universität koblenz
weiter:denken

Web Information Retrieval

Web Search

(SOSE 2023, 24.05.2023)

Frank Hopfgartner, Stefania Zourlidou
Institute for Web Science and Technologies

Credit for these slides

universität koblenz
weiter:denken

These slides have been adapted from

- Web IR (Zeyd Boukhers-WeST, SOSE 2020)

Recapitulation

universität koblenz
weiter:denken

- Language models
  - Query likelihood
  - Document likelihood
  - Comparison model

Objectives of this lecture

universität koblenz
weiter:denken

- Classical IR vs. Web IR
- Web IR, Web search basics
  - Ads
  - Spams
  - Duplicates
  - User Needs
  - Web Graph (anchor text)
  - Indexing

universität koblenz
weiter:denken

1. Classical IR

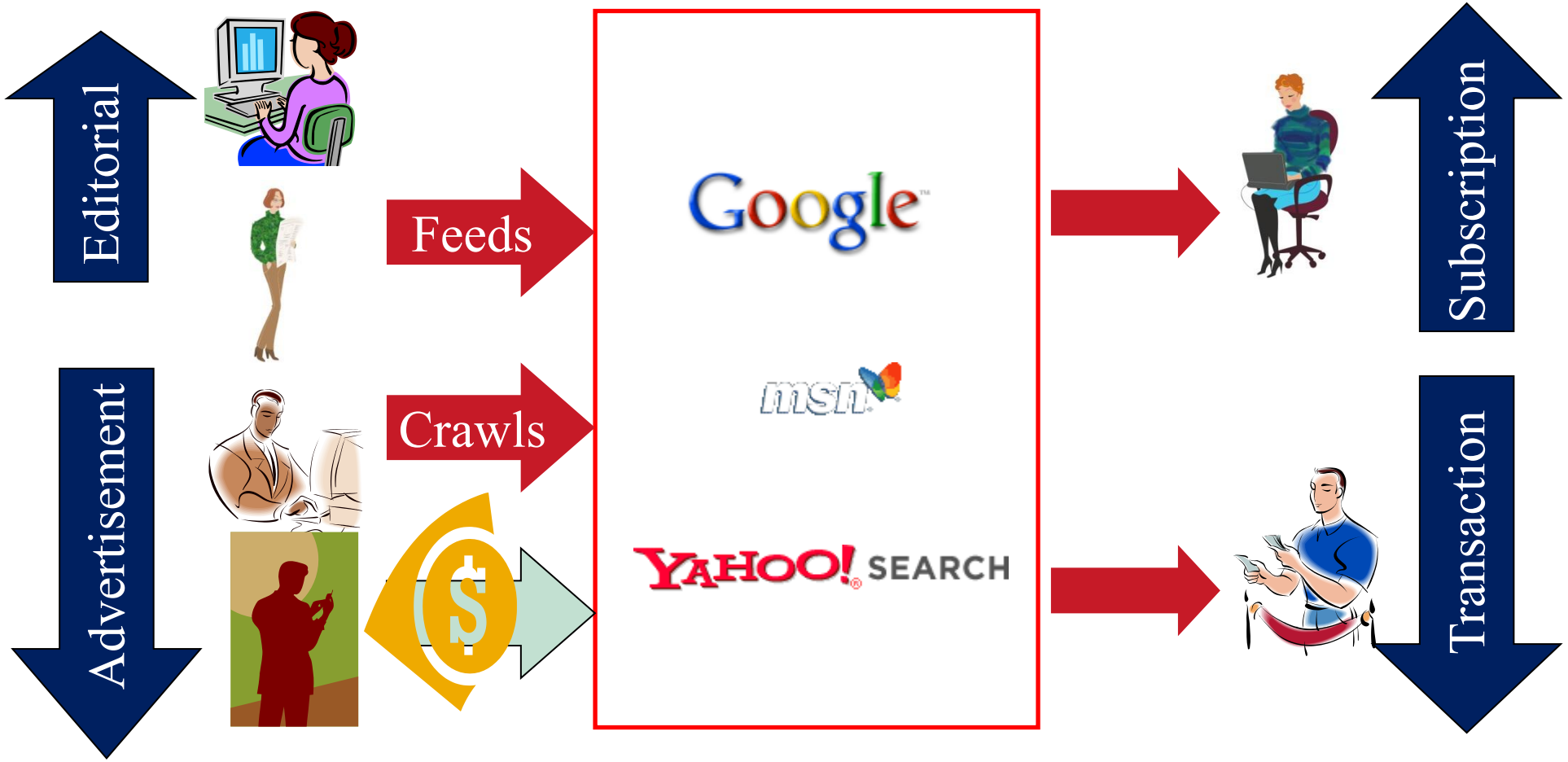
Classical IR

- Corpus
  - Fixed collection
  - Corpus is predetermined
- Goal
  - Retrieve documents with content relevant to user's information need
- Relevance
  - For every query q and a document d, there exists a relevance score $Score(q, d)$
    - Score is context independent
    - Score is user independent

universität koblenz
weiter:denken

2. Web IR

Web IR


Editorial
Feeds
Google
Subscription
Crawls
msn
Advertisement
YAHOO! SEARCH
Transaction
Content creators
Content aggregators
Content consumers

Web IR

- Corpus
  - Scale much larger than in classical IR
  - Volume is expanding
  - Spam: Billions of pages
- Content
  - Truth, lies, obsolete information, contradictions
  - Unstructured, semi-structured, structured
  - Can be *dynamically generated*

Web IR

universität koblenz
weiter:denken

- Other characteristics
  - **4.39 billion** internet users in 2019
    - Digital 2019 report
  - 1.7 billion websites in 2020
    - internetlivestats.com
  - High linkage: More than 8 links/page in the average
  - Significant duplication
    - Syntactic – 30%-40% (near) duplicates [Brod97, Shiv99b, etc.]
    - Semantic – ?

