

**LIGA DE ENSINO DO RIO GRANDE DO NORTE
FACULDADE NATALENSE PARA O DESENVOLVIMENTO DO RIO GRANDE DO
NORTE**

CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

**RECUPERAÇÃO DE INFORMAÇÕES EM PÁGINAS
DINÂMICAS NA WEB: UM ESTUDO DE CASO NO PORTAL
UOL**

Monografia apresentada ao Curso de
Graduação em Bacharelado em Sistemas de
Informação da Faculdade Natalense para o
Desenvolvimento do Rio Grande do Norte
como requisito parcial para obtenção do grau
de Bacharel em Sistemas de Informação.

Viviane de Menezes Ramalho

Orientador:

Bertrão Paraguaná Fátima Cristina Rocha, Msc

NATAL /RN

2004

VIVIANE DE MENEZES RAMALHO

RECUPERAÇÃO DE INFORMAÇÕES EM PÁGINAS DINÂMICAS NA WEB: UM ESTUDO DE CASO NO PORTAL UOL

Monografia apresentada ao Curso de Graduação em Bacharelado em Sistemas de Informação da Faculdade Natalense para o Desenvolvimento do Rio Grande do Norte como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Orientadora:

Bartira Paraguaçu Falcão Dantas Rocha, Msc



VIVIANE DE MENEZES RAMALHO

DINÂMICAS NA WEB: UM ESTUDO DE CASO NO PORTAL UOL
RECUPERAÇÃO DE INFORMAÇÕES EM PÁGINAS

Catálogo na Publicação – Biblioteca da FARN
Setor de Processos Técnicos

Ramalho, Viviane de Menezes

Recuperação de informações em páginas dinâmicas na
web: um estudo de caso no portal UOL /Viviane de Menezes
Ramalho. __ Natal: FARN, 2004.
37p.

Orientadora: Bartira Paraguaçu F. Dantas Rocha

Monografia (Graduação em Sistemas de Informação) –
Faculdade Natalense Para o Desenvolvimento do Rio Grande
do Norte.

1, Internet – Recuperação de informações – monografia
2. Linguagem de Programação – HTML – monografia I.Título

RN / FARN / BC

CDU 004.738.5

RB-21341

FARN

BIBLIOTECA

Reg M 000073

12/04/05

M
004.738.5
R 1362
2004

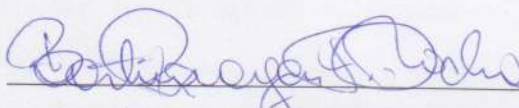
FARN

LIGA DE ENSINO DO RIO GRANDE DO NORTE
FACULDADE NATALENSE PARA O DESENVOLVIMENTO DO RIO GRANDE DO
NORTE
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

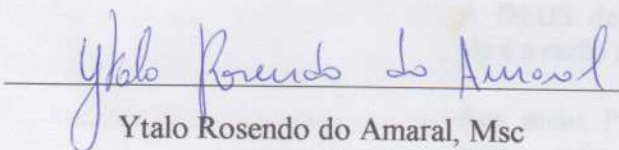
RECUPERAÇÃO DE INFORMAÇÕES EM PÁGINAS DINÂMICAS NA
WEB: UM ESTUDO DE CASO NO PORTAL UOL

Viviane de Menezes Ramalho

Monografia apresentada e aprovada em 11 de DEZEMBRO de 2004
Pela banca examinadora composta pelos seguintes membros:



Bartira Paraguaçu F. Dantas Rocha, Msc
Orientadora



Ytalo Rosendo do Amaral, Msc
Membro examinador



Antônio de Pádua de Miranda Henriques, Msc
Membro examinador

Natal, 11 de DEZEMBRO de 2004

AGRADECIMENTOS

Este trabalho foi construído com o apoio de muitas pessoas, e principalmente, com o desejo de que todo esforço, empenho e dedicação tenha obtido um resultado positivo. Sendo assim, venho sinceramente agradecer, esperando ter alcançado o meu objetivo com sucesso e atendido as expectativas de todos que contribuíram com este trabalho.

Assim, agradeço principalmente a DEUS, razão de minha vida, inspiração dos meus pensamentos, luz da minha sabedoria, origem das minhas vitórias, sem o qual eu não conseguiria.

Aos meus pais, riquezas inestimáveis, que sempre me dá muita vontade de aprender e vencer. Que se dispuseram a dar o melhor de si, para ajudar-me na construção da minha vitória.

Aos meus irmãos que com amor e incentivo, foram fundamentais em tudo que alcancei.

As minhas tias queridas Ester e Neutira, que apoiaram a minha vida desde o princípio, pelo incentivo, amor, ensinamento e atenção.

À Marcos, meu namorado e amigo, que se fez presente em todos os momentos dessa caminhada, me apoiando e ajudando em tudo que precisei.

As minhas amigas Raquel, Ana Helena, Nidia, Janyana, Kryssa, Patrícia e as amigas Josenilson, vocês foram importantes em cada momento, e com carinho, compreensão e amizade, me ajudaram a vencer.

Aos meus professores, em especial a Prof. Baitira, minha orientadora, professora e amiga, que incondicionalmente me destacou com sua atenção e ensinamento. A Livia, coordenadora, professora e amiga, alguém que me ensinou que a capacidade de vencer os desafios é uma lição de gratidão.

A Faculdade Natalense Para o Desenvolvimento, que me ofereceu para a realização e conclusão desta minha jornada, e a produção do conhecimento.

E a todos que, de alguma forma, contribuíram na elaboração deste trabalho.

A DEUS dedico este trabalho, pois ele é a razão da minha vida.

Aos meus Pais, como resultado de meu esforço e dedicação e principalmente pelo apoio, amor e carinho que deles recebi em todos os momentos da minha vida.

À minha Avó Juvina (*in memoriam*) por todo amor que ela dedicou a mim.

AGRADECIMENTOS

Este trabalho foi construído com o apoio de muitas pessoas, e principalmente, com o desejo de que todo esforço, empenho e determinação tenha obtido um resultado positivo. Sendo assim, venho sinceramente agradecer, esperando ter alcançado o meu objetivo com sucesso e atendido as expectativas de todos que contribuíram com esse trabalho.

Assim, agradeço primeiramente a DEUS, razão da minha vida, inspiração dos meus pensamentos, luz da minha sabedoria, origem das minhas verdades, dono do meu ser. Sem Deus eu não conseguiria.

Aos meus pais, riquezas imensuráveis, que acreditaram na minha vontade de aprender e vencer. Que se dispuseram a dar o melhor de si, para ajudar-me na construção da minha vitória.

Aos meus irmãos que com amor e incentivo, foram fundamentais em tudo que alcancei.

As minhas tias queridas Éster e Nautilia, que apoiaram a minha vida desde o princípio, pelo incentivo, amor, ensinamento e atenção.

Á Marcos, meu namorado e amigo, que se fez presente em todos os momentos dessa caminhada, me apoiando e ajudando em tudo que precisei.

As minhas amigas Raquel, Ana Helena, Niedja, Julyana, Kryssia, Patrícia e ao amigo Jossenilson, vocês foram importantes em cada momento, e com carinho, compreensão e amizade, me ajudaram a vencer.

Aos meus professores, em especial a Prof. Bartira, minha orientadora, professora e amiga, que incondicionalmente me distinguiu com sua atenção e ensinamento. A Lívia, coordenadora, professora e amiga, alguém que me ensinou muitas coisas, inclusive a entender que a capacidade de vencer os desafios é uma lição de grande valor.

A Faculdade Natalense Para o Desenvolvimento do Rio Grande do Norte, pelo suporte que me ofereceu para a realização e conclusão desta oportunidade produtiva e valorosa que é a produção do conhecimento.

E a todos que, de alguma forma, contribuíram na elaboração deste trabalho.

