INFORMATION RETRIEVAL

L11. QUERIES AND SESSIONS

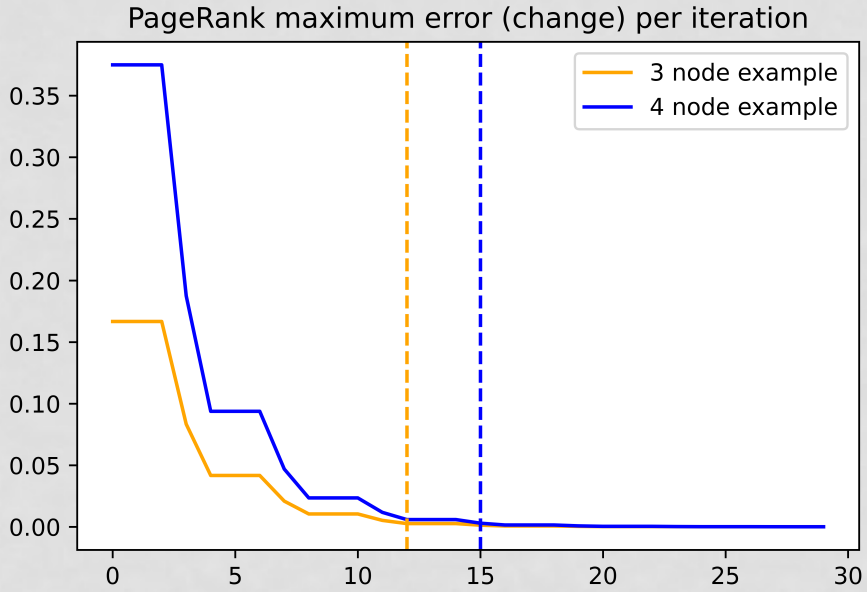
SUZAN VERBERNE 2022

Universiteit Leiden

LAST WEEK

- PageRank convergence (thanks to Job Mooij for the implementation and visualisation!)

```
import numpy as np
from matplotlib import pyplot as plt

graph_3 = {'A': ['B', 'C'], 'B': ['C'], 'C': ['A']}
graph_4 = {'A': ['B'], 'B': ['C'], 'C': ['B', 'D'], 'D': ['B']}


def pagerank(graph, iterations = 30):
    scores = {i: 1 / len(graph) for i in graph}
    scores_static = scores
    error = []
    for i in range(iterations):
        scores_new = {i: 0 for i in graph}
        for node in graph:
            outgoing_nodes = graph[node]
            # print('Node %s with %d outgoing\t%s' % (node, len(outgoing_nodes), scores_new))
            for outgoing in outgoing_nodes:
                scores_new[outgoing] += scores[node] / len(outgoing_nodes)
        old = np.array([scores[i] for i in scores])
        new = np.array([scores_new[i] for i in scores_new])
        scores = scores_new

        error.append(np.max(abs(old - new)))
    return error

plt.plot(pagerank(graph_3), label='4 nodes')
plt.plot(pagerank(graph_4), label='3 nodes')
plt.legend()

# plt.show()
plt.savefig('pagerank.png')
```



TODAY'S LECTURE

- Query log analysis
- Sessions
- Characteristics of queries
- Learn from user behaviour
- Simulation of user interaction
- Networks of queries
- Relevance feedback

1. IIR chapter 9: Relevance feedback and query expansion (not reading material)
2. "Mining Query Logs: Turning Search Usage Data into Knowledge" by Fabrizio Silvestri. Sections 2, 3, 4.1, 4.2 and 4.3 (p. 16-64)

QUERY LOG ANALYSIS

EXERCISE

- Suppose that you developed a search engine for the archives of a large governmental organization. Through a standard query interface the employees can search for information contained in the archive. The results are presented in a list, ranked by relevance.
- The client provides you with the logs of the search engine, with this information:
  - IP_address, timestamp, query, clicked_document, rank_of_clicked_document
- The client wants to get some insight in the use of the search engine.
- What information do you extract and how?
- Discuss with your neighbour

EXERCISE

- General statistics:
  - Number of queries, number of unique queries
  - Number of users, number of unique users
  - Most frequent queries, most frequent terms, most frequently clicked documents
  - Distribution of query length
- Session extraction
  - Number of sessions
  - Number of queries per session per user
- Evaluative aspects:
  - Proportion of queries with at least one click
  - Mean Average Precison based on position of clicked documents
  - Estimate of proportion of successful queries/sessions
  - Stopping behaviour: when do users leave the search

EVALUATION IN IR

Evaluation approaches:

- Cranfield paradigm — use of static test collections
- User observation
- Query logs
- Simulation

- (… or a combination of those)

LIMITATIONS OF THE CRANFIELD PARADIGM

- Static relevance judgments (independent of user need)
- Relevance assessments not created by searcher
- Per-query evaluation

- Context of a query = session
- Session: all interactions between user and search engine related to one information need

QUERY LOGS

Key aspects of query logs:

- Queries in context of other queries by the same user
- Documents clicked for each query (+ dwell time)
- If unique user ids are kept (of course no IP addresses) then user profiling for personalization is possible. But:
  - Keep the privacy of the users in mind
  - AOL data leak: No IP addresses but personally identifiable information

- How to prevent privacy leaks?

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

KEY ASPECTS OF QUERY LOGS

- How to prevent privacy leaks?
  - further anonymization (e.g replace terms by integers)
  - remove the long tail of relatively unique data
  - start a new user ID with every session

