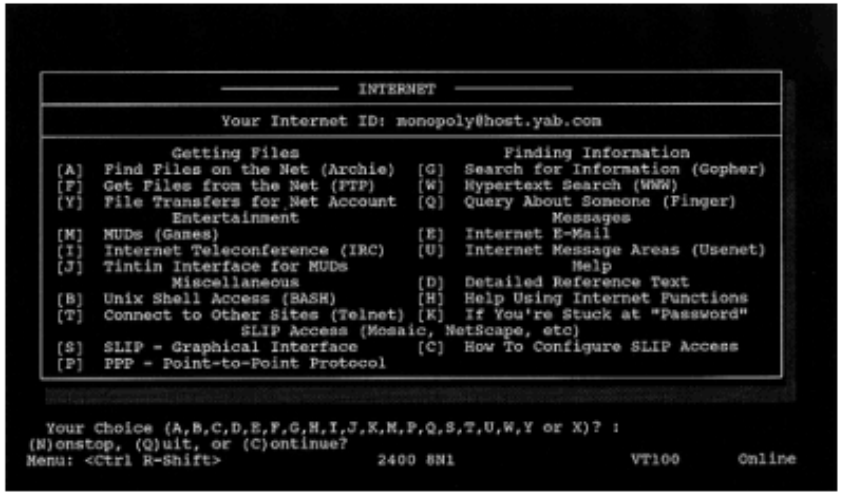
Web Search Engines

UNIVERSITY OF
TORONTO

Brief History of Search Engines

- Past
  - Before browsers
    - Gopher
  - Before the bubble
    - Altavista
    - Lycos
    - Infoseek
    - Excite
    - HotBot
  - After the bubble
    - Yahoo
    - Google
    - Microsoft



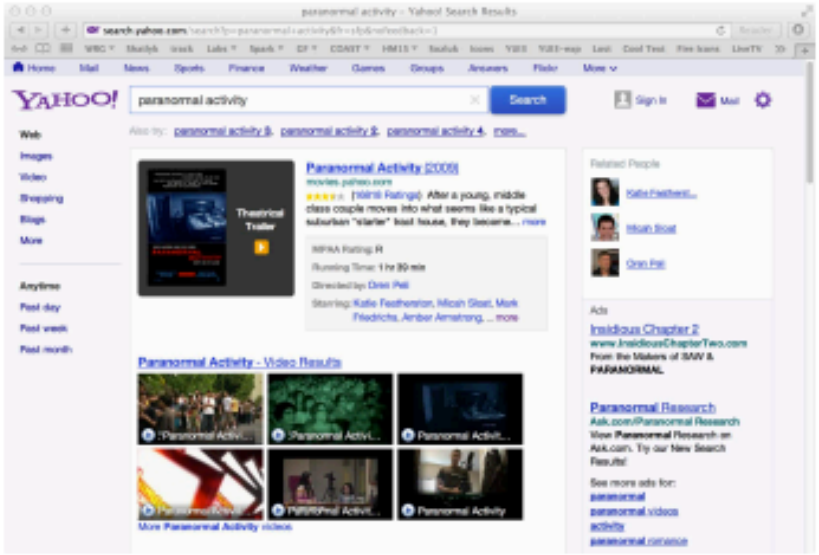
- Current
  - Global
    - Google, Bing
  - Regional
    - Yandex, Baidu

