

X-Informatics

Web Search and Text Mining Part II

2013

Geoffrey Fox

gcf@indiana.edu

<http://www.infomall.org/X-InformaticsSpring2013/index.html>

Associate Dean for Research,
School of Informatics and Computing
Indiana University Bloomington

2013

Big Data Ecosystem in One Sentence

Use **Clouds** running **Data Analytics Collaboratively**
processing **Big Data** to solve problems in
X-Informatics (or e-X)

X = Astronomy, Biology, Biomedicine, Business, Chemistry, Climate,
Crisis, Earth Science, Energy, Environment, Finance, Health,
Intelligence, Lifestyle, Marketing, Medicine, Pathology, Policy, Radar,
Security, Sensor, Social, Sustainability, Wealth and Wellness with
more fields (physics) defined implicitly
Spans Industry and Science (research)

Education: **Data Science** see recent New York Times articles
<http://datascience101.wordpress.com/2013/04/13/new-york-times-data-science-articles/>



Climate Informatics
network

How Wealth Informatics can help
with your financial freedom?



Xinformatics

xinfor
XIU TOU

Biomedical Informatics
Computer Applications in Health Care
and Biomedicine

AstroInformatics2012

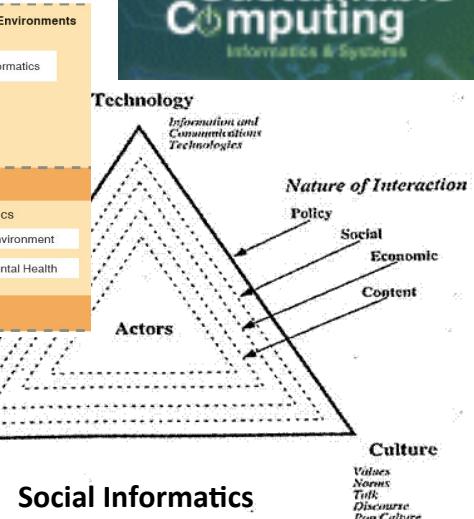
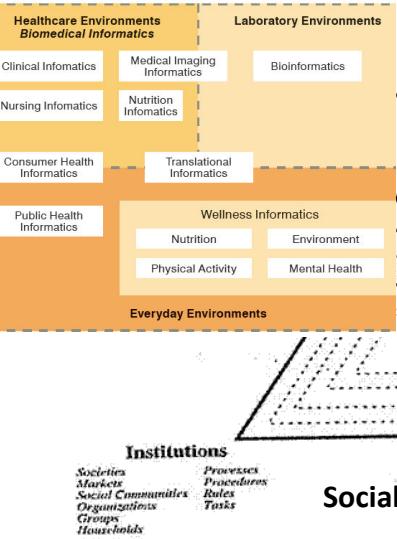
Redmond, WA, September 10 - 14, 2012

RICHARD E. NEAPOLITAN • XIA JIANG

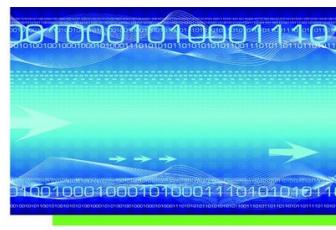
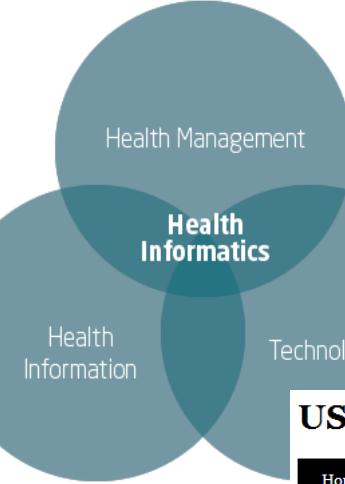
PROBABILISTIC
METHODS
FOR FINANCIAL AND
MARKETING
INFORMATICS



Sustainable
Computing
Informatics & Systems



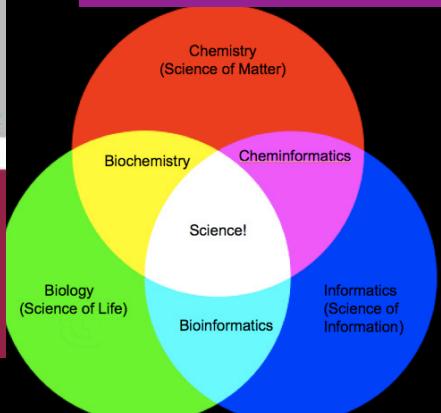
Journal of
Pathology
Informatics



Noelia Penelope Greer (Ed.)
Business Informatics
Information technology, Management,



LNCs LNCS
Intelligence and Security Informatics



Opportunities and Challenges
in Crisis Informatics

USC Center For Energy Informatics

Home Research Publications Smart Grids

GEO Informatics
Knowledge for Surveying, Mapping & GIS Professionals

About the Center

Welcome to the Center For Energy Informatics (CEI) at USC, an Organized Research Unit (ORU) housed in the [Viterbi School of Engineering](#). Energy Informatics is the application of informatics to energy systems.

Lifestyle Informatics

Applications of Lifestyle Informatics
How is the training classified
Occupation Professions
Further study
Student at the University
Watch the movie
Studying Abroad



Admission and registration
VU Honours Programme

ENVIRONMENTAL INFORMATICS



Lifestyle Informatics: Let people live longer

The study Lifestyle Informatics is about studying how people live. This bachelor including applied psychology knowledge about language and informatics can help people live longer. Lifestyle Informatics: let people live longer.



BACHELORVOORLEIDINGSDAG
ZATERDAG 3 NOVEMBER



LOOP EEN DAG MEE
MET EEN STUDENT

Data Analytics for Web Search

“Web Data Analytics”

- 1) Get the digital data (from web or from scanning)
 - 1) How to crawl web (? Solved “engineering” problem)
- 2) Pre process data to get searchable things (words, positions)
- 3) Form Inverted Index mapping words to documents
- 4) Rank relevance of documents: **PageRank**
- 5) Lots of technology for advertising, “reverse engineering”
“preventing reverse engineering”
- 6) **1) 2) 3) 5) above are easily parallel over documents**
 - **4) is a nontrivial parallel algorithm over PageRank as different PageRanks are closely coupled by iteration**
- 7) Structure of the Internet and its people and pages
(research but also commercially important)
- 8) Clustering of documents into topics (as in Google News)
 - 1) A lot of research – not clear what works
- 9) We noted value of Bayes converting Mathematics of frequency into Mathematics of belief

Document Preparation



Document Preparation

INPUT DOCUMENT



DOCUMENT TEXT

Y2K Around the World

As computers all over the world switched to 2000, few Y2K bugs were reported in several labs.
[...]



TOKENIZATION

y2k around the world as
computers all over the
world switched to 2000 few
y2k bugs were reported in
several labs [...]



FILTRATION

y2k world computers world
switched 2000 y2k bugs
reported labs [...]



DOCUMENT REPRESENTATION

y2k (2), world (2),
computer (1), switch (1),
2000 (1), bug (1), report (1),
lab (1)

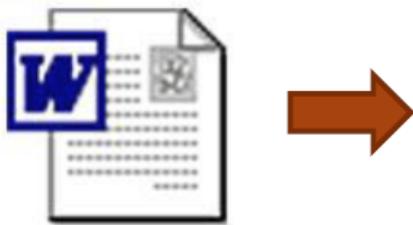


STEMMING

y2k world computer world
switch 2000 y2k bug report
lab [...]



Character Sequence Decoding



DOCUMENT TEXT

Y2K Around the World

As computers all over the world switched to 2000, few Y2K bugs were reported in several labs.

[...]

- Let's assume some document has to be indexed
- **First step: Getting a textual representation**
- Sounds easy, but might be pretty complicated
 - Many different document formats:
DOC, plain text, PDF, HTML, XLS, PPT, RTF, XML, ...

Some OCR pretty flaky!

<http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws>



Tokenization

DOCUMENT TEXT

Y2K Around the World

As computers all over the world switched to 2000, few Y2K bugs were reported in several labs.
[...]



TOKENIZATION

y2k around the world as computers all over the world switched to 2000 few y2k bugs were reported in several labs [...]

- **Tokenization:**
 - Remove formatting information (e.g. HTML tags)
 - Remove punctuation
 - Carry out basic normalization (e.g. remove capitalization)
 - **Goal:** Convert the text into a **sequence of “tokens”**



Tokenization (5)

- **Normalization** handles most of the problematic cases
- Define **equivalence classes** of character sequences that get mapped to the same token
 - U.S.A. and USA
 - naïve and naive
- Define these classes implicitly by **transformation rules**
 - **Omit all accents**
 - **Remove periods** between two characters, where there is no whitespace around (e.g. in U.S.A.)
 - Do **case folding**, i.e. reduce all letters to lower case
 - Maybe you need exceptions for names: windows ≠ Windows
(Not important, since users ask **queries in lowercase** anyway)



TOKENIZATION

y2k around the world as
computers all over the
world switched to 2000 few
y2k bugs were reported in
several labs [...]



FILTRATION

y2k world computers world
switched 2000 y2k bugs
reported labs [...]

- Removal of **stop words!**
- **Stop words:**

Extremely common words, which are of little value in selecting which documents match a user's query

- **Examples:** a, an, and, are, as, at, be, by, for, from, has, he, in, is, it, its, of, on, that, the, to, was, were, which, will, with
- “to be or not to be”?



Filtration (2)

<http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws>

- In classical IR systems, stop words have been rigorously deleted
- But stop words are needed for **phrase queries**, e.g. “King of Finland” or “As We May Think”
- For example, **Google** does not remove stop words:

The screenshot shows a Google search results page. At the top, there is a navigation bar with links for Web, Images, Maps, News, Shopping, Mail, and more. On the right side of the bar are Sign in and Preferences links. Below the navigation bar is the Google logo and a search bar containing the query "king of finland". To the right of the search bar are buttons for Search, Advanced Search, and Preferences. The main content area displays search results. The first result is a link to the Wikipedia article on the Monarchy of Finland, titled "Monarchy of Finland - Wikipedia, the free encyclopedia". It includes a snippet of text about King Charles XII of Sweden. The second result is a link to the Wikipedia article on the List of Finnish monarchs, titled "List of Finnish monarchs - Wikipedia, the free encyclopedia". It includes a snippet of text about Prince Frederick Charles of Hesse. The third result is a link to a TIME magazine article titled "Finland's King - TIME". It includes a snippet of text about Jean Julius Christian Sibelius. The search results page also shows the number of results (1 - 10 of about 19,400) and the time taken (0.18 seconds).

