Challenges for search engine retrieval effectiveness evaluations: Universal Search and user intents, and results presentation

Dirk Lewandowski

Hamburg University of Applied Sciences, Faculty DMI, Department of Information, Finkenau 35, D – 22081 Hamburg, Germany

dirk.lewandowski@haw-hamburg.de

Abstract. This chapter discusses evaluating the quality of Web search engines to effectively retrieve information. It identifies three factors that lead to a need for new evaluation methods: (1) the changed results presentation in Web search engines, called Universal Search, (2) the different query types that represent different user intentions, and (3) the presentation of individual results. It discusses implications for evaluation methodology and provides some suggestions about measures.

Keywords. Web search engines, retrieval effectiveness, evaluation, Universal Search, search engine results page (SERP), user behaviour

Introduction

Quality is important in all information retrieval (IR) systems, including Web search engines. The goal of this chapter is to discuss methods for evaluating Web search engines with a focus on the current standards for results presentation and on users' intentions.

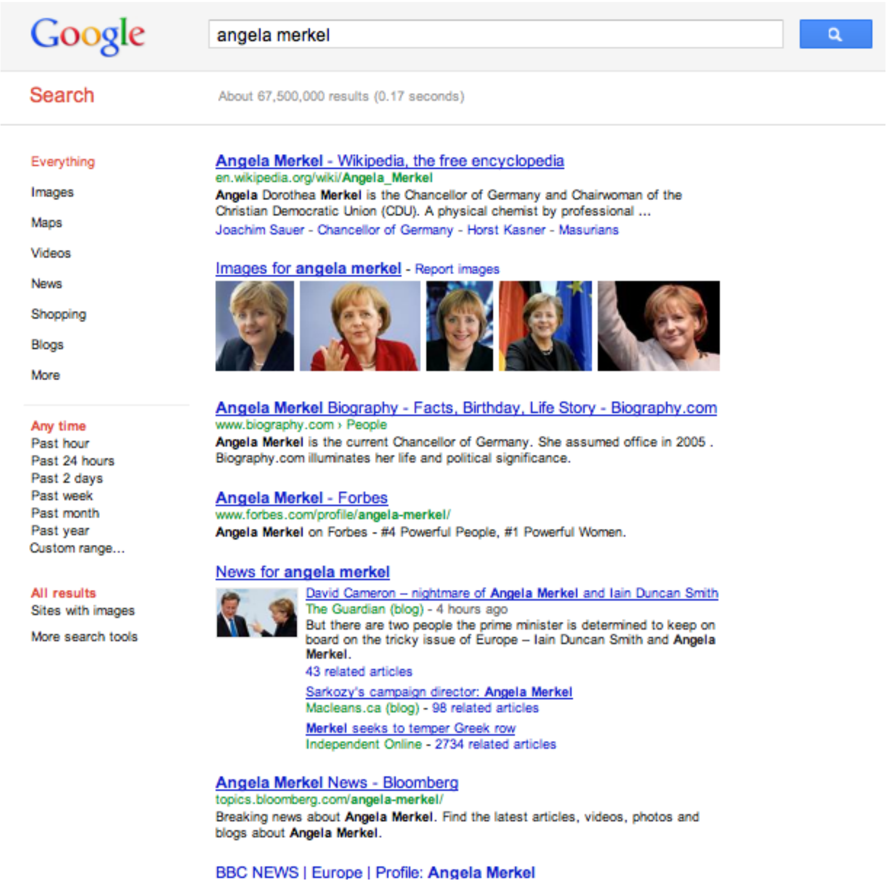
The quality of Web search engines is of great importance, as users may choose their preferred search engine based on its perceived quality. The quality of the different search engines is also of great interest to search engine vendors (to improve their system or to benchmark their system with others) and the general public. Search engines have become a major means to acquire knowledge, and the results they show in the first positions have a great influence on the information that users actually consume.

Evaluation is traditionally an integral part of information retrieval research, and it pays particular attention to a user's examination of the results list presented by the information system from top to bottom. Evaluators also assume that the user's decision to choose one site over another is based on reading the abstract (snippet) presented in the results list.

However, the presentation of search results in Web search engines has changed in recent years, and user behaviour has followed suit. While simple lists were dominant for several years, nowadays, results from different document collections (such as news, images, and video) are presented on the search engine results pages (SERPs) (Höchstötter & Lewandowski, 2009). This type of presentation is called Universal Search, which is the composition of search engine results pages from multiple sources. While in traditional results presentation, results from just one database (the Web index) are presented in sequential order and the presentation of individual results does not differ considerably, in universal search, the presentation of results from the different collections is adjusted to the collections' individual properties.

A search engine results page (SERP) is a complete presentation of search engine results; that is, it presents a certain number of results (determined by the search engine). To obtain more results, a user must select the "further results" button, which leads to another SERP. On a SERP, results from different collection, a.k.a. vertical search engines can be presented. Contrary to the general-purpose search engine, a vertical search engine focuses on a special topic.

The properties of the different results types lead not only to a different presentation of the results pages but also to a different presentation of the individual results. For example, it is clear that SERPs that include image and video results should show preview pictures. Figure 1 shows an example of a Universal Search results page, Fig. 2 provides examples of the presentation of an individual result.



