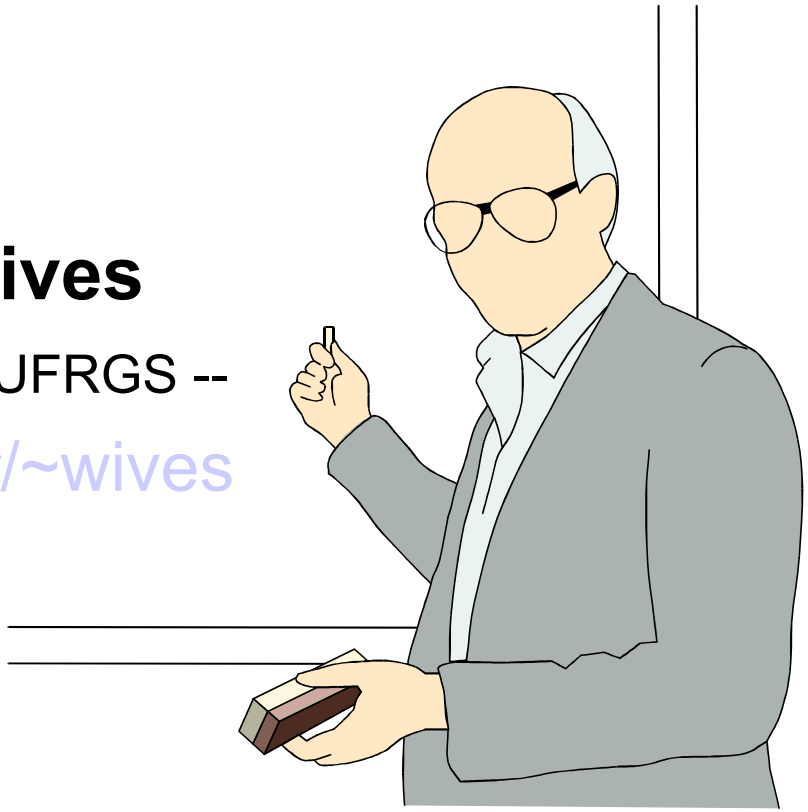
Recuperação da Informação na web: tecnologias e perspectivas

Leandro Krug Wives

-- Instituto de Informática – UFRGS --

http://www.inf.ufrgs.br/~wives



Agenda

- **Um pouco de história**
- **Panorama da web atual**
- **Tecnologias de busca**
- **Problemas, alternativas e perspectivas**

Um pouco de história...

Desde seus primórdios, o homem tem inventado diferentes meios de escrita a fim de registrar eventos (dados e informações) a fim de transmiti-los para as futuras gerações...

Um pouco de história...

- **Biblioteca de Alexandria:**
  - **A primeira biblioteca pública de larga escala**
  - **Catálogo de 700.000 manuscritos, já em 300 AC**
  - **Queimada 1600 anos atrás!**

Apesar de nos lembrar as bibliotecas atuais,
não se imaginava, na época, o que estava por vir...

Um pouco de história...

Não é preciso ir tão longe, pois ainda há quem duvide da capacidade do homem:

"*640KB devem ser suficientes para todos*"

Bill Gates, 1981

(1 MP3 ocupa 6x mais do que isso!)

"*Não há demanda no mundo para mais do que cinco computadores*"

Thomas Watson, IBM, 1943

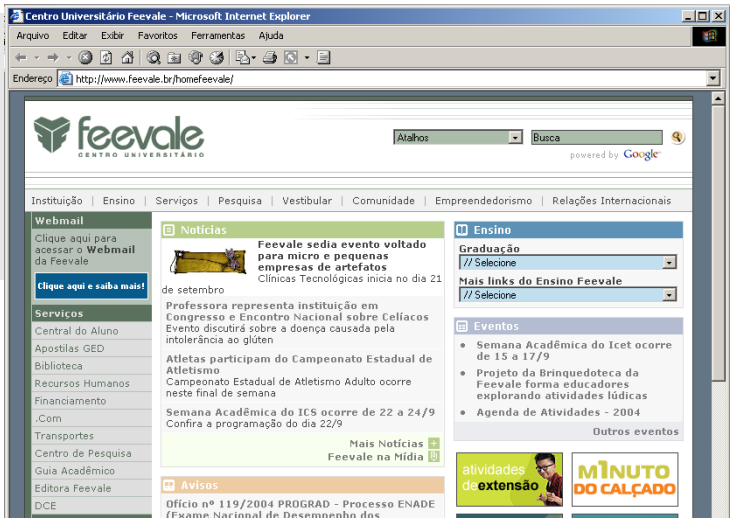
(graças a Deus ele estava errado!)

