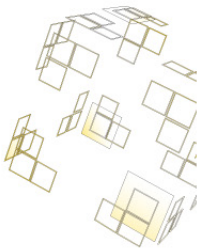
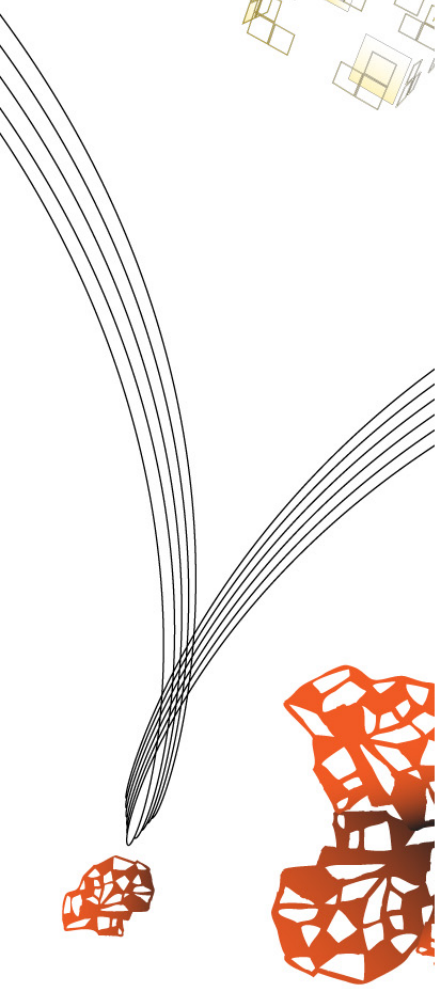
Displaying intermediate results for on-going searches

MSc Assignment

by

Sebastiaan Vercammen





Supervisors:
Dr. ir. Djoerd Hiemstra
Dr. Robin Aly

Database Group
Faculty of EEMCS
University of Twente

December 12, 2014

UNIVERSITEIT TWENTE.

Abstract

Distributed search introduces problems with resources that require time to process queries and produce results, and users waiting to get an answer to their query. The system could wait a maximum amount of time for every resource to produce its results or start displaying results the very moment they are retrieved by the distributed search engine. This thesis introduces a number of alternative display strategies and describes a method to research their effectiveness in providing the most relevant results, as quickly and as high in the combined results as possible, while maintaining a user-friendly search experience. It then continues by describing the performed research and its results. For each experiment, test participants are asked a number of questions, to describe their experience operating the search engine using the specific display strategy. Also recorded are statistics concerning test participants' clicks. These metrics are combined with the answers to the user questions and also used for determining the best display strategy. Observations were made of aspects that seemed to have influenced the experiment, such as the red color of the notifications used for one of the display strategies. The precise influence of these aspects should be further studied, by using A/B testing, as proposed in section 7.2. Finally, the conclusion is drawn that the *Screen fill with "endless" scrolling* display strategy (section 3.3.4) performed best when taking the test participants' answers into account.

Contents

Chapter 1

Introduction

With distributed search (e.g., federated search [1], meta search [2, 3] or peer-to-peer search [4]) new opportunities and challenges arise. In contrast to a centralized search engine, search results will be gathered from a host of resources that are beyond the control of the distributed search engine. Most situations will include resources (i.e., a secondary search engine, whose results are incorporated into the combined result of a distributed search engine) that require some amount of time to process a search request. This means a large proportion of the combined search is spent waiting for these resources to process the search query and to gather and combine their results. Since users typically abandon a search after a short amount of time [5, 6], a search engine cannot afford to wait this long and must start displaying search results as early as possible. However, when starting to display results before the overall combined search has completed, it is likely that more relevant results than the ones already displayed may be retrieved. Since users mostly focus on the first few results [7], a method is needed to quickly display new results in a user-friendly manner while insuring the most relevant results are displayed as high in the combined result list as possible. This need is further enhanced by the emergence of some highly specialized search systems, such as systems searching through enormous amounts of log files or systems that search for people using facial recognition, which at least nowadays takes a long time before results are found and presented

