Information Retrieval as Interaction

12th European Summer School in Information Retrieval

Maarten de Rijke

July 15, 2019

University of Amsterdam

derijke@uva.nl

Based on teaching materials developed with Evangelos Kanoulas, Ilya Markov, and Harrie Oosterhuis.

Introduction

We need information to make decisions . . .

. . . to identify or structure a problem or opportunity

. . . to put problem or opportunity in context

. . . to generate alternative solutions

. . . to choose the best alternative

THE UNIVERSAL DECLARATION OF Human Rights

WHEREAS *recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world,*

WHEREAS *disregard and contempt for human rights have resulted in barbarous acts which have outraged the conscience of mankind, and the advent of a world in which human beings shall enjoy freedom of speech and belief and freedom from fear and want has been proclaimed as the highest aspiration of the common people,*

WHEREAS *it is essential, if man is not to be compelled to have recourse, as a last resort, to rebellion against tyranny and oppression, that human rights should be protected by the rule of law,*

WHEREAS *it is essential to promote the development of friendly relations among nations,*

WHEREAS *the peoples of the United Nations have in the Charter reaffirmed their faith in fundamental human rights, in the dignity and worth of the human person and in the equal rights of men and women and have determined to promote social progress and better standards of life in larger freedom,*

WHEREAS *Member States have pledged themselves to achieve, in co-operation with the United Nations, the promotion of universal respect for and observance of human rights and fundamental freedoms,*

WHEREAS *a common understanding of these rights and freedoms is of the greatest importance for the full realisation of this pledge,*

NOW THEREFORE THE GENERAL ASSEMBLY

PROCLAIMS *this Universal Declaration of Human Rights as a common standard of achievement for all peoples and all nations, to the end that every individual and every organ of society, keeping this Declaration constantly in mind, shall strive by teaching and education to promote respect for these rights and freedoms and by progressive measures, national and international, to secure their universal and effective recognition and observance, both among the peoples of Member States themselves and among the peoples of territories under their jurisdiction.*

Access to information a basic human right

teaching, practice, worship and observance.

ARTICLE 19 —Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.

ARTICLE 20 —1. Everyone has the right to freedom

Information retrieval

Technology to connect people to the right information in the right way at the right time

- **Search engines**
  - Google, library search engine, desktop search engine, search engine on your mobile phone, product search...
- **Recommender systems**
  - Product recommendation, movie recommendation, date recommendation, news recommendation, music recommendation, ...
- **Digital assistants**
  - Alexa, Siri, Google, bots on web sites, bots for consumer services, ...

Search engines

User enters a query to express their information need

Broad collection of items ("the web") vs narrow collection (scientific articles, your email)

Understand user intent and rank items based on intent analysis, interaction data, behavioral analysis, analysis of document content, popularity, temporal factors, ...

Typically returns a ranked list of items, with the best items ranked at the top

Recommender systems

User logs on

In case there is a broad collection, facets may help to narrow down space of recommended items

Recommender system may use short-term interests (current session) and long-term user interests (previous sessions) and exploit implicit or explicit interaction data

Typically returns a broad range of options for user to select from

Digital assistants

There may or may not be a screen or keyboard

System asks clarification questions

- life is **easier** for the system: can ask for help
- life is **harder** for the system: needs to ask the right question and understand user response

Uses query information, profile information, information gleaned from interactions (implicit and explicit)

Typically returns a single result ("the answer") at a time

Getting the **right** information

to the **right** people

in the **right** way

Information retrieval

Making IR results as useful as possible possibly easy for some IR tasks

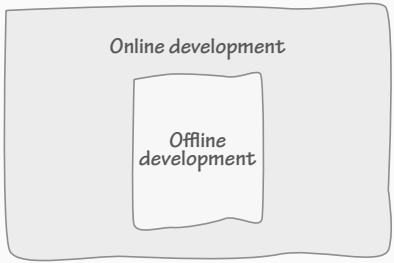
Dependence on context: user's background knowledge, age, location, their specific search goals

The more we learn about how much context influences peoples' search behavior and goals, the more it becomes clear that many hundreds of preference criteria play a role

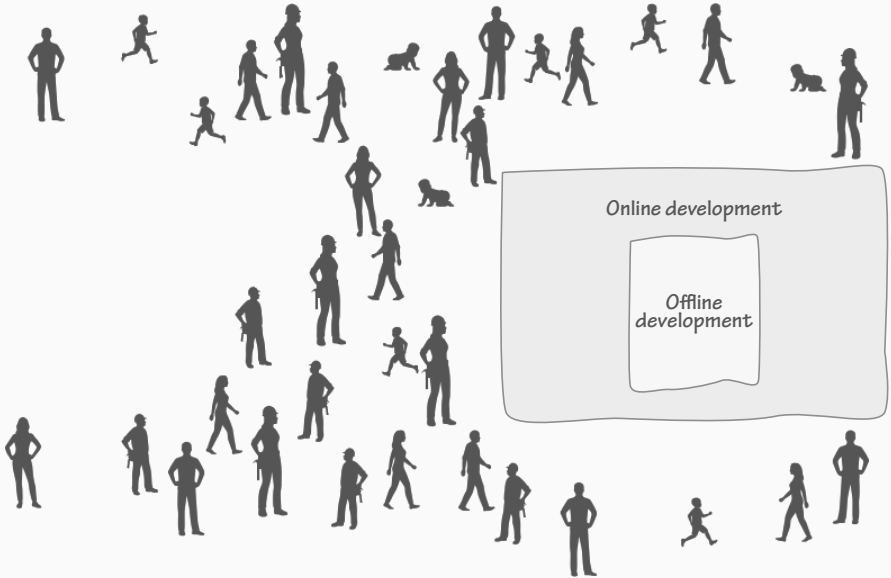
Addressing each optimal combination of IR result preferences individually not feasible → look for scalable methods that learn good IR results without expensive tuning – in search, recommendatino and conversations

Information retrieval system development– Two phases

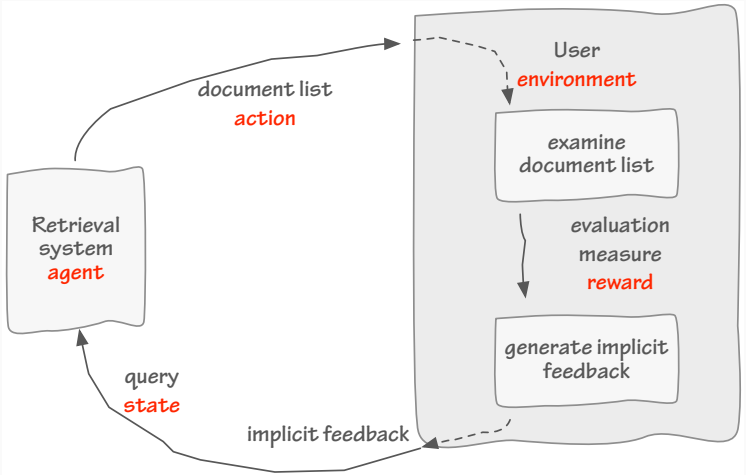
Information retrieval system development – Two phases



Information retrieval system development – Two phases



Information retrieval system development – The online phase



Offline vs. online system development

In CS, algorithms that receive their input sequentially operate in an **online** modality

- Typically includes tasks that involve sequences of decisions, like when you choose how to serve incoming queries in 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

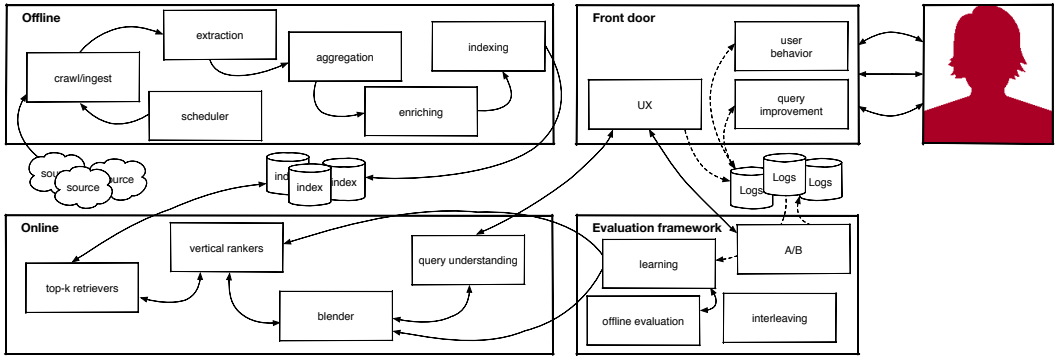
- Any processing that is not query dependent (crawling, indexing, . . .)