Web Images Maps News Shopping Mail more ▾ Sign in

Google™ "king of finland" Search Advanced Search Preferences

Web Results 1 - 10 of about 19,400 for "king of finland". (0.18 seconds)

[Monarchy of Finland - Wikipedia, the free encyclopedia](#)
(Redirected from **King of Finland**). Jump to: navigation, search ... of the late king Charles XII of Sweden, could be proclaimed as the **King of Finland** ...
en.wikipedia.org/wiki/King_of_Finland - 28k - [Cached](#) - [Similar pages](#)

[List of Finnish monarchs - Wikipedia, the free encyclopedia](#)
5 Jan 2009 ... During Svinhufvud's regency, Prince Frederick Charles of Hesse was elected as the **King of Finland** on October 9, 1918. ...
en.wikipedia.org/wiki/List_of_Finnish_rulers - 46k - [Cached](#) - [Similar pages](#)
[More results from en.wikipedia.org »](#)

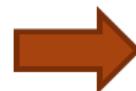
[Finland's King - TIME](#)
The old man was Jean Julius Christian Sibelius, most famous of present-day composers and "Uncrowned **King of Finland**"; the occasion was his seventieth ...
www.time.com/time/magazine/article/0,9171,758541,00.html - 37k - [Cached](#) - [Similar pages](#)



Stemming

FILTRATION

y2k world computers world
switched 2000 y2k bugs
reported labs [...]



STEMMING

y2k world computer world
switch 2000 y2k bug report
lab [...]

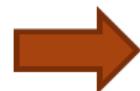
- **Lemmatization:**
The process of grouping together the different **inflected forms** of a word, e.g. walking → walk, better → good
- **Stemming:**
Trying to do lemmatization using crude heuristics, usually without knowledge of the word's context or any rules of grammar, e.g. walking → walk, but better → better



Further Normalization

FILTRATION

y2k world computers world
switched 2000 y2k bugs
reported labs [...]



STEMMING

y2k world computer world
switch 2000 y2k bug report
lab [...]

- There are **some additional** things apart from stemming that **could** be done during this step
- Take care of **umlauts** and **accents**
 - Usually, just remove them (e.g. ä → a) or transliterate them (e.g. ä → ae)
 - But be careful: unbeschränkt ≠ unbeschränkt
- Take care of **synonyms**, e.g. auto → car
 - Again, be **very careful** when doing this!



Document Representation

STEMMING

y2k world computer world
switch 2000 y2k bug report
lab [...]



DOCUMENT REPRESENTATION

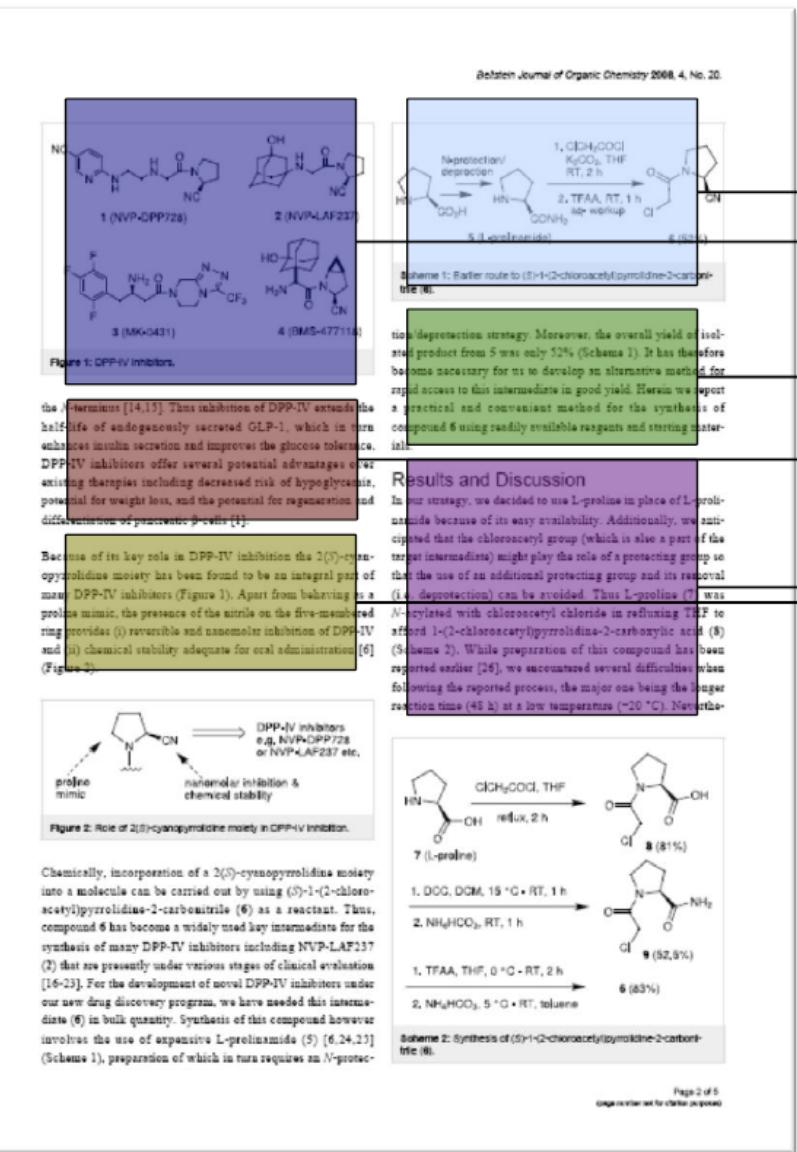
y2k (2), world (2),
computer (1), switch (1),
2000 (1), bug (1), report (1),
lab (1)

- Finally, we arrive at the **bag-of-words representation**
- The preparation of original documents is finished now, but we need to talk about **efficient data structures for managing the inverted indexes...**



Segmentation

Detour

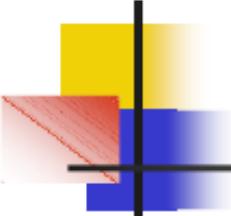


Needed in Google Scholar or CiteSeer

Textual output

<http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws>

Inverted Index



The Inverted Index Data Structure

- An Inverted Index consists of 2 elements:
 - The lexicon (AKA vocabulary, dictionary)
 - The inverted file (AKA postings file)
- The Lexicon is the set of all indexing units (terms) in a given collection
- The Inverted file is a set of *postings lists* – a list per term. The list consists of *posting elements*.
 - The list of term t holds the locations (documents + offsets therein) where t appeared. Offset = location in document for phrases
 - Additional data (*payload*) is often encoded per location
 - Many variations and degrees of freedom
- Supports efficient lookup from word to occurrences
- Essentially a way of representing and accessing (by row) a huge and very sparse matrix

Inverted Index Example

The good
1 2
the bad
3 4
and the ugly
5 6 7

Doc #1

As good as it gets, and more
1 2 3 4 5 6 7

Doc #2

Is it ugly and bad? It is, and more!
1 2 3 4 5 6 7 8 9

Doc #3

Lexicon

Term	DF	
The	1	→ (1; 1,3,6)
Good	2	→ (1; 2) – (2; 2)
Bad	2	→ (1; 4) – (3; 5)
And	3	→ (1; 5) – (2; 6) – (3; 4,8)
Ugly	2	→ (1; 7) – (3; 3)
As	1	→ (2; 1,3)
It	2	→ (2; 4) – (3; 2,6)
Gets	1	→ (2; 5)
More	2	→ (2, 7) – (3, 9)
Is	1	→ (3; 1,7)

Basic questions:

- How to efficiently build an inverted index?
- How does the inverted index efficiently support the various retrieval models (Boolean, Vector Space, etc.)?
- How does the inverted index handle complex operators, e.g. phrase queries?



Recap: Inverted Indexes

- In **Boolean retrieval**, queries have been evaluated using **inverted indexes** (aka inverted files)
- Document collection:
 - Document1 = {step, mankind}
 - Document2 = {step, China}
- Inverted index:
 - step: {Document1, Document2}
 - mankind: {Document1}
 - China: {Document2}
- Query:
 - “mankind AND step”





Indexing

Indexing:

“The process of assigning keywords to each document”

- These **keywords** are used as a (possibly intermediate) **representation** of each document
- **Some problems** we are faced with:
 - “3/12/91” vs. “Mar 12, 1991” vs. “12/3/1991”
 - “<h1>Document heading</h1>” vs. “Document heading”
 - computer vs. computers vs. Computer vs. computer’s
 - aren’t vs. are not

Index Construction



Index Construction

- **Building an inverted index looks easy:**
 1. Assign an ID to each document: **docID**
 2. Run the **document preparation** process on each document
 3. Compile a list of all **index terms**
 4. Assign an ID to each index term: **termID**
 5. Create a list of all **(termID, docID, tf) triplets**
 6. **Sort** this list: Primarily by termID, secondarily by docID
- Essentially, this corresponds to a **matrix transposition**
- Let's have a look at an example...



Example

- Our **example collection** of six documents:
 1. The old night keeper keeps the keep in the town
 2. In the big old house in the big old gown
 3. The house in the town had the big old keep
 4. Where the old night keeper never did sleep
 5. The night keeper keeps the keep in the night
 6. And keeps in the dark and sleeps in the light
- Case-folding, stopping, and stemming reduces the vocabulary to **ten index terms**:

1. big	4. house	7. night	10. town
2. dark	5. keep	8. old	
3. gown	6. light	9. sleep	



Example (2)

tf	1	2	3	4	5	6	7	8	9	10
1					3		1	1		1
2	2		1	1				2		
3	1			1	1			1		1
4					1		1	1	1	
5					3		2			
6		1			1	1			1	

- Then, we get the following (**termID**, **docID**, **tf**) triplets:

(5, 1, 3), (7, 1, 1), (8, 1, 1), (10, 1, 1), (1, 2, 2), (3, 2, 1), (4, 2, 1),
(8, 2, 2), (1, 3, 1), (4, 3, 1), (5, 3, 1), (8, 3, 1), (10, 3, 1), (5, 4, 1),
(7, 4, 1), (8, 4, 1), (9, 4, 1), (5, 5, 3), (7, 5, 2), (2, 6, 1), (5, 6, 1),
(6, 6, 1), (9, 6, 1)

Then sort by termID and then docID



Large Collections

- **Building the inverted index isn't difficult** if the whole document collection fits in **main memory**
- Now, let's get serious:
Typical collections are very large...
- What can we do?
 - **Sort-based inversion:**
Use a external (i.e. disk-based) sorting algorithm that works on compressed disk blocks (for performance reasons)
 - **Merge-based inversion:**
Read and index documents in memory until a fixed capacity is exceeded; when memory is full, the index is flushed to disk and merged with the index already stored on disk



Index Representations

- The problem of building the index essentially is solved
 - OK, Google uses massive replication and data distribution but these are very special requirements...
- Now, how to store an inverted index on disk?
- Since disk accesses are very expensive (e.g. compared to computations), there are **two major requirements**:

Keep the index as small as possible!

