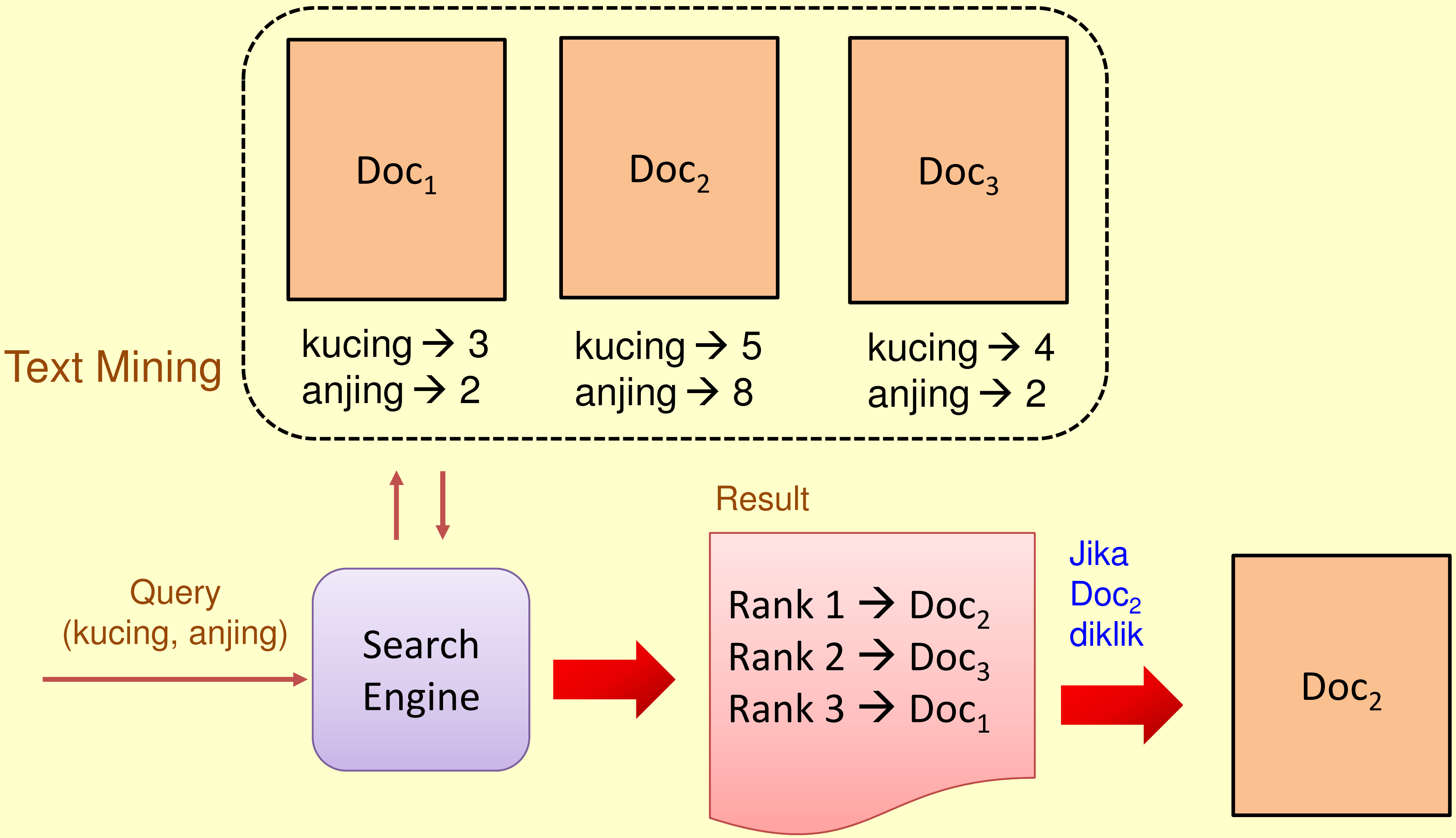
Text Mining & Search Engine

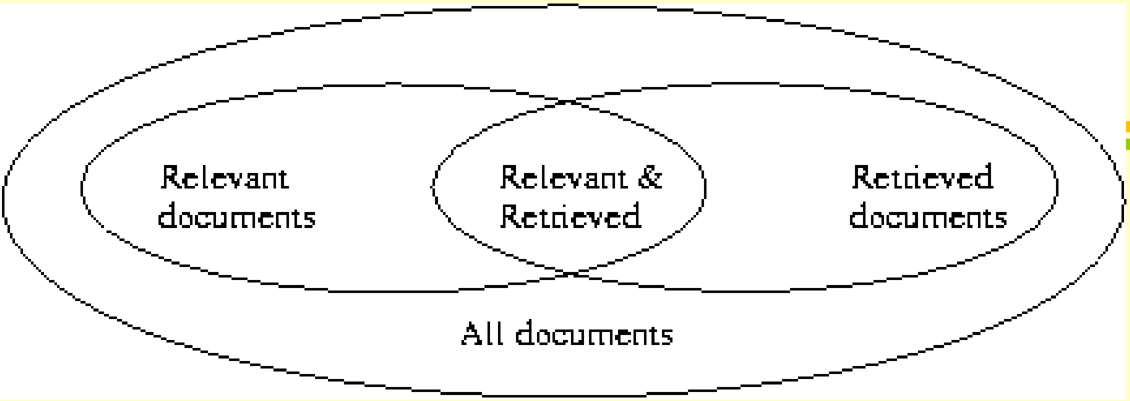
Ali Ridho Barakbah, Entin Martiana, Tri Hadiah

Knowledge Engineering Research Group
Department of Information and Computer Engineering
Politeknik Elektronika Negeri Surabaya

Text Mining & Search Engine



Basic Measures for Text Retrieval



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

