Evaluation in Contextual Information Retrieval: Foundations and Recent Advances within the Challenges of Context Dynamicity and Data Privacy

Lynda Tamine, Mariam Daoud

Evaluation in Contextual Information Retrieval: Foundations and Recent Advances within the Challenges of Context Dynamicity and Data Privacy

LYNDA TAMINE, University of Toulouse, IRIT Laboratory, Toulouse 31062, France
MARIAM DAOUD, Seneca College of Applied Arts and Technology, Toronto M5H 212, Canada

Context such as the user's search history, demographics, devices, and surroundings, has become prevalent in various domains of information seeking and retrieval such as mobile search, task-based search, and social search. While evaluation is central and has a long history in information retrieval, it faces the big challenge of designing an appropriate methodology that embeds the context into evaluation settings. In this article, we present a unified summary of a wide range of main and recent progress in contextual information retrieval evaluation that leverages diverse context dimensions and uses different principles, methodologies, and levels of measurements. More specifically, this survey article aims to fill two main gaps in the literature: First, it provides a critical summary and comparison of existing contextual information retrieval evaluation methodologies and metrics according to a simple stratification model; second, it points out the impact of context dynamicity and data privacy on the evaluation design. Finally, we recommend promising research directions for future investigations.

Key Words: Information retrieval, context, evaluation, relevance, users, tasks

https://doi.org/10.1145/3204940

1 INTRODUCTION

1.1 Background: Context Definition

The huge volume of digital information available on the World Wide Web (WWW), the diversity of web search tasks, the variety of user's profiles seeking information, as well as the rapid growth in the use of diverse devices made information seeking and retrieval more challenging. This has resulted in a growing demand for leveraging contextual knowledge to improve the search effectiveness. In an attempt to foster research in this field, several focused context-based information

Authors' addresses: L. Tamine, University of Toulouse, IRIT Laboratory, 118 Route de Narbonne, Toulouse 31062, France; email: tamine@irit.fr; M. Daoud, Seneca College of Applied Arts and Technology, 1750 Finch Ave E, North York, ON M2J 2X5, Toronto M5H 212, Canada; email: daoud.mariam@gmail.com.

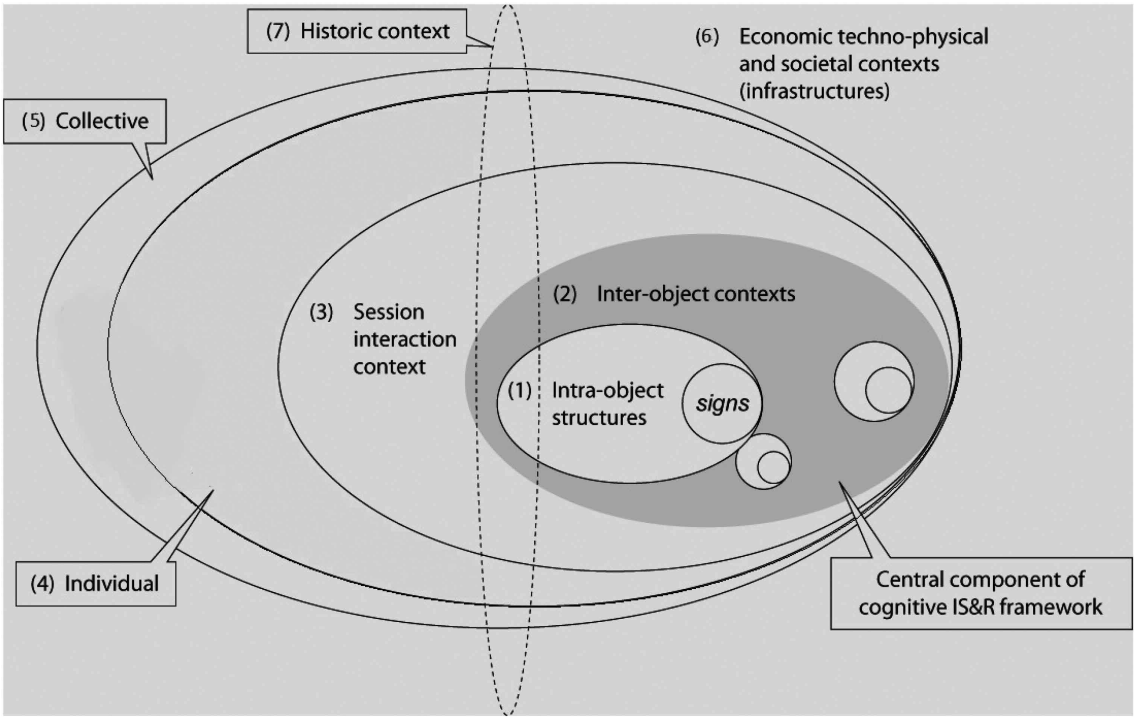

Fig. 1. Nested model of context stratification for IIR. From Revision of Ingwersen & Jarvelin (2004, 2005).

retrieval (IR) initiatives have been launched such as the Information Interaction in Context Symposium (IIIX) [42] that started in 2006, the ACM SIGIR 2005 Workshop on Information Retrieval in Context [66], and the Conference on Human Information Interaction and Retrieval (CHIIR) [107] that started in 2016 and represent a merger of IIIX and Human Computer Information Symposium (HCIR).

Several context definitions and taxonomies [62, 67, 101, 106, 135] were proposed in contextual IR literature where they mainly differ by their constituent elements. Early approaches in contextual IR focus on the cognitive context represented with the user profile. This context includes the user interests and background during the search. User interests have been shown to be the most important context element that helps disambiguate the search [112].

Multidimensional context definitions have been proposed through context taxonomies [27, 50, 67, 106, 121, 135]. Each of these taxonomies defines a stratification model of context features according to dimensions. Saracevic [121] and Ingwersen [62] were the first to introduce the context without distinction with the search situation. The context is defined based on several cognitive dimensions such as the user cognitive dimension, the social and organizational environment, the intention behind the search, the user goals and the system context. Based on Cool Taxonomy [27], a more comprehensive definition of context according to levels has been proposed. The first context level is the search environment that includes the cognitive, social and professional factors that impact the user perception of relevance. The second level is related to the user background, search goals, and intentions. The third context level is the user-system interaction that involves the impact of the environment on the user relevance assessments. The last context level concerns the query linguistic level that impacts the performance of the system in interpreting and disambiguating the queries. In Reference [63], a nested model of cognitive context stratification is defined based on seven dimensions as presented in Figure 1: (1) The intra-objects structures where objects could represent software entities, interface, or the document. Intra-object context

.

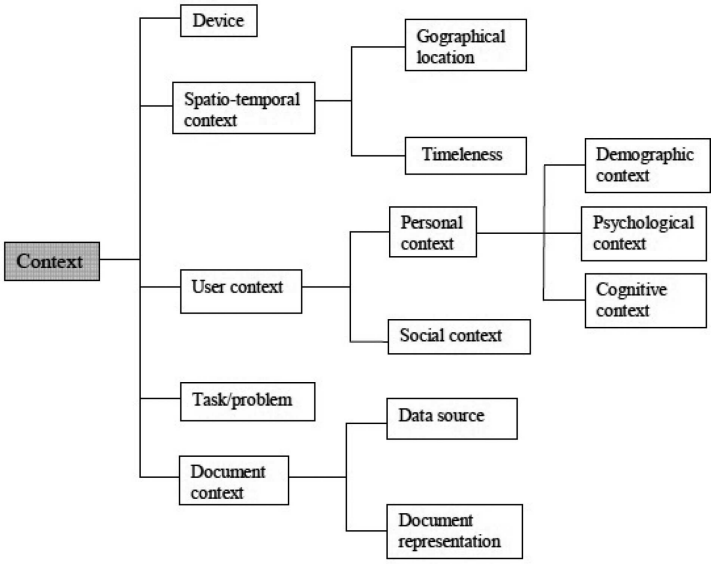

Fig. 2. Context taxonomy of Tamine et al. (2010).

represents a relevant information surrounding the object such as text surrounding an image or vice versa. They could also be intra-document features, such as words in context of phrases, in context of sentences, in context of paragraphs, and so on. (2) The inter-objects contexts where objects could be contextual to one another such as the document references/outlinks and citations/inlinks considered as giving context to or taking context from the content of other objects, respectively. (3) Interaction context (or session activity) consists of social interaction and interactive IR activities that can be made available for the system to help it interpret current searcher actions. When the interface is considered as the core of the model, interaction context refers to the retrieval session that includes intra-system interactions like query expansion, the searching actors and their emotional-conceptual traits, and the like that are considered nested within the interaction context. (4) An individual conceptual and emotional context (actor: searcher, author); and systemic (engine, interface, information object) and domain properties immediately surrounding the core actor or component (work task, interest, preference, product). (5) A collective conceptual and emotional context (actors: search teams, author groups), systemic (networks, meta-engines, information objects, information space), and sociocultural and organizational structures in local settings. (6) Techno-economic-politico-societal infrastructures influencing all actors, components, and interactive sessions. (7) The historic context operating across this stratification, which reflects the history of all participating actors' experiences, forming their expectations. All Interactive IR processes and activities are under the influence of this temporal form of context.

In Reference [136], a context taxonomy has been defined including mobile context, user context, task context, and document context as presented in Figure 2. The mobile context includes the search environment such as the use of mobile devices, personal agendas, network characteristics, and the like. The user context is defined along two main dimensions: (a) the personal context including demographic, psychological, and cognitive context and (b) the social context. The search task dimension is concerned with two main aspects: the type of user information needed behind the query such as informational, navigational, or transactional, and the domain of user interest specific to the search task. The document context dimension is defined according to three main sub-dimensions: (1) the document representation such as the structural elements, the citations,

and the metadata, (2) the characteristics of the data source such as credibility, and (3) the quality of the information including freshness, precision, coherence, security, and so on.

From an evolving perspective, a distinction has been made between static and dynamic context.

- —A static context includes persistent characteristics of the user, system, and physical environment that influence the search. At a given time of search, we identify the following context elements as static: user personal context, including the user background, expertise, and education level; the spatio-temporal context including location and time of search; document context including language and information quality and credibility, device, and network characteristics. In addition to the aforementioned static context elements, many traditional areas in IR typically consider retrieval tasks that are most often static, purely topical, content-only, well-defined, and exhaustive. Early and major evaluation methodologies relying on a static context are reported in Section 2.
- —A dynamic context reflects an evolving user information need along a search task. It includes a dynamic user request expressed through a series of query reformulations until the completion of the search task [18], viewed as a set of sub-tasks where each addresses one aspect of a main task [156], user feedback and click-through information that are accumulated through several rounds of user-system interactions until completion of the search task [161], or user-generated content and a log of user interactivity at different stages of the search process.[1] Context dynamicity raises new evaluation challenges and provides opportunities to design appropriate evaluation methodologies, in addition to those used for static context, as detailed in Section 4.

1.2 Basic Hypothesis of IR Evaluation: In-Context Debate and Challenges

On a fundamental and historical level, IR evaluation involves materials such as protocols, methods, data tests, and metrics that provide levels of qualitative and quantitative measurements of the IR system effectiveness. The primary goal of traditional IR evaluation studies is to basically answer the following question: *Is the system able to select relevant documents?* According to this perspective, system-centered IR evaluation has experienced great advancements through the traditional laboratory model initiated by Reference [26] in the Cranfield Project II. This model is widely used in IR evaluation campaigns such as TREC, launched in 1991 by the National Institute of Standards and Technology (NIST), and the Defense Advance Research Project Agency (DARPA) in USA, INEX,[2] CLEF,[3] NTCIR,[4] and so on.

The underlying evaluation approach abstracts neither the user, viewed as the primary or internal context, nor the surrounding internal or external context, by focusing on the topical or algorithmic relevance of the information [9], even though early studies highlighted the multidimensional aspect of relevance. Indeed, as stated by Reference [120]: "*Relevance has a context, external and internal … Context: the intention in expression of relevance always from context is directed toward context. Relevance cannot be considered without context.*" Dealing with relevance in IR evaluation through algorithms, measures, and test collections should then be shifted beyond the scope of topical relevance where the basic goal of evaluation becomes: *How can we make the system and user, within his surrounding context, work together to select relevant documents?*. This perspective, raised from the cognitive view of IR [62], launched the user-centered evaluation approach. With this goal on one hand and context definitions and models reported in the literature on the other hand,

[1] http://social-book-search.humanities.uva.nl/#/interactive.
[2] INitiative for the Evaluation of XML retrieval.
[3] Cross Language Evaluation Forum.
[4] NII Testbeds and Community for Information access Research.

traditional IR evaluation models are challenged. Researchers adopting the user-centered evaluation paradigm have to define more realistic hypothesis and design novel evaluation norms that fit context-aware evaluation [75, 87]. Evaluation studies designed at the system level adopt the laboratory-based model under controlled frameworks [67, 122]. The underlying evaluation setting is based on several hypotheses highlighted below along with the context challenges.

1.2.1 Relevance. Historically, relevance has been a core notion in IR, understood intuitively and less formally [122]. Early evaluation models were built based on the system or algorithmic interpretation of relevance. Consequently, the goal of traditional evaluation models is to measure the effectiveness of (topically) ranked relevant documents given topical queries.

Hypothesis 1. Relevance involved in IR evaluation [10] is algorithmic, expressing the relation between the query (terms) and the collection of information objects expressed by the retrieved results.

— *Arguments in favor*: algorithmic relevance is objective and measurable, leading to valuable comparisons. This aspect is actually the best claim issued by major evaluation campaigns such as TREC and CLEF.
— *Arguments against*: algorithmic relevance is the minimal level of relevance, which is overwhelmed by other relevance levels such as cognitive, situational, and affective relevance [10].

Hypothesis 2. Relevance is stable, independent, and consistent. Stability mainly refers to the low variation of relevance inferences for the same user involved in a search task. Relevance independence leads to the assertion that information objects are assessed independently. Consistency deals with the uniqueness of relevance judgments or a perfect relevance judgments' agreement between assessors.