In order to try to reduce this problem a number of display techniques will be proposed. They will range from naive, simply appending a set of results to the combined list, to interactive where the user is notified of newly retrieved results that are considered to be more relevant than the ones displayed. In order to test the effectiveness of these techniques, a number of test participants have been asked to use each strategy in a simulated environment which allows for measuring of a number of metrics, such as the time it takes a participant to click a result or the location of that result at the moment of the click. Using these measurements it can be tested whether any of the proposed techniques will yield an improvement over the baseline system described above. In addition, each participant will be asked questions regarding their experience using each proposed technique. Most of these questions ask for subjective judgements of the used display technique and will be answered on a closed Likert scale (i.e., numeric answers with an uneven number of options, allowing for an *average*

answer). The answers to these questions can then later be used to rate each technique and to see whether any yields a subjective advantage over the baseline system.

1.1 Research Question

In this thesis a number of different methods of displaying search results in distributed or peer-to-peer search engines will be tested. In this context, a distributed search engine is an engine that itself does not poses a database of documents in which to search, but rather outsources each search query to any number of other search engines (resources). For each of these resources, an answer to the query is collected and these answers are combined and then displayed in unison. A peer-to-peer search engine is a variation on this concept, in which the list of resources is not fixed. Peers ask each other which resources are recommended for a particular search query. Whenever a peer then learns of a new resource, suitable for a specific topic, it is from then on able to inform other peers of this resource.

Here, the main research question is: *On the subject of distributed search (e.g., peer-to-peer or federated search), do methods exist to present intermediate search results that allow early access to relevant results?* In order to answer this question, for this thesis, the following questions are researched:

- Whether there is a method of displaying results that provides greater user satisfaction than waiting for all results to be retrieved;
- Whether the user prefers more rapidly displayed search results, even if it means presenting a number of less relevant results first;
- Whether the proposed strategies improve relevant result distribution (Section 4.1).

1.2 Structure of thesis

This thesis is structured as follows: Chapter 2 details common problems on building and operating search engines, users using search engines, and considerations of data presentation. Chapter 3 describes what will be researched and which method will be used. Chapter 4 focuses on how the research question will be answered and which requirements answering the research question produces. Chapter 5 discusses the results of the research, which include the answers to user questions after experiments and recorded metrics. Chapter 6 then makes a number of observations that influenced the research and provides hints on how these influences could be negated in future research. Chapter 7 proposes how this study can be continued in future work. Chapter 8 then concludes with a summary of this thesis.

Chapter 2

Background

Related to search engine and this thesis' research question, *on the subject of distributed search (e.g., peer-to-peer or federated search), do methods exist to present intermediate search results that allow early access to relevant results?*, previous research has been done. This chapter discusses this previous work. It show aspects such as the importance of good search result order, bias of users towards result domains and the need for speed.

2.1 Results ranked 1–4

Eye tracking studies [8, 7] have been performed to measure where the user looks and what they see when using a search engine. Users look mostly at the top items on a search result page. The number of results that are considered vary. Joachims et al. [9] found that the first two results receive the most attention while Cutrell and Guan [10] found that users usually at least looked at the first three or four results. Lorigo et al. [11] found that, on average, about three results per page were viewed at all, although they do not specify the position of these results on the page, nor whether the starting point is always the same.

2.2 The number of results that generally are considered

Users generally consider about 6–8 search results before clicking a result [9, 10]. This number of results usually is the amount of results that fit on one page without scrolling [7]. The most time is spent reading results 1 and 2 [7]. Joachims et al. [9] found that users generally scan the viewable results thoroughly before resorting to scrolling. Furthermore, Lorigo et al. [11] found that only 4% of the users who do not click a result, scan all 10 results on a Google result page. They conclude that users generally quickly determine whether they will click a result or reformulate their query. According to Lorigo et al. [11], in 96% of queries the user only looked at the first page and no user went beyond the third result page. Their results indicate that only the most highly ranked search results are likely to be exposed to the user.

2.3 The importance of correct order

The order of results influences which search results the user will read and which they will probably not read. To ensure they read the most relevant results first, it is important to know in which order results are generally processed.