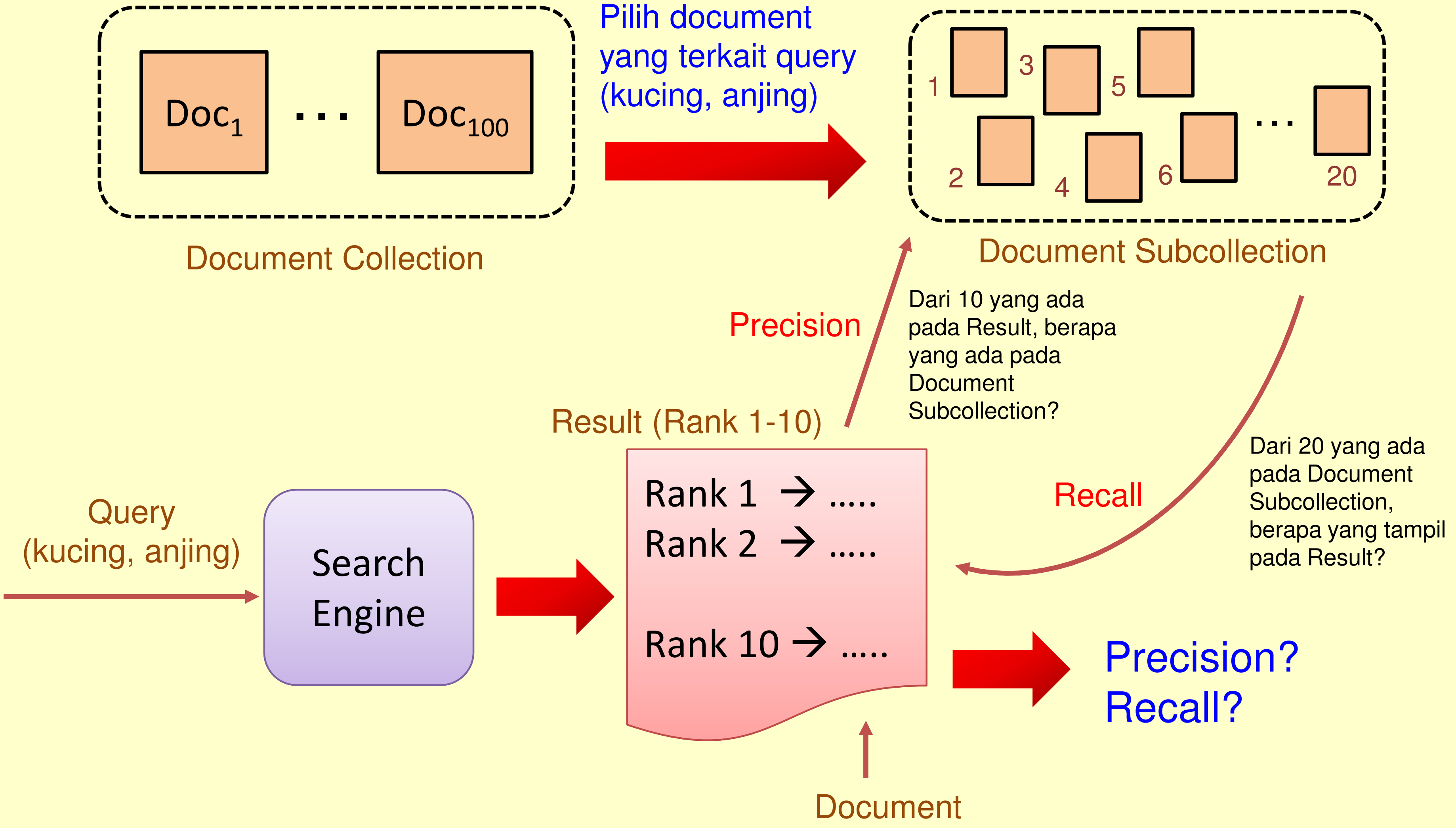
$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|}$$

- 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

Ilustrasi: Query → kucing anjing



Doc_1 . . . Doc_{100}

Document Collection

Pilih document yang terkait query (kucing, anjing)

1 3 5 2 4 6 . . . 20

Document Subcollection

Precision

Dari 10 yang ada pada Result, berapa yang ada pada Document Subcollection?

