Datenbank Spektrum (2020) 20:5–15
https://doi.org/10.1007/s13222-020-00331-1

SCHWERPUNKTBEITRAG

Studies on Search: Designing Meaningful IIR Studies on Commercial Search Engines

Dirk Lewandowski[1] · Sebastian Sünkler · Sebastian Schultheiß

Received: 14 October 2019 / Accepted: 11 January 2020 / Published online: 24 January 2020
© The Author(s) 2020

Abstract

The purpose of this paper is (1) to show which topics are especially fruitful for researchers interested in user behavior in commercial search engines, (2) to help researchers decide which data to collect and to what extent. We classify potential areas for IIR research along two dimensions, namely the type of interaction data used (small-scale or large-scale), and whether search engine companies are likely to publish research on the topic chosen (likely or unlikely). This results in a framework consisting of five areas, which are further detailed. In the second part of the paper, we present some empirical studies showing how researchers could approach relevant topics where no results from the search engine providers themselves are published. We also show how researchers can improve the evidential value of their work by going from small-scale to at least medium-scale studies.

1 Introduction

In the field of (Interactive) Information Retrieval (IR and IIR, respectively), there are many examples for fruitful collaborations between researchers from academia and researchers from the search engine industry, as well as collaborations where researchers from academia were given access to commercial search engines' data (e.g., [1–5]).

In cases where researchers do not have access to data from a search engine provider, however, we often have to question the evidential value of the results obtained. If researchers conduct studies on commercial search engines that would have been better done with access to (often large-scale) data from search engine providers, we have to ask for the reasons these studies have been conducted at all. In the field of interactive information retrieval, we often find small-scale studies addressing research questions that could have been better answered using a large-scale dataset from one of the search engine providers. Still, there are several reasons for conducting studies focusing on commercial search engines: (1) Search engine providers do not publish studies of some topics relevant to (interactive) information retrieval, (2) Some research questions can only be answered conducting lab studies, i.e., researchers from search engine companies do not have advantages over researchers from academia in conducting these studies, (3) The research topic is one where the search engine provider(s) have an interest in not having results published, i.e., results may contradict their self-interests. By self-interests, we refer to predominantly commercial interests of search engine companies that may be in conflict with decisions on relevance or user interests. For instance, as search engine companies generate the vast majority of their revenue through advertising, their self-interest is to generate clicks on ads. We cannot reasonably assume that they will publish research if their studies show that users misleadingly click on ads assuming they are organic results. Another example is the integration of vertical search engines into the result pages of general-purpose search engines. Here, search engine companies have an interest in keeping their users on their own properties (i.e., referring them to their verticals). This has been extensively debated in the context of the European Commission's competitive investigation against Google [6].

In this paper, we focus on how to conduct studies focusing on user behavior in commercial search engines like Google or Bing. A general assumption of our research is that search engine providers are not only interested in providing the most relevant results and the easiest interaction paths towards these results to their users, but that they have self-interests which lead them to prefer certain types of results, as well as influencing their users through designing interactions that help achieve search engine providers' self-interests.

A general decision researchers have to make is to explicitly state who should benefit from their research. In

✉ Dirk Lewandowski
dirk.lewandowski@haw-hamburg.de

[1] Department of Information, Hamburg University of Applied Sciences, Finkenau 35, 22081 Hamburg, Germany

interactive information retrieval, the goal most often is to better understand users' interactions with search systems in order to gain knowledge on *how to optimize these systems*. This means that apart from aiming to produce generalizable findings derived from users' interactions with commercial search engines, a goal of these studies is to help search engine providers (or other providers of IR systems) to improve their systems. The question, therefore, is whether interactive information retrieval researchers interested in users' interactions with commercial search engines without access to datasets from one of these engines, should strive for the second-best solution, i.e., collecting interaction data on a small scale, or whether they should instead decide to investigate topics that are either not researched by search engine companies, or topics where search engine providers do not publish results because these would harm their self-interests. In this paper, we will focus on the latter and show how interactive information retrieval research can on the one hand contribute to the understanding of user interactions with search engines and, on the other hand, help foster a public understanding of these interactions involving the role that search engine providers play in knowledge acquisition and also concerning these companies' self-interests. Therefore, the purpose of this paper is (1) to show which topics are especially fruitful for researchers interested in user behavior in commercial search engines, (2) to help researchers decide which data to collect and to which extent.

The rest of this paper is structured as follows: Firstly, we will show how IIR studies investigating commercial search engines can be classified along the two dimensions "interaction data" and "likeliness of search engine companies to publish this type of study." Using the resulting framework, we will show research areas that may be especially fruitful to researchers interested in investigating commercial search engines. Subsequently, we will argue for a better understanding of search engine result pages (SERPs) to make studies more realistic. Then, we will present some studies from our research group, showing how IIR research can tackle questions not yet addressed due to lack of interest by search engine companies, and how the sample sizes of IIR studies can be increased considerably through software tools, crowdsourcing approaches, and collaborations with market research firms. In the conclusion, we summarize our approach and show further directions for research.

