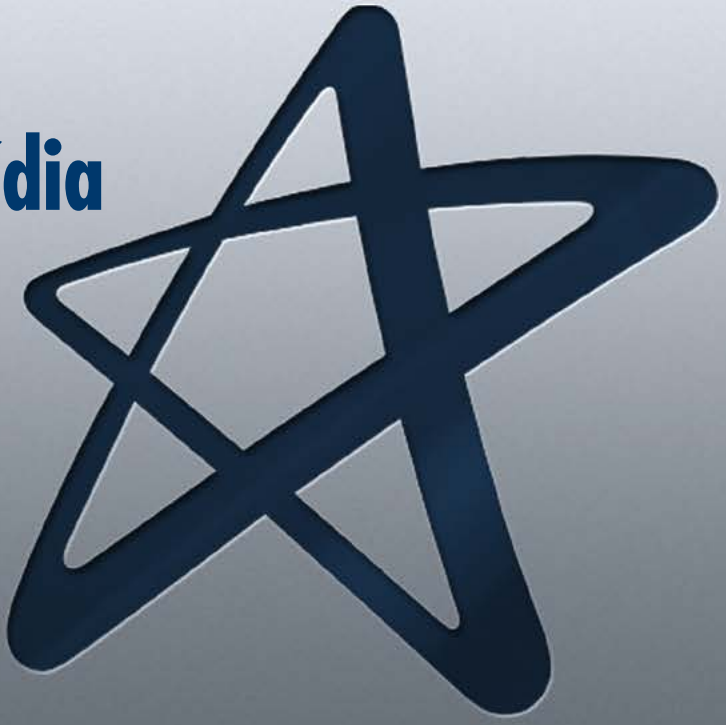
Sistemas de Hipermídia/Multimídia



Educação a Distância
Cruzeiro do Sul Educacional
Campus Virtual

Material Teórico

Spider e arquivo Robot

Responsável pelo Conteúdo:
Prof. Ms. Rolfi Cintas Luz Gomes

Revisão Textual:
Profª. Esp. Mácia Ota

UNIDADE Spider e arquivo Robot

- Redes Multimídia
- Busca e Recuperação da Informação

Objetivo de APRENDIZADO

Estudaremos Redes Multimídias, alguns conceitos de Rede de computadores como TCP/IP, UDP, http E FTP e protocolos multimídias como RTP, que são procotolos de apresentação de conteúdo em tempo-real, Arquitetura dos motores de busca e recuperação da informação no caos da Web.

Assim, iremos nos ater nos conceitos de:

- Redes Multimídias – formas de transmitir e receber pacotes multimidiáticos;
- Busca e Recuperação da Informação - arquitetura de recuperação de informações na Internet através dos motores de busca.

A organização deste material está agrupada por meio de diversos recursos, nos quais, você poderá ter uma maior compreensão do assunto abordado.

Desse modo, o assunto se tornará mais interessante, à medida que você tenha noções de arquitetura Web para sistemas hipermidiáticos.

Os materiais desta disciplina estão logo abaixo. Primeiramente, leia a contextualização da unidade, após, observe, no nosso esquema gráfico, a interligação dos tópicos que iremos estudar, seguindo do material de leitura.

Lembre-se de seguir o calendário e os prazos, pois a modalidade online exige organização e dedicação. Não se esqueça das participações nas Atividades de Aprofundamento e a interação nos Forúns com os colegas e o professor são de grande valia.

Contextualização

Que tal imaginarmos uma situação para dar início aos estudos desta unidade?

Veja:

Antigamente, na década de 90, havia milhares de tipos de documentos hipermidiáticos, nos quais para "descobri-los", era necessário um catálogo de endereços de websites que continham as listagens dos documentos a serem procurados.

Então, imagine um planeta que produz muitos arquivos e os postam na Internet, mas a única forma de encontrá-los seria pela "sorte" ou algum catálogo com imensas listagens de websites ou motores de busca, que não eram tão eficientes.

Assim sendo, esse aparente cenário de total desorganização e a necessidade de poder resgatar informações importantes em determinados locais e datas de forma rápida, foi "o ponta pé" inicial de Larry Page e Sergey Brin a fim de mudar a forma de buscas nos conteúdos da Internet, onde de 1996 até os dias de hoje, a Google passa de um projeto de doutorado para uma das empresas com maior visibilidade e sucesso no ramo da Computação e Internet no mundo. Pelo simples fato de buscar na "bagunça" de forma rápida e eficiente o que estamos procurando.