Eis que surge a web...

"***Todos nós já escutamos que um milhão de macacos batendo em um milhão de máquinas de escrever vão, eventualmente, produzir obras inteiras de Shakespeare. Agora, graças à Internet, sabemos que isso não é verdade.***"

Robert Wilensky, 1997

Eis que surge a web...



- A "**parte gráfica da Internet**" foi criada em 1989 por Tim Berns Lee (um físico)
- Compreende (hiper-)**documentos**, descritos em HTML, **que podem apontar para outros documentos** (escritos por outras pessoas em qualquer lugar do mundo), **formando a teia de alcance mundial**
- Ela "**deslanchou**" em 1993, **quando um grupo de estudantes** (média 23 anos), da University of Illinois, **desenvolveu o** navegador **Mosaic** (atual Netscape)

Eis que surge a web...

- **Para se ter uma idéia do que é a web:**
  - **Em 2004 o Google**
    - **Possuia mais de 6 bilhões de itens indexados (atualmente estima-se em mais de 20 bilhões)**
    - **Realizava mais de 200.000.000 (200 milhões) de buscas por dia**
- **Cerca de 80% desta informação está em Inglês!**

O problema não é só da web...

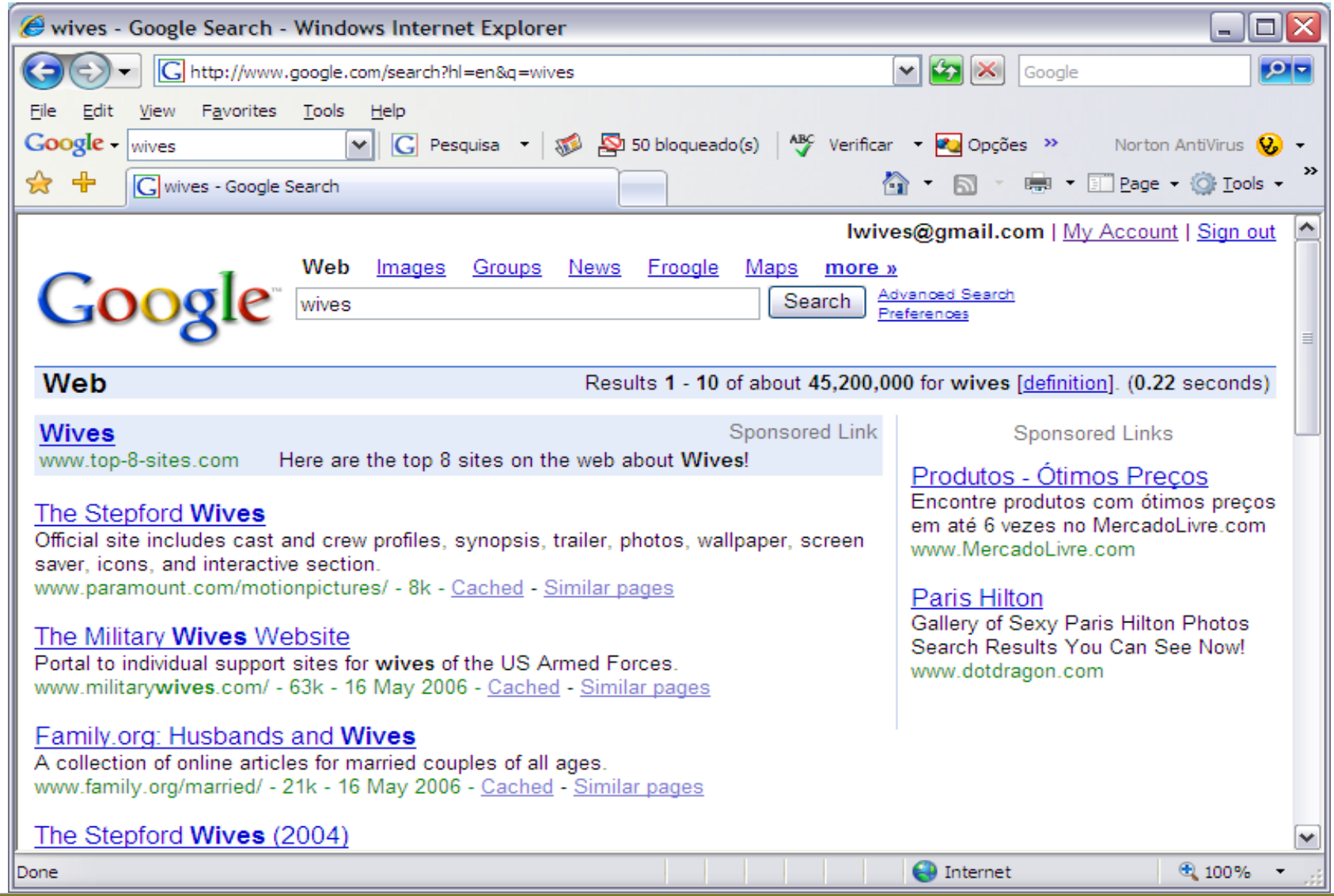
- Um computador pessoal suporta atualmente mais de 160 GB de informação em seu disco rígido;
- Com isso nossa capacidade de armazenamento de documentos (assumindo 1 byte por caractere e 3500 caracteres por página de texto puro) é muito maior do que 22 milhões de páginas de informação;
- Em escalas corporativas, a capacidade de armazenamento e a diversidade de formatos de informação é comparável ao número de funcionários multiplicado por 10 ou 100.