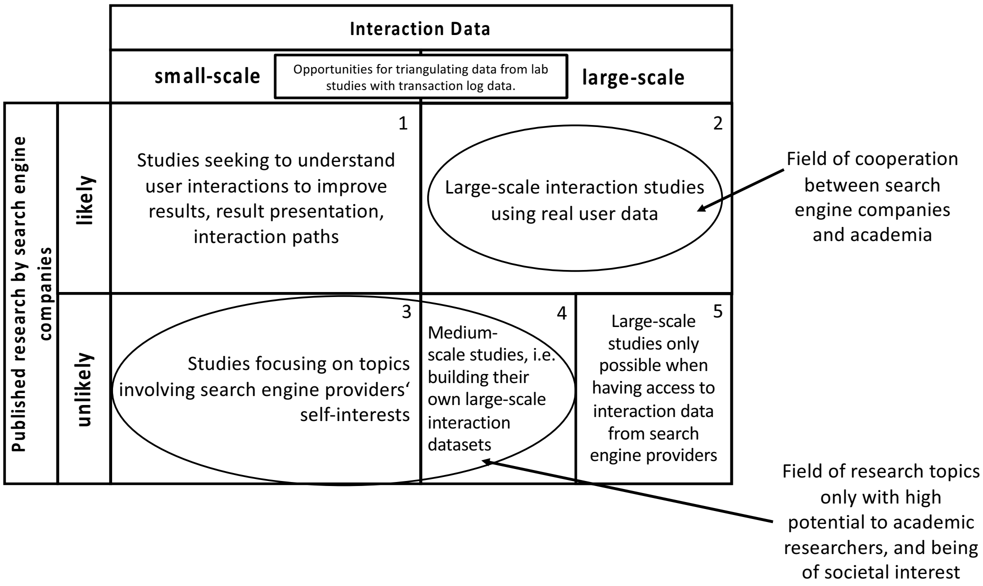
2 Classifying IIR Research on Commercial Search Engines

In Fig. 1, potential research areas are classified along two dimensions, namely the type of interaction data used (small-scale or large-scale) and whether search engine companies are likely to publish research on the topic chosen (likely or unlikely). This leads to four distinctive fields, one of which is further subdivided, resulting in a total of five distinctive fields. In the following, we will describe each field and in later sections, give an exemplary overview of research in these different fields, focusing mainly on studies conducted by our research group. The figure is intended to help researchers to decide which problems to tackle, and what type of interaction data to collect.

On the one hand, regarding the type of interaction data, research areas are divided into small-scale vs. large-scale studies. This admittedly is a rough differentiation, but still helps to get an understanding of the evidential value of different studies. In cases where researchers do not have access to data from search engine companies, we will introduce a further category, medium-scale studies. On the other hand, research can be divided into topics where it is likely or unlikely that search engine providers are interested in *publishing* research results. This does not mean that in the case of studies labeled as "unlikely," search engine providers are not interested in conducting such research, but rather that publishing the results may not be in their interest.

Distinguishing research topics/studies along the two dimensions leads to five areas (see Fig. 1), which we describe in the following.

1. *Likely/small-scale*: In this area we typically find research that is conducted in the laboratory, where usually only a small number of participants can be observed. With this type of research, it does not matter whether studies are conducted commercially or academically. While specific equipment such as eye-trackers or neurophysiological research tools [7] may be needed, search engine companies do not have an advantage over academic researchers. One should keep in mind, however, that when aiming to improve SERPs, commercial search engine providers like Google have already conducted extensive tests (mostly unpublished) to optimize user experience on SERPs. So, even if search engine providers did not publish research studies on certain elements of SERPs, it may still be better to follow their best practices than following results from small-scale academic studies, as one can assume that the design of SERPs has been tested extensively using A/B testing on a large scale whereas evidential value of academic lab studies is often restricted due to small sample sizes.
2. *Likely/large-scale*: In this area, we find studies using large-scale interaction data from real users. Obviously, only search engine providers collecting these data can carry out such research. Researchers from academia can only conduct such studies if they are granted access to datasets from search engine providers. There have been many studies where such access was granted. However,



Interaction Data

small-scale

Opportunities for triangulating data from lab studies with transaction log data.

large-scale

Published research by search engine companies

likely

unlikely

1 Studies seeking to understand user interactions to improve results, result presentation, interaction paths

2 Large-scale interaction studies using real user data

Field of cooperation between search engine companies and academia

3 Studies focusing on topics involving search engine providers' self-interests

4 Medium-scale studies, i.e. building their own large-scale interaction datasets

5 Large-scale studies only possible when having access to interaction data from search engine providers

Field of research topics only with high potential to academic researchers, and being of societal interest

Fig. 1 Dividing research on commercial search engines along the lines of interaction data and likeliness of search engine providers publishing research in that area

no search datasets are freely available to researchers[1], which means that researchers are not free in choosing their research topic when working with such data, but have to apply for access from a search engine provider stating their research topic and their need for the data. A further advantage to search engine providers is that they can triangulate their large-scale data with data from laboratory studies. For instance, they may derive groups from their transaction logs, and then further investigate users from these distinct groups in-depth in the lab.

3. *Unlikely/small-scale*: In the area where published research from search engine providers is unlikely, small-scale studies are those which focus on topics involving search engine providers' self-interests. Many such studies have been conducted (e.g., [8–11]). However, results from these studies are often taken into question due to small samples and thus not being representative, even more so when results from these studies are considered in policy decisions.
4. *Unlikely/medium-scale*: A possible solution to the ostensibly unsolvable problem of getting access to large-scale datasets is to conduct medium-sized, representative studies where researchers collect interaction data through task-based questionnaires and not through real user interactions from the field. This type of research addresses both the problem of obtaining interaction data from search engines as well as the criticism towards small-scale studies previously mentioned. A further advantage of medium-scale (and even small-scale) studies is that as researchers do not have to rely on a dataset from a single search engine comparisons between different engines are possible, as well.
5. *Unlikely/large-scale*: In this field, studies by academic researchers are, by definition, only possible when they have access to interaction data from search engine providers. However, as these types of studies are against search engine providers' self-interests, it is highly unlikely that these companies grant researchers access to their data, which has often been lamented.

In current research, there is more or less a blind spot in the lower-left quadrant and especially in the lower middle, i.e., medium-scale, representative studies. In the following, we will consider studies to be small-scale if they have less than 100 participants, medium-scale if they have participant numbers in the hundreds to thousands, and large-scale if they are based on thousands or even millions of user interactions. While these are only rough distinctions and categories may not be mutually exclusive, it will help exemplifying our framework. In Sect. 4, we will further detail this discussion by reviewing some studies in these different areas.

