Como funcionam as máquinas de busca?

Instituto de Computação uff

Vanessa Braganholo

vanessa@ic.uff.br

Quantas vezes você usou o Google hoje?

Quanto tempo o Google demorou para dar cada resposta?

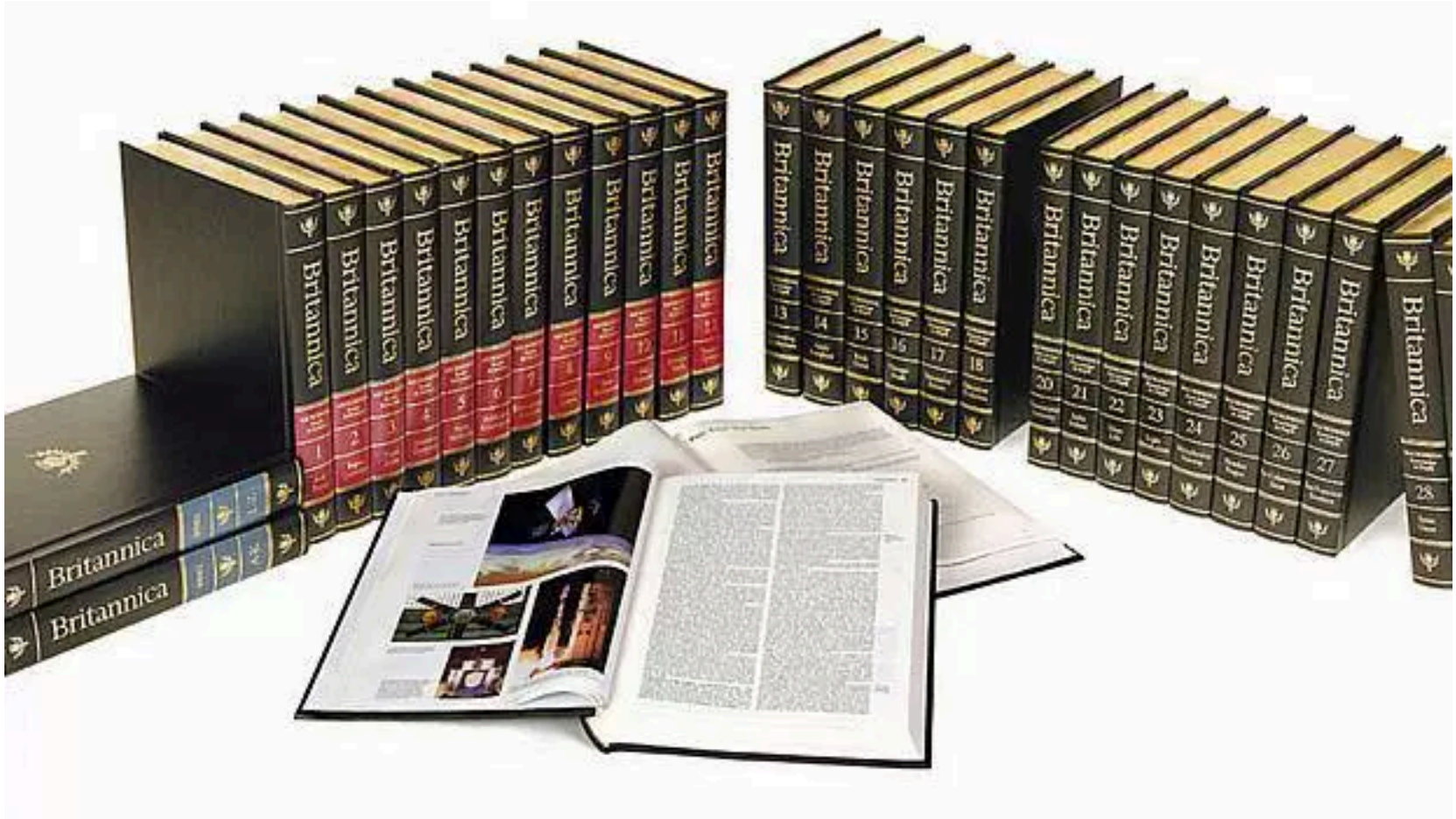
Como isso é possível?

Instituto de Computação

De volta ao passado 1980...

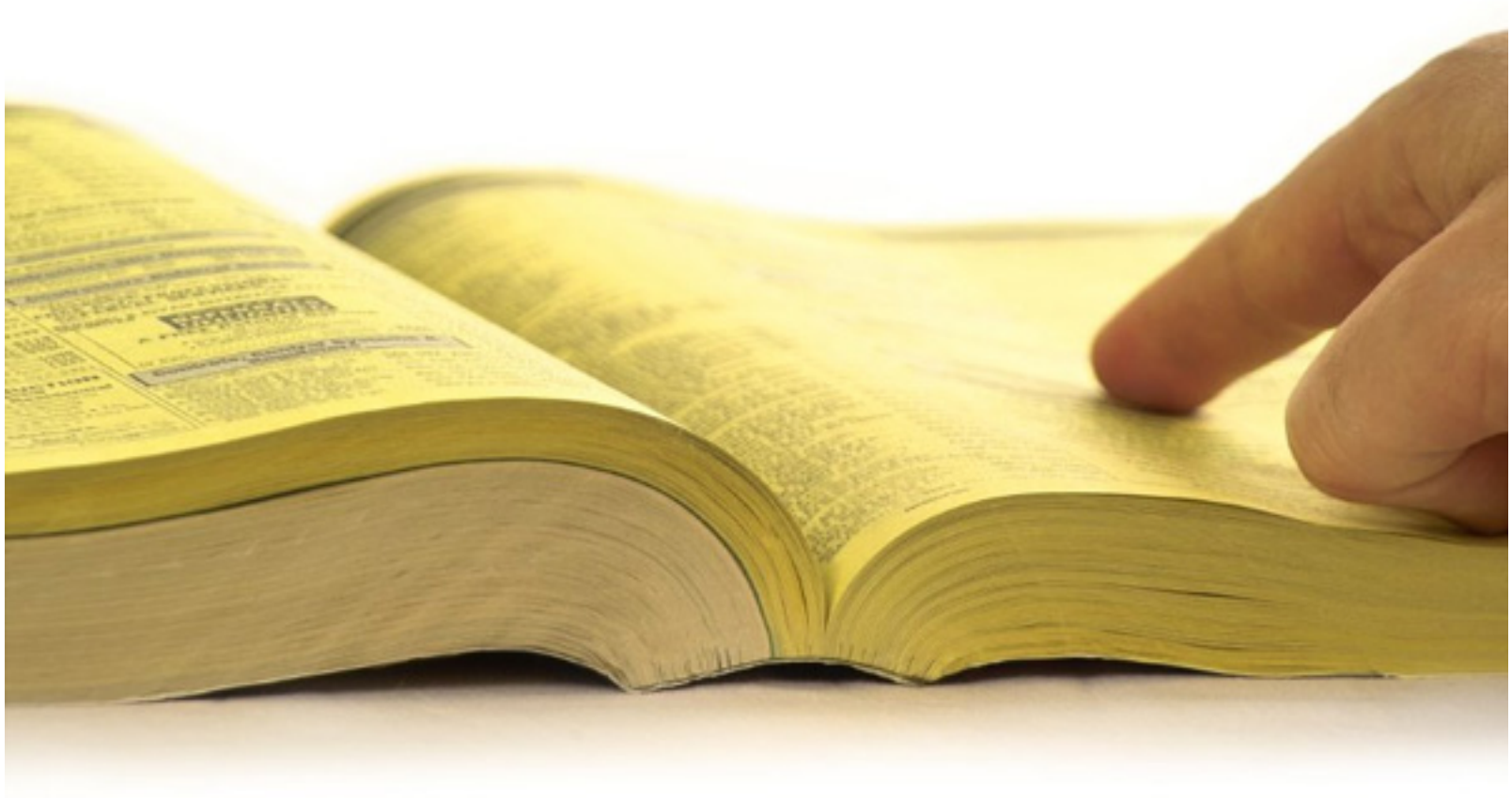
Quem foi Alan Turing?

