Towards a Living Lab for Information Retrieval Research and Development

A proposal for a living lab for product search tasks

Leif Azzopardi[1] and Krisztian Balog[2]

[1] School of Computing Science, University of Glasgow
Leif.Azzopardi@glasgow.ac.uk
[2] Dept. of Computer and Information Science, NTNU, Trondheim
Krisztian.Balog@idi.ntnu.no

Abstract. The notion of having a "living lab" to undertaken evaluations has been proposed by a number of proponents within the field of Information Retrieval (IR). However, what such a living lab might look like and how it might be setup has not been discussed in detail. Living labs have a number of appealing points such as realistic evaluation contexts where tasks are directly linked to user experience and the closer integration of research/academia and development/industry facilitating more efficient knowledge transfer. However, operationalizing a living lab opens up a number of concerns regarding security, privacy, etc. as well as challenges regarding the design, development and maintenance of the infrastructure required to support such evaluations. Here, we aim to further the discussion on living labs for IR evaluation and propose one possible architecture to create such an evaluation environment. To focus discussion, we put forward a proposal for a living lab on product search tasks within the context of an online shop.

1 Introduction

Evaluation is a key challenge within the field of Information Retrieval (IR) [14]. From early on in the history of IR, objective and precise ways to measure, compare and evaluate systems, methods and models have been central to the research conducted [13, 14]. The main advances have been through the dedicated efforts to form consortium that build and develop test collections, methodologies, and measures (such as CLEF, TREC and INEX). While test collection based research has been of great benefit to the IR community, allowing researchers to study a variety of task and domains, they do have a number of limitations [14]. The abstractions often lack realism, there is often no user/user model, nor any interaction [3, 9]. As such, ever more complicated measures that try to incorporate the user into the way that IR systems are evaluated have been developed [12]. However, to properly test IR systems, evaluation needs to be performed in context (i.e., with real users performing tasks using real-world applications). So one alternative that has been recently proposed is the introduction of "living labs"

that involve and integrate users within the research process [7, 9]. This would, not only, enable the capture of real interaction and usage data, but also provide a context for testing and evaluating IR models, methods and systems. In Kelly et al. [9], they outline what such a lab might offer, be, and enable:

> *A living laboratory on the Web that brings researchers and searchers together is needed to facilitate ISSS [Information-Seeking Support System] evaluation. Such a lab might contain resources and tools for evaluation as well as infrastructure for collaborative studies. It might also function as a point of contact with those interested in participating in ISSS studies.*

According to Pirolli [11], having such a living lab available for research purposes would be,

> *a great attractor for scientific minds in diverse areas ranging from behavioral economics, incentive mechanisms, network theory, cognitive science, and human computer interaction...*

From discussions at the SIGIR 2009 Future Information Retrieval Evaluation workshop [7] , there was a clear desire from participants to be able to understand user information-seeking behavior in situ and the idea of a living lab as a way to do this was generally endorsed. It was also seen as a way to bridge the **data divide** within the research community, because currently interaction data is often only available to those working within organizations that provide real-world IR applications. A living lab would provide a common data repository and evaluation environment giving researchers (in particular from academia) the data required to undertake meaningful and applicable research. More generally though, a living lab has been presented not just as a platform for collaborative research, but also as a platform where users co-create the product, application or service (i.e., users are not just subjects of observation, but also part of the creation). Essentially, the users explore emerging ideas and scenarios in situ, the evaluation process is then fed back into the design of the product to further enhance their user experience[3]. While living labs have lots of appeal offering a number of opportunities and benefits, the development and implementation throws up some difficult challenges and problems which need to be overcome before such an evaluation platform can be realized.