3 Understanding Search Engine Result Pages

When investigating user interactions in commercial search engines, it is crucial to understand the complex search result presentations these engines now provide. This relates to the design of the SERPs as well as to results coming from different sources (organic as well as paid-for), both of which are often poorly understood by users. While in some cases, it may be beneficial for researchers to strip

[1] When AOL released a search engine transaction log in 2006, it was soon found out that it was possible to identify individual users, and therefore, AOL retracted the dataset from public use.

down SERPs (e.g., only showing organic results to the participants of a study) for gaining control in experiments, we argue that in many cases, this leads to unrealistic user behavior which may not be transferrable to real search situations. When researchers are interested in the behavior of real users, they should use realistic SERPs. While stripped-down SERPs allow for more control, especially when the results for several tasks/SERPs are aggregated, this may result in unrealistic (or at least simplistic) models of search engine user behavior. For instance, when a researcher is interested in researching the influence of search results on users' opinion-forming, but strips advertisements from the SERPs, results may be biased because many users do not distinguish between ads and organic results when examining search results (see Sect. 3.4).

3.1 Design of Search Engine Result Pages

Search engine results pages (SERPs) consist of various result types. These are organic results, ads, universal search results, and knowledge-graph results. Organic results are generated by algorithms and ranked according to equal criteria. Text ads are context-based advertisements that match a query and closely resemble organic results. Google's Shopping ads differ from text ads in that they contain a product picture, the price and the name of the retailer, or other information. Universal search results are results from other, so-called vertical collections (e.g., news, maps, images, videos). Knowledge-graph results give factual information directly on the SERP in answer to various questions, such as questions about famous personalities [12, p. 422]. The design of SERPs is regularly revised by Google, e.g., by new functions like "infinite scroll" on mobile devices [13]. Besides, the elements of result snippets vary between results, i.e., some having additional information to the usual elements title, URL, and description.

3.2 Different Intents for Showing Results

A general assumption in IR is that the goal of every information system is to provide the user with the best, i.e., most relevant results. However, this does not necessarily apply to commercial search engines, as they have commercial interests in users clicking on certain results, whether it being ads which the search engine provider directly generates revenue from or results from the search engine provider's properties (like results from Google's subsidiary YouTube in the Google search engine).

As Ian Lurie [14] describes in his blog post on Search Engine Land, providing relevant results is a key component for search engines to increase their profits. Relevant results lead to satisfied users who will use the search engine again. As a result of more frequent use, the probability of ad clicks also increases. Also, Google is providing an increasing amount of information from external sources directly on its SERPs, such as information on symptoms and treatment options for diseases [15]. Thus, the user no longer necessarily needs to perform a click, which leads to a continually decreasing number of clicks on the organic results [16].

Commercial search engines such as Google are financed mainly through search-based advertisements. In 2018, Google generated a profit of 30.7 billion dollars on a turnover of 136.8 billion dollars. 83% of the turnover was generated by advertising [17]. Advertisers do not pay for the placement of the ad; instead, they pay only when a user interacts with it through a click [18]. Ads are labeled with a green ad label within a green frame (desktop search) or with a non-framed, black "ad" label (mobile search). Google regularly changes the labeling of ads, with a trend toward more subtle labels [19]. This trend is accompanied by an increase in the number of clicks on ads [20]. An additional info button is displayed for shopping ads on PCs and mobile devices as well as for mobile text ads, providing information on how the ads are generated [21].

3.3 External Influences on the Search Results (Through Search Engine Optimization)

Search Engine Optimization (SEO) describes external influences on the ranking of search engines: "SEO is the process of modifying a website in order to better satisfy a ranking algorithm and thus improving the chance of getting listed on the first search engine result page" [22, p. 1]. For the year 2020, a total turnover of 80 billion dollars is assumed for search engine optimization in the United States alone [23].

Some industry studies give insight into the influence search engine traffic has on the success of websites. An analysis of the 100 largest US shopping websites showed that around 45% of visits (traffic) was generated by search engines (organic: 21%) [24]. Also, many content providers acquire a significant proportion of their visits via search engines, especially via Google. The percentages for journalistic/informative content are often between 20–30% of all page impressions (e.g., nytimes.com (32%), washingtonpost.com (35%), and zeit.de (22%)). This traffic almost exclusively comes from organic clicks (99.3–99.9%) and is, therefore, influenced by SEO [25]. Very little is known about the actual impact of search engine optimization on search results. An ongoing research project ("SEO-Effekt[2]") funded by the German Research Foundation (DFG) aims to close this research gap.

[2] https://searchstudies.org/seo-effekt.

3.4 Users' Understanding of SERPs

On the one hand, users consider search engines as fair and unbiased sources of information [26]. They heavily trust in the search engines' ranking (e.g., [26–28]). On the other hand, the information literacy of search engine users can be regarded as rather low, as several studies have shown (e.g., users face difficulties in formulating precise queries [29] and solving complex tasks [30)]. As the results of a representative study show, the majority of users do not understand Google's business model and cannot reliably distinguish advertisements from organic results. The labeling of the advertisements thus does not seem to contribute to an understanding of this type of result [12].

This high level of trust in search engines combined with a poor understanding of SERPs and their elements can be considered problematic. Users with a low level of information literacy are confronted with highly professionalized external actors (in the areas of search engine advertising and search engine optimization), as well as the self-interests of the search engines as described above.

In summary, this brief analysis of SERPs shows that to achieve results on realistic user behavior, researchers need to better understand how SERPs are designed, how search engine providers use design features to influence user behavior, and which parties are interested in and capable of influencing search engine results.

