Search and Discovery in Personal Email Collections

Michael Bendersky
Marc Najork
Xuanhui Wang
Donald Metzler
(Google Research)

https://github.com/bendersky/sdpe-tutorial

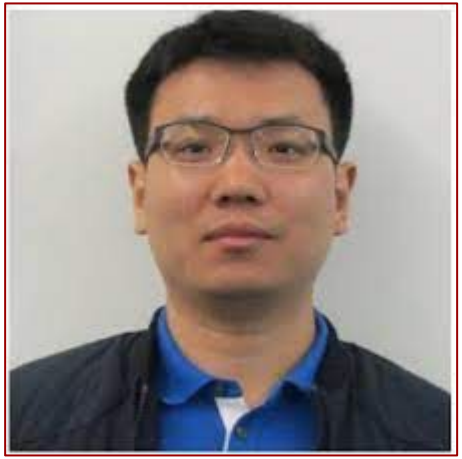
Presenters

Michael Bendersky

Donald Metzler

Co-authors

Marc Najork



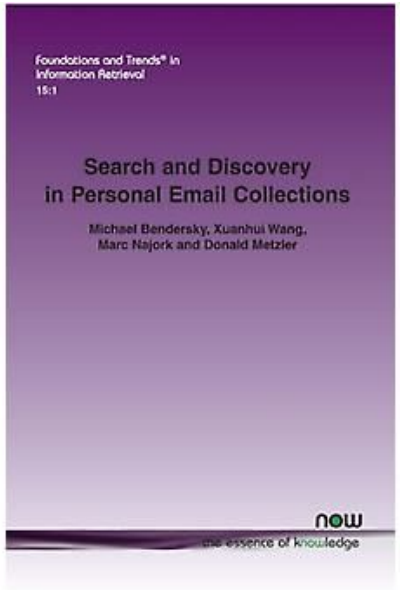
Xuanhui Wang

Foundations and Trends® in Information Retrieval Survey

Foundations and Trends® in Information Retrieval > Vol 15 > Issue 1

"Search and Discovery in Personal Email Collections"

By: Michael Bendersky, Xuanhui Wang, Marc Najork, Donald Metzler



- ***This tutorial covers most (but not all) of the main material of the survey***
- ***To improve coverage, we focus on breadth, rather than depth***
- ***We provide pointers to relevant papers***
- ***More in-depth technical details can be found in the survey***

Tutorial Outline

Part I

(1) Introduction (30 min)

(2) The Anatomy of an Email Search Engine (50 min)

Part II

(3) Managing and Learning from User Data (50 min)

(4) The Next Frontier (30 min)

Part III

(5) Discussion (20 min)

Introduction

Some Stats

- Estimated ***300B*** messages a day, ***4B*** users (Radicati Group Report, 2019)
- Personal communications (Maarek, 2017)
  - Recent decline in interpersonal communications
  - Large increase in ***machine-generated messages***
    - store promotions, newsletters, receipts and bills
- Enterprise communications (Naragon, 2018)
  - A survey of 1,000 US employees
  - On average, ***3*** hours checking work email during the week
  - Email is still the most preferred communication medium

Managing Email Overload

- Management strategies
  - **Pilers** – forego foldering and rely on search
  - **Filers** – minimize the number of inbox messages by foldering
  - **Spring Cleaners** – occasionally organize their email

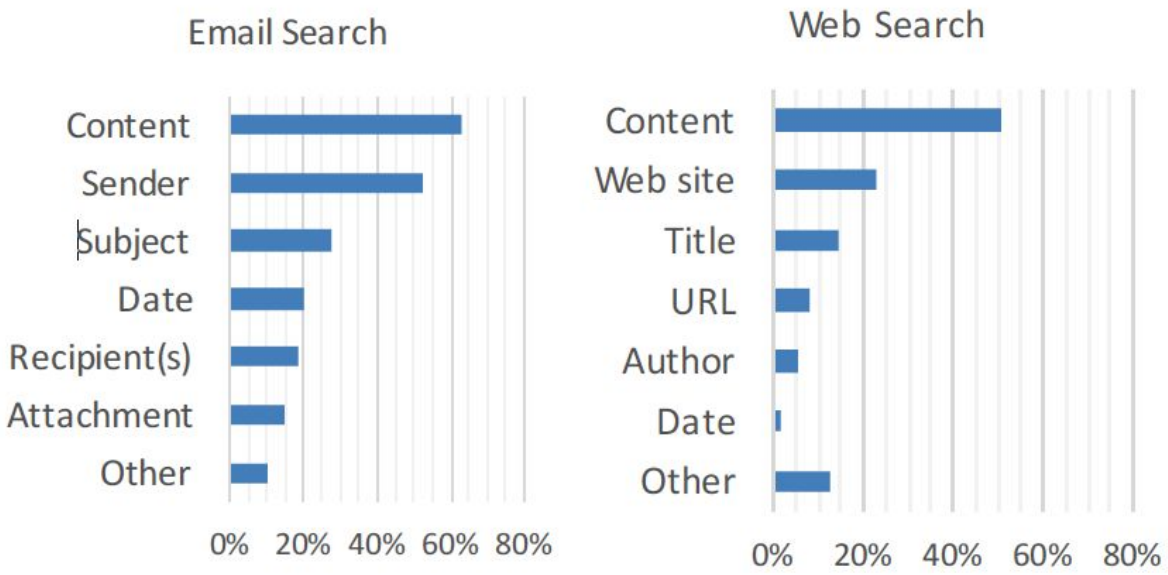
- Distribution of email activity for 300,000 Microsoft employees (Narang et al., 2017)
  - **20% – 35%** of all activity involves organization,
  - **10% – 20%** of all activity involves search
  - % of search activity increases proportionally to the mailbox size

Email Revisiting

Type of Information	Percent
Instructions to perform a certain task	24.1%
A document (e.g., attachment, link)	22.0%
An answer to a question that was previously asked	16.3%
status update	10.2%
A solution to a problem	9.0%
A task request to you	4.9%
A person/customer (e.g., contact information)	2.0%
An appointment/event	2.0%
Machine generated message (e.g., reservation)	0.8%
Other	8.6%

Distribution of information types users are looking for in email revisits, as reported in a survey of 395 corporate email users (Alrashed et al., 2018).

Attribute Recall

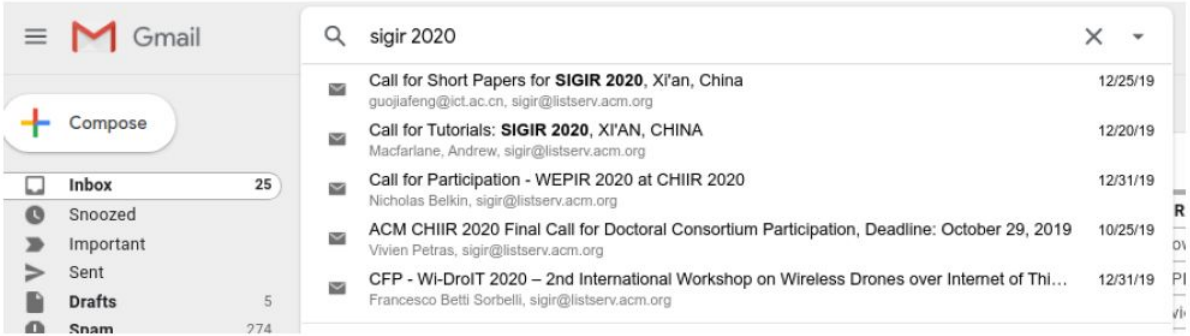


Percentage of searchers who remembered certain attributes, compared between email and web searches, based on a survey of 324 regular email users, conducted by Ai et al. (2017).

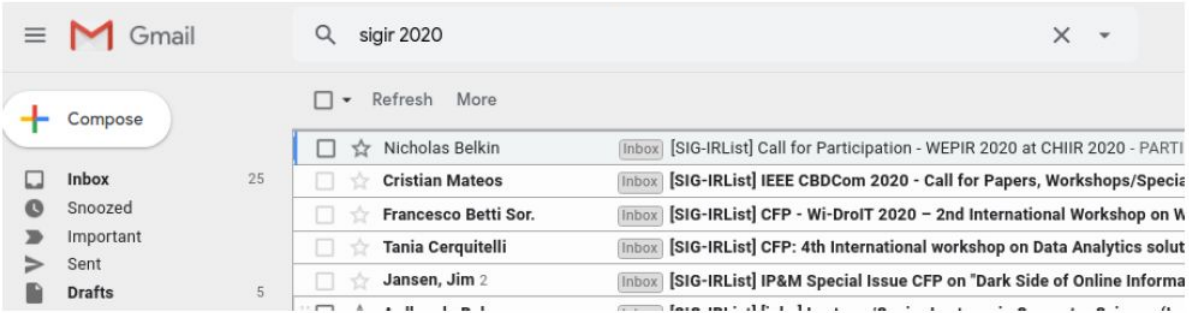
Search Interfaces

- Chronological ordering is quite central to email search experience
  - contextual cues play a critical role in recalling personal information
  - the most important contextual cue of is time
  - majority of users prefer chronological ordering regardless of the initial order (Dumais et al., 2003)

- Relevance ordering
  - Facilitates discovery of ***older messages***
    - hard to remember their context
    - ranked low chronologically
  - Allows inexact matching and increases recall



(a) **Relevance**

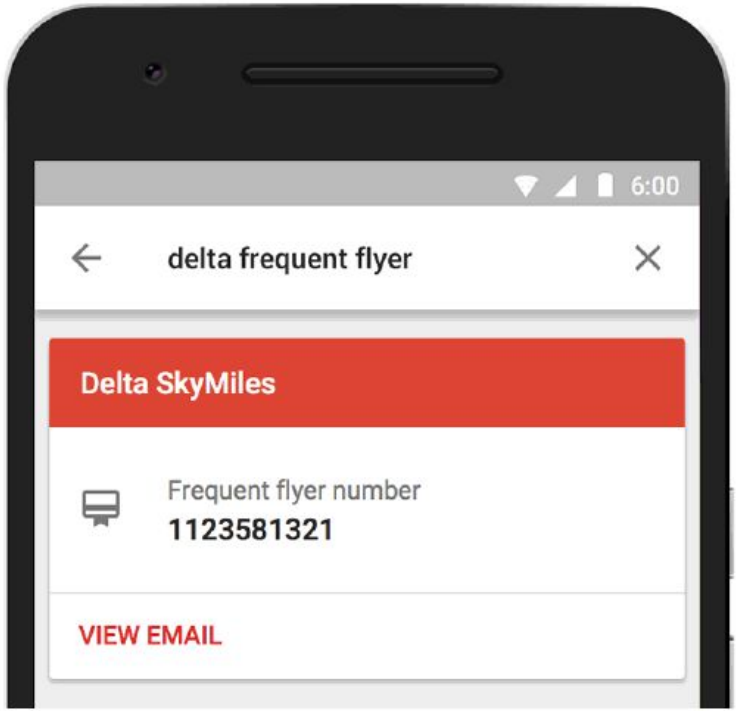


(b) **Chronological**

Relevance v. Chronological

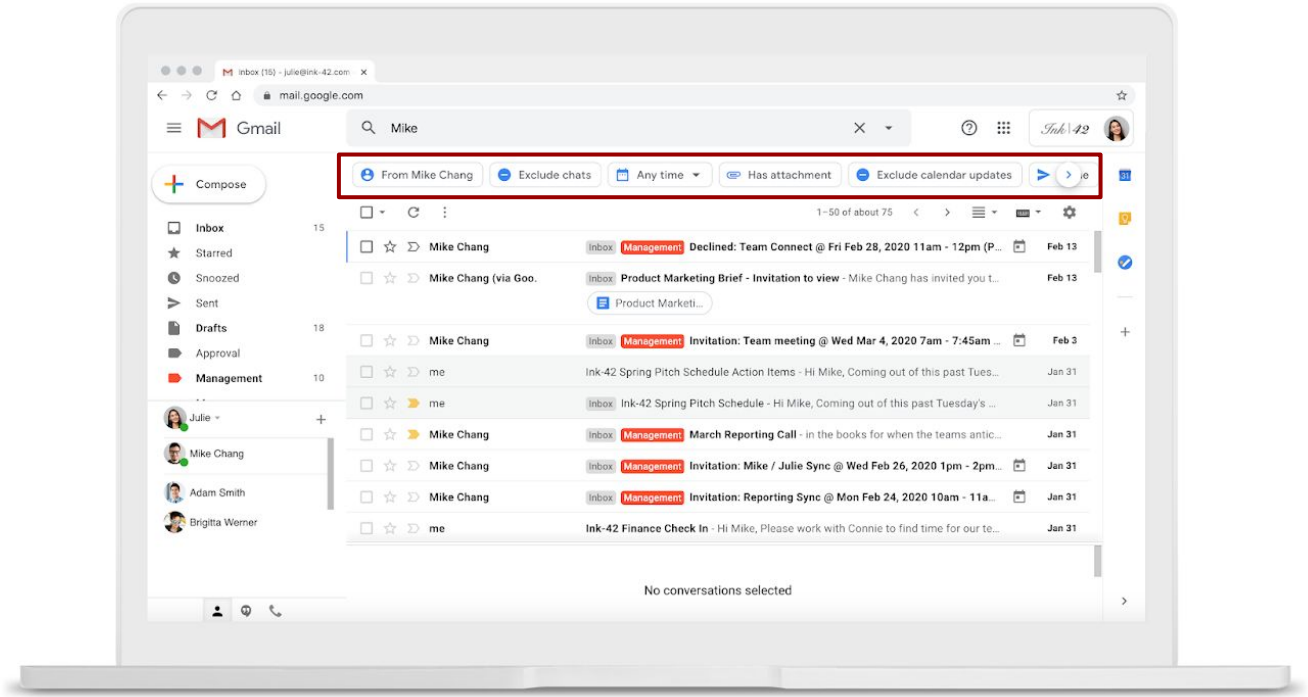
An illustrative example of hybrid "heroes" relevance results (HList) followed by chronological results (TList) displayed on a mobile device (Carmel et al., 2017a).