Como Page e Brin em 1995, você pode se perguntar: como isso é possível?

Esta é a Unidade III, na qual iremos apresentar alguns conceitos de rede, que são a base das **Redes Multimídias.** Dessa forma, iremos nos aprofundar nos conceitos principais dos motores de busca o **Web Crawling ou Spider!**

Redes Multimídia

Você sabia que o principal objetivo da rede é facilitar a comunicação?

Por isso, estão disponíveis vários componentes, protocolos e serviços que são responsáveis por essa comunicação. Assim sendo, devido à demanda x tempo x banda passante, temos vários tipos de compressão de dados que nos viabilizam a transferência de informações multimídia através da rede.

1.1 Introdução a Redes.

Ideias Chave

Redes é um conjunto de componentes interligados, comunicando-se entre si, geralmente, sem interferência humana. Para se comunicarem usam uma "linguagem" ou "forma de conversar" denominada protocolo.

O TCP/IP (Transmission Control Protocol/Internet Protocol) é o protocolo mais utilizado entre as redes em geral, devido sua difusão entre as redes, hardwares e serviços disponibilizados.

Para a comunicação entre os componentes da rede, é preciso abrir uma conexão, ou seja, uma "apresentação dos computadores" e do "que um deseja e do que o outro pode servir". Com isso, depois das conexões estabelecidas verificam a "forma de servir e receber".

As "formas de servir e receber" um arquivo são por **serviço não orientado à conexão (UDP)** e **serviço orientado à conexão (TCP).**

UDP(User Datagram Protocol): é um serviço não orientado à conexão, que tem por objetivo enviar o pacote e não preocupar foi efetivamente enviado ou não o mesmo. Usado para o envio de serviços/conteúdo com tolerância à perda, ou seja, não precisam dos dados de forma completa ou em tempo-real. Ex. teleconferências, programações de rádio e televisão disponíveis online,VoIP e etc.

TCP(Transmission Control Protocol): é um serviço orientado à conexão que tem como objetivo enviar pacotes, responsabilizando-se de seu envio através de técnicas e métodos para garantir sua integridade e confiabilidade do arquivo a ser enviado.Usado para o envio de serviços/conteúdo sem tolerância à perda, ou seja, obrigatoriamente, precisam dos dados de forma completa. Devido às suas características de verificações de como está sendo enviado o conteúdo/serviço, o TCP é mais lento que o UDP. Ex. e-mails, documentos, acesso remoto e outros.

Aplicação	Perda de dados	Largura de Banda	Sensível ao tempo
Transferência de arquivos.	Sem perda.	Elástica.	Não.
E-mail.	Sem perda.	Elástica.	Não.
Documentos Web.	Sem perda.	Elástica (alguns Kbps).	Não.
Áudio/Vídeo em tempo real.	Tolerante à perda.	Áudio: alguns Kbps – 1 Mbps. Vídeo: 10 Kbps – 5 Mbps	Sim, décimos de segundo.
Áudio/Vídeo armazenado.	Tolerante à perda.	Igual acima.	Sim, alguns segundos.
Jogos interativos.	Tolerante à perda.	Alguns Kbps – 10 Kbps.	Sim, décimos de segundo.
Aplicações financeiras.	Sem perda.	Elástica.	Sim e não.

Tabela 1- Principais serviços disponibilizados pelo TCP.

Aplicação	Protocolo de camada de aplicação.	Protocolo de transporte subjacente
Correio eletrônico.	SMTP (RFC 821).	TCP.
Acesso a terminal remoto.	Telnet (RFC 854).	TCP.
Web.	http (RFC 2616).	TCP.
Transferência de arquivos.	FTP (RFC 959).	TCP.
Servidor de arquivos remoto.	NFS (McKusik, 1996).	UDP ou TCP.
Recepção de multimídia.	Proprietário.	UDP ou TCP.
Telefonia por Internet.	Proprietário.	Tipicamente UDP.

Tabela 2 – Aplicações populares da Internet e seus protocolos.

