AALBORG UNIVERSITY

# Introduction to Web Intelligence

Peter Dolog
dolog@cs.aau.dk
http://people.cs.aau.dk/~dolog

Based (in part) on Stanford slides by Christopher Manning, Pandu Nayak & Prabhakar Raghavan and on slides provided by Bo Thiesson, AAU

# My background – your background

Web Intelligence
Recommender Systems
Personalization and User Modelling
Web Science
Web Engineering

# Course, Exercises, Hand-in, Exam…

- 11 lectures + exercises (**TUESDAYS MORNING**)
- 11 sessions (4-hour) for practical experience & making notes (**WEDNESDAYS Afternoon WITHOUT ME**)
  - Very important to get hands-on experience!
  - Content, Structure, Usage
  - Groups of 2-3
  - Exercises will support hands-on (i.e. you will start with your hands-on with a possibility to discuss with me or Felipe Costa)
  - I **DO NOT COLLECT** your notes
- No lectures and hands on sessions in week 42
- Exam
  - Oral
  - Your notes can serve you as a basis for examination

# Examples of WI applications

## Some of the big players

- Information retrieval & search
- Recommender systems
- Social network analytics
- Business intelligence from server logs, blogs. wikis. and reviews
- Crowd sourcing
- (Web-spam detection)

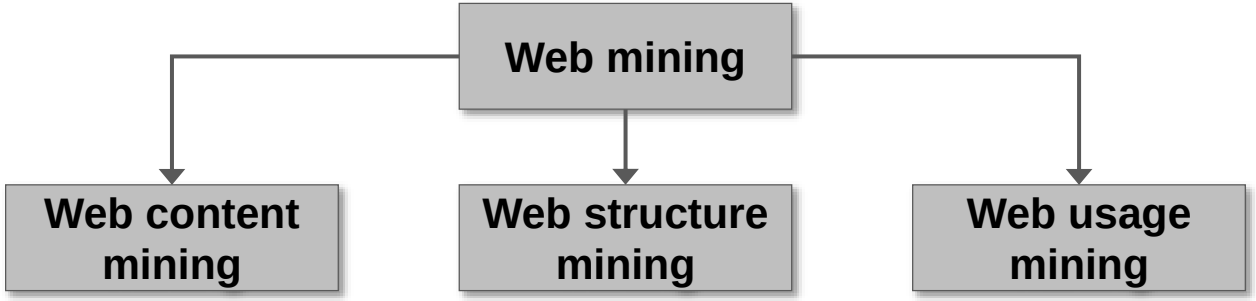
# Research areas related to WI

- Information Retrieval & Search
- Web Science
- Recommender Systems
- User Modelling and User Adapted Interaction
- Machine Learning
- Data Mining
- Natural Language Processing
- Human Computer Interaction
- Database Systems
- Distributed Systems
- Algorithms

# Web mining taxonomy

**Web data:**

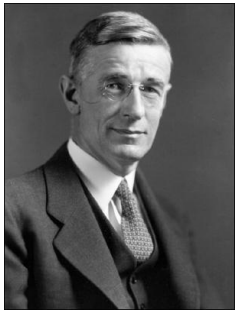
- **Content data**: the data contained in web pages (text, images, etc.)
- **Structure data**: data referring to the structure/organization of the content (e.g. hyperlinks connecting pages)
- **Usage data**: monitored information about the user's interaction with the web (e.g. web-server logs and cookies, called also implicit feedback)
- **User profile data**: demographic information about users of a web site (collected through user registrations, feedback questionairs or other explicit means – called also explicit feedback)



# Outline today

- Course overview √
- Brief history of the Web
- Web search basics
- Web spam basics
- Near-duplicate detection

# The Birth of the Web


Vannevar Bush


Tim Berners-Lee

- Vannevar Bush – 1940's – hypertext
- First working hypertext systems – 1970's
- In 1989, Tim Berners-Lee while working at CERN as independent contractor, proposed a project on the idea of ***hypertext***
- The World Wide Web (WWW) was born
- First web page online on the 6th August 1991

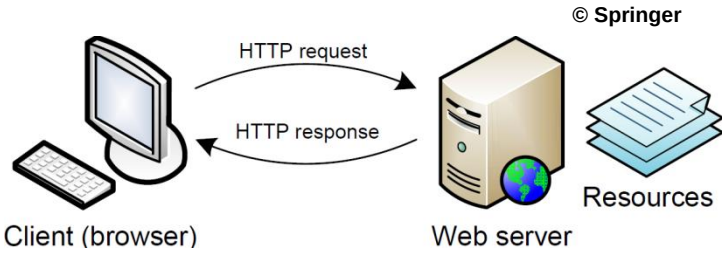

Early hypertext system

- **…20 years later**:
  - The Web has become ubiquitous
  - The Indexed Web contains close to 5 bil pages - BING(10 Sep, 2018, WorldWideWebSize.com).

# Web preliminaries

**Drivers for success**:

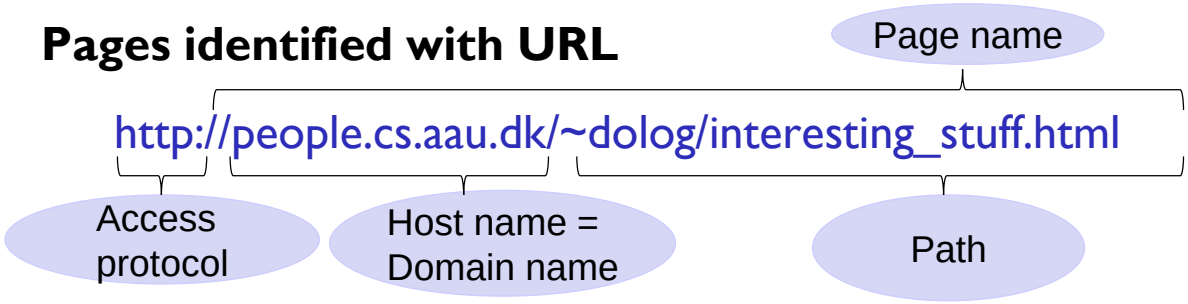
- ”Crowd sourcing” / no control on publishing
- Low barrier to entry
- No risk of ”breaking the system”