Hybrid



Inbox by Gmail (a defunct service) directly surfacing the relevant answer

Knowledge panels



Search chips that allow filtering by contextual cues

Search Chips

Personal Email Search v. Web Search

Corpus Size

- Billions of public pages v. individual mailboxes
- *Zero-result queries* and *recall* are important issues

Personal Email Search v. Web Search

Links and anchor text

- No anchor text / link graphs available
- No cross-reference across mailboxes
- Message metadata is more important
    - sender information, attachments, stars, labels, etc.

Personal Email Search v. Web Search

Implicit user feedback

- No cross-user interactions with the same email message
- Click data is highly sparse
  - techniques to aggregate it were developed[†]

—

[†]More on this later

Personal Email Search v. Web Search

Content and query dynamics

- web search exhibits a wide spectrum of content and query dynamics
  - (a) zoom – intent zooms in on to the current event
  - (b) shift – intent undergoes a gradual shift over time
  - (c) static – relatively stable intent over time.

- email search exhibits a very strong recency bias

Personal Email Search v. Web Search

Adversarial content, spam, low-quality

- spam, phishing and other adversarial content are grouped under "Spam", and are generally excluded from search
- promotional or group messages may be demoted

Personal Email Search v. Web Search

Search tasks

- **General tasks** - users have a broad idea of what they are trying to achieve, but do not have a particular document in mind.
- **Specific tasks** - users are looking for a particular document addressing their information need
- Web Search
  - a mixture of general and specific tasks
- Email Search
  - largely specific tasks

Personal Email Search v. Web Search

Data privacy

- Both email message and user queries are private
- Researchers ***must*** employ privacy-preserving techniques when examining any email or query content

The Rest of the Tutorial at a Glance

- The anatomy of an email search engine
  - Architecture
  - Indexing
  - Retrieval & Ranking
  - Query & Document Understanding
  - Metrics and Collections
- Managing and Learning from User Data
  - Anonymity Principles: de-identification, k-anonymity, and differential privacy
  - Learning from biased click data
- The New Frontier
  - Recommendations and task assistance
  - Assisted composition
  - New modes of personal search: conversational, multi-modal, etc...

Additional Material in the Survey

- Mailbox organization
  - Content filtering (spam, phishing)
  - Categorization
  - Clustering
- Content extraction
  - Task detection
  - Field extraction from templated emails
- Deeper dive-in on query understanding models for search
  - auto-complete
  - spelling correction
  - query expansion
- Advanced models on position bias correction in personal search

The Anatomy of an Email Search Engine

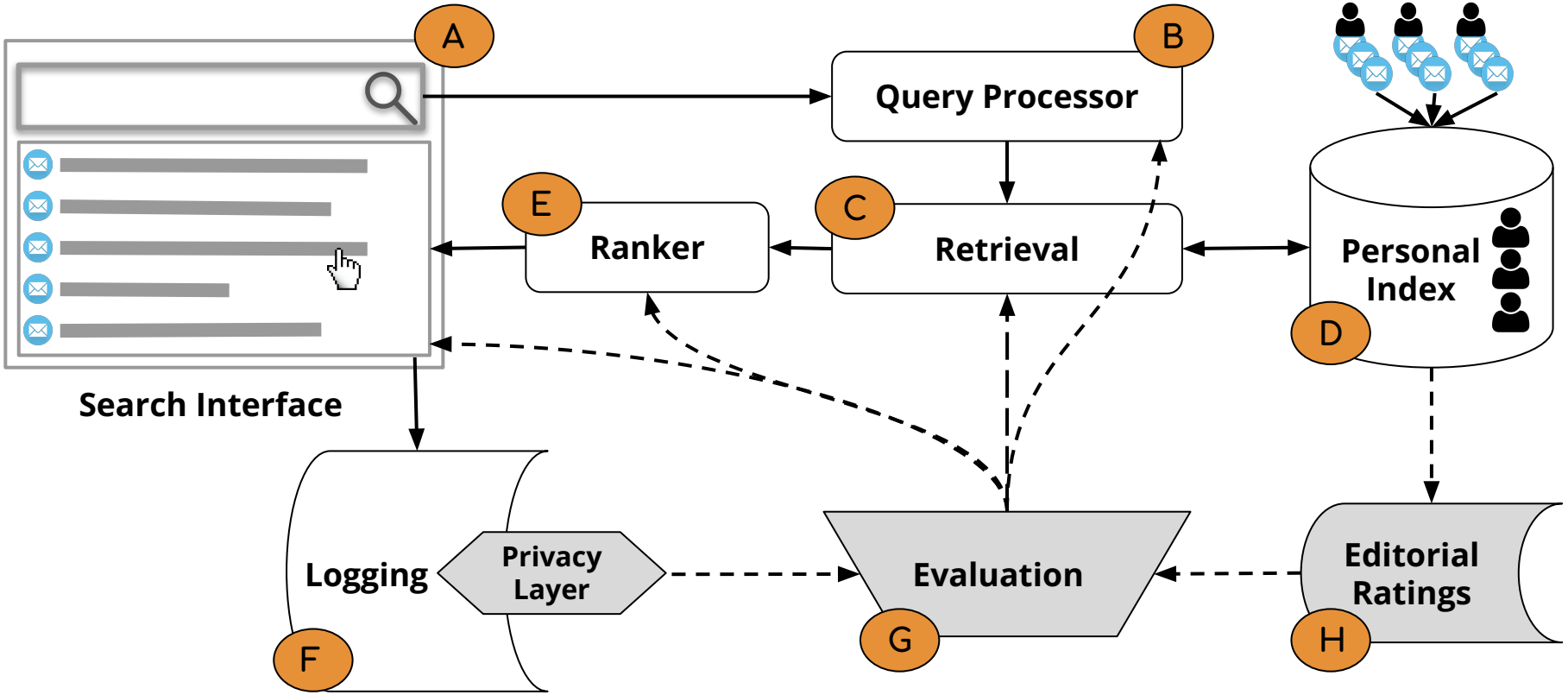


Figure 1: An end-to-end email search engine architecture. Solid arrows demonstrate the online flow of a search request. Dashed arrows and shaded shapes indicate the offline evaluation process.

Known-Item Search Metrics

- Often (but not always) email search can be modeled as **a known-item search**
  - *The searcher is attempting to recall a single, known in advance email message*
  - *May be measured through editorial judgments or clicks*

- $\texttt{rank}_i$ – the rank at which the relevant / clicked message is retrieved
- **N** – number of evaluated queries

$$MRR = \frac{\sum_{i=1}^{N} \frac{1}{rank_i}}{N} \qquad success@k^{\dagger} = \frac{\sum_{i=1}^{N} \mathbb{I}(rank_i \leq k)}{N}$$

—

† For click-based evaluation, `success@k` is equivalent to `CTR` when `#shown ≤ k`

Personalized Metrics

- User-centric measurement
  - *Normalizes a metric with respect to user's historical behavior*
  - *Rewards changes with the most effort savings*
  - *Shown to better discriminate between ranking changes in A/B experiments*

Average historic click rank for user issuing query *i*

$$pMRR = \frac{\sum_{i=1}^{N} \frac{p_i}{rank_i}}{\sum_{i=1}^{N} p_i}, \quad p_i = \log\left(\frac{\overline{rank_i}}{rank_i} + 1\right)$$

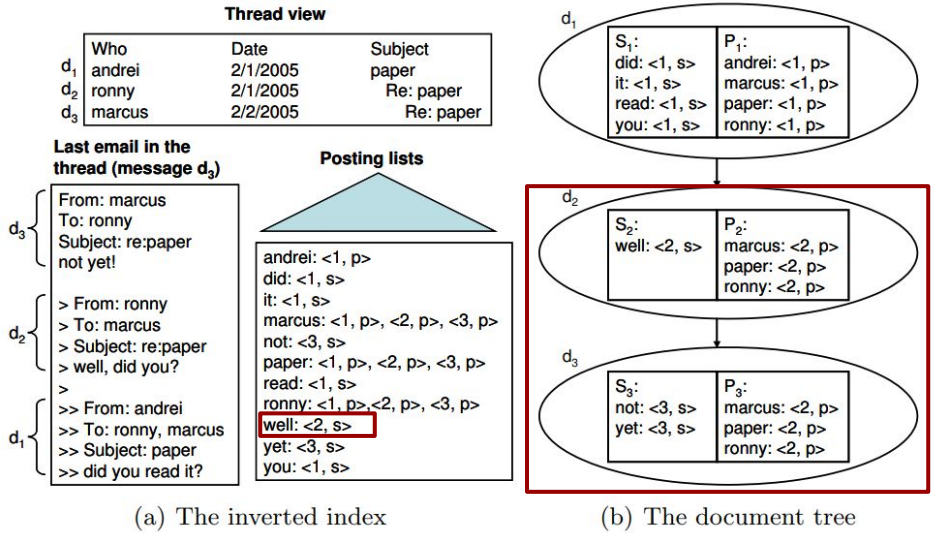
(Ashkan & Metzler, 2019)

Private Indexing

- The main difference from the web search is the ***privacy of the index***.
- Searchers have access only to their mailboxes, so the index should be partitioned
  - **Centralized Solution I** Maintain separate per-user indices
    - Short posting lists
    - A lot of vocabulary duplications
  - **Centralized Solution II** Maintain a single index, filter at retrieval time
    - More space efficient / better compression
    - Longer posting lists
  - **Distributed Solution** The user indices are stored locally
    - No need for "trusted host" and low latency search
    - Updates across devices are required
    - Harder to deploy algorithmic updates

Thread Indexing

Query: ***well***



Retrieved list: ***d2, d3***

<s> - shared content
<p> - private content

(Broder et al., 2006)

Retrieval – Chronological Sort

- One-pass matching algorithm
  a. retrieve all emails such that ***all*** query terms appear in at least one of the fields
  b. sorting the retrieved documents in descending received timestamp order

- Advantages
  a. simple, efficient & scalable
  b. generally high quality results that meet users expectations
     - *with multiple sort options, users still issue more queries in which they sort the results by* ***`date`*** *(Dumais et. al, 2003)*

- Disadvantages
  a. vocabulary mismatch may lead to zero-result queries
  b. may not work as well for longer, more complex queries
  c. low ranks for older relevant emails

Retrieval – Search Operators

`from:alice@mail.com subject:dinner photos after:2020/01/01 size:10M`

- Advantages
  a. may help to ***explicitly*** and ***precisely*** express complex information needs
  b. can be efficiently implemented using ***field restricts***

- Disadvantages
  a. rarely used explicitly
     - *<5% of queries use search operators beyond* ***`from:`*** *(Ai et al., 2017)*
  b. may not be adequate for all information needs
     - *e.g., I don't remember the exact subject, sender*