Quem foi Alan Turing?



Qual o endereço do restaurante Cheiro de Mar?

Qual o endereço do restaurante Cheiro de Mar?



Instituto de Computação UFF

Será que a pousada do Sol Nascente é boa?

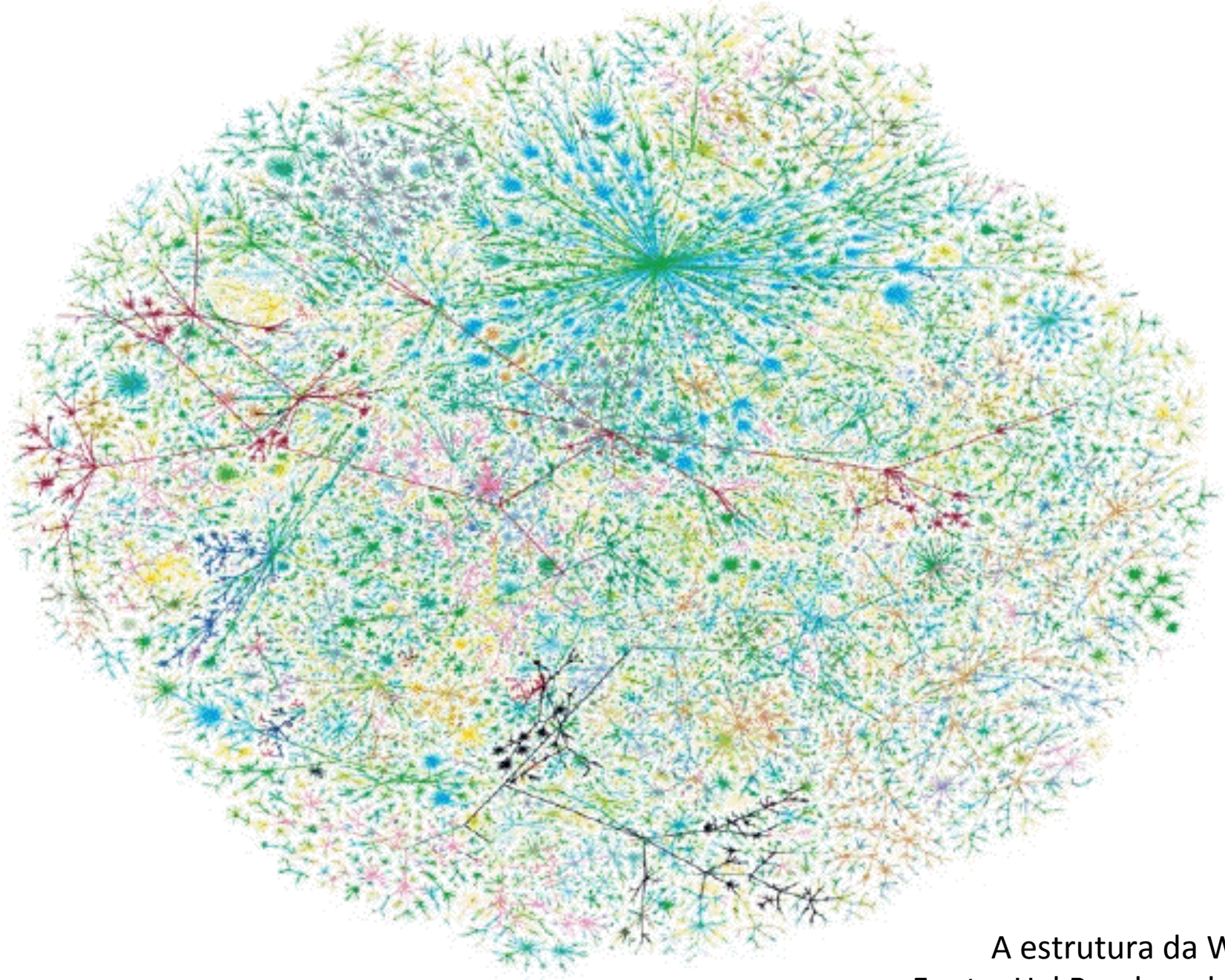
Será que a pousada do Sol Nascente é boa?