— *Arguments in favor*: useful when we address the system effectiveness. Indeed, although algorithmic relevance evaluation relies on subjective human relevance assessors, the differences in relevance assessments do not lead to changes in system ordering as stated in Reference [143] and Lesk and Salton's work [89]. Hence, any resulting disagreement between assessors can be solved, and all IR systems competing in the evaluation are treated equally.
— *Arguments against*: regarding stability, information science findings [120] report that relevance inference along a search task is accomplished within a dynamic, interacting process, in which the interpretation of other attributes may change, as context changes. The issue of relevance independence has already been addressed by several studies on the effect of several features of IR tasks, the user's relevance assessment such as result ranking [4, 61] or information format [69], and so on. The related finding clearly argues against the independence hypothesis. Recent studies on anchoring and adjustment in relevance estimation [88, 126] prove that the human annotators are likely to assign different relevance labels to a document, depending on the quality of the last document they had judged for the same query. Regarding consistency, it is well known today that TREC and CLEF style experiments are generally based on expert assessments seen as objective, while real-life IR settings are based on real users for whom assessments are seen as subjective [19, 60] and several contextual factors affect the users when judging document relevance. Several studies validate the fact that relevance assessments are not generalizable across users [40, 58, 132]. Recent studies [142] have also shown that the use of mobile devices affects the user notion of relevance where there are differences between mobile and desktop judgments.

1.2.2 Users.

Hypothesis 1. Users as black-boxes.

— *Arguments in favor*: abstracting users allows evaluating the effectiveness of the system within replicable experiments.
— *Arguments against*: the cognitive view of IR [62] places the user at the core of the IR task. Rather than studying the entire retrieval process, both internal and external contexts are studied and differentiated across users. In Reference [101], the author also identified user interaction variables and how to use them to evaluate contextual search.

Hypothesis 2. Assessors as users.

— *Arguments in favor*: Relevance assessments are useful when the focus is on the system outcomes but not on the search task. More specifically, the stability of TREC collections relies on the assumption that relevant documents are relevant to a single assessor, who functions as a user, at a single point of time [77].
— *Arguments against*: according to the hypothesis of multi-dimensional notion of relevance and in relation to individual differences, relevance assessments are not generalizable across users; they depend not only on the search topic but also on the individual contexts embedded within the retrieval task. Experimental and observational findings regarding this issue have been summarized in References [120, 122]. In support of collecting real relevance assessments, recent studies [19] argue on the use of implicit feedback from real users to replace the relevance judgments obtained by the experts as an alternative to the traditional Cranfield approach.

1.2.3 Levels of Information Needs: Topics versus Tasks.

Hypothesis 1. Information needs are equivalent to topics within constant tasks.

— *Arguments in favor*: coherent with the relevance hypothesis.
— *Arguments against*: generally speaking, a topic represents the purpose of the IR task; several topics within the same task may be the focus of specific searches. Thus, the task, as topic background, helps define the situational relevance that goes beyond topical relevance [10]. Several studies show that, in addition to topics, tasks affect users' information seeking, interactions, and relevance assessments; consequently, they are good candidate variables for an in-depth user and system analysis [76, 77, 81]. In addition, TREC task's track [161] argues on defining a task as a general topic that yields a set of subtopics to help achieving the task. Several studies [90, 96, 100] defined search tasks and sub-tasks while differentiating them from topics.

1.2.4 Summative versus Formative Evaluation.

Hypothesis 1. Users are situated at the end of the IR system evaluation design.

— *Arguments in favor*: supports the dichotomy system versus user-centered evaluation in the sense that while the evaluation focuses primarily on the outcome of the search, the user is considered as an ancillary facet.
— *Arguments against*: [98] argued that "*... we cannot discover how users can best work with systems until the systems are built, yet we built systems based on knowledge of users and how they work.*" Thus, formative evaluation of IR systems should run across the entire evaluation spectrum of the system, the search context and the interaction between the user and the system [37, 67].

1.3 Related Surveys and Differences

Although contextual IR has received great attention in recent research studies and that the IR community is aware that evaluation of context-based IR is a challenging task, there are few survey papers about the evaluation of contextual IR [51, 101, 135]. The review in Reference [51] focuses on the evaluation of context-based search in a particular application domain, namely mobile IR. Melucci [101] also addresses the issue of evaluation in a survey about contextual IR in general. However, the author briefly discusses the impact of interaction variables on user relevance assessment and retrieval effectiveness. In an attempt to address challenges in system- and user-centered evaluation approaches, a workshop [110] has been held to initiate productive knowledge exchange and partnerships by combining user and system-centered methodologies in meaningful ways that can respond to the increasing user, task, system, and contextual complexity of the IR field.

Our previous article [135], which reviews existing research progress on the evaluation of contextual IR, presents an overview of early evaluation methodologies and measures that mostly fall in the system-based category where some important aspects of user-based evaluation with related levels of measurements, as well as the diversity of context forms were missing. Besides, new evaluation approaches including new methodologies, new tasks, and new measurements have been developed since the publication of the aforementioned surveys. A significant part of the recent progress is mainly due to the rise of new challenges such as the consideration of context dynamicity or the emergence of a crucial need in the area to tackle well-known issues such as privacy-preserving data. Therefore, a new systematic review of the state of the art is needed.

We present several contributions in this survey. First, we perform a detailed and thorough investigation of the state of the art of system-based as well as user-based contextual IR evaluation methodologies and related levels of measurements. This study includes a comparative analysis of those methodologies and emphasizes the impact of each context form on the evaluation design. Second, we review the recent progress in the area to tackle two main challenges: context dynamicity and data privacy. For each challenge, we provide a detailed discussion of the different aspects of the problem and its implications on context-based evaluation. We also provide a list of techniques and methodologies that have been applied to address these challenges. Finally, we provide a list of promising research directions in the area.

1.4 Outline

The organization of the article is summarized as follows:

- In Section 2, we provide a synthesized and stratified review of the different evaluation methodologies in contextual IR that fall into two main classes: system-based methodologies and user-based methodologies. For each evaluation methodology, we present an overview of the related context data, tasks, topics, and relevance assessments. We also highlight their strengths and limitations and provide a compiled description to give insights on their appropriate use.
- In Section 3, we review the metrics used within those methodologies based on two main evaluation objectives: estimating the context accuracy and measuring the search effectiveness.
- In Section 4, we discuss recent trends in contextual IR evaluation. More precisely, we focus on the impact of context dynamicity on the evaluation design. We first highlight the emerging underlying challenges and then report an overview of the related evaluation methods and measures.
- In Section 5, we address the impact of context-based evaluation on privacy-preserving data. Therefore, we review major contextual IR applications with the related privacy risks, and

the techniques used for measuring such risks including differential privacy. Next, we explore the issue of evaluating task performance within private data.
—In Section 6, we discuss promising research directions.

2 MAJOR CONTEXTUAL IR EVALUATION FRAMEWORKS

Here we compile, criticize, and compare between representative evaluation methods of the literature in contextual IR areas according to two main classes of evaluation settings: system-based evaluation settings and user-based evaluation settings.

2.1 System-Based Evaluation Settings

This category of evaluation settings usually consists of a test bed comprising use cases and a fixed set of hypothetical context situations with corresponding relevance assessments. This category includes simulated context-based studies, log-based studies, and interface simulation studies. According to the nested model of context stratification proposed by Ingwersen and Jarvelin [63] shown in Figure 1, interface simulation studies mainly focus on dimension 3 (interaction context), simulated context-based studies, and log-based studies focus on dimension 4 (individual context) without taking interaction/dynamic context into account and the effect of a multi-stage search scenario on the evaluation design.

2.1.1 Simulated Context-Based Studies. A simulated context-based evaluation strategy [14, 28, 108, 128] is a typical system-based evaluation strategy relying on several criteria: users are abstracted, topics are previously created, and relevance assessments are provided by assessors—not by the participants involved in the search task. This class of evaluation studies includes evaluation tracks issued from major evaluation campaigns such as TREC, CLEF, and NTCIR.

2.1.1.1 Official Evaluation Tracks. Table 1 presents the major official evaluation tracks designed for static context-based evaluation.

TREC Contextual Suggestion track [36][5] started in 2012 and deals with complex information needs that depend on context and user interests. The main objective of the track is to recommend interesting venues and activities to users according to their profile. In TREC Contextual Suggestion track, the impact of integrating the user profile and the user location on IR evaluation design is reflected in a relevance judgments process that is completed to achieve the profile formation on one hand and the ground truth for evaluating the search results effectiveness with respect to user profile and location at a time on the other hand. To build the user profiles, assessors were asked to give two five-point ratings for each attraction, one for how interesting the attraction seemed to the assessor based on its description and one for how interesting the attraction seemed to the assessor based on its website. To build the ground truth, since the context is composed of user profile and user location, judgment was split up into two tasks: profile relevance and geographical relevance. For profile relevance, each user represented by a specific profile will rate each suggestion as −1 (negative preference), 0 (neutral), 1 (positive preference); for the geographic relevance, NIST assessors rate the suitability of each suggestion according to user location as 2 (appropriate), 1 (marginally appropriate), 0 (not appropriate). Context-profile pairs were also judged where assessors gave ratings for the attraction descriptions and websites of the top-five ranked suggestions for each run for their profile and one, two, or three randomly chosen contexts. In addition to the basic evaluation measures (P@5; Mean reciprocal rank (MRR)), a novel context-oriented measure was considered to evaluate the system effectiveness, namely Time-Biased Gain (TBG) detailed in Section 3.

[5] http://sites.google.com/site/treccontext/.

Table 1. Summary of Official Evaluation Tracks for Contextual IR Evaluation

Track	*Context form*	*Impact of context on the evaluation design*
TREC Contextual Suggestion track	Context includes two main components: (1) User location: it corresponds to a particular city location, described with longitude and latitude parameters. (2) User profile: a profile is a pair <user, suggestion> where the preference of the user for the given suggestion is assessed.	• Graded relevance judgments: (1) profile formation; (2) ground truth formation split into: (a) profile relevance, (b) geographical relevance, (c) context-profile relevance. • Context-oriented evaluation measure: Time-Biased Gain (TBG)
TREC Microblog track	TREC 2011 through TREC 2014: query time which represents the timestamp of the query; TREC 2015: 50 interest profiles, a profile is a combination of a narrative and a few (<10) sample tweets relevant to that narrative.	• Time-sensitive topic creation. • Graded relevance judgments. • Cluster-based tweet presentation for better judgment consistency. • Context-oriented evaluation measure: expected latency-discounted gain (ELG).
NTCIR GeoTime 2010–11	Geographic and temporal query constraints	• Topic development: geographic and temporal sensitive topics. • Graded relevance judgments.
GeoCLEF track 2005–2008	Geographic query constraints	• Topic development: geo-sensitive topics in a cross-language environment. • Relevance assessment complexity depending on the document length, language, and content.

TREC Microblog track [91][6] launched in 2011 and continued until 2015. The main task in TREC 2011 through TREC 2014 consists of a real-time ad hoc search task where context data includes the query time that represents the timestamp of the query in a human and machine readable ISO standard form. In TREC 2015, the main task is a real time filtering task where two sub-tasks are defined: Push notifications on a mobile phone and periodic email digest. For the push notification sub-task, the goal was to explore technologies for monitoring a stream of social media posts where the system has to recommend interesting content to a user based on the interest profile. In this task, 50 interest profiles were developed and assessed via the standard pooling methodology. In TREC Microblog track, the impact of integrating the query time on the evaluation design is reflected in the topic creation and relevance judgment process to ensure consistency. Indeed, topics are developed by NIST to represent an information need at a specific point in time where systems must rank tweets that are relevant to the user's information need and that are published prior to and including the query time defined by the topic. Tweets were judged on a three-way scale of "not relevant," "relevant," and "highly relevant." Although a standard pooling methodology was adopted, the main change in the process of providing relevance judgments was to cluster tweets so that textually similar tweets are presented to assessors in close proximity to enhance

[6] http://trec.nist.gov/data/microblog2015.html.

judgment consistency. In addition to basic evaluation measures (mean average precision (MAP), R-precision, and precision at rank 30, NDCG@K, normalized cumulative gain (nCG)), a context-oriented evaluation measure adopted as an official measure is the expected latency-discounted gain (ELG) detailed in Section 3.

In NTCIR GeoTime task [48], the focus is on search with geographic and temporal constraints. Geo-temporal IR is concerned with the retrieval of thematically, temporally, and geographically relevant information resources in response to a query of the form $< theme, what, where >$. The challenges of integrating time and location in IR evaluation design are related with topic development and relevance assessments. Topics were developed based on Wikipedia articles so that they include time and space aspects. Regarding the relevance judgments, judgment was graded in that a document could be assessed as "fully relevant" if it contained text that answered both the "when" and "where" aspects of the topic. The document was assessed as "partially relevant where" if it answered the geographic aspect of the topic and "partially relevant when" if it answered the temporal aspect of the topic. The official evaluation measures are Average Precision (AP), Q, and normalized Discounted Cumulative Gain (nDCG).

In GeoCLEF tracks (2005–2008) [47], the focus is to test and evaluate cross-language geographic IR (GIR), which consists of retrieval for topics with a geographic specification. Context is represented with geo-references within the query. The challenge for evaluation design relates to topic generation where the topic set should be geographically challenging, have relevant documents in other collections of different languages, and should not favor keyword-based approaches instead of profound geographic reasoning. This was achieved by including several difficulties into the topics such as ambiguity, vague geographic regions, cross-lingual issues, and so on. Another challenge at the evaluation level was related to the relevance assessment process, which is highly dependent of the document length, language, and content. For instance, short news agency articles were easier to judge with their headlines than long and essay-like stories covering multiple events without a specific narrow focus.

2.1.1.2 Track Campaign-Like-Based Evaluation. Evaluation settings under this evaluation strategy are proposed externally to official campaigns but reuse similar evaluation design and/or the related resources.

- *Context*: user contexts are generally simulated by hypothetic context situations [104, 128]. The user is simulated in Reference [28] by considering that a plausible interest could be represented by a TREC domain of Disk1&2 in which automatically generated subqueries define a search session. In Reference [128], a set of signal documents, called the profile set, classified under the concept/user interest from which the tested query is formulated, allows modelling the user profile/context used to personalize the search results for the query at hand. Context is simulated in References [13, 128] by considering that an ODP[7] concept represents a potential user interest. In Reference [104] click-through data are also hypothesized and used as additional part of the simulated contexts;
- *Topics and tasks*: topics could be predefined [104], represent TREC topics as in Reference [28], or automatically generated using the top terms of the DMOZ (directory.mozilla) ontology [128] or the top terms of DMOZ ontology selected in a location branch of the DMOZ (US cities) in Reference [14]; DMOZ is an open-content directory of WWW links maintained and known as the Open Directory Project (ODP).

[7] Open Directory Project.

— *Relevance assessments*: heavily depend on the query process generation, TREC relevance assessments [28], documents under the concept representing the context [14, 128], or expert human assessment [104].