Relevance Retrieval

- Partial query matching to increase the number of returned results
  - may surface irrelevant results, esp. in longer queries

- Fielded match (Ogilvie & Callan, 2005)
  - models an email as a structured document
  - allows for principled field weight assignment

$$P(w|\theta_e) = \sum_f \lambda_f P(w|\theta_{MLE(f)})$$

$$P(Q|\theta_e) = \prod_{i=1}^{|Q|} P(q_i|\theta_e)$$

  - still does not model **non-match** features

Relevance Ranking

- Apply a ***learning-to-rank*** (LTR) approach
  - Retrieve a ***high-recall*** initial result set using partial match
  - For each result extract a set of features
  - Re-rank the set using a machine-learned function defined over the features, optimizing ***precision***
  - Training data may be obtained using
    - explicit relevance judgments
      - *difficult to obtain, especially due to the personal nature of email*
    - click data
      - *may be biased,[†] but abundant*

—

[†]More on this later

(Carmel et al., 2015)

Relevance Ranking

LTR Features		
Feature Type	**Description**	**Example**
Sender	*sender-searcher affinity*	*communication counts, sender search counts*
Recipient	*characteristics of the recipient group*	*is searcher in to / cc / mailing list?*
Message	*attributes of the email message.*	*freshness, system labels, attachment types*
Action	*actions performed on the message*	*opens, replies, stars, spam assignments*
Query	*topical query-message similarity*	*BM25, term overlap, semantic similarity*
Searcher	*search personalization*	*situational context, e.g., time & location*
Interactions	*click-based features*	*clicks on this and similar messages*

(Carmel et al., 2015; Zamani et al., 2017)

Relevance Ranking

Editorial Evaluation

Algo	*NDCG@10 (+lift %)*	*p@1*	*p@3*	*p@5*	*p@10*
Time	0.4936	0.5540	0.4420	0.3920	0.2962
REX	0.6647 (+34.66%)	0.6960	0.6073	0.5352	0.4244

Click-based Evaluation

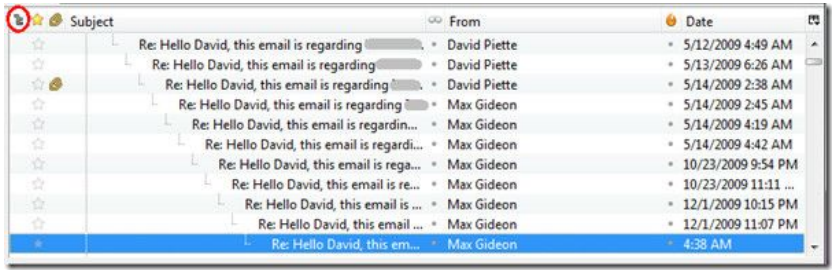
Algo	*MRR (+lift %)*	*success@1*	*success@3*	*success@5*	*success@10*
Corporate email dataset					
Time	0.3722	0.2238	0.4213	0.5416	0.7037
REX$(fresh. + sim.)$	0.4261 (+14.48%)	0.2748	0.4887	0.6028	0.7509
REX$(fresh. + sim. + actions)$	0.4550 (+22.24%)	0.2999	0.5253	0.6421	0.781
REX$(fresh. + sim. + actions + sender)$	0.4548 (+22.19%)	0.2994	0.5263	0.6419	0.7837
Web email dataset					
Time	0.3717	0.2282	0.4290	0.5264	0.6783
*$REX(fresh. + sim.)$	0.3785 (+1.81%)	0.2316	0.4406	0.5419	0.6909
REX$(fresh. + sim. + actions)$	0.4238 (+14%)	0.2711	0.4925	0.6004	0.7436
REX$(fresh. + sim. + actions + sender)$	0.4258 (+14.55%)	0.2731	0.4959	0.6000	0.7427

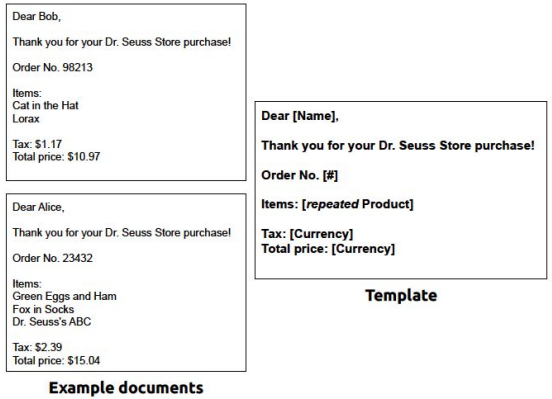
(Carmel et al., 2015)

Document Understanding

- Emails are similar to other structured documents
  - Have a structured header including sender information, sent date, subject, any attachments, etc.

- Except...





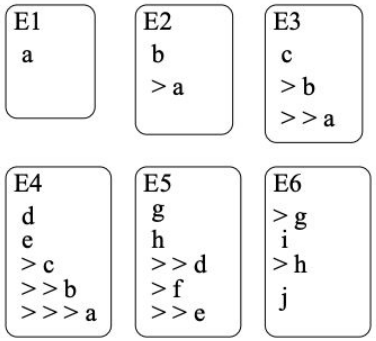
Threads & Other Unstructured Data

Structured Templates

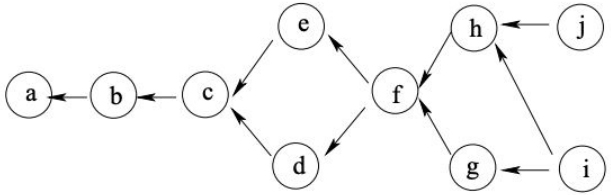
Latent Structure in Unstructured Emails

- **Signature Block Extraction**
  - *set of lines that contain sender signature, including name, phone number, affiliation, etc.*
- **Email Segmentation**
  - *Identifying various zones in an unstructured email*
- **Named Entity Extraction & Contact Resolution**
  - *e.g., Bob, Mr. Bruce → robert.bruce@enron.com*
- **Boilerplate Text Removal**
  - *headers, quotations and signature blocks*
- **Thread Structure Resolution**
  - *Converting unstructured threads to a structured representation, eg., a graph*

Complex Thread Resolution



(a) Conversation involving 6 Emails

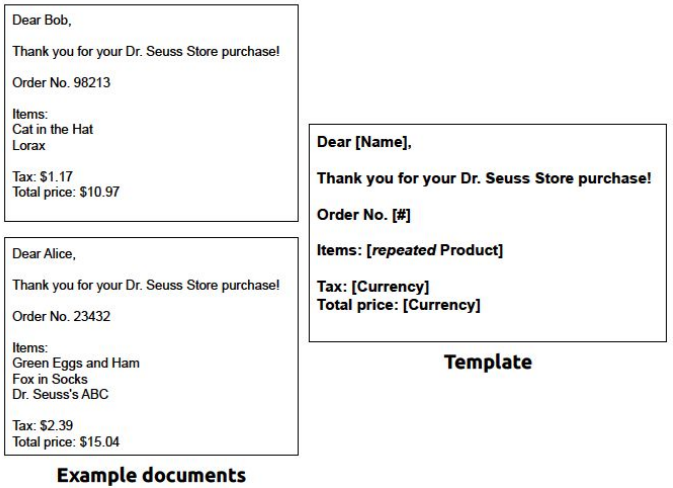


(b) Fragment Quotation Graph

- Lexical chaining via semantic similarity
  - Word co-occurrence (Carenini et al., 2007)
  - Latent topics, e.g., SVD (Wan and McKeown, 2004)

Templatization

- Common in B2C communications
  - *Travel itineraries*
  - *Bills*
  - *Receipts*
  - *...*
- Comprises a large (and growing) portion of personal email traffic



Web Mail is not Dead! It's Just Not Human Anymore

Yoelle Maarek
Yahoo Research
MATAM Park, Haifa 31905, Israel
yoelle@yahoo.com

WWW 2017 Keynote

Template Induction - Subject Templates

- Group emails by bulk senders
  - e.g., `shipment-tracking@amazon.com`
- Group subject lines into regexps
  - Replace with a wildcard
    - long numbers,
    - proper names,
    - unique identifiers
    - words with probability below a certain threshold per sender
  - `Your order #1123-222 has shipped on 01/02/2022`
    → `Your order * has shipped on *`
- Match emails to unique template IDs:
  - `hash(shipment-tracking@amazon.com, "Your order * has shipped on *")`

(Ailon et al., 2013)

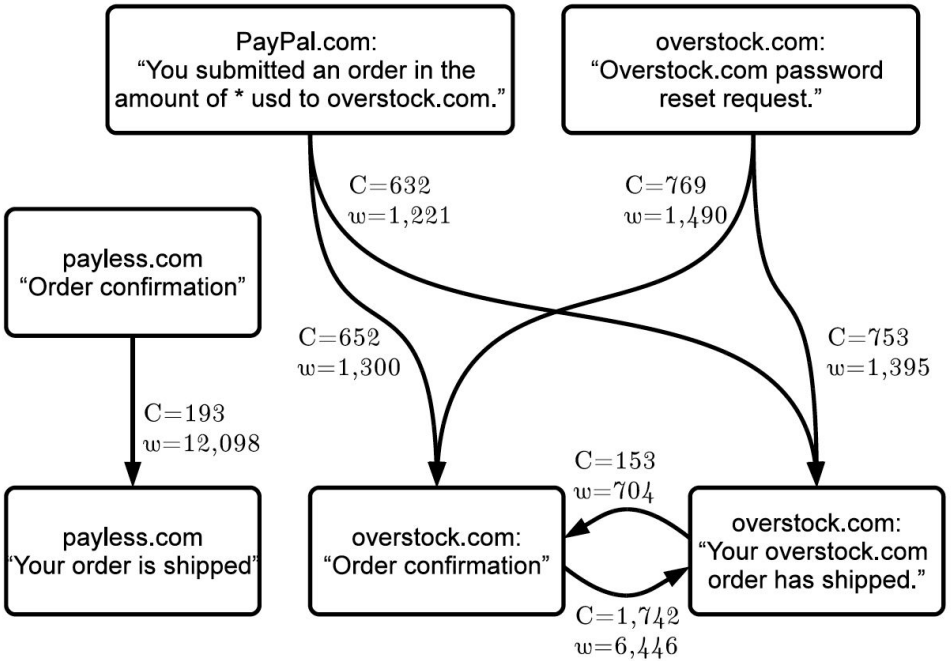
Template Induction – Body Templates

- Group emails by bulk senders
  - e.g., `shipment-tracking@amazon.com`
- Use body XPaths to group emails
  - expressions that specify a full path from the document root to some target node in an HTML document
- Group emails into `<sender, XPath list>` tuples
  - Collapse repetitions to group together receipts with different number of items
  - Further collapse XPaths with small pairwise edit distance
- Success Metrics
  - Increases extraction coverage by 25% → *how many emails can be templatized?*
  - Reduces extraction success rate by 2% → *how many template have useful information?*

(Avigdor-Elgrabli et al. 2016)

Template Uses – Threading

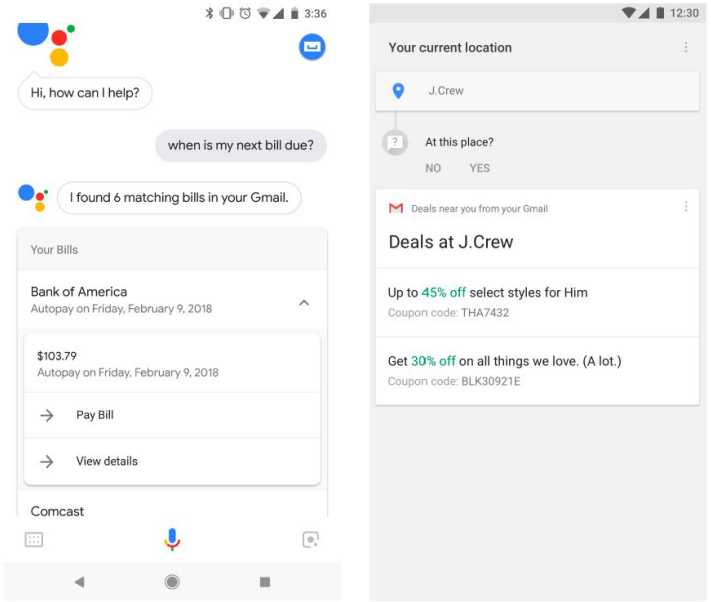
- Causal threading of sequences of machine-generated emails.