Dentro dos conceitos multimídia, destacam-se as aplicações HTTP e FTP:

HTTP (Hipertext Transfer Procol): Implementado no protocolo TCP/IP, utiliza a arquitetura cliente-servidor, onde o programa browser (navegador) representa a parte cliente e o servidor Web executa na máquina-servidor. Geralmente, disponível nas portas 80/TCP e 443/TCP (HTTP Security).

As páginas webs, serviços de busca de dados são alocados na máquina- servidor e os browsers solicitam os mesmos através das URL (Endereço Web de conteúdo ou serviço).

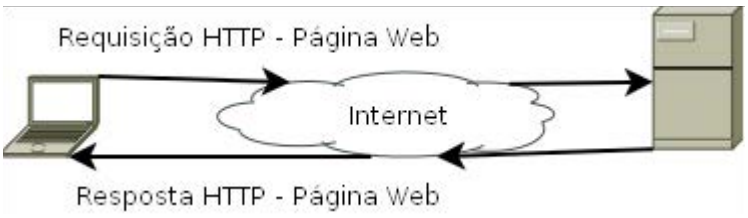


Figura: Requisição e Resposta HTTP

Através das solicitações HTTPs, podem ser repassadas vários tipos de informações multimídia, dentro de uma URL pode conter diversos hipertextos e hipermídias concentrados em um arquivo base, o HTML (Hyper Text Markup Language).

FTP (File Transfer Protocol): Implementado no protocolo TCP/IP, muito semelhante ao HTTP, salvo algumas modificações, entre elas, o controle de usuário, permissão e maior confiabilidade com relação à entrega de arquivos entre o cliente e servidor. Responsável por grandes tráfegos na rede. Geralmente, disponível na porta 21/TCP.

1.2 Qualidade de serviços (QoS) em Sistemas Multimídia

- QoS - Qualidade de Serviço (Quality of Service).
- Operação de rede para propiciar desempenho aceitável de serviço.
- Aplicações Multimídia possuem maior exigência de QoS.
- O QoS trabalha comumente com os parâmetros abaixo:

Atraso (Latência) fim-a-fim: Tempo de resposta de um ponto ao outro, considerando a soma dos atrasos dos equipamentos de rede. Seus principais fatores são: atraso de propagação, velocidade de transmissão e processamento nos equipamentos.

Jitter: Variação no atraso e na sequência de entrega dos pacotes(informações).

Vazão(Banda): Taxa de recurso mínimo aceitável para uma aplicação.

Ex. Vídeo - 100 Kbps a 1 Mbps.

Voz - 10 kbps a 120 Kbps.

Conferência - 500 kbps a 1 Mbps.

Taxa de perda de pacotes: É a taxa do limite de perda adequado da operação durante o processo de transporte.

Os maiores fatores que ocasionam perda de pacotes são os descartes de pacotes (pelos equipamentos de redes) e erros ocorridos no transporte (nas camadas ATM, frame relay, ethernet...).

Disponibilidade: Medida da garantia de execução da aplicação x tempo x equipamentos usados. Alguns fatores a serem considerados como ISPS (Provedoras de Internet) e Rede WAN, LAN e MAN.

1.3 Protocolos RTP/RTSP

RTP (Real Time Protocol) – Usado para garantir qualidade para conteúdos multimídia. Protocolo de tempo real fim-a-fim , transmite dados com prioridades de tempo.Projetado para ser flexível e não implementa QoS.

RTCP (Real Time Control Protocol) – trabalha junto com o RTP monitorando a qualidade do serviço.

RTSP (Real Time Streaming Protocol) – RFC 2326 , Protocol 554.

Protocolo de fluxo contínuo em tempo real, simplifica a comunicação entre 2 pontos, assim, usa menor tempo com verificações, aproxima-se, então, do tempo-real, permitindo controlar apresentações de mídia.

1.4 Streaming (fluxo de mídia)

Consiste na forma de disponibilizar o conteúdo de hipermídia através de pacotes em tempo real, sem arquivar os dados na máquina cliente, executando no Player o conteúdo na forma que é baixado, geralmente, utiliza-se do protocolo UDP e RTSP, tem suporte na arquitetura Multicast IP ou Broadcast, permitindo diversos tipos de serviços como WebTvs , Rádio Online e canais de TV Onlines.

