Search and Discovery in Personal Email Collections

Michael Bendersky, Xuanhui Wang, Marc Najork, Donald Metzler
{bemike,xuanhui,najork,metzler}@google.com
Google Research
USA

ABSTRACT

Email has been an essential communication medium for many years. As a result, the information accumulated in our mailboxes has become valuable for all of our personal and professional activities. For years, researchers have developed interfaces, models, and algorithms to facilitate email search, discovery, and organization. This tutorial brings together these diverse research directions and provides both a historical background, as well as a high-level overview of the recent advances in the field. In particular, we lay out all of the components needed in the design of email search engines, including user interfaces, indexing, document and query understanding, retrieval, ranking, evaluation, and data privacy. The tutorial also goes beyond search, presenting recent work on intelligent task assistance in email and a number of interesting future directions.

CCS CONCEPTS

• **Information systems → Email**; **Retrieval models and ranking**.

KEYWORDS

Personal collections, privacy-preserving systems, email search

ACM Reference Format:
Michael Bendersky, Xuanhui Wang, Marc Najork, Donald Metzler. 2022. Search and Discovery in Personal Email Collections. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM '22), February 21–25, 2022, Tempe, AZ, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3488560.3501393

1 MOTIVATION

Despite the widespread use of social networks and instant messaging, email still remains an important medium for personal and professional communication. For instance, in a recent survey, Naragon et al. [28] found that users spend more than 3 hours on a weekday checking their work email. Roughly 50% of survey participants check both their personal and work email at least every few hours. In addition, email has grown in importance as a B2C communication channel, and as a repository for recent and past commercial transactions [26].

In the past decade, there has been a resurgence of interest in email search and discovery research, mainly led by industrial research labs such as Yahoo! Research, Microsoft Research, and Google Research. This research has been accompanied by the development of new user-visible features, such as hybrid chronological-relevance search, knowledge panels, assistant integration, and smart composition features (see Figure 1). In a recent Foundations and Trends® in Information Retrieval article [8], we provide a comprehensive overview of this recent research, as well as a survey of decades of academic research that laid the foundation for these newer developments.

In the proposed half-day tutorial, our aim is to provide a high-level entry point for those interested in the topics of our survey. We will revisit the main themes of the survey, and provide pointers on the important papers in each of the fields. Our hope is that this tutorial will lead to better recognition of the unique technologies behind our everyday tools, and will also inspire researchers to think beyond the scope of the current paradigms.

To the best of our knowledge, this would be the first tutorial on email search and discovery held at WSDM or any other major information retrieval conference in recent years. Given the large body of recent research in this area, and our recent survey, we believe that WSDM 2022 is a timely and opportune venue for this tutorial.

2 TOPICS OUTLINE

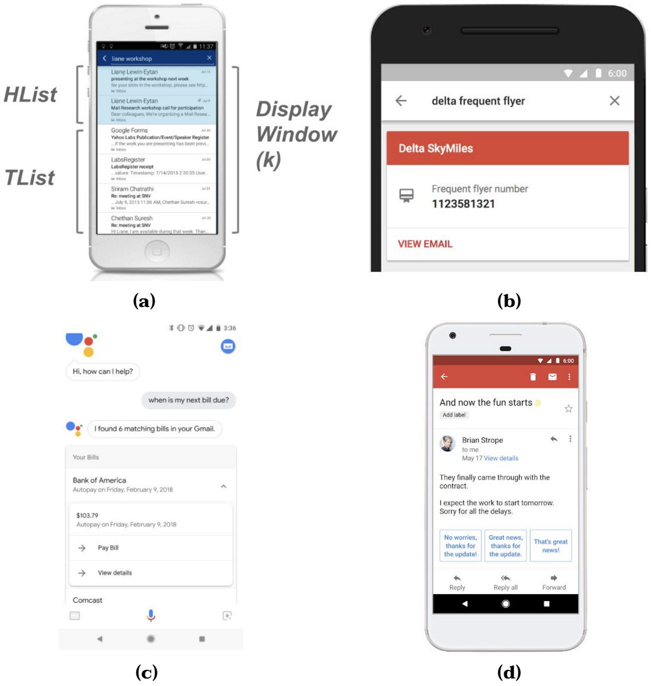
The topics of the tutorial will closely follow the contents of our survey [8]. Some material will be omitted due to time constraints, but we will provide pointers to the relevant resources as needed. In particular we will cover the following high-level topics. See more detailed description of what will be covered in Section 4.

