Information Retrieval

Weike Pan

The slides are **adapted from those provided by Prof. Hinrich Schütze** at University of Munich (http://www.cis.lmu.de/~hs/teach/14s/ir/).

Chapter 19 Web search basics

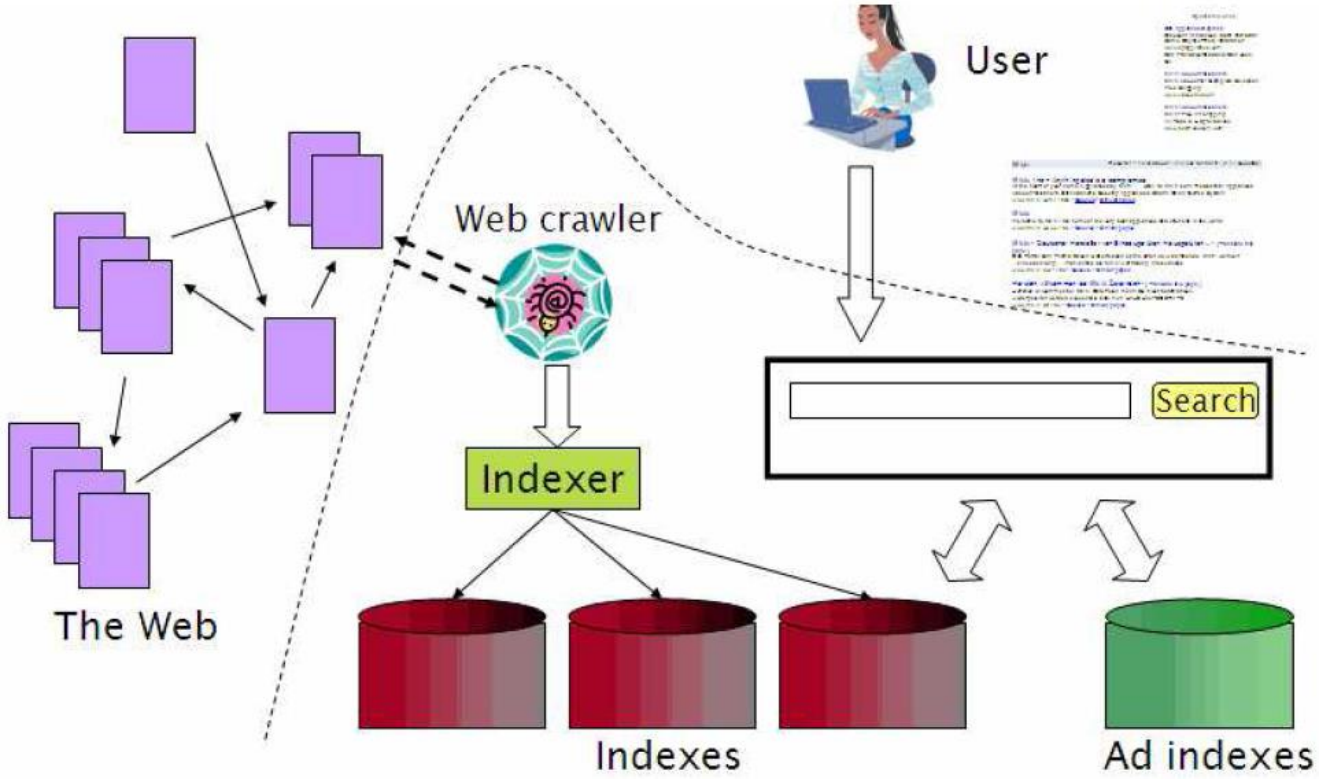
- 19.1 Background and history
- 19.2 Web characteristics
- 19.3 Advertising as the economic model
- 19.4 The search user experience
- 19.5 Index size and estimation
- 19.6 Near-duplicates and shingling
- 19.7 References and further reading

Outline

- 19.1 Background and history
- 19.2 Web characteristics
- 19.3 Advertising as the economic model
- 19.4 The search user experience
- 19.5 Index size and estimation
- 19.6 Near-duplicates and shingling
- 19.7 References and further reading

19.1 Background and history

Web search overview



19.1 Background and history

Search is a top activity on the web

- How often do you use search engines on the Internet?
  - Four or more times each day
  - At least once every day
  - Several times each week
  - At least once each week
  - Several times each month
  - Less frequently
  - Never

19.1 Background and history

Without search engines, the web wouldn't work

- Without search, content is hard to find
- Without search, there is no incentive (激励) to create content
  - Why publish something if nobody will read it?
  - Why publish something if I don't get ad revenue from it?