- Disponível em tecnologias móveis.
- Formatos de compressão comuns: AVI, Ogg Vorbis, Quick Time, RMVB.
- Players comuns: Real Player , Windows Media Player.
- Para utilizá-lo nas páginas Web é preciso apenas salvar o conteúdo multimídia nos formatos comuns, o próprio plug-in encarrega-se de executar o streaming.

1.5 VoIP e Telefonia IP.

Telefonia IP é um conjunto de tecnologias que usam a intenet ou redes IP privados (ip fixo) para a comunicação de voz, com qualidade equivalente aos serviços tradicionais de telefonia em geral.

Tecnologia com melhor custo x benefício para usuários que usam com muita frequência chamadas internacionais e teleconferência, e uso contínuo como escritório de telemarketing.

O funcionamento é basicamente a transformação de voz em pacotes Ip´s de dados do emissor e convertido de dados para voz no receptor.

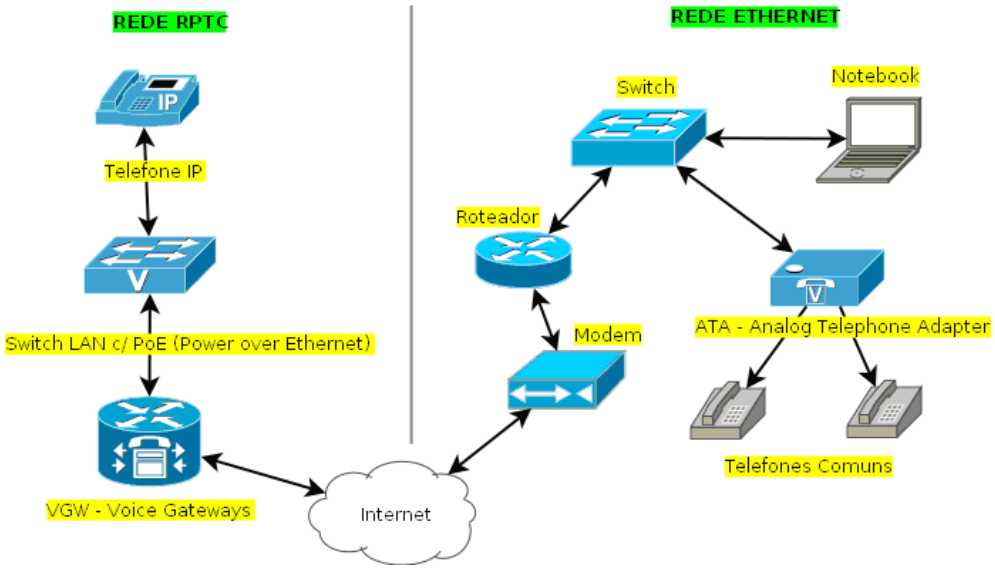


Figura: Diferença entre Rede RPTC e Ethernet *RPTC(Rede pública de telefonia comutada) ou PSTN(Public switched telephone network)

Explore

Para acessar o blog da certificação CCNA, acesse: blog.ccna.com.br, para saber mais detalhes sobre a rede IPT utilizada na Figura acima, acesse http://blog.ccna.com.br/2008/04/14/unified-communications-parte-ii/#hide

1.6 HDTV

HDTV (high-definition television) é um sistema televisivo de transmissão de resolução de tela, superior aos formatos (PAL,NTSC).

