

# Arsitektur Teknologi Informasi

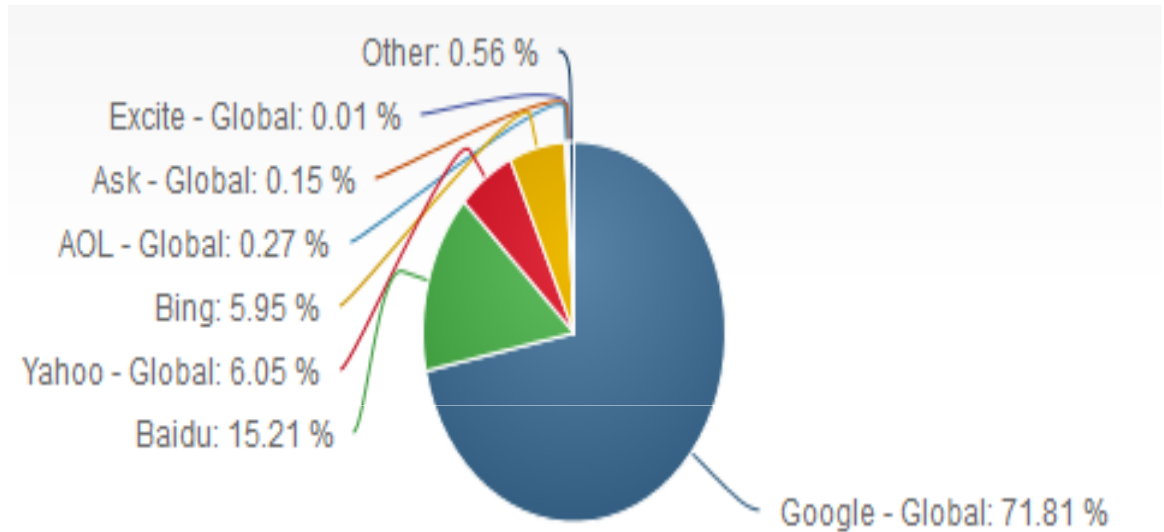
Arsitektur Search Engine & Information  
Retrieval

Antonius Rachmat C – [anton@ti.ukdw.ac.id](mailto:anton@ti.ukdw.ac.id)

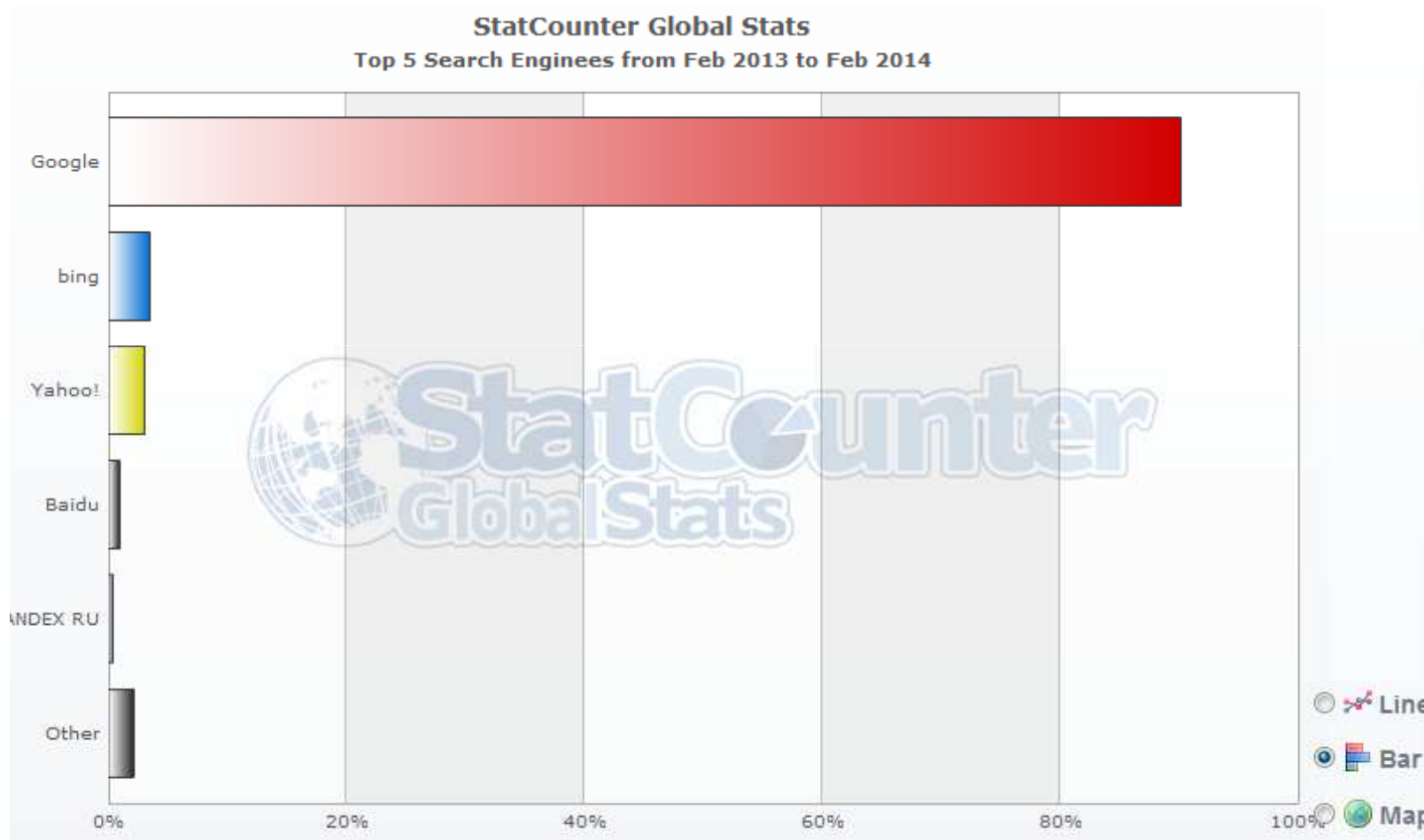
# Background

- Web data is very **large**
  - It's **dynamically** generated content
  - New pages get added all the time
    - Ex: Technorati has 50M+ blogs
    - Ex: The size of the blogosphere doubles every 6 months
- Yahoo deals with **12TB** of data per day (according to Ron Brachman)
- So we need **search engine** to search in the entire web!