Since users are abstracted and real contexts are not involved, simulated context-based studies allow obtaining an overview of the system performance with slight explanations of the impact of context on the system performance. Moreover, they generally require creating artificial information needs, which is a hard task [43]. Last but not least, relevance assessments are generally undertaken in hypothetic contextless settings [28, 104, 128], which highly impact their real value.

2.1.2 Log-Based Studies. A transaction log is an electronic record of interactions that results from the user interactions with a system [70]. Log analysis, which gives more specifically a realistic and longitudinal picture of user's searches on search engines, is widely exploited in contextual IR evaluation [1, 8, 86, 145, 155, 162].

— *Context*: various contextual features in users' log data are used for evaluation purposes, such as click-through data, browsing features, queries and related results, top search results [1, 134, 137, 149], GPS and mobile cellular data connection [8, 162], and so on;
— *Topics and tasks*: topics are generally randomly extracted from log data [1, 86], based on later queries in the sessions as in Reference [137]. They also refer to specific problems preprocessed within case studies [149];
— *Relevance assessments*: relevance data is not entailed in log files. In most cases, the top search results of the testing queries are assessed by assessors [1, 137] or relevance judgments can be inferred through behavioral evidences analysis such as in Reference [134] where document pair preferences are assumed on the basis of "download," "not download" a user's actions. In case studies, relevance assessments are provided by experts such as in Reference [149], where one relevant document is assigned to each case (query) being the problem's right answer.

The main advantage of log-based studies is that they provide a method for collecting data from a great number of users, namely queries and click-through information in a reasonable and nonintrusive manner. However, they present the following limitations:

— It is unclear for what task purpose the queries were written [77].
— The automatic identification of session boundaries is a hard problem, which makes the query sampling a time-consuming task. Session identification has been studied widely in the literature, where it has been clearly stated that identifying session boundaries is challenging [59] and should be complemented with manual intervention;
— Since log data does not contain explicit relevance assessments and the number of queries and users is uncertain, then a pool of queries should be first defined and an evaluation scenario should involve participants to collect relevance judgments.
— The log usage data has a limited availability.
— A user's log data may be insufficient within specific context applications such as mobile IR, where additional context features like individuals' identities, demographic data, nearby persons' movements, and external sensors' data are required.

2.1.3 Interface Simulation-Based Studies. According to the context stratification model proposed by Ingwersen and Jarvelin [63], interface simulation-based studies focus on dimension 3 (interaction context) of Figure 1. Interface simulation studies refer to the use of searcher simulations inferred through searcher interactions at the results interface for evaluating implicit feedback models [151, 153]. Specific search interfaces were developed to more engage the searchers in the

examination of search results than traditional styles of result presentation adopted by commercial search systems. More precisely, the developed search interfaces are designed to implement an aspect of polyrepresentation theory [65], and display multiple representations of the retrieved documents simultaneously in the results interface. Documents are represented at the interface by the document title and query-biased summary of the document: a list of top-ranking sentences (TRS) extracted from the top 30 documents retrieved, scored in relation to the query; a sentence in the document summary, and each summary sentence in the context it occurs in the document (i.e., with the preceding and following sentence).

- *Context*: Context is represented with interactive relevance paths that are composed when searchers travel between different representations of the same document. The paths provide searchers with progressively more information from the best documents to help them choose new query terms and select what new information to view. These paths are simulated by being extracted just from relevant documents, non-relevant documents, or from a mixture of relevant and non-relevant documents, depending on the simulated search scenario [153];
- *Topics and tasks*: TREC topics 101–150 were used and the query was taken from the short title field of the TREC topic description. For each of the 50 TREC topics used as queries, the simulation retrieved the top 30 results and so simulate the relevance paths;
- *Relevance assessments*: TREC relevance assessments were used. The use of TREC relevance assessments was still valid in this evaluation methodology as they were assumed to be independent of the interfaces and the systems that led to the documents being assessed.

The main advantage of interface simulation studies is that they enable more control over the experimental conditions in predetermined retrieval scenarios without the need for costly human experimentation. Additionally, it helps system designers in identifying the strengths and weaknesses of the system by allowing them to eliminate interactions that could cause problems, and the best solutions could be further tested with real test persons in the laboratory. One disadvantage of this type of evaluation study is that it assumes that all searchers in a scenario would interact in the same way.

2.2 User-Based Evaluation Settings

This category of evaluation settings includes user studies [94, 114, 115, 124, 139, 163] and diary studies [11, 23, 95]. According to the context stratification model of Ingwersen and Jarvelin [63] shown in Figure 1, user-based evaluation studies focus on dimension 4 without considering interaction/dynamic context.

2.2.1 User Studies. The basic underlying idea is to test and evaluate the system effectiveness along with the natural user interaction with the system [10]. A system that supports IIR-based user studies on the web has been proposed in Reference [148] where it includes web-event logging and usability/eye-tracking functionality. Also, user studies have been undertaken for the purpose of designing effective interfaces. The user study in Reference [92] has shown that users prefer paragraph-sized chunks of text over just an exact phrase as the answer to their questions in the context of using a question-answering system. In Reference [113], the focus of the user study was on the emotional impact of search tasks upon the searcher and how could it lead to improved experimental design. Another user study [3] was focused on exploring the relationship between relevance criteria use and human eye movements. In Reference [80], the purpose of the user study was to find the best snippet size needed in mobile web search.

In the literature of contextual IR evaluation, a wide range of user studies have been undertaken within several evaluations settings and protocols [38, 45, 108, 115, 124, 152]. These settings include the following:

- —User studies where users are involved mainly in controlled tasks and relevance assessment is considered as the basic and only interaction with the system (users as assessors);
- —Users studies where users are involved in controlled search tasks with predefined sets of queries and interact with the search engine through formulating a small number of queries to simulate a search session, examining the search results, terminating sessions, and assessing document relevance (users as evaluation participants);
- —User studies where users deal with simulated/predefined tasks [10] and interact "naturally" with search engines under laboratory conditions. Users are to formulate queries and interact naturally with the system to complete a search task and are also involved in the relevance assessment process.

2.2.1.1 User Studies: Controlled Tasks, Minimal Interaction with the System. This setting is the most basic way to extend simulated context-based studies [38, 45, 108]. By involving users as assessors, this allows us to evaluate the extent to which multidimensional relevance assessments provided by users can enhance the context and consequently evaluate the quality of the search engine outcomes.

- —*Context*: data context is limited to user's relevance feedback on the top search results.
- —*Topics and tasks*: topics are not user-generated. They are automatically created in a research step [108] or collected from web search-engine logs [45], thus ensuring reliable assessment.
- —*Relevance assessment*: users assess the top search results [45, 108] and/or assess relevant concepts related to the queries at hand [38].

This type of evaluation studies involves users to perform one main activity: assess document relevancy. The main advantage of this category of user studies is that relevance assessments provided by test persons/users are comparable to ground truth assessments such as TREC if the evaluation is performed with short term interaction (one to two retrieval runs). In other terms, the assessor/test person provides relevance feedback for only one retrieval run where no learning effects by test persons can influence the experiment [64]. Although the minimal interaction with the system under evaluation is useful, it is clearly limited in realism. Indeed, if a system is designed to support contextual search, then rich user interactions with the system during the evaluation process are important to allow naturalistic contextual search scenarios to happen and put more context features in play.

2.2.1.2 User Studies: Controlled Tasks, Rich Interactions with the System. This setting makes use of prior well-built test collections to launch the evaluation studies with real users where rich user interaction with the system at multiple stages of the evaluation process are involved [119, 123, 125]. Well-defined topics such as TREC topics are used as simulated work tasks [9] to generate queries.

- —*Context*: rich user interactions with the system such as click-through data, queries, and related search results [119, 123, 125].
- —*Topics and tasks*: topics in TREC 2004, 2003 TeraByte track [123] and TREC AP [125] are used as task descriptions. The users may generate a small number of queries within the same topic to simulate a search session and then allow a user's interaction with the system within a search session [125].

— *Relevance assessments*: In Reference [125], TREC relevance assessments are used, while, in Reference [123], users assess relevance of the top mixed rankings issued from web search engines. In Reference [123], users are provided with a predefined set of topic categories and related search tasks. During the relevance assessment process, in case the simulated session involves several interactions with the system, then TREC assessments cannot be applied and compared to test persons' assessments due to the learning effects by test persons on the experiment as stated in Reference [64].

The use of well-established experimental resources such as TREC in an evaluation study is definitely interesting: topics are well defined and relevance assessments are available. However, if the latter are used [125], this induces the assumption of relevance generalizability, which is far from being true [144]; if relevance assessments are ignored and new ones are provided by users involved in the evaluation [123], recall-oriented measures are not applicable but this would not be a limitation for high-precision-oriented evaluation studies. One can mismatch between TREC relevance judgments and real users' judgments by applying interactive recall and precision [141].

2.2.1.3 User Studies: Uncontrolled Tasks, Rich Interactions with the System. One of the most common ways to undertake a user study is through uncontrolled tasks, enabling naturalistic user interactions with systems [1, 94, 103, 104, 114]. Users are asked to complete main tasks and the study is designed to gather information context, either internal or external, to obtain more insights about the user's behavior for prediction purposes and, consequently, to enhance the search outcomes.

— *Context*: personal user information such as desktop index [1], queries, top results, and feedback on automatic annotations [94] or browsing actions on documents [104].
— *Topics and tasks*: self-generated queries from the diary log of users [1, 114], general [1] prior queries [104], free queries that seemingly interest the users [103].
— *Relevance assessment*: a user's relevance assessment of the top search results [1, 103] and annotations [104].

Evaluations based on user studies conducted using uncontrolled tasks with rich user interactions with the system are the most comprehensive type of user-centered evaluation. However, they are time consuming and researchers must be careful to what extent they can generalize their findings regarding several limits: topic familiarity differences between users, accurate and reliable interpretation of search behavior, appropriateness of queries and tasks.

2.2.2 Diary Studies. A diary study consists of a representative sample of test participants recording information about their daily lives in real situations for a given period of time. The data gathered from the study, structured through diary entries, can vary from simple format (date, time, description) to complex ones issued from questionnaires. Diary study entries can then be exploited to achieve different goals, depending on the nature of the data and the objectives of the study. Diary studies are presented in an early work by Reference [117] as a workplace-oriented tool to guide laboratory efforts in the human computer interaction (HCI) field and used later widely (in HCI field) for several evaluations tasks [15, 138]. In the contextual IR area, few evaluation studies are based on typical diary studies. Beyond analyzing mobile information needs [21, 24], diary studies were undertaken [11] to measure retrieval accuracy within mobile environments where location, time, and user interests are the key context attributes. In Reference [21], a diary study was conducted where the data collection is mainly a lifelog data that includes computer activities, mobile phone activities, photos, geo-location, Bluetooth, biometrics, and tweets. The purpose of the study was to investigate episodic context features for refinding tasks. A review of the

Table 2. Participants' Roles in the Evaluation Settings

	Evaluation focus		
	Participants create topics	Participants interact with the search engine	Participants assess relevance
Simulated context-based studies	N	N	N
Log-based studies	Y	Y	N
Evaluation based on user studies	Y/N	Y	Y
Evaluation based on diary studies	Y	Y	Y

different types of search contexts for the purpose of designing lifelogging systems is presented in Reference [95].

— *Context*: participants are asked to label their information needs with a narrative description, the explicit location, and time they issued their queries. The participants' interests are manually specified by participants themselves or automatically learned from their manual relevance assessments of the top returned documents in response to their past queries [11, 21].
— *Topics and tasks*: free topics expressed by the participants *in situ* within their current task (travelling, leisure, working, and so on) [11, 21].
— *Relevance assessments*: each participant who submitted a query (in the diary study), is asked to judge whether a document from the set of top N retrieved results in response to his/her query was relevant or not according to its underlying context. Relevance judgments are given using a three-level relevance scale: relevant, partially relevant, or not relevant.

Diary studies allow non-intrusive evaluation, in comparison to user studies or simulated context-based studies. Moreover, since diary study frameworks involve real users in situ within real contexts, they allow gathering various context dimensions in the testing data. However, they require information recording, which could be tedious in the case of complex entries. Another weakness of diary studies is the lack of methodological guidelines to achieve reliable outcomes [154], even though some advice is given in Reference [117].

2.3 Synthesis

We synthetize the four evaluation settings in contextual IR area described above, according to our subjective view through several dimensions that heavily impact the overall evaluation methodology: the role of participants, viewed as the core context in the evaluation settings, the evaluation mode, duration, and type. The role of the participants mainly consists of (1) creating testing topics, (2) interacting with the search engine, and (3) assessing the relevance of the search engine outcomes. Table 2 presents the role of the participants in each of the evaluation settings where Y denotes that the role is present and N denotes that the role is absent.

— Simulated context-based evaluation represents one extreme in Table 2 (N, N, N) regarding the user's role in the evaluation settings. No interaction is acquired with real users.
— In log-based studies, users create topics and interact with the search engine while achieving the search tasks. However, their relevance assessments of documents are not logged.

Table 3. Evaluation Setting Specifications

	Evaluation focus		
	Mode	Duration	Type
Simulated context-based studies	Laboratory	Narrow time	Batch
Log-based studies	Naturalistic	Longitudinal	Batch
User studies	Laboratory/Naturalistic	Longitudinal/Narrow time	Live
Diary studies	Naturalistic	Longitudinal	Live

Table 4. Experiment Features

	Repetitive	Large scale	Controlled task
Simulated context-based studies	Y	Y	Y
Log-based studies	Y	Y	Y
User studies	N	N	Y/N
Diary studies	N	N	N

—In user studies, participants may create the topics, interact with the search engine to achieve the search task, and also provide their relevance feedback on the result set.
—Diary studies represent another extreme of user-centered evaluation studies (Y, Y, Y) in Table 2 regarding the user's role in the evaluation settings. Participants are involved in recording diary studies' concurrent logs and also in the relevance assessment process.

From another point of view, we synthesize the four above evaluation settings (simulated context-based studies, log-based studies, user studies, and diary studies) using three main evaluation features as shown in Table 3: evaluation mode, duration, and type.

—*Mode*: the evaluation mode determines the conditions of the place where the evaluation happens, either controlled in a laboratory evaluation mode or within real life conditions in a naturalistic mode.
—*Duration*: we mainly distinguish longitudinal experiments that require occurring within a long period of time and narrow time experiments that take a relatively short slot of time. Longitudinal evaluation is required to integrate reliable experiment variables in contextual IR evaluation such as user's interests, behavior, task achievement, and local browsing traces. Evaluation measurements are usually taken within regular intervals of time.
—*Type*: evaluation type could be either batch or in-live. Batch evaluation refers to an evaluation setting where search tasks have already occurred and traced in files such as in track-campaign-like evaluation. In contrast, live evaluation involves users achieving search tasks in real time.