- Future
  - Facebook ?
  - …

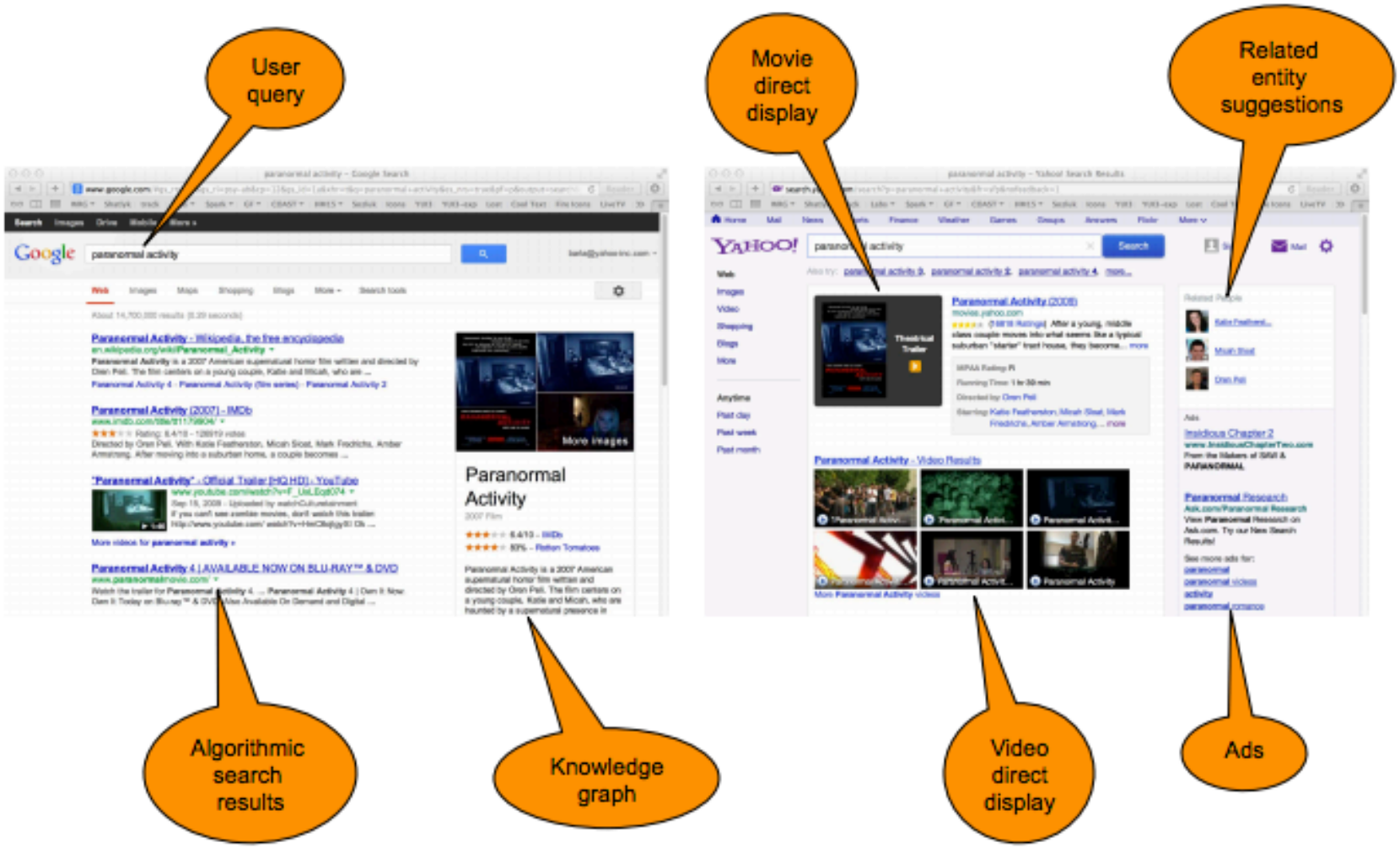
UNIVERSITY OF TORONTO

Anatomy of a Search Engine Result Page

- Web search results
- Direct displays (vertical search results)
  - image
  - video
  - local
  - shopping
  - related entities
- Query suggestions
- Advertisements



Anatomy of a Search Engine Result Page



UNIVERSITY OF TORONTO

Actors in Web Search

- User's perspective: accessing information
  - relevance
  - speed

- Search engine's perspective: monetization
  - increase the ad revenue
  - attract more users
  - reduce the operational costs

- Advertiser's perspective: publicity
  - attract more users
  - pay little

What Makes Web Search Difficult?

- Size

- Diversity

- Dynamicity

- All of these three features can be observed in
  - the Web
  - web users

What Makes Web Search Difficult?