# Search engine



# Search Engine Statistics Feb 2014








# Searches per Day 2010

Service	Searches Per Day	As Of/Notes
AltaVista	50 million	9/00 (as reported to me by AltaVista, for its site and queries through partners)
Inktomi	47 million	4/00 (still reflects queries from Yahoo, which will be handled by Google from July 2000).
Google	40 million	8/00 (14 million of these are at Google.com, 15 million are probably generated through Google's partnership with Yahoo, and the remainder come through Google partner sites, such as Netscape Search)
GoTo	5 million	4/00 (as reported by GoTo to a reader, who forwarded the information to me. Includes queries through affiliates and partners).
Ask Jeeves	4 million	3/00
Voila	1.5 million	1/00 (as reported to me by Voila, for its entire network of sites)

# Other SE Statistics






## Search Engine Analysis

The following report shows **search engines** for the industry '**All Categories**', ranked by **Volume of Searches** for the **4 weeks** ending **01/14/2012**.

Rank	Search Engine	Searches
1.	www.qooqle.com	62.40% 
2.	search.yahoo.com	16.13% 
3.	www.bing.com	14.64% 
4.	www.ask.com	3.88% 
5.	search.aol.com	2.43% 






## Industry Clickstream

The following report shows **downstream websites** for the industry '**Computers and Internet - Search Engines**', ranked by **Clicks** for the **week** ending **01/14/2012**.

Rank	Website	Clicks
1.	Facebook	6.29% 
2.	YouTube	3.48% 
3.	Gmail	2.36% 
4.	Wikipedia	1.46% 
5.	Yahoo! Mail	1.08% 






## Industry Rankings

The following report shows **websites** for the industry '**Computers and Internet - Search Engines**', ranked by **Visits** for the **week** ending **01/14/2012**.

Rank	Website	Visits Share
1.	Google	66.59% 
2.	Bing	10.99% 
3.	Yahoo! Search	10.90% 
4.	Ask	2.59% 
5.	AOL Search	1.63% 

## Search Engine Search Terms

The following report shows **search terms** for the search engine '**www.google.com**', ranked by **Volume of Searches** for the **4 weeks** ending **01/14/2012**.

Rank	Search Term	Volume
1.	facebook	3.11% 
2.	youtube	0.96% 
3.	yahoo	0.53% 
4.	yahoo mail	0.44% 
5.	craigslist	0.41% 

# Purpose of Search Engines

- Helping people **find** what they're looking for
  - Starts with an “information need”
  - It's convert into a query and then gets results
- SE materials are available in:
  - Web pages, documents
  - Image, Flash, Audio, Video
  - Any other format

# Examples of search engines

- Conventional (library catalog).  
Search by keyword, title, author, etc.
- Text-based (Lexis-Nexis, Google, Yahoo!).  
Search by keywords. Limited search using queries in natural language.
- Multimedia (QBIC, WebSeek, SaFe)  
Search by visual appearance (shapes, colors,... ).
- Question answering systems (Ask, Wolfram Alpha)  
Search in (restricted) natural language
- Research systems (Lemur, Nutch)
- Meta Search Engine (agregation search engine)



# What does it take to build a search engine?

- Decide what to index
- Collect it
- Index it (efficiently)
- Keep the index **up to date**
- Provide user-friendly query facilities

# Searching example

- pizza AND pepperoni AND ham AND NOT olives AND NOT garlic
- “Carilah link ke semua pages yang meliputi kata *pizza* seperti halnya kata *pepperoni* dan kata *ham*, tetapi *mengabaikan* pages yang mengandung kata *zaitun* atau kata *garlic*.”



tattoos self-esteem teens

Search

[Advanced Search](#)  
[Preferences](#)

Web

Results 1 - 1

### [Teen Tattoos & Piercings -- All About Parenting and Tweens & Teens ...](#)

"It's important for kids to forge a sense of **self**." So tread lightly and choose your battles ...  
**Teens, Tattoos**, and Piercings: More Than Meets the Eye ...  
[family.go.com/parentpedia/preteen-teen/behavior/teen-tattoos-piercings/](#) - 107k -  
[Cached](#) - [Similar pages](#)

### [For Teens - Convincing Your Parents to Let You Get a Tattoo or ...](#)

Without that acceptance, their **self-esteem** can suffer and they may carry ... For **Teens** -  
Convincing Your Parents to Let You Get a **Tattoo** or Piercing ...  
[tattoo.about.com/cs/articles/a/convinceparents.htm](#) - 26k - [Cached](#) - [Similar pages](#)

### [Teens, Tattoos and Body Piercing, Tattoo and Body Piercing](#)

Sep 25, 2006 ... while others take part to enhance their **self-esteem** and peer image. ...  
**Teens** who are considering a **tattoo** should realize the following: ...  
[www.emaxhealth.com/68/7541.html](#) - 19k - [Cached](#) - [Similar pages](#)

### [Teen Fashion, Body Piercing and Tattooing - Teen Behavior Article](#)

Body piercing, dyed hair, shaved heads and **tattoos**. Should **teen** fashion be cause ... Related  
Books. Safe **Teen** - 501 Ways to Boost Your Child's **Self-Esteem** ...  
[www.theparentreport.com/resources/ages/teen/behavior/100.html](#) - 41k -  
[Cached](#) - [Similar pages](#)



## Tattoos

What's the safest way to get a tattoo? Does it hurt? What can go wrong? ... Body Image and **Self-Esteem**. **Tattoos**. KidsHealth>**Teens**>Your Body>Skin Stuff>**Tattoos** ... [kidshealth.org/teen/your\\_body/skin\\_stuff/safe\\_tattooing.html](http://kidshealth.org/teen/your_body/skin_stuff/safe_tattooing.html) - [Cached](#)

## Behavior and Self-Esteem Articles

Articles on the behaviors, issues and trends influencing modern tweens and **teens**. ... An Expression Of Individuality - A parent's primer on piercings and **tattoos**. ... [twensandteensnews.com/archives/2006/main\\_behaviorselfesteem.php](http://twensandteensnews.com/archives/2006/main_behaviorselfesteem.php) - [Cached](#)

## Tattoos are a sign of low self-esteem - Health - Wellness - Lifestyle ...

**Tattoos** are a sign of low **self-esteem** ... Milkshakes can help anorexic **teens**. More >> More Lifestyle Stories. Infidelity On-Line ...