The contributions we make in this paper are twofold. First, we propose one possible system architecture for a living lab based on a number of distinct web based services that provide a level of independence between the different parties involved in the research and development cycle (i.e., academics, commercial organizations, evaluation forums and users). While this is a rather idealized architecture of an IR focused living lab, it provides a starting point for serious discussion about how to implement such an idea. Second, we propose a living

[3] The concept of living labs is attributed to Jarmo Suominen. See `http://staffnet.kingston.ac.uk/~ku07009/LivingLabs/PapersAndSlides/Day1RichardEnnals.pdf` for an explanation and some of the history regarding the concept of living labs.

lab evaluation platform for an online shopping scenario. This scenario provides: (i) a novel set of search tasks, which have not received much attention in current evaluation forums, (ii) a problem where the size and scale is significantly more tractable than other tasks, such as web search i.e., an online retailer houses information on only a few thousand products, for which there is lots of rich interaction data, whereas the web contains billions of documents and large volumes of interaction data, (iii) product search data is not as problematic when it comes to privacy of the user (i.e., product search is can be made anonymous much more easily), (iv) the tasks in this scenario have direct economic implications, and (v) it provides an incentive for smaller online retailers to participate as they can benefit from research and development activity they could not otherwise afford. We hope that this work stimulates interest in the development of a living lab and leads to the creation of such an evaluation platform.

The remainder of this paper is structured as follows. In Section 2 we consider some of the potential steps or stages from standard test collections to living labs. In Section 3 we present an idealized system architecture for the development of a living lab that would facilitate a closer integration between researchers and industry. Then, in Section 4 we describe our proposal for an online shopping living laboratory. Finally, we conclude with a discussion on the benefits and challenges involved, before outlining the next steps in developing a living lab for IR research and development in Section 5.

2 From Standard Test Collections to Living Labs

There has been a number of different developments and proposals for Information Retrieval evaluation platforms. We shall briefly present the main approaches. They range from the Standard Test Collection approach (which typically adopts the Cranfield paradigm) to Fully Intergrated Living Labs. At each step the platforms become more and more application/user focused.

- **Standard Test Collection.** A testbed containing documents, topics and relevance judgments which allows for rigorous and replicable testing of methods, models and theory. Most TREC/CLEF/INEX collections are representatives of this type of test set.
- **Extended Test Collections.** A test collection augmented and extended by conducting a series of experiments that involve users. The usage and interaction data is recorded and distributed as part of the collection. The TREC Interactive track [5] and later the HARD track [1] both attempted to bring in the user into the loop. Although these tracks struggled to establish comparability between experimental sites, they were successful at highlighting the importance of users in IR research [15].
- **Simulation of Interaction.** Following on from the extended test collection, users and interaction are seeded, simulated and validated against the usage data in the extended test collection [2]. Alternatively, an abstracted task model could be developed (ranging from a simple search task to a more complex exercise that might not be solved in a single session) and researchers submit "simulated users" to perform that task [2].

- **Observational Test Centers.** Here users of an application would be logged and monitored (and depending on the setup this may be without the user's consent or knowledge). An observational test center would be able to build up a rich set of usage and interaction data (such as query logs) which could be used for research purposes.
- **Sandboxes.** A fully working application which can be modified by researchers to facilitate different configurations and permutations. IR toolkits (such as Lucene, Lemur or Terrier) may be viewed as lab or system based sandboxes, where one can experiment with varying some components. In the setting of this paper, our primary focus is on application based and human focused sandboxes; these enable researchers to vary and change various components of interest in an application. These changes can then be evaluated with users who volunteer to trial a different version of the live application.
- **Fully Integrated Living Lab.** The ideal scenario where users are not only observed, and researchers change configurations to perform experiments, but they are also part of the research process, and co-create the application or service through their usage behavior. Arguably, web search engines are already living labs, though their experimentation is performed strictly behind closed doors.

The above steps represent the continuum from system focused to application/user focused research and map to the spectrum provided by Kelly [8] where test collections are largely system focused, while living labs are on the opposite end the spectrum and largely application focused. Our focus throughout this paper is on developing a fully integrated living lab, where we will primarily concentrate on the high-level design of the machinery required to facilitate a living lab[4]. In the next section we shall outline one possible system architecture to support a living lab evaluation platform, before describing how it would be applied in the context of an online shop in Section 4.