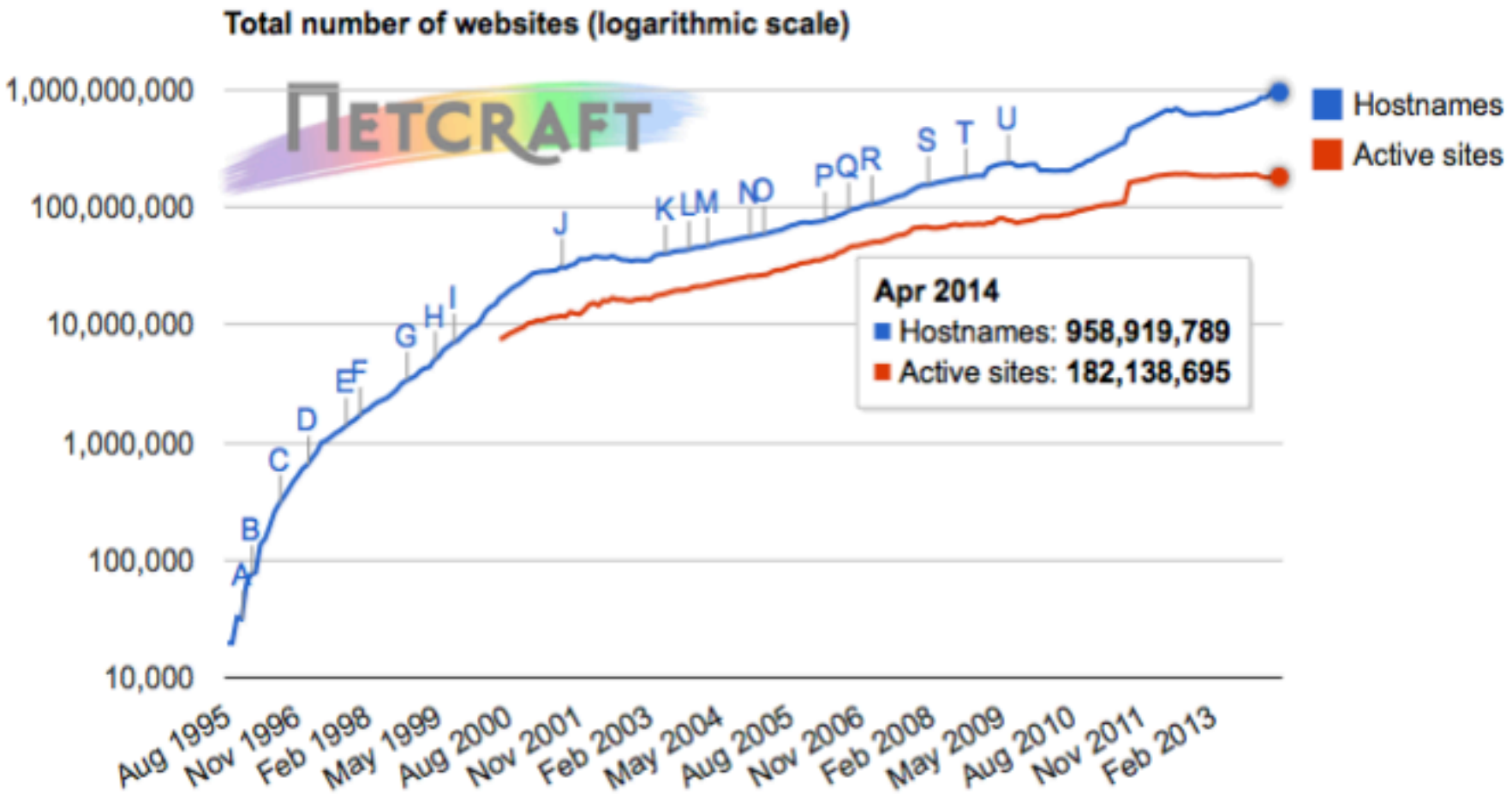
- The Web
  - more than 180 million Web servers and 950 million host names
    - compare with almost 1 billion computers directly connect to Internet
  - the largest data repository (estimated as 100 billion pages)
  - constantly changing
  - diverse in terms of content and data formats

- Users
  - too many! (over 2.5 billion at the end of 2012)
  - diverse in terms of their culture, education, and demographics
  - very short queries (hard to understand the intent)
  - changing information needs
  - little patience (few queries posed & few answers seen)

Expectations from a Search Engine

- Crawl and index a large fraction of the Web
- Maintain most recent copies of the content in the Web
- Scale to serve hundreds of millions of queries every day
- Evaluate a typical query under several hundred milliseconds
- Serve most relevant results for a user query

Web Growth



UNIVERSITY OF TORONTO

Search Data Centers

- Quality and performance requirements imply large amounts of compute resources, i.e., very large data centers
- High variation in data center sizes
  - hundreds of thousands of computers
  - a few computers