Result (Rank 1-10)

Rank 1 → .....

Rank 2 → .....

Rank 10 → .....

Recall

Dari 20 yang ada pada Document Subcollection, berapa yang tampil pada Result?

Query (kucing, anjing)

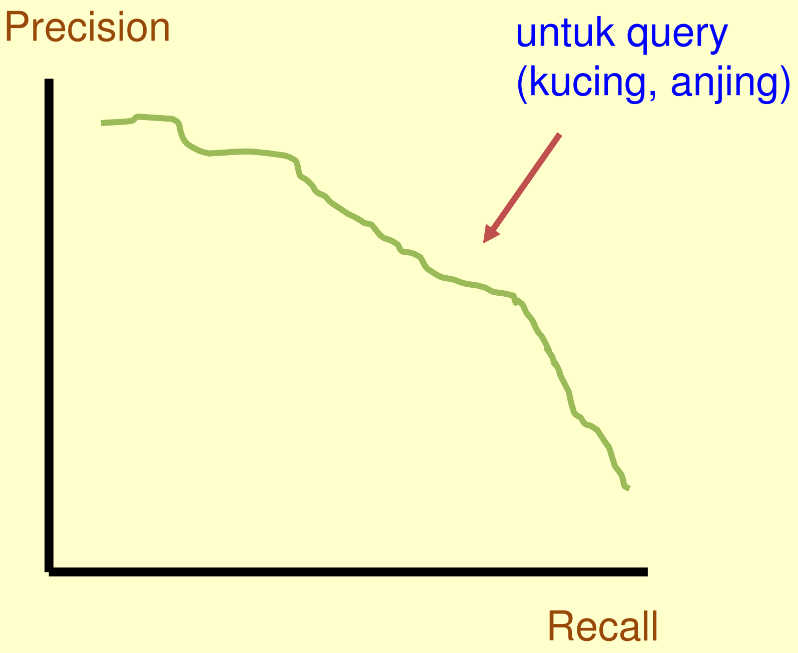
Search Engine

Precision?

Recall?

