Search Engines & Information Retrieval

406.424 Internet Applications

Jonghun Park

jonghun@snu.ac.kr

Dept. of Industrial Eng.

Seoul National University

9/1/2010

search and information retrieval

- search on the web is a daily activity for many people throughout the world
- **search** and **communication** are most popular uses of the computer
- applications involving search are everywhere
- the field that is most involved with R&D for search is **information retrieval (IR)**

information retrieval

- *“Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.”* (Salton, 1968)
- originated from “library science” community
- primary focus of IR since the 50s has been on **text** and **documents**

what is a document?

- examples:
  - web pages, email, books, news stories, scholarly papers, text messages, Word, Powerpoint, PDF, forum postings, patents, IM sessions, etc.
- common properties
  - significant text content
  - some structure (e.g., title, author, date for papers; subject, sender, destination for email)

Documents vs. Database Records

- Database records (or tuples in relational databases) are typically made up of well-defined fields (or attributes)
  - e.g., bank records with account numbers, balances, names, addresses, social security numbers, dates of birth, etc.
- Easy to compare fields with well-defined semantics to queries in order to find matches
- Text is relatively **unstructured**: more difficult
- Example bank database query
  - Find records with balance > $50,000 in branches located in Amherst, MA.
  - Matches easily found by comparison with field values of records
- Example search engine query
  - bank scandals in western mass
  - This text must be compared to the text of entire news stories

Comparing Text

- Comparing the query text to the document text and determining **what is a good match** is the core issue of information retrieval
- **Exact matching of words is not enough**
  - Many different ways to write the same thing in a “natural language” like English
  - e.g., does a news story containing the text “bank director in Amherst steals funds” match the query?
  - Some stories will be better matches than others

Dimensions of IR

- Any application that involves a collection of text or other unstructured information will need to organize and search that information
- IR is more than just text, and more than just web search
  - although these are central
- New applications increasingly involve **new media**
  - e.g., video, photos, music, speech
- Like text, media content is difficult to describe and compare
  - text may be used to represent them (e.g. tags)

IR Tasks

- Ad-hoc search: Find relevant documents for an arbitrary text query
- Filtering: Identify relevant user profiles for a new document
- Classification: Identify relevant labels for documents
- Question answering: Give a specific answer to a question

Content	Applications	Tasks
Text	Web search	Ad hoc search
Images	Vertical search	Filtering
Video	Enterprise search	Classification
Scanned docs	Desktop search	Question answering
Audio	Forum search	
Music	P2P search	
	Literature search	

web search engine market

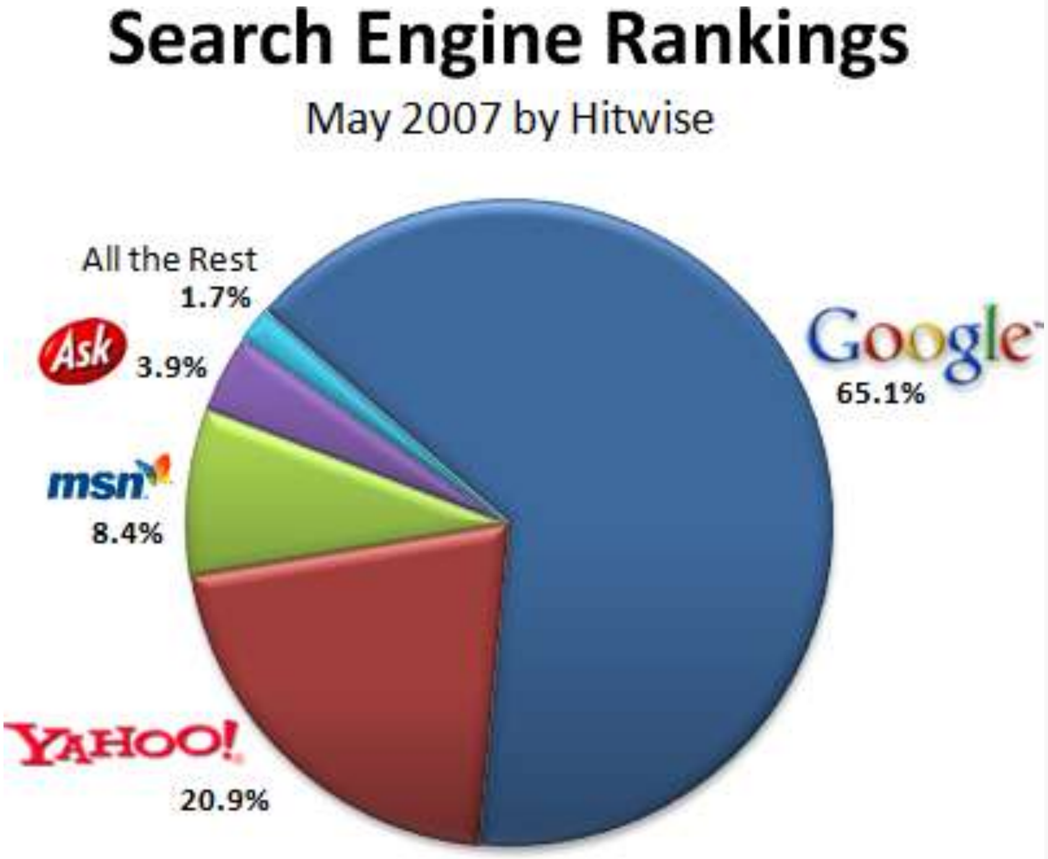
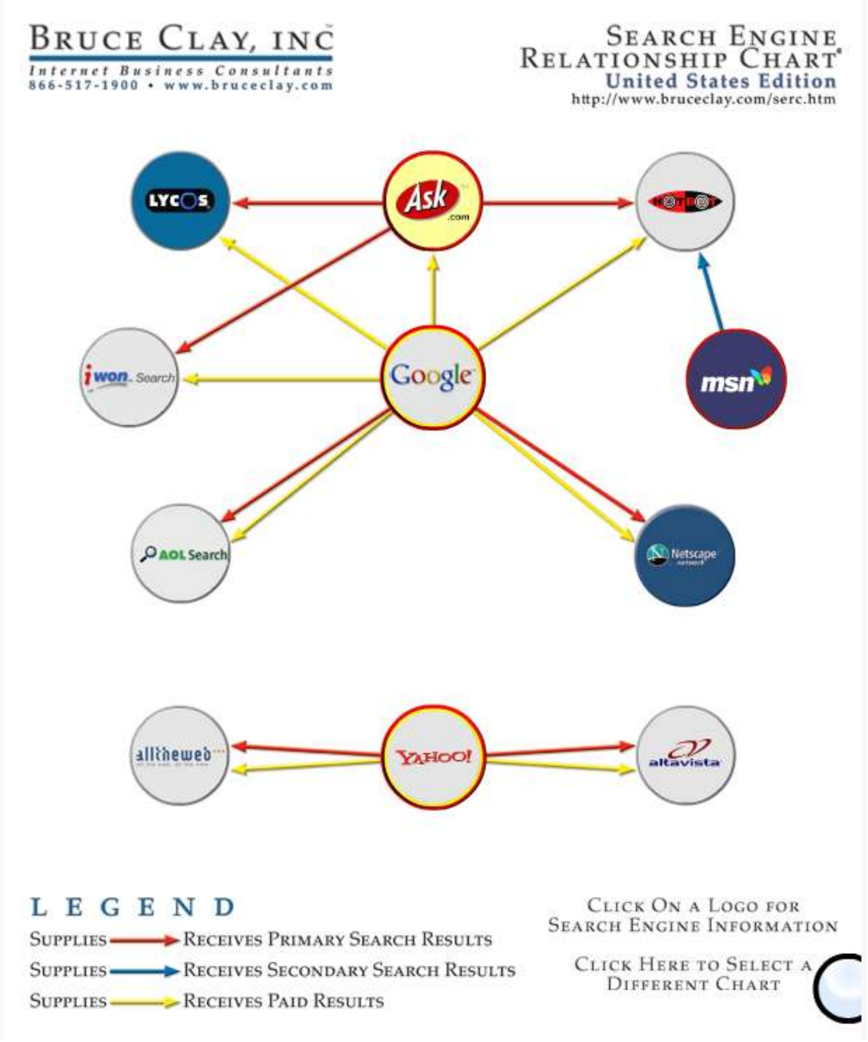