- Somebody needs to pay for the web
  - Servers, web infrastructure, content creation
  - A large part today is paid by search ads
  - Search pays for the web

19.1 Background and history

Interest aggregation

- Unique feature of the web: A small number of geographically dispersed (分散的) people with similar interests can find each other
  - Elementary school kids with hemophilia (血友病)
  - People interested in translating R5R5 Scheme into relatively portable C (open source project)
  - Search engines are a key enabler (使能者, 赋能者) for interest aggregation

19.1 Background and history

IR on the web vs. IR in general

- On the web, search is not just a nice feature
  - Search is a key enabler (使能者, 赋能者) of the web: financing, content creation, interest aggregation, etc.

- The web is a chaotic and uncoordinated collection → lots of duplicates -- need to detect duplicates
- No control/restrictions on who can author content → lots of spam -- need to detect spam
- The web is very large → need to know how big it is

Outline

- 19.1 Background and history
- 19.2 Web characteristics
- 19.3 Advertising as the economic model
- 19.4 The search user experience
- 19.5 Index size and estimation
- 19.6 Near-duplicates and shingling
- 19.7 References and further reading

19.2 Web characteristics

The goal of spamming on the web

- You have a page that will generate lots of revenue for you if people visit it.
- Therefore, you would like to direct visitors to this page.
- One way of doing this: get your page ranked highly in search results.
- Question: How can I get my page ranked highly?

19.2 Web characteristics

Spam technique: Keyword stuffing (堆砌)/hidden text

- Misleading meta-tags, excessive repetition (大量重复)
- Hidden text with colors, style sheet tricks, etc.
- Used to be very effective, most search engines now can catch these

19.2 Web characteristics

Spam technique: Doorway pages and lander pages

- Doorway page: optimized for a single keyword, redirects to the real target page

- Lander page: optimized for a single keyword or a misspelled domain name, designed to attract surfers who will then click on ads

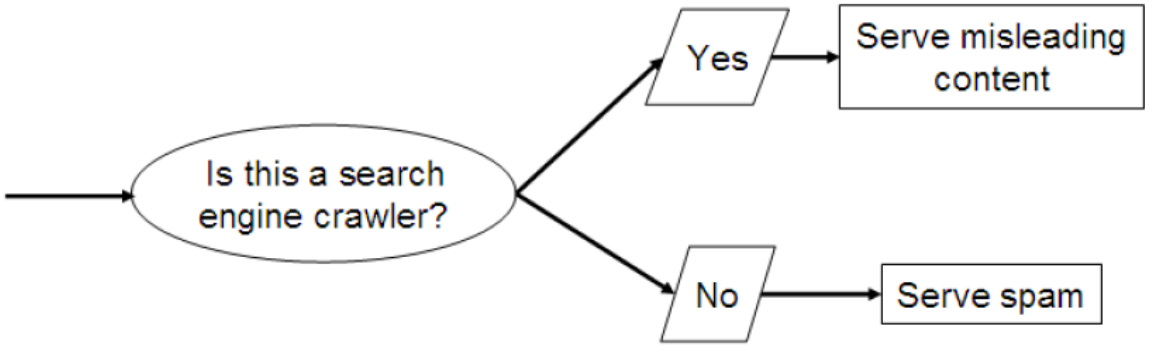
19.2 Web characteristics

Spam technique: Duplication

- Get good content from somewhere (steal it or produce it yourself)
- Publish a large number of slight variations of it

19.2 Web characteristics

Spam technique: Cloaking (隐藏页,障眼法,伪装技术)



- Serve fake content to search engine spider

19.2 Web characteristics

- Spam technique: Link spam

- Create lots of links pointing to the page you want to promote

- Put these links on pages with high (or at least non-zero) PageRank
  - Newly registered domains (domain flooding)
  - A set of pages that all point to each other to boost each other's PageRank (mutual admiration society)
  - Pay somebody to put your link on their highly ranked page
  - Leave comments that include the link on blogs

19.2 Web characteristics

SEO: Search engine optimization

- Promoting a page in the search rankings is not necessarily spam.
- It can also be a legitimate business (合法的商业) -- which is called SEO.