[timesofindia.indiatimes.com/Lifestyle/.../articleshow/4361981.cms](http://timesofindia.indiatimes.com/Lifestyle/.../articleshow/4361981.cms) - [Cached](#)

## Teenage Self-Esteem -- All About Parenting and Tweens & Teens Behavior ...

**Tattoos** & Piercings. Decoding Teen Lingo. Suggest A Topic. From Our Sponsors. What Experts Say ... 2007 Not Acceptable? **Teens** and **Self-Esteem**. 5 days ago Not ... [family.go.com/parentpedia/preteen-teen/behavior/teen-self-esteem](http://family.go.com/parentpedia/preteen-teen/behavior/teen-self-esteem) - 109k - [Cached](#)

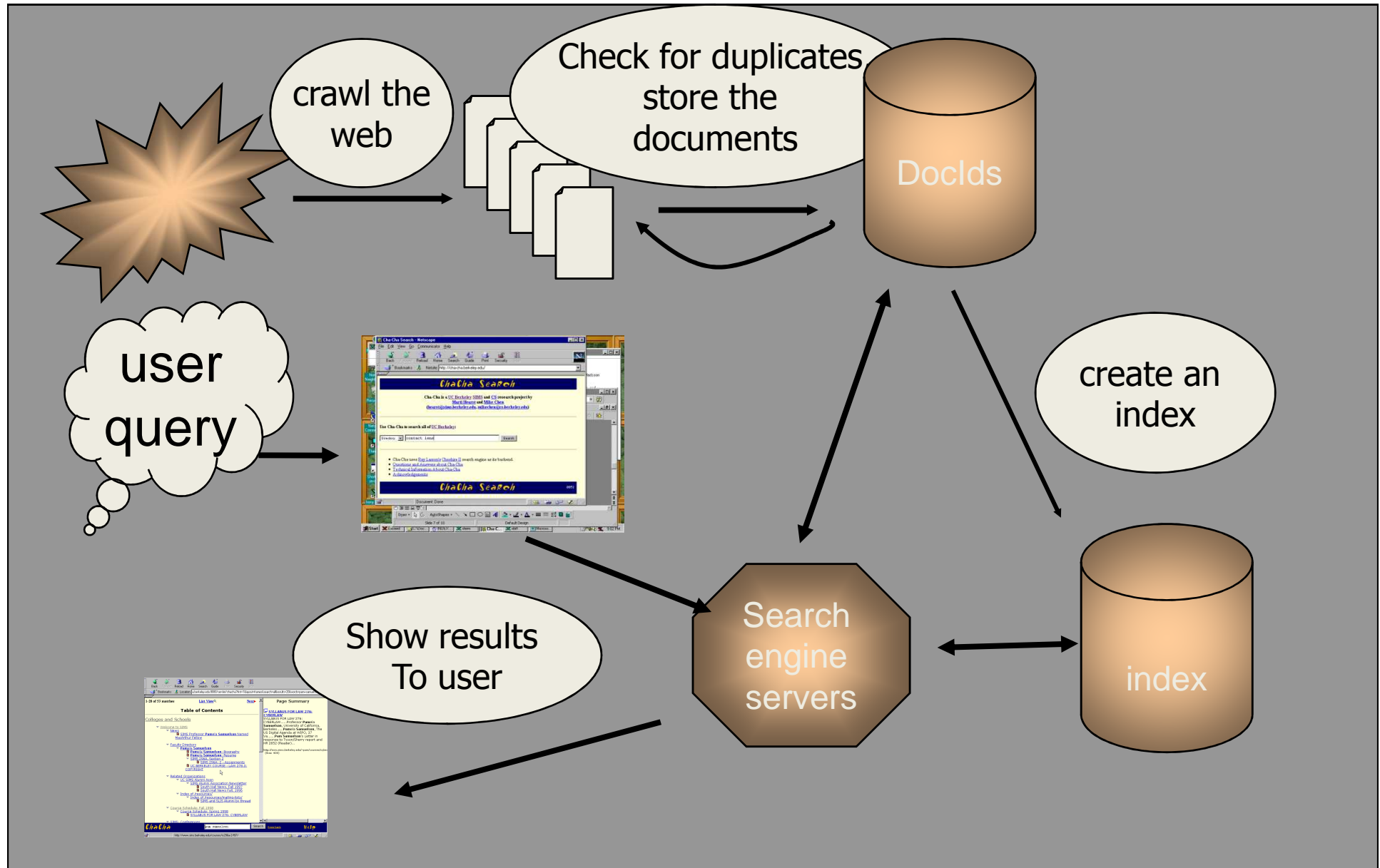
# *Why Searches Fail*

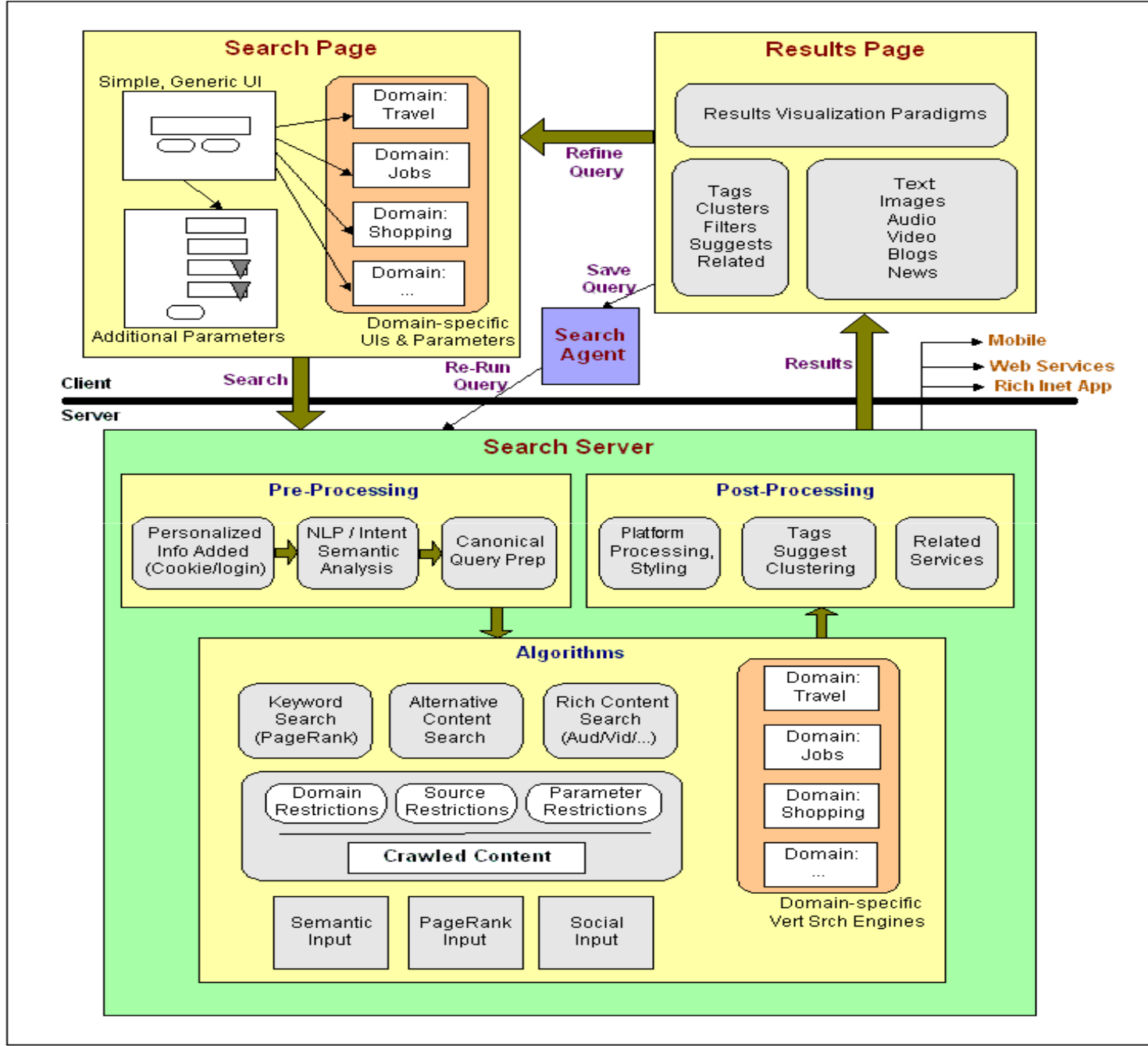
- Empty search
- Nothing on the site on that topic (scope)
- Misspelling or typing mistakes
- Vocabulary differences
- Missed search choices
- System fail

# Search engine problems

- Human maintenance
  - **Subjective**
    - Example: Ranking hits based on \$\$\$
- Automated search engines
  - **Quality of results**
- Searching process **algorithm**
  - High quality results aren't always at the top of the list

# Standard Web Search Engine Architecture





# Search Engine Architecture (2)



# Search Engine Characteristics

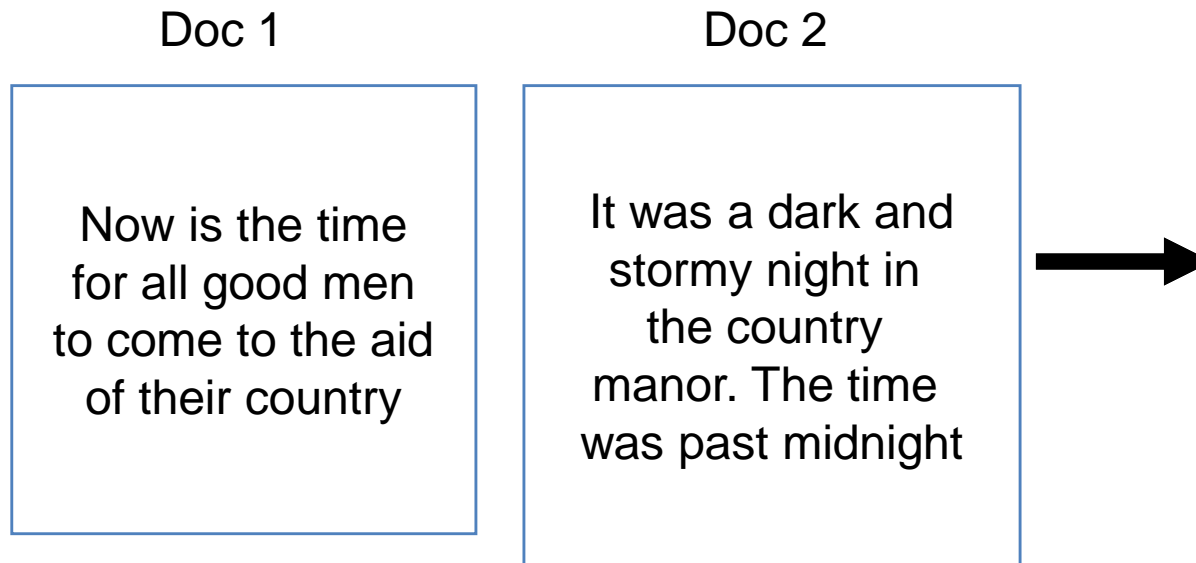
- **Unedited – anyone can not enter content**
  - Quality issues; Spam
- **Varied information types**
