

IN4325



Personalization in (web) search

Claudia Hauff (WIS, TU Delft)

Project interview questions

Your project in the course context.

You have used retrieval algorithm X in your project, can you tell me how it differs broadly from the vector space model?

You used relevance feedback in your project, could you use pseudo-relevance feedback as well?

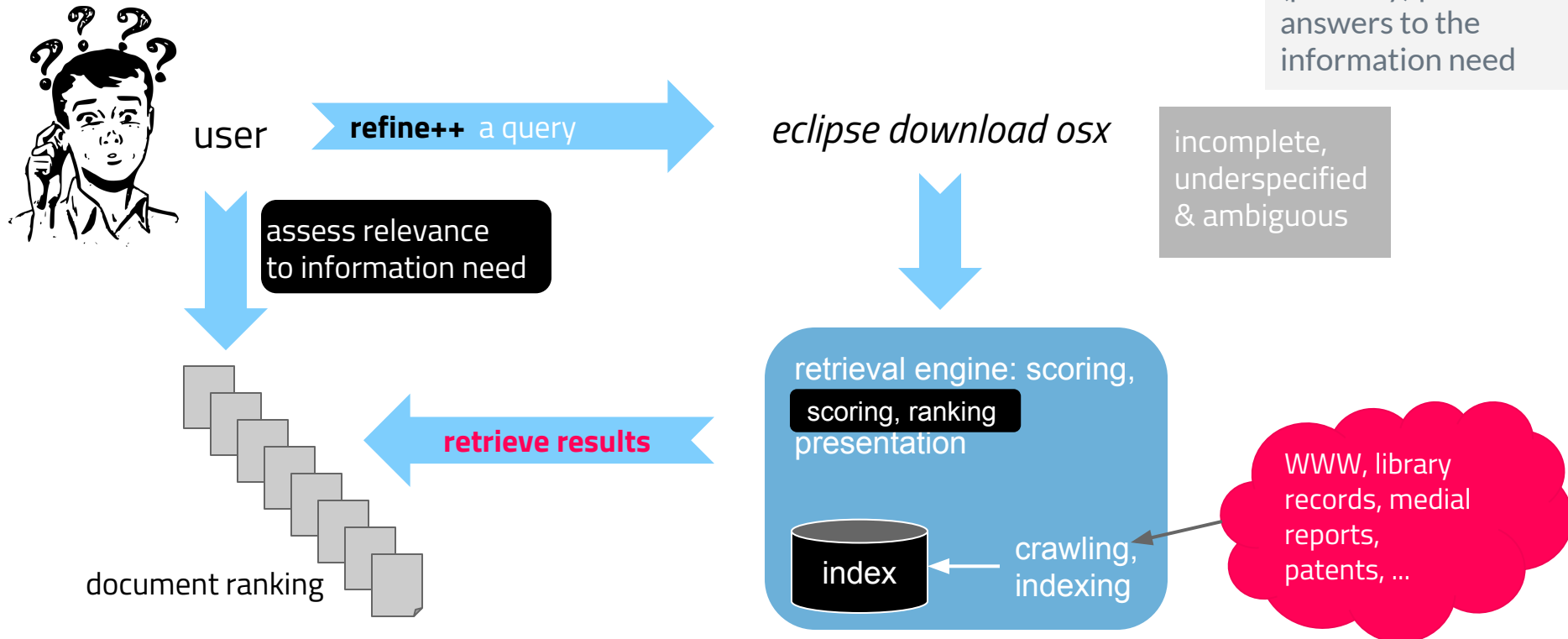
You have used metrics X, Y and Z. How do they differ from each other?

Your project took place on domain X, could you apply a similar pipeline to domain Y? What would you change?

The big picture

The essence of IR

Information need: *Looks like I need Eclipse for this job. Where can I download the latest beta version for macOS Sierra?*



Information need

Topic the user wants to know more about

Query

Translation of need into an input for the search engine

Relevance

A document is relevant if it (partially) provides answers to the information need

A hot topic for industry ...

"... The method comprises, at a computerized search engine system distinct from a client system: receiving a search request associated with a user from the client system, the search request having one or more search terms; obtaining a user profile corresponding to the user, where the user profile is generated based in part on the user's prior computing activities, comprising one or more of browsing, searching, and messaging; obtaining search results for the search request; generating a personalized snippet for at least one of the search results in accordance with the obtained user profile, the snippet comprising a text portion of the search result chosen based on at least one or more search terms and one or more terms of the obtained user profile; and transmitting the search results and personalized snippet to the client system for display." (ONE sentence)

patent retrieval

personalized search patent

About 17,600 results (0.11 sec)

System and method for **personalized** snippet generation

[TH Haveliwala, SD Kamvar](#) - US **Patent** 9,805,116, 2017 - Google Patents

... No. 14/154,071, filed Jan. 13, 2014 which is a continuation of US **patent** application Ser. No. ... FIG. 2 is a flow chart for a process for generating **personalized** snippets for a set of **search** results in accordance with some embodiments of the present invention. FIG ...

☆ ⓘ Cited by 16 Related articles All 4 versions Import into EndNote ⓘ

System and Method for **Personalized Search** While Maintaining Searcher Privacy

[PV Hayes](#) - US **Patent** App. 15/183,619, 2017 - Google Patents

... 0012]. The present disclosure relates to a system and method for **personalized search** while maintaining ... of ResultRank, indicated and/or inferred searcher satisfaction with the relevance of **search** result abstracts ... The term Result Rank was introduced in US **patent** application Ser. ...

☆ ⓘ All 2 versions Import into EndNote ⓘ

Adaptive Reading Level Assessment for **Personalized Search**

[DJ Weiss, E Mitsakaki](#) - US **Patent** App. 15/650,173, 2017 - freepatentsonline.com

... Title: Adaptive Reading Level Assessment for **Personalized Search**. Document Type and Number: United States **Patent** Application 20170372628 Kind Code: A1. Abstract: A system and associated methods are provided for generating ...

☆ ⓘ Import into EndNote ⓘ

Personalized search result summary

[BW Chang, SK Rakshit](#) - US **Patent** 9,779,170, 2017 - Google Patents

... a **search** query, if the personalization system 120 identifies the **search** result as related to a **patent**, then the personalization system 120 selects the **patent** summary template 200 and displays the **personalized** summary for the **search** result using the **patent** summary template 200 ...

☆ ⓘ Cited by 1 Related articles All 4 versions Import into EndNote ⓘ

Methods, systems and techniques for **personalized search** query suggestions

[S Zhu, CM Sze, H Su, H Wu, H Wu, J Gan...](#) - US **Patent** App. 14 ..., 2017 - Google Patents

Connect public, paid and private **patent** data with Google Patents Public Datasets Methods, systems and techniques for **personalized search** query suggestions. Download PDF Info. Publication number US20170097939A1. Authority ...

☆ ⓘ All 2 versions Import into EndNote ⓘ

Personalized search library based on continual concept correlation

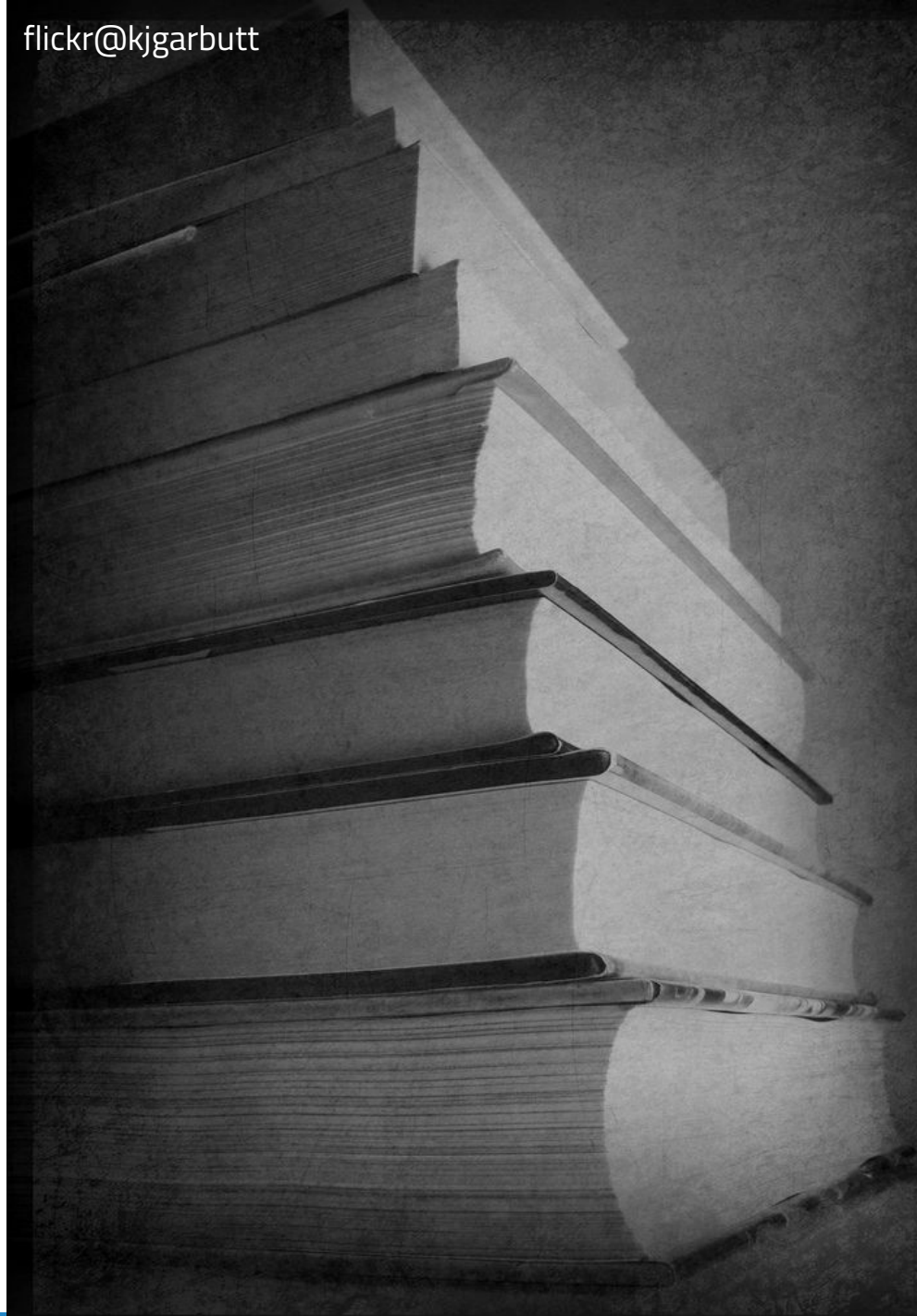
[S Mo, RH Wouhaybi, MA Mian, TM Kohlenberg...](#) - US **Patent** ..., 2017 - Google Patents

... Download PDF Info. Publication number US9582572B2. Authority US Grant status Grant. **Patent** type. Prior art keywords content **personalized search** user concept Prior art date 2012-12-19 Legal status (The legal status is an assumption and is not a legal conclusion ...