Instituto de Computação

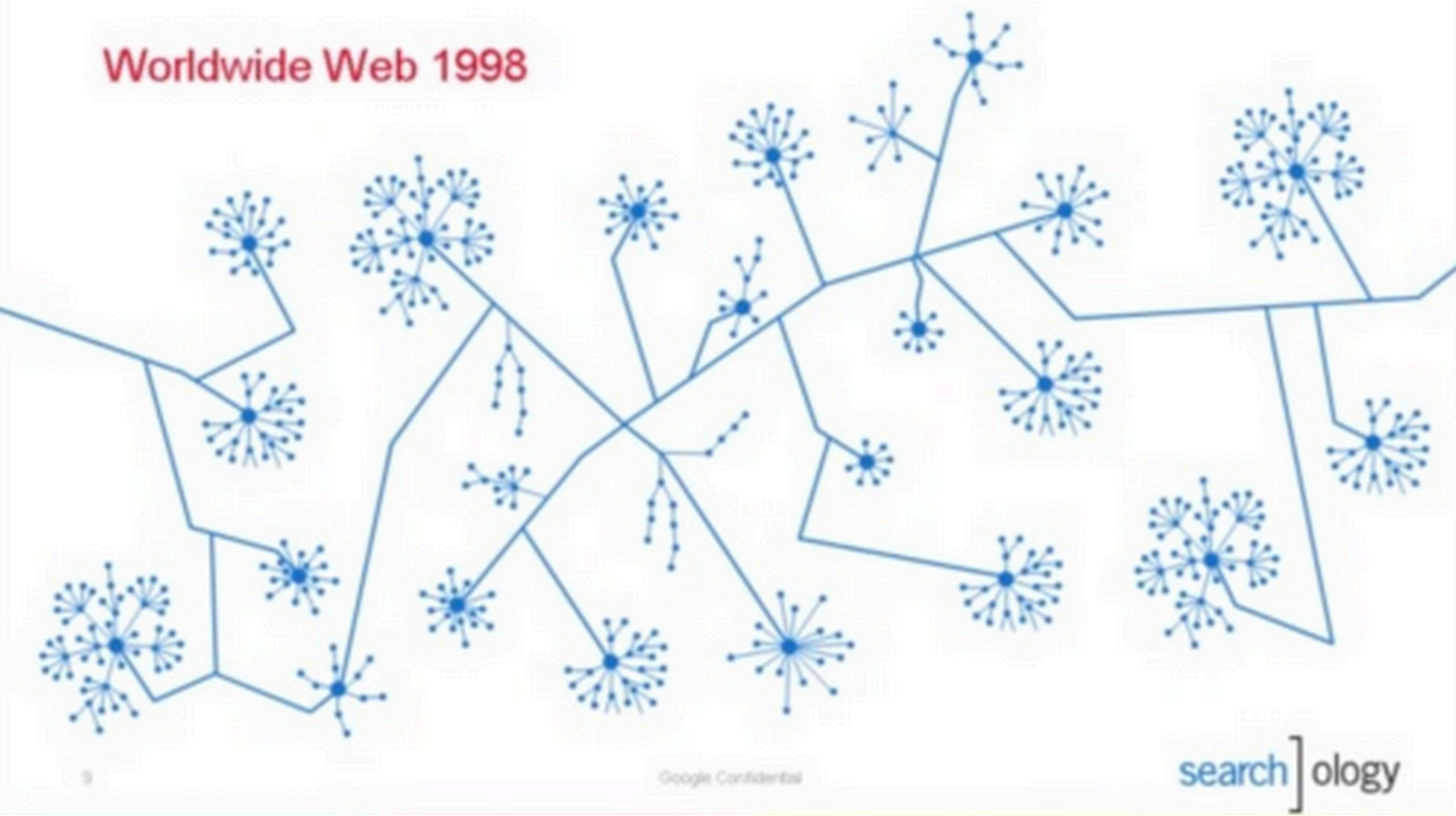
De volta para o futuro 1990...

Internet

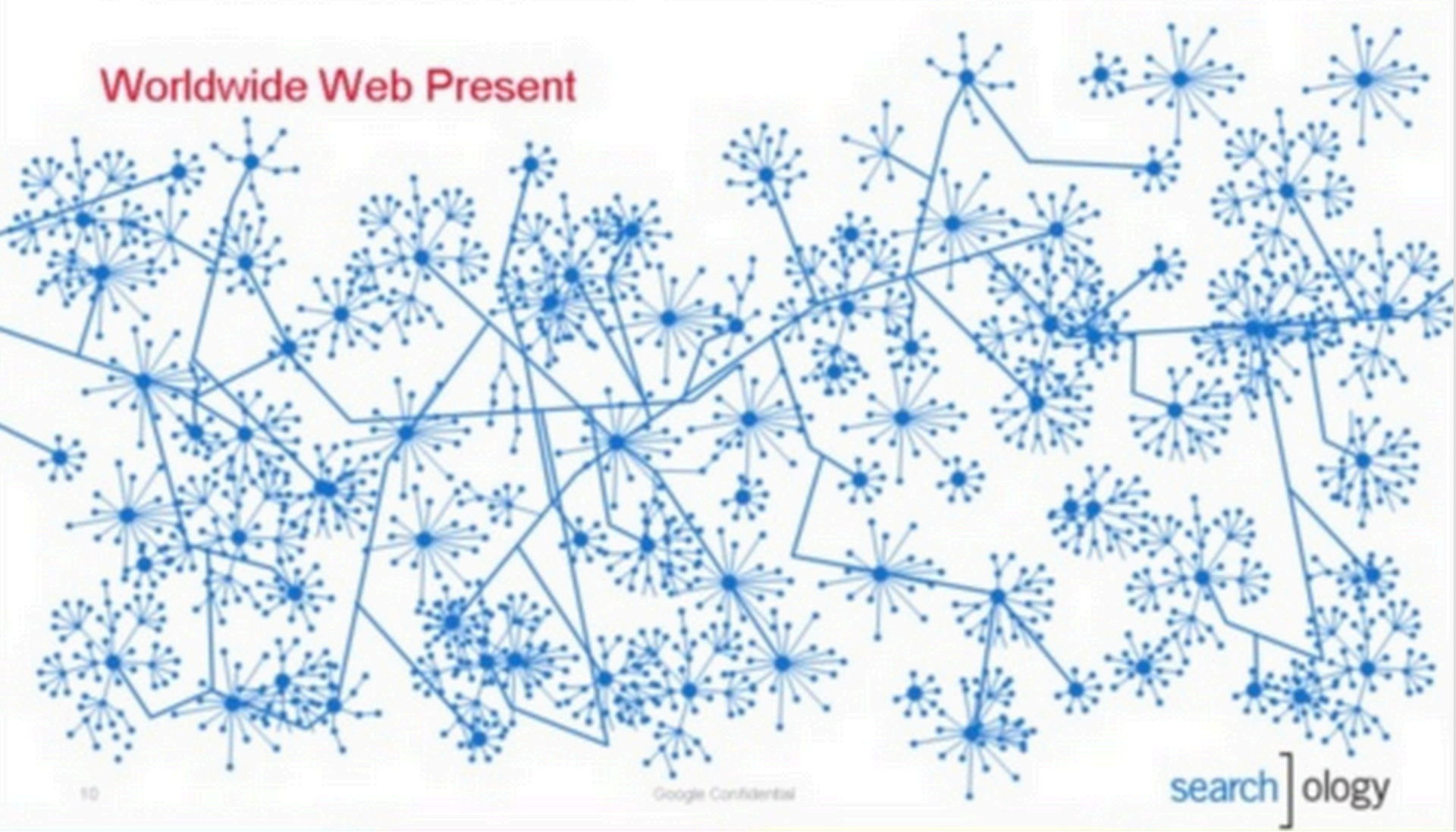
Revolução na forma como buscamos informação