Table 4 summarizes the experiments' features of each class of evaluation methodologies in terms of repeatability, large scale, controlled, or uncontrolled tasks.

3 LEVELS OF MEASUREMENTS

In this section, we focus on the measures used for evaluating contextual IR techniques and models' effectiveness. The state of the art in contextual IR reveals that there are two classes of measures for evaluating the retrieval effectiveness. The first class concerns the measurement of context accuracy with respect to real-life contexts involved in the evaluation through users and their corresponding

.

Table 5. Review of Measures for Context Accuracy Evaluation

Context form	*Metric and Description*	*Formula*
Cognitive context: a ranked set of ODP concepts	$P@X$ [30]: Measures the proportion of top relevant context items	$P@X = \frac{Rel_X}{X}$, Rel_X: Number of relevant context items among top X selected items
Cognitive context: Topic preference vector T of m topics with the user's degree of interest in each topic T= [T(1), …, T(m)]	Relative error [114]: Measures the difference between the estimated context vector and the reference context	$E(t_e) = \frac{\lvert T_e - T \rvert}{\lvert T \rvert}$, T_e: learned topic preference vector, T: actual topic preference vector
Cognitive context: Instance of the ODP ontology	Convergence [128]: Measures the relative stability of context scores within an iterative process of automatic scoring of context relevance. It is commonly computed using variance measure.	$Var(S) = \sqrt{\frac{(S-\overline{S})^2}{n}}$, S: context score, $\overline{S}$: average score, n: number of scores
Cognitive context: User preference vector P of n document pairs $P = p_1, p_2, \ldots, p_n$	Preference accuracy [134]: Measures the percentage of correctly ranked document pairs	$accuracy(U) = \frac{\lVert \{p_j \lvert Score(p_j, U) > 0 \rvert\} \rVert}{\lVert \{p_j \in P\} \rVert}$, p: page, P, set of pages, U: user model, Score(p_j, U): topical similarity between page P and user U
Cognitive context: A weighted category-term matrix where categories are issued from the top-three levels of the ODP ontology	Rank accuracy [94]: Measures the relative rank of estimated contextual rank	$accuracy(U) = \frac{\left(\sum \frac{1}{1+rank_{ei-ideal\_rank_i}}\right)}{n}$, n: number of related categories
Cognitive context: A set of ODP concepts	Accuracy [38]: Measures the proportion of correctly classified contextual items	$accuracy(U) = \frac{Nd_c}{Nd}$, Nd_c: number of correctly classified documents into ODP concepts.

internal and/or external contextual features. The second class of measures is performance-oriented and used to measure the relevancy of the search outcomes regarding the search task and the context in which it occurs.

3.1 Context Accuracy Evaluation Measures

Context accuracy measures allow comparing automatically created context and actual-context representations. Thus, the evaluation studies that report context accuracy measures rely on manual or hypothetical relevance assessment of the automatically created context. A review of context accuracy measures used in previous work is presented in Table 5. It is worth it to mention that there are no standard measures for context evaluation, thus the suitability and interpretation of any measure is highly dependent on the context representation.

3.2 Contextual Retrieval Effectiveness Evaluation

The evaluation measures used to compute the overall performance of contextual search highly depend on three main factors: the evaluation scenario whether it is contextless or not, the contextualization methodology whether it is click-based or profile-based, and the type of involved context. For contextless evaluation scenarios, the evaluation measures that are used to compute the overall performance of contextual search are commonly used to evaluate basic IR. Indeed, the effectiveness measurement requires baseline evaluation scenarios that are basically represented through contextless scenarios [1, 30, 104, 134, 137] involving topical ranking such as BM25 model and vectorial model [1, 134, 137], relevance feedback model [1], BM25 with relevance feedback re-ranking [1, 30], web search ranking [12, 29, 133], or close contextual ranking models [2, 30, 135]. In terms of the contextualization methodology, we distinguish between profile-based personalization and click-based personalization. Profile-based personalization aims at exploiting a user profile defined with topics of interest to improve the retrieval effectiveness. Click-based personalization aims at exploiting the previous user click-through information to enhance the rank of relevant documents in the search results and, consequently, the retrieval effectiveness.

Most existing works do not adapt evaluation measures with respect to the type of context. However, recent work showed that context dynamicity gave rise to specific measures as detailed in Section 4. In the following, we categorize the set of evaluation measures used to evaluate contextual retrieval effectiveness systems as follows: binary relevance measures, graded relevance measures, correlation-based measures, and context-oriented measures. These measures are presented in Table 6 along with the context of use, description, and mathematical form for each measure.

— *Binary relevance-based measures:* These are basic IR evaluation measures, including recall and precision-like measures such as MAP, P@X, and R@X. Most of the existing contextual IR approaches that use these measures rely on profile-based personalization and adopt contexless evaluation scenarios. The main feature of these measures is that they are used for evaluating overall topical ranking performance and are insensitive to the level of relevance of documents. These measures are generally used in both system-based and user-based evaluation methodologies.
— *Graded relevance-based measures:* These evaluation measures are designed for situations of non-binary notions of relevance such as discounted cumulated gain measures, namely DCG and nDCG. These measures are sensitive to both the rank of relevant documents and level of relevance and are mainly used to evaluate the relevance over some number k of top search results. Graded relevance-based measures are commonly used in simulated context-based studies, namely in official evaluation campaigns such as TREC contextual suggestion, TREC microblog search, CLEF social book search, TREC session track, TREC task track. They are also used commonly in click-based personalization approaches where the evaluation is conducted as track-like evaluation studies, log-based studies, user studies, or diary studies.
— *Correlation-based measures:* These measures allow estimating the correlation between context-based ranking and ideal ranking. Most of contextual IR approaches that use these measures rely on click-based personalization. These measures include page gain ratio, mean click position, average and weighted average rank. These measures are commonly used in click-based personalization approaches where click-through information is used to promote relevant results to higher ranks.
— *Context-oriented measures:* Context-oriented evaluation measures shift the focus of evaluation from topical to context-driven metrics. Context-oriented metrics include time-biased gain and expected latency gain that were used as official evaluation metrics in the TREC contextual suggestion track and Trec Microblog track, respectively.

Table 6. Review of Measures for Contextual Retrieval Effectiveness Evaluation

Context form	*Metric and Description*	*Formula*
	Binary relevance measures	
Cognitive context: User profile [30, 111]; Clickthrough data [1, 124]	Mean Average Precision (MAP): average precision over queries	MAP = $\frac{1}{\|Q\|}\sum_{q_j}\frac{1}{m_j}P(R_j)$, Q: set of queries, m_j: recall level, P(m_j): Precision at m_j recall level
Cognitive context: User profile [28, 128]; Clickthrough data [1, 124, 137]	P@X: Precision at top X documents computes the precision after X retrieved documents	P@X = $\frac{Rel_X}{X}$, Rel_X: number of relevant documents among top X returned documents
Cognitive context: User profile [28, 128]	R@X: Recall at top X documents computes the recall after X retrieved documents	R@X = $\frac{Rel_X}{R}$, Rel_X: number of relevant documents among top X returned documents, R: total number of relevant documents.
	Graded relevance measures	
Cognitive context: User interests built from search history [139]	Discounted Cumulated Gain (DCG): discounts the value of relevance documents according to their rank	$DCG(i) = G(i) when i < b, DCG(i-1) + \frac{G(i)}{log_b i}$ i: document rank, G(i): gain value at rank i, b: log base
Cognitive context: Clickthrough data [1, 124, 137]	Normalized DCG (nDCG): normalized DCG measure according to an ideal rank	$NDCG(i) = \frac{DCG(i)}{IDCG(i)}$, $IDCG(i)$: ideal rank of document at rank i
	Correlation based measures	
Cognitive context: User profile [114]; clickthrough data [39, 133]; Spatio-temporal context: User location, query time, weather and user activity [86];	Average Rank: computes the average rank of relevant results, Weighted average rank (WAvgRank): computes the weighted average rank issued from the reference ranking	$WavgRank(u,q) = \sum_{p\in S} p(q\|u) * R(p)$, S: set of pages user U selected for query $R(p)$: rank of page by the reference rank, p(q|u): probability that user U issues query q, $WavgRank(p,q) = \frac{1}{R(q)}\sum_{p\in R(q)} Rank(p)$, p: page, $R(q)$: relevant pages for query q, $Rank(p)$: Rank of page p in the outcome results' pages
Cognitive context: User profile [118]	Page gain ratio: Computes the relative gain induced from the number of browsing pages between the contextual retrieval reference outcome and ideal retrieval outcome	$RatioG = 1 - \frac{G_R}{G_{opt}}$, G_R: gain of reference rank, G_{opt}: optimal gain
Cognitive context: Clickthrough data [155]	Mean click position: Measures whether a re-ranking method promotes the search results which are clicked on by the users to higher positions	$MCP = \frac{\sum_i \sum_{u\in U_c^{(i)}} R(u)}{\sum_i U_c^{(i)}}$, $R(u)$ is the rank of u in a ranked list, $U_c^{(i)}$ is the set of the clicked URLs for the last query in the i^{th} test case.

(Continued)

Table 6. Continued

Context form	*Metric and Description*	*Formula*
	Context based measures	
Cognitive and geographic user context [35]	Time-Biased Gain (TBG) [131]: This measure estimates the amount of time it takes for the user to process a given document by considering aspects such as the document length, duplicates and summaries that influence the time required.	$TBG = \sum_{k=1}^{5} D(T(k))A(k)(1-\theta)(\sum_{j=1}^{k-1} Z(j))$ • D is a decay function. • T(k) is how long it took the user to reach rank k, calculated using the following two rules: —The user reads every description which takes time T_{desc}. —If the description judgment is 2 or above then the user reads the document which takes time T_{doc}. • A(k) is 1 if the user gives a judgment of 2 or above to the description and 3 or above to the document, otherwise it is 0. • Z(k) is 1 if the user gives a judgment of 1 or below to either the description or the document, otherwise it is 0. The four parameters for this metric are taken from Dean-Hall et al. [34]:$\theta = 0.5$, $T_{desc} = 7.45s$, and $T_{doc} = 8.49s$, and the half-life for the decay function $H = 224$.
Temporal context: query time in TREC microblog track [93]	Expected Latency Discounted Gain (ELG): this measure is used to accommodate the integration of query time into the mobile push notification of the TREC Microblog track.	$ELG = \frac{1}{N} \sum G(t)$ where where N is the number of tweets returned and G(t) is the gain of each tweet: • Not relevant tweets receive a gain of 0. • Relevant tweets receive a gain of 0.5. • Highly-relevant tweets receive a gain of 1.0. The measure considers a latency penalty= MAX(0, (100 - delay)/100) applied to all tweets where the delay is the time elapsed (in minutes, rounded down) between the tweet creation time and the putative time the tweet is delivered.

4 HOW CONTEXT DYNAMICITY MADE IR EVALUATION MORE CHALLENGING?

4.1 Basic Notion: Dynamic Context

Usually, it is assumed that the context surrounding a search session from which we extract evidence to evaluate contextual search effectiveness is fixed. However, the recent research literature in the IR area highlights that the range of search tasks grows with an increasing number of complex nature of search tasks [6]. The latter includes exploratory searches [150] where a user seeks to learn more about a topic through iterative and multitactical search processes, and multi-tasking searches [56] where the information needs to involve multiple and interrelated sub-tasks. Both types of search tasks span over multiple search sessions based on a dynamic search scenario embedding an open-ended context. It turns out that the search context is not static but rather evolves during the search. Thus evaluation is even more challenging since (1) the evaluation protocol might consider the context milestones and the coverage of the inherent sub-contexts to estimate the overall search performance and (2) beyond relevance of the system results, the evaluation is rather qualitative since it should provide insights on the usefulness of the system for users by helping them to accomplish the underlying complex search task. To the best of our knowledge, apart from theoretical investigations on the evaluation of IR systems from a task perspective to provide a framework for designing evaluation studies [72], so far there are no standard evaluation settings for dynamic context-based IR, and the few studies that tackled this issue specifically covered the search task as the core element of context. In this section, we first examine the additional requirements of an IR evaluation framework given a complex task and then review the recent specific evaluation methodologies and measures.

4.2 Impact of Context Dynamicity on Evaluation Design

4.2.1 Evaluating Context Coverage. According to the viewpoint theory [72], the dynamicity of contexts suggests the need of developing evaluation indicators that reflect the longitudinal and coherence aspect of the context built alongside the search task. Since evaluation requires comparisons against standards, there is a need to provide (1) baseline techniques for generating the expected sub-tasks given the original query that triggers the complex task, (2) qualitative human assessment of the automatically generated sub-contexts to build the ground truth, and (3) measures that evaluate the coherence and coverage of the dynamic context. For this purpose, both quantitative and qualitative evaluation are required. Baseline techniques for sub-task generation include topical clustering techniques such as LDA) [99] or adaptive- structural clustering [146], graphs linking task nodes based, for instance, on random walks [56, 57]. The ground truth is built based on human assessment according to diverse criteria judged based on a three-point scale such as: relatedness, interestingness, diversity, completeness, or coherence and coverage [56, 57]. The metrics used for context coherence and validity evaluation rely on recall-oriented measures based on the best alignment of the automatically expected sub-tasks with the ground truth [56, 99, 146].

4.2.2 Evaluating Task Completion. While the standard context-based evaluation has focused on assessing the performance of retrieval systems in returning the best results that match the user and the search context, the dynamicity of context puts the evaluation focus on the extent to which systems help the user in achieving a given complex search task. Thus, appropriate evaluation methodologies and measures are required to support success indicators in reaching this new goal. An overview of the major evaluation metrics used for this purpose is provided in Table 7.