  - Audio, video, flash, book, brochures, catalogs, dissertations, news reports, weather, all in one place!
- **Different kinds of users**
  - Lexis-Nexis: Paying, professional searchers
  - Online catalogs: Scholars searching scholarly literature
  - Web: Every type of person with every type of goal
- **Scale**
  - Hundreds of millions of searches/day; billions of docs

# Search Engine Methods

- Page “popularity” (e.g., DirectHit)
  - Frequently visited pages (in general)
  - Frequently visited pages as a result of a query
- Link “co-citation” (e.g., Google)
  - Which sites are linked to by other sites?
  - Draws upon sociology research on bibliographic citations to identify “authoritative sources”
- Using Information Retrieval Method
  - Tokenisasi, Stemming, Stopwords, Bobot, Feature Selection, Data mining method, evaluation

# How Index files Are Created

- Periodically rebuilt
- Documents are parsed to extract tokens. These are saved with the **Document ID**.



Term	Doc #
now	1
is	1
the	1
time	1
for	1
all	1
good	1
men	1
to	1
come	1
to	1
the	1
aid	1
of	1
their	1
country	1
it	2
was	2
a	2
dark	2
and	2
stormy	2
night	2
in	2
the	2
country	2
manor	2
the	2
time	2
was	2
past	2
midnight	2

# How Index Files are Created

- After all documents have been **parsed** the index file is **sorted** alphabetically.