4 Empirical Studies

In this section, we will go into more detail regarding the research areas identified through the framework presented in Sect. 2. We outline empirical studies that show how research questions where search engine companies do not publish results can be conducted when also aiming to overcome the limitations of small-scale studies. While briefly discussing studies in all areas of the framework, we will primarily focus on areas 3 and 4. One should note that the intention here is not to give a complete literature review but to give some examples of studies in the respective areas. We will predominantly use our own published research as examples.

4.1 Small-scale Studies Seeking to Understand User Interactions to Improve Results, Results Presentation, and Interaction Paths

As White [31] argues, different approaches in IIR evaluation have different strengths and weaknesses, and there is always a tradeoff in terms of scale and level of detail. So, for instance, log-based studies deliver a lot of interaction data from real users but fail in terms of understanding user intent, while more user-focused methods (such as interviews and focus groups) give such insights but with only a small amount of data. Thus, it is evident that there is a need for in-depth, small-scale studies in IIR. The question, however, is when to conduct such studies, i.e., not replacing large-scale log studies with small-scale lab studies just because of a lack of access to such data. Without wanting to shame any researcher in particular, we point out that we found many research studies in IIR and beyond that could have produced more meaningful results using transaction log data from search engine providers.

It may seem that when search engine companies are likely to conduct studies in a particular area, it may be useless for academic researchers to tackle that same area. However, as search engine companies do not have any advantages over academic researchers when conducting small-scale (lab) studies, we argue that it may be worthwhile for academic researchers to work in this field *as long as the results are at least to some extent generalizable, i.e., they are not only relevant to a particular search engine*. It should be noted here that questions of generalizability not only relate to whether results hold true for other search engines or search systems, as well, but also to sample sizes and the composition of the sample. As we will argue in Sect. 4.4, researchers should strive not only for larger but also for more diverse samples. In the following, the methods of eye-tracking and neuro-IR are briefly presented. Both approaches use devices where studies are difficult to scale but do not require large-scale interaction data for results that contribute to our understanding of user interactions.

Eye-trackers are devices that use infrared technology to measure what a person is looking at (fixations) and the sequence in which the eyes are shifting from one location to another (saccades). By using these devices, it is possible to determine the level of attention a person pays to a visual representation, such as a SERP [32]. In eye-tracking studies, various limitations have to be considered. A literature review on eye-tracking focusing on SERPs by Lewandowski and Kammerer [33] revealed three significant limitations. Firstly, most of the studies identified were low-scale studies with a median of 30 subjects, which is a threat to the validity of results. Secondly, there is a lack of comparability of stimuli. Thirdly, gaze data on small areas of interest (AOIs; e.g., ad label) are difficult to measure reliably due to measurement inaccuracies. According to the authors, larger sample sizes, controlled eye-tracking laboratory studies, and triangulation with other methods could be feasible solutions to these problems. Concerning these disadvantages, the question arises for which research questions the eye-tracking method provides an essential benefit. According to Lewandowski and Kammerer [33], this benefit is given in particular when current SERPs are investigated, which do not have to lead to user clicks. This is the case when the query is answered directly on the SERP, for example by instant answers [15]. Here, together with verbal protocols, eye-tracking can provide insights into user behavior that could not have been determined by using click-through data alone.

There is a growing interest in neuro-physiological (NP) methods in human-information information interaction (HII) and interactive information retrieval (IIR). This interest has been motivated by the limitations of traditional data collection methods. NP methods complement such traditional methods. They help to get a deeper understanding of HII, which can lead to new information search models that go beyond behavioral data, and to enable the development of neuro-adaptive IIR systems [7]. Laboratory experiments in this area are complex because of all the technologies needed to measure neurophysiological signals. Therefore, despite the often small sample size, the experiments are essential to develop extended information search models and to get more insights into search behavior. Examples of such studies are about neurophysiological signals regarding the relevance of information objects [34], predicting term relevance by brain signals [35], or creating neurophysiological models of the realization of information need [36].

As shown in this section, there is an opportunity for academic researchers conducting small-scale studies in an area where it is likely that search engine companies will publish research results when their research aims at achieving generalizable results about user interactions with search systems and not only focusing on a single search engine. Furthermore, using specialized equipment like eye-tracking and tools for measuring neuro activity could lead to deeper insights into search processes in general, taking a particular search engine as an example.

4.2 Large-scale Interaction Studies Using Real User Data

As already mentioned, in many cases, using large interaction datasets from search engine providers is the single best solution, whereas other data sources, i.e., collecting one's data, is only the second best. Large-scale datasets are also often used for conducting A/B tests, i.e., large-scale experiments. As these often only concern tiny changes to the SERPs, results are usually not published.

As a general rule, researchers without access to search engine data should not choose a research question that can better be answered using such transaction log data, at least unless one can assume that no research on the topic in question will be published either by the search engine companies or by researchers collaborating with them.

