KnowLib Projects: ALVIS

Anders Ardö

Anders.Ardo@it.lth.se

Digital Information Systems Group - KnowLib

Information Technology

Lund University

ALVIS

Superpeer Semantic Search Engine

(STREP EU-project 2004 – 2006)

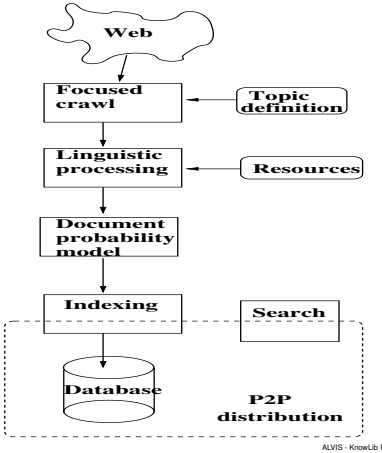
Next generation search engine

http://www.alvis.info/

ALVIS Goals

- semantic-based search engine
- natural language processing for English, French, (+ maybe Chinese, Slovenian)
- probabilistic document models
- scalable, distributed peer-to-peer search
- scientific, peer-reviewed studies of algorithms
- topic-specific customization
- support incremental growth, third-party involvement, low barrier to entry
- Open Source software development

ALVIS Overview



ALVIS Research Overview I

- design, use and interoperability of topic-specific search engines
- linguistic processing in the heart of the search engine
- probabilistic document model to support
  information retrieval (document topic, synonyms),
  categorization,
  relevance ranking
- allow users to formulate queries more easily

ALVIS Research Overview II

- advancing peer-to-peer technology
- scalable, distributed system
- query distribution and result merging
- topic-specific Web-crawling

ALVIS Partners

Finland – Helsinki Institute for Information Technology

France – Univ. Paris-Nord - Informatique, Exalead SA, INRA - Mathematique, Informatique et Genome

Switzerland – EPFL - Distributed Information Systems

Sweden – Lund Univ. - Information Technology

Denmark – DTU - CVT, IndexData Aps

Spain – ALMA Bioinformatica, S.L.

Slovenia – Josef Stefan Institute - Intelligent Systems

China – Tsinghua Univ. - Computer Science

Topic-specific Web-crawling

- **Problem**
  Construct a topic specific search-engine
  (ex. Carnivorous plants)

Topic-specific Web-crawling

- **Problem**
  Construct a topic specific search-engine
  (ex. Carnivorous plants)
- **Solution**
  Make a Web-crawler walk through Internet and colle
  all pages with topic 'Carnivorous plants'

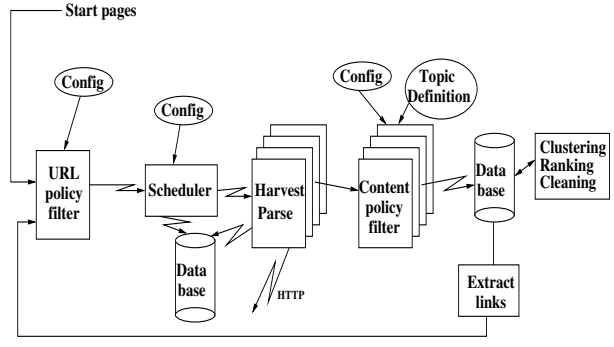
Topic-specific Web-crawling

Problem
Construct a topic specific search-engine
(ex. Carnivorous plants)

Solution
Make a Web-crawler walk through Internet and collect
all pages with topic 'Carnivorous plants'

easier said than done!

Focused Crawler



Focused crawler architecture

Automated Topic Classification

- list with topic terms
- are they present in the text?
- relevance: how many; where in the text

Automated Topic Classification

list with topic terms

are they present in the text?

relevance: how many; where in the text

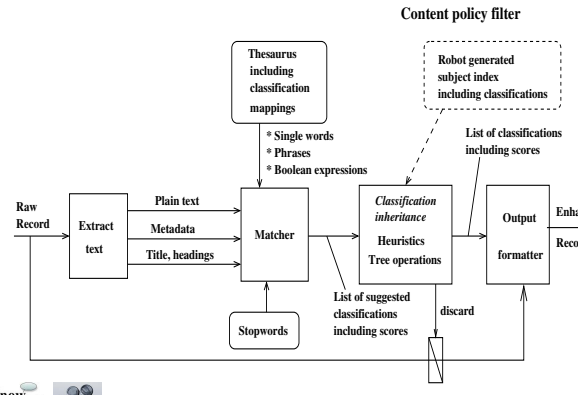
$$\text{Relevance} = \sum_{\text{terms, locations}} (hits[\text{term}_i] * weight[\text{term}_i] * weight[\text{location}_j])$$

rmalize with respect to document size

Classification precision

- Validate using topic hierarchy?
- Validate using other methods?
  - SVM, LSI, ...
  - hierarchical network of NN (one for each level in the topic hierarchy)
- Automatically detect/improve list with topic terms