- You can hire an SEO firm to get your page ranked highly
- Restructure your content in a way that makes it easy to index
- Talk with influential bloggers and have them link to your site
- Add more interesting and original content

19.2 Web characteristics

- The terms Google bomb and Googlewashing refer to the practice of causing a website to rank highly in web search engine results for irrelevant, unrelated or off-topic search terms by linking heavily.

- In contrast, search engine optimization (SEO) is the practice of improving the search engine listings of web pages for relevant search terms.

https://en.wikipedia.org/wiki/Google_bomb

19.2 Web characteristics

The war against spam

- Quality indicators
  - Links, statistically analyzed (PageRank, etc.)
  - Usage (users visiting a page)
  - No adult content
  - Distribution and structure of text (e.g., no keyword stuffing)
- Combine all of these indicators and use machine learning techniques

- Editorial intervention (干预): Blacklists, top queries audited (审核), complaints addressed, suspect patterns detected

19.2 Web characteristics

Webmaster guidelines

- Major search engines have guidelines for webmasters
- These guidelines tell you what is legitimate SEO and what is spamming
- Ignore these guidelines at your own risk
- Once a search engine identifies you as a spammer, all pages on your site may get low ranks (or disappear from the index entirely)
- There is often a fine line (明显的界限) between spam and legitimate SEO

- Scientific study of fighting spam on the web: adversarial information retrieval (对抗性信息检索)

Outline

- 19.1 Background and history
- 19.2 Web characteristics
- 19.3 Advertising as the economic model
- 19.4 The search user experience
- 19.5 Index size and estimation
- 19.6 Near-duplicates and shingling
- 19.7 References and further reading

19.3 Advertising as the economic model

First generation of search ads: Goto (1996)

- Buddy Blake bid the maximum ($0.38) for this search.
- He paid $0.38 to Goto every time somebody clicked on the link.
- Pages were simply ranked according to bid – revenue maximization for Goto.
- No separation of ads/docs. Only one result list!
- Upfront (坦率的) and honest. No relevance ranking, but Goto did not pretend there was any.

