A Internet e a Informação

Rubens Queiroz de Almeida
queiroz@unicamp.br

Introdução

A Comunicação

- **Descobrimento da América - 5 meses**
- **Assassinato de Lincoln na Europa - duas semanas**
- **Neil Armstrong na Lua - 1,3 segundos**

A Internet: Um Pequeno Histórico

- **1964** Concepção
- **1969** Início da Internet
- **1972** Criação email
- **1979** Criação USENET
- **1982** Protocolos TCP/IP e adoção do nome Internet
- **1984** 1.000 computadores
- **1989** 100.000 computadores
- **1991** 1.000.000 computadores Aparecimento da WWW
- **1993** Aparecimento primeiro browser gráfico, MOSAIC
- **1994** Aparecimento dos primeiros mecanismos de busca e catálogos

A Informação e a Internet

- No princípio era o caos
- Equivalente a uma biblioteca sem fichas catalográficas
- O acesso à informação era privilégio de experts
- Informação compartilhada principalmente por meio do correio eletrônico, usenet news e ftp
- Da mesma forma que o valor de uma biblioteca está diretamente relacionado ao índice que lista seus livros, o valor da Web é estreitamente dependente dos mecanismos de pesquisa que a servem
- Embora de forma rudimentar, a WWW contribuiu para a explosão da quantidade de informações disponíveis na Internet
- Tentativas de ordenar e classificar o caos

Tamanho da Web

- Estima-se que existam hoje cerca de 150 milhões de páginas na Web distribuídos em 650.000 servidores
- Em 1995, quando do lançamento do AltaVista, existiam 50 milhões de páginas distribuídas por 100.000 servidores

Recursos Disponíveis

- World Wide Web
- Web Browsers
- Índices
- Catálogos
- Meta Pesquisas
- Usenet News
- Listas de discussão
- Notícias
- FTP (File Transfer Protocol)

Catalogação do Conteúdo da Web

A Indexação da Web (1)

- **Tarefa bastante difícil**
  - Grande amplitude
  - Grande taxa de crescimento
  - Volatilidade dos documentos armazenados
    - Tempo médio de vida de um documento na Web é de 75 dias, com uma grande parcela de documentos sendo modificados a cada dez dias.
    - 10% dos links armazenados não mais existem
  - Necessidade de grandes investimentos em hardware

A Indexação da Web (2)

- **Mesmo os mecanismos de busca mais poderosos indexam apenas amostras de cada site**
  - AltaVista - máximo de 600 páginas por site
- **Os mecanismos de pesquisa cobrem a metade ou menos das páginas existentes**
- **Acesso negado por alguns sites aos mecanismos de busca**

A Indexação da Web (3)

- O mais importante hoje não é ser o maior e sim o melhor
- Tenta-se indexar inteiramente apenas os sites mais frequentemente visitados
- As respostas a 90% das perguntas são encontradas dentre apenas um milhão de páginas
- Mais de 90% das páginas nunca são acessadas
- Duplicação de informações
- Enorme trabalho para manter os índices “limpos”

A Indexação da Web (4)

- **Não existe nenhum mecanismo de busca que indexe toda a Web**
  - O avanço da tecnologia levou à criação de novos formatos de armazenamento de documentos que são difíceis ou impossíveis de se indexar criando sites “infinitos” em tamanho
    - Bancos de Dados
    - Conteúdo gerado dinamicamente
    - Arquivos PostScript, PDF
- **Spamdexing**
  - White on white
  - Jump Pages
  - Uso indevido de palavras chave

Catalogação do Conteúdo da Web

- **Realizada por seres humanos e não por computadores**
- **O principal e mais bem sucedido catálogo da Web é o Yahoo**
- **Classificação da informação em categorias**
  - Pollution
    - Society and Culture:Environment and Nature:Pollution
- **Dilema: Tamanho x Qualidade**
- **A diferença entre um catálogo e um índice: catálogos fornecem um contexto**
- **Missão impossível**