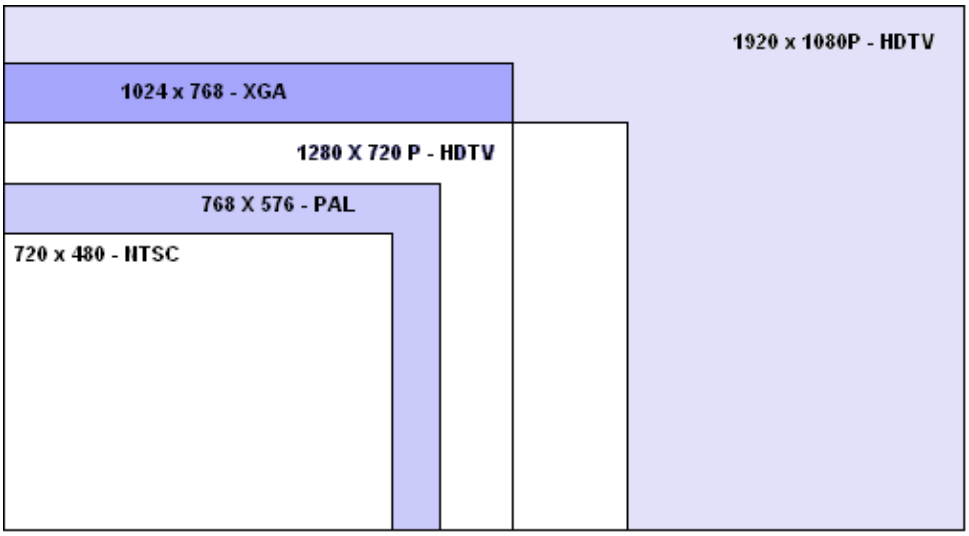


Figura: Resoluções de tecnologias para Televisão

- Os aparelhos tradicionais precisam de conversores para sintonizar a transmissão HDTV.
- Porém, devido às suas próprias limitações de resolução, 480 linhas não visualizam a transmissão em sua totalidade.
- O modo de varredura é a forma de mostrar a imagem na televisão, onde podemos ter:
  - "i" entrelaçado ou interlaced: primeiramente, são desenhadas as linhas ímpares e em seguida, as pares em todo o processo para um único quadro. Sendo 60 quadros por segundo, intercalando 30 quadros de linhas ímpares e 30 de linhas pares.

Impar par resultado na visão

Figura: Varredura entrelaçado (Adaptado de RIVED – MEC)

- “p” não-entrelaçado ou progressive : as linhas são desenhadas simultaneamente.

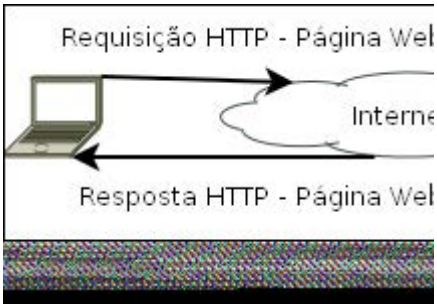


Figura: Varredura não entrelaçado

Nem todas as resoluções da HDTV são “p” ou “i”, segue abaixo o suporte das resoluções disponíveis no HDTV:

HDTV

Vantagens: Imagem com mais qualidade de nitidez, cores e brilho.
Sem Flicker (cintilação) - Espectros, rastros da imagem antiga.

Desvantagem: Necessidade de aparelhos com suporte a tecnologia.

NTSC/PAL-M - Tradicionais

Vantagens: Economicamente viável para transmissão; além disso, equipamentos antigos podem ser utilizados.

Desvantagem: Qualidade.

Busca e Recuperação da Informação

2.1 Arquitetura de ferramentas de busca

Após o surgimento da Internet e os Servidores Web, teve-se a preocupação dos pesquisadores da área de como localizar e organizar todas as informações disponíveis na Web de forma eficaz e rápida.

Para resolver esse problema, foram desenvolvidas as ferramentas de busca ou search engines, que são websites que disponibilizam uma lista de websites com relação à palavra-chave digitada. Entre as maiores empresas, encontram-se o Google, Yahoo, MSN, Baidu...

Grosso modo, as primeiras ferramentas de busca funcionavam deste modo:

A. São cadastrados ‘N’ Web sites no servidor.

B. Destes Websites são varridos seus links e cadastrados no servidor, essa é uma das funções dos softwares chamado spiders.

C. Os Websites são classificados através de seus títulos, cabeçalhos e meta-tag localizado na própria página HTML.

Com a comercialização da informação, surgiram diversos algoritmos e tecnologias proprietárias e sigilosas para as ferramentas de busca.

Porém, basicamente, as search engines estão fundamentadas sob três pilares:

1. Web crawling ou Spieders ou Web Robot - captura dos dados;
2. Indexação - organização e classificação dos dados;
3. Busca - resposta dos termos informado pelo usuário.

Existem vários conceitos de buscadores, tais como:

Globais: Pesquisa global e abrangente de palavras-chave. Ex: Google e Yahoo.

Verticais: Pesquisa especializada na própria base de dados. Ex: Buscapé e Catho.

