Introduction to Text Mining

Ali Ridho Barakbah

Knowledge Engineering Research Group
Soft Computing Laboratory
Department of Information and Computer Engineering
Electronic Engineering Polytechnic Institute of Surabaya

Definisi

- Menambang data yang berupa teks
- Sumber data biasanya didapatkan dari dokumen
- Tujuannya adalah mencari kata-kata yang dapat mewakili apa yang ada di dalam dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen

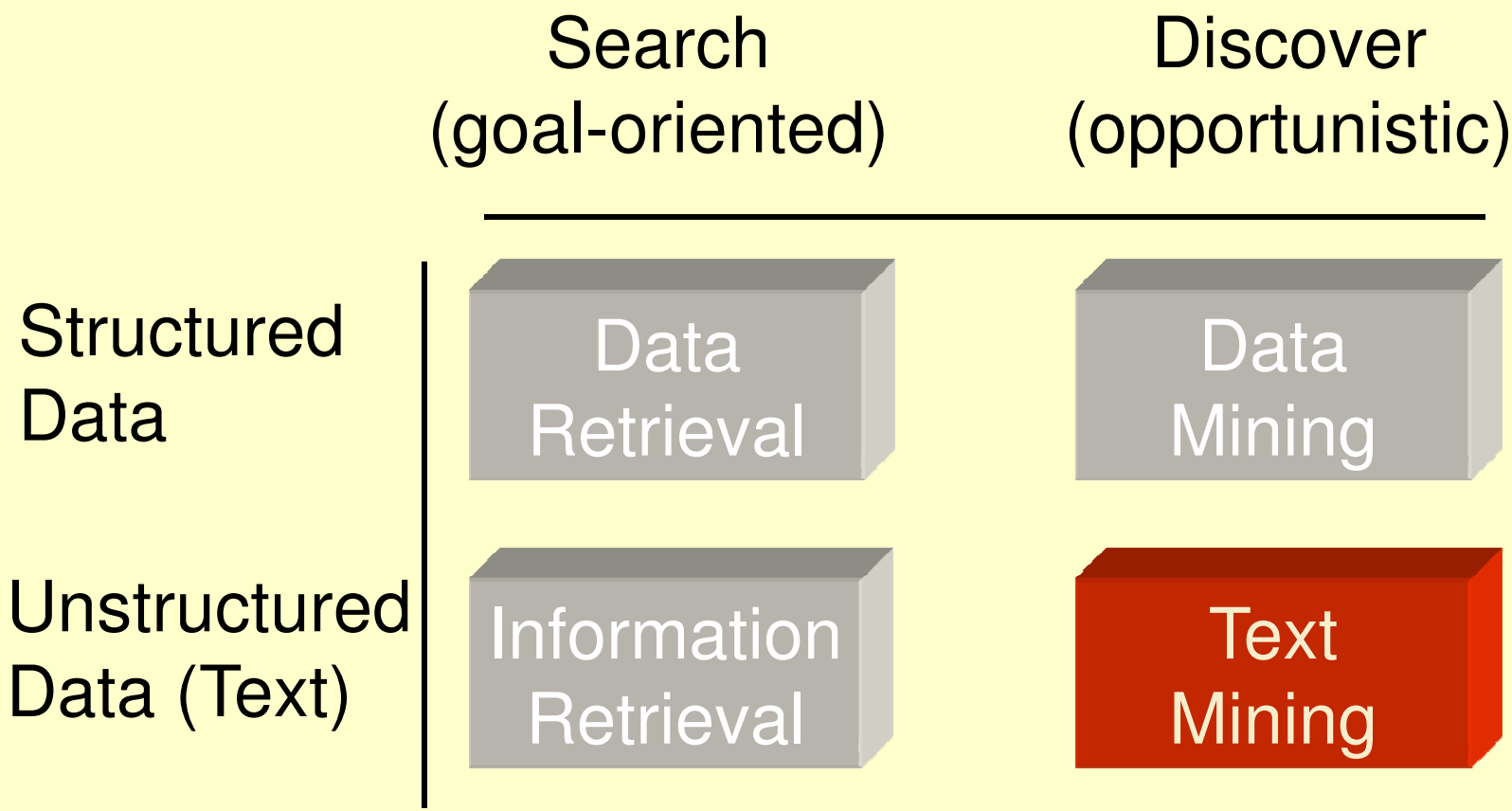
Keterkaitan Text Mining?