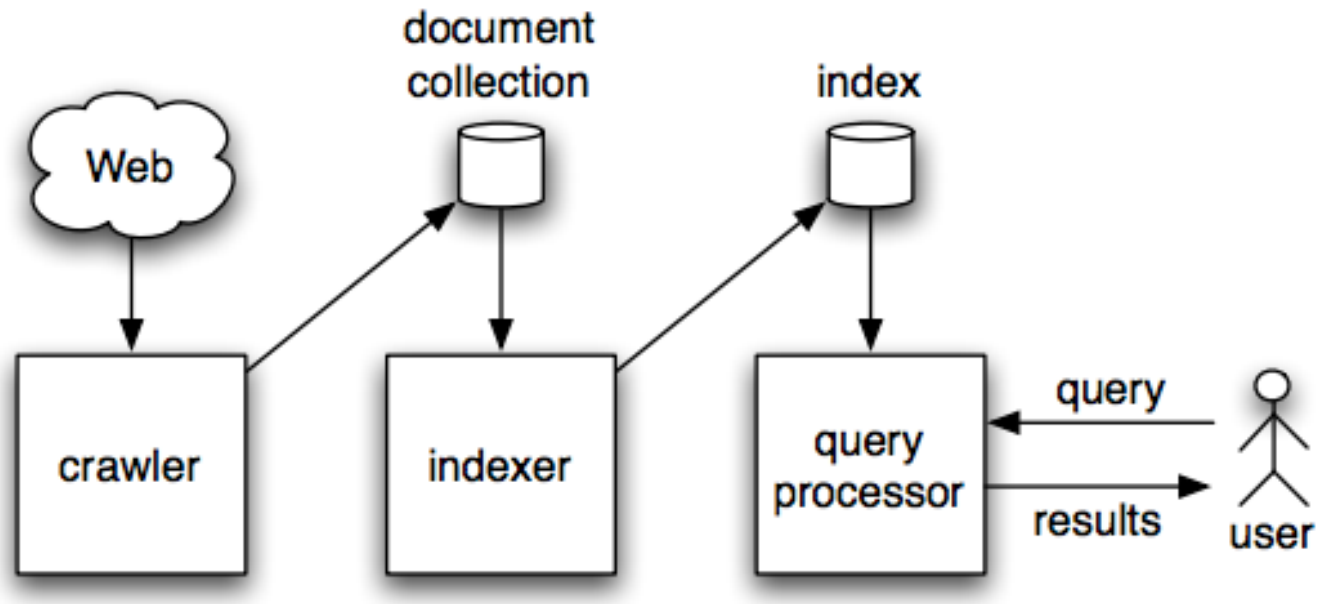
UNIVERSITY OF TORONTO

Cost of Data Centers

- Data center facilities are heavy consumers of energy, accounting for between 1.1% and 1.5% of the world's total energy use in 2010.
- depreciation: old hardware need to be replaced
- maintenance: failures need to be handled
- operational: energy spending need to be reduced