Objectives

- Ads
- Spams
- Duplicates
- User Needs
- Web Graph (anchor text)
- Indexing
- Empirical Evaluation

universität koblenz
weiter:denken

3. Ads Web IR

Web IR: brief (non-technical) history

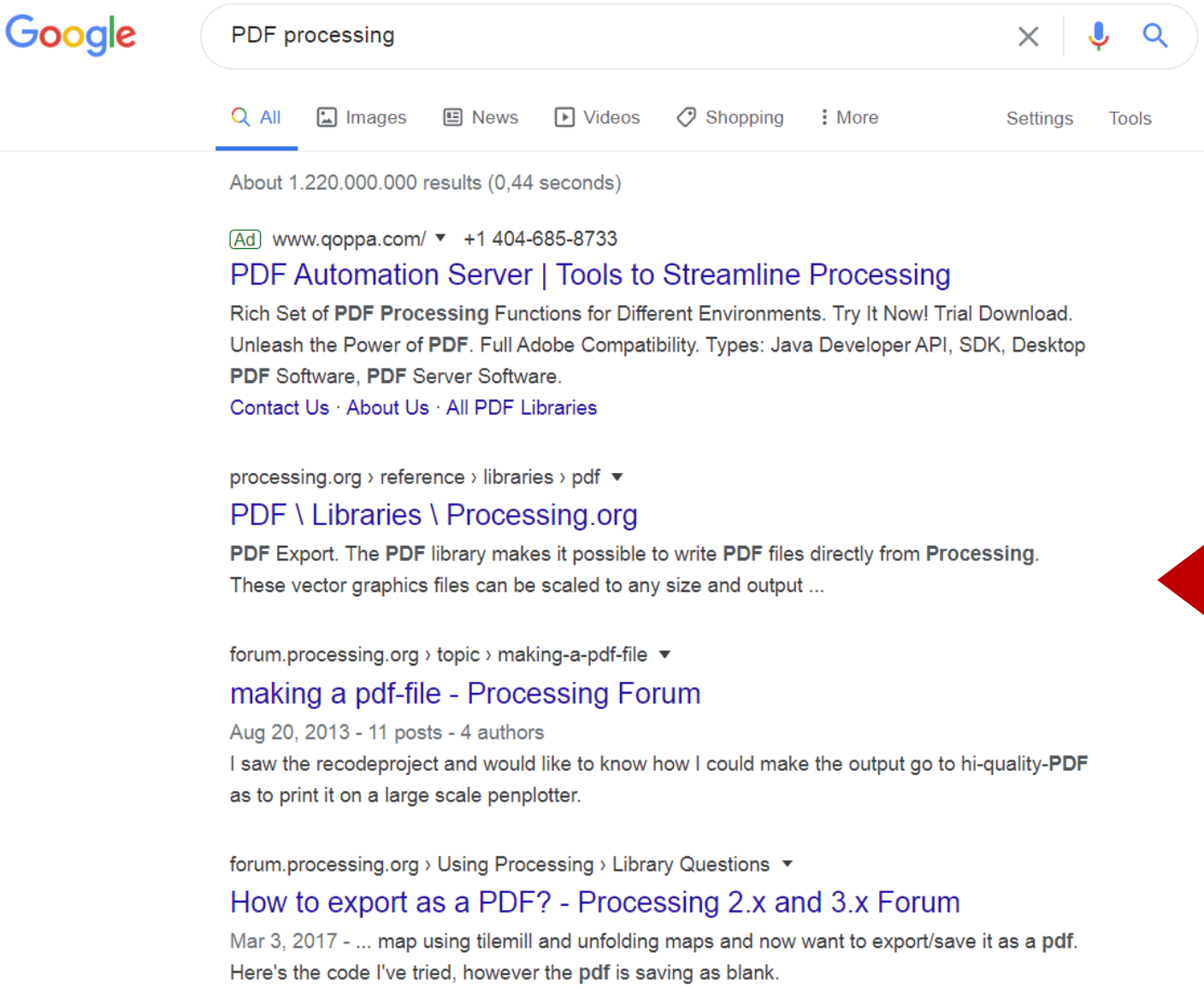
- Early keyword-based engines
  - Altavista, Excite, Infoseek, Inktomi, ca. 1995-1997

- Paid placement ranking: Goto.com (morphed into Overture.com → Yahoo!)
  - Search ranking depended on how much you paid
  - Auction for keywords: ***casino*** was expensive!

Brief (non-technical) history

- 1998+: Link-based ranking pioneered by Google
  - Blew away all early engines save Inktomi
  - Great user experience in search of a business model
  - Meanwhile Goto/Overture's annual revenues were nearing $1 billion
- *Result*: Google added paid-placement "ads" to the side, independent of search results
  - Yahoo follows suit, acquiring Overture (for paid placement) and Inktomi (for search)

Google Ads



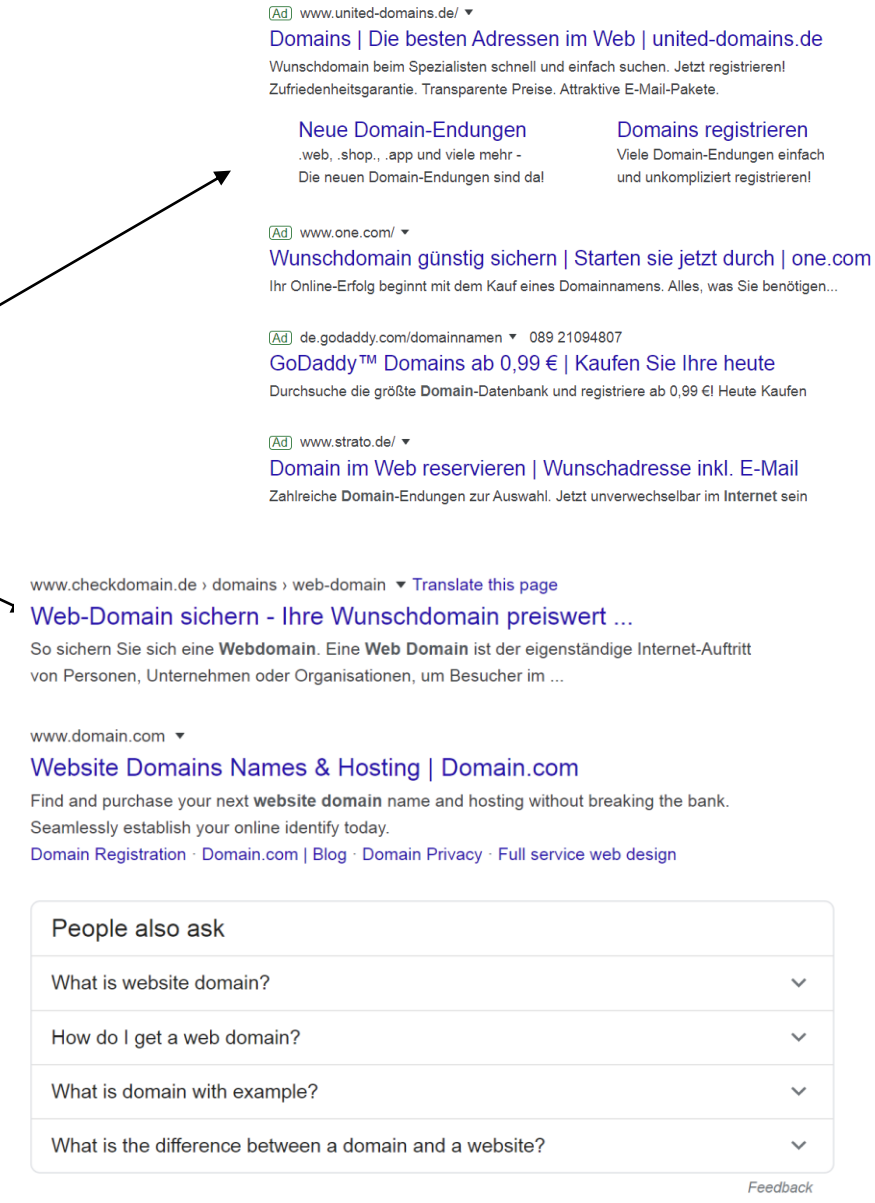
Ads

Algorithmic result

Ads vs. search results

Google has maintained that ads (based on vendors bidding for keywords) do not affect vendors' rankings in search results