Cutrell and Guan [10] found that users tend to view search results in roughly linear order, with most attention to the first results and the low ranked results (i.e., those at the bottom of page) viewed last and least.

Joachims et al. [9] found that on average, users tend to read results from top to bottom. They tend to view the first and second-ranked results within about three seconds and then this time increases to 7–13 seconds for results 3–6. From result 7 on this time increases significantly, likely as a result of the results not being visible without scrolling [7]. They also found that when swapping the first two results on a Google search result page, users still were biased towards the first result, even though the second result was then more relevant. Reversing the order of all results on a search result page caused users to significantly scan more results than in the unchanged situation; it also reduced the number of clicks. They conclude that the order in which search results are ranked influences the user's clicking behavior: if the relevance of the retrieved results decreases, users click on more results that are on average less relevant.

Granka et al. [7] too found that users spend the most time reading the first two results. After that, time spent reading results sharply decreases. They also found that users significantly more often click the first result[1].

Test participants spent the most time reading the first two result search results (about 0.8–0.9 seconds) and less than a tenth of a second from the sixth result on [7]. According to Cutrell and Guan [10], users arrive at the first result within a second, and spend, on average about 2.3 seconds reading that search result. The second search result is reached in, on average, less than 1.5 seconds with an average time spent reading of about 1.5 seconds. Time spent reading results keeps decreasing to less than 0.5 seconds from approximately the eight search result on. They also note that users changed their query after considering eight search results without clicking any.

Klöckner et al. [12] found that about 80% of their subjects evaluated search results in a strict or partial depth first strategy, meaning they choose whether to click a result after reading only a single or a small number of search results. Most subjects in that group did not scan ahead at all, with only 20% of the subjects scanning ahead only a few results. 15% of the users used an extreme breadth-first strategy, looking through the entire list before clicking a result. Their results underline speed is of the essence, since a user reading a search result may decide to explore (and thus, click) that result while more relevant results are still to be retrieved.

2.4 Search engine bias

Search engine users are biased towards search results and the relevance of search results. In addition to the order in which search results generally are considered

[1] The first result was selected in about 150 out of 397 experiments; the second search result was selected less than 40 times.

(Section 2.3), users tend to generally click the first result, even if the second result is generally considered to be more relevant or when the first two results are swapped [9]. They trust the search engine in producing the most relevant items in the correct order and do not question its results. This behavior increases the importance of displaying the most relevant result first.

In addition, Ieong et al. [13] found that in about 25% of the cases, users factor in the domain of a search result (e.g., `facebook.com` or `wikipedia.org`). In these cases, they prefer results on reputable domains over those of less reputable domains, even if the result on the less reputable domain is more relevant. They note that in these 25% of the cases, the behavior of users resembles a blind trust (i.e., following links without factual knowledge of the relevance of the document) in these reputable domains, rather than following the results listed as most relevant (i.e., at the top of the page) by the search engine. They also note that the concentration of search results on fewer domains increases click-through rates (i.e., more clicks to these reputable domains).

2.5 Search engine modifications and manipulations

Search engines are free to modify their results. They can, and do, this for a variety of reasons, which are not always publicly known. They can, for example, exclude items that conflict with terms of operation [14] or improve the position of advertised results. Modifications may also be court ordered [15] or may be subject to censorship [16].

Aside from active moderation of search results, search engines can make editorial choices in order to satisfy their audience [14] or give prominence to popular, wealthy and powerful sites at the expense of others [17]. Lastly, they can fall victim to their own success, with people or institutions learning how to influence search results [18, 19].

By utilizing and combining the results of a multitude of search engines, a more complete image is formed. A search engine omitting a result has less impact when another search engine does include that result. Likewise, the effect of one search engine pushing a result (e.g., an advertisement) to the top is lessened when other engines do not push this result. It may also help reduce the portions of the web that remain hidden from view [17].

2.6 The need for speed

As with most services on the internet, search engines need to quickly respond to user questions. As Brutlag [20] notes, a decrease in speed (in his case by introducing artificial delays) decreases the number of daily searches and user satisfaction. Granka et al. [7] measured the time in which a user selects a document to be between 5 and 11 seconds, averaging 7.78 seconds. In this thesis, and for the research to be done, it is assumed that these timings also apply for systems that incrementally produce results (i.e., that produce results over time, rather than all at once).