Chart by www.internetworldstats.com, Copyright © 2007



LEGEND

Supplies → Receives Primary Search Results
Supplies → Receives Secondary Search Results
Supplies → Receives Paid Results

Click On a Logo for Search Engine Information

Click Here to Select a Different Chart

niche search engines: by domain

wink People Search

People Search | Email Search | Find Friends | Background Check | Phone Number Search | Find a Person Online

Name: | Location: | Other:

Refine search:

People Search News

People Watch

Where are you online?

bebo

Local.com

Local Restaurants

El Cholo Restaurant

Local Reviews

Local Traffic

Find them on Local.com

Local Coupons

Local Favorites

OVER 20 MILLION POTENTIAL CUSTOMERS visit Local.com each month

REACH THEM TODAY

KAYAK

Hotels

Flights

Cars

Deals

Vacations

Cruises

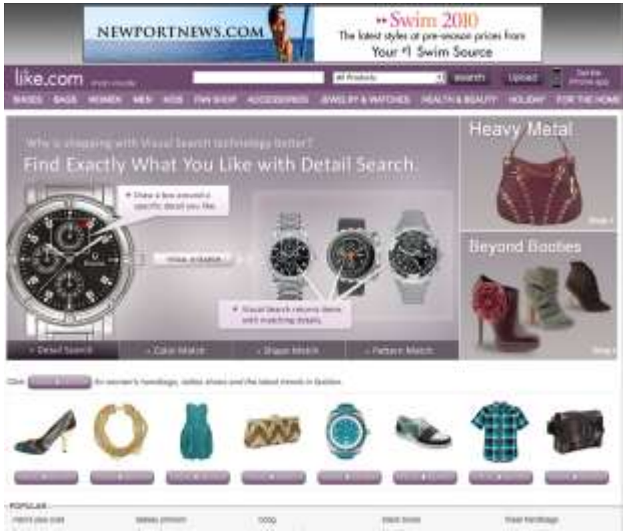
My Trips

Private Sale

More

Search

HOLIDAY INN
BEST WESTERN
BRITISH AIRWAYS
HAMPTON


NEWPORTNEWS.COM

Swim 2010

The latest styles at pre-season prices from Your #1 Swim Source

like.com

Find Exactly What You Like with Detail Search.

Heavy Metal

Beyond Booties

simplyhired. job search made simple

1,060,939 New Jobs In the Last Week

Keywords | Location

Find Jobs by Category

Job Search Made Simple

TIME | Forbes | PC | CNET.com

Yakaz

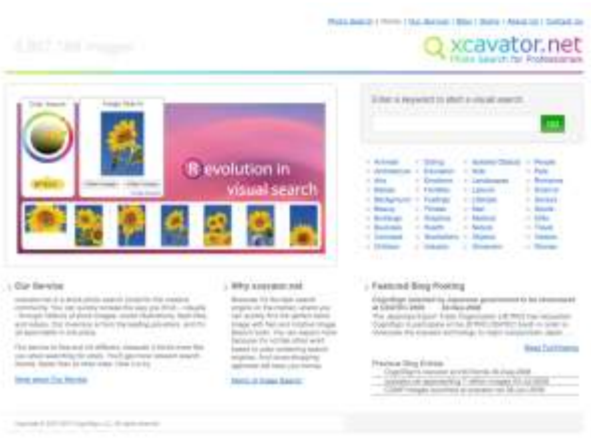
United States

Classified ads search engine

Search

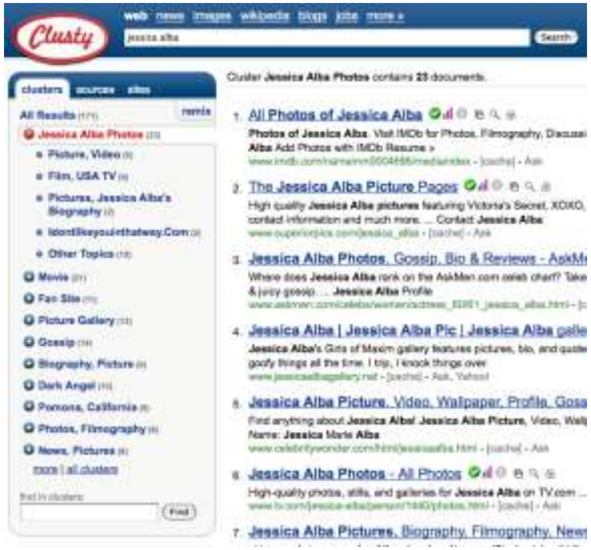
4,739,983 classified ads coming from 11,483 websites

niche search engines: by type



TRUVEO
SEARCH
lady gaga | tiger woods | ellen degeneres | sandra bullock | nascar
News
Sports
TV Shows
Music
Most Twittered

