRAS
MINAS GERAIS - BRASIL
2003

MICHELLE DE LOURDES PIMENTA

AVALIAÇÃO DA USABILIDADE DE *SITE*S DE BUSCA QUE UTILIZAM TÉCNICAS DE EXTRAÇÃO DE DADOS SEMI-ESTRUTURADOS

Monografia apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do Curso de Ciência da Computação, para obtenção do título de Bacharel.

APROVADA em 18 de Junho de 2003

_____

Prof. André Luiz Zambalde

_____

Profa. Olinda Nogueira Paes Cardoso
(Orientadora)

LAVRAS
MINAS GERAIS – BRASIL

DEDICATÓRIA

Dedico este trabalho aos meus pais, João e Cecília, aos meus irmãos, Alex e Marcelo; à Bruna e a Mica pela amizade, aos meus colegas de sala pelos ótimos momentos juntos, e a todos que de alguma forma contribuíram para mais essa etapa em minha vida.

AGRADECIMENTOS

À minha Orientadora Olinda, pela atenção.

Avaliação da Usabilidade de Sites de Busca que Utilizam Técnicas de Extração de Dados Semi-estruturados

A *Web* vem se tornando um enorme repositório de dados dos mais variados domínios. Na maioria das vezes esses dados não possuem estrutura definida, sendo chamados de dados semi-estruturados. Devido a essa heterogeneidade dos dados, é cada vez mais difícil recuperar e manipular informações da *Web*. Com o objetivo de diminuir essas dificuldades, várias ferramentas e técnicas de extração de dados semi-estruturados têm sido estudadas e desenvolvidas, além de modelos de dados e linguagens de consulta. Outra grande dificuldade por parte dos usuários é que os mecanismos de recuperação de dados existentes (máquinas de busca) não são construídos levando em consideração a usabilidade dos sistemas, e nem sempre retornam o que os usuários procuram, gerando inúmeras insatisfações. Daí a importância de se construir ferramentas que, além de eficientes e rápidas, sejam fáceis de utilizar e entender. O objetivo deste trabalho foi avaliar a usabilidade de um *site* de busca que utiliza técnicas de extração de dados semi-estruturados no processo de recuperação das informações, fazendo um comparativo com o modelo atual das máquinas de busca existentes.

Palavas-chave: Dados semi-estruturados, extração de dados, usabilidade, *Web*, recuperação de informação.

Usability Evaluation of Search Sites that use Semi-structered Data Extraction Techniques

The Web is turning in an enormous repository of data of the most varied domains. Most of the time those data don't possess defined structure, being called semi-structured data. Due to that heterogeneity of the data, it is more and more difficult to recover and to manipulate information of the Web. With the objective of reducing those difficulties, several tools and techniques of extraction of semi-structured data have been studied and developed, besides models of data and consultation languages. Another great difficulty on the part of the users is that the mechanisms of recovery of existent data (search machines) they are not built taking in consideration the usability of the systems, and not always they come back what the users seek, generating countless dissatisfactions. Then the importance of building tools that, besides efficient and fast, be easy to use and to understand. The objective of this work was to evaluate the usability of a search site that uses techniques of extraction of data semi-structured in the process of recovery of the information, making a comparative one with the current model of the existent search machines.

Keywords: semi-structured data, data extraction, usability, *Web,* information retrieval.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

OO	– Orientados a Objetos
HTML	– *HyperText Markup Language*
XML	– *Extensible Markup Language*
BD	– Banco de Dados
OEM	– *Object Exchange Model*
SQL	– *Structured Query Language*
DTD	– *Document Type Definition*
W3C	– *World Web Consortium*
SGML	– *Standard Generalized Markup Language*
OId	– *Object Identifier*
WWW	– *World Wide Web*

Capítulo 1-Introdução

Com a evolução da Internet, a *World Wide Web (Web)* tornou-se um vasto repositório de dados dos mais variados tipos e formatos. Entretanto, os dados disponíveis na *Web* são, em geral, difíceis de serem utilizados e manipulados pela maioria dos usuários da Internet. A principal dificuldade deve-se ao fato de que esses dados não podem ser consultados e manipulados através de técnicas de indexação e consulta como as encontradas em ambientes tradicionais de banco de dados (BD). Outra grande dificuldade dos usuários é que a maioria dos*sites* de recuperação de dados é difícil de entender e manipular; não ajudam o usuário a formular suas consultas, ou seja, não são muito amigáveis. Além disso, o resultado de uma busca geralmente não traz somente os dados mais relevantes para o usuário, gerando insatisfação e tédio no seu uso, tornando o processo de busca bastante exaustivo.

De uma maneira geral, as únicas formas utilizadas para recuperar informação na *Web* são através de buscas por palavras-chave e/ou navegação (*browsing*), o que limita o tipo de consulta que se pode fazer e, principalmente, a posterior manipulação dos dados obtidos para outra finalidade. *Browsing* não é conveniente para localizar um item de dados particular, porque a navegação através de *links* pode se tornar cansativa. Pesquisa por palavras-chave é algumas vezes mais eficiente que *browsing,* mas freqüentemente retorna uma quantidade de informação além da capacidade que o usuário pode manipular.

Como a grande maioria dos documentos disponíveis na *Web* não possui uma estrutura rígida, ou seja, não possui um esquema pré-definido e os dados não são fortemente tipados, é comum que haja

variação de tipos de dados que representam um mesmo atributo em fontes de dados diferentes. No caso dos currículos, pode ser que o atributo endereço em um documento esteja especificado como Rua, número e complemento em outro documento. Esses tipos de dados são chamados dados **semi-estruturados**.

Para recuperar os dados mais eficientemente da *Web*, alguns pesquisadores têm recorrido a técnicas de bancos de dados. No entanto, estas técnicas requerem que os dados sejam estruturados, o que não ocorre na verdade. Para resolver este problema, uma possível estratégia é a de extrair os dados de fontes de dados semi -estruturados para povoar um banco de dados num modelo tradicional para posterior manipulação. A extração de dados semi-estruturados na *Web* tem sido estudada sob diversos aspectos, seja através da construção manual de programas de extração, ou por meio de ferramentas semi-automáticas.

Para modelar esses dados semi-estruturados foram criados vários modelos e linguagens, dentre eles o OEM (*Object Exchange Model*) e o XML (*Extensible Markup Language*), já que os modelos e linguagens para Banco de Dados tradicionais não são adequados. Assim como a modelagem de dados semi-estruturados, foram surgindo mecanismos de consulta, onde o principal objetivo é consultar conjuntos de documentos como se fosse um BD, permitindo consultas mais rápidas e eficientes.

O uso do HTML como padrão para recuperar informações na *Web* vem se tornando bastante obsoleto devido às suas limitações. Com o objetivo de resolver essa questão um grupo de trabalho da *World Web Consortium* (W3C) desenvolveu o XML. O uso do XML é possível em aplicações totalmente diferentes daquelas hoje utilizadas por meio do padrão HTML, dentre elas para recuperação de dados semi-estruturados na *Web*.

Além de um melhor processo de buscas e recuperação de dados, também a Interface do usuário deve auxiliar o entendimento de expressões e informações necessárias. Ela deve também ajudar usuários a formular suas pesquisas, selecionar entre fontes de informação disponíveis, entender resultados encontrados e manter rastro ou histórico do progresso da sua busca [BAE 99].

Os estudos, pesquisas e, em conseqüência, as ferramentas que estão sendo desenvolvidas para recuperar dados semi-estruturados na *Web*, principalmente para extrair dados semi-estruturados, ainda não focalizam a importância da interface de um produto Como são ferramentas que ainda estão em fase de desenvolvimento e testes, elas exigem do usuário conhecimentos mais avançados em informática, ou seja, não estão sendo projetadas levando em consideração a usabilidade.

1.1 Motivação

A partir do momento em que o formato XML começou a firmar-se como um padrão para a representação e troca de dados na *Web*, pesquisas foram iniciadas no sentido de desenvolver mecanismos próprios para esta linguagem, incluindo modelos de dados, linguagens de consultas e mesmo sistemas gerenciadores de banco de dados específicos.

Paralelamente, algumas abordagens procuram desenvolver alternativas que tornem possível a aplicação de metodologias e conceitos tradicionais da área de banco de dados para o tratamento de dados semi-estruturados. Estas abordagens procuram, na sua grande

maioria, aproveitar o grande número de técnicas desenvolvidas nas últimas décadas para sistemas gerenciadores de bancos de dados relacionais, além de permitir que se utilizem recursos já disponíveis nas organizações, pela adaptação dos novos tipos de dados às estruturas existentes.

Como exemplo de aplicação que pode ser desenvolvida, suponhamos um número variado de *sites* de hotéis de uma cidade turística, cada um com informações referentes a um hotel. Atualmente para realizar uma busca do tipo: verificar dentre esses hotéis qual o mais barato para passar um feriado com a família, o usuário teria que utilizar um *site* de busca e navegar dentre as páginas retornadas analisando cada uma delas separadamente. Uma outra forma de extrair esta informação, poderia ser feita através de uma interface que realiza essa pesquisa de forma rápida e eficiente, e monta dinamicamente um novo *site* contendo informações de todos os hotéis, utilizando o padrão XML. Esse tipo de consulta pode ser realizado utilizando o padrão HTML, mas o nível de dificuldade é bem maior, se comparado com um construído com o XML.

Além da preocupação com a eficiência e rapidez na recuperação de informações na *Web*, o que tem sido alvo de estudos é a preocupação com a interface e facilidade de uso do *site*, pois de que adianta uma máquina de busca eficiente e rápida se para utilizá-la o usuário precisa de conhecimentos avançados em informática? Estamos caminhando para uma fase em que não mais vale só ter um *site* ou sistema que funcione adequadamente, com milhares de recursos avançados. Ele, além disso, tem que ser amigável ao usuário e possuir uma boa interface.

Hoje em dia, como a quantidade de produtos semelhantes no

mercado e de boa qualidade é grande, o fator que acaba sendo decisivo na hora de escolher o *site* a ser utilizado pelos usuários é a interface e as facilidades de uso que ele proporciona.

1.2 Objetivos do Trabalho

O objetivo geral desse trabalho é avaliar a usabilidade de um *site* de busca que utiliza técnicas de extração de dados semi-estruturados no processo de recuperação das informações, fazendo um comparativo com o modelo atual das máquinas de busca existentes.

O enfoque principal do projeto foi extração de dados semi-estruturados na *Web*, onde foram mostradas as principais técnicas e ferramentas utilizadas na extração de informações.

Uma visão geral sobre os seguintes tópicos foi abordada:

- Dados semi-estruturados na *Web*: principais características, como eles são organizados e a dificuldade de se recuperar informações na *Web*;
- Extração de dados semi-estruturados, mostrando as principais ferramentas utilizadas na extração de dados;
- Modelagem de dados semi-estruturados;
- Principais linguagens de consulta;
- Os padrões HTML e XML: suas principais características, vantagens e desvantagens de se utilizar cada um na extração de dados semi-estruturados. Um estudo comparativo será feito entre esses dois padrões (HTML e XML), mostrando as limitações do padrão HTML e as vantagens de se utilizar XML para extrair dados semi-estruturados de forma mais rápida e eficiente.

Também foi feito um estudo sobre as interfaces homem-computador, sua importância no mundo de hoje, analisando principalmente a usabilidade de *sites* de recuperação de dados na *Web*. A avaliação da usabilidade das interfaces foi feita através de questionários, onde o usuário visitou uma página de teste, que também foi desenvolvida, a fim de verificar a usabilidade da mesma comparada com as máquinas de recuperação de informações existentes hoje. Vale ressaltar que foi desenvolvida apenas a interface de um *site* de busca, simulando como seria na realidade um produto que utiliza técnicas de extração de dados semi-estruturados na recuperação de informações.

Capítulo 2- Dados semi-estruturados

Dados semi-estruturados apresentam uma representação estrutural bastante heterogênea, não sendo nem totalmente não-estruturados nem estritamente tipados. Dados *Web* se enquadram nessa definição: em alguns casos os dados possuem estrutura definida, uniforme, em outros, algum padrão estrutural pode ser identificado, mas na maioria dos casos praticamente não existem informações descritivas associadas [MEL 00].

Dados semi-estruturados são aqueles cujo esquema não é pré-definido como nos BD tradicionais, podendo inclusive estar implícito nos próprios dados, além do esquema ser relativamente grande e podendo freqüentemente mudar [SIL 00].

Como um exemplo, considere um usuário acessando um *WebSite* sobre informações de venda de carros usados. Ele deseja obter a seguinte informação: Quais são os carros fabricados entre 1996 e 2000, com preço abaixo de R$12.000,00? Esse tipo de consulta pode exigir uma busca exaustiva por várias páginas do *site*, já que esses tipos de dados geralmente estão disponíveis na forma textual ou de documentos HTML, tornando as buscas inviáveis. Se os dados referentes aos carros usados estivessem armazenados em BD tradicionais (na forma de tabelas, por exemplo), a busca seria realizada de forma rápida e eficiente, não ocorrendo problemas como o demonstrado anteriormente.

As Figuras 2.1, 2.2 e 2.3 mostram as diferenças de documentos (*sites*) onde os dados estão organizados de forma não estruturada (dados semi-estruturados) e dados que possuem estrutura fixa (dados

estruturados). Estas figuras foram extraídas de fontes reais da *Web*, onde os *site*s se referem a vendas de automóveis usados.

