ARTICLE

Research Frontiers in Information Retrieval

Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018)

Editors
J. Shane Culpepper, Fernando Diaz, and Mark D. Smucker

Authors and Participants (listed alphabetically)
James Allan, Jaime Arguello, Leif Azzopardi, Peter Bailey, Tim Baldwin, Krisztian Balog, Hannah Bast, Nick Belkin, Klaus Berberich, Bodo Billerbeck, Jamie Callan, Rob Capra, Mark Carman, Ben Carterette, Charles L. A. Clarke, Kevyn Collins-Thompson, Nick Craswell W. Bruce Croft, J. Shane Culpepper, Jeff Dalton, Gianluca Demartini, Fernado Diaz Laura Dietz, Susan Dumais, Carsten Eickhoff, Nicola Ferro, Norbert Fuhr, Shlomo Geva Claudia Hauff, David Hawking, Hideo Joho, Gareth Jones, Jaap Kamps, Noriko Kando, Diane Kelly, Jaewon Kim, Julia Kiseleva, Yiqun Liu, Xiaolu Lu, Stefano Mizzaro, Alistair Moffat, Jian-Yun Nie, Alexandra Olteanu, Iadh Ounis, Filip Radlinski, Maarten de Rijke, Mark Sanderson, Falk Scholer, Laurianne Sitbon, Mark Smucker, Ian Soboroff, Damiano Spina, Torsten Suel, James Thom, Paul Thomas, Andrew Trotman, Ellen Voorhees, Arjen P. de Vries, Emine Yilmaz, Guido Zuccon

Abstract

The purpose of the Strategic Workshop in Information Retrieval in Lorne is to explore the long-range issues of the Information Retrieval field, to recognize challenges that are on – or even over – the horizon, to build consensus on some of the key challenges, and to disseminate the resulting information to the research community. The intent is that this description of open problems will help to inspire researchers and graduate students to address the questions, and will provide funding agencies data to focus and coordinate support for information retrieval research.

1 Introduction

Over the past fifteen years, three Strategic IR Workshops have been organized in Lorne, Australia, all of which have had a singular vision – to look back at how research has evolved in the Information Retrieval community, and to look forward on where the research frontier is taking us. The first SWIRL workshop was organized by Alistair Moffat and Justin Zobel in 2004, and had 35 participants – several of which were PhD students. The major output of the meeting was the SIGIR Forum article "Recommended Reading for IR Research Students."[1]

In 2012, the second SWIRL workshop was organized by James Allan, Bruce Croft, Alistair Moffat, Mark Sanderson, and Justin Zobel. The theme of the workshop shifted away from previous work, and focused more on future directions for the IR research community. Together, the 45 attendees debated several possible research topics, and eventually converged on 6 main themes and 21 minor themes. These themes were then summarized and published in the SIGIR Forum article "Frontiers, Challenges, and Opportunities for Information Retrieval."[2]

Many of the themes described in the 2012 SWIRL report have seen significant progress in the ensuing years, but not all of them. At the 25th Anniversary TREC reception in 2016, several IR researchers reminisced about the SWIRL outcomes, and agreed that the major research directions in IR had evolved enough to warrant a third SWIRL. From these discussions, the main theme of the Third SWIRL emerged – How has research in IR evolved in the last five years, and where do we expect to be five years from now? In order to achieve this goal, a third SWIRL was organized by Shane Culpepper and Fernando Diaz. A total of 60 IR researchers, 20 from three regions (North/South America, Europe, Oceania) were invited to Lorne to discuss the future of IR research. This report captures the ensuing surveys and homework assignments in the lead up to SWIRL 2018, and summarizes the main outcomes of the meeting in Lorne.

1.1 Workshop Format

The workshop followed the format originally devised in the 2012 SWIRL meeting. On the first evening, a reception was held, and answers from the homework assignments were summarized and discussed. A bus then took all of the participants from Melbourne to Lorne the next morning. After lunch, six seed talks were given, and summarized below. On the second day of the workshop, the morning sessions were composed of six groups of ten participants breaking out and brainstorming about the future of IR based on the initial seed discussions. Each group then voted, and pitched three ideas that they thought were the most important. These 18 ideas were then grouped by similarity, and participants voted on the topics they were most interested in exploring further. The afternoon session then contained the breakout focus groups. A total of eight focus groups formed, and these make up the main sections of this report. Other topics that were proposed but that did not progress to the focus group stage are included at the end of the report as "Minor Topics". The final day of the workshop was a continuation of the focus groups. All participants finished up discussions on the topics, and worked together to produce a summary report of these discussions.

[1] https://doi.org/10.1145/1113343.1113344

[2] https://doi.org/10.1145/2215676.2215678

1.2 Invitation Questionnaire

As part of the initial RSVP for SWIRL, participants were asked what topics they thought were important. Table 1 shows the most common responses. The number of respondents suggesting the topic is shown in parenthesis. There was a strong consensus that Conversational Search, Machine Learning, and Fairness, Accountability, Confidentiality, and Transparency (FACT*) / Responsible IR were three important topics for discussion at the workshop.

1.3 Pre-Meeting Homework

1.3.1 Retrospective Questionnaire of Previous SWIRL Reports

The first homework task assigned to participants was to go back and read the SWIRL reports from 2004 and 2012, and asked three questions:

(1) What do you think previous SWIRL attendees accurately predicted about the future of Information Retrieval (i.e. true positives: what did we get right)?

(2) What do you think previous SWIRL attendees did not accurately predict about the future of Information Retrieval (i.e. false positives: what did we get wrong)?

(3) What do you think previous SWIRL attendees did not predict about the future of Information Retrieval (i.e. false negatives: what did we miss)?

From these questions, common themes were aggregated. In Table 2, we observed several interesting trends. While there was strong agreement about the second SWIRL missing Neural IR, but recognizing that Conversational IR would be important, perhaps the most interesting trend is the disagreement. For example, 10 participants thought that we were on target with the predictions about Mobile, while 11 others believed we got it wrong. Nevertheless, several important new trends were identified in this exercise, including the increasing importance of machine learning in IR, search bias & opinion engineering (broadly speaking – Fairness, Accountability, Confidentiality, and Transparency) Other topics of interest mentioned as important future directions included medical IR, monetization, video, green computing, efficiency (generally), session level search, ArXiV, reproducibility, cross device search, explainability of algorithms, and responsible IR.

conversational search (15)	responsible IR (5)
evaluation (13)	task-based IR (4)
IR and AI (12)	interpretability/decision support (4)
reproducibility (7)	virtual/augmented reality (3)
SIGIR organization (7)	user understanding (3)
new applications (6)	

Table 1: Important topics suggested during the initial RSVP process.

Question 1	**Question 2**	**Question 3**
Conversational IR (13)	Mobile (11)	Neural IR (24)
Structured (13)	Search as Learning (7)	ML Domination (11)
Mobile (10)	Zero Query Search (6)	Search bias (8)
Empowering Users (10)	Evaluation (6)	Online/User-Centered Evaluation (6)
Zero Query Search (7)	Simulated Interaction (5)	Opinion Engineering/Fake News (6)
Beyond Ranked Retrieval (6)	Axiometrics (5)	Virtual Assistants/Devices (4)
Search as Learning (6)	Personalization (5)	Social Media Search (4)
Opinion Engineering (6)		Whole Page Optimization (4)

Table 2: The most common responses from SWIRL participants on the retrospective questionnaire. Values shown in parenthesis are the total number of participants mentioning that item.

1.3.2 Important Papers Since SWIRL 2012

As part of the homework assignment, we asked participants to select one paper from within their area of expertise and one paper from outside of their area of expertise that they considered important for the information retrieval community. We manually classified all papers in order to understand the participants' perspective on recent research. The complete set of papers in these categories can be found at the end of this manuscript.

Table 3 shows the number of papers suggested by category in the homework responses. As expected based on the RSVP data, many participants selected papers from the machine learning community when asked about papers outside of the Core IR community. Deep learning was recognized as a fundamental tool that had powered significant advances in other fields. Sixteen participants cited Mikolov's word2vec paper as an important recent contribution [MSC$^+$13].

Another important theme was stateful search, defined to include conversational search and other multi-turn information access. Ten participants selected Radlinski and Craswell's theoretical model for conversational search [RC17].

The social implications of information access systems are beginning to get increased attention more broadly in the academic community [BS16]. Common themes in this area include algorithmic bias, ethics, and transparency. While there were no papers recommended by multiple participants, the subfield of *Responsible Information Retrieval* is growing.

There were two themes related to evaluation: experimentation and off-policy evaluation. While several of the participants selected Tetsuya Sakai's meta-analysis of previously-published results [Sak16a], many participants recognized issues with replicating and reproducing results, perhaps inspired by recent reproducibility concerns in psychology [Ope15]. The second theme concerned reuse of production log data for evaluating new treatments (e.g. algorithms, parameters). This problem occurs often when evaluating and training retrieval models in industry and has been receiving attention in the machine learning community. Several participants cited the work of Thorsten Joachims and his students as representative of this area [JSS17]. The issue of quantifying the effect of unjudged documents was also a common theme, with Rank-Biased Precision [MZ08] and its successor INST [MBST17] being examples of weighted-precision metrics in which this ability was specifically explored.

Finally, an efficiency theme also emerged. The most commonly referenced themes to watch in

Deep Learning (16)	Stateful Search (10)
Responsible IR (9)	Experimentation (7)
Efficiency (6)	Counterfactual/Off-policy Evaluation (5)
Cognitive Effects (5)	Temporal IR (4)
User/Topic Variability (3)	Social Effects (2)
Recommendation (2)	Rank-Biased Precision (2)
Marketplaces (2)	Brain (2)
Multiturn IR (1)	

Table 3: Important paper topics since SWIRL 2012.

this area were related to improving the efficiency in learning stages of multi-stage retrieval systems [LNO^{+}15], explicitly learning trade-off costs [CGBC17], and the exciting new area of combining learning and indexing [BNMN16, KBC^{+}17].

1.4 Summary of Seed Talks

Based on the RSVP questionnaires and the homework, six "fire starter" talks were proposed, with the goal of capturing the interests of the participants, and to be *provocative*. The seeded talk topics were stateful search (conversation, exploratory search, task-based), reproducibility (collection design, experimentation), evaluation metrics (online and offline measures), fairness/transparency (algorithmic bias), user issues (cognitive biases), and system performance/indexing (scalability, machine learning algorithms, ranking). A brief summary of each talk is provided here for future reference.

Stateful search. This talk focused primarily on the challenges with task-based search. The key problems identified were task extraction / representation; task-based evaluation; design considerations in task-based retrieval systems, and task-driven personalization. One of the key arguments was that search will be a many-device problem, and so better task abstractions are needed. Other thoughts on the increasing importance of conversational IR were presented and discussed.

Reproducibility. This talk explored the dilemma in IR on the problem of reproducibility. Everyone believes it is important, but it is not **new** research. So, an argument was made that we need to understand the link between reproducibility, validity, and performance prediction. More importantly, we need a shift in culture, where reproducibility studies are part of the research process, and this work is somehow acknowledged as part of career progression. Other important questions were deciding *what* should be reproduced, and having the proper assessment tools in place to help us know *when* we can consider something has been sufficiently reproduced.

Evaluation Metrics. This talk focused on the rift between resources and approaches between academic and industry practitioners. More specifically, how do we bridge the gap between online and off-line measurement of search quality? *Success* in complex systems is an end-to-end process, but many of our tools look at individual components at small scale. Interactions between components and people is often ignored. So the provocative question proposed was: Can we get rid of off-line evaluation all together? Since the future includes mobile, personal assistants, and "intelligent" systems, the notions of relevance and meta-relevance become even more muddled.

Fairness & Transparency. This talk focused on FACT* (Fairness, Accountability, Confidentiality, Transparency, Ethics, Bias, Explainability, Interpretability, ...) in IR. This is an area receiving a great deal of attention in the IR community at the moment. Several interesting problems were highlighted, including:

- **IR without bias.** How to avoid "unfair" conclusions even if they appear true?
- **IR that ensures confidentiality.** How to produce results without revealing secrets?
- **IR without guesswork.** How to produce results with a guaranteed level of accuracy? Would that help or harm? When and why?
- **IR that provides transparency.** How to clarify results such that they become trustworthy?

Long term problems in machine-altered reality, when questions should / should not be answered, and autonomous algorithmic intervention were described, in addition to shorter term research problems of documenting biases / risks in current datasets/tools, end-to-end analyses of bias, and explainable IR systems to help people make better decisions. This is clearly a multi-disciplinary problem affecting many research communities.

User Issues. This talk argued that we should be talking about people, and not users, who are not just actors who will stop doing what they are doing to engage in an IR system. The key point is that in the emerging technological and social-technical environment, people will be constantly and ubiquitously emerged in a sea of information. As such, several different future "users" were described. They were:

- Ubiquitous Users who are immersed in a sea of information from the Internet of Things;
- Thinking Users, where cognitive and neurophysiological conditions affect interactions with information;
- Working Users, where search is a complex combination of a multiplicity of tasks; and
- Social Users, which encompass how systems can be designed to respond to a persons social environment, in terms of supporting their interactions not only with information, but also with others.

The key overall argument was that information interaction should be the focus, and not the systems themselves.

Efficiency. This talk argued that the value of efficiency continues to be an important research area in IR. A total of three challenges were presented. The first challenge was at the systems level – How do we explore the trade-offs between efficiency and effectiveness as systems become increasingly more complex? The second challenge focused on efficient learning and NLP – How do we scale complex neural networking models, and find a balance between quality and cost in the NLP models being used in IR? The third challenge was around multimodal indexing – As we move beyond text, how do we efficiently combine, index, and search many different data formats?

1.5 Summary of Brainstorming Breakout Sessions

Six breakout groups discussed themes from the seed talks as well as any other topics that participants felt was not covered in those talks. After aggregation, the following themes emerged,

- **Decision Support over Pathways**: Understanding and designing systems to help people in making decisions.
- **Generating New Information Objects**: *Ad hoc* generation, composition, and summarization of new text, and layouts in response to an information need.
- **Transparent/Explainable Information Retrieval**: Explaining ranking decisions. Providing reliable and responsible information access.
- **Cognitive-aware IR**: Tracking and modeling user behavior and perception. Modeling political-correctness of decisions. Identifying fake news and provenance.
- **Societal impact of information retrieval**: Understanding the long term impact of IR on society and the economy.
- **Personal information access**: Federated personal information search and management (e.g. knowledge graphs). Biometrics for affective state.
- **Next Generation Efficiency-Effectiveness Issues**: Efficient machine learning inference. Resource-constrained search.
- **Machine Learning and Search**: Developing effective machine-learned retrieval models (e.g. neural networks, reinforcement learning, meta-optimization).
- **Personalized interaction**: Diversified and personalized interactions.
- **Conversational information access**: Information-seeking conversations. Learning representations for conversations.
- **New approaches to evaluation**: Moving beyond the Cranfield paradigm, topical relevance, and queries. Controlling for variability. Counterfactual evaluation and off-policy evaluation.
- **New interaction modes with information, multi-device search**: Multi-device search.
- **Blending online and physical**: Search in the context of mobile, smart environments, and augmented/virtual reality.
- **Task-specific representation learning**: Adapting machine learned models for new search domains.
- **Pertinent Context**: Surfacing and using the relevant contextual information for search.
- **Success prediction**: Formal models and principles to inform retrieval system design (build the right bridge instead of build six bridges and see which survives).