- *Introduction* to the key research challenges and unique properties of email search and discovery.
- *The Anatomy of an Email Search Engine* – a high level overview of the various components of email search engines, including user interfaces, ranking, query and document understanding, and evaluation.
- *Data Management* – an overview of techniques designed to keep user data private, while developing new models for effective search and discovery. We will also propose techniques for dealing with bias and sparsity in user interaction data (e.g., clicks).
- *Intelligent Task Assistance* – going beyond search, and discussing other modes of personal content discovery, including recommendation, activity prediction, and assistive composition, as well as new frontiers in email search and discovery.

3 RELEVANCE TO THE COMMUNITY

Email search is an important search domain that often gets less attention from the research community, in large part due to the difficulty of developing appropriate public test collections. Due

WSDM '22, February 21–25, 2022, Tempe, AZ, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9132-0/22/02.
https://doi.org/10.1145/3488560.3501393



11:37
liane workshop
Liane Lewin-Eytan Jul 15
presenting at the workshop next week
for your slots in the workshop, please see http...
Inbox
Liane Lewin-Eytan Jul 9
Mail Research workshop call for participation
Dear colleagues, We're organizing a Mail Resea...
Inbox
Google Forms Jul 24
Yahoo Labs Publication/Event/Speaker Register
...if the work you are presenting has been previ...
Inbox
LabsRegister Jul 23
LabsRegister receipt
...values: Timestamp: 7/14/2015 2:20:35 User...
Inbox
Sriram Chatrathi Jul 21
Re: meeting at SNV
...July 9, 2015 11:06 AM, Chethan Suresh <csur...
Inbox
Chethan Suresh Jul 20
Re: meeting at SNV
Hi Liane, I am available during that week. Than...

(a)

6:00
delta frequent flyer
Delta SkyMiles
Frequent flyer number
1123581321
VIEW EMAIL

(b)

3:36
Hi, how can I help?
when is my next bill due?
I found 6 matching bills in your Gmail.
Your Bills
Bank of America
Autopay on Friday, February 9, 2018
$103.79
Autopay on Friday, February 9, 2018
Pay Bill
View details
Comcast

(c)

6:00
And now the fun starts
Add label
Brian Strope
to me
May 17 View details
They finally came through with the contract.
I expect the work to start tomorrow. Sorry for all the delays.
No worries, thanks for the update!
Great news, thanks for the update.
That's great news!
Reply
Reply all
Forward

(d)

Figure 1: Examples of email search and discovery features developed in recent years (a) An example of hybrid relevance results (*HList*) followed by chronological results (*TList*) [13] (b) Knowledge panel in Inbox by Gmail (a now defunct service) [21] (c) Google Assistant responding to a user query for bills in their email [30] (d) Smart Reply feature in Gmail [31].

to its private nature, much of the more recent research has been conducted in industrial research labs (e.g. [5, 7, 12]). However, we believe that the topics of the tutorial will be interesting to all attendees, academics, and industrial researchers alike.

There are quite a few challenges in email search that can be carried over to other domains, such as deriving features from a corpus in a privacy-preserving manner, learning query intents over private and dynamic corpora, learning from biased user feedback, enhancing recall when retrieving from small corpora (private mailboxes), etc. We will touch upon all of these unique challenges in our tutorial. In addition, we will present some emerging research that goes beyond the established paradigms of how email (and other private content) can be stored, processed, and accessed (e.g., [22, 25, 27, 32]). We hope that this work will encourage more researchers in the information retrieval community to tackle the challenges of building the next generation of private content search.

4 DETAILED SCHEDULE

The tutorial schedule will closely follow the contents of our survey [8].

- *Introduction* – Chapters 1 – 3 [30 min].
  - Email finding strategies – how do people seek and manage information in their mailboxes (e.g., [3, 35]).
  - Search interfaces – foldering and labeling [9], chronological ordering [15], relevance-based [13].
  - Summary of key differences between email and web search.
- *The Anatomy of an Email Search Engine* – Chapters 2, 4 and 5 [1 hr]
  - Challenges in indexing email such as access-control, content duplication across threads [10], and real-time updates.
  - Retrieval – search operators; relevance-based retrieval [12].
  - Relevance ranking using learning-to-rank [12].
  - Query and document Understanding – query completion [18], expansion [24] and spell correction [17]; thread structure resolution and email templatization [4, 23].
  - Evaluation – publicly available email corpora and test collections [1, 29]; user-centric success metrics [6].
- *Managing and Learning from User Data* – Chapter 7 [1 hr]
  - Anonymity principles and their implementation in email search (e.g., [11, 14]) – data de-identification, k-anonymity, and differential privacy.
  - Transparent data access for limited user studies [22].
  - Dealing with bias in user interaction data – position bias estimation and correction, unbiased metrics, trust bias (e.g., [2, 20, 34]).
  - Cross-user aggregation to combat data sparsity [7].
- *The Next Frontier* – Chapters 6 and 8 [30 min]
  - Intelligent task assistance – personal content recommendation [33], assistive composition [36], and predicting user activity [16].
  - Other advanced topics – question answering, multi-modal search, on-device search (e.g., [19, 25, 32]).