C **- conditional template counts**
w **- causal connection weights**
- ***How much do the conditional counts deviate from a Poisson distribution (random co-occurrence)?***

(Ailon et al., 2013)

Template Uses – Extraction

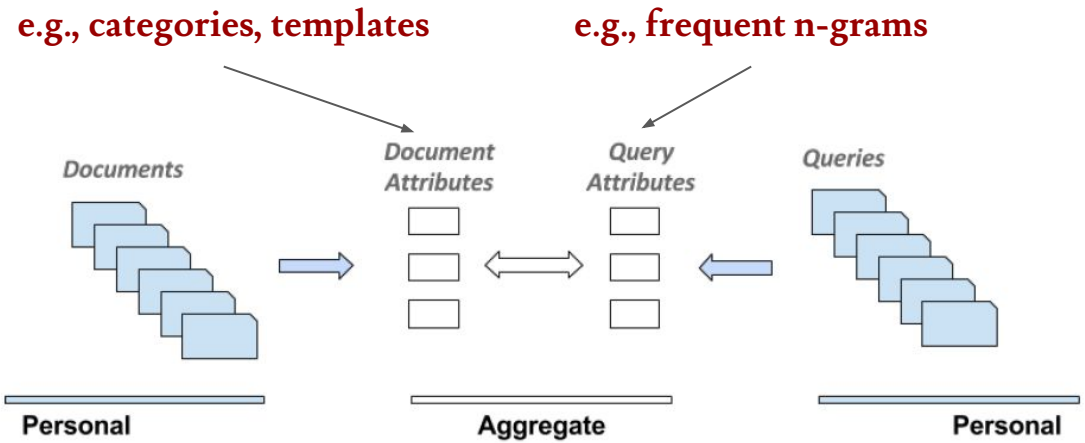


Google Assistant responding to a user query for their recent bills, and proactively displaying deals extracted from Gmail when the user enters the relevant store

(Sheng et al., 2013)

Template Uses – Search

- Aggregating sparse click data through templates (and other attributes)



(Bendersky et al., 2017)

Template Uses – Search

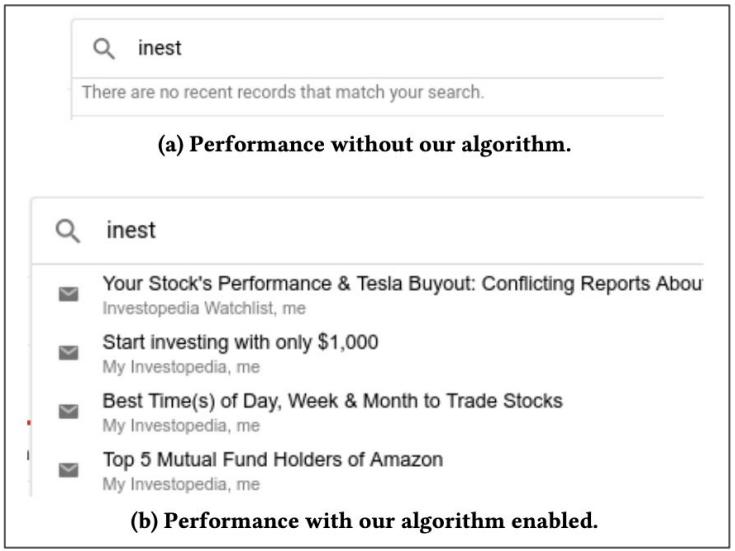
- Incorporating attribute-based aggregates into relevance ranking

	ΔMRR	
Attribute Type	Query-Independent	Query-Dependent
Categories	+0.48*	+0.80**
Structure	+1.56**	+1.22**
Content	+1.27**	+2.11**
All	+2.10**	+2.60**
Full Model	**+3.24****	

(Bendersky et al., 2017)

Query Understanding [Personal Mailbox]

- User mailbox is helpful in identifying words or names that are unique to the user
- E.g., can be leveraged for ***personalized spelling correction***



(a) Performance without our algorithm.

(b) Performance with our algorithm enabled.

Metrics	Email	Drive	Calendar
CTR	+31.55%	+16.04%	+38.27%
MRR	+26.00%	+10.17%	+30.41%
Has Result Rate	+15.55%	+8.51%	+43.34%
Number of Results	+31.13%	+14.53%	+76.61%

Huge performance gains for the affected queries (~3% of search traffic)

(Gupta et al., 2019)

Query Understanding [User Context]

- Personal query log
  - *identifies recurring searches, and unique search intents*
  - *may be very sparse*
- Demographic factors (Carmel et al., 2017b)
  - *age, income, gender may be useful for query auto-completion*
- Geolocation (Foley et al., 2018)
  - *semantic representations of the fine-grained user location*
  - *especially useful for mobile email search*

Gains drop off rapidly for longer prefixes

			$p = \epsilon; \lvert p \rvert = 0$			$\lvert p \rvert = 1$			$\lvert p \rvert = 2$			$\lvert p \rvert = 3$		
			MRR	mAP	P@1	MRR	mAP	P@1	MRR	mAP	P@1	MRR	mAP	P@1
QPM	§4.1	$P(q\|p)$	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x
DLM	§4.2	$P(q\|h,p)$	1.88x†	3.47x†	2.64x†	1.09x†	1.09x†	1.15x†	1.03x†	1.03x†	1.04x†	1.01x†	1.01x†	1.01x†
TLM	§4.4	$P(q\|L,p)$	4.08x†	7.71x†	5.21x†	1.19x†	1.20x†	1.31x†	1.07x†	1.07x†	1.10x†	1.02x†	1.02x†	1.03x†
CCLM	§4.5	$P_{CC}(q\|L,p)$	4.51x†	8.78x†	5.91x†	1.22x†	1.22x†	1.35x†	1.08x†	1.08x†	1.11x†	1.02x	1.02x	1.03x

† Represents statistical significance with $p < 0.0001$ with a pairwise randomization test over the entry in the previous row.

Query Understanding [Global Query Log]

- An anonymized global query log is useful for leveraging the universal email search trends
  - *Result still needs to be verified against personal mailbox*

- ***Horovitz et al. (2017)*** *find that a linear combination of mailbox and global logs models for query completion leads to optimal results*

Query Expansion

- Kuzi et al. (2017) experiment with
  - (1) **a translation model** based on global query log
  - (2) **an embedding model** based on the user's mailbox
  - (3) **a pseudo-relevance** feedback model
  - Overall, (1) performs the best; in some cases an interpolation is helpful

Query Expansion

- Li et al. (2019) confirm the usefulness of global email logs for query expansion
  - Propose a **multi-view multi-task NN model** leveraging clicks, sessions and user distribution for candidate generation
  - Followed by candidate ***fusion***, ***filtering*** and ***ranking*** stages

	MRR	CTR
Rank fusion	+0.97	+1.56*
Coordinate Ascent	+1.32*	+1.68*

Online evaluation, comparing to a baseline utilizing web-based synonyms

Enron Email Dataset

- Contains data from about 150 users, mostly senior management of Enron, organized into folders
- Not officially supported by any data consortium or institution,
- The privacy of the email correspondents has not been preserved through any reduction procedure.

Therefore, we discourage its use, and advise that analysis and algorithms preserve the privacy of Enron correspondents.

Avocado Dataset

- Distributed by LDC
- 1.3 million emails taken from 279 accounts of a defunct information technology company referred to as “Avocado”.
- Contains some additional information not available in Enron
  - *contact information, email attachments, etc.*
- Illustrative use cases
  - *Commitment detection*
  - *Attachment recommendation*
  - *Action item extraction*
  - *Search relevance*
  - *Email sensitivity*

Managing and Learning from User Data

Data Privacy Principles

1. **Encryption**

- All types of data are stored in an encrypted format
- Unredacted data cannot be directly accessed by the model developers

(Kannan et al., 2016)

Data Privacy Principles

2. **Aggregation**

- Model developers can only inspect aggregated model statistics
- These statistics cannot be associated or traced back to a particular user

(Kannan et al., 2016)

Data Privacy Principles

3. **Frequent Words**

- Words that occur across *multiple users*
- If the inspection of a particular text snippet is required
  - any user identifications are removed
  - only a bag of frequent words is retained.

(Kannan et al., 2016)

Anonymized Data

"... information which does not relate to an identified or identifiable natural person or ... personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."

General Data Protection Regulation (EU)

Data Anonymization Techniques

- **De-identification**
  - Removes any personally identifiable information from the stored content
  - May be error-prone
  - No provable guarantees for removing all personal identifiers

- ***k*-Anonymity**
  - Any combination of protected attributes can be indistinctly matched to at least *k* individuals
  - Difficult to maintain for high-dimensional datasets

- **Differential Privacy**
  - Provably ensuring that adding or removing a single database item does not substantially change the output distribution

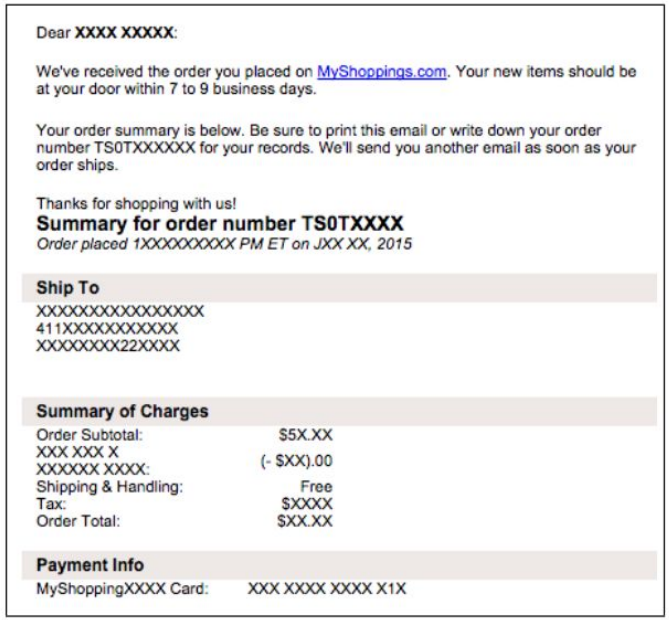
Data De-identification (Healthcare Example)

(a) Names	
(b) All geographic subdivisions smaller than a state, or the initial three digits of the ZIP code, if the geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people	
(c) All elements of dates (except year) for dates that are directly related to an individual	
(d) Telephone numbers	(k) License numbers
(e) Fax numbers	(l) Vehicles identifiers and serial numbers
(f) Email addresses	(m) Device identifiers and serial numbers
(g) Social security numbers	(n) URLs
(h) Medical record numbers	(o) IP addresses
(i) Health plan beneficiary numbers	(p) Biometric identifiers
(j) Account numbers	(q) Full-face photographs
(r) Any other unique identifying number, characteristic, or code	

Personal identifiers that need to be removed in order to de-identify health information.
According to the Safe Harbor method, as described by the US Department of Health and Human Services (2012) HIPAA regulation §164.514(b)(2).

k - Anonymity (Email Example)

- Deriving *k*-anonymized email samples for human inspection (Di Castro et al, 2016)



Dear **XXXX XXXXX**:

We've received the order you placed on MyShoppings.com. Your new items should be at your door within 7 to 9 business days.

Your order summary is below. Be sure to print this email or write down your order number TS0TXXXXXX for your records. We'll send you another email as soon as your order ships.

Thanks for shopping with us!