Aparecimento dos mecanismos de busca

Yahoo	1994
Lycos	1994
Altavista	1995
Infoseek	1995
Excite	1995
HotBot	1996
LookSmart	1996

Classificação dos mecanismos de busca

- **Genéricos**
  - Excite, Altavista, Lycos, OpenText
- **Especializados**
  - SMART (Ciência da Computação)
    - URL: http://www.ncstrl.org/
  - MedLine (Medicina)
    - http://www.nlm.nih.gov

Índices, Catálogos e Híbridos

- **Índices**
  - Dados obtidos por meio de programas de computadores
    - Spiders, Crawlers
  - Mais abrangentes
- **Catálogos**
  - Criados por pessoas
  - Informação classificada de acordo com categorias
  - Serviço personalizado, de menor abrangência
- **Híbridos**
  - Alguns índices oferecem também informação catalogada, como por exemplo, Excite e Infoseek

A Pesquisa da Informação

O que se busca na Internet

- Informação em geral
- Endereços e telefones de pessoas e empresas
- Entretenimento
- Comunicação
- Grupos de pessoas com interesses comuns

Termos mais pesquisados (1)

- **Pesquisas Simples (Uma palavra)**
  1. Hotels
  2. Sex
  3. Training
  4. Advertising
  5. Accountants
  6. Antiques
  7. Jobs
  8. Schools
  9. Wine
  10. Weather
- **Pesquisas Compostas**
  1. Estate Agents
  2. Search Engines
  3. Bargains Cheap
  4. Beauty Therapy
  5. Career Development
  6. Computer Security

Termos mais pesquisados (2)

- **Yahoo Top 200 Search Words**
  - *http://eyscream.com/yahootop200.html*
- **WebCrawler Search Ticker**
  - *http://webcrawler.com/WebCrawler/Fun/SearchTicker.html*
- **DogPile Top 200 Search Words**
  - *http://eyescream.com/dogpiletop200.html*

Características dos mecanismos de busca

- Tamanho
- Atualização
- Submissão de páginas
- Suporte a Frames
- Sites Protegidos
- Taxa de atualização
- Redirecionamento
- Taxa de Relevância
- Meta Tags
- Stemming
- Exibição dos resultados
- Páginas visitadas
- Data
- Profundidade
- Image Maps
- Popularidade de links
- Restrições de acesso
- Stop words
- Punição a spammers
- Remoção de sites
- Capitalização de palavras
- Descrições

Metasearching

- **Vantagens**
  - Pesquisas submetidas simultaneamente a vários mecanismos de busca
  - Economia de tempo
- **Desvantagens**
  - Pesquisas feitas segundo um mínimo denominador comum entre os serviços de busca
    - Exceção: ProFusion
  - Agrupamento das respostas segundo critérios diferentes

Metasearch

- **Metacrawlers não fazem a indexação da Web à semelhança dos índices. Eles enviam perguntas a vários mecanismos de busca e constroem suas listagens a partir das respostas obtidas**
- **ProFusion**
  - URL: http://www.designlab.ukans.edu/profusion/
- **AskJeeves**
  - URL:http://www.askjeeves.com
- **DogPile**
  - URL:http://www.dogpile.com
- **Inference Find**
  - URL: http://www.inference.com/ifind
- **MetaCrawler**
  - URL: http://www.metacrawler.com

Conceitos Úteis

- Concept Search
- Boolean Search
- Fuzzy Search
- Index
- Keyword Search
- Phrase Search
- Proximity Search
- Stemming
- Stop Words

Exemplos

- **+informação +tecnologia -futebol**
- **“Rubens Queiroz de Almeida”**
- **ProFusion**
- **futebol AND Pelé**
- **futebol OR Pelé (AND) NOT Garrincha**
- **Pelé NEAR Garrincha**
- **Where can I find information on how to read in English (Excite)**
- **comunic***
  - comunicação, comunicar, comunicador, etc.