- Keterkaitan dengan data mining?
- Keterkaitan dengan computational linguistics?
- Keterkaitan dengan information retrieval?

	Finding Patterns	Finding "Nuggets"	
		Novel	Non-Novel
Non-textual data	General data-mining	Exploratory Data Analysis	Database queries
Textual data	Computational Linguistics		Information Retrieval

Source: Rebecca Hwa, Overview of Text Mining, 2002

"Search" versus "Discover"



© 2002, AvaQuest Inc.

Data Retrieval

- Find records within a structured database.

Database Type	Structured
Search Mode	Goal-driven
Atomic entity	Data Record
Example Information Need	"Find a Japanese restaurant in Boston that serves vegetarian food."
Example Query	"SELECT * FROM restaurants WHERE city = boston AND type = japanese AND has_veg = true"

© 2002, AvaQuest Inc.

Information Retrieval

- Find relevant information in an unstructured information source (usually text)

Database Type	Unstructured
Search Mode	Goal-driven
Atomic entity	Document
Example Information Need	"Find a Japanese restaurant in Boston that serves vegetarian food."
Example Query	"Japanese restaurant Boston" or Boston->Restaurants->Japanese

© 2002, AvaQuest Inc.

Data Mining

- Discover new knowledge through analysis of data

Database Type	Structured
Search Mode	Opportunistic
Atomic entity	Numbers and Dimensions
Example Information Need	"Show trend over time in # of visits to Japanese restaurants in Boston "
Example Query	"SELECT SUM(visits) FROM restaurants WHERE city = boston AND type = japanese ORDER BY date"

© 2002, AvaQuest Inc.

Text Mining

- Discover new knowledge through analysis of text

Database Type	Unstructured
Search Mode	Opportunistic
Atomic entity	Language feature or concept
Example Information Need	"Find the types of food poisoning most often associated with Japanese restaurants"
Example Query	Rank **diseases** found associated with "Japanese restaurants"

© 2002, AvaQuest Inc.

Challenges of Text Mining

- Very high number of possible “dimensions”
  - All possible word and phrase types in the language!!
- Unlike data mining:
  - records (= docs) are not structurally identical
  - records are not statistically independent
- Complex and subtle relationships between concepts in text
  - “AOL merges with Time-Warner”
  - “Time-Warner is bought by AOL”
- Ambiguity and context sensitivity
  - automobile = car = vehicle = Toyota
  - Apple (the company) or apple (the fruit)

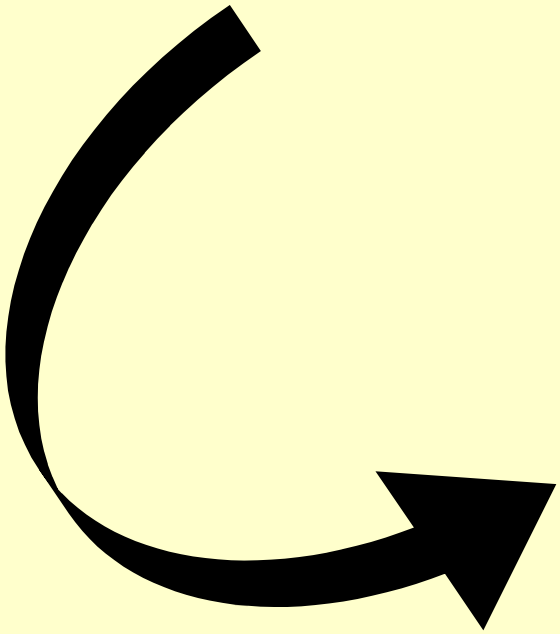
© 2002, AvaQuest Inc.