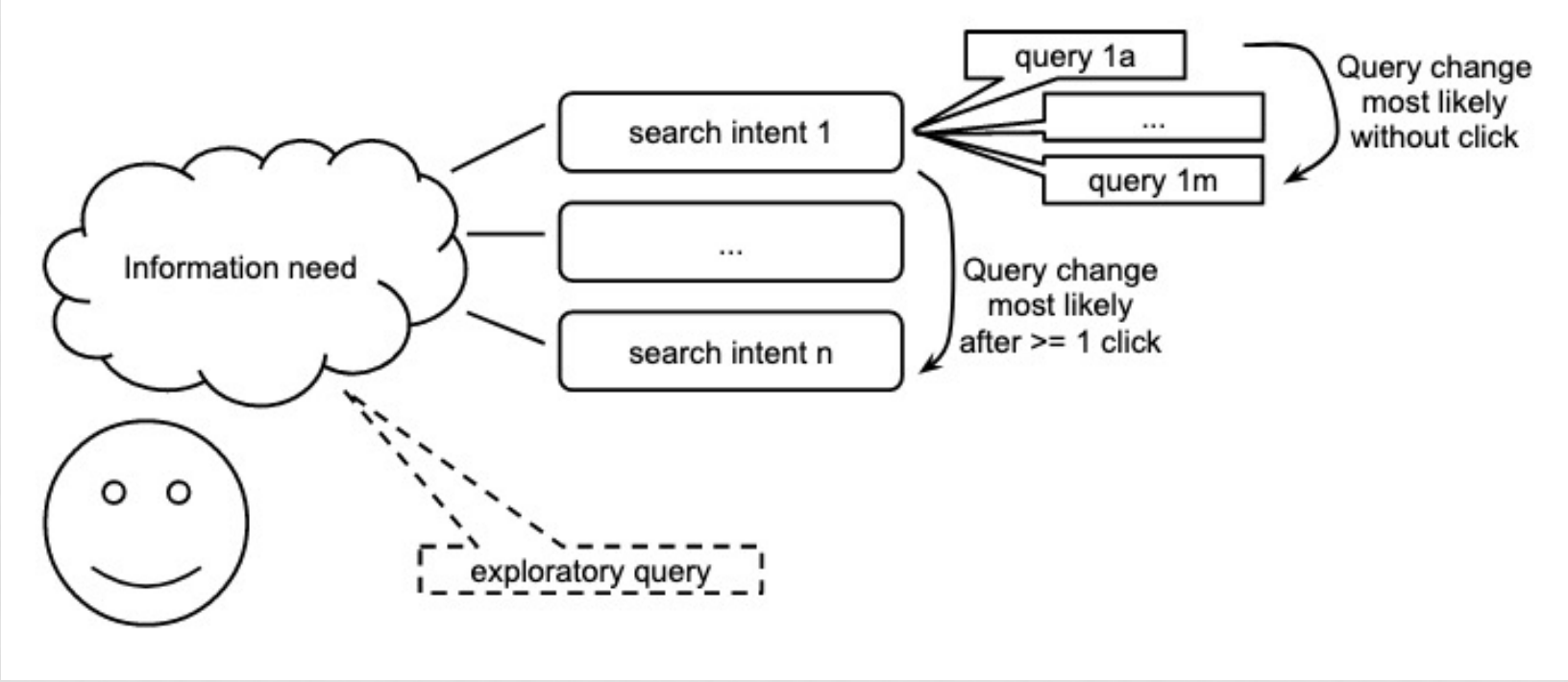
SESSIONS

SILVESTRI SECTION 3.1

SESSIONS

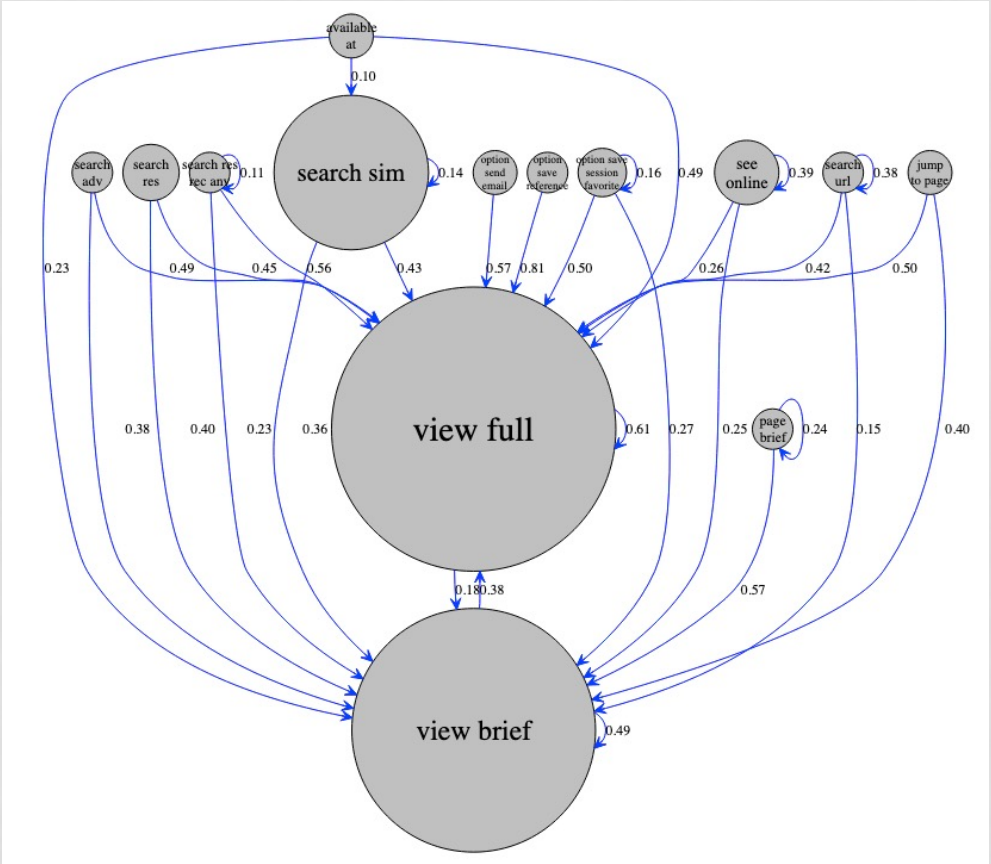
- Session: all interactions between user and search engine related to one information need
- How to define a session if we don't know the information need?
  - Get all interactions for one user
  - Split into sessions based on time between queries (e.g. 30 minutes)

SESSION BEHAVIOUR



SESSIONS

- Session information:
  - Series of queries
  - Per query:
    - Results examined
    - Results clicked
    - Query re-formulation
  - Session ends when:
    - user is satisfied or
    - gives up



Verberne et al. (2010). How does the Library Searcher Behave? A Contrastive Study of Library Search against Ad-hoc Search.