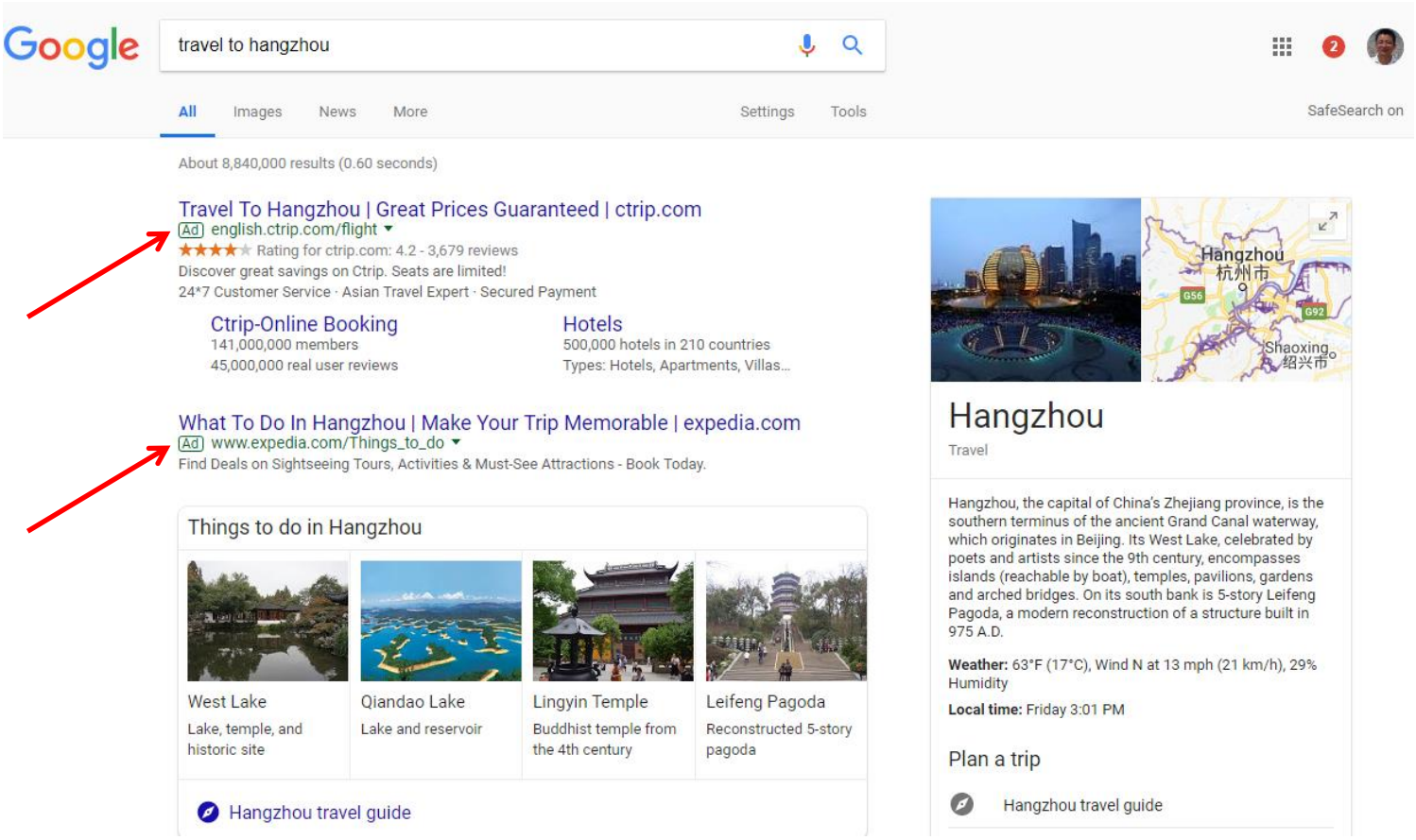
19.3 Advertising as the economic model

Second generation of search ads: Google (2000/2001)

- Strict separation of search results and search ads

19.3 Advertising as the economic model

- Example



19.3 Advertising as the economic model

- Example

19.3 Advertising as the economic model

Do ads influence editorial content?

- Similar problem at newspapers/TV channels.
- A newspaper is reluctant to publish harsh criticism of its major advertisers.
- The line often gets blurred (变得模糊) at newspapers/on TV.
- No known case of this happening with search engines yet?

19.3 Advertising as the economic model

How are the ads ranked? (1/3)

- Advertisers bid for keywords -- sale by auction (拍卖).
- Open system: Anybody can participate and bid on keywords.
- Advertisers are only charged when somebody clicks on your ad.
- How does the auction determine an ad's rank and the price paid for the ad?
  - Basis is a second price auction (次价拍卖)
  - For the bottom line, this is perhaps the most important research area for search engines -- **computational advertising (计算广告学)**

19.3 Advertising as the economic model

How are the ads ranked? (2/3)

- First cut: according to bid price
  - Bad idea: open to abuse
  - Example: query [treatment for cancer?] → how to write your last will (遗嘱)
  - We don't want to show nonrelevant or offensive (冒犯的) ads.

- Instead: rank based on bid price and relevance
  - Ad relevance: clickthrough rate (CTR) = clicks per impression
  - A nonrelevant ad will be ranked low. Even if this decreases search engine short-term revenue.
  - Hope: Overall acceptance of the system and overall revenue is maximized if users get useful information.

19.3 Advertising as the economic model

How are the ads ranked? (3/3)

- Other ranking factors: location, time of day, quality and loading speed of landing page

- The main ranking factor: the query

19.3 Advertising as the economic model

Google's second price (sealed-bid) auction (次价密封投标拍卖)

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

0.08/0.03+0.01=2.68

0.09/0.06+0.01=1.51

0.04/0.08+0.01=0.51

- bid: maximum bid for a click by advertiser
- CTR: click-through rate
- ad rank: bid * CTR
- rank: rank in auction
- paid: second price (sealed-bid) auction price paid by advertiser

https://www.wordstream.com/articles/what-is-google-adwords

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

19.3 Advertising as the economic model

- Keywords with high bids

$69.1	mesothelioma treatment options
$65.9	personal injury lawyer michigan
$62.6	student loans consolidation
$61.4	car accident attorney los angeles
$59.4	online car insurance quotes
$59.4	arizona dui lawyer
$46.4	asbestos cancer
$40.1	home equity line of credit
$39.8	life insurance quotes
$39.2	refinancing
$38.7	equity line of credit
$38.0	lasik eye surgery new york city
$37.0	2nd mortgage
$35.9	free car insurance quote

19.3 Advertising as the economic model

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 an ad.
- The advertiser finds new customers in a cost-effective way.