A estrutura da Web em 2004
Fonte: Hal Burch and Bill Cheswick



(Parte d)A estrutura dos documentos da Web em 1998
Fonte: Google, Searchology, 2007

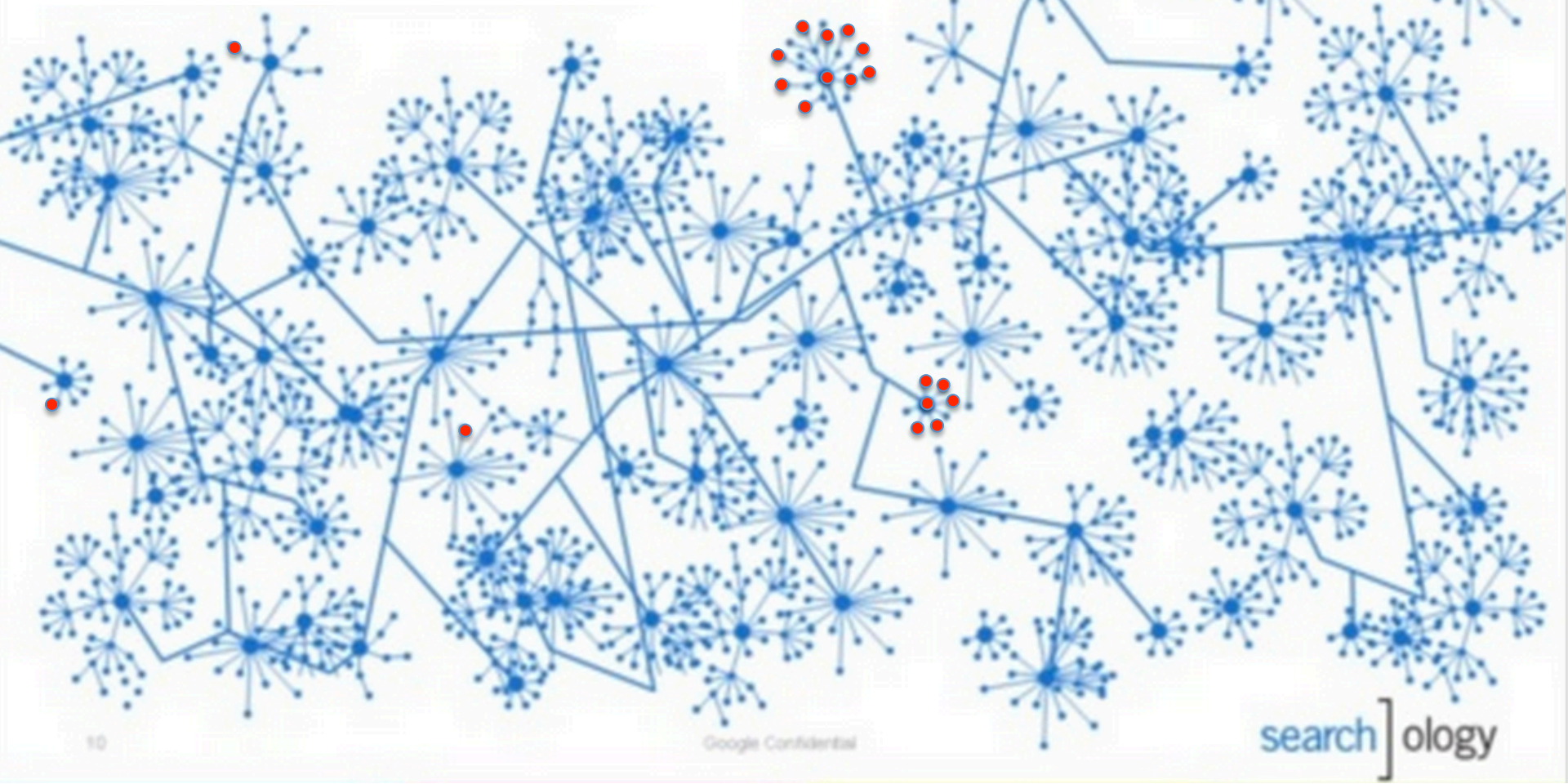


(Parte d)A estrutura dos documentos da Web em 2007
Fonte: Google, Searchology, 2007