Tahapan

- Tokenizing
- Filtering
- Stemming
- Tagging
- Analyzing

Tokenizing

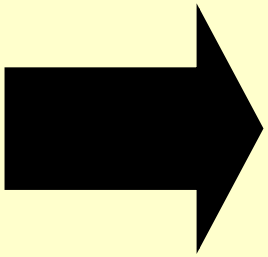
This lecture is talking about how to mine data



this
lecture
is
talking
about
how
to
mine
data

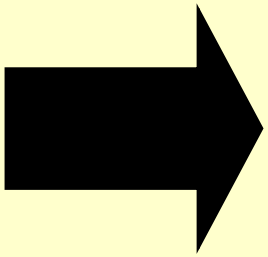
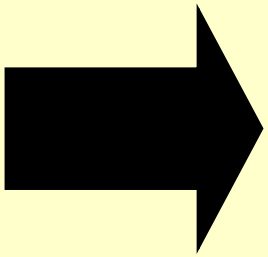
Filtering

this
lecture
is
talking
about
how
to
mine
data



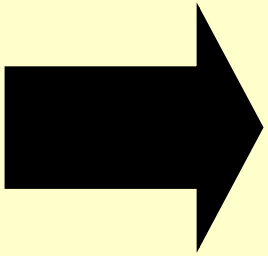
lecture
talking
mine
data

Stemming

lecture talking mine data		lecture talk mine data
reading stories		read stori

Tagging

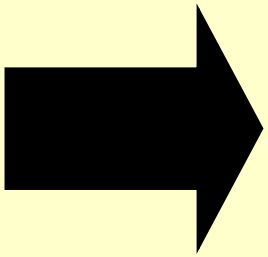
thougt was stori	think be story



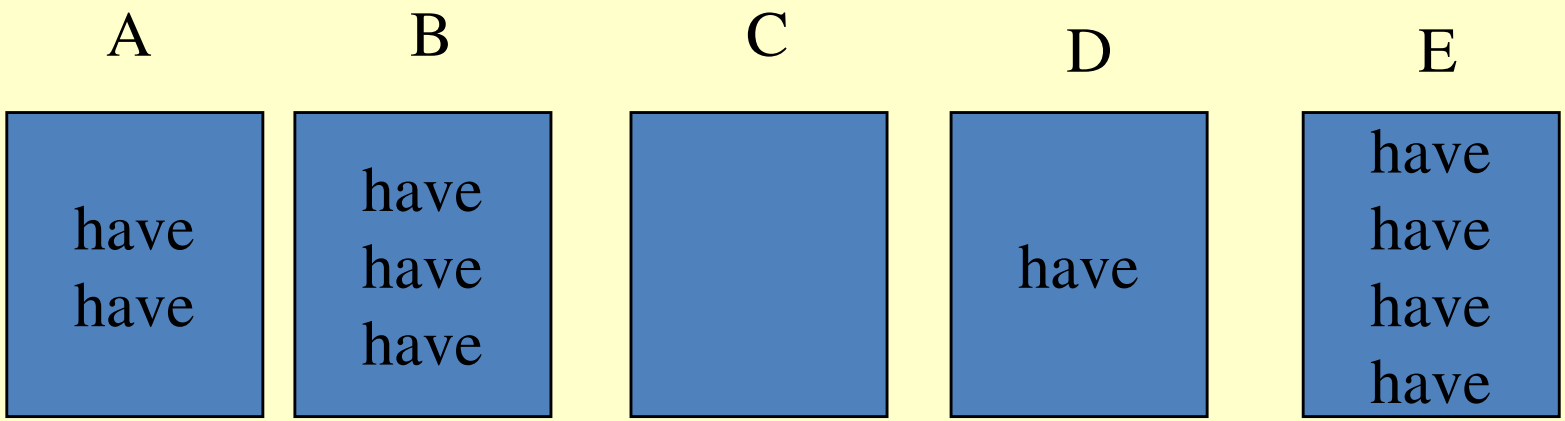
Analyzing

- Mencari seberapa jauh keterhubungan antar kata-kata antar dokumen
- Term Frequency-Inversed Document Frequency (TF-IDF) → Algoritma yang paling sederhana yang biasanya dipakai untuk scoring