3 A System Architecture for Living Labs

In Figure 1, we outline a high level system architecture that includes test centers, sandboxes and a fully integrated living lab. The architecture is somewhat idealized consisting of four independent web based services that would cooperate together for mutual benefit. Service **A** is the web based evaluation forum that coordinates evaluation efforts among researchers and acts a broker between the live applications provided by services **B** and **D** and the research services developed, **C**. Service **B** facilitates access to the commercial web application and would provide the interaction and usage data (this vetted data is then supplied to Service **A**). Services of type **C** are the web based services that researchers develop. They interact with **A** to obtain data for the particular evaluation search task. Service **D** encompasses non-commercial applications for testing and evaluation with users out with the commercial application. User would interacts with

[4] For an excellent survey and practical guide on running controlled experiments within a living lab, we refer the reader to work performed by Kohavi et al [10].

the lives applications (which are denoted by diamond shapes in Figure 1); these are the Live (or Living Lab) Applications, the Test Center Applications and the Sandbox Applications. The usage data produced would be stored by service **A**, potentially along with explicit relevance data, to enable evaluations to be conducted. Next, we shall describe each service in more detail.

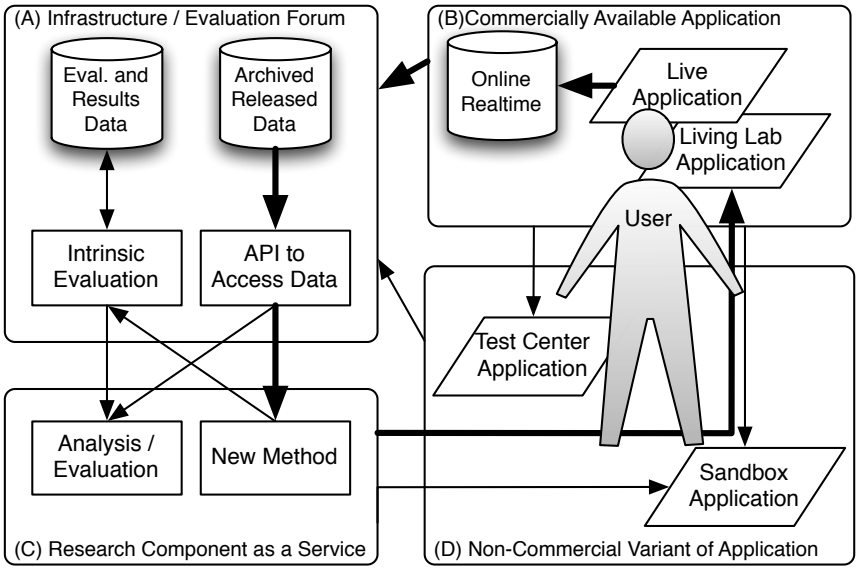

Fig. 1. A Possible System Architecture for a Living Laboratory. The thick black arrows denote the cycle of interactions required for a living lab.

A) Infrastructure/ Evaluation Forum. This web service provides a proxy between a Live Application (**B**) and the developers and researchers of various components (**C**). It connects to **B**'s API and receives updates on the data generated by the Live App (secure one way transfer). This could happen periodically, perhaps monthly, or even continually. This Archived and Released Data repository would provide the means to perform various evaluation tasks (and might include documents, query logs, click-through data, etc.). Service **A** would also collect and collate evaluation data of different research components (which are registered with the infrastructure). It would provide two main APIs to its users, i.e., the researchers and developers of new components (**C**). One API would provide access to the data that is housed in the archive ("Data API"). The other is to provide an API to Intrinsic Evaluations that can be performed using the data within **A**. For example, this may compare the differences and similarities of results produced by the new Research Component against other existing Research Components (via a "Task API"). Being an evaluation forum/platform, Service **A** would allow Research Component Services to be registered and evaluated. If a Test Center or Sandbox Application is used, then usage data and judgements

from particular tasks could also be included within **A** to facilitate evaluation (without a living lab, or full cooperation from the providers of a live application).