Placa/Foto	Modelo	Ano Mod	Cor	Combustível	Valor R$	Nome	Endereço	TelRes
GVR1965 	Seat Cordoba GLX 1.8	96/96	Cinza	Gasolina	15800,00	José Henrique	Rua Oscar Trompowisk, 1242/102, Gutierrez, Bh - MG	(031) 372-2843
BEG1900 	BMW 325 i A	93/93	Vinho	Gasolina	38000,00	Pedro / Ronaldo	Rua Frei Orlando, 1125, Caiçara, Bh - MG	(031) 464-1158
CMD5013 	Ford Taurus LX	94/95	Bege	Gasolina	19500,00	Márcio	Rua Prof. Pimenta da Veiga, 307/104, Cidade Nova, Bh - MG	(031) 486-7470

Figura 2.1- Exemplo de dados semi-estruturados na *Web*

Figura 2.2- Exemplo de dados semi-estruturados na *Web*

Nas Figuras 2.1 e 2.2 os dados referentes aos carros usados não estão organizados de forma estruturada. Eles estão dispostos na forma textual. Se eles estivessem organizados de forma estruturada, definida,

como em tabelas de banco de dados tradicionais, eles estariam dispostos da seguinte maneira:

Modelo	Ano	Cor	Combustível	Valor (R$)	Portas	Marca
Elba 1.5		branca	gasolina	8.500,00	4p	Fiat
Palio EDX	97	branco	gasolina	11.500,00	4p	Fiat
Palio Weekend	98	vermelho	gasolina	17.800,00		Fiat
Tempra	92	chumbo	gasolina	8.300,00		Fiat
Tipo 2.0	95	prata	gasolina	10.000,00	4p	Fiat
Tipo 1.6	95	chumbo	gasolina	9.800,00	4p	Fiat
Tipo 1.6	95	vermelho	gasolina	9.000,00		Fiat
Uno S	86	verde	alcool	4.500,00		Fiat
Uno CS	90	vermelho	gasolina	6.000,00		Fiat
Uno Ep	96	azul	gasolina	9.800,00	4p	Fiat
Uno Mile Fire	2002	branco	gasolina	15.000,00	4p	Fiat
Seat Cordoba GLX 1.8	96/96	Cinza	Gasolina	15800,00		Import.
BMW 325 i A	93/93	Vinho	Gasolina	38000,00		Import.
Ford Taurus LX	94/95	Bege	Gasolina	19500,00		Import.

Figura 2.3 – Exemplo de dados estruturados.

Podemos notar através destes exemplos que embora as Figuras 2.1 e 2.2 contenham basicamente informações sobre os carros que estão a venda, existem muitas informações que estão presentes em determinados *sites* e em outros não, reafirmando o que já foi dito em seções anteriores em que os dados semi-estruturados não possuem estrutura bem definida, dentre outras características também já citadas.

2.1 Características de dados semi-estruturados

As principais características de dados semi-estruturados são [MEL 00]:

- **A estrutura é irregular**: coleções extensas de dados semanticamente similares estão organizadas de maneiras diferentes, podendo algumas ocorrências ter informações incompletas ou adicionais em relação a outras. Em resumo, não existe um esquema padrão para esses dados. Como exemplo, podemos citar o *curriculun vitae.*Cada pessoa constrói seu currículo contendo basicamente as mesmas informações, porém com formatos distintos*;*
- **A estrutura é implícita:** muitas vezes existe uma estrutura básica para os dados, porém, essa estrutura está implícita na forma como os dados são apresentados. É necessário realizar uma computação para obter essa estrutura;
- **Estrutura parcial:** apenas parte dos dados disponíveis pode ter alguma estrutura, seja ela implícita ou explícita. Como conseqüência, um esquema para esses dados nem sempre é completo do ponto de vista semântico e nem sempre todas as informações estão presentes;
- **Estrutura extensa:** a ordem de magnitude de uma estrutura para esses dados é grande, uma vez que os mesmos são muito heterogêneos. Supondo diferentes formatos para um *curriculun*

vitae, uma união de atributos significativos em cada formato pode produzir um esquema muito extenso;

- **Estrutura evolucionária:** a estrutura dos dados modifica-se freqüentemente junto com seus valores. Dados na *Web* apresentam este comportamento, uma vez que existe o interesse em manter os dados sempre atualizados;
- **Estrutura descritiva e não prescritiva:** dada a natureza irregular e evolucionária dos dados semi-estruturados, as estruturas de representação implícitas ou explícitas normalmente se restringem a descrever o estado corrente de poucas ocorrências de dados similares. Desta forma, não é possível prescrever esquemas fechados e muitas restrições de integridade com relação à semântica dos atributos;
- **Distinção entre estrutura e dados não é clara:** como a estrutura está embutida na descrição dos dados, muitas vezes a distinção lógica entre estrutura e valor não é clara. Pode-se ter, por exemplo, um endereço representado como um valor atômico em uma ocorrência de dado (*string*) ou como um tipo definido pelo usuário em outra ocorrência. Esta característica torna mais complexo o projeto de um BD para tais dados;

Na pesquisa relacionada a dados semi-estruturados existem três tópicos que são considerados mais relevantes: extração, modelagem e consulta de dados [HEU 02].

O processo de **extração de dados** tem por objetivo construir uma visão estruturada de dados semi-estruturados para facilitar a sua manipulação. Este processo envolve mecanismos que identificam, recuperam e estruturam dados relevantes, transformando, em geral,

dados de uma fonte de dados como a *Web* para dados adequados a um modelo de dados, seja ele semi-estruturado, estruturado, ou algum formato com maior estruturação que possa ser entendido por uma gramática de processamento de consultas [HEU 02].

Modelos de dados semi-estruturados são grafos direcionados rotulados, onde os vértices representam objetos identificáveis e as arestas são arcos para objetos componentes. Como cada ocorrência de dado pode ter uma estrutura diferente, não há restrição no número de arcos que partem de um dado objeto. OEM *(Object Exchange Mode*l) é o modelo mais popular para representação de dados semi-estruturados. Objetos OEM podem ser atômicos *(intege*r, *strin*g, *gi*f, etc) ou complexos (conjunto de referências a objetos). Todos os vértices folha do grafo apresentam tipos atômicos [HEU 02].

Linguagens de consulta a dados semi-estruturados devem permitir pesquisas em grafos. Alguns requisitos para tais linguagens são: (i) **a especificação de expressões de caminho** que percorram a estrutura de objetos semi-estruturados para obter informação ou formular uma condição sobre um atributo. A forma mais usual de especificação é a concatenação de nomes de rótulos. Tais expressões podem ainda indicar padrões de busca, quando combinadas com certos caracteres especiais. Essa facilidade é útil quando se desconhece total ou parcialmente a estrutura dos objetos; (ii) **o uso de coerção** (critérios de conversão de tipos) na comparação entre atributos de objetos, para lidar com as heterogeneidades de tipo e estrutural dos dados; (iii) **consulta ao esquema de classes de dados semi-estruturados,** quando este existir, para a identificação de características a serem consultadas [HEU 02].

2.2 Dados tradicionais X Dados semi-estruturados

Pode-se notar que as características dos dados semi-estruturados diferem bastante das características dos dados de BD tradicionais. A Tabela 2.1 apresenta algumas dessas diferenças [MEL 00]:

Tabela 2.1- Comparação entre dados tradicionais e dados semi-estruturados

Dados Tradicionais	**Dados semi-estruturados**
Esquema predefinido	Nem sempre há um esquema predefinido
Estrutura regular	Estrutura irregular
Estrutura independente dos dados	Estrutura embutida nos dados
Estrutura reduzida	Estrutura extensa
Estrutura fracamente evolutiva	Estrutura fortemente evolutiva
Estrutura prescritiva	Estrutura descritiva
Distinção entre estrutura e dados é clara	Distinção entre estrutura e dado não é clara

Outra característica própria de dados semi-estruturados, a ser considerada, é que neste tipo de dado normalmente a noção de esquema é criada *a posterior*i, após a existência dos dados, com base em uma investigação de suas estruturas particulares e da análise de similaridades e diferenças. Em bancos de dados tradicionais, utiliza-se a noção de esquema *a prior*i, o qual é definido antes do armazenamento de qualquer dado no banco de dados.

2.3 Extração de dados semi-estruturados

A extração de dados semi-estruturados tem sido estudada sob diversos aspectos, seja através da construção manual de programas de

extração, ou por meio de ferramentas semi-automáticas. Identificam-se duas abordagens principais no processo de extração:

- **Extração sintática:** este processo se baseia na análise de padrões presentes em dados semi-estruturados, exigindo que o documento seja todo percorrido na busca da informação desejada. Nesse método de extração, cada alteração do documento exige uma alteração no programa de especificação.
- **Extração semântica:** esse método de extração é mais preciso que o sintático, pois, ele é capaz de identificar conceitos e suas relações em fontes de dados, sendo que as alterações nos processos de extração só ocorrem quando o esquema conceitual é alterado.

A Tabela 2.2 mostra um comparativo entre essas duas técnicas de extração: a sintática e a semântica:

Tabela 2.2- Comparação entre extração sintática e extração semântica

	Extração Semântica	**Extração Sintática**
Desenvolvimento	Mais Complexo	Menos complexo
Resultados	Precisos	Precisos
Base da Extração	Domínio do Problema	Instâncias de Dados
Alteração no Processo	Quando altera o esquema	Quando altera a instância

O processo de extração é responsável pela recuperação de informações contidas nas fontes de dados, mas sem retornar os documentos completos que contenham estas informações. Ele geralmente transforma dados semi-estruturados de uma fonte de dados para dados que são adequados a um determinado modelo, seja ele

estruturado ou não. O resultado é posteriormente processado de alguma forma [MEL 00].

Existem diferentes abordagens e técnicas para extração de dados semi-estruturados. Para que se possa fazer a extração dos dados de forma mais eficiente estão sendo desenvolvidas várias ferramentas, como as semi-automáticas, por exemplo.

A seguir serão apresentadas algumas dessas ferramentas e uma visão geral de cada uma delas [TEI 00]:

NoDoSE: (*Northwestern Document Structure Extractor*) é uma ferramenta interativa para determinar semi-automaticamente a estrutura de documentos que contém informação semi-estruturada e então extrair seus dados. Usando uma interface gráfica, o usuário indica regiões de interesse no texto e a partir daí o processo detecta automaticamente, em documentos com comportamento descritivo similar, os mesmos dados relevantes [MEL 00].

W4F: é uma *toolkit* (biblioteca de rotinas para construção de aplicações) que permite o projeto de *wrappers* (Programas que atuam como módulos intermediários entre aplicações e fontes de dados semi-estruturados, sendo responsável pelo acesso aos dados) para fontes *Web*. W4F divide o processo de desenvolvimento de um *wrapper* em três fases: primeiramente é necessário descrever como acessar o documento, em seguida, quais pedaços de informação extrair e finalmente qual estrutura alvo utilizar para armazenar os dados que foram extraídos.

XWRAP: é uma ferramenta para geração de *wrappers* que tem três características: (1) separar explicitamente as tarefas de construção de *wrappers* que são específicas para uma determinada fonte, daquelas que são repetidas para qualquer fonte; (2) fornecer uma interface amigável para permitir que os usuários realizem o processo de desenvolvimento de *wrappers* com poucos cliques do mouse; (3) a fase de geração de código é feita em duas fases.

2.3.1 A Ferramenta DEByE

Uma ferramenta bastante estudada e utilizada é a Ferramenta **DEByE** (*Data Extraction By Example*)[LAE 99]. É baseada em uma abordagem na qual o processo de extração de dados é totalmente guiado por exemplos fornecidos pelo usuário. A DEByE é constituída de dois módulos principais: um módulo de **Interface Gráfica** que é utilizada pelo usuário para especificação de exemplos dos objetos a serem extraídos e um módulo **Extrator** que, com as informações geradas pelo módulo de Interface Gráfica , realizará a extração propriamente dita.

2.4 Modelagem de Dados

Para se fazer modelagem de dados semi-estruturados, os modelos para BD tradicionais são inadequados, pois os dados não possuem a mesma estrutura. Assim, os modelos para dados semi-estruturados devem ser bastante flexíveis para suportar representações heterogêneas de dados semanticamente iguais.

Para representar dados semi-estruturados o mais usual é através de **grafos direcionados rotulados**, pois a estrutura particular de cada dado já está embutida no próprio grafo. Na estrutura de grafos rotulados, os vértices representam objetos e as arestas são arcos para outros objetos que fazem parte da sua estrutura, que é hierárquica.

Existem vários modelos que representam dados semi-estruturados. Dentre eles o modelo OEM (*Object Exchange Model*), que foi o pioneiro para dados semi-estruturados.

2.4.1 Modelo OEM

Esse modelo foi proposto pela Universidade de Stanford, sendo aplicado a projetos tais como o TSIMMIS, do departamento de Ciência da Computação da mesma universidade [MEL 00]. Ele é baseado em grafos, sendo os objetos os vértices do grafo.

Um objeto OEM é uma quádrupla (**label, oid, type, value**), onde :

- **label**: cadeia de caracteres
- **oid**: identificador do objeto
- **type**: complexo ou atômico (i*nt, string, gif, jpeg* …)
- **value**: se o tipo é complexo, conjunto de oid's; senão um valor atômico.

