Information Retrieval

Foundational Concepts and Models

Maarten de Rijke
derijke@uva.nl

University of Amsterdam

ESSIR 2015 Tutorial

Based on joint material/discussions with Fei Cai, Aleksandr Chuklin, Katja Hofmann, Xinyi Li, Ilya Markov, Daan Odijk, Anne Schuth, Shimon Whiteson, Masrour Zoghi.

- "**Information retrieval** (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text (or other content-based) indexing." [1]

- "**Information retrieval** (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text (or other content-based) indexing." [1]
- **Users** and **search engines** the essential agents
  - users' information needs
  - search engine results

- "**Information retrieval** (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text (or other content-based) indexing." [1]
- **Users** and **search engines** the essential agents
  - users' information needs
  - search engine results
- Search engine and users are agents that perform actions in response to each other: interactions, result list, interactions, result list, ...

[1] https://en.wikipedia.org/wiki/Information_retrieval

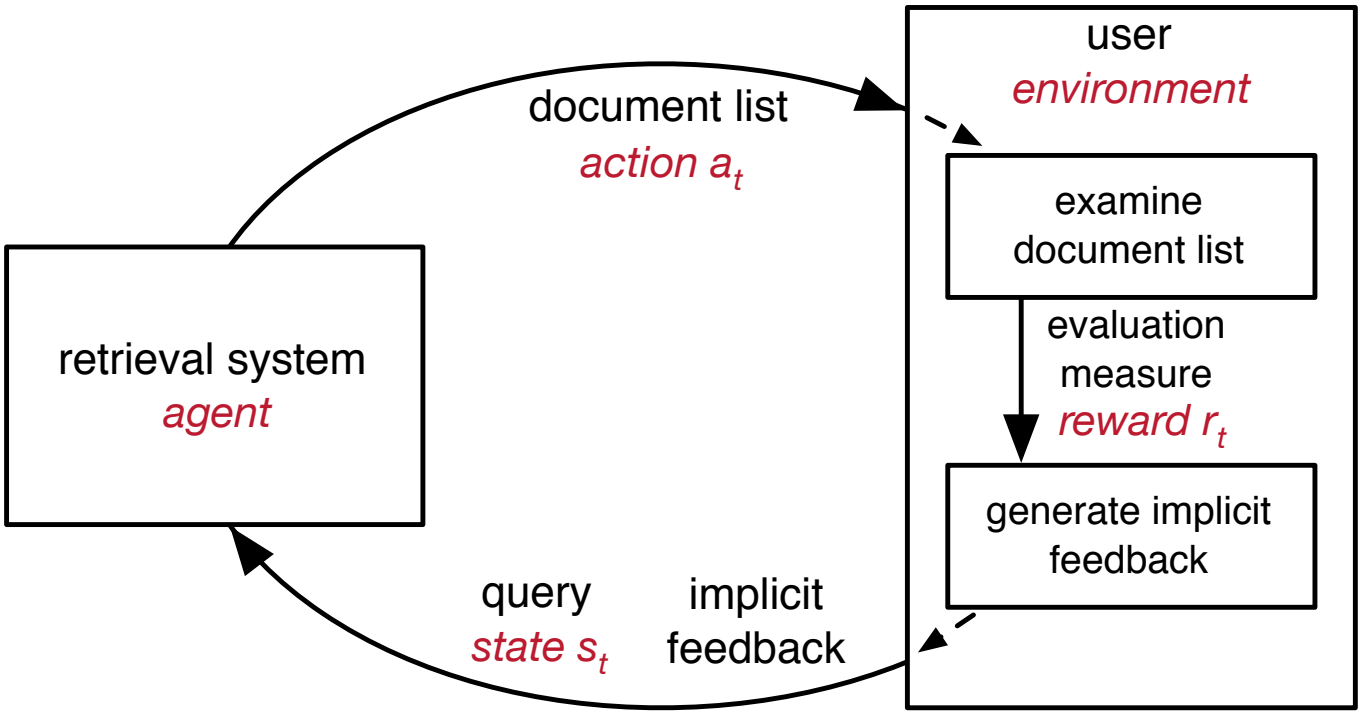
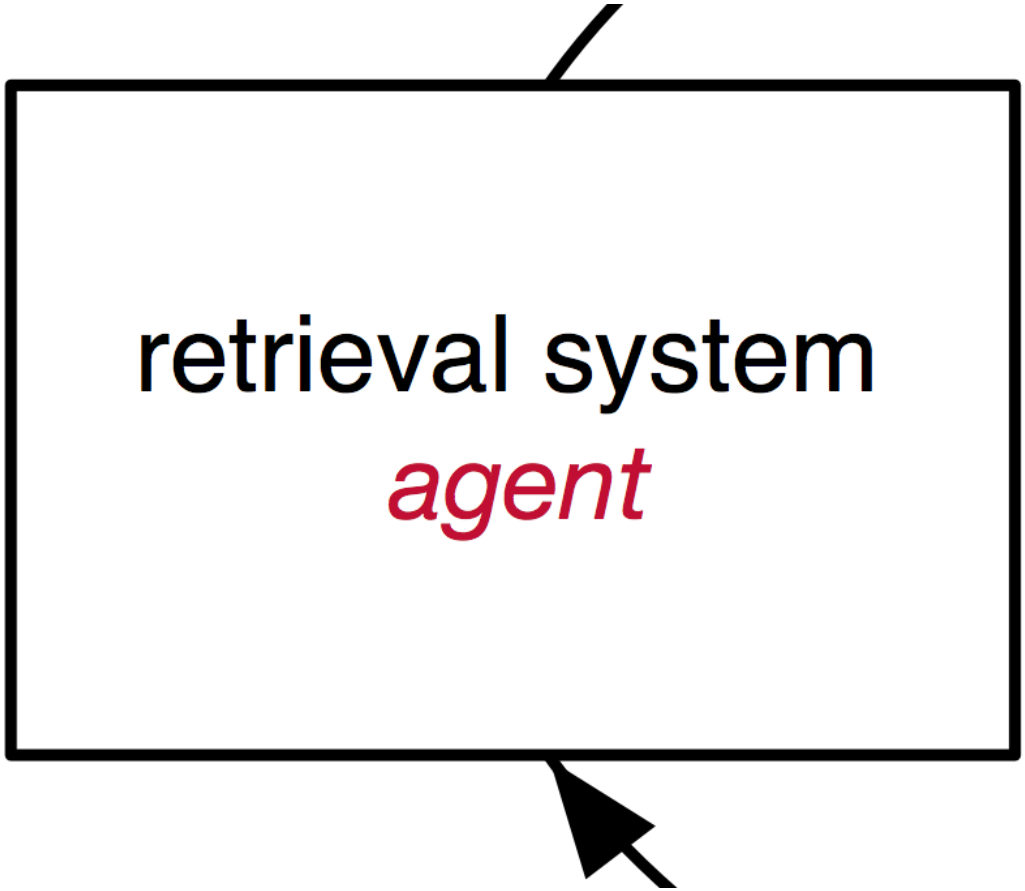
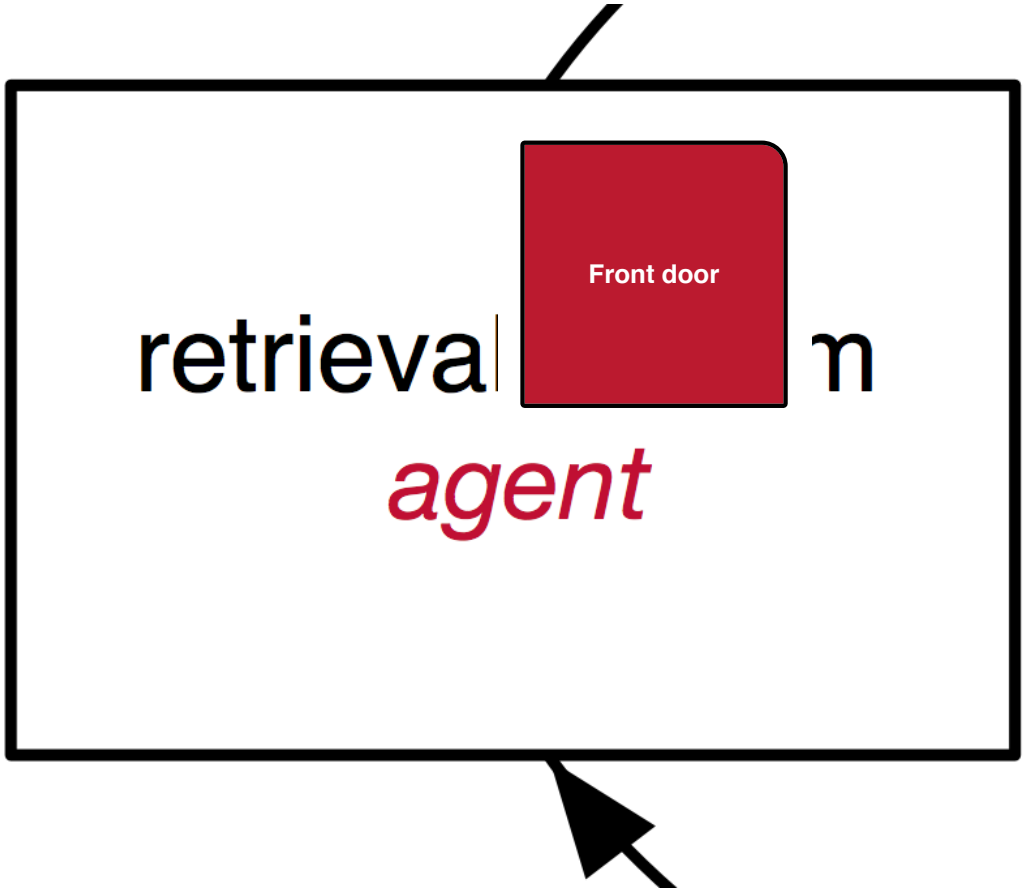


Image taken from K. Hofmann, S. Whiteson, and M. de Rijke. Contextual bandits for information retrieval. In NIPS 2011 Workshop on Bayesian Optimization, Experimental Design, and Bandits, December 2011.

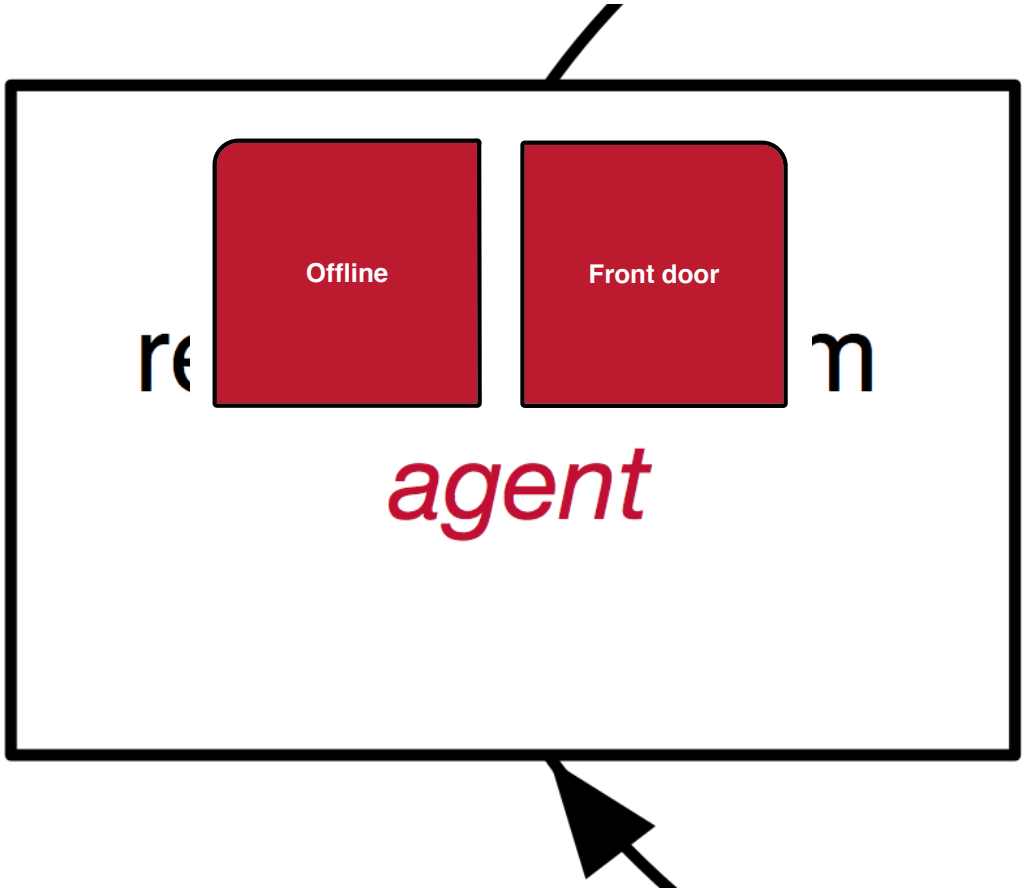
The retrieval system



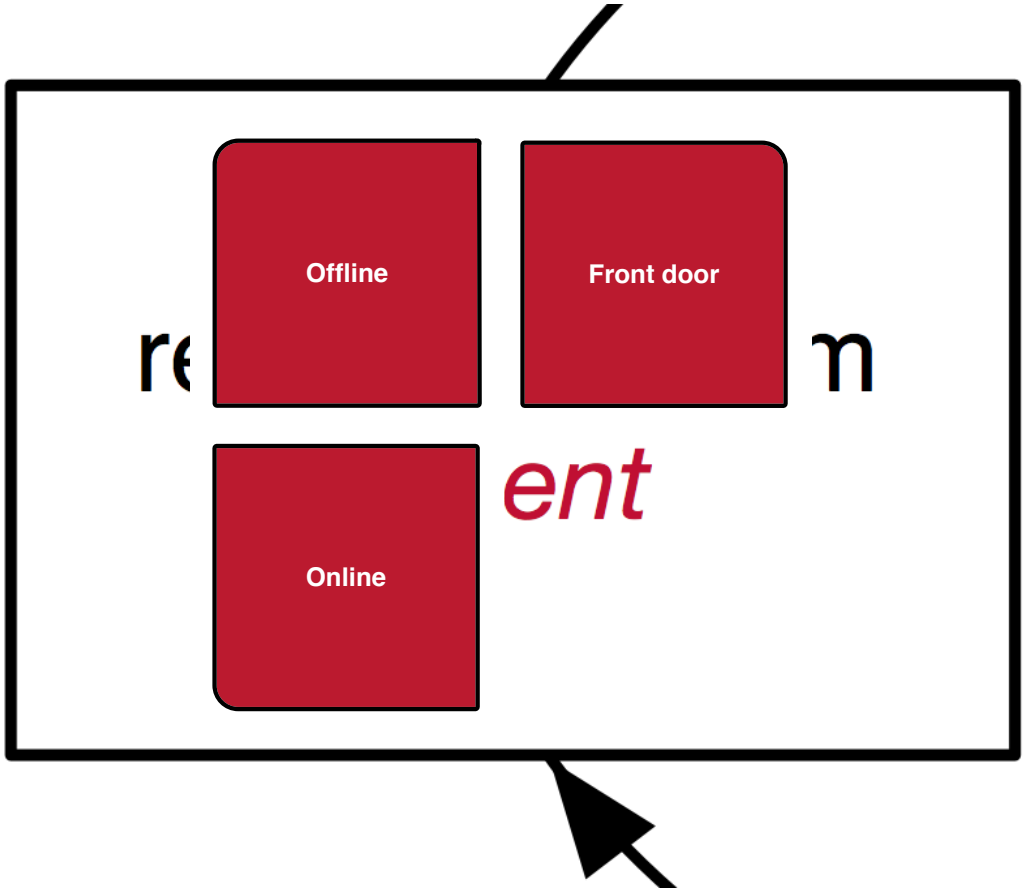
The retrieval system



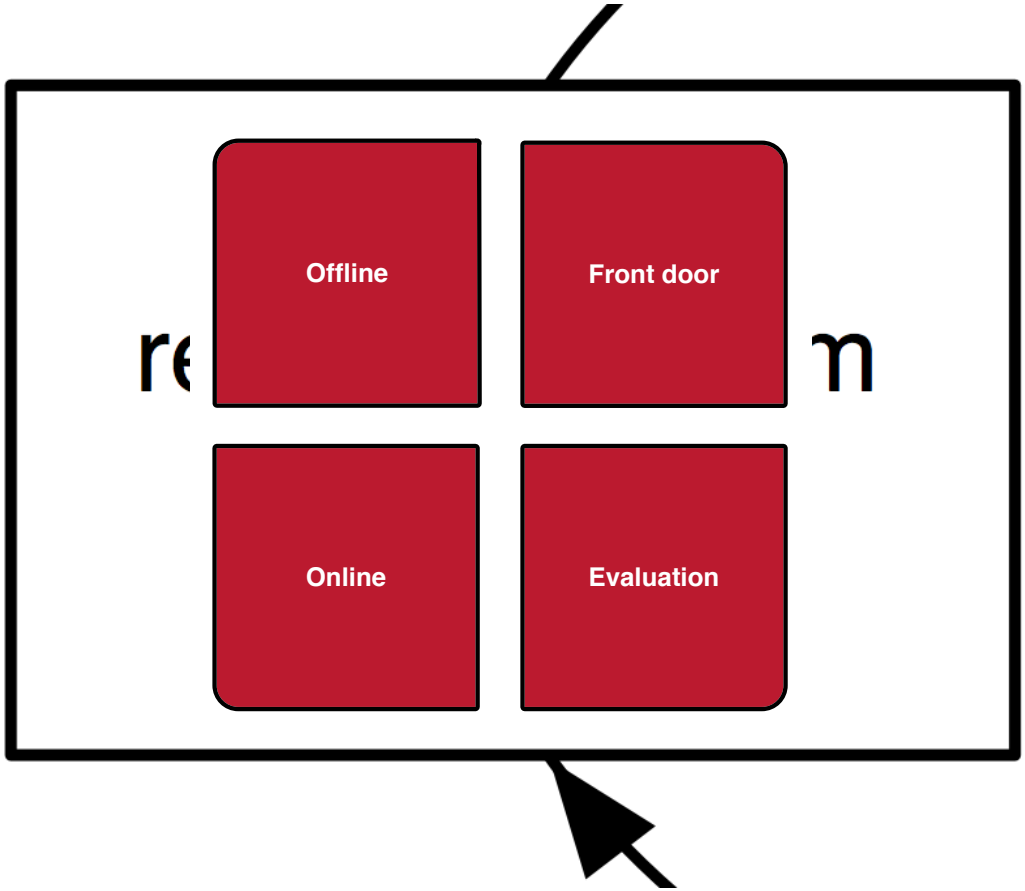
The retrieval system



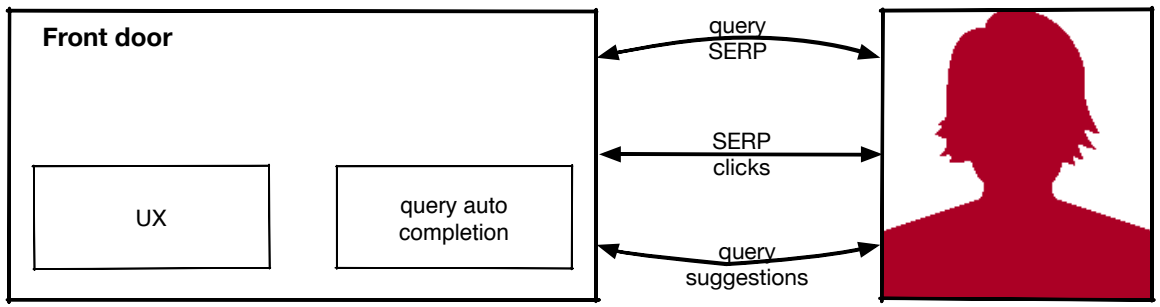
The retrieval system

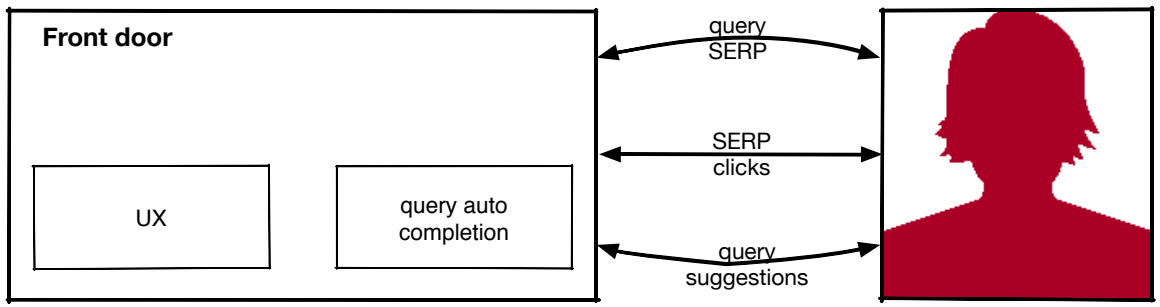


The retrieval system



The retrieval system


Front door
UX
query auto completion
query
SERP
SERP
clicks
query
suggestions

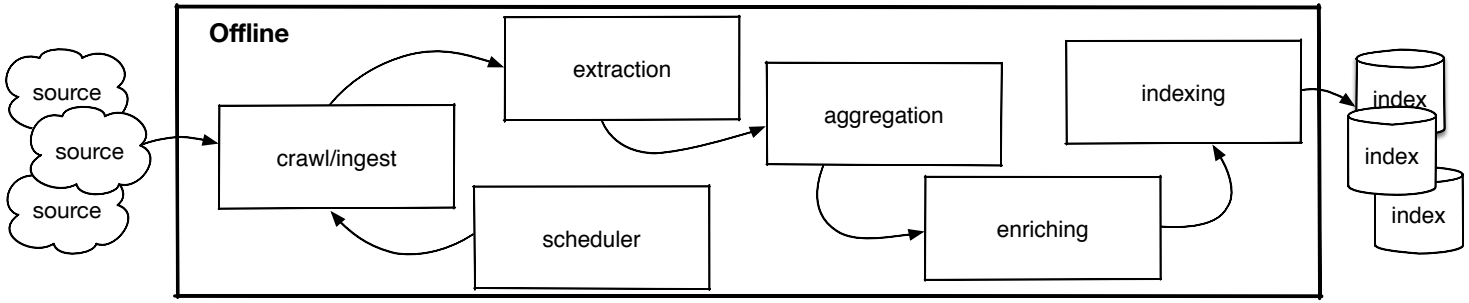


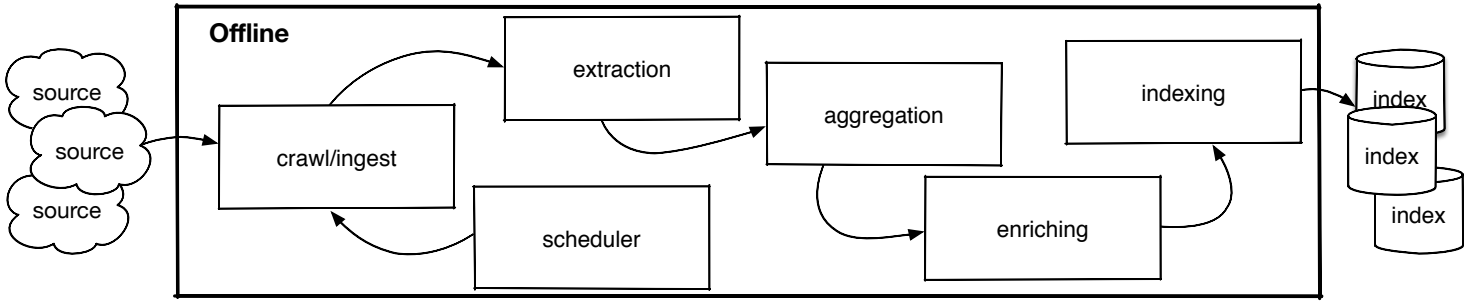
The front door determines the user experience, produces search engine result page (SERP).