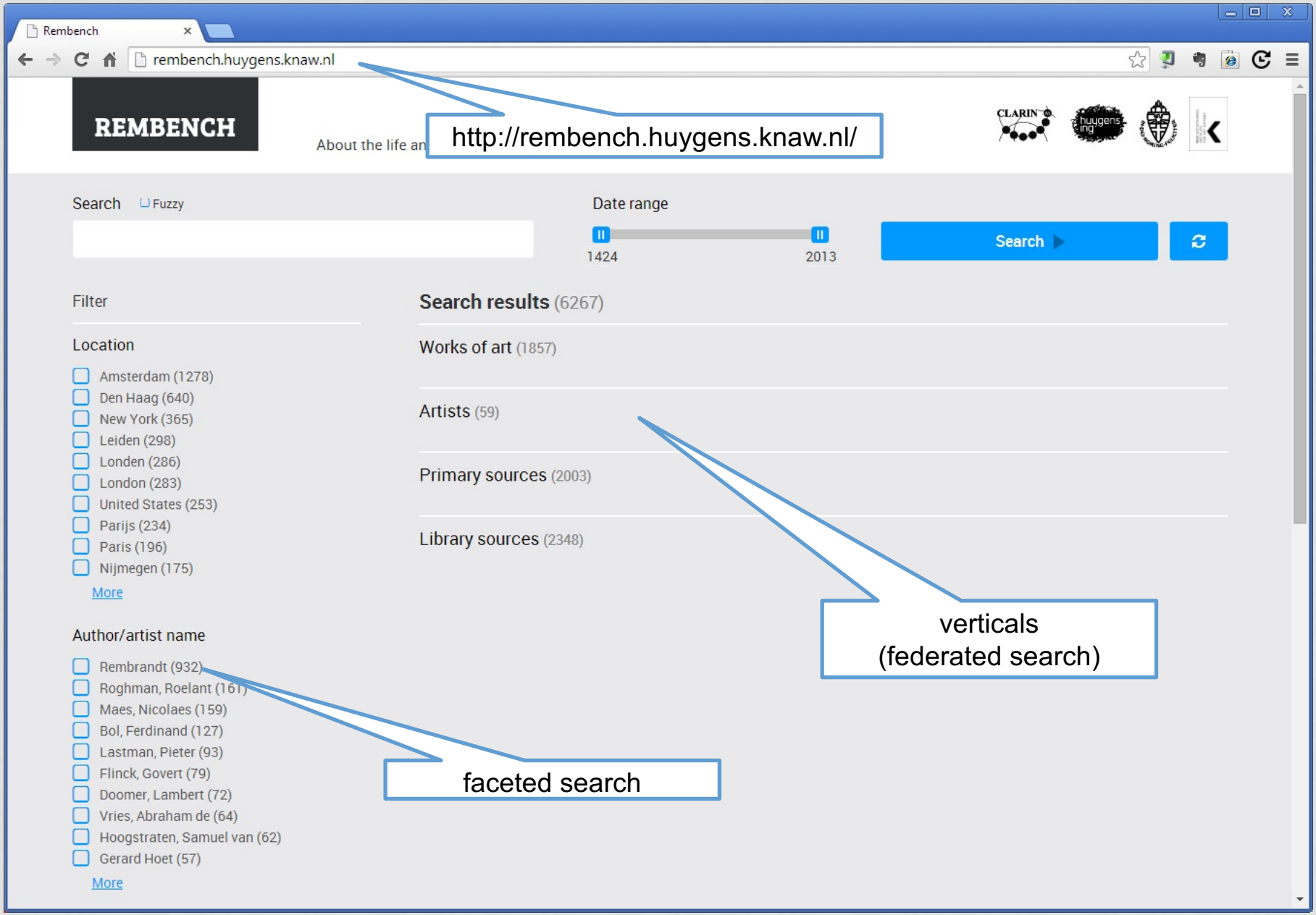
Universiteit Leiden

CHARACTERISTICS OF QUERIES

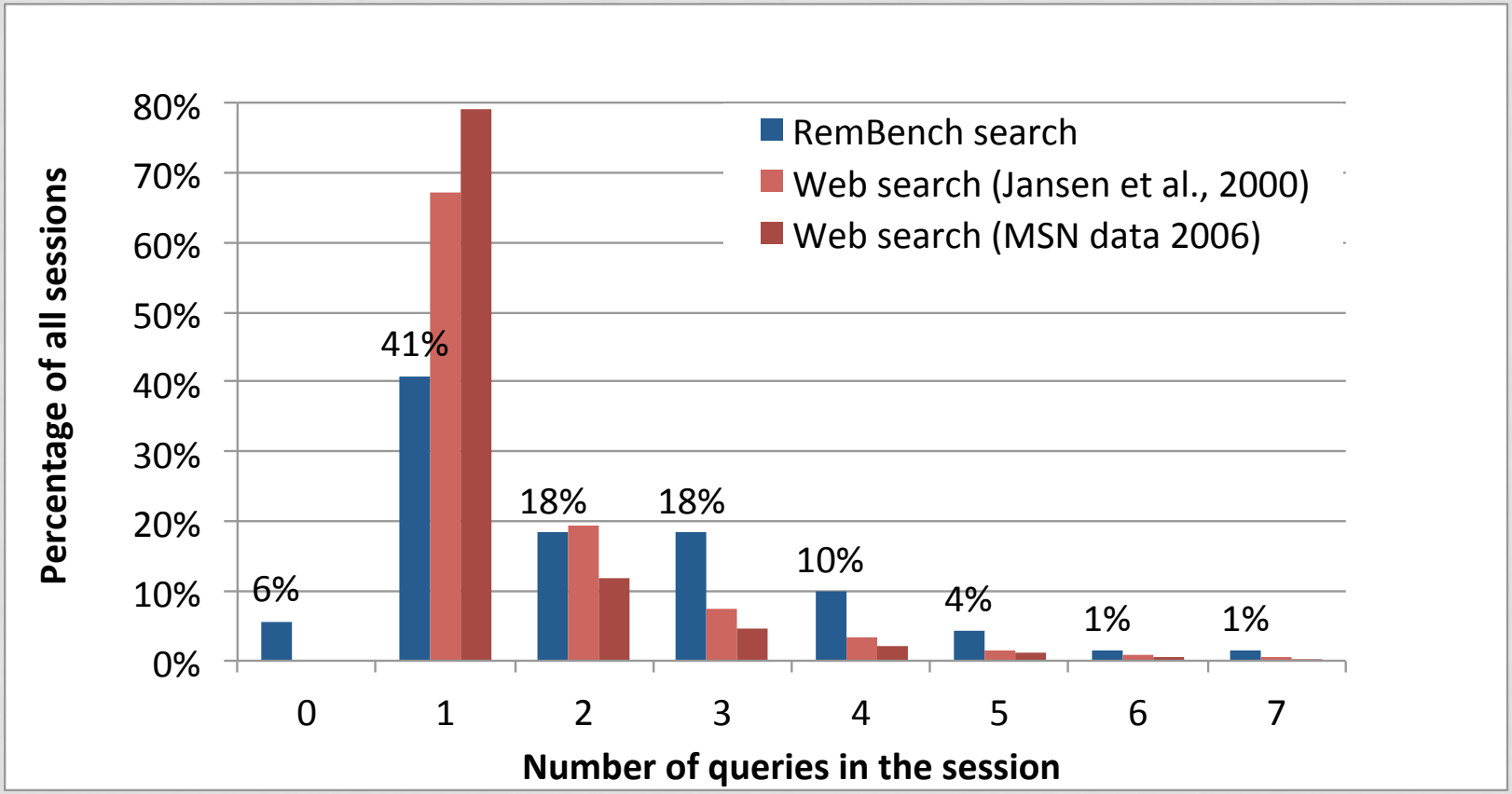
SILVESTRI SECTION 2

USER BEHAVIOUR DEPENDS ON THE SEARCH ENGINE

- Queries in large scale web search engines and more specific search environments are different
  - Web search:
    - queries are short (1-3 words. avg: 2.35 words)
    - queries are simple: hardly any Boolean operators or phrases
    - many queries are navigational in nature
  - Digital libraries or domain-specific search:
    - longer queries
    - more complex queries