☆ ⓘ Cited by 2 Related articles All 4 versions Import into EndNote ⓘ

Topics

- Evaluation strategies
- Search strategies
- Privacy and software architectures
- HITS
- (Personalized) PageRank



Personalized search: evaluation scenarios


Search tailored to a user's interests.

Common experimental approach

- 1) Retrieve the top N search results for query Q from a search engine (result list $R1$) [*personalization switched off]
- 2) Compute the personalized score for each document in $R1$ and **rerank** by it (result list $R2$)
- 3) Combine the rankings $R1$ and $R2$ through **Borda count** (generalization of majority vote), yielding the personalized result list R

User-based eval.

Click-based eval.



R1	R2
D1 (4p)	D3 (4p)
D2 (3p)	D4 (3p)
D3 (2p)	D2 (2p)
D4 (1p)	D1 (1p)



Can we build a reusable test collection (enabling offline evaluation)?

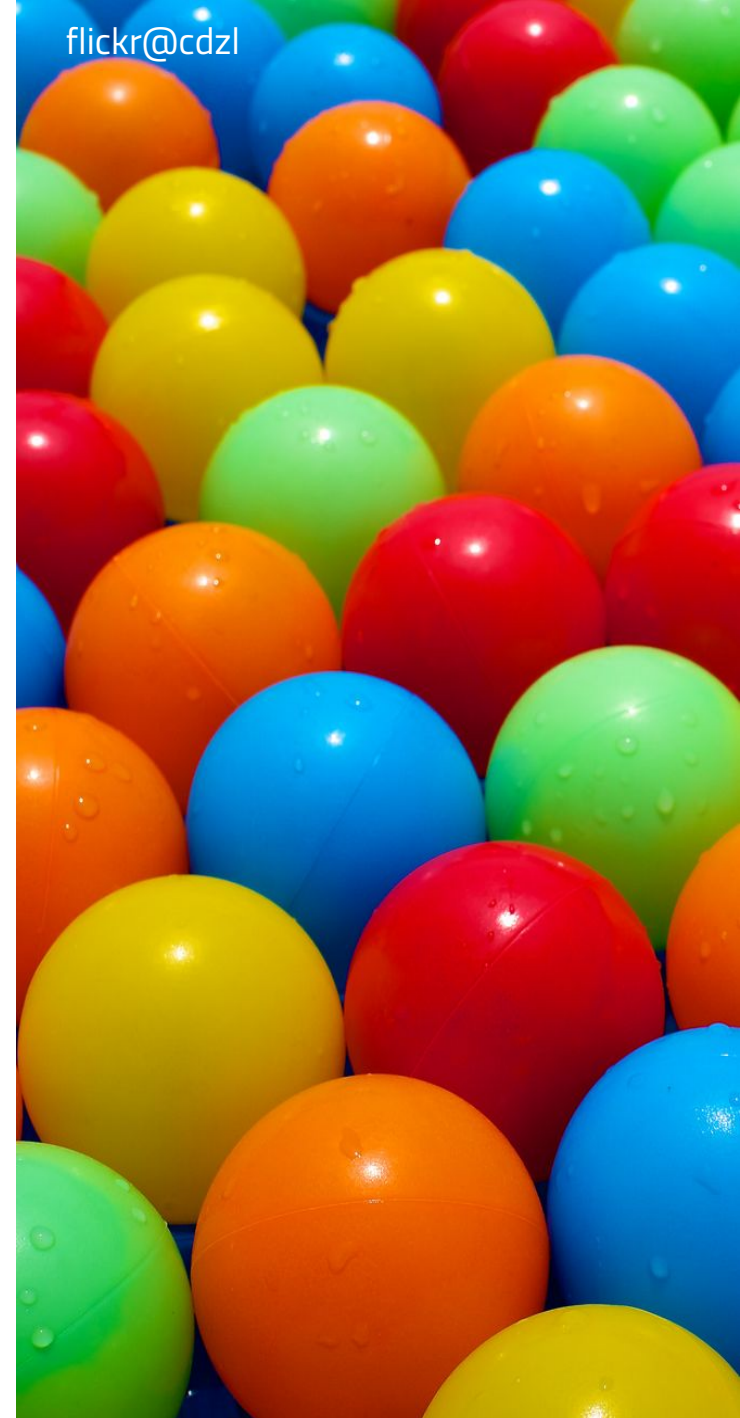
User evaluation

Participants use a **personalized search system** (online eval.) and compare the personalized and non-personalized SERP

Questionnaires and **browser histories** are used to create user profiles



advantages /
disadvantages?



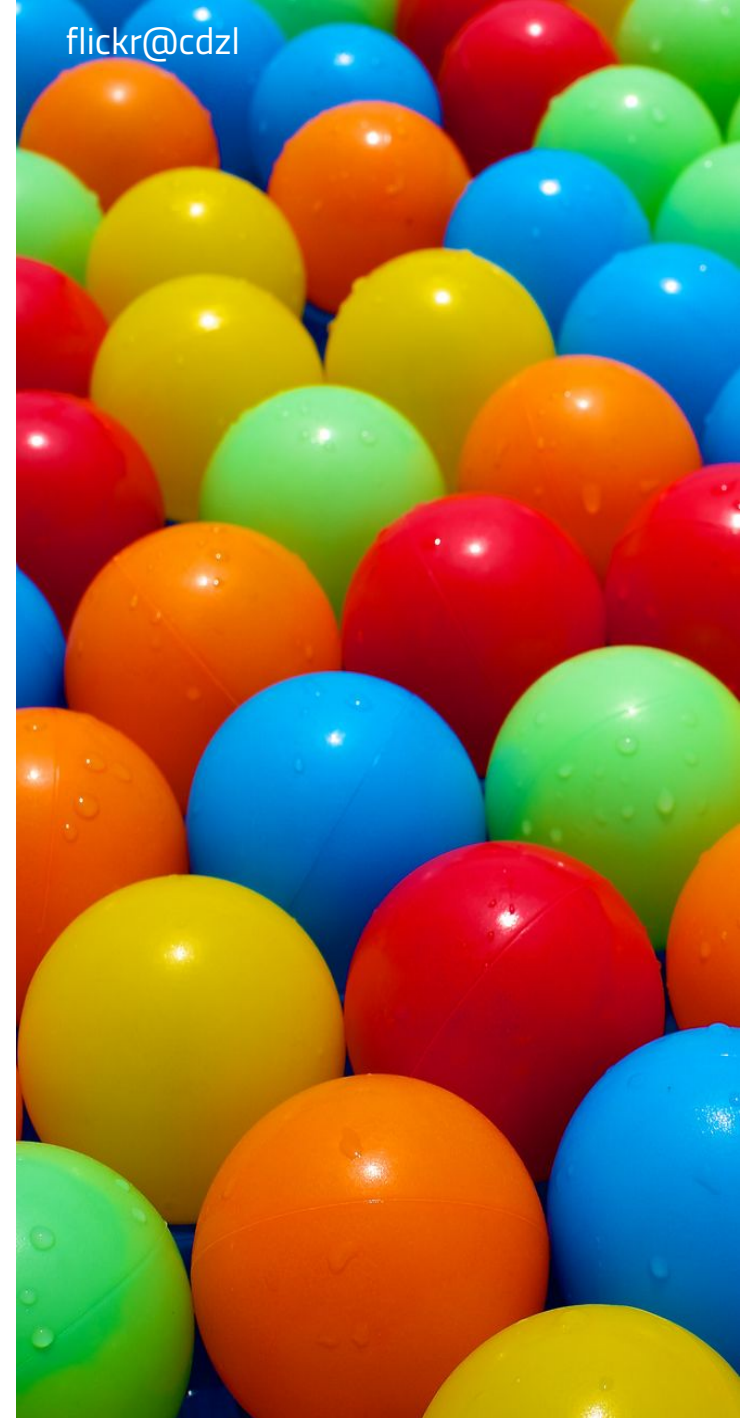
User evaluation

Participants use a **personalized search system** (online eval.) and compare the personalized and non-personalized SERP

Questionnaires and **browser histories** are used to create user profiles

What if a revised personalization approach needs to be tested?