Typical **online** computations in information retrieval

- Any processing that depends on users, their profile and their input

How does it all fit together?

A “spaghetti” picture for information retrieval



How does the ESSIR 2019 program fit?

Understand users and track intent

Update models and space of possible actions (answer, ranked list, SERP, . . .)

Select the best action and sense its effect

What does this mean for machines?

Life is easier for systems than in an offline trained query-response paradigm

- Engage with user
- Educate/train user
- Ask for clarification from user

Life is harder for systems than in an offline trained query-response paradigm

- **Safety** – Don't hurt anyone
- **Explicability** – Be transparent about model, about decisions

Introduction

Front door

Evaluation

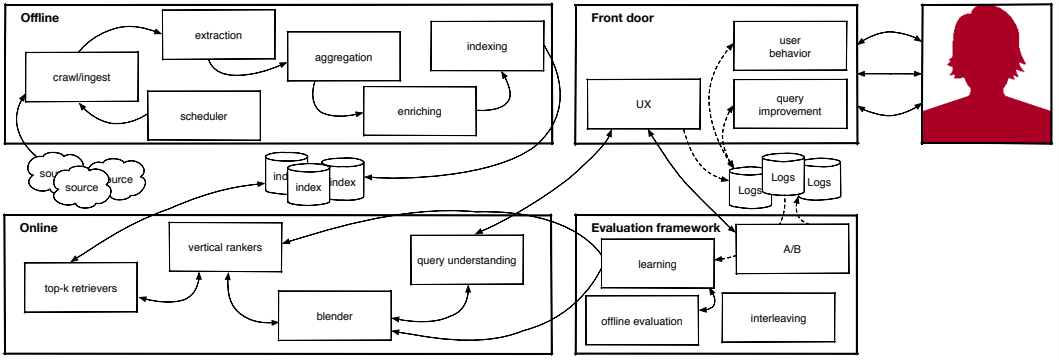
Online

Offline

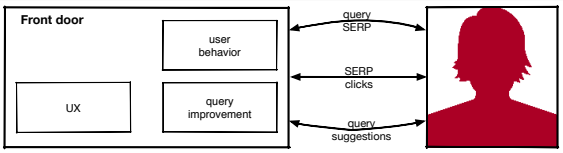
Wrap-up

Front door

The big picture



Zooming in on the front door



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

Receives query, may return query improvements or suggestions for improvements.

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

UX

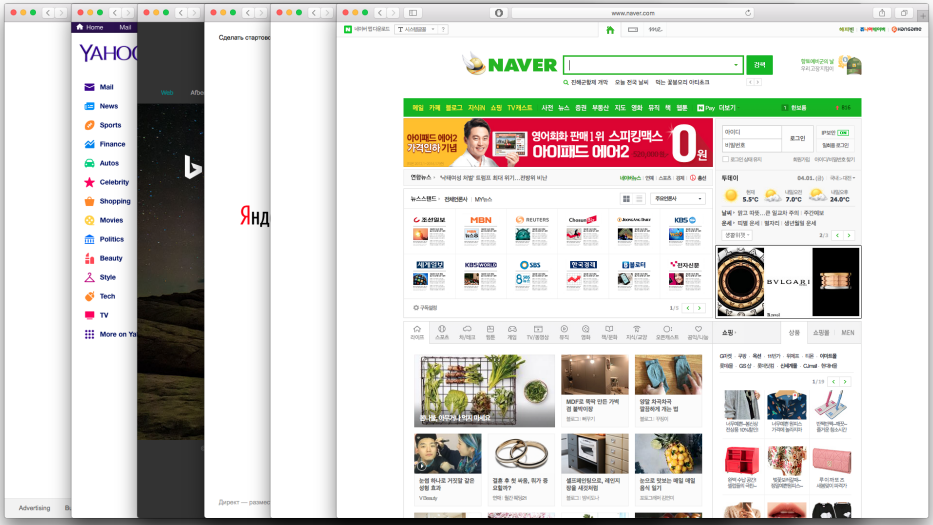
Search interface guidelines include:

- 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

(Hearst, 2009).

Example: Web search engines



Example: AVResearcherXL

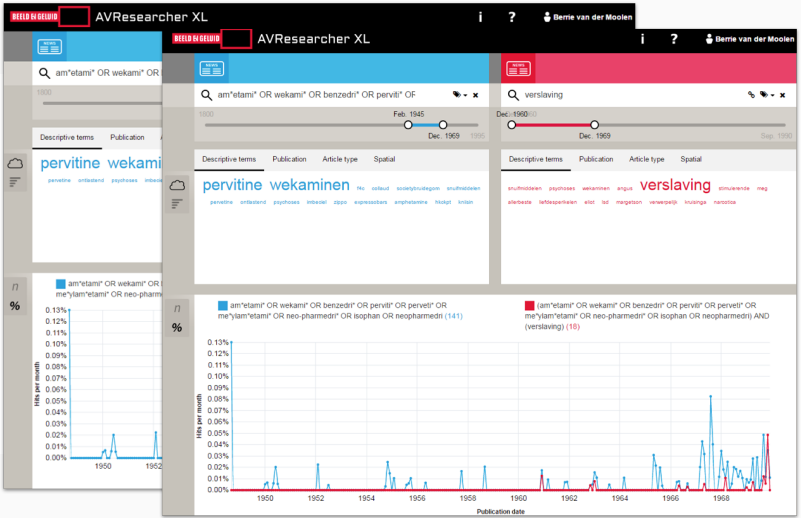
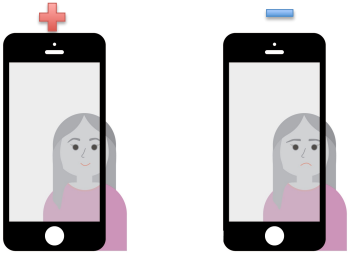


Image credits: (Bron et al., 2012).

New UX challenges

(A) User Satisfaction



(B) Clicks Heatmap

(C) Touch Interaction

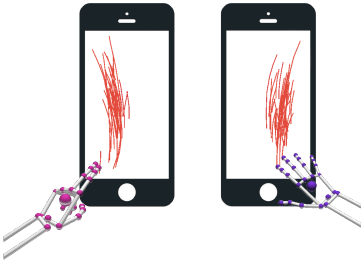
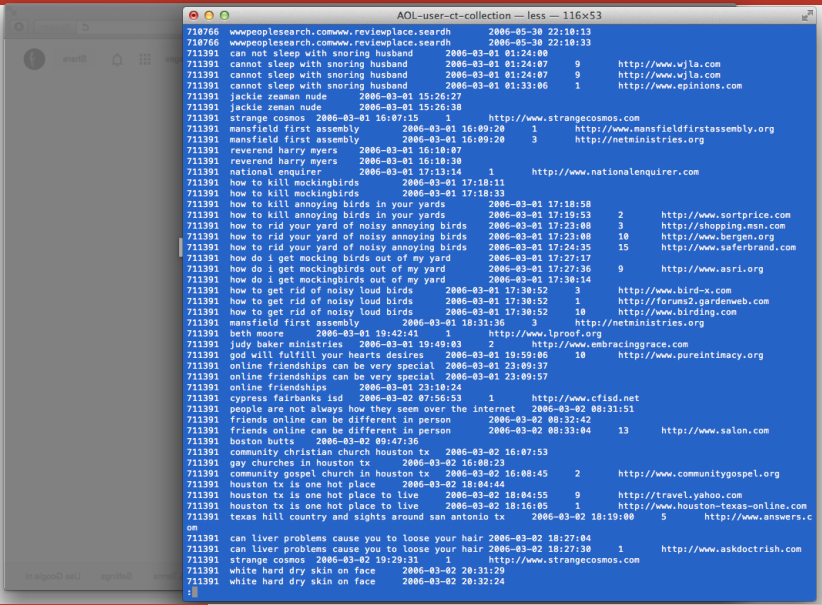


Image credits: (Kiseleva and de Rijke, 2017)

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*

In the logs . . .