Search = ***web domain***

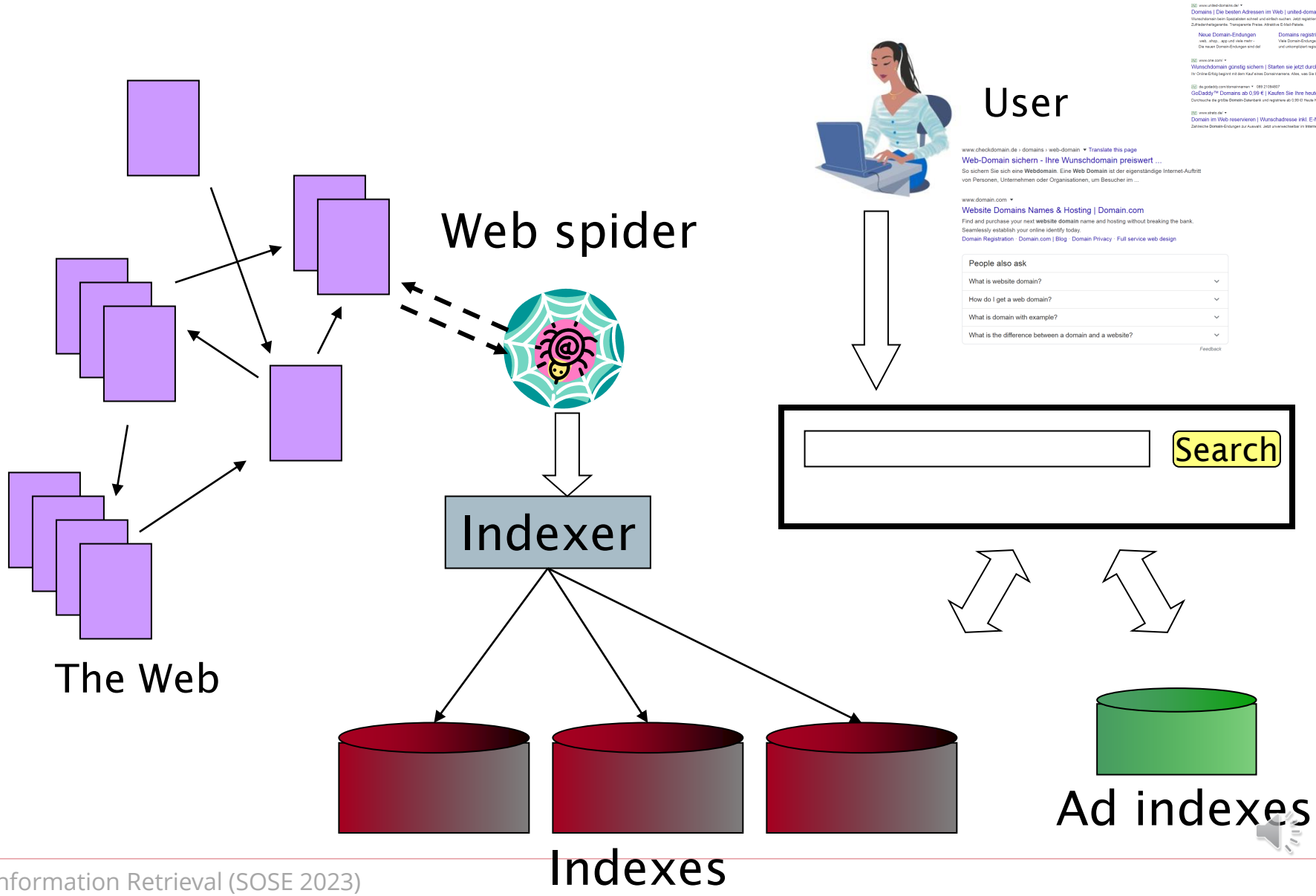


Ads vs. search results

- Other search engines (Yahoo, MSN) have made similar statements from time to time
  - Any of them can change anytime

- We will focus primarily on search results independent of paid placement ads
  - Although the latter is a fascinating technical subject in itself

Web search


User
Web spider
Indexer
Search
The Web
Indexes
Ad indexes
People also ask
What is website domain?
How do I get a web domain?
What is domain with example?
What is the difference between a domain and a website?

How are ads ranked?

- First cut: according to bid price à la Goto
  - Bad idea: open to abuse
  - Example: query [Buying fresh Chicken?] → ad for KFC
  - We don't want to show nonrelevant ads
- Instead: rank based on bid price and relevance
- Key measure of ad relevance: clickthrough rate
  - clickthrough rate = CTR = clicks per impressions
- Result: A nonrelevant ad will be ranked low.
  - Even if this decreases search engine revenue short-term
  - Hope: Overall acceptance of the system and overall revenue is maximized if users get useful information
- Other ranking factors: location, time of day, quality and loading speed of landing page
- The main ranking factor: the query

Google's second price auction

advertiser	bid	CTR	ad rank	rank	paid
A	$4.00	0.01	0.04	4	(minimum)
B	$3.00	0.03	0.09	2	$2.68
C	$2.00	0.06	0.12	1	$1.51
D	$1.00	0.08	0.08	3	$0.51

- bid: maximum bid for a click by advertiser
- CTR: click-through rate: when an ad is displayed, what percentage of time do users click on it? CTR is a measure of relevance.
- ad rank: bid $\times$ CTR: this trades off (i) how much money the advertiser is willing to pay against (ii) how relevant the ad is
- rank: rank in auction
- paid: second price auction price paid by advertiser

Google's second price auction

universität koblenz
weiter:denken

advertiser	bid	CTR	ad rank	rank	paid
A	$4.00	0.01	0.04	4	(minimum)
B	$3.00	0.03	0.09	2	$2.68
C	$2.00	0.06	0.12	1	$1.51
D	$1.00	0.08	0.08	3	$0.51

Second price auction: The advertiser pays the minimum amount necessary to maintain their position in the auction (plus 1 cent)

- $price_1 = bid_2 \times CTR_2 / CTR_1$
- $p_1 = bid_2 \times CTR_2/CTR_1 = 3.00 \times 0.03/0.06 = 1.50$
- $p_2 = bid_3 \times CTR_3/CTR_2 = 1.00 \times 0.08/0.03 = 2.67$
- $p_3 = bid_4 \times CTR_4/CTR_3 = 4.00 \times 0.01/0.08 = 0.50$

Keywords with high bids

- According to https://www.wordstream.com/articles/most-expensive-keywords

Insurance	$54.91
Loans	$44.28
Mortgage	$47.12
Attorney	$47.07
Credit	$36.06
Lawyer	$42.51
Donate	$42.02
Degree	$40.61
Hosting	$31.91
Claim	$45.51
Conference Call	$42.05
Trading	$33.19
Software	$35.29

Search ads: a win-win-win?

