Evaluation Methodologies in Information Retrieval Dagstuhl Seminar 13441

Maristella Agosti
University of Padua, Italy
agosti@dei.unipd.it

Norbert Fuhr
University of Duisburg-Essen, Germany
norbert.fuhr@uni-due.de

Elaine Toms
Sheffield University, UK
e.toms@sheffield.ac.uk

Pertti Vakkari
University of Tampere, Finland
pertti.vakkari@uta.fi

1 Introduction

This paper reports on the *Evaluation Methodologies in Information Retrieval* Seminar[1] [1] held from 27 October to 1 November 2013 at the *Schloss Dagstuhl* - Leibniz Center for Informatics that is a world-wide renowned venue for informatics where scientists come together to exchange their knowledge and to discuss their research findings. Dagstuhl offers modern facilities and is located in the beautiful countryside of Saarland, Germany.

Schloss Dagstuhl has previously hosted other seminars on information retrieval topics, e.g. the Seminar 09101 on *Interactive Information Retrieval* which took place in 2009[2], but this one was the first one devoted to the specific topic of information retrieval evaluation. The seminar was attended by 42 participants from thirteen different countries, including a large number of established researchers as well as some some promising young researchers, and also practitioners from industry[3].

As it is well-known evaluation of information retrieval (IR) systems has a long tradition (see, for example, [2]), however, the test-collection based evaluation paradigm is of limited value for assessing today's IR applications, since it fails to address major aspects of the IR process. Thus there is a need for new evaluation methodologies, which are able to deal with the following issues:

- In interactive IR, users have a wide variety of interaction possibilities. The classical paradigm only regards the document ranking for a single query. In contrast, new functions such as

[1] http://www.dagstuhl.de/en/program/calendar/semhp/?semnr=13441
[2] http://www.dagstuhl.de/en/program/calendar/semhp/?semnr=09101
[3] http://www.dagstuhl.de/program/calendar/partlist/?semnr=13441&SUOG

search term completion, query term suggestion, faceted search, document clustering, query-biased summaries also have a strong influence on the user's search experience and thus should be considered in an evaluation.

- From a user's point of view, evaluation of performance of IR systems should be in terms of how well they are supported with respect to whole search sessions. Typically, users initiate a session with a specific goal (e.g. acquiring crucial information for making a decision, learning about topics or events they are interested in, or just for getting entertained). Thus, the overall quality of a system should be evaluated with regard to the user's goal. However, it is an open research issue how this can be achieved.

- There is an increasing number of search applications (especially on mobile devices), which support specific tasks (e.g. finding the next restaurant, comparing prices for a specific product). Here goal-oriented evaluation may be more straightforward. From an IR researcher's point of view, however, we would like to learn about the quality of contribution of the underlying IR engine, and how it can possibly be improved.

- Besides ad-hoc-retrieval, also monitoring or filtering is an important IR task type. Here streams of short messages (e.g. tweets, chats) pose new challenges. It is an open question whether or not the classical relevance-based evaluation is sufficient for a user-oriented evaluation.

2 Motivation

To address and to solve the previously stated issues, there is a need for the development of appropriate methodologies such as:

- Evaluation infrastructures that provide test-beds along with evaluation methods, software and databases for computing measures, collecting and comparing results.

- Test-beds for interactive IR evaluation are hardly reusable at the moment (with the exception of simulation approaches like, for example, in the TREC session track). However, sharing data from user experiments might be an important step in this direction.

- Living labs use operational systems as experimental platforms on which to conduct user-based experiments at scale. To be usable, we need a site attracting enough traffic, and an architecture that allows for plugging in components from different researcher groups.

- Frameworks for modeling system-user interactions with clear methodological implications.

3 Approach

Before the event, each participant was asked to identify one to five crucial issues in IR evaluation methodology. A summary of these contributions presented by Pertti Vakkari showed that there are five major themes deemed relevant by the participants: 1) Evaluation frameworks, 2) Whole

session evaluation and evaluation over sessions, 3) Evaluation criteria: from relevance to utility, 4) User modeling, and 5) Methodology and metrics.

Based on the evaluation model proposed in [4], the seminar started with introductory talks covering major areas of IR evaluation.

Nick Belkin gave a survey over *Framework(s) for Evaluation (of whole-session) IR*, addressing the system components to be evaluated and the context to be considered.

In his presentation on *Modeling User Behavior for Information Retrieval Evaluation*, Charlie Clarke described efforts for improving system-oriented evaluation through explicit models of user behavior.

Kal Järvelin talked about *Criteria in User-oriented Information Retrieval Evaluation*, characterizing them as different types of experimental variables and distinguishing between output- and (task-)outcome related criteria.

Evaluation Measures in Information Retrieval by Norbert Fuhr outlined the steps necessary for defining a new metric and the underlying assumptions, calling for empiric foundation and theoretic soundness.

Diane Kelly presented problematic issues related to *Methodology in IR Evaluation*, such as the relationship between observation variables and criteria, the design of questionnaires, the difference between explanatory and predictive research and the appropriateness of statistical methods when dealing with big data.

The round of introductory talks was concluded with Maristella Agosti's presentation *Future in Information Retrieval Evaluation*, where she summarized challenges identified in three recent workshops in this area.

After a general discussion, the participants decided the topics to be addressed in small focused groups. For the rest of the week, the participants then formed the working groups described in the following.

4 Working Groups

From Searching to Learning

This working group focused on the learning as search outcome and the need for systems supporting this process. Learning may occur at two different levels, namely the content level and the search competence level. There is a need for understanding of the learning process, its relationship to the searcher's work task, the role of the system, and the development of appropriate evaluation methods. Approaches may address different aspects of the problem, such as the system, the interaction, the content, the user and the process. For evaluation, the framework in [3] can be used as a point of departure for developing criteria and measures at the levels of information retrieval, information seeking, the work task and the social-organizational and culture level.