About 67,500,000 results (0.17 seconds)

Angela Merkel - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Angela_Merkel
Angela Dorothea Merkel is the Chancellor of Germany and Chairwoman of the Christian Democratic Union (CDU). A physical chemist by professional ...
Joachim Sauer - Chancellor of Germany - Horst Kasner - Masurians

Images for angela merkel - Report images

Angela Merkel Biography - Facts, Birthday, Life Story - Biography.com
www.biography.com › People
Angela Merkel is the current Chancellor of Germany. She assumed office in 2005 . Biography.com illuminates her life and political significance.

Angela Merkel - Forbes
www.forbes.com/profile/angela-merkel/
Angela Merkel on Forbes - #4 Powerful People, #1 Powerful Women.

News for angela merkel
David Cameron – nightmare of Angela Merkel and Iain Duncan Smith
The Guardian (blog) - 4 hours ago
But there are two people the prime minister is determined to keep on board on the tricky issue of Europe – Iain Duncan Smith and Angela Merkel.
43 related articles
Sarkozy's campaign director: Angela Merkel
Macleans.ca (blog) - 98 related articles
Merkel seeks to temper Greek row
Independent Online - 2734 related articles

Angela Merkel News - Bloomberg
topics.bloomberg.com/angela-merkel/
Breaking news about Angela Merkel. Find the latest articles, videos, photos and blogs about Angela Merkel.

BBC NEWS | Europe | Profile: Angela Merkel

Figure 1: Search engine results page (example from Google)

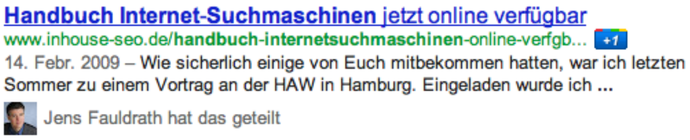

Handbuch Internet-Suchmaschinen jetzt online verfügbar
www.inhouse-seo.de/handbuch-internetsuchmaschinen-online-verfgb... +1
14. Febr. 2009 – Wie sicherlich einige von Euch mitbekommen hatten, war ich letzten Sommer zu einem Vortrag an der HAW in Hamburg. Eingeladen wurde ich ...
Jens Fauldrath hat das geteilt

Figure 2: Example of an individual result description (example from Google)

In Figure 1, we can see how results from different sources (i.e., specialized search engine indices or so-called vertical search engines) are injected into the general results list created from the Web index (i.e., the search engine's main index). Additional results in this case come from the image index and from the news index.
In Figure 2, we see a typical results description provided in a results list. It contains a title, a URL, a short description, and, in this case, a social recommendation.

While Universal Search is a concept usually associated with Web search, it may also be applied to such diverse offerings as Web portals, e-commerce websites and Intranets. Therefore, the discussion presented in this chapter may also be applied to search scenarios other than Web searches.

Along with the positioning of the results, the different representations of search results determine the search engine users' viewing and clicking behaviour. Another important factor is the user's goal when entering a query. The classic distinctions between query types posited by Broder (2002) are informational, navigational, and transactional intentions that are the basis for a further discussion on retrieval effectiveness and success.

To summarize, we will discuss search engine evaluation in the context of

- Results presentation (design of Universal Search results pages)
- Query types
- Results selection

It is obvious that the challenge when measuring the retrieval effectiveness of Web search engines is to develop evaluation methods that consider the three areas mentioned. This chapter provides methods used to evaluate Universal Search results pages and suggestions for designing retrieval effectiveness studies. The structure of this chapter is as follows: First, we will give a short overview of search engine evaluation, then we will discuss users' intentions (as expressed through different types of queries). After that, we will detail Web search engines' results presentations and users' selection behaviour on the search engine results pages. Bringing these three areas together, we will discuss points to consider when evaluating search engines with a Universal Search results presentation. The chapter closes with some conclusions and suggestions for further research.

Search engine evaluation

When discussing search engine evaluation, it is important to stress that quality measurement goes well beyond retrieval effectiveness, i.e., measuring the quality of the results. Some frameworks for a more complete search engine evaluation have been proposed (e.g., Xie, Wang & Goh, 1998; Mansourian, 2008). Lewandowski and Höchstötter's model (Lewandowski, Höchstötter, 2008) divides Web search engine quality into four major areas:

- Index Quality: This area of quality measurement indicates the important role that search engines' databases play in retrieving relevant and comprehensive results. Areas of interest include Web coverage (Gulli, 2005); country bias (Vaughan & Thelwall, 2004; Vaughan & Zhang, 2007), and freshness (Lewandowski, 2008a; Lewandowski, Wahlig, & Meyer-Bautor, 2006).
- Quality of the results: Derivates of classic retrieval tests are applied here. However, one needs to consider which measures should be applied and whether or not new measures are needed to satisfy the unique character of a search engines and its users (Lewandowski, 2008b).

- Quality of search features: A sufficient set of search features and a sophisticated query language should be offered and should function reliably (Lewandowski, 2004, 2008c).
- Search engine usability: The question is whether it is possible for users to interact with search engines in an efficient and effective way.