Este modelo possui algumas características, tais como:

- Facilidade de ser editado;
- Facilidade de leitura;
- Compatibilidade com atuais variações;
- Facilidade de gerar um *parsing* por um programa;

- Possibilidade de gerar extensões para um modelo futuro.

A Figura 2.4 ilustra um exemplo de um objeto semi-estruturado representado no modelo OEM:

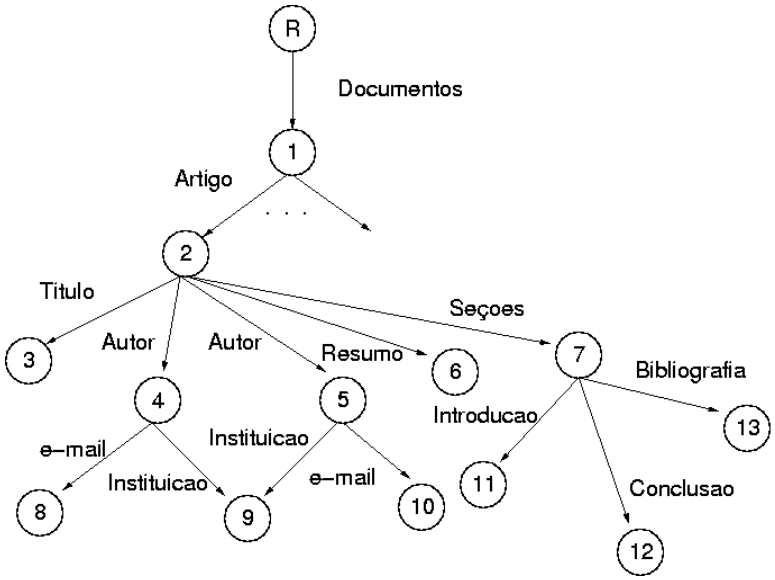


Figura 2.4- Representação do modelo OEM para um objeto semi-estruturado

Segundo [MEL 00], OEM é um modelo flexível, pois não considera a existência de um esquema fixo para um conjunto de dados. Toda informação esquemática está incluída nos rótulos, que podem mudar dinamicamente. OEM é considerado um modelo para instâncias de dados semi-estruturados, pois representa valores de dados e associa rótulos a cada valor para descrever o seu significado.

2.5 Linguagens de Consulta

Diversos tipos de linguagens de consultas para a *Web* têm sido propostos nos últimos anos, devido a crescente utilização de documentos semi-estruturados para representar informações. Os mecanismos de consultas existentes que utilizam buscas por palavras-chave, por exemplo, não são adequados para documentos semi-estruturados. Por esse motivo se tornou necessário desenvolver linguagens de consultas para que esses tipos de dados pudessem ser consultados de forma mais eficiente e abrangente.

Dentre as linguagens de consulta existentes algumas são extensões da linguagem OQL (versão da SQL para utilizar em bases de dados Orientados a Objetos - OO), como a UnQL e a StruQL, que também utilizam grafos rotulados como um modelo de dados flexíveis, enfatizando a característica de consultar o esquema dos dados, tratando suas irregularidades, como por exemplo, a falta ou repetição de campos e registros heterogêneos. Vale lembrar que essas linguagens não foram desenvolvidas para *Web* especificamente [SIL 00].

Outras propostas abordam linguagens de consulta para *Web*, como a W3QL, *Web*SQL e a *Web*OQL, que podem combinar condições de padrões de texto que aparecem no conteúdo das páginas com os padrões de gráficos que descrevem as estruturas de ligações das páginas.

Segundo [CAR 98], a W3QL é uma linguagem poderosa, que trata navegação, consultas e explora a estrutura interna dos objetos, porém seu uso está limitado ao ambiente UNIX.

A *Web*SQL tem como proposta modelar a *Web* como sendo um banco de dados relacional, composto por sua vez de duas relações

virtuais: *document* e *anchor*. A primeira relação apresenta uma tupla para cada documento na *Web*, e a segunda, apresenta uma tupla para cada âncora em cada documento da *Web*. Ela combina navegação e consultas de forma clara e eficiente, fazendo distinção ente objetos armazenados de forma remota ou local, porém não explora a estrutura interna dos documentos da *Web*.

Já a linguagem *Web*OQL provê acesso à estrutura interna dos objetos, permitindo a reestruturação dos dados semi-estruturados da *Web*, mas possui uma linguagem mais complexa e pouco poderosa.

Considerando a tendência de que XML venha a ser adotada como padrão para troca e intercâmbio de informações na *Web*, já existem propostas de linguagens como LOREL, XML-QL e XQL.

A linguagem de consulta LOREL utiliza como base a OQL, com modificações e extensões para suportar dados semi-estruturados, sendo que, originalmente, baseava-se no modelo OEM, que serve como base para várias outras linguagens. Hoje, a linguagem trabalha com XML.

XML-QL é capaz de realizar diversas tarefas como a extração de documentos extensos, conversão de dados entre banco de dados e documentos XML, mapeamento de dados XML entre diferentes DTD (*Document Type Definition*) e integração entre dados XML de múltiplas fontes, através da realização de junções e outras operações encontradas em SQL. Sua sintaxe combina elementos da sintaxe de XML com elementos de sintaxe tradicionais de linguagens de consulta de BD [GAG]. As consultas XML-QL podem tanto recuperar dados XML quanto construir novos documentos.

Já a XQL é uma proposta a fim de expressar consultas a documentos XML. As consultas em XQL sempre atuam sobre um determinado contexto. De uma forma sintética, pode-se dizer que o

contexto é o escopo da consulta dentro da árvore do documento XML [KAD 99].

Pode-se considerar que a principal diferença entre XML-QL e XQL é que a primeira é voltada à consulta a dados XML e a segunda é voltada à consulta a documentos XML [MEL 00].

2.6 Ontologias

Uma ontologia [GRU 93] é um conjunto de termos (palavras) hierarquicamente estruturados para a descrição de um domínio que pode ser usado como um esqueleto fundamental para uma base de conhecimento.

As ontologias, entre outras coisas, colaboram no sentido de se obter uma *Web* onde os recursos disponíveis são acessíveis não somente por seres humanos, mas também por processos automatizados.

Ontologias podem ser usadas para buscar em bases de informações (*Web*) recursos desejados, como por exemplo, documentos, páginas *Web*, etc. Fazendo o uso de ontologias na busca de informações a precisão dessas buscas aumenta consideravelmente. Além disso, o tempo total gasto na busca desejada é reduzido.

Quando não é encontrada uma resposta perfeita à consulta desejada, a estrutura semântica da ontologia capacita o sistema a retornar respostas que mais se assemelham com a busca desejada.

A necessidade de linguagens e ferramentas para construção, consulta e integração de ontologias é indispensável. Entretanto, somente a representação de conhecimento e informação não é

suficiente. Pessoas e/ou agentes da *Web* que buscam informações necessitam usar e consultar ontologias e os recursos inerentes a elas. Com isso, fazem com que a necessidade de ferramentas de armazenamento e consultas em ontologias seja cada vez maior.

2.6.1 Características

Algumas características desejadas de uma ontologia são [MEL 00]:

- **Aberta e dinâmica:** deve ser capaz de suportar ajustes decorrentes de mudanças na estrutura ou comportamento do domínio;
- **Escalável e interoperável:** deve ser facilmente escalável, considerando um domínio amplo, e adaptável a novos requisitos. Deve também ser possível integrar várias ontologias em uma nova ontologia quando o tratamento de diferentes vocabulários conceituais é requerido;
- **Fácil manutenção:** sua manutenção não deve ser complexa. Portanto, deve ser de fácil compreensão;
- **Coerente com o texto:** não devem existir termos muito específicos para não tornar complexa a associação com as fontes de dados e futuras integrações com outras ontologias.

A especificação de uma ontologia como modelo conceitual pode proporcionar muitas vantagens aos usuários de dados semi-estruturados, entre eles o conhecimento do domínio de interesse. Além disso, considerando que dados semi-estruturados, na sua maioria, são

documentos eletrônicos, a busca por seu conteúdo pode tornar-se menos exaustiva se o usuário puder utilizar uma ontologia para indicar a estrutura das informações desejadas. Outra vantagem é que o sistema utilizado para executar esta tarefa se beneficia tendo uma ontologia como suporte ao processamento de consultas e à extração de informações.

Desta forma, é possível a aplicação de ontologias para armazenar e gerenciar o vocabulário de um domínio específico, de forma a permitir que se defina um esquema para consultas sobre dados semi-estruturados, independentes de qualquer aplicação ou das fontes de dados consideradas.

Uma importante vantagem na aplicação de ontologias é que ela provê uma interpretação semântica unificada para diferentes representações de dados semi-estruturados referentes a um mesmo domínio.

Em [MEL 00], podemos verificar alguns exemplos de propostas de aplicação de ontologias a dados semi-estruturados, tais como: Observer, SHOE, Ontobroker, Proposta de Embley, MOMIS, *Web*KB.

Capítulo 3- HTML e XML

O HTML é uma linguagem de marcação desenvolvida pelo W3C (*World Wide Web Consortium*). É derivado da SGML (*Standard Generalized Mark-up Language*), que é uma Linguagem de Marcação Padrão Genérica. O SGML não é exatamente uma linguagem (como o HTML), e sim um conjunto de padrões para a criação de linguagens de formatação de documentos. Através do SGML são criados padrões de documentos, como tamanho de laudas e fontes de cartas comerciais ou relatórios oficiais. Enquanto o SGML padroniza, linguagens como o HTML implementa na prática esses padrões, ou parte deles.

O padrão HTML é formado por um conjunto de marcadores pré-determinados (*tags*) utilizados principalmente para publicação de documentos na *Web*. Foi desenvolvida para ser simples tanto para autores dos documentos de marcação de hipertextos como para os desenvolvedores de *browsers*, os quais são interpretadores de códigos HTML.

Essa linguagem possui algumas vantagens. São elas:

- Fácil de usar (proliferação de páginas *Web*);
- Bom suporte industrial para o usuário;
- Autores escrevem páginas mostrando informações;
- Portabilidade e liberdade através da rede.

Exemplo de um documento HTML:

```
<HTML>
<HEAD>
<TITLE>HTML Básico</TITLE>
```

```
</HEAD>
<BODY>
<H1>Este é o primeiro nível de cabeçalho</H1>
Bem-vindo ao mundo do HTML.
Este é o primeiro parágrafo.<P>
E este é o segundo.<P>
</BODY>
</HTML>
```

Apesar de sua grande popularização e de ser bastante simples, HTML contém várias restrições, devido principalmente a essa pré-determinação de seus marcadores (*tags*).

HTML possui várias desvantagens, principalmente com relação à sua flexibilidade. Dentre elas podemos citar:

- Apresenta *tags* fixas;
- O conteúdo da página e sua apresentação estão implementados juntos;
- Armazenamento de muitas informações pobres;
- Informações armazenadas em HTML e convertidas em SGML;
- Dificuldade de alterar ou manter os documentos;
- HTML possui pouca ou nenhuma estrutura semântica, pois seus elementos são agrupados sem seguir nenhuma estrutura pré-definida;
- Se houver necessidade de mudanças na forma de apresentação dos documentos, um novo documento HTML deve ser gerado.

3.1 XML (*Extensible Markup Language*) - Sucessor do HTML

A linguagem XML (*Extensible Markup Language*) surgiu como uma tentativa para cobrir as restrições do HTML, uma nova linguagem que fornece funcionalidades superiores às fornecidas pelo HTML. XML também é um sub conjunto da SGML. A diferença é que HTML é uma linguagem de marcação específica enquanto XML é uma linguagem de marcação genérica para documentos de hipertexto. HTML é específico, pois, seus elementos são pré-definidos. XML por outro lado, não define seus elementos. Por isso ele é chamado genérico. É o autor que define os elementos de seu documento.

A linguagem XML tem alcançado crescente aceitação como um padrão não só para a descrição de documentos, como também como uma linguagem de descrição de dados para todos os tipos de informações disponíveis. Desenvolvida pelo *World Web Consortium* (W3C), caracteriza-se como uma linguagem de marcação de documentos, cujo objetivo principal é possibilitar a entrega de estruturas de dados auto-descritivas de diferentes complexidades para aplicações que requerem tais estruturas [FLO 99].

Estruturalmente, cada documento XML consiste de um conjunto de elementos, delimitados por uma *tag* inicial e uma *tag* final. Cada elemento tem um tipo, identificado por um nome, e pode ter um conjunto de especificações de atributos. Cada especificação de atributo tem um nome e um valor. Adicionalmente cada elemento pode ter também uma lista arbitrária de subelementos [FLO 99]. XML pode ainda ser vista como uma linguagem de modelagem de dados.

Dentre as vantagens oferecidas pela XML podemos citar:

- Extensível;
- Não há grupos fixos de *tags*;
- Bom para processamento de informações (igual a SGML);
- Adequado para rede;
- Processamento pode ser feito ao lado do cliente;
- Não é tão completo como a SGML e, por pouco tempo, os *browsers*.