Principais Mecanismos de Busca

Yahoo (1)

- **URL:** http://www.yahoo.com
- **Lançado em abril de 1994 por David Filo e Jerry Yang, estudantes de doutoramento em Stanford**
- **25 a 40 milhões de acessos mensais**
- **50 milhões de páginas servidas diariamente**
- **70 servidores rodando FreeBSD**
- **Tráfego de 200MB/s em horários de pico**
- **Informações catalogadas por uma equipe de aproximadamente 50 pessoas**

Yahoo (2)

- **Uma das mais empresas mais bem sucedidas na Internet**
  - Valorização de 511% de suas ações em 1997
  - Valor de mercado de 2.8 bilhões de dólares
  - E sem cobrar um centavo de seus usuários
- **Considerado por muitos como o ponto de entrada da Internet**
- **Ponto de partida obrigatório para exploração de qualquer assunto**
- **Integrado com Altavista**
- **Problemas**
  - Não consegue acompanhar o crescimento da Web
  - Demora de meses, até mesmo anos, para catalogar novas entradas

Altavista

- **URL:** http://altavista.digital.com
- **100 milhões de páginas indexadas**
- **10 milhões de páginas consultadas diariamente**
- **Atualização: Um a trinta dias**
- **Submissão de URLs: um dia**

Excite

- **URL:** http://www.excite.com
- **55 milhões de páginas indexadas**
- **3 milhões de páginas consultadas diariamente**
- **Atualização: uma a três semanas**
- **Submissão de URLs: três semanas**
- **Tamanho do índice: 50GB**

HotBot

- **URL:** http://www.hotbot.com
- **110 millhões de páginas indexadas**
- **Até 10 milhões de páginas consultadas diariamente**
- **Atualização: um dia a duas semanas**
- **Submissão de URLs: um a dois dias**

Infoseek

- URL: http://www.infoseek.com
- 30 milhões de páginas indexadas
- 6 a 10 milhões de páginas visitadas diariamente
- Atualização: uma a duas semanas
- Submissão de URLs: minutos

HyperLink

- **URL:** http://rankdex.gari.com
- **Serviço experimental**
- **O que um site diz a respeito de si próprio não conta, o que conta são os links para este site e descrições associadas**
- **Possibilidade maior de se obter documentos mais relevantes**
- **Não aceita submissão de URLs**

Pesquisa de Notícias

- **Possibilitam o acesso a notícias veiculadas em centenas de jornais e agências de notícias. Resultados excepcionalmente bons e direcionados.**
  - NewsTracker
    - http://nt.excite.com
  - NewsBot
    - http://www.newsbot.com
  - News Index
    - http://www.newsindex.com
  - NewsHub
    - http://www.newshub.com
  - PaperBoy
    - http://www.paperboy.net
  - TotalNews
    - http://www.totalnews.com

Qual o melhor mecanismo de busca

- **Escolhas se alternam entre Infoseek, HotBot e AltaVista**
- **Altavista, Excite, HotBot, Infoseek**
  - Competição promovida pela revista PC Computing, em maio de 1997, em Las Vegas
    - Tema da Competição: “Qual o melhor mecanismo de busca para uso comercial?”
    - Classificação: HotBot, Excite, Altavista, Infoseek
- **Opções pessoais**
  - Escolha o seu e aprenda o máximo que puder sobre ele

Perspectivas

Novas Tendências (1)

- **Filtragem colaborativa**
  - Alexa
    - URL: http://www.alexa.com
- **Crescimento de mecanismos de pesquisa especializados**
- **Metasearching baseado no cliente**
  - WebSleuth
    - URL: http://www.promptsoftware.com

Novas Tendências (2)

- **Agentes**
  - Inquisit
    - URL: http://www.inquisit.com
- **Colaboradores Humanos**
  - The Mining Company
    - URL: http://www.miningco.com
  - Human Search
    - URL: http://www.humansearch.com