Issues: small number of participants and **potential bias** through self-selected test queries



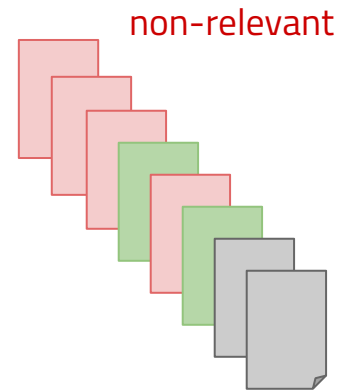
Click-log evaluation

Click-through data in query logs is used to **simulate** user experience in Web search

Assumption: users click on search results from top to bottom

Idea: if the user **clicks** on one or more documents and **skips over others**, those **clicked documents** are more relevant to her

Personalized retrieval is evaluated by **reranking** the result list; ideally those clicked results should appear at the top of the ranking → *click decisions become qrels*



advantages /
disadvantages?

Click-log evaluation metrics

Rank scoring

1, if click on j given q ;
0 otherwise

Rank of the page

$$R_q = \sum_j \frac{\delta(q, j)}{2^{(j-1)/(\alpha-1)}}$$

Expected utility
of a ranked list
for test query q

Rank scoring across all test queries

$$R = 100 \frac{\sum_q R_q}{\sum_q R_q^{Max}}$$

Max. possible utility
when all clicked pages
appear at the top

Click-log evaluation metrics

Rank scoring

1, if click on j given q ;
0 otherwise

Rank of the page

$$R_q = \sum_j \frac{\delta(q, j)}{2^{(j-1)/(\alpha-1)}}$$

Expected utility
of a ranked list
for test query q

Rank scoring across all
test queries

$$R = 100 \frac{\sum_q R_q}{\sum_q R_q^{Max}}$$

Average rank

Rank of page p

$$AvgRank_q = \frac{1}{|\mathcal{P}_q|} \sum_{p \in \mathcal{P}_q} R(p)$$

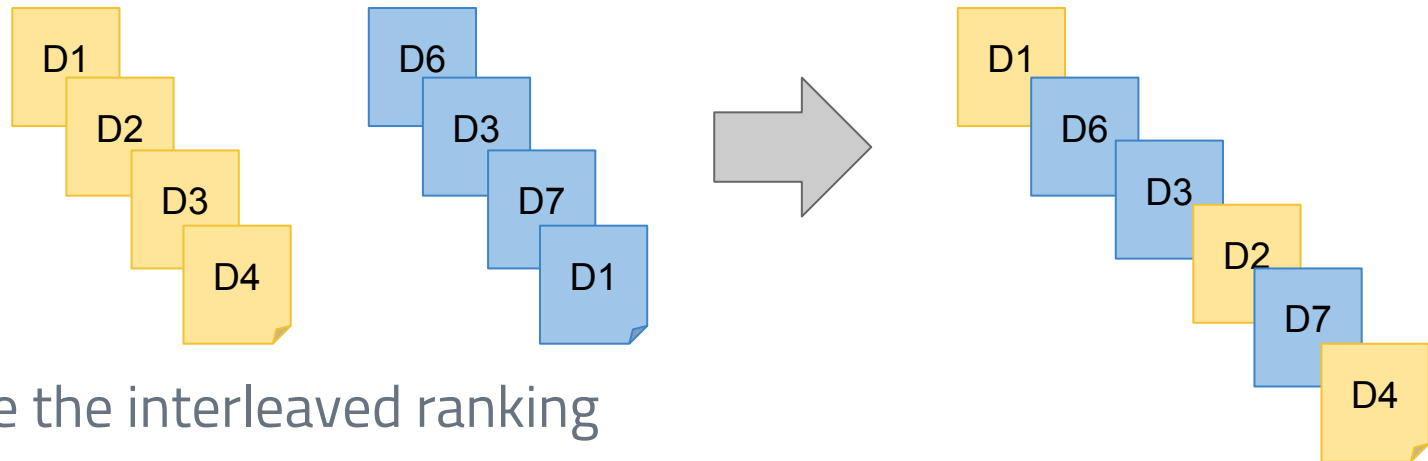
clicked pages for q

AvgRank across all test
queries

$$AvgRank = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} AvgRank_q$$

Interleaved evaluation

Combines the ranked lists of two (or more) rankings by alternating between the result lists and ignoring duplicates



Users receive the interleaved ranking

Ranking that received most clicks over many queries is considered of higher quality. More sensitive than other approaches.

Requires a substantial amount of user data

Personalized search: strategies

click-based

profile-based

Search tailored to a user's interests.

Person-level reranking strategy

Intuition: given a user's query, web pages that the user clicked on in the past are more relevant to her than those rarely clicked

$$S^{P-click}(q, p, u) = \frac{|\text{Clicks}(q, p, u)|}{|\text{Clicks}(q, \bullet, u)| + \beta}$$

Score of page p given query q by user u

user

#clicks in the past

#clicks in the past on any page given q and u

Smoothing factor

Person-level reranking strategy

Intuition: given a user's query, web pages that the user clicked on in the past are more relevant to her than those rarely clicked

$$S^{P-click}(q, p, u) = \frac{|\text{Clicks}(q, p, u)|}{|\text{Clicks}(q, \bullet, u)| + \beta}$$

Score of page p given query q by user u

user

#clicks in the past

#clicks in the past on any page given q and u

Smoothing factor

Issue: reranking fails if the user never issued the query before; luckily users are prone to repeat searches over time ("**re-finding**" of Web pages constitutes a large part of Web search traffic).

Re-finding

Definition: a user clicking a URL following a search, and then later clicking the same URL via another search

	Label	Query	Click
Monday	Q ₁	<i>swine flu incidence</i>	
	C ₁₁		healthmap.org/swineflu
	C ₁₂		www.swine-flu-map-animation.com
	C ₁₃		www.cdc.gov/H1N1Flu **
	Q ₂	<i>swine flu deaths</i>	
	Q ₃	<i>h1n1</i>	
	C ₃₁		en.wikipedia.org/wiki/H1N1
	C ₃₂		www.cdc.gov/H1N1Flu **
Tues.	Q ₄	<i>h1n1</i>	
	C ₄₁		www.cdc.gov/H1N1Flu **
	C ₄₂		h1n1.nejm.org
Wed.	Q ₅	<i>swine flu</i>	
	Q ₆	<i>cdc swine flu</i>	
	C ₆₁		www.cdc.gov/H1N1Flu **
Sat.	Q ₇	<i>cdc swine flu</i>	
	C ₇₁		www.cdc.gov/H1N1Flu **

One month of MSN log data ...

- **22%** of all of the queries sampled were instances of re-finding
- **30%** of all single-click queries were re-finding queries (5% for a multi-click query)
- **66%** of re-finding queries were also previous queries for a later re-finding
- **48%** of all re-finding instances occurred within a single session

Person-level reranking based on user interests

Intuition: given a user's query, web pages that are covering topics of interest to the user (based on her history) are more likely to be relevant

$$S^{L-profile}(q, p, u) = \frac{c_l(u) \times c(p)}{\|c_l(u)\| \|c(p)\|}$$

Reranking based on user profile and document profile vector similarity

User profile: weighted vector of topic categories

Document profile: weighted vector of topic categories

Unpopularity of p across users

How can we build $c(u)/c(p)$?



What issues does this profile have?

$$c_l(u) = \sum_{p \in \mathcal{P}(u)} P(p|u) w(p) c(p)$$

Pages visited in the past

Click prob. of p

Short- and long-term profiles

Fact: short-term user profiles tend to be more useful to improve search in the current session

$$S^{S-profile}(q, p, u) = \frac{c_q(u) \times c(p)}{|| c_q(u) || || c(p) ||}$$

Scoring **only**
dependent on the
clicks of the current
search session

Short- and long-term profiles can be combined:

$$S^{LS-profile}(q, p, u) = \theta \times S^{S-profile}(q, p, u) + (1 - \theta) S^{L-profile}(q, p, u)$$

Group-level re-ranking

Intuition: a single user may only have a relatively sparse user profile, we can benefit from combining her profile with that of similar users

Set of nearest user profile neighbours

similarity between two user profiles

$$S^{G-click}(q, p, u) = \frac{\sum_{u_s \in S_u(u)} Sim(u_s, u) \times |Clicks(q, p, u_s)|}{\beta + \sum_{u_s \in S_u(u)} |Clicks(q, \bullet, u_s)|}$$

Intuition: the more similar a user profile x to the current user, the more important the clicks of x are to the current user

Dataset

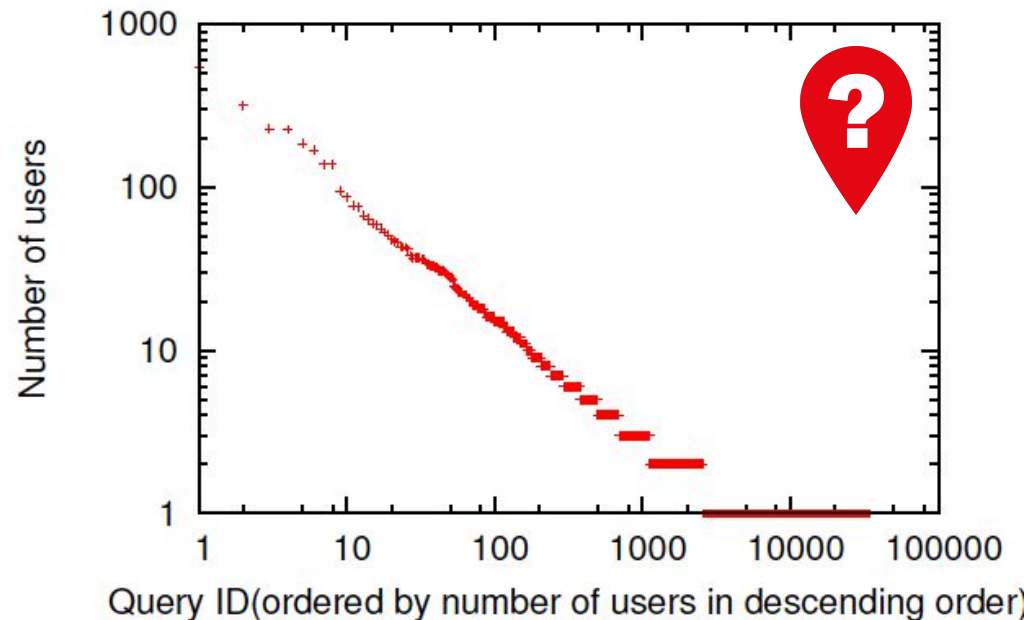
12 days of MSN log data
(2006) with 10K users
randomly sampled from it

11 days to **train**, **1** day to **test**

The 3% most popular distinct
queries are issued by 47% of
users

46% of all queries in the test
set also appear in training

Item	ALL	Training	Test
#days	12	11	1
#users	10,000	10,000	1,792
#queries	55,937	51,334	4,639
#distinct queries	34,203	31,777	3,465
#Clicks	93,566	85,642	7,924
#Clicks/#queries	1.6727	1.6683	1.7081
#sessions	49,839	45,981	3,865



Query click variation

Click entropy captures how uniform or divergent the clicks following a specific query are:

$$CE(q) = \sum_{p \in \mathcal{P}(q)} -P(p|q) \log_2 P(q|p)$$

Pages clicked for
query q

%of clicks on p
for all clicks by
users issuing q



When is the
entropy zero?
When is it largest?