Receives query, may return query auto completion suggestions.

Receives other user signals (clicks, shares, . . .).

Should be connected to evaluation framework and to online module.


Offline
source
source
source
crawl/ingest
extraction
scheduler
aggregation
enriching
indexing
index
index
index



Content crawling/ingestion

- Scheduling (freshness, e.g., based on social media) and discovery

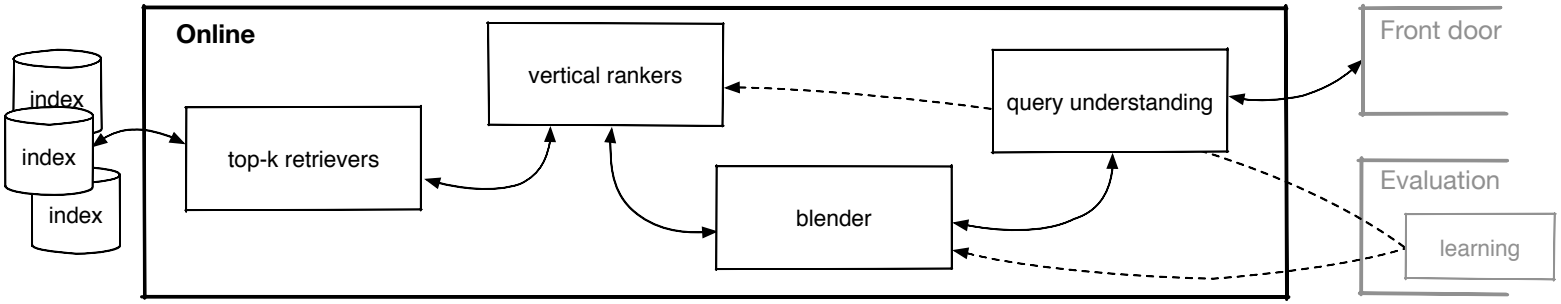
Enriching

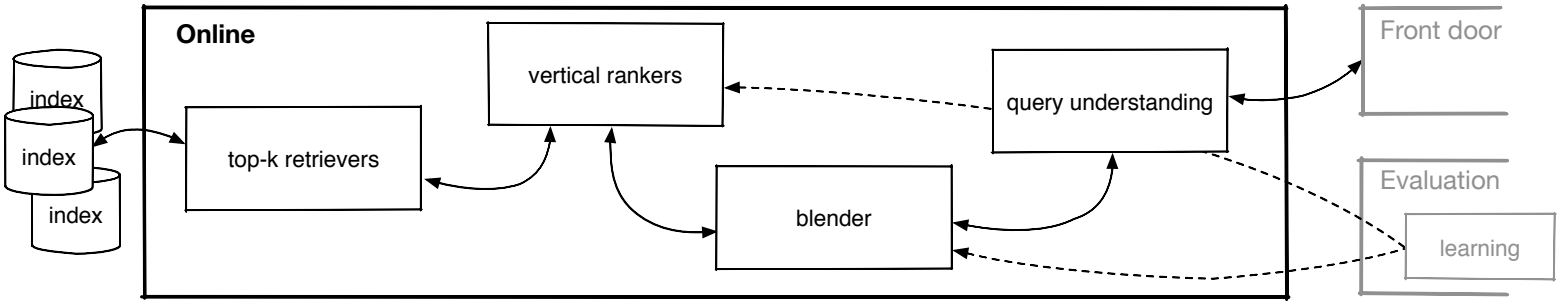
- Classification (spam, adult, . . .), extraction (entities, multimedia, . . .), and annotators (document expansion, translation, . . .)

Aggregation of sources

- Interaction features (clicks, . . .), social features (Twitter, . . .)
- Graph-based computations (anchor text, PageRank, HITS, . . .)

Indexing


Online
index
index
index
top-k retrievers
vertical rankers
blender
query understanding
Front door
Evaluation
learning



Query auto completion

Query understanding

- Down: Alterations (suggestions, translation, ...), classification (entities, intent, performance prediction, task detection, ...)
- Up: SERP generation (snippet generation, device tailoring, answer insertion, suggestions, ...)

Blender

- Down: ranker type and parameter selection (web, fresh, news, image, video, apps, social, ...)
- Up: merging results (interleaving, diversity) and UX selection

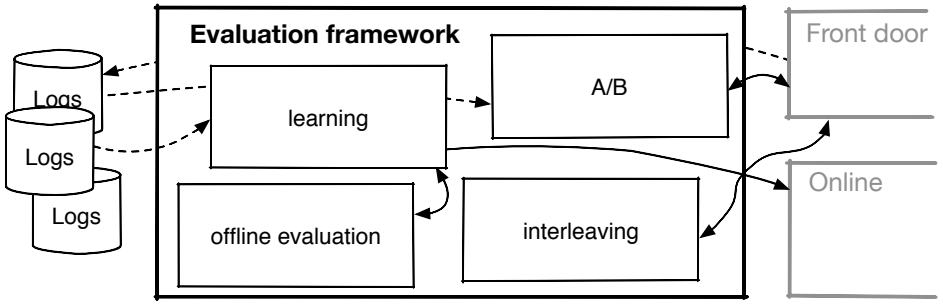
Vertical ranking

- Hotfixes (personalized); compute Q+D features; apply rankers

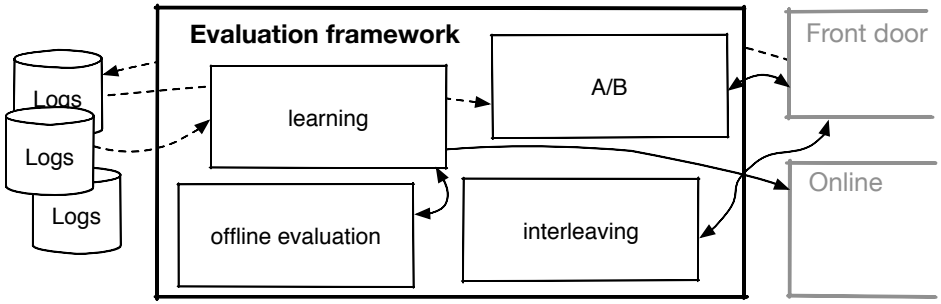
Top-k retrieval

- Keyword matching, retrieval of document features

Evaluation framework



Evaluation framework



Metrics

- Online, offline

Flighting

- Bucketing, A/B testing, interleaving

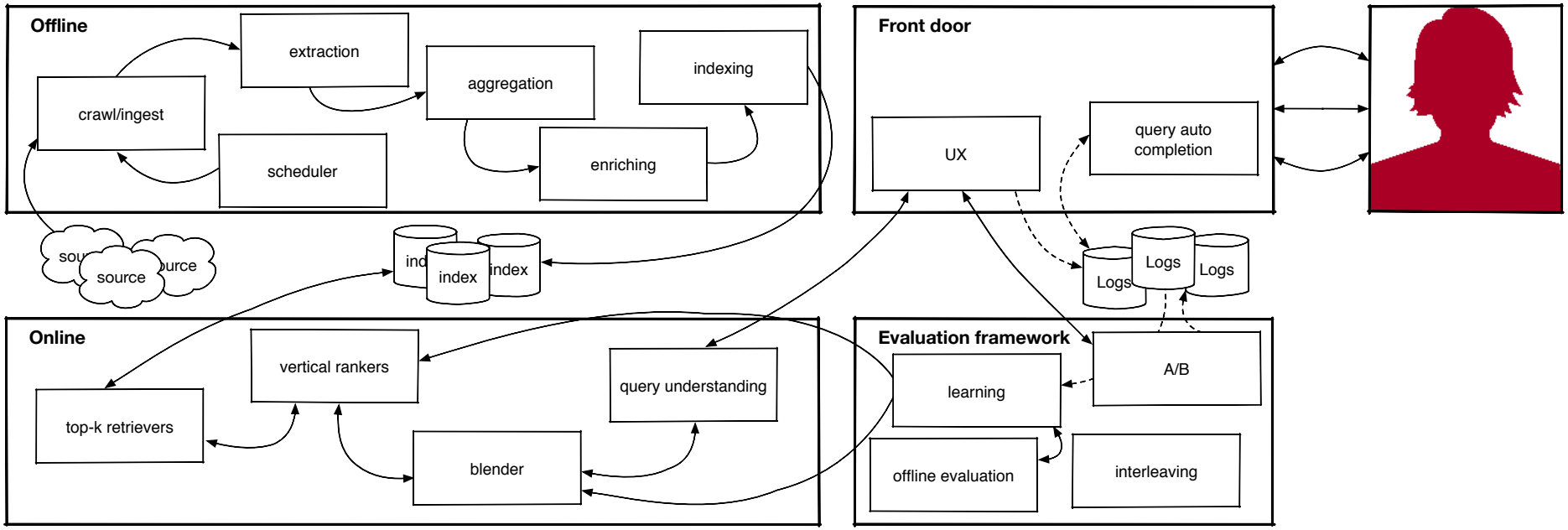
Learning

Logging

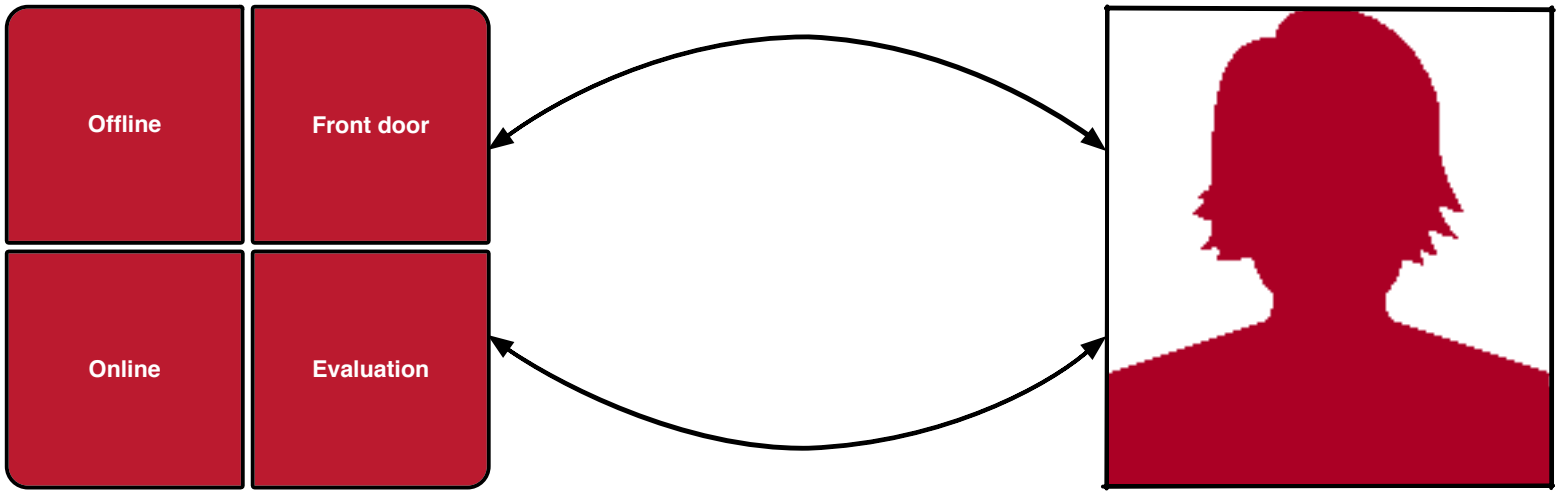
Annotations

- Experts, crowd

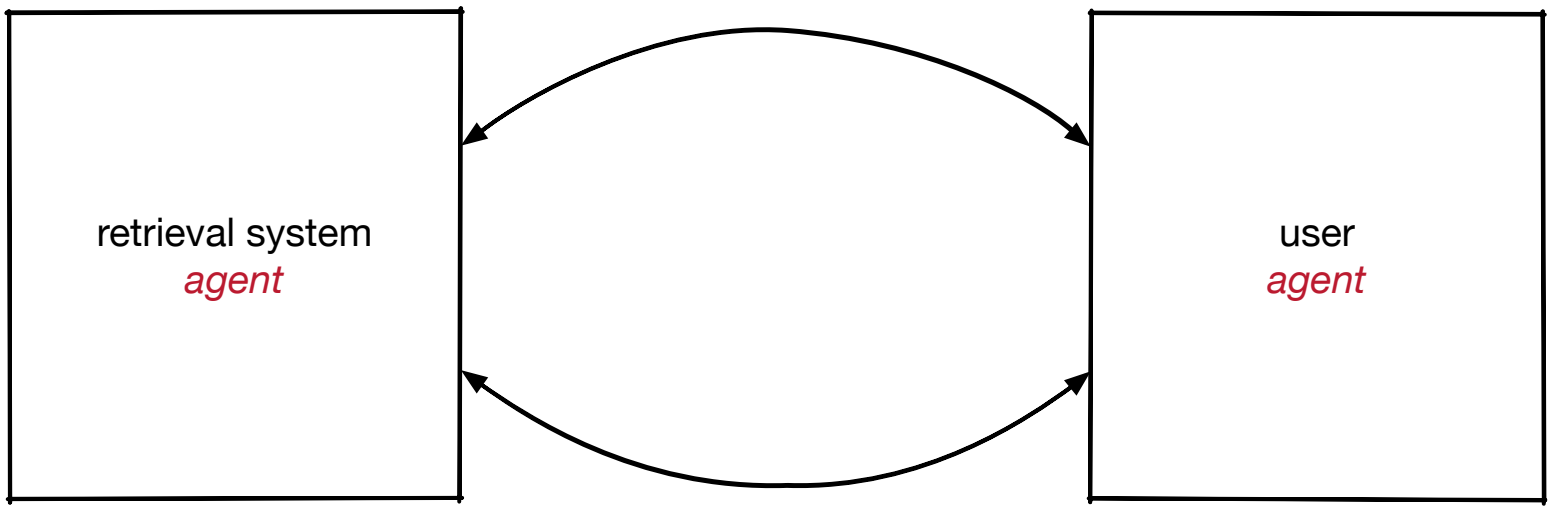
The big picture



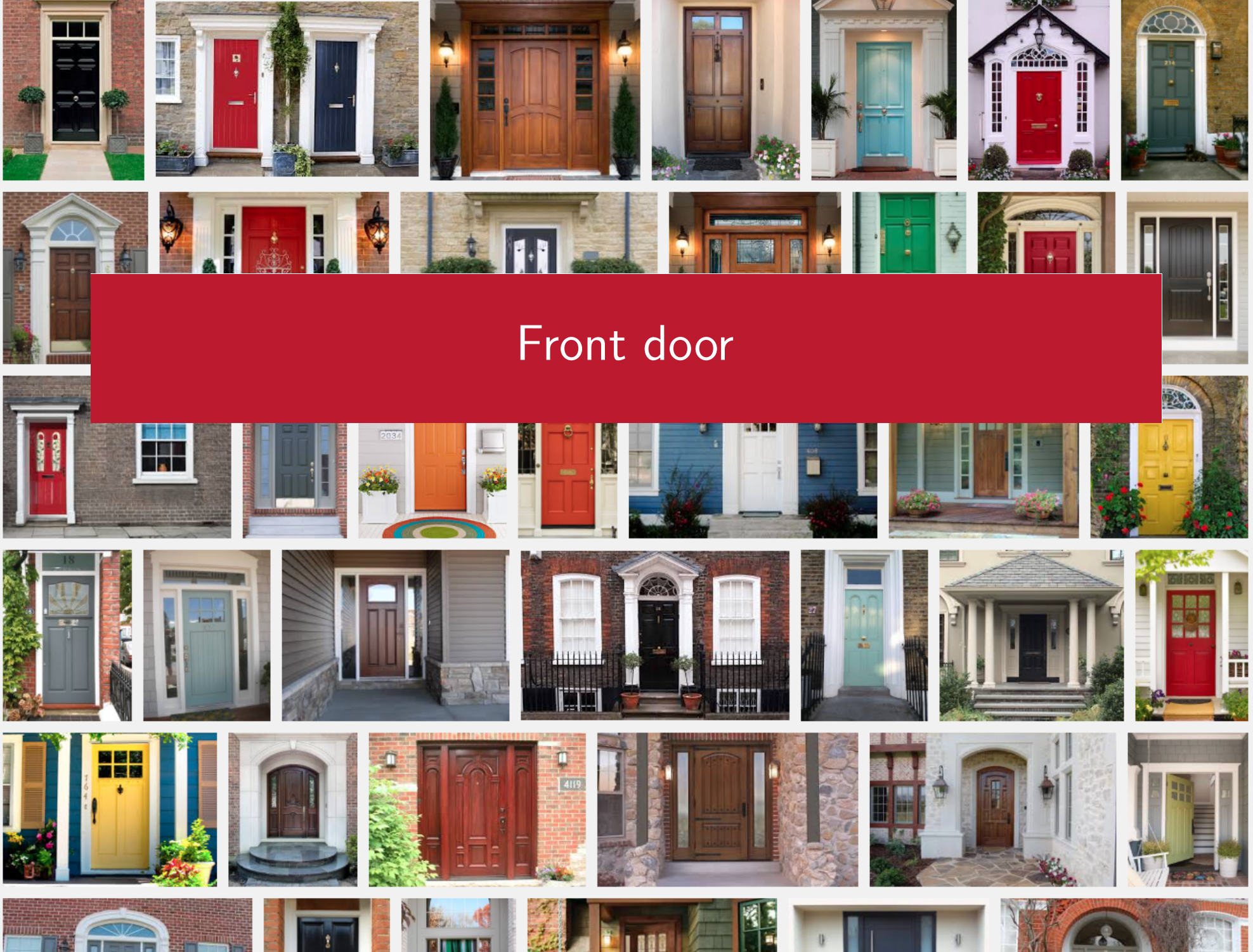
The bigger picture



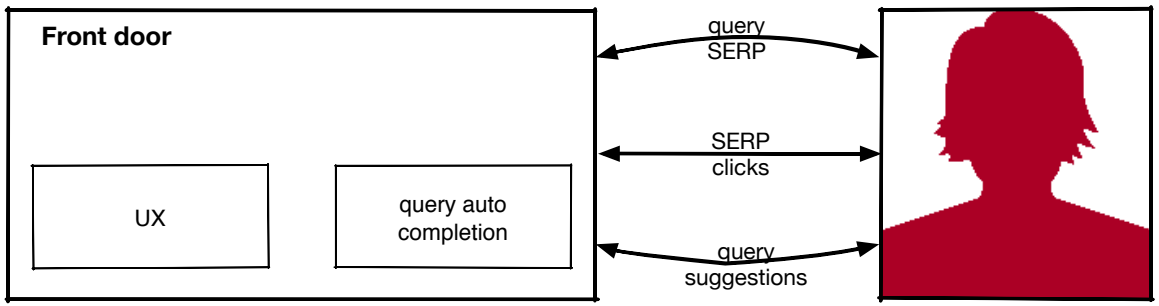
The biggest picture



Outline



Front door



The front door determines the user experience, produces search engine result page (SERP).

Receives query, may return query auto completion suggestions.

Receives other user signals (clicks, shares, . . .).