An extension of the *nDCG* metric to multi-query sessions, called the *session nDCG* has been proposed by Reference [71]. The authors introduce a cost for reformulating a query as well as scanning down a ranked list by applying a penalty to those documents that appear at the bottom for later

Table 7. Review of Measures for Dynamic Context-Based Retrieval Effectiveness Evaluation

Context form	*Metric and Description*	*Formula*
User's search history based on sessions	*Session nDCG* [71]: extends the DCG measure by introducing the cost of reformulating a query	$DG(i) = \frac{2^{rel(i)-1}}{(log_b(i+(b-1)))}$, b: log base chosen to be 2, Additional discount is applied to documents retrieved for later reformulations. For rank i between 1 and k, there is no discount. For document at rank i that came after j^{th} reformulation, the discount is $sDG(i) = \frac{1}{(log_{bq}(j+bq-1))} DG(i)$, bq: log base chosen to be 2 session DCG use the sum over sDG(i): $sDCG(k) = \sum_{i=1}^{mk} \frac{2^{rel(i)-1}}{(log_{bq}(j+(bq-1))log_b(i+(b-1)))}$
User's search history based on sessions	*µERR* [20]: Measures the length of time that the user will take to reach a relevant document	$ERR = \sum_{r=1}^{n} \frac{1}{r} \prod_{i=1}^{r-1}(1-R_i)R_r$, Where R_i is the probability that a user finds a document relevant as a function of the editorial grade of that document calculated as follows: $R_i = R(g_i)$, where g_i is the grade of the i^{th} document, and R is a mapping from relevance grades to probability of relevance.
User's search history based on sessions	*Expected session measure esM* [74]: Extends the interpolated precision and recall by considering the relevant count of each possible browsing path of the user	$esM = \sum_{\omega \in \Omega} P(\omega) M_\omega$; $P(\omega) = P(r_i) \cdot P(ref \mid r_i)$, where Ω is the set of all possible paths that follow the user model, ref is the set of reformulations at different points of path w, r_i is the reformulation at point i.
User's search history based on sessions, user's search task	*Cube Test* [97]: measures the speed of reaching a relevant document with regard to the task structure (sub-task)	$Gain(Q, d_j) = \sum_i area_i height_{ij} keepFilling_i$, $KeepFilling_i$ is a function specifying the need of "document water"; $CT(Q, D) = \frac{Gain(Q,D)}{Time(D)}$.

query reformulations. *ERR* [20] is an extension of the classical reciprocal rank (*RR*) to the graded relevance of documents. Basically, it measures the length of time that the user will take to reach a relevant document. Unlike the *RR* metric, the *ERR* is based on a cascade model that considers the dependency between documents in a result page beyond their position. The *esM* [74] metric relies on a user model over the space of browsing paths built upon query reformulations, scrolling down and abandoning. For instance, in the model-free session model, the authors define session versions of interpolated precision and recall by dividing the relevant counts for each possible path through the results. Recently, a new metric, called the *Cube Test*, has been proposed by Reference [97]. The peculiarity of this metric is that it is based on a water-filling model that considers both the speed of gaining relevant information to answer the overall information need and the subtopic relevance.

Table 8. Summary of Evaluation Tracks Dealing with Dynamic Context-Based Search

Track	*Context form*	*Evaluation measure*
TREC Session	The user's search history which consists of (1) the current query q_t submitted at time t, (2) the set of past queries in the session, (3) the ranked list of URLS for each past query, and (4) the set of clicked URLs/snippets and the time spent by the user reading the corresponding webpage	nDCG@10, Session nDCG
TREC Dynamic Domain	The user's search task: A set of subtopics, each of which addresses one aspect of a main topic that is the key search target for one complete run of dynamic search. Passages from the relevant documents are also identified and assigned to the subtopics with a graded relevance rating	Cube Test [97] and μ-ERR [19]
TREC tasks	The user's search task: The implicit task the user attempts to achieve	Diversity metrics such as ERR-IA and $\alpha - NDCG$ [25]
CLEF interactive Social Book Search	The user's search task: An implicit multi-step or exploratory search task	Session length, number of submitted queries and number of books collected

It includes a function specifying whether more "document water" is needed for a subtopic, which depends on the current amount of document water in a subtopic cuboid.

4.3 Evaluation Methodologies

To the best of our knowledge, we can learn from the literature that most of evaluation methodologies used so far for evaluating dynamic-search-based scenarios are the traditional log-based studies described in Section 2 [56, 57, 99, 146] and specific system-based evaluation methodologies launched recently within official evaluation campaigns [52, 74, 157]. We report in Table 8 the major recent IR evaluation tracks, including scenarios of complex and longitudinal search tasks.

The TREC Session track [18][8] is the first one that specifically deals with search activities over sessions; a session is a sequence of query search activities related to the same information need. The main underlying goal is to test whether a document ranking is improved by considering a past user's queries and interactions (clicked URL webpages) over sessions. The Dynamic Domain track [156][9] is a young track that started in TREC 2015. This track focuses on domain-specific search algorithms that adapt to the dynamic information needs of professional users as they explore in complex domains. The main task is a dynamic search across multiple interesting domains where

[8] http://ir.cis.udel.edu/sessions.
[9] http://trec-dd.org/.

the search includes multiple runs of interactions and the participating systems are expected to dynamically adjust their systems based on the relevance judgments provided along the way. Systems should stop when they believe they have covered all the user's subtopics sufficiently. The subtopics are not known by the system in advance; systems must discover the subtopics from the user's responses.

The TREC Tasks track [160][10] is another young track that was started in 2015 as an intersection of the diversity and session tracks. This track puts the user's task at the center of the evaluation process by exploring system's understanding of tasks users aim to achieve and measures relevance of output documents according to evolving tasks in a given query.

The interactive track of CLEF Social Book Search [82][11] was started in 2014. The goal of the track is to investigate interactive multi-step or exploratory searches using book search systems that rely on different types of book metadata and social-generated content. The latter includes opinionated descriptions and user-supplied tags that provide users with new criteria of search. The organizers claim that the long-term objective of the tasks is to investigate user behavior through a range of user tasks and identify the influence of different types of metadata on the multi-step search process.

5 IMPACT OF CONTEXT-BASED EVALUATION ON PRIVACY-PRESERVING DATA

Within the objective of evaluating contextual IR, several contributions have focused on constructing datasets based on topical profiles of a user's short-term and long-term search history, a user's interactions with the system or surrounding users through social interactions, or a user's location history. However the availability and use of such datasets by third parties would raise user concerns from a privacy perspective. For instance, the AOL data released in 2006 is a well-known privacy incident.[12] AOL released 20 million search keywords and search history for over 650,000 users over a 3-month period intended for research purposes. However, *The New York Times* was able to de-identify users from the released dataset by cross referencing the data with phone book listings. Another well-known example of privacy concerns in datasets is Netflix, which released an anonymous dataset of user movie ratings used for evaluating recommendation tasks. Unfortunately, users in the dataset have been re-identified [105] by linkage with the Internet Movie Database (IMDb), leading to a lawsuit. More specifically, in the IR evaluation area, TREC Medical Record Retrieval tracks have been suspended because of the privacy issue and the TREC Microblog track provided the test collection for participants only through the use of the Twitter API. As stated by TREC organizers, the use of the API for creating the ground truth may violate the Twitter's terms of service in case it is used to collect the entire content of the Twitter stream. Moreover, the NTCIR lifelog task [53] organizers release only data gathered for few individuals due the the highly personal nature of the datasets and the inherent legal issues around the release. This limitation implies the challenge of sourcing datasets and test collections that could support comparative evaluation of tasks using lifelog datasets [54].

All the above examples highlight the barriers that could be faced when evaluating IR tasks embedding contextual and personal data. The privacy-preserving data issue in IR has been explored in recent venues such as privacy preserving IR (PIR) workshops launched since 2014 [127] in conjunction with the SIGIR conference. Some of the critical aspects that have attracted considerable debate over the third edition are (1) How to detect personal/private information and measure the privacy and security risks? (2) How to ensure private IR dataset release for evaluation purpose?

[10] http://www.cs.ucl.ac.uk/tasks-track-2016/.
[11] http://social-book-search.humanities.uva.nl/#/interactive.
[12] http://www.nytimes.com/2006/08/09/technology/09aol.html.

Furthermore, various recent works have focused on the design of suitable evaluation frameworks and measures under the challenge of privacy issues; accordingly, another relevant question is (3) to what extent is the evaluation of contextual IR performance feasible and reliable when using private data? In what follows, we summarize the relevant state-of-the-art findings that attempt to answer each of the above questions.

5.1 Overview of Contextual IR Applications and Related Privacy Risks

The growth and the diversity in the types of context used in a wide range of IR applications have created new risks to user privacy. The main issue is that a data releaser could not tell what information an attacker would cross-reference in the released data to infer other private data. In the following, we review the major contextual IR applications that involve sensitive data and identify the nature of privacy risks followed by the techniques used for measuring such risks.

5.1.1 *Major Contextual IR Applications and Related Sensitive Data.*

— *Personalized search*: E-commerce engines and medical search applications mainly rely on the use of collections of contextual data related to the personal user activities such as clickthrough, queries, browsing information, and demographics data to learn user models with the purpose of enhancing the search and/or maximizing revenues. Personalized search can potentially lead to a privacy violation by revealing such data, even if they are masked with random identifiers or hashed for anonymization purposes. For instance, Refernce [85] showed that publishing tokenized query logs does not preserve privacy in the case where the adversary has access to another log that can be used for reverse-engineering the tokens based on query frequency in the log. Reference [73] also demonstrated that privacy leaks are possible using query logs through the use of basic classifiers that are able to connect demographic data to individuals. From a commercial side, in addition to AOL and Netflix incidents, in October 2010, Facebook acknowledged that its top-10 popular applications including FarmVille and Texas Hold'em shared personal demographic user data and clicked ads with advertisers without users' agreement.

— *Location-based search*: Location is another element of context that is collected at a large scale using different devices such as smartphones, tablets and game consoles. However, the privacy concerns are significant when users share their location data with location-based service providers via queries and check-ins [102] since location traces of mobile users can be linked, even if anonymized, to other sources of user profiles including a few spatio-temporal points of the location trajectory of a target user. For instance, Reference [32] showed that only a few samples of GPS data are sufficient for identifying 95% of all users involved in a 500k-user phone log. In the same spirit, Reference [33] previously demonstrated that even though cell locations do not reveal the exact locations of users, a third party could rely on a sequence of cells to reveal the individuals with a high level of confidence. One well-known case of privacy violation that attracted high attention is Apple storing and collecting location data from its user iPhones, backed up with iTunes and sent to Apple without user consent [116].

— *Social search*: The widespread adoption of social platforms by users has renewed the problem of data privacy particularly because users naturally share information using these platforms and that information aggregation across platforms allows inferring personal invasive information [7] as well as linking user profiles across different social communities [49] or de-anonymizing users [165]. Several studies have also shown that user characteristics such as demographic and professional data [140, 147] can be inferred from users' social contexts by combining that information with third-party background knowledge of correlation

between different attributes. However, users remain unaware of the potential implications of their sharing activities on privacy controls.

5.1.2 Measuring Sensitivity and Privacy Risks. With considering the risks of privacy violation users are faced with, a natural question that rises is: How to provide privacy risks assessment to users and making them aware of possible exposure? Previous approaches for designing privacy-preserving data methodologies that provide control for users relied on the incorporation of user preferences over what type of data can be logged [84]. Recent approaches rather provide users with a measure that is able to estimate the risk of de-identification based on raw anonymized data. Within this realm, authors in Reference [129] introduced a new approach called stochastic privacy, which is fundamentally based on communicating a privacy risk measure to users, computed as the probability that their private data will be shared with third parties will not exceed a given bound. The general methodology is based on the optimization of a utility function constrained by both data accessibility with the purpose of maximizing the service quality (e.g., personalization) and privacy risk. The study by Reference [7] particularly estimates the risk assessment measure emerging from users' postings in online communities by relying on a ranking-based approach to the privacy-risk estimation. The general model first identifies sensitive topics and related adversary background knowledge and then computes the privacy risk score of a user with respect to a given sensitive topic. The latter is formulated as the maximal entropy estimated for a given user over the set of users in the community by means of masked versus unmasked attributes.

5.2 Evaluating Task Performance Using Private Data

Based on relevant literature, we can clearly see that while the release of personal and contextual data for evaluation purpose would be a great benefit to the research community, it could be harmful from a user data privacy perspective. The core question that naturally rises is: How to protect privacy by providing uncertain personal data, while still ensuring their utility and reliability for both evaluation and service design? This is the well-known privacy-utility tradeoff. Regarding the privacy-protection question, early techniques for privacy protection led to the notion of *anonymization*. The latter refers to removing personally identifiable information from data such as names, addresses and affiliations. Several models have been proposed for structured data such as k-anonymity [5], relying on the property that each record is indistinguishable from at least $k-1$. However, several previous studies and incidents (e.g., AOL and Netflix incidents) showed that such techniques fail under the information linkage process performed by third parties by using non-anonymized resources. A recent and strong formal privacy approach that emerged is *differential privacy* that became the standard of privacy-preserving data analytics. Differential privacy formalizes the constraint that the addition or removal of data record does not impact the result of statistical queries on the data. Accordingly, record linkage would not be possible, at least it would achieve inaccurate results. For a systematic review on differential privacy, the interested reader might refer to Reference [41]. Differential privacy has been successfully performed on diverse types of contextual data including user browsing log [44], query logs [46, 166], query and click logs [83], and location data [79].

The privacy-utility tradeoff specifically addresses the usefulness of safe data resulting from the application of privacy-preserving techniques. The review of the literature highlights that (1) this question has been experimentally evaluated with respect to the comparative performance of target tasks using the original data versus released data and that (2) the results generally vary jointly with the nature of data being perturbed and privacy budget values. For instance, in Reference [55], the authors studied the privacy-preserving issue within the task of people search in a social network. The authors showed that the task performance is significantly sensitive to the type of social

network attributes being masked: local versus global social network. In contrast, Fan et al. [44] showed that their technique achieves comparable results on both original browsing behavior logs and released logs for different tasks namely *top-K mining* and *web page counts* but varies according to privacy budget. Larger noise might lower task performance and hinders utility. The same conclusion has been found by Zhang et al. [166] within a adhoc web search task. Beyond the trade-off between data privacy and data utility for evaluation purpose, recent studies advocate the design of novel evaluation frameworks appropriate for searching safe data. Oard et al. [109] argue that sensitivity is a relevant criteria that could be incorporated into the nDCG measure. A system is evaluated with respect to its ability to *searching among secrets*, in other words, retrieving both relevant and non-sensitive information. In the same spirit, Fang and Zhai [55] propose the VIRlab system, which is characterized by a data-centric evaluation approach where the evaluation algorithms move from the sites of algorithm to the sites of data. This architecture provides the users with a better control on the types of data used for evaluation purpose but the underlying challenge is the design of a modular evaluation allowing to separate between the system and the user requirements.