Query click variation

Click entropy captures how uniform or divergent the clicks following a specific query are:

$$CE(q) = \sum_{p \in \mathcal{P}(q)} -P(p|q) \log_2 P(q|p)$$

Pages clicked for query q

%of clicks on p for all clicks by users issuing q



When is the entropy zero?
When is it largest?

		Click Entropy		
		Low	Mid	High
Clicks/User	Low	www.schoolloop.com usps.gov men's health magazine espn2	fox news network ontario airport wvu larry king	ecw fcc arrow internet explorer update
	Mid	corvette america cleartype petfinder.org pfchang	michigan state football alaska cruise trivia quiz knee injury	toyota camry rachel ray recipes bruce springsteen lyrics stress hormones
	High	(no queries)	restaurant guide famous poems calculate bmi woodrow wilson	first aid hand foot mouth disease cupcake recipes house spiders

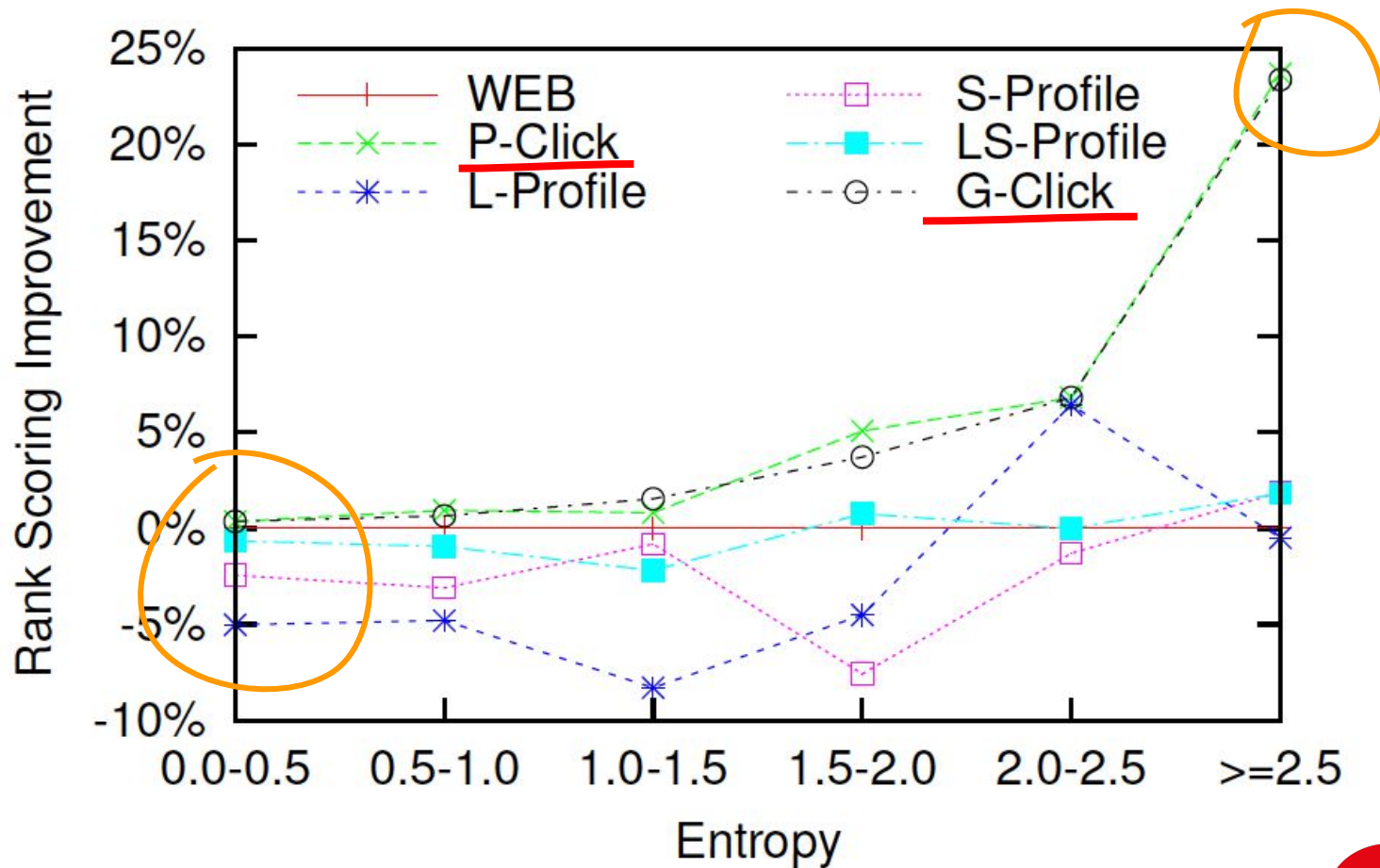
Personalization strategies evaluated

Queries for which reranking may help

method	all		not-optimal	
	R.S.	A.R.	R.S.	A.R.
WEB	69.4669	3.9240	47.2623	7.7879
P-Click	70.4350	3.7338	49.0051	7.3380
L-Profile	66.7378	4.5466	45.8485	8.3861
S-Profile	66.7822	4.4244	45.1679	8.3222
LS-Profile	68.5958	4.1322	46.6518	8.0445
G-Click	70.4168	3.7361	48.9728	7.3433

- **Click-based personalization** outperforms the WEB baseline
- **Profile-based personalization** degrades on average (large performance deviation across queries)

Evaluation across click entropies

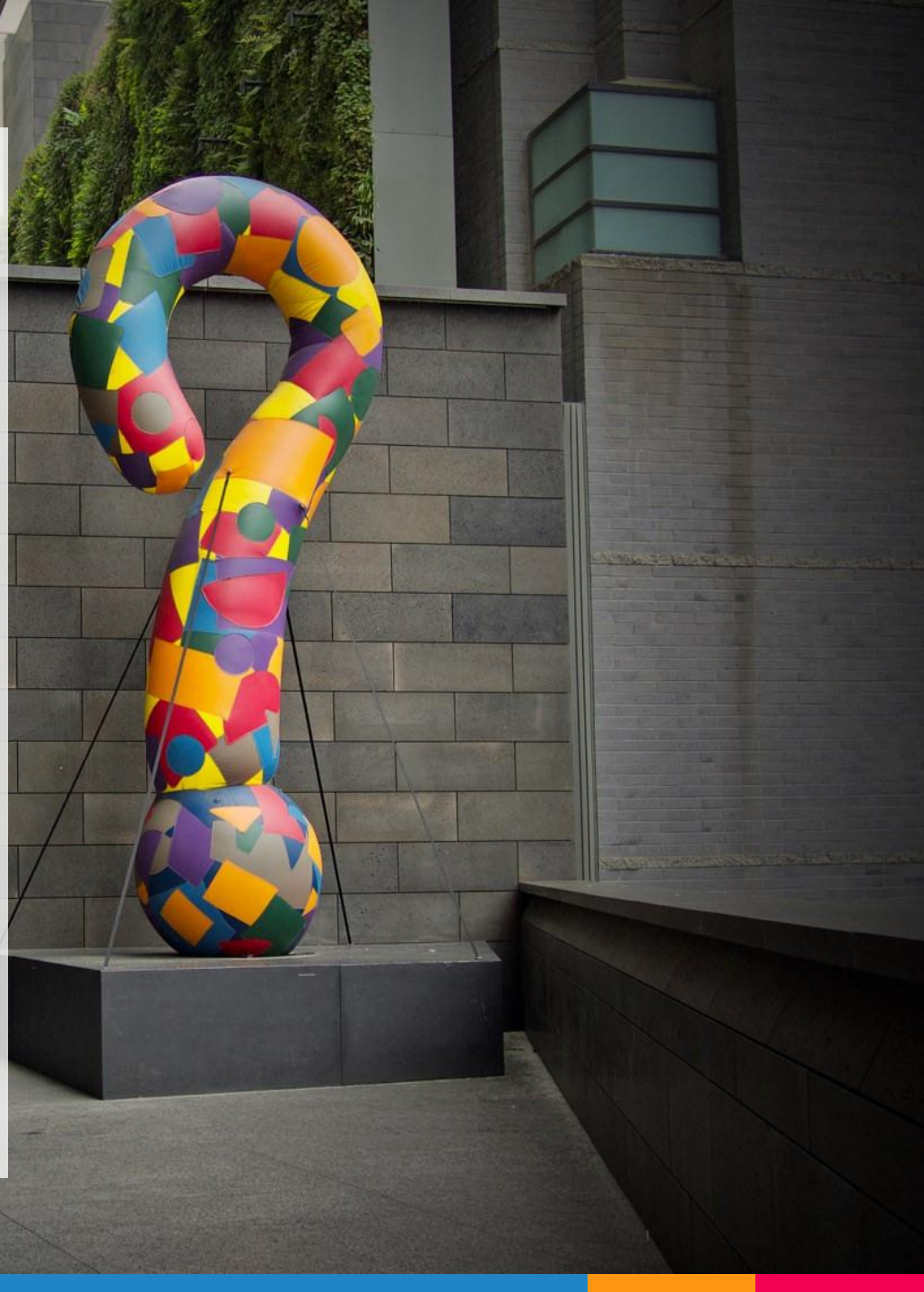


Personalization hurts queries with low entropy.