Instituto de Computação

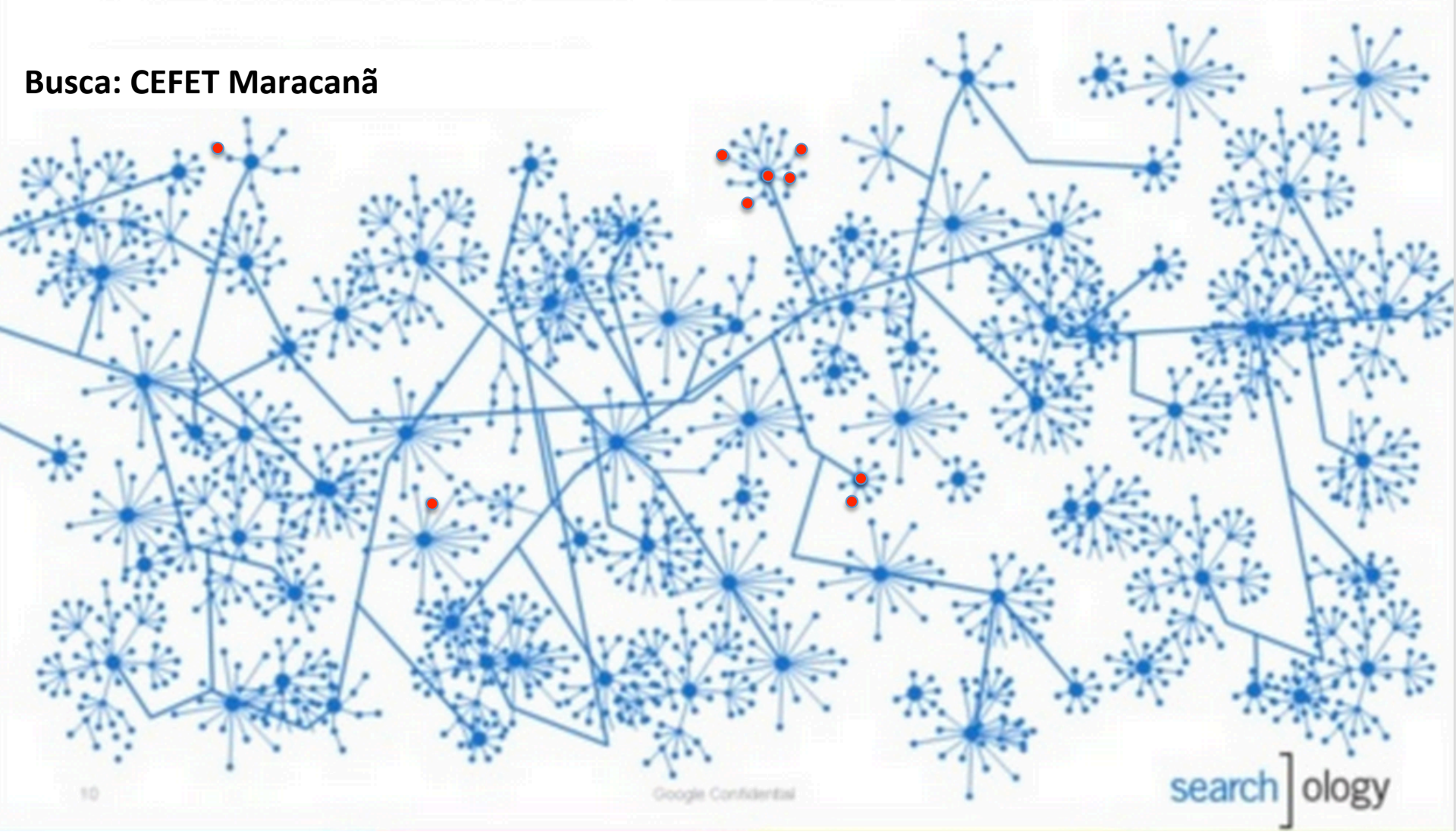
Como encontrar conteúdo?

Busca: CEFET



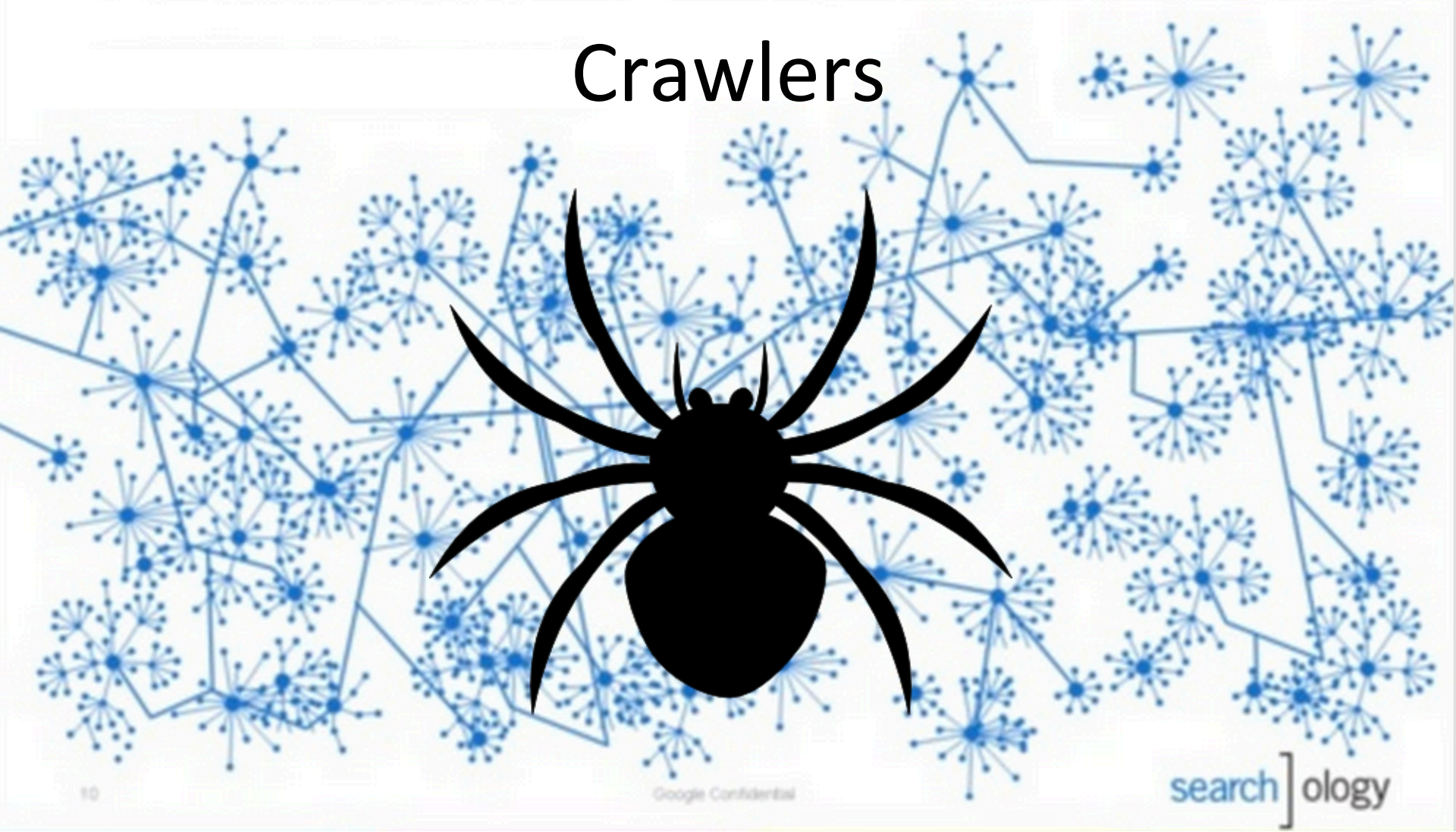
(Parte d)A estrutura dos documentos da Web em 2007
Fonte: Google, Searchology, 2007