Rembench
rembench.huygens.knaw.nl
REMBENCH
About the life an
http://rembench.huygens.knaw.nl/
Search
Fuzzy
Date range
1424
2013
Search
Filter
Search results (6267)
Location
Amsterdam (1278)
Den Haag (640)
New York (365)
Leiden (298)
Londen (286)
London (283)
United States (253)
Parijs (234)
Paris (196)
Nijmegen (175)
More
Works of art (1857)
Artists (59)
Primary sources (2003)
Library sources (2348)
verticals
(federated search)
Author/artist name
Rembrandt (932)
Roghman, Roelant (161)
Maes, Nicolaes (159)
Bol, Ferdinand (127)
Lastman, Pieter (93)
Flinck, Govert (79)
Doomer, Lambert (72)
Vries, Abraham de (64)
Hoogstraten, Samuel van (62)
Gerard Hoet (57)
More
faceted search

USER BEHAVIOUR DEPENDS ON THE SEARCH ENGINE



Universiteit Leiden

Verberne et al. (2016). Information access in the art history domain: Evaluating a federated search engine for Rembrandt research. Digital Humanities Quarterly.

OTHER USER AND QUERY ASPECTS

- User interest: categorize queries in broad topics
  - Can be used for results diversification
- Query behaviour per time of day
  - Number of queries per hour
  - Distribution of topics
  - The distribution of queries over time is ‘bursty’
- Specificity: the more specific the information need, the longer the query
- Query modification behaviour

QUERY MODIFICATION BEHAVIOUR

- 7 types of query modification:
  - New: A query for a topic not previously searched for by this user
  - Generalization: A query on the same topic as the previous one, but seeking more general information than the previous one
  - Specialization: A query on the same topic as the previous one, but seeking more specific information than the previous one
  - Reformulation: A query on the same topic that can be viewed as neither a generalization nor a specialization, but a reformulation of the prior query
  - Interruption: A query on a topic searched on earlier by a user that has been interrupted by a search on another topic
  - Request for additional results: A request for another set of results on the same query Blank queries: Log entries containing no query (e.g. filters)
- We can use these patterns to compare user behaviour between search engines

LEARN FROM USER BEHAVIOUR

SILVESTRI SECTION 3

LEARN FROM USER BEHAVIOUR

- How can be decided
  - what the underlying intent of the query was?
  - if a query has been correctly answered?
  - if a user is satisfied by the search results?

- How does a user re-formulate a query if they are not satisfied yet?
- What is the goal/intent of a user?

QUERY INTENT

- Query intent categories according to Broder (2002):
  - Navigational
  - Informational
  - Transactional

In one of the first paper devoted to discovery user intent behind queries, Andrei Broder [48] studies the goal a user wants to reach when submitting a query to a search engine. Following Broder's formulation a query can be either a *Navigational query* — where the immediate intent is to reach a particular site (e.g. *Greyhound Bus*, *american airlines home*, or *Don Knuth*); an *Informational query* — where the intent is to acquire some information assumed to be present on one or more web pages (e.g. *San Francisco* or *normocytic anemia*); a *Transactional queries* — where the intent is to perform some web-mediated activity (e.g. *online music*, or *online flower delivery service*).

IIR section 19.4

QUERY INTENT

- Experiment:
  - Ask users to label their own queries with the intent
  - Ask external assessors to do the same
  - Calculate the agreement among the assessors, and between the user (query owner) and each assessor

- Findings:
  - Agreement among external assessors: Cohen's kappa = 0.29
  - Agreement with query owner: Cohen's kappa = 0.09

- Query intent cannot be determined from the query log by external assessors

LEARN FROM USER INTERACTIONS

SILVESTRI SECTION 3.2

LEARN FROM USER INTERACTIONS

What can the search engine learn from:

- Query modification behaviour
  - query suggestions
- Clicked documents for queries
  - improved ranking (cf. lecture 9)
  - similarity between queries
  - evaluation

QUERY SUGGESTION

- Use information on past users' queries to propose a particular user with a list of queries related to the one submitted
- The user can select the best similar query to refine their search
- Goal: find related queries in the query log
  - based on common substring
  - based on co-occurrence in a session
  - based on term clustering (embeddings)
  - based on clicks
  - → or a combination of these

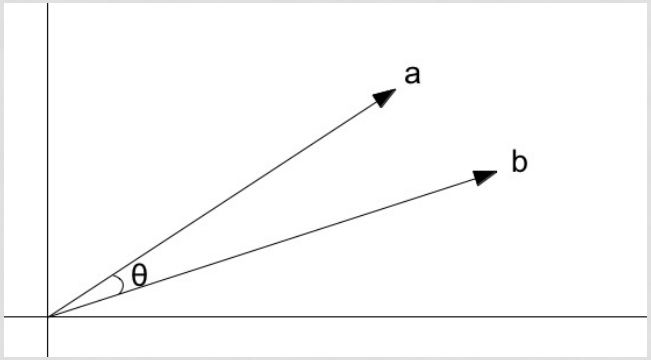
INFORMATION RE-FINDING

- An interesting event to detect is information re-retrieval
- It is common that users search instead of bookmark URLs
  - repeated queries by the same user are almost 50% of the total number of queries (numbers differ between search engines)