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 revered 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.com
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
:

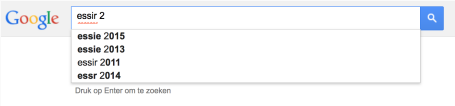
Query improvement

Help users formulate better queries

- Query auto completion
- Query suggestion
- Query expansion
- Query correction

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

(Mitra et al., 2014).

In class exercise

Propose a method plus architecture for query auto completion

- What data to learn from?
- Which features?
- How to evaluate?

Solving the query auto completion problem

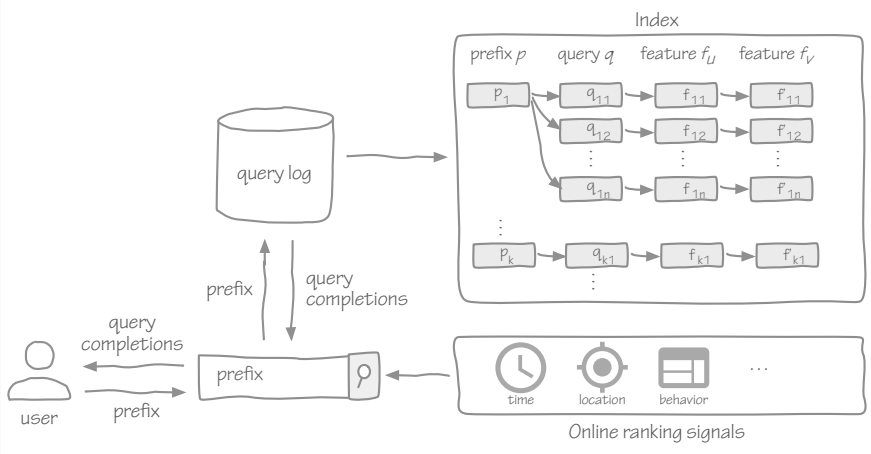


Image credits: (Cai and de Rijke, 2016).

Solving the query auto complextion problem

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

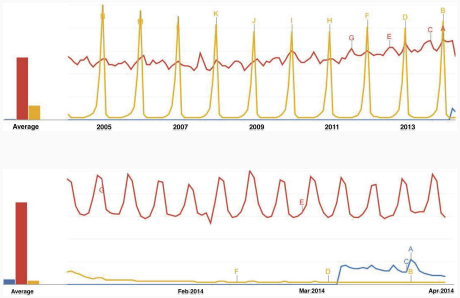


Image credits: Cai et al. (2014)

movie, christmas, MH370

Solving the query auto completion problem

Observations

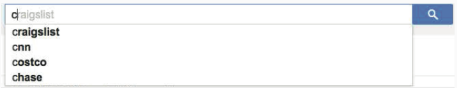
- Recency; Whiting et al. (WWW '14)
- Specific temporal intervals; Shokouhi et al. (SIGIR '12)

Predictions

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

Solving the query auto completion problem

Add personalization



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

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

Solving the query auto completion problem

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)

Recent trends

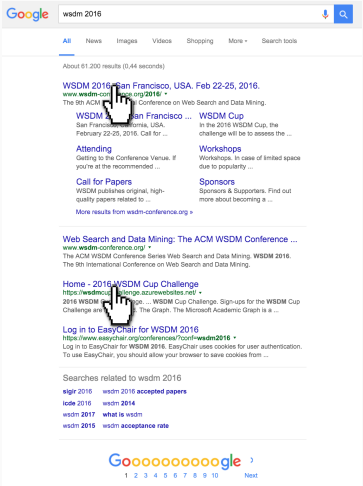
- Semantics (distributed representations; Mitra, 2015)
- Adaptive models

(Cai and de Rijke, 2016)

Many other query improvement tasks

- Query suggestions
- Query expansion
- Query correction
- Semantic analysis of queries
- ...

Interaction signals



Why clicks?

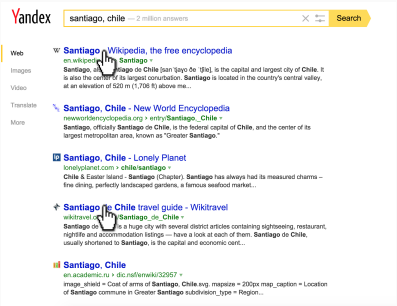
Reflect user interests

Help to improve search

Help to evaluate search

Example: Position-based model

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



Basic click models

CTR models: counting clicks

Position-based model (PBM): examination and attractiveness

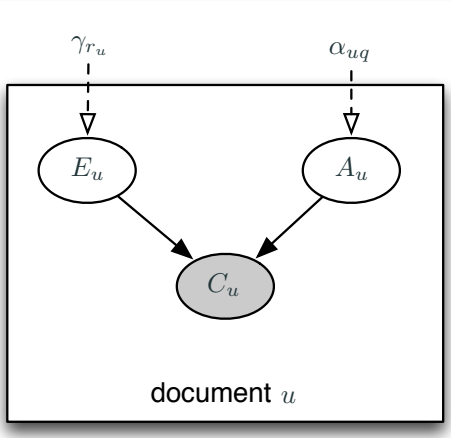
Cascade model (CM): previous examinations and clicks matter

Dynamic Bayesian network model (DBN): satisfactoriness

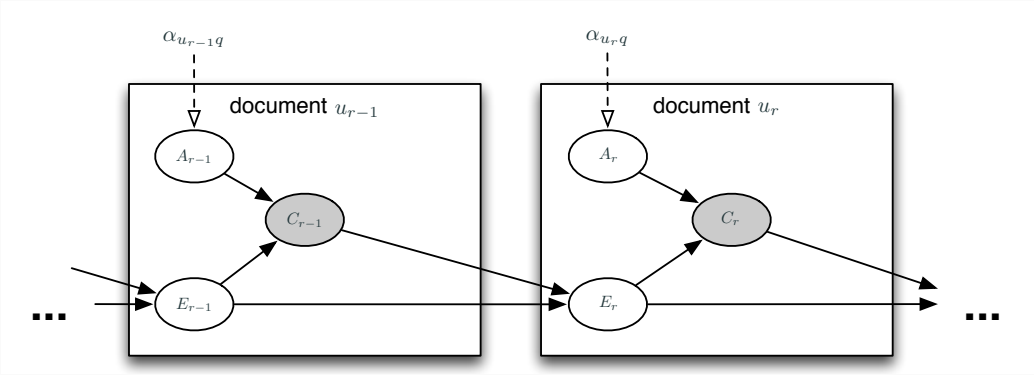
User browsing model (UBM): rank of previous click

...