Summary for order number TS0TXXXX

Order placed 1XXXXXXXXX PM ET on JXX XX, 2015

Ship To

XXXXXXXXXXXXXXX
411XXXXXXXXXX
XXXXXXXX22XXXX

Summary of Charges

Order Subtotal:	$5X.XX
XXX XXX X XXXXXX XXXX:	(- $XX).00
Shipping & Handling:	Free
Tax:	$XXXX
Order Total:	$XX.XX

Payment Info

MyShoppingXXXX Card: XXX XXXX XXXX X1X

1. **Grouping**
   a. Group messages by MD5 hash of its DOM-tree signature
   b. Delete any group with less than *k* unique users

2. **Masking**
   a. Delete from each of the DOM-tree entries, up to the point where all messages in the group are identical
   b. Replace deleted entries with XXX

3. **Assignment**
   a. Each user is associated with at most one template per assessor

(Di Castro et al., 2016)

Differential Privacy (Formal Definitions)

$$\frac{P(\mathcal{A}(D) \in S)}{P(\mathcal{A}(D') \in S)} \leq \exp(\epsilon)$$

- D & D' differ in at most one element
- Provably ensures that adding or removing a single database item does not substantially change the output distribution of $\mathcal{A}$
- $\varepsilon \in [0, 0.01]$ required for strong privacy guarantees
- Generally implemented by adding small amounts of random noise to either
  - dataset D
  - algorithm $\mathcal{A}$

Differential Privacy – Adding noise to algorithm $\mathcal{A}$

- **DP-SGD** – A popular differentially private SGD algorithm (Abadi et al., 2016)

Algorithm 1 Differentially private SGD (Outline)

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample L_t with sampling probability L/N

Compute gradient

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

Descent

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

Differential Privacy – Adding noise to dataset *D*

- **d□-privacy** – anonymizes an input string **x** by perturbing each word to a word close in the embedding space (Feyisetan et al., 2020)
  - d□-private strings preserve semantics with provable privacy
  - e.g., were shown to be useful for BERT pre-training (Qu et al., 2021)

Algorithm 1: Privacy Preserving Mechanism

Input: string $x = w_1 w_2 \cdots w_\ell$, privacy parameter $\varepsilon > 0$

for $i \in \{1, \ldots, \ell\}$ **do**

- Compute embedding $\phi_i = \phi(w_i)$
- Perturb embedding to obtain $\hat{\phi}_i = \phi_i + N$ with noise density $p_N(z) \propto \exp(-\varepsilon \|z\|)$
- Obtain perturbed word $\hat{w}_i = argmin_{u \in W} \|\phi(u) - \hat{\phi}_i\|$
- Insert $\hat{w}_i$ in ith position of $\hat{x}$

release $\hat{x}$

	the	emotions	are	raw	and	will	strike	a	nerve
Perturbed tokens	the	emotions	are	raw	and	will	strike	a	nerve
	a	emotion	were	smackdown	or	would	strikes	the	rebels
	its	emotionally	is	matt	but	can	attack	an	reason
	and	hormones	being	##awa	,	may	drop	his	cells
	his	organizations	re	unused	-	better	##gen	its	spirits
	her	emotional	have	division	as	must	aim	her	bothering
	some	moods	of	protection	"	self	stroke	one	communications

Transparent Data Access

- Under what circumstances would users be comfortable sharing some of their mailbox data with researchers?

- EmailValet – users knowingly share data with crowdsource workers for a limited duration of time.

- The crowdsource workers assist users in managing their inbox, e.g., via task extraction

- Initially all users had concerns, but:
  - "over half of those with concerns (10 of 18) ... reported that they felt more comfortable with the service over time, while no one reported a decrease in comfort".

(Kokkalis et. al, 2013)

Transparent Data Access (Takeaways)

Users may be willing to provide limited access to their mailbox as long as

(a) the users can clearly specify the time limits for workers access;

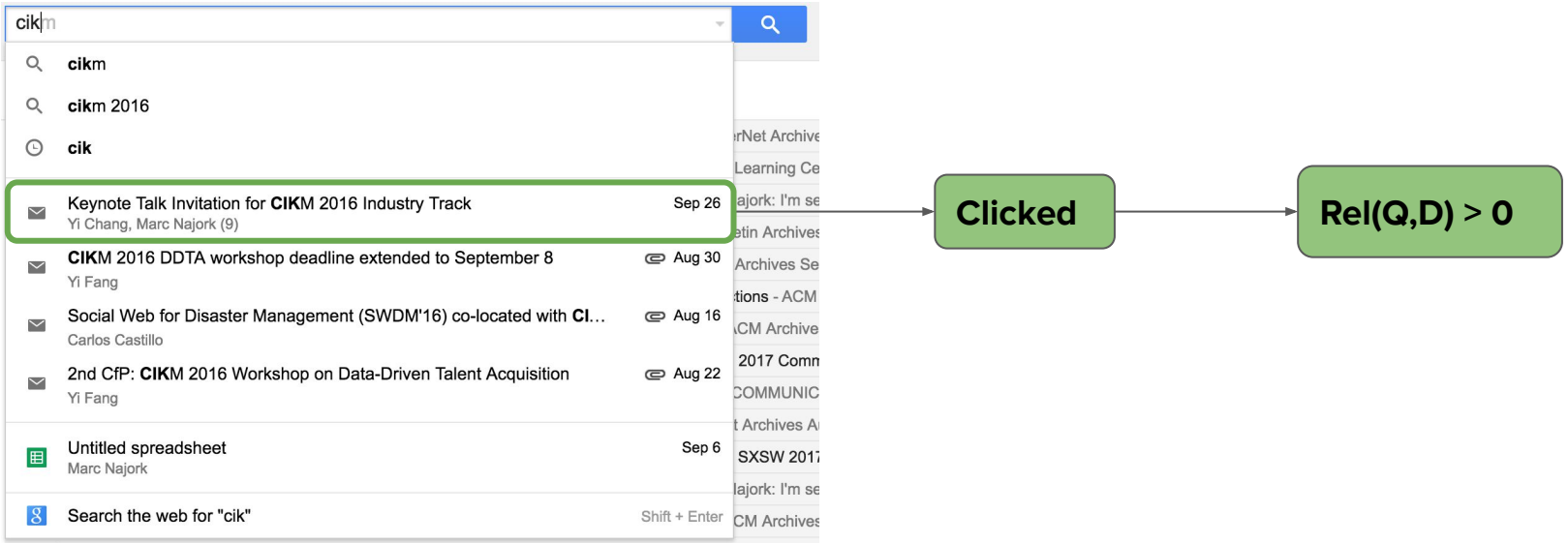
(b) the users have full control over what data is being shared;

(c) the users perceive the services enabled by data sharing valuable.

Learning from Clicks

- Human judged relevance is hard to acquire
  - Due to personal corpora
  - Unique in personal search, different from web search or TREC
- Approach: learning from implicit click data
  - Abundant and easily available
  - It is inherently **biased**
  - It is highly sparse
- How to effectively learn from clicks?

Learning from Clicks



Data Bias

Position Bias

- Users are more prone to click on the first few results, *independent of relevance*
- Varies between retrieval systems, evolves over time

Selection Bias

- Difficult queries will result in fewer clicks, therefore the data set is skewed

Presentation Bias

- Click probability is influenced by quality of result snippet, visual appeal

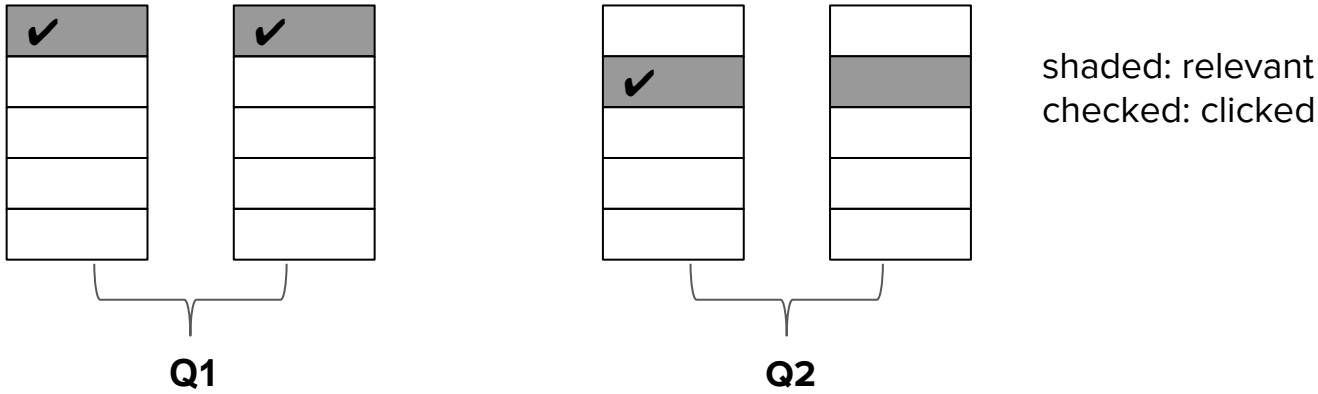
Click Noise

- Clicked results are (often) assumed to be relevant, sometimes incorrectly so

Click Sparsity

- Each user has their own corpus, so cannot aggregate across all users

The Data Bias Problem in Click Data



shaded: relevant
checked: clicked

- $\boldsymbol{U}$ - total query universe (approximated by query log)
  $\boldsymbol{S}$ - collection of queries **with** clicks (useable portion of query log)

- $\boldsymbol{S}$ is biased: $P(Q \mid \boldsymbol{S}) \neq P(Q \mid \boldsymbol{U})$
  - $P(Q_1 \mid \boldsymbol{U}) = P(Q_2 \mid \boldsymbol{U})$
  - $P(Q_1 \mid \boldsymbol{S}) = 2 * P(Q_2 \mid \boldsymbol{S})$

(Wang et al., 2016)

Bias Correction

- Propensity score: $P(Q|\boldsymbol{S})$
- Inverse propensity weights
  - $w_Q = P(Q|\boldsymbol{U}) / P(Q|\boldsymbol{S})$
  - Similar to importance sampling
- Loss function

$$L_{\mathcal{S}}(f) = \frac{1}{|\mathcal{S}|} \sum_{Q \in \mathcal{S}} w_Q \cdot l(Q, f)$$ rather than $$L_{\mathcal{U}}(f) = \frac{1}{|\mathcal{U}|} \sum_{Q \in \mathcal{U}} l(Q, f)$$

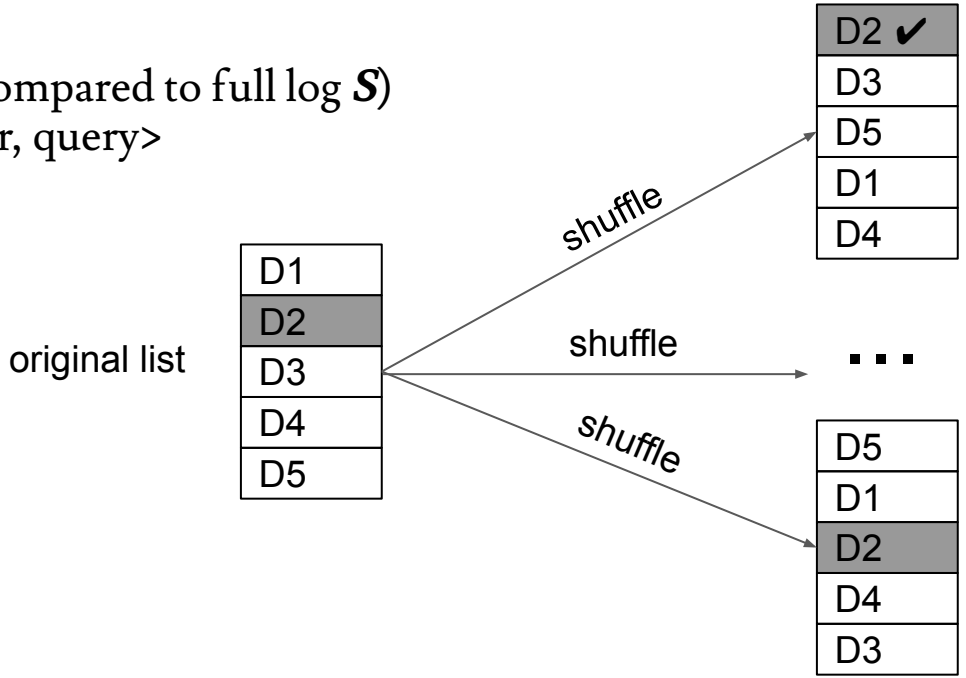
- How to estimate w_Q?