Devlin: Bert: Pre-training of deep bidirectional... - Google Scholar	23/03/2022,
Devlin: Bert: Pre-training of deep bidirectional... - Google Scholar	23/03/2022,
Devlin: Bert: Pre-training of deep bidirectional... - Google Scholar	23/03/2022,
devlin bert - Google Scholar	23/03/2022,
devlin bert - Google Scholar	16/03/2022,
Devlin: Bert: Pre-training of deep bidirectional... - Google Scholar	30/11/2021,
Devlin: Bert: Pre-training of deep bidirectional... - Google Scholar	30/11/2021,
bert devlin - Google Scholar	30/11/2021,
devlin - Google Scholar	26/10/2021,
bert devlin - Google Scholar	10/10/2021,
bert devlin - Google Scholar	21/04/2021,
Devlin: Bert: Pre-training of deep bidirectional... - Google Scholar	21/04/2021,
Devlin: Bert: Pre-training of deep bidirectional... - Google Scholar	08/04/2021,
devlin bert - Google Scholar	16/10/2020,
bert devlin - Google Scholar	07/10/2020,
Devlin: Bert: Pre-training of deep bidirectional... - Google Scholar	07/10/2020,
Devlin: Bert: Pre-training of deep bidirectional... - Google Scholar	07/10/2020,
bert devlin - Google Scholar	07/10/2020,
bert devlin - Google Scholar	05/10/2020,
Devlin: Bert: Pre-training of deep bidirectional... - Google Scholar	15/09/2020,
Devlin: Bert: Pre-training of deep bidirectional... - Google Scholar	15/09/2020,

SIMILARITY BETWEEN QUERIES

- Represent queries as vectors in a vector space where the dimensions are clicked URLs
- Calculate cosine similarity between two queries
- Example:
  - vector space with 5 URLs
  - feature values are click counts
  - a and b are queries
    - a = (0,3,0,4,0)
    - b = (1,2,0,0,1)
  - What is the cosine similarity?



SIMILARITY BETWEEN QUERIES

- a and b are queries
  - a = (0,3,0,4,0)
  - b = (1,2,0,0,2)
- What is the cosine similarity?

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

- inner product of a and b: $0 * 1 + 3 * 2 + 0 * 0 + 4*0+0*2=6$
- square root of sum of squares for a and b:
  - $\sqrt{0^2 + 3^2 + 0^2 + 4^2 + 0^2} = \sqrt{25} = 5$
  - $\sqrt{1^2 + 2^2 + 0^2 + 0^2 + 2^2} = \sqrt{9} = 3$
- $sim = \frac{6}{5*3} = 0.4$

SIMILARITY BETWEEN QUERIES

Table 3.3. Equivalent queries.

Query	Sim	Type of Equivalence
tcfu ↔ teachers federal credit union	1.0	acronym
fhb ↔ first hawaiian bank	1.0	acronym
wtvf ↔ new channel 5	1.0	synonyms (Nashville TV channel)
ccap ↔ wcca	1.0	synonyms (Wisconsin court Internet access)
free hispanic chat↔ latinopeoplemeet	1.0	synonym for domain name
lj ↔ www.livejournal.com	1.0	acronym for URL
babel fish ↔ altavista babel fish	1.0	synonyms
aj ↔ askgeeves	1.0	synonyms with misspell
yahoo for kids ↔ yahooligains	0.9	synonym for misspelled domain name
unit conversion ↔ online conversion	0.85	synonym
merriam↔m-w.com	0.84	name for domain name
yahoo directions↔maps.yahoo.com	0.48	synonym for URL

USE LOG DATA FOR EVALUATION

- How can we use log data for evaluation?

- Use clicking and browsing behaviour in addition to queries:
  - the click-through rate: the number of clicks a query attracts
  - time-on-page: the time spent on the result page
  - scrolling behaviour: how users interact with the page
  - stopping information: does the user abandon the search engine after the click?

LIMITATIONS OF QUERY LOGS

- Information need is unknown
  - although it can partly be deduced from previous queries
  - e.g. recognizing that a user is planning a trip

- Relevance assessments are unknown
  - deduce from clicks + dwell time
  - stay on a result page → result probably relevant
  - abandon the search engine → user satisfied?

SIMULATION OF USER INTERACTION

SIMULATION OF INTERACTION

- User simulations as What-if Experiments (observational)
  - To observe what happens when a user behaves in different ways
  - E.g. investigate the effect of query modification behaviour on the effectiveness of the search engine

- Session simulation:
  1. Simulate queries
  2. Simulate clicks
  3. Simulate user satisfaction

Azzopardi et al. (2011). Report on the SIGIR 2010 workshop on the simulation of interaction.

USER MODELLING

- Session simulation:
  1. Simulate queries
  2. Simulate clicks
  3. Simulate user satisfaction

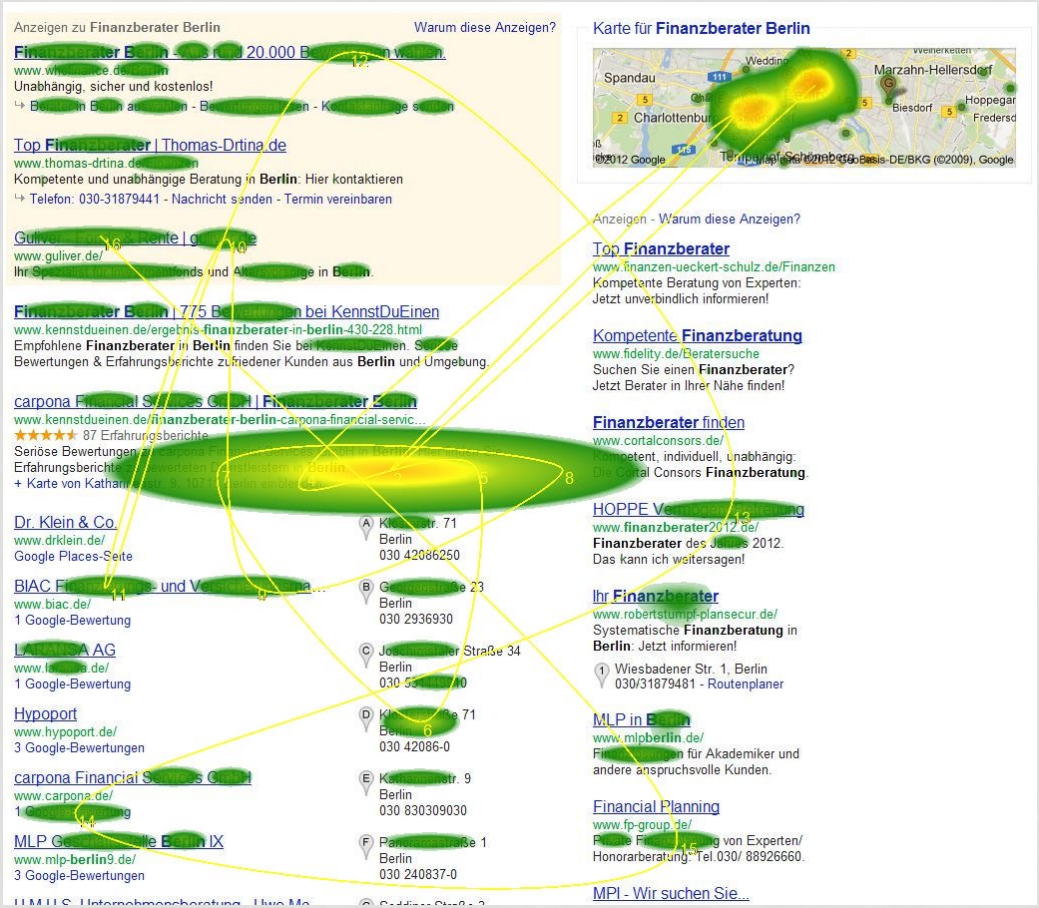
- Each of these require a model for range of user behaviour
  - Users do not always behave deterministically
  - They might make non-optimal choices
  - Models need to contain noise

