Online Evaluation for Information Retrieval

Katja Hofmann
Microsoft
katja.hofmann@microsoft.com

Lihong Li
Microsoft
lihongli@microsoft.com

Filip Radlinski
Microsoft
filip.radlinski@microsoft.com
Editorial Scope

Topics

Foundations and Trends® in Information Retrieval publishes survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation, and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2016, Volume 10, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.
Online Evaluation for Information Retrieval

Katja Hofmann  
Microsoft  
katja.hofmann@microsoft.com

Lihong Li  
Microsoft  
lihongli@microsoft.com

Filip Radlinski  
Microsoft  
filip.radlinski@microsoft.com
## Contents

1 Introduction 2  
   1.1 Terminology 3  
   1.2 Motivation and Uses 4  
   1.3 This Survey 5  
   1.4 Organization 6  

2 Controlled Experiments 7  
   2.1 Online Controlled Experiments in Information Retrieval 7  
   2.2 Planning Controlled Experiments 10  
   2.3 Data Analysis 16  
   2.4 Between-subject Experiments 20  
   2.5 Extensions to AB testing 22  
   2.6 Within-subject Experiments 26  
   2.7 Extensions to Interleaving 29  

3 Metrics for Online Evaluation 31  
   3.1 Introduction 31  
   3.2 Absolute Document-level Metrics 33  
   3.3 Relative Document-level Metrics 36  
   3.4 Absolute Ranking-level Metrics 37  
   3.5 Relative Ranking-level Metrics 39
Abstract

Online evaluation is one of the most common approaches to measure the effectiveness of an information retrieval system. It involves fielding the information retrieval system to real users, and observing these users’ interactions in-situ while they engage with the system. This allows actual users with real world information needs to play an important part in assessing retrieval quality. As such, online evaluation complements the common alternative offline evaluation approaches which may provide more easily interpretable outcomes, yet are often less realistic when measuring of quality and actual user experience.

In this survey, we provide an overview of online evaluation techniques for information retrieval. We show how online evaluation is used for controlled experiments, segmenting them into experiment designs that allow absolute or relative quality assessments. Our presentation of different metrics further partitions online evaluation based on different sized experimental units commonly of interest: documents, lists and sessions. Additionally, we include an extensive discussion of recent work on data re-use, and experiment estimation based on historical data.

A substantial part of this work focuses on practical issues: How to run evaluations in practice, how to select experimental parameters, how to take into account ethical considerations inherent in online evaluations, and limitations. While most published work on online experimentation today is at large scale in systems with millions of users, we also emphasize that the same techniques can be applied at small scale. To this end, we emphasize recent work that makes it easier to use at smaller scales and encourage studying real-world information seeking in a wide range of scenarios. Finally, we present a summary of the most recent work in the area, and describe open problems, as well as postulating future directions.

Introduction

Information retrieval (IR) has a long and fruitful tradition of empirical research. Since early experiments on indexing schemes, and the development of the Cranfield paradigm, researchers have been striving to establish methodology for empirical research that best supports their research goals – to understand human information seeking, and to develop the most effective technology to support it.

In the past decade, IR systems, from large-scale commercial Web search engines to specialized analysis software, have become ubiquitous. They have transformed the way in which we access information, and are for many an integral part of their daily lives. This shift towards everyday, ubiquitous IR systems is posing new challenges for empirical research. While it was previously possible to substantially improve IR systems by measuring and optimizing reasonably objective criteria, such as topical relevance, this is no longer sufficient. IR systems are becoming increasingly contextual and personal. They take into account information about their users’ current situation as well as previous interactions, and aim to predict their users’ requirements and preferences given new contexts. No longer can users or experts be asked to provide objective assessments of retrieval quality for such complex scenarios.
1.1. Terminology

Online evaluation for IR addresses the challenges that require assessment of systems in terms of their utility for the user. The current state of the art provides a set of methods and tools, firmly grounded in and informed by the tradition of controlled experimentation. Giving an overview of these methods and their conceptual foundations, as well as guiding the reader on how to run their own online evaluations are the purposes of this survey.

In the next section, we define key concepts and terminology used throughout this survey. Then, we closely examine the motivations for online evaluation, and provide example use cases. Finally, we outline the scope and organization of the remainder of this survey.

1.1 Terminology

For the purpose of this survey, we adopt the following definition of online evaluation.

**Definition 1.1.** Online evaluation is evaluation of a fully functioning system based on implicit measurement of real users’ experiences of the system in a natural usage environment.

The first key to the definition is implicit measurement, which we take to include any measurements that can be derived from observable user activity that is part of users’ natural or normal interaction with the system [Kelly and Teevan 2003]. Implicit measurements can range from low-level and potentially noisy signals, such as clicks or dwell-times, to more robust signals, such as purchase decisions. The key distinction we make between implicit and explicit measurements is that implicit measurements are a by-product of users’ natural interaction, while explicit ones are specifically collected for feedback purposes. Both types of measures can also be combined into composite metrics capturing higher-level concepts, such as user satisfaction. These considerations give rise to a wide range of metrics, as discussed in Chapter 3.