2.7 Types of search

For every search engine, it is important to produce the most relevant result possible. In order to define relevancy, we need to distinguish the different types of search task a user is performing. Broder [21] specifies three types of search tasks: *navigational*, *informational* and *transactional*. The goal of a navigational search is to find a specific website (e.g., the website of the Dutch Rijksmuseum). Navigational queries usually have only one "right" result. Informational items (e.g., Amsterdam landmarks) are assumed to be available in static form (i.e., not created in response to the user query) and may be available on more than one location. Transactional searches lead to sites where further interaction will happen (e.g., downloading of software, shopping for goods, finding servers for gaming). Rose and Levinson [22] redefine the transactional type to a more broad *resource* type, defined as: "[...] to obtain a resource (not information) available on web pages". This type of search task is also not necessarily restricted to a single location.

Broder [21] performed a query log analysis and a user survey on search engine use and found that about 20–24.5% of the search tasks were navigational, 39–48%[2] were informational and about 30–36%[3] were either informational or transactional. Rose and Levinson [22], using a different definition of the *transactional* category which they call *resource*, found that 11.7–15.3% of searches were navigational, 61.3–63.0% was informational and 21.7–27.0% was resource oriented. This means that navigational tasks constitute a minority of web searches. As only navigational tasks target specific resources, this means that for the majority of tasks it does not matter as much to which web site a user is sent, as long as that site offers the requested information [10].

2.8 Why multiple sources

In distributed information retrieval systems, the task is to search a group of independent collections, and to effectively merge the results they return for queries [1]. This is useful in examples such as in an environment where a general search engine (e.g., Google or Yahoo) can not access the documents for indexing and retrieval, either because of a physical (i.e., no access to the target network) or a virtual limitation (i.e., the standard `robots.txt` file). It also allows a search engine to specialize in one of the three search tasks [21].

It is important to consider that the indexable web is big and keeps on growing. In 1997, Bharat and Broder [23] estimated the number of pages in the static web, to be at least 200 million. In 2005, it consisted of an estimated 11.5 billion indexable pages [24]. In 2008 Google had found 1 trillion unique pages [25]. It is not likely that the number of indexable pages has since stopped increasing. According to Gulli and Signorini [24], search engines do not index the same collection. By combining the results of multiple search engines, a better coverage of the web is achieved.

By combining the results of multiple search engines, results will be smoothed out. As results that occur in multiple result sets rise to the top of the combined

[2] 39% percent is an estimation

[3] 36% is an estimation, he states the value to be at least 22%

result page, (self) promotions, advertisements and other modifications (Section 2.5) will fall to the lower ranks in the combined search results, producing a more neutral view.

2.9 Summary

In this chapter a number of subjects influencing search engines and their use were discussed. Section 2.1 finds the number of results that users generally consider to be about four, with most attention drawn to the first two results. Section 2.2 explained that users generally only consider results on the first result page, with the most attention given to about 6–8 results before clicking a result. This amount of results is usually the amount of results that fit on one page without scrolling. Section 2.3 discusses the importance of correct result order. Search engine users generally tend to scan search results top to bottom, in linear order, giving the most attention to the first few results. They see the first two results in about three seconds and results 3–6 in 7–13 seconds. From the seventh result on this time increases significantly. Section 2.4 briefly touches on the risks of search engine bias, where users tend to click on results that stem from familiar websites, even when more relevant results are available. Section 2.5 explains how search engines are free to modify their results. They may be coerced into doing so, or may change their results for financial gain. When combining search results from multiple search engines, this risk is somewhat reduced. It may also help reduce the portions of the web that remain hidden from view. Section 2.6 briefly discusses the need for speed, as the user usually selects a document in 5–1 seconds. It was also found that slowing down search engines reduces the number of daily searches. Section 2.7 then discusses different types of search and what these types of search usually mean for the number of results expected from each type. Finally, section 2.8 shows why it is a good idea to use multiple resources, rather than a single search engine. A single search engine will probably not be able to index the web in its entirety or be specialized to a small number of topics. Multiple search engines may fill each others gaps in this respect.