Term	Doc #
now	1
is	1
the	1
time	1
for	1
all	1
good	1
men	1
to	1
come	1
to	1
the	1
aid	1
of	1
their	1
country	1
it	2
was	2
a	2
dark	2
and	2
stormy	2
night	2
in	2
the	2
country	2
manor	2
the	2
time	2
was	2
past	2
midnight	2



Term	Doc #
a	2
aid	1
all	1
and	2
come	1
country	1
country	2
dark	2
for	1
good	1
in	2
is	1
it	2
manor	2
men	1
midnight	2
night	2
now	1
of	1
past	2
stormy	2
the	1
the	1
the	2
the	2
their	1
time	1
time	2
to	1
to	1
was	2
was	2

# How Index Files are Created

- Multiple term entries for a single document are **merged**.
- Within-document term **frequency** information is compiled.

Term	Doc #
a	2
aid	1
all	1
and	2
come	1
country	1
country	2
dark	2
for	1
good	1
in	2
is	1
it	2
manor	2
men	1
midnight	2
night	2
now	1
of	1
past	2
stormy	2
the	1
the	1
the	2
the	2
their	1
time	1
time	2
to	1
to	1
was	2
was	2



Term	Doc #	Freq
a	2	1
aid	1	1
all	1	1
and	2	1
come	1	1
country	1	1
country	2	1
dark	2	1
for	1	1
good	1	1
in	2	1
is	1	1
it	2	1
manor	2	1
men	1	1
midnight	2	1
night	2	1
now	1	1
of	1	1
past	2	1
stormy	2	1
the	1	2
the	2	2
their	1	1
time	1	1
time	2	1
to	1	2
was	2	2

# Storage Strategy

- The indexes are still used, even though the web is so **huge**.
- Some systems **partition** the indexes across different machines.
  - Using **distributed** database
  - Each machine handles **different** parts of the data.
  - Other systems **duplicate** the data across many machines; queries are distributed among the machines.
  - Most do a combination of these.

# Ranking system: **Link Analysis**

- Assumptions:
  - If the pages **pointing** to this page are **good**, then this is also a **good** page
  - The **words** on the links pointing to this page are useful indicators of what this page is **about**
- Why does this work?
  - The official Toyota site will be linked to by lots of other official (or high-quality) sites
  - The best Toyota fan-club site probably also has many links pointing to it

# Google Page Rank

- Sebuah situs akan semakin **populer** jika semakin banyak situs lain yang meletakkan **link** yang mengarah ke situsnya, dengan asumsi isi/content situs tersebut lebih berguna dari isi/content situs lain.
- PageRank dihitung dengan skala **1-10**.
- Sebuah halaman juga akan menjadi semakin penting jika halaman lain yang memiliki rangking (**pagerank**) **tinggi** mengacu ke halaman tersebut.
- Jika sebuah situs yang mempunyai Pagerank 9 akan di urutkan lebih **dahulu** dalam list pencarian Google daripada situs yang mempunyai Pagerank 8 dan kemudian seterusnya yang lebih kecil.

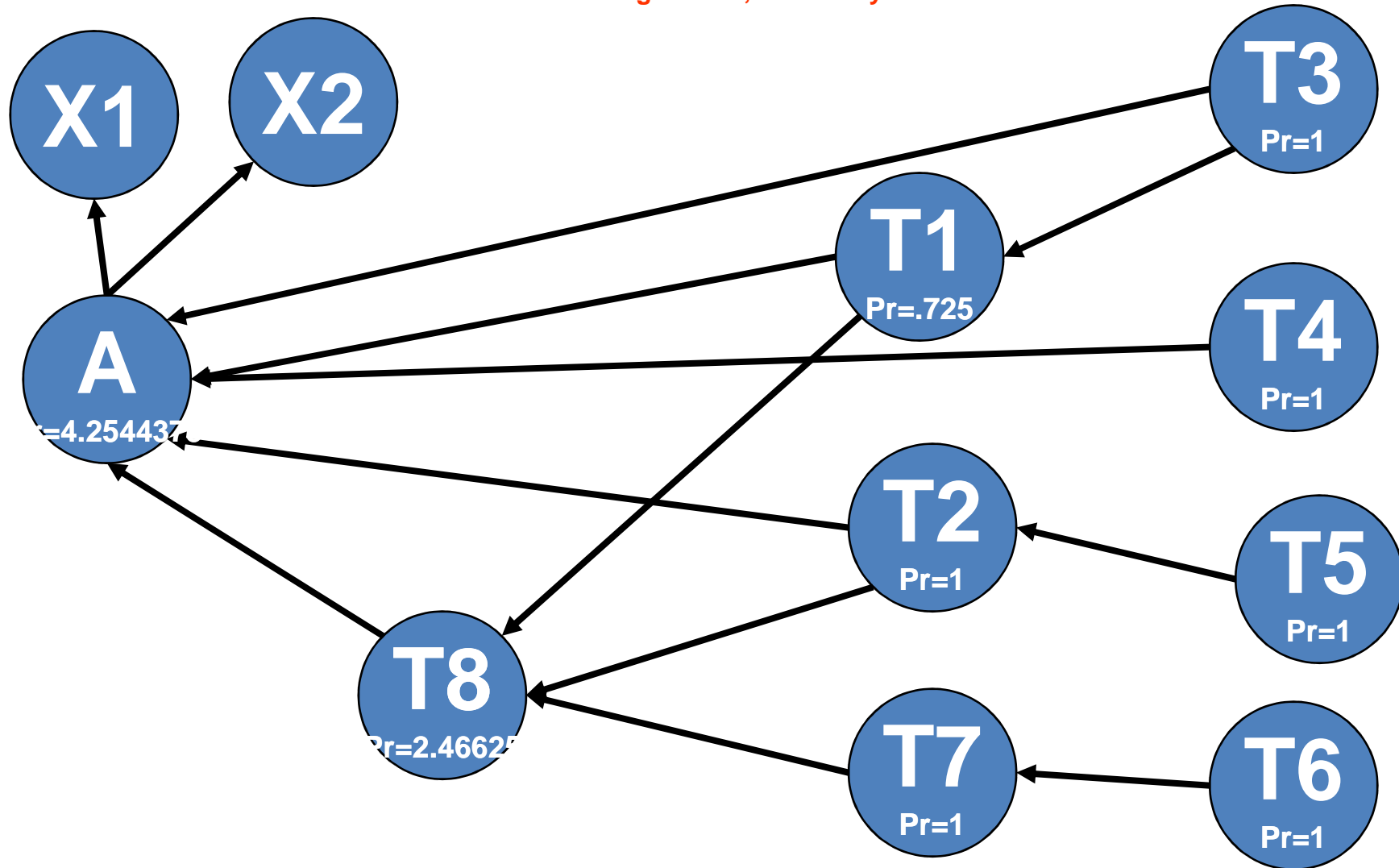


# Google Page Rank

- We assume **page A** has **pages T1...Tn** which point to it (i.e., are citations).
- **The parameter d is a damping factor** which can be set between 0 and 1. d is usually set to 0.85.
- C(A) is defined as the **number of links going out of page A**.
- **$PR(A) = (1-d) + d (PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn))$**
- So to be included in Google's top ranked results a page:
  - must have lots of votes from outside
  - votes cast by pages that have received many votes of their own.