Busca: CEFET Maracanã



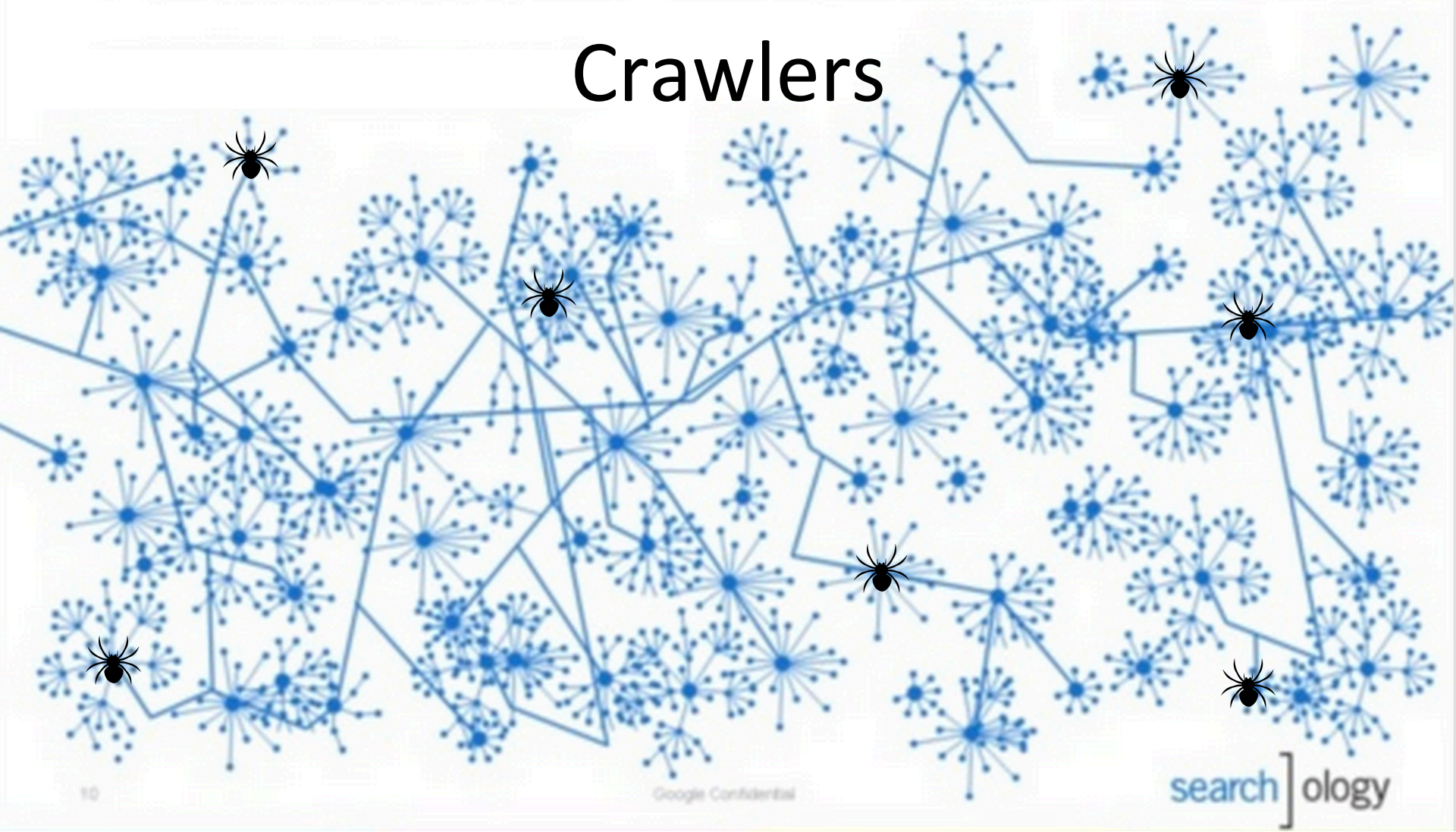
(Parte d)A estrutura dos documentos da Web em 2007
Fonte: Google, Searchology, 2007

Crawlers



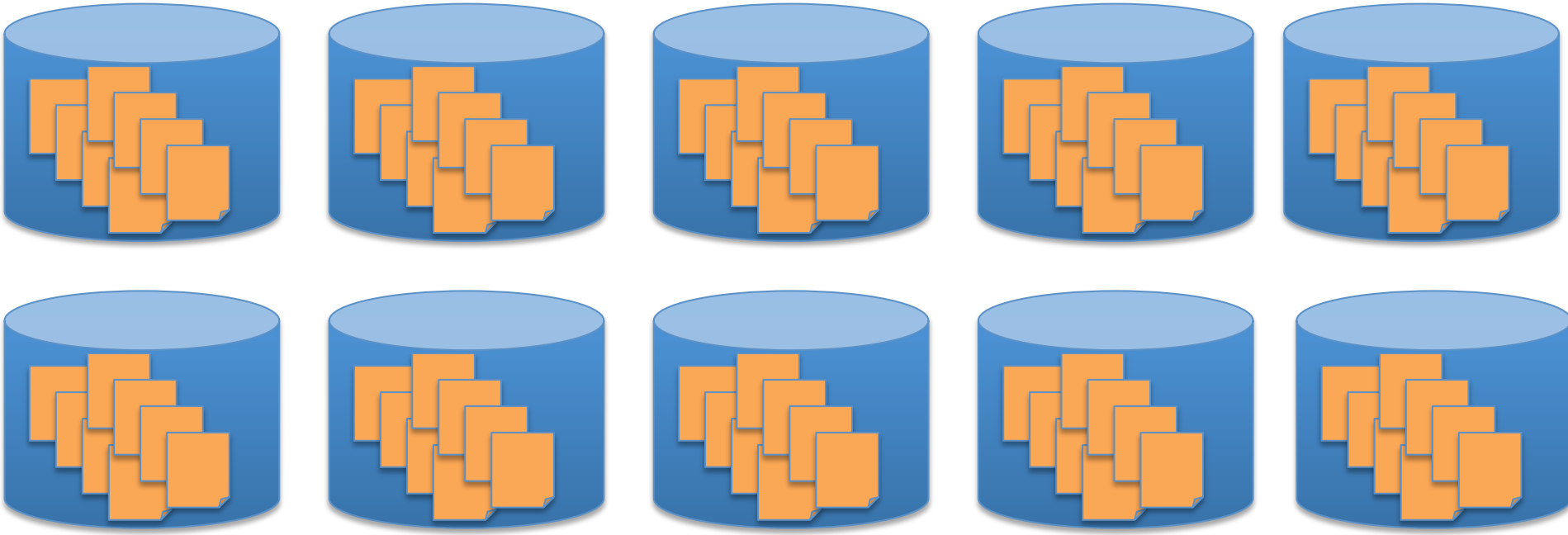
(Parte d)A estrutura dos documentos da Web em 2007
Fonte: Google, Searchology, 2007