4.3 Small-scale Studies Focusing on Topics Involving Search Engine Providers' Self-interests

As said above, small-scale studies are often the only option, especially when they are hard to scale due to restrictions in lab time and equipment. As the first example of a small-scale laboratory study that investigates a topic targeting search engine providers' self-interest, we chose an eye-tracking study investigating the effect of knowledge of Google Ads on user behavior on SERPs [37]. As our prior research had found, users are not well able to distinguish between ads and organic results, and are not well informed about Google's business model, either. We found that users click on ads more often when their knowledge of Google's business model and ads is low (see Lewandowski et al., [12, 38], also see Sect. 4.4 Medium-scale studies focusing on topics involving search engine providers' self-interests). Based on these prior studies, eye-tracking was employed to investigate not only result selection but also users' gaze behavior, in order to gain a deeper understanding of users selecting or not selecting ads. The study considered behavior on the desktop screen as well as on the smartphone. Major results were that subjects with a low level of knowledge on search advertising are more likely to click on ads than subjects with a high level of knowledge. Moreover, participants with little knowledge showed less willingness to scroll down to the organic results. While the first result is a confirmation of prior research, the second result adds to the body of knowledge in that it shows that the two groups differ in their gaze behavior, which in turn influences what they click. Furthermore, the study revealed that there are significant differences in viewing behavior between SERPs on the desktop and on the smartphone screen. These can be attributed to the influence of the direct visibility of search results on both devices tested. To strengthen the evidential value of this study, we, one the one hand, used a sample of 100 participants, which could be considered large compared to other eye-tracking studies (cf. Lund [39]; Lewandowski and Kammerer [33]). On the other hand, we aimed for a diverse sample consisting not only of students (as is common in laboratory research) but people from outside the university, as well. Results indicate that the behavior of non-students was quite different from that of the students, and therefore, had we only considered students, results would also have been different.

The second example of a study in Sect. 3 of our framework is a comparison of the relevance of Google and Bing results using small-scale data collected in the lab [40]. The importance of the topic lies in the fact that Google is by far the most frequently used search engine, but it is not known whether its results are better since there are hardly any suitable methods to carry out realistic and reliable comparisons of search engines in a natural setting. We proposed a method and conducted a user study aiming to allow for such comparisons. The aim was to find out how Google and Bing perform when users work freely on pre-defined tasks, and then judge the relevance of the results immediately after finishing their search session. In a lab-based user study, 64 participants each performed two search tasks and then assessed the quality of the results (1) they had selected, (2) they were presented but did not click, and (3) from a competing search engine whose results they did not see in their search session. The subjects were free to choose the search engine as well as to formulate their queries. After each task, participants assessed the relevance of the first ten search results generated by Google and Bing for their previous query. For querying the search engines, collecting the results, and allowing the subjects to assess the relevance of the results, we used a self-programmed tool (Relevance Assessment Tool[3]; [41]). One key finding was that Google's results were considered slightly more relevant than the results from Bing. In addition, in most cases users selected results that they later considered relevant. While the laboratory study with a small sample is subject to limitations, our research offers a framework for studying actual users' behavior in correlation with relevance judgments in more depth than before. With the Relevance Assessment Tool, it is easy to significantly increase retrieval studies in any environment and thus transfer small-scale studies (Sect. 3) into medium-scale studies (see Sect. 4.4).

4.4 Medium-scale Studies Focusing on Topics Involving Search Engine Providers' Self-interests

Increasing sample sizes in lab-based studies always has its limitations in the time needed per participant and in the lab situation itself, especially when particular equipment is needed. In other types of studies, where participants do not need to come to the lab but can participate online, the bigger question is how researchers can achieve samples that are of comparable quality to transaction log data from search engine providers. It should be mentioned here that, notwithstanding limitations in these approaches, they could even result in better (i.e., more representative) samples than data from the search engine companies, which are, by definition only representative of that company's users. In this section, we will show how on the one hand, sample sizes in more traditional settings can be increased considerably, and on the other hand, how medium-sized samples can be built for research in IIR settings.

Firstly, we describe a study with the aim to improve methods for search engine retrieval effectiveness studies [42] through increasing data sets using crowdsourcing. It should be noted, however, that for this study, we had access to some data from a search engine provider. We included it nevertheless as the aim in the context of this paper is to show how the number of (relevance) judgments in more traditional juror-based IR studies can be increased. Two random representative samples with 1000 informational queries and 1000 navigational queries, respectively, were used, both from the German search portal T-Online. For the informational queries, the top 10 results were collected from Google and Bing. Jurors were given all results for a query in random order and judged their relevance. For collecting the results, as well as distributing them to users and collecting their judgments, the Relevance Assessment Tool [41, 43] was used. For navigational queries, only the first result was collected, since navigational queries can be defined as those where a clear distinction can be made between one correct and other irrelevant results. A research assistant classified the results as either correct or incorrect. The results show that although Google outperforms Bing in both query types, the difference in the performance for informational queries was rather low. However, for navigational queries, Google found the correct answer in 95.3% of cases, whereas Bing only found the correct answer 76.6% of the time. It can be concluded that search engine performance on navigational queries is of great importance because users, in this case, can clearly identify queries that have returned correct results. As described in the previous section, results from small-scale studies are often questioned due to their sample sizes. Also, most studies use a small number of queries selected by the researchers, while the selection is not representative of the queries entered in real search engines. Both problems are addressed in this study, using the Relevance Assessment Tool to analyze 1000 queries of each sample efficiently. This study thus well illustrates the transition from a small- to a medium-scale study.

The second example is a set of representative online studies assessing the commitments proposed by Google as part of an EU competition investigation. The study was carried out identically for Germany, Spain, France, and Italy (for a country comparison report and individual reports of the countries, see [44]). The charge was that Google abused its market dominance to promote its own comparison shopping service in the search results while demoting those of rivals. Google attempted to address the Commission's concerns by proposing changes in the presentation of the offerings [6]. The aim of the studies, therefore, was to analyze whether the proposals made by Google could address the Commission's concerns. Google's suggestions for labeling its vertical search services and the placement of rival offerings formed the basis of the investigation. In each country, $N = 1000$ internet users took part in the study, whose

[3] https://searchstudies.org/relevance-assessment-tool/.

core was a click study based on screenshots provided by Google. Findings for Germany are that users to a large degree clicked on Google's vertical results, ignored the info icon, and to a large extent clicked on Google's vertical results even if they were explicitly asked to click on a rival result (Report for Germany, see [41]). The results show that the effectiveness of the proposals must be considered as insufficient. This opinion was also shared by the European Commission, which sentenced Google to a fine of 2.42 billion Euros [6]. The studies and their legal consequences for Google presented above show the regulatory impact that medium-scale studies can achieve. Similar studies in the laboratory with much smaller samples (in this context, e.g., [45]) would not have guaranteed the necessary foundation of the results. Increasing the sample size in our studies to 1000 participants per country was possible through hiring a market research firm. To our knowledge, this approach had not been taken in IIR research so far, even though it could lead to an increased external validity of results. Furthermore, online market research using representative samples can be conducted at a reasonable cost nowadays.