- The search engine company gets revenue every time somebody clicks on an ad
- The user only clicks on an ad if they are interested in the ad
  - Search engines punish misleading and nonrelevant ads
  - As a result, users are often satisfied with what they find after clicking on an ad
- The advertiser finds new customers in a cost-effective way

Search ads

- Why is web search potentially more attractive for advertisers than TV spots, newspaper ads or radio spots?
- The advertiser pays for all this. How can the advertiser be cheated?
- Any way this could be bad for the user?
- Any way this could be bad for the search engine?

Not a win-win-win: keyword arbitrage

universität koblenz
weiter:denken

- Buy a keyword on Google
- Then redirect traffic to a third party that is paying much more than you are paying Google
  - E.g., redirect to a page full of ads
- This rarely makes sense for the user
- Ad spammers keep inventing new tricks
- The search engines need time to catch up with them

universität koblenz
weiter:denken

4. Spam

The trouble with paid placement

- It costs money. What's the alternative?
- *Search Engine Optimization*
  - "Tuning" your web page to rank highly in the search results for select keywords
  - Alternative to paying for placement
  - Thus, intrinsically a marketing function
- Performed by companies, webmasters and consultants ("Search engine optimizers") for their clients
- Some perfectly legitimate, some very shady

Simplest forms

- First generation engines relied heavily on *tf/idf*
  - The top-ranked pages for the query **maui resort** were the ones containing the most **maui**'s and **resort**'s
- SEOs responded with dense repetitions of chosen terms
  - e.g., **maui resort maui resort maui resort**
  - Often, the repetitions would be in the same color as the background of the web page
    - Repeated terms got indexed by crawlers
    - But not visible to humans on browsers

Pure word density cannot be trusted as an IR signal

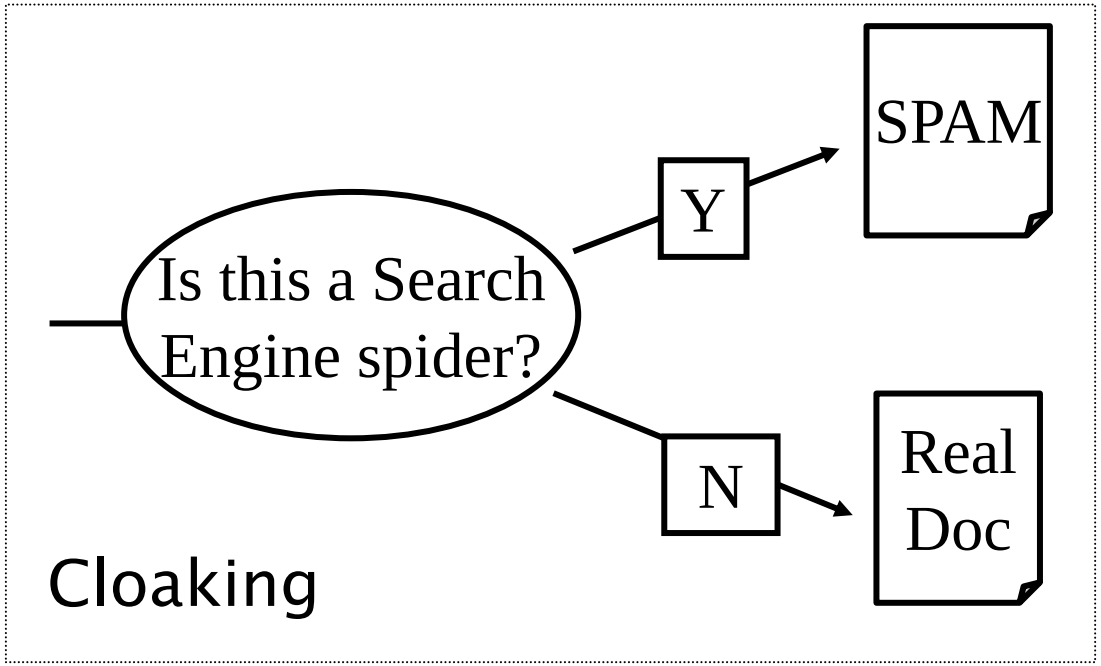
Variants of keyword stuffing

- Misleading meta-tags, excessive repetition
- Hidden text with colors, style sheet tricks, etc.

Meta-Tags =
"… London hotels, hotel, holiday inn, hilton, discount, booking, reservation, sex, mp3, britney spears, viagra, …"