Read as little data as possible from disk!

- Since computational power comes at (almost) no cost, **effective data compression** is our first way to go!



Dynamic Indexing

- How to handle changes to document collection?
 - New documents
 - Updated documents
 - Deleted documents
- **Simple (but inefficient) solution:**
Rebuild the index from scratch
- **Better solution:**
Keep an auxiliary in-memory index that keeps track of all changes
 - If the auxiliary index gets too large, it is merged with the main index

Query Structure and Processing

Query Result Caching

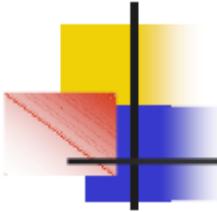
Exploits locality of reference in the query streams that are submitted to search engines:

- Query popularities vary widely, from the extremely popular to the very rare
- Although a minority, significant number of users still view multiple *result pages* per query (motivates prefetching)



Successful caching and prefetching of results can:

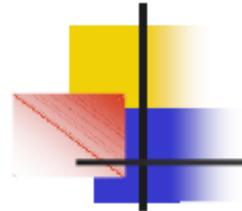
- Lower the number/cost of query executions
- Shorten the engine's response time and lower its hardware requirements



Query Log Analysis

Analysis of 7160190 queries submitted to AltaVista during September 2001:

- The 7160190 queries requested 4496503 distinct result pages, belonging to 2657410 distinct query strings
- 67% of the 2657410 strings were only requested once; the most popular string was queried 31546 times
- 48% of the result pages were only requested once; the 50 most requested pages account for almost 2% of all requests
- Qualitative behavior consistent with analyses of query logs of other engines



Result Caching and Prefetching

- Simple LRU-based caches of query results can achieve cache hit-rates of about 30% (Markatos, 2000)
- Prefetching deeper result pages, along with more complex caching policies, can achieve hit rates of over 50%
- What is the cost associated with prefetching additional results?
- What happens when the index changes?
- BTW, what about caching of other data in search engines, e.g. postings lists?

Least Recently Used (LRU): discards the least recently used items first



Query Processing

- **Boolean retrieval:**

Process queries as we already have discussed it

- **Vector space retrieval:**

Answer the query “dark, keep, night” by scanning through the postings lists for “dark,” “night,” and “keep,” while accumulating scores for each document

- Let's have a look at an example...



Vector Space Retrieval

- **Query = “dark keep night”**
- Vector representation: Simple term frequencies
 - Query vector = $(0, 1, 0, 0, 1, 0, 1, 0, 0, 0)$

- Similarity measure:
Simple scalar product

Term	Posting List
2: dark	$(6, 1)$
5: keep	$(1, 3), (3, 1), (4, 1), (5, 3), (6, 1)$
7: night	$(1, 1), (4, 1), (5, 2)$

- Process query by scanning through the three lists and add up term frequencies for each occurring document
- This gives the following final scores:

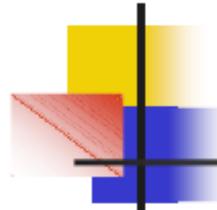
Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6
$3 + 1 = 4$	0	1	$1 + 1 = 2$	$3 + 2 = 5$	$1 + 1 = 2$



Phrase Queries

- A special type of queries are **phrase queries**
- **Example:** “King of Finland”
- **Three strategies** to process phrase queries:
 - **Postprocessing:**
Initially, ignore the word order and do Boolean retrieval;
In a second step, search through the documents found and return only the ones containing the phrase
 - **Store word positions:**
Add word positions to each posting so that the locations of terms in documents can be checked during query evaluation
 - **Partial phrase indexes:**
Create a (partial) index containing phrases

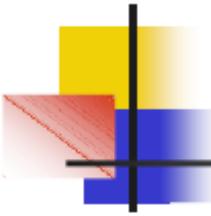
Link Structure Analysis including PageRank



The Information Need Behind the Query

[Broder, SIGIR Forum 36(2), 2002]

- **Informational** Find Topics
 - I want to learn more about the “roman empire”
 - Many possible fine results
- **Navigational**
 - Find me the home page of “el al”
 - In principle, a single correct result
 - In practice, correct result may depend on language/locale
- **Transactional**
 - I want to buy a “digital camera”
 - Good results: e-tailors that sell digital cameras



Navigational Queries & Anchor Text

- Anchor text – the highlighted clickable text of a link
- Physically in page A, but actually describes the content of page B
- Many times, concisely defines page B
 - Often, better than the text present on page B itself
- Anchor text is the most important factor in treating navigational queries
- Case study: the query “engine” on Microsoft’s Live Search
- Famous pop-culture examples: “French Military Victories”, “miserable failure”



A screen capture showing what you find when you go to Google, type in "french military victories," and click the "I'm feeling lucky" button

See Also: [Funniest Political Pictures of All Time](#)

< – [Previous Picture](#) | [Next Picture](#) – >

[BACK TO INDEX: Funny Political Pictures](#)

[More French Jokes](#)

[Email This Pic to a Friend](#) | [Subscribe to Newsletter](#)

Related Articles

- [CustomizeGoogle Firefox Extension Review - Review of CustomizeGoogle Firefo...](#)
- [Getting a Big Company to Do the Right Thing - Reader Stories: Entrepreneurs...](#)
- [Why Hyperlink Names Matter to Google Search Results](#)

Fr

Er

Di

Si

•

•

•

•

•

Case study: the query “engine” on Microsoft’s Live Search

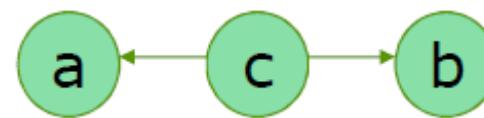
The screenshot shows the Microsoft Internet Explorer interface with the title bar "Live Search: engine - Microsoft Internet Explorer". The address bar contains the URL "http://search.live.com/results.aspx?q=engine&go=Search&lm=en-us&scop=8FORM=LMSOP". The main content area displays search results for "engine". At the top left is a "Web results" section stating "Page 1 of 810,000,000 results". Below it is a "See also" section with links to "Images", "Video", "News", "Maps", "MSN", and "More". A prominent result is a link to "Engine - Wikipedia, the free encyclopedia" with a brief description: "An engine in the broadest sense, is something that produces an output effect from a given input. The origin of engineering however, came from the design, building and working of ...". Further down are links to "List of search engines - Wikipedia, the free encyclopedia" and "Google". The bottom of the page includes a sidebar with "Sponsored sites" like "Rebuilt Engines" and "Engine replacement parts".

The screenshot shows the Microsoft Internet Explorer interface with the title bar "Google - Microsoft Internet Explorer". The address bar contains the URL "http://www.google.com". The main content area displays search results for "engine". The top result is a large, bold "Google" logo. Below it is a search bar with "Google Search" and "I'm Feeling Lucky" buttons. Navigation links for "Web", "Images", "Video", "News", "Maps", "Gmail", and "more" are visible. At the bottom of the page, there are links for "Advertising Programs", "Business Solutions", "About Google", "Go to Google Israel", and "Make Google Your Homepage". The copyright notice "©2007 Google" is at the very bottom.

The word “engine” does not appear
Anywhere on the page, nor in its source...

Link-Structure Analysis

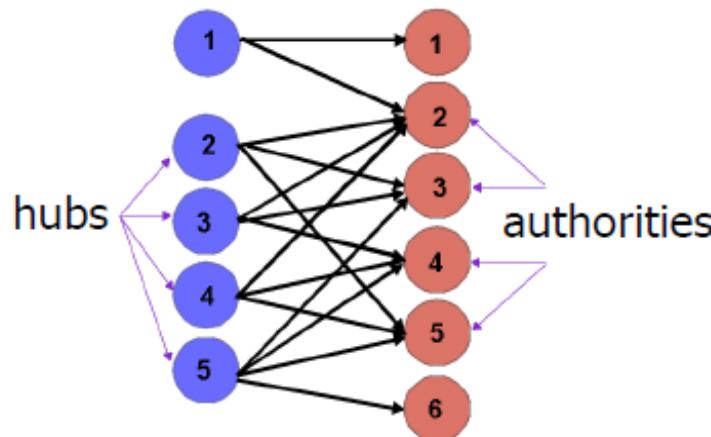
- Beyond anchor-text: the connectivity patterns between Web pages contain a gold mine of information
- A link from page **a** to page **b** can often be interpreted as:
 1. A recommendation, by **a**'s author, of the contents of **b**.
 2. Evidence that pages **a,b** share some topic of interest.
- A co-citation of **a** and **b** (by a third page **c**) may also constitute evidence that **a,b** share some topic of interest



Hubs and Authorities

- Notions proposed by Jon Kleinberg (1998)
- Two types of quality Web pages that pertain to a given topic