5 SUPPORT MATERIALS

The tutorial attendees will be provided with free digital copies of our survey, on which the tutorial is based [8]. In addition, we will make the tutorial slides available for download.

REFERENCES

[1] Samir AbdelRahman, Basma Hassan, and Reem Bahgat. 2010. A New Email Retrieval Ranking Approach. *International Journal of Computer Science and Information Technology* 2, 5 (2010), 44–63. https://doi.org/10.5121/ijcsit.2010.2504

[2] Aman Agarwal, Xuanhui Wang, Cheng Li, Michael Bendersky, and Marc Najork. 2019. Addressing Trust Bias for Unbiased Learning-to-Rank. In *Proceedings of the World Wide Web Conference (WWW '19)*. Association for Computing Machinery, New York, NY, USA, 4–14. https://doi.org/10.1145/3308558.3313697

[3] Qingyao Ai, Susan T. Dumais, Nick Craswell, and Daniel J. Liebling. 2017. Characterizing Email Search using Large-scale Behavioral Logs and Surveys. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*. 1511–1520. https://doi.org/10.1145/3038912.3052615

[4] Nir Ailon, Zohar Shay Karnin, Edo Liberty, and Yoelle Maarek. 2013. Threading Machine Generated Email. In *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013*. 405–414. https://doi.org/10.1145/2433396.2433447

[5] Tarfah Alrashed, Ahmed Hassan Awadallah, and Susan Dumais. 2018. The Lifetime of Email Messages: A Large-Scale Analysis of Email Revisitation. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. ACM, New York, NY, USA, 120–129. https://doi.org/10.1145/3176349.3176398

[6] Azin Ashkan and Donald Metzler. 2019. Revisiting Online Personal Search Metrics with the User in Mind. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*. Association for Computing Machinery, 625–634. https://doi.org/10.1145/3331184.3331266

[7] Mike Bendersky, Xuanhui Wang, Marc Najork, and Don Metzler. 2018. Learning with Sparse and Biased Feedback for Personal Search. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*. 5219–5223.

[8] Michael Bendersky, Xuanhui Wang, Marc Najork, and Donald Metzler. 2021. Search and Discovery in Personal Email Collections. *Foundations and Trends® in Information Retrieval* 15, 1 (2021), 1–133. https://doi.org/10.1561/1500000069

[9] Andrew D. Birrell, Edward P. Wobber, and Michael D. Schroeder. 1997. *Pachyderm*. http://birrell.org/andrew/pachywww/

[10] Andrei Z. Broder, Nadav Eiron, Marcus Fontoura, Michael Herscovici, Ronny Lempel, John McPherson, Runping Qi, and Eugene Shekita. 2006. Indexing Shared Content in Information Retrieval Systems. In *Advances in Database Technology - EDBT 2006*, Yannis Ioannidis, Marc H. Scholl, Joachim W. Schmidt, Florian Matthes, Mike Hatzopoulos, Klemens Boehm, Alfons Kemper, Torsten Grust, and Christian Boehm (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 313–330.

[11] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. In *Proceedings of the 28th USENIX Conference on Security Symposium (SEC 2019)*. 267–284.

[12] David Carmel, Guy Halawi, Liane Lewin-Eytan, Yoelle Maarek, and Ariel Raviv. 2015. Rank by Time or by Relevance?: Revisiting Email Search. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 283–292.

[13] David Carmel, Liane Lewin-Eytan, Alex Libov, Yoelle Maarek, and Ariel Raviv. 2017. Promoting Relevant Results in Time-Ranked Mail Search. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1551–1559. https://doi.org/10.1145/3038912.3052659

[14] Dotan Di Castro, Liane Lewin-Eytan, Yoelle Maarek, Ran Wolff, and Eyal Zohar. 2016. Enforcing k-anonymity in Web Mail Auditing. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, 327–336.

[15] Susan Dumais, Edward Cutrell, JJ Cadiz, Gavin Jancke, Raman Sarin, and Daniel C. Robbins. 2003. Stuff I've Seen: A System for Personal Information Retrieval and Re-use. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 72–79.

[16] Iftah Gamzu, Zohar Shay Karnin, Yoelle Maarek, and David Wajc. 2015. You Will Get Mail! Predicting the Arrival of Future Email. In *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*. 1327–1332. https://doi.org/10.1145/2740908.2741694

[17] Jai Gupta, Zhen Qin, Michael Bendersky, and Donald Metzler. 2019. Personalized Online Spell Correction for Personal Search. In *The World Wide Web Conference (WWW '19)*. ACM, New York, NY, USA, 2785–2791. https://doi.org/10.1145/3308558.3313706

[18] Michal Horovitz, Liane Lewin-Eytan, Alex Libov, Yoelle Maarek, and Ariel Raviv. 2017. Mailbox-Based vs. Log-Based Query Completion for Mail Search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 937–940.