Resumindo...

- **Começamos com os instrumentos de escrita;**
- **Criamos as bibliotecas;**
- **Passamos pela a digitalização (computador e os dispositivos de armazenamento);**
- **Criamos novos sistemas de manipulação de informação e mecanismos rápidos de transmissão de informação...**

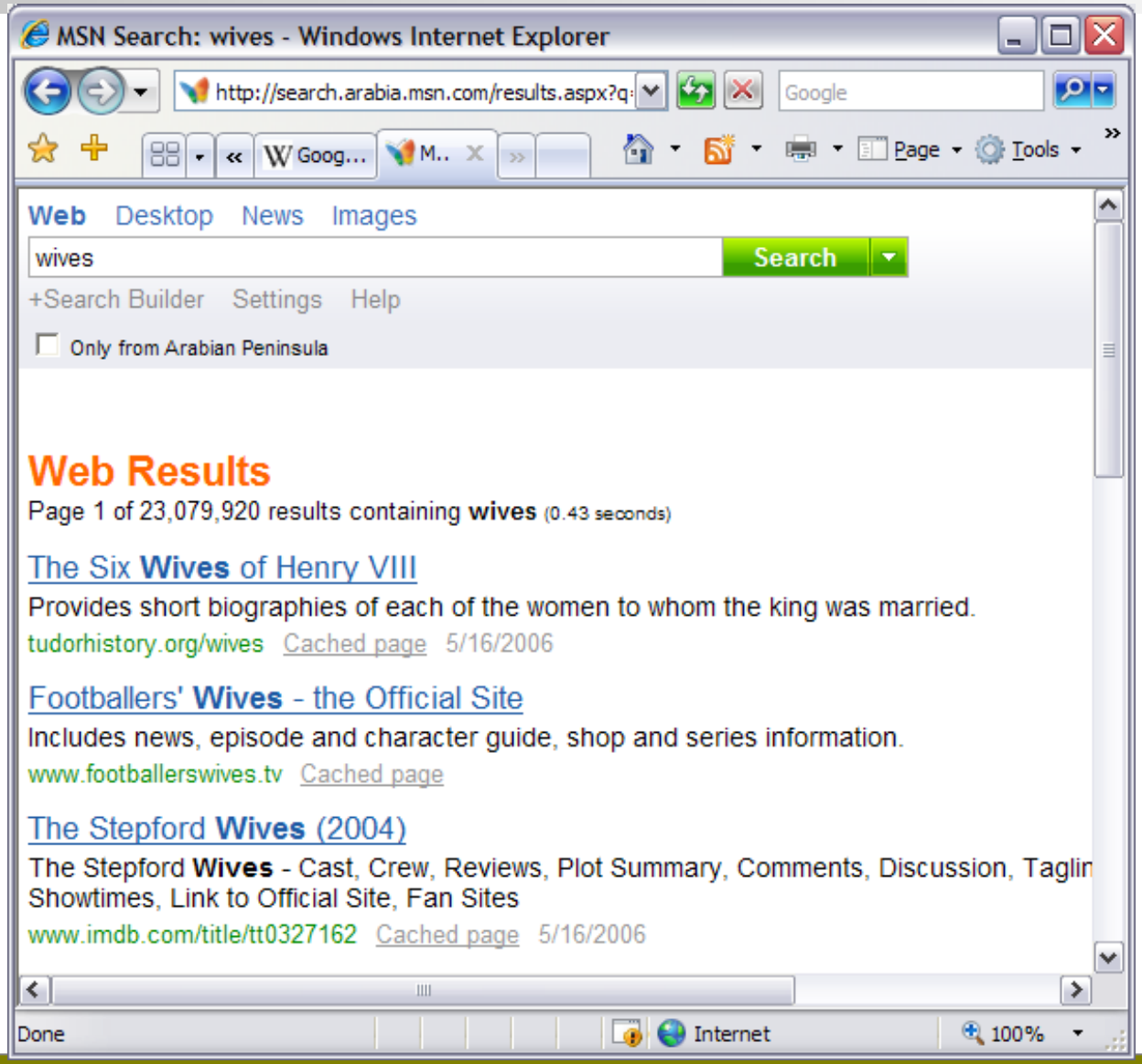
E conseguimos achar o que precisamos???

Alguns exemplos...