**Web pages structured in Standard HTML**

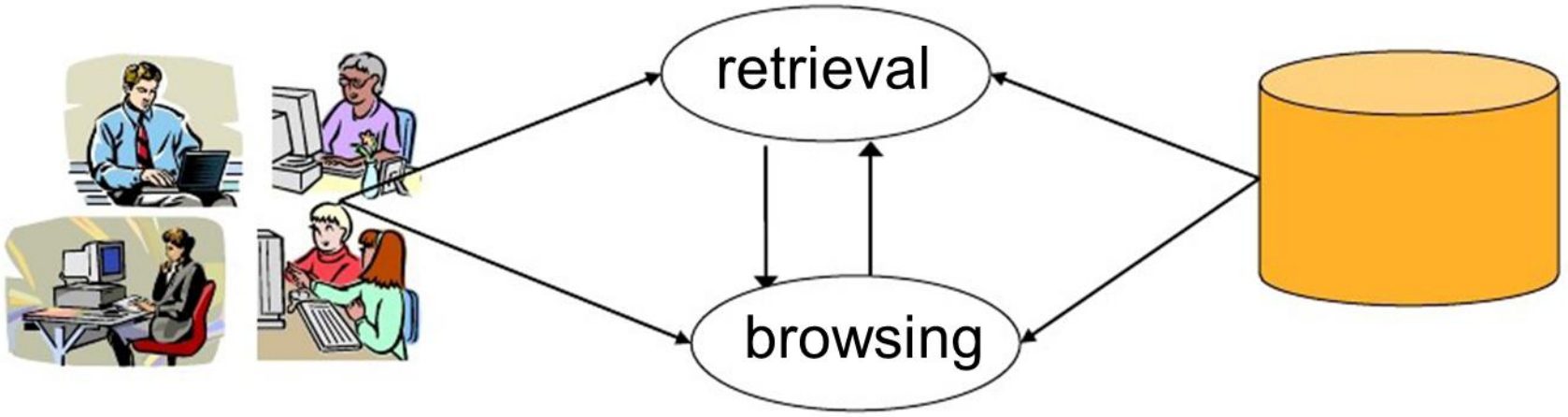
- Information content
- Hyperlinks = pointer from one page to another, loads second page if clicked on
- Formatting rules for rendering

**Pages identified with URL**




Client can address and render pages – a **browser**
The browser can ignore what he does not recognize