While all quality factors mentioned are important, results quality is still the major factor in determining which search engine performs best. A good overview of newer approaches in Web search engine retrieval effectiveness evaluation is provided by Carterette, Kanoulas, and Yilmaz (2012).

In retrieval effectiveness evaluation, two approaches need to be differentiated:

1. "Classic" retrieval effectiveness tests use a sample of queries and jurors to judge the quality of the individual results. These studies use explicit relevance judgements made by the jurors. An overview of search engine retrieval effectiveness studies using explicit relevance judgements is provided by Lewandowski (2008b).
2. Retrieval effectiveness studies analyse click-through data from actual search engine users (e.g., Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., & Gay, G., 2007; Jung, Herlocker & Webster, 2007). As users give their relevance judgements only through their selection behaviour, we speak of implicit relevance judgements here.

Both approaches have their merits. When using click-through data, researchers can rely on large quantities of data and can determine which results are preferred by the actual users of a search engine. The drawback, however, is that these decisions are based on the results descriptions on the SERPs that heavily influence users' results selections, and users choose only from some of the results presented. For example, a user would not read all the results descriptions and then choose a result from the third results page. On the contrary, he will rely on the first results presented by the search engine and choose from them.

The main advantage of classic retrieval effectiveness tests are that no data from the search engine providers is needed, and jurors can be asked for their opinions, so a researcher can go beyond decisions about whether an individual result is relevant or not. The drawback of such tests, however, is that such studies usually rely on a relatively low number of queries and jurors, and results are seen as independent of one another. This can be illustrated by a user who chooses a completely relevant result and will therefore not need another relevant result that just repeats the information already given.

From this short discussion, one can readily see that a combination of the two approaches described would be the best way to go. However, researchers usually do not have access to click-through data from real search engines, so only the search engine vendors themselves usually do this type of evaluation. However, in recent years, some search engine providers have made datasets including click-through data available to the research community, but there are still many cases where search results need to be evaluated and the researcher does not have access to such data. Furthermore, it is highly unlikely that a researcher will get this data from more than one search engine for benchmarking purposes.

We would like to stress that a basic assumption in this discussion on search engine evaluation is that the researchers do not have access to data owned by the search engines considered. Therefore, we will discuss the use of click-through data merely as an addition to other approaches.

As this short discussion reveals, a challenge in measuring the retrieval effectiveness of Web search engines is to develop evaluation methods that consider both the results presentation and user behaviour. However, search engine retrieval effectiveness studies to date still lack the integration of explicit user models. Based on the Cranfield paradigm, many evaluation models have been built that provide a robust framework for conducting search engine retrieval effectiveness tests. These evaluations are based on retrieval measures, which largely measure the performance of systems, even though newer approaches try to integrate explicit user models into such evaluations (Carterette et al., 2012). What even integrating users models into the evaluation frameworks may not be enough; furthermore, an explicit *results presentation model* is needed.

User intents

Search engines are used for a multitude of purposes, including navigating to an already known website, simple fact-checking, complex research tasks, and even entertainment purposes. It is clear that, depending on the task, a user will enter different types of queries into the search box. The difference between Web search engines and other information retrieval (IR) systems is usually seen as search engines having no thematic focus and being used by laypersons. But it must be stressed that another important point that differentiates them from other IR systems is that they are used for different user intents, expressed through query types.

A simple, yet powerful classification of user intents is Broder's (2002) query type classifications: informational, navigational and transactional.

Navigational queries are used by a person who knows about a Web page or assumes that it exists (for example, the homepage of a company like ebay or people like Angela Merkel). Such queries normally terminate in one correct result. The information need is satisfied when the requested page is found.

In contrast, informational queries require more than one document. The user requires problem-oriented information (see Frants, Shapiro, & Voiskunskii, 1997). The user wishes to become informed about a topic and therefore intends to read several documents. Informational queries aim at static documents to acquire the desired information, which makes further interaction with the Web page unnecessary.

Transactional queries, however, aim at Web pages offering the possibility of a subsequent transaction, such as the purchase of a product, the download of data, or the search of a database.

Broder's query type classification has been refined over the years by researchers like B. J. Jansen, Booth, and Spink (2008), Kang and Kim (2003) and Rose and

Levinson (2004). However, all rely on the basic distinction between the three types mentioned. While it is perfectly legitimate to use a more fragmented query type classification, the general query type classification is sufficient for our purposes.

Evaluating Web search engines' performance for the query types mentioned requires adjusting methods and retrieval measures. Many studies on informational queries have been conducted (for an overview, see Lewandowski, 2008b), but only a few studies have examined navigational queries (Lewandowski, 2011; for an overview of measures, see MacFarlane, 2007). Evaluating search engines' performance on transactional queries is especially difficult, as they require an interaction on the results.

Results presentation in Web search engines

Search engine results pages are usually simple lists of search results ("10 blue links"). Each result is presented in the same format, and, apart from a position bias, every result has the same chance of being selected. However, this is not the case, as different result types are now presented on the SERPs, results are given different amounts of space, and some results are highlighted in either through the use of a different background colour or additional graphical elements. This section provides a discussion of the major elements of a SERP. For a more detailed discussion, see Höchstötter and Lewandowski (2009) and Lewandowski and Höchstötter (2009).