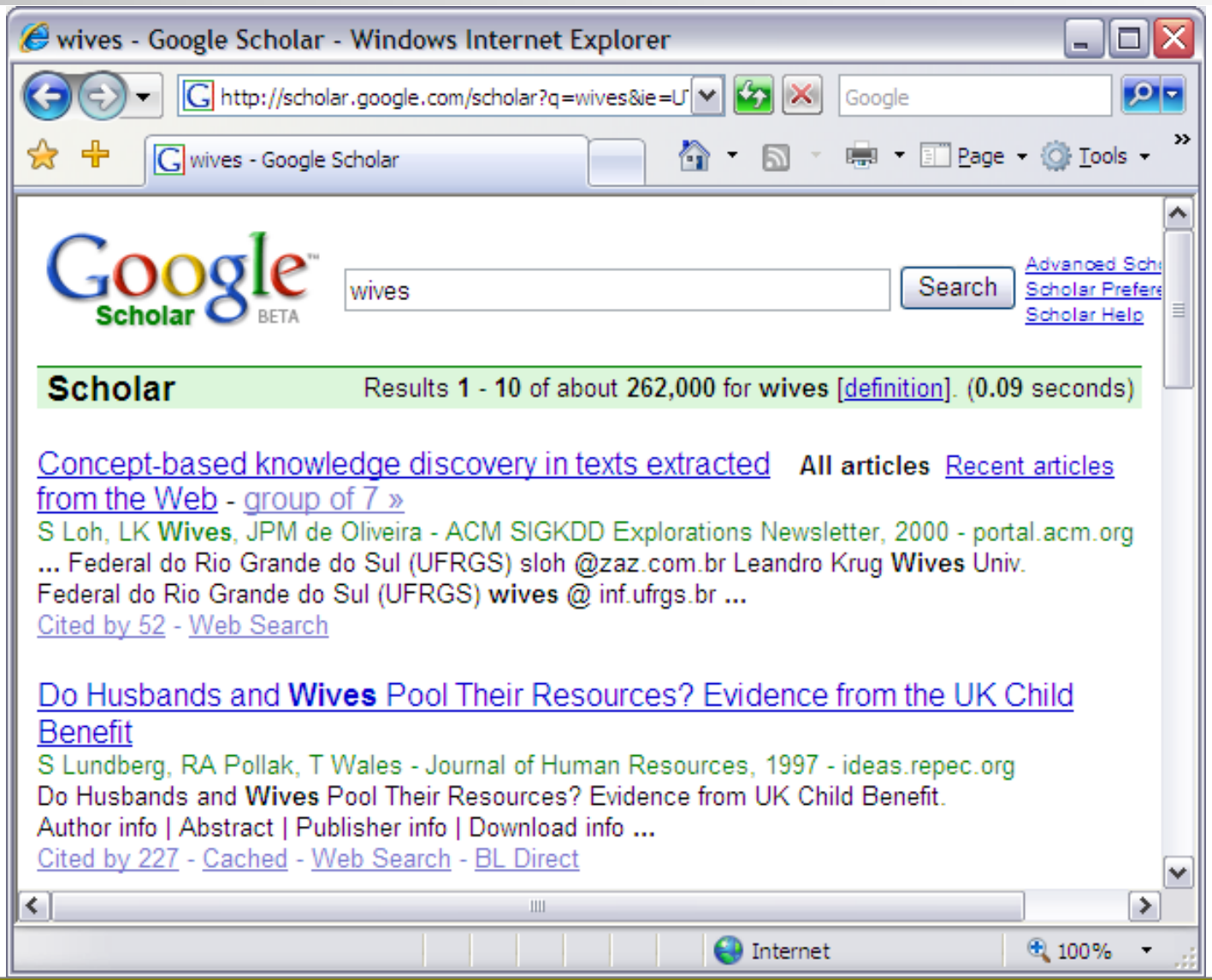
A procura por “wives” retornou páginas sobre esposas (em inglês), o que seria de esperar, dentre 45 milhões de resultados.

Alguns exemplos...



Aqui os resultados são outros. O sistema de ordenação (ranking) é diferente.

Alguns exemplos...