SUMÁRIO

1. INTRODUÇÃO.....	10
1.1. PROBLEMA DE PESQUISA.....	10
1.2. Objetivos.....	11
1.2.1. Objetivo Geral.....	11
1.2.2. Objetivos Específico.....	11
1.3. Justificativa.....	11
2. FUNDAMENTAÇÃO TEÓRICA.....	13
2.1. INTERNET.....	13
2.1.1. WWW.....	15
2.1.2. Páginas Dinâmicas	16
2.1.3. HTML.....	17
2.1.4. Recuperação de Informações em Páginas Dinâmicas	19
2.1.5. Mineração de Dados na Web.....	21
2.1.6. Indexação de Páginas HTML.....	21
3. METODOLOGIA.....	23
4. ANÁLISE DOS RESULTADOS	24
4.1. ESTUDO DE CASO.....	24
4.1.1. Código Fonte.....	26
4.1.2. AD1.....	27
4.1.2.1 Código Fonte Associado a AD1.....	28
4.1.3 Técnicas de Mineração de Dados na Página UOL.....	31
4.1.4. Desenvolvimento de um Sistema de RI.....	31
4.1.4.1. Sistema de Recuperação de Informações.....	32
5. CONCLUSÃO.....	33
REFERÊNCIAS BIBLIOGRÁFICAS.....	35
BIBLIOGRAFIA CONSULTADA.....	37

LISTA DE FIGURAS

FIGURA 1 PÁGINA PRINCIPAL DO UOL.....	25
FIGURA 2 ÁREA DINÂMICA 1 (AD1).....	27

Palavras-chave: WWW, Páginas Dinâmicas, HTML, Mineração de Dados e Recuperação de Informações.

RESUMO

A Web é hoje a maior fonte de informação eletrônica que dispomos. Entretanto, por causa da sua natureza dinâmica, recuperar determinadas informações se torna uma tarefa difícil. As páginas dinâmicas alteram seu conteúdo constantemente e, muitas vezes, as informações que estavam disponíveis há alguns minutos desaparecem. Analisaremos o Portal Universo on-line (UOL), caracterizado por seu dinamismo, e a partir daí desenvolveremos uma estratégia para definirmos uma forma de estruturação do conteúdo da página, a partir da estrutura html, e desenvolveremos uma ferramenta com o intuito de recuperar essas informações e disponibiliza-las através de uma estrutura de dados adequada.

Palavras chaves: *WWW, Páginas Dinâmicas, HTML, Mineração de dados e Recuperação de Informações.*

1. INTRODUÇÃO

ABSTRACT

The Web is today the biggest source of electronic information that we make use. However, because of its dynamic nature, to recoup definitive information if becomes a difficult task. The dynamic pages constantly modify its content and, many times, the information that were available have some minutes disappear. We will analyze the Vestibule Universe on-line (UOL), characterized for its dynamism and from there, we will develop a strategy we will define a form of structure of the content of the page, from structure HTML, and will develop a tool with intention to recoup these information and to get available them later, through a structure of data.

Words-keys: WWW, Dynamic Pages, HTML, Mining of data, Recovery of information.

A Web não foi planejada e projetada para manter informações de forma organizada e ordenada, não obedecendo estrutura lógica pedida de suas páginas.

A Internet não possui normas e regulamentação na disponibilização de arquivos e nem oferece indexação e recuperação "semântica" de suas páginas. Diante do volume de informações manipuladas pelo usuário, do dinamismo da Web, ou seja, conteúdos sendo alterados constantemente, e da necessidade de captar e encontrar adequadamente as informações, analisaremos técnicas para recuperação de informações que contribuam com o ganho de eficiência, relevância e resultado.

Esta pesquisa aborda a utilização de técnicas de várias áreas de pesquisa, tais como Técnicas de Recuperação de Informações, Modelos / Estruturas de dados para armazenamento de informações "descontínuas" (árvore, listas), dentre outras, para a recuperação de informações na Web.

1.1. PROBLEMA DE PESQUISA

Recupere Informações em páginas dinâmicas: Como as páginas dinâmicas (como a página do Portal UOL) alteram seu conteúdo constantemente e, muitas vezes, as informações que estavam disponíveis há alguns minutos desaparecem, como fazer para recuperá-las?

1. INTRODUÇÃO

A criação de ferramentas para a localização e recuperação de informações na web é uma preocupação que surgiu desde o princípio da Internet. A grande quantidade de informações disponíveis na web tem gerado uma falta de precisão nas buscas efetuadas. Diante disso, tem sido propostos vários mecanismos de extração, mineração e recuperação de informações mais eficazes a serem usados na web.

A situação tornar-se ainda mais complexa quando a informação encontra-se em páginas que possuem o seu conteúdo constantemente alterado. Será considerada uma página dinâmica aquela que tem o seu conteúdo frequentemente atualizado. Exemplos de páginas dinâmicas são os portais de informações que fornecem variados serviços.

A Web não foi planejada e projetada para conter informações de forma organizada e ordenada, não obedecendo estrutura lógica padrão de suas páginas.

A Internet não possui normas e regulamentos na disponibilização de arquivos e nem oferece indexação e recuperação “semântica” de suas páginas. Diante do volume de informações manipuladas pelo usuário, do dinamismo da Web, ou seja, conteúdos sendo alterados constantemente, e da necessidade de captar e encontrar adequadamente as informações, analisaremos técnicas para recuperação de informações que contribua com o ganho de eficiência, relevância e resultado.

Essa pesquisa aborda a utilização de técnicas de várias áreas de pesquisa, tais como Técnicas de Recuperação de Informações, Modelos / Estruturas de dados para armazenamento de informações “desconexas” (árvores, listas), dentre outros, para a recuperação de informações na Web.

1.1. PROBLEMA DE PESQUISA

Recuperar Informações em páginas dinâmicas. Como as páginas dinâmicas (neste caso a página do Portal UOL) alteram seu conteúdo constantemente e, muitas vezes, as informações que estavam disponíveis há alguns minutos desaparecem, como fazer para recuperá-las?

1.2. OBJETIVOS

1.2.1. Objetivo geral

Desenvolver uma estratégia de análise de páginas dinâmicas na web, com o intuito de recuperar informações e propor uma forma de extração dessas informações com base em técnicas de Mineração de Dados.

1.2.2. Objetivos Específicos

- Analisar e definir uma forma de estruturação do conteúdo da página dinâmica a partir da estrutura HTML da mesma;
- Desenvolver uma metodologia de varredura utilizando técnicas de Mineração de Dados e recuperação de informações no Portal UOL.

1.3. JUSTIFICATIVA

A grande explosão da Internet e sua rapidez de renovação das páginas, têm aumentando ainda mais a dificuldade de busca de informações específicas. Estudos realizados em 500.000 páginas, concluiu que 23% dessas páginas mudavam diariamente. Essa constante alteração nas páginas requer novas técnicas de recuperação de informações (FREITAS, 2002)

A popularização dos servidores www (Word Wide Web) é um dos maiores responsáveis por este crescimento, que ocorre de forma desordenada.

A geração de ferramentas para criação de agentes que minimizem a dificuldade das pesquisas na Internet é o ponto principal para resolução desse problema. Muitas pesquisas têm sido feitas com o objetivo de facilitar a busca na Internet.

Essa enorme quantidade de informações disponíveis na web tem gerado a necessidade do desenvolvimento de ferramentas para agilizar o processo de busca. Alguns mecanismos de busca têm surgido com o intuito de oferecer mais precisão e rapidez no que se refere as informações que esperamos encontrar em páginas na web. Para MACHADO (2004) alguns mecanismos tem sido utilizados com o objetivo de sanar esse problema de falta de precisão na busca da informação.

2- **RE** De acordo com FREITAS (2002) a utilização da Mineração dos Dados, é um processo capaz de explorar uma enorme base de dados em busca de informações importantes, onde tem como objetivo principal minimizar a busca e trazer benefícios quanto a tarefa de encontrar informações relevantes.

Para ALCANTARA, ETAL (1997, p.01) define o que é a Internet como um conceito, mesmo para os seus entusiastas. Eles pensam que saber o que significa, hoje, uma estrutura tão abrangente e inovadora que é preciso cuidado para não se deixar de fora muitas de suas várias facetas.

Internet é acesso de host uma de telecomunicação, possuem nos meios jurídicos, mídia, lazer, cultura, educação, marketing e uma série de outras conexões. A Internet é a maior rede de computadores do mundo, integrando informações localizadas em diferentes países, de continentes a continente. São milhões de computadores e dezenas de milhões de pessoas trocando informações em todo o mundo. Todas as estimativas atuais do número de usuários da Internet foram rapidamente superadas.

A Internet é simplesmente uma biblioteca, um acervo literário e a toda, seja bem ou ruim. A Internet é o telefone, a Internet é um círculo de amigos. A Internet é quadro de alunos a Internet é o rádio, a Internet é a televisão, a Internet são os livros, a Internet são os filmes, a Internet é a discoteca ou um bar, a Internet é um lugar de igreja. A Internet é qualquer coisa. Ela não é nada do que um amigo (LUNDHOLM apud ANDRADE, 2004).