As shown in Figure 1, the search engine results pages contain a variety of elements. In the following paragraphs, we will discuss these elements in detail.

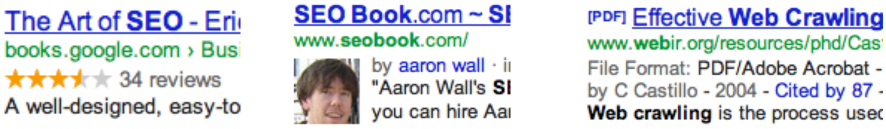
Organic results

Organic results (also called algorithmic results) are the listings on a search engine results page that are not paid for (Jansen, 2011). The organic results list forms the core of the results presentation, and its ranking is produced by algorithms that aim to determine their relevance. Thereby, all documents in the search engine's index are treated the same.

There is no direct human intervention in the results listings, although in recent years there has been some discussion about whether search engines favour their own offerings (Edelman, 2010). Making an analogy to journalism, Nicholson et al. (2006) spoke of the organic results as "editorial listings."

The basic elements of a description used in an organic result are a heading, a short description, and a URL. Other elements, such as a date and a link to a cache copy, can also be shown. In recent years, search engines have further enriched their results descriptions with ratings (Figure 3a), pictures (Figure 3b), information from social networking sites (Figure 2), and citation information for scholarly articles (Figure 3c). This means that results are no longer equally represented and that more factors than the results position and the contents of the description are influencing users' results selections. Studies show that such "rich snippets" can help users make quicker and more relevant decisions (Li & Shi, 2008), but they can also distract users from relevant results.

The so-called site links are particularly important in organic results. Here, a search engine enriches the result on the first position of the organic list in a special way. The search engine provides links to popular destinations within the website and describes the homepage. Such a description occupies a lot of space on the SERP.



The Art of SEO - Eri
books.google.com › Bus
★★★½☆ 34 reviews
A well-designed, easy-to

SEO Book.com ~ SI
www.seobook.com/
by aaron wall · i
"Aaron Wall's SI
you can hire Aa

[PDF] Effective Web Crawling
www.webir.org/resources/phd/Cas
File Format: PDF/Adobe Acrobat -
by C Castillo - 2004 - Cited by 87 -
Web crawling is the process use

Figure 3: Added information shown within results descriptions: (a) picture and author information and (b) author and citation information (c)

Sponsored results

Sponsored results are also called sponsored links or AdWords (named after Google's sponsored results product). These are text-based advertisements that are presented on the SERPs and are a context response to the query entered by the user. Advertisers are charged for clicks on their ads, and the price per click (PPC) is determined through an auction process (J. Jansen, 2011).

Sponsored results are usually shown on the right side of and on top of the organic results. This means that a user scanning the SERP from top to bottom first sees the sponsored results, which are labelled as advertisements with words such as "ads", "sponsored", etc.), and are often highlighted through the use of a different background colour, a technique not used in the organic results. For example, Google uses white as the background colour for the organic results and a light yellow background for the sponsored results.

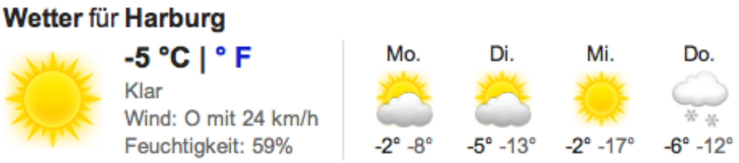
The basic elements used in a sponsored result are a heading, a short description and a URL. The descriptions of sponsored results can also be enriched by ratings or other graphical information.

One can easily see that the descriptions of sponsored results are built using techniques similar to those used in organic results. As sponsored results are also shown in response to a query and can therefore be relevant to the query (B. J. Jansen, 2007), users might confuse sponsored results and organic results. However, there are only a few studies that have examined this (Bundesverb & Digitale Wirtschaft, 2009; Fallows, 2005), so no clear judgement on this can be given here. However, as we know that users do consider sponsored results (for whatever reason), this results type should be considered in a search engine evaluation.

Shortcuts

Shortcuts provide a direct answer to the query on the SERP itself, and the user does not need to click on a result on the SERP. Shortcuts come from special databases maintained by the search engines, and their inclusion is usually triggered by query words. Search engines have been criticised for injecting results from their own collections into the SERPs (Edelman & Lockwood, 2011), but others have argued that this practice is completely understandable and does not negatively influence the quality of the results (Sullivan, 2011).

Figure 4 shows an example of a shortcut for a search for the weather in Hamburg. The basic elements of a shortcut cannot be provided as the SERPs vary and they are fitted to the individual topic.



Wetter für Harburg

-5 °C | ° F

Klar

Wind: O mit 24 km/h

Feuchtigkeit: 59%

Mo.	Di.	Mi.	Do.
-2° -8°	-5° -13°	-2° -17°	-6° -12°

Figure 4. Example of a shortcut

Results from special collections

Results from special collections (sometimes just called "universal results") are results that do not come from the general Web index of the search engine but from specialized databases, such as the news index, the video index, or the image index. These results are injected into the list of organic results but are often presented differently. A basic element of results descriptions from special collections is a clickable heading that leads to results obtained only from the vertical index in question. For example, when users click on the heading for the news results, they get a results list containing only results from the news index.