Major Components in a Web Search Engine

- Web crawling
- Indexing
- Query processing



Web Crawling

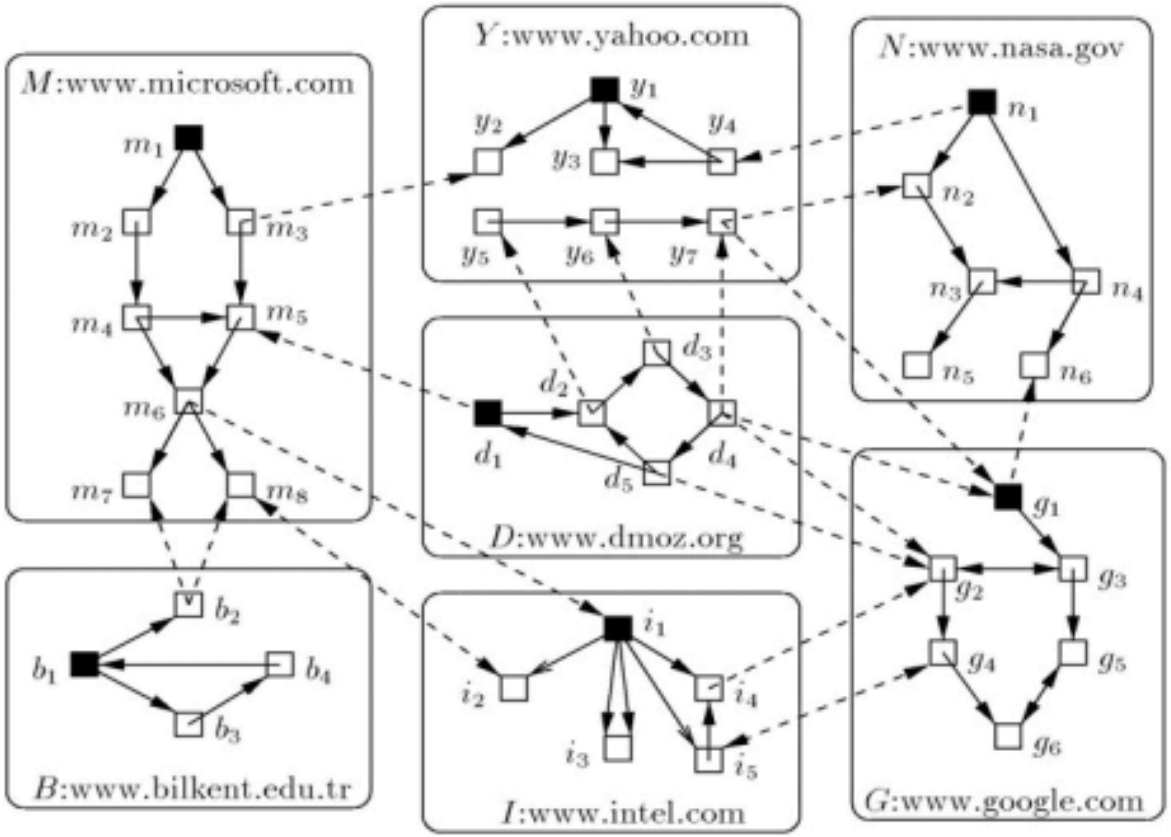
• Web crawling is the process of locating, fetching, and storing the pages available in the Web

• Computer programs that perform this task are referred to as

- crawlers
- spider
- harvesters

Web Graph

- Web crawlers exploit the hyperlink structure of the Web



Web Crawling Process

- A typical Web crawler
  - starts from a set of seed pages,
  - locates new pages by parsing the downloaded seed pages,
  - extracts the hyperlinks within,
  - stores the extracted links in a fetch queue for retrieval,
  - continues downloading until the fetch queue gets empty or a satisfactory number of pages are downloaded.

Issues in Web Crawling

- Dynamics of the Web
  – Web growth
  – content change

- Malicious intent
  – hostile sites (e.g., spider traps, infinite domain name generators)
  – spam sites (e.g., link farms)

- Web site properties
  – sites with restricted content (e.g., robot exclusion),
  – unstable sites (e.g., variable host performance, unreliable networks)

Robot Exclusion Protocol

- A standard from the early days of the Web
- A file (called robots.txt) in a web site advising web crawlers about which parts of the site are accessible
- Crawlers often cache robots.txt files for efficiency purposes

```
User-agent: googlebot          # all services
Disallow: /private/            # disallow this directory

User-agent: googlebot-news     # only the news service
Disallow: /                    # on everything

User-agent: *                  # all robots
Disallow: /something/          # on this directory

User-agent: *                  # all robots
Crawl-delay: 10                # wait at least 10 seconds

Disallow: /directory1/                 # disallow this directory
Allow: /directory1/myfile.html         # allow a subdirectory

Host: www.example.com                  # use this mirror
```

Published Web Crawler Architectures

- Bingbot: Microsoft's Bing web crawler
- FAST Crawler: Used by Fast Search & Transfer
- Googlebot: Web crawler of Google
- PolyBot: A distributed web crawler
- RBSE: The first published web crawler
- WebFountain: A distributed web crawler
- WebRACE: A crawling and caching module
- Yahoo Slurp: Web crawler used by Yahoo Search

Open Source Web Crawlers

- DataparkSearch: GNU General Public License.
- GRUB: open source distributed crawler of Wikia Search
- Heritrix: Internet Archive's crawler
- ICDL Crawler: cross-platform web crawler
- Norconex HTTP Collector: licensed under GPL
- Nutch: Apache License
- Open Search Server: GPL license
- PHP-Crawler: BSD license
- Scrapy: BSD license
- Seeks: Affero general public license
- WIRE: Carlos Castillo's PhD thesis