A característica mais surpreendente dessa rede, chamada de interoperabilidade, é a capacidade de conectar computadores bastante diferentes entre si, feitos por vários fabricantes, como PCs, Macintoshes, cartões de trabalho, laptops, etc. Por isso ela é referenciada como uma rede de redes (ALCANTARA, 1997).

A Internet, essa rede mundial de computadores, tornou-se uma fonte de pesquisa indispensável ao dia a dia das pessoas, apresentando um grande acervo de dados e informações em todos os campos de conhecimento.

O desenvolvimento da tecnologia torna os dados cada vez mais disponíveis aos usuários, contudo a dificuldade de encontrar informações precisas é bastante perceptível.

Com o crescimento da Internet, um considerável aumento no volume de dados tem sido observado nos numerosos sites Web (MACHADO, 2004). Esse grande volume de informações e seu dinamismo têm dificultado a busca na Internet, gerando a necessidade de desenvolver técnicas de recuperação de informações, que ajudem na minimização, quanto a tarefa de encontrar informações precisas.

2 – REFERENCIAL TEÓRICO

2.1 – INTERNET

Para ALCÂNTARA, ETAL. (1997, p.01) definir o que é a Internet constitui um desafio, mesmo para os seus visionários. Isso porque essa super-rede apresenta, hoje, uma estrutura tão abrangente e inovadora que é preciso cuidado para não se deixar de fora muitas de suas varias facetas.

Internet é assunto de bar, tema de telenovela, polêmica nos meios jurídicos, mídia, lazer, cultura, educação, marketing e uma série de outros conceitos. A Internet é a maior rede de computadores do mundo, integrando informações localizadas em diferentes países, de continente a continente. São milhões de computadores e dezenas de milhões de pessoas trocando informações em todo o mundo. Todas as estimativas iniciais do número de usuários da Internet foram rapidamente superadas.

A Internet é simplesmente uma biblioteca, um acesso instantâneo a tudo, seja isso bom ou ruim. A Internet é o telefone, a internet é um círculo de amigos. A internet é quadro de avisos, a internet é o rádio, a internet é a televisão, a internet são os livros, a internet são os diários, a internet é a discussão em um bar, a internet é um bazar de igreja. A internet é qualquer coisa. Ela não é mais do que um meio. (LUNDBERG apud ANDRADE, 2004)

A característica mais surpreendente dessa rede, chamada de interoperabilidade, é a capacidade de conectar computadores bastante diferentes entre si, feitos por vários fabricantes, como PCs, Macintoshes, estações de trabalho, laptops, etc. Por isso ela é referenciada como uma rede de redes (ALCÂNTARA,1997).

A Internet, essa rede mundial de computadores, tornou-se uma fonte de pesquisa indispensável ao dia a dia das pessoas, representando um grande acervo de dados e informações em todos os campos do conhecimento.

O desenvolvimento da tecnologia torna os dados cada vez mais disponíveis aos usuários, contudo a dificuldade de encontrar informações precisas é bastante perceptível.

Com o crescimento da Internet, um considerável aumento no volume de dados tem sido observado nos numerosos sites Web (MACHADO, 2004). Esse grande volume de informações e seu dinamismo têm dificultado a busca na Internet, gerando a necessidade de desenvolver técnicas de recuperação de informações, que ajudem na minimização, quanto a tarefa de encontrar informações precisas.

O surgimento da Internet começou desde o ano de 1969, no Departamento de Defesa dos Estados Unidos, através do Advanced Research Projects Agency (ARPA – Agência de Projetos de Pesquisa Avançada). Um dos mais importantes avanços da Internet veio em 1986, Quando o National Science Foudation (NSF) criou a NSFNET para conectar diversos super computadores de alta velocidade ao longo do país (HONEYCUTT, 1998).

“A Internet é um vasto conjunto de redes diferentes que utilizam certos protocolos comuns e fornecem determinados serviços comuns” (TANENBAUM, 2003, p.53).

O grande poder da Internet está na capacidade de interligar qualquer tipo de computador em qualquer lugar do mundo – de um PC até um supercomputador.

A Internet, da forma que hoje se concebe, surgiu em meados da década de 80, apresentando, a partir de então, um crescimento surpreendente, não apenas em número de usuários, mas também em volume de informações disponível.

O crescimento da Internet tem sido explosivo. No ano de 1985 havia mais de 2.000 computadores hosts na Internet. Hoje são mais de 9 milhões de computadores hosts que conectam milhões de usuários. A cada dia a Internet adiciona novos usuários (HONEYCUTT,1998).

A Internet hoje é usada não só por cientistas, mas por várias outras pessoas. A popularidade da rede está fundamentada no princípio da *liberdade*. Não existe censores, chefes, “donos” ou normas no espaço da Internet.

A evolução dos serviços prestados na Internet também é um determinante da sua popularização. Inicialmente, o serviço principal era o correio eletrônico, baseado em recursos não-gráficos. O surgimento de interfaces agradáveis, on-line, multimídia e a filosofia hipertexto da WWW provocaram a adesão em massa de milhões de usuários, seduzidos pela interatividade da mídia emergente.

Para RANDALL (1997) a popularidade da Internet pode ser explicada ainda mais facilmente. Ela é enorme e global. Entre na Rede (como ela é chamada) e, de repente, você poderá trocar mensagens e arquivos com mais de um milhão de computadores, em dezenas de países. A cada 10 minutos, mais ou menos, outra rede de computadores se liga à Rede, o que se significa que esse milhão de computadores está ganhando novos companheiros muito mais rapidamente.

A Internet continua a evoluir sob diversas formas, principalmente no que diz respeito à criação de novos serviços e ao aprimoramento dos padrões utilizados atualmente na Web. A Web é hoje a maior fonte de informação eletrônica que dispomos. Entretanto, por causa da

sua natureza dinâmica, a tarefa de achar informações relevantes se torna muitas vezes uma experiência frustrante (MARINHO; GIRARDI, 2004).

2.1.1. WWW (World Wide Web)

WWW (World Wide Web, W3 ou Web) é o mais poderoso sistema de informação da Internet e o que mais tem se desenvolvido. Ele integra recursos multimídia e hipertexto em uma interface gráfica bastante interativa.

A Web é a primeira ferramenta da Internet. Ao contrário do restante da Internet, a www oferece sua própria atratividade. Quando você acessa uma página Web – elas se chamam páginas, freqüentemente pelo fato de se assemelharem às páginas de um livro ou revista. Uma página da www tira proveito do fato de que em nossa vida estamos sempre interagindo com uma página impressa (RANDALL, 1997).

No contexto de serviços de informação, a Web funciona como um superserviço, pois consegue prover informações que os outros sistemas de informação (FTP, gopher, etc) já proviam, com a vantagem de possuir uma interface agradável e simples de usar, o que lhe garante uma grande popularidade.

O poder de sedução da Web não vem do fato de ela ser um superserviço, mas dos recursos singulares que apresenta. Podemos citar algumas de suas aplicações:

- Divulgação de produtos para consumidores em potencial, compra e venda via web;
- Revistas eletrônicas, jornais, rádios podem ser acessados;
- Museus, galerias, agendas culturais, mapas;
- Informações com imagens, som, vídeo, texto.

A www consiste em um compacto conjunto de documentos estáticos e interativos que são vinculados entre si. Você utiliza um browser (navegador) para visualizar essas páginas Web. As páginas Web são encontradas em centenas de servidores Web espalhados por todo o mundo e se move para cada uma clicando nos vínculos que você encontra em todas as páginas (HONEYCUTT, 1998).

[...] as páginas da web tem evoluído muito rapidamente, desde quando começaram a idealizar a internet. Tendo essa evolução quatro fases diferentes: a internet estática, aplicações com CGI, utilização de linguagens de script e programas de interface com o servidor e a criação de páginas dinâmicas com acesso a bancos de dados através de ODBCs [...](OLIVEIRA, 2004).

Essa evolução pode ser melhor explicada por Oliveira (2004) através dessas 04 (quatro) fases, conforme apresentado abaixo:

1ª fase - a internet estática: onde não havia nenhum tipo de interatividade.

2ª fase – aplicações com CGI. através do CGI (Common Gateway Interface), foi criado um mecanismo que permitia ao browser se comunicar com o servidor, solicitando-lhe a execução de uma aplicação. Essa tecnologia foi o ponto de partida para a abertura do comércio eletrônico na web.