Guias Locais: Pesquisa local ou regional de algum assunto, palavra-chave ou serviço.Ex: Listão, Guia Mais e MEC.

Guias de Busca Local: São buscas baseadas na pesquisa de Guias Locais com abrangência Nacional, que baseia sua consulta tanto no assunto, quanto no CEP. Ex: Google Maps.

Diretório de WebSites: São sites organizados por categoria e subcategoria em forma de árvore ou grafo.

2.2 Web crawling ou Spiders

São softwares que percorrem as âncoras de um documento para posterior processamento. Também chamado de web robot, ants, automatic indexers, bots e worms. Usados para :

- criar cópia parcial das informações relevantes de todas as páginas visitadas;
- verificar os links e conteúdo da página;
- obter informações específicas de páginas ou busca de e-mails.

O processo web crawling ou spidering inicia-se pela semente, recursivamente, sob diversas políticas de busca e direitos autorais.

- **Sementes:** uma lista de URLs a serem percorridas.
  - Normalização da URLs – Padronização da URL para não gravar certos caracteres, e certificar-se de que uma URL será gravada apenas uma vez.
  - Identificação de Web Crawlers – Através da tag “User-agent” na requisição HTTP HEAD do crawling, busca o arquivo robots.txt do servidor para saber as páginas do servidor e as meta tags referente a esta página, sua requisição é apenas para páginas em HTML.

- **Fronteira de percorrimento:** lista de URLs dentro das páginas sementes.
  - Políticas de busca:
    - seleção: seleciona as páginas que deve capturar através de algoritmos e métricas;
    - re-visitação: verifica as mudanças em uma página para revisitá-la;
    - cortesia: para Websites muito extensos, organiza o spyder para não causar sobrecarga;
    - paralelização: organiza e coordena spyders recursivos.

Algoritmos e métricas spyders:

a) breadth-first: considera que as páginas importantes são as que possuem diversos links, e vasculham pelos hosts procurando páginas importantes.

b) Pagerank: Algoritmo criado por um dos fundadores do Google, determina como páginas 'confiáveis', as que contém diversos links e várias páginas, 'confiáveis' apontando para ela.

c) OPIC(On-line Page Importance Computation): Semelhante ao Pagerank, porém veloz.

2.2.1 Diversos tipos de Crawling.

Crawling de caminho ascendente - São crawler que "listam" todo o conteúdo da URL, independente, de página Web ou não, forcendo também arquivos como imagens, documentos de textos entre outros.

Crawling focado - busca páginas com conteúdo similar através das páginas que a mesma está ancorada.

Crawling na Web Profunda ou Web invisível - são todas as páginas isoladas que não possuem links ou âncoras para nenhuma outra página, ou através de várias resultantes de busca de banco de dados.

Algumas empresas estão pesquisando formas de encontrar as Web invisíveis, porém um dos passos seria uma listagem disponibilizada pelo servidor robots.txt ou pelo próprio web master.

Existem diversos tipos de crawlers desenvolvidos, sendo executados, constantemente, na Web, porém seguindo essa mesma técnica, temos os crawlers, que ao invés de apenas indexar e disponibilizar o Website para os usuários, utilizam esse programa para fazerem Spans em blogs, fórum e etc , e para evitar este tipo de inconveniente , na tag abaixo, é possível deixar a página "invisível" para o spyder, pode-se colocar a tag :

```
<meta name="robots" content="NOINDEX, NOFOLLOW">
```

Uma dica de segurança em relação à Crawler, seria renomear documentos importantes que contenham algo secreto no nome para qualquer outro nome e a página de acesso restrito para outro nome, impedindo assim Crawlers mal intencionados de detectar sua página.

2.3 Indexação ou Web Indexing

Responsável pela coleta, análise, recuperação e armazenamento da informação fornecida pelo Spyder.

A forma de armazenar e analisar considera diversos fatores da linguística e semântica de cada área e palavra-chave, para tanto, utiliza-se de vários algoritmos de compressão e estrutura de dados entrelaçado a uma série de técnicas da lingüística e computação para a indexação dos dados.

2.4 Busca

A busca é a resultado das páginas maior relevância, encontrado no banco de dados da Indexação de acordo com a palavra-chave do usuário. Geralmente, a busca encontra suporte as ligações NÃO, E,OU e " " junto com as palavras-chave.