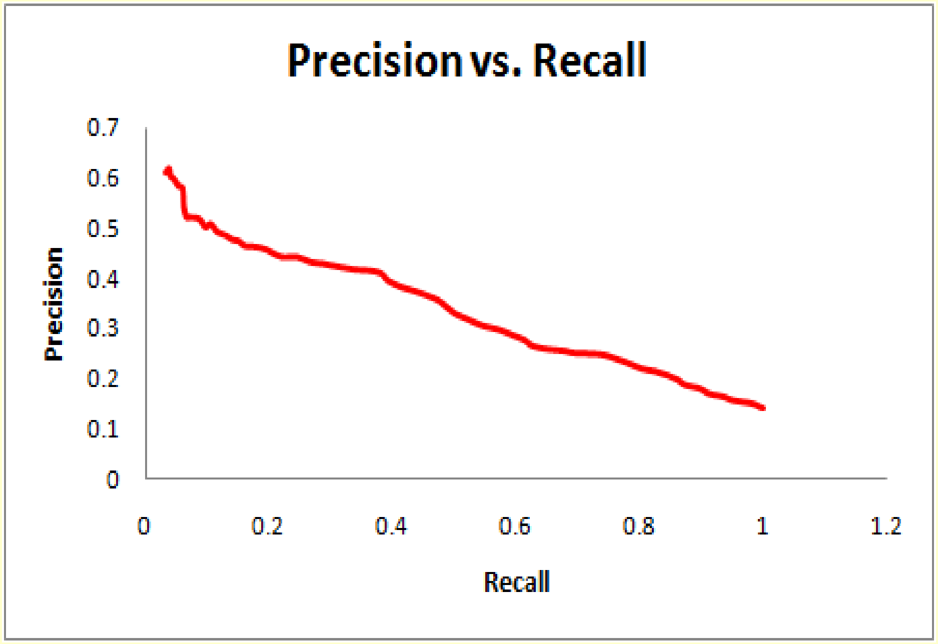
Document

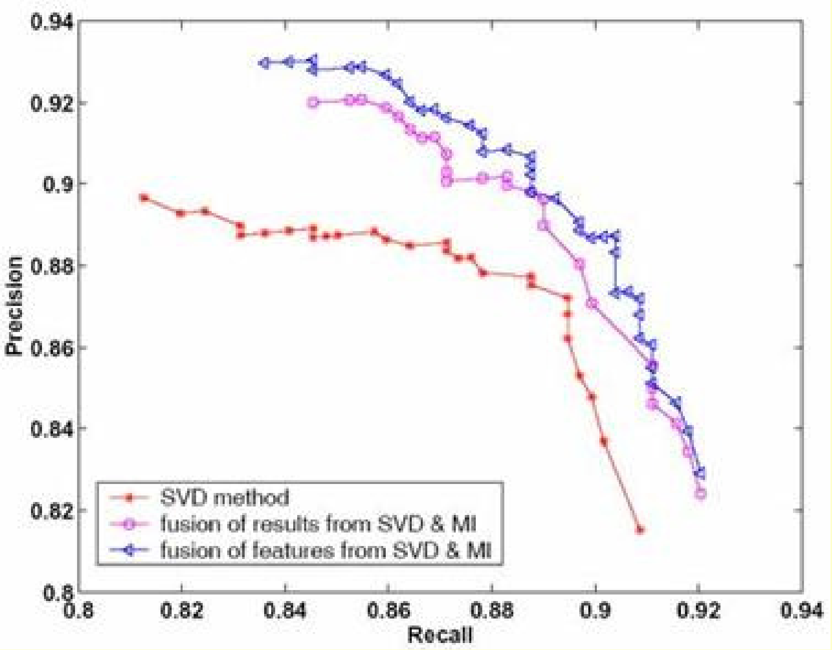
Rank 1	Recall?	Precision?
Rank 1-2	Recall?	Precision?
Rank 1-3	Recall?	Precision?
Rank 1-4	Recall?	Precision?
Rank 1-5	Recall?	Precision?
Rank 1-6	Recall?	Precision?
Rank 1-7	Recall?	Precision?
Rank 1-8	Recall?	Precision?
Rank 1-9	Recall?	Precision?
Rank 1-10	Recall?	Precision?
…	Recall?	Precision?



untuk query
(kucing, anjing)

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

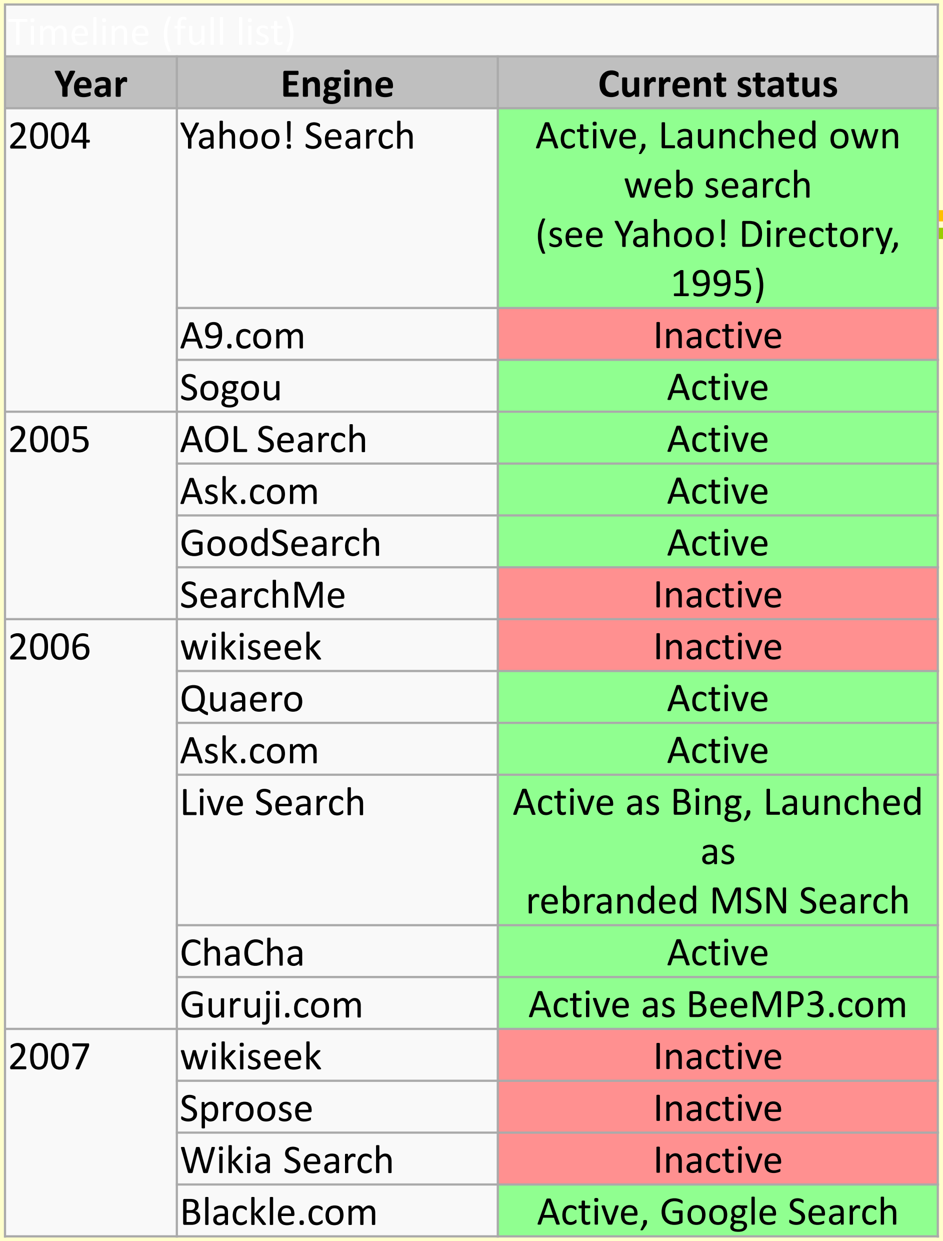
Source: Saeed El-Darahali, Search Engines & Search Engine Optimization (SEO), 7th World Congress on the Management of e-Business