3ª fase - utilização de linguagens de script e programas de interface com o servidor: através da criação de scripts de cliente, passou a ser possível habilitar a máquina do cliente para realizar parte do processamento das tarefas incluídas nas páginas. Além disso, a Microsoft foi pioneira introduzindo um programa de interface com o servidor (ISAPI - Internet Server Application Programming Interface) que solucionou uma das principais limitações do CGI.

4ª fase - criação de páginas dinâmicas com acesso a bancos de dados através de ODBCs. O surgimento de linguagens de script com recursos para acessar informações em bancos de dados, assim como a definição do padrão XML (eXtensible Markup Language), um complemento à linguagem HTML (HyperText Markup Language), que permite a estruturação de dados na web, constituem-se no estágio atual das páginas da web. Através de drivers ODBC (Open DataBase Connectivity), instalados junto com o programa de interface com o servidor, unidos ao próprio programa servidor web e ao interpretador da linguagem script, é possível criar páginas dinâmicas com conteúdo atualizado diretamente a partir de um banco de dados, com ricos recursos e alta interatividade com o usuário (OLIVEIRA, 2004).

2.1.2. Páginas Dinâmicas

A expansão do ambiente de informação global causado pelo surgimento da Web permitiu a integração e ampliação dos serviços oferecidos pela Internet e outras redes. Um aspecto essencial nessa ampliação de serviços oferecidos foi o surgimento das chamadas "páginas dinâmicas" na Web. Essas páginas são criadas dinamicamente no servidor através de programas que têm como entrada dados recebidos do cliente (via preenchimento de formulários, por exemplo) e retornam ao cliente o resultado na forma de dados que ele é capaz de entender (uma página HTML ou outro tipo MIME) (BRENNER, 1996)

Páginas WWW dinâmicas são de fundamental importância para Intranets e para o comércio eletrônico da Internet, onde a natureza da informação muda constantemente e necessita de atualizações imediatas e automáticas, normalmente executada através do acesso a grandes bancos de dados. (MENDES, 2004)

Consideramos neste trabalho o dinamismo existente na exibição do conteúdo que certas páginas apresentam, independente de interação com o usuário. Tais páginas, ao serem acessadas, apresentam o seu conteúdo de uma forma e, depois de um certo intervalo de tempo, modifica a apresentação dos mesmos, podendo até mesmo, que o que ocorre na maioria das vezes, substituir integralmente o conteúdo, ou parte dele.

As informações continuam disponíveis no site, mas perdem o destaque original da página principal. Exemplos de páginas que dispõem desse dinamismo são os Portais de informação, onde as principais informações são modificadas em um curto espaço de tempo.

O Portal Universo On-line (UOL) é considerado uma página dinâmica por possuir tais características. Se observarmos a sua página principal, perceberemos que algumas áreas principais (Manchetes) são atualizadas com frequência, caracterizando o seu dinamismo.

2.1.3. HTML

Segundo (BARBOSA, Et Al. (1997,)) HTML (Hiptertext Markup Language) é a linguagem padrão adotada pela Web desde 1990 para criar e reconhecer documentos. Todos os documentos hipertexto da Web são escritos nessa linguagem. Esses documentos frequentemente são chamados páginas HTML, ou páginas Web, ou ainda home pages.

Como o próprio nome indica, é uma linguagem de “marcação”. Uma linguagem de marcação se caracteriza por dividir o texto em vários pedaços funcionais, que são identificados por elementos de marcação ou tags. Alguns desses pedaços podem ser, por exemplo, o nome do autor da página, uma âncora, um cabeçalho, um texto que deve ser enfatizado, e assim por diante.

A tarefa de um servidor Web é fornecer documentos HTML ao cliente. O cliente é um aplicativo responsável pela visualização desses documentos. A tarefa de reconhecer os elementos de marcação e formatar o texto de acordo com ele é deixada para esse programa conhecido como browser ou navegador. O navegador pode oferecer opções de configuração que permitam ao usuário controlar a forma como o documento será apresentado na tela. Portanto, quando um navegador vai apresentar um documento escrito em HTML, ele interpreta as marcações de cada parte do texto, aplicando a formatação respectiva.

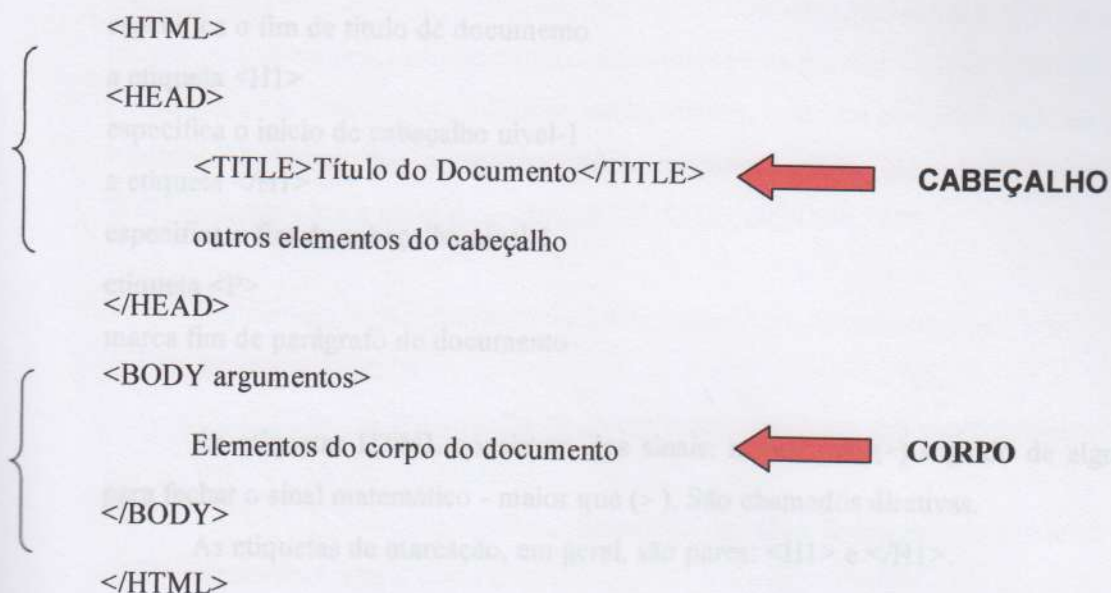
HTML é composta basicamente de elementos, que são instruções para os navegadores realizarem uma determinada tarefa como, por exemplo, tornar o texto *itálico*, exibir uma figura ou criar um link de hipertexto para outro documento. Os elementos de marcação de HTML, ou tags, consistem em palavras-chave delimitadas pelos símbolos "<" e ">".

Alguns fatores consagram HTML como a linguagem de autoria de hipertexto mais utilizada na Web. A flexibilidade, o alto nível de abstração e a simplicidade da linguagem são alguns dos motivos que explicam o seu sucesso.

A principal característica da linguagem HTML é que ela oferece uma maneira de descrever a informação independente da plataforma de visualização. Os elementos de HTML descrevem o que é a informação, não a forma específica como ela deve ser mostrada.

A maioria das páginas HTML divide-se em *cabeçalho*, *corpo* e *rodapé*.

ESTRUTURA DE UM DOCUMENTO HTML



A container tgs <HTML> delimita o documento. É importante que esse elemento esteja presente porque, futuramente, é possível que em um mesmo arquivo sejam transmitidos o documento HTML e mais alguma forma de dado, onde as tags <HTML> e </HTML> seriam utilizadas para definir a área do documento.

O documento é dividido em uma área de cabeçalho, delimitada pelos elementos <HEAD> e </HEAD>, onde estão informações sobre o documento e, no corpo, delimitado por <BODY> e </BODY>, que apresenta o conteúdo do documento.

Exemplo de um Documento HTML

<TITLE> Este é um exemplo de um simples HTML</TITLE>

<H1>Este e' um cabeçalho de nivel-1</H1>

Bem vindo ao mundo de HTML,

este e' um parágrafo. <P>

E este e' um outro parágrafo.

O HTML usa etiquetas de marcação para dizer ao browser Web, como deve mostrar o texto no video (display).

No exemplo acima:

a etiqueta <TITLE>

especifica o inicio de titulo de documento

a etiqueta </TITLE>

especifica o fim de titulo de documento

a etiqueta <H1>

especifica o inicio de cabeçalho nivel-1

a etiqueta </H1>

especifica o fim de cabeçalho nivel-1

etiqueta <P>

marca fim de parágrafo de documento

As etiquetas HTML consistem dos sinais: menor que (<) seguido de algum texto, e para fechar o sinal matemático - maior que (>). São chamados diretivas.