The third example is a medium-scale study on users' understanding of search-based advertising. It consists of a survey, a task-based user study, and an online experiment with $N = 1000$ German search engine users. First, we will describe the survey and the task-based study [12]. Search engine companies generate their revenues through user clicks on search-based advertisements. These paid results are very similar in appearance to organic results, as Google's search results show. Except for an ad label, whose design has become increasingly subtle in recent years, no difference is noticeable. The research questions we therefore raised were how search engine users think search engine companies make money and whether users can distinguish between paid advertising and organic results. To answer both questions, we conducted a survey with questions about Google's business model while also containing tasks that asked users to differentiate between ads and organic results. In the latter, participants were asked either to mark all ads or all organic results on SERP screenshots. The results show that users' knowledge of Google's business model is very limited. Only 38.8% of users correctly stated that it is possible to pay Google for a preferred listing of one's company on the SERPs and also knew how ads differ from organic results. Also, the results show that only a small percentage of users can reliably distinguish ads from organic results. 1.3% of participants were able to identify all results correctly, while 10.9% of users made no incorrect identifications but did not mark all results that should have been marked. We conclude that ads are insufficiently labeled as such and that many users may click on ads assuming that they are selecting organic results. Based on the previously described survey and task-based study [12], an online experiment was conducted using the same representative sample of $N = 1000$ German search engine users [38]. In the experiment, users' selection behavior was compared on two versions of the same Google SERP, one showing advertisements and organic results, the other showing organic results only. For both versions, only organic results were used, i.e., the ads shown in the experimental condition were actually organic results, only with an ad labeling. We investigated whether users' knowledge of Google's business and ads and users' ability to correctly marking ads (survey and task-based study, see [12]) influences their selection behavior (online experiment). We found that users who were not able to mark ads correctly in the task-based study selected ads around twice as often in the experiment as users in the knowledgeable group. In the control condition (where only organic results were shown), users who knew how Google makes money chose the first position significantly more often than users without that knowledge. This may be explained by the fact that these users noticed that there were no ads on the SERP and therefore regarded the first result as trustworthy. Regarding the result of significantly more clicks on ads by users with little knowledge, we argue that ads need to be labeled more clearly and that more information literacy is needed among search engine users. As with the study in the context of the antitrust case described above, users' understanding of ads is a question of societal interest. Contrary to legal requirements, the ad label is not understood by users, and thus ads are clicked on under false assumptions. It can be assumed that search engines deliberately blur the lines between paid and unpaid search results in order to achieve higher revenues through ad clicks. Obviously, it is not in the interest of search engines to publish such results themselves. Hence, it should be an objective of IR research to identify such problems and to investigate the impact this has on knowledge acquisition in society.

4.5 Large-scale Studies Focusing on Topics Involving Search Engine Providers' Self-interests

As per definition, large-scale studies using data from search engine providers but still investigating topics where these providers have an interest not to have results published are highly unlikely. Large-scale transaction log studies are often regarded as being the gold standard for investigating interactions with search systems. One should keep in mind, however, that these studies are only based on data from one search engine, and it may be questionable whether the results obtained apply to other search systems, as well. Furthermore, transaction log studies can by definition only address active users of a particular system.

5 Suggestions for Future Research

In the following, we provide some suggestions for further research. Due to the focus of our paper and the high academic potential, we focus on areas 3 and 4 of Fig. 1. As mentioned in Sect. 4.1, small-scale studies employing eye-tracking are particularly beneficial when no clicks are available for analysis. Using this approach, researchers could investigate in area 3 (focusing on topics involving search engine providers' self-interests) how intensively the direct answers (featured snippets) included in the SERPs are perceived and how this affects the distribution of views on the other parts of the SERP.

In area 4, we suggest studies that further shed light on users' understanding of the SERP. For instance, in the "SEO-Effekt" research project, we are currently working on a representative online survey with $N = 2000$ German internet users, examining the users' perspectives and their understanding of search engine optimization. Further surveys, which also cover the understanding of elements such as universal search results, would lead to a better overall picture of user behavior with regard to the growing complexity of SERPs. Another field that could be well addressed by representative surveys is the topic of trustworthiness of information that is found online. By means of a task-based questionnaire, it could be experimentally investigated whether internet users trust information found, for example, on Facebook more than the same information received via active search in a search engine.

Of course, a systematic review of papers in the research areas mentioned in this paper would also be desirable. This could include an analysis of the evidential value of the studies conducted and, therefore, identify fruitful areas not only for future work but also for areas where evidence needs to be strengthened (e.g., through larger-scale replications of existing research).

6 Conclusion

In this article, we showed how researchers in the field of (interactive) information retrieval who do not have access to data from a commercial search engine can still address meaningful research questions related to these engines and collect data for their purposes. We stressed that this does not necessarily lead to small-scale studies, but that, primarily through hiring market research firms, larger (and representative) sample sizes can be achieved. Furthermore, market research firms can address specific socio-demographic groups at a reasonable cost. The basis for designing IIR in the context of commercial search engines is a solid understanding of the composition of search engine result pages. The evidential value of studies on IIR can be improved when realistic SERPs are used.