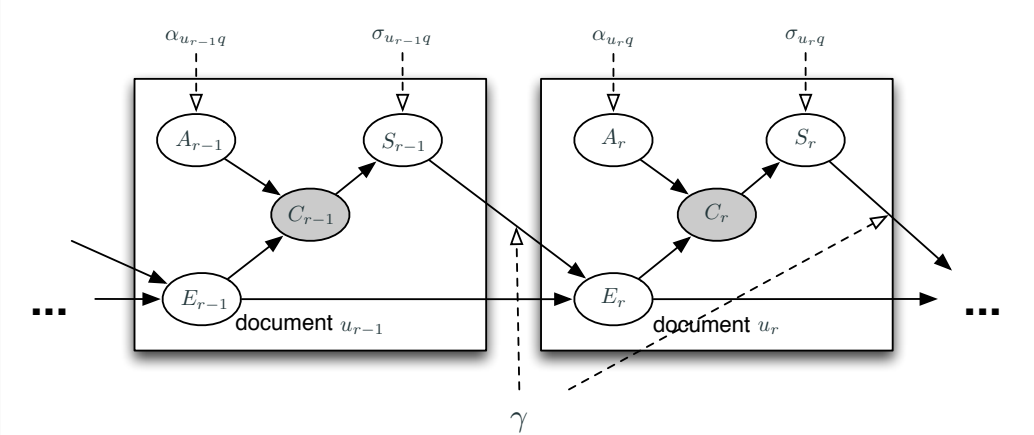
Position-based model: probabilistic graphical model



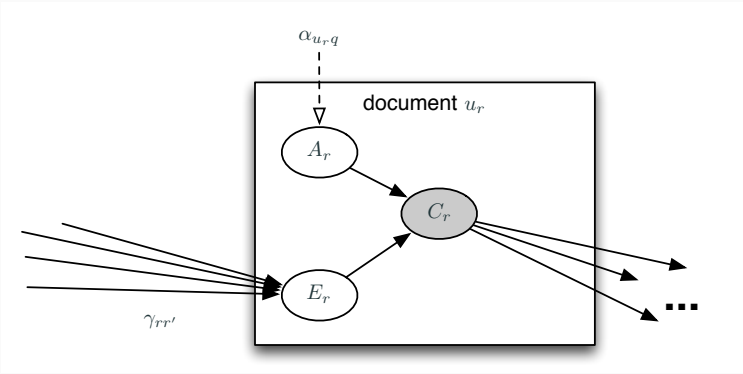
Cascade model: probabilistic graphical model



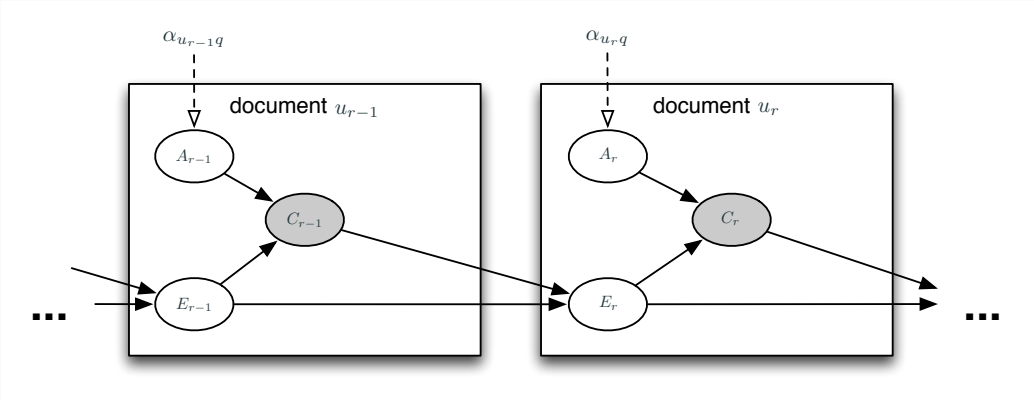
Dynamic Bayesian network: probabilistic graphical model



User browsing model: probabilistic graphical model



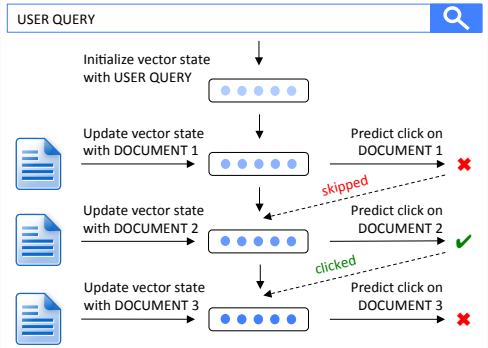
PGM-based click models



Pros: Based on the *probabilistic graphical model* (PGM) framework

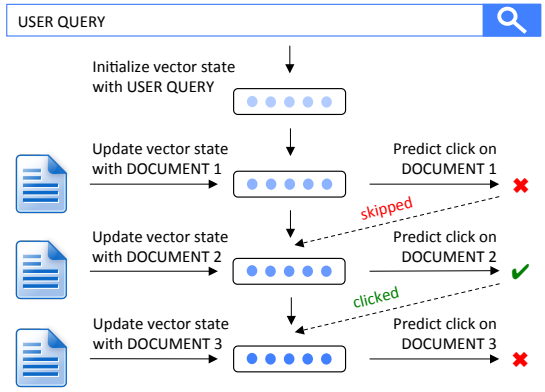
Cons: Structure of the underlying PGM has to be set manually

Alternative click modeling framework



Learns patterns of user behavior directly from click-through data

Distributed representations $(\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2, \ldots)$



Model user browsing behavior as sequence of vector states $(\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2, \ldots)$ that captures the user's information need and information consumed by user during search

Mappings I, U and Function F

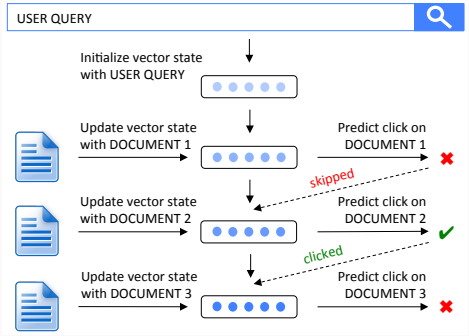
q – user query

d_r – document at rank r

i_r – user interaction with document at rank r

$$\mathbf{s}_0 = \mathcal{I}(q)$$

$$\mathbf{s}_{r+1} = \mathcal{U}(\mathbf{s}_r, i_r, d_{r+1})$$



$$P(C_{r+1} = 1 \mid q, i_1, \ldots, i_r, d_1, \ldots, d_{r+1}) = \mathcal{F}(\mathbf{s}_{r+1})$$

Click modeling upshot

Neural models beat probabilistic graphical models

- Behavior prediction (perplexity, log-likelihood)
- Extraction relevance signal (NDCG)

Neural models have been extended to capture other behavior signals too

- Time between actions
- Journey prediction
- Removing context bias
- ...

(Borisov et al., 2016)

Related tutorial(s) for more in-depth treatment at ESSIR

- Approaches to Research in IR, W. Bruce Croft (next)
- Foundations of User-oriented IR, Nick Belkin & Diane Kelly (Tue)
- Foundations of Machine Learning for IR, Claudia Hauff (Tue)
- Design and Evaluation of Recommender Systems, Paolo Cremonesi (Thu)
- Understanding & Inferring User Tasks and Needs, Emine Yilmaz (Fri)