Further elements of the results descriptions are dependent on the document type. While in video results, a title, a still and the source may be shown, results from the database of scientific articles provide very different information. This distinction between the presentations of different document types makes evaluation difficult.

To summarize, there are different results types presented on the search engine results pages. Furthermore, while the presentations of these results are in some ways similar, the presentations differ according to types.

When evaluating Web search results, the amount of space given to an individual result and the graphic presentation of the result should be considered, as both influence users' results selections. When results presentation changes, so needs evaluation methodology.

Users' results selection

While search engines usually return thousands of results, users are not willing to view more than a few (Jansen & Spink, 2006; Keane, O'Brien, & Smyth, 2008; Lorigo et al., 2008; Machill, Neuberger, Schweiger, & Wirth, 2004), so, for all practical purposes, the first page of a search engine's results is considered most relevant. Even in cases when users are willing to view more than the first few results, the initial ranking heavily influences their perceptions of the results set. That is, when a user does not consider the first few results to be relevant, he or she usually modifies the query or abandons the search.

The first results page must also be used to present results from additional collections, as users usually do not follow links to these additional collections, the so-called "tabs". Search engine expert Danny Sullivan (2003) coined the term "tab blindness" to name this phenomenon. Results beyond organic results also require space on the results page. With the addition of more and more results from special collections (and, in some cases, the addition of more and more ads above the organic results), we have observed a general change in results presentation (Lewandowski, 2008d). Organic results become less important as additional results take their place.

The changed design of the search engine results pages also leads to a larger number of results presented on these pages. While in the classic list-based results presentation, ten results per page (+ads) were displayed, a typical Universal Search results pages can easily contain 30 or even more clickable results links. When considering this more complex results presentation, users' results selection is based on a multitude of factors (see Kammerer & Gerjets, 2012).

An important question behind users' results selection behaviour is whether it is based on an informed decision. By this we mean whether the user actually makes a cognitive judgment about the quality of the individual results and selects the most appropriate result(s) or whether he just clicks on a link shown prominently. In the ongoing discussion on "search neutrality") the position of the search engine vendors is that users indeed make an informed decision (Granka, 2010; Grimmelmann, 2010). If that were not the case, one might question the search engines' revenue model, and if users were not able to distinguish between organic and sponsored results and therefore clicked on ads without knowing that they were doing so, regulation authorities would have to force search engines to label sponsored results more clearly than they are currently doing.

In the following section, we will discuss search results presentation factors influencing users' results selections. We will consider the results position, the results description content, its size and its design.

Results position

When considering the probability of a result being clicked, one has to first consider whether the result is included in the visible area of the SERP. The visible area is defined as the part of the SERP that is directly visible to the user, without scrolling down the results list. The size of the visible area depends on the user's screen size, the size of the browser window and the size of the browser's navigation elements, as well as browser toolbars that a user may have installed.

Research shows that users first and foremost consider results within the visible area (also called "above the fold"). However, it is important to understand that navigational queries make up a large proportion of queries, and if the search engine works well, the one desired result will be shown within the visible area, so no scrolling will be necessary. Even when making informational queries, many users desire only one or a few results, and such queries can be satisfied with the results in the visible area.

The second point to consider is the actual ranking of the results. Results presented first get considerably more attention than results presented lower in the ranked lists (Cutrell & Guan, 2007; Hotchkiss, 2006; Joachims, Granka, Pan, Hembrooke, & Gay, 2005; Pan et al., 2007). However, when considering the results positions, one must also decide whether to see the list of sponsored results and the list of organic results as two separate lists or as one continuous list. Eye-tracking research suggests that at least some users consider these two areas as one continuous list (Wirtschaft, 2009).

Results description content

It is evident that users are influenced by the actual contents of the results descriptions. There is a high probability that users will click on results whose descriptions contain the query words and have meaningful titles. However, there is little research on the importance of the results descriptions' content. Evidence is mainly given by search engine optimizers who emphasize the influence of good description copy for generating clicks on the results (Thurow, 2007; Thurow & Musica, 2009) Pirolli (2009) used the concept of information scent to explain the influence of the description content, stating that users follow trails if they are given hints about where to find the desired information.

Results description size

As stated above, the results descriptions occupy different amounts of space on the search engine results page. We can measure this space in pixel size and weigh the probability that a user will click it. We assume that the more space a results description occupies, the higher the probability that a user will click on it.

Results description design

Some results may be presented with additional images or may be emphasized by the use of colour. For example, sponsored links may be highlighted in yellow. Users may perceive these results as being more important and are therefore more likely to click on them. Moreover, given that users know that results highlighted in such a way are advertisements, they may avoid them. Therefore, it is not easy to make a decision about whether to ascribe more or less weight to these results rather than organic results.

While sponsored results are usually highlighted in colour, other results (whether organic or from a special collection) may carry images that draw the users' attention. However, this attention is hard to measure because if the information is displayed on a page cluttered with enriched results, it may get less attention than a result that is the only one highlighted.