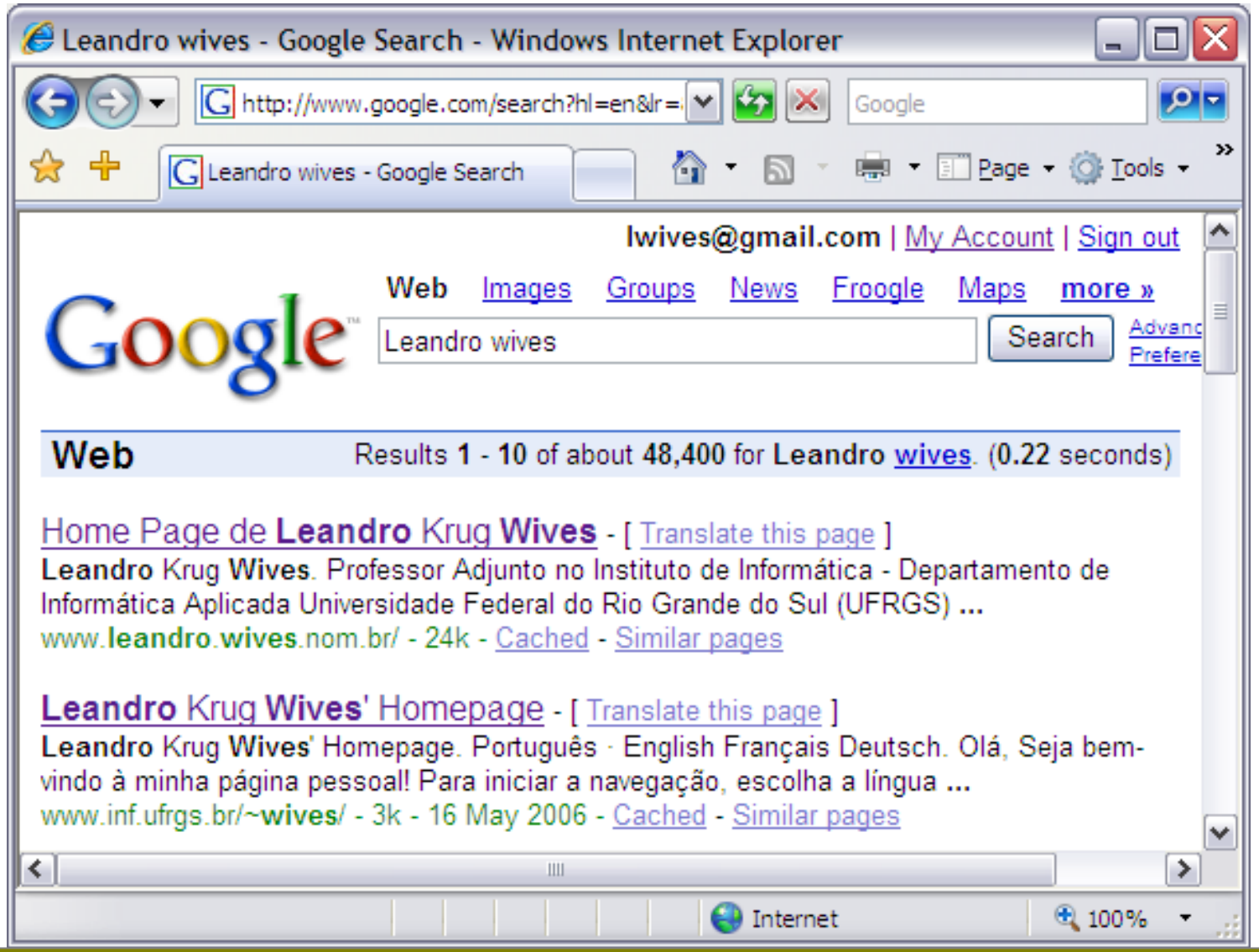


Na busca especializada (acadêmica), a situação muda!

Alguns exemplos...

Não só pessoa é chamada de “Leandro”!

Alguns exemplos...



Quantos Leandros com sobrenome Wives você conhece?

Os problemas, então...

- **Definição do "foco" ou tema:**
  - A linguagem de busca oferece auxílio ou os operadores necessários?
  - O usuário sabe (especificar) o que quer? Tem dificuldade de abstração e seleção de palavras significativas!
- **Representação da informação/ambigüidade:**
  - É possível representar o que o usuário deseja?
  - Não há outros significados para os signos usados?
  - E se o internauta desejasse procurar algo sobre "wives" (esposas, em inglês), mesmo no Scholar Google ou no MSN?
- **Sobrecarga:**
  - Quantidade muito grande de resultados desordenados

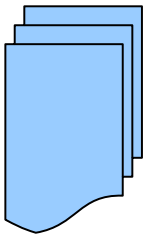
Tecnologias de busca

- **Fases:**
  - Processo de indexação
  - Processo de busca

Tecnologias de busca

- **Processo de indexação:**
  - Identificação de palavras
  - Remoção de stopwords
  - Análise de radicais (stemming), dicionários de sinônimos e de erros freqüentes
  - Análise de freqüência
  - Indices invertidos