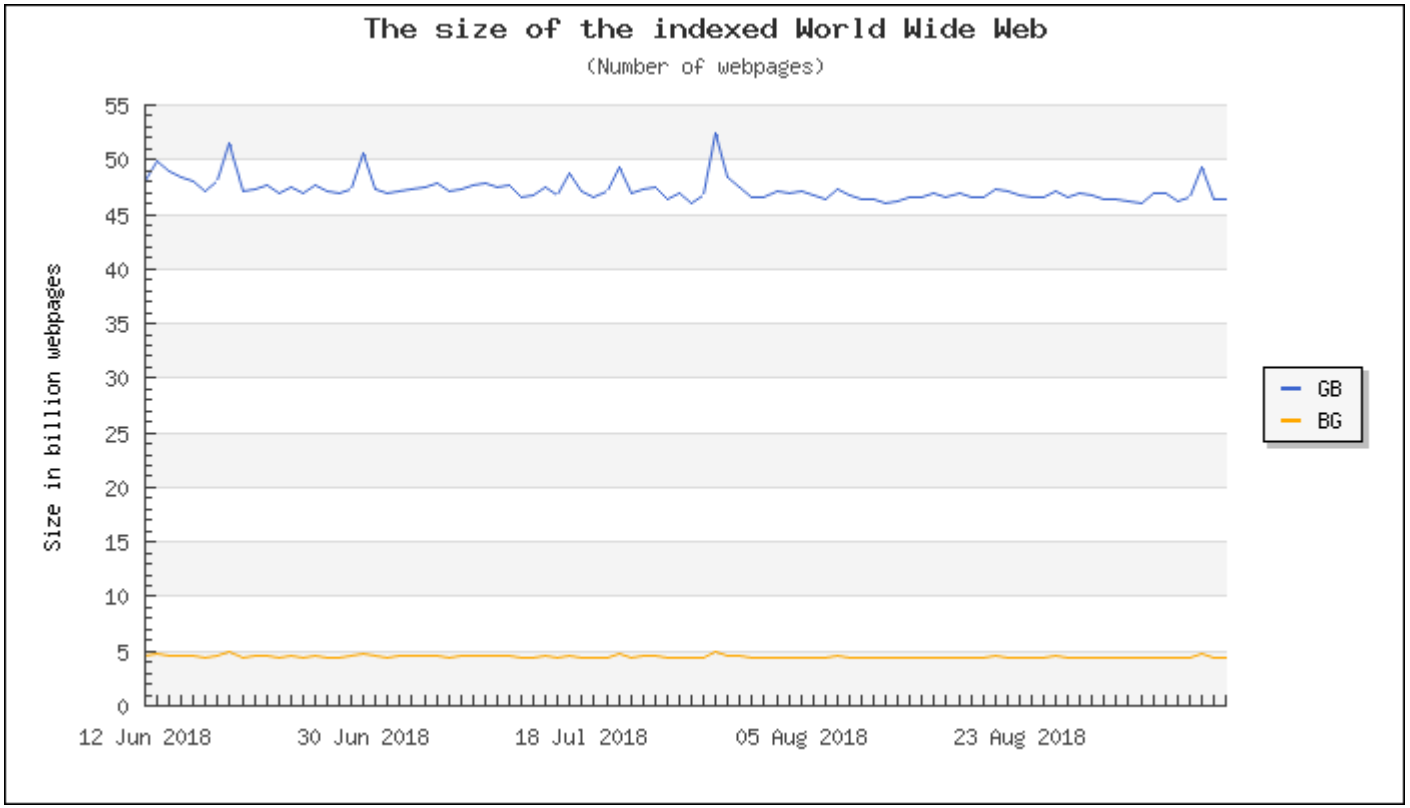
# The User Tasks




**Retrieval**: search for relevant information (usually focused)

**Browsing**: ”looking around” for information

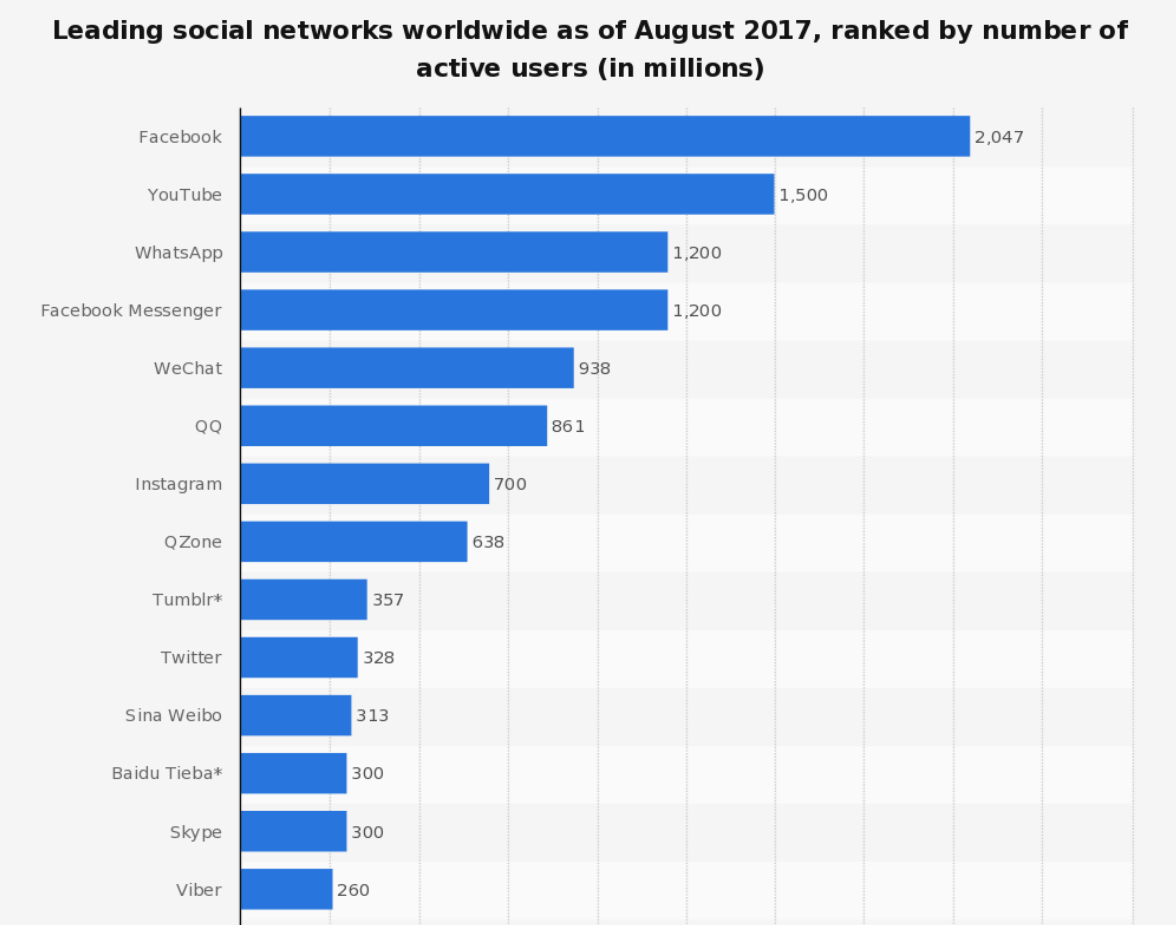
# Indexed Web



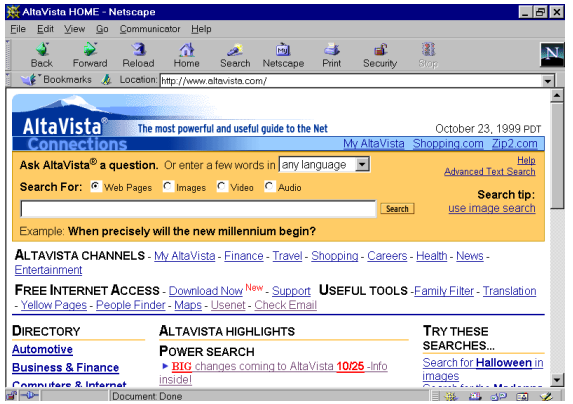

HTTP://WWW.WORLDWIDEWEBSIZE.COM/ **ON 10/9/2018**

# Social Network Active Users

HTTPS://WWW.STATISTA.COM/STATISTICS/272014/GLOBAL-SOCIAL-NETWORKS-RANKED-BY-NUMBER-OF-USERS/
**ON 15/8/2017**



# First Search Engines



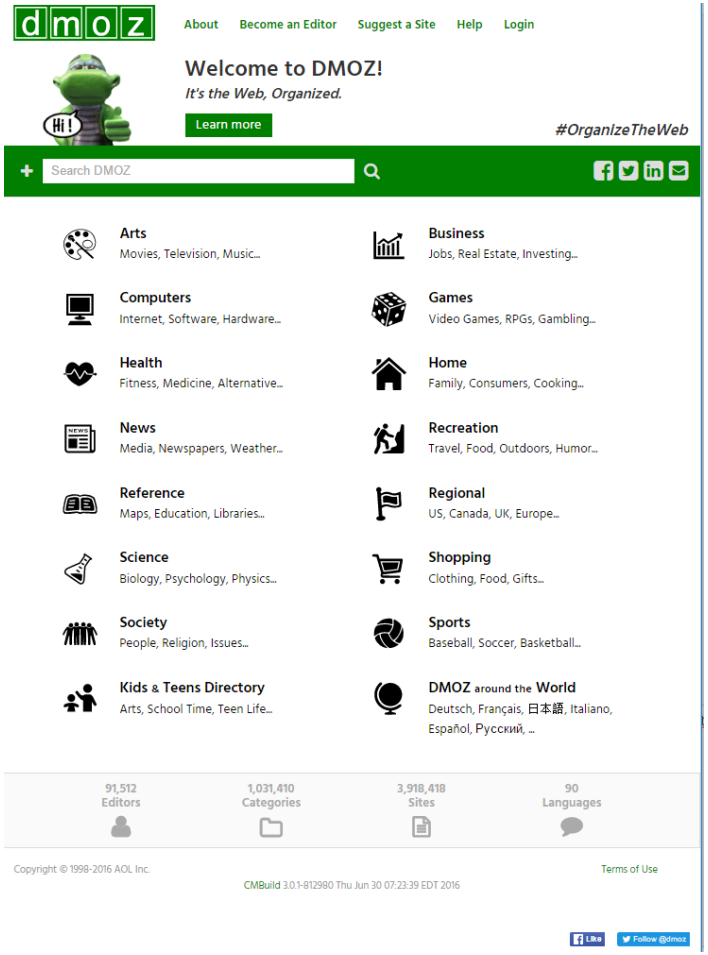

**General purpose, Full text indexed search**: Altavista, Infoseek, Excite, Google, Bing, …

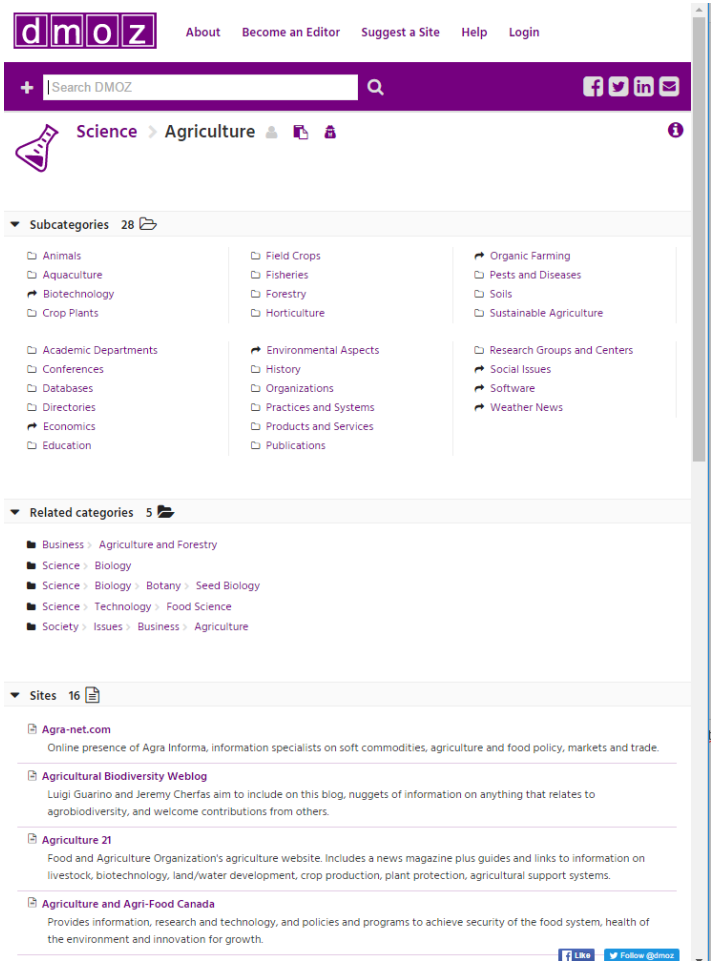
**Taxonomies populated with pages and categories**: Yahoo!

- Handmade taxonomies are hard to maintain
- User & Editors may have different taxonomic semantics
- Users may have difficulty in matching too many categories (Yahoo! reached 100 very early)
- BUT may be good in restricted domains (Shopping robots, applet finders!!)

Open directory project (ODP)… very nice but closed on March 17, 2017. Static mirror still available

# Was available at DMOZ.org (directory.mozilla.org) / ODP (Open Directory Project)
# Static mirror: http://dmoztools.net/



# Brief (Incomplete) History of Search Engines

- First web server and first html web page published
- Yahoo! (www.yahoo.com) - (1994-) directory service and search engine. - 16 mil pages online estimated
- Infoseek (1994-2001) search engine
- Inktomi (1995-) search engine infrastructure, acquired by Yahoo! 2003.
- AltaVista (1995-) search engine, acquired by Overture in 2003.
- AlltheWeb (1999-) search engine, acquired by Overture in 2003.
- Ask Jeeves (www.ask.com) - (1996-) Q&A and search engine, acquired by IAC/InterActiveCorp in 2005.
- Overture (GoTo) (1997-) pay-per-click search engine, acquired by Yahoo! 2003.
- Bing (www.bing.com) – (2009-) Microsoft rebranded search engine, was Live in 2006 and MSN search before.
- Google (www.google.com) – (1998-) – search engine.

# Example of search results

princess diana

## Engine 1

**Princess Diana Memorial WebRing**
Follow the WebRing for a tour of memorial site
**87%** http://www.geocities.com/RainForest/Vines/1009/diana
1998