Social Media

Social media allow users to create and share content, with a strong focus on personal connections. While web search engines are still the primary starting point for many information seeking activities, information access activities are shifting to more personalized services taking into account

social data. This trend leads to new IR-related research issues, such as e.g. utility, privacy, the influence of diverse cultural backgrounds, data quality, authority, content ownership, and social recommendations. Traditional assumptions about information seeking will have to be revised, especially since social media may play a role in a broad range of information spaces, ranging form everyday life and popular culture to professional environments like journalism and research literature.

Graph Search and Beyond

The work of this working group started from the observation that an increasing amount of information on the Web is structured in terms of entities and relationships, thus forming a graph, which, in turn allows for answering more complex information needs. For handling these, search engines should support incremental structured query input and dynamic structured result set exploration, Thus, in contrast to the classical search engine result page, graph search calls for an incremental query exploration page, where entries represent the answers themselves (in the form of entities, relationships and sub-graphs). The new possibilities of querying and result presentation make necessary the development of adequate evaluation methods.

Reliability and Validity

The concepts of reliability and validity are considered as the most central issue in IR evaluation, especially in the current situation where there is increasing discussion in the research community about reproducibility and generalizability of experimental results. Thus, this working group decided to start the preparation of a book on best practices in IR evaluation, which will cover the following aspects: basic definitions and concepts, reliability and validity in experimentation, reporting out experiments, failure analysis, definition of new measures and methods, guidelines for reviewing experimental papers.

Domain Specific Information Retrieval

Information retrieval in specific domains like, e.g., in cultural heritage, patents and medical collections is not only characterized through the specifics of the content, but also through the typical context(s) in which this information is accessed and used, which requires specific functionalities that go beyond the simple search interaction. Also context often plays an important role, and thus should be considered by the information system. However, there is a lack of appropriate evaluation methods for considering contexts and new functions.

Task-Based IR

Task-based IR typically refers to research focusing on the task or goal motivating a person to invoke an IR system, thus calling for systems being able to recognize the nature of the task and to support the accompanying search process. As task types, we can distinguish between motivating, seeking, and search tasks. Task-based IR approaches should be able to model people as well as the process, and to distinguish between the (task-related) outcome and the (system) output.

Searching for Fun

The searching for fun subject calls attention to the interaction with an information system without a specific search objective, like, e.g., online window shopping, watching pictures or movies, or reading online. This type of activity requires different evaluation criteria, e.g., with regard to stopping behavior, dwell time and novelty. Also, it is important to distinguish between system criteria and user criteria, where the latter may be subdivided into process criteria and outcome ones. A major problem in this area is the design of user studies, especially since the starting points (e.g. casual or leisure needs) are difficult to create under experimental conditions. A number of further related issues was also identified.

The Significance of Search, Support for Complex Tasks, and Searcher-aware Information Access Systems

This working group addressed three loosely related challenges. The first challenge addresses the definition of IR in the light of the dramatic changes during the last two decades, and the limited impact of our research. The second challenge considers the development of tools supporting more complex tasks, and their evaluation. Finally, information systems should become more informed about the searcher and the progress in user's task.

Interaction, Measures and Models

This last working group discussed the need for a common framework for user interaction models and associated evaluation measures, especially as a means for achieving a higher degree of reliability in interactive IR experiments. This would allow for evaluating the effect of the interaction and the interface on performance. A possible solution could consist of three components, namely an interaction model, a gain model and a cost model.

5 Outcomes and Future Work

The seminar participants operated to increase the understanding of the central problems in evaluating information retrieval, trying to achieve a cross-fertilization of ideas in the evaluation approaches from the different IR evaluation communities, identifying relevant aspects to take into account for creating new methodologies and approaches for solving existing problems. As main objectives, they worked on enhancing the validity and reliability of future evaluation experiments, and examining how to extract pertinent IR systems design elements from the results of evaluation experiments, in the long run.

Finally, many of the attendees were planning to continue to collaborate on the topics addressed during the seminar since the fruitful discussions were a useful base for future cooperation, so we expect to see relevant outcomes to be produced by groups of participants in the next future.

Acknowledgments

We would like to thank all the participants to the seminar who gave an extremely valuable contribution and made the seminar a success.

We also thank the *Schloss Dagstuhl* - Leibniz Center for Informatics who has made this seminar possible.

The work of Maristella Agosti has been partially supported by the CULTURA project, as part of the Seventh Framework Programme of the European Commission, Area Digital Libraries and Digital Preservation (ICT-2009.4.1), grant agreement no. 269973[4].

References

[1] M. Agosti, N. Fuhr, E. Toms, and P. Vakkari. *Evaluation Methodologies in Information Retrieval*. Dagstuhl Reports, Volume 3, Issue 10, 2013. `http://drops.dagstuhl.de/opus/institut_dagrep.php?fakultaet=07`.

[2] D. Harman. *Information Retrieval Evaluation*. Morgan & Claypool Publishers, 2011.

[3] P. Ingwersen, and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*. The Information Retrieval Series, Vol. 18, Springer, 2005.

[4] T. Saracevic, and L. Covi. Challenges for Digital Library Evaluation. In D. H. Kraft, editor, *Knowledge Innovations: Celebrating Our Heritage, Designing Our Future. Proceedings of the 63rd Annual Meeting of the American Society for Information Science*, pages 341–350. American Society for Information Science, Washington, D.C., USA, 2010.

[4] `http://www.cultura-strep.eu/`