What now?

- Why was no significance test performed?
- What do you expect the reranking effectiveness to be on repeated queries? (and what does this mean?)
- What do you expect from users with more/less search activity in the training days?
- What kind of features could a click entropy classifier have?



	Click entropy	
	All	Low RE
Query length (words)	0.20	0.16
Query length (chars)	-0.04	0.03
URL fragment	-0.36	-0.23
Location mentioned	-0.03	-0.04
Advanced query	-0.01	-0.02
# of query suggestions	0.12	0.15
# of times issued	0.00	-0.01
# of distinct users	-0.01	0.00
Avg. # of results	0.03	-0.02
% issued during work	-0.10	-0.04
Query clarity	0.02	-0.02
Category entropy	-0.01	0.01
# of distinct categories	0.01	-0.02
# of URLs in ODP	0.09	0.05
Top level domain entropy	0.02	0.04
# of distinct hosts	0.19	0.17
Click entropy	1.00	1.00
Potential at 10	0.87	0.86
Result entropy (RE)	0.53	-0.04
Avg. clicks per user	0.73	0.69
Avg. click position	0.90	0.86
Avg. seconds to click	0.03	0.05



Personalized search: privacy & software architectures

Tension

Personalized search requires the collection of information about users (the more the better)

Privacy preservation requires us to reveal as little as possible to the search provider

- Collected information reveals a lot about a user's private life
- **How can we preserve privacy in personalized search?**



Formally

User identifying
information (IP
address, user ID)

Text description of
information need N (related
queries, viewed results ...)

$$P(U) = \{ID(U, i), TEXT(N, i)\}$$

personal
information of
user U
(needed for
personalized
search)

where $i = 1, \dots, k$

k search activities

Who should $P(U)$ be revealed to?

Only a "trusted" party (e.g. a search engine company with clear privacy protection rules) or some "untrusted" parties (e.g. a third party with access to the Web search log) as well?

Level 1: pseudo identity

AOL debacle ...

$ID(U)$ is replaced
by pseudo identity
 $IDP(U)$

$IDP(U)$ contains
less personally
identifiable
information than
 $ID(U)$

Content of user
profile information
remains intact

Level 1: pseudo identity

AOL debacle ...

$ID(U)$ is replaced
by pseudo identity
 $IDP(U)$

$IDP(U)$ contains
less personally
identifiable
information than
 $ID(U)$

Content of user
profile information
remains intact

Level 2: group identity

A group of users
share a single
identity $ID(U)$

The description of
user information
needs $TEXT(N,i)$ is
aggregated at the
group level

To enable effective
personalized search,
group members
should share
interests

Implementable
through a proxy or
implicitly through
TrackMeNot

Level 1: pseudo identity

AOL debacle ...

$ID(U)$ is replaced
by pseudo identity
 $IDP(U)$

$IDP(U)$ contains
less personally
identifiable
information than
 $ID(U)$

Content of user
profile information
remains intact

Level 2: group identity

A group of users
share a single
identity $ID(U)$

The description of
user information
needs $TEXT(N,i)$ is
aggregated at the
group level

To enable effective
personalized search,
group members
should share
interests

Implementable
through a proxy or
implicitly through
TrackMeNot

Level 3: No identity

The user identity
 $ID(U)$ is not
available to the
search engine

The information
need descriptions
 $TEXT(N,i)$ cannot be
aggregated on the
search engine side

To enable
personalized search,
client-side
personalization is
necessary.

Level 1: pseudo identity

AOL debacle ...

$ID(U)$ is replaced
by pseudo identity
 $IDP(U)$

$IDP(U)$ contains
less personally
identifiable
information than
 $ID(U)$

Content of user
profile information
remains intact

Level 2: group identity

A group of users
share a single
identity $ID(U)$

The description of
user information
needs $TEXT(N,i)$ is
aggregated at the
group level

To enable effective
personalized search,
group members
should share
interests

Implementable
through a proxy or
implicitly through
TrackMeNot

Level 3: No identity

The user identity
 $ID(U)$ is not
available to the
search engine

The information
need descriptions
 $TEXT(N,i)$ cannot be
aggregated on the
search engine side

To enable
personalized search,
client-side
personalization is
necessary.

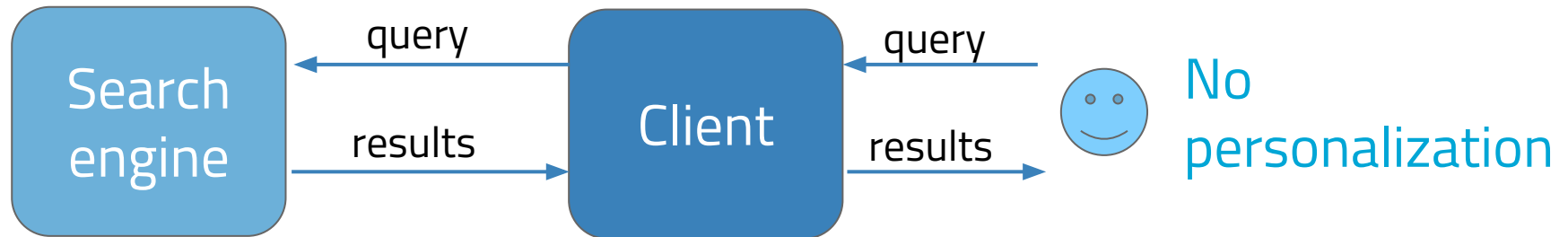
Level 4: No personal information

Neither the user
identity $ID(U)$ nor
the description of
information needs
 $TEXT(N)$ are
available to the
search engine

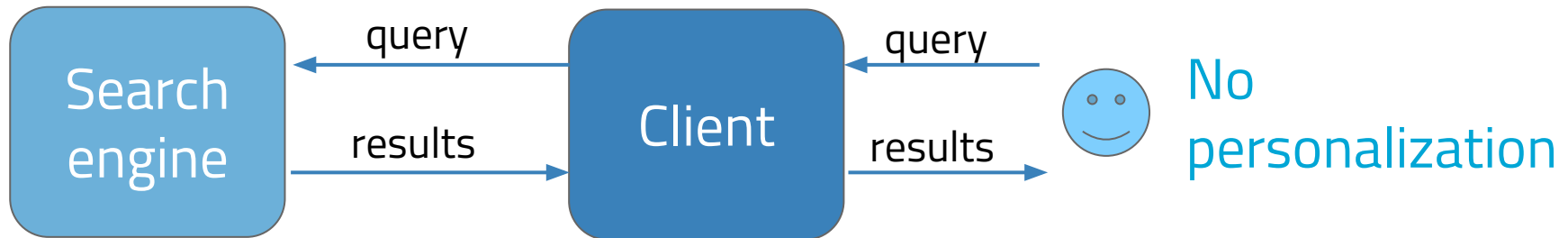
Ultimate privacy

Hard to guarantee
in practice
(cryptography,
laws, ...)