B) **Commercially Available Application.** A company that delivers online services, like a search engine, online shop, etc. runs a live application, and to participate needs to supply data to **A** (once the data is vetted and moderated). End users interact with their Live Application to produce usage data. **B** may also incorporate into their application a living lab, where the live application is augmented by utilizing Research Components developed and accessible via web services of type **C** (assuming that these services can reliably and robustly handle the request and demands placed upon it by **B**). Feedback and usage data collected from the users would be collected and again exported to the evaluation forum (for the developers of **C** to analyze).

C) **Research Components as Web Services.** The developers of new components would interact with Service **A** to obtain the latest data available. A new component would use this data to perform the particular task (e.g., estimate the ranking of documents, summarize sentences, etc.). The Research Component web services would provide an API that exposes their method in a standard way for a particular task. For example, it accepts a query, and responds with a set of results in a pre-defined format. The new method developed could then be utilized by the other web services (**A**, **B** and/or **D**) as part of the Intrinsic Evaluation, Sandbox testing, or even used within a Living Lab Application.

D) **Non-Commercial Variants of Live Applications.** Here two types of applications could be created. One is a Test Center Application, which is essentially the Live Application provided by **B**, but which has been instrumented to obtain usage data (i.e., a client is created that exposes the functionality of **B** and decorates it with logging functionality). The usage data collected is exported to **A** for research and evaluation purposes. The other alternative is the Sandbox Application, where the API of **B** enables researchers to configure a variant of the Live Application to include the Research Components available through services of type **C**. Again, usage data from the Sandbox Application would be logged and provided to **A**.

These services could reside within one organization (to support in-house research), or may be distributed between the evaluation forum efforts, commercial organizations and research institutes. By breaking up the cycle into four major parts different organizations can be responsible for providing different services to facilitate research and development. This has the advantage that independence is maintained between parties. For example, the researcher of a new method can experiment and develop their algorithm without disclosing details of the algorithm, which they may wish to patent at some point. Alternatively, existing methods can be tested by invoking the API's of services of type **C** for the given task (assuming the web service is up and running) or the evaluation forum can collect evaluation results so that the performances of existing methods is available for comparison. Since researchers have access to the data, they can process

the data on their own machines, using their own representations, and with their preferred programming language(s)/toolkit(s)/etc.

Commercial organizations running a live application are also buffered from potential security risks because the access to the data is via a third party/proxy. While, test centers and sandboxes can be created where users can be recruited to test variants of commercially available applications. It should be noted that test centers could be run without any direct commercial involvement. For example, companies like Bing and Google provide APIs to their search engine, so it would be quite feasible to create a web interface that connects to their search API and logs all the interactions.

While, there are advantages to separating out these concerns, this architecture does introduce a number of overheads, such as creating the services, conforming to the defined APIs for tasks, and the increased complexity in development. However, web applications are becoming substantially easier to develop, and skeleton code could be made available to help researchers expose their components as services. So far we have talked very generally about a living lab; to help focus the discussion and make the problem more tractable we shall describe how the architecture would enable evaluation of product related search tasks.

4 A Living Lab for Online Shopping

Online shopping is an activity that is commonly and frequently performed. It is attractive for customers because of the high level of convenience, broader selection, competitive pricing and greater access to information [6]. Part of the process of shopping is finding vendors, browsing and searching for products, researching products, finding reviews about products, comparing prices and buying. These tasks are performed through an intermediary search service (such as a web search engine, or portal like eBay or Amazon) and/or through direct search services provided by the online vendor. Here, we shall consider the search and browsing performed within an online shop: where the vendor's main goal is to support customers to find the products that they are interested in, the related products and similar products, to improve their online shopping experience (and ultimately to drive and increase sales).