Introduction

Front door

Evaluation

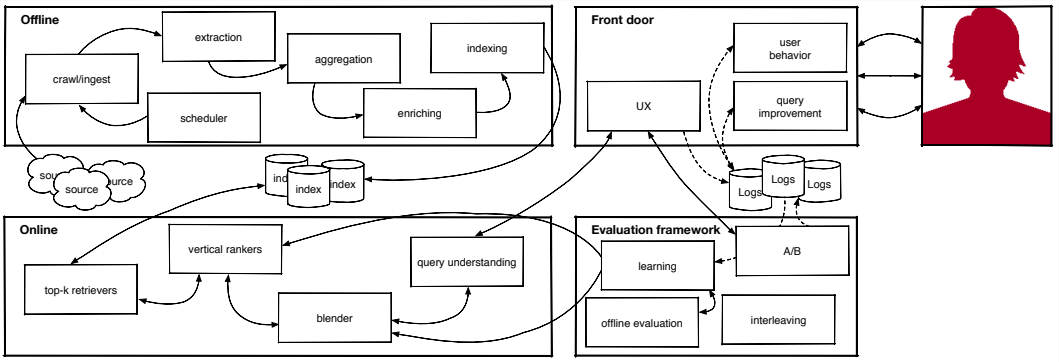
Online

Offline

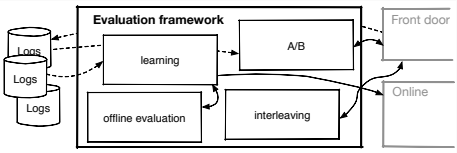
Wrap-up

Evaluation

The big picture



Zooming in on evaluation



Evaluation framework ⇒ 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 bullets

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

(Sanderson, 2010)

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

(Sanderson, 2010)

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

(Kelly, 2009)

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

(Kelly, 2009)

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

(Hofmann et al., 2016)

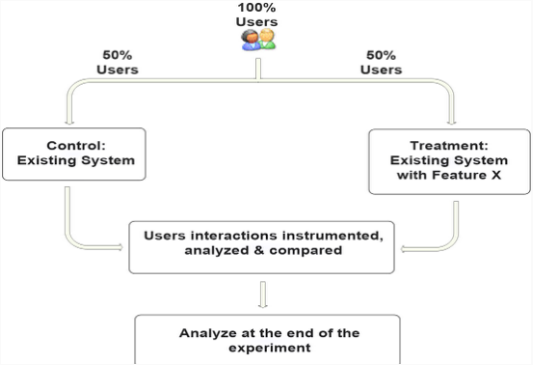
Online evaluation in 3 words

- 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

(Hofmann et al., 2016)

A/B testing and interleaving

- Concept of A/B testing 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 way to prove causality (metrics changes caused by treatment changes)



(Kohavi, 2013)

Advantage of A/B testing

- When the variants run concurrently, only two things could explain a change in metrics:
  - The "features" (A vs. B)
  - Random choice
- Everything else happening affects both conditions
- For #2, conduct statistical tests for significance
- A/B experiments are not the panacea for everything
  - (Kohavi, 2013) has interesting anecdotes

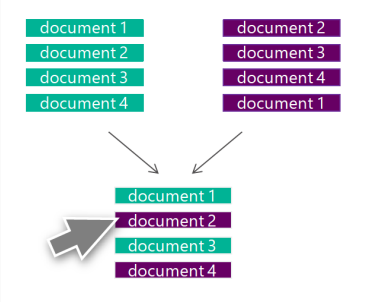
A/B beware

- 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 not necessary, 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 1–2 AM November, 2013 relative to the same hour a week prior. Why?
- Run an A/A test
- ...

Interleaving

Task: Find out which of two systems is best (A or B), based on natural interactions with system

How can we reliably determine preferences from noisy, relative feedback?



- Goal: compare two result lists using click data
- Procedure:
  - Generate interleaved result list
  - Observe user clicks
  - Credit clicks to original rankers to infer outcome: $o \in \{-1, 0, +1\}$

(Joachims, 2003)

Interleaving pros and 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
  - From interleaving to multileaving (Schuth et al., 2014, 2015)
- Drawbacks
  - Benchmark: No absolute number for benchmarking
  - Interpretation: Unable to interpret much at the document-level, or about user behavior
- Hofmann et al. (2016)

Related tutorial(s) for more in-depth treatment at ESSIR

- Foundations of Evaluation 1, Nicola Ferro & Julio Gonzalo (today)
- Foundations of Machine Learning for IR, Claudia Hauff (Tue)
- Foundations of Contextual IR, Gabriella Pasi (Tue)
- Design and Evaluation of Recommender Systems, Paolo Cremonesi (Thu)
- Understanding & Inferring User Tasks and Needs, Emine Yilmaz (Fri)

Introduction

Front door

Evaluation

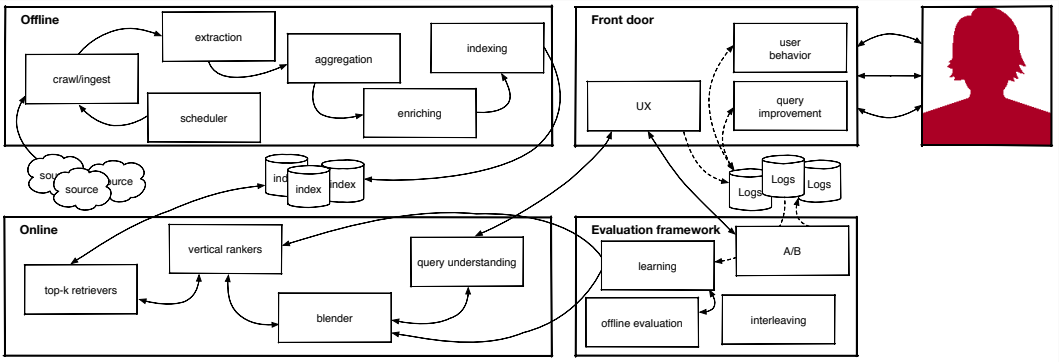
Online

Offline

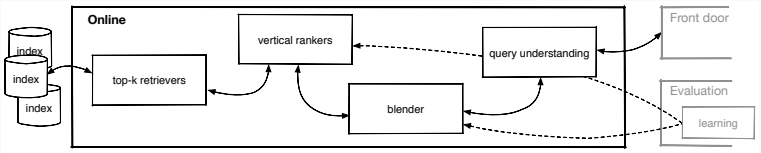
Wrap-up

Online

The big picture



Zooming in on online



Query understanding

Blending

Vertical rankers

Top-k retrieval

Learning to rank

Click models

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
  - *Milan* vs *Milan shopping*

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

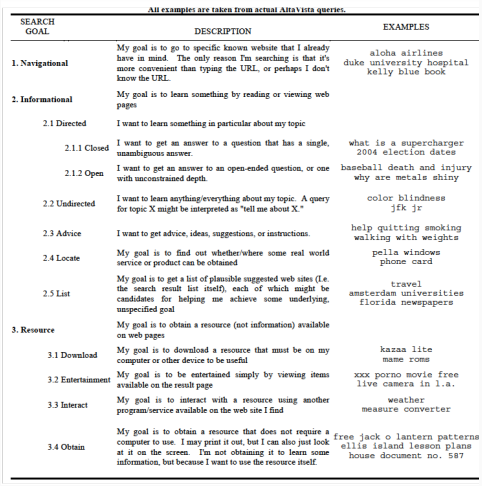
- Intent classification
- Intent discovery

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"

Understanding user goals in web search

- 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 credits: (Rose and Levinson, 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

- 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*

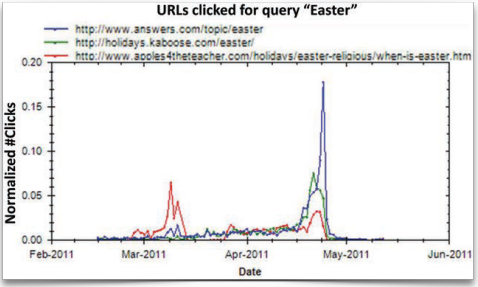


Image credits: (Radinsky et al., 2013)

Learning to detect intent shifts

- Queries whose intent shifts from non-fresh to fresh
- Aggregated search approach to freshness
  - A "fresh" vertical with a 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
- ...

(Lefortier et al., 2014)

Learning to detect intent shifts

- ...
- Online exploration for quick adaptation
- Multi-armed bandits
  - SERP = action; consider only actions that integrate fresh results on SERP differently
  - Action $\approx$ decide 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

(Lefortier et al., 2014)

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, KB panel (entity enrichment)
- Device tailoring

Diversity

- 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

(Radlinski et al., 2008)

Diversion

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.