19.3 Advertising as the economic model

Question

- Why is web search potentially more attractive for advertisers than TV spots (电视广告), 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?

19.3 Advertising as the economic model

Not a win-win-win: Keyword arbitrage (套利)

- 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.

19.3 Advertising as the economic model

Not a win-win-win: Violation of trademarks (商标)

- Example: geico
  - During part of 2005: The search term “geico” on Google was bought by competitors.
  - Geico lost this case (没有打赢官司) in the United States.

- Louis Vuitton lost a similar case in Europe.

- It’s potentially misleading to users to trigger an ad off of a trademark if the user can’t buy the product on the site.

Outline

- 19.1 Background and history
- 19.2 Web characteristics
- 19.3 Advertising as the economic model
- 19.4 The search user experience
- 19.5 Index size and estimation
- 19.6 Near-duplicates and shingling
- 19.7 References and further reading

19.4 The search user experience

- Query distribution (1/2)

- More than 1/3 of these are queries for adult content.

19.4 The search user experience

Query distribution (2/2)

- Queries have a power law distribution

- Recall Zipf's law: a few very frequent words, a large number of very rare words
- Same here: a few very frequent queries, a large number of very rare queries

19.4 The search user experience

Types of queries/user needs in web search (1/4)

- **Informational user needs**: I need information on something.
  - Rhinallergosis (过敏性鼻炎)

19.4 The search user experience

Types of queries/user needs in web search (2/4)

- **Navigational user needs**: I want to go to this web site.
  - "hotmail", "myspace", "United Airlines"

19.4 The search user experience

Types of queries/user needs in web search (3/4)

- **Transactional user needs**: I want to make a transaction.
  - Buy something: “MacBook Air”
  - Download something: “Acrobat Reader”
  - Chat with someone: “live soccer chat”

19.4 The search user experience

Types of queries/user needs in web search (4/4)

- Difficult problem: How can the search engine tell what the user need or intent for a particular query is?

19.4 The search user experience

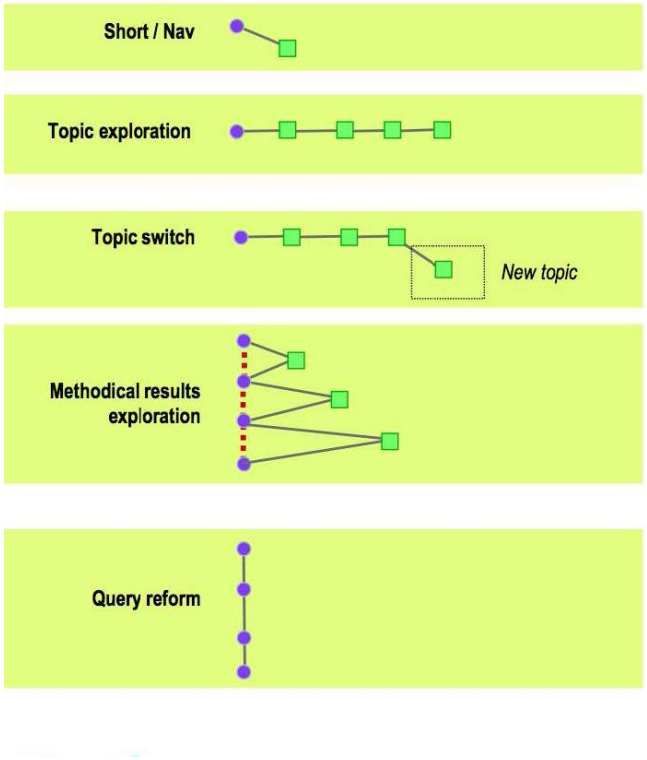
Search in a hyperlinked collection

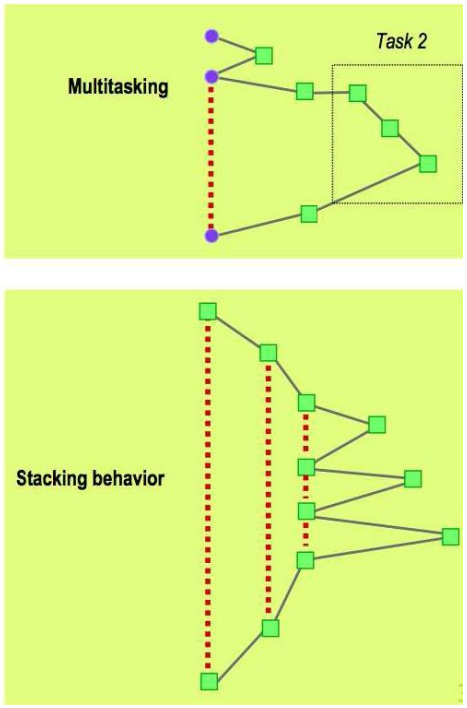
- Web search in most cases is interleaved (交错) with navigation … i.e., with following links.