Processo de indexação



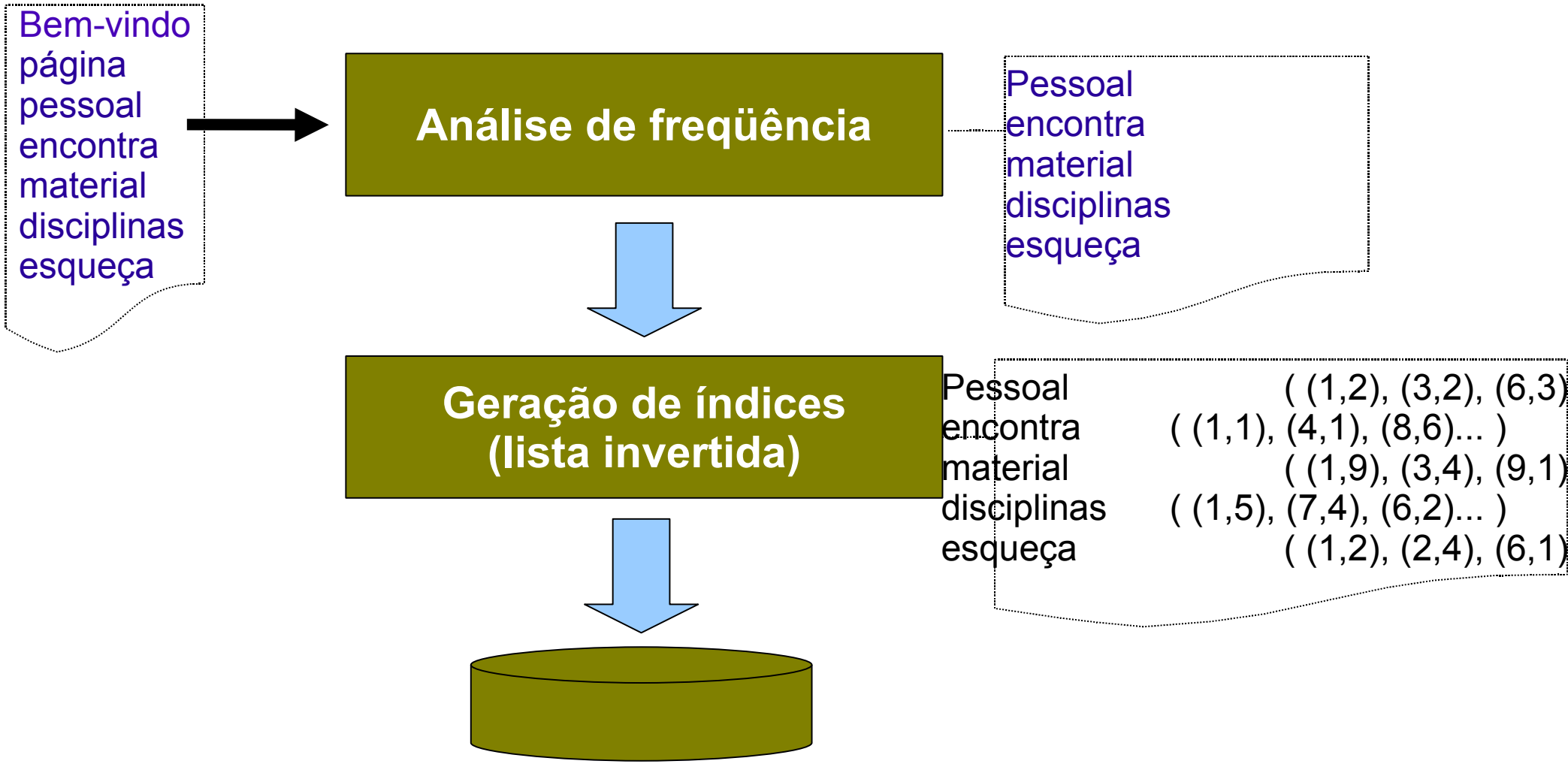
Identificação de palavras

<body>Bem-vindo! Esta é minha página pessoal. Aqui
 você encontramaterial de disci-plinas. Não esqueça de...

Identificação de Stopwords

bem-vindo esta é minha página pessoal aqui você encontra material de disciplinas não esqueça de...

Processo de indexação



Tecnologias de busca

- **Processo de busca:**
  - Localização (*matching*) no gigantesco e distribuído índice (google tem muitos clusters espalhados e com objetivos diferentes)
  - Ordenação (ranking) por freqüência, citações, links
  - Agrupamento por semelhança
  - Retroalimentação, refinamento (sinônimos, identificação de erros, ignorar palavras freqüentes)
  - Interação (uma só consulta não adianta)!