Nodejs **crawler**

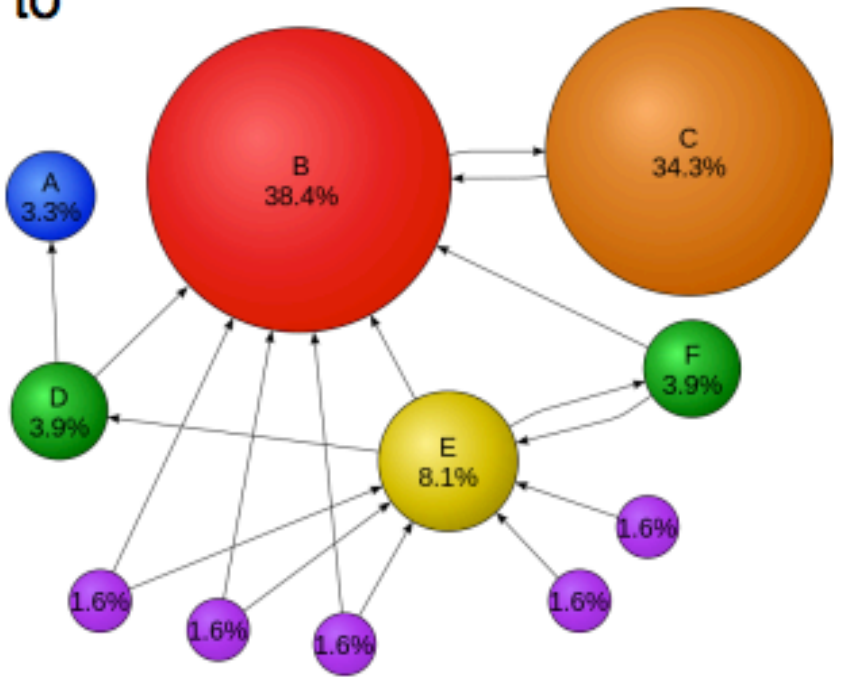
Indexing

- Indexing is the process of converting crawled web documents into an efficiently (compressed) searchable form.

- An index is a representation for the document collection over which user queries will be evaluated.

PageRank

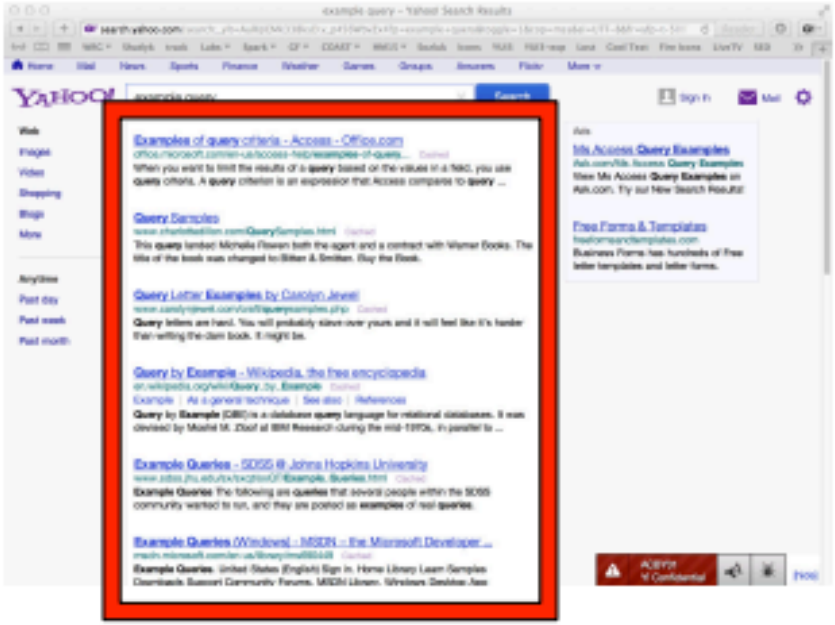
- A link analysis algorithm that assigns a weight to each web page indicating its importance
- Iterative process that converges to a unique solution
- Weight of a page is proportional to
  - number of inlinks of the page
  - weight of linking pages
- Other algorithms
  - HITS
  - SimRank
  - TrustRank



Query Processing

- Query processing is the problem of generating the best-matching answers (typically, top 10 documents) to a given user query, spending the least amount of time

