Um estudo de caso:

os motores de busca estruturando a teia mundial

Imre Simon
Universidade de São Paulo
São Paulo, Brasil

<is@ime.usp.br>

http://www.ime.usp.br/~is/

Um estudo de caso

Case study: tentativa de inferir do particular para o geral

O curso: Inovação e Cooperação na Internet

O caso: Como organizar gigantescas massas de dados

A conclusão: Vitória retumbante de métodos sintáticos

O paradigma básico: o índice de um livro técnico

A teia: um verdadeiro palheiro (dados de 2001)

repositório (aparentemente) caótico de informação

Google indexa 1,3 G páginas (3 TB de texto)

- um livro comum = 1MB
- Encyclopaedia Britannica (texto) = USP online = 240 MB
- todobr = 15 GB
- Biblioteca do IME = 40 GB

Google = 200 todobr = 12K USPonline = 3M livro

Um ano depois o Google indexa o dobro de páginas

Achando agulhas no palheiro da teia

Motores de busca

Google, AltaVista, todobr, radix, ...

Alguns exemplos de busca:

- Google: livro verde
- Google: MAC 339
- Google: Estruturas de Dados
- Google: Estruturas de Dados site:usp.br

Outros índices e compilações

Usenet : groups.google.com

Yahoo! : índice temático, feito à mão
Veja também: directory.google.com

Imagens: images.google.com

Notícias: news.google.com

ResearchIndex : Uma biblioteca do IME, só de CC

Amazon.com, Dedalus, Library of Congress, ...

Dicionários: www.dict.org, FOLDOC, br.ispell

Enciclopédias: www.wikipedia.com

Compilações especializadas: TechWeb, Debian, ...

Qual é a mágica?

Motores são mecanismos sintáticos.
Deonde vem a sua inteligência?

AltaVista (1995) -> Google (2000): ordenação por reputação

Exemplos de critérios de reputação:

- página mais referenciada
- página mais procurada
- localização do objetivo da busca : na URL, no título, ...
- artigo mais referenciado (em ResearchIndex)

Google, restrito a ResearchIndex: search engine

O motor de busca ajuda a auto-organização da teia

Auto-organização de um processo caótico

Centenas de milhões de usuários,
filtrando a informação continuamente

os motores percolam (e direcionam)
estas filtragens

eles organizam a informação em conhecimento

permitem a criação de especialistas eletrônicos
que conhecem tudo numa dada área

Conhecem tudo mesmo?

Há um tremendo buraco a negociar:

a instituição da propriedade intelectual

(but that's another story: "Irma la Douce", 1963)

- Google: "Jack LemmonShirley MacLaine"

O (difícil) caso do amplo acesso à literatura científica

- Public Library of Science
- BOAI: Budapest Open Access Initiative
- Debate Nature: e-daccess
- http://www.ime.usp.br/ is/PLoS/

GoogleWhacking

Ache uma busca que retorna uma só página

exemplos

- washingtonian hippopotami
- imre figura aspectos tropical waterloo people

Uma bela diversão: google: whacking

A última (e incrível) novidade de 2002

News: news.google.com

Integralmente gerado por computador

Inovação científica (secreta por enquanto)

Potencial valor como um modelo de negócio

Qual o interesse dos sites monitorados?

Como se faz um motor destes?

O mecanismo básico é este:

- robot percorre a teia que nem uma aranha
- debulhador retém apenas o texto
- indexador inverte as informações
- mecanismo de recepção das consultas
- buscador localiza as páginas a indicar
- ordenador prioriza estas páginas

Google usa dezenas de milhares de computadores com Linux

uma busca típica leva meio segundo

aspecto importante: use buzilhões de memória

dá para fazer tudo com componentes abertos,
de software livre

À guisa de conclusão

A estrutura de uma massa de dados pode esconder verdadeiros tesouros

Os tesouros podem ser recuperados sintáticamente de forma automatizada

Os índices e coleções passam a ferramentas indispensáveis

Que organizam e consolidam a massa de dados

E que estruturam a informação em conhecimento

Podem, talvez, a levar a novos modelos de negócios na rede

a