Cloaking

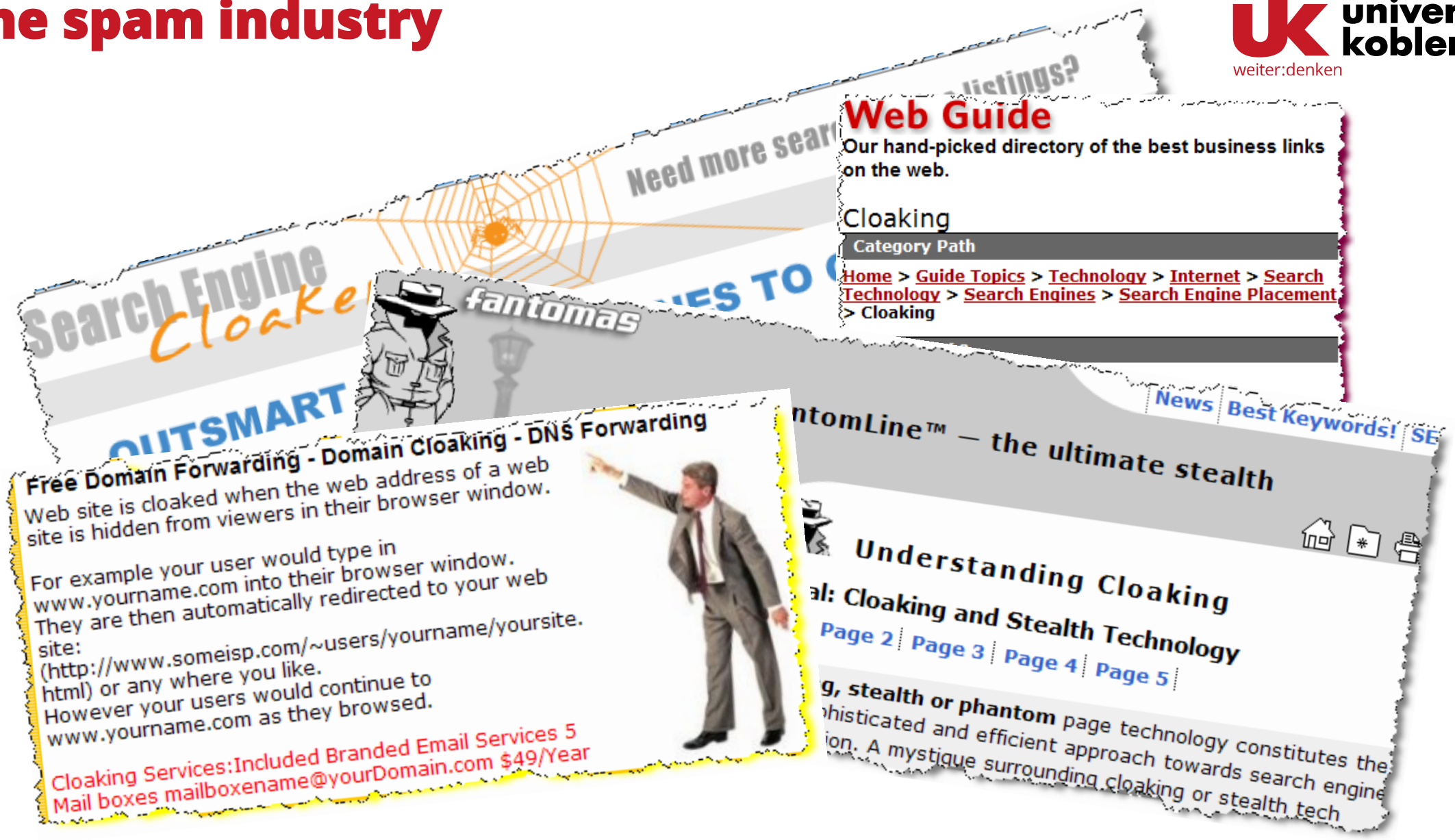
- Serve fake content to search engine spider
- DNS cloaking: Switch IP address. Impersonate



Search engine optimization (Spam)

- Motives
  - Commercial, political, religious, lobbies
  - Promotion funded by advertising budget
- Operators
  - Contractors (Search Engine Optimizers) for lobbies, companies
  - Web masters
  - Hosting services
- Forums
  - E.g., Web master world (www.webmasterworld.com)
    - Search engine specific tricks

The spam industry



The war against spam

- Quality signals - Prefer authoritative pages based on
  - Votes from authors (linkage signals)
  - Votes from users (usage signals)
- Limits on meta-keywords
- Robust link analysis
  - Ignore statistically implausible linkage (or text)
  - Use link analysis to detect spammers (guilt by association)

- Spam recognition by machine learning
  - Training set based on known spam
- Family friendly filters
  - Linguistic analysis, general classification techniques, etc.
  - For images: flesh tone detectors, source text analysis, etc.
- Editorial intervention
  - Blacklists
  - Top queries audited
  - Complaints addressed
  - Suspect pattern detection

More on spam

- Web search engines have policies on SEO practices they tolerate/block
  - http://help.yahoo.com/help/us/ysearch/index.html
  - http://www.google.com/intl/en/webmasters/
- Adversarial IR: the unending (technical) battle between SEO's and web search engines
- Research http://airweb.cse.lehigh.edu/

universität koblenz
weiter:denken

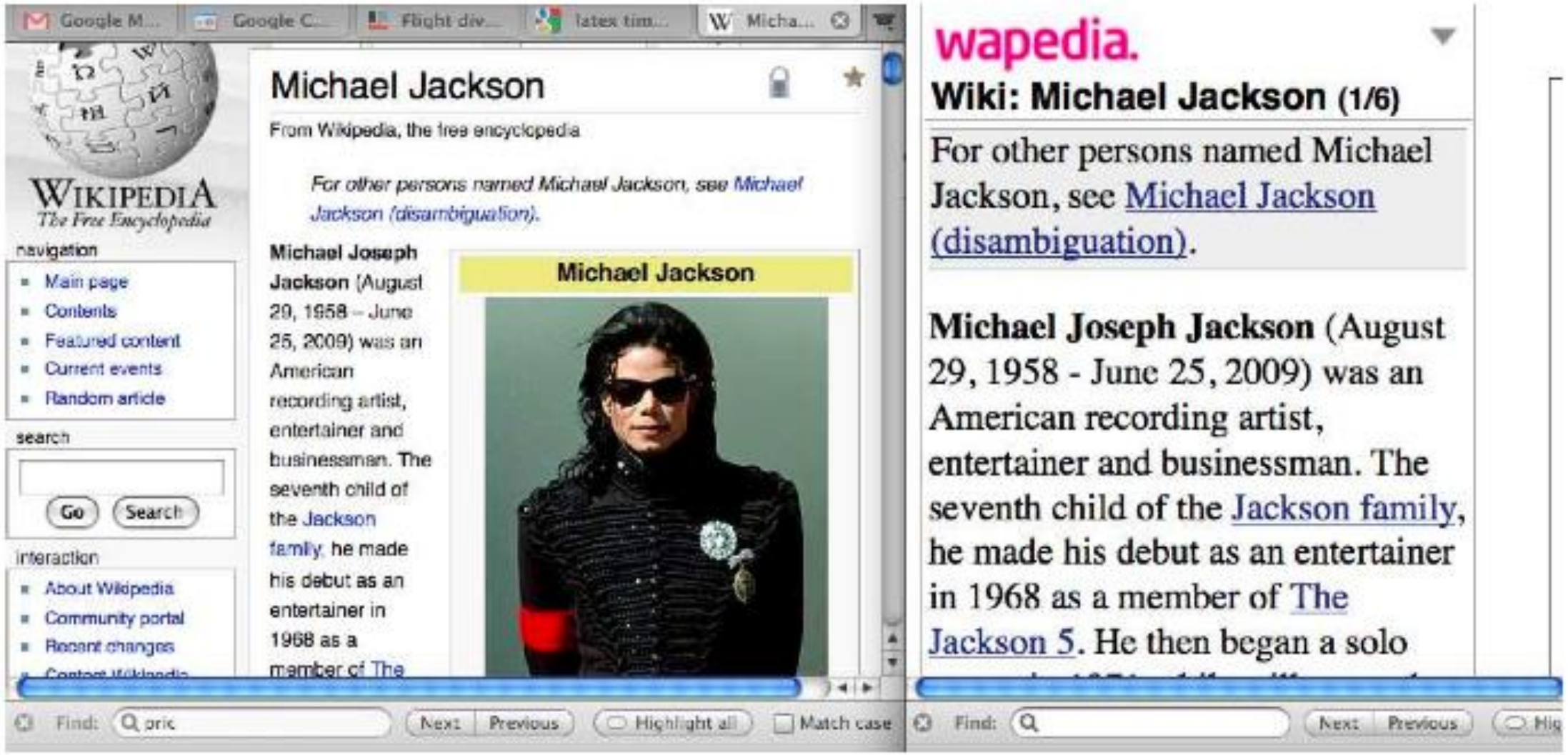
5. Duplicates – WEB IR

Duplicate detection

universität koblenz
weiter:denken

- The web is full of duplicated content
- More so than many other collections
- Exact duplicates
  - Easy to eliminate
  - E.g., use hash/fingerprint
- Near-duplicates
  - Abundant on the web
  - Difficult to eliminate
- For the user, it's annoying to get a search result with near-identical documents
- Marginal relevance is zero: even a highly relevant document becomes nonrelevant if it appears below a (near-)duplicate
- We need to eliminate near-duplicates