1.6 Summary of Focus Group Breakouts

A straw poll was held for participants to identify the three topics they found most interesting. This resulted in eight topics which formed the final breakout focus groups. The focus groups spent the final day of the workshop discussing their topic, and developing the summary reports found in the following sections. The eight themes that emerged were:

Section 2: Conversational Information Seeking

Section 3: Fairness, Accountability, Confidentiality and Transparency in Information Retrieval

Section 4: IR for Supporting Knowledge Goals and Decision-Making

Section 5: Evaluation

Section 6: Machine Learning in Information Retrieval (Learnable IR)

Section 7: Generated Information Objects

Section 8: Efficiency Challenges

Section 9: Personal Information Access

In addition two minor themes emerged from the workshop, and were included in the report in Section 10. These were "IR for an IoT World" and "Impact of IR Systems in Society". The remaining sections summarize the findings for all ten of these themes.

In the remainder of this report, each of the above themes is detailed in its own section. Each section follows a standard format with subsections of: description, motivation, proposed research, research challenges, broader impact, broadening SIGIR, and obstacles and risks.

2 Conversational Information Seeking

2.1 Description

Conversational information seeking (CIS) is concerned with a task-oriented sequence of exchanges between one or more users and an information system. This encompasses user goals that include complex information seeking and exploratory information gathering, including multi-step task completion and recommendation. Moreover, CIS focuses on dialog settings with variable communication channels, such as where a screen or keyboard may be inconvenient or unavailable.

Building on extensive recent progress in dialog systems, we distinguish CIS from traditional search systems as including capabilities such as long term user state (including tasks that may be continued or repeated with or without variation), taking into account user needs beyond topical relevance (how things are presented in addition to what is presented), and permitting initiative to be taken by either the user or the system at different points of time. As information is presented, requested or clarified by either the user or the system, the narrow channel assumption also means that CIS must address issues including presenting information provenance, user trust, federation between structured and unstructured data sources and summarization of potentially long or complex answers in easily consumable units.

2.2 Motivation

Conversations are a natural form for humans to seek information, and there are decades of study on formal dialogues and interactions of users with reference librarians. The natural next step is to design automated systems that are 'virtual librarians', eliciting information needs, correcting misconceptions, and providing the right amount of information at the right time across all possible domains. Multi-turn conversations should also become more natural in the digital environment today due to the increasing variety of devices that are accessible anytime/anywhere (perhaps without screen or keyboard), the maturity of speech interfaces, and recent developments in general representation learning. Today's digital assistants are only capable of very basic "conversations", which usually means a single user question ("What's the weather like today?" or "When does my flight leave tomorrow?"), followed by a single system answer. In contrast, this research direction will lead to multi-turn, multi-user, multi-task and multi-domain conversational information seeking systems.

2.3 Proposed Research

Development of conversational information seeking systems requires new research on a broad range of topics related to information elicitation, user modeling, precision-oriented search, exploratory search, generated information objects (Section 7), description of retrieval results, session-based search, dialog systems capable of sustained multi-turn conversations, and evaluation. The IR community is well-positioned to work on these issues due to its deep roots in studying elicitation, information seeking, information organization, and what makes search difficult. Meaningful progress is likely to require partnering with colleagues in research areas such NLP, dialog, speech, HCI, and information science that have complementary skills, thus broadening and enriching the field. Several promising research directions are described briefly below, to give a sense of what this topic entails.

User Models. User modeling in conversational information seeking systems involves inferring, representing, and updating information about a person (from general information about their tastes and conversational style to their current cognitive and emotional state), their state of knowledge surrounding the current topic, their current goal(s), and their previous interactions with the system. The user model informs predictive tasks. For example, based on the user model, the system can decide what information to elicit from the user, how to elicit the information, and what information to provide. We note that elicitation is one key difference from traditional search engines, allowing the system to proactively focus on resolving uncertainties in a person's information need, both for the system and for the user. It also allows a person to explicitly refer to previous conversations with the system as a form of grounding or disambiguation.

Important research questions involve knowing when to take the initiative; inferring satisfaction; understanding which attributes of conversational interactions influence outcomes related to engagement and/or mental workload; and knowing when the information seeking session has concluded.

Finding Information. Conversational information seeking systems will require distinct search strategies for different conversational states, for example, precision-oriented search when the information need is specific or focused, and diverse recall-oriented search when the information need is uncertain or exploratory. Natural conversational delays create opportunities for anticipatory search or deeper analysis of search results to prepare for likely next states in the dialog. After the system gathers information, it must organize, summarize, and describe what it found. The type of organization and summarization depends upon the user's state, the state of the dialog and the mode of communication. For example, it may be organized to provide a broad overview of the key concepts and to elicit additional information from the user by supporting drilling down into specific topics or information sources; or when the focus is narrow and specific, it may be an abstractive summarization that covers key information supported from multiple sources.

Engagement. In order to make a conversational information seeking system engaging to a wide variety of people over a prolonged period of time, the system should exhibit affective traits: it should be able to convey humor, sympathy and other traits in its interactions with its users in a personalized manner. At the same time, the interactions need to enable people to build an accurate mental model of the system's abilities to avoid causing disappointment, for example when having repeated conversational turns that lead to unsatisfactory outcomes. We also note that response time is likely to be a critical component of engagement in CIS systems; they must respond in a tolerable time, otherwise people will discontinue use of the system.

Domain generality and specificity. Like traditional web search engines, some conversational information seeking systems will support conversations across diverse domains – potentially all domains of human knowledge. For example, conversational information seeking may begin with the request "Tell me about dementia". Such systems will require development of general-purpose methods of eliciting, describing, and engaging. This type of generality is not yet possible with task-oriented dialog systems.

General open-domain systems will lack the depth and domain expertise that is possible in domain-specific conversational systems. A domain-specific application may define focused domain-specific intents on classes of entities in a specialized knowledge base curated by experts. This parallels current work in dialog systems focused on domain-specific models of intents with predefined schema and slots. We envision that specialized conversational models are needed to perform deep conversational tasks, for example a doctor performing more detailed research ("Tell me about the relationship between dementia and thyroid problems.")

Failure modes. Given the complexity of multi-turn conversations, failures will occur that may cause the user to end the conversation, for example, inability to refer to an earlier conversational turn, failure to find information, or failure to understand retrieved information. A failure may be caused by an individual component or by interactions across components. When failure happens, the system should guide a person to provide information that allows recovery. For example, instead of saying, "I can't help you with that" or falling back to reading web results, the system should engage the user to recover or avoid the issue.

A key challenge is that the system should "know what it doesn't know" to express gaps in understanding of the request or the underlying information. This means that the system necessarily needs to quantify its confidence in the responses generated, in terms of i) whether the system properly understood the utterance/request, ii) whether it was able to retrieve appropriate information, iii) whether it was able to properly organize and aggregate retrieved items into generated information objects (Section 7), and iv) whether it was able to render results (through best answer selection, summarization, etc.) clearly to the user.

When failure occurs, the user may correct the system, for example by issuing a comment of the form “No, what I meant was ...”. Such corrective feedback offers an opportunity for the system to both reinterpret the current dialog state, (realigning the current information seeking process toward successful conclusion), as well as provide useful training data for improving the system for subsequent interactions.

Evaluation. Developing successful systems requires further understanding of what constitutes a successful information seeking conversation. One starting point could be to create collections of human-human information seeking conversations in which one person plays the role of the system (with access to one or more information services and/or domain expertise) and the other person plays the role of the user. So-called Wizard of Oz studies can be used to gather example conversations and use questionnaires to measure outcomes such as task completion, workload, and perceived usability. By doing so, we may be able to gain insights about the correlations between specific conversation characteristics and different outcomes.

Evaluating a conversational information seeking system requires component-wise and end-to-end evaluation methods. CIS systems involve several components that can be evaluated individually using specialized evaluation methodologies. First, given the input from the user, the system needs to accurately detect the task the user is trying to perform and the state they are in with respect to the task. The system should also be able to detect the possible next state(s) of the user, or their possible next task(s)/goal(s). Hence, evaluation methodologies need to be designed in order to evaluate how well the system understands the user's task and state, and predicts future user needs. Since the information delivered will be personalized, we also need methods to evaluate the quality of personalization. Finally, some conversational responses will involve summarization. Prior research developed metrics for evaluating summarization quality; however, summarization in the context of CIS systems is likely to require different metrics because the response the user expects from such systems is quite different than typical document summarization.

Besides component-wise evaluation methods, it is also critical to have an end-to-end evaluation approach. End-to-end evaluation is necessary to compare between different systems and to determine whether dialog is the appropriate mode of interaction (e.g., compared to a more traditional mode). In this respect, we need methods and metrics that can be compared across different modes of interaction. These metrics may need to consider outcomes such as user engagement, satisfaction, and task completion time.

Multi-modal conversations. Informed conversations between humans often involve supplementary evidence such as documents, images, or videos. The current multiplicity of devices could support such multi-modality provided that we can gain a better understanding of what are the appropriate modes depending on the device and user context (previous/current activity, location), and potential materials available to the user (for example, the query may include an image) and to the system (for example, the system response may include audio: “Do you mean a sound like this?”). Models of turn taking (feedback, granule of information) to drive conversations will need to incorporate the possibilities for a variety of devices and modalities.

Cross-device conversations. When a conversational information seeking system is designed to support complex exploratory tasks, one needs to take cross-device behavior into account, since the information seeking session might continue for a long time under different circumstances. A CIS system should be able to optimize the query acquisition, clarification, and presentation methods based on a device at hand. Supporting users to effectively resume an ongoing task across different devices with multiple modality (e.g., audio, text, multimedia) can be challenging.

Collaborative information seeking. Conversation/dialog does not necessarily occur between a CIS system and a single user. The system should also be able to support multiple users or a group of users who engage in a collaborative task. This involves user identification and tracking during an information seeking session, mining information needs, relevance feedback, pertinent contextual factors from collaborative conversations, and personalization / diversification of results for the group or for the individual members of the group. Sensing the state of the discussion in the group can also be an important signal to optimize the seeking session.

2.4 Research Challenges

Conversational IR systems can be seen as a federation of agents or subsystems, but they will also be inherently complex systems, with models that will reach beyond the boundaries of individual components. With that will arise challenges such as how to bootstrap such systems with reasonable effort, how to ensure they are responsive as a whole, how to perform component-wise diagnosis, and at what level to consider their robustness.

Ethical challenges arising across the field of information retrieval, such as trust in information, biases, and transparency, will likely be exacerbated by the inherent narrowing of the communication channel between the systems with their users.

2.5 Broader Impact

Current search engines are widely used in many settings. Conversational IR systems could replace or augment these for many tasks, reducing the cognitive burden on the user and potentially supporting them to achieve success more often or to improve the efficiency or ease of their search. Effective search agent design will enable a greater level of control and transparency of search process and outputs. Conversational IR can provide support to users who are initially unable to express their information need sufficiently well to properly begin a search task, for example by providing feedback after a vague initial search and eliciting more information to progressively build a meaningful expression of the information need. The ability of a search system to remember all or part of previous search sessions may prevent the user from needing to repeat previous search tasks or to provide support by reminding about previous search activities (e.g., "When you looked for this before, you were interested in these items").

Proactivity by the search agent could provide people with details about their search topics, retrieved documents, or opinions expressed in the documents. Proactive analysis and reporting can enable broader, less-biased perspectives of a given topic, leading to improved information literacy of end-users. Search currently requires a person to break off from their current activity or task and to engage in a separate activity. Conversational search may be more fully integrated in their work. For example, maintaining details of previous search inputs, results obtained, and monitoring of ongoing work to be able to provide context relative search. In addition, rich modalities in conversational search interaction (e.g., speech, sound, text, multimedia) can achieve an inclusive system for a wide range of users and situations including low literacy, disability, in hands-busy environments.

2.6 Broadening SIGIR

Development of successful conversational IR systems will require significant expertise in eliciting, finding, and delivering information, which are core strengths of the information retrieval research community. It will also require user modeling, dialog systems, speech interfaces, and HCI skills that provide opportunities for collaboration with colleagues in other areas of computer science. People tackling this research problem will need to work across disciplines.

2.7 Obstacles and Risks

There are several obstacles and risks to research on this topic. Reusable datasets may be difficult to design or acquire due to the personalized, interactive nature of the task and the detailed temporal user models it develops. Conversational information seeking systems may retain information about a person over long periods of time, which raises privacy and legal issues. People could be uncomfortable with systems learning and retaining detailed information about what they know and how they acquire information. There is a possibility that successful systems might expose people to a broader range of information than they consider now; and a risk that more effective organization and filtering of information might discourage critical thinking. Finally, the level of language understanding required to provide useful assistance might be too difficult to enable more than shallow systems during the next 5-10 years.

3 FACT IR: Fairness, Accountability, Confidentiality and Transparency in Information Retrieval

3.1 Description

IR is about connecting people to information. However, as with all software-based systems, IR systems are not free of human influence; they embed the biases of those that create, maintain and use them. Empirical evidence suggests that certain communities have differential access to information; in other words, their needs might not be equally well supported or certain information types or sources might be more or less retrievable or might not be well represented. In addition, as we increasingly rely on the outcome of IR systems such as search engines, recommender systems, and conversational agents for our decision making, there is a growing demand for these systems to be explainable. Such problems are related to many fundamental aspects of information retrieval, including information representation, information or answer reliability, information retrievability and access, evaluation, and others. While, traditionally, the IR community has been focused on building systems that support a variety of applications and needs; it is becoming imperative that we focus as much on the **human**, **social**, and **economic** impact of these systems as we do on the underlying algorithms and systems.

We argue that an IR system should be **fair** (e.g., a system should avoid discriminating across people), **accountable** (e.g, a system should be reliable and be able to justify the actions it takes), **confidential** (e.g., a system should not reveal secrets), and **transparent** (e.g., a system should be able to explain why results are returned). Judgment is needed sometimes to balance these four considerations (e.g., it is responsible to bias against unreliable sources). Other communities, such as the machine learning, artificial intelligence, and computational social science, are already focusing on these and other related issues, including how human behavior online is confounded by algorithmic systems, how we can audit black box models, and how can we maximize benefit and reduce risks. The research directions we describe here aim to increase these efforts as they apply to IR systems.

3.2 Motivation

Why does it matter for IR? IR systems often capture associations between entities and/or properties, and depending on the semantic connotations of such relationships they might lead to reinforcing current stereotypes about various groups of people, propagating and amplifying harm. For example, these associations may originate from the data used to train the ranking models, which may not provide enough coverage for all possible associations such that they can all be learned. Certain groups of individuals may be over- or under-represented in the data, which could be a reflection of greater societal disparities (e.g., unequal access to health care can result in unequal representation in health records) or of the types of people who are able to contribute content, including the rate at which these contributions are made (e.g., women tend to be over-represented in Instagram data, but under-represented in StackOverflow data). Representation is also affected by the quality of the tools used to capture the data. For example, it is more difficult to do facial recognition of dark-skinned people in video surveillance footage because of limitations with how cameras are calibrated. As a result, an image retrieval system might fail to properly identify images related to darker-skinned people, while an image assessment system might flag them more often for security interviews, or to scrutinize them in more detail.