Tasks in an Online Shopping Environment. Let us imagine an online toy shop which has a large catalogue of products. We would like to be able to: (1) let customers find the products easily through a site-search component by (a) providing some query assistance, and (b) a good ranking of products that are "relevant" to the user given the query; (2) when a customer is viewing a product provide product recommendations such as displaying related and/or similar products. For example, Patrick visits this online toy shop and would like to purchase a remote control helicopter. He queries for "RC helicopters," the system provides a number of suggestions "RC helicopters valkyre," "RC helicopters apache," "RC helicopters parts," etc., where he chooses the first suggestion. The system then returns a set of products relevant and related to this query (i.e., a number of valkyre helicopter versions and models, perhaps the competing helicopter of the same type, and commonly purchased add-ons

such as batteries and blades). Patrick then selects a recent model. The system displays the web page for this product, which contains information about the product, price, ratings, etc. as well as related products such as batteries, body kits, blades, etc. for the currently selected product. Along on the page, similar products to the valkyre helicopter might also be displayed such as the apache helicopter, and other competing versions.

From this scenario there are three main search and recommendation tasks which are common to most online shops:

Query suggestion. We differentiate between two types of query suggestion. *Auto completion* refers to the functionality that recommends queries (displayed often in a drop-down list) as the user types in the search phrase; the feature is usually activated automatically after a certain number of characters are entered. *Query recommendation* is presented along with search results and offers alternative formulations of the original query; typical examples include spell correction and related searches. Displaying query recommendations is optional; as shown in sponsored search advertising, it is acceptable, and occasionally even desirable, not to show any suggestions [4].

Product search. The ability to search for products is a basic functionality that is essential for the ease and convenience of online shopping. Following the practice of web search engines, it is common to provide users with a single text input field (basic search). Many sites offer the option of advanced search, where users may put in additional filtering or matching criterion. Given the information need entered in either basic or advanced form, the search result page returned in response presents a ranked list of products, typically, along with the total number of hits and controls for paging between multiple pages of records.

Product recommendation. We distinguish between two product recommendation exercises. *Recommending similar products* is the task of offering various alternatives for a given product, which typically are displayed on the product's page. *Recommending related products* is the task of finding products that might be purchased along with the goods already selected by the user; such recommendations might be presented on any page, including product and category pages, search results (separated from organic results), and even the homepage. Product recommendations can be based on search keywords, similar items, cross-sell (related products), and up-sell (higher priced products).

How would the scenario of product search fit with the proposed architecture? Below we describe how each service might look or act given the system architecture described in Section 3.

A) Infrastructure / Evaluation Forum. The API for data would enable researchers to obtain: (1) product information, (2) usage and query log data, (3) anonymized user information, and potentially (4) trading logs. The high-level functionality for each data source is as follows.

Table 1. Search and recommendation tasks in an online shopping environment.

Tasks	Inputs	Outputs
Basic product search	keyword query	ranked list of products
Query auto-completion	keyword query	ranked list of query suggestions
Query recommendation	keyword query	ranked list of query suggestions (or none)
Similar products	current product	ranked list of products
Related products	current product	ranked list of products (or none)
Product Recommendation	previous products	ranked list of products

- *Product information*: access to the list and properties of products and product categories. Certain product attributes are common to all webshops (such as name, description and price), while others can be specific to the commercial segment or to the given vendor. The product description, ratings and perhaps even product reviews might also be included.
- *Usage and query log*: number of times a query was issued, follow-up queries, search results clicked, number of times a product/category page was visited, average time spent on a product/category page, etc.
- *Anonymized user information*: information about the current user, including pages visited so far and time spent on each, and content of the shopping basket. If the user can be identified (i.e., is logged in) also historical data for this person, such as previous purchases and favourited products.
- *Trade log*: number of purchases/pieces sold for a given product, cross-sales, most popular products, etc.