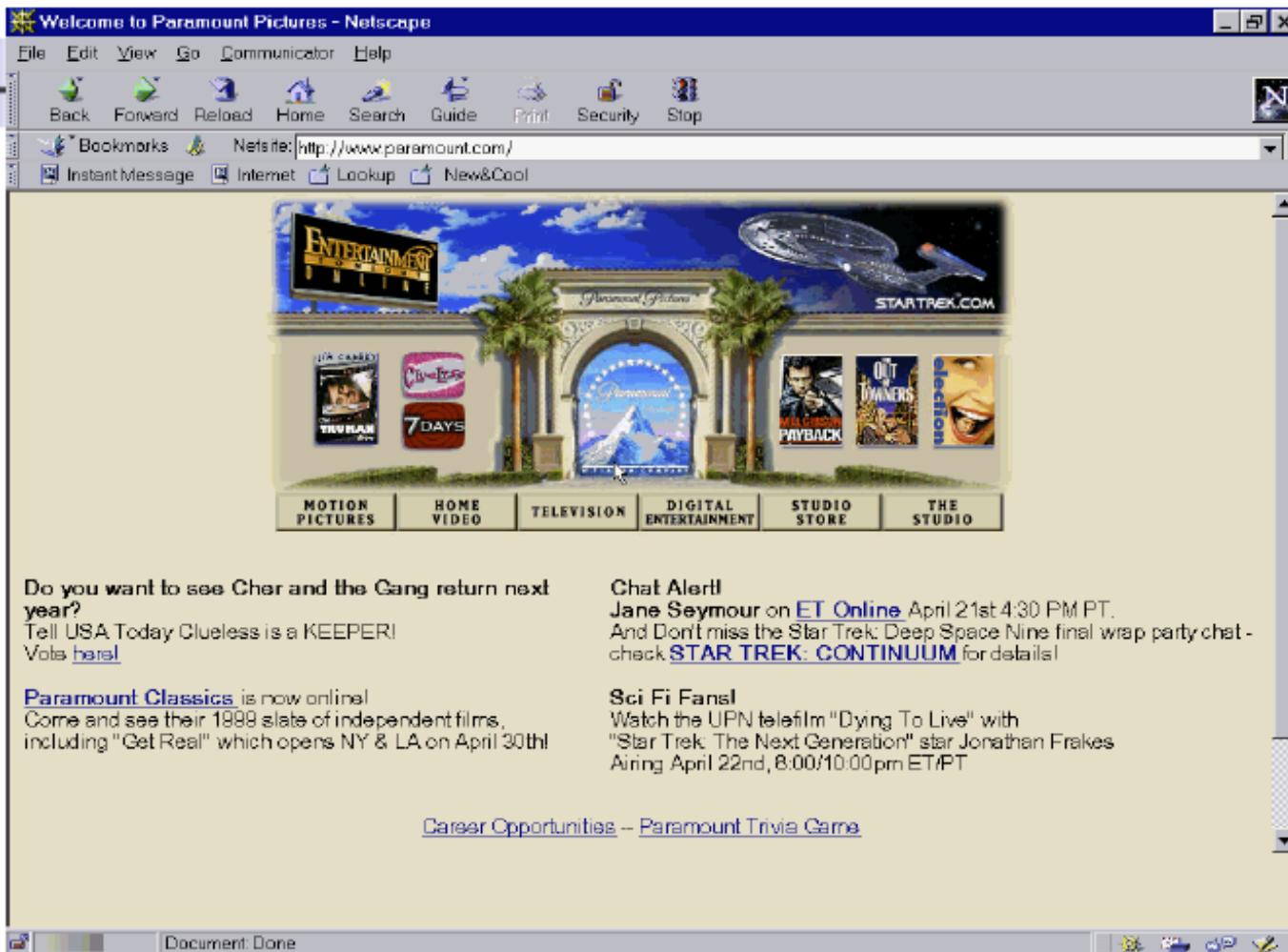
hubs: resource lists,
containing links to
many authorities



Authorities: pages
containing authoritative
content on the topic

- HITS (Hypertext Induced Topic Search): an algorithm that identifies hubs and authorities, given a topic of interest t
- Based on the Perron-Frobenius theory of non-negative matrices

Example – an Authority for the Query “movies”

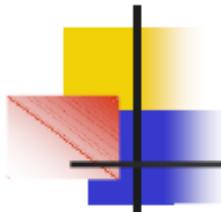


Example – a Hub for the Query “movies”

The screenshot shows a Netscape browser window titled "ENN at the Movies - Netscape". The address bar displays the URL <http://www.enn2.com/movies.htm>. The page content lists various movie-related websites, each preceded by a small logo:

- [CNN Showbiz News](#)
- [Weekly Box-Office](#)
- [Mr. Showbiz](#)
- [People Magazine Weekly](#)
- [Entertainment Weekly News](#)
- [MovieLink Local What's Playing](#)
- [Paramount Pictures](#)
- [Buena Vista Movieplex \(Disney\)](#)
- [20th Century Fox](#)
- [MGM/United Artists](#)
- [SONY/Columbia Pictures/Tri-Star](#)
- [Universal Pictures](#)
- [Turner Movie Classics](#)
- [Internet Movie Database \(U.S.\)](#)
- [Internet Movie Database \(U.K.\)](#)
- [Blockbuster Movies on Video](#)
- [Showtime Online Movies on TV](#)
- [Hollywood Supports \(AIDS care\)](#)
- [**wildwildweb Entertainment**](#)
- [Galaxy Film & Video Database](#)
- [Welcome to the Biz](#)

The "wildwildweb Entertainment" link is highlighted with a yellow box.

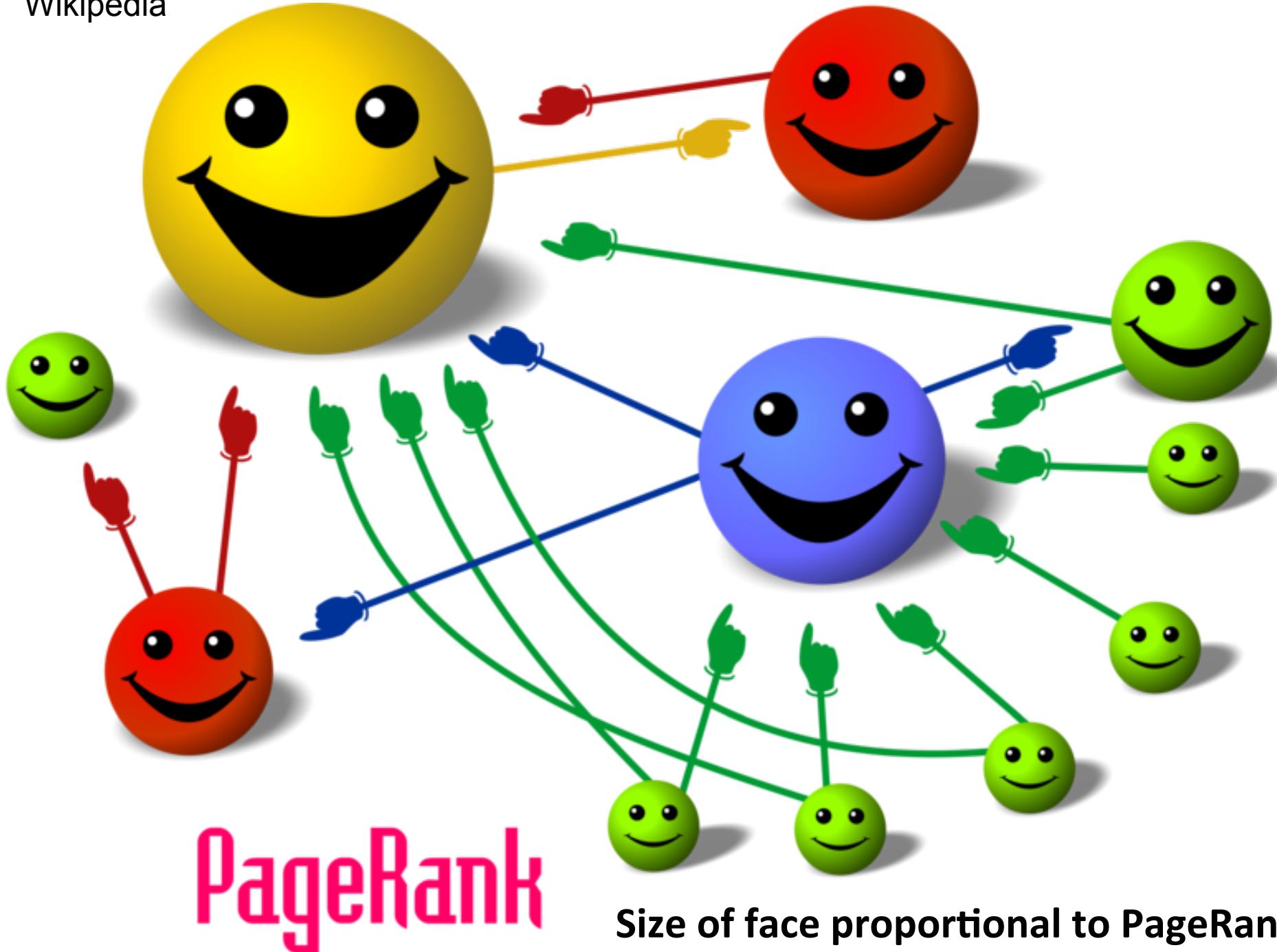


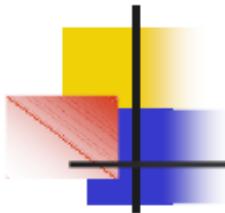
PageRank (Brin & Page, 1998)

- Named after Google's co-founder, Larry Page
- A global, query independent importance measure of Web pages
- A page is considered important if it receives many links from *important* pages
- Based on Markov chains and random walks

Markov Chains are roughly random walks

i.e. Sequence of random browsing selected either from current page (0.85)
or from world (0.15)





PageRank: “Random Surfer” Model

A random surfer moves from page to page.