What makes this specific to IR? Given the ubiquitous usage of IR systems, often broadly construed (e.g., search, recommendation, conversational agents), their impact — negative included — is potentially wide ranging. For instance, research has shown that people trust more sources ranked higher in the search results, but the ranking criteria may rather rely on signals indicative of user satisfaction, than on those indicative of factual information. For consequential searching tasks, such as medical, educational, or financial, this may raise concerns about the trade-offs between satisfying users and providing reliable information.

The SIGIR community has the responsibility to address fairness, accountability, confidentiality and transparency in all aspects of research and in the systems built in industry. Similar responsibility issues are addressed in related fields, however, there are specific issues in IR stemming from the characteristics of, and reliance on document collections and the often imprecise nature of search and recommendation tasks. IR has a strong history of using test collections during evaluation. As evaluation tools, test collections also have certain types of bias built-in. For example, the people who construct topics and make relevance assessments arguably are not representative of the larger population. In some cases, they have not been representative of the type of users who are being modeled (e.g., having people who do not read blogs evaluate blogs). Evaluation measures are designed to optimize certain performance criteria and not others, and either implicitly or explicitly have built-in user models. Systems are then tested and tuned within this evaluation framework, further reinforcing and rectifying any existing biases. For example, in building test collections, bias should be avoided by ensuring diversity in the sources of documents included and using people from diverse backgrounds to create topics.

What are examples of human, social, and economic impact? Infrastructure and accessibility variations may introduce differential representation in training data. For example, research has shown that social media accounts with non-Western names are more likely to be flagged as fraudulent, and argue that this is because the classifiers have been trained on Western names. Bias can also be introduced by the interfaces and tools that are presented to users. For example, query autocompletion is a common feature of search systems that learn suggestions from past behaviors of users; however, often the people who type queries about particular topics are from a specific segment of the population and the intent behind their queries is often unclear. For example, the query prefix "transgenders are ..." results in offensive autocomplete suggestions of "transgenders are freaks" and "transgenders are sick".

Another motivation for this work is the growing concern about the understandability, explainability and reliability of deep learning methods including the complexity of the parameter space. These techniques are being used in a variety of domains to assist with a range of high-impact tasks such as in the medical domain for diagnosis and the intelligence community for detecting threats and combating terrorism. Many of the domain experts working in these fields are not satisfied with a simple answer, but rather desire to know about the reasoning and evidence behind the answer that the system produces because the decisions they are making can have significant consequences. Moreover, the engineers who create systems often do not understand which parts of the system are responsible for failures, and it can be difficult to trace the origins of errors in such complex parameter spaces. However, it is unclear how such explanations, evidence-trails and provenance might be communicated to the various user groups and how such communications might change behaviors, and the quality, quantity, and nature of human-computer interaction.

We, the IR community, should take the initiative before others do in the face of changing legal frameworks. For example, in Europe with the General Data Protection Regulation, individuals have a right to erasure of personal information and a right of explanation. IR systems need to incorporate these rights. Thus, an indexing scheme needs to be able to delete information and search results may require explanation.

3.3 Proposed Research

We propose an agenda driven by the ideal of incorporating social and ethical values into core information retrieval research and development of algorithms, systems, and interfaces. This necessitates a community effort and a multi-disciplinary approach. We focus on fairness, accountability, confidentiality, and transparency in IR:

- Fair IR
  - How to avoid "unfair" conclusions even if they appear true?
  - For instance, in the case of people search, how do we make sure that results do not suffer from being underrepresented in the training data?
  - Avoid "discrimination" even when attributes such as gender, nationality or age are removed. And even when the vox populi dictates a certain ranking. Avoid selection bias and ensure diversity.
  - To what extent is the assortment of information objects presented to us representative of all such objects 'out there'?

  - How can we measure and quantify fairness of an IR system?
  - Evaluation of fairness vs. fair evaluation. How can we measure ‘harm’, and variations in ‘harm’ across verticals?

- Accountable IR
  - How can we avoid guesswork and produce answers and search results with a guaranteed level of accuracy?
  - Would providing such a guaranteed level of accuracy help or harm? When and why?
  - Attach meaningful confidence levels to results. Handling veracity issues in data. When to roll out hot-fixes? Rankings with solid guarantees on the reliability of the displayed answers and results.
  - How might the assortment of information objects presented to us impact our perceptions of reality and of ourselves?

- Confidential IR
  - How to produce results without revealing secrets?
  - Personalization without unintended leakage of information (e.g., filter bubbles) by randomization, aggregation, avoiding overfitting, etc.

- Transparent IR
  - How to clarify results such that they become trustworthy?
  - Automatically explaining decisions made by the system (e.g. retrieved search results, answers, etc.) allowing users to understand “Why am I seeing this?”
  - Traceability of results (e.g., link to raw data underlying entity panels).

3.4 Research Challenges

These general research questions manifest themselves along the entire information retrieval “stack” and motivate a broad range of concrete research directions to be investigated:

Does the desire to present fair answers to users necessitate different content acquisition methods? If traceability is essential, how can we make sure that basic normalization steps — such as content filtering, named entity normalization, etc. — do not obfuscate this? How can we give assurances in terms of fairness towards novel retrieval paradigms (e.g., neural retrieval models being trained and evaluated on historic relevance labels obtained from pooling mainly exact term-matching systems)?

How should we design an information retrieval system’s logging and experimental environment in a way that guarantees fair, confidential, and accurate offline and online evaluation and learning? Can exploration policies be designed such that they comply with guarantees on performance? How are system changes learned online made explainable?

Indexing structures and practices need to be designed/revisited in terms of their ability to accommodate downstream fairness and transparency operations. This may pose novel requirements towards compression and sharding schemes as fair retrieval systems begin requesting different aggregate statistics that go beyond what is currently required for ranking purposes.

Interface design is faced with the challenge of presenting the newly generated types of information (such as provenance, explanations or audit material) in a useful manner while retaining effectiveness towards their original purpose.

Retrieval models are becoming more complex (e.g., deep neural networks for IR) and will require more sophisticated mechanisms for explainability and traceability. Models, especially in conversational interaction contexts, will need to be "interrogable", i.e., make effective use of users' queries about explainability (e.g., "why is this search result returned?").

Recommender systems have a historic demand for explainability geared towards boosting adoption and conversion rates of recommendations. In addition to these primarily economic considerations, transparent and accountable recommender systems need to advance further and ensure fair and auditable recommendations that are robust to changes in product portfolio or user context. Such interventions may take a considerably different shape than those designed for explaining the results of ranked retrieval systems.

User models will face the novel challenges of personalizing retrieval services in a fair, explainable, and transparent manner. This is particularly relevant in the context of diversity and the way in which biased or heavily polarizing topics and information sources are handled. Additionally, transparent retrieval systems will require new personalization techniques that determine the right level of explanation that fits different sets of requirements (e.g., explanations that are effective for novice searchers, professional journalists or policy makers vs. explanations for highly technology-affine search engineers investigating system failures). Finally, such personalization should be reliable in terms of robustness to confounding external context changes.

Efficiency will be a key challenge in serving explanations at real time. Structures and models will need to accommodate for on-demand calculation as well as caching or approximate explanations in order to meet run time and latency goals. In addition, a key challenge will be the design of indexing structures and models that are fair without compromising efficiency.

There may surface a need for new **evaluation metrics** that capture the quality of an explanation and that understand user satisfaction as a composite of immediate goal accomplishment as well as fairness, trustworthiness and transparency considerations. Depending on the concrete application, there are different trade-offs between the severity of different error types (e.g., misses vs. false alarms). Such investigations of failure consequences can be conducted at different temporal resolutions, ranging from immediate short-term effects on single users to long-term consequences at a population-wide resolution. Finally, this line of evaluation motivates a stronger separation between organic traces of user behavior and users' reactions to platform and algorithm properties.

There are numerous challenges that the desire to realize fair, accountable, confidential and transparent IR systems poses that cut across the IR stack. This includes, for instance, maintaining confidentiality of information when providing explanations involves indexing structures, models, and evaluation.

3.5 Broader Impact

As information retrieval systems (search engines, recommender systems, conversational agents) touch on every aspect of our life, the technology that we help develop should be informed by, and inform, the world around us. This starts with understanding and replying to immediate stakeholders — users, decision makers, and engineers — across applications such as web search, information discovery in social networks, HR analytics, medical search, and e-commerce.

But there are broader concerns. Studies suggest that online platforms have impacted society by leading to increasing polarization; changing the metrics can help to begin to assess and understand such issues. The IR field needs to recruit more diverse people, and not just as collaborators, but also students. These changes might help IR researchers better understand the methods (and communicate about them), which in turn could lead to additional insights and theoretical developments.

Finally, a stronger emphasis on transparency will likely force us to document our methods and experiments in a better and more systematic ways. This, in turn, will positively impact teaching and reproducibility in IR.

3.6 Broadening SIGIR

Many questions pertaining to responsible and accountable technology originate in other scientific communities. Often, they are social, ethical, or legal in nature rather than purely technical. We need technical skills to solve them but we should collaborate with social scientists, psychologists, economists and lawyers, e.g., to understand the impact of using FACT IR systems in society, to be exposed to suitable ethical frameworks, and to anchor the definitions of the core concepts in FACT IR, such as what is an explanation in scientific discourses that have considered such notions for decades.

3.7 Obstacles and Risks

To enable this research we need broad collaborations between IR researchers and communities outside IR. Finding effective ways of collaborating and finding a shared language requires considerable effort and investment that may not be properly "rewarded" by funding bodies and evaluation committees.

An important **risk** concerns the diversity of perspectives on the **definition** of core concepts such as fairness, ethics, explanation or bias across scientific and engineering disciplines, governments or regulating bodies. Having more transparent IR systems could make systems more vulnerable for adversaries as knowledge about the internals of systems need to be shared through explanations.

A potential obstacle is initial resistance from system developers and engineers, who might have to change their workflows in order for systems to be more transparent. Another possible obstacle is the tension between transparency and fairness, and an enterprise's commercial goals.

An inadvertent risk is introducing a new type of bias into our systems about which we are unaware.

4 IR for Supporting Knowledge Goals and Decision-Making

4.1 Description

IR systems should support complex, evolving, or long-term information seeking goals such as acquiring broad knowledge either for its own sake or to make an informed decision. Such support will require understanding what information is needed to accomplish the goal, scaffolding search sessions toward the goal, providing broader context as information is gathered, identifying and flagging misleading or confusing information, and compensating for bias in both information and users. It requires advances in algorithms, interfaces, and evaluation methods that support these goals. It will be most successful if it incorporates growing understanding of cognitive processes: how do people conceptualize their information need, how can contrasting information be most effectively portrayed, how do people react to information that flies in the face of their own biases, and so on.

4.2 Motivation

People seek to satisfy various information needs that involve acquiring knowledge and/or making decisions, such as learning about world affairs from reading news articles, understanding their medical problems and possible treatments, or training for a job. Invariably, retrieval systems fall short of the best possible outcome, or even user expectations. The user may have had to expend more effort than ideally needed, or ended up with information that is inaccurate, biased, or lacking utility.

In order to successfully accomplish such knowledge-seeking and decision-making tasks, users often need more support than that currently offered by information systems. This support needs to be offered at different stages in the information seeking process, starting even before an information need is expressed: a search system should be aware of the context of the user in which the information need is to be placed and of the user's existing skills and knowledge. If a more complex task is to be accomplished (such as gathering different forms of evidence for a decision involving multiple constraints or aspects), the system may help by scaffolding the task at every step, as needed by the user. The system needs to be aware of biases of the user and/or the search results and take those into account when presenting these results to end up with the best possible outcome. Similarly, the user should be made more aware of the broader context in which the returned information exists. Ideally, a system should also be aware of and be able to competently deal with distractions or lack of motivation of the user.

While these demands on a retrieval system in a sense have always existed, it is more pertinent than ever that these are incorporated in the information retrieval process. Technology is much better suited now to help fulfill these requirements on the one hand, and on the other, there is greater scope for the user to end up more misinformed after a search than before. To give an example, search systems (and related algorithms, such as ranking algorithms employed by social media systems) contributed to large amount of misinformation during the 2016 presidential election cycle in US politics.

Finally, as learning is supplemented more and more with online technology, improved methods for getting students the right information for their learning goals could help increase student engagement, curiosity, and retention, as well as, in the longer-term, enable better knowledge transfer to other courses. As we describe in 'Broadening SIGIR' below, advances in this research area could also lead to advances in psychology and learning sciences as systems are used to give insights into, e.g., the relative fidelity of cognitive models in predicting outcomes or enhancing learning.

Developing solutions in these areas also includes significant challenges in evaluation. Current evaluation metrics and methods do not adequately capture notions of users' satisfaction, confidence, and trust, or the quality of the outcomes or decisions made based on the search process. Users may be trying to fill different types of knowledge gaps and their goals may evolve. Methods are needed to evaluate whether they fill these gaps or not and whether they fill them with correct, adequate, and contextualized information. Evaluation needs to consider these complex, longer-term (and possibly on-going) aspects of the users' information goals.

4.3 Proposed Research

The proposed research consists of several main threads: (1) understanding cognitive aspects of users that are relevant to their information seeking, (2) investigating ways that search systems can provide information (beyond ranked lists and underlying documents) that will aid searchers in evaluating and contextualizing search results, (3) exploring ways that search systems can help users move through a learning or decision-making process, and (4) overcoming challenges in evaluating how well systems support users in learning and decision making.

The first research area is concerned with understanding cognitive aspects of users that may influence their interactions with information returned from search systems. For example, users may have pre-existing knowledge and biases, differing levels of curiosity and trust, or even different learning strengths. Research needs to explore how search systems can detect, represent, and productively utilize cognitive aspects of users to help support learning and decision support during search processes. Specific research questions include: (a) How do cognitive models and processes affect searching and vice-versa? What cognitive biases make content more difficult to absorb? (b) How do people assess content (e.g., Is this information true/factual versus opinion/biased? How does this information relate to other content I've seen before?), (c) How do we detect and represent users' knowledge and knowledge states, cognitive processes, and the effort and difficulty of processing information?, and (d) How do we represent different information facets for users to support meta-cognition?

The second area focuses on investigating ways that search systems can represent and provide information so as to aid searchers in evaluating and contextualizing search results. Research questions in this area include: (a) what information or sources of information can be provided to help users overcome their cognitive biases (e.g. teenage moms might trust other teenage moms); (b) what visualizations or presentations are useful to convey relationships between known and new information? (c) what interface choices leverage a user's cognitive biases in order to lead them to better learning or more informed decisions; and (d) what types of metadata can be presented (and how can it be presented) to help a user understand the biases, trustworthiness, provenance, or utility of information?

The third area focuses on exploring ways that search systems can provide more explicit interaction/interface support to users who are searching in order to help make a decision or to engage in ongoing learning about a topic. Research questions in this area include (a) How can we help users assess and contextualize information returned by search systems (e.g., quality, trustworthiness, opinion vs. fact, position of the information in the domain space), (b) How do we go beyond topical clustering to uncover structure relevant to users' knowledge goals (e.g., alternative views), (c) How can systems encourage meta-cognition and reflection, thus providing scaffolding and mentoring toward their goals?, (d) how well does IR technology support decision making (comparing items, understanding dimensions, testing hypotheses).