6 CONCLUSION AND PROMISING PERSPECTIVES

In this survey, we reviewed the literature surrounding contextual IR evaluation. After presenting the different context definitions, we showed the influence of context on the validity of basic hypothesis of traditional IR evaluation through an argumentative debate and highlight the need for novel and more realistic hypothesis and norms that fit context-aware evaluation. Then, we presented an overview of major contextual IR evaluation methodologies and categorized them according to (1) system-based studies including official evaluation campaigns, track-like evaluation methods, and log-based studies, and (2) user-based evaluation methods including user studies and diary studies. We particularly reviewed this work by focusing on the impact of context on IR evaluation design and highlight the strengths and limitations of each category of evaluation methodologies. A second part of our review concerns the evaluation challenges accompanying the shift to dynamic IR on one hand and the privacy concerns within contextual IR evaluation on another hand.

The review of the literature in the area of contextual IR evaluation reveals that remarkable advances were developed this last decade in defining evaluation tasks that involve dynamic context, mobile context, search task, user sessions, geographic and temporal context. Although the advances in benchmarking evaluation frameworks for contextual IR within predefined tasks, IR evaluation becomes more and more challenging due to the timely convergence of several factors such as the rapid growth of devices, the multiplicity of contextual signals, the increasing complexity of search tasks, while guaranteeing the evaluation of the system through both controlled and uncontrolled experiments, thus to enable the evaluation of the system in real settings and to ensure repeatability of the experiments.

In what follows, we present some open issues that may contribute to relevant investigations in the field of contextual IR evaluation.

New Evaluation Design for Dynamic Context-Based IR. Dynamic IR is an emerging and promising sub-field of IR research [130, 158, 158]. It deals with the dynamics of real world search settings, responds to adverse changes, learns and adapts. Evaluation of dynamic IR is challenged by the need of a dynamic multi-stage search scenario that concerns the ranking of documents over multiple stages of search results. Although most user-based studies have assumed the context/relevance as dynamic/changing during session time, they are commonly situated outside the TREC/CLEF frameworks except a few studies that have exploited the TREC framework, more precisely the

HARD track of TREC collection for evaluating contextual IR systems [31, 78]. For the purpose of finding relationships between usefulness assessments and perceptions of work task complexity and search topic specificity, an exploratory user-based study has been conducted [68] using *i*Search Test Collection in Physics that includes structured and comprehensive user-generated information on the search situation such as the description of task, search topic, the perceived task complexity, and search topic specificity. By using datasets from official evaluation campaigns like Dynamic Domain TREC Track, a drawback is that they lack interactive data like a user's query reformulations as additional and contextual knowledge in evaluating a realistic dynamic IR evaluation framework. Moreover, the framework provides fine-grained feedback data on specific subtopics that are not available in real search settings. Another important missing aspect in dynamic IR evaluation framework design is the implementation of the principle of polyrepresentation, which is a theory that informs about the structured merge of a range of contextual elements involved in IR [65]. Polyrepresentation of the interaction process has been investigated through interface simulations studies [151–153].

One promising research opportunity in dynamic IR would be the use of the living lab evaluation as an alternative approach. This approach provides a benchmarking platform for researchers to train and then test their ranking systems in a live setting with real users in their natural task environments. From the dynamic search evaluation perspective, the main benefit of living lab experiments is the access to real interaction data within real applications on one hand and the ability to evaluate average system effectiveness regardless of any baseline ranker on another hand. Moreover, it opens up new evaluation perspectives related, for instance, to the impact of aspect modeling or query prediction on system effectiveness, at least at late search stages. This would lead to a better achievement of an optimal tradeoff between gain and effort. However, although the living lab approach has showed success in its first editions, it is challenged by a number of concerns regarding the tasks, design, development, maintenance, and security of the infrastructure required to support such evaluations. We highlight possible research perspectives for key concerns in benchmarking living lab evaluation for dynamic search: (1) As commercial organizations participate by providing access to their data and business processes, a first issue in living lab evaluation is related with legal and ethical issues such as the user consent, legalities regarding the release of data, copyright issues, commercial sensitivity of interaction data, and so on. This would motivate research in investigating privacy-preserving techniques that we reviewed in Section 5 to be integrated in the design and implementation of the living lab framework. (2) To make the evaluation of dynamic search more user-oriented, it is worth measuring the evolving user satisfaction over the search process. Beyond quantitative measurement that could be extracted from ground truth built by the users themselves rather than independent assessors, the living-lab framework naturally offers the opportunity of learning and testing qualitative indicators. However, because of the reproducibility issue, further investigation is needed to ensure the generalizability of experimental findings.

On the Consistence of Evaluation Metrics Used for Contextual Search. The literature review about contextual IR evaluation highlights that a metric generally encompasses different criteria such as novelty, freshness, ranking, gain, and cost to cite but a few. For instance, while the Session nDCG [71] considers the cumulative gain and document ranking, the Cube Test metric [97] considers a novelty discount per subtopic as well as the gain and cost per document. Jointly optimizing such factors is known as an NP-hard problem [17]. More particularly, these multi-criteria-based measures are heavily dependent on the topic (and subtopics) being evaluated, which makes the computation of averaged scores across topics not accurate and lead to their inappropriate use for systems comparison. Thus, an interesting research direction would be the design of appropriate optimization methods that rely on the aggregation and combination of partial optimal measures

per criteria that allow generating a global measure that does not necessarily require the individual measures to be comparable.

Another issue that rises from such measures is the fact that they require fine-grained relevance judgments from users at the subtopic level, which are costly in terms of both time and money. Prior works have investigated the reliability of results with incomplete judgments and have proposed appropriate measures such as the bPref [16]. However, they were performed using traditional precision-recall measures. Thus, another relevant investigation would be the study of the robustness of such measures in presence of incomplete judgments with the aim of revising them if needed to make dynamic test collections reliable evaluation resources.

Evaluation of Contextual Lifelogging Systems. The aim of lifelogging is to timely record multimodal digital traces from users performing their everyday activities. The traces include diverse forms of contextual data including time, location and actions. Thus, lifelogging technologies manage past and current historical and personal data, which is the ultimate form of context [54]. This invasive nature of lifelogging naturally impacts the personal privacy of lifeloggers. Thus, the emergence of lifelog data and related applications in information access and retrieval would renew the debate around the social, legal, and psychological aspects of personal privacy [54] and we believe that this would give rise to research opportunities in the area of contextual IR evaluation, which is initiated by the NTCIR 12 Lifelog Semantic Access Task. The core questions that are worth it to investigate are the following: (1) What would be the common evaluation guidelines to set up experimental evaluation protocols ensuring the lifeloggers' privacy preserving? (2) How to allow large-scale evaluation and the generalizability of the evaluation outcomes using limited samples of safe data? Preliminary investigations that attempt to answer the above questions have been recently undertaken by Chowdhury et al. [22] but the research field is still young and, to gain maturity, it requires the collaboration between researchers from two research disciplines, namely IR and information security as claimed in the Privacy-Preserving IR (PIR) workshop series [55].

In summary, evaluation of contextual IR seems to be in a significant and continuous progress, even if an important amount of related research has achieved maturity. Given its importance in many applications, new and more realistic frameworks that move beyond static frameworks with sampled and limited data need to be explored in the forthcoming future.

REFERENCES

Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06)*. ACM, New York, NY, 19–26.

Mehmet S. Aktas, Mehmet A. Nacar, and Filippo Menczer. 2006. Using hyperlink features to personalize web search. In *Proceedings of the 6th International Conference on Knowledge Discovery on the Web: Advances in Web Mining and Web Usage Analysis (WebKDD'04)*. Springer-Verlag, Berlin, 104–115.

Panos Balatsoukas and Ian Ruthven. 2010. What eyes can tell about the use of relevance criteria during predictive relevance judgment. In *Proceedings of the Information Interaction in Context Symposium (IIiX'10), New Brunswick, NJ, August 18–21, 2010*, Nicholas J. Belkin and Diane Kelly (Eds.). ACM, 389–394. DOI : http://dx.doi.org/10.1145/1840784.1840844

Carol L. Barry. 1998. Document representations and clues to document relevance. *J. Am. Soc. Inform. Sci.* 49, 14 (1998), 1293–1303.

Roberto J. Bayardo and Rakesh Agrawal. 2005. Data privacy through optimal k-anonymization. In *Proceedings of the 21st International Conference on Data Engineering (ICDE'05)*. IEEE Computer Society, Washington, DC, 217–228.

Paul N. Bennett, Kevyn Collins-Thompson, Diane Kelly, Ryen W. White, and Yi Zhang. 2015. Overview of the special issue on contextual search and recommendation. *ACM Trans. Inf. Syst.* 33, 1 (2015), 1e:1–1e:7.

Joanna Asia Biega, Krishna P. Gummadi, Ida Mele, Dragan Milchevski, Christos Tryfonopoulos, and Gerhard Weikum. 2016. R-susceptibility: An IR-centric approach to assessing privacy risks for users in online communities. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'16)*. ACM, New York, NY, 365–374.fi

Nilton Bila, Jin Cao, Robert Dinoff, Tin Kam Ho, Richard Hull, Bharat Kumar, and Paulo Santos. 2008. Mobile user profile acquisition through network observables and explicit user queries. In *Proceedings of the 9th International Conference on Mobile Data Management (MDM'08)*. IEEE Computer Society, Washington, DC, 98–107.

Pia Borlund. 2003. The concept of relevance in IR. *J. Am. Soc. Inf. Sci. Technol.* 54, 10 (Aug. 2003), 913–925.

Pia Borlund. 2003. The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Inf. Res.* 8, 3 (2003).

Ourdia Bouidghaghen, Lynda Tamine, and Mohand Boughanem. 2010. A diary study-based evaluation framework for mobile information retrieval (poster). In *Proceedings of the Asia Information Retrieval Society Conference (AIRS'10), Taipei, Taiwan, 01/12/2010-03/12/2010*, CHENG P.-J. et al. (Eds.), Vol. 6458/2010. Springer-Verlag, Berlin, 389–398.

Ourdia Bouidghaghen, Lynda Tamine, and Mohand Boughanem. 2011. Context-aware user's interests for personalizing mobile search, see Reference [164], 129–134.

Ourdia Bouidghaghen, Lynda Tamine, and Mohand Boughanem. 2011. Personalizing mobile web search for location sensitive queries, see Reference [164], 110–118.

Ourdia Bouidghaghen, Lynda Tamine, Gabriella Pasi, Guillaume Cabanac, Mohand Boughanem, and Célia Da Costa Pereira. 2011. Prioritized aggregation of multiple context dimensions in mobile ir (regular paper). In *The 7th Asia Information Retrieval Societies Conference (AIRS'11)* (LNCS), Mohamed Salem, Khaled Shaalan, Farhad Oroumchian, Azadeh Shakery, and Halim Khelalfa (Eds.), Vol. 7097. Springer-Verlag, 169–180.

Barry A. T. Brown, Abigail J. Sellen, and Kenton P. O'Hara. 2000. A diary study of information capture in working life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'00)*. ACM, New York, NY, 438–445.

Chris Buckley and Ellen M. Voorhees. 2004. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'04)*. ACM, 25–32.

Ben Carterette. 2011. An analysis of NP-completeness in novelty and diversity ranking. *Inf. Retr.* 14, 1 (Feb. 2011), 89–106.

Ben Carterette, Paul Clough, Mark Hall, Evangelos Kanoulas, and Mark Sanderson. 2016. Evaluating retrieval over sessions: The TREC session track 2011-2014. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'16)*. ACM, New York, NY, 685–688.

Olivier Chapelle, Thorsten Joachims, Filip Radlinski, and Yisong Yue. 2012. Large-scale validation and analysis of interleaved search evaluation. *ACM Trans. Inf. Syst.* 30, 1 (2012), 6:1–6:41.

Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09)*. ACM, New York, NY, 621–630.

Yi Chen and Gareth J. F. Jones. 2010. Augmenting human memory using personal lifelogs. In *Proceedings of the 1st Augmented Human International Conference (AH'10)*. ACM, New York, NY, Article 24, 9 pages.

Soumyadeb Chowdhury, Md. Sadek Ferdous, and Joemon M. Jose. 2016. Bystander privacy in lifelogging. In *HCI 2016 - Fusion! Proceedings of the 30th International BCS Human Computer Interaction Conference, BCS HCI 2016, Bournemouth University, Poole, UK, 11-15 July 2016 (Workshops in Computing)*, Shamal Faily, Nan Jiang, Huseyin Dogan, and Jacqui Taylor (Eds.). BCS, 1–3.

Karen Church, Antony Cousin, and Nuria Oliver. 2012. I wanted to settle a bet! Understanding why and how people use mobile search in social settings. In *Proceedings of the 14th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI'12)*. ACM, 393–402.

Karen Church and Barry Smyth. 2009. Understanding the intent behind mobile information needs. In *Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI'09)*. ACM, New York, NY, 247–256.

Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*. ACM, 659–666.

Cyril Cleverdon. 1967. The Cranfield tests on index language devices. *Aslib Proceedings* 19, 6 (1967), 173–194. DOI: http://dx.doi.org/10.1108/eb050097

Colleen Cool and Amanda Spink. 2002. Issues of context in information retrieval (IR): An introduction to the special issue. *Inf. Process. Manag.* 38, 5 (Sept. 2002), 605–611.

Mariam Daoud, Lynda T. Lechani, and Mohand Boughanem. 2008. Using a graph-based ontological user profile for personalizing search. In *CIKM'08: Proceeding of the 17th ACM Conference on Information and Knowledge Mining*. ACM, New York, NY, 1495–1496.