- Different from most other IR collections

19.4 The search user experience

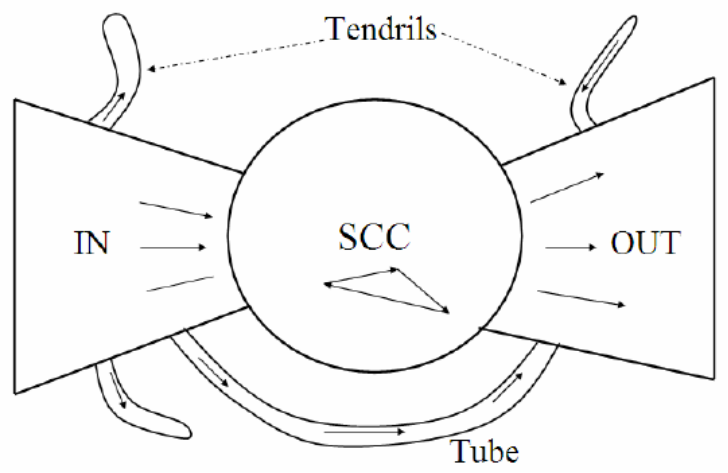
- Kinds of behaviors





19.4 The search user experience

Bowtie (领结) structure of the web



- Strongly connected component (SCC) in the center
- Lots of pages that get linked to, but don't link (OUT)
- Lots of pages that link to other pages, but don't get linked to (IN)
- Tendrils (卷须), tubes, islands

19.4 The search user experience

User intent: Answering the need behind the query

- What can we do to guess user intent?
- Guess user intent independent of context
  - Spell correction
  - Precomputed "typing" of queries
- Better: Guess user intent based on context
  - Geographic context
  - Context of user in this session (e.g., previous query)
  - Context provided by personal profile (用户画像)

19.4 The search user experience

Guessing of user intent by “typing” queries

- Calculation: 5+4
- Unit conversion (单位换算): 1 kg in pounds
- Currency conversion (货币兑换): 1 euro in kronor
- Tracking number: 8167 2278 6764
- Flight info: LH 454
- Area code: 650
- Map: columbus oh (俄亥俄州哥伦布)
- Stock price: msft (微软公司的股票代码)
- Albums/movies etc.: coldplay (a British rock band formed in London in 1996)

19.4 The search user experience

The spatial context: Geo-search (1/2)

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania (阿尔巴尼亚))
  - User (located in Palo Alto (帕罗奥多))

19.4 The search user experience

The spatial context: Geo-search (2/2)

- Locating the user
  - IP address
  - Information provided by user (e.g., in user profile)
  - Mobile phone

- Geo-tagging: Parse text and identify the coordinates of the geographic entities
  - Example: East Palo Alto CA → Latitude: 37.47 N, Longitude: 122.14 W

19.4 The search user experience

How do we use context to modify query results?

- Result **restriction**: Don't consider inappropriate results
  - For user on google.fr … only show .fr results

- Ranking modulation: use a rough generic ranking, **re-ranking** based on personal context

- **Contextualization/personalization** is an area of search with a lot of potential for improvement.

19.4 The search user experience

Users of web search

- Use short queries (average < 3)
- Rarely use operators
- Don't want to spend a lot of time on composing a query
- Only look at the first couple of results
- Want a simple UI, not a search engine start page overloaded with graphics
- Extreme variability in terms of user needs, user expectations, experience, knowledge: Industrial/developing world, English/Estonian (爱沙尼亚语), old/young, rich/poor, differences in culture and class
- One interface for hugely divergent needs

19.4 The search user experience

How do users evaluate search engines?