Grouped results from http://www.geocities.com

**FOR DIANA, PRINCESS OF HEART - Dr. K**
...
Dr. Kate Wachs Comments on Princess Diana T
**84%** http://www.therelationshipcenter.com/diana.shtml (Si

**Princess Diana Editorial Cartoons! Cartoons**
The Professional Cartoonists Index is the most
cartoonists on the internet. www.cagle.com is th
daily cartoons. ...
**82%** http://www.cag (Size 8.8K)

**Diana, Princess of Wales**
1 July 1961 - 31 August 1997 The BBC Web sit
Camera Press/Snowdon
**79%** http://www.royal.gov.uk/start.htm (Size 2.3K) Doc
Grouped results from http://www.royal.gov.uk

**Relevant high quality**

## Engine 2

1. **Re: Lost in the shadow of Princess Diana**
[**URL:** www.spiceisle.com/talkshop/messages/6232.htm]
The Spicelslander TalkShop. [ Follow Ups ] [ Pos
The Spicelslander TalkShop ] Date: September
00:54:03 From: Sno,...
Last modified 12-Sep-97 - page size 4K - in English [ Tran

2. **Re: Princess Diana's gown auction**
[**URL:** www.elle.com/textes/blablabla/forum/messages1/15
Re: Princess Diana's gown auction. [ Follow Ups
Followup ] [ Elle International - Blablabla ] Posted
September 07, 1997 at 02:15:26:...
Last modified 30-Mar-98 - page size 2K - in English [ Tran

3. **Re: Princess Diana**
[**URL:** spicyhot.com/gaynet/messages/1053.html]
Re: Princess Diana [ Follow Ups ] [ Post Followu
Maine Gaynet Personals ] [ FAQ ] Posted by blai
November 09, 1997 at 03:35:41 In...
Last modified 30-Mar-98 - page size 2K - in English [ Tran

4. **Re: Princess Diana - Queen of Hearts**
[**URL:** www.elle.com/textes/blablabla/forum/messages1/28
Re: Princess Diana - Queen of Hearts. [ Follow U
Followup ] [ Elle International - Blablabla ] Posted
on August 31, 1997 at...
Last modified 30-Mar-98 - page size 4K - in English [ Tran

**Relevant low quality**

## Engine 3

1. Free Passwords To Adult Sites ...
**99% - Articles & General info:** Free Passwords
Sites ..... warez princess diana demi moore
magazine kathy ireland lingerie jennifer aniston cook
warez princess diana demi moore... 03/09/98
**Commercial site:** http://www.prurient.com /warez

2. SEX CHAT XXX NUDE PORNO PLAYBOY P...
[redacted]
**Personal page:** http://www.connix.com /~wgonzo
/sex/slidesuperall.htm