As etiquetas de marcação, em geral, são pares: <H1> e </H1>.

2.1.4. Recuperação de Informações em Páginas Dinâmicas

A tarefa de encontrar documentos pertinentes às necessidades dos usuários é o que pode ser definido como Recuperação de Informações (RI), tendo como o mais conhecido sistema de recuperação os mecanismos de busca utilizados na Word Wide Web (RUSSEL; RUSSEL, 2004).

A recuperação de informação, ou descobrimento de recursos, trata da automatização do processo de recuperação de documentos relevantes, que inclui, principalmente, representação, indexação e busca por documentos (MARINHO; GIRARDI, 2004).

A criação de ferramentas para a localização de informações na web é uma preocupação que surgiu desde os primórdios da Internet. A grande quantidade de informações disponíveis na web tem gerado uma falta de precisão nas buscas efetuadas. Diante disso, tem sido propostos vários mecanismos de extração e recuperação de informações.

A RI fornece técnicas para procurar informações relevantes em grandes coleções de dados, inclusive não estruturados.

A recuperação de informações na web é feita, principalmente, através de mecanismos de busca. Muito se tem pesquisado sobre a área conhecida como *tratamento de informações*, onde se utiliza processos de indexação, representações e classificação de informações.

A utilização de ferramentas na Internet para a recuperação de informações tem evoluído numa grande proporção. Duas ferramentas utilizadas como RI são os chamados “Mecanismos de Busca” e “Máquinas de Busca”. Os *mecanismos de busca* constroem seus índices através da descrição de páginas fornecidas pelas pessoas no momento de submetê-las. As pesquisas se baseiam nas informações fornecidas, e não no conteúdo real das páginas. Nas máquinas de busca, elas criam seus índices automaticamente, percorrendo continuamente a Web, visitando sites e indexando páginas (RUSSEL; RUSSEL, 2004).

Um sistema de Recuperação de Informações pode ser caracterizado por:

- 1- Uma coleção de documentos: cada sistema deve decidir o que deseja tratar como um documento;
- 2- Uma consulta apresentada em uma linguagem de consulta: a consulta especifica o que o usuário quer;
- 3- Um conjunto de resultados: subconjunto de documentos que o sistema de RI julga serem relevantes para a consulta;
- 4- Uma apresentação do conjunto de resultados.

Para implementarmos um sistema de RI, são necessárias e fundamentais duas estruturas de dados chamadas de léxico e índice invertido. O léxico testa todas as palavras na coleção de documentos e o índice invertido testa todos os lugares onde cada palavra aparece na coleção dos documentos (RUSSEL; RUSSEL, 2004).

O léxico é uma estrutura de dados que admite uma única operação: dada uma palavra, ele retorna a posição de índice invertido que armazena as ocorrências da palavra. Em algumas implementações, ele também retorna o número total de documentos que contém a palavra. Às vezes, um conjunto de palavras comum com pouco conteúdo de informações será omitido de léxico. Essas palavras ou stops words ("o", "de", "se", "ser", "um", etc) ocupam espaço no índice e não melhoram a pontuação dos resultados. A única razão de usá-las é para auxiliar sistemas que admitem consultas de frases.

O índice invertido consiste em um conjunto de listas de assentos: lugares em que cada palavra ocorre. Para o modelo booleano de palavra-chave, uma lista de documentos.

2.1.5. Mineração de Dados na Web

É evidente o enorme aumento de informações e recursos na Web nos últimos anos. Mais de um bilhão de páginas são indexadas pelos motores de busca (PAL, 2000 apud MARINHO, GIRARDI) e achar a informação desejada pode algumas vezes se tornar uma tarefa penosa. Essa abundância de informações e recursos instigou a necessidade do desenvolvimento de ferramentas automáticas de mineração e descoberta de informações na Web.

A mineração de dados refere-se ao processo não trivial de identificação de padrões válidos, previamente desconhecidos e potencialmente úteis dos dados (FRAWLEY, et al, 1992). Entretanto, utilizar e compreender os dados disponíveis na Web não é uma tarefa simples, pois esses dados são muito mais sofisticados e dinâmicos do que os sistemas de armazenamento de bancos de dados tradicionais. Enquanto estes últimos utilizam estruturas de armazenamento bem definidas e estruturadas, a Web não possui qualquer controle sobre a estrutura ou o tipo dos documentos que virtualmente armazena (MARINHO, GIRARDI, 2004).

2.1.6. Indexação de Páginas Html

Para CAMPOS, BAX (2000) as máquinas de busca são hoje um dos métodos mais utilizados para recuperação de informações na Internet. Várias pesquisas são feitas nesta área, procurando incrementar os processos automáticos de indexação e classificação de informações utilizados, a fim de melhorar a relevância e a velocidade da recuperação de informações na Web.

Se a página (documento HTML referenciado por uma URL¹) a ser indexada contiver suas informações organizadas de maneira que as máquinas possam acessá-las e compreendê-las, os processos de indexação e classificação são então potencializados, resultando em índices de melhor qualidade.

Quando se pesquisa em uma máquina de busca por páginas de um determinado assunto, utiliza-se palavras-chave relacionadas com este assunto. Palavras-chave são palavras que possuem a capacidade de descrever semanticamente o conteúdo de uma página, sob o ponto de vista dos usuários dos mecanismos de busca (CAMPOS, BAX (2000))

Os mecanismos de busca são bem menos eficientes que os humanos para analisar semanticamente uma página HTML e determinar qual é o assunto tratado, assim como sua relevância.

Ao se construir um documento HTML, deve-se procurar posicionar as palavras-chave estratégicas nestas posições cruciais, para garantir que os mecanismos conseguirão “entender” o significado da página.

Os mecanismos de busca consideram o cabeçalho da página (textos entre tags `<head>..</head>`, `<h1>..</h1>`, `<h2>..</h2>` ou `<strong>..</strong>`) e as primeiras linhas de texto, uma região importante para a classificação.

Pode-se pensar que a máquina de busca vê o texto da página como alguém que lê um artigo de um jornal. O primeiro parágrafo de um artigo sempre diz (ou deveria dizer) quais são os seus pontos principais. As palavras-chave estratégicas devem aparecer tanto no cabeçalho da página, quanto no primeiro parágrafo do texto.

URL – “Uniform Resource Locator”, modo de endereçamento formal de um recurso na Web

3 - METODOLOGIA

Este estudo classifica-se como Pesquisa Exploratória, pois proporciona uma compreensão maior do estudo. Segundo Oliveira (1997), este tipo de pesquisa requer teorias como ponto de partida e, objetiva pesquisar, comprovar ou rejeitar hipóteses e fazer sua aplicação às diferentes necessidades humanas.

O Método utilizado foi o Histórico e Dialético, através dos levantamentos das técnicas de recuperação de informações em páginas dinâmicas na web e a partir daí uma análise.

As fontes de pesquisa classificam-se como Pesquisas Bibliográficas, através de livros, artigos, ou seja, com base em material já elaborado.

Os dados deste estudo classificam-se como Pesquisa Qualitativa, pois será analisada uma série de leituras e documentação sobre o assunto, para estabelecer relações e, a partir daí, darmos nossa conclusão. Identificaremos causas, efeitos e conseqüências na utilização de Técnicas de Recuperação de Informações em Páginas Dinâmicas na Web.

Os procedimentos técnicos classificam este estudo como um Estudo de Caso, pois será feito um levantamento de dados no Portal UOL, além de uma análise na sua estrutura permitindo assim, o entendimento no seu comportamento.

4. ANÁLISE DOS RESULTADOS

4.1. ESTUDO DE CASO

O Portal Universo On-line (UOL) é um dos principais portais de conteúdo pago de acesso à Internet do país.

Lançado em abril de 1996, o UOL provê acesso em mais de 2000 localidades brasileiras e oferece também números locais de conexão em mais de 14 mil cidades no exterior. Tem hoje cerca de 1,3 milhão de assinantes pagantes. Desde setembro de 1999, atua também como portal e provedor de acesso na Argentina. O UOL reúne o mais extenso conteúdo em língua portuguesa do mundo. Está organizado em 42 estações temáticas, com mais de mil diferentes canais de notícias, informação, entretenimento e serviços, somando mais 7 milhões de páginas.