YAHOO!
GO 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
dmoz open directory project
ZenSearch
Netscape
DOGPILE
Mamma.com
Go.com
ToggleBot!
Questel • Orbit
France Telecom Group

Timeline (full list)

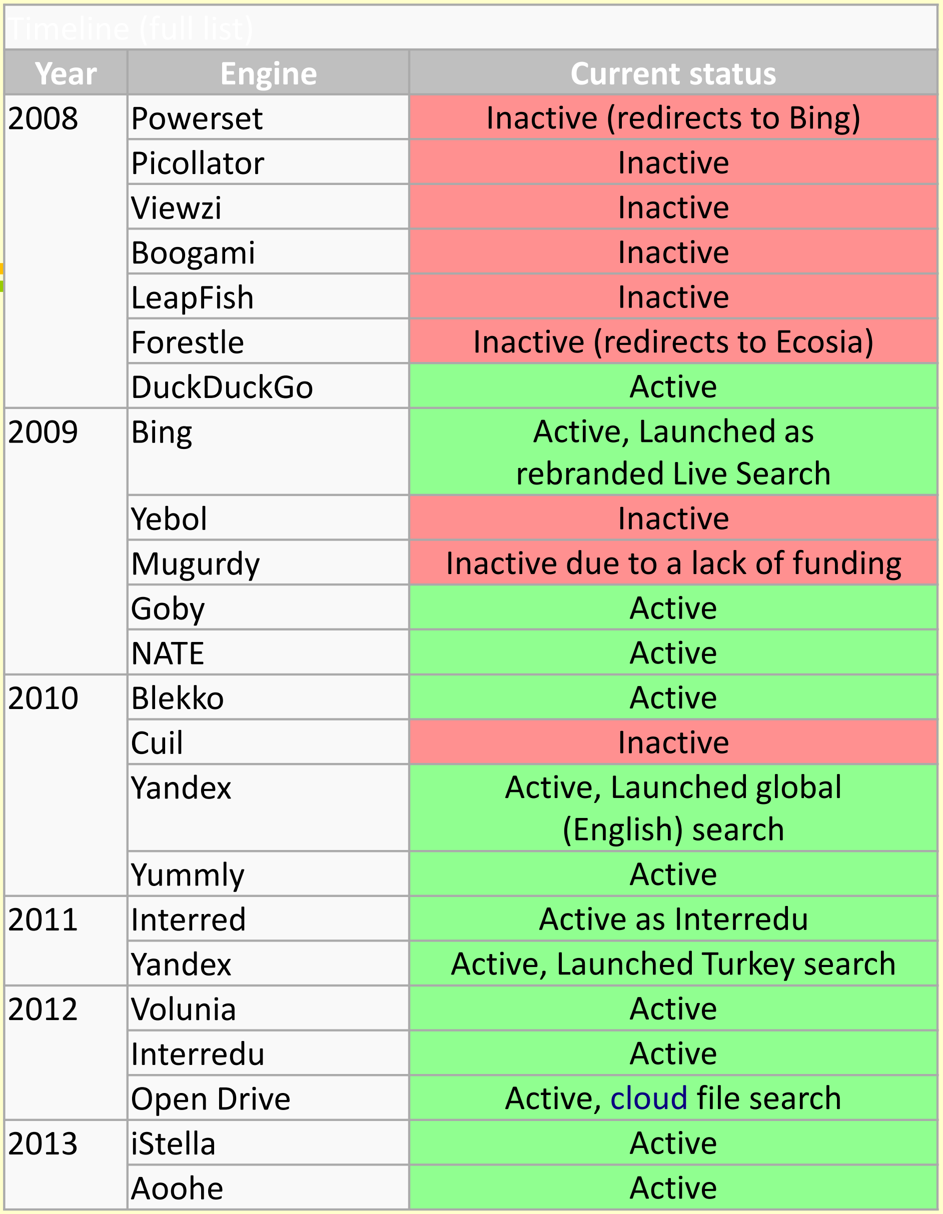
Year	Engine	Current status
1993	W3Catalog	Inactive
	Aliweb	Inactive
1994	WebCrawler	Active, Aggregator
	Go.com	Active, Yahoo Search
	Lycos	Active
1995	AltaVista	Active, Yahoo Search
	Daum	Active
	Magellan	Inactive
	Excite	Active
	SAPO	Active
	Yahoo! 2008	Active, Launched as a directory
1996	Dogpile	Active,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