# PageRank

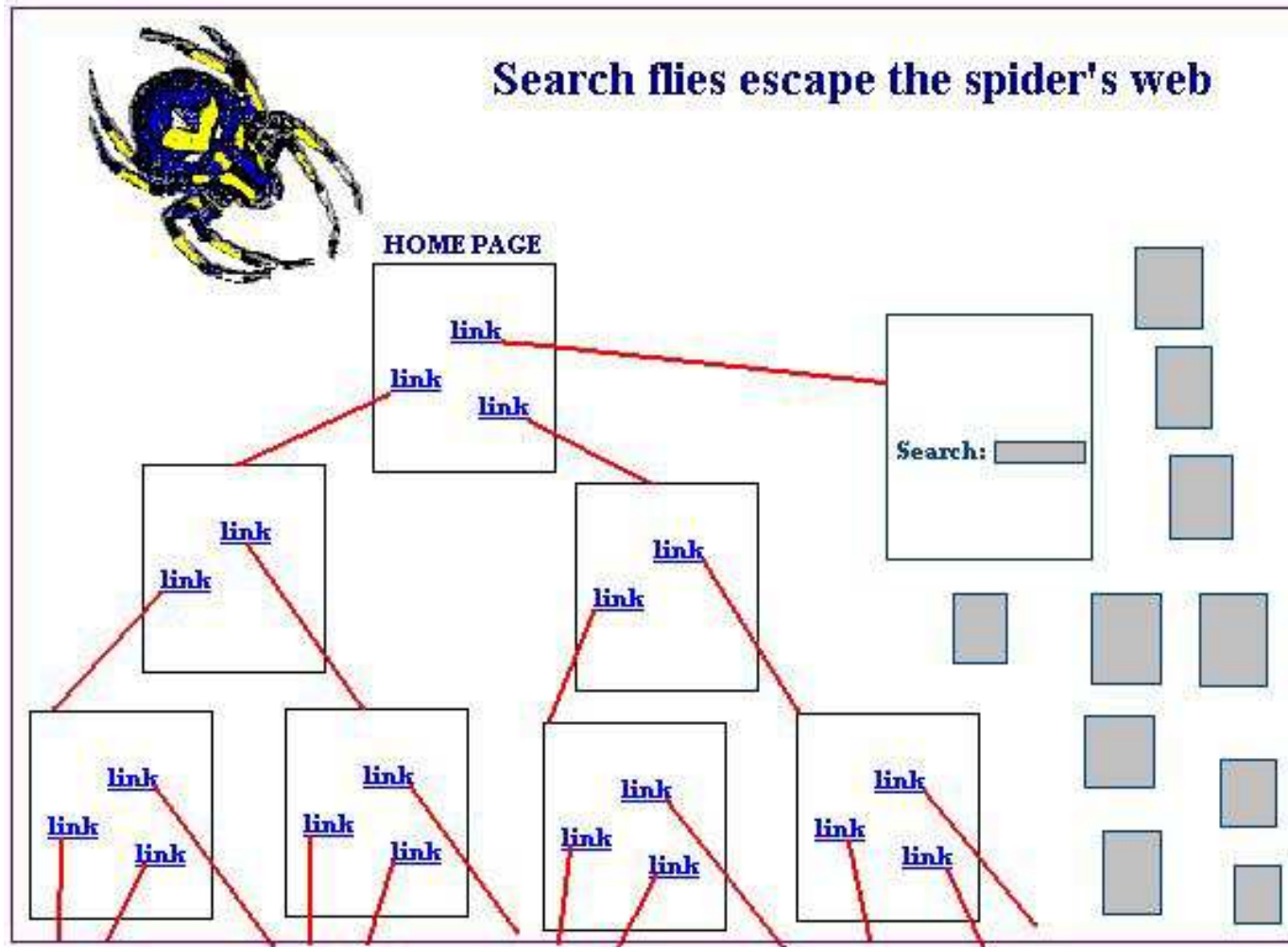
Note: these are not real PageRanks, since they include values  $\geq 1$



# Web Crawlers

- How do the web search engines get all of the items they index? **Automatic Web Crawlers**
- **Main idea:**
  - Start with **known** sites
  - Record information for **these sites**
  - **Follow the links** from each site
  - Record information found at **new sites**
  - **Repeat**
- 1 doc per minute per crawling server

# Crawler Indexing Diagram



# What the Index Needs

- Basic information for document or record
  - File name / URL / record ID
  - Title or equivalent
  - Keywords
  - Size, date, MIME type
- Full text of item
- More metadata
  - Product name, picture ID
  - Category, topic, or subject
  - Other attributes, for relevance ranking and display

# Web Crawling Issues

- **Keep out signs**
  - A file called `norobots.txt` / `robots.txt`
  - Figure out which pages change often, and recrawl these often.
- **Duplicates, virtual hosts, etc.**
  - Convert page contents with a hash function
  - Compare new pages to the hash table
- **Lots of problems**
  - Server unavailable; incorrect html; missing links;

# Robots.txt

- Protocol for giving spiders (“robots”) **limited** access to a website, originally from 1994
  - [www.robotstxt.org/wc/norobots.html](http://www.robotstxt.org/wc/norobots.html)
- Website announces its request on what **can(not)** be crawled
  - For a URL, create a file `URL/robots.txt`
  - This file specifies access **restrictions**