Chapter 3

Research method

In order to answer the research question, *on the subject of distributed search (e.g., peer-to-peer or federated search), do methods exist to present intermediate search results that allow early access to relevant results?*, a number of techniques for displaying search results have been tested (Section 3.3, Display strategies). Queries have been run on a simulated federated search engine, using a data set of queries and results for these queries gathered for earlier research (Section 3.8). In order to better simulate reality, each of the resources exhibited a delay before presenting its results (Section 3.1). This delay is partially based on delays observed when querying real-life search engines, in addition to custom delays to better show the effect of each display strategy.

3.1 Assumptions

In order to limit the number of variables that influence the research, a number of assumptions have been made. First, when conducting user testing, all search results were served from a data set of search results gathered for *FedWeb 2012* (Section 3.8). This eliminated the variable amount of time each questioned resource takes to answer a query as well as the time required for the transportation of the answers.

To simulate the nature of the Internet, each result was served following a short delay. To ensure that repeating tests would yield the same delays for each result, each delay has been predetermined in an attempt to evenly distribute response times of relevant and non-relevant resources (e.g., when searching for *jaguar*, search engines representing zoos and car manufacturers are considered to be more relevant).

Search rating schemes generally are not available. For this research though, the FedWeb 2012 data set (Section 3.8) provides relevancy judgements for each snippet and target page with respect to the search task. These judgements are aggregated for each search task/resource combination. The aggregated scores can be used to order the resources by relevance. Then, for each result the snippet score is multiplied with the peer score, resulting in an individual score for a search result. These individual scores are then used to order the results within the combined result set. Instead of using round robin merging, this

ensures that non-relevant results are less likely to creep up to the top of the combined results.

3.2 Data presentation

When using multiple search engines to compile one combined overview of results, one of the questions is: how to display the results in such a manner that relevant results are shown early (i.e., high in the list) while reducing the time a user waits for search results to be retrieved.

Since there is a variance in response speeds between search engines (caused both by processing time as well as network transport) some search results will be retrieved sooner than others. Unfortunately, the relevant search results are not always the first to be retrieved. While it is important to display these relevant search results as soon as possible, a typical user is unlikely to be willing to wait for a long time while the search is carried out and responses are being processed. It is therefore important to display some search results soon, so that the user both knows that the search is indeed being executed and has some search results they can assess soon after starting the search. Likewise, it is important to keep working to display relevant results as high in the combined list as possible, even when less relevant search results already have been displayed.

3.3 Display strategies

Search results are presented using a number of strategies, which represent a diverse range of levels of interaction, ranging from fully passive, to requiring user interaction in order to display additional results. Each of these strategies is explained later in this section:

1. Wait for all results to be retrieved; this is the baseline system;
2. Directly display result sets when retrieved;
3. Re-order displayed search results when results are retrieved;
4. Screen fill with “endless” scrolling;
5. Visual indication of more relevant results.

The list is limited to these options because there is little overlap between the strategies. Though variations can be thought of, they would be just that: variations on strategies that already are proposed. Such variations would only increase experiment complexity and would in a sense not add any useful data. They would also require more testing in order to achieve the same level of details and would either increase the workload of test participants or increase the number of test participants required.

An interesting case is a result that is initially ranked low in the combined result list but becomes more relevant as more resources produce that same result. When multiple resources produce the same result, it is deemed to be more relevant than results that only a single resource produced. The problem is that certain strategies (e.g., *screen fill with “endless” scrolling*) assume results

do not move once displayed. Some strategies however (e.g., *re-order displayed search results when results are retrieved*), do allow for results moving throughout the combined result list. When preparing the experiments, an assessment was made that re-ordering items already displayed would add little improvement over only reordering items to allow for new results to be displayed. Therefor, it was determined to remove duplicate results and to let the slowest resource produce the duplicate result. Usually a result becomes a duplicate result when a slower resource yields a result that was already produced by a faster resource. In order to help make the treatment more easily visible, this order was reversed such that the slower resource will be the one producing the result, instead of the faster resource. Being produced by multiple resources, the result will likely be considered to be relevant. Being produced by the slowest resource, this yields a worst-case scenario as the result would normally be produced more quickly. This worst-case scenario though, is best for testing whether the proposed display strategies yield improvements over the baseline system (see section 3.6).