presentation of search results



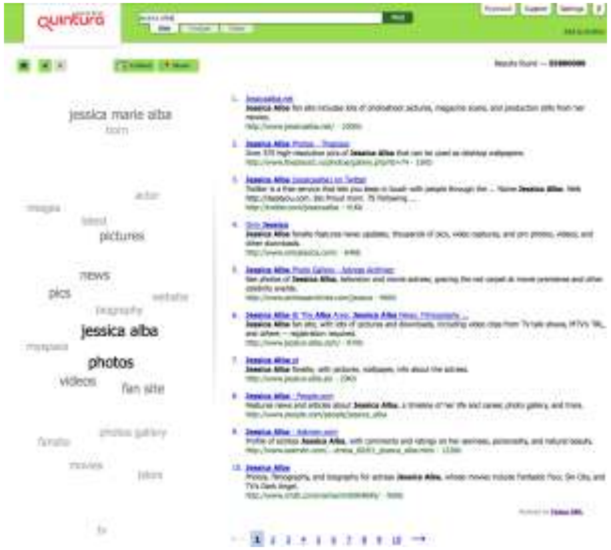
All Results (171)
- Jessica Alba Photos (23)
  - Picture, Video (6)
  - Film, USA TV (4)
  - Pictures, Jessica Alba's Biography (2)
  - Idontlikeyouinthatway.Com (2)
  - Other Topics (10)
- Movie (21)
- Fan Site (15)
- Picture Gallery (13)
- Gossip (19)
- Biography, Picture (9)
- Dark Angel (10)
- Pomona, California (8)
- Photos, Filmography (8)
- News, Pictures (8)

more | all clusters

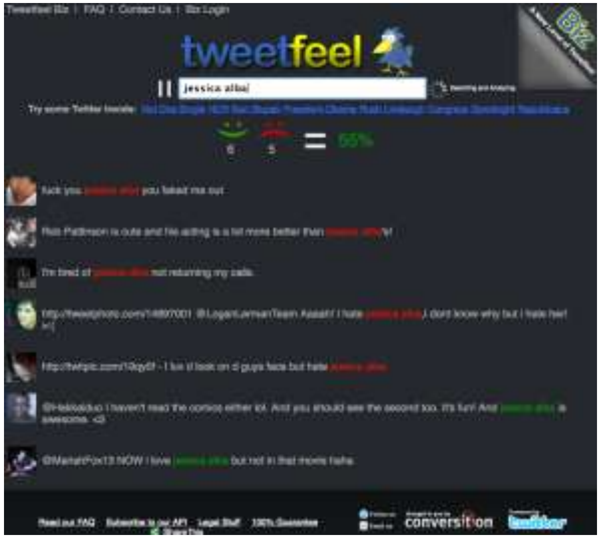
find in clusters:

Cluster **Jessica Alba Photos** contains **23** documents.

1. All Photos of Jessica Alba
   Photos of Jessica Alba. Visit IMDb for Photos, Filmography, Discussi... **Alba** Add Photos with IMDb Resume »
   www.imdb.com/name/nm0004695/mediaindex - [cache] - Ask
2. The Jessica Alba Picture Pages
   High quality **Jessica Alba pictures** featuring Victoria's Secret, XOXO, ... contact information and much more. ... Contact **Jessica Alba**
   www.superiorpics.com/jessica_alba - [cache] - Ask
3. Jessica Alba Photos, Gossip, Bio & Reviews - AskM...
   Where does **Jessica Alba** rank on the AskMen.com celeb chart? Take... & juicy gossip. ... **Jessica Alba** Profile
   www.askmen.com/celebs/women/actress_85951_jessica_alba.html - [c...
4. Jessica Alba | Jessica Alba Pic | Jessica Alba galle...
   Jessica Alba's Girls of Maxim gallery features pictures, bio, and quote... goofy things all the time. I trip, I knock things over
   www.jessicaalbagallery.net - [cache] - Ask, Yahoo!
5. Jessica Alba Picture, Video, Wallpaper, Profile, Goss...
   Find anything about **Jessica Alba**! **Jessica Alba Picture**, Video, Wall... Name: **Jessica** Marie **Alba**
   www.celebritywonder.com/html/jessicaalba.html - [cache] - Ask
6. Jessica Alba Photos - All Photos
   High-quality photos, stills, and galleries for **Jessica Alba** on TV.com ...
   www.tv.com/jessica-alba/person/1440/photos.html - [cache] - Ask
7. Jessica Alba Pictures, Biography, Filmography, News...



1. jessicaalba.net
2. Jessica Alba Photos - ...
3. Jessica Alba (jessicaalba) on Twitter
4. Only Jessica
5. Jessica Alba Photo Gallery - Adnews Archives
6. Jessica Alba @ The Alba Area: Jessica Alba News, Filmography ...
7. Jessica Alba.ro
8. Jessica Alba - People.com
9. Jessica Alba : Askmen.com
10. Jessica Alba



fuck you jessica alba you faked me out

Rob Pattinson is cute and his acting is a lot more better than jessica alba's!

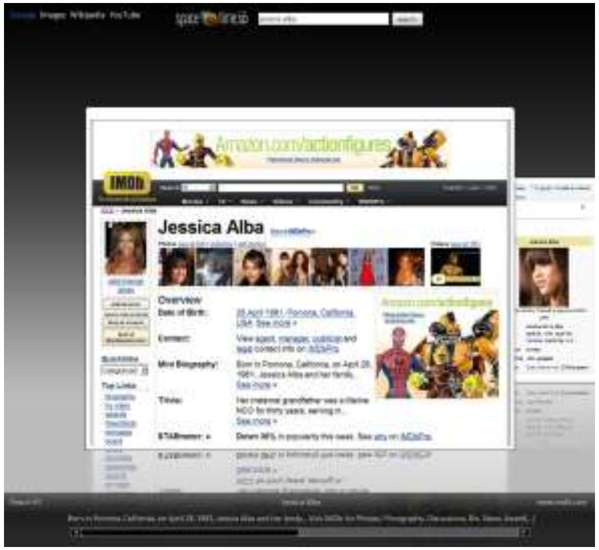
I'm tired of jessica alba not returning my calls.

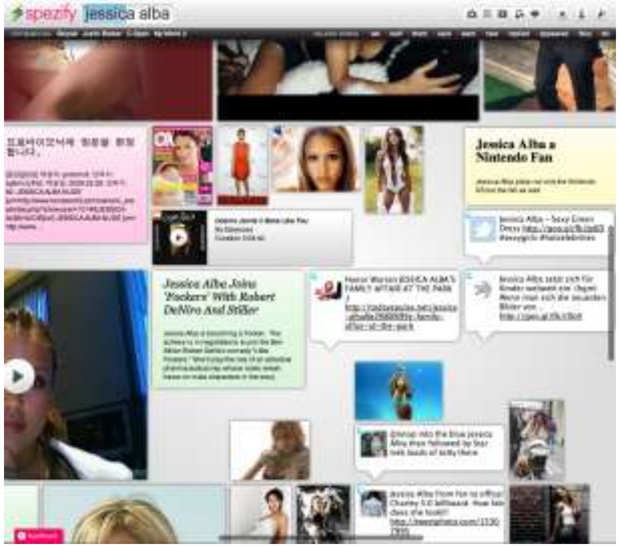
http://tweetphoto.com/14897001 @LoganLermanTeam Aaaah! I hate jessica alba,I dont know why but I hate her! lol]