This data API can be used to develop services for the various product search and recommendation tasks (i.e., **C**). Table 1 summarizes the possible interfaces for these tasks, where in the case of the product search task, the task API takes a query as input and returns a ranked list of product ids as output. Once such a service is developed, researchers could then invoke the evaluation forum's intrinsic evaluation, that calls on their services, to evaluate the given component.

B) Commercially Available Application. The data from the live applications is supplied via this web service. This requires an online shop to participate in the living lab, supply the usage data and to trial research methods.

C) Research Components as Web Services. This web service defines a task API for each of the search and recommendation tasks addressed (i.e., Table 1). The developed methods then can be tested either using intrinsic evaluation (by using web services of **A**) or within the living lab application (utilized by services of **B**). Intrinsic evaluation allows for the component to be compared against other methods, as well as tested against any judgements acquired for the task from sandbox evaluation performed with assessors. However, evaluation within the living lab makes it possible to measure the user experience of customers over entire sessions, quantified e.g., in terms of time spent on the site, conversion rate or the sum of purchases made.

D) Non-Commercial Variants of Live Applications. The first variant is the Test Center Application, which allows users to observe and examine users' interactions with the live online shop system through usage logs collected. The second variant is the Sandbox Application; it is the implementation of a given task submitted by a researcher to be evaluated in the live system.

We have described the high level interactions between services for an online shop based living lab. Realization of this vision will require a substantive amount of negotiation between evaluation organizers, an online retailer and researchers to come to agreement about what can and can not be accommodated. The specific details about what data can be provided by commercial organizations will invariably determine what tasks can be acceptably outsourced to external researchers and under what conditions. If the conditions and restrictions are too great, then an alternative solution may involved setting up a dedicated commercial web application for research purposes. Assuming it is possible to amicably involve a commercial web application in the process, decisions about what search tasks and measures can be undertaken to define the APIs and types of intrinsic evaluations.

Here, we have only covered the high-level aspects regarding the design and development of a living lab for online shopping. While, still quite abstract, we hope this leads to some meaningful dialogue within the community and facilitates the development of a living lab for product search evaluation.

5 Discussion and Future Plans

In this paper we have outlined a potential architecture for developing the infrastructure to support a living lab in the context of IR evaluation. To provide a concrete example and propose a new evaluation track, we discussed what a lab might look like in an online shopping environment, for product search and related tasks. Central to the design is a distributed and flexible web based architecture (i.e., service oriented), and this means that a number of parties can cooperate in an independent fashion. However, there are a number of issues that such an evaluation platform would need to address.

What are the problems and challenges that face the development and use of living labs? There are a number of legal and ethical issues that need to be considered (such as, user consent and ethics approval of such research, legalities regarding the release of data, copyright issues, commercial sensitivity of interaction data, trust between parties), as well as privacy and security issues for the users and the commercial organizations (think AOL query log fiasco). A concern that may put off commercial organizations releasing data is the (perceived) commercial value of the said data and exposing part of their business processes. This may lead to competitors gaining an advantage. Legal, ethical and business issues aside, there are also a number of technical challenges which arise and range from design and implementation issues, to the cost of implementation, maintenance and adoption, to the reliability, robustness and provision of services. Here we have focused on the architecture to provide a possible design.

Once the machinery is designed and created, the next barriers are in terms of its adoption and use by researchers and developers, and importantly the cooperation and support of the commercial organizations involved. Management issues will invariably arise in the how the evaluation forum and infrastructure is managed, and who should be responsible for maintaining such as service. To resolve these issues will either require a dedicated group of volunteers and/or long-term funding to maintain, organize and coordinate services. However, it may be feasible to prototype living lab on a small budget, if it was run for a short duration and a limited number of participants.