Software architectures

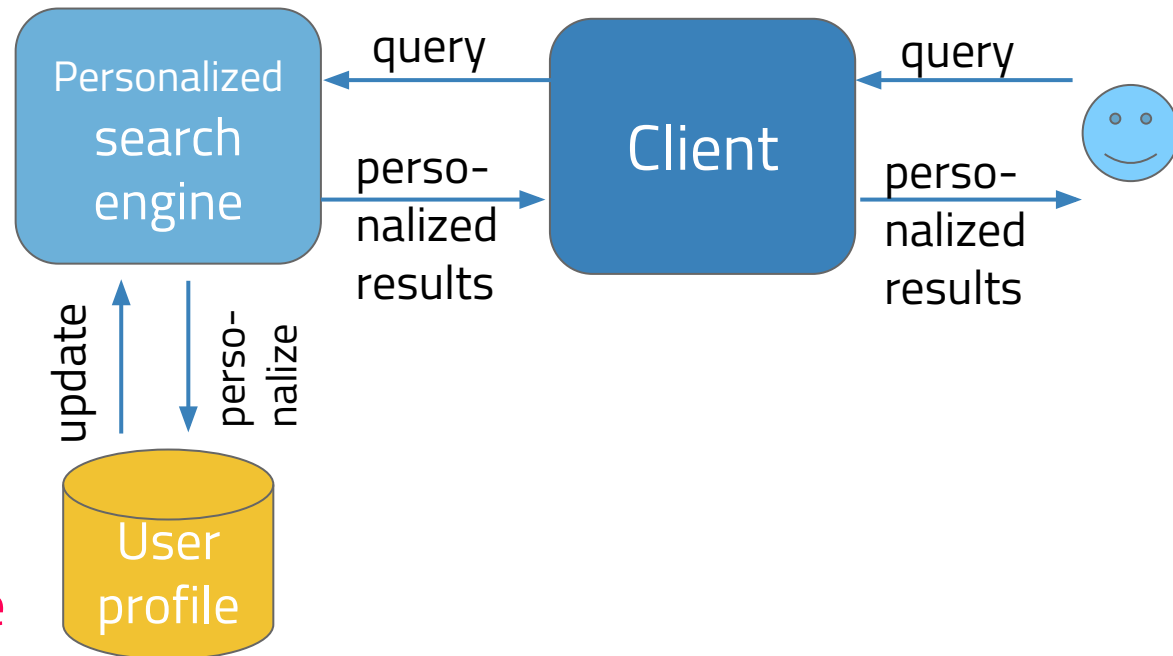


Software architectures



Server-side personalization

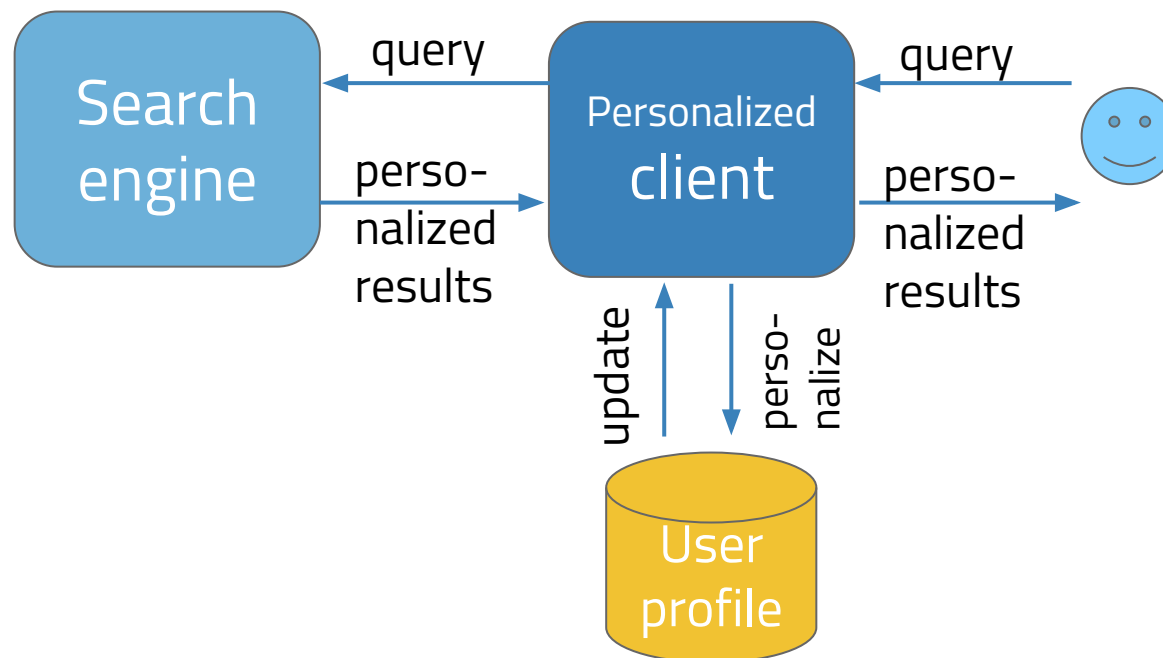
- $P(U)$ on the server
- Full power of personalized algs., client unchanged
- Level 1 privacy, with proxy level 2;
levels 3/4 impossible



Software architectures

Client-side personalization

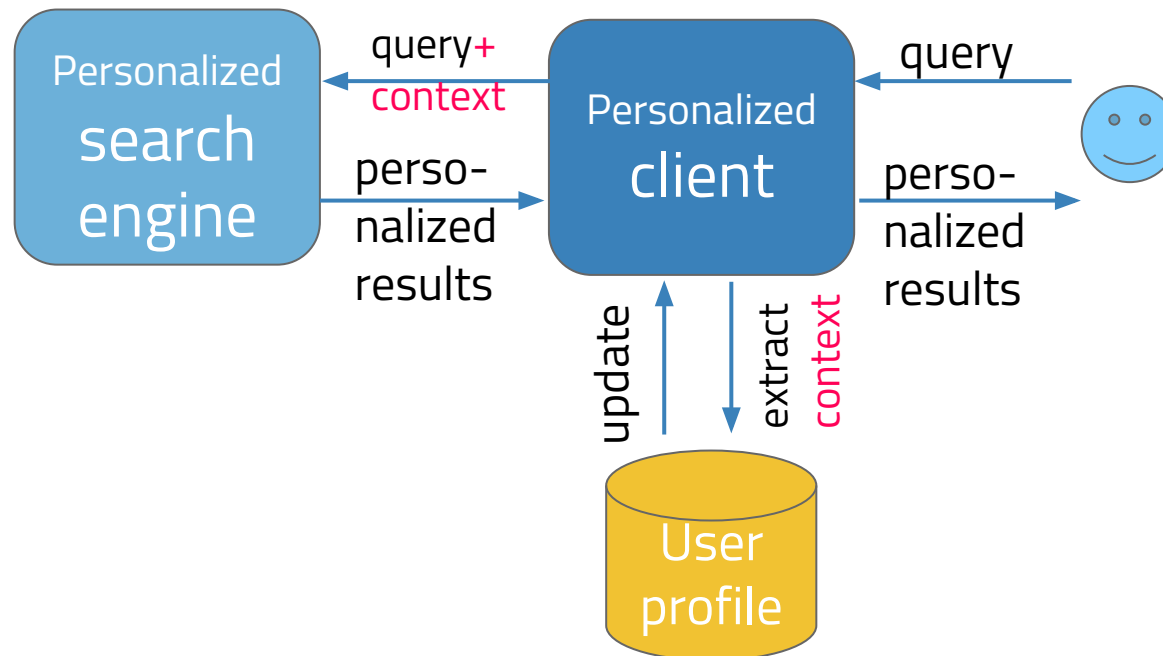
- Reranking based on $P(U)$ or automatic query expansion
- Allows more than search activities into the profile
- Some knowledge is only available on the server
- Usage of an anonymous network enables **privacy level 3**



Software architectures

Client-server collaborative personalization

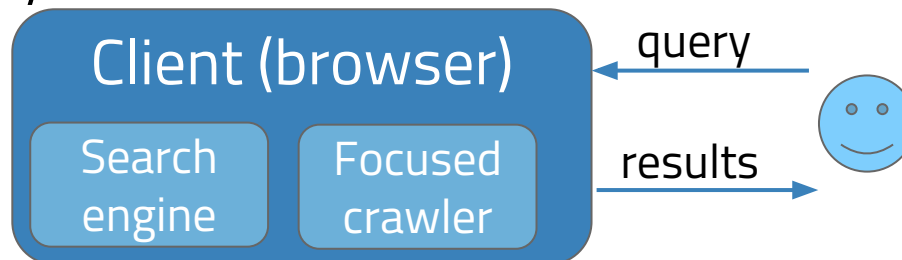
- Compromise: enables the use of the search engine's resources to personalize while not storing $P(U)$
- Context example: query expansion terms, topic weights, ...



Software architectures

Private search

- Search engine self-contained in the browser; **level 4 privacy**
- No third party logging queries, clicks, page visits
- Useful for sensitive searches (medical conditions, etc.)
- Three stages:
 1. Focused crawler activated (can take a day)
 2. Index creation (JScene prototype: ~10h for 1M tweets)
 3. User interacts with the in-browser search engine
- Degree of privacy determined by breadth of crawl; time vs. privacy tradeoff



Towards personalised PageRank

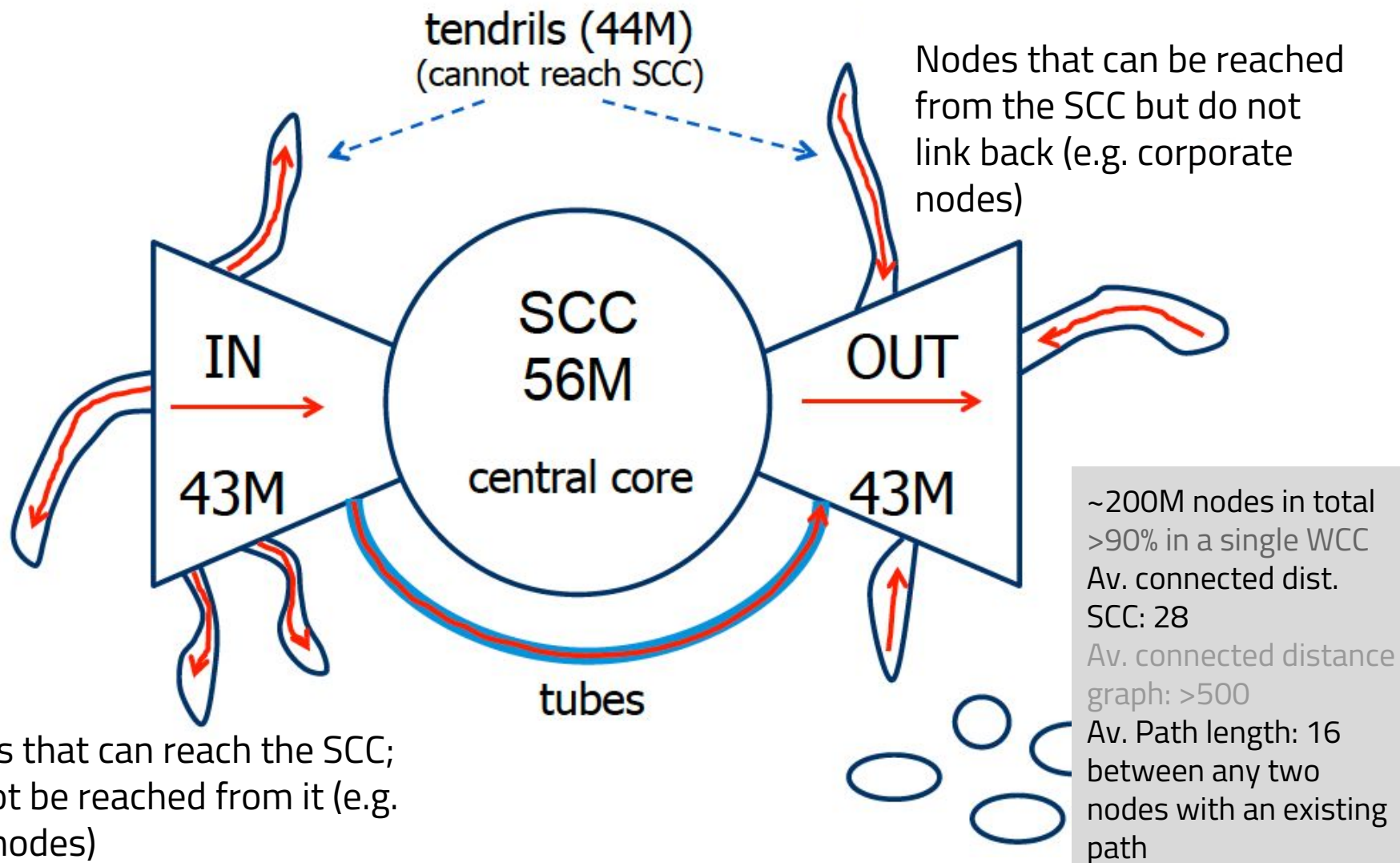
Detour: the web

Vannevar Bush (1947)