lecture
talk
mine
data



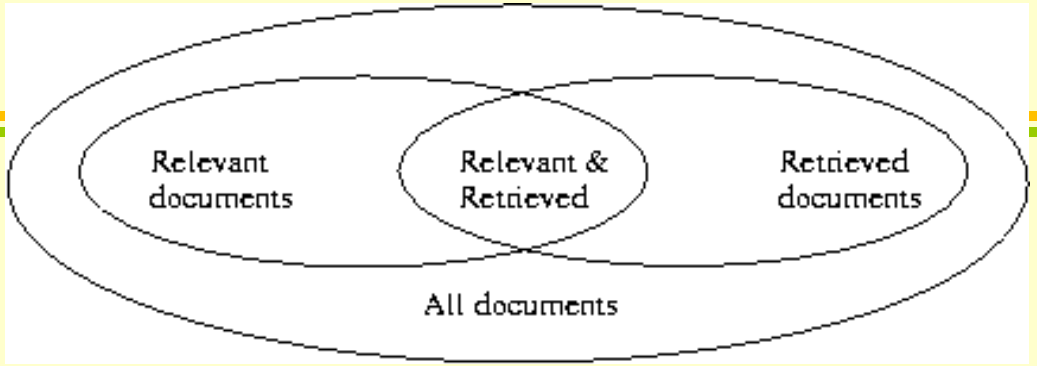
Lecture → 0.8
Talk → 0.34
Mine → 0.7
Data → 0.45



$$TFIDF_{d,t} = FREQ_{d,t}(1 + \log \frac{N}{DFREQ_t})$$

$$\text{TFIDF}_{have,B} = 3 \text{ x } (1 + \log(5 / 4))$$

Basic Measures for Text Retrieval



- Precision: the percentage of retrieved documents that are in fact relevant to the query (i.e., "correct" responses)

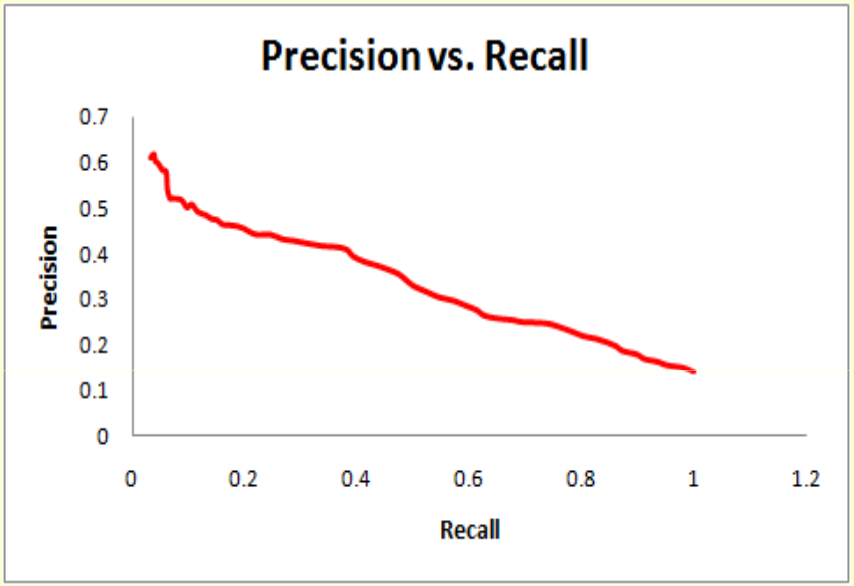
$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|}$$