Upon leaving page p , the surfer chooses one of two actions:

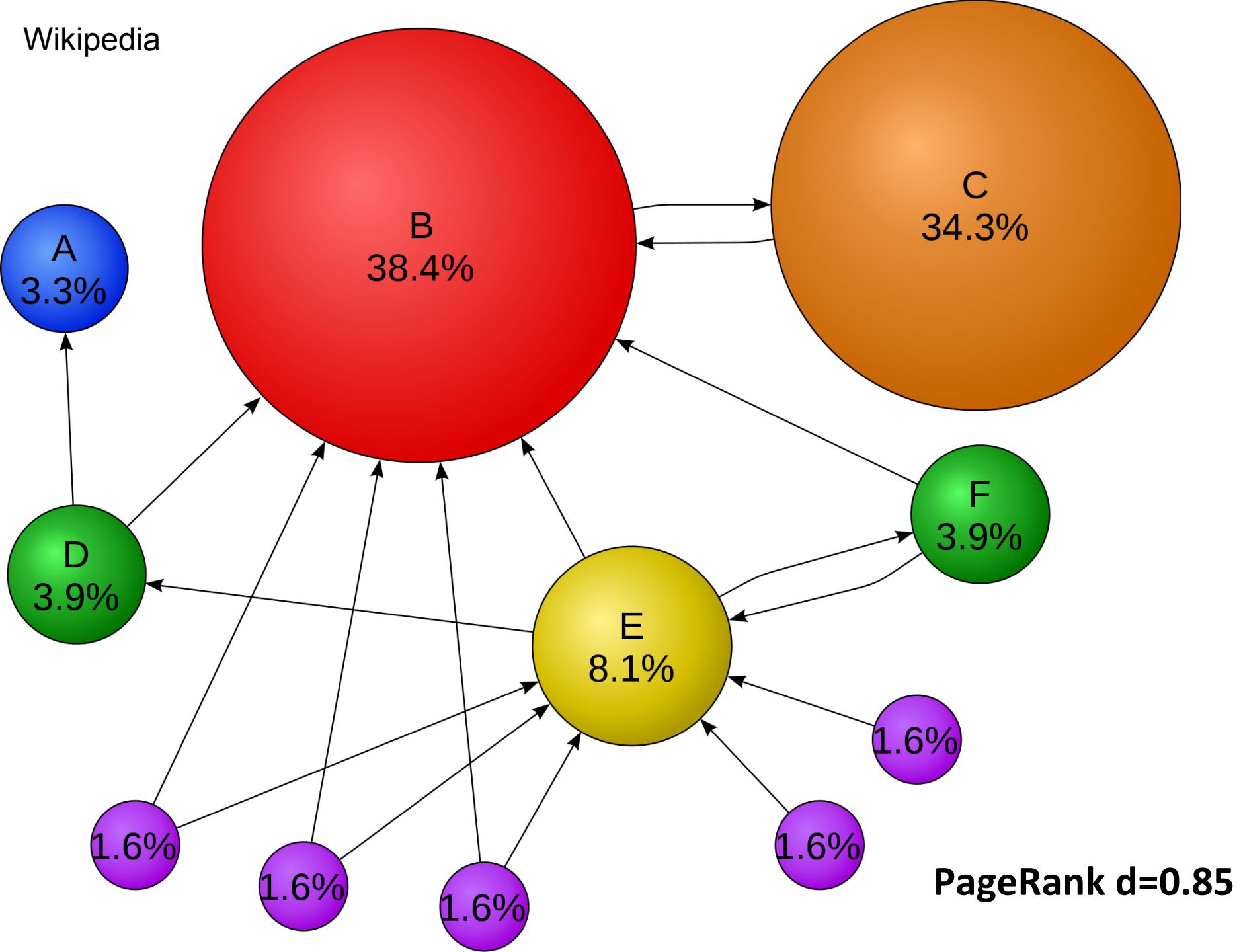
1. Follows an outgoing link of p (chosen uniformly at random), w.p. d
2. Jumps to an arbitrary Web page (chosen uniformly at random), w.p. $1-d$

The vector of PageRanks is the stationary distribution of this (ergodic) random walk

$$d = 0.85$$

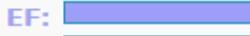
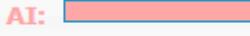
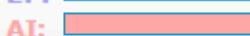
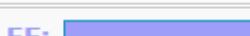
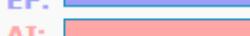
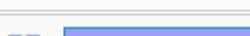
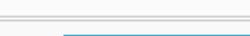
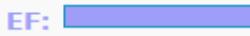
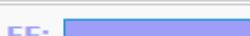
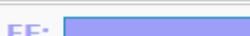
PageRank

- PageRank is probability that Page will be visited by a surfer is clicks each link on page with equal probability
 - minor corrections for pages with no outgoing links
- Found Iteratively with each page getting at each iteration a contribution equal to its page rank divided by #Links on page
- $\text{PR}(\text{Page } i) = \sum_{\text{Page } j \text{ pointing at } i} \text{PR}(\text{Page } j) / (\text{Number of Pages linked on Page } j)$
- One adds to this the chance $1-d$ that surfer types a random URL into web browser.
- That takes PageRank to d times above plus $(1 - d)$ divided by total number of pages on web
- On general principles, this will converge whatever the starting point
 - It can be written as iterative matrix multiplication



Related Applications

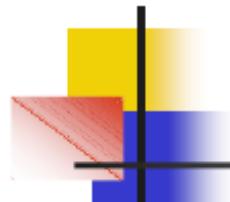
- Thinking of Page Rank as reputation
- A version of PageRank has recently been proposed as a replacement for the traditional Institute for Scientific Information (ISI) impact factor, and implemented at eigenfactor.org. Instead of merely counting total citation to a journal, the "importance" of each citation is determined in a PageRank fashion.
 - Impact Factor is number of citations of each article
 - The Eigenfactor score of a journal is an estimate of the percentage of time that library users spend with that journal. The Eigenfactor algorithm corresponds to a simple model of research in which readers follow chains of citations as they move from journal to journal.
- A similar new use of PageRank is to rank academic doctoral programs based on their records of placing their graduates in faculty positions. In PageRank terms, academic departments link to each other by hiring their faculty from each other (and from themselves).

Order	Journal	Percentile	EF ↓	AI ↓
1	NATURE ISSN: 0028-0836	EF:  100 AI:  100	1.65524	20.373
2	P NATL ACAD SCI USA ISSN: 0027-8424	EF:  100 AI:  99	1.60168	4.8961
3	SCIENCE ISSN: 0036-8075	EF:  100 AI:  100	1.41162	17.5248
4	PHYS REV LETT ISSN: 0031-9007	EF:  100 AI:  98	1.14457	3.5175
5	J AM CHEM SOC ISSN: 0002-7863	EF:  100 AI:  97	0.817303	2.7989
6	PHYS REV B ISSN: 1098-0121	EF:  100 AI:  89	0.756039	1.4281
7	J BIOL CHEM ISSN: 0021-9258	EF:  100 AI:  94	0.742127	2.0307
8	APPL PHYS LETT ISSN: 0003-6951	EF:  100 AI:  89	0.675752	1.3875
9	NEW ENGL J MED ISSN: 0028-4793	EF:  100 AI:  100	0.66383	21.304
10	CELL ISSN: 0092-8674	EF:  100 AI:  100	0.660816	20.554

EF= Eigenfactor
 AI = Article Influence
 over the first five
 years after publication

Eigenfactor scores are scaled so that the sum of the *Eigenfactor* scores of all journals listed in Thomson's Journal Citation Reports (JCR) is 100

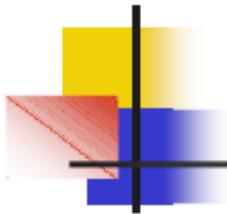
Article Influence scores are normalized so that the mean article in the entire Thomson Journal Citation Reports database has an article influence of 1.00



Link Analysis Algorithms

- Many variants and refinements of both HITS and PageRank have been proposed
 - Course will cover SALSA & Topic-Sensitive PageRank
- Other approaches include:
 - Max-flow techniques [Flake et al., SIGKDD 2000]
 - Machine learning and Bayesian techniques
- Examples of applications:
 - Ranking pages (topic specific/global importance)
 - Categorization, clustering, finding related pages
 - Identifying virtual communities
- A wealth of literature

**None done
here!**

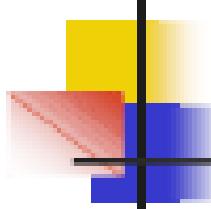


Stability of Link Analysis

- IR algorithms should be robust – small changes in the input should not dramatically alter the results
 - The Web changes constantly: pages and links are created, deleted and modified
- ?
- Do the known link-based algorithms produce stable scores/rankings?
 - ?
 - Does stability of scores imply stability of rankings?

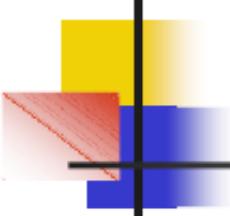
Different research approaches have been applied to these questions, some problems are still unsolved

Summary Issues



Comparing the Models

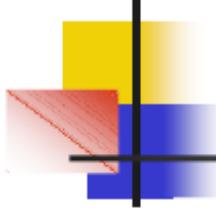
- Differ primarily in their theoretical basis, and in how “relevance” is defined and measured
 - Boolean model assumes relevance is related to term occurrence
 - Vector space model assumes relevance is related to similarity
 - Probabilistic model relies on probability estimates
- The Boolean model does not provide for varied relevance scores and is considered to be the weakest classic model
- In theory, the probabilistic model should supply the best predictions of relevance given enough information
- Experimentally, the vector space model often outperforms the probabilistic model on general collections



Classic and Modern IR Issues

- Polysemy (ambiguity): does “Jordan” refer to Michael Jordan, the Jordan River, or the Hashemite Kingdom of Jordan?
 - Can we ensure that our SERP* will represent all aspects?
- Synonymy: elevator/lift, movie/film
- Do the models hold when queries are very short?
- Short queries require the addition of proximity considerations (query term “hits” should be in proximity to each other) – how does this fit with the models?
- Information Retrieval merits a course of its own!