- Study the search behavior of media studies scholars

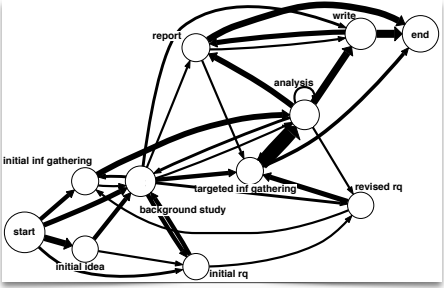


Image credits: (Bron et al., 2013, 2012)

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”)?

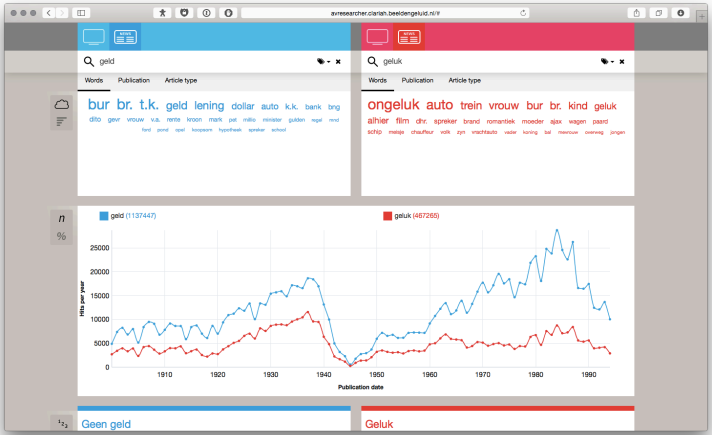


Image credits: (Bron et al., 2013, 2012)

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

Image credits: (Bron et al., 2013, 2012)

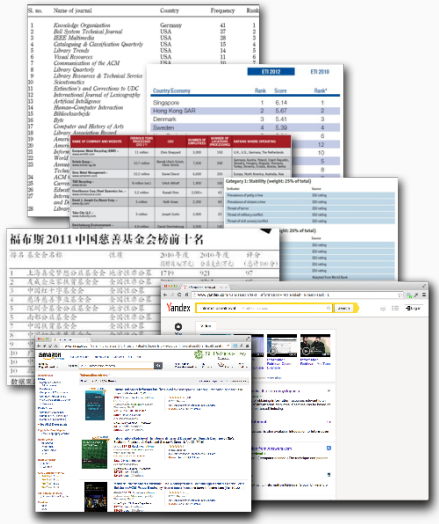
Vertical rankers

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, dwell time, . . .
- **Documents represented by feature vectors**
  - Combine features extracted for every (q, d) pair (i.e., score output by ranker)
  - Incorporate new retrieval model by including the model's output

Learning to rank

- 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)
- …

Learning to rank

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

Limitations of the Annotated Datasets

Some of the most substantial limitations of **annotated datasets** are:

- **expensive** to make (Qin and Liu, 2013; Chapelle and Chang, 2011).
- **unethical** to create in **privacy-sensitive settings** (Wang et al., 2016).
- **impossible** for small scale problems, e.g., **personalization**.
- **stationary**, cannot capture **future changes in relevancy** (Lefortier et al., 2014).
- **not necessarily aligned with actual user preferences** (Sanderson, 2010),
  i.e. annotators and users often disagree.

Limitations of the Supervised Approach

Annotated datasets are **valuable** and have an **important place in research and development**.

However, the supervised approach is:

- **Unavailable** for practitioners without a **considerable budget**.
- **Impossible** for certain ranking problems.
- Often **misaligned** with *true* user preferences.

Therefore, there is a **need** for an **alternative** learning to rank approach.

Learning from User Interactions: Advantages

Learning from user interactions solves the problems of annotations:

- Interactions are **virtually free** if you have users.
- User **behaviour** is indicative of their **preferences**.

User interactions also bring their **own difficulties**:

- Interactions give **implicit feedback**.

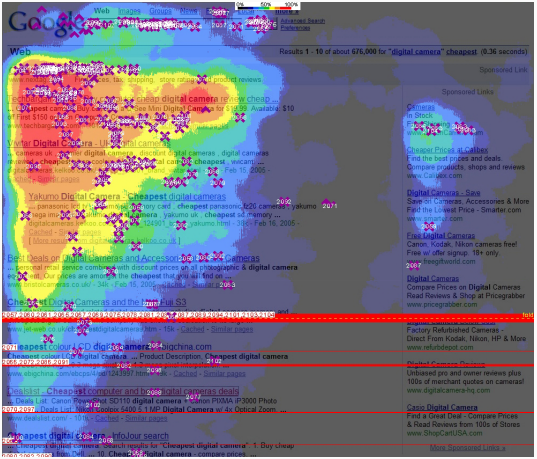
Learning from User Interactions: Difficulties

User interactions bring their **own difficulties**:

- **Noise:**
  - Users click for **unexpected reasons**.
  - Often clicks occur **not because** of relevancy.
  - Often clicks do not occur **despite** of relevancy.

- **Bias:** Interactions are affected by **factors other than relevancy**:
  - **Position bias:** **Higher ranked** documents get more attention.
  - **Item selection bias:** Interactions are **limited** to the **presented** documents.
  - **Presentation bias:** Results that are **presented differently** will be **treated differently**.
  - ...

The Golden Triangle



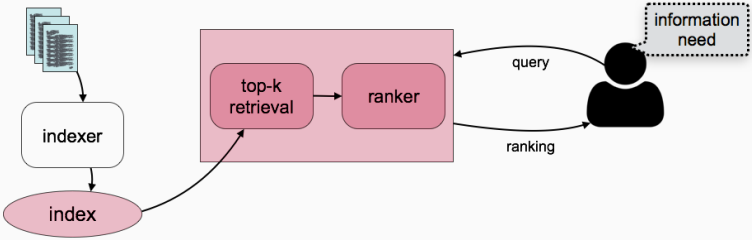
Source: http://www.mediative.com/

Learning from User Interactions: Goal

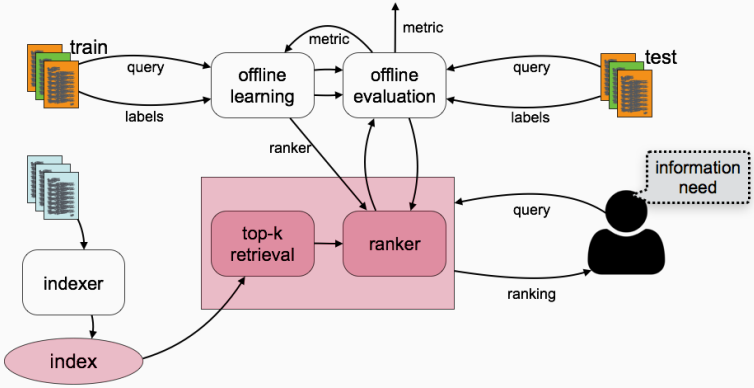
Goal of unbiased learning to rank:

- Optimize a ranker w.r.t. **relevance preferences** of users from their interactions.
- **Avoid** being **biased by other factors** that influence interactions.