3. Rourke was getting [redacted]
[redacted]
**Personal page:** http://www.octet.com /~gonzo/jy

4. Sunday, 18-Jan-98
**99% - Articles & General info:** Sunday, 18-Jan-
CHAT XXX NUDE PORNO PLAYBOY PAME

**Not relevant Index pollution**

# Web Information Retrieval

- **Input**: The accessible part of Web
- **Goal**: Retrieve high quality pages that is relevant to user's need.
  - Static (files: text, audio, …)
  - Dynamic (generated on request; mostly database access)
- **Two aspects**:
  - Processing the corpus
    - Collecting static pages
    - "Learning" about dynamic pages
  - Processing the queries (searching)

# Web IR: Challenges – pages

- Scale …….. > 6 billion pages – if BING considered (WorldWideWebSize.com)
- Highly dynamic …….. Estimates 23%/day, 38%/week
- Heterogeneity
  - Type ……Text,audio,pictures,scripts,…
  - Quality …… From spam to hardcore technical documents
  - Language ……..…... 100+
- Duplication
  - Syntactic …… 30% (near) duplicates
  - Semantic ……....….. ???
- Ambiguity …….. ???
- Search engine persuation (the Search Engine Optimizer – SEO industry)

# Web IR: Challenges – users

- Not information science professionals, large variance in needs, knowledge, and mental bandwidth
- Make poor queries
  - Short (2.35 terms avg.)
  - Imprecise terms
  - Search operators rarely used
  - Low effort
- Behavior (Google reports)
  - 85% only look at first page of results (some only the top)
  - 78% of all queries are not modified
  - Follow links

# Web IR: the Combined Challenge

Retrieve high quality pages that is relevant to user's need.

given

**extreme scale and heterogeneity of Web pages**

and

**poorly made queries**

# Web IR: Evolution

**First Generation**

- IR Classical approaches applied to page content
- Scale & content diversity
- Lycos, Excite, Altavista

**Second Generation**

- Web as a graph
- Authoritativeness (substance)
- Google, Bing (Live, MSN search)

**Third Generation**

- Computational Advertisement
- Mobile Information Search
- Matching and discovery of web Services (think applets! – less about the searching and more about the getting)

# 1. Generation

# Web IR: Evolution

**First Generation**

- IR Classical approaches applied to page content
- Scale & content diversity
- Lycos, Excite, Altavista

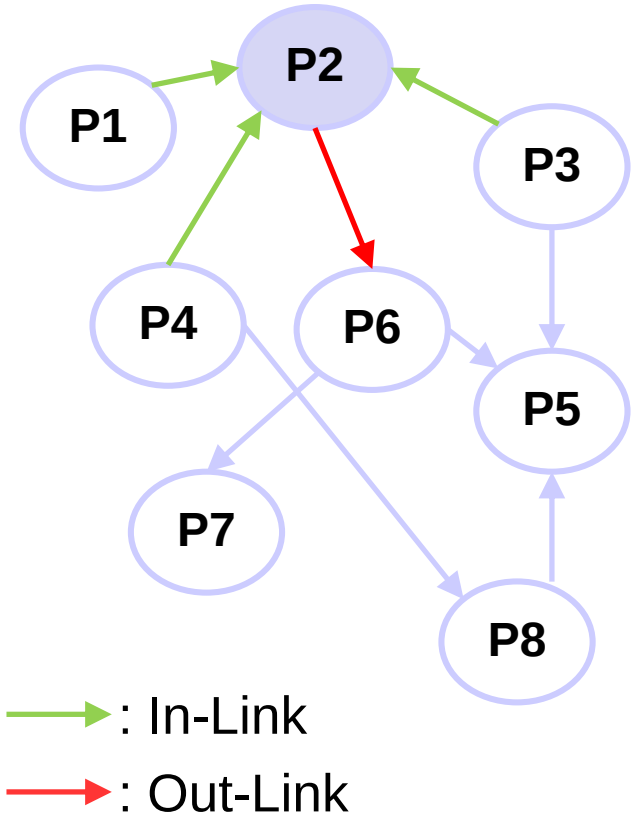
**Second Generation**

- Web as a graph
- Authoritativeness (substance)
- Google, Bing (Live, MSN search)

**Third Generation**

- Computational Advertisement
- Mobile Information Search
- Matching and discovery of web Services (think apps! – less about the searching and more about the getting)

# 2. Generation: Web Graph




**Characteristics**

- Not strongly connected
- #In-Links follows the Power Law. That is,
- (Fraction of pages with In-degree i is $1/i^{\alpha}$)
- Studies report that $\alpha = 2.1$

# Web IR: Evolution

**First Generation**

- IR Classical approaches applied to page content
- Scale & content diversity
- Lycos, Excite, Altavista