UX

Search interface guidelines include:

- Offer efficient and informative feedback
- Balance user control with automated actions
- Reduce short-term memory load
- Provide shortcuts
- Reduce errors
- Recognize the importance of small details
- Recognize the importance of aesthetics

To design successful search user interfaces, understand human information seeking process, including strategies people employ when engaged in search

M. Hearst. *Search User Interfaces*, CUP, 2009

User data

	Observational	Experimental
User studies Controlled interpretation of behavior with detailed instrumentation	In-lab behavior observation	Controlled tasks, controlled systems, laboratory studies

User data

	Observational	Experimental
User studies Controlled interpretation of behavior with detailed instrumentation	In-lab behavior observation	Controlled tasks, controlled systems, laboratory studies
User panels In the wild, real-world, tasks, probe for detail	Ethnography, field studies, case reports	Diary studies, critical incident surveys

User data

	Observational	Experimental
User studies Controlled interpretation of behavior with detailed instrumentation	In-lab behavior observation	Controlled tasks, controlled systems, laboratory studies
User panels In the wild, real-world, tasks, probe for detail	Ethnography, field studies, case reports	Diary studies, critical incident surveys
Log analysis No explicit feedback but lots of implicit feedback	Behavioral log analysis	A/B testing, interleaved results

User data

	Observational	Experimental
User studies Controlled interpretation of behavior with detailed instrumentation	In-lab behavior observation	Controlled tasks, controlled systems, laboratory studies
User panels In the wild, real-world, tasks, probe for detail	Ethnography, field studies, case reports	Diary studies, critical incident surveys
Log analysis No explicit feedback but lots of implicit feedback	Behavioral log analysis	A/B testing, interleaved results
	understand behavior	*contrast approaches*

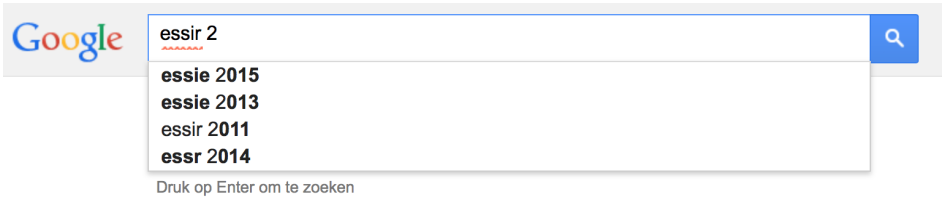
pizza

AOL-user-ct-collection — less — 116×53

```
710766  wwwpeoplesearch.comwww.reviewplace.seardh       2006-05-30 22:10:13
710766  wwwpeoplesearch.comwww.reviewplace.seardh       2006-05-30 22:10:33
711391  can not sleep with snoring husband      2006-03-01 01:24:00
711391  cannot sleep with snoring husband       2006-03-01 01:24:07     9       http://www.wjla.com
711391  cannot sleep with snoring husband       2006-03-01 01:24:07     9       http://www.wjla.com
711391  cannot sleep with snoring husband       2006-03-01 01:33:06     1       http://www.epinions.com
711391  jackie zeaman nude      2006-03-01 15:26:27
711391  jackie zeman nude       2006-03-01 15:26:38
711391  strange cosmos  2006-03-01 16:07:15     1       http://www.strangecosmos.com
711391  mansfield first assembly        2006-03-01 16:09:20     1       http://www.mansfieldfirstassembly.org
711391  mansfield first assembly        2006-03-01 16:09:20     3       http://netministries.org
711391  reverend harry myers    2006-03-01 16:10:07
711391  reverend harry myers    2006-03-01 16:10:30
711391  national enquirer       2006-03-01 17:13:14     1       http://www.nationalenquirer.com
711391  how to kill mockingbirds        2006-03-01 17:18:11
711391  how to kill mockingbirds        2006-03-01 17:18:33
711391  how to kill annoying birds in your yards        2006-03-01 17:18:58
711391  how to kill annoying birds in your yards        2006-03-01 17:19:53     2       http://www.sortprice.com
711391  how to rid your yard of noisy annoying birds    2006-03-01 17:23:08     3       http://shopping.msn.com
711391  how to rid your yard of noisy annoying birds    2006-03-01 17:23:08     10      http://www.bergen.org
711391  how to rid your yard of noisy annoying birds    2006-03-01 17:24:35     15      http://www.saferbrand.com
711391  how do i get mocking birds out of my yard       2006-03-01 17:27:17
711391  how do i get mockingbirds out of my yard        2006-03-01 17:27:36     9       http://www.asri.org
711391  how do i get mockingbirds out of my yard        2006-03-01 17:30:14
711391  how to get rid of noisy loud birds      2006-03-01 17:30:52     3       http://www.bird-x.com
711391  how to get rid of noisy loud birds      2006-03-01 17:30:52     1       http://forums2.gardenweb.com
711391  how to get rid of noisy loud birds      2006-03-01 17:30:52     10      http://www.birding.com
711391  mansfield first assembly        2006-03-01 18:31:36     3       http://netministries.org
711391  beth moore      2006-03-01 19:42:41     1       http://www.lproof.org
711391  judy baker ministries   2006-03-01 19:49:03     2       http://www.embracinggrace.com
711391  god will fulfill your hearts desires    2006-03-01 19:59:06     10      http://www.pureintimacy.org
711391  online friendships can be very special  2006-03-01 23:09:37
711391  online friendships can be very special  2006-03-01 23:09:57
711391  online friendships      2006-03-01 23:10:24
711391  cypress fairbanks isd   2006-03-02 07:56:53     1       http://www.cfisd.net
711391  people are not always how they seem over the internet   2006-03-02 08:31:51
711391  friends online can be different in person       2006-03-02 08:32:42
711391  friends online can be different in person       2006-03-02 08:33:04     13      http://www.salon.com
711391  boston butts    2006-03-02 09:47:36
711391  community christian church houston tx   2006-03-02 16:07:53
711391  gay churches in houston tx      2006-03-02 16:08:23
711391  community gospel church in houston tx   2006-03-02 16:08:45     2       http://www.communitygospel.org
711391  houston tx is one hot place     2006-03-02 18:04:44
711391  houston tx is one hot place to live     2006-03-02 18:04:55     9       http://travel.yahoo.com
711391  houston tx is one hot place to live     2006-03-02 18:16:05     1       http://www.houston-texas-online.com
711391  texas hill country and sights around san antonio tx     2006-03-02 18:19:00     5       http://www.answers.c
om
711391  can liver problems cause you to loose your hair 2006-03-02 18:27:04
711391  can liver problems cause you to loose your hair 2006-03-02 18:27:30     1       http://www.askdoctrish.com
711391  strange cosmos  2006-03-02 19:29:31     1       http://www.strangecosmos.com
711391  white hard dry skin on face     2006-03-02 20:31:29
711391  white hard dry skin on face     2006-03-02 20:32:24
:
```

Query auto completion

Helps users formulate query when they have an intent in mind but not a clear way to express it.



Typical query completion service of modern search engine takes initial characters entered by user and returns matching queries to automatically complete search clue.

Where offered, query completion is heavily used by visitors and highly influential on search results

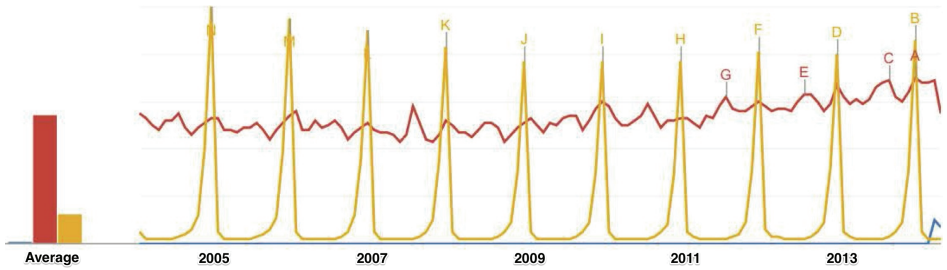
B. Mitra, M. Shokouhi, F. Radlinski, K. Hofmann. On user interactions with query auto-completion. In SIGIR '14, 2014

Useful and straightforward approach to rank QAC candidates is to use Maximum Likelihood Estimation (MLE) based on the past popularity of queries

- Assumes that the current query popularity distribution is the same as what was previously observed

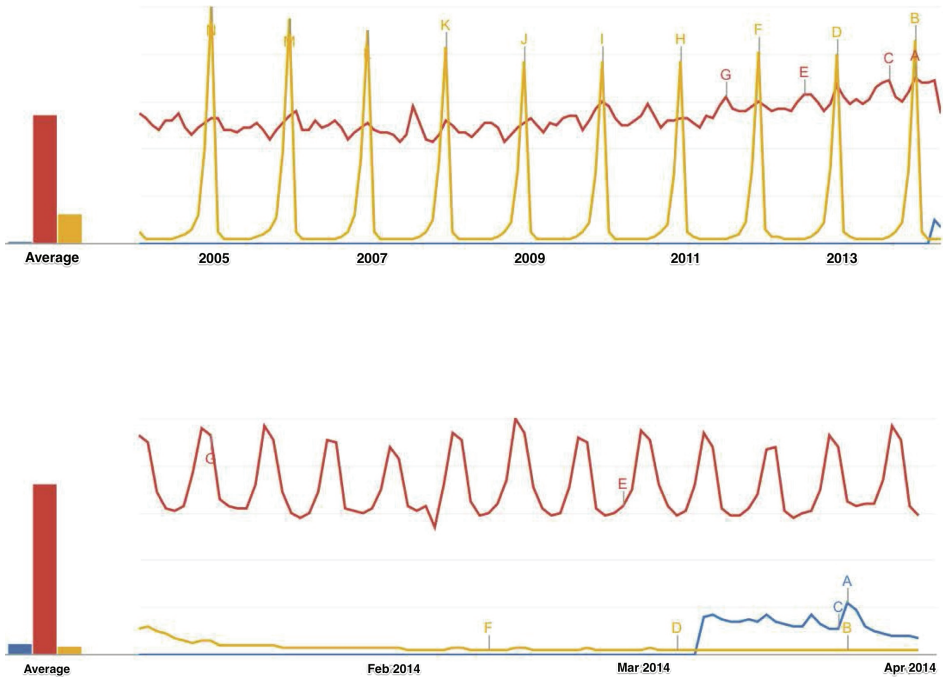
Useful and straightforward approach to rank QAC candidates is to use Maximum Likelihood Estimation (MLE) based on the past popularity of queries

- Assumes that the current query popularity distribution is the same as what was previously observed



Useful and straightforward approach to rank QAC candidates is to use Maximum Likelihood Estimation (MLE) based on the past popularity of queries

- Assumes that the current query popularity distribution is the same as what was previously observed



movie, christmas, MH370

F. Cai, S. Liang, M. de Rijke. Time-sensitive personalized query auto-completion. In CIKM '14, 2014

Observations

- Recency; Whiting et al. (WWW '14)
- Specific temporal intervals; Shokouhi et al. (SIGIR '12)

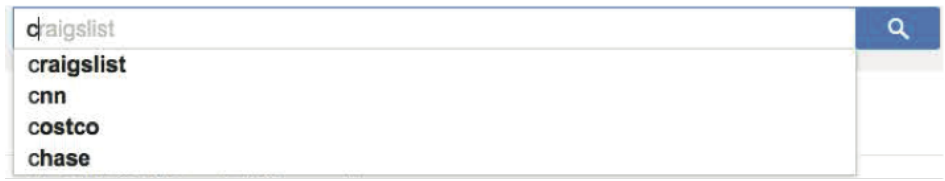
Observations

- Recency; Whiting et al. (WWW '14)
- Specific temporal intervals; Shokouhi et al. (SIGIR '12)

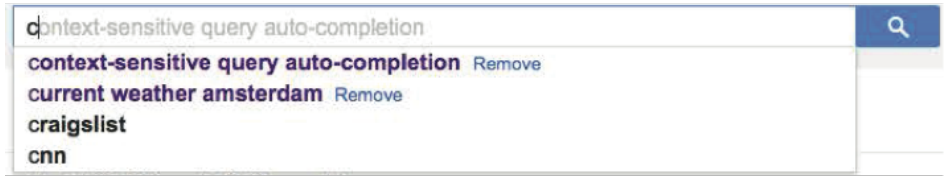
Predictions

- Time-series modeling; Shokouhi et al. (SIGIR '12)
- Regression; Whiting et al. (WWW '14)

Add personalization



QAC of typed prefix c **without logging in.**



QAC of typed prefix c **after logging in.**

Context-aware

- Previous queries; Bar-Yossef et al. (WWW '11)
- Click graph; Cao et al. (KDD '08)

Context-aware

- Previous queries; Bar-Yossef et al. (WWW '11)
- Click graph; Cao et al. (KDD '08)

Learning to personalize

- Demographics + MPC + history; Shokouhi (SIGIR '13)
- Query co-occurrence; Ozertem (SIGIR '12)

Context-aware

- Previous queries; Bar-Yossef et al. (WWW '11)
- Click graph; Cao et al. (KDD '08)

Learning to personalize

- Demographics + MPC + history; Shokouhi (SIGIR '13)
- Query co-occurrence; Ozertem (SIGIR '12)

Personalized + time-sensitive

- Learning to rank based approach; Cai et al. (CIKM '14)

Context-aware

- Previous queries; Bar-Yossef et al. (WWW '11)
- Click graph; Cao et al. (KDD '08)

Learning to personalize

- Demographics + MPC + history; Shokouhi (SIGIR '13)
- Query co-occurrence; Ozertem (SIGIR '12)

Personalized + time-sensitive

- Learning to rank based approach; Cai et al. (CIKM '14)

Trends at SIGIR '15

- Semantics (distributed representations; Mitra, 2015)
- Adaptive models

Once we receive clicks, how can we make sense of it?

Click models

- Probabilistic graphical models of user interaction behavior
- (Chuklin et al., 2015)

Once we receive clicks, how can we make sense of it?

Click models

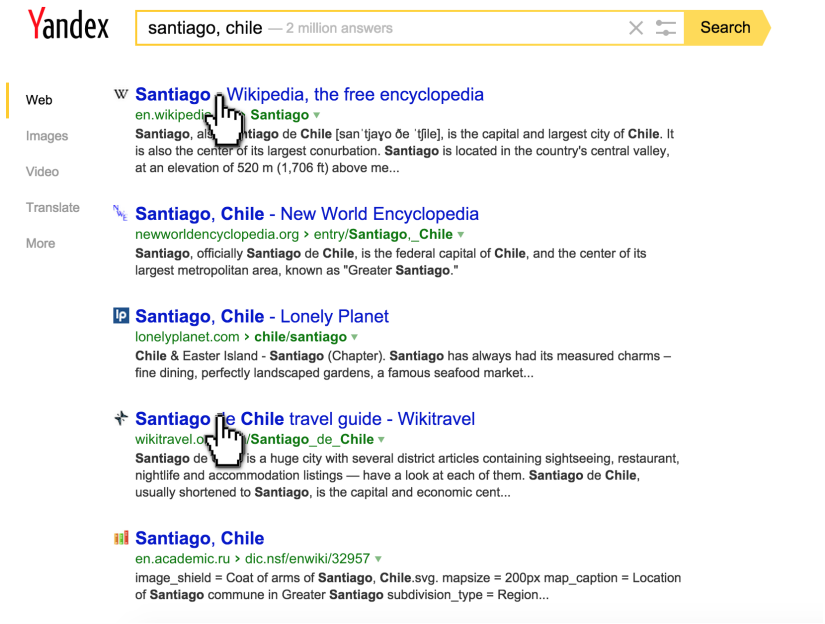
- Probabilistic graphical models of user interaction behavior
- (Chuklin et al., 2015)

Why?

- Understand users
- Simulate users
- Evaluate search
- Improve search

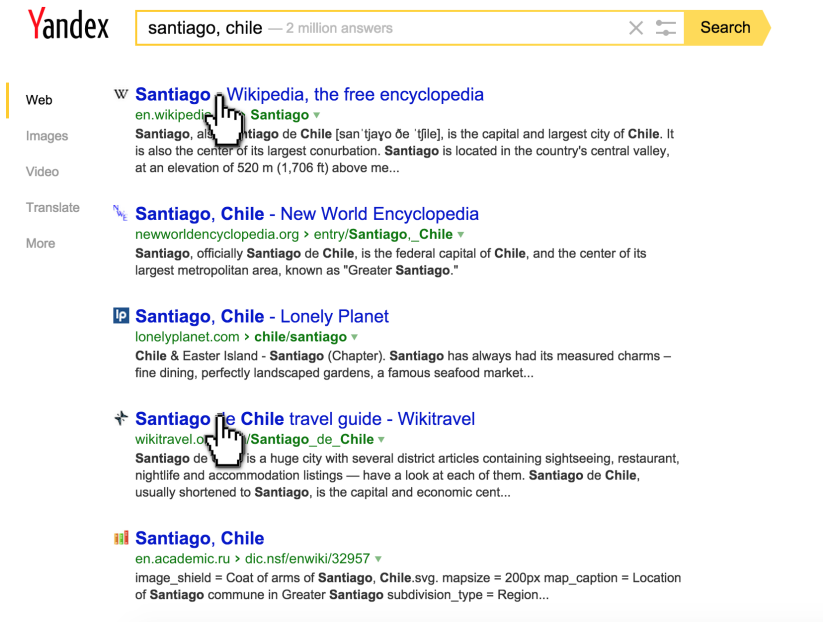
Random click model

?



Random click model

$$P(C_u = 1) = \rho$$



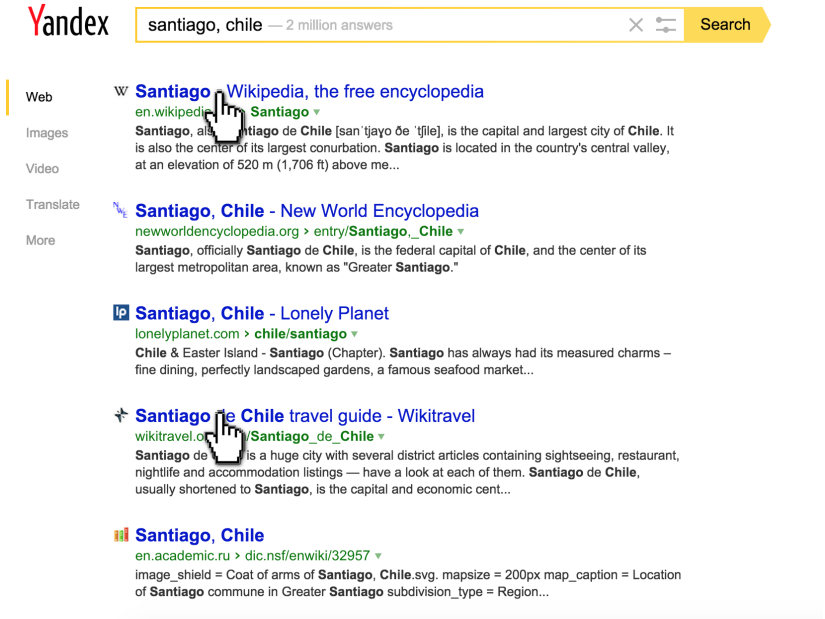
CTR models

Rank-based CTR:

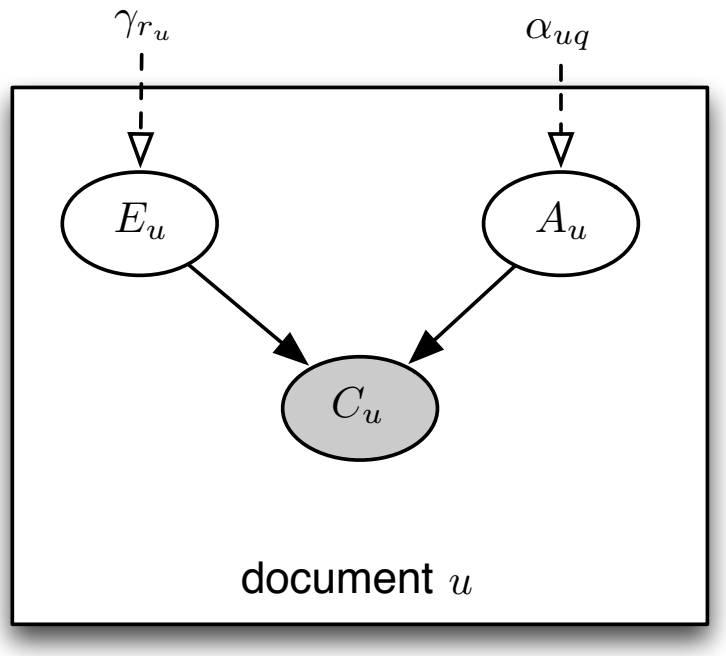
$$P(C_r = 1) = \rho_r$$

Document-based CTR:

$$P(C_u = 1) = \rho_{uq}$$

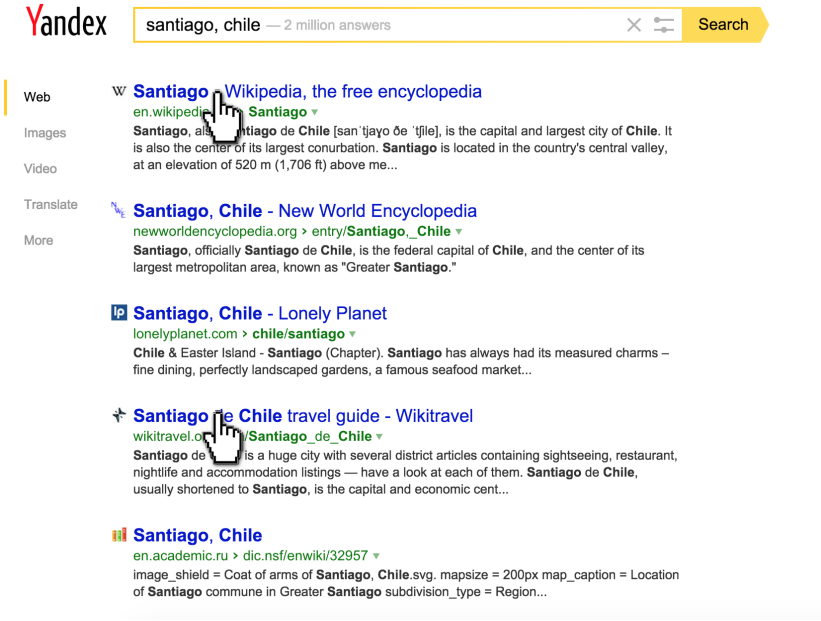


Position-based model

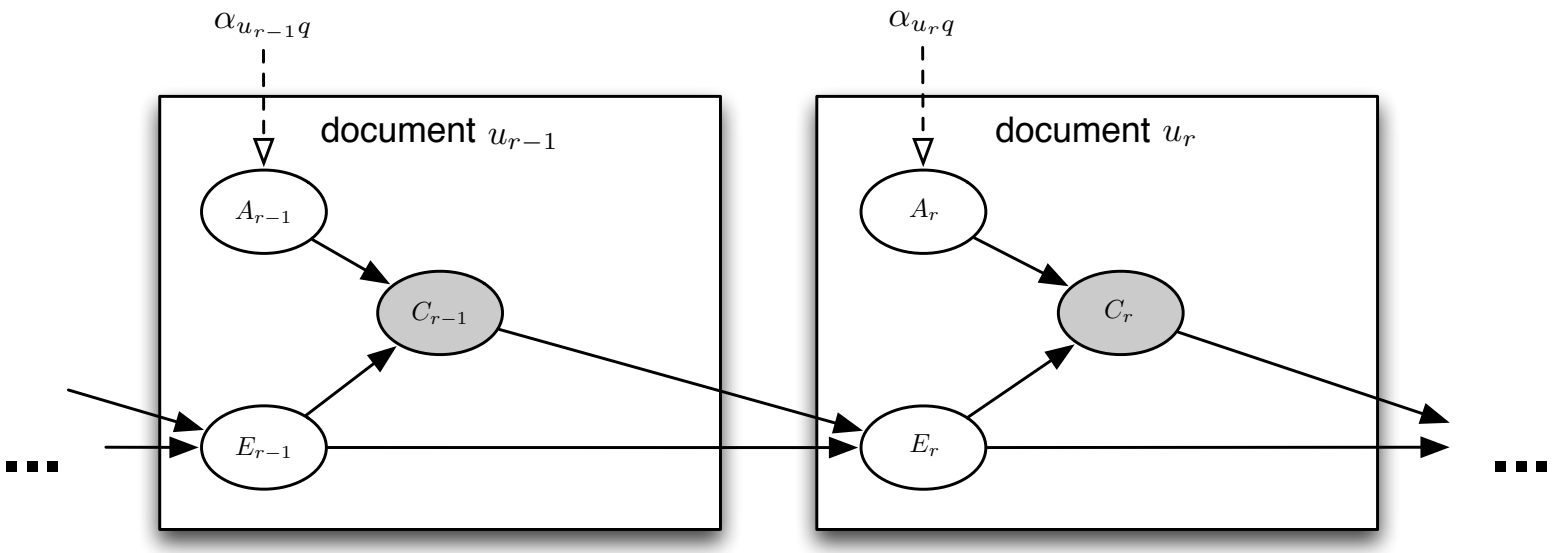


Position-based model

$$P(C_u = 1) = P(E_u = 1) \cdot P(A_u = 1)$$
$$P(A_u = 1) = \alpha_{uq}$$
$$P(E_u = 1) = \gamma_{r_u}$$



