Web Crawling

(Cap. 6, Alag, 2008)

Mauricio Aniche
Mauricio De Diana

A web é...

Composta por bilhões de páginas.
- 1 trilhão de páginas em julho/08 (Google)

Dinâmica.
Cresce rápido.

... como encontrar
informação de interesse?

Web crawling

Processo automatizado.

Várias formas.

Web spiders e bots.

Por quê?

Agregação de conteúdo.

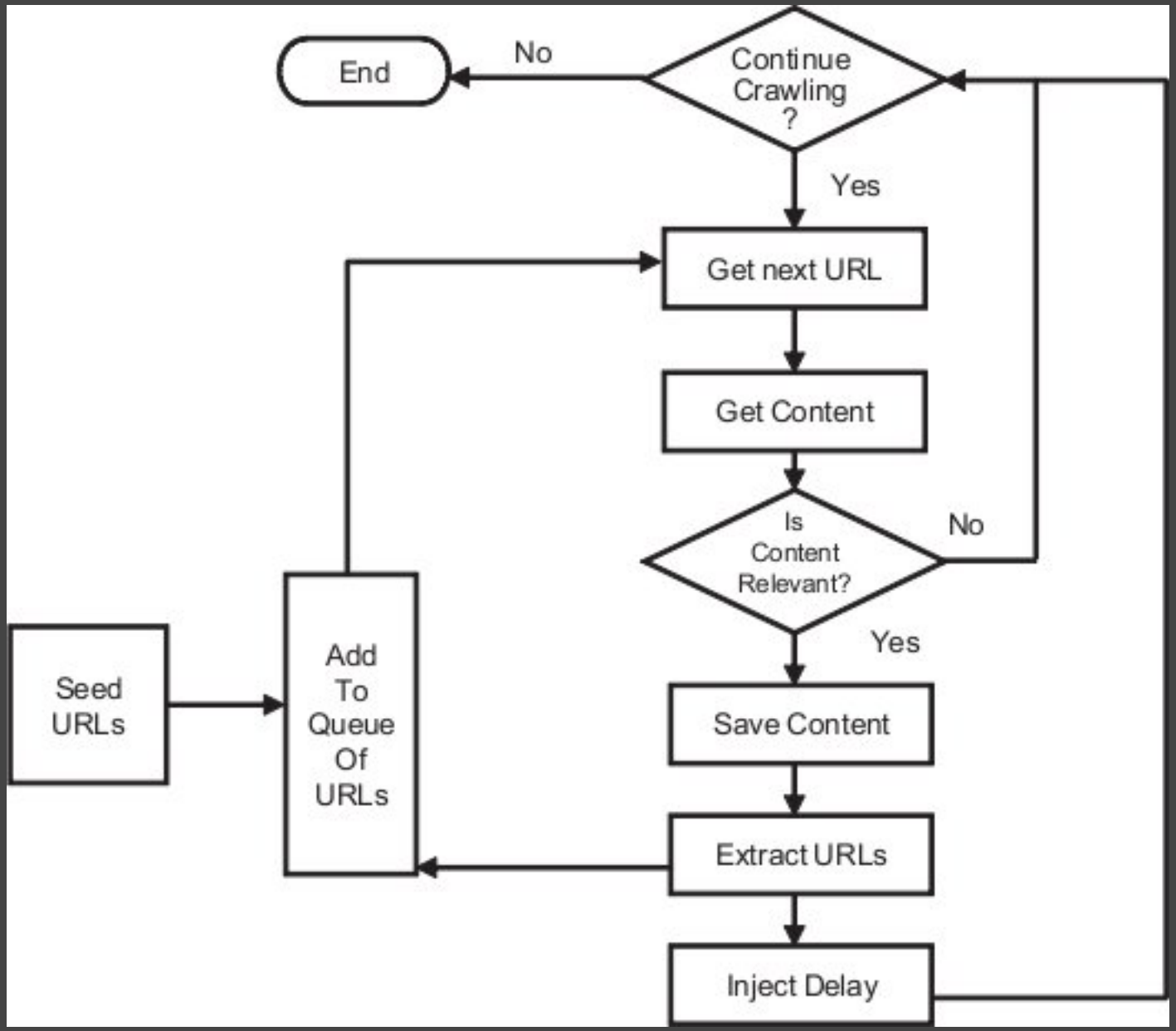
Busca de informação específica.

Disparar eventos.

Detectar links quebrados.

Procurar por quebras de direitos autorais.

Algoritmo


End
No
Continue Crawling ?
Yes
Get next URL
Get Content
Is Content Relevant?
No
Yes
Save Content
Extract URLs
Inject Delay
Seed URLs
Add To Queue Of URLs

Código

Complicações

Páginas adicionadas, alteradas e deletadas.

Mirror sites.

Spider trap.

Spammers.

Conteúdo específico.

Páginas inacessíveis.

Melhorias

Threads/Paralelismo.

Algoritmo de relevância.

Periodicidade.

Delay inteligente.

Apache Nutch

Lucene

Processamento distribuído

Crawler e searcher
Hadoop / MapReduce

Conclusão

Ideia é simples, implementações refinadas.

Implementações existentes merecem ser estudadas a fundo.

Links úteis

http://www.sitemaps.org/protocol.php

http://www.robotstxt.org/

http://www.
manageability.org/blog/stuff/open-source-web-crawlers-
java/view

Dúvidas?