Mudança no perfil dos serviços

- **Objetivo: capturar audiências maiores com mudança do perfil de serviços, mudando de ponto de trânsito para ponto de destino**
  - Oferta de endereços eletrônicos gratuitos
  - Chat
  - Áreas para discussão
  - Canais (Channels)
- **Pagamento para inclusão nas listagens**

Marketing

Marketing na Internet (1)

- Boas colocações de seu site nos mecanismos de busca
- Divulgação paga
- Listas de discussão
- Anúncios em grupos na Usenet

Marketing na Internet (2)

- **Links a partir de outros sites**
- **Criação de listas próprias**
- **Como está o seu site?**
  - Did-It
    - URL: http://www.did-it.com
  - PositionAgent
    - URL: http://www.positionagent.com
- **Promoção de seu site**
  - Do-It-Yourself
    - URL: http://www.mmgco.com/top100.html
  - WebStep
    - http://www.mmgco.com/top100.html

Estratégias Eficientes de Marketing (1)

- **A maioria das pessoas olha as páginas superficialmente**
  - Escreva de forma a conseguir transmitir a maior quantidade de informação rapidamente
    - textos em destaque (negrito, links, variações de fonte e cores)
    - listas (bullets)
    - Uma idéia por parágrafo
    - Estilo da pirâmide invertida, conclusão em primeiro lugar
    - Utilize a metade das palavras que normalmente usa

Estratégias Eficientes de Marketing (2)

- **Credibilidade**
  - páginas bem feitas, com imagens de qualidade
  - uso de links para páginas de outros sites (por que não?)
  - evitar uso excessivo de termos de marketing (o melhor, mais bonito, mais inteligente, etc.)

Erros de digitação ...

- www.infosek.com
- www.webcralwer.com
- www.whitehouse.com

Publicação Eletrônica e o Futuro do Livro

Características do Livro Tradicional

- Portabilidade
- Acesso randômico
- Multimedia
- Fácil acesso
- Baixo consumo de energia

Características do Livro Eletrônico

- Produção e disseminação rápida
- Fácil de atualizar ou corrigir
- Colaborativo e interativo
- Eliminação de intermediários
- Maior interação com consumidor final
- Alcance limitado
- Necessidade de domínio de tecnologias avançadas, tanto pelo consumidor como pelo produtor
- Limitações quanto ao tipo de ambiente em que pode ser utilizado
- User-unfriendly
- Consome mais energia que um livro

O Futuro do Livro

- **A curto prazo, a substituição total do livro convencional pela publicação eletrônica é pouco provável**
- **Tendência a substituir certos tipos de publicações ou jornais**
  - Dados ou informações que possuem um tempo de vida mais curto, necessitam ser atualizadas mais frequentemente e direcionadas a audiências limitadas e bem conhecidas
    - Livros de referência, informações de negócios, estatísticas e enciclopédias
    - Publicações acadêmicas

Porque o livro não vai acabar (tão cedo)

- A publicação eletrônica é conveniente para quem conhece a tecnologia e incômoda para quem não usa computadores diariamente
- Informalidade, conveniente em todos os ambientes
- Problemas na manutenção de documentos digitais

Ferramentas de Acesso a Informação

Os Dez Idiomas mais Usados na Internet

Inglês	82,3%
Alemão	4%
Japonês	1,6%
Francês	1,5%
Espanhol	1,1%
Italiano	0,8%
Português	0,7%
Sueco	0,6%
Holandês	0,4%
Noruegues	0,3%

A Língua Inglesa

- **Ferramenta para acesso à informação**
- **Estrutura simples**
- **Muitos pontos de afinidade com a língua portuguesa**
- **Facilidade de aprendizado para leitura (inglês instrumental)**
  - 250 palavras mais comuns - 57% total
  - 1000 palavras mais comuns - 99,25% total
- **Domínio para leitura alcançado em um prazo muito mais curto**
- **Habilidade mais duradoura e importante**

Rubens Queiroz de Almeida
email: queiroz@unicamp.br

http://www.dicas-l.unicamp.br