Cascade model



$$E_r = 1 \text{ and } A_r = 1 \Leftrightarrow C_r = 1$$

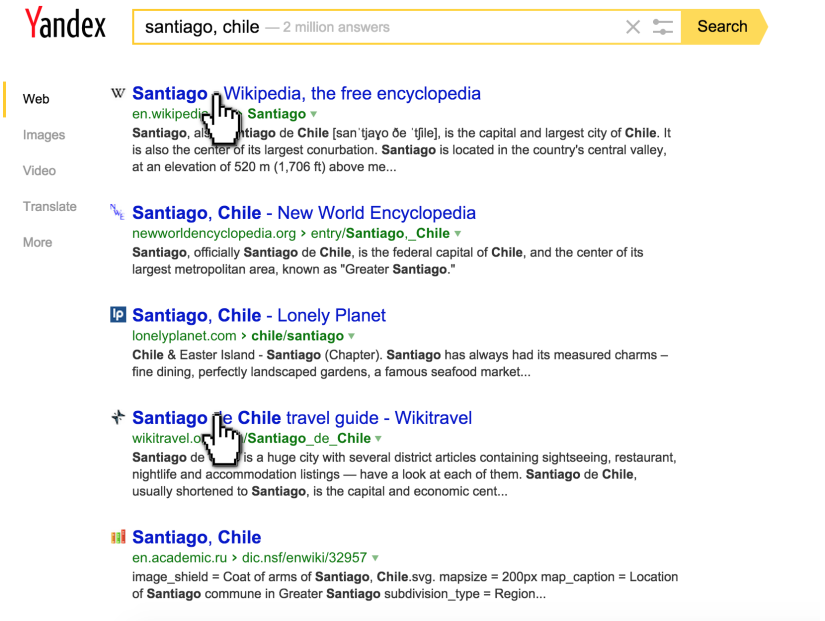
$$P(A_r = 1) = \alpha_{u_r q}$$

$$P(E_1 = 1) = 1$$

$$P(E_r = 1 \mid E_{r-1} = 0) = 0$$

$$P(E_r = 1 \mid C_{r-1} = 1) = 0$$

$$P(E_r = 1 \mid E_{r-1} = 1, C_{r-1} = 0) = 1$$



Query suggestions

"Did you mean?"

- Catching zero result queries
- More popular queries

Query suggestions

"Did you mean?"

- Catching zero result queries
- More popular queries

Most work on query suggestions exploits query logs

- Exploiting consecutive queries during sessions combined with a content-based method using search frequency and query frequency (Zhang et al., 2006)
- Perform random walk on bipartite graph consisting of queries and documents, with transition probabilities derived from the number of clicks between queries and documents: the click graph (Craswell et al., 2007)
- Query-flow graph inferred from reformulation patterns in search sessions, and uses random walk on the graph to obtain suggestions (Boldi et al., 2008)

Query suggestions

"Did you mean?"

- Catching zero result queries
- More popular queries

Most work on query suggestions exploits query logs

- Exploiting consecutive queries during sessions combined with a content-based method using search frequency and query frequency (Zhang et al., 2006)
- Perform random walk on bipartite graph consisting of queries and documents, with transition probabilities derived from the number of clicks between queries and documents: the click graph (Craswell et al., 2007)
- Query-flow graph inferred from reformulation patterns in search sessions, and uses random walk on the graph to obtain suggestions (Boldi et al., 2008)

Most of these developed and tested methods for so-called head queries

Click graph

- Bipartite with two types of nodes: queries and documents
- Edge connects a query and a document if a click for that query-document pair is observed
- The edge may be weighted according to the total number of clicks from all users

Click graph

- Bipartite with two types of nodes: queries and documents
- Edge connects a query and a document if a click for that query-document pair is observed
- The edge may be weighted according to the total number of clicks from all users

Original model uses click data alone, without considering document content or query content

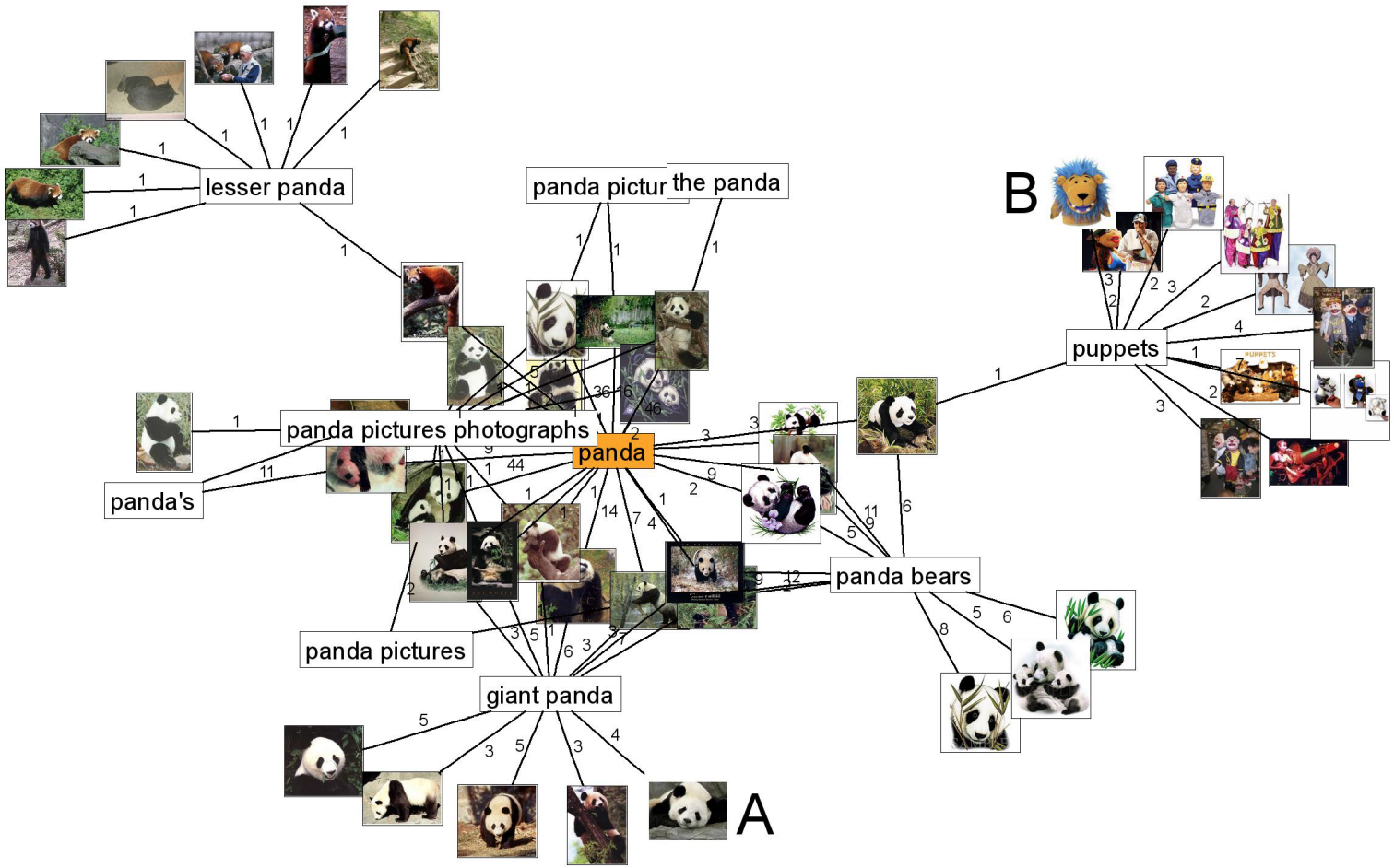
Click graph

- Bipartite with two types of nodes: queries and documents
- Edge connects a query and a document if a click for that query-document pair is observed
- The edge may be weighted according to the total number of clicks from all users

Original model uses click data alone, without considering document content or query content

Use to find query suggestions (given a query as input)

- Random walk on click graph
- Query-document transitions: prefer the most-clicked document for the query
- Document-query transitions: original model treats documents and queries symmetrically so prefers query with the most clicks
  - $+$: model will prefer to follow edges where we have the most evidence of relevance
  - $-$: model will prefer popular queries



Nodes are queries or images, edges indicate clicks. Images A and B are equidistant from the query 'panda' (distance=3), so retrieval based on a naive shortest-path algorithm could not distinguish them. Markov random walk approach sums over paths, so image A benefits from having 7 distinct paths of length 3. Nodes A and 'panda' are connected by a large "volume" of paths.

Image taken from N. Craswell, M. Szummer. Random walks on the click graph. *SIGIR '07*.

Query flow graph is a directed graph

- Nodes are queries
- Arcs are reformulations: non-symmetrical
- Arcs have annotations: frequencies, similarities, etc.

Query flow graph is a directed graph

- Nodes are queries
- Arcs are reformulations: non-symmetrical
- Arcs have annotations: frequencies, similarities, etc.

Perform random walk to identify most probable suggestions

Query flow graph is a directed graph

- Nodes are queries
- Arcs are reformulations: non-symmetrical
- Arcs have annotations: frequencies, similarities, etc.

Perform random walk to identify most probable suggestions

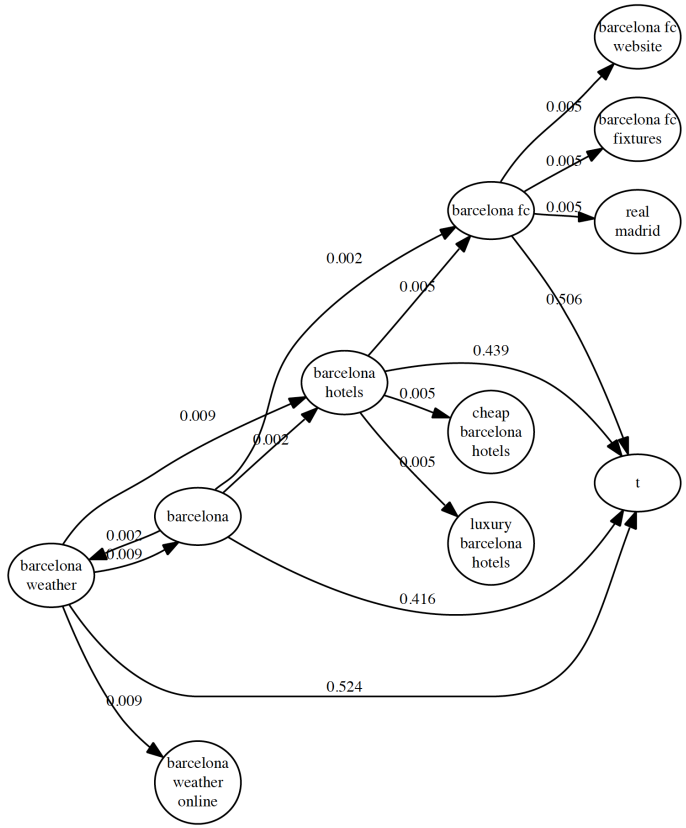


Image taken from P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, S. Vigna. The query-flow graph: model and applications. *CIKM '08*.

banana → apple	banana
banana	banana
apple	eating bugs
usb no	banana holiday
banana cs	opening a banana
giant chocolate bar	banana shoe
where is the seed in anut	fruit banana
banana shoe	recipe 22 feb 08
fruit banana	banana jules oliver
banana cloths	banana cs
eating bugs	banana cloths

beatles → apple	beatles
beatles	beatles
apple	scarring
apple ipod	paul mcartney
scarring	yarns from ireland
srg peppers artwork	statutory instrument A55
ill get you	silver beatles tribute band
bashles	beatles mp3
dundee folk songs	GHOST'S
the beatles love album	ill get you
place lyrics beatles	fugees triger finger remix

Recommendations for the query "apple" considering that the previous query was "banana" (top) or "beatles"? (bottom).

banana → apple	banana
banana	banana
apple	eating bugs
usb no	banana holiday
banana cs	opening a banana
giant chocolate bar	banana shoe
where is the seed in anut	fruit banana
banana shoe	recipe 22 feb 08
fruit banana	banana jules oliver
banana cloths	banana cs
eating bugs	banana cloths

beatles → apple	beatles
beatles	beatles
apple	scarring
apple ipod	paul mcartney
scarring	yarns from ireland
srg peppers artwork	statutory instrument A55
ill get you	silver beatles tribute band
bashles	beatles mp3
dundee folk songs	GHOST'S
the beatles love album	ill get you
place lyrics beatles	fugees triger finger remix

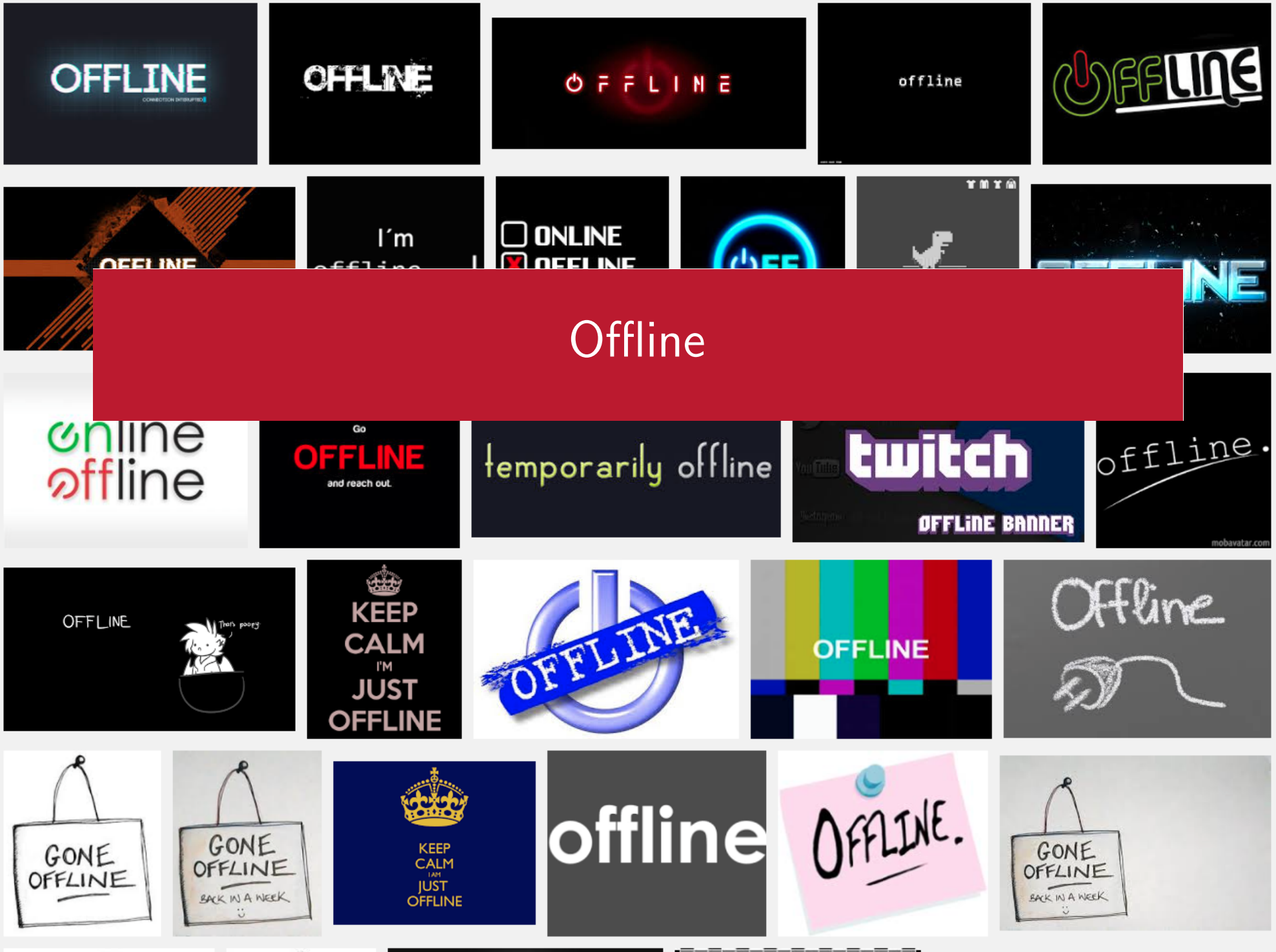
Recommendations for the query "apple" considering that the previous query was "banana" (top) or "beatles"? (bottom).

The query-flow graph also facilitates classifying types of reformulation behavior:

- Parallel moves (50%–60%)
  - *amsterdam* → *berlin*
  - The most frequent class
- Specializations (30%–40%)
  - *amsterdam soccer* → *amsterdam arena*
- Generalizations (5%–10%):
  - *amsterdam hotels* → *amsterdam*
  - Specialization and generalization frequently appear together in alternating order
- Corrections (5%–10%):
  - *masterdam* → *amsterdam*

Image taken from P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, S. Vigna. The query-flow graph: model and applications. *CIKM '08*.

Outline


OFFLINE
OFFLINE
OFFLINE
offline
OFFLINE
ONLINE
OFFLINE
I'm
online
offline
Go
OFFLINE
and reach out
temporarily offline
twitch
OFFLINE BANNER
offline.
mobavatar.com
OFFLINE
KEEP
CALM
I'M
JUST
OFFLINE
OFFLINE
OFFLINE
Offline
GONE
OFFLINE
GONE
OFFLINE
BACK IN A WEEK
KEEP
CALM
I AM
JUST
OFFLINE
offline
OFFLINE.
GONE
OFFLINE
BACK IN A WEEK

Offline

Offline vs. online

In computer science, algorithms that receive their input sequentially operate in an online modality.

- Typical application areas include tasks that involve sequences of decisions, like when you chooses how to serve each incoming query in a stream

Offline vs. online

In computer science, algorithms that receive their input sequentially operate in an online modality.

- Typical application areas include tasks that involve sequences of decisions, like when you chooses how to serve each incoming query in a stream

Batch or offline processing does not need human interaction

- E.g., batch learning proceeds as follows:
  - Initialize the weights
  - Repeat the following steps:
    - Process all the training data
    - Update the weights

Offline vs. online

In computer science, algorithms that receive their input sequentially operate in an online modality.

- Typical application areas include tasks that involve sequences of decisions, like when you chooses how to serve each incoming query in a stream

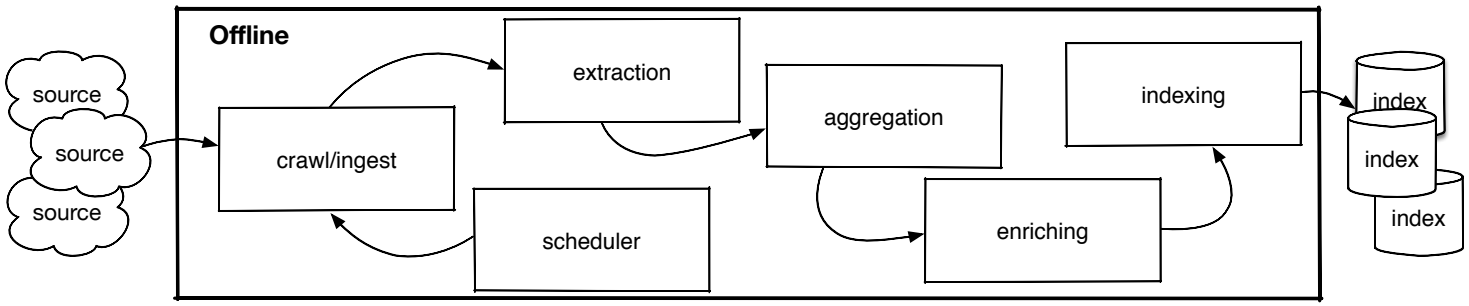
Batch or offline processing does not need human interaction

- E.g., batch learning proceeds as follows:
  - Initialize the weights
  - Repeat the following steps:
    - Process all the training data
    - Update the weights

Typical offline computations in information retrieval

- Any processing that is not query dependent (crawling, document enriching, aggregation, indexing, . . .)

Offline



Getting content

Document classification

Duplicate detection

Document enrichment

Content aggregation

Indexing

Getting content: many scenarios

- Desktop search
  - Recursive descent on file system
- Search on your phone
  - Recursive descent on file system (if battery permits?)
- Library search
  - Nightly ingestion
- Enterprise search
  - Nightly ingestion
- Twitter search
  - Near real-time availability
- Web search
  - Getting the content of the documents takes longer
  - Operate at variable speeds, with different priorities...

Crawling

- Initialize queue with URLs of known seed pages
- Repeat
  - Take URL from queue
  - Fetch and parse page
  - Extract URLs from page
  - Add URLs to queue

Fundamental assumption: The web is well linked.