* SERP – Search Results Page



Size & Dynamics of the Web

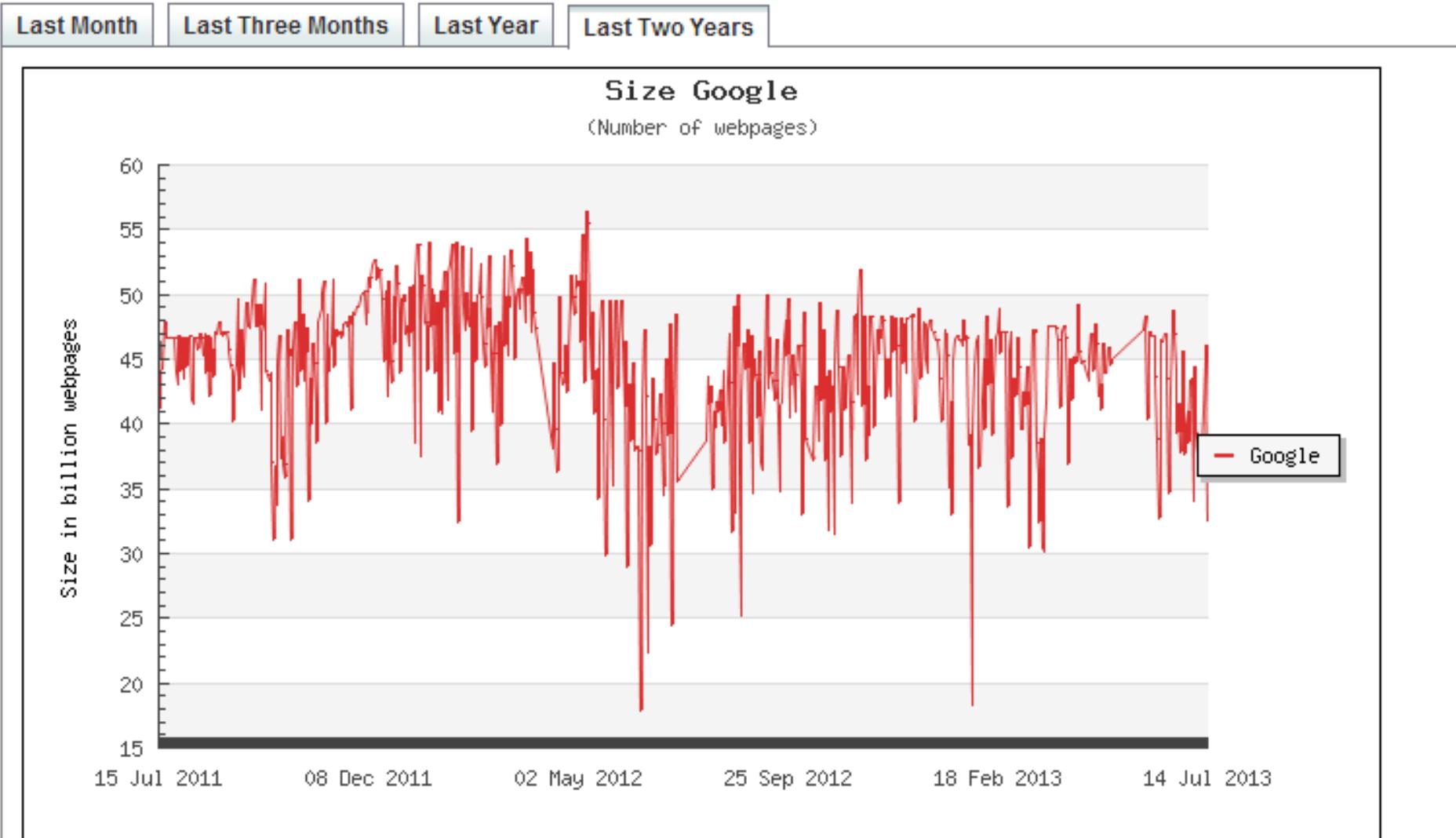
- Billions of pages, tens of millions of sites
- Hundreds of terabytes (10^{14}) of data
- Projected growth rate: doubling in size every two years (some estimate rate is much quicker)
- New pages are added; existing pages are modified and deleted
- Unstructured information in chaotic disorder
- Many media types
- Widely distributed, both geographically and organizationally

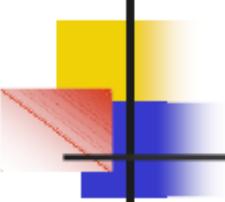
<http://www.worldwidewebsize.com/>

Estimated from number of pages returned to simple queries and fraction of documents satisfying queries



The size of the World Wide Web: Estimated size of Google's index





The Web as a Graph

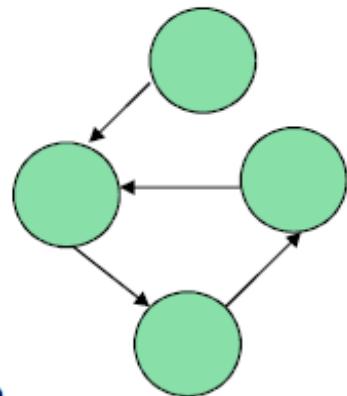
See start of previous lesson

Pages as graph nodes, hyperlinks as edges

- Sometimes sites are taken as the nodes

Some natural questions:

1. Distribution of the number of in-links to a page
2. Distribution of the number of out-links from a page
3. Distribution of the number of pages in a site
4. Connectivity: is it possible to reach most pages from most pages?
5. Are there models that explain how the Web graph evolved to its current shape?



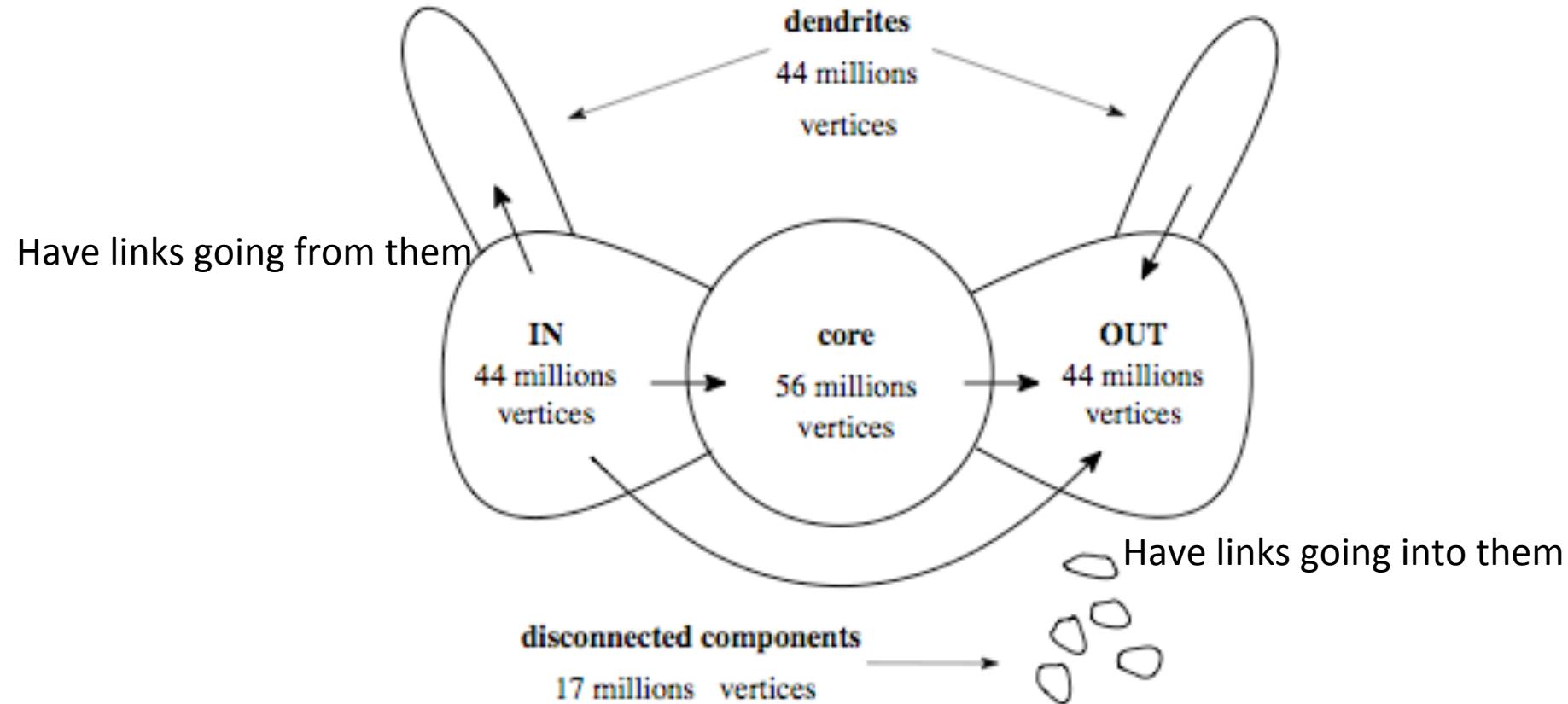
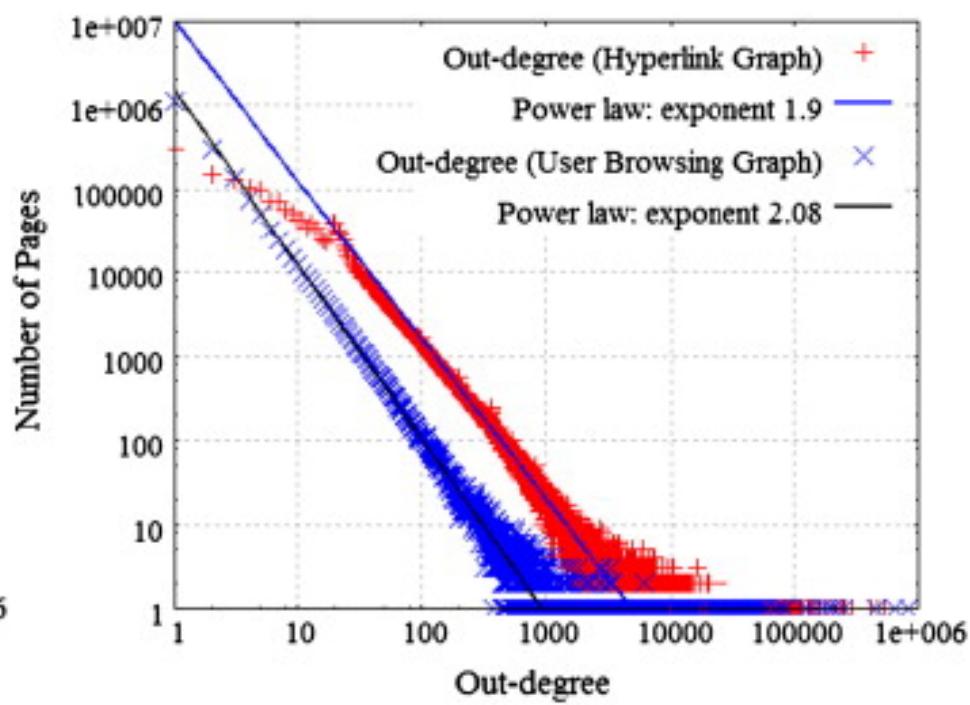
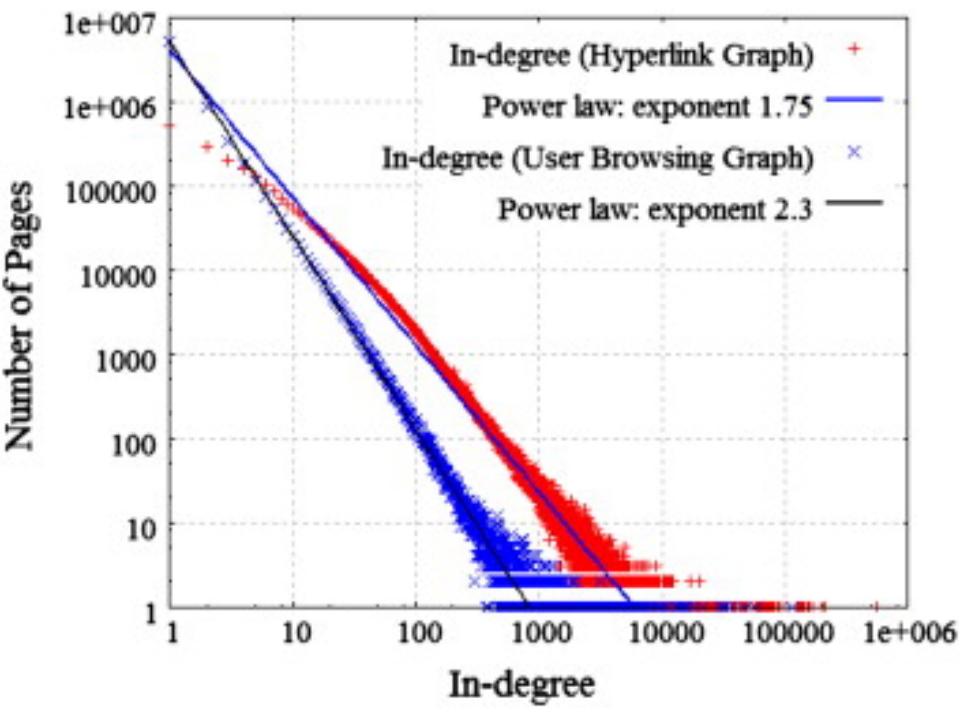