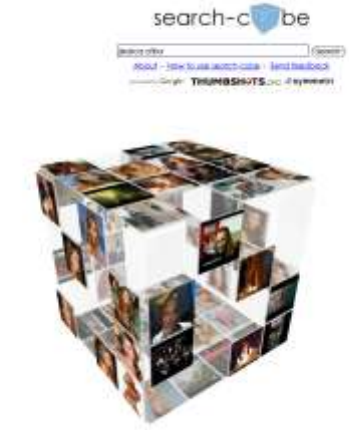
http://twitpic.com/14qy0f - I luv d look on d guys face but hate jessica alba

@Helloduo I haven't read the comics either lol. And you should see the second too. It's fun! And jessica alba is awesome. <3

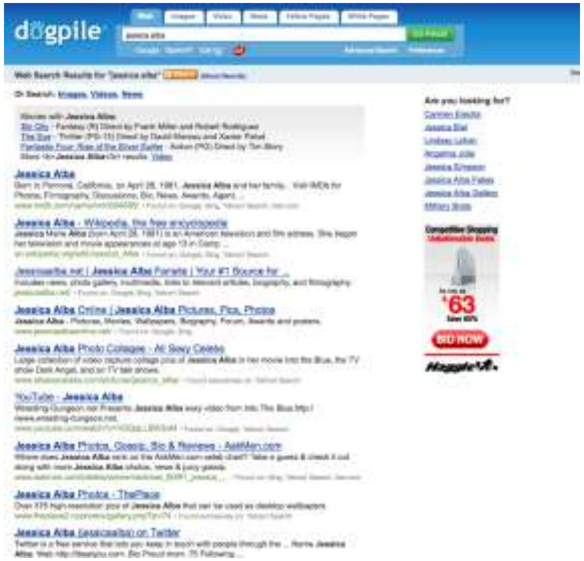
@MariahFox13 NOW I love jessica alba but not in that movie haha

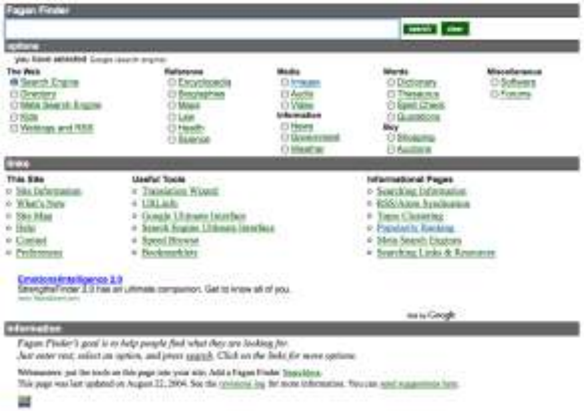



spezify
jessica alba
Jessica Alba a Nintendo Fan
Jessica Alba Joins 'Fockers' With Robert DeNiro And Stiller

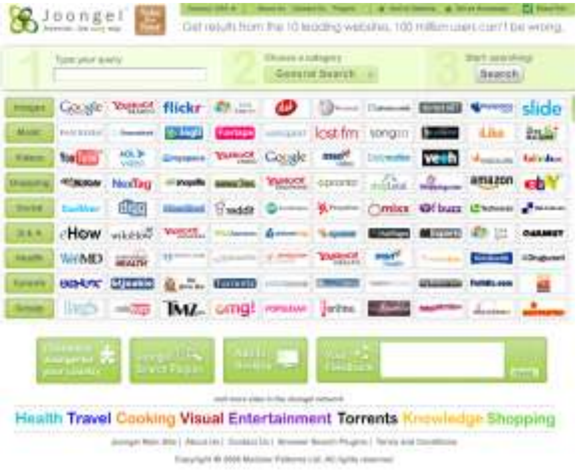


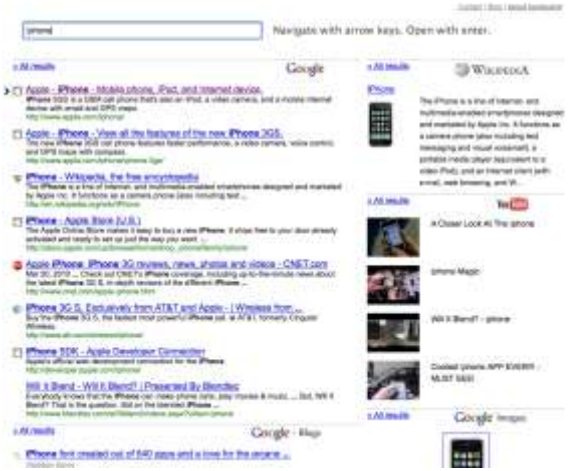
enhancing search engine usability

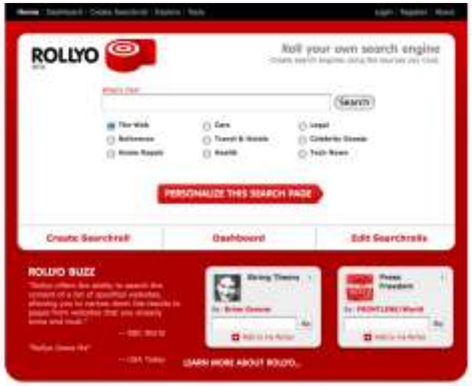

dogpile
Web Search Results for "Jessica alba"
Are you looking for?
Jessica Alba
Jessica Alba - Wikipedia, the free encyclopedia
Jessica Alba Online | Jessica Alba Pictures, Pics, Photos
Jessica Alba Photo Collage - All Sexy Celebs
YouTube - Jessica Alba
Jessica Alba Photos, Gossip, Bio & Reviews - AskMen.com
Jessica Alba Photos - ThePlace
Jessica Alba (jessicaalba) on Twitter
$63
BID NOW

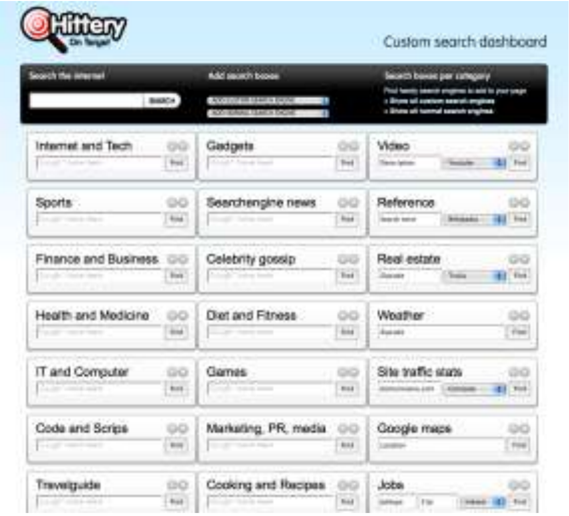


Fagan Finder's goal is to help people find what they are looking for.
Just enter text, select an option, and press search. Click on the links for more options.