O UOL é formado por páginas estáticas e dinâmicas. As páginas dinâmicas são aquelas em que os seus conteúdos são alterados constantemente. Podemos observar esse comportamento na página principal, principalmente nas seções onde encontramos as manchetes diárias. A página principal do UOL é dividida em várias seções, sendo: Serviços, Publicidade, Manchetes, Conteúdos UOL.

As seções onde encontramos as manchetes principais é o que nos referimos de ÁREAS DINÂMICAS. Essas áreas dinâmicas foram divididas em diversas partes e definidas numa sequência, como a que se segue: AD1, AD2...ADn (Fig.1). Essa divisão foi feita para facilitar o estudo do comportamento dessas seções. Apesar dessas seções terem seu conteúdo modificado com frequência, pudemos observar um padrão na disposição desses conteúdos no código html da página principal. A partir daí, analisamos como os seus conteúdos estão dispostos e fizemos um estudo específico na AD1, analisando a estrutura do seu código fonte, observando-a e tomando como parâmetros as tags HTML.

Assim, a partir das informações adquiridas e de estudos sobre mecanismos e técnicas de recuperação de informações em páginas dinâmicas na web, é proposto o desenvolvimento de um software capaz de recuperar essas informações com base na análise desenvolvida.

LEGENDA



ÁREAS DINÂMICAS
PUBLICIDADE



SERVIÇOS
CONTEÚDO UOL

FIGURA 1: Página Principal do UOL

The image shows the UOL homepage layout. At the top, there's a navigation bar with links like ASSINE, BANDA LARGA, BATE-PAPO, CENTRAL DO ASSINANTE, DISCADOR, FOLHA ONLINE, and ÍNDICE. Below this is a banner for '12% de desconto IMPERDÍVEL' with the text 'em todo o site à vista no boleto'. The main content area is divided into several sections: a top news section with a headline about Colin Powell and 3 secretaries, a middle section with various articles and images, and a bottom section with more news and a sidebar. Red starburst labels (AD1-AD10) are placed over specific areas to indicate dynamic content or advertisements. The sidebar on the left contains a list of services and categories. The right sidebar features promotional offers from various companies like MercadoLivre.com, Notebooks Dell, and Magazine Luiza.

ASSINE **BANDA LARGA** **BATE-PAPO** **CENTRAL DO ASSINANTE** **DISCADOR** **FOLHA ONLINE** **ÍNDICE**

UOL **12% de desconto IMPERDÍVEL** em todo o site à vista no boleto

ASSINE UOL Conteúdo Exclusivo

BATE-PAPO **Por 1000** **IMM** **BUSCA** **UOL** **STÁDIO** **UOL.com.br** **SENHA** **UOL**

Amigos Virtuais
Bate-papo
Biblioteca
Bichos
Blog / Fotoblog
Buzza
Cantos
Cartões
Central do Assinante
Cinema
Corpo e Saúde
Crianças
Diversão e Arte
Economia
Educação / Vestibular
E-Mail
Empregos
Esporte
Folha Online
Fórum
Gay
Horóscopo
Humor
Jogos
Jornais

Ução de Casa
Mundo Digital
Música
Personalidades
Rádio UOL
Revistas
Serviços / Classificados
Sexo
Shopping
Sites Pessoais
Teen
Televisão
Tempo Trânsito Mapas
TV UOL
Últimas Notícias
UOL News
Viagem

ASSINANTE UOL
► Clube UOL
► Discador UOL
► Troque sua senha
► AntiSpam
► Acelerador de E-Mail

SERVIÇOS GRATUITOS
► E-mail grátis
► Multibúscas
► AntiPeppup

Após reeleição de Bush Segunda-feira, 16 de novembro de 2004
Colin Powell e mais 3 secretários pedem demissão

Folha Online
Estados Unidos sofrem novos ataques no Iraque

Protesto contra Lula
Estudantes jogam ovos e tomates em palanque

Rio Polícia prende chefe do tráfico

Fim de feriado Motorista sem carteira volta para SP

Apoena Meireles Acusado de matar senista é condenado

Flamengo Após briga com a torcida, time beira no interior

AD1

AD2

UOL Música
Carta de Kurt Cobain que descreve viagem de LSD vai ser leiloadada

FGV-SP
Veja correção completa do vestibular para direito

AD3

AD4

Black Jack
21

AD5

AD6

AD7

AD8

AD9

AD10

JORNAL E REVISTA
Veja Deputado diz, em gravação, que matou oito

ASTRAL
Tarô de Marselha Faça a sua pergunta e jogue as cartas

FOTOBLOG
Big Flash! por que nem tudo é preto no branco?

MUNDO DIGITAL
AntiXorcto Para ninguém saber por onde você vai

AntiFraude
Proteja-se de sites suspeitos e spam

AD10

20:31 Itaú

ASSINE JA O LIDER

Pontofilo.com
12% no Boleto
Out 12 vezes sem juros. Confira aqui!

MercadoLivre.com
PlayStation 2 a partir de R\$ 700. Clique e confira!

Notebooks Dell
Pague em 6 sem juros no cartão. Aproveite as ofertas

Magazine Luiza
Televisor Gradiente 14" Next em 12 de R\$ 35,75

Submarino
Todo o site em 12 vezes sem juros. Aproveite!

130.000 empregos
Inclua seu currículo por 7 dias sem pagar nada!

Americanas.com
Eletrrodomésticos com frete grátis em 12 vezes. Confira

Shoptime.com
Multifuncional HP de R\$ 600 por R\$ 400!

Submarino
Webcam Dlink em 6 de R\$ 18,17 sem juros

Pague suas dívidas
Solte até R\$ 5.000 e pague em até 36 vezes

Compra internacional
Aproveite as ofertas como se você estivesse nos EUA!

Notebooks Dell
Pague em 6 sem juros no cartão. Aproveite as ofertas

Polishop
AB Swing autêntico em 10 de R\$ 19,90 sem juros

MercadoLivre.com
CD players portáteis MP3 a partir de R\$ 129,99!

Claro Nokia 3100
Claro Conta a partir de 10 de R\$ 10,00

Sack's Perfumaria
CK Eternity Moment: lançamento mundial!

Camo Unime
Cadastre seu currículo: 130 mil empregos!

Submarino
CD player portátil Philips em 4 vezes de R\$ 48,75

Anna Pegova
Lançamento bioprotetor antidade Pagotan 80+

Americanas.com
Telefone sem fio 2,4 GHz com bina: 10 de R\$ 19,90

Sobre o UOL • About UOL • Assine UOL • Anúncio no UOL • Segurança e privacidade • Trabalhe no UOL • UOL Argentina

© 1996-2004 UOL - O melhor conteúdo. Todos os direitos reservados

LEGENDA



ÁREAS DINÂMICAS
PUBLICIDADE



SERVIÇOS
CONTEÚDO UOL

4.1.1. Código Fonte

A estrutura geral do código fonte da página é apresentado abaixo. Nela podemos observar o *cabeçalho*, o *corpo* e o *rodapé*. Dentro do *corpo*, na área em destaque, observamos parte da AD1, que é a área objeto do nosso estudo. A AD1 é a área onde aparecem as principais manchetes jornalísticas do Portal UOL.