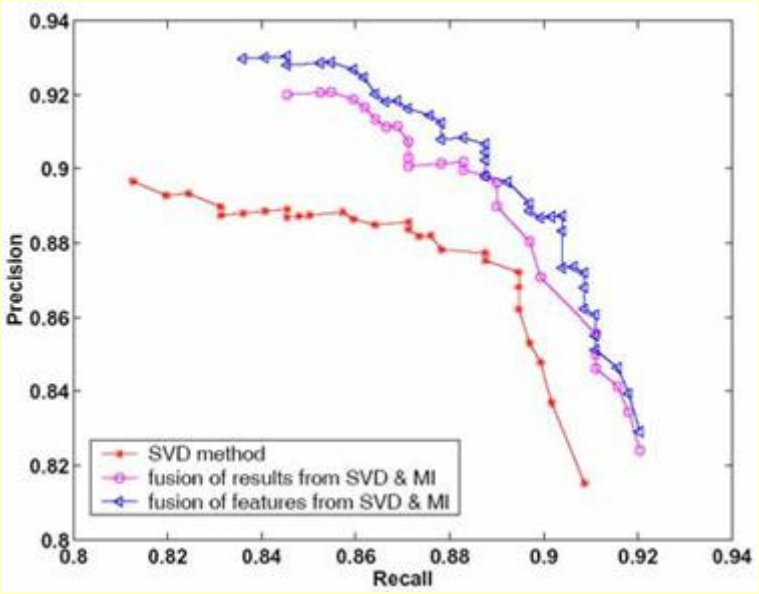
- Recall: the percentage of documents that are relevant to the query and were, in fact, retrieved

$$recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}$$

Source: Data Mining -Volinsky - 2011 - Columbia University

Precision Recall Curves





Apa itu Search Engine?

- Software code that is designed to search for information on the World Wide Web. (Wikipedia)
- Programs that search documents for specified keywordsand returns a list of the documents where the keywords were found. (Webopedia)
- Computer software used to search data (as text or a database) for specified information; also : a site on the World Wide Web that uses such software to locate key words in other sites. (Merriam Webster)

Common Characteristics

- Spider, Indexer, Database, Algorithm
- Menemukan dokumen yang tepat dan menampilkannya sesuai kondisi yang terakhir
- Proses update yang sering terhadap dokumen web pada pencarian dan membuat pemodelan terhadap dokumen
- Berusaha menyajikan hasil yang lebih presisi dibandingkan dengan kompetitor

Source: Saeed El-Darahali, Search Engines & Search Engine Optimization (SEO), 7th World Congress on the Management of e-Business


YAHOO!
Network
MAGELLAN
galaxy
LYCOS
msn
Alexa
Scrub The Web!
direct hit
HOTBOT
WEBCRAWLER
"It's that Simple."
SearchHound
infoHiway
excite
AOL
Northern Light
whatUseek
ANZWERS
AUSTRALIA & NZ
Google
Ask Jeeves
SCOUR
open directory project
ZenSearch
Netscape
DOGPILE
Mamma.com
ToggleBot!
Questel • Orbit
France Telecom Group

Timeline (full list)

Year	Engine	Current status
1993	W3Catalog	Inactive
	Aliweb	Inactive
1994	WebCrawler	Active, Aggregator
	Go.com	Active, Yahoo Search
	Lycos	Active
1995	AltaVista	Active, Yahoo Search
	Daum	Active
	Magellan	Inactive
	Excite	Active
	SAPO	Active
	Yahoo! 2008	Active, Launched as a directory
1996	Dogpile	Active, Aggregator
	Inktomi	Acquired by Yahoo!
	HotBot	Active (lycos.com)
	Ask Jeeves	Active (rebranded ask.com)

Timeline (full list)

Year	Engine	Current status
1997	Northern Light	Inactive
	Yandex	Active
1998	Goto	Inactive
	Google	Active
	MSN Search	Active as Bing
	empas	Inactive (merged with NATE)
1999	AlltheWeb	Inactive (URL redirected to Yahoo!)
	GenieKnows	Active, rebranded Yellowee.com
	Naver	Active
	Teoma	Active
	Vivisimo	Inactive
2000	Baidu	Active
	Exalead	Inactive
2002	Inktomi	Acquired by Yahoo!
2003	Info.com	Active
	Scroogle	Inactive