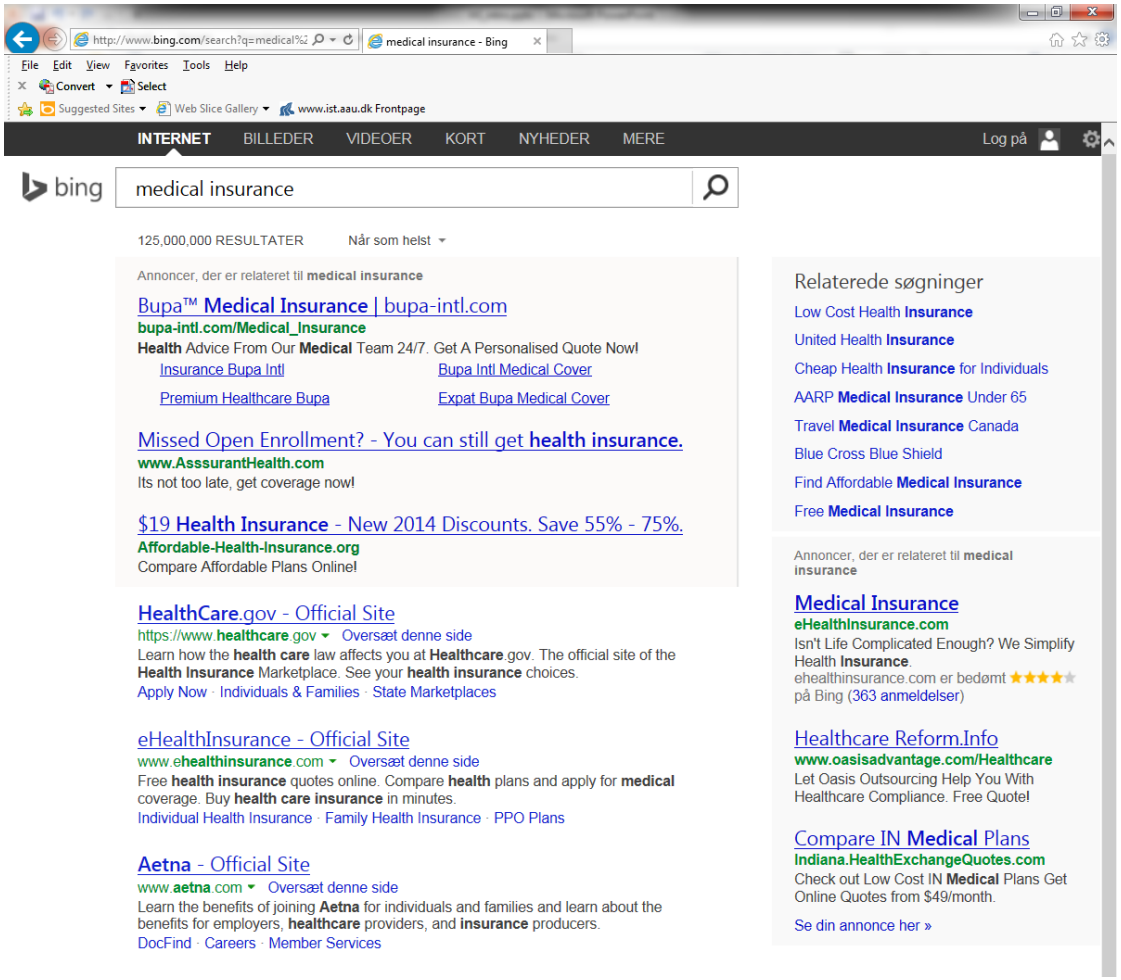
**Second Generation**

- Web as a graph
- Authoritativeness (substance)
- Google, Bing (Live, MSN search)

**Third Generation**

- Computational Advertisement
- Mobile Information Search
- Matching and discovery of web Services (think apps! – less about the searching and more about the getting)

# Third Generation: Organic & Paid search: Bing example



# Web IR: Evolution

**First Generation**

- IR Classical approaches applied to page content
- Scale & content diversity
- Lycos, Excite, Altavista

**Second Generation**

- Web as a graph
- Authoritativeness (substance)
- Google, Bing (Live, MSN search)

**Third Generation**

- Computational Advertisement
- Mobile Information Search
- Matching and discovery of web Services (think apps! – less about the searching and more about the getting)

# Web Spam (spamdexing, search spam…) - Why

- Search engines are the primary tools that people use to find information on the web
- Exclusion of a site from search engines cuts off the site from its intended audience.
- ⇒Search Engine Optimizers (SEO) is big industry!

**Goal:** Deliberate manipulation of search engine indexes

- Tricks search engine to rank relevant (commercial) web site higher than competitors.
- Tricks users to visit site that is substantially different from search engine description (e.g. delivering pornographic content cloaked within non-pornographic search results)

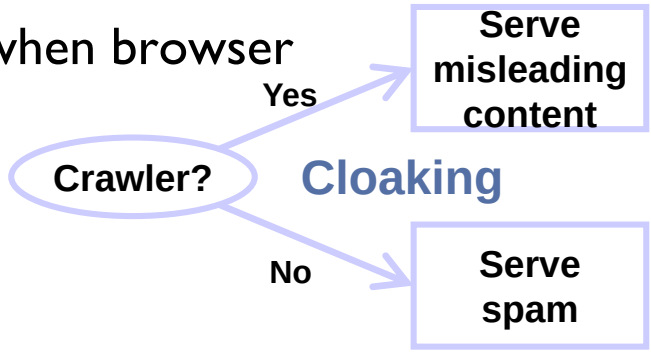
# Web Spam - How

**First Generation**

- Manipulation of web page content
- Key-word stuffing (user will not see, search indexer will see), misleading meta-tags, excessive repetition,…

**Second Generation**

- Cloaking
  - Doorway page will serve search indexer well-selected content for ranking high on selected (query) key-words.
  - Doorway page will present different content when browser connects to it.
- Web farms (link-spam)
  - Manipulation of web page authoritativeness




**Third Generation**

- ?Computational Advertisement?
- Relevance ↔ $$$
- …

# Taxonomy of Web Search

Often the user needs are not informational in nature.

[Broder, 2002] classifies web queries according to their intent into 3 classes:

- **Navigational**: the immediate intent is to reach a particular site (20%)

  q = "aalborg university"

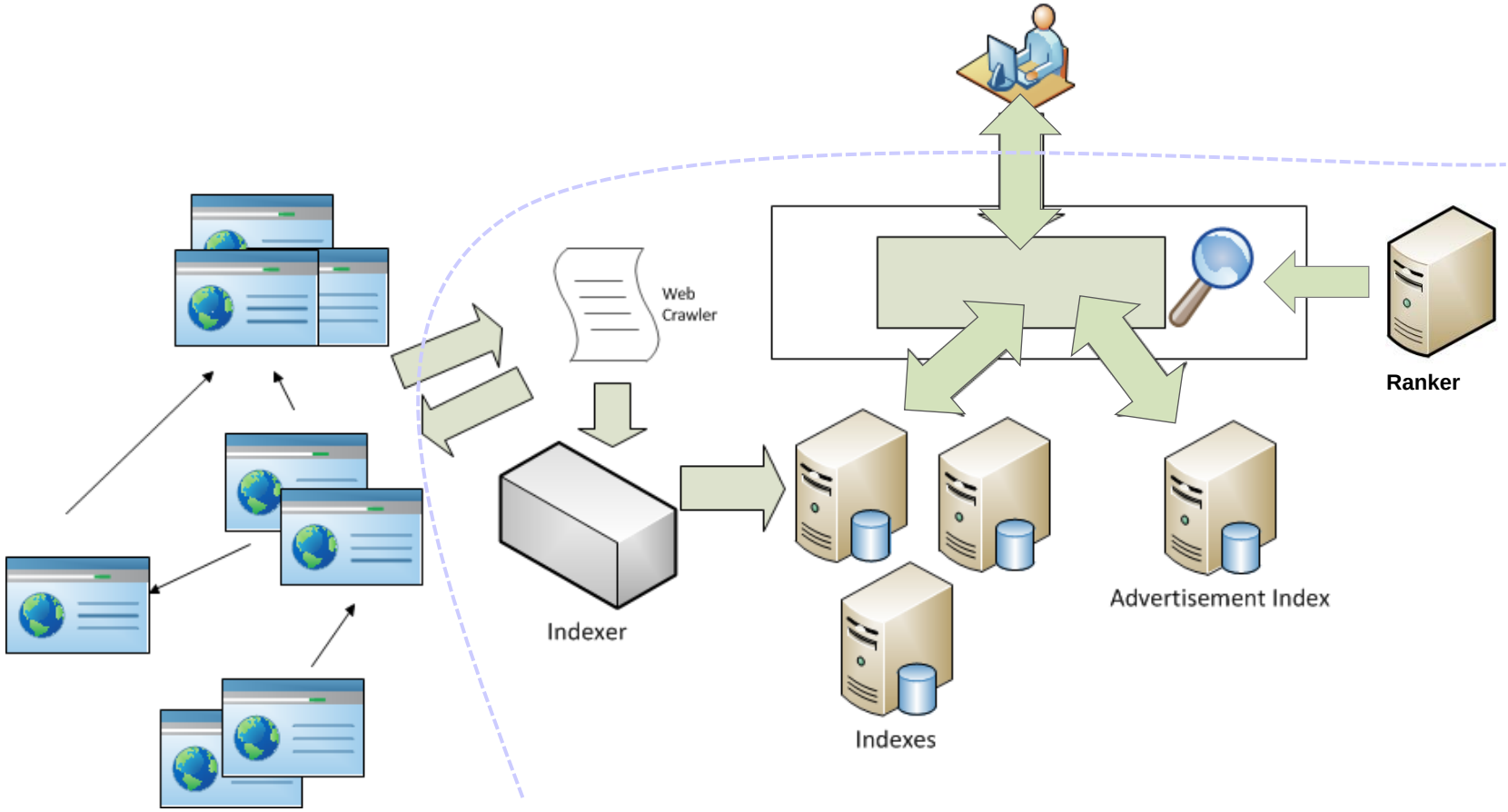
- **Informational**: the intent is to acquire some information assumed to be present on one or more web pages (50%)