In summary, we can see that users' results selections are influenced by a variety of factors. However, the importance of these factors is not easy to measure. Apart from large-scale click-through studies that are usually provided by the search engine vendors themselves and do not allow for comparing search engine, lab-based experiments that track eye movements are used to research selection behaviour (Beiler, 2005; Pan et al., 2007). The advantage of lab-based studies is that the researchers can also ask participants about their reasons for selecting certain results.

Approaches to weighting results on the SERPs

In a previous study (Lewandowski, 2012), we presented a general framework for measuring the quality of search engines' results. While the influence of results presentation on users' results selections is a part of the model, we intend in this chapter to focus on the weighting of individual results within the results presentation. The factors to be weighted are discussed in the following sections.

Results to be considered

First, one needs to decide whether all results presented on the SERP should be considered or not. The main concern is usually whether one should consider the results in the visible area vs. results in the invisible area and organic vs. sponsored results.

A researcher might decide that results in the invisible area will not be considered as most users do not consider them. However, one should also consider that the size of the visible area varies from user to user. Therefore, one can classify users according to their screen sizes (e.g., "W3 Schools Browser Statistics," 2011).

When deciding whether to include sponsored results that are shown above the organic results in the evaluation, one must decide whether to see them as separate from the organic results or to consider sponsored and organic results as one continuous list. When giving weights to sponsored results, one must decide whether highlighting leads users to this type of results.

Position

Weighting results according to position is an element included in most retrieval effectiveness measures. However, in the case of evaluating Universal Search SERPs, one must decide which results types are to be considered as an element of the results list, i.e., those that actually have a position.

Screen real estate

"Screen real estate" is a term introduced by Nielsen & Tahir (n.d.) to measure the amount of space taken up by a Web page that is filled with content in relation to the part that is left blank. The term has been used in different scenarios (e.g., Dewan, Freimer, & Zhang, 2002). Nicholson et al. (2006) used the screen real estate measurement to calculate the space that search engines grant organic results vs. paid results on the first results screen shown to the user. While this is a specific application in the context given, screen real estate can be calculated for every result presented by a search engine. The use of screen real estate may seem obvious when results are presented in different formats; however, it can also be used for list-based results presentation if all results are not equally presented.

Weighting screen real estate can either be applied to the whole SERP or just the visible area. However, one must consider that the size of the visible area is different, depending on the browser window size utilised by individual users.

Click-through rates

As mentioned above, click-through data is usually not available to researchers if they are not affiliated with a search engine vendor. However, when available, the importance of individual results can be weighted according to their CTR. In other words, results with a higher CTR are assigned a higher importance.

As CTRs for the individual queries are often not available, one could also use the average CTR for certain query types. When considering navigational queries, one can see that a large ratio of clicks is aggregated to just one result position. The distribution is wider for informational queries, so one must give more value to more results in the evaluation.

Combining Universal Search SERPs and user intent

When dealing with navigational queries in list-based results pages, one assumes there is one relevant result that should be displayed in the first position of the list, but this raises two problems. First, there is not always a clear distinction between query types. A query can be both navigational and informational. The query type may differ from user to user. Even in the eBay example given above, there may be users who wish to get information about the company and do not wish to navigate to Ebay's website. By using click-through data, however, researchers can see that the overwhelming ratio of users do see this query as navigational.

Considering a more complex example, the query about Angela Merkel depicted in Figure 1 can be navigational or informational. The user may wish to navigate to her personal website, but he may also want to get information about her life and work. Furthermore, the user may be interested in current news.

This example shows that informational and navigational intentions can be considered in one results list. Furthermore, even when the intent is informational, it may matter considerably whether the user is seeking current information or a general overview of the life and work of Angela Merkel. In such cases, Universal Search results pages can satisfy diverse user intentions.

In evaluating results quality in Web search engines, it is always important to distinguish between query types. It becomes even more important in the context of a Universal Search, as multiple query intentions can be considered within the same results presentation. Therefore, we suggest collecting information on the underlying query intentions for every query one evaluates. If the researcher has access to click-through data from a search engine, navigational queries can be identified reliably. If such data is not available, we suggest asking users about query intentions and using this information to derive weighted query intentions, as Huffman and Hochster (2007) did in their study.

Conclusion

In this chapter, we showed that current approaches to Web search engine retrieval effectiveness evaluation have shortcomings and discussed new factors that can be used in evaluations. Often, studies do not differentiate between query types and results that are presented differently so they also ignore the probability of their being selected. We propose an approach that considers all these elements. However, while we were able to show how search engine evaluation is influenced by the decisions of the researcher in this regard, we were not able to provide an empirical evaluation.

Further research needs to examine users' approaches to different results. In particular, the status of sponsored results has been ignored by researchers. While search engine vendors make an overwhelming proportion of their incomes from sponsored results, we still do not know how users actually perceive these results. Some studies suggest that a fairly large ratio of users is not able to differentiate between organic and sponsored results (e.g., Bundesverband Digitale Wirtschaft, 2009), but we need a thorough study on this topic.

In addition, studies that ignore the different results types do not exactly measure what the user gets to see in his searches. As search engine evaluation at least tries to model the user's behaviour, researchers should consider empirical studies of all results types.