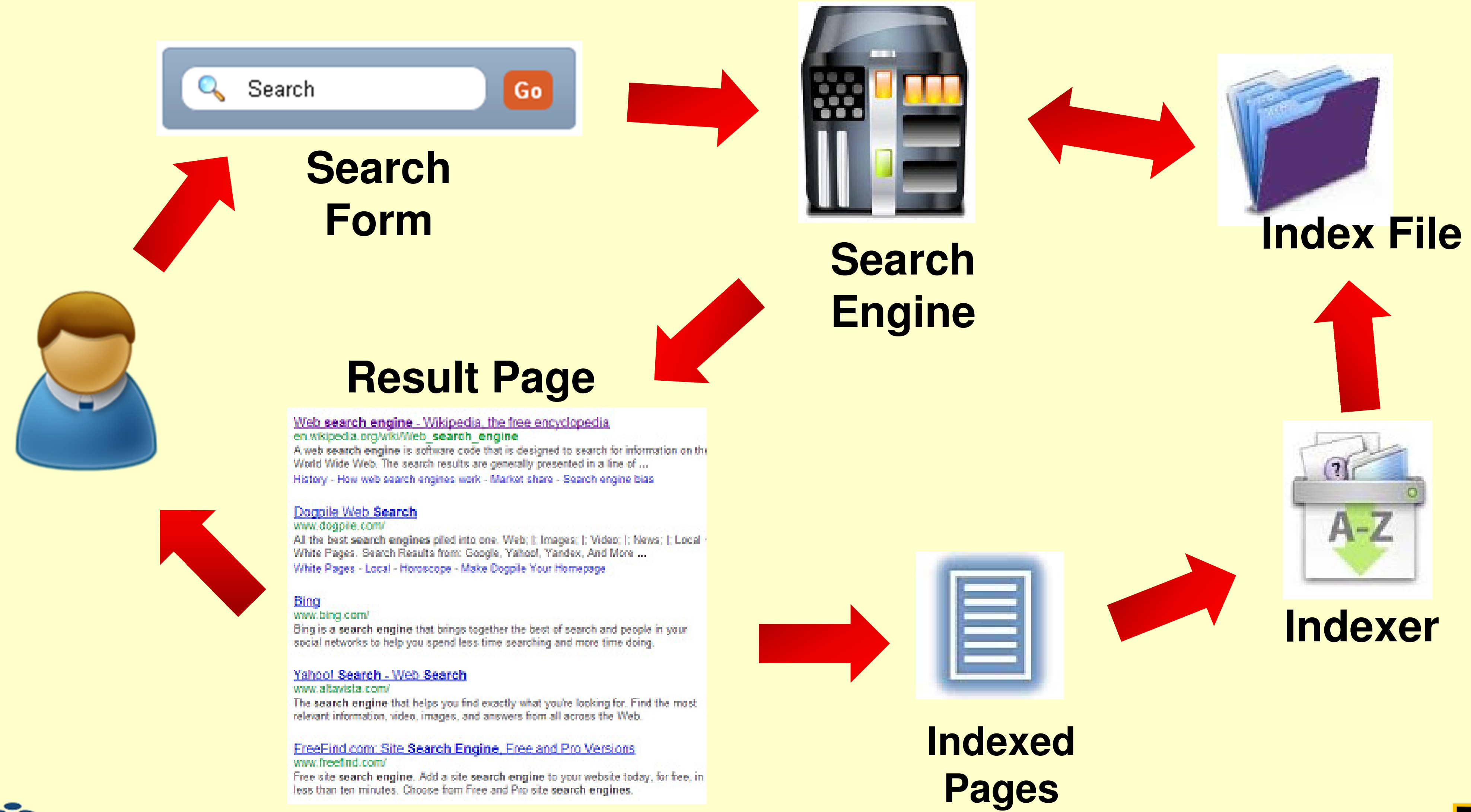
Year	Engine	Current status
2008	Powerset	Inactive (redirects to Bing)
	Picollator	Inactive
	Viewzi	Inactive
	Boogami	Inactive
	LeapFish	Inactive
	Forestle	Inactive (redirects to Ecosia)
	DuckDuckGo	Active
2009	Bing	Active, Launched as rebranded Live Search
	Yebol	Inactive
	Mugurdy	Inactive due to a lack 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?