The fourth area focuses on developing new evaluation models suited for evolving, complex, and longer term information seeking. Research questions in this area include (a) understanding how well algorithms and systems support users in such tasks? (b) how do we measure the impact of differing cognitive processes on information seeking? (c) how do we measure success for long-term tasks where satisfaction may be ephemeral or may change in light of information acquired later? (d) how do we measure the quality of the ultimate decision, the user's satisfaction with the decision or the process, or the depth of the user's learning during or after looking for information.

Prioritization/Progression: Near-term work in this area could focus on understanding and supporting specific types of learning and decision-making tasks. For example, work could investigate (a) how people assess whether information is true or not, (b) how interfaces can provide scaffolding to guide a search, and (c) how to convey where information is situated in a space or along a particular set of dimensions. Longer-term work should consider (a) broader goals to understand how to represent knowledge, biases, and cognitive processes in users, (b) how documents, rankings and interactions operate as functions that change users' knowledge states and beliefs, and (c) how users could use search systems to formulate hypotheses and understand options.

Finally, this proposed research connects with multiple SWIRL themes, including evaluation, fairness and accountability, and past themes like search as learning (2012). In particular, while the latter focused on developing users' search skills using a variety of tools and interfaces, we look at broader support of knowledge goals and incorporate cognitive aspects, including bias, as part of automatically improving retrieval processes and outcomes.

4.4 Research Challenges

Challenges are faced on each of the areas that the proposed research covers. The proposed research touches on the collection over which the search engine operates, the user's interaction with the search system, the user's cognitive processes, and the evaluation of the changes to the user's knowledge state or performance on tasks.

At the level of the collection, we are concerned with the mix of information that is available. For large scale collections, such as the web, it is very difficult to understand the amount of material on a given topic, and thus is it hard to know what the existing biases are in the collection. For example, we might be interested in measuring the accuracy of decisions that users make after using a search engine. Collections of interest will contain a mix of correct and incorrect information, but the scale of the collection will make it difficult to understand the amount of correct and incorrect information in the collection apriori to the user's search session.

The field of IR is still in its infancy with respect to understanding user interaction and user cognitive processes. For us to be able to design systems that lead users to their desired knowledge state or decision, we will need to better understand how their cognitive processes affect their interaction with the system and how the stream of information that they consume changes their mental state. A challenge here will be a lack of expertise in cognitive science and psychology (how people learn, how people make decisions, biases). Progress in this area will likely require collaboration outside of IR and require input from and engagement of other communities, including: cognitive science, human-computer interaction, psychology, behavioural economics, and application/domain specific communities (e.g., intelligence community, clinical community). The envisioned systems may require radical changes to aspects of user interfaces. Uptake of new UI solutions, however, is often difficult and poses extra onus on users, thus creating a high barrier to entry for the proposed new systems.

Finally, evaluation ranges from the simple to the complex. We are interested both in simple measures such as decision accuracy, and complex measures such as increases in curiosity. Evaluation is envisioned to embrace larger aspects of the user-system interaction than just the information seeking phase, e.g., evaluation of decisions users take given the information systems provided. Given that almost all evaluation will be with respect to changes in the user, evaluation will be as costly in time and effort as all user studies and human research is. Evaluation may also be hindered by difficulties in evaluating aspects such as learning, or the unavailability of a normative reference to evaluate decisions. Indeed, there are many circumstances in which the "right decision" or the "right knowledge" depends on personal circumstances, or cultural/societal frameworks.

4.5 Broader Impact

The proposed research will make users better informed and more aware of information quality and its broader context, by providing a broader, more balanced view of the information space and meta-cognition for the further process of information seeking or decision making. It also connects a user's information seeking behavior to a growing understanding of the cognitive processes that underlie a person's searching, learning, and decision-making. This should lead to users being more confident and efficient in their learning and decision-making, to improvement in the overarching task of connecting people with the right information, to support for complex matching tasks such as expert-finding or peer matching, and to enabling people to learn more and to learn more effectively.

4.6 Broadening SIGIR

There is potential for cross-disciplinary collaboration and impact with a number of scientific fields, including psychology, economics, learning sciences, and robotics. In fact, some IR advances described in this report will require interdisciplinary solutions that draw from paradigms and methods in multiple areas.

4.7 Obstacles and Risks

With existing search systems, we currently know little about the actual extent to which users are helped or hurt in their ability to reach their desired knowledge state or make decisions. As we attempt to measure and improve performance, we risk making systems worse. What if systems lead users to the wrong answers or to bad, possible harmful decisions (e.g., bad health decisions)?

Another risk involves systems that deliberately or accidentally over-represent or promote the values of certain cultural groups/majorities, and discard the values, opinions and conventions of minorities.

Adversarial aspects are a serious risk: in principle, systems using the proposed technology could deliberately introduce biased, incomplete, or fraudulent information. Moreover, people who know the algorithms used in these systems could potentially design their material to work around the safeguards and thus spam the users that the systems are designed to support. To minimize these risks, our evaluation methods will need to be designed to cover both offline and online evaluation that includes adversarial scenarios.

5 Evaluation

We describe three elements of evaluation research: online evaluation, developing methods to predict evaluation results, and the ever present challenge of interactive evaluation.

5.1 Research opportunities arising from online evaluation

5.1.1 Description

For more than a decade, online evaluation has proved itself to be a challenging, but powerful, research methodology. Evaluating a fully functioning system based on implicit measurement of users is a process that has transformed the way that companies manage, trial, and test innovations for their respective systems.

5.1.2 Motivation

While there has been much publication by both companies and academic groups in this area, trends in search interfaces as well as techniques that connect online with offline evaluation mean there are rich new opportunities for researchers to contribute to this critical area of evaluation.

5.1.3 Proposed Research

To illustrate the range of possibilities of this broad agenda, we list the following suggested projects:

(1) **Counterfactual analysis** lies at the junction of online and offline evaluation. It is a tool from causal reasoning that allows the study of what users would do if the retrieval system, they interact with, was changed. Drawing on a system interaction log, one can (offline) "re-play" the log, re-weighting interactions according to their likelihood of being recorded under the changed system. From the re-played interactions, an unbiased estimator of the "value" of the changed system can be calculated. Value metrics are typically based on user interactions (e.g. clicks, dwell time, scrolling, etc) but can incorporate editorial judgments of relevance or other factors. Because the user/information need sample is the same in every experiment, variance due to those factors can be more controlled than in open-ended interaction studies.

Counterfactual analysis relies on a rich log that captures a wide range of interactions. Typically some fraction of users must be shown results that have been perturbed in a systematic way, but may not be optimal for them. The main challenge is balancing the counterfactual need for perturbed results against the need to show users optimal results. There is extensive opportunity for research on means to minimize both the degree of perturbation of system results and the amount of log data required to produce low-variance, unbiased estimates.

(2) **Define the axiometrics of online evaluation metrics.** In the 2012 SWIRL report, determining the axioms of offline metrics was proposed and soon after the meeting two SWIRL colleagues were granted a Google Faculty Award to explore this research idea further. We propose that axioms for online metrics be determined. Already some axioms of such measures have been defined (e.g. directionality, sensitivity) but it is clear that such work is not yet complete.

(3) **New online metrics from new online interactions.** Current online metrics mainly draw on naïve user interactions. There is a growing concern that determining value from such interactions misses important information from users, producing systems that optimize short term benefits rather than long term goals. Additionally, new modes of interaction, such as conversational systems as well as smaller interface forms such as smart watches won't capture clicks or scrolls.

It is necessary to move to more sophisticated interaction logging and understanding. Background ambient noise or richer understanding of context or user session (see 5.3), as well as technologies such as eye tracking, could be used to determine how users are reacting and benefiting from an online system.

5.1.4 Research Challenges

Some may think that online evaluation is off limits to academia because of a need to 'get' live users. However TREC, NTCIR, and CLEF have explored ways of making such a provision. In addition, smaller-scale evaluation in laboratory or live-laboratory settings, or in situ, could lead to advances in evaluation taking account of rich contextual and individual data. We believe that it may also be possible to simulate user bases with recordings of user interaction in conjunction with counterfactual logging. Such collections may include logs, crowd-sourced labels, and user engagement observations. Such data may be collected by means of user-as-a-service components that can provide IR systems with on-demand users who can interact with the system (e.g., given a persona description) to generate logs and the context where online evaluations can be carried on.

5.1.5 Broader Impact

Online evaluation is not just the domain of a few global search brands, it is an industry. For example, the online evaluation/optimization company Optimizely has gone from 0 to 500+ employees in about six years. Such companies enable smaller companies to perform online evaluation and test changes. Work in online evaluation of search will have a substantial impact on search as well as related topics, such as recommender systems.

5.1.6 Broadening SIGIR

Papers on offline evaluation through test collections dominate SIGIR evaluation papers. While such work is important, there are other research challenges to address. Venues like KDD, NIPS, WSDM, and ICML are publishing much work in online evaluation, and SIGIR-focused researchers should stake more of a claim. We have the expertise in large-scale reusable experimental design that will be necessary to harness the full power of these methods for retrieval systems. If we can encourage more IR focused online evaluation research, we hope this will create a bridge between SIGIR and the other more ML focused conferences as well as attracting new SIGIR participants from those communities.

5.1.7 Obstacles and Risks

A common cry from academic evaluation researchers is a request for logs, but many years of asking have provided few widely available datasets of user interactions. We have to be more creative than calling for others to help us. A key risk is that the smaller scale research and results that we conduct will not translate to the large scale problems of the search engines. However the only way to understand such risks will be to try.

5.2 Performance Explanation and Prediction

5.2.1 Description

Despite the wide success of IR systems, their design and development is a complex process, mostly driven by an iterative trial-and-error approach. It is impossible to make strong predictions on how a system will work until it is tested in an operational mode. This is because IR lacks any comprehensive model able to describe, explain, and predict the performance of an IR system in a given operational context.

5.2.2 Motivation

There are new IR applications launched every day (e.g. online shops, enterprise search, domain-specific information services), which often require substantial investments. IR systems are complex: made up of pipelines of heterogeneous components. They are used together with other technologies, for example, natural language processing, recommender systems, dialogue systems, etc., and they serve complex user tasks in a wide range of contexts. However, each new instantiation of these applications can only be evaluated retrospectively.

There is a growing need to predict the expected performance of a method for an application before it is implemented and to have more sophisticated design techniques that allow us to ensure that IR systems meet expected performance in given operational conditions. We cannot postpone any further the development of techniques for explaining and predicting performance, if we wish to be able to improve and make more effective the way in which we design IR systems in order to keep pace with the challenges the systems have to address.

5.2.3 Proposed Research

We need a more insightful and **richer explanation of IR system performance**, which not only allows us to account for why we observe given performance: e.g. **failure analysis**. We also need to decompose a performance score into the different components of an IR system, how the components interact, and how factors external to the system also impact overall performance.

Richer explanations will provide the basis for strengthening the investigation of the **external validity** of experimental findings, i.e. how much can findings be generalized? This, in turn, this will foster accurate **performance prediction** of IR systems.

With such research in place, stronger links with **interactive IR** will be possible: testing different types and degrees of comparability for their suitability for evaluation of interactive IR. This will also involve constructing and testing simulations of user models, to see if they can be used for traditional comparative evaluation – calling for much more empirical work on characteristics of users, their tasks, their contexts and situations.

5.2.4 Research Challenges

There have been past initial attempts to build **explanatory models of performance** based on linear models validated through ANOVA but they are still far from satisfactory. Past approaches typically relied on the generation of all the possible combinations of components under examination, leading to an explosion in the number of cases to consider. Therefore, we need to develop greedy approaches to avoid such a combinatorial explosion. Moreover, the **assumptions** underlying IR models and methods, datasets, tasks, and metrics should be identified and explicitly formulated, in order to determine how much we are departing from them in a specific application and leverage this knowledge to more precisely explain observed performance.

We need a better **understanding of evaluation metrics** Not all the metrics may be equally good in detecting the effect of different components and we need to be able to predict which metric fits components and interaction better. Sets of more specialized metrics representing different user standpoints should be employed and the relationships between system-oriented and user-/task-oriented evaluation measures (e.g. satisfaction, usefulness) should be determined.

A related research challenge is how to exploit richer explanations of performance to design better and **more re-usable experimental collections** where the influence and bias of undesired and confounding factors is kept under control. Most importantly, we need to determine the **features** of datasets, systems, contexts, and tasks that affect the performance of a system. These features together with the developed explanatory performance models can be eventually exploited to train **predictive models** able to anticipate the performance of IR systems in new and different operational conditions.

5.2.5 Broader Impact

A better understanding and a more insightful explanation of IR system performance opens up new possibilities in terms of reproducibility, external validity, and generalizability of experimental results since it provides the means to understand what succeeded or failed, especially if linked to failure analysis. Better analytic tools are also an indispensable basis for moving IR toward becoming a predictive science.

5.2.6 Broadening SIGIR

There are neighbourhood areas, such as Natural Language Processing and Recommender Systems, which suffer from similar issues in terms of explanation and prediction of the performance of their systems. These areas could benefit from an advancement within SIGIR and, at the same time, SIGIR could benefit from teaming up with these areas to jointly address these issues and come to more general and robust solutions.

5.2.7 Obstacles and Risks

While some of the proposed research activities (metrics, performance analysis, assumptions) can already be carried out with existing resources, the identification of performance-critical application features and the development of performance models require empirical data from a larger variety of test collections. Thus, researchers should share their test collections both for supporting reproducibility and research on prediction. Indeed, while individual contributions to such an effort might not seem worthwhile for researchers, collaborative approaches in the form of evaluation campaigns might be more promising. Another potential obstacle is the need for more sophisticated competencies in data modelling, statistics, and data analysis, and so on. Moreover, both the explanatory and the predictive performance models may be quite demanding in terms of computational resource needed to train and compute them.

5.3 Measures and Methods for Evaluating Interactive IR

5.3.1 Description

All IR is, to some degree, inherently interactive with the interaction taking place among a person seeking information for some goal / task / purpose, some corpus of information objects (including the objects themselves), and some intermediary (e.g. an IR system) acting to support the person's interaction with the information object(s). Methods for evaluating system support for persons engaged in interaction must be developed in order for IR systems to continue to improve. Such methods may be similar to those of the test collection model, but, given experience to date, it is clearly necessary to consider quite different alternatives.

5.3.2 Motivation

The classical IR evaluation model was designed to evaluate the performance of the IR system with respect to just one interaction instance: the response that the system provides to one query put to that system. The model has been extended in various ways, to differential effect. Test collections have used a surprisingly wide range of labeling criteria: topical relevance, home-page-for, key page, spam, opinionated, a-venue-I-would-go-to, novelty, and others. Cranfield assumes an atomic preference criterion: that is, an individual document's preference label is defined with respect to the document and topic only. Atomicity allows us to build test collections scalably because documents can be labeled in a single pass.

Other kinds of criteria for building test collections should be explored. For other atomic qualities we need to understand how to define them, how to develop labeling guidelines that are understandable enough for separate sites to label items comparably, how to measure the consistency and reliability of those labels, and how to measure the impact of label disagreements. As research problems these questions deserve more attention.

Although there have been serious attempts to design methods to evaluate system support for **information search sessions**, these have uniformly failed. There are various reasons for this failure. The atomic criterion of relevance, basic to the model, does not easily apply to the evaluation of the success of a whole session, and the presence of human beings, having varied intentions during the information search session, making individual decisions during the search session, and having varied individual characteristics, has made comparability of performance of different systems with different persons, as required by the classic model, seemingly impossible.