Researchers also need to consider browser sizes. While it surely is valid to calculate retrieval effectiveness measures for the whole SERP, the user usually only recognizes a part of it. Users focus on the visible area of the results pages (Höchstötter & Lewandowski, 2009), so results in this area should be considered foremost. Going even further, one could also use eye-tracking research to determine which sections of the results pages are actually seen or considered by the users.

Finally, we must say that there are more open questions in search engine evaluation than answers. We think that in this chapter we have raised more important questions rather than providing answers. Nonetheless, we can see that Web search engine evaluation is not merely a technical issue but also has a societal dimension. When we consider the billions of queries entered into Web search engines each day (ComScore, 2009) and understand that search engines influence searchers' selection behaviours through their results presentations, we recognize that search engine evaluation techniques need to be applied to measuring "search neutrality", i.e., a fair representation of the Web's contents in search engines.

References

Beiler, M. (2005). Selektionsverhalten in den Ergebnislisten von Suchmaschinen. Modellentwicklung und empirische Überprüfung am Beispiel von Google. In M. Machill & N. Schneider (Eds.), *Suchmaschinen : Neue Herausforderuingen für die Medienpolitik* (Vol. 50, pp. 165-189). Berlin: VISTAS Verl.

Broder, A. (2002). A taxonomy of Web search. *ACM Sigir Forum*, *36*(2), 3-10.

Bundesverband Digitale Wirtschaft. (2009). *Nutzerverhalten auf Google-Suchergebnisseiten: Eine Eyetracking-Studie im Auftrag des Arbeitskreises Suchmaschinen-Marketing des Bundesverbandes Digitale Wirtschaft (BVDW) e.V.*

Carterette, B., Kanoulas, E., & Yilmaz, E. (2012). Evaluating Web retrieval effectiveness. In D. Lewandowski (Ed.), *Web Search Engine Research* (pp. 105-137). Bingley: Emerald.

ComScore. (2009). Global search market draws more than 100 billion searches per month. Retrieved from http://www.comscore.com/Press_Events/Press_Releases/2009/8/Global_Search_Market_Draws_More_than_100_Billion_Searches_per_Month

Cutrell, E., & Guan, Z. (2007). *Eye tracking in MSN Search: Investigating snippet length, target position and task types. Technical Report, TR-2007-01.* Retrieved from http://research.microsoft.com/pubs/70395/tr-2007-01.pdf

Dewan, R., Freimer, M., & Zhang, J. (2002). Managing Web sites for profitability: Balancing content and advertising. *System Sciences, 2002. HICSS*. Proceedings from the *35th Annual Hawaii International Conference on* (Vol. 0, pp. 2340-2347).

Edelman, B. (2010). Hard-coding bias in Google "algorithmic" search results. Retrieved from http://www.benedelman.org/hardcoding/

Edelman, B., & Lockwood, B. (2011). Measuring bias in an "organic" Web search. Retrieved from http://www.benedelman.org/searchbias/

Fallows, D. (2005). Search engine users: Internet searchers are confident, satisfied and trusting–but they are also unaware and naive. *Pew Internet & American Life Project*, 1-36. Washington, DC: Pew Internet & American Life Project.

Frants, V. I., Shapiro, J., & Voiskunskii, V. G. (1997). *Automated information retrieval: Theory and methods*. San Diego, CA: Academic Press.

Granka, L. (2010). The politics of search: A decade retrospective. *The Information Society*, *26*(5), 364-374.

Grimmelmann, J. (2010). Some skepticism about search deutrality. In B. Szoka & A. Marcus (Eds.) *The next digital decade: Essays on the future of the Internet*, 435-460. Washington, DC: TechFreedom.

Gulli, A. (2005). The indexable Web is more than 11.5 billion pages. *14th international conference on World Wide Web* (pp. 902-903). New York: ACM.

Hotchkiss, G. (2006). *Eye tracking report: Google, MSN, and Yahoo! Compared*. British Columbia, Canada: Enquiro, Kelown,.

Huffman, S. B., & Hochster, M. (2007). How well does result relevance predict session satisfaction? *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 567–574). New York: ACM.

Höchstötter, N., & Lewandowski, D. (2009). What users see – Structures in search engine results pages. *Information Sciences*, *179*(12), 1796-1812.

Jansen, B J, Booth, D. L., & Spink, A. (2008). Determining the informational, navigational, and transactional intent of Web queries. *Information Processing and Management*, *44*(3), 1251-1266.

Jansen, B.J. (2007). The comparative effectiveness of sponsored and nonsponsored links for Web e-commerce queries. *ACM Transactions on the Web, 1*(1), 1-25.

Jansen, B.J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management*, *42*(1), 248-263.

Jansen, J. (2011). *Understanding sponsored search: Core elements of keyword advertising*. Cambridge, MA: Cambridge University Press.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting click-through data as implicit feedback. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 154–161). New York: ACM.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., & Gay, G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. *ACM Transactions on Information Systems*, *25*(2), article 7.

Jung, S., Herlocker, J. L., & Webster, J. (2007). Click data as implicit relevance feedback in web search. *Information processing & management*, *43*(3), 791–807.