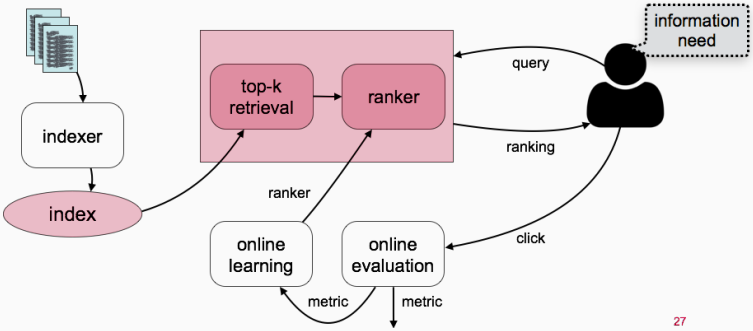
Zooming in


indexer
index
top-k
retrieval
ranker
query
ranking
information
need

Zooming in


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

Zooming in


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

Online learning to rank

Goal: let search engine take actions that maximize (discounted) cumulative reward

- action $\sim$ result ranking a_t
- context $\sim$ query (plus user?) x_t
- reward $\sim$ clicks r_t

$$C = \sum_{t=0}^{\infty} \gamma^{t-1} r_t(a_t)$$

$\gamma \sim$ discount factor

Challenges

- Need to generalize across context
- Need to infer feedback from noisy biased interactions (relative feedback)
- Need to deal with very large action spaces

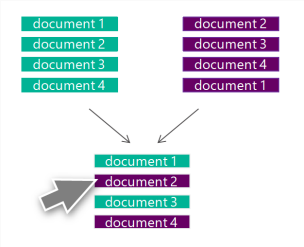
Online learning to rank

Example approach: learning from relative, listwise feedback:

- Comparison: interleaved comparisons to infer listwise relative feedback
- Learning: Dueling bandit gradient descent optimizes a weight vector for weighted linear combinations of ranking features
- Data reuse: use probabilistic interleave and importance sampling to compare candidate rankers on historical data and focus exploration

Interleaved comparisons

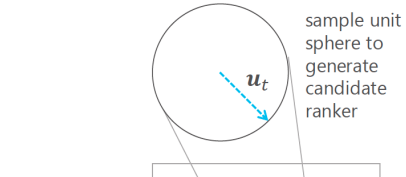
- Generate interleaved (combined) ranking
- Observe user clicks
- Credit clicks to original rankers to infer outcome $o \in \{-1, 0, 1\}$



Online learning to rank

Learning

- **Dueling bandit gradient descent** (DBGD) optimizes a weight vector for weighted linear combinations

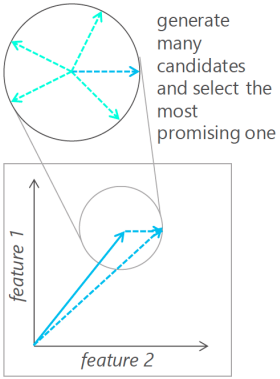


compare w_t and w'_t using interleaving; if w'_t wins, update: $w_{t+1} = w_t + \alpha u_t$

Online learning to rank

Data reuse

- Use probabilistic interleave and importance sampling to compare candidate rankers on historical data and focus exploration



Generate several candidate rankers, and select the best one by running a tournament on historical data

Use probabilistic interleave and importance sampling for ranker comparisons during the tournament

Recent developments

Dueling bandit gradient descent

- Unable to reach optimal performance in ideal settings
- Strongly affected by noise and position bias

Pairwise differentiable gradient descent

- Considerably outperforms DBGD
- Capable of reaching optimal performance in ideal setting
- Robust to noise and position bias

Oosterhuis and de Rijke (2018)

Counterfactual learning to rank

Can we evaluate a ranker **without deploying** it or **annotated data**?

- **Counterfactual evaluation**: Evaluate a new ranking function using historical interaction data (e.g., clicks) collected from a previously deployed ranking function
- **Counterfactual learning**: Learn a new ranking function using historical interaction data (e.g., clicks) collected from a previously deployed ranking function

CLTR – reaches higher level of performance in absence of position bias; and when there is little interaction noise

OLTR – More robust to noise; is able to address item selection bias as well

Jagerman et al. (2019b)

Related tutorial(s) for more in-depth treatment at ESSIR

- Approaches to Research in IR, W. Bruce Croft (next)
- Foundations of Machine Learning for IR, Claudia Hauff (Tue)
- Foundations of Personalized IR, Gabriella Pasi (Tue)
- Efficiency and Scalability in IR, Nicola Tonellotto (Wed)
- Recommender Systems, Paolo Cremonesi (Thu)
- Biases on Web Search and Recommender Systems, Ricardo Baeza-Yates (Fri)

Introduction

Front door

Evaluation

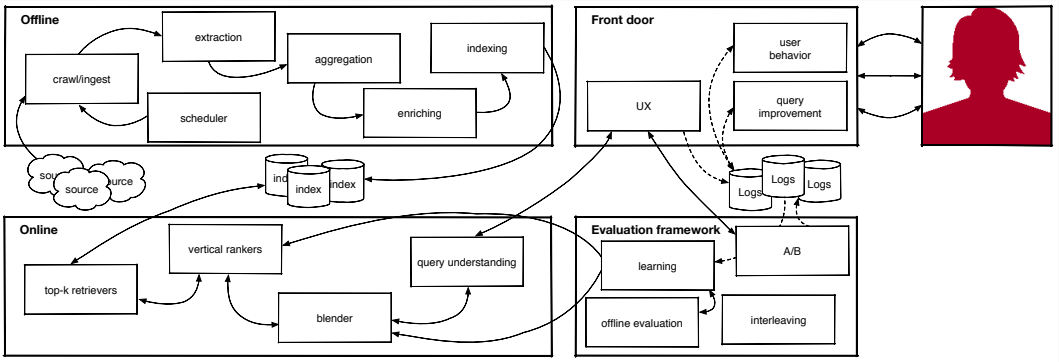
Online

Offline

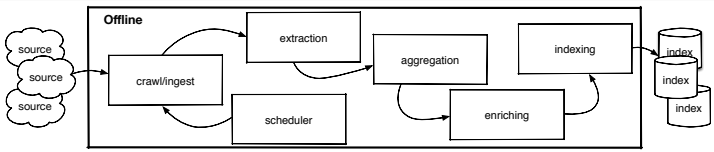
Wrap-up

Offline

The big picture



Offline



Getting content

Enriching

- 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...

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

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

Document analysis

Recognizing meaningful expressions

- People, products, locations, . . . , relations, properties

Sentiment analysis

- Opinions, emotions, framing

Typically tackled as a **supervised** learning problem

- Labeled data often language and genre dependent

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

Link-based analysis of document, item, . . . collections

Examples

- Hypertext links
- Citation links
- Visitation patterns
- Basket analysis
- Implicit links (e.g., based on co-occurring terms or entities)

Purpose

- Quality assurance
- Navigational suggestions
- Query modeling
- Recommendation
- Intent disambiguation

What to put in the index?

Text extraction and normalization

- File formats, encodings, accents and diacritics, determine useful content

Lexical analysis

- Tokenization, extract words, split compounds

Removing stopwords

- No-content words: *the*, *it*, *of*, ...

Morphological normalization

- Stemming lemmatization, n-gramming

Term selection

- "Nouns" only, controlled vocabulary, ...

Related tutorial(s) for more in-depth treatment at ESSIR

- Approaches to Research in IR, W. Bruce Croft (next)
- Efficiency and Scalability in IR, Nicola Tonellotto (Wed)
- Social Media Analytics, Carlos Castillo (Wed)
- Medical (text and image) IR, Henning Müller (Thu)
- Design and Evaluation of Recommender Systems, Paolo Cremonesi (Thu)
- Biases on Web Search and Recommender Systems, Ricardo Baeza-Yates (Fri)

Introduction

Front door

Evaluation

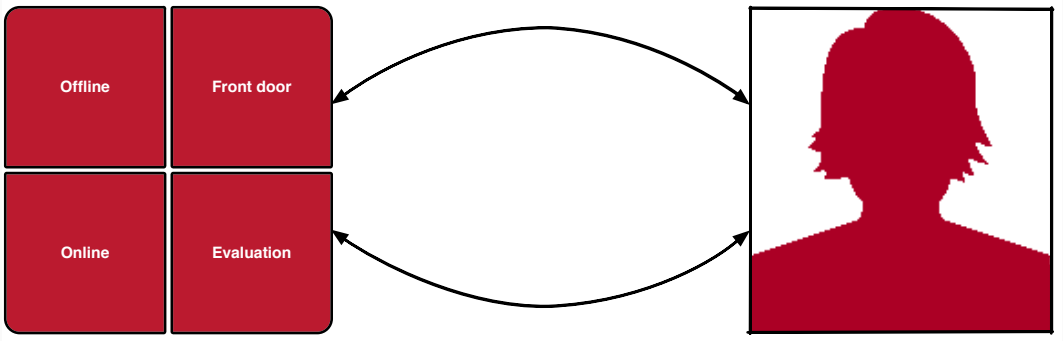
Online

Offline

Wrap-up

Wrap-up

The bigger picture: IR systems as interactive systems



What does this mean for machines?