3.3.1 Wait for all results to be retrieved

This strategy is the slowest of all. The display strategy will wait until *all* results are retrieved, and will then display a combined list of results. The main advantage is that the most relevant items can always be displayed at the top of the list, where user attention is mostly focused. The main disadvantage is that the display strategy will wait until every resource responded or has timed out. This can take significant amounts of time, especially when the time-out is set to a time that is as long, or longer, than the amount of time a user is willing to wait [5, 6]. Because the user may not be aware that a search is actually happening, visual feedback using a progress bar is provided. This progress bar indicates the *maximum* amount of time remaining and will fill up when resources return their results.

3.3.2 Directly display result sets when retrieved

This strategy appends results to the combined result list as soon as they are retrieved, without considering result order or merging results of other sources. This means that every set of results is displayed in whole and relevant results may be spread over the entire list of results in a saw-like pattern, because every first item in a result set will be considered to be more relevant than the last result of the previous result set, even though this first result is printed below the entire previous result.

3.3.3 Re-order displayed search results when results are retrieved

Results will be displayed as soon as they are retrieved, but instead of appending them to the combined result list, the combined list will be re-ordered in order to accommodate the new results. In order to give visual feedback of the ordering process, search results will move using simple animations as described in the jQuery manual[1].

[1] `http://jqueryui.com/show/`, accessed February 26, 2014

This strategy allows testing whether users find lists with moving results to be confusing and whether they think that results keep improving while they are evaluating them. A major downside is that, when the user is scrolled some way down the result list and a new set of results comes in, the user may not be aware that a high ranking result is inserted into the page, above their viewport. This can cause them to miss some results.

3.3.4 Screen fill with "endless" scrolling

This strategy consists of three states: screen not yet filled, screen filled, and user scrolling. Whenever the screen is not yet filled, the first strategy (*Directly display result sets when retrieved*, Section 3.3.2) is utilized. As soon as the screen is filled (with some overflow to allow for scrolling), results are not displayed but rather cached while they are retrieved. When the user starts scrolling down and the overflow reaches a minimal threshold, new items are displayed from the result cache.

This strategy allows for items to stay in transit for longer times while ensuring that the highest scoring items (as determined by their weighted round robin scores) are presented early in the result list. The major benefit of this strategy is that the user needs to wait only for a minimal amount of time before items are displayed, while trying to present more relevant items as early as possible. This mostly resembles search engines the user is already familiar with (e.g., Google, Bing, Yahoo). The downside of this strategy is that the first batch of items to be presented may contain less relevant results as the first set of results is directly displayed. Also, since users have been observed to not usually scroll through search results [9] this strategy may never reach its full potential.

3.3.5 Visual indication of more relevant results

This strategy does, unlike the other strategies, not automatically restructure displayed results when new results are retrieved. Rather, it displays a visual cue that more relevant results are available but only displays those results when the user indicates they want to see these new results. This will be implemented by tinting the background for a result x for which more relevant results x' are available, and adding a textual cue above result x. The text of this textual cue is *Found N more results at this position; click to display* (see Section 6.2). Only when the user clicks this textual cue, will the new results x' be displayed by inserting the new results x' above result x.

3.4 Presentation delay

Aside from the simulated delay due to network transport, each of the strategies will be subjected to a base delay. This delay is measured in milliseconds and would in a real-life scenario, allow a strategy to collect some small amount of data before presentation. This delay should not be longer than a typical search engine will exhibit and must not exceed the simulated network transport time as this will effectively reduce every strategy to that of waiting for all resources to return their results. To find an acceptable delay a number of search engines have been queried and have their response time measured. The average response

times have been used in this experiment. Figure 3.1 shows the measured delays, omitting the outlier values. These outlier values are a minority of measurements that are distant from the other observations, with one response time beyond ten minutes. They are omitted because they will otherwise render the graph useless by pushing most data to the very bottom of the graph.