A limitation of conducting medium-sized online studies (and experiments in particular) is that researchers have less control over their participants and the research setting. While there is certainly a tradeoff between increasing the sample size and providing more realistic SERPs on the one hand, and control, on the other hand, we are confident that the evidential value of IIR studies can be increased by aiming for medium-sized, representative online studies.

Another limitation is that in the studies reported, interaction was mainly modeled as user clicks on the SERPs. Further interaction was only considered in the interactive study on search engine comparisons [40]. Nevertheless, modeling more complex user interactions in online studies is not impossible. As market research firms allow researchers to incorporate HTML code and the like into their questionnaires, researchers should be free to model more complex interactions. We intend to move further into that direction in future studies.

Funding Open Access funding provided by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Agarwal A, Zaitsev I, Wang X, Li C, Najork M, Joachims T (2019) Estimating Position Bias without Intrusive Interventions. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. WSDM, vol 19. ACM Press, New York, New York, USA, pp 474–482 https://doi.org/10.1145/3289600.3291017
2. Chuklin A, Schuth A, Hofmann K, Serdyukov P, de Rijke M (2013) Evaluating aggregated search using interleaving. In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. CIKM, vol 13. ACM Press, New York, New York, USA, pp 669–678 https://doi.org/10.1145/2505515.2505698
3. Goel S, Broder A, Gabrilovich E, Pang B (2010) Anatomy of the long tail. In: Davison BD, Suel T, Craswell N, Liu B (eds) (Eds.), Proceedings of the third ACM international conference on Web search and data mining. WSDM, vol 10. ACM Press, New York, New York, USA, p 201 https://doi.org/10.1145/1718487.1718513
4. Montanez GD, White RW, Huang X (2014) Cross-Device Search. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. CIKM, vol

14. ACM Press, New York, New York, USA, pp 1669–1678 https://doi.org/10.1145/2661829.2661910
5. Nikolov D, Flammini A, Menczer F (2019) Quantifying Biases in Online Information. Exposure 70(3):218–229. https://doi.org/10.1002/asi.24121
6. European Commission (2017) Antitrust: Commission fines Google €2.42 billion for abusing dominance as search engine by giving illegal advantage to own comparison shopping service—Factsheet. http://europa.eu/rapid/press-release_MEMO-17-1785_en.htm. Accessed 2 Jun 2018
7. Gwizdka J, Moshfeghi Y, Wilson ML (2019) Introduction to the special issue on neuro-information science. J Assoc Inf Sci Technol. https://doi.org/10.1002/asi.24263
8. Ballatore A (2015) Google chemtrails: A methodology to analyze topic representation in search engine results. First Monday, 20(7). http://www.firstmonday.org/ojs/index.php/fm/article/view/5597/4652. Accessed 17 Nov 2015
9. Jansen BJ (2007) The comparative effectiveness of sponsored and nonsponsored links for Web e-commerce queries. ACM Transactions on the Web, 1(1), article 3. https://doi.org/10.1145/1232722.1232725
10. Liu Z, Liu Y, Zhou K, Zhang M, Ma S (2015) Influence of Vertical Result in Web Search Examination. In: Baeza-Yates IR, Lalmas M, Moffat A, Ribeiro-Neto B (eds) (Eds.), Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR, vol 15. ACM Press, New York, New York, USA, pp 193–202 https://doi.org/10.1145/2766462.2767714
11. Otterbacher J, Bates J, Clough P (2017) Competent Men and Warm Women. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems—CHI '17 (pp. 6620–6631). New York, New York, USA: ACM Press. https://doi.org/10.1145/3025453.3025727
12. Lewandowski D, Kerkmann F, Rümmele S, Sünkler S (2018) An empirical investigation on search engine ad disclosure. J Assoc Inf Sci Technol 69(3):420–437. https://doi.org/10.1002/asi.23963
13. Schwartz B (2018) Google Expands 'More Results' Button To Remove Paginated Search Results. https://www.seroundtable.com/google-expands-more-results-button-test-25542.html. Accessed 14 May 2018
14. Lurie I (2010) 3 Lies The Search Engines Will Tell You - Search Engine Land. https://searchengineland.com/3-lies-the-search-engines-will-tell-you-45828. Accessed 10 Oct 2019
15. Darko B (2019) Google Click Analysis: More SERP Features, Less Traffic for Websites. https://blog.searchmetrics.com/us/google-click-analysis-serp-traffic/. Accessed 10 Oct 2019
16. Fishkin R (2019) How Much of Google's Search Traffic is Left for Anyone But Themselves? https://sparktoro.com/blog/how-much-of-googles-search-traffic-is-left-for-anyone-but-themselves/. Accessed 10 Oct 2019
17. Alphabet Inc. (2019) Alphabet Announces Fourth Quarter and Fiscal Year 2018 Results. https://abc.xyz/investor/static/pdf/2018Q4_alphabet_earnings_release.pdf. Accessed 4 Oct 2019
18. Google (2019a) Get Results With An Advertising Budget That Works For You – Google Ads. https://ads.google.com/intl/en/home/pricing/. Accessed 10 Oct 2019
19. Marvin G (2019) Updated: A visual history of Google ad labeling in search results. https://searchengineland.com/search-ad-labeling-history-google-bing-254332. Accessed 12 Aug 2019
20. Edelman B (2014) Google's Advertisement Labeling in 2014. http://www.benedelman.org/adlabeling/google-colors-oct2014.html. Accessed 19 Apr 2018
21. Google (2019b) Why you're seeing an ad – Ads Help. https://support.google.com/ads/answer/1634057#info. Accessed 3 Jul 2019
22. Neethling R (2007) Search engine optimisation or paid placement systems – user preference. Thesis, Cape Peninsula University of Technology
23. McCue T (2018) SEO Industry Approaching $80 Billion But All You Want Is More Web Traffic. https://www.forbes.com/sites/tjmccue/2018/07/30/seo-industry-approaching-80-billion-but-all-you-want-is-more-web-traffic/. Accessed 1 Oct 2019
24. Adobe (2018) Adobe Digital Insights: Holiday Recap Report 2017. https://de.slideshare.net/adobe/adobe-digital-insights-holiday-recap-report-2017. Accessed 3 Jun 2019
25. Similarweb.com (2019) SimilarWeb | Website Traffic Statistics & Market Intelligence. https://www.similarweb.com/. Accessed 31 May 2019
26. Purcell K, Brenner J, Rainie L (2012) Search Engine Use 2012. https://www.issuelab.org/resources/12470/12470.pdf. Accessed 12 Apr 2016
27. Pan B, Hembrooke H, Joachims T, Lorigo L, Gay G, Granka L (2007) In Google We Trust: Users' Decisions on Rank, Position, and Relevance. J Comput Commun 12(3):801–823. https://doi.org/10.1111/j.1083-6101.2007.00351.x
28. Schultheiß S, Sünkler S, Lewandowski D (2018) We still trust in google, but less than 10 years ago: An eye-tracking study. Inf Res 23(3). http://www.informationr.net/ir/23-3/paper799.html. Accessed 10 Jan 2019
29. Stark B, Magin M, Jürgens P (2014) Navigieren im Netz Befunde einer qualitativen und quantitativen Nutzerbefragung. In: Stark B, Dörr D, Aufenanger S (eds) (Eds.), Die Googleisierung der Informationssuche. DE GRUYTER, Berlin, Boston, pp 20–74 https://doi.org/10.1515/9783110338218.20
30. Singer G, Norbisrath U, Lewandowski D (2012) Ordinary Search Engine Users assessing Difficulty, Effort, and Outcome for Simple and Complex Search Tasks. In Proceedings of the Fourth Information Interaction in Context Symposium. ACM, New York, pp 110–119. https://doi.org/10.1145/2362724.2362746
31. White RW (2016) Interactions with Search Systems. Cambridge University Press, New York
32. Ball LJ, Poole A (2010) Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects. In: Ghaoui C (ed) Encyclopedia of Human-Computer Interaction. Idea Group, Inc, Pennsylvania, pp 211–219
33. Lewandowski D, Kammerer Y (2019) Factors Influencing Viewing Behaviour on Search Engine Results Pages: A Review of Eye-Tracking Research. Manuscript submitted for publication
34. Jacucci G, Barral O, Daee P, Wenzel M, Serim B, Ruotsalo T, … Blankertz B (2019) Integrating neurophysiologic relevance feedback in intent modeling for information retrieval. Journal of the Association for Information Science and Technology, 70(9), 917–930. https://doi.org/10.1002/asi.24161
35. Eugster MJA, Ruotsalo T, Spapé MM, Kosunen I, Barral O, Ravaja N, … Kaski S (2014) Predicting term-relevance from brain signals. In: Geva S, Trotman A, Bruza P, et al (eds) Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval—SIGIR '14 (pp. 425–434). New York, New York, USA: ACM Press. https://doi.org/10.1145/2600428.2609594
36. Moshfeghi Y, Pollick FE (2019) Neuropsychological model of the realization of information need. Journal of the Association for Information Science and Technology, asi.24242. https://doi.org/10.1002/asi.24242
37. Schultheiß S, Lewandowski D (2019) How users' knowledge of advertisements influences their viewing and selection behaviour in search engines. Manuscript submitted for publication
38. Lewandowski D, Sünkler S, Kerkmann F (2017) Are Ads on Google Search Engine Results Pages Labeled Clearly Enough? The Influence of Knowledge on Search Ads on Users' Selection Behaviour. In M. Gäde, V. Trkulja, & V. Petras (Eds.), Everything