Exemplo de um documento XML:

```
<?xml version='1.0' ?>
<!DOCTYPE livro SYSTEM "livro.dtd">
<livro>
  <titulo> XML e Java  </titulo>
            <autor> João Silva </autor>
  <conteudo>
                        <capitulo focus="XML">      Introdução </capitulo>
              <capitulo focus="XML"> DTD </capitulo>
                        <capitulo focus="XML"> Elementos </capitulo>
                        <capitulo focus="Java"> SAX </capitulo>
                        <capitulo focus="Java"> DOM </capitulo>
            <conteudo>
            <bibliografia> bla bla bla </bibliografia>
  <copyright> &ufpbCopyright; </copyright>
</livro>
```

Apesar das inúmeras vantagens do XML ela possui algumas desvantagens, como por exemplo, a dificuldade de ser programada e a falta de *browsers* que suportam essa tecnologia.

3.1.1 DTD (Document type Definition)

Segundo [PIT 00], DTD (Document type Definition) é um arquivo que é separado do restante do documento XML principal e que fornece um conjunto de regras para o documento XML ao qual ele é anexado. DTD é o que realmente distingue o XML das demais linguagens de marcação. Uma DTD fornece uma lista de elementos, atributos, notações e entidades contidas em um documento, além dos relacionamentos entre eles. DTD especificam um conjunto de regras para a estrutura de um documento. É apresentado a seguir um exemplo de DTD de uma bibliografia que mostra um elemento LIVRO com cinco subelementos (**nome_autor, titulo, editora, evento, ano_publicacao)**. Todos os subelementos são do tipo PCDATA (*parsed character data*), ou seja, é o tipo de conteúdo de elemento constituído apenas por texto puro.

Exemplo de uma DTD:

```
<!ELEMENT Livro (nome_autor+, titulo, editora, evento?, ano_publicacao)+>
<!ELEMENT nome_autor (#PCDATA)>
<!ELEMENT titulo (#PCDATA)>
<!ELEMENT editora (#PCDATA)>
<!ELEMENT evento (#PCDATA)>
<!ELEMENT ano_publicacao (#PCDATA)>
```

No exemplo acima, o elemento **nome_autor** é seguido pelo sinal de “+” indicando que um livro pode conter mais de um autor. O elemento **evento** é opcional, isto é, ele está sucedido pelo caracter “*?*”, indicando que um livro pode não ter sido gerado de um evento. Assim, esse elemento ficará vazio dentro do arquivo de código-fonte que marca o documento. Já o sinal “+” que aparece após o último subelemento do elemento LIVRO diz ao analisador ou processador que o elemento

LIVRO pode aparecer uma ou mais vezes no documento, ou seja, qualquer disciplina pode ter em sua bibliografia diversos livros.

A utilização de Sistemas Gerenciadores de Banco de Dados Relacionais comerciais para manipulação e consulta a documentos XML tornou-se possível devido à existência das DTD. A partir do processamento da DTD, o esquema relacional é gerado, e em seguida, o documento XML é analisado e os dados nele contidos são armazenados em tabelas de um banco de dados relacional. Estando os dados armazenados em tabelas, será possível então a manipulação e consulta desses dados pelos usuários.

3.2 XML comparada com HTML

HTML define o modo como o *browser* mostra a informação. Esta é a finalidade de uma linguagem de marcação. Editores de texto, por exemplo, tem sua própria linguagem de marcação para descrever como o texto será apresentado. HTML porem é fixa em relação ao que as *"tags"* significam e devem ser interpretadas.

XML, como o nome diz, é EXTENSÍVEL. O usuário pode escolher qualquer marcação que deseje para descrever e definir os dados. Como XML define a estrutura e adiciona significado aos dados, os projetistas de aplicações podem fazer melhor uso desses dados do que hoje é possível.

Segue abaixo algumas características que diferem HTML da XML:

→ **XML**

- Define o conteúdo (dados);

- *Tags* descrevem os dados, como temperatura, umidade, etc;
- *Tags* definidas pelo criador do documento;
- Apresentação definida por folhas de estilo;
- Dados separados da apresentação e do processamento dos dados.

→ HTML

- Descreve o formato de apresentação;
- Número limitado e não extensível de *tags*;
- Inadequado para gerenciamento de grande volume de dados;
- Não oferece a funcionalidade requerida pelo comércio eletrônico.

→ O Papel da XML

A XML será usada por pessoas e organizações que possuem recursos de informação que não se encaixam no HTML. O papel da XML será maximizado em situações em que os recursos de informação são de longo prazo. Uma possibilidade é XML ser usada para melhorar a própria HTML.

→ O Papel da HTML

Em muitos casos a XML não valerá o esforço. A HTML é uma aplicação "pronta para usar". Além disso, existe uma infinidade de *software* para criação de páginas em HTML.

A XML procura ser uma linguagem mais completa e de uso mais amplo que a HTML, eliminando as complexidades existentes na elaboração de documentos com SGML. Apesar da grande popularidade do HTML pela sua simplicidade e dos inúmeros *softwares* disponíveis no mercado para sua edição, a XML tende a se tornar uma linguagem bastante difundida entre aplicações científicas, comerciais e voltada à educação.

HTML é utilizada como a linguagem universal de apresentação de documentos na *Web*, mas não é adaptada para descrever a estrutura destes documentos. As bases de dados atuais são muito rígidas para manipular dados cuja estrutura é irregular e evolui com o tempo, como é o caso dos dados semi-estruturados.

Desta maneira, procuramos demonstrar algumas das flexibilidades que a XML oferece, em comparação com as restrições impostas pelo HTML em aplicações distribuídas entre clientes e servidores *Web*.

Capítulo 4- Interfaces

As formulações de consultas sobre fontes de dados sejam eles estruturados ou não, não é uma tarefa fácil. A maioria das ferramentas que existem na atualidade não fornece facilidades de uso para a maioria dos usuários no sentido de orientá-los na especificação da consulta desejada. Sendo assim, mesmo que o usuário saiba exatamente o que deseja consultar, ele não sabe especificar, de forma precisa, os dados cujos valores devem estar no resultado da consulta, muitas vezes devido a uma interface mal projetada e desenvolvida.

De maneira geral, uma interface compreende os comportamentos do usuário e as características e facilidades do sistema, do equipamento e do ambiente. Interface não é só o que se vê na tela, mas também os periféricos, os manuais, o local de trabalho, materiais impressos e até o suporte técnico e de treinamento oferecido pelo fabricante [ZAM 01].

Para o usuário, a interface é o sistema. É parte de um sistema interativo responsável por traduzir ações do usuário em ativações das funcionalidades do sistema, permitir que os resultados possam ser observados e coordenar esta interação. É responsável pelo mapeamento das ações do usuário sobre dispositivos de entrada em pedidos de processamento à aplicação e pela apresentação em forma adequada dos resultados produzidos.

A quantidade de usuários leigos em conhecimentos de informática é grande, diferentemente dos usuários com experiência, e as dificuldades na interação com as máquinas são enormes. Estas dificuldades são geralmente provenientes da falta de experiência, das

diferenças individuais e das funções cognitivas exigidas na tarefa de interação, forçando, assim, o desenvolvimento de interfaces cada vez mais amigáveis [SHN 98].

4.1 Objetivos da Interface com o usuário

A interface homem-computador é bem menos entendida que outros aspectos dos sistemas de informações, em parte por que os seres humanos são bem mais complexos que os sistemas computacionais, e suas motivações e comportamentos são mais difíceis de compreender e caracterizar [BAE 99].

Para a construção de uma interface, várias alternativas de projeto devem ser avaliadas para comunidades específicas de usuários e para tarefas específicas. Um bom projeto para uma comunidade de usuários pode ser inapropriado para outra comunidade. O cuidadoso estudo da comunidade de usuários e do conjunto de tarefas é a base para o estabelecimento dos objetivos da interface com o usuário. Existem cinco fatores humanos que são imprescindíveis na avaliação de um sistema interativo por parte dos usuários. São eles: tempo de aprendizado, taxa de erros dos usuários, retenção com o tempo, satisfação e velocidade de resposta [ZAM 01].

A diversidade das habilidades, experiência, motivações, personalidade e estilo de trabalho dos seres humanos desafiam os projetistas dos sistemas interativos. Entender as diferenças físicas, intelectuais e pessoais dos diferentes usuários é extremamente importante para os profissionais da área de interface homem-máquina [ZAM 01].

A interface deve possuir duas características essenciais: amigabilidade e usabilidade. O termo amigável é comumente atribuído a interfaces. Para ser amigável o sistema deve ser fácil de usar, aprender, a taxa de erros deve ser mínima, o usuário, principalmente o esporádico, deve ter recordação rápida quando for utilizar o sistema, não precisando recorrer a ajudas ou manuais para realizar as tarefas de que necessita.

4.2 A importância das Interfaces nos dias atuais

Nos dias atuais, com os avanços sofridos principalmente na área de Engenharia de *Software*, encontrar no mercado *softwares* confiáveis e que funcionem corretamente não é mais tarefa difícil. Como existem muitos produtos semelhantes e de boa qualidade, o fator que acaba sendo decisivo na hora de adquirir um produto é a interface. O mesmo ocorre na *Web*. Como existe uma enorme variedade de *sites*, documentos, várias máquinas de recuperação de dados, o que faz com que o usuário seja fiel a um determinado *site*, por exemplo, é a facilidade de aprendizado, a eficiência de uso, dentre outros. Se o usuário perceber que o que ele está utilizando não mais satisfaz às suas necessidades, seja ela qual for, ele simplesmente começa a utilizar outro produto.

Por esses e outros motivos é extremamente impo rtante projetar e desenvolver produtos cada vez mais amigáveis, onde o usuário se sinta confortável e satisfeito com seu uso, principalmente na *Web,* onde a variedade e a concorrência são muito grandes.

4.3 Porque Interfaces são difíceis de serem projetadas

Embora os benefícios da melhora da usabilidade de uma interface sejam indiscutíveis, não estão resolvidos os problemas que levam à dificuldade do projeto de uma Interface. Discute-se algumas dificuldades abaixo segundo [MAI 93]:

- **A dificuldade em entender as tarefas e os usuários**
  A necessidade da interface estar diretamente ligada ao modo com que será usado requer compreensão profunda dos usuários e de suas habilidades e expectativas. Levantar este tipo de informação é difícil, sobretudo porque programadores têm dificuldade de se imaginarem efetivamente na condição de usuários comuns.

- **A complexidade inerente às tarefas e aplicações**
  Em geral, o domínio da aplicação a ser criado envolve situações de difícil modelagem - seja porque a tarefa em si é complicada, seja porque a aplicação se propõe a resolver problemas de gama extensa.

- **A variedade de aspectos e requisitos diferentes**
  Além das limitações inerentes a qualquer projeto, interfaces com o usuário envolve questões como padrões, *design* gráfico, documentação, internacionalização, e performance, entre outras. Estas questões associadas contribuem para aumentar a complexidade do desenvolvimento da interface.

- **Teoria e métodos não são suficientes para resolver o problema**
  Embora existam muitas metodologias para a criação de uma

interface boa, a maior parte dos estudos feitos a seu respeito revela que a habilidade dos projetistas é o fator primário para a qualidade das interfaces geradas. O fato de existir grande proporção de casos que sejam exceções às regras propostas nos métodos contribui para a dificuldade de se criar um método abrangente.

- **Dificuldade de se fazer um projeto iterativo**
  Embora se veja como ideal o processo de se refinar ciclicamente uma interface, este processo em si já é difícil de ser executado - muitas vezes, as modificações trazem uma piora de usabilidade, e é difícil saber quando a interface está realmente bem-elaborada. Além disso, é difícil obter resultados do uso da interface diretamente dos usuários primários, que muitas vezes não são os seus compradores e nem responsáveis.

4.4 Interfaces para *sites* de recuperação de informações na *Web*

As interfaces de busca devem prover aos usuários bons caminhos para dar início a sua busca. Uma tela vazia ou uma entrada de forma branca não ajuda o usuário a decidir como começar o processo de busca. Usuários normalmente não são muito criativos e não detalham suas expressões de busca por informações necessárias. Estudos mostram que usuários tendem a começar com perguntas muito curtas, examinam os resultados e então modificam aquelas perguntas num ciclo de retorno incremental. A pergunta inicial pode ser vista como um tipo de teste para ver que tipos de resultados são retornados e dar uma idéia de como reformular a questão. Assim, uma das tarefas de interface de

acesso a informação é para ajudar usuários a selecionarem as fontes e as coleções para sua procura [BAE 99].

A maioria das máquinas de recuperação de informações que existem hoje é difícil de ser entendida e utilizada por grande parte dos usuários da internet. O que ocorre muitas vezes é que quando o usuário pretende utilizar um sistema de busca para um objetivo qualquer, ele espera que os resultados recuperados sejam exatamente o que ele deseja e especificou na busca. Porém, o que ele recebe como resultado é geralmente um conjunto de assuntos completamente diferentes do procurado.

Isto mostra a importância de se projetar máquinas de busca que além de rápidas e eficientes, sejam também fáceis de utilizar, entender, ou seja, que sejam amigáveis.

4.4.1 Modelos de interação no processo de acesso à informação

Muitos dos processos de acesso à informação assume um ciclo de interação consistindo de especificação de consulta, recibo e análise de resultados retornados e então parando ou reformulando a consulta e repetindo o processo até que um perfeito resultado seja encontrado. O processo padrão pode ser descrito de acordo com a seqüência de passos abaixo:

(1) Começar com uma informação necessária;
(2) Selecionar um sistema e coleções para serem buscados;

(3) Formular uma pergunta;
(4) Enviar a pergunta para o sistema;
(5) Receber os resultados na forma de itens de informação;
(6) Avaliar e interpretar os resultados;
(7) Parar, ou;
(8) Reformular a questão e voltar ao passo 4.