Kammerer, Y., & Gerjets, P. (2012). How search engine users evaluate and select Web search results: The impact of the search engine interface on credibility assessments. In D. Lewandowski (Ed.), *Web Search Engine Research* (pp. 251-279). Bingley: Emerald.

Kang, I. H., & Kim, G. C. (2003). Query type classification for Web document retrieval. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (pp. 64–71). New York: ACM.

Keane, M. T., O'Brien, M., & Smyth, B. (2008). Are people biased in their use of search engines? *Communications of the ACM*, *51*(2), 49-52.

Lewandowski, D. (2004). Date-restricted queries in Web search engines. *Online Information Review, 28*(6) 420-427.

Lewandowski, D. (2008a). A three-year study on the freshness of Web search engine databases. *Journal of Information Science*, *34*, 817-831.

Lewandowski, D. (2008b). The retrieval effectiveness of Web search engines: Considering results descriptions. *Journal of Documentation*, *64*(6), 915-937.

Lewandowski, D. (2008c). Problems with the use of Web search engines to find results in foreign languages. *Online Information Review*, *32*(5), 668-672.

Lewandowski, D. (2008d). Search engine user behaviour: How can users be guided to quality content? *Information Services & Use*, *28*, 261-268.

Lewandowski, D. (2011). The retrieval effectiveness of search engines on navigational queries. *ASLIB Proceedings, 61*(4), 354-363.

Lewandowski, D. (2012). A framework for evaluating the retrieval effectiveness of search engines. In C. Jouis, I. Biskri, G. Ganascia, M. Roux (Eds.). *Next generation search engines: Advanced models for information retrieval* (pp. 456-479). Hershey, PA: IGI Global.

Lewandowski, D., & Höchstötter, N. (2008). Web searching: A quality measurement perspective. In Amanda Spink & M. Zimmer (Eds.), *Web search: Multidisciplinary perspectives* (pp. 309-340). Berlin: Springer.

Lewandowski, D., & Höchstötter, N. (2009). Standards der Ergebnispräsentation. In D Lewandowski (Ed.), *Handbuch Internet-Suchmaschinen* (pp. 204-219). Heidelberg, Germany: Akademische Verlagsgesellschaft Aka.

Lewandowski, D., Wahlig, H., & Meyer-Bautor, G. (2006). The freshness of Web search engine databases. *Journal of Information Science*, *32*(2), 131-148.

Li, Z., & Shi, S. (2008). Improving relevance judgment of Web search results with image excerpts. *17th international conference on World Wide Web* (Vol. 17, pp. 21-30).

Lorigo, L., Haridasan, M., Brynjarsdóttir, H., Xia, L., Joachims, T., Gay, G., Granka, L., et al. (2008). Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology*, *59*(7), 1041-1052.

MacFarlane, A. (2007). Evaluation of web search for the information practitioner. *Aslib Proceedings: New Information Perspectives*, *59*(4-5), 352-366.

Machill, M., Neuberger, C., Schweiger, W., & Wirth, W. (2004). Navigating the Internet: A study of German-language search engines. *European Journal of Communication*, *19*(3), 321-347.

Mansourian, Y. (2008). Web search efficacy: definition and implementation. *Aslib Proceedings*, *60*(4), 349–363.

Nicholson, S., Sierra, T., Eseryel, U. Y., Park, J.-H., Barkow, P., Pozo, E. J., & Ward, J. (2006). How much of it is real? Analysis of paid placement in Web search engine results. *Journal of the American Society for Information Science and Technology*, *57*(4), 448-461.

Nielsen, J., & Tahir, M. (n.d.). *Homepage Usability: 50 websites deconstructed*. Indianapolis, IN: New Riders Publishing.

Pan, B, Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In Google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication*, *12*, 801-823.

Pirolli, P. (2009). *Information foraging theory: Adaptive interaction with information*. London: Oxford University Press.

Rose, D. E., & Levinson, D. (2004). Understanding user goals in Web search. *Proceedings of the 13th international conference on World Wide Web* (pp. 13–19).

Sullivan, D. (2003). Searching with invisible tabs. *Search Engine Watch*. Retrieved from http://searchenginewatch.com/showPage.html?page=3115131

Sullivan, D. (2011). Study: Google "favors" itself only 19% of the time. *Search Engine Land*. Retrieved from http://searchengineland.com/survey-google-favors-itself-only-19-of-the-time-61675

Thurow, S. (2007). *Search engine visibility*. Berkeley, CA: New Riders.

Thurow, S., & Musica, N. (2009). *When search meets Web usability*. Berkeley, CA: New Riders.

Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: Evidence and possible causes. *Information Processing & Management*, *40*(4), 693-707.

Vaughan, L., & Zhang, Y. (2007). Equal representation by search engines? A comparison of websites across countries and domains. *Journal of Computer-Mediated Communication*, *12*(3), 888-909.

W3 Schools Browser Statistics. (2011). Retrieved from http://www.w3schools.com/browsers/browsers_stats.asp

Xie, M., Wang, H., & Goh, T. N. (1998). Quality dimensions of Internet search engines. *Journal of Information Science*, *24*(5), 365-372.