Source: Wikipedia

Timeline (full list)

Year	Engine	Current status
2004	Yahoo! Search	Active, Launched own web search (see Yahoo! Directory, 1995)
	A9.com	Inactive
	Sogou	Active
2005	AOL Search	Active
	Ask.com	Active
	GoodSearch	Active
	SearchMe	Inactive
2006	wikiseek	Inactive
	Quaero	Active
	Ask.com	Active
	Live Search	Active as Bing, Launched as rebranded MSN Search
	ChaCha	Active
	Guruji.com	Active as BeeMP3.com
2007	wikiseek	Inactive
	Sproose	Inactive
	Wikia Search	Inactive
	Blackle.com	Active, Google Search

Timeline (full list)

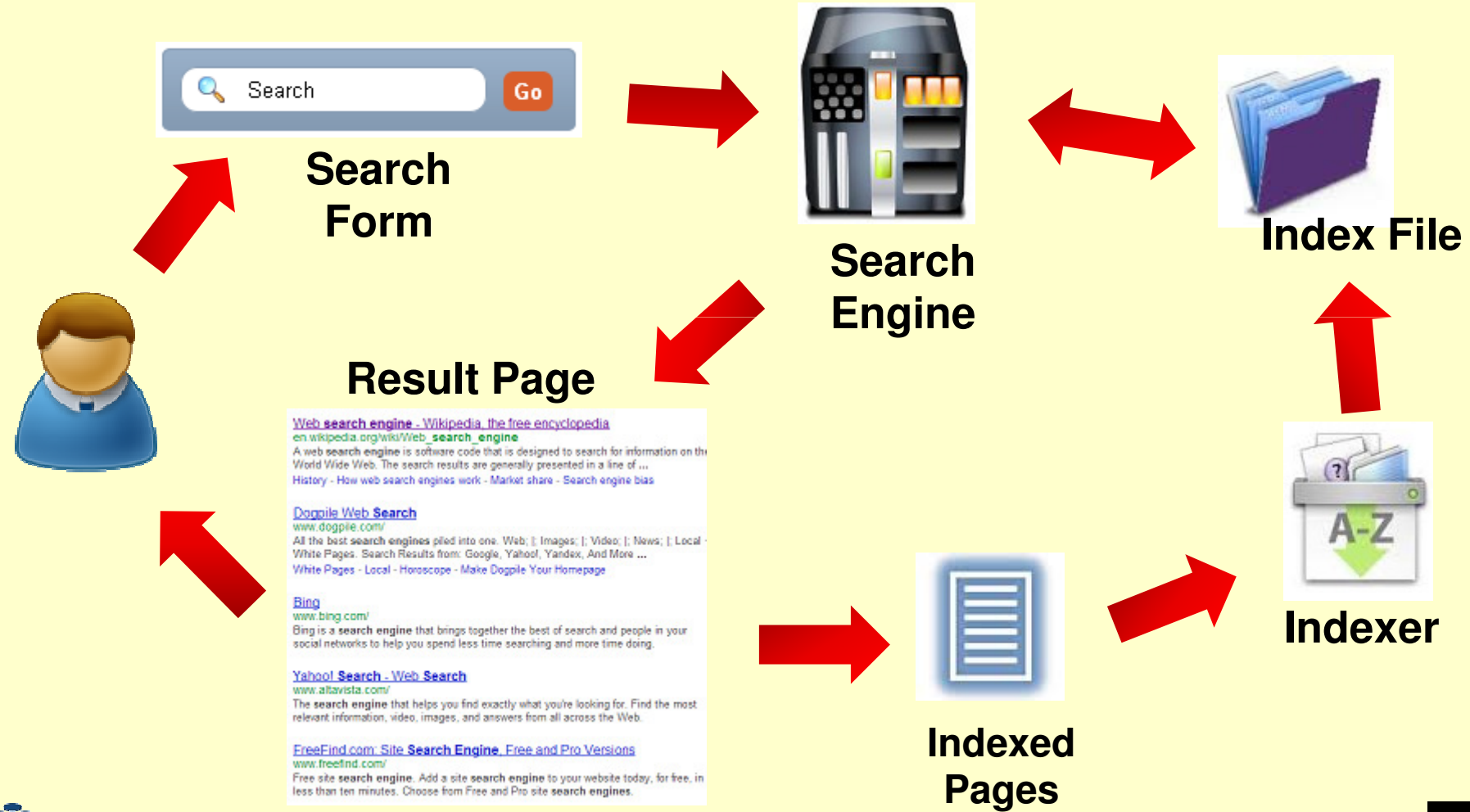
Year	Engine	Current status
2008	Powerset	Inactive (redirects to Bing)
	Picollator	Inactive
	Viewzi	Inactive
	Boogami	Inactive
	LeapFish	Inactive
	Forestle	Inactive (redirects to Ecosia)
	DuckDuckGo	Active
2009	Bing	Active, Launched as rebranded Live Search
	Yebol	Inactive
	Mugurdy	Inactive due to a lack of funding
	Goby	Active
	NATE	Active
2010	Blekko	Active
	Cuil	Inactive
	Yandex	Active, Launched global (English) search
	Yummly	Active
2011	Interred	Active as Interredu
	Yandex	Active, Launched Turkey search
2012	Volunia	Active
	Interredu	Active
	Open Drive	Active, cloud file search
2013	iStella	Active
	Aoohe	Active