Near-duplicates: example



Detecting near-duplicates

- Compute similarity with an edit-distance measure
- We want “syntactic” (as opposed to semantic) similarity
  - True semantic similarity (similarity in content) is too difficult to compute
- We do not consider documents near-duplicates if they have the same content, but express it with different words
- Use similarity threshold θ to make the call “is/isn’t a near-duplicate“
- E.g., two documents are near-duplicates if similarity

$$> \theta = 80\%.$$

Represent each document as set of shingles

universität koblenz
weiter:denken

- A shingle is simply a word n-gram
- Shingles are used as features to measure syntactic similarity of documents
- For example, for $n = 3$, “a rose is a rose is a rose” would be represented as this set of shingles
  - { a-rose-is, rose-is-a, is-a-rose }
- We define the similarity of two documents as the Jaccard coefficient of their shingle sets

Recall: Jaccard coefficient

- A commonly used measure of overlap of two sets
- Let A and B be two sets
- Jaccard coefficient

$$\text{JACCARD}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

  - $(A \neq \emptyset \text{ or } B \neq \emptyset)$
  - JACCARD(A,A) = 1
  - JACCARD(A,B) = 0 if $A \cap B = 0$
  - A and B don't have to be the same size
  - Always assigns a number between 0 and 1

Jaccard coefficient: example

- Three documents
  - *d*1: “Jack London traveled to Oakland”
  - *d*2: “Jack London traveled to the city of Oakland”
  - *d*3: “Jack traveled from Oakland to London”
- Based on shingles of size 2 (2-grams or bigrams), what are the Jaccard coefficients $J(d1, d2)$ and $J(d1, d3)$?
  - $J(d1, d2) = 3/8 = 0.375$
  - $J(d1, d3) = 0$
  - Note: very sensitive to dissimilarity

universität koblenz
weiter:denken

6. User need – WEB IR

User needs

- **Informational** – want to learn about something (~40% / 65%) `Low hemoglobin`
- **Navigational** – want to go to that page (~25% / 15%) `United Airlines`
- **Transactional** – want to do something (web-mediated) (~35% / 20%)
  - Access a service `Seattle weather`
  - Downloads `Mars surface images`
  - Shop `Canon S410`
- **Gray areas**
  - Find a good hub `Car rental Brasil`
  - Exploratory search “see what’s there”
  - Need [Brod02, RL04]

Answering “the need behind the query”

- Semantic analysis
  - Query language determination
    - Auto filtering
    - Different ranking (if query in Japanese do not return English)
  - Hard & soft (partial) matches
    - Personalities (triggered on names)
    - Cities (travel info, maps)
    - Medical info (triggered on names and/or results)
    - Stock quotes, news (triggered on stock symbol)
    - Company info
    - Etc.
  - Natural Language reformulation
  - Integration of Search and Text Analysis

Answering "the need behind the query": context

universität koblenz
weiter:denken

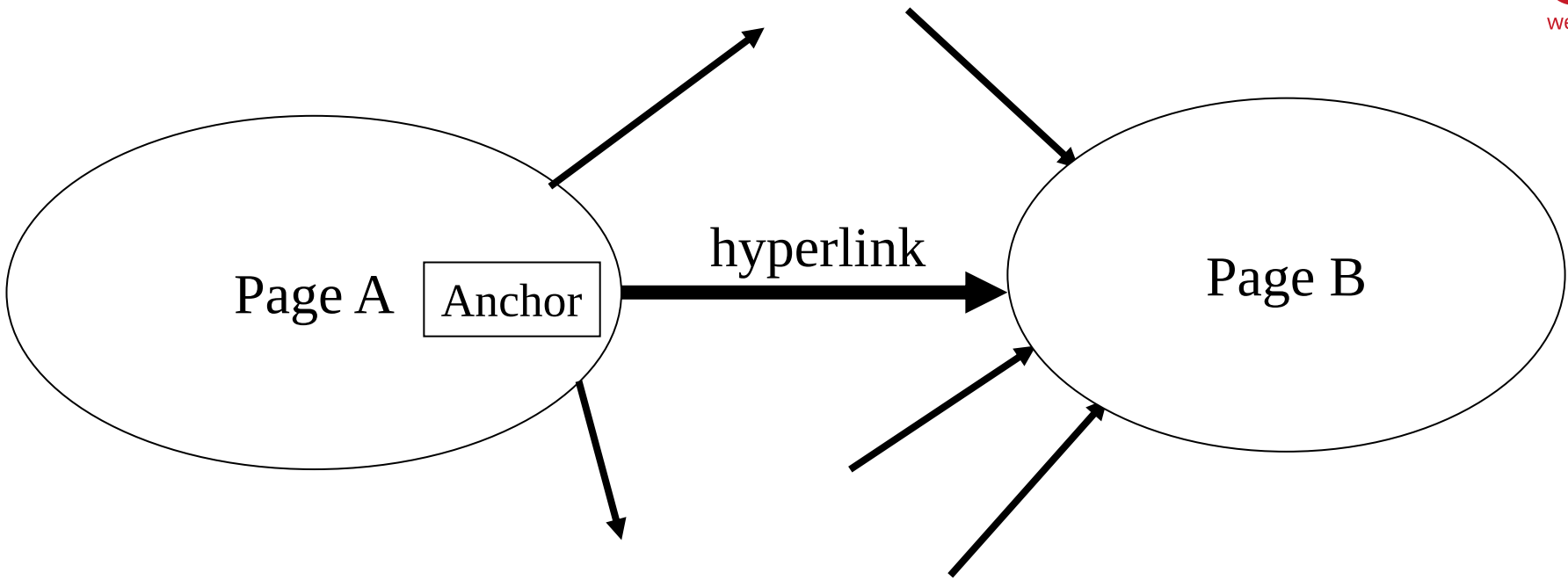
- Context determination
  - spatial (user location/target location)
  - query stream (previous queries)
  - personal (user profile)
  - explicit (user choice of a vertical search)
  - implicit (use Google from France, use google.fr)