- Classic IR relevance (as measured by F) can also be used for web IR
- Equally important: Trust, duplicate elimination, readability, loads fast, no pop-ups

- On the web, **precision** is more important than recall
  - Precision at 1, precision at 10, precision on the first 2-3 pages

- But there is a subset of queries where **recall** matters

19.4 The search user experience

Web information needs that require high recall

- Has this **idea** been patented (已经申请专利了) or published (已经发表了)?
- …

19.4 The search user experience

Web documents: different from other IR collections

- Distributed content creation: no design, no coordination
  - Result: extreme heterogeneity (异构/异质) of documents on the web

- Unstructured (text, html), semistructured (html, xml), structured/relational (databases)
- Dynamically generated content

19.4 The search user experience

Dynamic content

- Dynamic pages are generated from scratch when the user requests them – usually from underlying data in a database
  - Example: current status of flight LH 454

- Most (truly) dynamic content is ignored by web spiders. It's too much to index it all. (暗网, e.g., 百度的阿拉丁计划)

19.4 The search user experience

Multilinguality

- Documents in a large number of languages
- Queries in a large number of languages
- First cut: Don't return English results for a Japanese query

- However: Frequent mismatches query/document languages
- Many people can understand, but not query in a language
- Translation is important

- Google example: "Beaujolais Nouveau -wine"

19.4 The search user experience

Duplicate documents

- Significant duplication – 30%–40% duplicates in some studies
- Duplicates in the search results were common in the early days of the web
- Today's search engines eliminate duplicates very effectively
- Key for high user satisfaction

19.4 The search user experience

Trust

- For many collections, it is easy to assess the trustworthiness of a document.
- Web documents are different: In many cases, we don't know how to evaluate the information

Outline

- 19.1 Background and history
- 19.2 Web characteristics
- 19.3 Advertising as the economic model
- 19.4 The search user experience
- 19.5 Index size and estimation
- 19.6 Near-duplicates and shingling
- 19.7 References and further reading

19.5 Index size and estimation

Size of the web: Who cares?

- Media
- Users
  - They may switch to the search engine that has the best coverage of the web.
  - Users (sometimes) care about recall. If we underestimate (低估) the size of the web, search engine results may have low recall.
- Search engine designers (how many pages do I need to be able to handle?)
- Crawler designers

19.5 Index size and estimation

What is the size of the web? Any guesses?