CLICK MODELS

- How do users examine the result list and where do they click?
- How to create a model that can be used for simulation?

CLICK MODELS

- Most models assume that the user examines the result list from top to bottom: the cascade assumption
- This is also what ranked evaluation metrics (MAP, nDCG) assume
- But real user behaviour is more complex



CLICK MODELS

- Deterministic models for result examination
- What determines if a user stops?
  - the frustration point rule
  - the satisfaction stopping rule
  - the combination rule
  - the representational stability rule
  - the difference threshold rule
  - …

CLICK MODELS

- Probabilistic models for result examination
- Dependent Click Model (DCM):
  1. users traverse result lists from top to bottom
  2. users examine each document as it is encountered
  3. user decides whether to click on the document or skip it
  4. after each clicked document the user decides whether or not to continue examining the document list
  5. relevant documents are more likely to be clicked than non-relevant documents

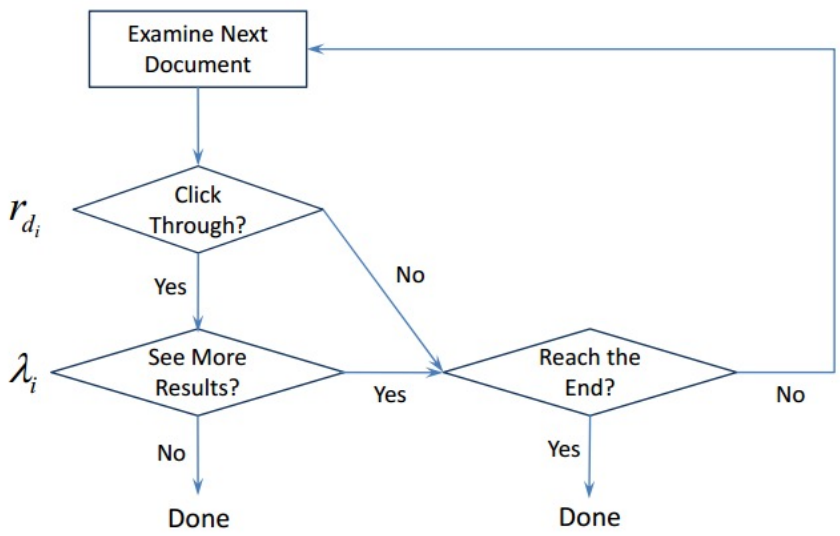


Figure 1: The user model of DCM, in which r_{d_i} is the document relevance of d_i, and λ_i is the user behavior parameter for position i.

Guo et al. (2009). Efficient multiple-click models in web search.

CLICK MODELS

$e_i = 1$: user examines rank i

$c_i = 1$: user clicks on rank i

$$P(C_i = 1|E_i = 0) = 0,$$
$$P(C_i = 1|E_i = 1) = r_{d_i},$$

$$P(E_1 = 1) = 1,$$
$$P(E_{i+1} = 1|E_i = 0) = 0.$$

$$P(E_{i+1} = 1|E_i = 1, C_i = 1) = \lambda_i$$
$$P(E_{i+1} = 1|E_i = 1, C_i = 0) = 1,$$

cascade assumption

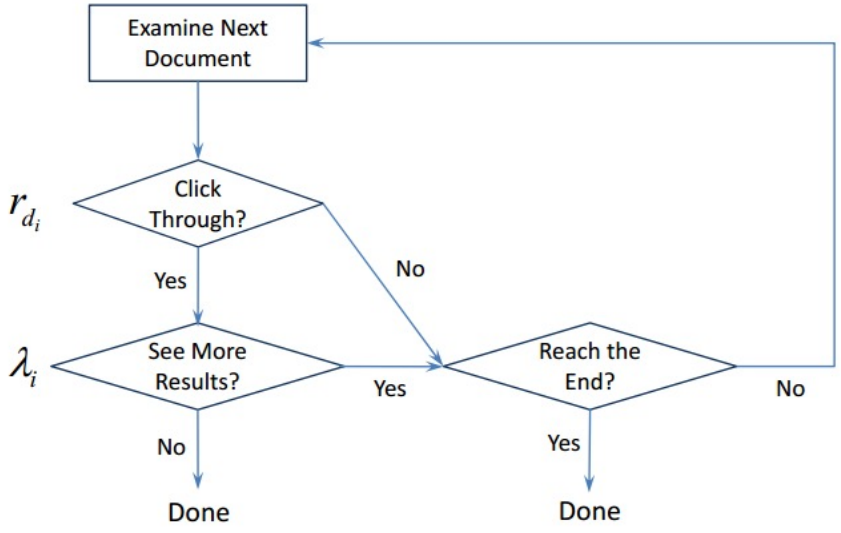


Figure 1: The user model of DCM, in which r_{d_i} is the document relevance of d_i, and λ_i is the user behavior parameter for position i.

Guo et al. (2009). Efficient multiple-click models in web search.

CLICK MODELS

- Adding noise to the Dependent Click Model:

relevance grade	0	1	2	3
perfect	0.00	0.33	0.67	1.00
informational	0.40	0.60	0.75	0.90
navigational	0.05	0.33	0.67	0.95

- Adapting examination probability for click bias:

$$P(E_{i+1} = 1 | E_i = 1, C = 0) = \frac{1}{1 + e^{k(i-y)}}$$

Hofmann et al. (2011). Balancing exploration and exploitation in learning to rank online.
Verberne et al. (2015). User simulations for interactive search: evaluating personalized query suggestion.