2.4.1 Ranking de documentos

Geralmente, o ranking de documentos se dá sobre vários critérios na busca, discutidos no Web Crawling, porém também temos algumas relevâncias comerciais do ranking de documentos, ou seja, sobre algumas palavras-chave é determinado um valor por clique e o responsável pelo Website paga pelo acesso do usuário; tornando assim concorridas as primeiras colocações nos sites de busca, ou seja, as primeiras posições são as mais caras.

Dentro desse conceito, temos várias empresas que faturam milhões por anos e possuem vários acessos como a Google Inc. e o Website brasileiro Buscapé que foi pioneiro no Brasil na disponibilização do serviço de indexação de produtos por palavras-chave e listagem por preço/marca, entre outros atributos.

2.5 Semantic Web

Semantic Web (Web semântica [significado na Web]) é uma tendência da futura Web 3.0 que não leva apenas em conta a Interação com o conteúdo da Web 2.0, mas também sua relação de conteúdo com significado das palavras.

Será um processo da passagem de dados para informações, ou seja, a própria Web relacionará conteúdos semelhantes e com relações através dos significados que elas trazem. E, agrupamento das "ilhas de dados" organizados por agentes através do significado dos dados, embora seja uma tendência ainda não difusa já fora definido que se deve utilizar o padrão W3C.

Material Complementar

Caro aluno, seguem alguns materiais que serão de grande valia para quem quiser se aprofundar nos assuntos abordados.

- Transmissão Multimídia em redes de computadores:
http://www.rnp.br/newsgen/0007/art2.html

- Multimedia Networks and Communication:
http://multimedia.ece.uic.edu/04-8.pdf

- A Review of Multimedia Networking – Introduction:
http://www.agocg.ac.uk/reports/mmedia/network/intro.htm

- Sistemas de busca e recuperação de informação:
http://www2.dbd.puc-rio.br/pergamum/tesesabertas/0313143_06_cap_05.pdf

- Estratégia de busca na recuperação da informação: revisão da literatura:
http://www.scielo.br/pdf/ci/v31n2/12909.pdf:

- Saiba mais sobre protocolos de Streaming : RTP, RTCP, RTSP:
http://wiki.icmc.usp.br/images/4/43/RTP-RTSP-slides.pdf
http://www.cse.wustl.edu/~jain/cis788-97/ftp/ip_multimedia.pdf

- Para se aprofundar nos conceitos de DTV, HTDV segue o link abaixo:
http://eletronicos.hsw.uol.com.br/hdtv.htm

- High Definition Television (HDTV) : Difference Between High & Standard Definition Video:
http://www.youtube.com/watch?v=DULT4L8c8IM

- O Básico Sobre TVs de Alta Definição (HDTV):
http://www.clubedohardware.com.br/artigos/1253

- Tudo sobre HDTV:
http://www.clubedohardware.com.br/printpage/Tudo-sobre-HDTV/1011

Referências

GLOOR, P. A. **Elements Of Hypermedia Design**: Techniques For Navigation And Visualization In Cy. Boston: Birkhauser - Publishers For Ar, 1997.

HALSALL, F. **Multimedia Communications**: Applications, Networks, Protocols And Standards. 5. ed. England: Addison Wesley Longman, 2001.

NIELSON, J.; LORANGER, H. **Usabilidade na Web.** Rio de Janeiro: Campus, 2007.

PAULA FILHO, W. P. **Multimídia:** Conceitos e Aplicações. 2. ed. Rio de Janeiro: LTC, 2011.

SAYOOD, K. **Introduction to Data Compression**. 3. ed. New York: Morgan Kaufmann Publishers, 2006.

BAEZA-YATES, R.; RIBEIRO NETO, B. **Modern Information Retrieval**. Harlow: Pearson Addison Wesley, 1999.

LOWE, D.; HALL, W. **Hypermedia & The Web**: An Engineering Approach. New York: John Wiley & Sons, 1999.