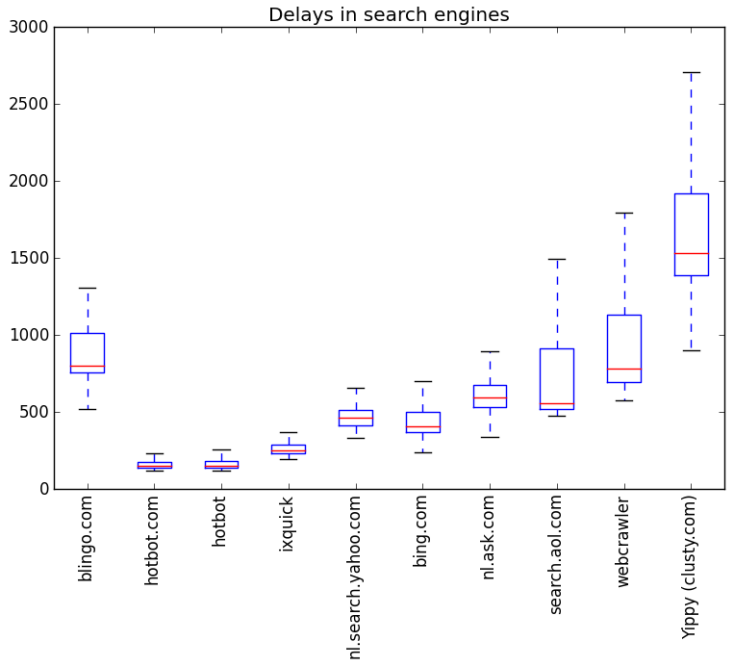


Figure 3.1: Measured delays, omitting outlier values

Since the observed delays were too low to be used for this research — any noticeable effect would be over before the test participant would know what happened — a mix of observed and artificial delays has been used. Figure 3.2 has a list of these delays. As explained in section 3.6, the delays are applied in order of least relevant resource to the most relevant resource, the least relevant resource being fastest.

3.5 Eye tracking

Eye tracking is a very useful tool for investigating user behaviour in computer systems [8, 7]. It can, for example, be used to gain insight into the number of search results a user generally considers [9, 10, 11] before trying a different query. It can also be used to determine the order in which search results are considered and how much time a user spends reading each result and to test whether this amount of time is constant or changes as more results are considered.

For this research, it is a useful tool to use to gain insight into the way test participants experience the tested display strategies. As some of these display strategies display animations or notifications it is useful to determine whether the user really sees the intended visual cue. The downside of eye tracking testing

Delay (msec.)	Measured on
629	nl.ask.com
1087	hotbot.com
1131	Blingo.com
1152	search.aol.com
1408	webcrawler
1899	Yippy (clusty.com)
2000	*custom*
3000	*custom*
4000	*custom*
5000	*custom*

Figure 3.2: Used delays for this experiment

is that equipment is still quite expensive and not available everywhere. This is also the case for this research. Therefor, the use of user questions, such as *I noticed the "found x more results" notice being displayed*, is opted. Answers to such questions do not offer results as definitive as those of eye tracking but provide a good approximation and can be used to determine work following this research. Eye tracking can then be used to further test the best performing alternative as proposed in this thesis.

3.6 Worst case scenario

When implementing display strategies such as the ones proposed in this thesis, response times of queried resources will usually vary (see section 3.4). In a best-case scenario, every resource — or at the very least the most relevant ones — will respond instantaneous. In such a case, the search engine can simply wait and collect all responses before displaying these responses in a list, ordered by relevance. In such a best-case scenario none of the proposed display strategies will provide improvements over the baseline system, which will wait until all responses are received before displaying them. As such, improvements by the proposed systems can not be tested in a best-case scenario. Therefore, for this research a worst-case will be simulated. This will be implemented by simulating the least relevant resources to be the fastest to respond, and the most relevant resources being last to respond.