- Context use
  - Result restriction
    - Kill inappropriate results
  - Ranking modulation
    - Use a "rough" generic ranking, but personalize later

universität koblenz
weiter:denken

7. Web graph

The Web as a directed graph

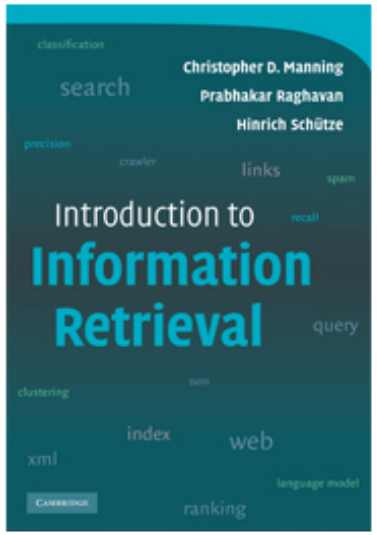


Hypothesis 1: A hyperlink between pages denotes a conferral of authority (quality signal)

Hypothesis 2: The text in the anchor of the hyperlink on page A describes the target page B

Assumption 1: reputed sites

Introduction to Information Retrieval



This is the companion website for the following book.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Informat*

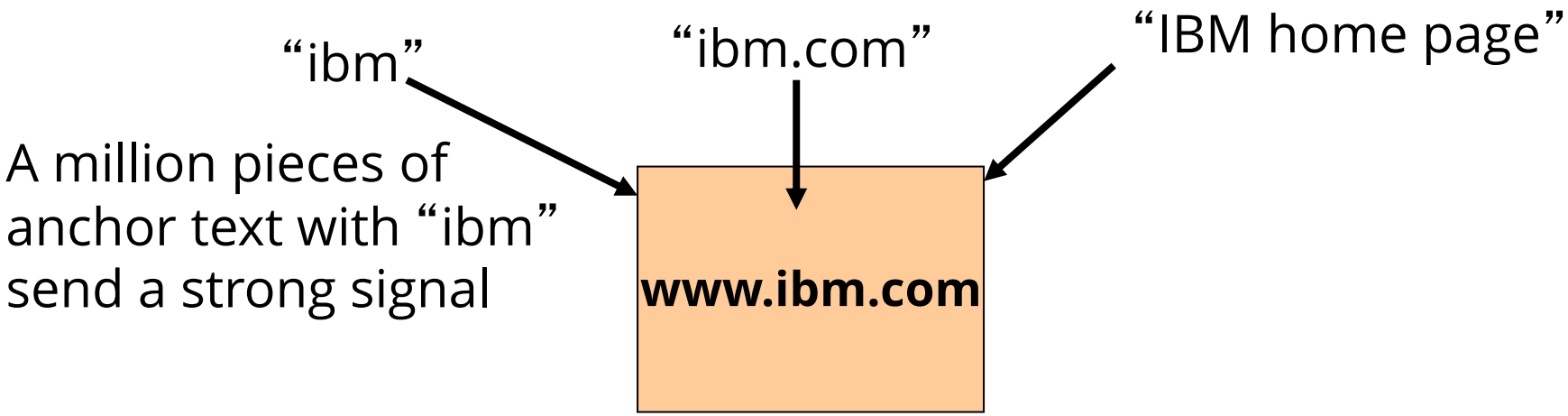
You can order this book at CUP, at your local bookstore or on the internet. The best search

The book aims to provide a modern approach to information retrieval from a computer scie
University and at the University of Stuttgart.

We'd be pleased to get feedback about how this book works out as a textbook, what is m
comments to: informationretrieval (at) yahoogroups (dot) com