A seguir apresentamos a AD1 mais detalhadamente.

```
<html> <head>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<title>UOL - O melhor conteúdo</title>
.
.
.
<script>
.
.
.
</script>
<style>
.
.
.
</style>
</head>
<body bgcolor=white>

<div id=c0>
<div id=cd>Segunda-feira, 15 de novembro de 2004</div>

<a
href=http://click.uol.com.br/?rf=hard_manchete&u=http://www1.folha.uol.com.br/folha/mun
do/ult94u78483.shtml><b>Após reeleição de Bush</b><h3>Colin Powell e mais 3
secretários pedem demissão</h3></a><!--SKa1/2-->
<span id=c1><a
.
.
<script>CkA("c2","hard_chamada",1)</script>
```

CABEÇALHO

CORPO

```

<script>CkA("des","hu_dest",0);CkA("rev","hu_rev",0);CkA("m3","hu_mm_texto",0);CkA("
tf1","hu_tfoto1",0);CkA("tf2","hu_tfoto2",0);CkA("tt1","hu_ttexto1",0);CkA("tt2","hu_ttexto
2",0);CkA("tt3","hu_ttexto3",0);</script>

```

```

</body>

```

```

</html>

```



4. 1. 2. AD1

A AD1, como foi dito anteriormente, possui as principais manchetes da página do UOL. Nela encontramos links para diversos conteúdos dispostos em locais diferentes dentro do próprio portal. É uma área extremamente dinâmica, tendo seu conteúdo alterado constantemente. Isso gera um problema de recuperação de informações, visto que, uma determinada informação que se encontra atualmente disponível em alguns instantes pode não mais estar. Com base no código html da página pudemos observar um padrão na disposição dessas manchetes. O código, assim como a análise de sua estrutura são apresentadas a seguir:

FIGURA 2 - AD1 (Área Dinâmica 1)



4.1.2.1 - Código Fonte Associado a AD1

```
<div id=c0>
```

```
<div id=cd>Segunda-feira, 15 de novembro de 2004</div>
```



Segunda-feira, 15 de novembro de 2004

```
<a
```

```
href=http://click.uol.com.br/?rf=hard_manchete&u=http://www1.folha.uol.com.br/folha/mundo/ult94u78483.shtml>
```

```
<b>
```

Após reeleição de Bush

Após reeleição de Bush



Colin Powell e mais 3 secretários pedem demissão

```
</b>
```

```
<h3>Colin Powell e mais 3 secretários pedem demissão</h3>
```

```
</a><!--SKa1/2-->
```

```
<span id=c1>
```

```
<a
```

```
href=http://click.uol.com.br/?rf=hard_fotoa&u=http://noticias.uol.com.br/ultnot/assuntododia/041115_album.htm?abre=5><img src=http://home.img.uol.com.br/0411/j/15arafat.jpg alt="Reuters" width=152 height=107> Palestinos visitam o túmulo de Iasser Arafat na Cisjordânia
```

```
</a>
```

```
</span><!--/SKa-->
```



Palestinos visitam o túmulo de Iasser Arafat na Cisjordânia

Folha Online

Estados Unidos sofrem novos ataques no Iraque

Folha Online

Estados Unidos sofrem
novos ataques no Iraque

Protesto contra Lula

Estudantes jogam ovos e tomates em palanque

Protesto contra Lula

Estudantes jogam ovos e
tomates em palanque

<a id=c2c1

href=http://www1.folha.uol.com.br/folha/cotidiano/ult95u101900.shtml>

Rio

 Polícia prende chefe do tráfico do morro da Providência

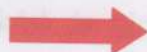
Rio Polícia prende chefe do tráfico do morro da Providência

```
<a
href=http://www1.folha.uol.com.br/folha/cotidiano/ult95u101919.shtml>
```

```
<b>Fim de feriado</b>
```

```
Motorista enfrenta lentidão na volta para SP
```

```
</a>
```



Fim de feriado Motorista enfrenta lentidão na volta para SP

```
<a
href=http://www1.folha.uol.com.br/folha/cotidiano/ult95u101918.shtml>
```

```
<b>Apoena Meireles</b>
```

```
Acusado de matar sertanista é condenado
```

```
</a>
```



Apoena Meireles Acusado de matar sertanista é condenado

```
<a
href=http://noticias.uol.com.br/pelenet/flamengo/ultimas/2004/11/15/ult1190u2757.jhtm>
```

```
<b>Flamengo</b>
```



Flamengo Após briga com a torcida, time treina no interior

```
Após briga com a torcida, time treina no interior
```

```
</a>
```

```
</span></span><script>CkA("c2","hard_chamada",1)</script>
```

Podemos observar, a partir do código acima, que existe uma padronização na estrutura html. Em qualquer dia ou hora que a página é acessada, o conteúdo das manchetes terá mudado, mas a sua estrutura continua a mesma. Isto é, as tags html que delimitam as

manchetes são sempre as mesmas, permitindo dessa forma, que seja definido um mecanismo da mineração e recuperação dessas informações.

4.1.3. Técnicas de Mineração de Dados na Página UOL

Segundo (FRAWLEY, et al, 1992) a Mineração de Dados acontece através de um processo não trivial de identificação de padrões válidos, previamente desconhecidos e potencialmente úteis dos dados.

Através da análise realizada na AD1 do UOL identificamos que esta padronização se faz presente na estrutura das páginas, na qual se utiliza as tags HTML. A identificação desta padronização é um dos recursos que utiliza-se para a criação de um mecanismo para a recuperação das informações.

Essa padronização pode ser observada através das seguintes informações:

- ✓ A quantidade das manchetes é sempre a mesma;
- ✓ Toda a AD1 é delimitada pelo <div id=c0>;
- ✓ A data é delimitada pelo <div id=cd>;
- ✓ A AD1 é composta por uma manchete principal que é delimitada pelo <H3>, três sub-manchetes delimitadas por , e respectivamente e quatro outras manchetes de menor destaque delimitadas pelo

4.1.4. Desenvolvimento de Um Sistema de RI

A estrutura de dados utilizada no processo de recuperação das informações estará fundamentada no que chamamos de *léxico*. O *léxico* será usado de forma modificada da sua definição original, pois não fará uso do *índice invertido*.

O *léxico* realiza testes nas tags html da página em busca de tags específicas que delimitam a área analisada (AD1). A partir do momento que essas tags são encontradas, consultas SQL irão inserir em um banco de dados as informações pertinentes da AD1 (data, hora de acesso, endereço, manchete). O papel do *índice invertido* não se tornará crítico nesta aplicação, porque já se sabe exatamente quais tags serão procuradas, assim como a sua disposição na página.

4.1.4.1 – Sistema de Recuperação de Informações

O *Sistema de Recuperação de Informações* proposto foi desenvolvido de acordo com as especificações abaixo:

- **Definição da Coleção de Dados:** no nosso estudo será a página principal do Portal UOL
- **Consulta:** realizada por um programa em JAVA usando SQL
- **Resultados:** o conteúdo recuperado é formado por, principalmente ,links e manchetes
- **Apresentação:** será feita na forma de um banco de dados contendo data, hora, links e manchetes

O único ponto não desenvolvido nesse Sistema foi a sua implementação, que não ocorreu devido a falta de tempo hábil para ser concluído. Mas, alguns pontos observados na tentativa de implementação pode ser úteis para uma futura implementação:

1. Apesar de Java possuir todo o perfil necessário para o desenvolvimento de uma aplicação desse natureza, ainda não existe na API do JDK 1.2.4 (Plataforma utilizada) nenhuma classe que faça uso de *tokens*.
2. A classe Java existente para o tratamento de *tokens*, que seria a forma de "quebrar" a página em palavras, considera como palavras as strings separadas por espaço em branco " " e tabulação.

Dessa forma se faz necessário o desenvolvimento de uma classe que trabalhe com *tokens* html, isto é, strings considerados como *tokens* aqueles delimitados por "<" e ">"

Não seria um trabalho tão simples devido a variedade de tags html existentes, pois além de trabalhar com estas tags delimitadas por "<" e ">" seria necessário ainda um tratamento das strings existentes entre o início e o fim de tag. A partir é que teremos de obter as informações desejadas, que no nosso caso seriam os links e suas respectivas manchetes.

5 – CONCLUSÃO

A Internet hoje tornou-se uma realidade presente em nosso cotidiano. Independente de qual área você trabalhe, ou qual o tipo de informação você procure, a Internet se apresenta como uma fonte preciosa de dados e serviços imprescindível para quem quer manter-se atualizado. O seu crescimento é notório, e a forma como as informações são dispostas não seguem nenhum padrão comum a todos. Os portais de informação são um exemplo clássico de como as informações podem ser fáceis de serem encontradas, mas difíceis de serem recuperadas a posteriormente.

A análise e o desenvolvimento de uma ferramenta que permita a recuperação de informações em sites com essa natureza dinâmica é bastante útil. Podemos observar, a partir das técnicas de Mineração de Dados, que, apesar do Portal UOL, que foi estudo de caso, ter seu conteúdo alterado dinamicamente, sua estrutura de informações era sempre a mesma. A partir da detecção desse padrão, é possível propor um Sistema de recuperação de Informações.

O único ponto não desenvolvido nesse Sistema foi a sua implementação, que não ocorreu devido a falta de tempo hábil para ser concluído. Mas, alguns pontos observados na tentativa de implementação pode ser úteis para uma futura implementação:

1. Apesar de Java possuir todo o perfil necessário para o desenvolvimento de uma aplicação dessa natureza, ainda não existe na API do JSDK 1.2.4 (Plataforma utilizada) nenhuma classe que faça uso de tokens html.
2. A classe Java existente para o tratamento de tokens, que seria a forma de “quebrar” a página em palavras, considera como palavras as strings separadas por espaço em branco “enter” e tabulação.