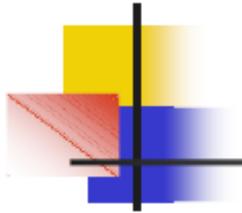
Fig. 3: The bow-tie macroscopic structure of the Web graph [BKM⁺00]: the core, the IN component, the OUT component and the dendrites. Each of these parts contains around one quarter of the pages, the disconnected part being reduced to less than 10% of the whole.

Universal Power Laws for Links

<http://www.sciencedirect.com/science/article/pii/S0167923612001844>

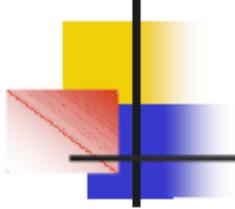


Crawling the Web



Crawlers - Framework

- The World Wide Web can be viewed as a directed graph
- Nodes are web objects, such as html pages, image files, pdf files, forms, etc...
- Nodes are uniquely identified by URLs
- Edges are html hyperlinks, pointing from an html page to its embedded objects or to other web objects
- Edges are uniquely identified by their source and target nodes



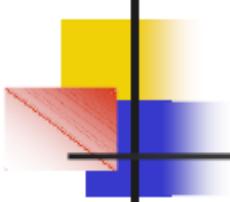
Generic Web Crawling Algorithm

Given a root set of distinct URLs:

- Put the root URLs in a queue
- While queue is not empty:
 - Take the first URL x out of the queue
 - Retrieve the contents of x from the web
 - Do whatever you want with it...
 - Mark x as visited
 - If x is an html page, parse it to find hyperlink URLs
 - For each hyperlink URL y within x :
 - If y is not marked visited, put y into queue



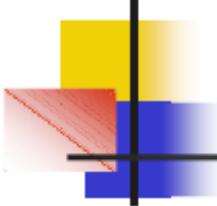
This is often the computational bottleneck



Crawler Design Issues

- Writing a simple crawler is a fairly easy task – one can be written in Java in a few lines of code
- However, there are many issues that need to be considered when writing a robust Web-scale crawler:
- Redirections
- Encrypted and password protected data
- Politeness - avoiding server load and respecting Robot Exclusion Protocol
- Script handling
- Language and encoding issues
- The challenge of dynamic pages
- Duplicate pages
- Crawl prioritization and variable revisit rates to ensure freshness
- Focused crawling
- Concurrency and distribution issues
- And many many more...

Web Advertising and Search

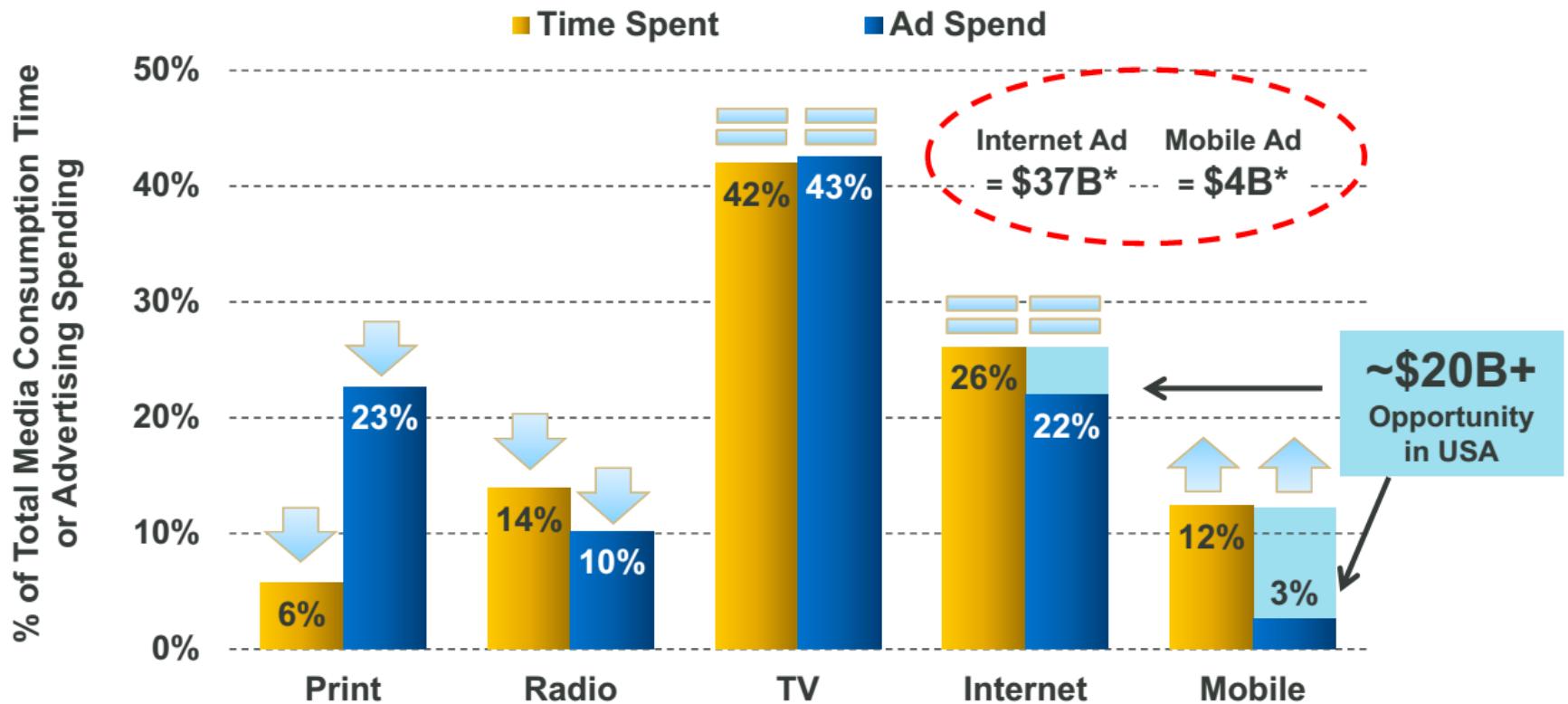


The Advertising Market

- Internet advertising is a $\$10^{10}$ business that is growing faster than old media advertising (radio, TV, newspapers & magazines, mail, outdoors)
- Online advertising budgets still lagging in proportion to time spent online, which is growing fast at the expense of old media
 - Seen as a driver to continued growth of online advertising market
- Traditional advertising:
 - Few, expensive opportunities
 - Targeting en-masse, by immediate context only
 - Difficult to measure effectiveness
- Internet advertising:
 - Billions of opportunities daily
 - Open to personalization via rich context of impression
 - Effectiveness is measurable: can measure click-through rates as % of impressions, and conversions as % of clicks

Material Upside for Mobile Ad Spend vs. Mobile Usage

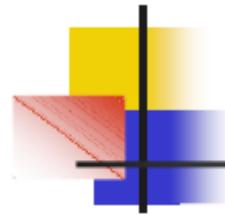
% of Time Spent in Media vs. % of Advertising Spending, USA 2012



Meeker/Wu May 29 2013 Internet Trends D11 Conference

Note: *Internet advertising reached \$37B in USA in 2012 per IAB, Mobile advertising reached \$4B per eMarketer. Print includes newspaper and magazine. \$20B opportunity calculated assuming Internet and Mobile ad spend share equal their respective time spent share. Source: Time spent and ad spend share data based on eMarketer (adjusted to exclude outdoors / classified media spend), 12/12.

Internet Advertising Types: Sponsored Search



textual ads served on search results pages, triggered by query terms on which advertisers bid

The screenshot shows a Mozilla Firefox browser window with the title "ski italy - Yahoo! Search Results - Mozilla Firefox". The address bar indicates the URL is <http://search.yahoo.com/search?p=ski+italy&oe=UTF-8&fr=yfp-t-501&n>. The search bar contains the query "ski italy". The results page includes a "Web" tab and a "Search" button. Below the search bar, there's a link to "Options" and "Customize". The main content area shows search results:

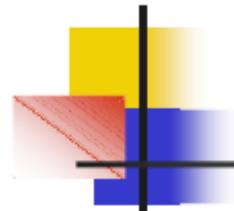
- Ski Italy and More 2009**
Love to Ski? See our 2009 Hosted Vacations to the Italian Alps.
www.skiing-italy.com
- Ski Italy resort and accommodation information**
... Italy, take your advice from the people that know Activelifestyle, 20 years plus of skiing in Italy ... In Italy probably more than any other ski country, it ...
www.skitaly.com - 106k - Cached
- Italy Skiling and Winter Sports - Top Ski Resorts in Italy**
Italy has good skiing and winter sports. From the Alps and Dolomites to Sicily ... Here are some of Italy's top ski resorts. Ski Piemonte - Home of the 2006 ...
goitaly.about.com/ad/wintersportsinitaly/ski_italy.htm - Cached

On the right side of the results, there are sections for "SPONSOR RESULTS" and "SPONSOR RESU". The "SPONSOR RESULTS" section lists:

- New Travel Ticker Site**
The Site for Insider Travel Deals. New Specials Added Hourly.
www.Travel-Ticker.com
- Ski in Italy**
Compare airfare prices from over 120 top websites and save up to 70%.
Flights.SideStep.com
- Italy Ski**
Ski Reports, Deals, Reviews For Italy Ski Resorts.
www.OnTheSnow.com

At the bottom of the browser window, the URL <http://help.yahoo.com/l/us/yahoo/search/basics/basics-23.html> is visible.