We specifically include methods for offline estimation, i.e., the estimation of online evaluation metrics based on past observations of users’ behavior, in Chapter 4. Such estimation substantially increases the flex-
ibility of online evaluation and facilitates theoretically well-founded end-to-end evaluation of system components.

1.2 Motivation and Uses

Online evaluation is often seen as a set of methods that are particularly applicable in industry and industrial research. In these settings, a fully functioning IR system is typically available and in need of constant innovation. These factors have significantly contributed to the rapid adoption of online evaluation techniques in these settings. In industry, online evaluation approaches such as AB tests (c.f., Section 2.4) and interleaved comparisons (Section 2.6) are now the state of the art for evaluating system effectiveness [Kohavi et al., 2009, Radlinski and Craswell 2010, Li et al. 2011, Bendersky et al. 2014].

However, it is important to recall that much of the initial work on online evaluation originated in academic settings. Important motivations here were the need for reliable measurement of search quality of specialized search services [Radlinski et al. 2008c]. This line of work originated in the tradition of interactive IR. The fruitful exchange of ideas between applications and research continues today. On the one hand, practical challenges from IR applications motivate the development of online evaluation methodology; Chapter 2 gives a few examples. On the other hand, lessons learned in practical applications make their way into the state-of-the-art methodological tool set of IR researchers.

In the context of both practical applications and basic research, a key aspect of online evaluation is its reliance on controlled experiments. This allows the researcher to answer explanatory questions, which can explain causal relations in observed phenomena. In practical settings, this is crucial for correctly attributing observed changes in user behavior to system behavior. In research, this allows the development of theory in terms of causal concepts. More details on controlled experiments for online evaluation are provided in Chapter 2.

Finally, in Chapter 5, we discuss pros and cons of online evaluation, compared with more traditional offline evaluation methodology. This will help guide the reader to understand when an online evaluation is suitable, and when it is not.
1.3 This Survey

Online evaluation comprises a specific set of tools and methods that we see as complementary to other evaluation approaches in IR. In particular, online evaluation addresses questions about users’ experience with an IR system that are quite distinct from those answered by offline evaluation using a test-collection-based approach, surveyed by Sanderson [2010]. Test-collection-based evaluation models users at varying levels of abstractions, and uses explicit assessments and offline metrics to assess system performance under these abstractions. Questions that are more appropriate for offline evaluation are those for which reliable and unbiased judgments can be collected from assessors (be they trained experts or crowdsourced representative users), but would be hard to infer from user interactions; an example being the quality of a document. Vice versa, online evaluation is more appropriate when the opposite is the case: for example, which of two topically relevant documents users find more interesting.

This survey does not discuss test-collection-based approaches in any detail, but points out conceptual differences when deemed appropriate. Furthermore, Chapter 5 focuses on online evaluation and test-collection-based approaches along a few dimensions.

Closely related to online evaluation is the long tradition of interactive IR (IIR) and the experimental framework developed for it, as surveyed by Kelly and Gyllstrom [2011]. We see online evaluation as a continuation of the IIR tradition, with considerable overlap. However, online evaluation extends to the specific requirements, limitations, and opportunities afforded by the scale, natural settings, and levels of control that are available in online settings. Generally speaking, IIR approaches, such as lab studies, are more appropriate for answering questions that require a high level of experimental control: for example, which tasks or queries a study participant is asked to solve. Conversely, online evaluation is preferred when researchers aim to study natural interactions at scale. This survey necessarily overlaps with some of the material that is relevant in the IIR setting, and we endeavor to point out connections as much as feasible. Our main focus will be on methodological questions that are specific to online evaluation settings.
Introduction

Throughout this survey, we consider IR in a broad sense, including for instance recommender systems and advertisement placement. Many aspects of online evaluation are shared across these areas. For example, early work on using historical information for estimating online performance focused on ad placement [Langford et al., 2008] and news recommendation [Li et al., 2011]. We cover work in all these areas, and emphasize work that is specific to IR, such as search result ranking evaluation, as appropriate.

We have also highlighted particular places in the text with tips (such as the one below) that may be particularly useful for experimenters performing online evaluation without having access to very large user bases. While at first glance online evaluation may appear to be best suited to settings such as commercial search engines, in fact it has been widely used in academic settings as well.

1.4 Organization

We start in Chapter 2 by motivating the need for controlled experiments and detailing common experiment designs used in online evaluation, with a focus on experimentation methodologies that are particularly useful for IR. Following this, Chapter 3 gives an extensive overview of the variety of metrics that have been proposed for different tasks and research questions. Considering how to re-use online measurement data, Chapter 4 details offline estimation of online metrics from historical data. Turning to more practical issues, Chapter 5 discusses advantages and limitations of online evaluation, while Chapter 6 discusses practical issues around running online experiments. Finally, Chapter 7 concludes this survey with an outlook on emerging trends and open challenges.
References


References


References


References


References


References


References


References


References