Extending the Cranfield model into full interactions is hard because it violates the atomicity criterion. To consider an interaction where a user starts from different queries, encounters documents differently, and moves towards completion of the task along multiple paths, a test collection would need, at a minimum, to define the relevance of each document with respect to all documents already seen. Without constraining this within some sort of structure, there would be an exponential number of relevance judgments needed. Taking a further step and allowing the user's understanding of the task to evolve and criteria for successful completion of that task to change during the interaction adds another exponent.

5.3.3 Proposed Research

- Identifying criteria and metrics that can/should be used to evaluate:
  - Support by the system toward accomplishing that which led the person to engage in information seeking, i.e. evaluation of success of the search session as a whole.
  - Support by the system with respect to what the person is trying to accomplish at each stage in the information searching process (search intentions).
  - Contribution of the activity of each stage of the information searching process to the ultimate success of the search session as a whole.
- Creating metrics that are sensitive to different **types** of motivating goals/tasks, and to different **types** of search intentions – we need to learn about the types, and desired outcomes for the types.
- Investigating how to apply those criteria and measures through user studies and test collections that are aligned, so that researchers can benefit from both.

There is also ample opportunity to incorporate these more detailed investigations of users into online evaluation.

5.3.4 Broader Impact

This research presents an incredible opportunity to broaden the community, because it will open a wide range of research questions, which have been largely ignored, yet are of central concern to the evaluation of, for instance, support for complete search sessions, or of personalization of search.

5.3.5 Broadening SIGIR

Accomplishing the research program will require collaboration among researchers from different disciplinary, theoretical and methodological traditions, e.g. computer scientists, information scientists, human-computer interaction researchers, cognitive and experimental psychologists. The SIGIR community needs to ensure that its core venues support the growth of research bridging interactive IR and test collection-based experimentation. There is a great deal of foundational work on methodologies, and that work is best conducted where research ideas are taken note of, in the conferences of record for the community.

5.3.6 Obstacles and Risks

The problem of evaluation of interactive search support is extremely difficult to solve, if comparison and generalization of results is to take place. There does not currently exist a sound, generally accepted theoretical understanding or model of interactive IR, on the basis of which the evaluation criteria, measures and methods can be derived.

6 Learnable Information Retrieval

6.1 Description

The availability of massive data and powerful computing has the potential to significantly advance almost every aspect of information retrieval. While these methods have been very successful in some domains – such as vision, speech, audio, and NLP – these successes have not been observed for information retrieval tasks. This research area analyzes some of the reasons and proposes to investigate artificial intelligence approaches to representation learning and reasoning for i) core retrieval problems, ii) robust, cross-domain ranking, and iii) novel or intractable retrieval scenarios; and to deal with limited training data by i) a community effort to build labeled data sets that are an order or magnitude larger than existing ones, ii) improving low sample complexity models, and iii) automatically generating training data from scavenged public data. Work in this research area will not only lead to more effective retrieval systems, but also provide new insights into the fundamental problems underlying search and relevance matching. While deep learning has led to some level of concern and even suspicion in the academic community, who have seen previous instances of “hype”, the impact of neural net approaches in many fields such as vision and NLP is undeniable and is well-documented in many peer-reviewed articles. In summary, the neural revolution in IR empowers the end-to-end learning of an entire search engine from data.

6.2 Motivation

The information retrieval community has a proud history of developing algorithms for efficient and effective information access. However, these systems are shallow in their representation and comprehension of text and other media, resulting in a disconnect between where search is now and where it could be. This shallow understanding limits our ability to perform more complex IR tasks such as conversational search, summarization, and multimodal interaction, as well as search tasks that require deeper understanding of documents contents and the user information need.

While users have mostly been content with shallow search in the past, they are expecting the next generation of IR systems to be more intelligent. This expectation is amplified by recent developments in artificial intelligence. The traditional models with manually designed representations, features and matching functions are likely unable to cope with this demand.

At the same time, the traditional models may also be less able to be adapted to new domains, as our existing approaches would require a nontrivial amount of feature engineering and retraining for the new domains. Similarly, modern-day deep learning methods are highly versatile and adaptable, and can be used to combine multimodal data inputs and heterogeneous data views, and can be trained jointly over multiple tasks simultaneously (possibly with partial labelling).

Recent advances in artificial intelligence have resulted in performance improvements in several areas such as computer vision, speech recognition and NLP. The new approaches based on machine learning, and more particularly, deep learning, offer new opportunities to IR to design and learn new models. However, IR tasks have their specificities. A naive utilization of deep learning approaches developed for other areas may not be a good fit for IR problems. In addition, existing deep learning approaches often require a massive amount of training data to generalize suitably, which is hard to obtain in IR area, suggesting that we should investigate methods for developing models with limited training data. Intensive investigations in this area are thus required.

6.3 Proposed Research

The proposed research can be divided into six areas: data efficiency, core ranking, representation learning, reinforcement learning, reasoning, and interpretability. We anticipate these advances complementing, rather than replacing, current approaches to information retrieval.

Data Efficiency. Limited data access has limited the ability for investigators to study deep learning approaches to information retrieval. Unfortunately, although this data exists in industry, distributing it to the academic community would incur substantial risks to intellectual property and user privacy. As a result, the community needs to conduct research into:

- training robust, accurate models using small collections,
- developing new techniques to expand current labeled datasets (such attempts have been implemented, e.g., with weak supervision),
- dealing with incomplete and noisy data,
- simulating user behavior (e.g., using RL),
- developing robust global models effective for data-poor domains, and
- reusing trained models for new tasks (e.g., for domain adaptation). Current approaches includes progressive NN and transfer learning.

Advanced retrieval and ranking models. One of the core information retrieval problems involves the representation of documents and queries and comparing these representations to produce a ranking based on estimated relevance. Neural information retrieval models have the potential of improving all aspects of this task by offering new methods of representing text at different levels of granularity (sentences, passages, documents), new methods of representing information needs and queries, and new architectures to support the inference process involved in comparing queries and text to find answers that depend on more than topical relevance. For example, hybrid models combining different structures such as CNNs and LSTMs can capture different linguistic and topical structures, attention mechanisms can capture relative term importance, and GANs may be able to lead to ranking models that require less training for a new corpus. It is not yet known which architectures are the most effective for a range of information retrieval tasks, but their potential is driving an increasing amount of research. As new models are developed, it will be critical that they are accompanied by in-depth analysis of how different aspects of the models lead to success or failure. Models that work with existing feature-based approaches, such as learning to rank, will have a critical role in producing systems that impact current search applications.

Data Representation. Current deep learning techniques for data representation are not directly suitable for IR models, as we deal with multimodal input, e.g. text documents, user features, music, images, videos. Therefore, we need to work on new ways of data representation specific to IR problems. There may be other, external information of value available, that needs to be combined with these dense representations in more effective ways than a hard filter. Another aspect of this is semantic emergence: semantic properties emerging during training for a particular task that are not directly related to the task and were not explicitly planned to emerge. An example for this is the emergence of a "sentiment neuron" when learning a simple language model on a large set of reviews. In image classification with deep neural networks, edge detection emerges on a certain level of the network. It will be interesting to find out which other semantic concepts can emerge in this way when training for basic (or not so basic) information retrieval tasks.

Reinforcement Learning. Because information access is often situated in an interactive search task, the ability to perform intelligent sequential decision-making is a fundamental — yet under-explored — area of information retrieval. Recent advances in reinforcement learning suggest that techniques are ready to be applied to complex domains like search. That said, applying these techniques to information retrieval requires substantial research into:

- acting in extremely large, non-stationary state and action spaces, and
- developing effective unsupervised objective functions for multi-turn retrieval, and
- modeling interactions through RL has a high potentials for user simulation task.

End to end learning. Certain complex information retrieval problems might be learnable in a completely end-to-end fashion. For example, input the query, output the set of relevant documents. Or input a question in natural language, and output an answer. There is already a fair amount of work in that direction, for example: given a question in natural language and a text, output the passage or passages of the text that answer the question.

Machine Reasoning. There has recently been significant progress on machine reasoning in the context of tasks such as text understanding and reasoning, e.g. bAbi, and dialogue state tracking, focusing on "memory" architectures for selectively capturing dialogue/document context as needed for long-distance inference. There are also many attempts to integrate domain knowledge or knowledge graphs in NLP (e.g. QA). There are direct applications for this style of model (and unique application areas) in information retrieval, including:

- tracking “state” in multi-turn information retrieval (e.g. conversational, session-based),
- smoothing document- and term-level predictions across a session/collection,
- interpreting complex search requests,
- supporting question answering, and
- implementing domain-specific information retrieval.

Error Analysis/Explainability. It is important to advance the current error analysis techniques and to make our model explainable for:

- finding the errors in training sets that cause problems on output level,
- understanding for what cases old models work better and when NN models show better results (e.g. NN work better for long queries), which may lead hybrid neural architecture, and
- making results explainable for users/system designers.

6.4 Research Challenges

Existing high baselines: Over the long history of IR, we have developed models and approaches for ad-hoc and other types of search. These models are based on human understanding of the search tasks, the languages and the ways that users formulate queries. The models have been fine-tuned using test collections. The area has a set of models that work fairly well across different types of collections, search tasks and queries. Compared to other areas such as image understanding, information retrieval has very high baselines. A key challenge in developing new models is to be able to produce competitive or superior performance with respect to the baselines. In the learning setting, a great challenge is to use machine learning methods to automatically capture important features in representations, which have been manually engineered in traditional models. While great potential has been demonstrated in other areas such as computer vision, the advantage of automatically learned representations for information retrieval has yet to be confirmed in practice. The current representation learning methods offer a great opportunity for information retrieval systems to create representations for documents, queries, users, etc. in an end-to-end manner. The resulting representations are built to fit a specific task. Potentially, they could be more adapted to the search task than a manually designed representation. However, the training of such representation will require a large amount of training data.

Low data resources: representation learning, and supervised machine learning in general, is based heavily on labeled training data. This poses an important challenge for using this family of techniques for IR: How can we obtain a sufficient amount of training data to train an information retrieval model? Large amounts of training data usually exist only in large search engine companies, and the obstacle to making the data available to the whole research community seems difficult to overcome, at least in the short term. A grand challenge for the community is to find ways to create proxy data that can be used for representation learning for IR. Examples include the use of anchor texts, and weak supervision by a traditional model.

Data-hungry learning methods have inherent limitations in many practical application areas such as IR. A related challenge is to design learning methods that require less training data. This goal has much in common with that of the machine learning area. The information retrieval community could target learning methods specifically designed for information retrieval tasks that require less labeled data.

6.5 Broader Impact

The development of machine learning and AI "algorithms" for new products and services, and the success of deep learning methods in high visibility competitions, have received enormous attention in the media. New courses on machine learning and neural models are proliferating and many new papers related to these topics are appearing every day on arXiv. Given that search is nearly ubiquitous and that search engines are the best example of IR and NLP in action, it is not surprising that there has been a significant upsurge in interest in applying neural models to many aspects of IR and search, both by graduate students and by companies working on a broad range of applications. Although this research is still at an early stage and much remains to be done to demonstrate the levels of improvement in effectiveness seen in other fields, the potential impact of new retrieval approaches developed using learning techniques is enormous. If we are successful, for example, we will make significant progress towards achieving one of the long-standing goals of IR - an "intelligent" conversational search assistant similar to those seen in the early Star Trek episodes that inspired so many future computer scientists. On a more mundane level, we should also be able to develop search techniques that are substantially more portable and effective across tasks, domains, applications, and languages. Search will be substantially more effective and available to all segments of society.

6.6 Broadening SIGIR

This area draws heavily in the first instance on work done in machine/deep learning and statistical natural language processing, and as such, any activity in this space will naturally lead to stronger connections with these fields through cross-fertilization of ideas and greater visibility for SIGIR research. Beyond this, there are unique characteristics/challenges in IR that we can expect to give rise to methodological breakthroughs with broader implications including:

- IR has a very mature understanding of what types of document/collection representation are needed for retrieval (e.g. inverted file indexing, positional indexing, document zoning, document graphs), more so than fields such as speech, NLP and computer vision, and developing representation learning methods that are able to capture these rich data structures will have implications well beyond IR.

- IR has decades of experience in assigning, interpreting and learning from document-level relevance judgments; there is considerable scope to transfer this expertise beyond the bounds of IR.

- IR has a rich history of multimodality (including images, speech, video, and (semi-)structured data) with well-established datasets, and a relatively mature understanding of how to harness that multimodality to draw inspiration from when developing new models.

- There is deep knowledge of methods for attaining run-time and storage efficiency in IR, both of which are critical issues in machine/deep learning research at present, and any advances on the part of the IR community would have far-reaching implications beyond SIGIR.
- IR expertise in evaluation, especially focusing on the user experience, has the potential to significantly shape research on tasks such as question-answering, summarization and machine reading, where current evaluation practices are narrowly focused on string matching with a gold standard.

6.7 Obstacles and Risks

Data requirements: the strong results that have been achieved by (deep) learning approaches in recent years are usually predicated on large amounts of training data. Training data is notoriously hard to get by in academia, especially in large quantities and of a quality that reflects real use cases. It is one of our challenges to figure out how to get by with less training data or to automatically generate training data in an unsupervised fashion. However, one outcome of these attempts may be that large amounts of explicit training data are indispensable.

Resource requirements: training effective models usually requires large amounts of computing resources (time and/or number of CPUs/GPUs/TPUs), even for only moderately large datasets. Getting access to these resources may be an obstacle for groups in academia, especially smaller groups or groups endowed with less money. Training these models on very large datasets may still be outside of our reach for a number of years.

Efficiency: the effectiveness of learned models often comes at the price of an increased processing time when using them for prediction or classification or whatever it is that they were trained to do. If these processing times are several orders of magnitude slower than those of state-of-the-art approaches, these approaches may be of little use in production systems.

Explainability: learned models may achieve an effectiveness that is superior to classical approaches, but it may be hard or impossible to explain why a particular results was computed. It may also be hard to provide guarantees or explain what went wrong in case something went wrong. These drawbacks may be major obstacles to adopting these techniques in production systems.

7 Generated Information Objects

7.1 Description

In modern devices, search results may be presented on a small screen or as a spoken response. This increases the importance of generating a search result in a form that is most helpful to the user. In such situations, and in general, it may be less than ideal for the retrieval system to present raw information as-is. Better would be a summarization of existing information to support absorbing complex material efficiently. Current data processing pipelines include an increasing amount of annotation, filtering, and aggregation steps. These will lead to new interaction paradigms beyond search engine result pages (SERPs). A conceptual framework to discuss such future research is centered around Generated Information Objects (GIOs).