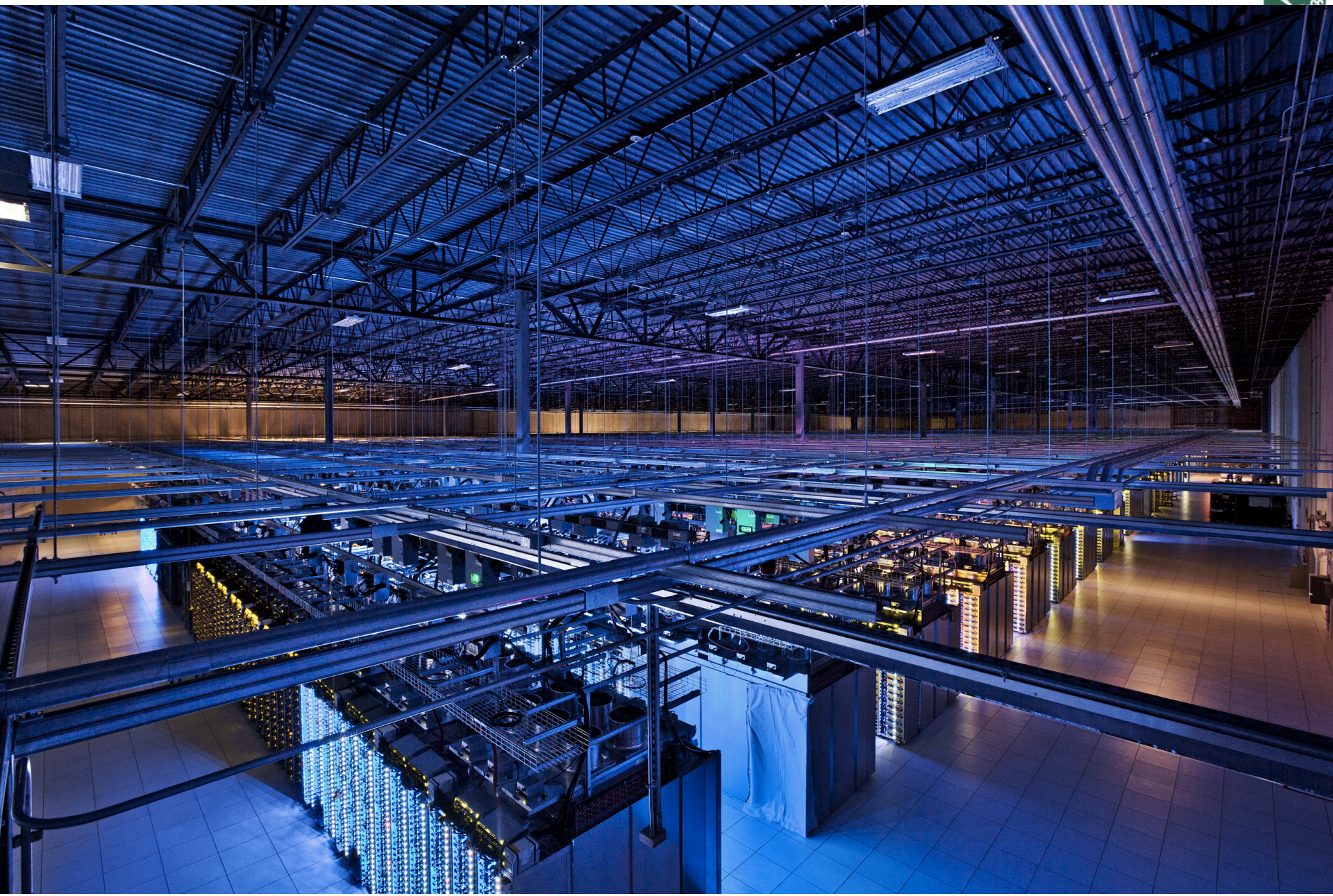
Crawlers



(Parte d)A estrutura dos documentos da Web em 2007
Fonte: Google, Searchology, 2007

Cópias da Web!





Índice

Exemplo de Documentos a serem indexados

Doc1	A casa amarela é bonita
Doc2	O carro amarelo está amassado
Doc3	Os carros são velozes
Doc4	A febre amarela é uma doença

Passo 1

- Palavras que não possuem significado (artigos, preposições) são eliminadas (**remoção de *stop words***)

Doc1	casa amarela é bonita
Doc2	carro amarelo está amassado
Doc3	carros são velozes
Doc4	febre amarela é doença

Passo 2

- Cada palavra é reduzida ao seu radical
- Se não fizermos isso, ao procurar pela palavra "carro", o usuário não encontraria o Doc3

Doc1	casa amarela é bonita
Doc2	carro amarelo está amassado
Doc3	carros são velozes
Doc4	febre amarela é doença

Depois da radicalização (exemplo simplificado)

Doc1	casa amarel ser bonit
Doc2	carro amarel estar amassad
Doc3	carro ser veloz
Doc4	febre amarel ser doença

Doc1	casa amarel ser bonit
Doc2	carro amarel estar amassad
Doc3	carro ser veloz
Doc4	febre amarel ser doença

Passo 3

- Construir o índice invertido

amarel	3	(Doc1) (Doc2) (Doc 4)
amassad	1	(Doc 2)
bonit	1	(Doc1)
casa	1	(Doc1)
carro	2	(Doc2) (Doc3)
doença	1	(Doc4)
estar	1	(Doc2)
febre	1	(Doc4)
ser	3	(Doc1) (Doc3) (Doc4)
veloz	1	(Doc3)

Consultas

- As consultas passam pelo mesmo processo de eliminação de stop words e radicalização
- Depois, basta ir direto ao índice

Exemplo

- Consulta: “carros AND amarelos”

amarel	3	(Doc1) (Doc2) (Doc 4)
amassad	1	(Doc 2)
bonit	1	(Doc1)
casa	1	(Doc1)
carro	2	(Doc2) (Doc3)
doença	1	(Doc4)
estar	1	(Doc2)
febre	1	(Doc4)
ser	3	(Doc1) (Doc3) (Doc4)
veloz	1	(Doc3)

Doc1	casa amarel ser bonit
Doc2	carro amarel estar amassad
Doc3	carro ser veloz
Doc4	febre amarel ser doença

E a ordem das respostas?