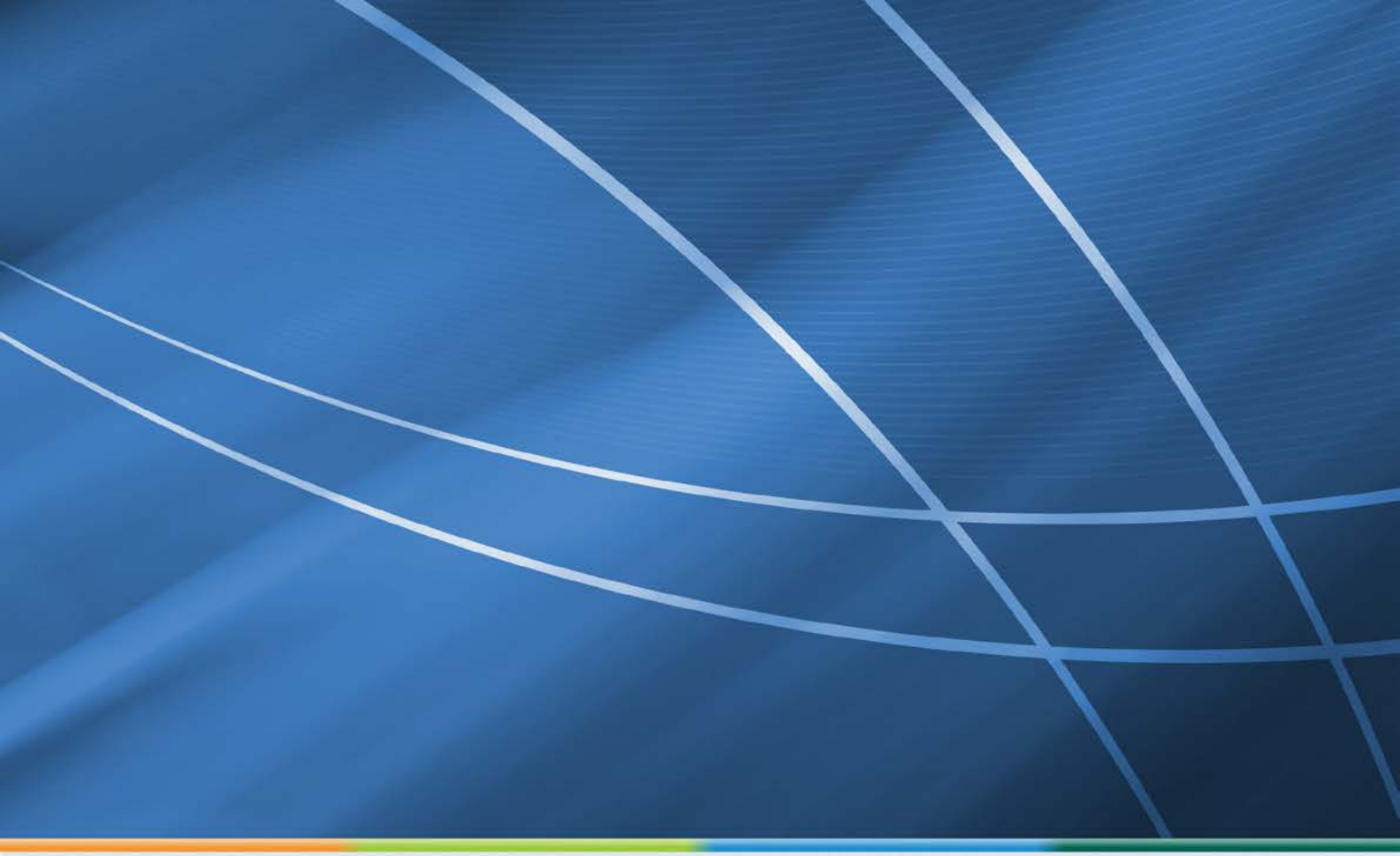
NIELSEN, J. **Projetando Websites.** Rio de Janeiro: Campus, 2000.

SAUCIER, C. **Animação e Interatividade na Web**. São Paulo: Market Books Brasil, 2000.

Anotações

Educação a Distância
Cruzeiro do Sul Educacional
Campus Virtual

www.cruzeirodosulvirtual.com.br
Campus Liberdade
Rua Galvão Bueno, 868
CEP 01506-000
São Paulo SP Brasil
Tel: (55 11) 3385-3000



Universidade Cruzeiro do Sul

Módulo Centro Universitário

UDF Centro Universitário

UNICID Universidade Cidade de S. Paulo