Este simples modelo de interação usado pelas máquinas de busca *Web* é o único modelo que os buscadores de informação vêem hoje. Este modelo não leva em conta o fato que muitos usuários não agradam ao serem confrontados com uma longa lista desorganizada de resultados recuperados que não trazem endereços diretos de suas informações necessárias.

Na atualidade, os usuários aprendem durante o processo de busca. Eles analisam as informações, lendo os títulos do conjunto de resultados retornados, lendo os documentos recuperados, visualizando listas de tópicos relatados para seus termos de consultas e navegando entre *hiperlinks* de *sites Web*. O recente advento de *hyperlinks* como parte pivô do processo de busca de informação torna muito fácil a análise dos resultados e navegação sem o seu próprio processo de busca.

Uma situação relatada ocorre quando usuários encontram um termo em seus resultados que o faz mudar por um tempo a sua forma de consultar um determinado assunto, para retornar a atividade corrente inacabada mais tarde. Uma implicação destas observações é que a interface de usuário pode suportar estratégias de busca tornando-se fácil seguir rastros com resultados inesperados.

Isto pode ser aperfeiçoado, em parte, por meio de prover caminhos para gravar o progresso da estratégia corrente e para armazenar, encontrar e recarregar resultados intermediários e suportar perseguição de múltiplas estruturas simultaneamente [BAE 99].

4.5 Avaliação de Interfaces - Usabilidade

Existem várias maneiras de se avaliar interfaces. Muitas são realizadas durante o processo de desenvolvimento, onde os problemas são verificados e corrigidos antes que a aplicação termine, ou até mesmo antes de ser implementada. Mas, a maior parte dos métodos de avaliação existentes é baseada na observação e monitoração de usuários, experimentos, testes e coleta de opiniões dos usuários.

Uma das formas de se avaliar interfaces é através do teste de usabilidade.

A propriedade de uma interface que permite classificá-la quanto à sua qualidade é conhecida como **usabilidade.** Este conceito é definido tradicionalmente como a conjunção de cinco atributos: facilidade de aprendizado, eficiência de uso, retenção, minimização de erros e a satisfação [ZAM 01].

A usabilidade, portanto, preocupa-se com a interação entre o sistema e o usuário através da interface.

4.5.1 Princípios básicos da usabilidade

Segundo Nielsen, citado em [ZAM 01], os princípios básicos que devem ser seguidos durante o projeto da interface com o usuário para obter um produto final usável são os seguintes:

- Diálogo simples e natural
- Fale a língua do usuário
- Minimize a carga de memória do usuário
- Consistência
- Retorno
- Saídas claramente marcadas
- Atalhos
- Boas mensagens de erro
- Prevenção de erros
- Ajuda e documentação

4.5.2 Avaliando a usabilidade

A avaliação de usabilidade envolve complexidade por natureza. Avaliar facilidade de uso, facilidade de aprendizagem, eficácia e eficiência do sistema, além da satisfação do usuário não é uma tarefa simples. Geralmente a avaliação da usabilidade é realizada com a observação da interação de usuários no mundo real ou sob condições controladas. Os avaliadores reúnem os dados dos problemas detectados no uso e verificam se a interface suporta o ambiente e as tarefas do usuário.

Os testes de usabilidade que são realizados em laboratório, sob condições controladas, apresentam a vantagem de ter maior disponibilidade de equipamentos e de infra-estrutura, o que facilita as

observações. Por outro lado, a situação não natural pode registrar situações que não aparecem no mundo real.

Já os testes realizados em campo, com a observação da interação de usuários no mundo real apresenta como desvantagem a exigência de um tempo maior devido às condições não controladas de ruído, movimento e interrupções, mas elas são preferíveis pois as interações entre sistema e indivíduo que são perdidas em laboratório podem ser observadas. O contexto é preservado e o usuário é visto em seu ambiente natural.

4.5.3 Importância da realização de testes de usabilidade

A realização de teste de usabilidade com usuários é uma etapa imprescindível do processo de projetar interfaces. É impossível ao projetista, por exemplo, prever o comportamento dos usuários diante de uma interface, mesmo que o projeto tenha sido elaborado visando a usabilidade. Outro ponto a favor da realização de testes de usabilidade é que a facilidade de aprendizagem de *software* não pode ser julgada por avaliadores que estejam intimamente envolvidos no seu desenvolvimento. É através dos testes de usabilidade que são mostrados os problemas ou as falhas do sistema, assim como onde o sistema funciona bem. Ajuda a avaliar as características do projeto e também ajuda a fornecer idéias para o projeto através das sugestões dos usuários.

Mas como em qualquer técnica de avaliação, as vantagens dos testes de usabilidade vêm acompanhadas de algumas desvantagens

como exemplo o alto custo e a necessidade de especialistas em interfaces.

4.5.4 Técnicas para avaliação da usabilidade

Existem várias técnicas para avaliação da usabilidade. Uma delas é a que busca a opinião do usuário sobre a interação com o sistema através da aplicação de questionários e entrevistas. Ela se mostra bastante pertinente na medida em que é o usuário a pessoa que melhor conhece o *softwar*e, seus defeitos e qualidades em relação aos objetivos em suas tarefas.

A elaboração de questionários constitui um passo importante no planejamento da pesquisa. Ele consiste numa técnica para coleta de dados, com uma série de perguntas que o entrevistado deve responder.

Um questionário pode ser estruturado, semi-estruturado ou misto. O questionário estruturado é formado por questões fechadas e o semi-estruturado é formado por questões abertas. No questionário semi-estruturado as questões são padronizadas, mas as respostas ficam a critério do entrevistado. Uma vantagem deste modelo é que ele permite que o entrevistado manifeste suas opiniões, seus pontos de vista e argumentos, o que não ocorre com o questionário estruturado.

Já o questionário misto pode conter tanto questões abertas quanto fechadas.

É importante salientar que os questionários de satisfação têm uma taxa de devolução reduzida (máximo 30% retornam), o que indica a necessidade de elaboração de um pequeno número de questões

sucintas. Um espaço para opiniões e sugestões livres deve sempre ser proposto ao usuário.

O que determina o tipo de questionário a ser elaborado é o propósito da pesquisa a ser realizada.

Capítulo 5- Metodologia

Neste trabalho foi desenvolvido um modelo de interface para *site* de busca na *Web*, que utiliza técnicas de extração de dados semi-estruturados no seu processo de busca, já descritas no Capítulo 2. Vale ressaltar que apenas a interface foi desenvolvida, fazendo uma simulação de como seria na realidade um *site* de busca que utiliza essas técnicas de extração.

Este trabalho, a interface e o questionário para avaliação da usabilidade, foram realizados nos laboratórios de computação da Universidade Federal de Lavras no período de março a maio de 2003.

O objetivo principal dessa interface foi analisar a usabilidade do *site* de busca criado, visto que as ferramentas que existem para recuperar dados semi-estruturados na *Web* ainda não levam em consideração a usabilidade, pois ainda estão em fase de estudos e desenvolvimentos.

Para que a simulação fosse o mais real possível, telas de retorno foram utilizadas, tentando mostrar a rapidez e eficiência na recuperação de informações na *Web* com a utilização das técnicas vistas no Capítulo 2.

As interfaces foram construídas com o programa Dreamweaver 4.0 da Macromedia. Ele foi escolhido por que é fácil de manipular e possui uma grande variedade de recursos para a construção de *Web sites.*

5.1 Construção das Interfaces

Para a construção das interfaces, a primeira tarefa realizada foi verificar qual seria o propósito da sua construção, ou seja, os objetivos que se desejaria alcançar com as interfaces. A partir dessa análise, pôde-se então iniciar o desenvolvimento das telas.

5.2 Interfaces desenvolvidas

O Localizador, nome dado ao *site* de busca criado com técnicas de extração de dados, foi feito para compra e venda de carros, ou seja, ele é um *site* de busca específico para carros.

Para a sua construção foram visitados vários *sites* de carros reais do país objetivando verificar os tipos de consultas e recursos que cada um utilizava. O que havia de comum entre eles foi então selecionado para fazer parte da tela de busca do Localizador.

Estas interfaces possuem um fundo de tela branco para uma maior legibilidade do texto. A cor verde foi utilizada nos textos, pois além de ser uma cor escura e se destacar no branco do fundo da tela, descansa os olhos, não causando tensão no usuário como as cores vermelhas e pretas, por exemplo.

A Figura 5.1 mostra a tela principal do Localizador. Ela é uma interface simples, onde não foi usado nenhum tipo de animação, som, para não dispersar a atenção do usuário.

Ela possui uma âncora para **Ajuda**, onde o usuário pode tirar suas dúvidas a respeito de como utilizar o Localizador. Possui além da Ajuda, uma âncora com informações do Localizador e outra para Contatos. Para fazer a pesquisa desejada, existe um *link* chamado

Localizar, que irá fazer a busca de acordo com os dados selecionados para a consulta.

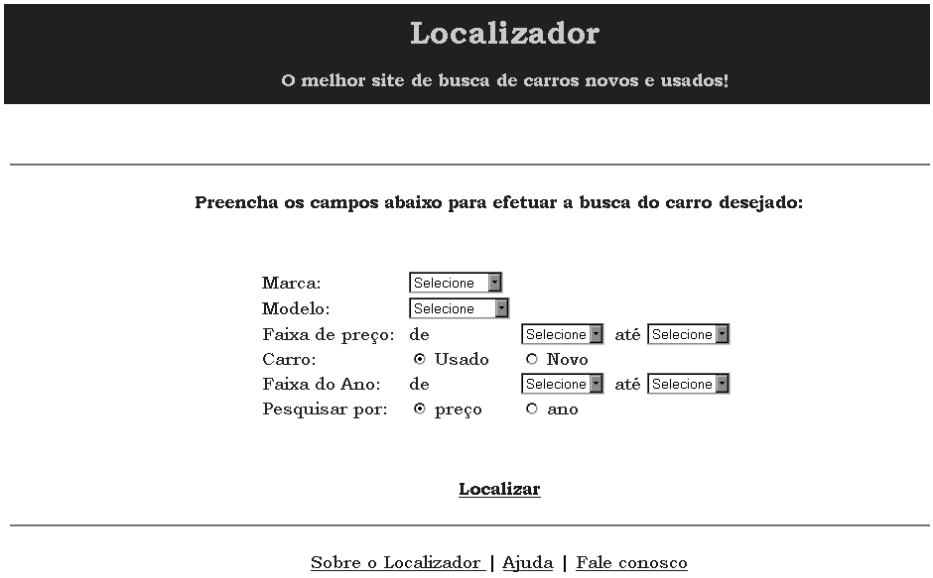
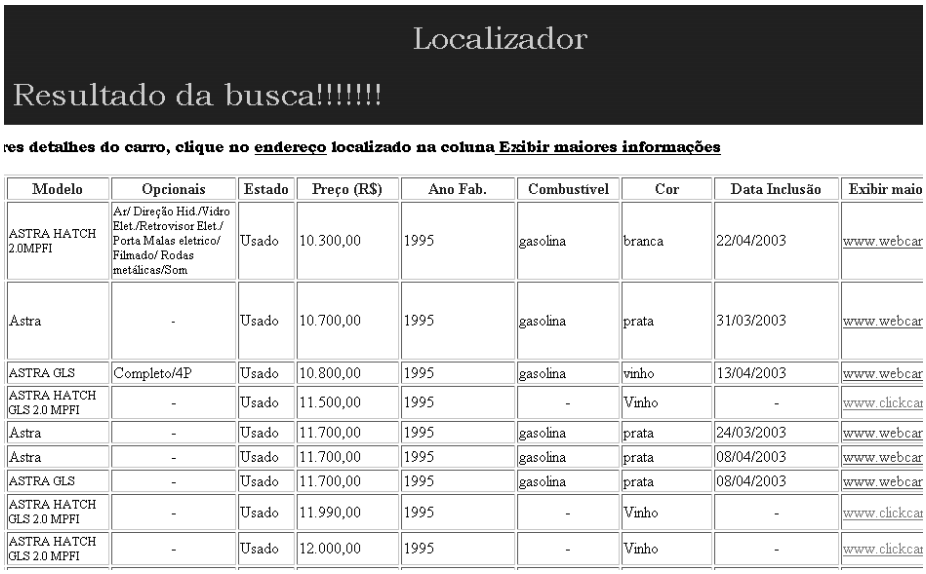


Figura 5.1 – Tela principal do Localizador

As Figuras 5.2 e 5.3 mostram uma simulação da tela de retorno do Localizador. Para nosso exemplo a consulta desejada foi de carro usado da marca Chevrolet, modelo Astra, cujo preço estivesse entre R$ 10.000,00 e R$ 20.000,00 e o ano de fabricação entre 1995 e 2000. A pesquisa foi realizada por ordem crescente de preço.

Na tela de retorno do Localizador, os resultados já aparecem diretamente na tela, não precisando que o usuário vis*ite* cada página retornada procurando o que deseja. No final de cada item retornado existe um *link* para o *site* original onde foi encontrado o item em questão.