Life is easier for systems than in an offline trained query-response paradigm

- Engage with user
- Educate/train user
- Ask for clarification from user

Life is harder for systems than in an offline trained query-response paradigm

- **Safety** – Don't hurt anyone
- **Explicability** – Be transparent about model, about decisions

Unpacking safety

Safely learn to re-rank online by combining strengths of online and counterfactual LTR – (Jagerman et al., 2019a)

Provably safe online LTR – (Li et al., 2019a)

Deep learning from logged feedback to learn better/faster – (Joachims et al., 2018)

Learn more accurate reward signals for response generation to ensure informativeness – (Li et al., 2019b)

Come up with realistic theoretical guarantees for high performing online LTR methods – (Oosterhuis and de Rijke, 2019)

...

Unpacking explicability

How does it work? vs. **How did we arrive at this decision?**

Faithfully explaining rankings in a news recommender – (Ter Hoeve et al., 2017)

Contextualize knowledge graph facts – (Voskarides et al., 2018)

Visually explain outfit recommendations in fashion search – (Lin et al., 2019)

Find influential training samples for your learning to search, recommend, converse method – (Sharchilev et al., 2018)

Explain the errors in your search, recommendation, conversation method – (Lucic et al., 2019)

...

Stuff you should work on

Think of information retrieval systems as **interactive systems**

- Large-scale understanding of users and **user behavior**
- Higher level **models** of any aspect of search, recommendation, conversation
- Online **anything**
- Buy in to deep learning
- Work on **safety**
- Work on **explicability**
- Be creative

References i

A. Borisov, I. Markov, M. de Rijke, and P. Serdyukov. A neural click model for web search. In *WWW '16*, pages 531–541. ACM, April 2016.

M. Bron, M. de Rijke, J. van Gorp, A. Vishneuski, F. Nack, and 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, and M. de Rijke. Aggregated search interface preferences in multi-session search tasks. In *SIGIR '13*, 2013.

F. Cai and M. de Rijke. A survey of query auto completion in information retrieval. *Foundations and Trends in Information Retrieval*, 10(4):273–363, September 2016.

F. Cai, S. Liang, and M. de Rijke. Time-sensitive personalized query auto-completion. In *CIKM 2014: 23rd ACM Conference on Information and Knowledge Management*. ACM, November 2014.

O. Chapelle and Y. Chang. Yahoo! Learning to Rank Challenge Overview. *Journal of Machine Learning Research*, 14:1–24, 2011.

M. A. Hearst. *Search User Interfaces*. Cambridge University Press, 2009.

References ii

K. Hofmann, L. Li, and F. Radlinski. Online evaluation for information retrieval. *Foundations and Trends in Information Retrieval*, 10(1):1–117, 2016.

R. Jagerman, I. Markov, and M. de Rijke. When people change their mind: Off-policy evaluation in non-stationary recommendation environments. In *WSDM 2019: 12th International Conference on Web Search and Data Mining*, pages 447–455. ACM, February 2019a.

R. Jagerman, H. Oosterhuis, and M. de Rijke. To model or to intervene: A comparison of counterfactual and online learning to rank from user interactions. In *42nd International ACM SIGIR Conference on Research & Development in Information Retrieval*, page (to appear). ACM, 2019b.

T. Joachims. Evaluating retrieval performance using clickthrough data. In *Text Mining*. Physica/Springer, 2003.

T. Joachims, A. Swaminathan, and M. de Rijke. Deep learning with logged bandit feedback. In *ICLR 2018*, April 2018.

D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1–2):1–224, 2009.

References iii

J. Kiseleva and M. de Rijke. Evaluating personal assistants on mobile devices. In *1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17)*. ACM, August 2017.

R. Kohavi. Online Controlled Experiments, 2013.

D. Lefortier, P. Serdyukov, and M. de Rijke. Online exploration for detecting shifts in fresh intent. In *CIKM 2014: 23rd ACM Conference on Information and Knowledge Management*. ACM, November 2014.

C. Li, B. Kveton, T. Lattimore, I. Markov, M. de Rijke, C. Szepesvari, and M. Zoghi. Bubblerank: Safe online learning to re-rank via implicit click feedback. In *UAI 2019: Conference on Uncertainty in Artificial Intelligence*, July 2019a.

Z. Li, J. Kiseleva, and M. de Rijke. Dialogue generation: From imitation learning to inverse reinforcement learning. In *AAAI 2019: 33rd AAAI Conference on Artificial Intelligence*. AAAI, January 2019b.

References iv

Y. Lin, P. Ren, Z. Chen, Z. Ren, J. Ma, and M. de Rijke. Improving outfit recommendation with co-supervision of fashion generation. In *The Web Conference 2019*, pages 1095–1105. ACM, May 2019.

A. Lucic, M. de Rijke, and H. Haned. Contrastive explanations for large errors in retail forecasting predictions through monte carlo simulations. In *XAI: IJCAI 2019 Workshop on Explainable Artificial Intelligence*, August 2019.

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

H. Oosterhuis and M. de Rijke. Differentiable unbiased online learning to rank. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1293–1302. ACM, 2018.

H. Oosterhuis and M. de Rijke. Optimizing ranking models in an online setting. In *ECIR 2019: 41st European Conference on Information Retrieval*, pages 382–396. Springer, April 2019.

T. Qin and T.-Y. Liu. Introducing letor 4.0 datasets. *arXiv preprint arXiv:1306.2597*, 2013.

References v

K. Radinsky, F. Diaz, S. T. Dumais, M. Shokouhi, A. Dong, and Y. Chang. Temporal web dynamics and its application to information retrieval. In *WSDM 2013*, pages 781–782, 2013.

F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *ICML 2008*, pages 784–791, 2008.

D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, pages 13–19. ACM, 2004.

M. Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010.

A. Schuth, F. Sietsma, S. Whiteson, D. Lefortier, and M. de Rijke. Multileaved comparisons for fast online evaluation. In *CIKM*, pages 71–80. ACM, 2014.

A. Schuth, R.-J. Bruintjes, F. Büttner, J. van Doorn, et al. Probabilistic multileave for online retrieval evaluation. In *SIGIR*, pages 955–958. ACM, 2015.

B. Sharchilev, Y. Ustinovsky, P. Serdyukov, and M. de Rijke. Finding influential training samples for gradient boosted decision trees. In *ICML 2018: International Conference on Machine Learning*, pages 4584–4592, July 2018.

M. Ter Hoeve, M. Heruer, D. Odijk, A. Schuth, M. Spitters, R. Mulder, N. van der Wildt, and M. de Rijke. Do news consumers want explanations for personalized news rankings? In *FATREC Workshop on Responsible Recommendation*, August 2017.

N. Voskarides, E. Meij, R. Reinanda, A. Khaitan, M. Osborne, G. Stefanoni, K. Prabhanjan, and M. de Rijke. Weakly-supervised contextualization of knowledge graph facts. In *SIGIR 2018: 41st international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 765–774. ACM, July 2018.

X. Wang, M. Bendersky, D. Metzler, and M. Najork. Learning to rank with selection bias in personal search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 115–124. ACM, 2016.

Acknowledgments