Bias Estimation with Randomization

- Randomized data collection ***R*** (small compared to full log ***S***)
  - Shuffle function is fixed per <user, query>
  - Single click scenario

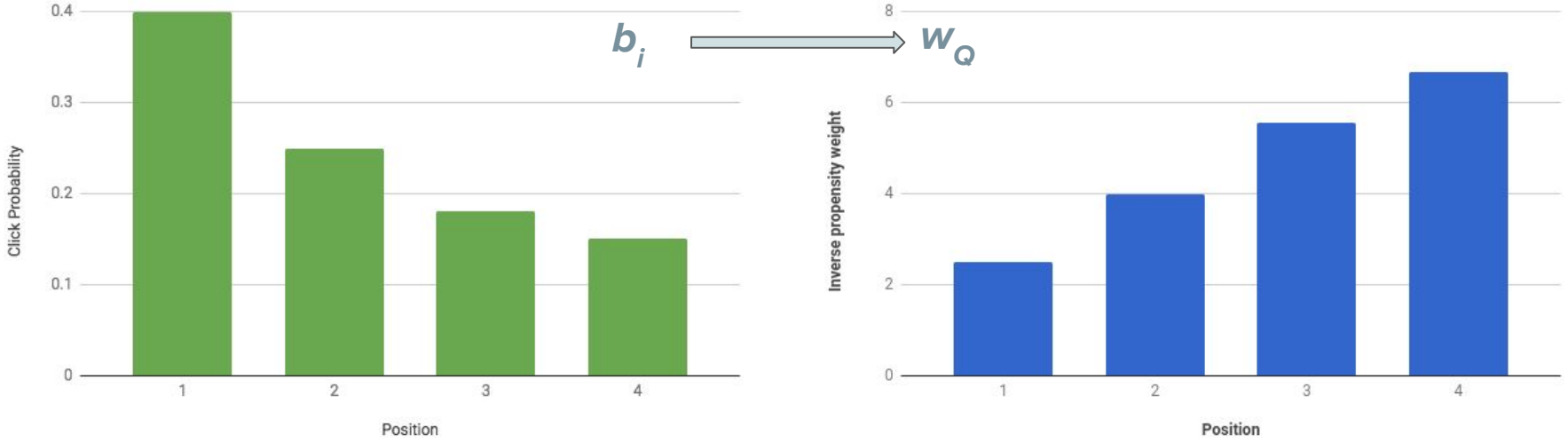
- Click probability $c^Q_{\mathbf{x}i} = r^Q_{\mathbf{x}} \cdot b_i$
  - Probability of relevance Independent of rank due to shuffling

- Methods
  - Global bias model
  - Segmented bias model
  - Generalized bias model

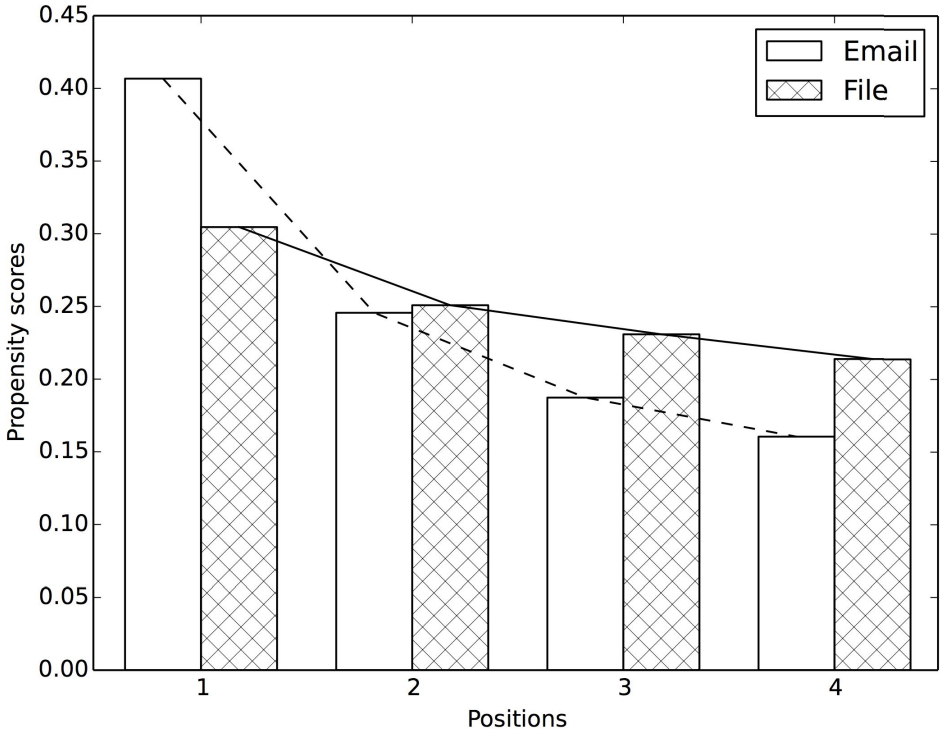


Global Bias Model

- Bias depends on position *i* only:
  - b_i = *Probability of seeing the click at position i*
  - Assigns higher weights w_Q to queries with clicks at lower positions



Position Bias Values are Specific to Retrieval System



Advanced Bias Models

- Segmented Bias Model
  - Group queries into different segments
  - In our data, we use the email categories as the segments
  - Estimate segment-dependent biases
  - Weigh each query according to its segment during LTR training

- Generalized Bias Model
  - **Position bias prediction:** Given Q, *predict* the click probability at position *i* IF we could show its set of documents in a uniformly random order
  - **Label**: the clicked position
  - **Features**: v(Q), e.g., query segment, query length, etc.

Bias Estimation without Randomization

	Email (N=3)	File Storage (N=5)
RandTopN	-13.94%*	-31.04%*
RandPair(1, 2)	-6.80%*	-12.44%*
RandPair(2, 3)	-0.56%	+3.75%
RandPair(3, 4)	+0.20%	+1.09%
RandPair(4, 5)	+0.38%	+0.36%

Table 1: The negative effect due to result randomization measured by the relative change of MRR against the production. * means statistically different.

(Wang et al., 2018)

Expectation-Maximization from Regular Clicks

- Log likelihood over data

$$\log P(\mathcal{L}) = \sum_{(c,q,d,k)\in\mathcal{L}} c \log \theta_k \gamma_{q,d} + (1-c)\log(1-\theta_k\gamma_{q,d})$$

- E-step: estimate the distribution of hidden variable

$$P(E=1,R=1|C=1,q,d,k) = 1$$

$$P(E=1,R=0|C=0,q,d,k) = \frac{\theta_k^{(t)}(1-\gamma_{q,d}^{(t)})}{1-\theta_k^{(t)}\gamma_{q,d}^{(t)}}$$

$$P(E=0,R=1|C=0,q,d,k) = \frac{(1-\theta_k^{(t)})\gamma_{q,d}^{(t)}}{1-\theta_k^{(t)}\gamma_{q,d}^{(t)}}$$

$$P(E=0,R=0|C=0,q,d,k) = \frac{(1-\theta_k^{(t)})(1-\gamma_{q,d}^{(t)})}{1-\theta_k^{(t)}\gamma_{q,d}^{(t)}}$$

Expectation-Maximization from Regular Clicks

- M-step: update the parameters

$$\theta_k^{(t+1)} = \frac{\sum_{c,q,d,k'} \mathbb{I}_{k'=k} \cdot (c + (1-c)P(E=1|c,q,d,k))}{\sum_{c,q,d,k'} \mathbb{I}_{k'=k}}$$

$$\gamma_{q,d}^{(t+1)} = \frac{\sum_{c,q',d',k} \mathbb{I}_{q'=q,d'=d} \cdot (c + (1-c)P(R=1|c,q,d,k))}{\sum_{c,q',d',k,} \mathbb{I}_{q'=q,d'=d}}$$

- Problem
  - Need the query and document identifiers
  - Not available in personal search

Regression-based EM

- Modify the M-step
  - No need of query and document identifiers *q, d*
  - Work in a feature space $\mathbf{x}_{q,d}$
  - Relevance estimated by a regression function: $\gamma_{q,d} = f(\mathbf{x}_{q,d})$
- Algorithm
  - For each data point,
    - Get P(R=1 | c, q, d, k) from E-step
    - Sample r in {0, 1}
  - From the training data {(**x**, r)} train a function f (e.g., using GBDT)
  - Iterate the EM procedure until convergence

Embedded in Discriminative Methods

- For a training instance for each data point: (position, features, click)
- Build a click prediction function on (position,features)
- Logistic regression
  - The coefficient of “position” can be used as bias
  - The rest of the model on “features” can be used to predict relevance
- This model can not distinguish “position” and “features”

EM vs Embedded

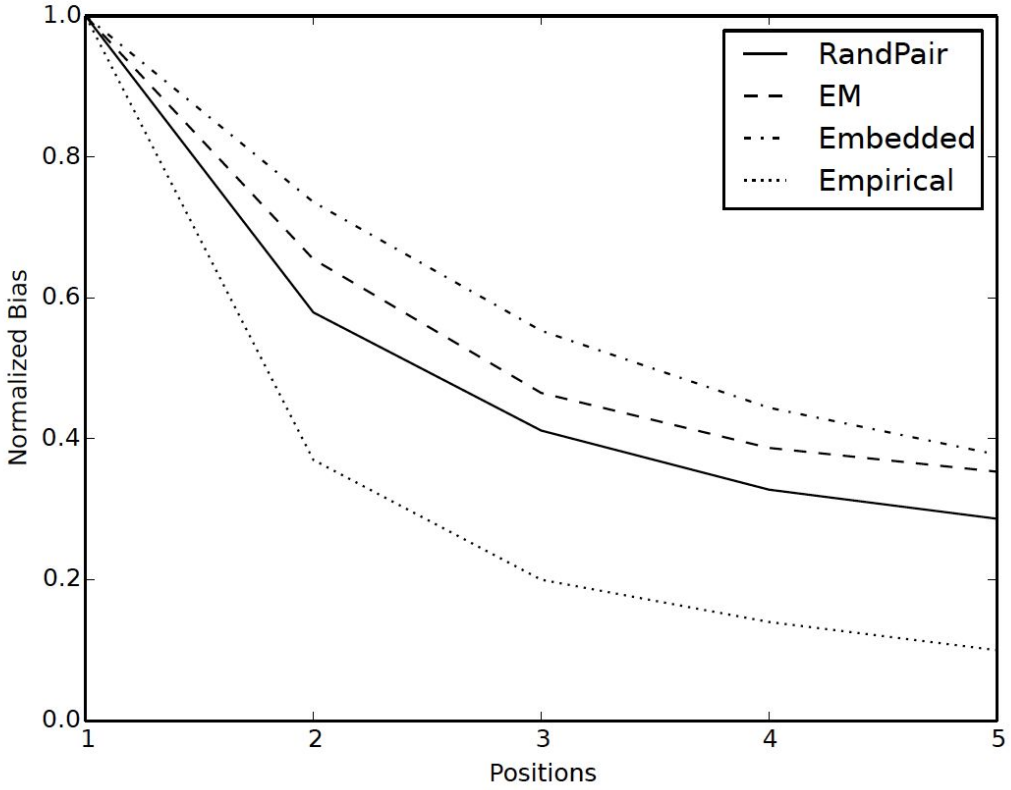
	EM	Embedded
Email	**-0.124**	-0.130
File Storage	-0.121	**-0.116**

Table 2: The average LogLikelihood on the test data set. A larger number means a better fit.

	EM	Embedded
Email	**+0.50%***	-4.44%*
File Storage	**+0.11%**	-3.10%*

Table 3: The relative difference of MRR against the baseline for the relevance components.