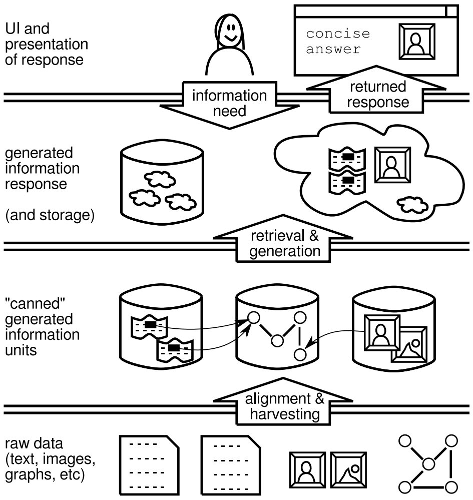
UI and presentation of response
concise answer
information need
returned response
generated information response
(and storage)
retrieval & generation
"canned" generated information units
alignment & harvesting
raw data (text, images, graphs, etc)

Figure 1: Example of a system using the GIO framework.

Given an information need the goal is to generate a single response that consists of multiple information units. These units can be derived from raw information but undergo a transformation/generation process. For instance, preprocessing, semantic annotation, segmentation, clean-up, and recombination are possible generation operations, but raw information can also be provided as-is (as a trivial generation operation). It shall also be possible to generate information objects through logical reasoning or natural language generation. Information objects can be "canned", i.e., preprocessed and indexed in an appropriate representation, or generated "on-the-fly" at query time. The generated response can further be archived, shared, made available for browsing and be recycled for the next response.

Both information units and the response are considered generated information objects (GIOs). Provenance information about the original information source and generation operations shall be preserved in the process. In the context of a response, comprising information units are further associated with an explanation of relevance (or usefulness) for a response. We expect users to question *why* information is important and have a subsequent conversation about the provided information.

The central research question: What is the best way to store generated information units, and how to make best use of information units and previous responses when generating a response?

The question on data structure and representation of generated information objects is related to the question of appropriate forms of presentation of the information. The same generated response object might be represented in different modalities depending on the situation, user's preferences, and situational context – as text, image, or voice.

7.2 Motivation

So far retrieval means to preserve information as found, but not to modify or change it. Novel approaches towards semantic annotations, recombination of information, and summaries of information are already breaking with this old paradigm, as the following examples illustrate.

Generating abstractive summaries from text retrieval was studied at the NTCIR One Click track. Combination of different heterogeneous source archives is discussed in federated search. The recombination of information into "information bundles" or "entity cards" is another form of generation. Diversification of search results is a form of generating rankings. When utilizing entity links and semantic annotations in retrieval, original information undergoes extraction and processing steps such as tokenization, entity linking, and entity type prediction. Such derived information needs to be stored in the right representation to maximize usefulness for downstream tasks. Often information changes over time, demanding approaches that ensure the consistency and freshness of information.

Considering how transformations of raw source information leads to generated information objects (GIO), gives us the option to discuss best practices for representation, storage, and access. Importantly, this discussion can be independent of the chosen presentation form. We suggest a conceptual framework of generated information objects of which all aforementioned examples are special cases. We believe the GIO framework helps to identify underlying patterns and share best practices regarding representation, storing of derived information and retrieval, recombining and recycling partial answers.

7.3 Proposed Research

The framework is intended to support a critical discussion about the state-of-the-art regarding:

(1) Harvesting: How to segment input data into units to that they are likely to be reused?

(2) Representation: What is the most effective representation of information units to be maximally amenable and effective for the downstream task?

(3) Dynamic changes: How to represent information units whose content changes over time? Information may change and canned GIOs may become out-of-date.

(4) Connections: How to preserve connections between information units that can be retrieved (not only "boring" connections)? How to make connections accessible through retrieval models?

(5) Storage: How to store the content while preserving provenance information across different generation, recombination, and modification operators?

(6) Efficient access: How to index information units so that candidate sets for complex queries (i.e., many query terms) for efficient retrieval of diverse canned GIO units?

(7) Condensing: How to represent information units so that redundant, entailing, similar, and related content can be efficiently identified? How to conflate different information units?

(8) Text generation: How to generate text for that is most appropriate for different devices and modes of interaction?

(9) Ease of use: How to optimize information structure and flow of GIOs? For example structure chronological or topical order of facts, salience, lack of redundancy, and readability.

(10) Recycling: How to recycle previous responses, especially in a turn-based dialog?

(11) Privacy: When reusing and recycling GIOs that are created in response to user interactions, this may leak private information into the system. How are user's privacy concerns appropriately addressed?

7.4 Research Challenges

Knowledge Graph Representation in GIOs. The goal is to represent open domain information for any information need. Current knowledge graph schemas impose limitations on the kinds of information that can be preserved. Schuhmacher et al. found that many KG schemas are inappropriate for open information needs. OpenIE does not limit the schema, but only low-level information (sub-sentence) is extracted. In contrast, semi-structured knowledge graphs such as DBpedia offer a large amount of untyped relation information which is currently not utilizable. A challenging question is how to best construct and represent knowledge graphs so that they are maximally useful for open domain information retrieval tasks. This requires new approaches for representation of knowledge graphs, acquisition of knowledge graphs from raw sources, and alignment of knowledge graph elements and text. This new representation requires new approaches for indexing and retrieval of relevant knowledge graph elements.

Adversarial GIOs. Not all GIOs are derived from trustworthy information. Some information ecosystem actors are trying to manipulate the economics or attention within the ecosystem. It is impossible to identify "fake" information in objects without good provenance. To gain the user's trust, it is important to avoid bias in the representation which can come from bias in the underlying resources or in the generation algorithm itself. To accommodate the former, the GIO framework enables provenance tracing to raw sources. Additionally, contradictions of information units with respect to a larger knowledge base of accepted facts need to be identified. Such a knowledge base needs to be organized according to a range of political and religious beliefs, which may otherwise lead to contradictions. The research question is how to organize such a knowledge base, and how to align it with harvested information units. Finally approaches for reasoning within a knowledge base of contradicting beliefs need to be developed. Equally important is to quantify bias originating from machine learning algorithms which may amplify existing bias.

Merging of Heterogeneous GIOs. To present pleasant responses, it is important to detect redundancy, merging units of information, such as sentences, images, paragraphs, knowledge graph items. For example, this includes detecting when two sentences are stating the same message (i.e., entailment). For example "the prime minister visited Paris" from a document about Margaret Thatcher, and an identical sentence "the prime minister visited Paris" from a document about Tony Blair, should not be conflated. Even more challenging is the detection of information units that are vaguely related (according to a relation that is relevant for the information need). The availability of such approaches would allow for structuring and organizing content. Provenance references of information units and textual context can potentially help the integration of information units. Record merging in data bases is achieved by counting agreements versus disagreements. The research challenge is to perform such a merge in a multi-modal and open-domain setting.

Resource Location Across Turn-based Conversational Information Seeking. In multi-turn interactions, people engage with a GIO as a response. A user asks a question and is presented with a response that is generated of multiple parts. It is likely that this user would like to interact with one part of the response. For example, the user may ask a follow-up question about one part or may reconsider a part of an earlier response at a later time. The research challenge is to provide a representation of multi-part answers and an appropriate presentation so that the user can refer to a part. This is even more challenging for voice interactions for which new kinds of anaphora resolution need to be developed. Resource location is difficult even in click-and-point presentations, especially when the answer arises from summarization of different information units. We suspect that this requires aligning information units into groups that the user intuitively interprets as one concept.

This research is related to previous approaches of identifying information "nuggets" as an input to summarization. The challenge is to identify nuggets without human involvement, by training algorithms that identified re-usable informative components that make for useful GIO information units in different context.

Deriving Explanations from GIOs. The rationale of a generated response (with respect to the information need) needs to be explainable to the user. In search snipped generation, such explanations are typically identified through high-density regions of keyword matches. For complex generated information responses this will not be sufficient. We envision that such an explanation entails two parts: 1) Explain how the given information need was interpreted, and 2) explaining every part of the generated response.

For example, in a system that jointly reasons about relevant entities and relevant text, the system may be asked why a particular entity is relevant. While the relevant text related to this entity may be one first approach, one can imagine that a user would rather hear a direct explanation such as "This entity is relevant because ... ". In order to give such explanations, a system must be able to understand the concepts contained in the text and their relation to the information need. The research challenge is how to understand both the information need and response text on a conceptual level that would allow such explanations.

Context and Personalization. Any information-seeking behavior is dependent on context: the user's prior knowledge, the task to be accomplished, and the previous interaction. The GIO framework allows for modeling, storing, and considering any user context that is available. While studied for many decades, it is still open which representation of the user's context is optimal for which task, and how to derive representations that are versatile across different retrieval tasks at once. Even more important is to model shifts in user's context and interest over time. This affects both the selection of information unit, the appropriate generation of the response object, and the effectiveness of the chosen presentation.

Evaluation. It is difficult to create reusable test collections when answers are summarizations of different parts. The entirety of a GIOs presentation may need to be evaluated in a holistic end-to-end information seeking context. However, early experiments arising from the TREC Complex Answer Retrieval evaluation demonstrate that it is practical and feasible to evaluate individual parts of the GIO response: Manual assessment of passages in the Cranfield paradigm correlated extremely well with alternative ground truths derived from gold articles.

Furthermore, we hypothesize that GIO representations that allow systems to generate explanations for the relevance/suitability of the GIO response or its parts, assist in creating assessor-independent datasets. Asking assessors to only consider parts as relevant if the submitted explanation would support this claim, and not use external world knowledge, may lead to a high inter-annotator agreement.

One key research challenge in generated information objects is the transformation of raw data through cleanup, combination, and modification to the point where many challenges of summarization also apply to GIOs. Evaluation of summaries (and thereby processes that generated them) are typically be based on the readability, cohesiveness, informativeness of content (intrinsic) or based on its helpfulness in the context of completing a task, such as reading comprehension, learning (extrinsic). At the intrinsic level, automated approaches have included measures such as BLEU and ROUGE, where a new generated summary is compared against a human-generated set of reference objects based on n-gram overlap. However, one may imagine an equally good or better summary that uses different words, which would lead to a bad score under the ROUGE measure, or equally a summary that is nonsensical from a human perspective but happens to score highly. While some studies have been shown that ROUGE can be correlated with human judgments of summary quality, others use cases identified limitations in terms of reflecting meaningful differences in algorithms.

These downsides of summarization evaluation are a major obstacle towards summarization-based approaches in the IR community. The same issue holds for many other forms of generation, such as redundancy-removal, grouping of ranked results, or presentation of heterogeneous information. Therefore, more research on alternatives for generated results are important. For example the NTCIR One-Click track evaluated summaries by the number of unique and positive facts included. The TREC Complex Answer Retrieval track, evaluated the relevance of parts (i.e., entities) by the system provided explanation of their relevance. The approach taken by One-Click and CAR is to evaluate a part of the response, which does not fit into the intrinsic/extrinsic scheme used for summarization. More research in this direction of IR-based summarization will not just provide an avenue for evaluating GIOs, but also have an impact on the NLP/Summarization community.

7.5 Broader Impact

Several sub-communities in information retrieval have covered various aspects of the GIO framework. We think that this conceptual framework can facilitate a spread of best practices across the field. New indexing approaches and search models will likely arise from applications that use generated information objects to the fullest.

As the GIO framework requires us to rethink many moving parts of typical IR systems, it bears the potential for more collaboration across typical IR subfields. Additionally, by opening up the definition of IR to explicitly include summarization, natural language generation, and computer vision, it enhances the ties between the IR community and communities on information extraction, knowledge graph construction, text and vision, spoken dialogs, and databases.

7.6 Broadening SIGIR

An explicit goal of research with generated information objects (GIOs) is to broaden the scope of the SIGIR discipline. To address open domain information needs, GIOs will be combined from rich and varied generations of new information objects. In the long term, GIOs are a stepping stone towards the synthesis of new information from existing sources.

This research effort will fuel a wide range of obvious cross-field collaborations: 1) NLP to understand linguistic representations, summarization, and discourse and dialog, 2) HCI to create user-information interactions that are natural and help the user accomplish her task effectively, and 3) vision and language to enable efficient presentation and interaction when multi-modal information is used as information units. We believe that creating challenge tasks with GIOs in mind will clear up some risks and concerns regarding feasibility and evaluation, and spur increased collaboration within and across the community.

7.7 Obstacles and Risks

The conceptual framework is quite general, and while many tasks can be addressed with this framework, this generality might make it difficult to pin down all possible challenges and see which practices can be shared.

Some of the engineering structures required to support GIOs are much less developed and investigated than conventional IR system components. Particular challenges lie in an algorithmic understanding of information needs and response text. This requires a representation and interaction mechanism that allows referring to generated response parts, giving relevance explanations for generated information units, and reasoning about conflicts and trustworthiness of harvested information units.

Industrial applications of GIOs for task-specific purposes are likely to push the development of this area quite quickly ahead of the research community. We run the risk of falling behind rather than leading this effort.

8 Next-Gen Efficiency Challenges: Smaller Faster Better

8.1 Description

Ensuring both effectiveness and efficiency is paramount for the practical deployment of web (and other) search engines. The last few years have seen the IR community tackle more complex search tasks and information needs coupled with a marked increase in the size of collections. New ranking paradigms based on machine learning (for example, learning to rank and neural networks) and new applications (for example, conversational information seeking or searching on the cloud) are pushing the boundaries of existing IR architectures, raising new challenges, but at the same time offering opportunities for the IR community to investigate new alternative architectures, revisit existing index structure assumptions, optimize trade-offs between effectiveness, efficiency and costs, and investigate new efficiency problems and paradigms.

8.2 Motivation

While there is a history of strong research on efficiency issues in the IR community, much attention has focused on a few established and often narrow problems and setups motivated by standard IR architectures and systems, leaving many other cases largely unexplored. There have also been a number of emerging changes in modern search systems that call for new directions and approaches. In particular, efficiency researchers need to look at issues related to Multi-Stage Search Systems (MSSs), which are increasingly being deployed with machine-learned models, and examine the end-to-end performance of methods under MSSs. There are also opportunities for further efficiency improvements that require the application of machine learning and data mining techniques, including methods that learn index structures, learn how to optimize queries, or that estimate query distributions and optimize for these. There might also be ways to completely bypass the currently used index structures via neural nets, structures from the Combinatorial Pattern Matching community, or FPGA-based systems. Finally, new emerging IR applications also require attention. In summary, it is time to reconsider some basic assumptions, and to step away (at least partially) from the comfortable world of inverted lists and simple ranking functions that so many (often strong) papers have addressed.

8.3 Proposed Research

We organize the proposed work into four major streams of research:

(1) **Efficiency and MSSs**: Search engines have been using highly complex rankers for quite some time, but the efficiency community has been slow to adapt. The last few years have seen some initial attempts to address this, but there are many remaining opportunities. Future work should evaluate new and existing end-to-end performance optimizations in the context of MSSs. We need automatic ways to create optimized index structures and to optimize query execution for a given MSS. We need new measures and objectives to optimize for in the early stages of cascading systems, and efficient ways to index and extract features for use in the cascades. Finally, we need to look at the impact of search results diversification and personalization in such systems.

(2) **ML for efficiency**: Researchers are increasingly using machine learning and data mining to improve algorithms and systems. Examples include learning of index structures for particular datasets and ranking functions, modeling of query distributions to build small index structures for likely queries, learning of query routing and scheduling policies to satisfy service level agreements (SLAs) on latency and quality while maximizing throughput, or prediction of query costs and the best execution strategy for particular queries. One major challenge is how to formalize and guarantee performance bounds on such machine-learned components, which will enable reasoning about guarantees for the overall system. In short, ML and data mining techniques are popping up everywhere in the search engine architecture, and will drive future performance improvements, sometimes in unexpected ways. Conversely, IR efficiency researchers should also use their skills to make machine learning tools more efficient, as training and evaluation currently require huge amounts of resources, e.g., deep neural nets.