# Contoh

```
# robots.txt for http://www.example.com/

User-agent: *
Disallow: /cyberworld/map/ # This is an infinite virtual URL space
Disallow: /tmp/ # these will soon disappear
Disallow: /foo.html
```

```
# robots.txt for http://www.example.com/

User-agent: *
Disallow: /cyberworld/map/ # This is an infinite virtual URL space

# Cybermapper knows where to go.
User-agent: cybermapper
Disallow:
```



# The Google Search Engine

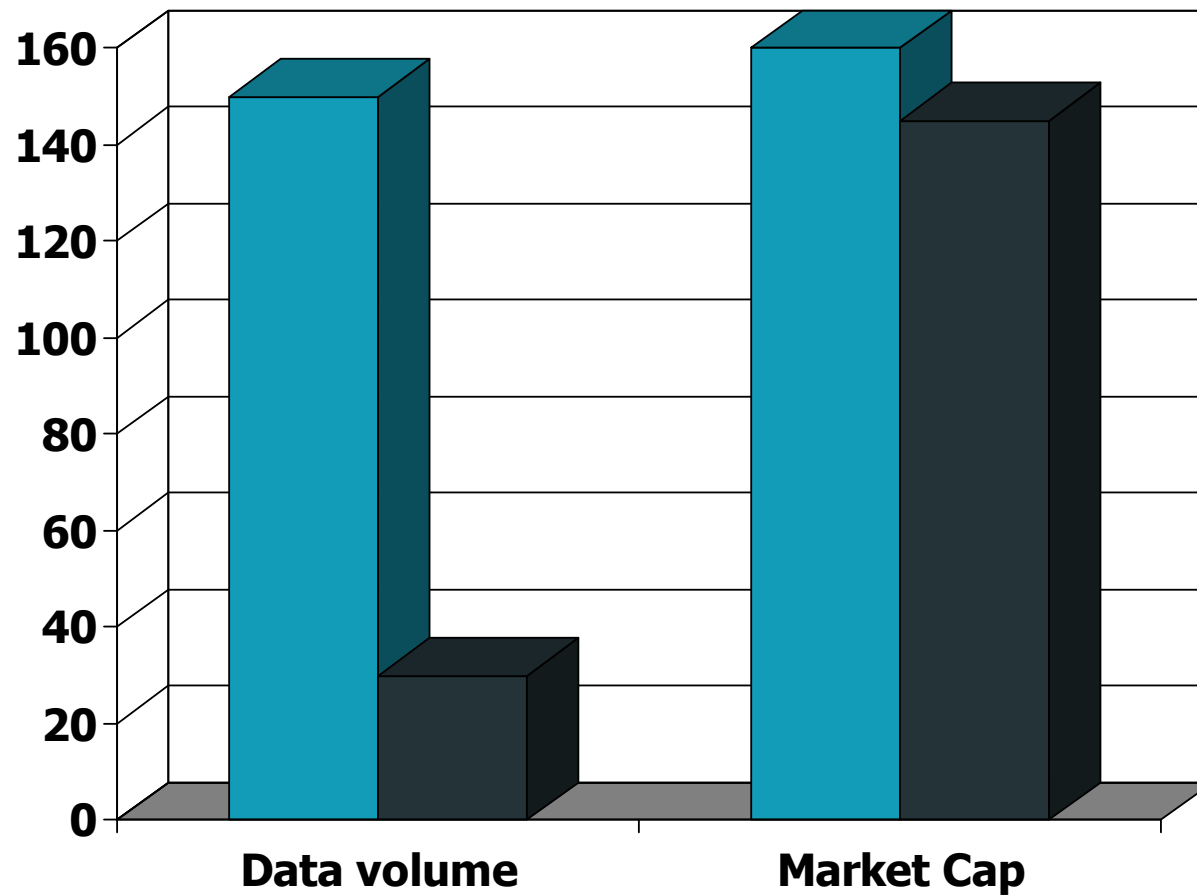
- Invented by **Larry Page** and **Sergey Brin**
- Online in **1998**
- Mission: to organize the world's information and make it universally accessible and useful
- How google finds data?
  - Using spider **GoogleBot**
  - Using site **submission**



# Information Retrieval

- Pencarian materi (biasanya dokumen) dari sesuatu yang sifatnya **tak-terstruktur** (*unstructured*, biasanya teks) untuk memenuhi kebutuhan informasi dari dalam koleksi besar (biasanya disimpan dalam komputer).
- Representasi, penyimpanan, organisasi, pencarian dan akses ke item informasi untuk memenuhi kebutuhan informasi pengguna.
- Penekanan pada proses **retrieval** informasi (bukan data).