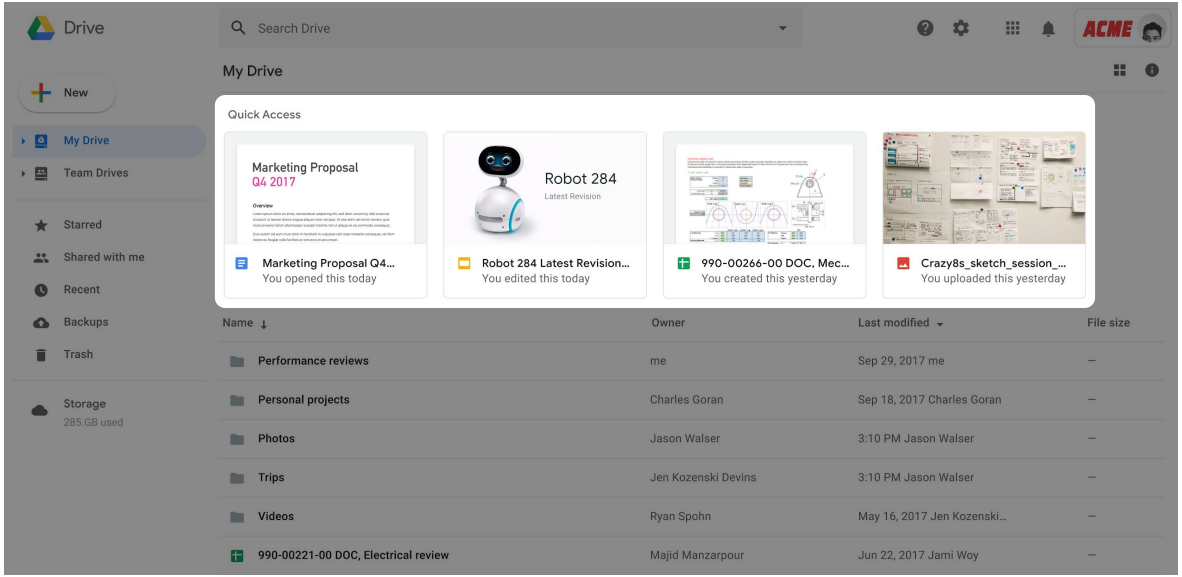
Estimated Bias

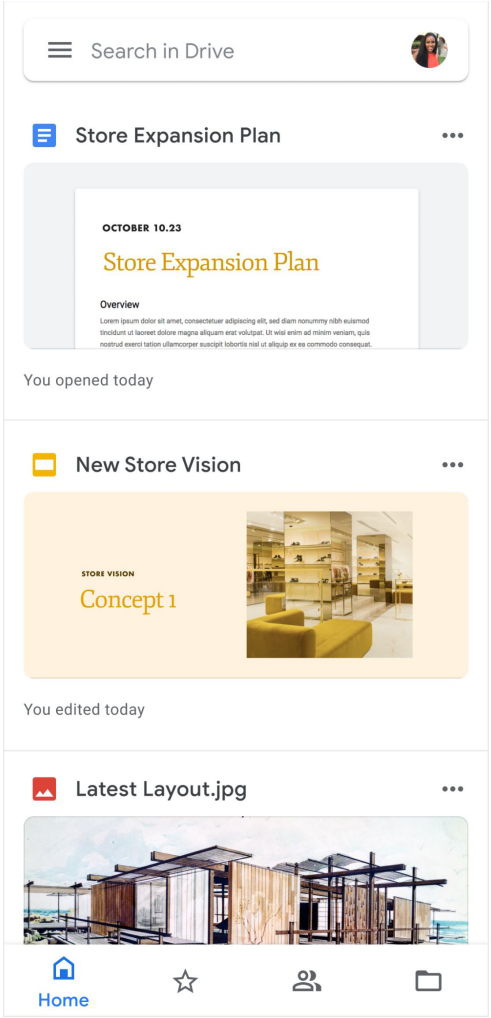


The New Frontier

Intelligent Task Assistance

Personal Content Recommendation





Personal Content Recommendation

- Personal content recommender system presents unique challenges
    - Private or shared documents
        - The (user, item, rating) matrix is extremely sparse
        - Information ‘smearing’ needs to be done carefully
    - Algorithms and models must respect user privacy
        - Data aggregation and anonymization
        - Challenges when it comes to debugging
    - Novel UI challenges
        - How recommendations interact with other personalized finding features (e.g., search)
        - Explainatations

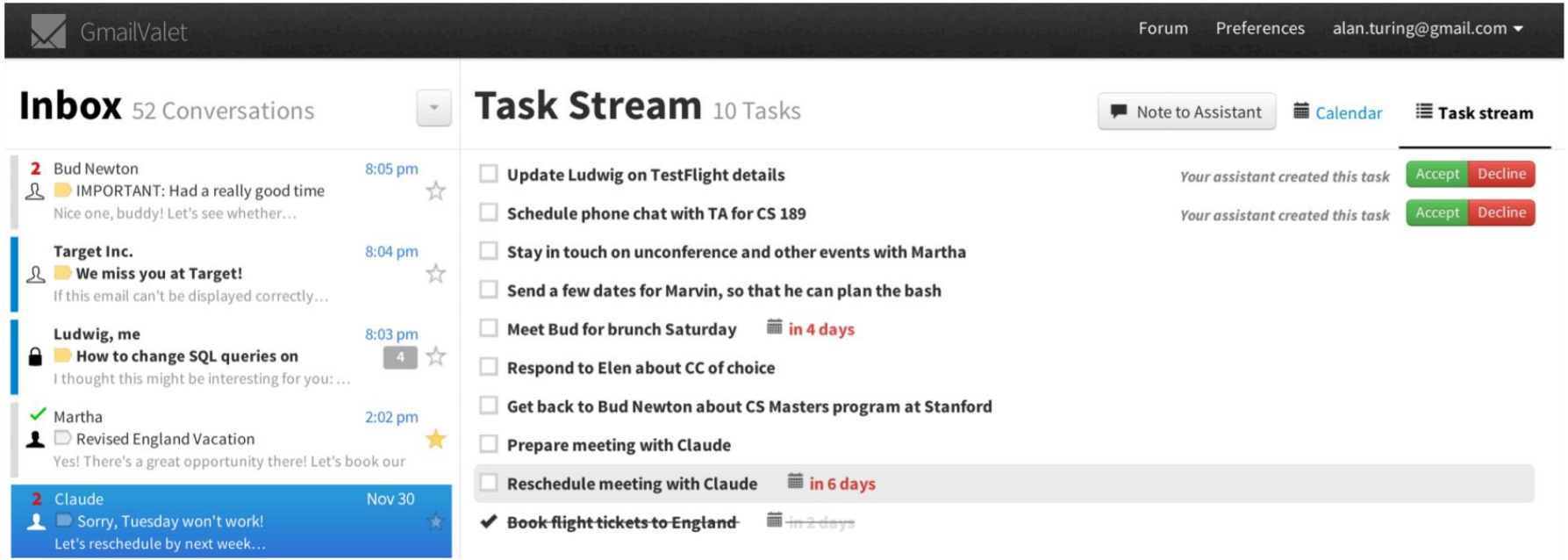
Cross-Platform and Cross-Device Assistance

- Users interact with many different personal content systems (e.g., email, calendar, photos, etc.) possibly across multiple devices (e.g., desktop, mobile, etc.) on a daily basis

- Understanding the user's **context** across all such systems and devices is important when providing assistance

- Many interesting research, privacy, and systems challenges

Cross-Platform and Cross-Device Assistance

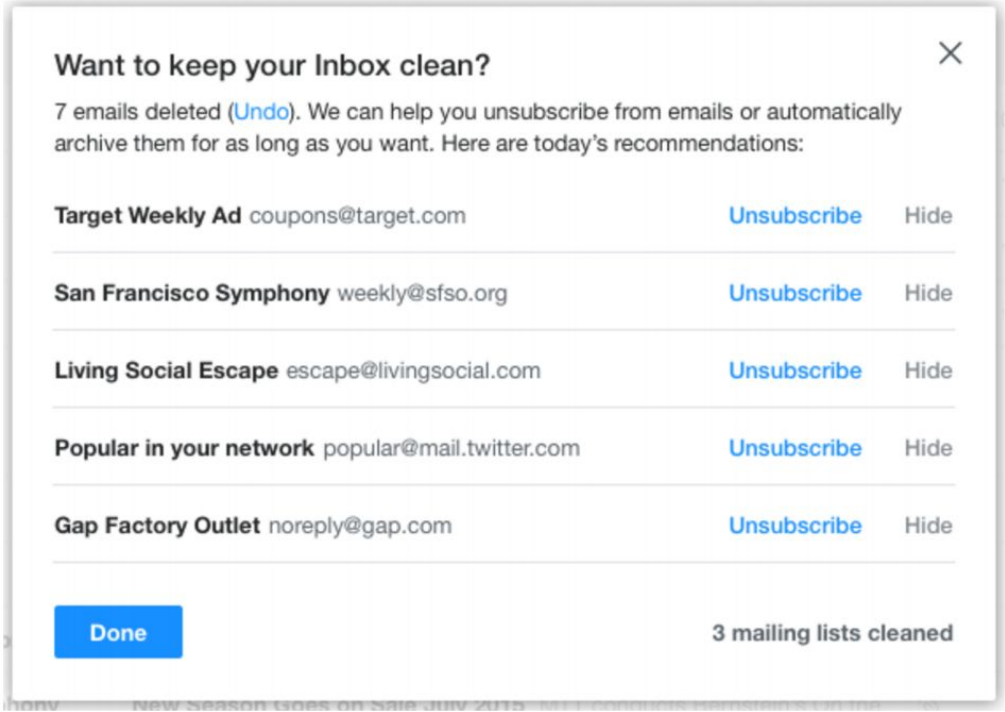


Email Valet (*Kokkalis et al., 2013*) transforms a cluttered mailbox into a succinct, mobile-friendly stream of individual tasks.

Activity Prediction

- Analyzing and learning from aggregated user actions can reduce information overload by proactively predicting time saving activities

- Many possible prediction tasks
  - Email importance
  - Bulk email unsubscribe
  - Email arrival time
  - Email labels



Unsubscribe dialog to facilitate easier unsubscription (*Gamzu et al., 2018*).

Assisted Composition

- Assisted composition technologies have two primary goals
    - Eliminate repetitive and often mundane content creation tasks
        - Generating canned email replies
        - Suggesting likely email recipients based on context and history
    - Improve the quality and readability of content
        - Suggesting relevant email attachments
        - Improving readability via better language clarity or grammar

Assisted Composition

- **Binary prediction**
  - Missing attachment detection
  - Unintentended email recipients
  - Request detection
- **Item ranking**
  - Likely email recipients
  - Documents to attach to an email
  - Email folder suggestion
  - Reply suggestion
- **Content generation**
  - Smart Reply (short canned responses)
  - Smart Compose (word and sentence completions as you type)

Other Advanced Topics

Domain-specific Search and Domain Adaptation

- **Enterprise search**
  - # of enterprise emails > # of non-machine-generated consumer emails (*Radicati Group, 2015*)
  - Average enterprise user receives ~130 emails per day (*Radicati Group, 2015*)
  - Workers spend ~20% of their time on search and information gathering activities (*Chui et al, 2012*)
- **Domain adaptation**
  - Monolithic solutions are unlikely to work well for all enterprises
  - Training specialized models per domain is challenging for many reasons
  - Some research has been done to address this, but there are still many unsolved problems

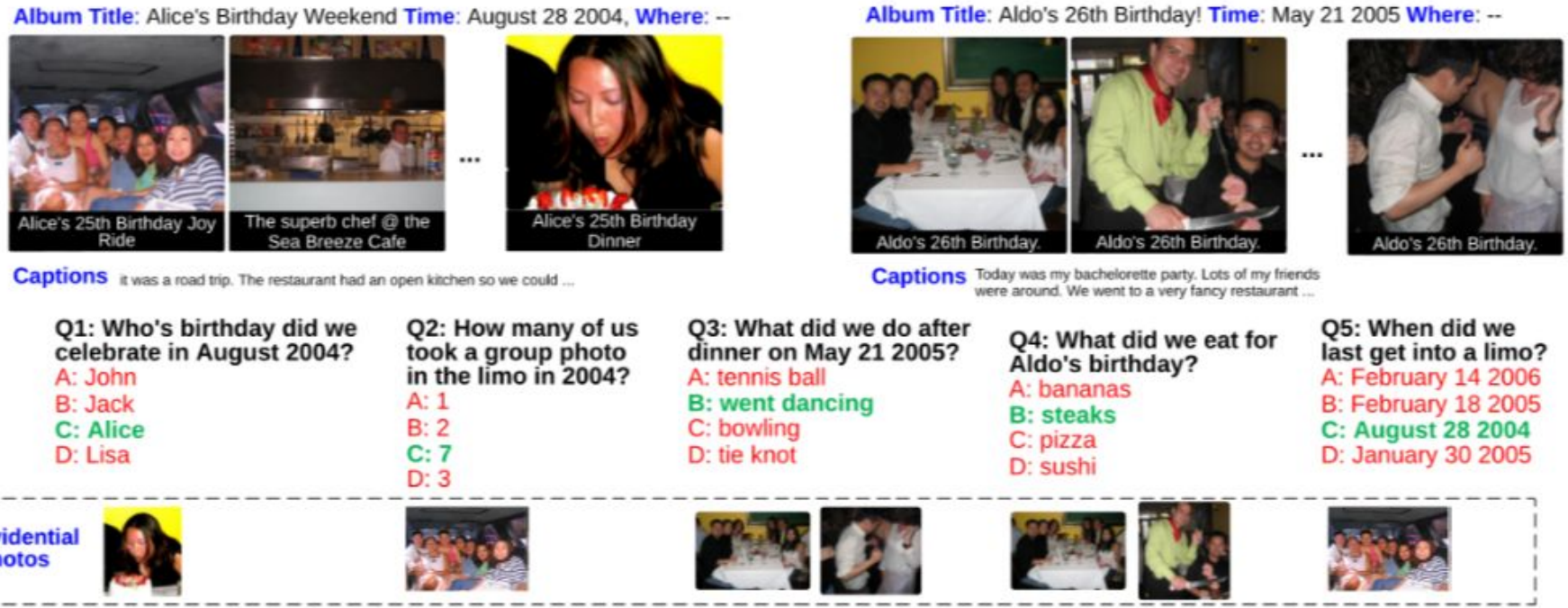
Question Answering Systems for Personal Content