Source: Wikipedia

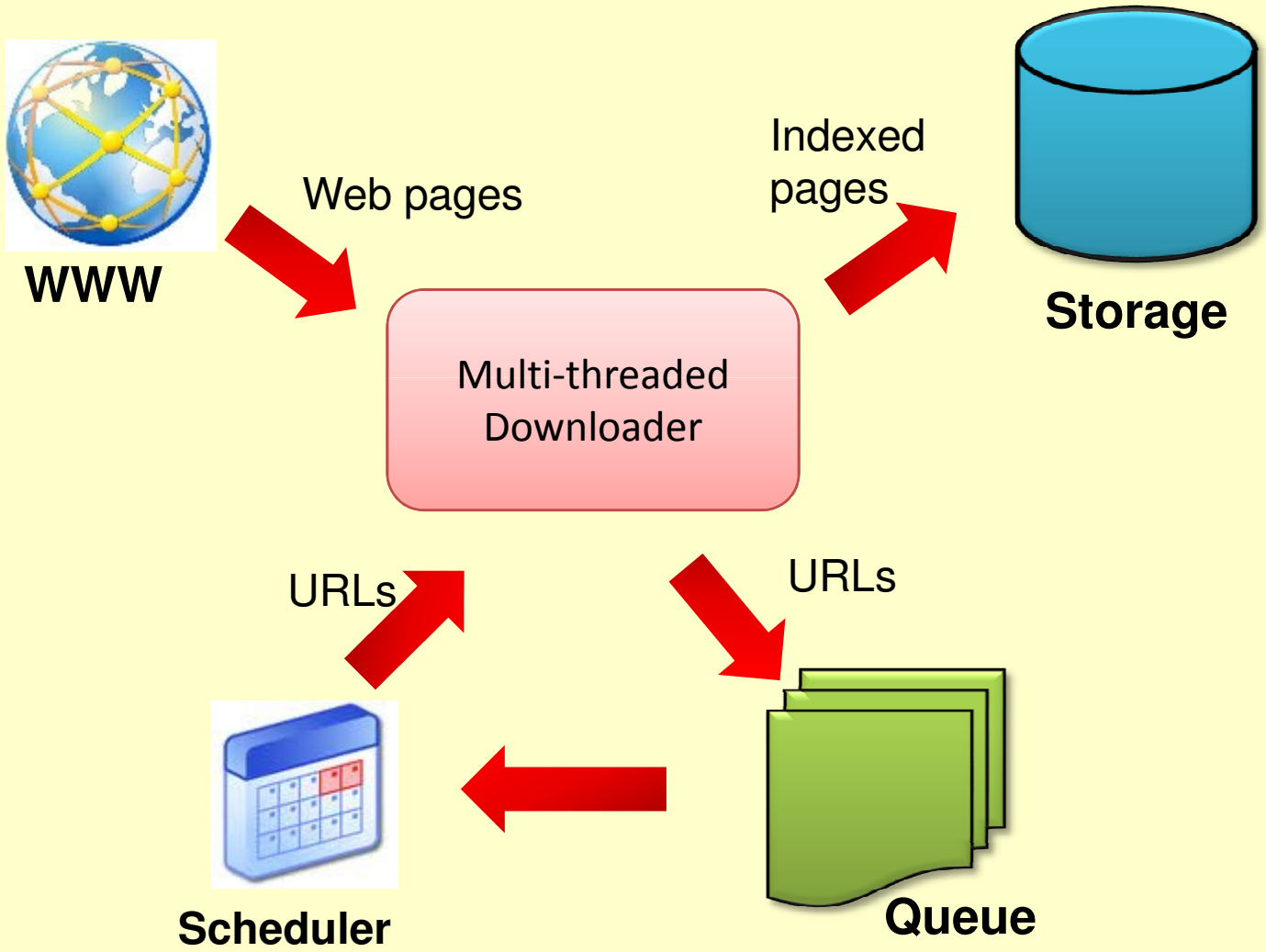
Bagaimana Search Engine Bekerja?