What are the benefits of a developing a living lab? One of the key benefits for researchers would be the access to real interaction data and (a variety of) real application contexts (like product search). Evaluations would become more user focused, and enable many more tasks to be evaluated and explored. The methods developed by researchers would have the potential to improve business processes. Thus labs lend themselves to being a bridge between academia and industry providing a direct route to commercialization for researchers. Besides access to more data and commercialization, another benefit of a living lab is that it can facilitate the independent verification of research results. This is because the evaluation forum services and commercial organizations can validate the research independently. Commercial organizations that participate in such initiatives could also benefit from having access to research and development teams without the associated overheads. Improvements to the provision of their service could lead to substantial improvements to their bottom line. In particular, for smaller organizations (such as independent online shops) and non-profit organizations (such as ACM Portal, citeseer.com, etc.) that cannot afford research staff, participation means having access to expertise with minimal investment. Also, participation would enable organizations to perform controlled experiments with good return-on-investment [10]. With appropriate infrastructure that facilitates experimentation and evaluation organizations could also innovate faster and more effectively [10].

Outlook and future directions. These are only some of the challenges, issues and benefits regarding the creation and development of a living lab. The major problems that need to be overcome are: (1) the initial design and development of the infrastructure to support a lab, and (2) the commitment of an organization and access to their data. While, the costs of building and developing the infrastructure are likely to be quite high for a fully integrated living lab, it may be possible to create a light-weight or scaled-down version on a smaller budget. Secondly, having an organization agree and commit to providing the tasks and data to support those tasks being performed is required. This is where we believe focusing on smaller online vendors or services would be more successful, than trying to develop living lab for a web search engine. Smaller vendors have specific problems and rich interaction data and are often without the resources to invest heavily in research. To this end, we are currently discussing with small online retailers about participating in such an initiative. However, before we continue to develop this initiative further we would like to discuss the proposed

living lab on product search with the wider community; ascertain the level of interest, the potential concerns and inevitable constraints, as well as discuss the possibility of developing and organizing a product search evaluation campaign as part of a forum such as CLEF.

Bibliography

[1] J. Allan. Hard track overview in trec 2003: High accuracy retrieval from documents. In *Text REtrieval Conference*, 2003.

[2] L. Azzopardi, K. Järvelin, J. Kamps, and M. D. Smucker. SIGIR 2010 workshop on the simulation of interaction. *SIGIR Forum*, 44:35–47, 2011.

[3] N. J. Belkin. Some(what) grand challenges for information retrieval. *SIGIR Forum*, 42(1):47–54, 2008.

[4] A. Broder, M. Ciaramita, M. Fontoura, E. Gabrilovich, V. Josifovski, D. Metzler, V. Murdock, and V. Plachouras. To swing or not to swing: learning when (not) to advertise. In *Proc. of the 17th ACM conf. on Information and knowledge management*, CIKM '08, pages 1003–1012, 2008.

[5] S. T. Dumais and N. J. Belkin. The trec interactive tracks: Putting the user into search. In *Text REtrieval Conference*, 2005.

[6] S. L. Jarvenpaa and P. A. Todd. Consumer reactions to electronic shopping on the world wide web. *Int. J. Electron. Commerce*, 1:59–88, 1996.

[7] J. Kamps, S. Geva, C. Peters, T. Sakai, A. Trotman, and E. Voorhees. Report on the SIGIR 2009 workshop on the future of IR evaluation. *SIGIR Forum*, 43:13–23, 2009.

[8] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3:1–224, 2009.

[9] D. Kelly, S. Dumais, and J. O. Pedersen. Evaluation challenges and directions for info. seeking support systems. *Computer*, 42(3):60–66, 2009.

[10] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: survey and practical guide. *Data Min. Knowl. Discov.*, 18:140–181, 2009.

[11] P. Pirolli. Powers of 10: Modeling complex information-seeking systems at multiple scales. *Computer*, 42:33–40, 2009.

[12] M. Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval (FnTIR)*, 4(4): 247–375, 2010.

[13] K. Sparck-Jones. *Information Retrieval Experiment*. Butterworth & Co, 1981.

[14] E. M. Voorhees and D. K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[15] R. W. White, G. Muresan, and G. Marchionini. SIGIR 2006 workshop on evaluating exploratory search systems. *SIGIR Forum*, 40:52–60, 2006.