Crawling

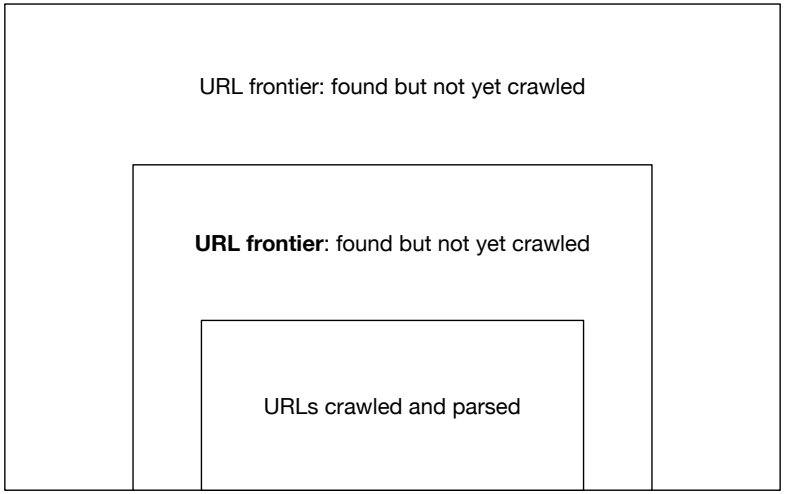
- Initialize queue with URLs of known seed pages
- Repeat
  - Take URL from queue
  - Fetch and parse page
  - Extract URLs from page
  - Add URLs to queue

Fundamental assumption: The web is well linked.

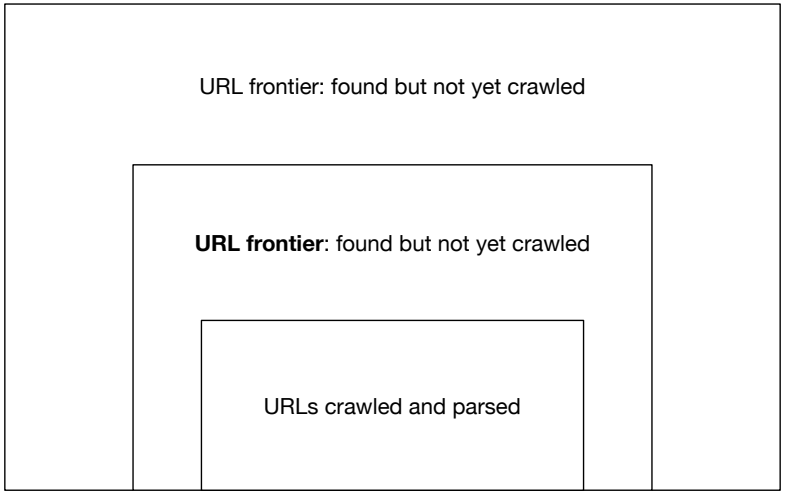
What's wrong with this simple crawler?

- Scale: we need to distribute
- We cannot index everything: we need to select. How?
- Duplicates: need to integrate duplicate detection
- Spam and spider traps: need to integrate spam detection
- Politeness: we need to be ?nice? and space out all requests for a site over a longer period (hours, days)
- Freshness: we need to re-crawl periodically
- Prioritize highly frequent re-crawls only for a small subset, frequent re-crawls for . . .

URL frontier



URL frontier



Data structure that holds and manages

- URLs we have seen, but which have not been crawled yet
- Can include multiple pages from same host
- Avoid trying to fetch them all at the same time
- Keep all crawling threads busy

Basic crawling architecture

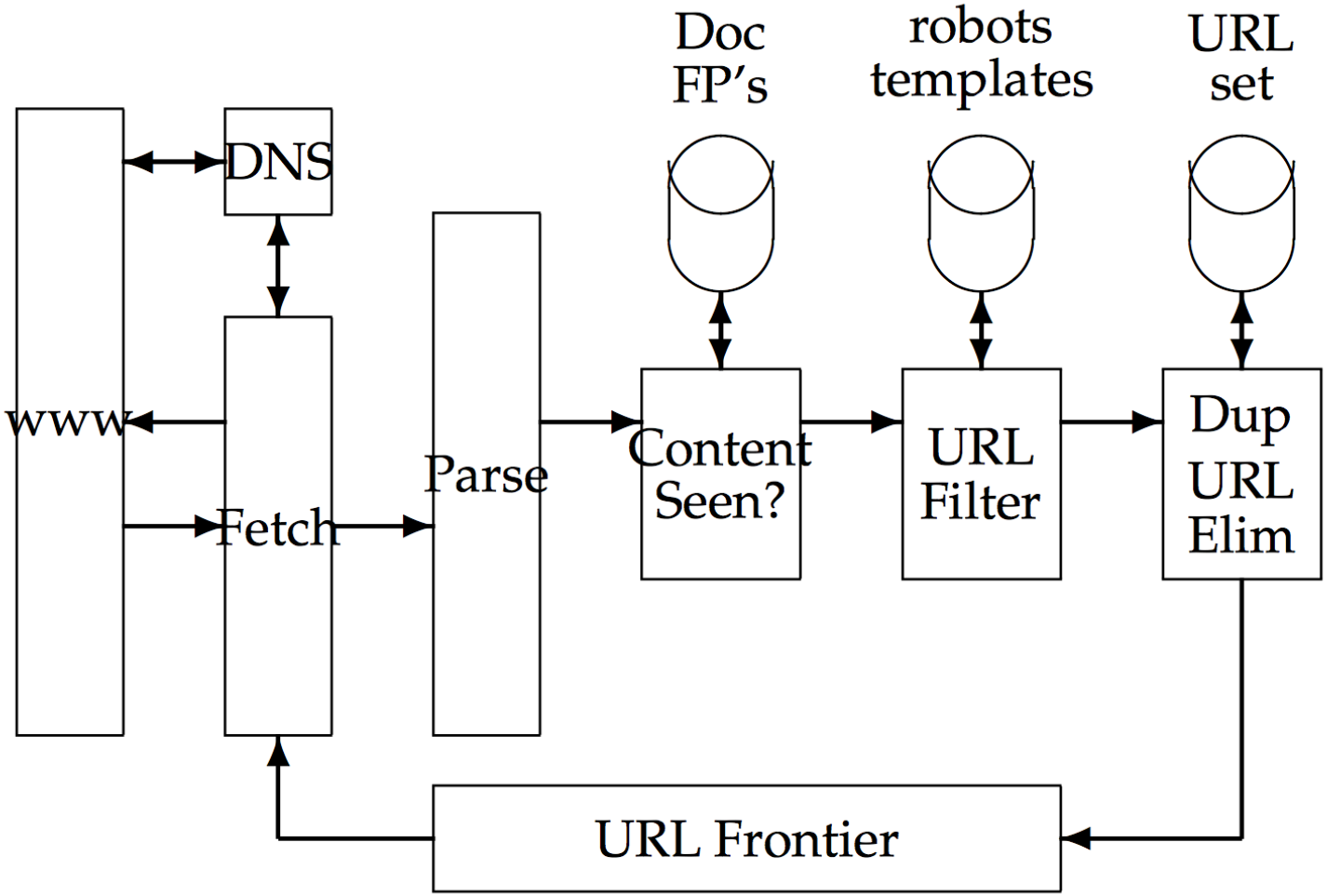


Image taken from C. Manning, P. Raghavan, H. Shütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008

Duplicate detection

Duplicates occur in most collections of reasonable size

- But 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 is annoying to get a search result with near-identical documents
- Marginal relevance is zero: even a highly relevant document becomes non-relevant if it appears below a (near-)duplicate

Compute similarity with an edit-distance measure

We want "syntactic" (as opposed to semantic) similarity.

We do not consider documents near-duplicates if they have the same content, but express it with different words.

Use similarity threshold θ to decide "is/is not a near-duplicate"

- E.g., two documents are near-duplicates if similarity $> \theta = 80\%$.

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:

$$\{\text{a-rose-is, rose-is-a, is-a-rose}\}$$

We can map shingles to $1..2^m$ (e.g., $m = 64$) by fingerprinting.

Define the similarity of two documents as the Jaccard coefficient of their shingle sets.

For efficiency, define sketches (well chosen subsets of shingles) and compute Jaccard coefficient for two sketches

- Index only document per equivalence class of similar documents

Spam detection

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

Exercise: How can I get my page ranked highly? ("Search engine optimization")

Spam detection

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

Exercise: How can I get my page ranked highly? ("Search engine optimization")

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

Spam technique: Duplication

- Get good content from somewhere (steal it or produce it yourself)
- Publish a large number of slight variations of it
- And include profitable links to ads

Spam technique: Duplication

- Get good content from somewhere (steal it or produce it yourself)
- Publish a large number of slight variations of it
- And include profitable links to ads

Spam technique: Cloaking

- Serve fake content to search engine crawler
- So do we just penalize this always?
- No: legitimate uses (e.g., different content to US vs. European users)

Spam technique: Duplication

- Get good content from somewhere (steal it or produce it yourself)
- Publish a large number of slight variations of it
- And include profitable links to ads

Spam technique: Cloaking

- Serve fake content to search engine crawler
- So do we just penalize this always?
- No: legitimate uses (e.g., different content to US vs. European users)

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; Pay somebody to put your link on their highly ranked page; Leave comments that include the link on blogs

Aggregation

Gather content that appears to belong together

- Anchor text on the web
  - Anchor text is often a better description of a page?s content than the page itself
  - Anchor text can be weighted more highly than the text on the page
- Information around an entity (person, organization, location, cultural artefact, . . .)
  - Large portion of queries are entity oriented
  - Information about "tail entities" is initially sparse (by definition) but may explode when entity hits the news
    - E.g., MH370, Ferguson, . . .
  - Aggregate content from news, wikipedia, social, Twitter, . . .
  - Spam, short-term interest, long-term interest

Inverted index construction

1. Collect the documents to be indexed:
   Friends, Romans, countrymen. | So let it be with Caesar | ...
2. Tokenize the text, turning each document into a list of tokens:
   Friends | Romans | countrymen | So | ...
3. Do linguistic preprocessing, producing a list of normalized tokens, which are the indexing terms: friend | roman | countryman | so | ...
4. Index the documents that each term occurs in by creating an inverted index, consisting of a dictionary and postings.

Material from C. Manning, P. Raghavan, H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008

Tokenization and preprocessing

Doc 1. I did enact Julius Caesar: I was killed i' the Capitol; Brutus killed me.
Doc 2. So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious:

$\Rightarrow$

Doc 1. i did enact julius caesar i was killed i' the capitol brutus killed me
Doc 2. so let it be with caesar the noble brutus hath told you caesar was ambitious

Material from C. Manning, P. Raghavan, H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008

Generate postings

Doc 1. i did enact julius caesar i was killed i' the capitol brutus killed me
Doc 2. so let it be with caesar the noble brutus hath told you caesar was ambitious

$\Longrightarrow$

term	docID
i	1
did	1
enact	1
julius	1
caesar	1
i	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

Material from C. Manning, P. Raghavan, H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008

Sort postings

term	docID
i	1
did	1
enact	1
julius	1
caesar	1
i	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

$\Longrightarrow$

term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
i	1
i	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2

Material from C. Manning, P. Raghavan, H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008

Create postings lists, determine document frequency

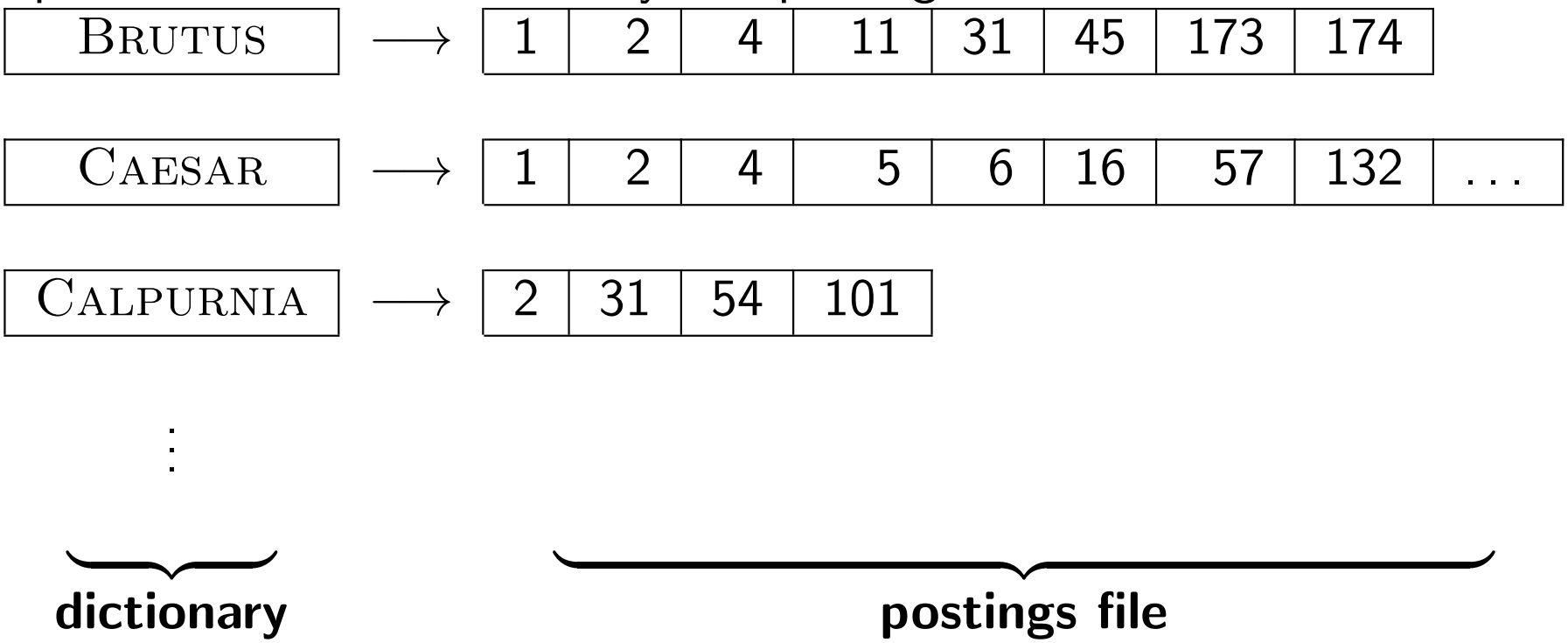
term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
i	1
i	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2

⟹

term	doc. freq.	→	postings lists
ambitious	1	→	2
be	1	→	2
brutus	2	→	1 → 2
capitol	1	→	1
caesar	2	→	1 → 2
did	1	→	1
enact	1	→	1
hath	1	→	2
i	1	→	1
i'	1	→	1
it	1	→	2
julius	1	→	1
killed	1	→	1
let	1	→	2
me	1	→	1
noble	1	→	2
so	1	→	2
the	2	→	1 → 2
told	1	→	2
you	1	→	2
was	2	→	1 → 2
with	1	→	2

Material from C. Manning, P. Raghavan, H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008

Split the result into dictionary and postings file



Material from C. Manning, P. Raghavan, H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008

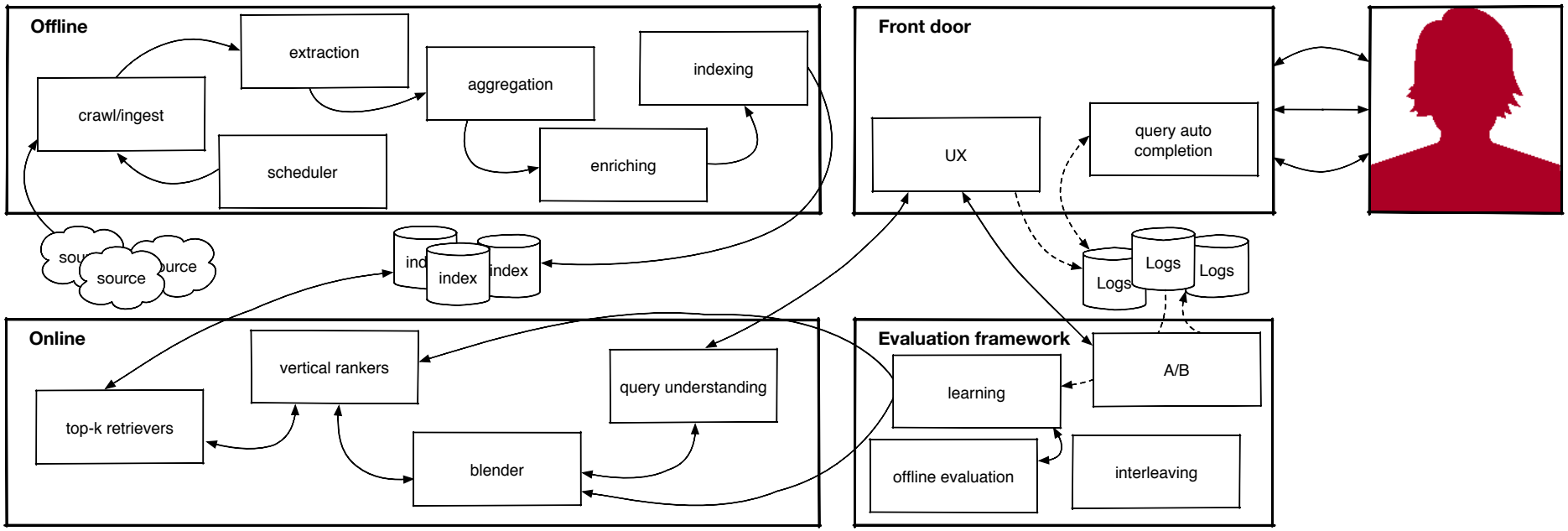
Sort postings in memory (if infrastructure allows)

External sorting for disk-based set-ups

Distributed indexing for very large collections (MapReduce)

For dynamic collections, maintain main index on disk and separate auxiliary index in memory, search across both and merge results. Periodically merge auxiliary index into main index

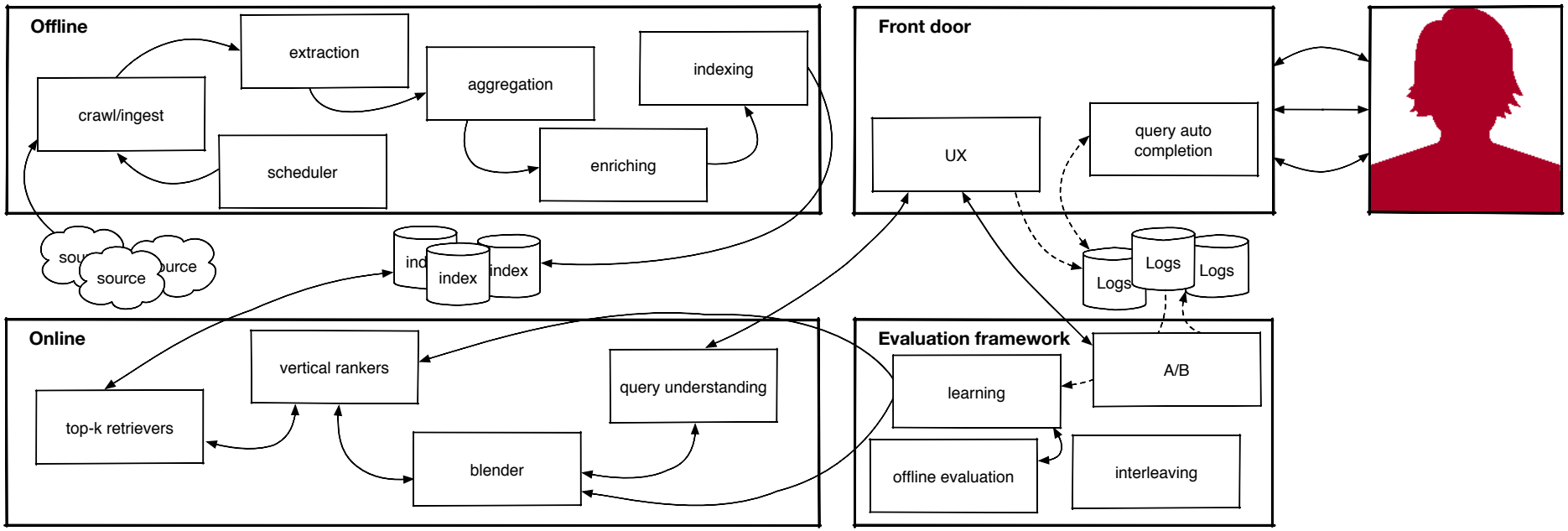
The big picture



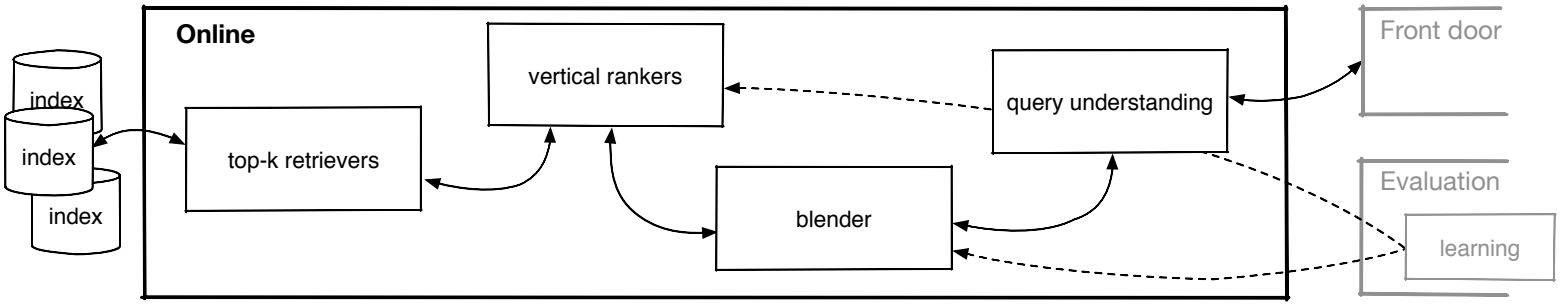
Break

Outline

The big picture



Online



Query understanding

Blending

Vertical rankers

Top-k retrieval

Query understanding

Down

- Alterations
  - Confident suggestions, structured query generation, hotfixes (rules), advanced search syntax, query translation, qa understanding, stopword handling (term weighting)
- Classifiers
  - **Query intent**, topic, directly answerable, query performance prediction, language classifier, task detection, device detection
- Annotators
  - User modeling, localization, session, conversation?, entity extractors
- Aggregators
  - Query stats, tail/head

Query understanding

Down

- Alterations
  - Confident suggestions, structured query generation, hotfixes (rules), advanced search syntax, query translation, qa understanding, stopword handling (term weighting)
- Classifiers
  - **Query intent**, topic, directly answerable, query performance prediction, language classifier, task detection, device detection
- Annotators
  - User modeling, localization, session, conversation?, entity extractors
- Aggregators
  - Query stats, tail/head

Up

- SERP generation
  - Snippet generation, device tailoring, translation, answer insertion

Intent

With the help of intent identification, search engines can perform intent-aware result ranking, or provide accurate results for specific types of query

- If an image-oriented search intent is identified, invoke image search module so as to show a few image results along with general web results
  - Thessaloniki vs Thessaloniki image

Intent

With the help of intent identification, search engines can perform intent-aware result ranking, or provide accurate results for specific types of query

- If an image-oriented search intent is identified, invoke image search module so as to show a few image results along with general web results
  - Thessaloniki vs Thessaloniki image

Two main lines of work on intent identification, according to whether the intent label is predefined or not

- Intent classification
- Intent discovery

Many flavors of intent classification

- Search goal
  - Navigational, informational, transactional
- Search task
  - "purchase computer", "job-finding query"
- Semantic topics
  - "Cars", "NBA" (DMOZ, ODP, Wikipedia)
- Vertical-oriented intents
  - Image, video, apps, ...
- Time-sensitivity
  - News-sensitive queries
- Location-sensitivity
  - "coffee"

Classifying queries into pre-defined intent classes is challenging since queries are short and ambiguous

Click-through data, session data, and search result data are widely used for the query classification tasks?