Joongel
Get results from the 10 leading websites. 100 million users can't be wrong.
Search
Google
flickr
slide
last.fm
YouTube
amazon
ebY
How
WebMD
TMZ
omg!
Health Travel Cooking Visual Entertainment Torrents Knowledge Shopping


iphone
Navigate with arrow keys. Open with enter.
Google
Wikipedia
Apple - iPhone - Mobile phone, iPod, and Internet device.
iPhone - Wikipedia, the free encyclopedia
YouTube
A Closer Look At The iphone
iphone Magic
Will it Blend? - iphone
Google Images
Google Blogs


ROLLYO
Roll your own search engine
Search
PERSONALIZE THIS SEARCH PAGE
Create Searchroll
Dashboard
Edit Searchrolls
ROLLYO BUZZ
LEARN MORE ABOUT ROLLYO...



big issues in IR

- **relevance**
  - simple definition: a relevant document contains the information that a person was looking for when s/he submitted a query to the search engine
  - many factors influence a person’s decision about what is relevant: e.g., task, context, novelty, style
  - **topical relevance** (same topic) vs. **user relevance** (everything else)

big issues in IR

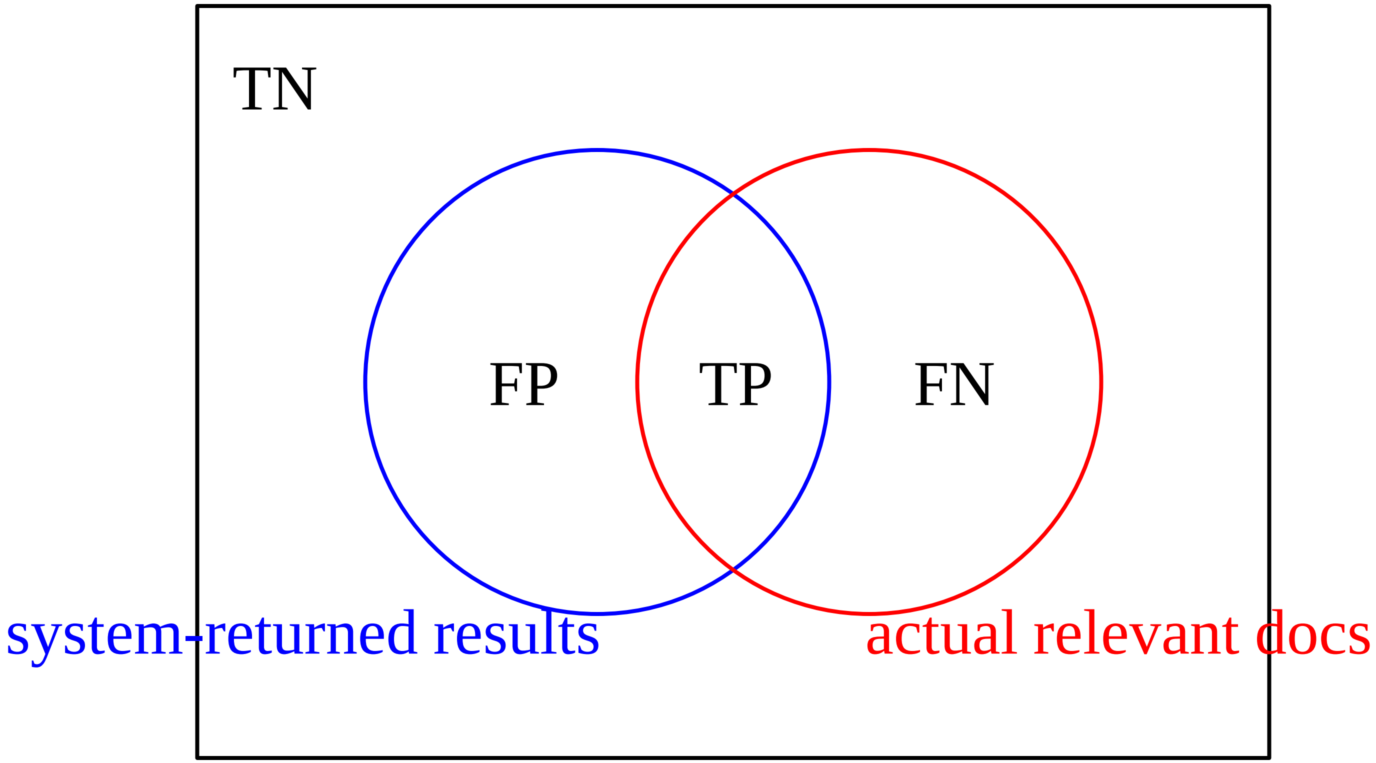
- relevance
  - **retrieval models** define a view of relevance
    - retrieval model: a formal representation of the process of matching a query and a document
  - ranking algorithms used in search engines are based on retrieval models
  - most models describe **statistical properties** of text rather than linguistic
    - i.e. counting simple text features such as words instead of parsing and analyzing the sentences
    - **statistical approach** to text processing started with Luhn in the 50s
      - e.g., TF
    - linguistic features can be part of a statistical model

big Issues in IR

- evaluation
  - experimental procedures and measures for comparing IR system output with user expectations
    - originated in Cranfield experiments in the 60s
  - IR evaluation methods now used in many fields
  - typically use test collection of documents, queries, and relevance judgments
    - most commonly used are **TREC collections**
  - recall and precision are two examples of effectiveness measures
    - precision: proportion of retrieved documents that are relevant
    - recall: proportion of relevant documents that are retrieved
    - assumption: all the relevant documents for a given query are known