[19] Lu Jiang, Yannis Kalantidis, Liangliang Cao, Sachin Farfade, Jiliang Tang, and Alexander G Hauptmann. 2017. Delving Deep into Personal Photo and Video Search. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 801–810.

[20] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased Learning-to-rank With Biased Feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 781–789.

[21] Govind Kaushal. 2016. Inbox by Gmail: Find Answers Even Faster. https://www.blog.google/products/gmail/inbox-by-gmail-find-answers-even-faster/

[22] Nicolas Kokkalis, Thomas Köhn, Carl Pfeiffer, Dima Chornyi, Michael S Bernstein, and Scott R Klemmer. 2013. EmailValet: Managing Email Overload Through Private, Accountable Crowdsourcing. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 1291–1300.

[23] Andrew Lampert, Robert Dale, and Cécile Paris. 2009. Segmenting Email Message Text into Zones. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 919–928. https://www.aclweb.org/anthology/D09-1096

[24] Cheng Li, Mingyang Zhang, Michael Bendersky, Hongbo Deng, Donald Metzler, and Marc Najork. 2019. Multi-view Embedding-based Synonyms for Personal Search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*. ACM, 575–584. https://doi.org/10.1145/3331184.3331250

[25] J. Liang, L. Jiang, L. Cao, Y. Kalantidis, L. Li, and A. G. Hauptmann. 2019. Focal Visual-Text Attention for Memex Question Answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), 1–1. https://doi.org/10.1109/TPAMI.2018.2890628

[26] Yoelle Maarek. 2017. Web Mail is not Dead!: It's Just Not Human Anymore. In *Proceedings of the 26th International Conference on World Wide Web (WWW 2017)*. International World Wide Web Conferences Steering Committee, Association for Computing Machinery, 5. https://doi.org/10.1145/3038912.3050916

[27] Marc Najork. 2018. Training On-Device Ranking Models from Cross-User Interactions in a Privacy-Preserving Fashion. In *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems, Bertinoro, Italy, August 28-31, 2018*. 108. http://ceur-ws.org/Vol-2167/short11.pdf

[28] Kristin Naragon. 2018. We Still Love Email, But We're Spreading the Love with Other Channels. https://theblog.adobe.com/love-email-but-spreading-the-love-other-channels/ Accessed: 2020-01-06.

[29] Douglas Oard, William Webber, David Kirsch, and Sergey Golitsynskiy. 2015. Avocado Research Email Collection. *Philadelphia: Linguistic Data Consortium* (2015).

[30] Ying Sheng, Sandeep Tata, James B. Wendt, Jing Xie, Qi Zhao, and Marc Najork. 2018. Anatomy of a Privacy-Safe Large-Scale Information Extraction System Over Email. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2018)*. ACM, 734–743. https://doi.org/10.1145/3219819.3219901

[31] Brian Strope and Ray Kurzweil. 2017. Efficient Smart Reply, now for Gmail. https://www.blog.google/products/gmail/inbox-by-gmail-find-answers-even-faster/

[32] Saiganesh Swaminathan, Raymond Fok, Fanglin Chen, Ting-Hao (Kenneth) Huang, Irene Lin, Rohan Jadvani, Walter S. Lasecki, and Jeffrey P. Bigham. 2017. WearMail: On-the-Go Access to Information in Your Email with a Privacy-Preserving Human Computation Workflow. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST '17)*. ACM, 807–815. https://doi.org/10.1145/3126594.3126603

[33] Christophe Van Gysel, Bhaskar Mitra, Matteo Venanzi, Roy Rosemarin, Grzegorz Kukla, Piotr Grudzien, and Nicola Cancedda. 2017. Reply With: Proactive Recommendation of Email Attachments. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM 2017)*. Association for Computing Machinery, 327–336. https://doi.org/10.1145/3132847.3132979

[34] Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. 2016. 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 (SIGIR 2016)*. 115–124. https://doi.org/10.1145/2911451.2911537

[35] Steve Whittaker, Tara Matthews, Julian Cerruti, Hernan Badenes, and John Tang. 2011. Am I Wasting My Time Organizing Email?: A Study of Email Refinding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2011)*. ACM, 3449–3458. https://doi.org/10.1145/1978942.1979457

[36] Yonghui Wu. 2018. Smart Compose: Using Neural Networks to Help Write Emails. https://ai.googleblog.com/2018/05/smart-compose-using-neural-networks-to.html