Dessa forma se faz necessário o desenvolvimento de uma classe que trabalhe com tokens html, isto é, seriam considerados como tokens aqueles delimitados por “<” e “>”.

Não seria um trabalho tão simples devido a variedade de tags html existentes, pois além de trabalhar com estas tags delimitadas por “<” e “>” seria necessário ainda um tratamento das strings existentes entre o início e o final da tag. A partir de agora é que teremos de fato as informações desejadas, que no nosso caso seriam os links e suas respectivas manchetes.

Propomos como trabalho futuro a implementação definitiva do Sistema de Recuperação de Informações, definido nesse trabalho, assim como um mecanismo de monitoramento que permita que seja feito um acompanhamento e recuperação de uma informação específica definida pelo usuário em tempo integral.

ANDRADE, C. A. V. Framework de um Modelo de Desenvolvimento para a Construção e Atualização de Portais Verticais de Conhecimento aplicados ao tempo de Engenharia de Software. Disponível em: <novodoc.net/eng/dissertacao.htm> Acesso em: 19 mai 2004.

BRENNER, J. H., AOKI, E. Introduction to CGI / Perl. M&T books, 1996. Disponível em: <<http://www.dca.fee.unicamp.br/courses/TA368F/1s1998/ProgWeb/g1.html#rs001>> Acesso em: 19 mai 2004.

CAMPOS, F., BAX, M. P. Como os Mecanismos de Busca da Web Indexam Páginas HTML, 2002. Disponível em: <<http://www.bax.com.br/bscarch/publications/Public/ComoMajornasBuscadedeasPaginasWeb.pdf>> Acesso em: 19 mai 2004.

FREITAS, F. L. G. Sistemas Multigentes Cognitivos para a Recuperação, Classificação e Extração de Informação da Web. UFSC, 2002. Disponível em: <<http://www.dca.ufsc.br/~fredl/fred6.pdf>> Acesso em: 19 mai 2004.

HONEYCUTT, J. Usando a Internet. Rio de Janeiro: Campus, 1998.

MACHADO, L. S. Recuperação de Sites Web: Uma Abordagem da Mineração de Dados Web, 2004. Disponível em: <www.inf.pucrs.br/~lmachado/PEP_Lufim.pdf> Acesso em: 23 mai 2004.

MARINHO, L. B., GIRARDI, Mineração na Web, 2004. Disponível em: <<http://www.sbc.org.br/relicodicoes/2004s2/artigos/MineracaodaWeb.pdf>> Acesso em: 27 mai 2004.

MENDES, M. A. 1998. Comparação e Análise de Requisições Dinâmicas em Servidores WWW. Disponível em: <<http://www.dca.ufmg.br/pos/comp/ing98/engs/cosello/accs/06/tese/>> Acesso em: 19 mai 2004.

OLIVEIRA, E., 2004. Evolução da Web e a estrutura cliente/servidor. Disponível em: <<http://www.dsp.com.br/revista.asp?acao=artigo14&versao=2&titulo=>> Acesso em: 19 mai 2004.

OLIVEIRA, S. L. de. Tratado de metodologia científica: projetos de pesquisas, TGI, TCC, monografias, dissertações e teses. São Paulo: Pioneira, 1997.

RANDALL, N. Aprenda em 1 Semana INTERNET. Rio de Janeiro, Campus, 1997.

RUSSELL, S. J., RUSSELL, P. N. Inteligência Artificial, Rio de Janeiro: Campus, 2004.

REFERÊNCIAS BIBLIOGRÁFICAS

- ALCÂNTARA, A. A., FILHO, C. S. F., BRASIL, C. C. A., ARANHA, D. C. S., BARBOSA, P. G. Home Pages Recursos e Técnicas para Criação de Páginas na WWW. Rio de Janeiro: Campus, 1997.
- ANDRADE, C. A. V., Framework de um Modelo de Desenvolvimento para a Construção e Atualização de Portais Verticais do Conhecimento aplicado ao tema de Engenharia de Software. Disponível em: <novainter.net/arq/dissertacao.doc> Acesso em 18.mai.2004.
- BRENNER, S. E., AOKI, E. Introduction to CGI / Perl. M&T books, 1996. Disponível em: <<http://www.dca.fee.unicamp.br/courses/IA368F/1s1998/ProgWeb/c5.html#ref001>> Acesso em: 19 mai.2004.
- CAMPOS, F., BAX, M. P. Como os Mecanismos de Busca da Web Indexam Páginas HTML, 2000. Disponível em: <<http://www.bax.com.br/research/publications/Publis/ComoMaquinasBuscaIndexamPaginasWeb.pdf>> Acesso em: 19 mai.2004.
- FREITAS, F. L. G. Sistemas Multiagentes Cognitivos para a Recuperação, Classificação e Extração de Informação da Web. UFSC. 2002. Disponível em: <<http://www.das.ufsc.br/~fred1/fred6.pdf>> Acesso em: 19 mai.2004.
- HONEYCUTT, J. Usando a Internet. Rio de Janeiro: Campus, 1998.
- MACHADO, L. S. Reestruturação de Sites Web: Uma Abordagem da Mineração de Dados Web, 2004. Disponível em: <www.inf.pucrs.br/~lmachado/PEP_Leticia.pdf> Acesso em: 28.mai.2004.
- MARINHO, L. B., GIRARDIR..Mineração na Web, 2004. Disponível em: <<http://www.sbc.org.br/reic/edicoes/2003e2/tutoriais/MineracaoNaWeb.pdf>> Acesso em: 27 mai. 2004.
- MENDES, M. A. 1998. Comparação e Análise de Requisições Dinâmicas em Servidores WWW. Disponível em: <<http://www.dcc.ufmg.br/pos/html/spg98/anais/corelio/corelio/html>> Acesso em: 19 mai.2004.
- OLIVEIRA, E., 2004. Evolução da Web e a estrutura cliente/servidor. Disponível em: <<http://www.fiap.com.br/secoes.asp?secao=artigo14&versao=2&titulo=>>> Acesso em: 19 mai.2004.
- OLIVEIRA, S. L. de. Tratado de metodologia científica: projetos de pesquisas, TGI, TCC, monografias, dissertações e teses. São Paulo: Pioneira, 1997.
- RANDALL, N. Aprenda em 1 Semana INTERNET. Rio de Janeiro. Campus, 1997.
- RUSSELL, S. J., RUSSELL P.N. Inteligência Artificial. Rio de janeiro: Campus, 2004.

BIBLIOGRAFIA CONSULTADA

TANENBAUM, A.S. Redes de Computadores. Rio de Janeiro: Campus, 2003.

BARROS, F. NUNES, C. Agentes para Extração de Informação, In: Universidade Federal do Pernambuco. Agentes de DI e IA. Disponível em:
<<http://www.di.ufpe.br/~campus/projeto-agentes.html>> Acesso em: 25 mar 2004.

COSTA, P. C. Uma Interface para Criação de Agentes de Consulta. In: SEMANA DA PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO 2 (SPG'98), 1998, Anais, MG, 1998. Disponível em:
<<http://www.dcc.ufmg.br/~pos/htm/spg98/anal/poporta/poporta.html>> Acesso em 27 mar 2004.

MOREIRA, J. L. Guia de Iniciação ao HTML. Disponível em:
<<http://www.sci.ufpb.br/~moreira/criilha.html>> . Acesso em 24 nov. 2004.

BIBLIOGRAFIA CONSULTADA

BARROS, F. NUNES, C. Agentes para Extração de Informação. In: Universidade Federal do Pernambuco. Agentes no DI e IA. Disponível em:
<<http://www.di.ufpe.br/~compint/projetos-agentes.html>> Acesso em: 25 mar.2004.

COSTA, P. C. Uma Interface para Criação de Agentes de Consulta. In: SEMANA DA PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO, 2 (SPG'98), 1998, Anais...MG, 1998. Disponível em:
<<http://www.dcc.ufmg.br/pos/html/spg98/anais/pccosta/pccosta.html>> Acesso em 27 mar. 2004

MOREIRA, J.A., Guia de Iniciação ao HTML. Disponível em:
<<http://www.mat.unb.br/~moreira/cartilha.html>> . Acesso em 28.nov.2004.