CLICK MODELS

User assumptions made in probabilistic click models:

- cascade assumption: examination is in strictly sequential order with no breaks
- click probability depends on document relevance
- examination probability depends on click
- users are homogeneous: their information needs are similar given the same query

SIMULATION OF INTERACTION

Advantages

- Investigate how the system behaves under certain behaviour ("what if...")
- Potentially a large amount of data (1000s of differently behaving users)
- Relatively low cost to create and use
- Enable the exact same circumstances to be replicated, repeated, re-used.
- Encapsulates our understanding of the process

Disadvantages

- Models can become complex if we want to mirror realistic user behaviour
- Simulations enable us to explore many possibilities
  - But which ones? And why?
  - How do we make sense of all the data?
- Does it represent actual user behavior/performance?
- What claims can we make? In what context?

THE COMPLEX SEARCHER MODEL

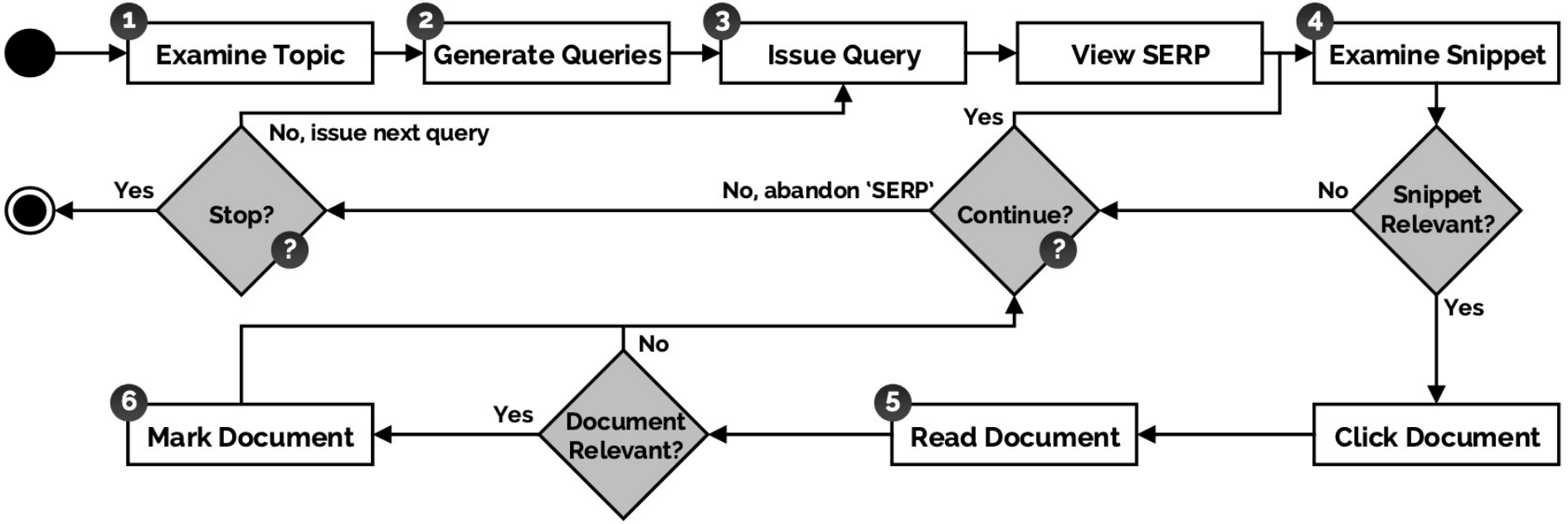


Figure 1: A flowchart of decisions (shown in grey) and tasks (shown in white) undertaken by simulated searchers 'participating' in this study. The model is adapted from Baskaya et al. [4] and Thomas et al. [30]. Numbers associated with tasks correspond to the steps detailed in Section 4.2. The two *?* markers indicate the decisions that are encoded within each implemented stopping strategy (refer to Section 4.5).

Universiteit Leiden

Maxwell et al. (2015). Searching and stopping: An analysis of stopping rules and strategies.

RELEVANCE FEEDBACK

IIR SECTION 9.1

QUERY REFINEMENT

- It can be difficult to formulate a good query
- But it is relatively easy to see if the retrieved documents are not what you were looking for
- Users often need multiple reformulations to get to the best formulation of their information need
- = iterative query refinement
- Idea of relevance feedback: the search engine helps the user to refine the result set

RELEVANCE FEEDBACK

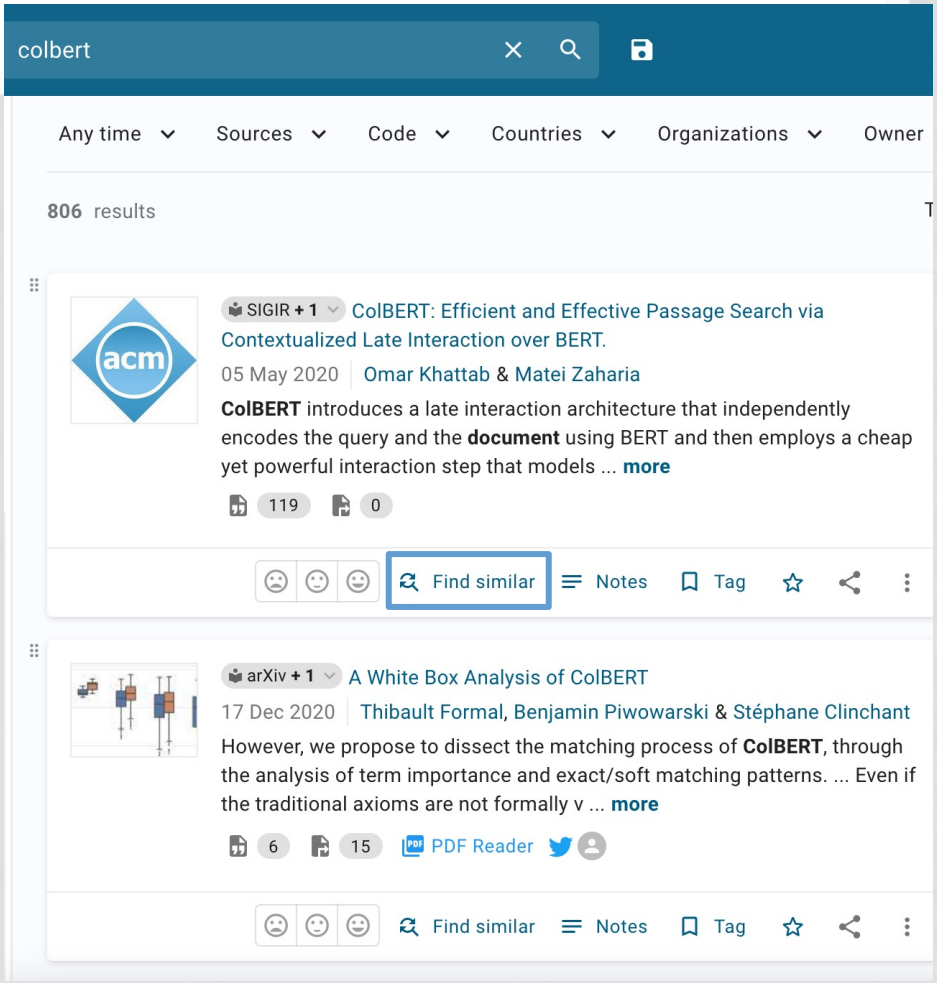
Involve the user in the retrieval process to improve the final result set