- Challenges
  - Information needed to answer questions is a mix of public and private knowledge
  - May need to synthesize information across multiple sources (e.g., email, calendar, photos, etc.)
  - Difficult to build test collections
  - Difficult to evaluate
- Personal Knowledge Graph (*Balog and Kenter, 2019*)
  - Structure that captures entities, attributes, and their relations for an individual user
  - Goal is to satisfy requests such as *"schedule an appointment with the dentist that was recently recommended by Alice"*
- Many interesting research opportunities

Multi-modal Search

- Email is more than just text
  - Embedded/attached images
  - Multi-modal attachments, like PDFs
  - Links to sites containing multimedia (e.g., YouTube)
- Visual Query Embeddings (*Jiang et al., 2017*)
  - Maps query terms to related visual concepts using click data
- MemexQA (*Liang et al., 2019*)
  - Question answering over personal media collections
- Mobile Multimodal Photo Annotation and Retrieval (*Anguera et al, 2008)*
  - Multi-modal search on mobile devices
- More research needed to develop better ‘holistic’ email models

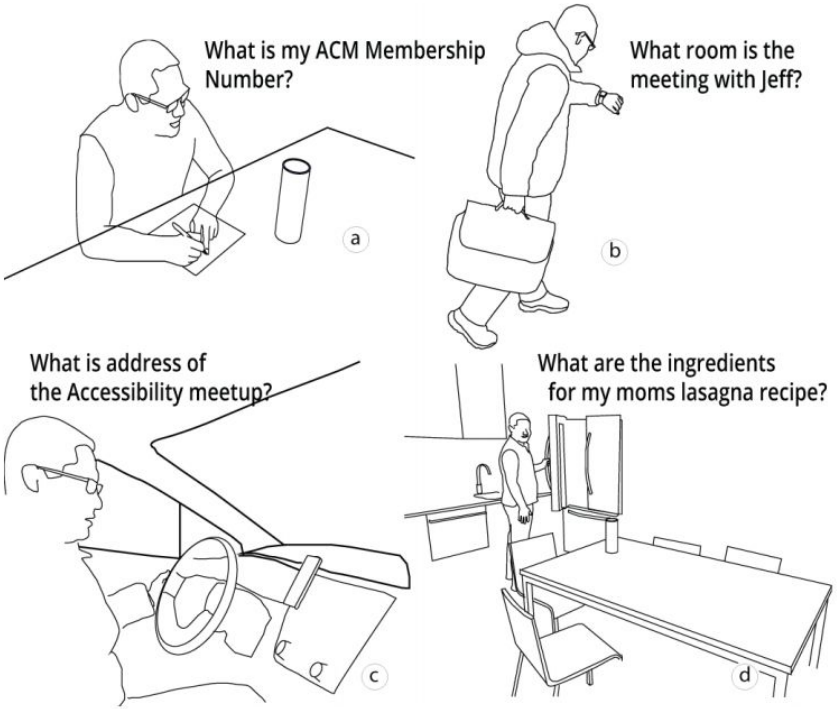
Multi-modal Search



Questions, multiple-choice answers and supporting evidence photos in the MemexQA system (the correct answers are marked in green)
(*Liang et al., 2019*)

Search on Mobile and Wearable Devices

- More and more search is happening on mobile and wearable devices
- Ubiquitous computing will result in new opportunities and challenges for personal search
- Dialog understanding and situational context (time, location, etc.) are critical components



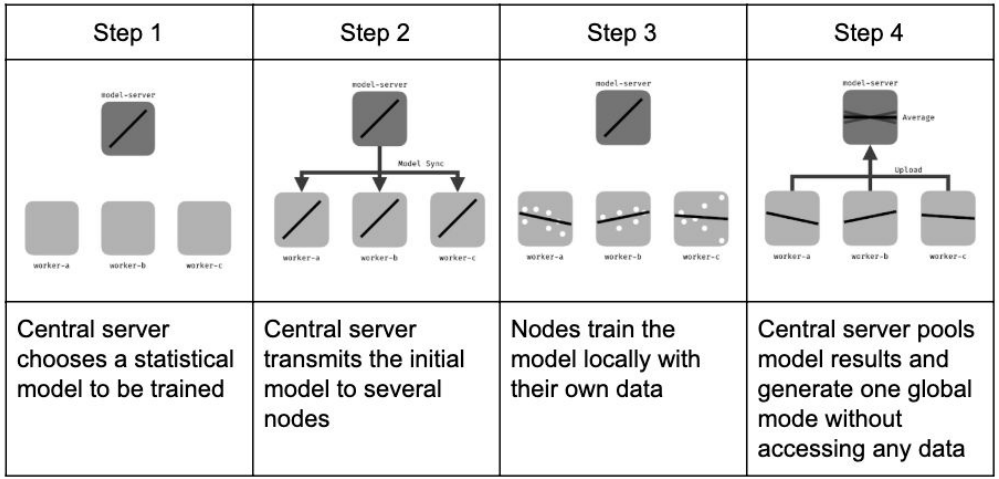
Illustrative examples of email search in various on-the-go settings (*Swaminathan et al., 2017*).

Beyond Relevance Ranking

- **User interfaces**
  - Chronological ordering, relevance ordering, hybrid, etc.
  - More advanced presentation layouts
  - Better snippet design
- **Federated search**
  - Merging results from multiple corpora on a single results page
  - Must consider overall presentation, diversity, relevance, etc. to optimize whole page experience
- **Fairness**
  - No known research on fairness in the context of personal search
  - Important topic as it is easy for bias to find its way into such systems

Federated Learning

- Federated learning is a promising direction that can help mitigate user data protection and privacy risks
- Non-aggregated data never leaves a user's device, but locally trained model parameters are shared with a central server
- Important direction of research given shift to mobile and increasing privacy concerns



By Jeromemetronome - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=79649308

What do you think is next?

References

- **Abadi**, M., A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K.Talwar, and L. Zhang. (2016). "Deep Learning with Differential Privacy"
- **Ai**, Q., S. T. Dumais, N. Craswell, and D. J. Liebling. (2017). "Characterizing Email Search using Large-scale Behavioral Logs and Surveys".
- **Ailon**, N., Z. S. Karnin, E. Liberty, and Y. Maarek. (2013). "Threading Machine Generated Email".
- **Alrashed**, T., A. H. Awadallah, and S. Dumais. (2018). "The Lifetime of Email Messages: A Large-Scale Analysis of Email Revisitation".
- **Anguera**, X., J. Xu, and N. Oliver. (2008). "Multimodal Photo Annotation and Retrieval on a Mobile Phone".
- **Ashkan**, A. and D. Metzler. (2019). "Revisiting Online Personal Search Metrics with the User in Mind".
- **Avigdor-Elgrabli**, N., M. Cwalinski, D. Di Castro, I. Gamzu, I. Grabovitch-Zuyev, L. Lewin-Eytan, and Y. Maarek. (2016). "Structural Clustering of Machine-Generated Mail".
- **Balog**, K. and T. Kenter. (2019). "Personal Knowledge Graphs: A Research Agenda".
- **Bendersky**, M., X. Wang, D. Metzler, and M. Najork. (2017). "Learning from User Interactions in Personal Search via Attribute Parameterization"
- **Broder**, A. Z., N. Eiron, M. Fontoura, M. Herscovici, R. Lempel, J. McPherson, R. Qi, and E. Shekita. (2006). "Indexing Shared Content in Information Retrieval Systems".
- **Carenini**, G., R. T. Ng, and X. Zhou. (2007). "Summarizing Email Conversations with Clue Words".
- **Carmel**, D., G. Halawi, L. Lewin-Eytan, Y. Maarek, and A. Raviv. (2015). "Rank by Time or by Relevance?: Revisiting Email Search".
- **Carmel**, D., L. Lewin-Eytan, A. Libov, Y. Maarek, and A. Raviv. (2017a). "Promoting Relevant Results in Time-Ranked Mail Search".
- **Carmel**, D., L. Lewin-Eytan, A. Libov, Y. Maarek, and A. Raviv. (2017b). "The Demographics of Mail Search and Their Application to Query Suggestion".
- **Chui**, M., J. Manyika, J. Bughin, R. Dobbs, C. Roxburgh, H. Sarrazin, G. Sands, and M. Westergren. (2012). "The Social Economy: Unlocking Value and Productivity Through Social Technologies".
- **Di Castro**, D., L. Lewin-Eytan, Y. Maarek, R. Wolff, and E. Zohar. (2016). "Enforcing k-anonymity in Web Mail Auditing".
- **Dumais**, S., E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. (2003). "Stuff I've Seen: A System for Personal Information Retrieval and Re-use".
- **Feyisetan**, O., Balle, B., Drake, T., & Diethe, T. (2020). "Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations".
- **Foley**, J., M. Zhang, M. Bendersky, and M. Najork. (2018). "Semantic Location in Email Query Suggestion".

References (Cont.)

- **Gamzu**, I., L. Lewin-Eytan, and N. Silberstein. (2018). "Unsubscription: A Simple Way to Ease Overload in Email".
- **Gupta**, J., Z. Qin, M. Bendersky, and D. Metzler. (2019a). "Personalized Online Spell Correction for Personal Search".
- **Horovitz**, M., L. Lewin-Eytan, A. Libov, Y. Maarek, and A. Raviv. (2017). "Mailbox-Based vs. Log-Based Query Completion for Mail Search".
- **Jiang**, L., Y. Kalantidis, L. Cao, S. Farfade, J. Tang, and A. G. Hauptmann. (2017). "Delving Deep into Personal Photo and Video Search".
- **Joachims**, T., A. Swaminathan, and T. Schnabel (2017). "Unbiased Learning-to-rank With Biased Feedback".
- **Kannan**, A., K. Kurach, S. Ravi, T. Kaufmann, A. Tomkins, B. Miklos, G. Corrado, L. Lukacs, M. Ganea, P. Young, and V. Ramavajjala. (2016). "Smart Reply: Automated Response Suggestion for Email".
- **Kokkalis**, N., T. Kohn, C. Pfeiffer, D. Chornyi, M. S. Bernstein, and S. R. Klemmer. (2013). "EmailValet: Managing Email Overload Through Private, Accountable Crowdsourcing".
- **Liang**, J., L. Jiang, L. Cao, Y. Kalantidis, L. Li, and A. G. Hauptmann. (2019). "Focal Visual-Text Attention for Memex Question Answering".
- **Maarek**, Y. (2017). "Web Mail is not Dead!: It's Just Not Human Anymore".
- **Naragon**, K. (2018). "We Still Love Email, But We're Spreading the Love with Other Channels".
- **Narang**, K., S. T. Dumais, N. Craswell, D. Liebling, and Q. Ai. (2017). "Large-Scale Analysis of Email Search and Organizational Strategies".
- **Ogilvie**, P. and J. Callan. (2005). "Experiments with Language Models for Known-Item Finding of E-mail Messages".
- **Qu**, C., Kong, W., Yang, L., Zhang, M., Bendersky, M., & Najork, M. (2021). Natural Language Understanding with Privacy-Preserving BERT.
- **Swaminathan**, S., R. Fok, F. Chen, T.-H. Huang, I. Lin, R. Jadvani, W. S. Lasecki, and J. P. Bigham. (2017). "WearMail: On-the-Go Access to Information in Your Email with a Privacy-Preserving Human Computation Workflow".
- **Wan**, S. and K. McKeown. (2004). "Generating Overview Summaries of Ongoing Email Thread Discussions".
- **Wang**, X., M. Bendersky, D. Metzler, and M. Najork (2016). "Learningto Rank with Selection Bias in Personal Search".
- **Wang**, X., N. Golbandi, M. Bendersky, D. Metzler, and M. Najork (2018). "Position Bias Estimation for Unbiased Learning to Rank in Personal Search".
- **Zamani**, H., M. Bendersky, X.Wang, and M. Zhang. (2017). "Situational Context for Ranking in Personal Search".