(3) **Challenging the current setup**: The ready availability of alternative architectures such as vector processors (SSE, AVX instructions, and the like) and FPGAs provide opportunities to examine IR efficiency from a new angle — research on these devices as well as GPUs is in its infancy. The introduction of General Purpose Graphics Processing Units (GPGPUs) and Tensor Processing Units (TPUs) into general purpose CPUs will provide entirely unexplored avenues for research. These hardware architectures will soon be available on all users' clients (from phone to tablet to laptop to desktop), providing opportunities to off-load work from the data center and onto the user's device. Beyond web search, there remain several unsolved fundamental problems. In an environment with high document turnover, new index structures, beyond inverted files, may be called for. What is clear is that there is demand for such structures, and updating an inverted file with additions and deletions is not necessarily an efficient process. In summary, the time has come to reconsider the standard index setup. Indeed, the index of the future might in extreme cases be just a set of weights in a neural network.

(4) **New applications**: In a truly interactive IR system, such as a conversational information seeking system, information needs are complex, typically requiring iterative user dialogue with the system, with each iteration encompassing query reformulation and access to the index. The costs of these iterations might be reduced in several dimensions. For example, we could leverage incremental computations across the sequence of iterations to enhance scalability and efficiency through suitable caching or prediction of the remaining dialogue the user will engage in. The development of real-time search applications on the Internet of Things (IoT) infrastructure also requires new indexing and search architectures, to allow the seamless ingestion, indexing and querying of data in real-time. Related to this, the emergence of new search services in the cloud, with service level guarantees coupled with limited resources and various constraints, opens up a number of unexplored directions in search efficiency and effectiveness trade-offs. For example, the search engine has to continuously analyze and decide on the best configuration of its system given the available resources and the guaranteed level of services. Moreover, the increasing importance of ensuring accountability, transparency, and explainability in machine learned MSSs entails additional costs for the search engine, beyond the actual task of retrieving information. Such costs include the generation of the explanations for the retrieved results and their visualization, which require new efficiency research directions including the revisiting of data structures to cope with such additional costs.

8.4 Research Challenges

Several interesting research challenges continue to exist when building traditional efficient and effective IR systems (such as compression, first stage query resolution, and so on). In multi-stage retrieval systems the complexity is substantially higher and new areas need addressing. For example, at present we do not even know where and why these systems are slow.

As mentioned above, exciting new challenges exist in the areas of conversational IR and learned data structures. While the notion of combining learning with efficient indexing is not an entirely new idea, recent advances in neural IR models have shown that learned data structures can in fact be faster, smaller, and as effective as their exact solution counterparts. However, enforcing performance guarantees in learned data structures is still a research problem requiring work. Likewise, as search becomes even more interactive, new opportunities for efficient indexing and ranking are emerging. For example, virtual assistants can leverage iterations on complex information in order to improve both effectiveness and efficiency in the interaction. But how to evaluate iterative changes for interactive search tasks is a significant challenge, and very few collections currently exist to test new approaches, let alone to test the end-to-end efficiency performance of such systems.

8.5 Broader Impact

The goal of efficiency is to make it possible to process (in the general sense) more data with fewer resources. For search this has the effect of reducing the overall quantity of hardware (and associated infrastructure). In a data center with 100,000 machines, a reduction in execution time by 5% is the equivalent of 5,000 machines, together with the power, cooling, and maintenance of those machines. This is a green computing contribution (aka green IR). The released resources could be used to absorb more user growth without requiring additional resources, could be directed towards other workloads, or could be simply turned off until needed.

A secondary effect is an increase in the amount of data that can be processed on a single resource, lowering the barrier to entry for researchers in the broader field of IR, and for companies entering the marketplace. A cluster of machines would not be needed if we make the work efficient enough such that it can be carried out on a laptop.

In the longer term, the quest for efficient IR has the potential to impact the design of CPUs and other components in the computing system. CPUs with vector processors (AVX, etc) are already in desktops – and the instructions are already being used in IR applications. FPGAs are now being integrated into CPUs, and we know that Microsoft is already using FPGAs for search. Future directions include GPGPU instruction integration into the main core.

Decreased costs might be achieved by offloading some of the work to the user's device (edge computing). For example, the last phase of reranking using an already learned (or adaptive) ranking function might even be performed on the user's mobile phone. Such a change might be disruptive to the standard model of internet or other search, and might also enable new privacy-preserving mechanisms.

8.6 Broadening SIGIR

Several opportunities exist to broaden the IR Community and interact with other communities on emerging efficiency challenges. Specific examples include the embedded / distributed computing community for cross-device search and machine-driven search in IoT devices, the NLP community for conversational IR, the ML community for complex ranking function optimization, the CPM community for future index structures, and the database community for combining structured and unstructured resources and for query optimization ideas.

8.7 Obstacles and Risks

For academics seeking to undertake research in large-scale IR systems there are obvious risks, primarily in regard to achieving genuine scale. Many of the research questions that offer the greatest potential for improvement – and the greatest possibilities for economic savings – involve working with large volumes of data, and hence significant computational investment. Finding ways of collaborating across groups, for example, to share hardware and software resources, and to amortize development costs, is a clear area for improvement.

Current practice in academic research in this area tends to revolve around one-off software developments, often by graduate students who are not necessarily software engineers, as convoluted extensions to previous code bases. At the end of each student's project, their software artifacts may in turn be published to GitHub or the like, but be no less a combination of string and glue (and awk and sed perhaps) than what they started with. Agreeing across research groups on some common data formats, and some common starting implementations, would be an investment that should pay off relatively quickly. If nothing else, it would avoid the ever-increasing burden for every starting graduate student to spend multiple months acquiring, modifying, and extending a code base that will provide baseline outcomes for their experimentation.

Harder to address is the question of data scale and hardware scale. Large compute installations are expensive, and while it remains possible, to at least some extent, for a single server to be regarded as a micro-unit of a large server farm, there are also interactions that cannot be adequately handled in this way, including issues associated with the interactions between different parts of what is overall a very complex system. Acquiring a large hardware resource that can be shared across groups might prove difficult. Perhaps a combined approach to, for example, Amazon Web Services might be successful in being granted a large slab of storage and compute time to a genuinely collaborative and international research group.

Harder still is to arrange access to large-scale data. Public web crawls such as the Common Crawl can be used as a source of input data, but query logs are inherently proprietary and difficult to share. Whether public logs can be used in a sensible way is an ongoing question. Several prior attempts to build large logs have not been successful: the logs of CiteSeer and DBLP are heavily skewed towards titles and authors, and academic groups have been unable to mobilize sufficiently large volumes of users to adopt instrumented toolbar and browser plugins. Attempts to use institutional proxy logs have shown that even with tens of thousands of users, the log is relatively sparse.

While efficiency does not automatically demand relevance judgments or similar "quality of retrieval" resources, there is a need for at least some level of quality assurance to be provided as there is often an efficiency / effectiveness trade-off to be quantified. Obtaining access to assessments at the required scale may also become a problem. To date, TREC resources have typically been used, noting that it is acceptable practice to measure effectiveness using one set of queries and documents, and then throughput using another set.

9 Personal Information Access

9.1 Description

Information created by, connected to, or consumed by an individual resides across a great number of separate information silos: personal devices (laptops, smart phones, watches, etc.); the web; personal or enterprise file shares; messaging systems and social media; and systems from external parties including medical doctors, bank records, employer and government records, and many others. Today, in order to search and make sense of this heterogeneous and disconnected set of data we depend on an equivalently large number of independent sources and access mechanisms, and rely on knowing where something is stored and how to get to it.

In what follows we discuss two, mostly orthogonal, challenges in accessing personal information: retrieval over personal information and personalized retrieval of information. These can be addressed separately, but each informs the other.

9.2 Motivation

People are producing and consuming more and more information. That information is stored in a multitude of places (cloud, computer, phone) and in a multitude of formats (e.g. mail, docs, slack, twitter, Facebook, apps, web searches, fitbits, sense cams, etc.). Without an integrated personal IR service, users must resort to multiple interactions over their own data. This is a time consuming process and prone to error. The time spent wading through and trying to find information in personal repositories results in wasted time, and in frustration. In spite of these evident shortcomings of current information technologies, little support exists to help people find, re-find, manage, organize, and share their personal information.

9.3 Proposed Research

There are at least four broad research questions which we need to address. First, **how can I find stuff that I've seen/interacted with before**, or should see; efficiently, effectively, and while preserving privacy? Second, **are there abstract representations of content and access patterns** which we can share - without violating privacy - to help design systems, to train machine learners, or to distribute computation? How can we safely generalize what we learn from one person to another? Third, if we have a rich model of a person, based on personal data and interactions, how can we use this to **personalize content or presentation**? When should we? What should we consider? And finally, **how can we search private information resources owned by others**, as distinct from searching our own information in other collections?

9.4 Research Challenges

Understanding (and Anticipating) Needs. A key challenge is understanding what information needs users have that the system should support. What these needs are, and how they are expressed, often depend on the device as well as the data. A personal digital assistant may also address these information needs proactively, e.g., identify routing tasks and pulling up all the related material, like preparing a travel expense declaration.

Task Representation, Identification & Abstraction. Once we understand something of personal search tasks, we need to represent access patterns, information needs, and behaviours in a way that existing systems can use to reason, and researchers can use to investigate new systems. This would involve extracting tasks from private data; abstracting them to allow insight; and developing a common protocol for describing tasks, or classes of tasks, without violating privacy or security.

Index and Schema Representation. A similar early research challenge is aggregation and representation of heterogeneous data sources and formats. One has to come up with a representation generic enough for data that ranges from unstructured to fairly structured, and sometimes with comprehensive metadata. Efforts on data integration (e.g., by the database community) and common vocabularies such as schema.org and Dublin Core (e.g., by the semantic web community) may be helpful to this end. As a community, we could distinguish better between the logical and physical representations of information, and express retrieval models at the logical level while delegating the actual relevance estimation over heterogeneous information in various data silos to the underlying physical layer.

Once the data has been represented in a common format, one has to think about suitable ways of querying it and interfaces to expose to users. While a rich query language might be helpful as an intermediary, it is unlikely to be apt for common users who would rather express their information needs using natural language.

Linking and Disambiguation across Silos. Once representational issues are addressed, a higher order research challenge is extracting entities from the heterogeneous data repositories - performing disambiguation, if required, and then linking entities within the collection. A more difficult challenge is linking of entities/objects to particular tasks, i.e. finding all the relevant artifacts associated with a given task, or set of tasks. This will enable a *personal knowledge graph* that contextualizes the relationships in user data.

One challenge here is that entities in the personal knowledge graph are unlikely to be generally popular, so that signals commonly used for named entity disambiguation (e.g., popularity and coherence) may just not work. Alternative signals (e.g., co-access patterns, similarity of usernames and email addresses) can potentially serve as replacements.

Ranking and Retrieval. Challenges related to search in personal data include the heterogeneity of information access tasks (ranking, summarization, etc.), of data sources, and of type of interactions (depending on device and modality). Retrieval methods will have to take the different characteristics of the data into account and also make use of metadata (e.g., creation times of files). It is foreseeable that the type of query result will depend on the information need at hand and may take the form of a list of files, bundles of interlinked files, or even a summary generated from the contents of relevant files.

In a new opportunity for IR, the searcher in a personal system may also be the author or curator. Their work, e.g. in filing into folders, can inform the retrieval process.

Computing over Aggregate Personal Data. Personal data provides a very rich representation of an individual's content and behavioral interaction patterns. Aggregating across individuals could augment this allowing generalization to new contexts.

For example, user interaction data is an important signal for learning-to-rank models in web search; these models require observing interactions across many users for the *same* query-document pair. This is challenging when considering personal (and private) data: the documents (e.g., emails or private files) are not shared across users, and queries are personal (e.g., “Sam’s email address”) and may not generalize well across users.

Personalization. As well as searching personal data, an agent tightly tied to a user (for example one running on a phone) can be greatly personalized. It may be possible to build per-person models of reading, cognition, biometrics, eye tracking, face recognition, or emotion. Challenges here include understanding the costs and benefits of such hyper-personalization, suggesting search interfaces or contents, and classifying users.

Privacy, Security, and Trust. A key challenge in PIA research is ensuring that the privacy and security of personal data is maintained and that the users trust that the system maintains their privacy. There is a wealth of work on privacy-preserving methods (for example in record linkage, enterprise search, and in masking queries to IR systems), but we must also understand how to distinguish private from shared from public data, what granularity to work at, which data to draw on in which circumstances, and how to explain these rules to both the subjects and the owners of information. Knowing how much information to disclose, to whom, and under what circumstances, is a tremendous challenge when even the fact of a query itself may cause harm if known (for example, in medical or patent cases). IR systems working with private data may also need to forget (or, forever hide) otherwise-accessible information under some circumstances, and both policies and mechanisms need to be developed for this.

Architecture and Applications. Along the way, there are many problems to solve if we are to build working tools either for research or general use. The basic architectural choices are open: where should the index live, where should the search happen, how do we aggregate data across silos? If we allow brokers to control access to silos, who gets to know what’s asked of a silo, and how can we route requests without knowing what holdings each broker has? Interface choices are also wide open and range from conventional metasearch, through PIM tools, intelligent and proactive personal assistants, to visualization and exploration tools to explore the data a person has.

9.5 Broader Impact

The most obvious, immediate, impact is the reduction in the cost of (re-)finding and (re-)using one’s own data which exists in many information silos (e.g., is the SWIRL document on my laptop, on a shared drive (which one?), or in my email (which account?)). Improved access to personal information will also reduce the frustration in finding and thus reduce the friction in working with personal data. This impacts everyone with a connected device - virtually every human on the planet. Improvements in this area will also translate to finding in the enterprise, where information workers spend large amounts of their time (re-)finding existing organizational knowledge within their own corporate repository and across all the repositories within the organization.

9.6 Broadening SIGIR

There is a lot of related work in other research communities, and opportunities for collaboration as well as starting points for IR research. This includes work on dataspaces, the EU NEPOMUK project on semantic infrastructure for desktop retrieval,[3] and work on semantics and retrieval in the lifelogging community, for example in the Lifelogging Tools and Applications workshops. Research in privacy-preserving data linkage and data mining will also be relevant to problems in linking and sharing data.

9.7 Obstacles and Risks

Personal information access, under different names, has been an outstanding problem for some time. Why hasn't the IR community made more progress? There are significant obstacles to even starting a research programme.

First, we note that the cost of entry is high. There has to be a lot of working parts to even be slightly useful, although it may be possible to start small by aggregating a few, related, silos. There is substantial engineering required for a minimal working system: to fetch data from different silos, parse different data formats, and monitor user activity. Further, access to the personal data and interaction data of a set of users is required in order to develop, test, and debug even a minimal working system.

Second, experimental evaluation for approaches in this research direction is highly challenging and many of our common practices are hard to apply. Information needs, for instance, are specific to users, entailing that only the user itself can judge the relevance of results. Likewise, the confidentiality of the data impedes creating and sharing test collections, limiting the reproducibility of experimental results. At least initially, case studies may be the only way to evaluate the approaches developed.

There are also obvious risks involved with searching among private stores or personalizing search.