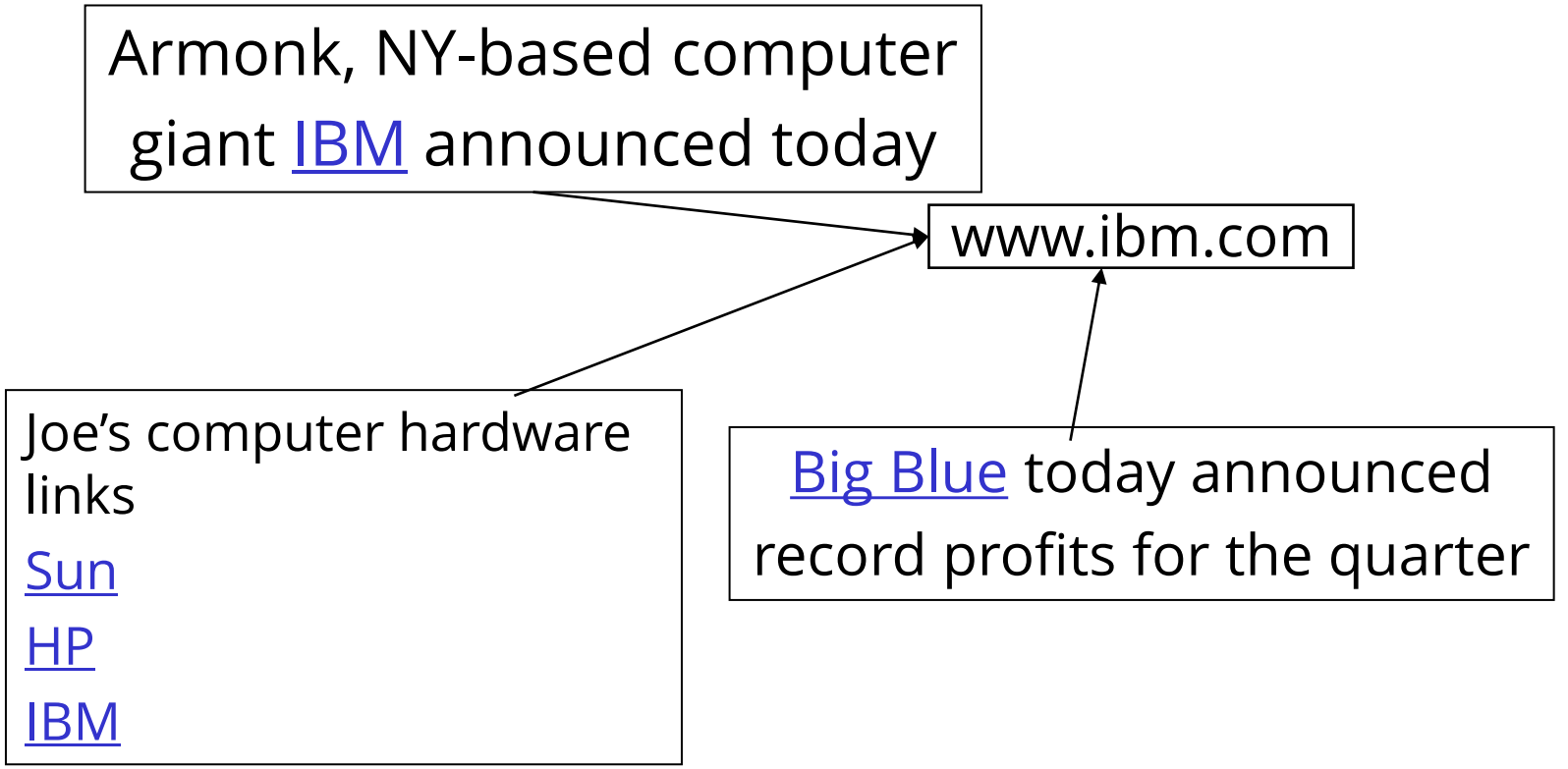
Anchor text

- For ***ibm*** how to distinguish between
  - IBM's home page (mostly graphical)
  - IBM's copyright page (high term freq. for 'ibm')
  - Rival's spam page (arbitrarily high term freq.)



Indexing anchor text

- When indexing a document *D*, include (with some weight) anchor text from links pointing to *D*



Indexing anchor text

- Thus: Anchor text is often a better description of a page's content than the page itself
- Anchor text can be weighted more highly than document text

universität koblenz
weiter:denken

8. Indexing

Zone index

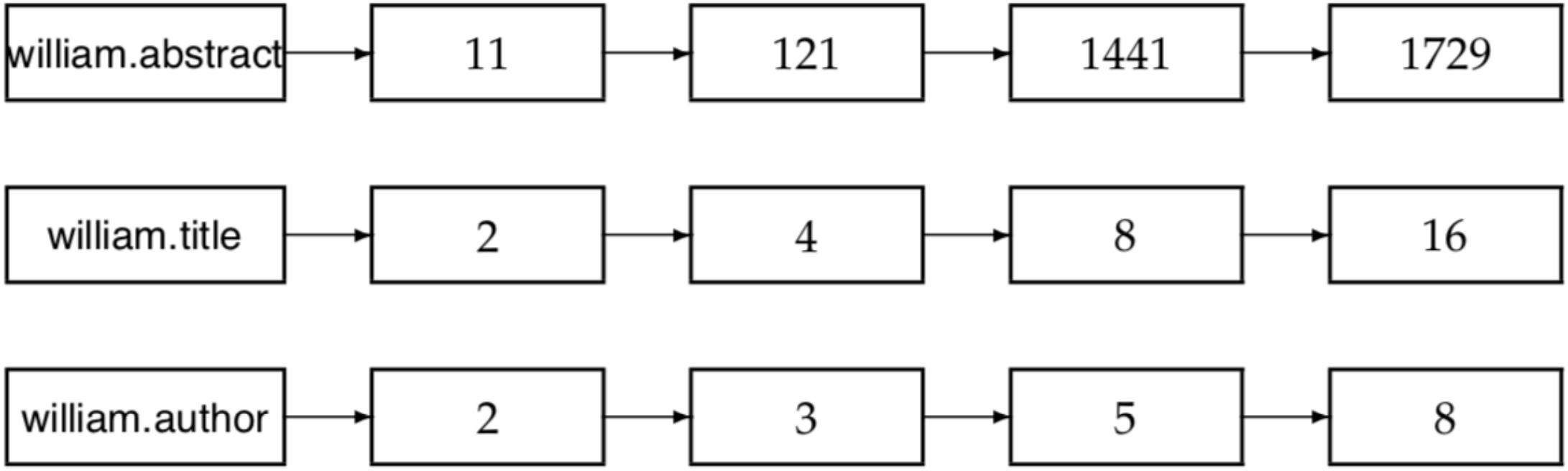
- Web pages with different zones
  - Title, metadata, author

- How to do weighted ranking

"find documents with *merchant* in the title and *william* in the author list and the phrase *gentle rain* in the body"

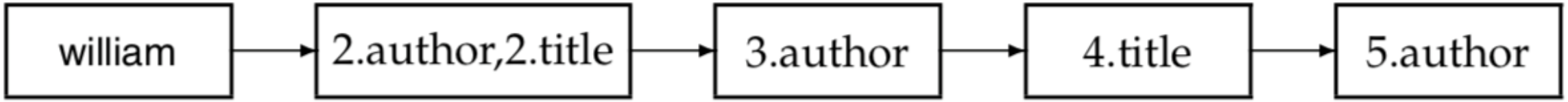
Basic zone index

- zones are encoded as extensions of dictionary entry



Advanced zone index

- zone is encoded in the postings rather than dictionary



Weighted zone scoring

- Consider a set of documents each of which has l zones

$$g_1, \cdots, g_l \in [0,1] \quad \sum_{i=1}^{l} g_i = 1$$

- s is the Boolean score denoting a match (or absence thereof) between *query* and the zone

$$\sum_{i=1}^{l} g_i s_i$$

Distributed indexing

universität koblenz
weiter:denken

- Web is large

- Term-partitioned index
- Document-partitioned index

- Map-reduce phenomena for the computation

universität koblenz
weiter:denken

9. User centric evaluation measures

Users' empirical evaluation of results

universität koblenz
weiter:denken

- Quality of pages varies widely
  - Relevance is not enough
  - Other desirable qualities (non IR!!)
    - Content: Trustworthy, diverse, non-duplicated, well maintained
    - Web readability: display correctly & fast
    - No annoyances: pop-ups, etc.
- Precision vs. recall
  - On the web, recall seldom matters
- What matters
  - Precision at 1? Precision above the fold?
  - Comprehensiveness – must be able to deal with obscure queries
    - Recall matters when the number of matches is very small
- User perceptions may be unscientific, but are significant over a large aggregate

Users' empirical evaluation of engines

- Relevance and validity of results
- UI – Simple, no clutter, error tolerant
- Trust – Results are objective
- Coverage of topics for polysemic queries
- Pre/Post process tools provided
  - Mitigate user errors (auto spell check, search assist,…)
  - Explicit: Search within results, more like this, refine …
  - Anticipative: related searches
- Deal with idiosyncrasies
  - Web specific vocabulary
    - Impact on stemming, spell-check, etc.
  - Web addresses typed in the search box
- "The first, the last, the best and the worst …"

universität koblenz
weiter:denken

10. Summary

Summary

- Web search basic
  - Ads
  - Spams
  - Duplicates
  - User Needs
  - Web Graph (anchor text)

References

[1] https://olat.vcrp.de/auth/RepositoryEntry/4071063853

[2] https://nlp.stanford.edu/IR-book/information-retrieval-book.html
Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008

- Chapter 4 (Distributed indexing)
- Chapter 6 (Parametric and zone indexes)
- Chapter 19 (Web search basics)