"When the user is building a **trail**, he names it, inserts the name in his code book, and taps it out on his keyboard. Before him are the **two items to be joined**, projected onto adjacent viewing positions. At the bottom of each there are a number of blank code spaces, and a pointer is set to indicate one of these on each item. The user taps a single key, and the items are **permanently joined**."

Detour: the structure of the web





“In a sense the web is much like a **complicated organism**, in which the local structure at a **microscopic scale** looks very regular like a biological cell, but the **global structure** exhibits interesting morphological structure (body and limbs) that are not obviously evident in the local structure.”

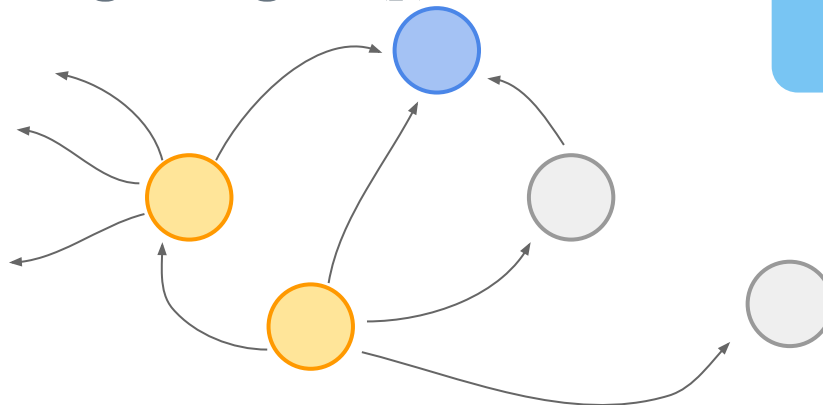
HITS: Hyperlink-Induced Topic Search

Intuition: two broad types of useful web pages for ad-hoc search queries

- **Authoritative pages:** pages containing a lot of relevant content (e.g. Wikipedia page); high weight $a(p)$
- **Hub pages:** pages containing a large number of useful hyperlinks pointing to pages with relevant content; high weight $h(p)$

A page pointed to by many **hubs**.

A page pointed to by many **authorities**.





Do you see an issue?

HITS: Hyperlink-Induced Topic Search

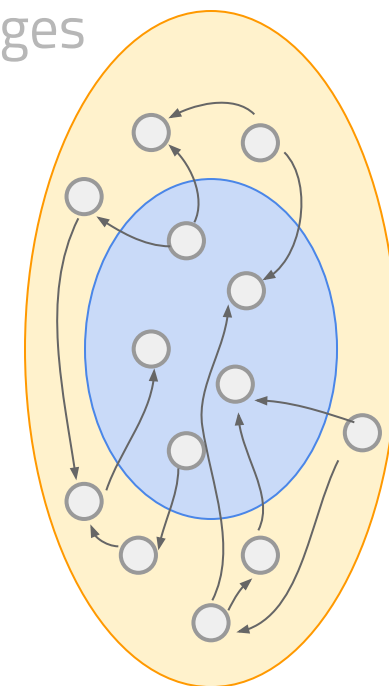
1. **Root set (RS)**: retrieve the top N results for a given keyword query
2. **Base set (BS)**: expand RS by including all pages that link to pages in RS or are linked-to by pages in RS
3. Clean hyperlink structure (link removal between pages belonging to the same web site)
4. **Initialize** all hub/authority weights to 1.0
5. **Iteratively update** hub/authority weights and normalize

$$a(p) = \sum_{q \rightarrow p} h(q)$$

Authority weight increases if good hubs point to p.

$$h(p) = \sum_{p \rightarrow q} a(q)$$

Hub weight increases if p points to good authorities.



>10,000 citations

PageRank

A **topic independent** approach to page importance, computed **once** per crawl



Which retrieval models can we easily adapt to include PageRank?

Every document of the corpus is assigned an **importance score**

Today, just one of hundreds/thousands of features in a modern web search engine.

PageRank takes the importance of the page where the link originates from into account (intuition: one link from `google.com` is better than 100 links from unpopular blogs)

“To test the utility of PageRank for search, we built a web search engine called Google.”

Paper rejected from SIGIR 1998, accepted at WWW 1998.

PageRank

Each page distributes importance through its out-links

Simple PageRank, iteratively defined:

Problem: pages that are sinks.
PageRank mass vanishes.

$$PageRank_{i+1}(v) = \sum_{u \rightarrow v} \frac{PageRank_i(u)}{N_u}$$

eventual
convergence

all nodes linking to v

A page with many
out-links has little
influence on one
particular page.

out-degree of node u

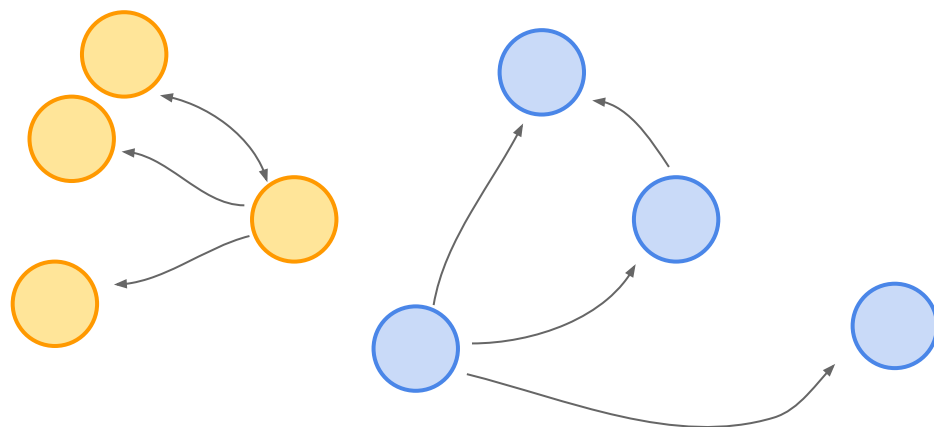
Random surfer model: PageRank score of v is the probability that a random surfer starts at any page and ends up at v (random surfer follows each link at a page with equal probability)

PageRank

Each page distributes importance through its out-links

PageRank, iteratively defined with a decay/damping factor:

$$PageRank_{i+1}(v) = p \sum_{u \rightarrow v} \frac{PageRank_i(u)}{N_u} + (1 - p)$$



Probability that the random surfer "teleports" instead of following an outlink.

PageRank applications

Search: re-rank the top retrieved documents of a content retrieval technique according to the pages' PageRank score

Search: filter out pages with low PageRank scores

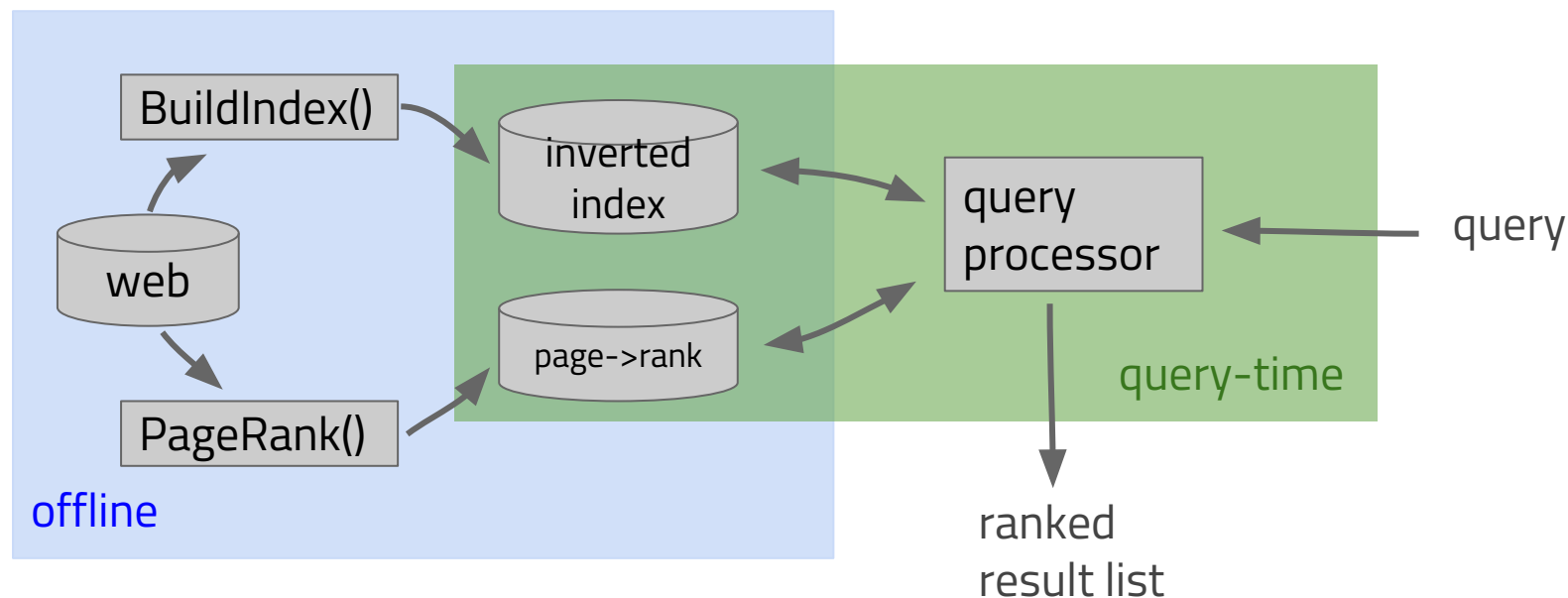
Personalized PageRank: instead of random teleporting, bias the teleport locations