1. The user issues a (short, simple) query
2. The system returns an initial set of retrieval results
3. The user marks some returned documents as relevant or nonrelevant
4. The system computes a better representation of the information need based on the user feedback
5. The system displays a revised set of retrieval results

Explicit relevance feedback

RELEVANCE FEEDBACK ON THE WEB

- 'Find similar': This can be viewed as a form of relevance feedback
  - The user indicates which document is the most relevant
- However, relevance feedback has been little used in web search:
  - On the web, most people want to complete their search in a single interaction
  - Relevance feedback is hard to explain to the average user
  - Relevance feedback is mainly a recall enhancing strategy, and web search users are only rarely concerned with high recall



'Find similar' in domain-specific search: https://search.zeta-alpha.com/

PSEUDO RELEVANCE FEEDBACK

Alternative: Pseudo Relevance Feedback (PRF):

- Automates the manual part of relevance feedback, so that the user gets improved retrieval performance without an extended interaction
- Assumption: the top k ranked documents are relevant

PSEUDO RELEVANCE FEEDBACK

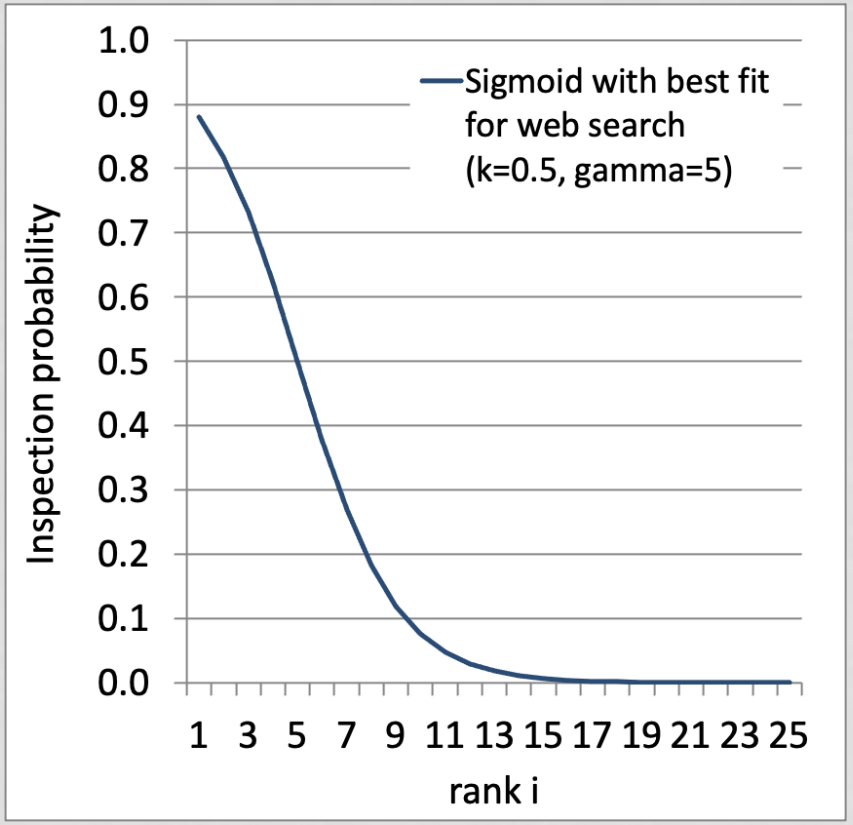
PRF procedure:

1. Take the top 5-20 results returned by the initial query
2. Select top 20-30 terms from these documents (e.g. tf-idf)
3. Add these terms to query, and then return a new ranking

- This automatic technique mostly works. What is the risk?
  - Query drift. “For example, if the query is about copper mines and the top several documents are all about mines in Chile, then there may be query drift in the direction of documents on Chile.”

USER BEHAVIOUR IS BIASED

- Remember the position bias we discussed in lecture 9?
- Challenge: deduce an unbiased signal from biased interactions
- All signals in the click log have a form of bias in them
- But user interactions do provide us with relevant information



CLICKS AND DWELL TIME AS USER FEEDBACK

- "Looking for 'quality clicks', this headline can also be very different"
  - NRC, 29 April 2022
  - "This form of testing is called 'A/B testing'. An article is presented to different readers with different headlines. Some of the users will see headline A, others will see headline B. The headline that generates the most clicks per view 'wins' and becomes the headline that all readers will see afterwards. Your collective reading and clicking behaviour of readers are therefore (anonymously) analyzed for this test. NRC uses the Chartbeat analysis platform for this."
  - "But clicks are not the silver bullet, says Sophie van Oostvoorn, who analyzes reading behaviour from the reader's desk. Chartbeat also records 'quality clicks'. A click counts as a quality click if a reader spends at least fifteen seconds reading the piece after clicking. If a roaring headline creates expectations that the article doesn't live up to, a reader is quickly gone. Van Oostvoorn is therefore not afraid of proliferation of clickbait."

https://www.nrc.nl/nieuws/2022/04/29/op-zoek-naar-kwaliteitskliks-kan-deze-kop-ook-heel-anders-a4118997

CONCLUSIONS

HOMEWORK

- Read:
  - “Mining Query Logs: Turning Search Usage Data into Knowledge” by Fabrizio Silvestri. Section 2, 3, 4.1, 4.2 and 4.3 (p. 16-64)

- Exercises about query log analysis (using real log data)
- See Brightspace -> assignments
- Deadline: Sunday May 8, 23.59

AFTER THIS LECTURE…

- You can describe what type of information can be extracted from query logs
- You can name three strategies to prevent privacy leaks when log data is used
- You can define a session in two ways: conceptually and pragmatically
- You can define the query intent categories by Broder
- You can define and explain the Dependent Click Model and its assumptions
- You can list advantages and disadvantages of using simulation of interaction
- You can explain in what way log data that can be used for query suggestion
- You can calculate the similarity between queries based on their clicks
- You can define and explain pseudo-relevance feedback