Web search engine - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Web_search_engine
A web search engine is software code that is designed to search for information on the World Wide Web. The search results are generally presented in a line of ...
History - How web search engines work - Market share - Search engine bias

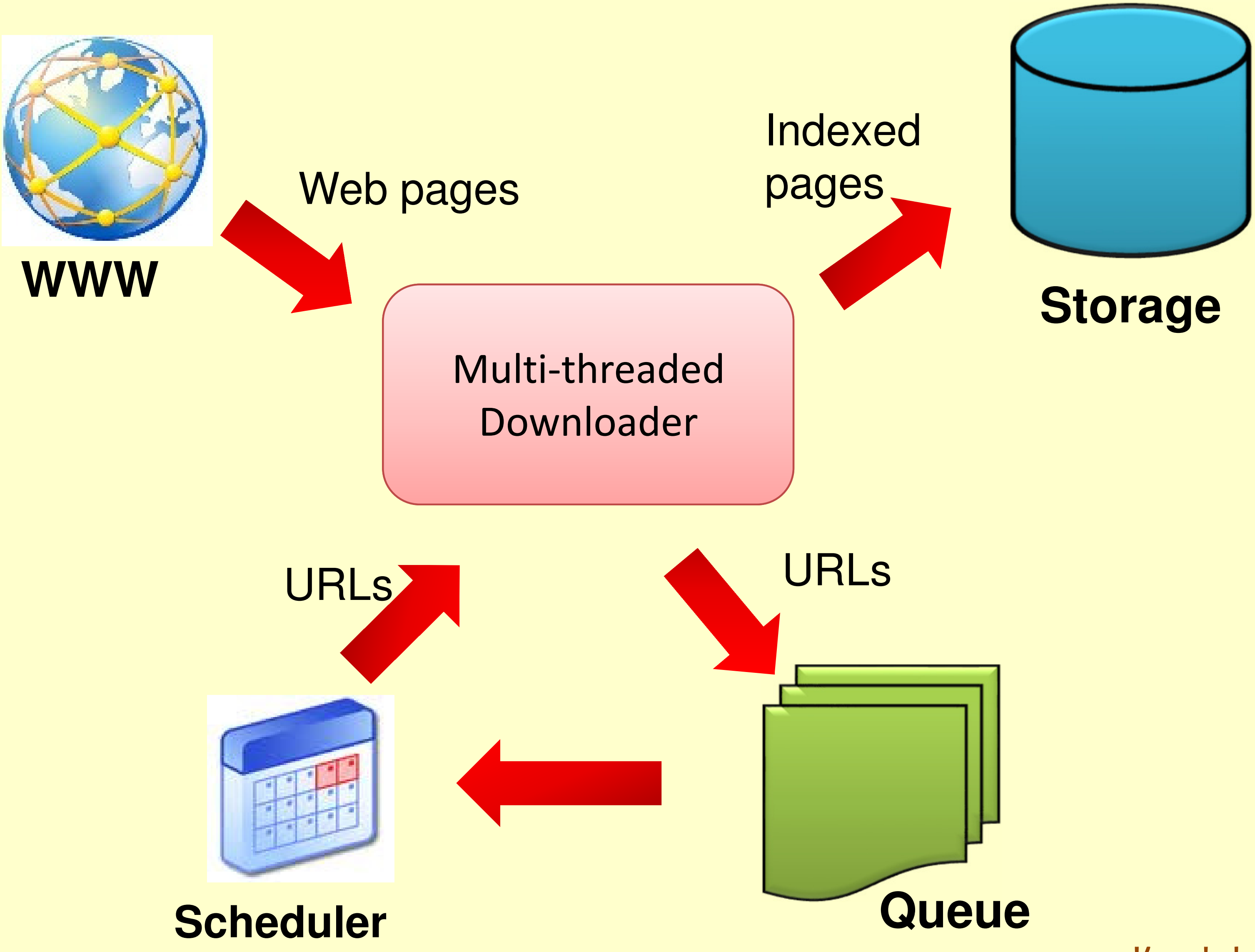
Dogpile Web Search
www.dogpile.com/
All the best search engines piled into one. Web; |; Images; |; Video; |; News; |; Local · White Pages. Search Results from: Google, Yahoo!, Yandex, And More ...
White Pages - Local - Horoscope - Make Dogpile Your Homepage

Bing
www.bing.com/
Bing is a search engine that brings together the best of search and people in your social networks to help you spend less time searching and more time doing.