Na interface de retorno existe um *link* para fazer uma nova consulta e outro para retornar à página inicial do Localizador.



Localizador

Resultado da busca!!!!!!!

res detalhes do carro, clique no endereço localizado na coluna Exibir maiores informações

Modelo	Opcionais	Estado	Preço (R$)	Ano Fab.	Combustível	Cor	Data Inclusão	Exibir maio
ASTRA HATCH 2.0MPFI	Ar/ Direção Hid./Vidro Elet./Retrovisor Elet./ Porta Malas eletrico/ Filmado/ Rodas metálicas/Som	Usado	10.300,00	1995	gasolina	branca	22/04/2003	www.webcar
Astra	-	Usado	10.700,00	1995	gasolina	prata	31/03/2003	www.webcar
ASTRA GLS	Completo/4P	Usado	10.800,00	1995	gasolina	vinho	13/04/2003	www.webcar
ASTRA HATCH GLS 2.0 MPFI	-	Usado	11.500,00	1995	-	Vinho	-	www.clickcar
Astra	-	Usado	11.700,00	1995	gasolina	prata	24/03/2003	www.webcar
Astra	-	Usado	11.700,00	1995	gasolina	prata	08/04/2003	www.webcar
ASTRA GLS	-	Usado	11.700,00	1995	gasolina	prata	08/04/2003	www.webcar
ASTRA HATCH GLS 2.0 MPFI	-	Usado	11.990,00	1995	-	Vinho	-	www.clickcar
ASTRA HATCH GLS 2.0 MPFI	-	Usado	12.000,00	1995	-	Vinho	-	www.clickcar

Figura 5.2- Parte da tela de resultados do Localizador

Localizador

Resultado da busca!!!!!!!

ASTRA HATCH GLS 2.0 MPFI	-	Usado	12.000,00	1995	-	Prata	-	www.clickcar
ASTRA HATCH SPORT 2.0 MPFI	-	Usado	12.300,00	1995	-	vermelho	-	www.clickcar
ASTRA HATCH GLS 2.0 MPFI	-	Usado	12.990,00	1999	-	Vinho	-	www.clickcar
ASTRA HATCH GL 1.8 MPFI	-	Usado	14.500,00	1999/2000	-	Azul	-	www.clickcar
ASTRA GL 1.8	dir hidr/ limpador traseiro	Usado	17.300,00	1995	gasolina	Vermelho	13/04/2003	www.webcar
ASTRA SEDAN GL 1.8 MPFI	-	Usado	17.500,00	1998/1999	-	Prata	-	www.clickcar
Astra 1.8 GLS	Ar/ Vidro Elétrico/ Trava automática	Usado	19.800,00	1998/1999	-	verde	17/03/2003	www.bolsaca
ASTRA HATCH	-	Usado	20.000,00	1998	gasolina	azul	31/03/2003	www.webcar

Fazer nova busca | Voltar à página inicial

Figura 5.3- Continuação da tela de resultados do Localizador

5.3 A avaliação das interfaces

A avaliação da tela de busca pelos usuários foi feita através de um questionário, elaborado com o propósito de avaliar o *site* em termos da usabilidade do sistema.

Para fazer a pesquisa, foi elaborado o questionário misto, com questões fechadas e abertas, onde o usuário pôde expor suas opiniões a respeito do *site* que foi avaliado.

O público escolhido para fazer a avaliação da usabilidade do Localizador foi formado por alunos e funcionários da UFLA com diferentes conhecimentos na área de informática, num total de 50 entrevistados.

O teste de avaliação das interfaces foi realizado no período de 15 a 30 de abril.

O questionário foi divido em duas partes (ver em Anexo A): a primeira com dados pessoais do entrevistado, e a segunda com a avaliação da usabilidade de acordo com os *sites* de busca existentes e o *site* construído especificamente para carros.

Capítulo 6- Resultados e Discussão

Para que a avaliação da usabilidade do *site* de busca se tornasse o mais real possível, foi colocada na Internet uma página onde existia um link para o Localizador, e o entrevistado poderia responder as questões sem a interferência de nenhuma pessoa que pudesse afetar nas suas respostas.

Dos 50 questionários distribuídos, somente 31 foram retornados.

Nas questões onde havia uma escala de valores de 1 a 10, eles foram agrupados de forma a facilitar nas análises dos resultados. O agrupamento foi feito da seguinte maneira:

1 2 3 4 5 6 7 8 9 10

Os valores 1 2 3 foram agrupados para o extremo inferior, 8 9 10 foram para o extremo superior e os valores 4 5 6 7 foram analisados como um meio termo entre os dois extremos.

6.1 Resultados Obtidos

A primeira parte do questionário de avaliação são informações pessoais referentes ao entrevistado, como idade, sexo, grau de escolaridade, freqüência com que utilizam *site*s de busca e os *site*s que mais utilizam. A finalidade dessa etapa do questionário foi avaliar o perfil dos entrevistados.

Faixa etária: 20 – 46 anos

Sexo: Masculino: 65.38%
Feminino: 34.62%

1) Escolaridade: Superior incompleto: 85.1%
Segundo grau: 11.1%
Superior completo: 3.8%
Primeiro grau: 0%
Pós-graduação: 0%

2) Freqüência que utilizam *site*s de busca: Freqüentemente: 88.8%
Raramente: 11.2%
Nunca utilizei: 0%

3) *Site*s de busca que costumam utilizar: Google: 96.2%
Cadê?: 3.8%
Não costumam utilizar: 0%
Yahoo: 0%
Alta Vista: 0%
Outro: 0%

A segunda parte do questionário foi sobre os *site*s de busca existentes na atualidade e sobre o Localizador (*site* de busca específico

para carros). As principais questões estão descritas no ANEXO A deste trabalho.

Com relação à facilidade de aprender, entender e utilizar, os *site*s de busca existentes na atualidade são: fáceis de usar para 77,7% dos entrevistados. Para 7,5% eles são difíceis de usar e 14,8% acham os *site*s de busca existentes um meio termo entre os dois. Este resultado pode ser mais bem visualizado na Figura 6.1.

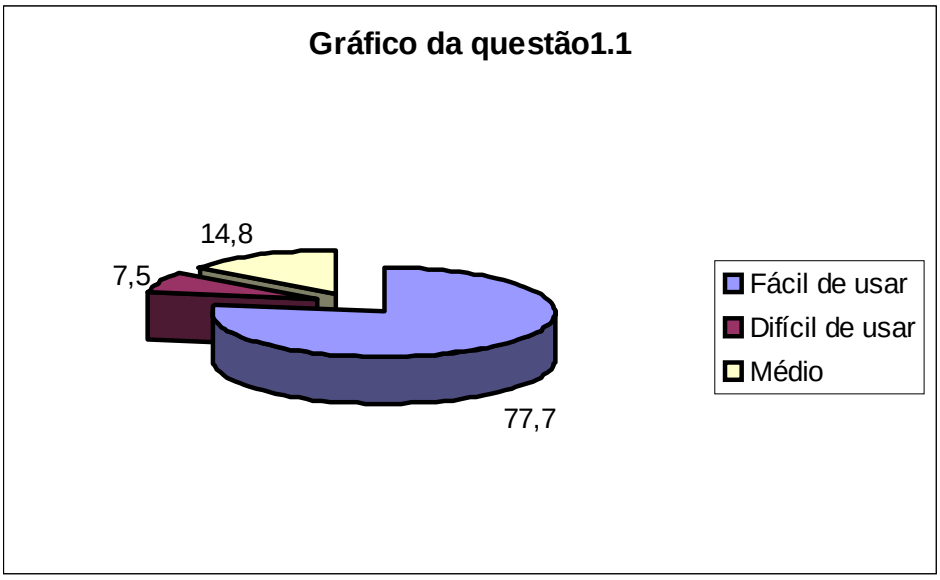


Figura 6.1- Gráfico da questão 1.1

Analisando o gráfico da questão 1.2 pode-se notar que 59,2% dos entrevistados acham que os *site*s de busca existentes são rápidos e eficientes no processo de busca. Para 3,8% eles são insatisfatórios e 37% acham que eles são medianamente satisfatórios com relação à rapidez e eficiência.

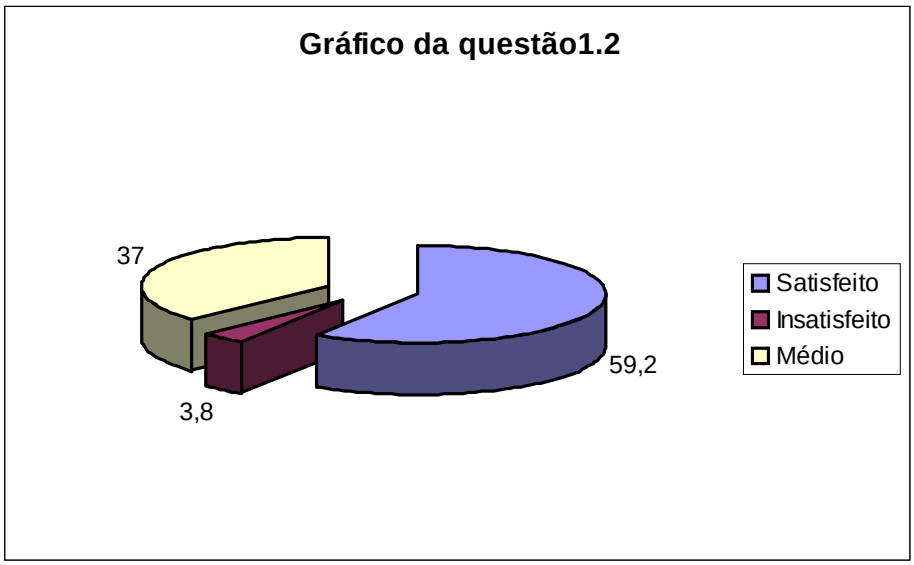


Figura 6.2- Gráfico da questão 1.2

A ajuda dos *site*s de busca são: fáceis de usar para 18,5% dos entrevistados, difíceis de usar para 3,7%, confusa para 14,8% e 63% deles nunca utilizaram a ajuda dos *site*s de busca que existem na *Web.*

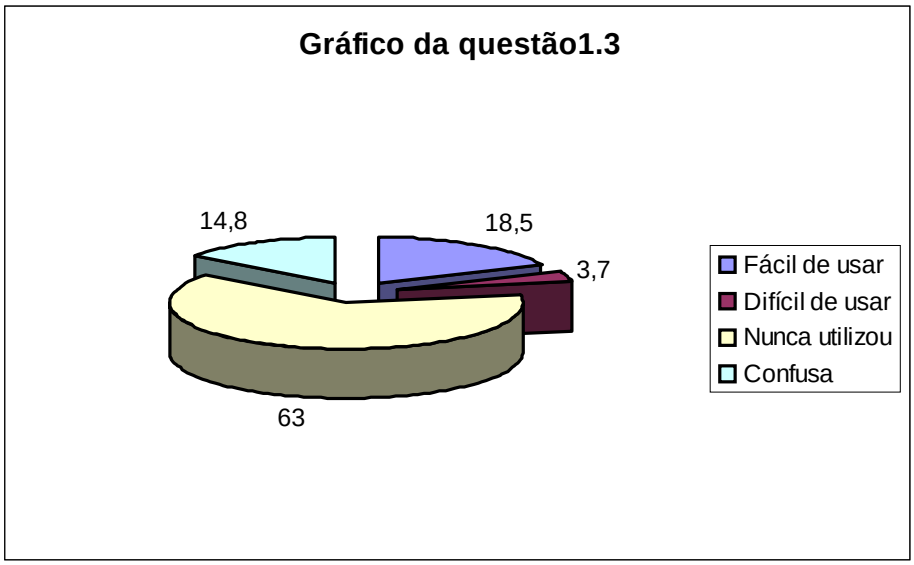


Figura 6.3- Gráfico da questão 1.3

O tempo médio gasto para efetuar a pesquisa nos três *site*s de exemplo foi de 7 min e 50 segundos.

Conforme a Figura 6.4, o grau de satisfação dos usuários ao utilizar os *site*s de busca existentes na atualidade é: ótimo para 7,4% dos entrevistados, bom para 51,9% deles e regular para 40,7% dos entrevistados.

Alguns dos entrevistados responderam que os *site*s de busca cobrem suas necessidades na maioria das vezes, encontrando resultados satisfatórios. Outros disseram que encontram o que procuram, mas não na primeira tentativa. E ainda, alguns responderam que são retornadas muitas informações irrelevantes, ficando o processo de verificação de cada resultado cansativo, além de nem sempre retornar o que desejam.

Pode-se notar que as respostas dos usuários estão de acordo com o estudo feito a respeito dos *site*s de busca existentes. Eles geralmente são difíceis de utilizar e entender, o modo de exibição dos resultados faz da busca um processo extremamente demorado devido ao grande número de informações que são retornados, nem sempre sendo o que o usuário deseja, além da inconveniência de ter que visitar página por página realizando a pesquisa manualmente.