Generally, hand-crafted training data and hand-crafted intent inventory

Classifying queries into pre-defined intent classes is challenging since queries are short and ambiguous

Click-through data, session data, and search result data are widely used for the query classification tasks?

Generally, hand-crafted training data and hand-crafted intent inventory

All examples are taken from actual AltaVista queries.

SEARCH GOAL	DESCRIPTION	EXAMPLES
1. Navigational	My goal is to go to specific known website that I already have in mind. The only reason I'm searching is that it's more convenient than typing the URL, or perhaps I don't know the URL.	aloha airlines duke university hospital kelly blue book
2. Informational	My goal is to learn something by reading or viewing web pages	
2.1 Directed	I want to learn something in particular about my topic	
2.1.1 Closed	I want to get an answer to a question that has a single, unambiguous answer.	what is a supercharger 2004 election dates
2.1.2 Open	I want to get an answer to an open-ended question, or one with unconstrained depth.	baseball death and injury why are metals shiny
2.2 Undirected	I want to learn anything/everything about my topic. A query for topic X might be interpreted as "tell me about X."	color blindness jfk jr
2.3 Advice	I want to get advice, ideas, suggestions, or instructions.	help quitting smoking walking with weights
2.4 Locate	My goal is to find out whether/where some real world service or product can be obtained	pella windows phone card
2.5 List	My goal is to get a list of plausible suggested web sites (I.e. the search result list itself), each of which might be candidates for helping me achieve some underlying, unspecified goal	travel amsterdam universities florida newspapers
3. Resource	My goal is to obtain a resource (not information) available on web pages	
3.1 Download	My goal is to download a resource that must be on my computer or other device to be useful	kazaa lite mame roms
3.2 Entertainment	My goal is to be entertained simply by viewing items available on the result page	xxx porno movie free live camera in l.a.
3.3 Interact	My goal is to interact with a resource using another program/service available on the web site I find	weather measure converter
3.4 Obtain	My goal is to obtain a resource that does not require a computer to use. I may print it out, but I can also just look at it on the screen. I'm not obtaining it to learn some information, but because I want to use the resource itself.	free jack o lantern patterns ellis island lesson plans house document no. 587

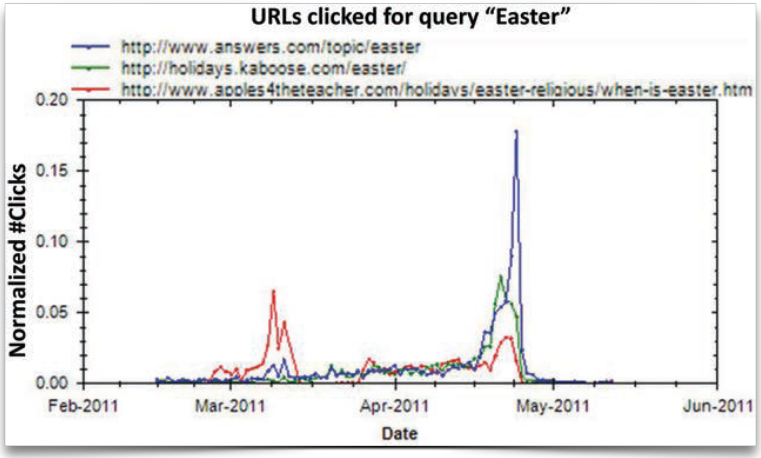
Image taken from D.E. Rose and D. Levinson. Understanding User Goals in Web Search. *WWW '04*, 2004

Intent discovery

- Another viewpoint of intent, not dependent on pre-defined intent categories
- Users with similar information needs click the same group of URLs, even though queries issued may vary
  - Query or URL clusters express highly similar information needs or intent.
  - Click-through bipartite graph often used in query clustering studies
  - A large fraction of queries follow some templates in most examined domains
    - Intent detection $\sim$ a problem of template (or structure) detection among queries
    - Queries that fall into the same or synonymous templates are regarded as having the same intent
  - Alternatively, detect different intents of an ambiguous query through query refinements queries or the clicked URLs
- Intent is often assumed to be static
  - but see examples to come
- Intent is often assumed to be binary (yes or no) for a small number of intents
  - but see challenge to come

Shifting intents

- Radinsky et al. (2013)
  - When users? information needs change over time, ranking of results should also change to accommodate these needs
- Query "easter" at different times during year
  - Few weeks prior: *When?*
  - Few days prior: *What to do?*
  - During: *Meaning of easter*



Learning to detect intent shifts (Lefortier et al., 2014)

- Queries whose intent shifts from non-fresh to fresh
- Aggregated search approach to freshness
  - A "fresh" vertical
  - Fresh intent detector (MSE $\sim$0.025)
- Intents may shift from non-fresh to fresh
  - $\sim$7% of queries display a shift
  - Fresh intent detector needs time to catch up
  - On average 7.9h on a sample
- Can we do better?
  - Without throwing the fresh intent detector away

Learning to detect intent shifts (Lefortier et al., 2014)

- Queries whose intent shifts from non-fresh to fresh
- Aggregated search approach to freshness
  - A "fresh" vertical
  - Fresh intent detector (MSE ∼0.025)
- Intents may shift from non-fresh to fresh
  - ∼7% of queries display a shift
  - Fresh intent detector needs time to catch up
  - On average 7.9h on a sample
- Can we do better?
  - Without throwing the fresh intent detector away
- Online exploration for quick adaptation
- Multi-armed bandits
  - Consider SERP as an action and consider only actions that integrate fresh results on SERP differently
  - Each action corresponds to deciding how many fresh results to integrate on SERP and where
- ExploreOnTop
  - Integrate one fresh result at the top of the SERP at the first position and gather user feedback and then re-estimate freshness
  - Reduce time delay by 57%, positive impact on 74% of SERPs, on average just 11 impressions of each selected query needed

Blender

Down (ranker parameterization)

- Ranker type selection
  - web, fresh, news, image, video, entity, apps
- Ranker parameter selection
  - production
  - for interleaving
  - for A/B testing
- Direct answers
  - Maps, facts, weather, qa...

Blender

Down (ranker parameterization)

- Ranker type selection
  - web, fresh, news, image, video, entity, apps
- Ranker parameter selection
  - production
  - for interleaving
  - for A/B testing
- Direct answers
  - Maps, facts, weather, qa...

Up (SERP generation)

- Merging results
  - Interleaving
  - Diversity
- UX selection
  - For A/B testing
  - Production
  - KB panel (entity enrichment)
- Device tailoring

Diversity (Radlinski et al., 2009)

- Extrinsic diversity: diversity as uncertainty about the information need
  - Ambiguity: "jaguar"
  - Different aspects: "ebola"

Diversity (Radlinski et al., 2009)

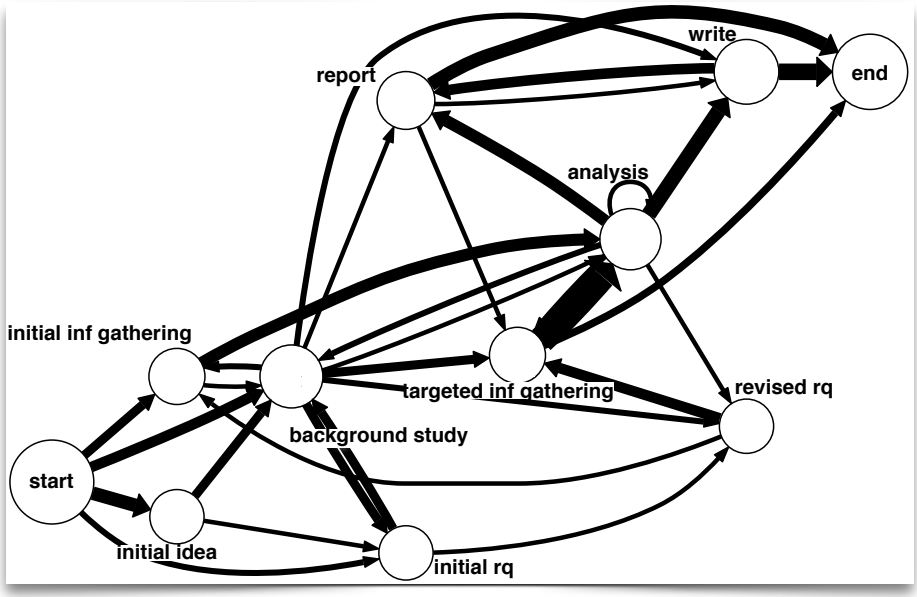
- Extrinsic diversity: diversity as uncertainty about the information need
  - Ambiguity: "jaguar"
  - Different aspects: "ebola"

- Intrinsic diversity: diversity as part of the information need
  - No single result that provides fully answers information need
  - User desires different views
  - User desires different options
  - Information need is to get an overview
  - Different results are needed from different sources to build confidence in correctness of answer

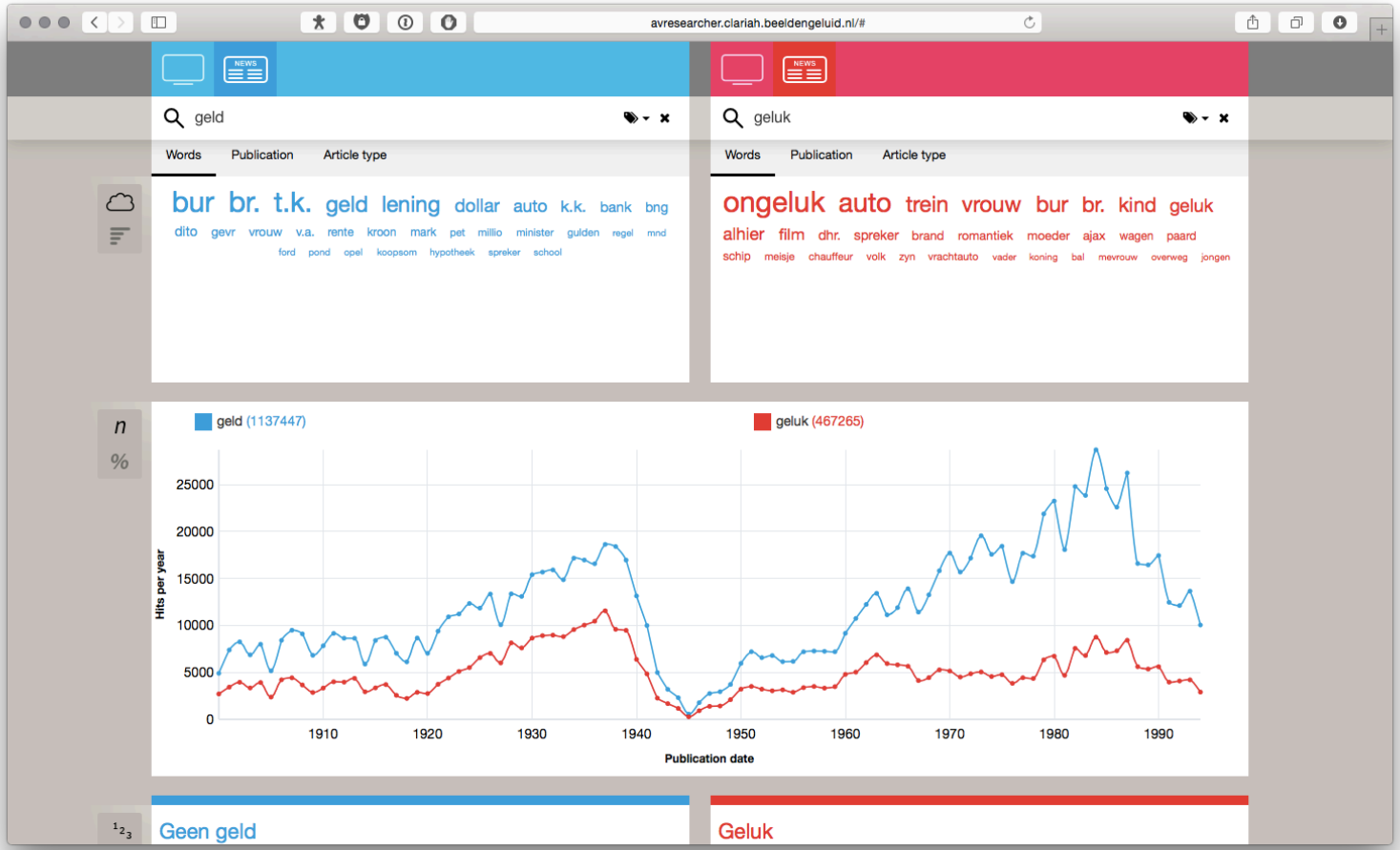
code (abbreviation)	description
initial idea (ii)	An idea, observation, or proposal that starts a project.
background study (bg)	Identify literature and background material for a topic.
initial research questions (ir)	Identify research questions or instruments, e.g., sampling.
initial data gathering (ig)	Initial search, exploration, or collection of data.
revised research questions (rr)	Revision of research questions and instruments.
targeted data gathering (tg)	Collect, search, or select data following guidelines.
analysis (an)	Inspect, read, code, compare, or organize data.
write (wr)	Write, select examples, drawing of conclusions.
report (rp)	Integrate findings into articles, chapter, or presentation.

Diversion (Bron et al., 2012, 2013)

- Study the search behavior of media studies scholars



Supports manual intrinsic diversity by offering a a subjunctive interface through which researchers can compare alternative queries ("successor queries") around a topic of their interest ("initiator queries")?



Support for manual intrinsic diversity search?

Research questions may undergo changes during a research project:

- questions become more specific;
- additional questions are added;
- or a changed perspective in the research question?

Reasons for the changes in research questions: researchers learn about the availability of material, discover trends in the material or gain alternative views on a topic

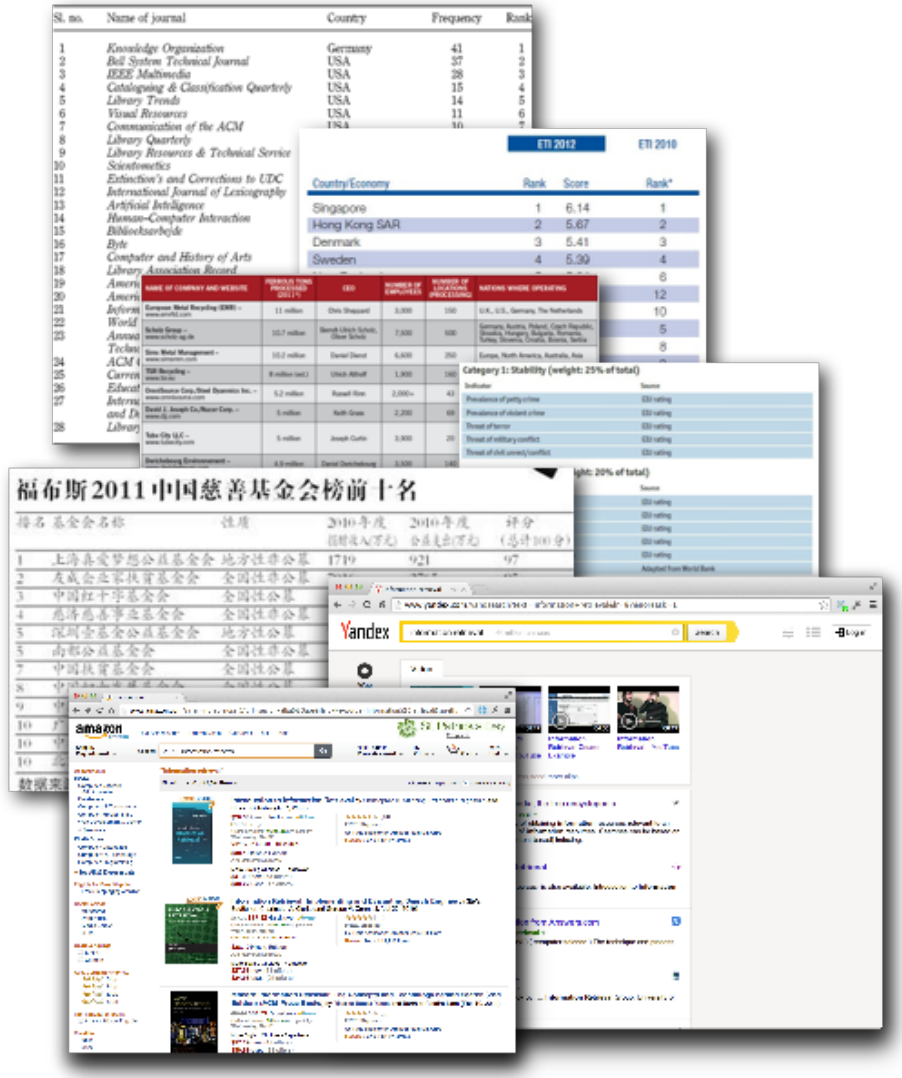
Vertical rankers

Ranker

- hotfixes
  - personalized
- compute query document features
  - geo spatial
  - bm25
  - ...
- apply ranking model to
  - query features
  - document features
  - query document features

So many criteria

- Aboutness
- Potential impact on reputation
- Importance
- Timeliness
- Quality
- Bias
- Fit with task/background
- Freshness
- Interestingness
- ...



Ranker development

- Traditionally, manual labor
- Think about what it means for a document to match a query
- Combination of term frequency, document frequency, document length E.g.,

$$BM25(q,d) = \sum_{q_i : tf(q_i,d) > 0} idf(q_i) \cdot \frac{tf(q_i,d) \cdot (k_1 + 1)}{tf(q_i,d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{avgdl})} \cdot \frac{(k_3 + 1) \cdot qtf(q_i,q)}{k_3 + qtf(q_i,q)}$$

So many rankers . . .

- Content-based
  - Boolean model, extended Boolean model, . . .
  - Vector space model, latent semantic indexing, . . .
  - BM25 model, statistical language model, . . .
  - Span-based model, distance-aggregation model, . . .
- Structure-based
  - Document structure
  - Site structure
  - Link structure
- Based on interaction behavior
  - Number of visits, . . .
  - Clicks, . . .

So many rankers ...

- Content-based
  - Boolean model, extended Boolean model, ...
  - Vector space model, latent semantic indexing, ...
  - BM25 model, statistical language model, ...
  - Span-based model, distance-aggregation model, ...
- Structure-based
  - Document structure
  - Site structure
  - Link structure
- Based on interaction behavior
  - Number of visits, ...
  - Clicks, ...
- $\Rightarrow$ Documents represented by feature vectors
  - Features extracted for every query-document pair (e.g., score output by a traditional retrieval model)
  - Combine a large number of features
  - Incorporate new retrieval model by including the model's output

- Least square retrieval function (TOIS 1989)
- Query refinement (WWW 2008)
- ListNet (ICML 2007)
- SVM-MAP (SIGIR 2007)
- Nested Ranker (SIGIR 2006)
- Pranking (NIPS 2002)
- LambdaRank (NIPS 2006)
- MPRank (ICML 2007)
- Frank (SIGIR 2007)
- MHR (SIGIR 2007)
- RankBoost (JMLR 2003)
- Learning to retrieval info (SCC 1995)
- LDM (SIGIR 2005)
- Large margin ranker (NIPS 2002)
- RankNet (ICML 2005)
- Ranking SVM (ICANN 1999)
- IRSVM (SIGIR 2006)
- Discriminative model for IR (SIGIR 2004)
- ...

- SVM Structure (JMLR 2005)
- OAP-BPM (ICML 2003)
- Subset Ranking (COLT 2006)
- GPRank (LR4IR 2007)
- QBRank (NIPS 2007)
- GBRank (SIGIR 2007)
- Constraint Ordinal Regression (ICML 2005)
- McRank (NIPS 2007)
- SoftRank (LR4IR 2007)
- AdaRank (SIGIR 2007)
- CCA (SIGIR 2007)
- ListMLE (ICML 2008)
- RankCosine (IPM 2007)
- Supervised Rank Aggregation (WWW 2007)
- Relational ranking (WWW 2008)
- Learning to order things (NIPS 1998)
- Round robin ranking (ECML 2003)
- ...