Mariam Daoud, Lynda Tamine, and Mohand Boughanem. 2010. A personalized graph-based document ranking model using a semantic user profile. In *Proceedings of the Conference on User Modeling, Adaptation and Personalization (UMAP'10)*. Springer, 171–182.fi

Mariam Daoud, Lynda Tamine, and Mohand Boughanem. 2011. A personalized search using a semantic distance measure in a graph-based ranking model. *J. Inform. Sci.* 37, 6 (2011), 614–636.
Mariam Daoud, Lynda Tamine-Lechani, and Mohand Boughanem. 2009. Towards a graph-based user profile modeling for a session-based personalized search. *Knowl. Inf. Syst.* 21, 3 (2009), 365–398. DOI: http://dx.doi.org/10.1007/s10115-009-0232-0
Yves-Alexandre De Montjoye, A. Hidalgo Cesar, Verleysen Michel, and D. Blondel Vincent. 2013. Unique in the crowd: The privacy bounds of human mobility. *Sci. Rep.* 3, 1376 (2013).
Yoni De Mulder, George Danezis, Lejla Batina, and Bart Preneel. 2008. Identification via location-profiling in GSM networks. In *Proceedings of the 7th ACM Workshop on Privacy in the Electronic Society (WPES'08)*. ACM, 23–32.
Adriel Dean-Hall. 2014. *An Evaluation of Contextual Suggestion*. Master's thesis. University of Waterloo.
Adriel Dean-Hall, Charles L. A. Clarke, Jaap Kamps, Julia Kiseleva, and Ellen M. Voorhees. 2015. Overview of the TREC 2015 contextual suggestion track. In *Proceedings of the 24th Text REtrieval Conference (TREC'15)*, Gaithersburg, Maryland, 1–2.
Adriel Dean-Hall, Charles L. A. Clarke, Jaap Kamps, Paul Thomas, and Ellen Voorhes. 2014. Overview of the TREC 2014 contextual suggestion track. In *Proceedings of Text REtrieval Conference (TREC'14)*. NIST, 1–2.
Alberto Díaz, Antonio García, and Pablo Gervás. 2008. User-centred versus system-centred evaluation of a personalization system. *Inf. Process. Manag.* 44, 3 (2008), 1293–1307.
Chen Ding and Jagdish C. Patra. 2007. User modeling for personalized web search with self-organizing map: Research articles. *J. Am. Soc. Inf. Sci. Technol.* 58, 4 (Feb. 2007), 494–507.
Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. 2007. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th International Conference on World Wide Web (WWW'07)*. ACM, New York, NY, 581–590.
Susan Dumais, Edward Cutrell, J. J. Cadiz, Gavin Jancke, Raman Sarin, and Daniel C. Robbins. 2003. Stuff I've seen: A system for personal information retrieval and re-use. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'03)*. ACM, New York, NY, 72–79.
Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* 9, 3&4 (2014), 211–407.
David Elsweiler, Bernd Ludwig, Leif Azzopardi, and Max L. Wilson (Eds.). 2014. *Proeedings of the 5th Information Interaction in Context Symposium, IIiX'14, Regensburg, Germany, August 26-29, 2014*. ACM. http://dl.acm.org/citation.cfm?id=2637002
David Elsweiler and Ian Ruthven. 2007. Towards task-based personal information management evaluations. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)*. ACM, New York, NY, 23–30.
Liyue Fan, Luca Bonomi, Li Xiong, and Vaidy Sunderam. 2014. Monitoring web browsing behavior with differential privacy. In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14)*. ACM, 177–188.
Paolo Ferragina and Antonio Gulli. 2005. A personalized search engine based on web-snippet hierarchical clustering. In *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web (WWW'05)*. ACM, New York, NY, 801–810.
Marco Gaboardi, Emilio Jesús Gallego Arias, Justin Hsu, Aaron Roth, and Zhiwei Steven Wu. 2014. Dual query: Practical private query release for high dimensional data. In *Proceedings of the 31st International Conference on Machine Learning (ICML'14)*. 1170–1178.
Fredric Gey, Ray Larson, Mark Sanderson, Hideo Joho, Paul Clough, and Vivien Petras. 2006. GeoCLEF: The CLEF 2005 cross-language geographic information retrieval track overview. In *Proceedings of the 6th International Conference on Cross-Language Evalution Forum: Accessing Multilingual Information Repositories (CLEF'05)*. Springer-Verlag, Berlin, 908–919.
Fredric C. Gey, Ray R. Larson, Noriko Kando, Jorge Machado, and Tetsuya Sakai. 2010. NTCIR-GeoTime overview: Evaluating geographic and temporal search. In *NII Testbeds and Community for Information Access Research*, Vol. 10. 147–153.
Oana Goga, Howard Lei, Sree Hari Krishnan Parthasarathi, Gerald Friedland, Robin Sommer, and Renata Teixeira. 2013. Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*. ACM, 447–458.
Ayse Göker and Hans I. Myrhaug. 2002. User context and personalisation. In *Workshop on Case Based Reasoning and Personalization*, Mehmet H. Göker and B. Smyth (Eds.). Aberdeen, Scotland, 1–7.
Ayse Göker and Hans I. Myrhaug. 2008. Evaluation of a mobile information system in context. *Inf. Process. Manage.* 44, 1 (2008), 39–65.
Dongyi Guan, Sicong Zhang, and Hui Yang. 2013. Utilizing query change for session search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*. ACM, New York, NY, 453–462.fi

Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, and Rami Albatal. 2016. Overview of NTCIR-12 lifelog task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*. 354–360.
Cathal Gurrin, Alan F. Smeaton, and Aiden R. Doherty. 2014. LifeLogging: Personal big data. *Found. Trends Inf. Retr.* 8, 1 (2014), 1–125.
Shuguang Han, Daqing He, and Zhen Yue. 2014. Benchmarking the privacy-preserving people search, see Reference [127], 13–18.
Ahmed Hassan and Ryen W. White. 2012. Task tours: Helping users tackle complex search tasks. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM'12)*. ACM, New York, NY, 1885–1889.
Ahmed Hassan Awadallah, Ryen W. White, Patrick Pantel, Susan T. Dumais, and Yi-Min Wang. 2014. Supporting complex search tasks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM'14)*. ACM, New York, NY, 829–838.
R. B. Haynes, K. A. McKibbon, C. J. Walker, N. Ryan, D. Fitzgerald, and M. F. Ramsden. 1990. Online access to MEDLINE in clinical settings. A study of use and usefulness. *Ann. Intern. Med.* 112, 1 (1990), 78–84.
Daqing He, Ayse Göker, and David J. Harper. 2002. Combining evidence for automatic web session identification. *Inf. Process. Manage.* 38, 5 (2002), 727–742.
Birger Hjørland. 2010. The foundation of the concept of relevance. *J. Am. Soc. Inf. Sci.* 61, 2 (2010), 217–237.
Mu-hsuan Huang and Hui-yu Wang. 2004. The influence of document presentation order and number of documents judged on users' judgments of relevance. *J. Am. Soc. Inf. Sci. Technol.* 55, 11 (Sept. 2004), 970–979.
Peter Ingwersen. 1996. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Document.* 52 (1996), 3–50.
Peter Ingwersen. 2007. Context in information interaction revisited 2006. In *Proceedings of the 4th Biennial DISSAnet Conference*. Bothma, T.J.D. & Kaniki, A. 2007, 13–23.
Peter Ingwersen. 2011. The user in interactive information retrieval evaluation. In *Advanced Topics in Information Retrieval*, Massimo Melucci and Ricardo A. Baeza-Yates (Eds.). The Information Retrieval Series, Vol. 33. Springer, 83–107. DOI : http://dx.doi.org/10.1007/978-3-642-20946-8_4
Peter Ingwersen. 2012. *Scientometric Indicators and Webometrics and the Polyrepresentation Principle in Information Retrieval.* Ess Ess Publications, New Delhi, India.
Peter Ingwersen and Kalervo Järvelin. 2005. Information retrieval in context: IRiX. *SIGIR Forum* 39, 2 (Dec. 2005), 31–39. DOI : http://dx.doi.org/10.1145/1113343.1113351
Peter Ingwersen and Kalervo Järvelin. 2005. *The Turn: Integration of Information Seeking and Retrieval in Context*, The Information Retrieval Series. Springer-Verlag New York, Inc., Secaucus, NJ.
Peter Ingwersen and Peiling Wang. 2012. *Relationship between Usefulness Assessments and Perceptions of Work Task Complexity and Search Topic Specificity: An Exploratory Study.* 19–23.
Joseph W. Janes. 1991. Relevance judgments and the incremental presentation of document representations. *Inf. Process. Manag.* 27, 6 (1991), 629–646.
Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. 2008. Determining the informational, navigational, and transactional intent of web queries. *Inf. Process. Manag.* 44, 3 (May 2008), 1251–1266.
Kalervo Järvelin, Susan L. Price, Lois M. L. Delcambre, and Marianne Lykke Nielsen. 2008. Discounted cumulated gain based evaluation of multiple-query IR sessions. In *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval (ECIR'08)*. Springer-Verlag, Berlin, 4–15.
Kalervo Järvelin, Pertti Vakkari, Paavo Arvola, Feza Baskaya, Anni Järvelin, Jaana Kekäläinen, Heikki Keskustalo, Sanna Kumpulainen, Miamaria Saastamoinen, Reijo Savolainen, and Eero Sormunen. 2015. Task-based information interaction evaluation: The viewpoint of program theory. *ACM Trans. Inf. Syst.* 33, 1, Article 3 (March 2015), 30 pages.
Rosie Jones, Ravi Kumar, Bo Pang, and Andrew Tomkins. 2007. I know what you did last summer: Query logs and user privacy. In *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management (CIKM'07)*. ACM, 909–914.
Evangelos Kanoulas, Ben Carterette, Paul D. Clough, and Mark Sanderson. 2011. Evaluating multi-query sessions. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'11)*. ACM, 1053–1062.
J. Kekalainen and K. Jarvelin. 2004. Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance. In *Proceedings of the 4th CoLIS Conference*. 253–270.
Melanie Kellar, Carolyn Watters, and Michael Shepherd. 2007. A field study characterizing web-based information-seeking tasks. *J. Am. Soc. Inf. Sci. Technol.* 58, 7 (May 2007), 999–1018.
Diane Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Found. Trends Inf. Retr.* 3 (2009), 1–224.fi

Diane Kelly and Xin Fu. 2007. Eliciting better information need descriptions from users of information search systems. *Inf. Process. Manag.* 43, 1 (Jan. 2007), 30–46. DOI: http://dx.doi.org/10.1016/j.ipm.2006.03.006

Ali Khoshgozaran and Cyrus Shahabi. 2009. Private information retrieval techniques for enabling location privacy in location-based services. In *Privacy in Location-Based Applications*. Springer, 59–83.

Jaewon Kim, Paul Thomas, Ramesh Sankaranarayana, Tom Gedeon, and Hwan-Jin Yoon. 2017. What snippet size is needed in mobile web search? In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR'17)*. ACM, New York, NY, 97–106. DOI: http://dx.doi.org/10.1145/3020165.3020173

Kyung-Sun Kim and Bryce Allen. 2002. Cognitive and task influences on web searching behavior. *J. Am. Soc. Inf. Sci. Technol.* 53, 2 (2002), 109–119.

Marijn Koolen, Toine Bogers, Maria Gäde, Mark Hall, Iris Hendrickx, Hugo Huurdeman, Jaap Kamps, Mette Skov, Suzan Verberne, and David Walsh. 2016. *Overview of the CLEF 2016 Social Book Search Lab*. Springer International Publishing, Cham, 351–370.

Aleksandra Korolova, Krishnaram Kenthapadi, Nina Mishra, and Alexandros Ntoulas. 2009. Releasing search queries and clicks privately. In *Proceedings of the 18th International Conference on World Wide Web (WWW'09)*. ACM, 171–180.

Andreas Krause and Eric Horvitz. 2008. A utility-theoretic approach to privacy and personalization. In *Proceedings of the 23rd National Conference on Artificial Intelligence—Vol. 2 (AAAI'08)*. AAAI Press, 1181–1188.

Ravi Kumar, Jasmine Novak, Bo Pang, and Andrew Tomkins. 2007. On anonymizing query logs via token-based hashing. In *Proceedings of the 16th International Conference on World Wide Web (WWW'07)*. ACM, 629–638.

Nicholas D. Lane, Dimitrios Lymberopoulos, Feng Zhao, and Andrew T. Campbell. 2010. Hapori: Context-based local search for mobile phones using community behavioral modeling and similarity. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp'10)*. ACM, New York, NY, 109–118.

Effie Lai-Chong Law, Tomaž Klobučar, and Matic Pipan. 2006. User effect in evaluating personalized information retrieval systems. In *Proceedings of the 1st European Conference on Technology Enhanced Learning: Innovative Approaches for Learning and Knowledge Sharing (EC-TEL'06)*. Springer-Verlag, Berlin, 257–271.

Chia-Jung Lee, Jaime Teevan, and Sebastian de la Chica. 2014. Characterizing multi-click search behavior and the risks and opportunities of changing results during use. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'14)*. ACM, New York, NY, 515–524.

Michael E. Lesk and Gerard Salton. 1968. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval* 4, 4 (1968), 343–359. DOI: http://dx.doi.org/10.1016/0020-0271(68)90029-6

Liangda Li, Hongbo Deng, Anlei Dong, Yi Chang, and Hongyuan Zha. 2014. Identifying and labeling search tasks via query-based hawkes processes. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*. ACM, New York, NY, 731–740.

Jimmy Lin, Miles Efron, Yulu Wang, Garrick Sherman, and Ellen Voorhees. 2015. Overview of the TREC-2015 microblog track. In *Proceedings of Text REtrieval Conference (TREC'15)*. NIST.

Jimmy Lin, Dennis Quan, Vineet Sinha, Karun Bakshi, David Huynh, Boris Katz, and David R. Karger. 2003. What makes a good answer? The role of context in question answering. In *Proceedings of Interact 2003*. 25–32.

Jimmy J. Lin and Miles Efron. 2013. Overview of the TREC-2013 microblog track. In *Proceedings of the 22nd Text REtrieval Conference (TREC'13), Gaithersburg, Maryland, November 19–22, 2013*, Ellen M. Voorhees (Ed.), Vol. Special Publication 500-302. National Institute of Standards and Technology (NIST), 1–5.

Fang Liu, Clement Yu, and Weiyi Meng. 2004. Personalized web search for improving retrieval effectiveness. *IEEE Trans. on Knowl. and Data Eng.* 16, 1 (2004), 28–40.

Ying-Hsang Liu and Ralf Bierig. 2014. A review of users' search contexts for lifelogging system design. In *Proceedings of the 5th Information Interaction in Context Symposium (IIiX'14)*. ACM, New York, NY, 271–274.

Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei. 2013. Discovering tasks from search engine query logs. *ACM Trans. Inf. Syst.* 31, 3 (2013), 14:1–14:43.

Jiyun Luo, Christopher Wing, Hui Yang, and Marti Hearst. 2013. The water filling model and the Cube Test: Multi-dimensional evaluation for professional search. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM'13)*. ACM, New York, NY, 709–714.

Gary Marchionini. 1995. *Information Seeking in Electronic Environments*. Cambridge University Press, New York, NY.

Rishabh Mehrotra, Prasanta Bhattacharya, and Emine Yilmaz. 2016. Deconstructing complex search tasks: A Bayesian nonparametric approach for extracting sub-tasks. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 599–605.