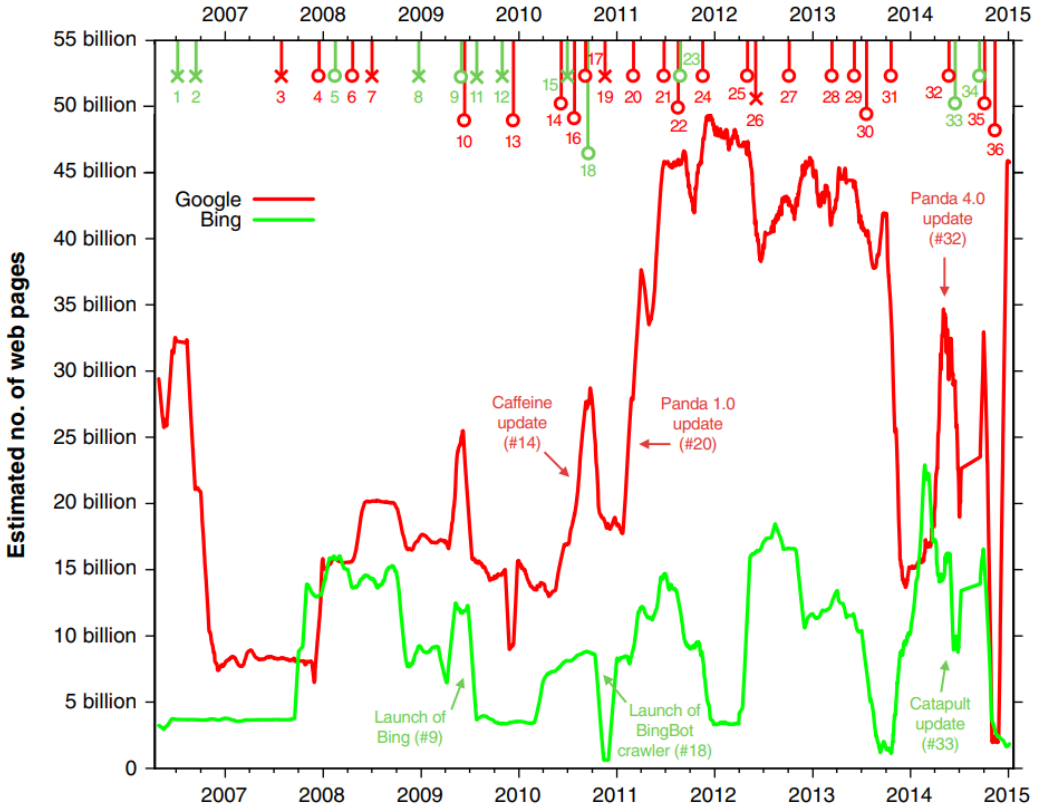
19.5 Index size and estimation

Sampling methods

- Random queries
- Random searches
- Random IP addresses
- Random walks

19.5 Index size and estimation

- Antal van den Bosch, Toine Bogers, Maurice de Kunder. **A Longitudinal Analysis of Search Engine Index Size**. In: Proceedings of ISSI 2015, June 2015



19.5 Index size and estimation

- The Indexed Web contains at least 4.26 billion pages (Friday, 06 April, 2018).
- The Indexed Web contains at least 5.64 billion pages (Monday, 10 June, 2019).
- The Indexed Web contains at least 5.48 billion pages (Monday, 15 June, 2020).
- The Indexed Web contains at least 2.4 billion pages (Thursday, 10 June, 2021).

http://www.worldwidewebsize.com/

Outline

- 19.1 Background and history
- 19.2 Web characteristics
- 19.3 Advertising as the economic model
- 19.4 The search user experience
- 19.5 Index size and estimation
- 19.6 Near-duplicates and shingling
- 19.7 References and further reading

19.6 Near-duplicates and shingling

Duplicate detection

- The web is full of duplicated content.
- Exact duplicates: Easy to eliminate, e.g., use hash/fingerprint
- Near-duplicates: Difficult to eliminate

- For the user, it is 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.
- Hence, we need to eliminate near-duplicates.

19.6 Near-duplicates and shingling

Exercise

- How would you eliminate near-duplicates on the web?

19.6 Near-duplicates and shingling

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 to make the call "is/isn't a near-duplicate", e.g., two documents are near-duplicates if similarity > = 80%.

19.6 Near-duplicates and shingling

Recall: Jaccard coefficient

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

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

19.6 Near-duplicates and shingling

Represent each document as a set of shingles

- 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.

19.6 Near-duplicates and shingling

Example

- Three documents:
  - d1: “Jack London traveled to Oakland”
  - d2: “Jack London traveled to the city of Oakland”
  - d3: “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 (对差异非常敏感)

19.6 Near-duplicates and shingling

Represent each document as a **sketch**

- The number of shingles per document is large.
- To increase **efficiency**, we will use a sketch, a cleverly chosen subset of the shingles of a document.
- The size of a sketch is, say, n = 200 (即进行200次的随机排序)…

- But how do we compute the Jaccard coefficient?
  - The **proportion** of successful permutations is the Jaccard coefficient (see Theorem 19.1)

19.6 Near-duplicates and shingling

Permutation and minimum

d_1	d_2	d_3
0	1	1
1	0	1
0	1	0
1	0	0

Two permutations: [2 3 0 1], [0 3 2 1]

- $\hat{J}(d_1, d_2) = 0/2$
- $\hat{J}(d_1, d_3) = 0/2$
- $\hat{J}(d_2, d_3) = 1/2$

For d1 and d2:
d1(2)=0, d1(3)=1, d2(2)=1;
d1(0)=0, d1(3)=1, d2(0)=1;
hence, 0/2

For d1 and d3:
d1(2)=0, d1(3)=1, d3(2)=0, d3(3)=0, d3(0)=1;
d1(0)=0, d1(3)=1, d3(0)=1;
hence, 0/2

For d2 and d3:
d2(2)=1, d3(2)=0, d3(3)=0, d3(0)=1;
d2(0)=1, d3(0)=1;
hence, 1/2

19.6 Near-duplicates and shingling

Efficient near-duplicate detection

- Now we have an extremely efficient method for estimating a Jaccard coefficient for two documents.
- But we still have to estimate O(N^2) coefficients where N is the number of web pages -> Still intractable
  - One solution: locality sensitive hashing (LSH)
  - Another solution: sorting (Henzinger 2006)

Summary

- 19.1 Background and history
- 19.2 Web characteristics
- 19.3 Advertising as the economic model
- 19.4 The search user experience
- 19.5 Index size and estimation
- 19.6 Near-duplicates and shingling
- 19.7 References and further reading