Changes, Everything Stays the Same? Understanding Information Spaces. Proceedings of the 15th International Symposium of Information Science. ISI 201(7):62–74 (Glückstadt: Verlag Werner Hülsbusch)

39. Lund H (2016) Eye tracking in library and information science: a literature review. Libr Hi Tech 34(4):585–614. https://doi.org/10.1108/LHT-07-2016-0085
40. Sünkler S, Lewandowski D (2017) Does it matter which search engine is used? A user study using post-task relevance judgments. Proceedings of the Association for Information Science and Technology, 54(1), 405–414. https://doi.org/10.1002/pra2.2017.14505401044
41. Lewandowski D, Sünkler S (2013a) Designing search engine retrieval effectiveness tests with RAT. Inf Serv Use 33(1):53–59. https://doi.org/10.3233/ISU-130691
42. Lewandowski D (2015) Evaluating the retrieval effectiveness of web search engines using a representative query sample. J Assoc Inf Sci Technol 66(9):1763–1775. https://doi.org/10.1002/asi.23304
43. Lewandowski D, Sünkler S (2019) Das Relevance Assessment Tool: Eine modulare Software zur Unterstützung bei der Durchführung vielfältiger Studien mit Suchmaschinen. Information – Wissenschaft & Praxis, 70(1), 46–56. https://doi.org/10.1515/iwp-2019-0007
44. Lewandowski D, Sünkler S (2013b) Representative online study to evaluate the commitments proposed by Google as part of EU competition investigation AT. 39740-Google: Report for Germany. http://searchstudies.org/wp-content/uploads/2015/10/Google_Online_Survey_DE.pdf. Accessed 20 May 2018
45. Möller C (2013) Attention and selection behavior on 'universal search' result pages based on proposed Google commitments of Oct. 21, 2013: Report about an eye tracking pilot study commissioned by ICOMP Initiative for a Competitive. https://de.slideshare.net/gesterling/bericht-icomp-vol4. Accessed 1 May 2019