Category	Algorithms
Pointwise approach	**Regression-based**: Least square retrieval (TOIS 1989), Regression tree for ordinal class prediction (FI 2000), … **Classification**: Discriminative model for IR (SIGIR 2004), … **Ordinal regression**: Pranking (NIPS 2002), OAP-BPM (ECML 2003), Ranking with large margin principles (NIPS 2002), …
Pairwise approach	Learning to retrieve information (SCC 1995), Learning to order things (NIPS 1998), Ranking SVM (ICANN 1999), Rankboost (JMLR 2003), LDM (SIGIR 2005), RankNet (ICML 2005), Frank (SIGIR 2007), MHR (SIGIR 2007), GBRank (SIGIR 2007), QBRank (NIPS 2007), MPRank (ICML 2007), …
Listwise approach	**Non-measure specific**: ListNet (ICML 2007), ListMLE (ICML 2008), BoltzRank (ICML 2009), … **Measure-specific**: AdaRank (SIGIR 2007), SVM-MAP (SIGIR 2007), SoftRank (LR4IR 2007), RankGP (LR4IR), …

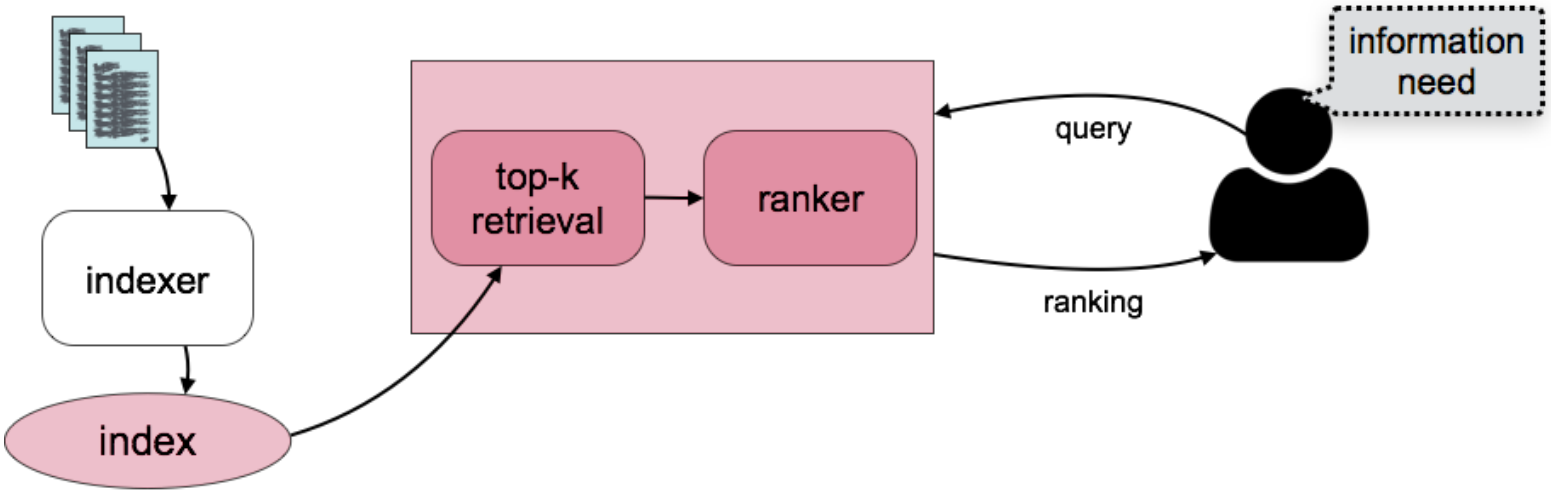
Ranker development

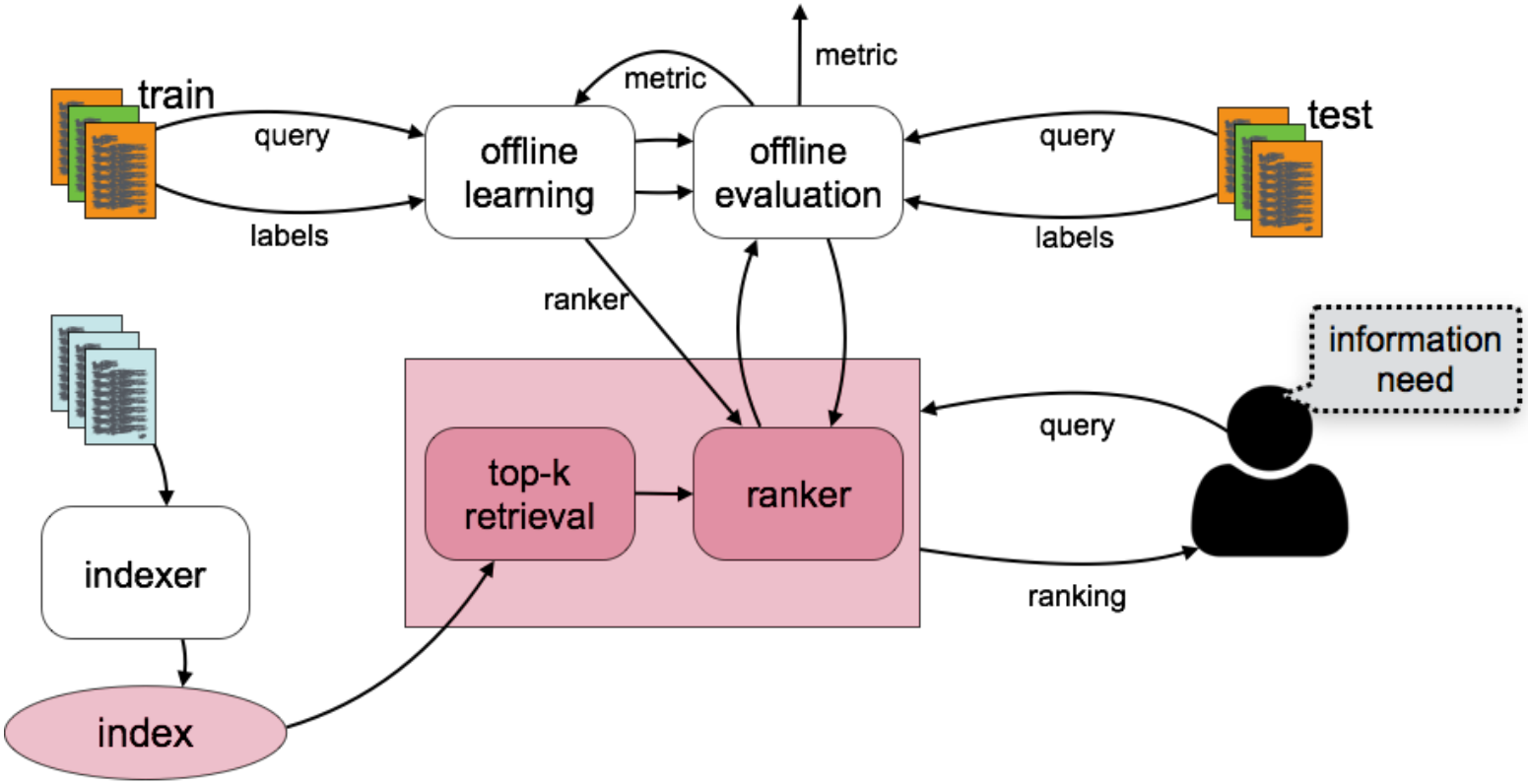
- Traditionally
  - Train and tune offline, then deploy online
  - Supervised learning paradigm

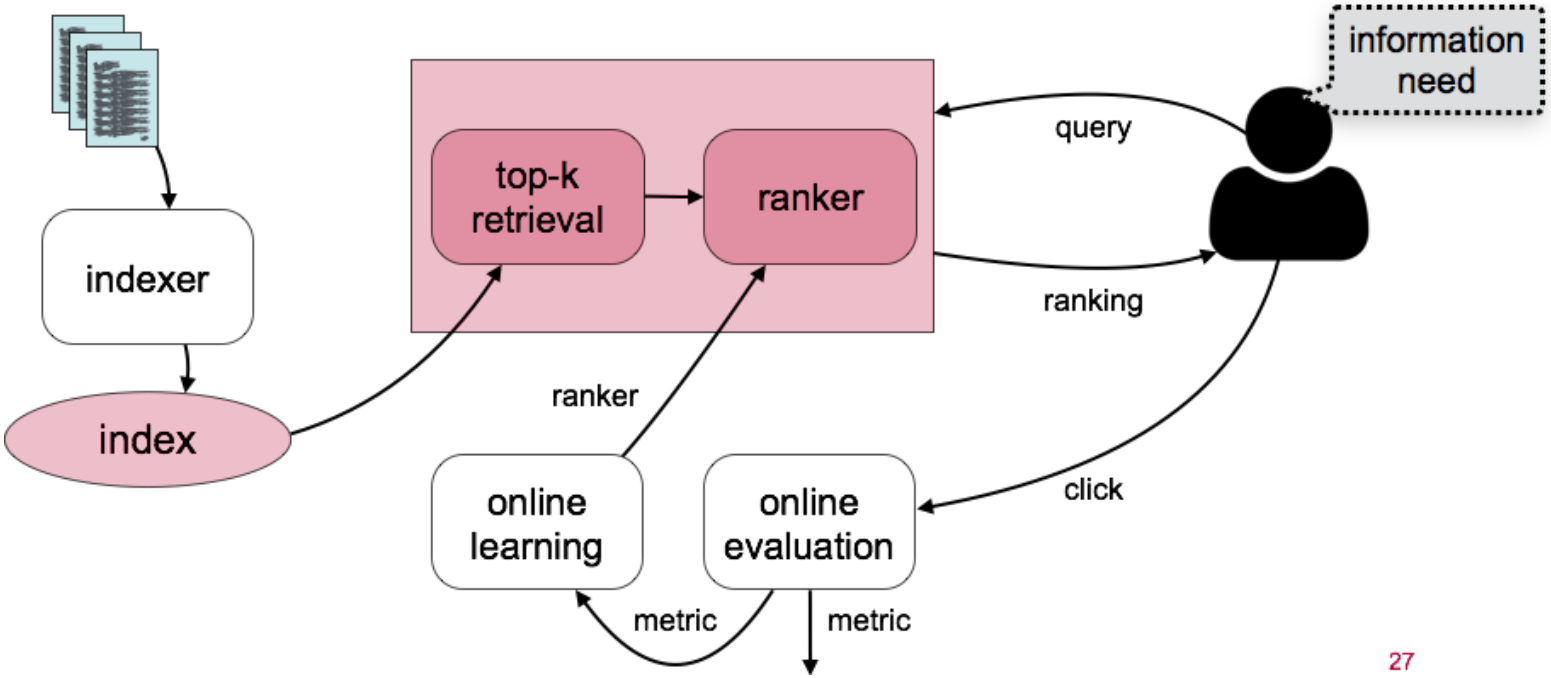
Ranker development

- Traditionally
  - Train and tune offline, then deploy online
  - Supervised learning paradigm

- Move away from supervised paradigm
  - Weakly supervised rankers?
  - A search engine that improves by being used, not in a supervised manner but in a weakly supervised way?
  - Learn from the natural interactions with users
    - To evaluate rankers
    - To combine rankers
    - To create individual rankers
  - Why is this a good idea


indexer
index
top-k retrieval
ranker
query
ranking
information need


train
query
labels
offline learning
metric
metric
offline evaluation
query
labels
test
ranker
information need
query
top-k retrieval
ranker
ranking
indexer
index


information need
query
top-k retrieval
ranker
indexer
index
ranking
ranker
online learning
online evaluation
click
metric
metric
27

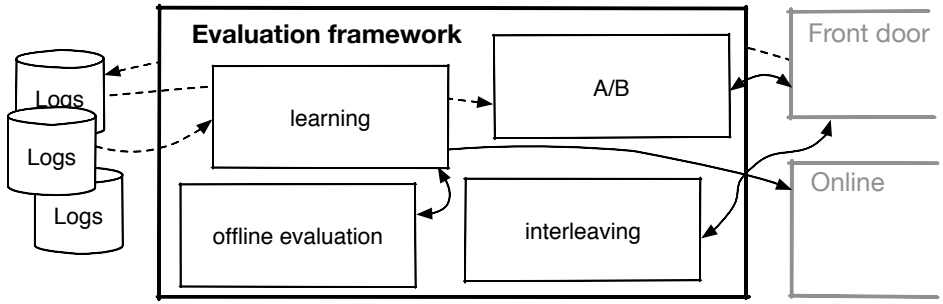
See Katja Hofmann's lecture on Thursday

Top-k retrieval

Skipped.

Outline


Evaluation framework



Evaluation framework $\Rightarrow$ Experimental Framework

Metrics

Flighting

Learning

Logging

Annotations

Three families of evaluation method

- In the literature and in practice
  - Offline evaluation
  - User-study evaluation
  - Online evaluation
- Each method has advantages and disadvantages

Offline evaluation in 3 words

Collect a set of queries

For each query, describe the information being sought

Have assessors determine which documents are relevant

Evaluate systems based on the quality of their rankings

- Evaluation metric: describes quality of ranking with known relevant/non-relevant docs

D. Kelly. Methods for Evaluating Interactive Information Retrieval Systems with Users. Foundations and Trends in Information Retrieval 3(1-2): 1-224, 2009

Offline evaluation in 3 bullets

- Advantages
  - the experimental condition is fixed; same queries, and same relevance judgements
  - evaluations are reproducible; keeps us "honest"
  - by experimenting on the same set of queries and judgements, we can better understand how system one system is better than another

Offline evaluation in 3 bullets

- Advantages
  - the experimental condition is fixed; same queries, and same relevance judgements
  - evaluations are reproducible; keeps us "honest"
  - by experimenting on the same set of queries and judgements, we can better understand how system one system is better than another

- Disadvantages
  - Human assessors that judge documents relevant/non-relevant are expensive
  - Human assessors are not the user; judgements are made ?out of context?
  - Assumes that relevance is the same for every user

Go and attend Stefano, Enrique, Julio and Evangelos's lectures!

User studies in 3 bullets

Provide a small set of users with several retrieval systems

Ask them to complete several (potentially different) search tasks

Learn about system performance by

- Observing what they do
- Asking why they did it

User studies in 3 bullets

- Advantages
  - Very detailed data about users? reaction to systems
  - In reality, a search is done to accomplish a higher-level task
  - In user studies, this task can be manipulated and studied
  - In other words, the experimental ?starting-point? need not be the query

User studies in 3 bullets

- Advantages
  - Very detailed data about users? reaction to systems
  - In reality, a search is done to accomplish a higher-level task
  - In user studies, this task can be manipulated and studied
  - In other words, the experimental ?starting-point? need not be the query
- Disadvantages
  - User studies are expensive (pay users/subjects, scientist?s time, data coding)
  - Difficult to generalize from small studies to broad populations
  - The laboratory setting is not the user?s normal environment
  - Need to re-run experiment every time a new system is considered

Go and attend Diane Kelly's lecture!

Online evaluation in 3 words

See how normal users interact with your live retrieval system when just using it

Observe implicit behavior

- Clicks, skips, saves, forwards, bookmarks, "likes", etc.

Try to infer differences in behavior from different flavors of the live system

- A/B testing
  - Have x% of query traffic use system A and y% of query traffic use system B
- Interleaving
  - Expose a combination of system versions to users

Online evaluation in 3 words

- Advantages
  - System usage is naturalistic; users are situated in their natural context and often don't know that a test is being conducted
  - Evaluation can include lots of users

Online evaluation in 3 words

- Advantages
  - System usage is naturalistic; users are situated in their natural context and often don't know that a test is being conducted
  - Evaluation can include lots of users
- Disadvantages
  - Requires a service with lots of users (enough of them to potential hurt performance for some)
  - This is often referred to as the "cold-start problem" requires a good understanding on how different implicit feedback signals predict positive and negative user experiences
  - Experiments are difficult to repeat

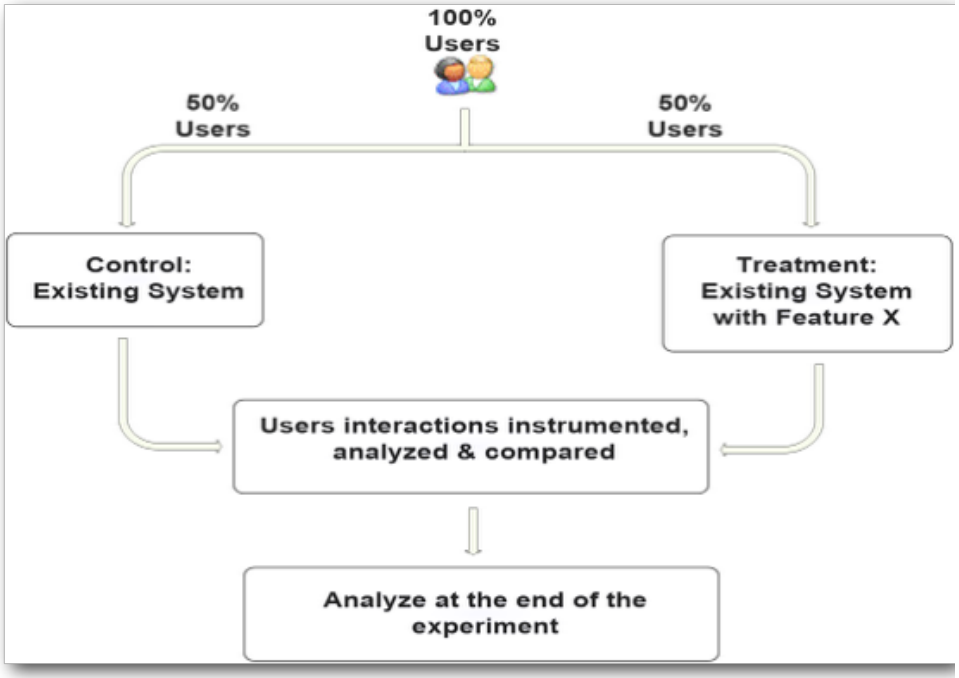
Go and attend Katja Hofmann's lecture!

Relative / SERP level

- A/B testing
- Interleaving

A/B testing

- Concept is trivial
  - Randomly split traffic between two (or more) versions
    - A (Control)
    - B (Treatment)
  - Collect metrics of interest
  - Analyze
- Must run statistical tests to confirm differences are not due to chance
- Best scientific way to prove causality, i.e., the changes in metrics are caused by changes introduced in the treatment(s)



Kohavi et al, 2009

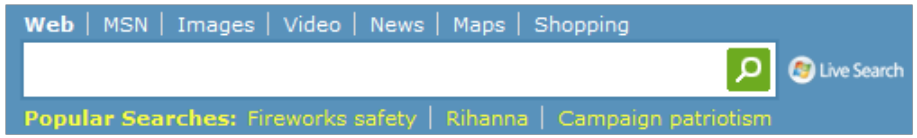
Advantage of A/B testing

- When the variants run concurrently, only two things could explain a change in metrics:
  1. The "feature(s)" (A vs. B)
  2. Random chance
- Everything else happening affects both the variants
- For #2, conduct statistical tests for significance ("Student's t-test")
- A/B experiments are not the panacea for everything
  - Issues discussed in survey paper by Kohavi et al., 2009

Example

- (Kohavi et al., 2013)
- Clickthrough rate for search box and popular searches

☐ **A**



☐ **B**

- Which one wins?

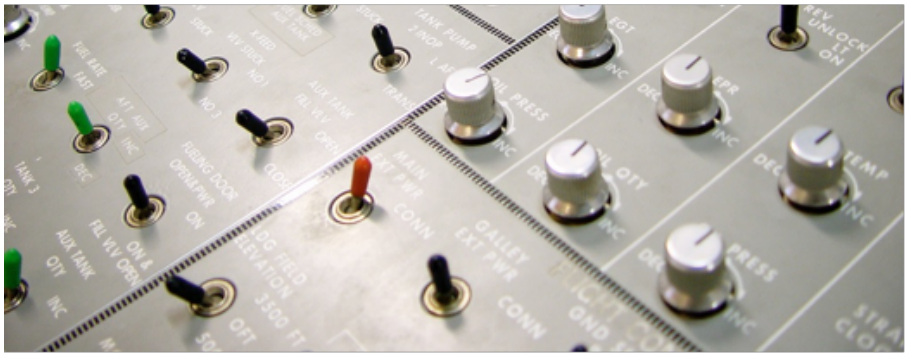
Kohavi et al, 2013

Beware

- Perform many sanity checks
- If something is “amazing,” find the flaw!
  - Examples
    - If you have a mandatory birth date field and people think it’s unnecessary, you’ll find lots of 11/11/11 or 01/01/01
    - If you have an optional drop down, do not default to the first alphabetical entry, or you’ll have lots of: jobs = Astronaut
    - For most web sites, traffic will spike between 1-2AM November 3, 2013, relative to the same hour a week prior. Why?
- Run an A/A test
- ...

Kohavi et al, 2013

Interleaving



http://www.flickr.com/photos/sameli/540933604/

Task
Combine hundreds of ranking features to get the best ranking for each search task / user

Approach

Today

Offline – use manual annotations for manual tuning or supervised learning, **problems:** resources, fidelity, scale

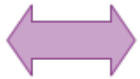
Tomorrow?

Online – learn directly from natural user interactions with the search system

Learning from natural user interactions with an IR system

(e.g., clicks on search results)

- Easy to collect while system operates
- Reflect natural user behavior and preferences
- Enable online learning



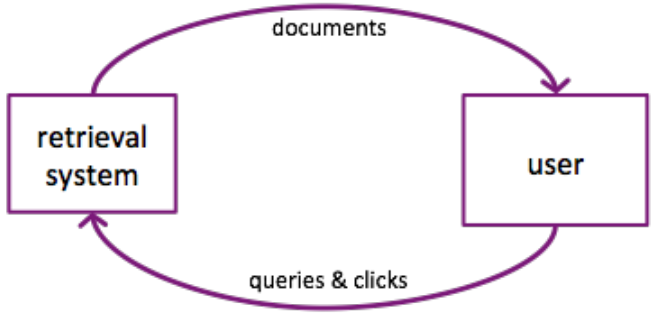
- Noisy
- Provide only relative preference indications

How can IR systems learn reliably and efficiently from noisy, relative feedback?