# Unstructured (text) vs. structured (database) data in 2006



Google™

YAHOO!®

■ Unstructured  
■ Structured



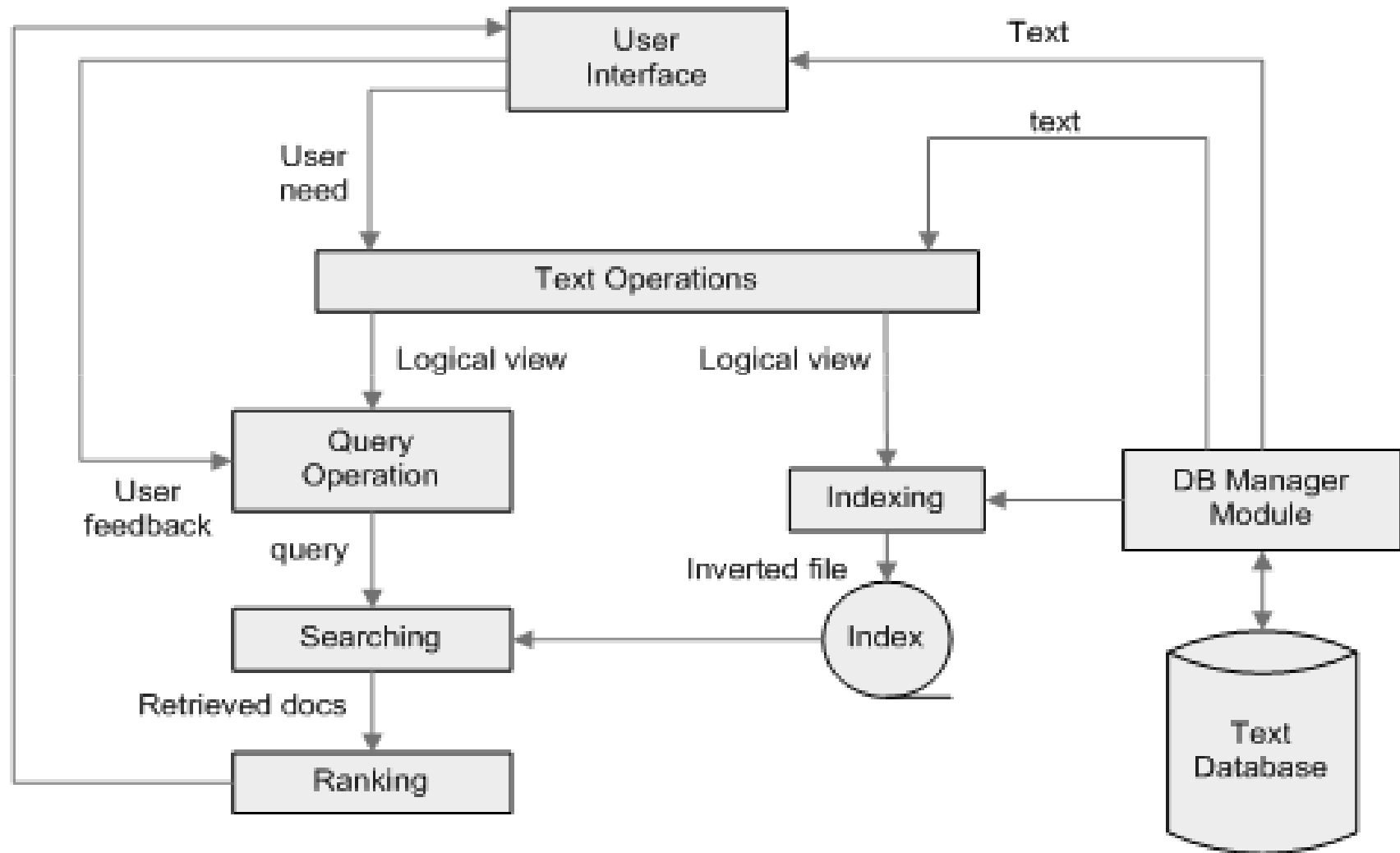
# IR vs. databases: Structured vs unstructured data

- Structured data bisa ditabelkan!

Employee	Manager	Salary
Smith	Jones	50000
Chang	Smith	60000
Ivy	Smith	50000

Typically allows numerical range and exact match  
(for text) queries, e.g., *Salary < 60000 AND Manager = Smith.*

# Sistem IR



# Main problems in IR

- Document and query indexing
  - How the best represents their contents?
- Query evaluation (or retrieval process)
- System evaluation
  - How good is a system?
  - Are the retrieved documents relevant? (**precision**)
  - Are all the relevant documents retrieved? (**recall**)

# Masalah dengan Keyword

- Mungkin tidak me-*retrieve* dokumen relevan yang menyertakan ***synonymous terms***.
  - “restaurant” vs. “café”
  - “UKDW” vs. “Universitas Kristen Duta Wacana”
- Mungkin me-*retrieve* dokumen tak-relevan yang menyertakan ***ambiguous terms***.
  - “bat” (baseball vs. mamalia)
  - “Apple” (perusahaan vs. buah-buahan)
  - “bit” (unit data vs. perilaku menggigit)

# Document indexing

- **Goal** = Find the important **meanings** and create an internal representation
- Factors to consider:
  - **Accuracy** to represent meanings (semantics)
  - Facility for computer to **manipulate**
- What is the best representation of contents?
  - **Char. string** (char trigrams): not precise enough
  - **Word**: good coverage, not precise
  - **Phrase**: poor coverage, more precise
  - **Concept**: poor coverage, precise

Coverage  
(Recall)



String

Word

Phrase

Concept

Accuracy  
(Precision)



# Bobot : tf\*idf weighting schema

- **tf = term frequency**
  - frequency of a term/keyword in a document
- **df = document frequency**
  - no. of documents containing the term
  - distribution of the term
- **idf = inverse document frequency**
  - the unevenness of term distribution in the corpus
  - the specificity of term to a document

The higher the tf, the higher the importance (weight) for the doc.

The more the term is distributed evenly, the less it is specific to a document

$$\text{weight}(t,D) = \text{tf}(t,D) * \text{idf}(t)$$

# Stopwords / Stoplist

- function words **do not bear useful** information for IR  
of, in, about, with, I, although, ...
- Stoplist: contains stop words, not to be used as index
  - Awalan, akhiran, sisipan, angka, kata sambung, kata-kata umum lainnya
- The removal of stopwords usually improves IR effectiveness
- A few “standard” stoplists are commonly used.

# Stemming

- Reason:
  - Different word forms may bear similar meaning
- Stemming:
  - Removing some **endings of word** -> **mencari kata dasar**

compute  
computes  
computing  
computed  
computation

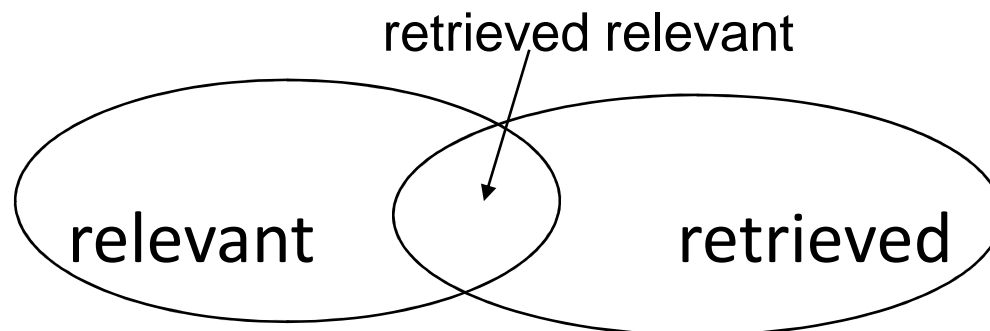
} **comput**

# IR Algorithms

- Boolean model
- Vector Space model
- Probabilistic model

# System evaluation

- Efficiency: **time, space**
- Effectiveness:
  - How is a system capable of **retrieving relevant** documents?
  - Is a system better than another one?
- Metrics often used (together):
  - Precision = retrieved relevant docs / retrieved docs
  - Recall = retrieved relevant docs / relevant docs



# Next

- Tidak ada TTS
- Jadwal Remidi saat jadwal TTS
  - Remidi TK1 dan Remidi TK2 (gantian)
- After TTS: **Business Process and Information Systems**