PageRank as future **inlink count predictor**: re-order crawling list accordingly (crawl better pages first)

Personalized PageRank

Desired: **query-time information** should be able to influence the PageRank score while still requiring **minimal computation**

Idea: compute offline a **set of PageRank scores** per document, each biased towards a topic (here: 16 ODP topics)



Detour

DMOZ

From Wikipedia, the free encyclopedia

DMOZ (from *directory.mozilla.org*, an earlier [domain name](#)) was a multilingual [open-content directory](#) of [World Wide Web](#) links. The site and community who maintained it were also known as the **Open Directory Project (ODP)**. It was owned by [AOL](#) (now a part of [Verizon's Oath Inc.](#)) but constructed and maintained by a [community](#) of volunteer editors.

DMOZ used a hierarchical [ontology](#) scheme for organizing site listings. Listings on a similar topic were grouped into categories which then included smaller categories.

DMOZ closed on March 17, 2017 because AOL no longer wished to support the project.^{[2][3]} The website became a single landing page on that day, with links to a static archive of DMOZ, and to the DMOZ discussion forum, where plans to rebrand and relaunch the directory are being discussed.^[3]

As of September 2017, a non-editable mirror remained available at [dmoztools.net](#),^[4] and it was announced that while the DMOZ URL would not return, a successor version of the directory named **Curlie** would be provided.^{[5][6]}

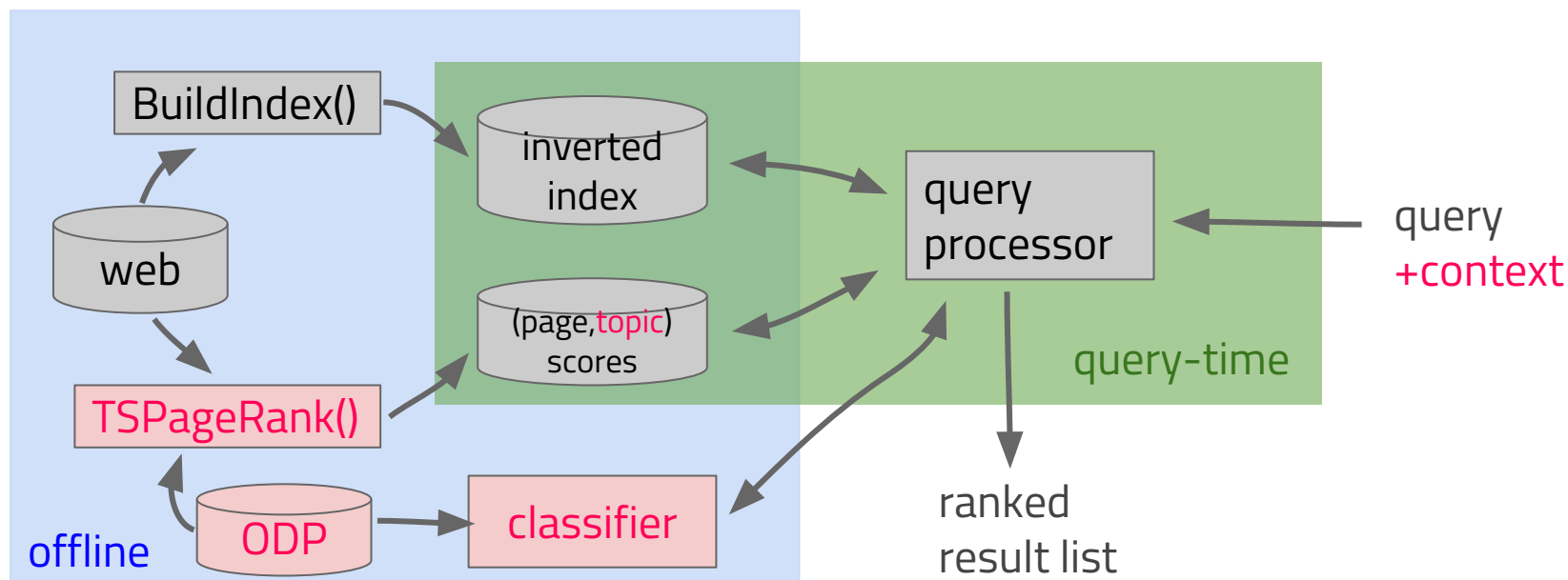
[DMOZ/ODP Wikipedia entry](#)



Personalized PageRank

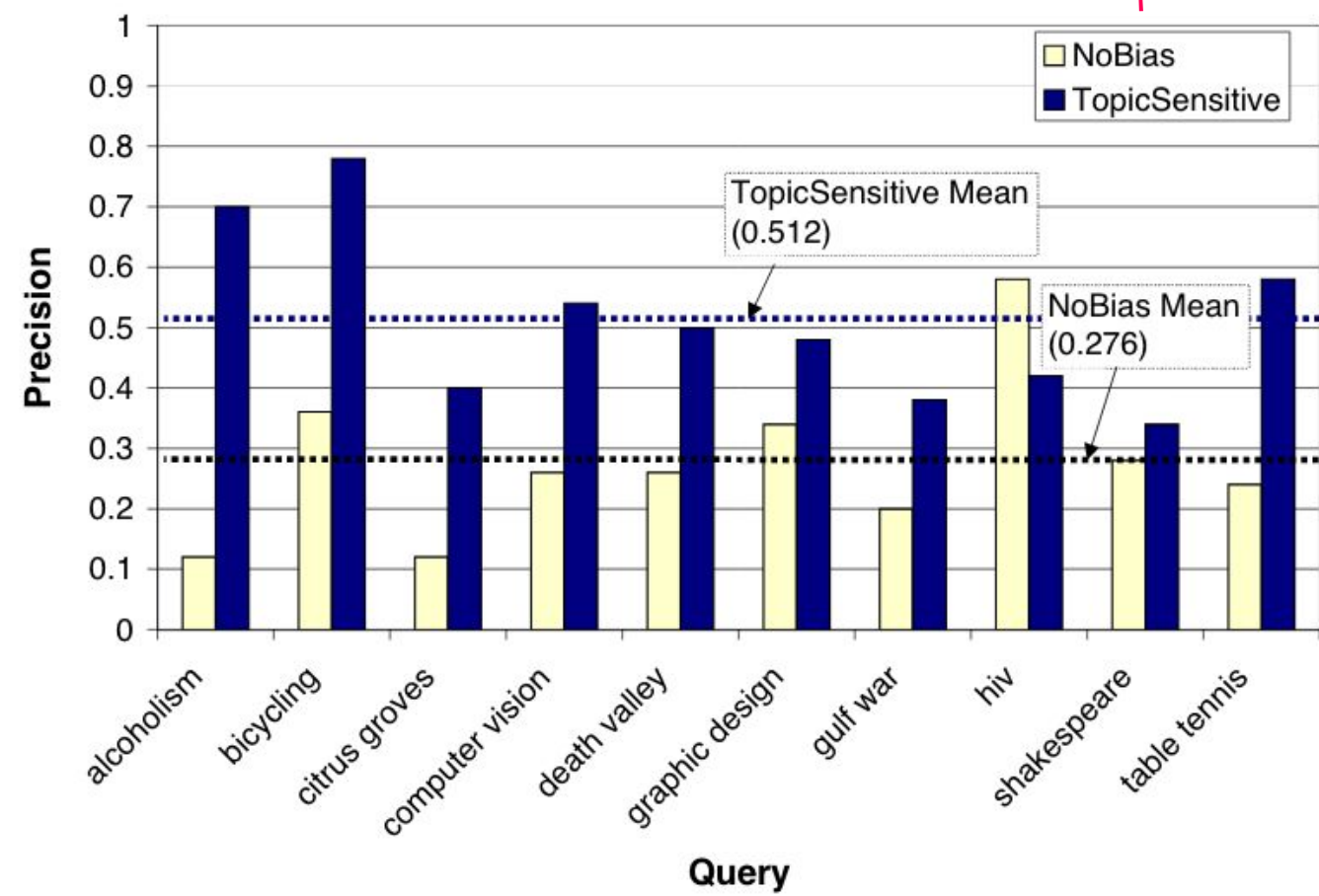
Desired: **query-time information** should be able to influence the PageRank score while still requiring **minimal computation**

Idea: compute offline a **set of PageRank scores** per document, each biased towards a topic (here: 16 ODP topics)



Personalized PageRank

10 test queries



That's it for today!

**Next week Friday: submit
your intermediate report!**

Slack: in43252019.slack.com

Email: in4325-ewi@tudelft.nl