big issues in IR

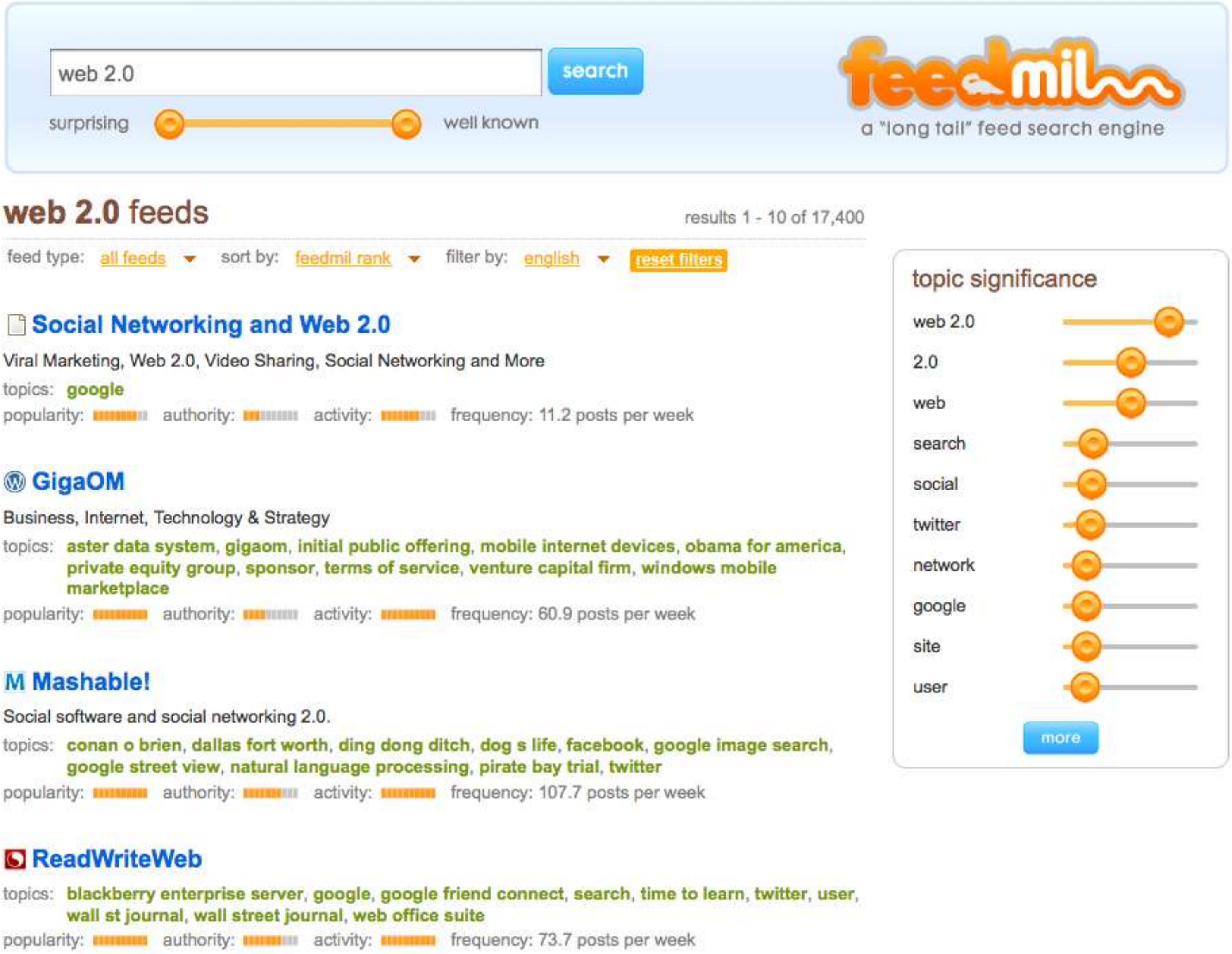
- precision = TP / (TP + FP)
- recall = TP / (TP + FN)



big issues in IR

- users and information needs
  - search evaluation is **user-centered**
  - keyword queries are often poor descriptions of actual information needs
    - e.g., “apple”
  - interaction and context are important for understanding user intent
    - analysis of clickthrough data
  - query refinement techniques such as **query expansion**, **query suggestion, relevance feedback** improve ranking

PRF example



web 2.0 search

surprising — well known

feedmil

a "long tail" feed search engine

web 2.0 feeds

results 1 - 10 of 17,400

feed type: all feeds sort by: feedmil rank filter by: english reset filters

Social Networking and Web 2.0

Viral Marketing, Web 2.0, Video Sharing, Social Networking and More

topics: google

popularity: authority: activity: frequency: 11.2 posts per week

GigaOM

Business, Internet, Technology & Strategy

topics: aster data system, gigaom, initial public offering, mobile internet devices, obama for america, private equity group, sponsor, terms of service, venture capital firm, windows mobile marketplace

popularity: authority: activity: frequency: 60.9 posts per week

Mashable!

Social software and social networking 2.0.

topics: conan o brien, dallas fort worth, ding dong ditch, dog s life, facebook, google image search, google street view, natural language processing, pirate bay trial, twitter

popularity: authority: activity: frequency: 107.7 posts per week

ReadWriteWeb

topics: blackberry enterprise server, google, google friend connect, search, time to learn, twitter, user, wall st journal, wall street journal, web office suite

popularity: authority: activity: frequency: 73.7 posts per week

topic significance

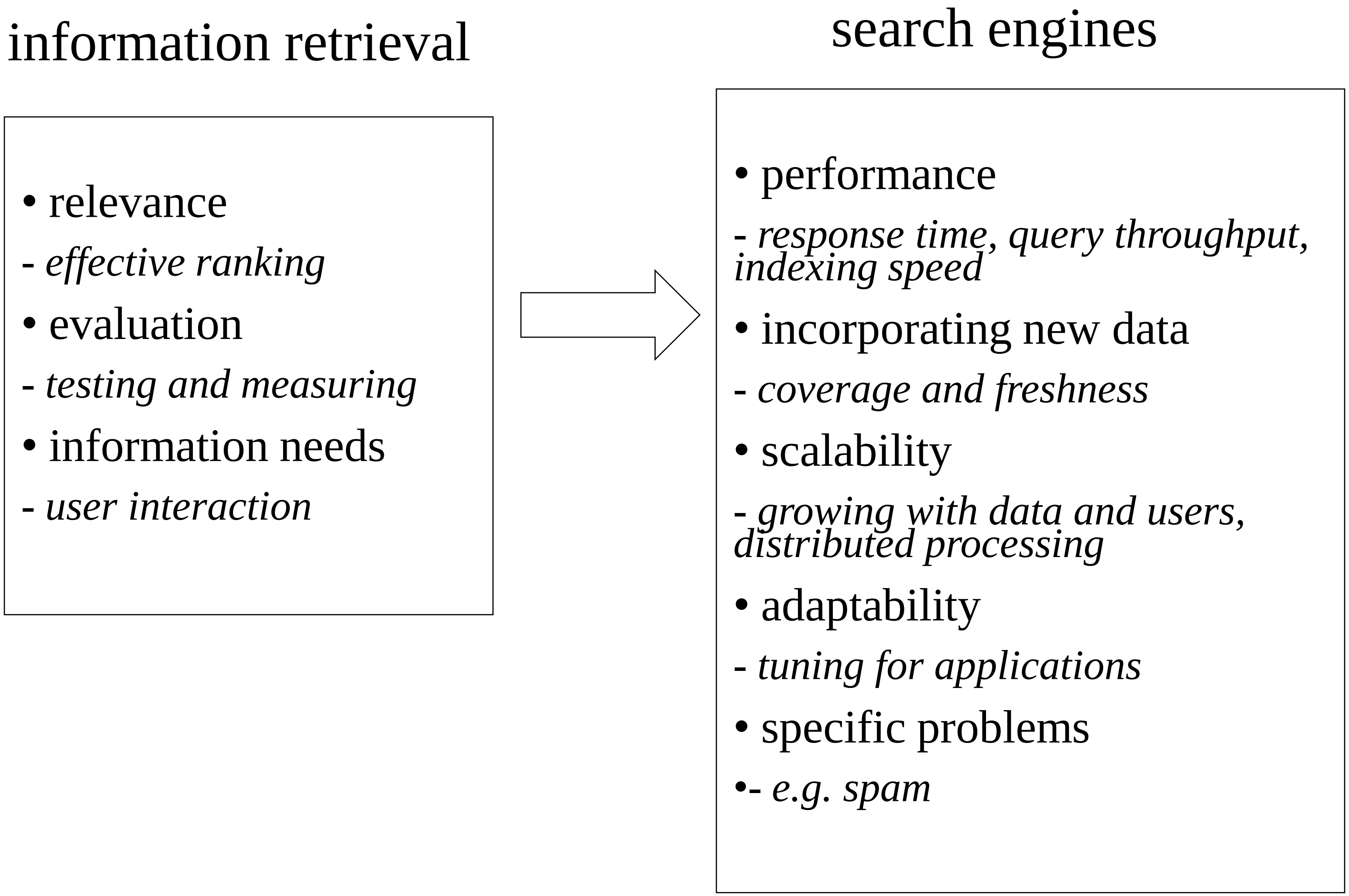
- web 2.0
- 2.0
- web
- search
- social
- twitter
- network
- google
- site
- user