  q = "hp envy review"

- **Transactional**: The intent is to perform some web-mediated activity (30%)

  q = "hotel in Barcelona"

# Search Engine Architecture



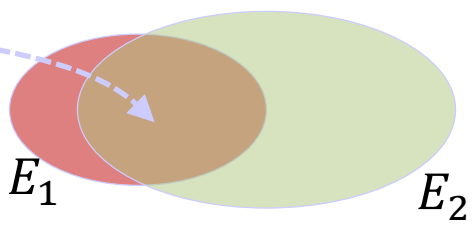
# Component Technologies for Web Search

- Query Understanding
- Document Understanding
- Query-Document Matching
- Ranking
- Crawling
- Indexing
- Search Result Presentation
- Anti-Spam
- Search Log Mining

# Index Size and Estimation

- The relative size of the index is commonly used as indirect measure of search engine's comprehensiveness/quality
- There are several issues about *what* the index covers and computation of its size
- But with some assumptions we can use **capture-recapture estimation method** (relative comparison):
  - Select randomly from one index and test whether it appears in another and vise versa – $x$ and $y$ are fractions of pages in $E_1$ and $E_2$ appearing in $E_2$ and $E_1$, respectively
  - $x|E_1| \approx y|E_2|$

$$\Rightarrow \frac{|E_1|}{|E_2|} \approx \frac{y}{x}$$



# Estimation – Pragmatic solution

**Method**:

- Build a dictionary from small set of pages crawled
- Consider conjunctive queries with 2-3 words from this dictionary
- Use the random queries on $E_1$ and pick randomly a page from top 100
- Pick 6-8 low frequency terms for query against $E_2$
- Repeat large enough number of times
- Researchers focus on improving number of biases in this approach

# (Near) Duplicate Detection

# Duplicate Detection -- Why

- By some estimates, 25-40% of the web is near-duplicate
  - Mirrors (e.g. LaTeX manual pages)
  - Same review with different boilerplates (online shopping)
  - Spam
  - …
- Web host: conserve resources
  - memory, computatations, …
- User: better experience
  - diversity, response time, …

# (Near-) Duplicate Detection

- Duplication: Exact match can be detected with fingerprints ("the hashing trick")
- Near-duplication: Approximate match
  - Compute syntactic similarity with an edit-distance measure
  - Threshold determines near-duplication
    - E.g., Similarity > 90% $\Rightarrow$ near-duplication
    - E.g., identical pages may differ only on date-time for last modification.

# Shingles
## (aka. Word N-Grams)

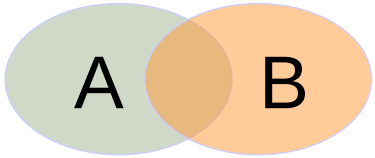
- N-Shingle = Fixed sized sequence of N sequential "words"
- E.g., 4-shingling

  **"Do not worry about your difficulties in Mathematics. I can assure you mine are still greater."**

  **Albert Einstein**

  **Do not worry about**
  **not worry about your**
  **worry about your difficulties**
  **about your difficulties in**
  **…**
- Represent document as set of N-shingles
- Intuitively, two documents are near duplicates if shingle sets are nearly the same

# Jaccard Similarity




Similarity measure between documents $A$ and $B$

$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad \left(\frac{\text{Overlap}}{\text{Union}}\right)$$

**A:** "do not worry about your difficulties in mathematics"

{do not worry, not worry about, worry about your, about your difficulties, your difficulties in, difficulties in mathematics}

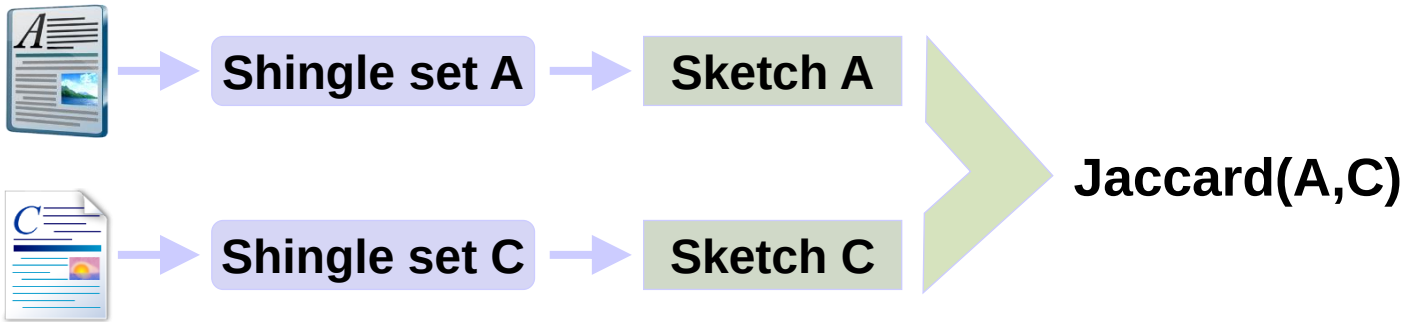
**B:** "i would not worry about your difficulties, you can easily learn what is needed."