Topic Filter



Conditions

Page is about Carnivorous plants
$\Longrightarrow$ automated topic classification

There are many pages on the Internet
$\Longrightarrow$ where to start?
$\Longrightarrow$ look only at interesting links
$\Longrightarrow$ take the most important pages first

Internet is Big

- Start page

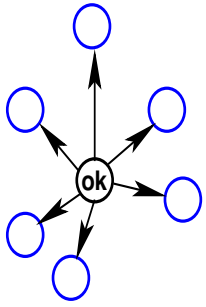
?

Internet is Big

- Start page
- OK, save

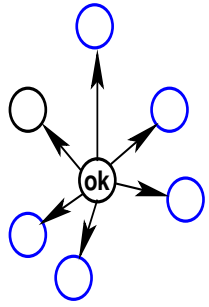
ok

Internet is Big



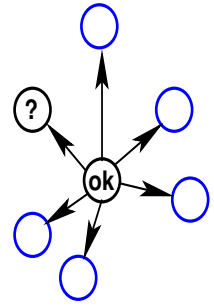
- Start page
- OK, save
- Links

Internet is Big



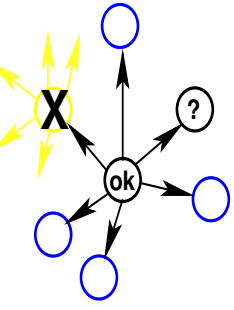
- Start page
- OK, save
- Links
- Choose

Internet is Big



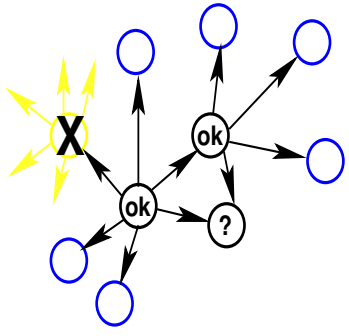
- Start page
- OK, save
- Links
- Choose
- Page OK?

Internet is Big



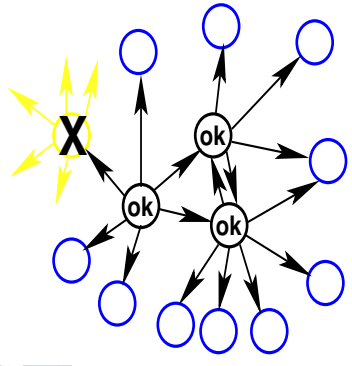
- Start page
- OK, save
- Links
- Choose
- Page OK?
- New page

Internet is Big



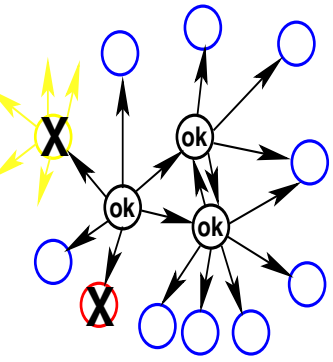
- Start page
- OK, save
- Links
- Choose
- Page OK?
- New page
- Page OK?

Internet is Big



- Start page
- OK, save
- Links
- Choose
- Page OK?
- New page
- Page OK?
- Save

Internet is Big



- Start page
- OK, save
- Links
- Choose
- Page OK?
- New page
- Page OK?
- Save
- New page

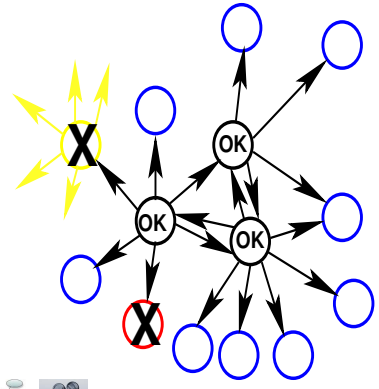
Basic Algorithm

Add links to good start pages
Start:

- Choose a page among links
- Page OK?
  - Save page
  - Add all links
- Go to Start

Save (database):
All relevant pages (search engine)
All analyzed pages
All new links

Problems



- Which new page? $\Longrightarrow$ use ranked scheduling for crawling

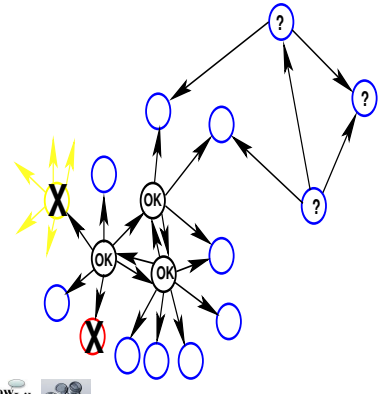
Link ranking

Topic score from automated classification
Topic score for link anchor text
Random surfer (PageRank) based ranks

$$\mathbf{x}^{(k)T} = \alpha \mathbf{x}^{(k-1)T} \mathbf{P} + (\alpha \mathbf{x}^{(k-1)T} \mathbf{a} + (1-\alpha)) \mathbf{v}^T$$

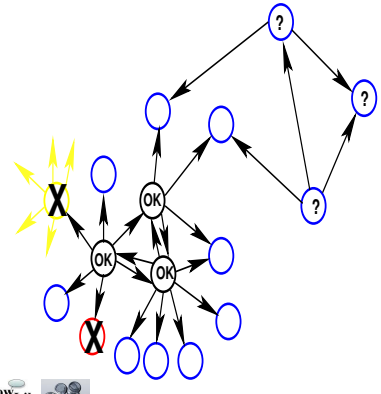
$\mathbf{x}$: PageRank vector
$\mathbf{P}$: Transition probability matrix
$\mathbf{a}$: Dangling (no outlinks) pages
α: 'teleporting' factor normally 0.85
$\mathbf{v}$: Personalization (Topic) vector

Problems



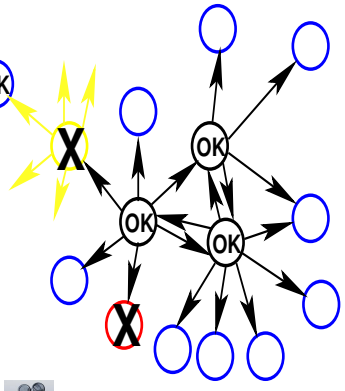
- Isolated pages

Problems



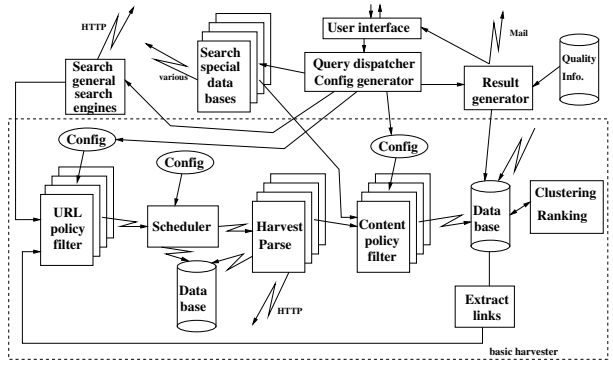
- Many pages $\Longrightarrow$ distribute crawl in a P2P or Grid fashion

Problems



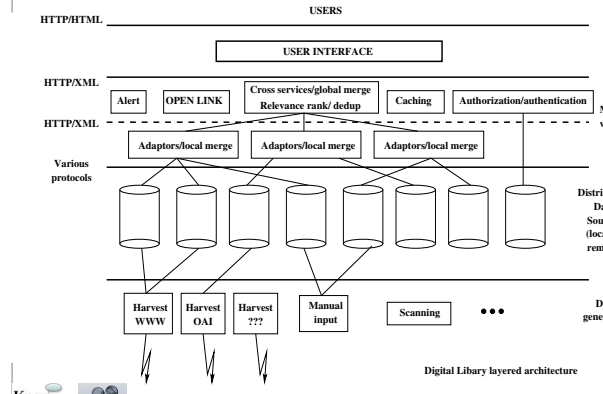
- Non relevant pages "blocking"

Interactive Agent



Semi–intelligent information search agent

Digital library architecture



Digital Libary layered architecture