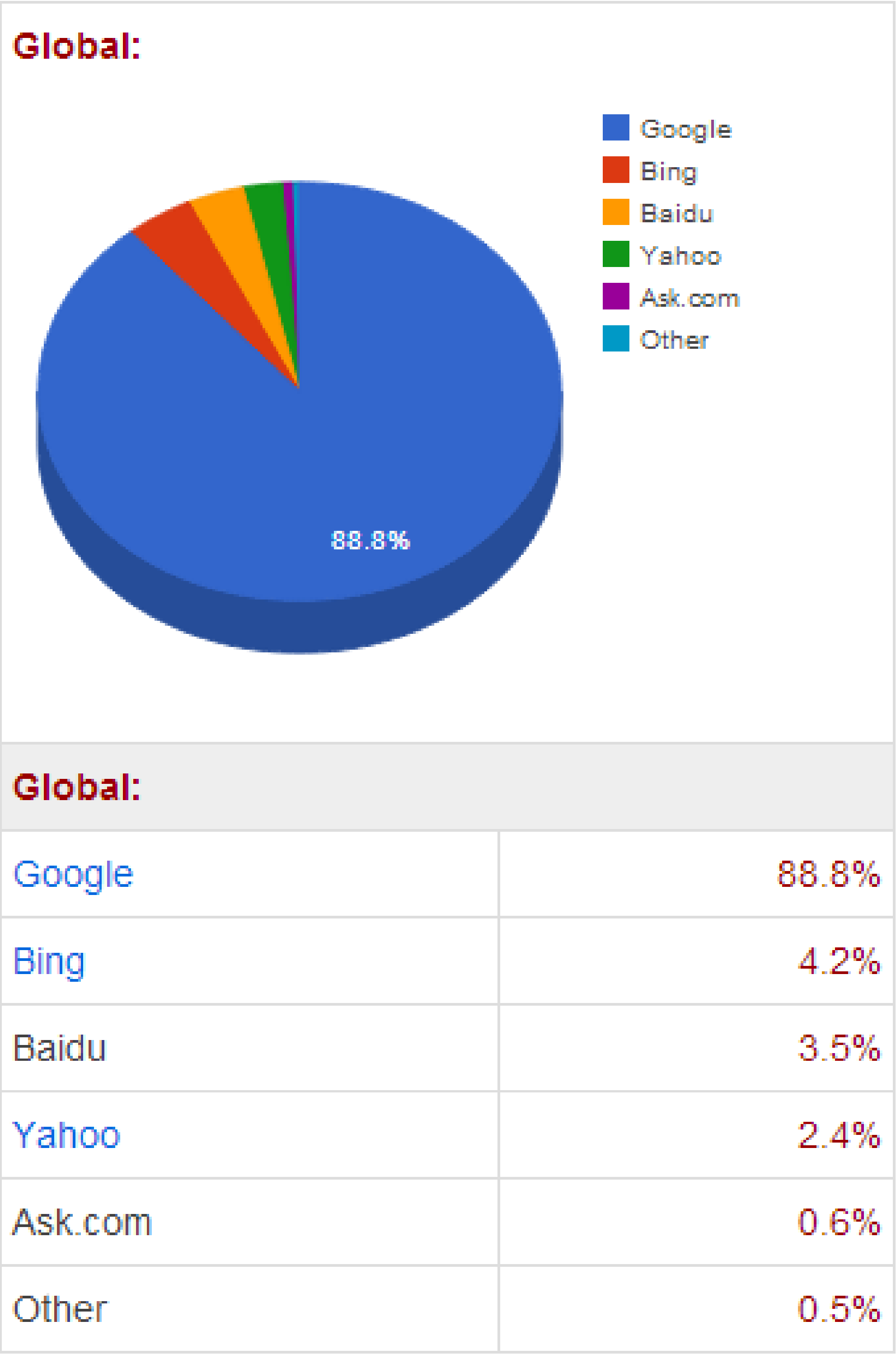
Yahoo! Search - Web Search
www.altavista.com/
The search engine that helps you find exactly what you're looking for. Find the most relevant information, video, images, and answers from all across the Web.

FreeFind.com: Site Search Engine, Free and Pro Versions
www.freefind.com/
Free site search engine. Add a site search engine to your website today, for free, in less than ten minutes. Choose from Free and Pro site search engines.

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/