{i would not, would not worry, not worry about, worry about your, about your difficulties, your difficulties you, difficulties you can, you can easily, can easily learn, easily learn what, learn what is, what is needed}

- Overlap: $|A \cap B| = 3$
- Union: $|A \cup B| = 15$
- Jaccard similarity: 3/15 = 0.2

# Shingling & Sketches -- Intuition

- Computing exact Jaccard similarity is relatively expensive. If done for all pairs of documents it becomes intractable!
- Trick 1: Approximate by using a cleverly chosen subset of shingles from each document (a sketch)




- Trick 2: Cheap pre-clustering: Group sketches into non-overlapping super-shingles; only compare documents that agree on super-shingles
- Algorithm due to Broder et al. (WWW '97), used in the Altavista search engine and all search engines since.

# Trick I -- Algorithm

- Hash each shingle with (64bit) hashing function:

  {do not worry, not worry about, worry about your, about your difficulties, your difficulties in, difficulties in mathematics}

  { 456, 183, 201, 123, 973, 778 }

  {i would not, would not worry, not worry about, worry about your, about your difficulties, your difficulties you, difficulties you can, you can easily, can easily learn, easily learn what, learn what is, what is needed}

  { 420, 911, 201, 123, 973, 106, 739, 205, 494, 332, 199, 380 }

- Store the minimum hash

  {123 }

  {106 }

# Trick I – Algorithm (cont.)

- Repeat many times with different hashing function (or random permutations of hash-table):

| | Min-hash1 | Min-hash2 | Min-hash3 | Min-hash4 | Min-hash5 |
|---|---|---|---|---|---|
| Doc A | 123 | 155 | 165 | 148 | 235 |
| Doc B | 106 | 210 | 166 | 148 | 155 |

- Theorem:
  - $\alpha = \text{min-hash}(A)$
  - $\beta = \text{min-hash}(B)$
  - $\Pr(\alpha = \beta) = \frac{|A \cap B|}{|A \cup B|}$
- Hence:
  - Jaccard(A,B) $\approx$ % of time the hashes agree!
  - $(= 1/5)$

Sketch
Typically 672 bytes
(84 64bit values)

# Trick2 – Super-shingles

- Problem: Doing all pairwise comparisons is still too expensive
  - For 1B documents →$10^9 \times 10^9 \times 10^2 = 10^{20}$ operations
- Solution: We only care about high similarity documents
- Sketch:

  $\{123, 155, 165, 148, 235, 174, 199, 287, \ldots, 155\}$
- Group into non-overlapping super-shingles

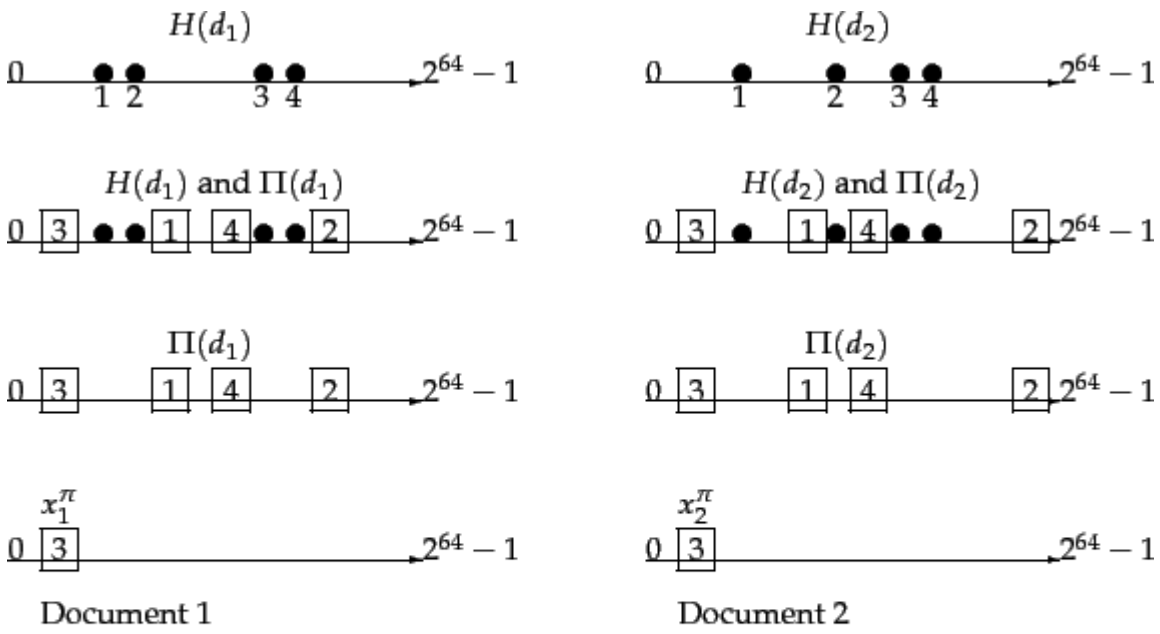
  $\{123, 155, 165, 148\}, \{235, 174, 199, 287\}, \ldots, \{\ldots, 155\}$
- Hash each super-shingle
  - $\{1003, 6505, \ldots, 8155\}$
- Use hashed set of super-shingles for cheap pre-clustering
  - E.g. same documents in same cluster if they agree on at least 2 super-shingles
  - Only compare documents in same cluster

# Real implementation

- Similarity = 90%. In a 1000 word page with shingle length = 8 this corresponds to
  - Delete a paragraph of about 50-60 words.
  - Change 5-6 random words.

- For sketch size t = 84, divide super-shingles into k = 6 groups of s = 14 samples
- Use 8 bytes hash/fingerprints → we store only 6 x 8 = 48 bytes/document
- Threshold for super-shingle similarity r = 2

# Alternative illustration of Trick 1




► **Figure 19.1** Illustration of shingle sketches. We see two documents going through four stages of shingle sketch computation. In the first step (top row), we apply a 64-bit hash to each shingle from each document to obtain $H(d_1)$ and $H(d_2)$ (circles). Next, we apply a random permutation $\Pi$ to permute $H(d_1)$ and $H(d_2)$, obtaining $\Pi(d_1)$ and $\Pi(d_2)$ (squares). The third row shows only $\Pi(d_1)$ and $\Pi(d_2)$, while the bottom row shows the minimum values $x_1^\pi$ and $x_2^\pi$ for each document.

http://nlp.stanford.edu/IR-book/html/htmledition/near-duplicates-and-shingling-1.html

# Outline today

- Brief history of the Web
- Web search basics
- Web spam basics
- Near-duplicate detection