- Spider melakukan crawling halaman-halaman web untuk menemukan dokumen-dokumen baru, biasanya dengan mengikuti hyperlinks dari web yang sudah ada di database
- Search engine melakukan indexing terhadap halaman web dan menambahkannya ke dalam database. Ada proses update secara berkala.
- Search engine melakukan pencarian pada database berdasarkan query yang dimasukkan oleh user (bukan langsung pencarian pada halaman web)
- Search engine melakukan ranking dari hasil pencarian dokumen dengan menggunakan algoritma tertentu

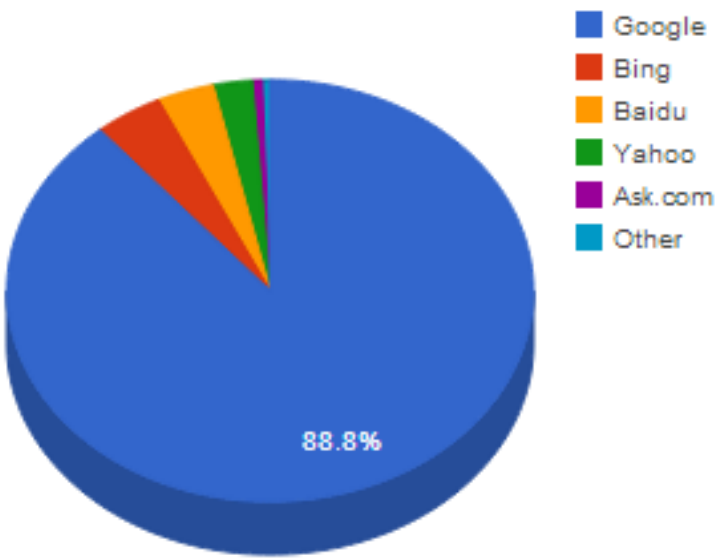
Bagaimana Search Engine Bekerja?



Bagaimana Web Crawler bekerja?



Market Share dari Search Engine



Global:

Global:	
Google	88.8%
Bing	4.2%
Baidu	3.5%
Yahoo	2.4%
Ask.com	0.6%
Other	0.5%

Source: www.karmasnack.com/about/search-engine-market-share/