Approach – Overview

Reinforcement Learning Approach

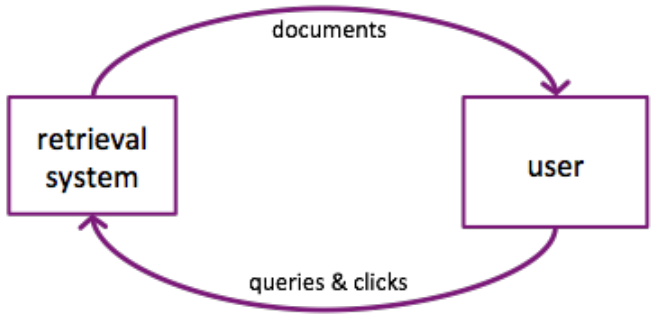
- Learn by trying out actions (document lists), and observing feedback
- Follow a listwise learning approach (compare two ranking functions per round)
- Assume independent queries (contextual bandit problem)



Approach – Overview

Reinforcement Learning Approach

- Learn by trying out actions (document lists), and observing feedback
- Follow a listwise learning approach (compare two ranking functions per round)
- Assume independent queries (contextual bandit problem)

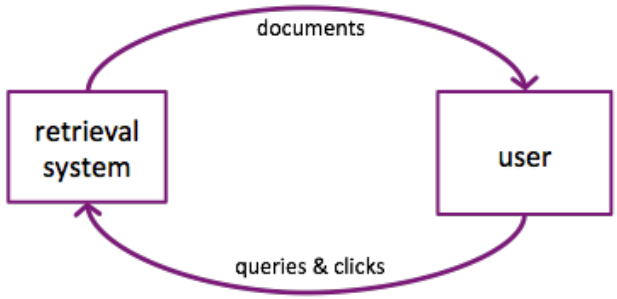


1. observe query
2. generate result list
3. infer feedback from clicks
4. update retrieval function

Approach – Overview

Reinforcement Learning Approach

- Learn by trying out actions (document lists), and observing feedback
- Follow a listwise learning approach (compare two ranking functions per round)
- Assume independent queries (contextual bandit problem)



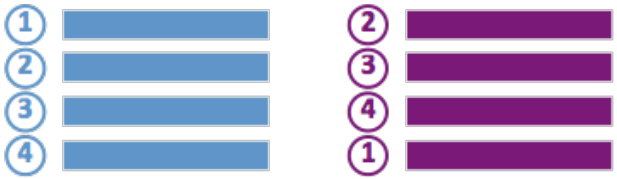
1. observe query
2. generate result list
3. infer feedback from clicks

How to reliably infer feedback?

4. update retrieval function

How to learn efficiently online?

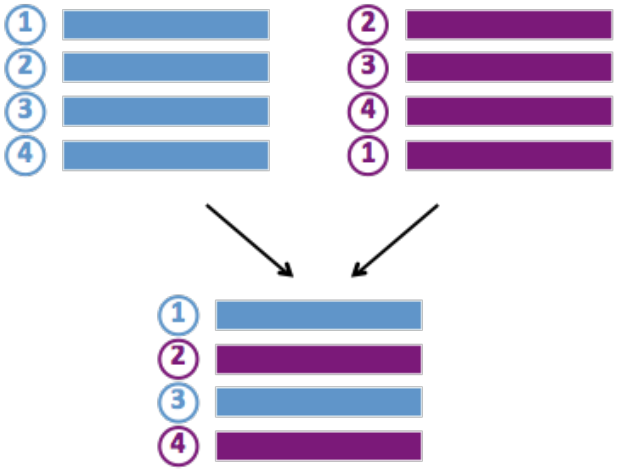
Interleaved comparisons method — online evaluation



- Goal: Compare two result lists using click data
- Procedure:
  1) Generate interleaved result list (randomize per pair of ranks)
  2) Observe user clicks
  3) Credit clicks to original rankers to infer outcome $o \in \{-1, 0, +1\}$

T. Joachims. *Evaluating Retrieval Performance using Clickthrough Data.* (Text Mining, Springer, 2003)

Baseline: Team draft

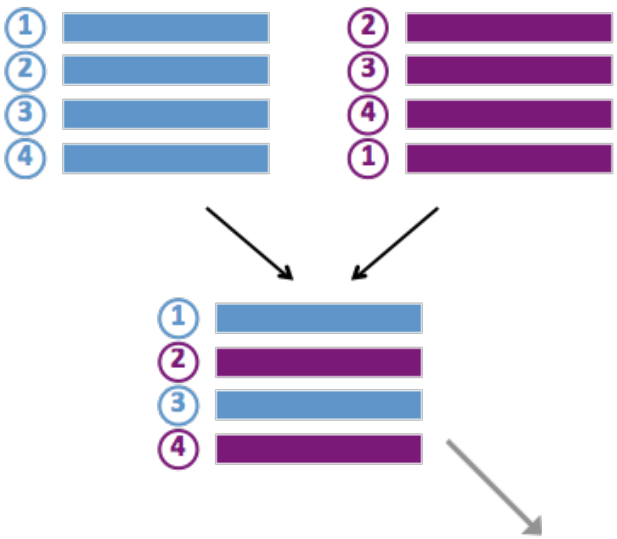


- Goal: Compare two result lists using click data
- Procedure:
  1) **Generate interleaved result list** (randomize per pair of ranks)
  2) Observe user clicks
  3) Credit clicks to original rankers to infer outcome

$$o \in \{-1, 0, +1\}$$

F. Radlinski, M. Kurup, and T. Joachims. *How does clickthrough data reflect retrieval quality?* (CIKM 2008)

Baseline: Team draft

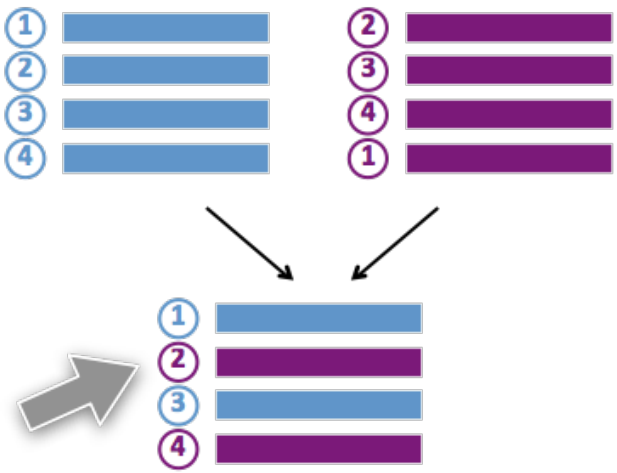


- Goal: Compare two result lists using click data
- Procedure:
  1) **Generate interleaved result list** (randomize per pair of ranks)
  2) Observe user clicks
  3) Credit clicks to original rankers to infer outcome $o \in \{-1, 0, +1\}$

Key idea: Keep track of **assignments** (which list contributed which document)

F. Radlinski, M. Kurup, and T. Joachims. *How does clickthrough data reflect retrieval quality?* (CIKM 2008)

Baseline: Team draft



- Goal: Compare two result lists using click data
- Procedure:
  1) Generate interleaved result list (randomize per pair of ranks)
  2) **Observe user clicks**
  3) Credit clicks to original rankers to infer outcome

$$o \in \{-1, 0, +1\}$$

F. Radlinski, M. Kurup, and T. Joachims. *How does clickthrough data reflect retrieval quality?* (CIKM 2008)

Interleaving pros & cons

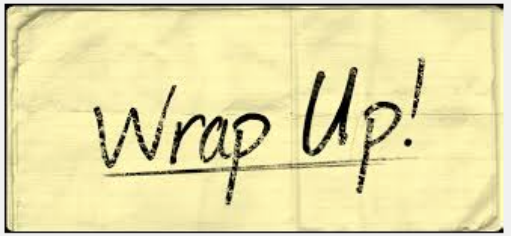
- Benefits
  - A direct way to elicit user preferences
  - More sensitive than many other online metrics
  - Deals with issues of position bias and calibration
  - Reusability recently addressed and partially solved
- Drawbacks
  - Benchmark: No absolute number for benchmarking
  - Interpretation: Unable to interpret much at the documentl-level, or about user behavior

Outline

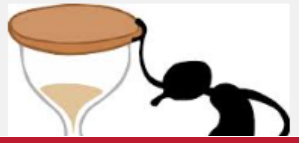

WRAP UP

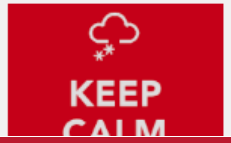

KEEP
CALM
AND
WRAP IT
UP


the
WRAP
UP


Wrap Up!


Weekly wrap up




KEEP


be inspired

Wrap-up


The
WRAP
UP


2011


THE END


2015


the
weekly
wrapup


What would you change about the program?
projects
Longer
days
Shorter
time
building
location
models


WRAP
IT
UP


WEEKLY
WRAP-UP


IT'S A
WRAP


THE WEEKEND
WRAP-UP


Thank you!

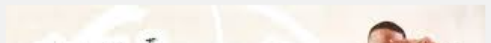

WRAP-UP
400 × 300 - dribbble.com


THAT'S A WRAP

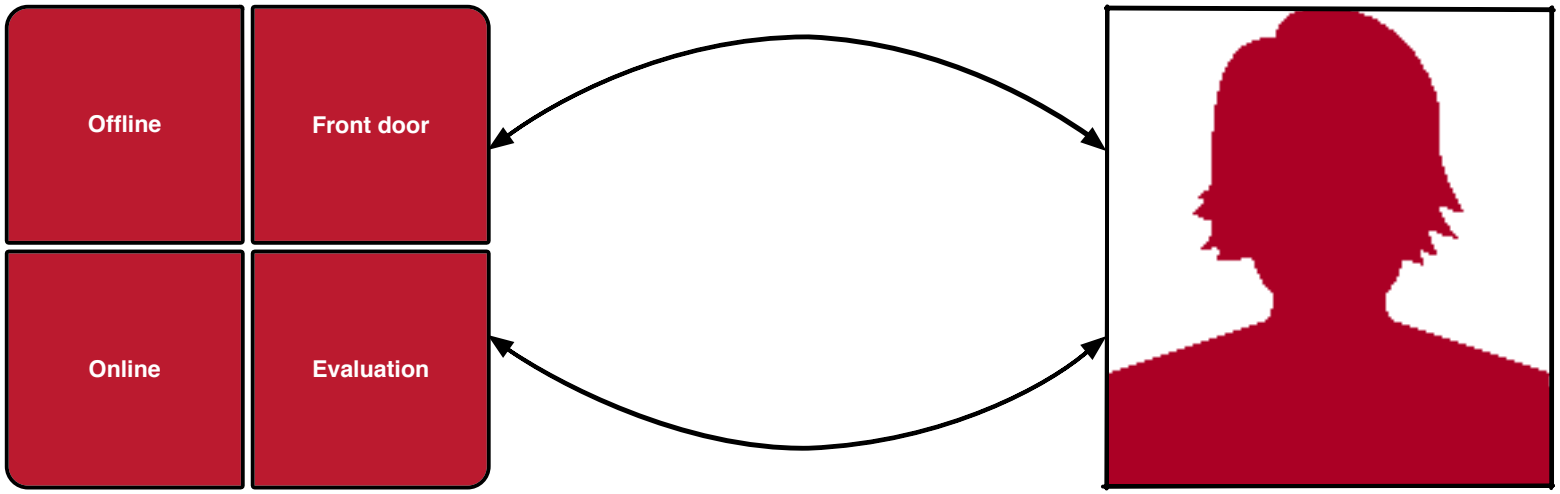

WRAP UP


THANK YOU
VERY MUCH
ONCE AGAIN
KINDLY, SIR
FOR NOTHIN


THE END



The bigger picture



Responsible information retrieval

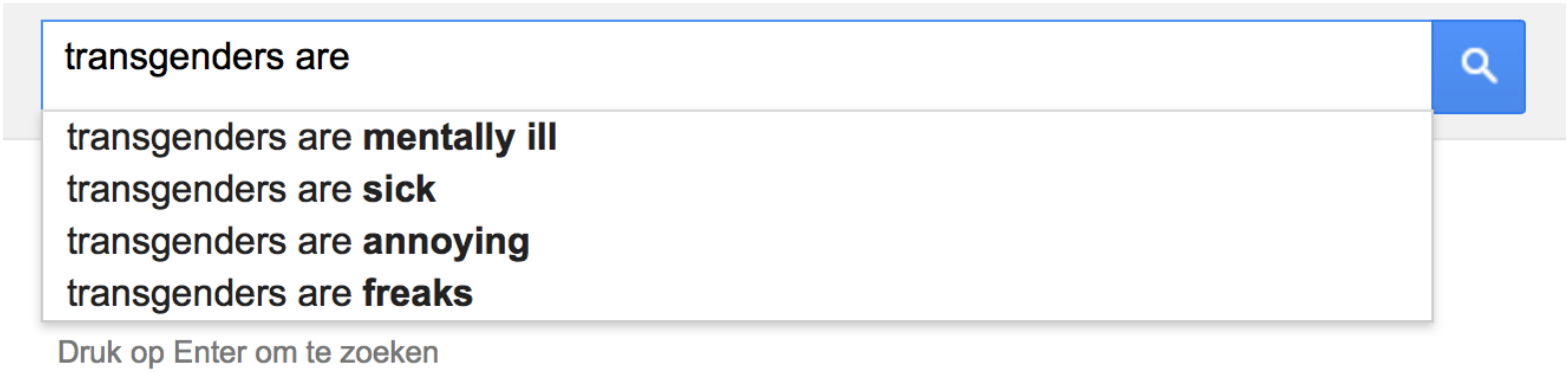
As user data becomes more and more important . . .

- Privacy
- Fairness

Responsible information retrieval

As user data becomes more and more important . . .

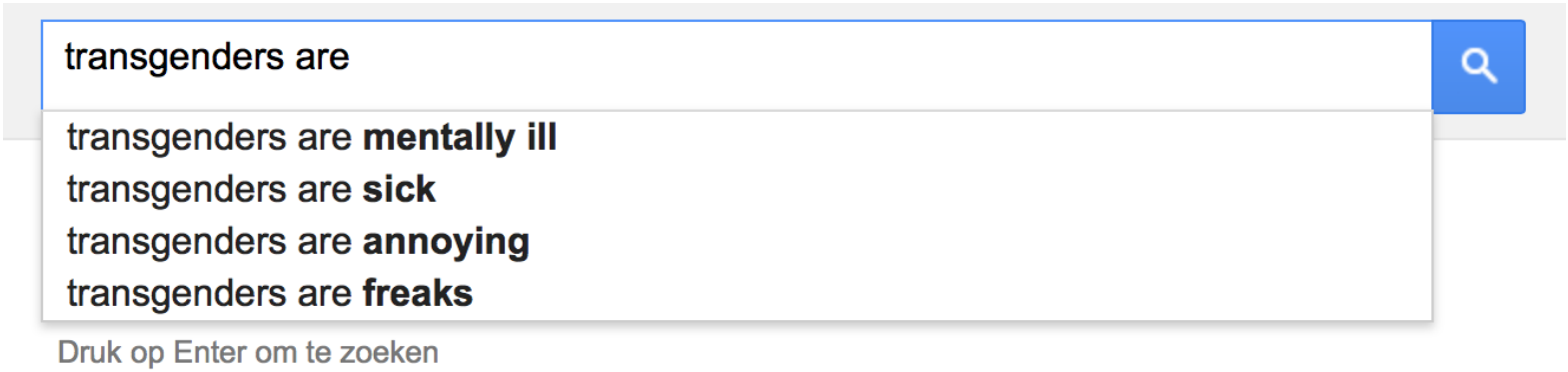
- Privacy
- Fairness



Responsible information retrieval

As user data becomes more and more important . . .

- Privacy
- Fairness

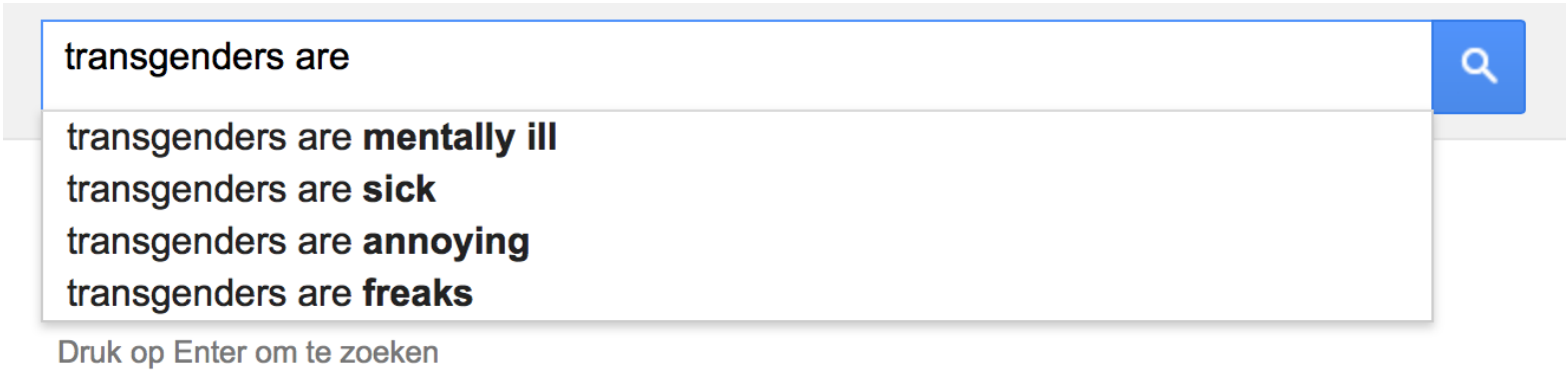


- Accuracy
  - More dominant groups have bigger counts, more accurate estimates

Responsible information retrieval

As user data becomes more and more important . . .

- Privacy
- Fairness



- Accuracy
  - More dominant groups have bigger counts, more accurate estimates
- Transparancy
  - Let the system explain why it is showing certain results

Stuff you should work on

- Large-scale understanding of users and user behavior
- Higher level models of any aspect of search
- Online anything
- Responsible IR

Acknowledgments

All content represents the opinion of the author, which is not necessarily shared or endorsed by his employer and/or sponsors.

Resources

Text books

- R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999
- S. Buettcher, C.L. A. Clarke and G.V. Cormack. *Information Retrieval: Implementing and Evaluating Search Engines*. MIT Press, 2010
- W.B. Croft, D. Metzler and T. Strohman. *Search Engines: Information Retrieval in Practice*. Pearson, 2010
- D. Grossman and O. Frieder. *Information Retrieval: Algorithms and Heuristics*. Springer, 2004
- M. Hearst. *Search User Interfaces*, CUP, 2009
- C. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008
- I. Witten, A. Moffat, and T. Bell. *Managing Gigabytes*. Morgan Kaufmann, 1999

Evaluation campaigns

- CLEF: Conference and Labs of the Evaluation Forum, `http://www.clef-campaign.org`
- FIRE: Forum for Information Retrieval Research, `http://fire.irsi.res.in/fire/home`
- MediaEval Benchmark: MediaEval Benchmarking Initiative for Multimedia Evaluation, `http://www.multimediaeval.org`
- NTCIR: NII Testbeds and Community for Information Access Research, `http://research.nii.ac.jp/ntcir/index-en.html`
- TREC: Text Retrieval Conference, `http://trec.nist.gov`

Papers cited

- Z. Bar-Yossef, N. Kraus. Context-sensitive query auto-completion. In WWW '11, 2011
- P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, S. Vigna. The query-flow graph: model and applications. In CIKM '08, 2008
- M. Bron, M. de Rijke, J. Van Gorp, A. Vishneuski, F. Nack, S. de Leeuw. A subjunctive exploratory search interface to support media studies researchers. In SIGIR '12, 2012
- M. Bron, J. Van Gorp, F. Nack, L.B. Baltussen, M. de Rijke. Aggregated search interface preferences in multi-session search tasks. In SIGIR '13, 2013
- F. Cai, S. Liang, M. de Rijke. Time-sensitive personalized query auto-completion. In CIKM '14, 2014
- H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, H. Li. Context-Aware Query Suggestion by Mining Click-Through and Session Data. In KDD '08, 2008
- A. Chuklin, I. Markov, M. de Rijke. Click Models for Web Search. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, July 2015
- N. Craswell, M. Szummer. Random walks on the click graph. In SIGIR '07, 2007
- A. Grotov, S. Whiteson, M. de Rijke. Bayesian ranker comparison based on historical user interactions. In SIGIR '15, 2015
- K. Hofmann, S. Whiteson, M. de Rijke. Contextual bandits for information retrieval. In NIPS 2011 Workshop on Bayesian Optimization, Experimental Design, and Bandits, December 2011
- D. Kelly. Methods for Evaluating Interactive Information Retrieval Systems with Users. Foundations and Trends in Information Retrieval 3(1-2): 1-224, 2009
- R. Kohavi, R. Longbotham, D. Sommerfield, R.M. Henne. Controlled experiments on the web: survey and practical guide. Data Min. Knowl. Discov. 18(1): 140-181, 2009
- R. Kohavi, A. Deng, B. Frasca, T. Walker, Y. Xu, N. Pohlmann. Online controlled experiments at large scale. In KDD '13, 2013
- D. Lefortier, P. Serdyukov, M. de Rijke. Online exploration for detecting shifts in fresh intent. In CIKM '14, 2014
- B. Mitra. Exploring session context using distributed representations of queries and reformulations. In SIGIR '15, 2015
- B. Mitra, M. Shokouhi, F. Radlinski, K. Hofmann. On user interactions with query auto-completion. In SIGIR '14, 2014

- U. Ozertem, O. Chapelle, P. Donmez, E. Velipasaoglu. Learning to suggest: a machine learning framework for ranking query suggestions. In SIGIR '12, 2012
- K. Radinsky, K.M. Svore, S.T. Dumais, M. Shokouhi, J. Teevan, A. Bocharov, E. Horvitz. Behavioral dynamics on the web: Learning, modeling, and prediction. ACM Trans. Inf. Syst. 31(3): 16, 2013
- F. Radlinski, P.N. Bennett, B. Carterette, T. Joachims. Redundancy, diversity and interdependent document relevance. SIGIR Forum 43(2): 46-52, 2009
- F. Radlinski, M. Kurup, T. Joachims. How does clickthrough data reflect retrieval quality?. In CIKM '08, 2008
- D.E. Rose and D. Levinson. Understanding User Goals in Web Search. In WWW '04, 2004
- M. Shokouhi. Learning to personalize query auto-completion. In SIGIR '13, 2013
- M. Shokouhi, K. Radinsky. Time-sensitive query auto-completion. In SIGIR '12, 2012
- S. Whiting, J.M. Jose. Recent and robust query auto-completion. In WWW '14, 2014
- Z. Zhang and O. Nasraoui. Mining search engine query logs for query recommendation. In WWW '06, 2006.
- M. Zoghi, S. Whiteson, M. de Rijke, R. Munos. Relative confidence sampling for efficient on-line ranker evaluation. In WSDM '14, 2014