Internet Advertising Types: Contextual Ads (a.k.a. Content Match)



textual ads served on third party ("publisher") pages, triggered by content of page on which advertisers bid

The screenshot shows a Microsoft Internet Explorer window displaying a news article from SI.com. The article is titled "He's back: Lakers top Sonics in Kobe debut". The text discusses Kobe Bryant's performance, mentioning he had 34 points and 13 rebounds in a win over Phoenix, and 22 points in a second victory against Golden State. It also quotes him saying he worked hard in the offseason and stepped up at the beginning of the season. The article ends with a quote from Seattle coach Bob Hill about free throws.

On the right side of the screen, there is a sidebar titled "advertiser links" containing four contextual ads:

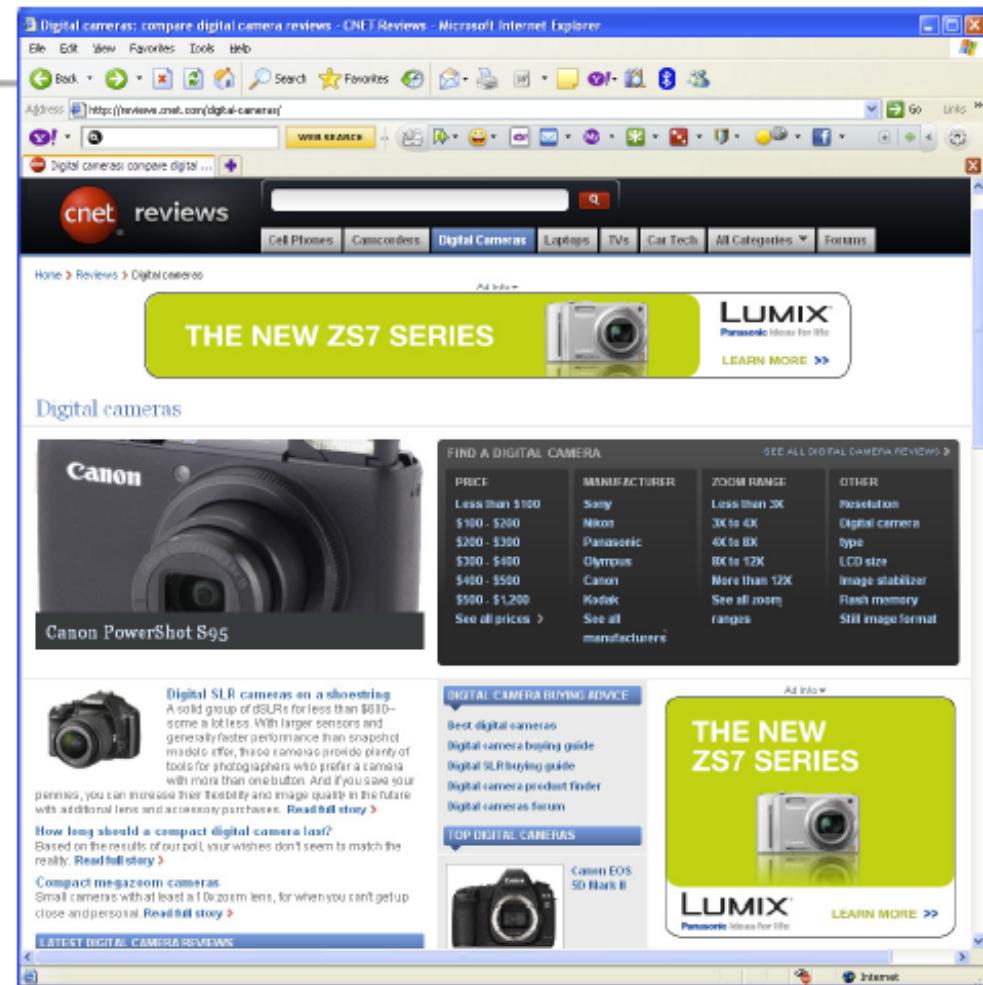
- Coldwell Banker - Los Angeles, CA**
Coldwell Banker Residential Brokerage helps you come home to Los Angeles, California...
www.californiarmoves.com
- Central LA Hotel**
Official Site. Stay with Hyatt Regency Century Plaza centrally located. Book Online and...
www.centuryplaza.hyatt.com
- Los Angeles Maps**
Explore Los Angeles with maps, satellite images, directions and more.
[localive.com](http://www.localive.com)
- Visiting Los Angeles, CA?**
Find affordable hotel rates and travel deals for Los Angeles, CA from over 120 top...
www.kayak.com

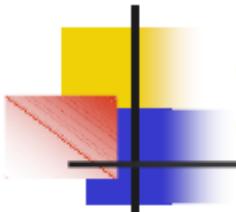
At the bottom of the browser window, there is a JavaScript code: "javascript:CNN_OpenPopup('/services/overture/d/frameset.exclude.html','620x430','toolbar=no,location=no,directories=no...")".

Internet Advertising Types: Display Ads

Graphical elements (e.g. banners) of standard sizes, served on publisher pages, mainly targeting audience demographics

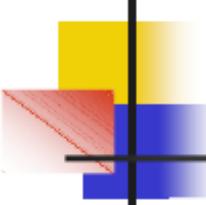
- Guaranteed delivery (GD): advertiser buys a display campaign months in advance, requesting ## impressions to certain demographics in a given time period; ad agency pays fine for under-delivery
- Non-guaranteed delivery (NGD): advertiser bids in real time for impressions, with no guarantees that ads will be shown





Internet Advertising Models

1. Cost per Impression (sometimes called Cost per *Mille*, i.e. 1000 impressions) – advertiser pays for ad to be shown
 - Main model for display advertising
2. Cost per Click – advertiser only pays when ad is clicked by user
 - Main model for sponsored search and content match
 - Sponsored search, 2005: 50% of clicks < \$0.25, 80% < \$0.50
3. Cost per Action – advertiser only pays when the user's interaction with the ad is followed by some transaction (not necessarily purchase)



Computational Advertising Challenges

- Conflicting goals:
 - Advertiser: campaign effectiveness – maximize ROI
 - Publisher: revenue without jeopardizing image of site
 - User: relevant ads, or at least ads that do not harm the experience
 - Ad agency (search engine): own revenue, while balancing goals of other parties; sometimes ad agency is also the publisher
- Ad agency challenges:
 - Which ads to trigger per query/page/impression? How to order them? What to charge for them?
 - In general, how to set the rules of the marketplace so as to ensure goals of parties are met (Game Theory/Mechanism Design)?
 - Which guaranteed delivery contracts to accept? How to balance GD with NGD?
 - Helping advertisers choose bid phrases effectively
 - Combating spam and fraud
 - Scale!!



The Long Tail

How the infinite choice available in e-commerce and the ease of publishing online have redefined the business of media & entertainment

Mottos:

- “The biggest money is in the smallest sales”
- “Embrace niches”

How Endless Choice Is Creating Unlimited Demand

The Long Tail

Enter

Why the Future of Business
Is Selling Less of More

CHRIS ANDERSON

“Anderson’s insights influence Google’s strategic thinking in a profound way.”

READ THIS BRILLIANT AND TIMELY BOOK.”

—ERIC SCHMIDT, CEO, GOOGLE

Long Tail Economics and Recommender Systems

- People cannot intelligently navigate product or media catalogs consisting of millions of items
 - e-Inventories of books, songs, movies, albums are practically infinite
 - In particular, niche audiences cannot find niche content
- Recommender systems, collaborative filtering and social tagging services are key to Long Tail economics
 - Find an item you like, and we'll point you to similar items
- Pioneered by Amazon; today, used "everywhere"

Customers who bought items in your Recent History also bought:



I Own It Not Interested
 Rate it
[Add to Cart](#) [Add to Wish List](#)



I Own It Not Interested
 Rate it
[Add to Cart](#) [Add to Wish List](#)



I Own It Not Interested
 Rate it
[Add to Cart](#) [Add to Wish List](#)

Clustering and Topic Models

Grouping Documents Together

- The **responses to a search query** give you a group of documents
- If we represent documents as points in a space, we can try to **identify regions**
 - **Clustering**: Nearby regions of points
 - **Support Vector Machine**: Chop space up into parts
 - **(Gaussian) Mixture Models**: A type of fuzzy clustering
 - **K-Nearest Neighbors** (if have examples)
- Alternatively we can determine “hidden meaning” with a **topic model** (defined by dynamic small bags of words)
 - Latent Semantic Indexing
 - Latent Dirichlet Allocation
 - With lots of variants of these methods to find “**latent factors**”

Topic Models

- Illustrated by Google News
- These try to group documents by Topics such as “Presidential Election” and not by inclusion of particular phrases
- You imagine each document is a set of topics (the latent factors) and each topic is a bag of words.
- Find the best set of topics and best set of words in topics



Yet Another Example

- Reuters-21578 collection
 - 21578 short newswire messages from 1987
- Top-3 results when querying for **taxes reagan** using LSI:

FITZWATER SAYS REAGAN STRONGLY AGAINST TAX HIKE
WASHINGTON, March 9 - White House spokesman Marlin
Fitzwater said President Reagan's record in opposing tax hikes is
"long and strong" and not about to change.

ROSTENKOWSKI SAYS WILL BACK U.S. TAX HIKE, BUT
DOUBTS PASSAGE WITHOUT REAGAN SUPPORT

WHITE HOUSE SAYS IT OPPOSED TO TAX INCREASE AS
UNNECESSARY

A Latent
Factor Finding
Method

- **The last document doesn't mention the term “reagan”!**

An example of DA-PLSI/PLSA

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
percent	stock	soviet	bush	percent
million	market	gorbachev	dukakis	computer
year	index	party	percent	aids
sales	million	i	i	year
billion	percent	president	jackson	new
new	stocks	union	campaign	drug
company	trading	gorbachevs	poll	virus
last	shares	government	president	futures
corp	new	new	new	people
share	exchange	news	israel	two

Top 10 popular words for each of 5 topics found from the AP news dataset divided into 30 topics. Processed by DA-PLSI and showing only 5 of 30 topics



<https://portal.futuregrid.org>