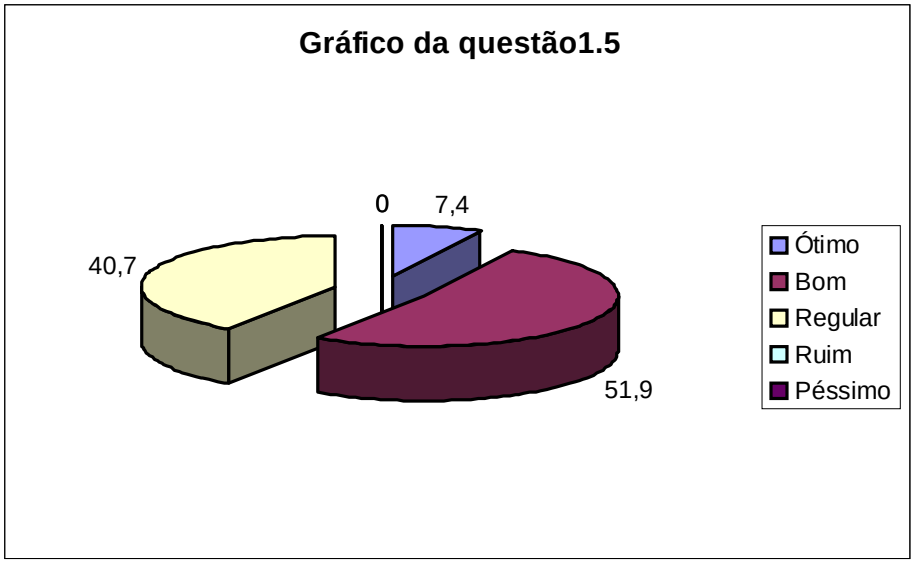


Figura 6.4- Gráfico da questão 1.5

As questões seguintes são referentes ao *site* de busca específico - o Localizador, que como já foi dito anteriormente é uma simulação de um *site* de busca que utiliza técnicas de extração de dados semi-estruturados no seu processo de busca. As principais questões estão descritas em ANEXO A deste trabalho.

Quanto à facilidade de aprender, entender e utilizar, o *site* de busca específico foi fácil de utilizar para 92,6% dos entrevistados. Para 3,7% deles o *site* foi difícil de utilizar e para outros 3,7% o *site* foi medianamente fácil de utilizar.

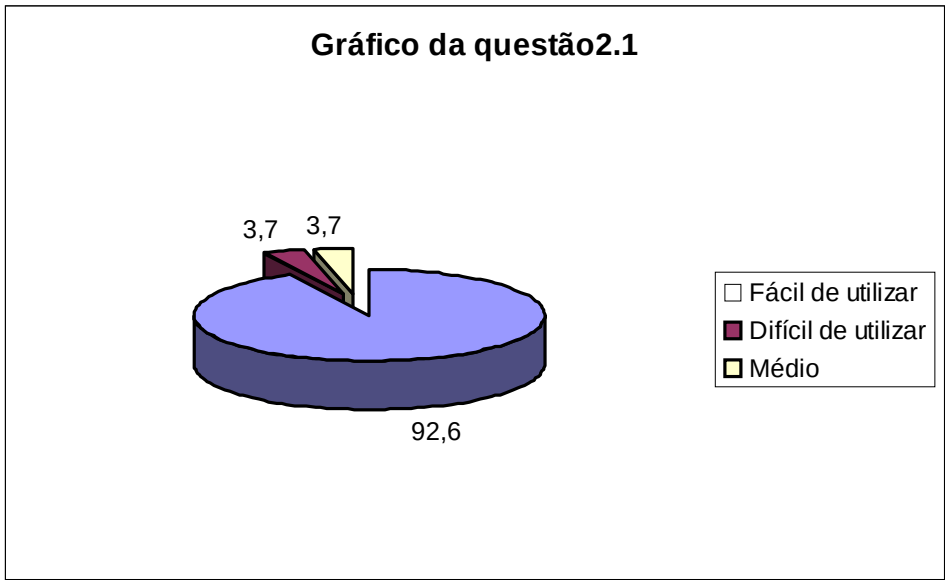


Figura 6.5- Gráfico da questão 2.1

Este resultado já era de se esperar, pois o *site* foi construído levando em consideração a usabilidade, ou seja, a facilidade de usar, entender e manipular, além de ser um *site* de busca onde a probabilidade de erro no preenchimento dos campos é bem reduzida, visto que possuem uma lista de valores para ajudar o usuário a realizar a busca desejada.

Com relação à quantidade de informações nas telas, as expressões utilizadas e a disposição das informações, o *site* específico foi: inadequado para 0% dos entrevistados, adequado para 92,6% deles e 7,4% achou um meio termo entre os extremos.

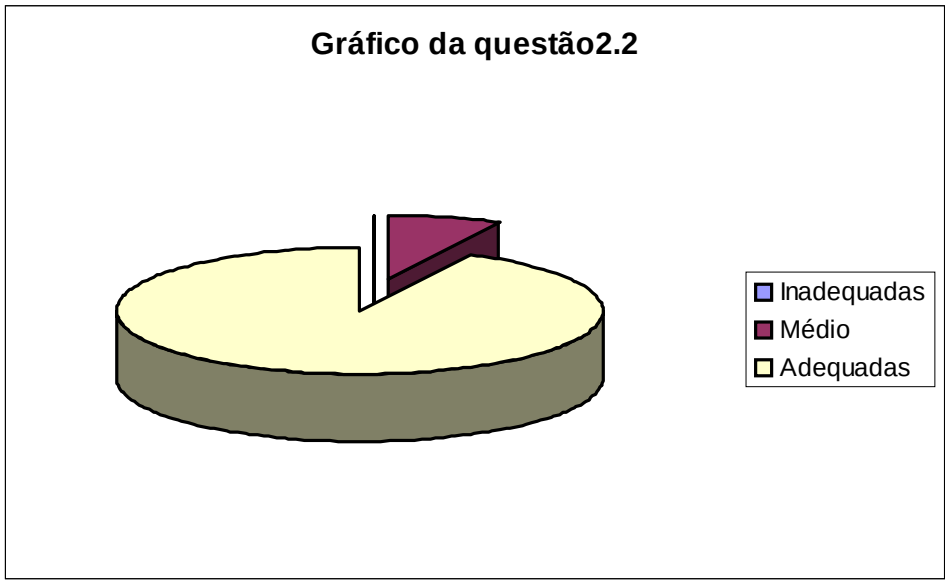


Figura 6.6- Gráfico da questão 2.2

Analisando a rapidez e eficiência no processo das buscas, 100% dos entrevistados acharam o Localizador satisfatório, conforme se pôde notar na Figura 6.7.

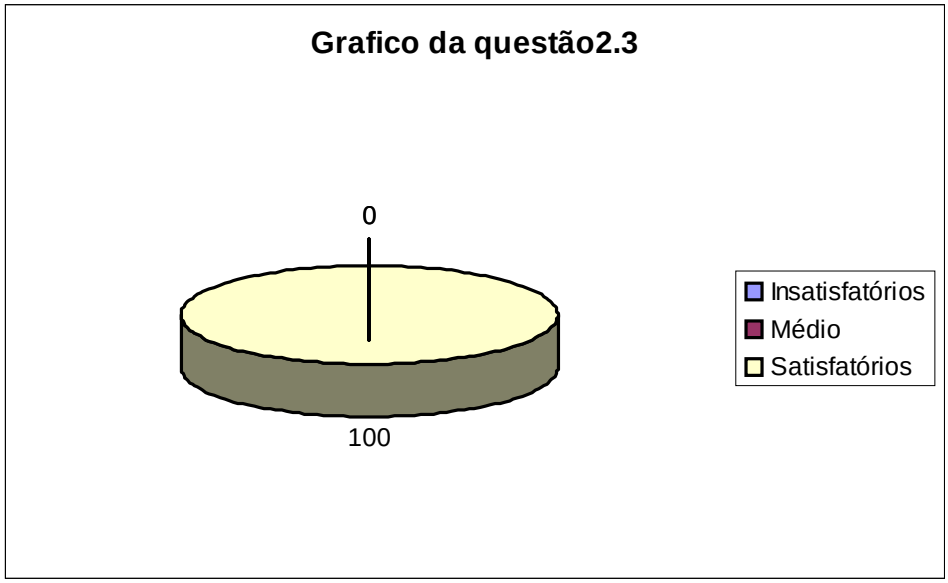


Figura 6.7- Gráfico da questão 2.3

Analisando o resultado dessa questão, onde 100% dos usuários responderam a favor do Localizador, pode-se perceber que *site*s de busca que utilizam técnicas de extração de dados semi-estruturados, apesar de serem bem específicos, são mais rápidos e eficientes que os que utilizam técnicas tradicionais no processo de busca.

O tempo médio que os usuários gastaram para efetuar a consulta pelo Localizador foi de 72 segundos, o que já era de se esperar, pois, neste método os resultados já aparecem diretamente, não precisando que o usuário fique procurando página por página o que deseja, como ocorre nos métodos tradicionais de busca.

A ajuda do *site* foi: clara para 92,6% dos entrevistados e 7,4% deles achou que ela poderia ser melhor.

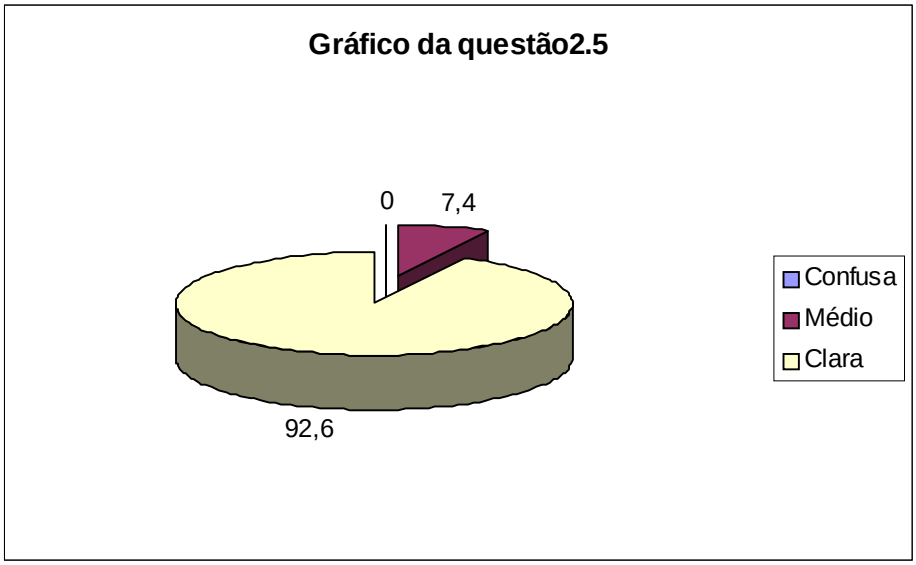


Figura 6.8- Gráfico da questão 2.5

A terceira questão faz um comparativo sobre os dois métodos de busca: o dos *site*s da atualidade e o específico.

Nesta questão 100% dos entrevistados acharam o *site* específico – o Localizador, mais fácil de realizar a busca desejada. Vários foram os motivos: por ser mais claro; rápido; fácil de entender e encontrar o que deseja, além de mostrar os resultados diretamente e somente informações relevantes, ou seja, a busca é mais precisa. Este resultado está de acordo com o que foi estudado em capítulos anteriores.

Com relação à rapidez e eficiência da busca, o método do *site* específico é para 92,6% dos entrevistados melhor que os *site*s de busca tradicionais e somente 7,4% deles não notaram diferença.

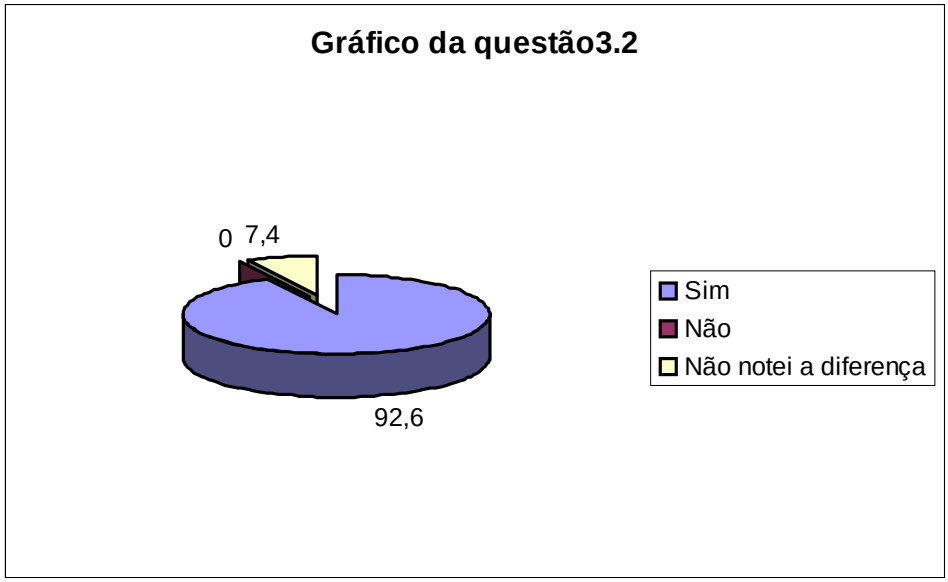


Figura 6.9- Gráfico da questão 3.2

Através dos estudos feitos em capítulos anteriores sobre as vantagens de utilizar técnicas de extração de dados semi-estruturados em processos de recuperação de dados da *Web*, pode-se comprovar através desta questão que realmente a rapidez e eficiência da busca aumenta, se comparada com os métodos tradicionais.