Rishabh Mehrotra and Emine Yilmaz. 2015. Towards hierarchies of search tasks & subtasks. In *Proceedings of the 24th International Conference on World Wide Web (WWW'15 Companion)*. ACM, New York, NY, 73–74.

Massimo Melucci. 2012. *Contextual Search*. Now Publishers Inc., Hanover, MA.fi

S. Merrill, N. Basalp, J. Biskup, E. Buchmann, C. Clifton, B. Kuijpers, W. Othman, and E. Savas. 2013. Privacy through uncertainty in location-based services. In *Proceedings of the 2013 IEEE 14th International Conference on Mobile Data Management*, Vol. 2. 67–72.

Alessandro Micarelli and Filippo Sciarrone. 2004. Anatomy and empirical evaluation of an adaptive web-based information filtering system. *User Modeling and User-Adapted Interaction* 14, 2–3 (June 2004), 159–200.

Ph. Mylonas, D. Vallet, P. Castells, M. Fernández, and Y. Avrithis. 2008. Personalized information retrieval based on context and ontological knowledge. *Knowl. Eng. Rev.* 23, 1 (2008), 73–100.

Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*. IEEE Computer Society, 111–125.

Fuhr Nobert. 2000. *Information Retrieval: Introduction and Survey*. University of Duisburg-Essen, Germany.

Ragnar Nordlie, Nils Pharo, Luanne Freund, Birger Larsen, and Dan Russel (Eds.). 2017. *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR'17), Oslo, Norway, March 7–11, 2017.* ACM. DOI: http://dx.doi.org/10.1145/3020165

H. O. Nyongesa and S. Maleki-Dizaji. 2006. User modelling using evolutionary interactive reinforcement learning. *Inf. Retr.* 9, 3 (2006), 343–355.

Douglas W. Oard, Katie Shilton, and Jimmy J. Lin. 2016. Evaluating search among secrets, see Reference [159].

Heather L. O'Brien, Nicola Ferro, Hideo Joho, Dirk Lewandowski, Paul Thomas, and Keith van Rijsbergen. 2016. System and user centered evaluation approaches in interactive information retrieval (SAUCE 2016). In *Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval, CHIIR 2016, Carrboro, North Carolina, March 13-17, 2016*, Diane Kelly, Robert Capra, Nicholas J. Belkin, Jaime Teevan, and Pertti Vakkari (Eds.). ACM, 337–340. DOI: http://dx.doi.org/10.1145/2854946.2886106

Maria Papadogiorgaki, Vasileios Papastathis, Evangelia Nidelkou, Simon Waddington, Ben Bratu, Myriam Ribière, and Ioannis Kompatsiaris. 2008. Two-level automatic adaptation of a distributed user profile for personalized news content delivery. *Int. J. Digital Multimedia Broadcasting* 2008 (2008), 863613:1–863613:21.

Taemin Kim Park. 1994. Toward a theory of user-based relevance: A call for a new paradigm of inquiry. *J. Am. Soc. Inf. Sci.* 45, 3 (April 1994), 135–141.

Arti Poddar and Ian Ruthven. 2010. The emotional impact of search tasks. In *Proceedings of the 3rd Symposium on Information Interaction in Context (IIiX'10)*. ACM, New York, NY, 35–44. DOI: http://dx.doi.org/10.1145/1840784.1840792

Feng Qiu and Junghoo Cho. 2006. Automatic identification of user interest for personalized search. In *Proceedings of the 15th International Conference on World Wide Web (WWW'06)*. 727–736.

M. M. Rahman, S. Yeasmin, and C. K. Roy. 2014. Towards a context-aware IDE-based meta search engine for recommendation about programming errors and exceptions. In *Proceedings of the 2014 Software Evolution Week—IEEE Conference on Software Maintenance, Reengineering and Reverse Engineering (CSMR-WCRE'14),* 194–203.

Report. 2014. Apple Q&A on location data. *Press Release* 48, 2 (2014), 83–88.

John Rieman. 1993. The diary study: A workplace-oriented research tool to guide laboratory efforts. In *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems (CHI'93)*. ACM, New York, NY, 321–326.

Jean-David Ruvini. 2003. Adapting to the user's internet search strategy on small devices. In *Proceedings of the 2003 International Conference on Intelligent User Interfaces*. 284–286.

Camilla Sanvitto, Debasis Ganguly, Gareth J. F. Jones, and Gabriella Pasi. 2016. A laboratory-based method for the evaluation of personalised search, see Reference [159].

Tefko Saracevic. 1975. RELEVANCE: A review of and a framework for the thinking on the notion in information science. *JASIS* 26, 6 (1975), 321–343.

Tefko Saracevic. 1997. The stratified model of information retrieval interaction: Extension and applications. In *Proceedings of the Annual Meeting-American Society for Information Science*, Vol. 34. Learned Information, Europe, LTD., 313–327.

Tefko Saracevic. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. *J. Am. Soc. Inform. Sci. Technol.* 58, 13 (2007), 1915–1933.

Xuehua Shen, Bin Tan, and ChengXiang Zhai. 2005. Context-sensitive information retrieval using implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05)*. ACM, New York, NY, 43–50.

Xuehua Shen, Bin Tan, and ChengXiang Zhai. 2005. Implicit user modeling for personalized search. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM'05)*. ACM, 824–831.

Xuehua Shen, Bin Tan, and ChengXiang Zhai. 2005. Implicit user modeling for personalized search. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM'05)*. ACM, New York, NY, 824–831.fi

Milad Shokouhi and Qi Guo. 2015. From queries to cards: Re-ranking proactive card recommendations based on reactive search history. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*. 695–704.

Luo Si and Hui Yang (Eds.). 2014. *Proceeding of the 1st International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security Co-located with 37th Annual International ACM SIGIR Conference, PIR at SIGIR 2014, Gold Coast, Australia, July 11, 2014*. CEUR Workshop Proceedings, Vol. 1225. CEUR-WS.org.

Ahu Sieg, Bamshad Mobasher, and Robin Burke. 2007. Web search personalization with ontological user profiles. In *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management (CIKM'07)*. ACM, New York, NY, USA, 525–534.

Adish Singla, Eric Horvitz, Ece Kamar, and Ryen White. 2014. Stochastic privacy. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI'14)*. AAAI Press, 152–158.

Marc Sloan and Jun Wang. 2016. Dynamic information retrieval: Theoretical framework and application. *CoRR* abs/1601.04605 (2016).

Mark D. Smucker and Charles L. A. Clarke. 2012. Time-based calibration of effectiveness measures. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*. ACM, New York, NY, 95–104.

Eero Sormunen. 2002. Liberal relevance criteria of TREC: Counting on negligible documents? In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'02)*. ACM, New York, NY, 324–330.

Micro Speretta and Susan Gauch. 2005. Personalized search based on user search histories. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*. IEEE Computer Society, Washington, DC, 622–628.

Yang Sun, Huajing Li, Isaac G. Councill, Jian Huang, Wang-Chien Lee, and C. Lee Giles. 2008. Personalized ranking for digital libraries based on log analysis. In *Proceedings of the 10th ACM Workshop on Web Information and Data Management (WIDM'08)*. ACM, New York, NY, 133–140.

Lynda Tamine and Mohand Boughanem. 2010. Inferring document utility via a decision-making based retrieval model. *Int. J. Know.-Based Intell. Eng. Syst.* 14, 2 (2010), 73–93.

Lynda Tamine-Lechani, Mohand Boughanem, and Mariam Daoud. 2010. Evaluation of contextual information retrieval effectiveness: Overview of issues and research. *Knowl. Inf. Syst.* 24, 1 (July 2010), 1–34.

Bin Tan, Xuehua Shen, and ChengXiang Zhai. 2006. Mining long-term search history to improve search accuracy. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*. ACM, New York, NY, 718–723.

Desney S. Tan, Darren Gergle, Regan L. Mandryk, Kori Inkpen, Melanie Kellar, Kirstie Hawkey, and Mary Czerwinski. 2008. Using job-shop scheduling tasks for evaluating collocated collaboration. *Pers. Ubiquitous Comput.* 12, 3 (2008), 255–267.

Jaime Teevan, Susan T. Dumais, and Eric Horvitz. 2005. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05)*. ACM, New York, NY, 449–456.

Hu Tianran, Xiao Haoyuan, Luo Jiebo, and Thi Nguyen Thuy-vy. 2016. What the language you tweet says about your occupation. In *Proceedings of the 10th International AAAI Conference on Web and Social Media (ICWSM'16)*. AAAI, 181–190.

Aravindan Veerasamy and Russell Heikes. 1997. Effectiveness of a graphical display of retrieval results. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*. ACM, New York, NY, 236–245.

Manisha Verma and Emine Yilmaz. 2016. Characterizing relevance on mobile and desktop. In *Advances in Information Retrieval—38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016*, Lecture Notes in Computer Science, Vol. 9626. Springer, 212–223.

Ellen M. Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*. ACM, New York, NY, USA, 315–323. DOI: http://dx.doi.org/10.1145/290941.291017

Ellen M. Voorhees and Donna K. Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. The MIT Press.

Thanh Vu, Alistair Willis, Udo Kruschwitz, and Dawei Song. 2017. Personalised query suggestion for intranet search with temporal user profiling. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR'17)*. ACM, New York, NY, USA, 265–268. DOI: http://dx.doi.org/10.1145/3020165.3022129

Hongning Wang, Yang Song, Ming-Wei Chang, Xiaodong He, Ryen W. White, and Wei Chu. 2013. Learning to extract cross-session search tasks. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*. 1353–1364.fi

Pengfei Wang, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2016. Your cart tells you: Inferring demographic attributes from purchase data. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM'16)*. ACM, 173–182.

Xing Wei, Yinglong Zhang, and Jacek Gwizdka. 2014. YASFIIRE: Yet another system for IIR evaluation. In *Proceedings of the 5th Information Interaction in Context Symposium (IIiX'14)*. ACM, New York, NY, 316–319. DOI: http://dx.doi.org/10.1145/2637002.2637051

Ji-Rong Wen, Ni Lao, and Wei-Ying Ma. 2004. Probabilistic model for contextual retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'04)*. ACM, New York, NY, 57–63.

Ryen White and Resa Roth. 2009. *Exploratory Search: Beyond the Query-Response Paradigm. Synthesis Lectures on Information Concepts, Retrieval, and Services*. Morgan & Claypool.

Ryen W. White. 2006. Using searcher simulations to redesign a polyrepresentative implicit feedback interface. *Inf. Process. Manage.* 42, 5 (Sept. 2006), 1185–1202. DOI: http://dx.doi.org/10.1016/j.ipm.2006.02.005

Ryen W. White, Joemon M. Jose, and Ian Ruthven. 2006. An implicit feedback approach for interactive information retrieval. *Inf. Process. Manage.* 42, 1 (Jan. 2006), 166–190. DOI: http://dx.doi.org/10.1016/j.ipm.2004.08.010

Ryen W. White, Ian Ruthven, Joemon M. Jose, and C. J. Van Rijsbergen. 2005. Evaluating implicit feedback models using searcher simulations. *ACM Trans. Inf. Syst.* 23, 3 (July 2005), 325–361. DOI: http://dx.doi.org/10.1145/1080343.1080347

Peter J. Wild, Chris McMahon, Mansur Darlington, Shaofeng Liu, and Steve Culley. 2010. A diary study of information needs and document usage in the engineering domain. *Design Studies* 31, 1 (2010), 46–73.

Biao Xiang, Daxin Jiang, Jian Pei, Xiaohui Sun, Enhong Chen, and Hang Li. 2010. Context-aware ranking in web search. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*. ACM, New York, NY, 451–458.

Hu Yang, Jan Franck, and Ian Soboroff. 2015. TREC 2015 dynamic domain track overview. In *Proceedings of Text REtrieval Conference (TREC'15)*. NIST, 1–28.

Hui Yang, Dongyi Guan, and Sicong Zhang. 2015. The query change model: Modeling session search as a Markov decision process. *ACM Trans. Inf. Syst.* 33, 4 (2015), 20:1–20:33.

Hui Yang, Marc Sloan, and Jun Wang. 2014. Dynamic information retrieval modeling. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'14)*. ACM, New York, NY, 1290–1290.

Emine Yilmaz and Charles L. A. Clarke (Eds.). 2016. *Proceedings of the 7th International Workshop on Evaluating Information Access, EVIA 2016, a Satellite Workshop of the NTCIR-12 Conference, National Center of Sciences, Tokyo, Japan, June 7, 2016*. National Institute of Informatics (NII), 1–32.

Emine Yilmaz, Evangelos Kanoulas, Manisha Verma, Ben Cartette, and Nick Craswell. 2015. TREC 2015 tasks track overview. In *Proceedings of Text REtrieval Conference (TREC'15)*. NIST, 1–7.

Emine Yilmaz, Manisha Verma, Rishabh Mehrotra, Evangelos Kanoulas, Ben Carterette, and Nick Craswell. 2015. Overview of the TREC 2015 tasks track. In *Proceedings of the 24th Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015*. National Institute of Standards and Technology (NIST).

Josh Jia-Ching Ying, Eric Hsueh-Chan Lu, Wang-Chien Lee, Tz-Chiao Weng, and Vincent S. Tseng. 2010. Mining user similarity from semantic trajectories. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks (LBSN'10)*. ACM, New York, NY, 19–26.

Gae-won You and Seung-won Hwang. 2007. Personalized ranking: A contextual ranking approach. In *Proceedings of the 2007 ACM Symposium on Applied Computing (SAC'07)*. ACM, New York, NY, 506–510.

Arkady B. Zaslavsky, Panos K. Chrysanthis, Dik Lun Lee, Dipanjan Chakraborty, Vana Kalogeraki, Mohamed F. Mokbel, and Chi-Yin Chow (Eds.). 2011. *Proceedings of the 12th IEEE International Conference on Mobile Data Management, MDM 2011, Luleå, Sweden, June 6-9, 2011, Volume 1*. IEEE Computer Society, 1–108.

Aston Zhang, Xing Xie, Kevin Chen-Chuan Chang, Carl A. Gunter, Jiawei Han, and XiaoFeng Wang. 2014. Privacy risk in anonymized heterogeneous information networks. In *Proceedings of the 17th International Conference on Extending Database Technology, EDBT 2014, Athens, Greece, March 24-28, 2014*. 595–606.

Sicong Zhang, Hui Yang, and Lisa Singh. 2016. Anonymizing query logs by differential privacy. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'16)*. ACM, New York, NY, 753–756.fi