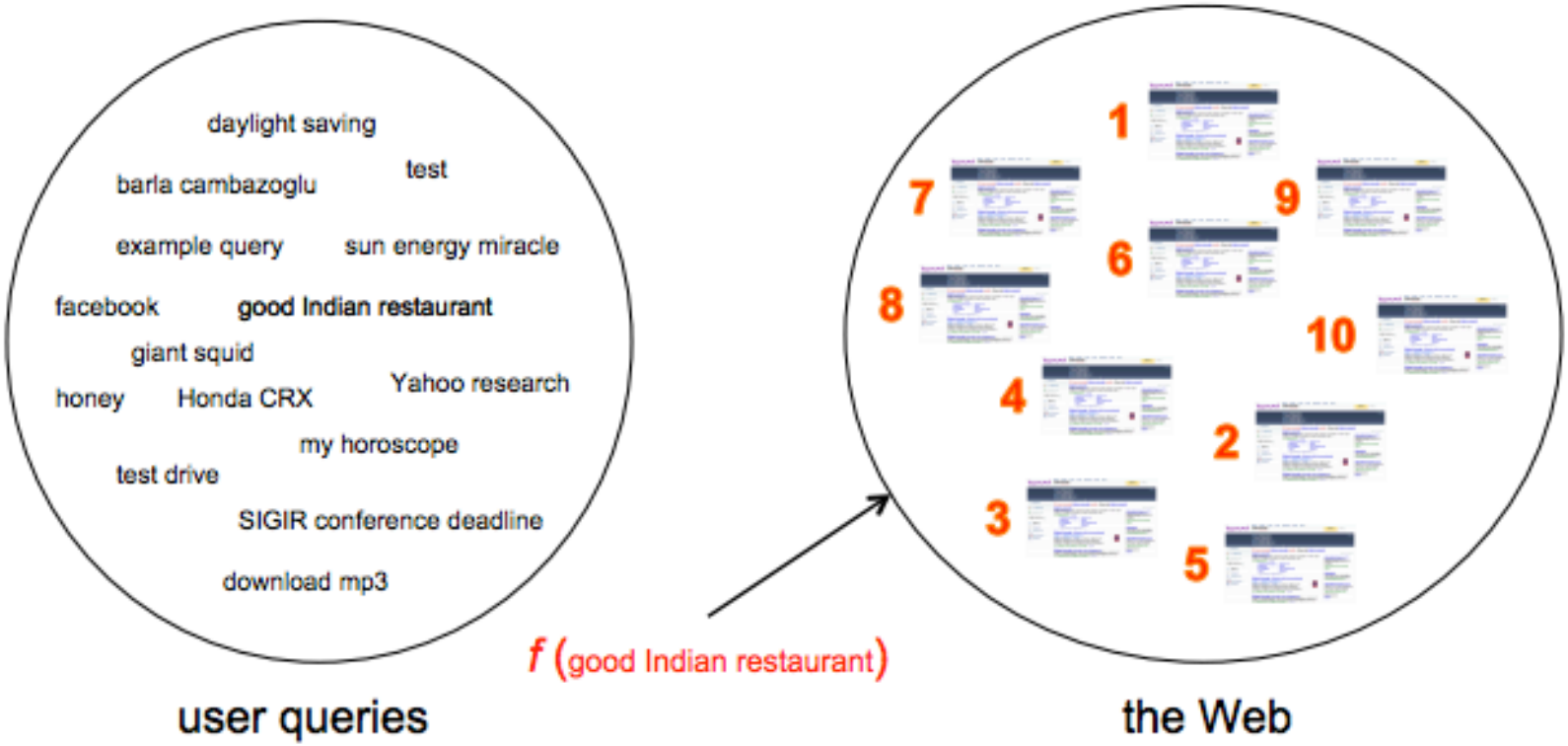
- Our focus: creating 10 blue links as an answer to a user query



UNIVERSITY OF TORONTO

Query Processing

- Web search is a sorting problem!



Metrics

- Quality metrics
  - relevance: the degree to which returned answers meet user's information need.

- Performance metrics
  - latency: the response time delay experienced by the user
  - peak throughput: number of queries that can be processed per unit of time without any degradation on other metrics

References

1. nginx.org/en/docs/http/load_balancing.html
2. http://www2014.kr/asset/slide/Scalability+and+Efficiency.pdf
3. http://www2014.kr/asset/slide/Social+Spam,+Campaigns,+Misinformation+and+Crowdturfing-www2014-tutorial.pdf
4. http://www.slideshare.net/freekbijl/web-30-explained-with-a-stamp?from=ss_embed

UNIVERSITY OF TORONTO