Ranking

- Cálculo da relevância
  - Conceito de Hub e Autoridade
  - Frequência dos termos da consulta nos documentos (tf, idf)
  - ??

- Documentos são retornados em ordem decrescente de relevância

eduardo ogazawara

Web | Imagens | Mapas | Vídeos | Notícias | Mais | Ferramentas de pesquisa

Aproximadamente 81.500 resultados (0,45 segundos)

Exibindo resultados para eduardo ***ogasawara***
Em vez disso, pesquisar por eduardo ogazawara

Eduardo Soares **Ogasawara** - Plataforma Lattes - CNPq
lattes.cnpq.br/0528303491410251

| **Eduardo Ogasawara**
eogasawara.wordpress.com/
Sobre. É professor do CEFET/RJ e atualmente é líder do grupo de pesquisa em Computação Aplicada do CEFET/RJ. É bacharel em informática pela UFRJ ...

Imagens de **eduardo ogasawara** Denunciar imagens

Mais imagens para **eduardo ogasawara**

Eduardo Ogasawara - Google Scholar Citations
scholar.google.com.br/citations?user=uloLWtgAAAAJ&hl=en
Professor of Computer Science, CEFET/RJ - cefet-rj.br
Scicumulus: A lightweight cloud middleware to explore many task computing paradigm in scientific workflows. D de Oliveira, E **Ogasawara**, F Baião, M Mattoso.

dblp: **Eduardo** S. **Ogasawara**
www.informatik.uni-trier.de/.../**Ogasawara**:**Eduardo**_... Traduzir esta página
Eduardo S. **Ogasawara**, Marta Mattoso: Uma Avaliação Experimental sobre Técnicas de Indexação em Bancos de Dados Orientados a Objetos. SBBD 1999: ...

eduardo ogazawara

Web | Imagens | Mapas | Vídeos | Notícias | Mais | Ferramentas de pesquisa

Aproximadamente 81.500 resultados (0,45 segundos)

Exibindo resultados para eduardo ***ogasawara***
Em vez disso, pesquisar por eduardo ogazawara

1 **Eduardo** Soares **Ogasawara** - Plataforma Lattes - CNPq
lattes.cnpq.br/0528303491410251

2 | **Eduardo Ogasawara**
eogasawara.wordpress.com/
Sobre. É professor do CEFET/RJ e atualmente é líder do grupo de pesquisa em Computação Aplicada do CEFET/RJ. É bacharel em informática pela UFRJ ...

Imagens de **eduardo ogasawara** Denunciar imagens

Mais imagens para **eduardo ogasawara**

3 **Eduardo Ogasawara** - Google Scholar Citations
scholar.google.com.br/citations?user=uloLWtgAAAAJ&hl=en
Professor of Computer Science, CEFET/RJ - cefet-rj.br
Scicumulus: A lightweight cloud middleware to explore many task computing paradigm in scientific workflows. D de Oliveira, E **Ogasawara**, F Baião, M Mattoso.

4 dblp: **Eduardo** S. **Ogasawara**
www.informatik.uni-trier.de/.../**Ogasawara**:**Eduardo**_... Traduzir esta página
Eduardo S. **Ogasawara**, Marta Mattoso: Uma Avaliação Experimental sobre Técnicas de Indexação em Bancos de Dados Orientados a Objetos. SBBD 1999: ...

eduardo ogazawara

Web | Imagens | Mapas | Vídeos | Notícias | Mais | Ferramentas de pesquisa

Aproximadamente 81.500 resultados (0,45 segundos)

Exibindo resultados para eduardo ***ogasawara***

Em vez disso, pesquisar por eduardo ogazawara

Eduardo Soares **Ogasawara** - Plataforma Lattes - CNPq
lattes.cnpq.br/0528303491410251

| **Eduardo Ogasawara**
eogasawara.wordpress.com/
Sobre. É professor do CEFET/RJ e atualmente é líder do grupo de pesquisa em Computação Aplicada do CEFET/RJ. É bacharel em informática pela UFRJ ...

Imagens de **eduardo ogasawara** — Denunciar imagens

Mais imagens para **eduardo ogasawara**

Eduardo Ogasawara - Google Scholar Citations
scholar.google.com.br/citations?user=uloLWtgAAAAJ&hl=en
Professor of Computer Science, CEFET/RJ - cefet-rj.br
Scicumulus: A lightweight cloud middleware to explore many task computing paradigm in scientific workflows. D de Oliveira, E **Ogasawara**, F Baião, M Mattoso.

dblp: **Eduardo** S. **Ogasawara**
www.informatik.uni-trier.de/.../**Ogasawara**:**Eduardo**_... Traduzir esta página
Eduardo S. **Ogasawara**, Marta Mattoso: Uma Avaliação Experimental sobre Técnicas de Indexação em Bancos de Dados Orientados a Objetos. SBBD 1999: ...

eduardo ogazawara

Web | Imagens | Mapas | Vídeos | Notícias | Mais | Ferramentas de pesquisa

Aproximadamente 81.500 resultados (0,45 segundos)

Exibindo resultados para eduardo ***ogasawara***
Em vez disso, pesquisar por eduardo ogazawara

Eduardo Soares **Ogasawara** - Plataforma Lattes - CNPq
lattes.cnpq.br/0528303491410251

| **Eduardo Ogasawara**
eogasawara.wordpress.com/
Sobre. É professor do CEFET/RJ e atualmente é líder do grupo de pesquisa em Computação Aplicada do CEFET/RJ. É bacharel em informática pela UFRJ ...

Imagens de **eduardo ogasawara** Denunciar imagens

Mais imagens para **eduardo ogasawara**

Eduardo Ogasawara - Google Scholar Citations
scholar.google.com.br/citations?user=uloLWtgAAAAJ&hl=en
Professor of Computer Science, CEFET/RJ - cefet-rj.br
Scicumulus: A lightweight cloud middleware to explore many task computing paradigm in scientific workflows. D de Oliveira, E **Ogasawara**, F Baião, M Mattoso.

dblp: **Eduardo** S. **Ogasawara**
www.informatik.uni-trier.de/.../**Ogasawara**:**Eduardo**_... Traduzir esta página
Eduardo S. **Ogasawara**, Marta Mattoso: Uma Avaliação Experimental sobre Técnicas de Indexação em Bancos de Dados Orientados a Objetos. SBBD 1999: ...

Como funcionam as máquinas de busca?

Instituto de Computação uff

Vanessa Braganholo

vanessa@ic.uff.br