Most obviously, when operating on personal data and considering sharing subsets of it among users, there is an inherent risk of breaching private information. Additionally, making use of personal data for personalization of search results may lead the user into a "filter bubble", showing only results reflecting its own opinions. When taking into account only the personal data of a single user, this could be even more grave than in more traditional settings with many users.

A final risk is that the business case for developing a "unified" PIA system is unclear, especially given the difficulty of accessing data in the first place. Unlike web search where the business case is driven around advertising, a PIA system, despite its obvious benefits, may not be worth implementing (without the right business model in place). Research into this area is therefore not only fraught with technical and evaluation difficulties, but also may not actually lead to a viable product in the short term. Within particular eco-systems (Google, Microsoft, Apple, etc) there is some integration between services - as each attempts to add value to their services, and each works with data it stores or controls - but this will invariably mean that such research will predominantly be performed within such companies. It will be important to break down the problem into specific tasks which researchers can pursue, rather than trying to tackle the whole system.

[3] `http://www.semanticdesktop.org`

10 Minor Topics

10.1 IR for an IoT World

This project aims to understand, design, implement, and evaluate an IR system for the Internet of Things (IoT) world, i.e., a world in which almost everything is connected and produces/contains data and information.

Motivation. IoT is a growing field, also under the "Industry 4.0" label. New devices and technologies are going to hit the market soon. Many funding agencies are devoting a large amount their funds to IoT-related projects. As a community, IR has the potential to address issues that other communities (like machine learning, data mining) will most likely not address: to study users and their needs, to better model specific needs exploiting sensor data, to devise novel and effective interaction modalities with information, and, overall, to apply results from specific IR subareas. The IoT situation is not much different from two previous waves that hit the IR field: the Web and mobile devices. Mobile IR was a successful story in SWIRL 2012. Papers on Location- and Context-awareness for Mobile IR are now being published in IR venues. IoT might well be the next wave after Web and Mobile.

Proposal. We foresee three main research directions:

(1) Novel data and collections. Since IoT sensors and devices contain and share data, new notions of collections, documents, and GIOs are likely to arise. There will be different protocols and formats, and a mixture of structured and unstructured data. Typical collections would contain for example, Lifelog data, sport tracking data, personal car and driving habits data, but more exotic scenarios might arise, like Thing retrieval (where it is not information but something in the real world that is retrieved). Efficiency will be very important (IoT devices need to be low energy) and all the research done in distributed/federated IR will need to be both taken into account and extended.

(2) New interaction modalities, for both information access and presentation. Because of the proliferation of sensors, context-aware IR systems (a.k.a. zero query) will have much more data available and exploitable to better model user needs. Conversely, the new devices will allow to present information in novel ways, using not only various devices (e.g., glasses, augmented objects, information augmented reality) but also modalities (e.g., Map- or Geocentric presentation modalities like www.thingful.net). This will open the possibility for information access and presentation to be combined by having users seamlessly browsing the information space by moving in the physical space, thus realizing the vision of a disappearing / ubiquitous IR system allowing to situate information in immersive spaces.

(3) Understanding users and needs. It is unclear, and worth studying, if current user needs will simply "scale-up" to the new collections and modalities, or if radically different needs will arise. The opportunity for this new wave of information access to be inclusive and supportive of neurodiversity will need to be supported by a wide range of user studies.

Challenges. The main challenge will be to integrate the "syntactic" level of rough sensor data within a broader concept of information. This research will move away from search engines that find IoT devices (`https://censys.io/`, `https://www.shodan.io/`), towards search engines that allow users to satisfy their needs also using information contained/produced by IoT devices and sensors. We are not interested in retrieving a “thing” via its name or other keywords specific to that thing, nor most sensor values of interest to data miners such as temperature, velocity, etc., but likely on its relevance/usefulness to current needs of the user (e.g., GPS coordinates). Only a deep understanding of the nature of “social sensors”, can provide richer ways of deriving contextual information from the activities of the users on social media (e.g., Foursquare Check-ins vs. Mobile phone position data).

Related efforts. There are some recent research trends in the IR community, like Context- and Location-aware retrieval, Geographic IR, cross-device search, that will likely be exploited and adapted to a higher level. The research will mutually benefit all other domains of IR: efficiency, conversational, interactive, GIOs, etc. Privacy will of course be a primary concern. Finally, related fields like ML, data mining, Ubicomp, HCI, will likely address similar issues.

10.2 Impact of IR Systems in Society

IR systems and tools intermediate most of the information consumed today, be it top sources on a given topic or the top news stories everyday. This project is concerned with assessing both short and long term impacts of the IR artifacts that our IR community has developed and studied over time.

Motivation. For over two decades, IR systems have influenced the way people around the world work, communicate, learn, and even how they live. Search engines have eased the way we access information. Recommender systems have changed the way we select what products and services we buy and consume. Social networks have changed how we keep in touch with family, friends, and acquaintances. Personal and conversational assistants are increasingly supporting us in our day-by-day tasks through reminders or contextual interventions such as heads-ups about traffic or weather. In essence, IR systems aim to empower individuals through access to information. However, do these systems always deliver positive outcomes to individuals, society, politics, the economy, and the environment? Information scientists with researchers from other disciplines should study the long-term and large-scale impacts of IR systems and technologies. Previous research indicates some of the areas that IR technologies impact, including:

- **Human cognitive processes.** Psychologists have been studying and raised concerns about the effect that easy access to via search engines might be having on, e.g., how people think and what people remember.

- **Individuals from minority communities.** Latanya Sweeney showed that querying by names predominantly used by black Americans is more likely to return results associated to arrest records than when querying by names predominantly used by white Americans. How many people might have been denied employment as a result? Similarly, search results for images of doctors or engineers are typically dominated by pictures of white men. How many girls of color might have been discouraged to pursue interests in these fields?

- **Social Communities.** Given the political climate around the world, many have raised concerns about the potential of highly personalized consumption of information and the filter bubbles that current IR systems can create, to drive increased polarization along social and political lines within and across local and regional communities.

- **Businesses.** Search platforms have disrupted the news media ecosystem through digital ads and targeting, being able to tailor content to each of their user interests, viewpoints, and beliefs. But, what are the costs of, for instance, accessing newsworthy information without the traditional journalistic curation?
- **Environment.** Providing the energy required to run the data centres that support large-scale search engines can be at a significant cost to the environment but IR technology can also contribute to smarter more efficient infrastructure for cities and transportation.

Proposal. While there is some consensus around the broad influence of IR, we often lack hard numbers, and more efforts should made towards identifying and quantifying specific (both positive and negative) outcomes of IR systems on society, beyond the short term goals of their users or their clients. Research directions include:

- What is the long-term impact of the ubiquitous access to information (anywhere, anytime) via web search engines on how people think, how people learn, and what they remember?
- How to educate developers of IR algorithms to avoid unfair bias that may have negative consequences for some individuals?
- Does widespread access to information improve or damage social cohesion?
- What is the economic impact of various developments in information retrieval?
- Can IR technologies have a net positive impact on the environment?

Challenges. Multidisciplinary research that needs to involve other disciplines such as social sciences, media studies, environmental science, political science, psychology, socio-psychology, psycho-sociology, economics, and maybe even ethnography and anthropology. What is an appropriate framework to evaluate the long term impact of developments in IR? How to isolate the effect of IR systems from other informational or societal-related factors?

Related efforts. There is a growing number of efforts in neighboring fields looking either broadly at the impact of computing systems, or at that of specific ML or AI tools, including prominent initiatives such as dedicated research institutes, e.g., Data & Society and AI Now; as well as workshops and conferences, e.g., Conference on Fairness, Accountability, and Transparency (FAT*) or AAAI/ACM Conference on AI, Ethics, and Society (AIES). These are also related with existing efforts in the space of FACT IR, discussed earlier in this report.

11 Conclusion

Information Retrieval remains a vital and active area of research in both academia and industry. Satisfying people's information needs is a fundamental, multi-disciplinary problem, and this reports captures the many important research themes in this important research area. The findings are in no way prescriptive. That is, many more important research themes were suggested than could be explored in our brief time in Lorne. We hope that SWIRL can inspire future strategic workshops which continue to shape and grow the Information Retrieval research community.

Deep Learning

[DZS+17] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 65–74, New York, NY, USA, 2017. ACM.

[HKG+15] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 1693–1701, Cambridge, MA, USA, 2015. MIT Press.

[HXTS16] Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. A joint many-task model: Growing a neural network for multiple NLP tasks. *CoRR*, abs/1611.01587, 2016.

[MC18] Bhaskar Mitra and Nick Craswell. *An introduction to neural information retrieval.* Foundations and Trends in Information Retrieval. Now Publishers Inc., 2018.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[MDB17] Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. *CoRR*, abs/1707.05589, 2017.

[MDC17] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. Learning to match using local and distributed representations of text for web search. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1291–1299, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

[MSC+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[SS17] Ryan Spring and Anshumali Shrivastava. Scalable and sustainable deep learning via randomized hashing. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 445–454, New York, NY, USA, 2017. ACM.

[SSWF15] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 2440–2448, Cambridge, MA, USA, 2015. MIT Press.

[SWM17] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *CoRR*, abs/1708.08296, 2017.

[WCB14] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *CoRR*, abs/1410.3916, 2014.

[XCL17] Chenyan Xiong, Jamie Callan, and Tie-Yan Liu. Word-entity duet representations for document ranking. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 763–772, New York, NY, USA, 2017. ACM.

[XDC$^+$17] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 55–64, New York, NY, USA, 2017. ACM.

[XMS16] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 2397–2406. JMLR.org, 2016.

Stateful Search

[CRH16] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. Towards conversational recommender systems. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 815–824, New York, NY, USA, 2016. ACM.

[Guy16] Ido Guy. Searching by talking: Analysis of voice queries on mobile web search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 35–44, New York, NY, USA, 2016. ACM.

[HTML12] Brent Hecht, Jaime Teevan, Meredith Ringe Morris, and Dan Liebling. Searchbuddies: Bringing search engines into the conversation. In *ICWSM 2012 - Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, pages 138–145, 12 2012.

[JHAJ$^+$15] Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. Automatic online evaluation of intelligent assistants. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 506–516, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee.

[LMR+16] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202. Association for Computational Linguistics, 2016.

[RC17] Filip Radlinski and Nick Craswell. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 117–126, New York, NY, USA, 2017. ACM.

[SSB+16] Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3776–3783. AAAI Press, 2016.

[VSAC17] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles L. A. Clarke. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, pages 2187–2193, New York, NY, USA, 2017. ACM.

[WVM+17] Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M. Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449. Association for Computational Linguistics, 2017.

[YSW16] Rui Yan, Yiping Song, and Hua Wu. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 55–64, New York, NY, USA, 2016. ACM.

Responsible Information Retrieval

[BCZ+16] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc., 2016.

[BS16] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104, 2016.

[CBN17] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, April 2017.

[DTD15] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *PoPETs*, 2015(1):92–112, 2015.

[Eic18] Carsten Eickhoff. Cognitive biases in crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 162–170, New York, NY, USA, 2018. ACM.

[JKB+15] Lilli Japec, Frauke Kreuter, Marcus Berg, Paul Biemer, Paul Decker, Cliff Lampe, Julia Lane, Cathy O'Neil, and Abe Usher. Big data in survey researchaapor task force report. *Public Opinion Quarterly*, 79(4):839–880, 2015.

[MSA+17] Rishabh Mehrotra, Amit Sharma, Ashton Anderson, Fernando Diaz, Hanna Wallach, and Emine Yilmaz. Auditing search engines for differential satisfaction across demographics. In *Proceedings of the 26th International Conference on World Wide Web*, 2017.

[OBC17] Jahna Otterbacher, Jo Bates, and Paul Clough. Competent men and warm women: Gender stereotypes and backlash in image search results. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 6620–6631, New York, NY, USA, 2017. ACM.

[OCDK17] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. Social data: Biases, methodological pitfalls, and ethical boundaries. Available at SSRN, 2017.

Evaluation

[ABSJ17] Aman Agarwal, Soumya Basu, Tobias Schnabel, and Thorsten Joachims. Effective evaluation using logged bandit feedback from multiple loggers. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 687–696, New York, NY, USA, 2017. ACM.

[BPnC+13] Léon Bottou, Jonas Peters, Joaquin Qui nonero Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14:3207–3260, 2013.

[Car12] Ben Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Trans. Inf. Syst.*, 30(1):4:1–4:34, March 2012.

[Car15a] Ben Carterette. Bayesian inference for information retrieval evaluation. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, ICTIR '15, pages 31–40, New York, NY, USA, 2015. ACM.

[Car15b] Ben Carterette. The best published result is random: Sequential testing and its effect on reported effectiveness. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 747–750, New York, NY, USA, 2015. ACM.

[JSS17] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 781–789, New York, NY, USA, 2017. ACM.

[MBST17] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Transactions on Information Systems (TOIS)*, 35(3):24, 2017.

[MZ08] Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1):2, 2008.

[Ope15] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251), 2015.

[Sak16a] Tetsuya Sakai. Statistical significance, power, and sample sizes: A systematic review of sigir and tois, 2006-2015. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 5–14, New York, NY, USA, 2016. ACM.

[Sak16b] Tetsuya Sakai. Topic set size design. *Inf. Retr.*, 19(3):256–283, June 2016.

[SKA$^+$17] Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3632–3642. Curran Associates, Inc., 2017.

[VSS17] Ellen M. Voorhees, Daniel Samarov, and Ian Soboroff. Using replicates in information retrieval evaluation. *ACM Trans. Inf. Syst.*, 36(2):12:1–12:21, August 2017.

[WBMN16] Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc 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*, SIGIR '16, pages 115–124, New York, NY, USA, 2016. ACM.

Efficiency

[BCR16] Xiao Bai, B. Barla Cambazoglu, and Archie Russell. Improved caching techniques for large-scale image hosting services. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 639–648, New York, NY, USA, 2016. ACM.

[BNMN16] Leonid Boytsov, David Novak, Yury Malkov, and Eric Nyberg. Off the beaten path: Let's replace term-based retrieval with k-nn search. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1099–1108. ACM, 2016.

[CGBC17] Ruey-Cheng Chen, Luke Gallagher, Roi Blanco, and J Shane Culpepper. Efficient cost-aware cascade ranking in multi-stage retrieval. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 445–454. ACM, 2017.

[GHL+17] Bob Goodwin, Michael Hopcroft, Dan Luu, Alex Clemmer, Mihaela Curmei, Sameh Elnikety, and Yuxiong He. Bitfunnel: Revisiting signatures for search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 605–614, New York, NY, USA, 2017. ACM.

[KBC+17] Tim Kraska, Alex Beutel, Ed H. Chi, Jeff Dean, and Neoklis Polyzotis. The case for learned index structures. 2017. ArXiV – `https://arxiv.org/abs/1712.01208`.

[LNO+15] Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, Nicola Tonellotto, and Rossano Venturini. Quickscorer: A fast algorithm to rank documents with additive ensembles of regression trees. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 73–82, New York, NY, USA, 2015. ACM.

[OV14] Giuseppe Ottaviano and Rossano Venturini. Partitioned elias-fano indexes. In *Proceedings of the 37th International ACM SIGIR Conference on Research &#38; Development in Information Retrieval*, SIGIR '14, pages 273–282, New York, NY, USA, 2014. ACM.