more

IR and search engines

- a search engine is the practical application of information retrieval techniques to **large scale** text collections
- web search engines are best-known examples, but many others
  - Google web search engine must be able to crawl many **terabytes of data**, and then provide **subsecond response times** to millions of queries submitted everyday from around the world
  - open source search engines are important for R & D
    - e.g., Lucene, Lemur/Indri, Galago

IR and search engines



information retrieval

- relevance

- *effective ranking*

- evaluation

- *testing and measuring*

- information needs

- *user interaction*

search engines

- performance

- *response time, query throughput, indexing speed*

- incorporating new data

- *coverage and freshness*

- scalability

- *growing with data and users, distributed processing*

- adaptability

- *tuning for applications*

- specific problems

•- *e.g. spam*

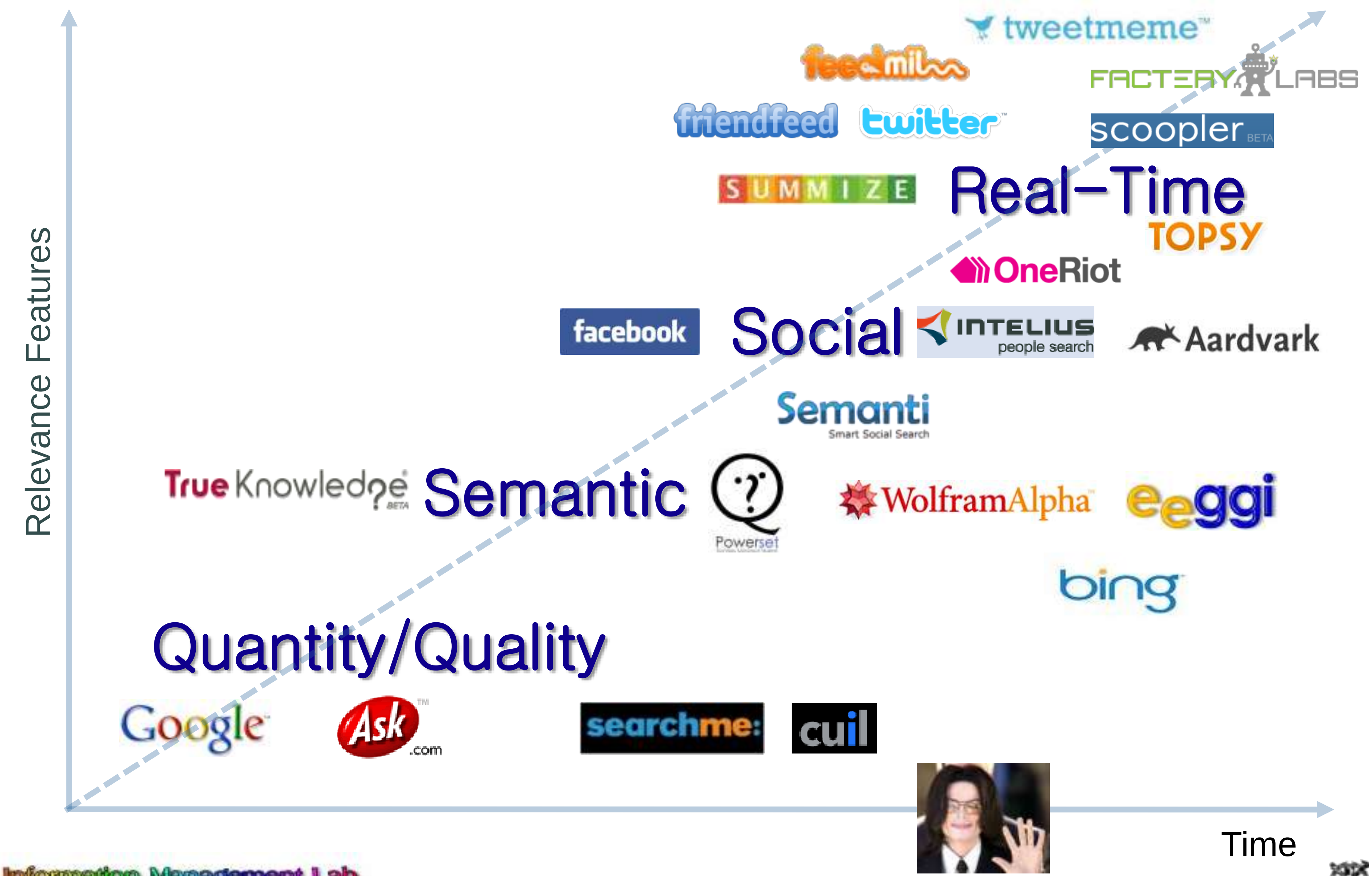
search engine issues

- **indexes** are data structures designed to improve search efficiency
  - designing and implementing them are major issues for search engines
- **dynamic** data
  - “collection” for most real applications is **constantly changing** in terms of updates, additions, deletions
    - e.g., web pages
  - acquiring or “crawling” the documents is a major task
    - typical measures are **coverage** (how much has been indexed) and **freshness** (how recently was it indexed)
  - updating the indexes while processing queries is also a design issue

spam

- for web search, spam in all its forms is one of the major issues
- affects the efficiency of search engines and, more seriously, the effectiveness of the results
- many types of spam
  - e.g. spamdexing or term spam, link spam, “optimization”
- new subfield called **adversarial IR**, since spammers are “adversaries” with different goals