Para 92,6% dos entrevistados a forma de exibição dos resultados no *site* específico facilitou a procura do carro desejado. Os 7,4% restantes acharam que não facilitou a busca.

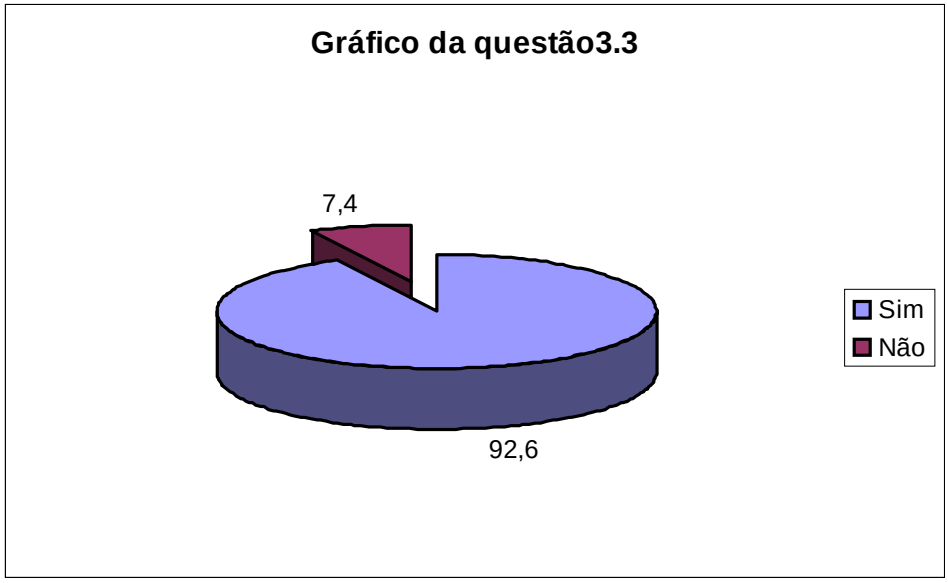


Figura 6.10- Gráfico da questão 3.3

A forma de exibição dos resultados na tela foi bem aceita pela maioria dos entrevistados, estando de acordo com o que foi estudado, pois ele mostra os resultados todos na mesma página, fazendo com que o processo de busca fique mais rápido e menos cansativo.

Esta questão pretendia avaliar a opinião dos usuários com relação ao *site* de busca específico - o Localizador.

Para muitos usuários entrevistados, o Localizador se destacou dos *site*s de busca convencionais por vários motivos, dentre eles: os campos para preenchimento são bastante intuitivos, ajudando o usuário

a selecionar o que deseja e evitando assim erros; os resultados são retornados com maior rapidez e eficiência; não precisam fazer a busca em outros *site*s para obter as informações, pois o resultado já aparece direto na tela de resultados, evitando assim que as buscas se tornem cansativas; as informações estão dispostas de forma clara; o número de informações irrelevantes é muito inferior.

Outros usuários disseram que a interface pode ser melhorada e que na tela principal onde são feitas as especificações da consulta a ser realizada faltou opções como cor do carro, número de portas, acessórios de carros no geral.

Capítulo 7- Conclusões

A cada dia que passa nos tornamos mais dependente da Internet, seja para desenvolver trabalhos, pesquisar dados ou para entretenimento pessoal. A quantidade de informações dos mais variados tipos é muito grande, e o conteúdo dessas informações nem sempre são confiáveis e fáceis de recuperar e manipular. Cabe a cada usuário saber o que usar e como usar para tirar o melhor proveito de tudo que ela vem oferecendo.

Com o estudo realizado neste trabalho podemos perceber as dificuldades de se recuperar e manipular dados da *Web* principalmente através das máquinas de busca existentes. Os resultados retornados por elas nem sempre são o que os usuários desejam, além da imensa quantidade de informações irrelevantes que são trazidas juntas. Isso torna o processo de busca e recuperação de dados extremamente cansativo e frustrante.

Além disso, elas são em geral difíceis de entender e utilizar para a maioria dos usuários da Internet, por que não são construídas levando em consideração a usabilidade, e atualmente não vale mais ter um produto que funcione adequadamente, realizando todas as tarefas e sem erros. Ele também deve ser fácil de usar e aprender.

Os estudos sobre recuperação de dados semi-estruturados e as ferramentas para extração de dados que vêm sendo desenvolvidas por vários pesquisadores visam resolver esses problemas, ou seja, construir máquinas de recuperação de informação que sejam rápidas e eficientes, onde os resultados sejam mais precisos.

Com a avaliação feita neste trabalho sobre a usabilidade de *sites* de busca que utilizam técnicas de extração de dados semi-estruturados, pode-se notar que a rapidez na hora de executar a tarefa aumenta, os resultados são mostrados diretamente na tela, não precisando que o usuário fique visitando página por página procurando o que deseja faz com que a busca não se torne um processo cansativo e exaustivo e em conseqüência, o tempo total gasto na pesquisa é bem menor que nos métodos tradicionais. No que se refere à usabilidade, pode-se perceber o aumento do grau de satisfação dos usuários ao utilizar uma máquina de busca que foi construída pensando em "facilitar a sua vida", com mais qualidade.

Este resultado foi devidamente comprovado através do teste de usabilidade feito com a simulação do *site* de busca para carros. Uma desvantagem dessas ferramentas no momento é que elas são bem específicas, mas como estes projetos e ferramentas ainda estão em fase de desenvolvimento e testes, pode ser que num futuro bem próximo existam máquinas de busca ainda melhores para recuperação de dados e nem tão específicas como estas.

7.1 Trabalhos Futuros

Pensando em trabalhos que possam ser realizados futuramente, seria interessante a construção de máquinas de busca geral utilizando as ferramentas de extração de dados descritas nos capítulos anteriores. Além disso, realizar outros testes de avaliação de interfaces, como por exemplo, os baseados na observação e monitoração de usuários e experimentos.

Capítulo 8- Referências Bibliográficas

[BAE 99] BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval.** New Iork: Ed. A. Wesley, 1999.

[CAR 98] CARDOSO, Olinda N. P. **Linguagens de Consulta para a *Web*.** Ano 1998.

[FLO 99] FLORESCU, D.; KOSSMANN, D. **A Performance Evaluation of Alternative Mapping Schemes for Storing XML Data in a Relational Database.** INRIA: [s.n.], 1999.
Disponível em: <http://rodin.inria.fr/Epubsbyyear.html>.
Consultado em outubro de 2002.

[GRU 93] GRUBER, T. R. **Towards principles for the design of ontologies used for knowledge sharing**. Stanford University, Knowledge Systems Laboratory Technical Report KSL93-04.
Disponível em:
<http://gicl.mcs.drexel.edu/people/regli/Classes/KBA/Readings/KSL-93-04.pdf>.
Consultado em Outubro de 2002.

[HEU 02] HEUSER, Carlos Alberto; MELLO, Ronaldo. **Aplicação de Ontologias a Dados Semi-Estruturados**.
Disponível em:
<http://www.inf.ufrgs.br/~ronaldo/a000139.pdf>.
Consultado em outubro de 2002.

[KAD 99] KADE, Adrovane M; HEUSER, Carlos Alberto. **Tendências em linguagens de consulta para documentos XML.**
Ano 1999
Disponível em:
http://www.fw.uri.br/~adrovane/widoc1999.pdf
Consultado em novembro de 2002.

[LAE 99] LAENDER, Alberto H. F; SILVA, Elaine E; ALTIGRAN S. da Silva. **DEByE - uma ferramenta para extração de dados semi-estruturados** Ano 1999.

[MAI 93] MYERS, A. B. **Why are Human-Computer Interfaces Difficult to Design and Implement?**, Carnegie Mellon University School of Computer Science Technical Report CMU-CS-93-183, Julho 1993.

[MEL 00] MELLO, Ronaldo; DORNELES, Carina; KADE, Adrovane; BRAGANHOLO, Vanessa; HEUSER, Carlos. **Dados Semi-Estruturados.** In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 15. 2000, João Pessoa, PB. **Anais**... João Pessoa: [s.n.], 2000.

[PIT 00] PITTS-MOULTIS, bN. e Kirk, C. **XML Black Book - Solução e Poder.** Makron Books, 2000. 627 p.

[SHN 98] SHNEIDERMAN, B. **Designing the User Interface: Straategies for Effective Human-Computer Interaction. Ed. A. Wesley** , 1998.628p.

[SIL] SILVEIRA, Fábio F. **Técnicas para Banco de Dados para World Wide *Web*.**
Disponível em:
http://www.alternet.com.br/users/ffs/download/Art_BD_ffs.pdf . Consultado em outubro de 2002.

[TEI 00] TEIXEIRA, Juliana Santiago. **Análise Comparativa de Diferentes Abordagens para Extração de Dados Semi-Estruturados.** Ano 2000.
Disponível em:
<http://www.dcc.ufmg.br/pos/html/spg2000/anais/juliana/juliana.htm> Consultado em outubro de 2002.

[ZAM 01] ZAMBALDE, A. L.; **Notas de aula: Interface homem-máquina.** Lavras/MG: DCC-UFLA 2001. 72 p.

ANEXOS - ANEXO A

Questionário para avaliação da Usabilidade de *site*s de busca que utilizam diferentes técnicas no processo de recuperação das informações

Para responder o questionário vi*site* a página
www.comp.ufla.br/~michellep/localizador.html

Primeira Parte

Idade: _____ Sexo: ( ) Feminino ( ) Masculino

1)Qual seu grau de escolaridade:
( ) 1º grau ( ) superior incompleto ( ) pós-graduação
( ) 2º grau ( ) superior completo

2) Com que freqüência você utiliza *site*s de busca?
( ) Nunca usei ( ) Raramente ( ) Freqüentemente

3) Quais *site*s de busca você costuma utilizar? (marcar somente uma alternativa)
( ) Não costumo usar este tipo de *site* ( ) Google
( ) Cadê? ( ) AltaVista
( ) Yahoo ( ) Outro. Qual? _____

Segunda Parte

Suponha que você queira fazer uma pesquisa detalhada sobre vendas de carros usados para saber qual o melhor preço dentre os *site*s de automóveis existentes na *Web* onde a marca procurada seja Chevrolet, modelo Astra, cujo preço esteja entre R$ 10.000,00 e R$ 20.000,00 e o ano de fabricação entre 1995 e 2000.

1) Faça essa consulta utilizando qualquer *site* de busca existente e responda às seguintes questões:

Obs.: Para facilitar o processo suponha que como resultado da busca foram retornados os seguintes *site*s: **www.bolsacar.com.br**, **www.clickcar.com.br**, **www.*Web*car.com.br**, além de outros de pouca relevância. Assim basta você visitar cada um destes *site*s separadamente fazendo a pesquisa, como é geralmente feito nos processos de busca convencionais.

1.1) Com relação à facilidade de aprender, entender e utilizar, os *site*s de busca existentes (google, cadê, etc) são:
Difíceis de usar 1 2 3 4 5 6 7 8 9 10 Fáceis de usar

1.2) Analisando a rapidez e eficiência no processo das buscas, eles são:
Insatisfatórios 1 2 3 4 5 6 7 8 9 10 Satisfatórios

1.3) A ajuda dos *site*s de busca que você costuma utilizar é:
( ) Fácil de usar ( ) Nunca utilizei
( ) Difícil de usar ( ) Confusa

1.4) Qual o tempo médio que você gastou para efetuar a pesquisa nos três *site*s exemplos?
_____
_____

1.5) Qual seu grau de satisfação ao utilizar os *site*s de busca que existem na atualidade:
( ) Ótimo ( ) Regular ( ) Péssimo
( ) Bom ( ) Ruim
Por que?
_____
_____
_____
_____
_____.

2) Agora, faça essa mesma consulta, mas utilizando o *site* que foi construído com técnicas de extração de dados no processo de busca, respondendo às seguintes questões:

Obs.: Este *site* é apenas uma simulação de como seria na realidade um *site* de busca que utiliza técnicas de extração de dados semi-estruturados na recuperação de informações.

2.1) Quanto à facilidade de aprender, entender e utilizar, o *site* de busca analisado é:
Difícil de utilizar 1 2 3 4 5 6 7 8 9 10 Fácil de utilizar

2.2) A quantidade de informações nas telas, as expressões utilizadas e a disposição das informações estão:
Inadequadas 1 2 3 4 5 6 7 8 9 10 Adequadas

2.3) Analisando a rapidez e eficiência no processo das buscas, ele é:
Insatisfatório 1 2 3 4 5 6 7 8 9 10 Satisfatório

2.4) Qual o tempo médio que você gastou para efetuar a pesquisa no *site* específico?

_____

2.5) A ajuda do *site* é:
Confusa 1 2 3 4 5 6 7 8 9 10 Clara

3) Fazendo um comparativo entre os dois métodos testados:

3.1) Em qual dos dois métodos você teve maior facilidade para realizar a busca desejada? Por que?

_____
_____
_____

3.2) Com relação à rapidez e eficiência da busca, o segundo método é melhor que o tradicional?
( ) Sim ( ) Não ( ) Não notei diferença

3.3) A forma de exibição dos resultados no *site* específico facilitou a procura do carro desejado?
( ) Sim ( ) Não

3.4)Qual a sua opinião em relação ao *site* de busca específico? Você o achou melhor que os *site*s tradicionais de busca em que aspectos?

_____
_____
_____
_____.