Algumas considerações

- **Os recursos que desenvolvemos produzem e propagam ainda mais a informação**
- **Continuamos perdidos (*lost in the cyberspace*) e sobrecarregados por uma imensa quantidade de informação**
- **Não somos capazes de administrar a quantidade de informações que é gerada**
- **Nosso poder de comunicação e manipulação é similar ao dos tempos das cavernas!**

Alternativas e perspectivas

- Há muito investimento na área
- Pesquisadores e centros de pesquisa renomados trabalhando com estas companhias
- Novas tecnologias e ferramentas: text mining, web semântica etc.
- Desktop search (Microsoft, Google, Yahoo)
- Bancos de Dados com extensões para a manipulação de textos/dados semi-estruturados (DB2, Postgres etc.)

Alternativas e perspectivas

- **Dar mais inteligência às ferramentas de comunicação e geração de informação:**
  - bate-papo (*chat*);
  - correio eletrônico (*email*, *webmail*);
  - grupos de notícias, listas e fóruns de discussão;
  - às comunidades de relacionamento.

Eles permitem a troca direta e indireta de informações entre pessoas ou grupos de indivíduos, oferecendo um espaço onde as pessoas podem colocar seus problemas ou auxiliar na solução dos problemas dos outros.

Alternativas e perspectivas

- A inteligência é dada através de métodos de IA, estatística e matemática (área de *text mining*):
  - **Extração de informações:**
    - Encontra valores de atributos implícitos nos textos: idade de pessoas, datas.
    - Extrai partes significativas de textos: “o que os jornais falam sobre mim?”
  - **Lexicometria e análise de distribuição:**
    - Encontra temas presentes e sua freqüência: “*educação* presente em 85% dos textos.”
  - **Recomendação de conteúdo**

Alternativas e perspectivas

- **Descoberta de associações:**
  - Regras associativas para identificar a probabilidade condicional entre temas:
    “atendimento” --> “demora”;
- **Classificação ou categorização:**
  - Encontrar a categoria ou o assunto de um texto: “saúde”;
- **Análise de similaridades e de diferenças**
  - Separar (agrupar) textos por similaridade (*clustering*);
  - Encontrar temas ou regras exclusivos: “só o texto do João fala sobre poesia”.

Mais problemas...

- Ainda há um grande *gap* entre a informação disponível para ferramentas que possam trabalhar com estes problemas e a informação mantida em uma forma compreensível para o homem!

->Enriquecer a informação disponível com semântica que possa ser processada pela máquina!

Semantic web!

- **Contém**:
  - meta-Informação capaz de ser “lida” pela “máquina”
  - Representação explícita do conteúdo e da semântica dos documentos e das informações
- **Permite**:
  - web baseada em conhecimento
  - construção de serviços *web* que amplifique suas capacidades atuais
  - prover um novo nível de serviços qualitativamente melhores

Uma luz no fim do túnel!

- **Este suporte é essencial para levar a web ao seu potencial máximo**
- **Assim serviços automatizados vão melhorar sua capacidade de assistir os seres humanos em atingir seus objetivos a partir da "compreensão" do conteúdo da web, provendo mecanismos mais acurados de filtragem, categorização e busca.**

Algumas aplicações

- **Inteligência do Negócio** (*Business Intelligence*)
  - Análise de problemas descritos em *Call Centers,* identificação das melhores soluções para os mesmos, segmentação de Mercado, Marketing de Precisão.
- **Intel. sobre o consumidor** (Customer Intel.)
  - Conhecer necessidades, valores e satisfação de clientes.
- **Inteligência Competitiva** (*competitive Intel.*)
  - Conhecer elementos do ambiente empresarial (players) e suas estratégias.
- **Gestão de Conhecimento e de Documentos** (*Knowledge Management* / *Document Management*)

Recuperação da Informação na web: tecnologias e perspectivas

Leandro Krug Wives

-- Instituto de Informática – UFRGS --

http://www.inf.ufrgs.br/~wives