search engine mega trend



Real–Time

Social

Semantic

Quantity/Quality

Relevance Features

Time

Information Management Lab

real-time search examples

real-time search engines

twitter™

bing tweets BETA

friendfeed

facebook

OneRiot

COLLECTA beta

TOPSY

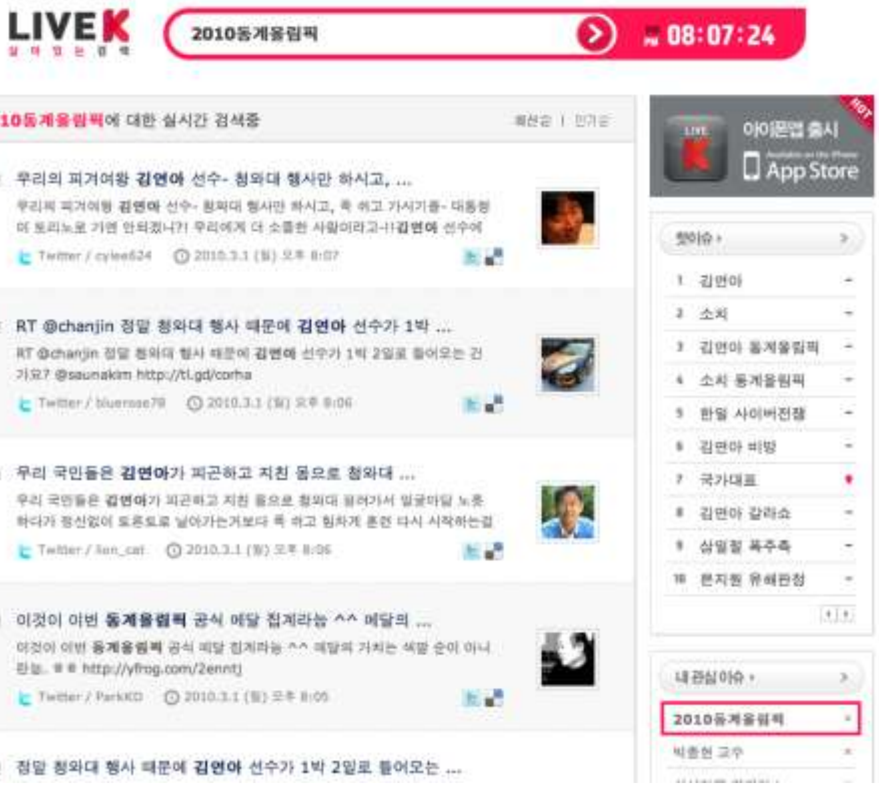
scoopler BETA

twazzup

beta crowdeye

pulse

almost.at



LIVEK
실시간 검색

2010동계올림픽

08:07:24

2010동계올림픽에 대한 실시간 검색중 최신순 | 인기순

27초전 우리의 피겨여왕 김연아 선수- 청와대 행사만 하시고, ...
우리의 피겨여왕 김연아 선수- 청와대 행사만 하시고, 푹 쉬고 가시기를- 대통령이 토리노로 가면 안되겠니?! 우리에게 더 소중한 사람이라고-!!김연아 선수에
Twitter / cylee624 2010.3.1 (월) 오후 8:07

59초전 RT @chanjin 정말 청와대 행사 때문에 김연아 선수가 1박 ...
RT @chanjin 정말 청와대 행사 때문에 김연아 선수가 1박 2일로 들어오는 건가요? @saunakim http://tl.gd/corha
Twitter / bluerose79 2010.3.1 (월) 오후 8:06

1분전 우리 국민들은 김연아가 피곤하고 지친 몸으로 청와대 ...
우리 국민들은 김연아가 피곤하고 지친 몸으로 청와대 불려가서 얼굴마담 노릇 하다가 정신없이 토론토로 날아가는거보다 푹 쉬고 힘차게 훈련 다시 시작하는걸
Twitter / lion_cat 2010.3.1 (월) 오후 8:06

1분전 이것이 이번 동계올림픽 공식 메달 집계라는 ^^ 메달의 ...
이것이 이번 동계올림픽 공식 메달 집계라는 ^^ 메달의 가치는 색깔 순이 아니란늘.. ㅎㅎ http://yfrog.com/2enntj
Twitter / ParkKD 2010.3.1 (월) 오후 8:05

2분전 정말 청와대 행사 때문에 김연아 선수가 1박 2일로 들어오는 ...

핫이슈 ›

1	김연아
2	소치
3	김연아 동계올림픽
4	소치 동계올림픽
5	한일 사이버전쟁
6	김연아 비방
7	국가대표
8	김연아 갈라쇼
9	삼일절 폭주족
10	본지원 유해판정

내 관심이슈 ›

2010동계올림픽

박종현 교수

실시간 검색중

history of search services in Korea

2001 NAVER 통합검색

2002 NAVER 지식검색

2003 empas 지식검색

2004 Google 한국어 검색

2005 Daum 신지식 검색 empas 열린검색

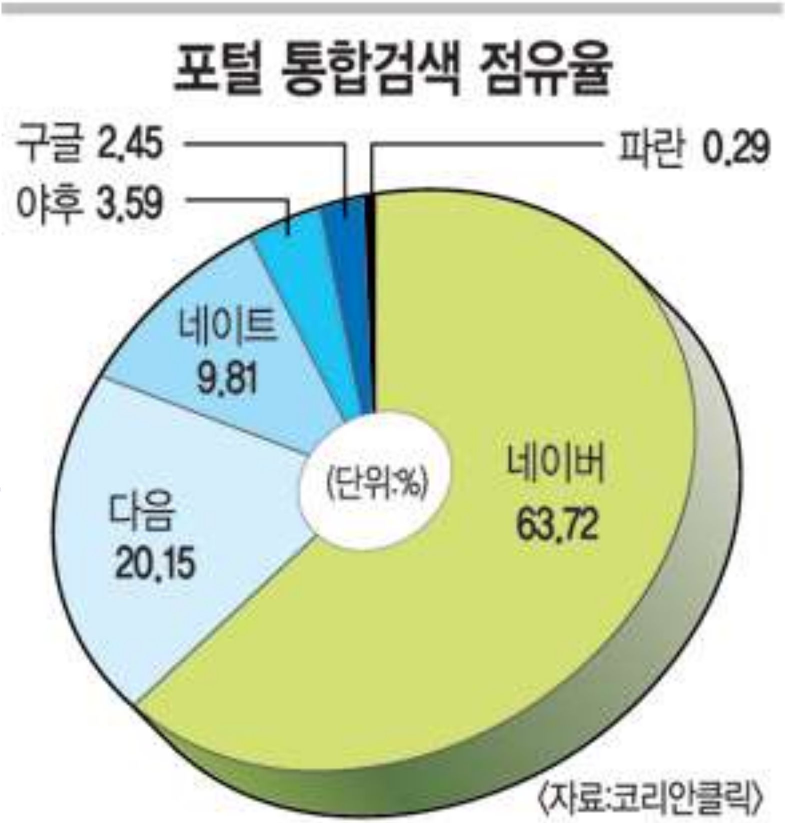
2006 블로그, 까페, 게시판 검색

2007 동영상검색 Daum 자체검색엔진 개발

2008 empas NATE 합병

2009 NATE 시맨틱 검색 Daum 소셜 검색

2010 LIVEK 살아있는 검색 실시간 검색



IR related conferences

- SIGIR: ACM Special Interest Group on IR
- WWW: World Wide Web Conference
- ECIR: European Conference on IR
- CIKM: Conference on Information and Knowledge Management
- WSDM: Web Search and Data Mining Conference
- WISE: Web Information System Engineering
- ICWE: International Conference on Web Engineering
- HT: Hypertext and Hypermedia