3.7 What to measure

In addition to eye tracking, click tracking is useful too. When a user clicks on an element that only recently appeared, this implicitly indicates they saw that UI change and are reacting to it. Recording the time between the UI element appearing and the user clicking it can provide further indication whether they see the information directly or only after some time. This however assumes that they click an element as soon as they notice it and ignores the fact that they might first evaluate all visible information and only then resort to clicking.

Finally, after testing each of the variations, the user will be asked to rate some questions on a Likert scale (Section 4.5). Questions include *rate your experience using this interface (1=bad, 7=good).* These questions could not be answered by technical means and using a numerical scale allows for averaging user experience and clearly ranking strategies.

In addition, for each experiment, the following metrics will be recorded:

- The time of each click from the start of the search (i.e., for three clicks, three times are recorded);
- The displayed position of each result at the moment it was clicked;

3.8 Data set

For the user experiment data gathered for *FedWeb 2012* [26] will be used. This data set contains a number of queries and has, for each of these queries, a number of search results from a number of search engines as well as relevance assessments for each result with respect to the query for which that result was retrieved. The use of this data set eliminates the need for a crawler and provides more stable search results. It also provides a challenge though, as it only provides search results for a limited number of queries, from which a selection must be made for the experiment described in this thesis. These queries should be understood by a broad public, as test participants need to use these queries and the gathered results in order to answer the search tasks they will be asked to perform during user testing. The queries also should be close to the query the user would use to answer the search task, preferably the same as they would use if they would have been free to formulate a query themselves.

Chapter 4

Validation

To answer the research question and validate our approach, a number of people will be asked to use the system and perform a series of search tasks using predetermined queries and query results. For each of these searches a presentation technique (Section 3.3) will be selected in a semi-random fashion, ensuring each technique is equally tested. Results are selected from the FedWeb data set (Section 3.8) and will be pre-loaded (i.e., through a JSON[1] file). JavaScript on the test participant's computer will then simulate the federated search engine and the search process by applying a series of events, either timed or manually triggered, specific to the selected presentation technique. In addition, a predetermined delay will be applied in order to simulate delays as experienced on the web (Section 3.4). The test participant will not be made aware of the selected display technique or that no actual live search is executed.

4.1 Distribution of relevant results within one set of results, or one "page"

When a search query is executed, a number of results will ultimately be provided. Not every result is as relevant as every other. In some cases, the search engine will put relevant links on the top of the results page, in some cases the relevant results will be more spread out. By recording the location of clicked links within one page of results, it can be determined how each of the proposed display strategies performs in this respect. In order to better prepare for this, the relevance of search results, as marked in the FedWeb 2012 (Section 3.8) data set, will be used to rate each resource with respect to the search query. In the data set, both snippets describing a search result as well as the page the snippet links to are scored for relevancy. Possible snippet relevancy scores are `junk`, `non`, `unlikely`, `maybe`, `sure` and `answered`. Page relevancy scores are `junk`, `non`, `rel`, `hrel`, `key` and `nav`. These scores are averaged for each snippet and page, eliminating some levels. These averages are re-used, using an incremental numeric score. For snippets, these average scores are: `no (0)`, `unlikely (1)`, `maybe (2)`, and `sure (3)`. For pages, they are: `non (0)`, `rel (1)`, `hrel (2)`, `key (3)`, and `nav (4)`. These snippet and page scores are aggregated then for

[1] JavaScript Object Notation: `http://json.org/`, accessed February 26, 2014

every search task/resource combination such that a relevancy ordering can then be made of resources with respect to a particular search task. This ordering is used to select resources for each search task and to later determine the delay with which the resource is simulated to operate (section 3.4). As snippet scores are natural numbers in the range of 0–3, it is very likely that two results from separate resources have an equal snippet score. This may cause search results from different resources to be essentially sorted in a round robin order. As some resources are considered to be more relevant than orders, this behaviour is not wanted. In order to prevent it, each snippet score is multiplied by the score of their resource. This ensures that results from a more relevant resource will generally score higher than results from a less relevant resource.

For each query, the top five and bottom five resources will be selected (see appendix A), which effectively will cause each result set to consist of roughly half relevant results, and half non-relevant results. There is one exception to this, namely the *When did South Africa gain its independence from the United Kingdom?* search task, where erroneously only one good resource was selected. In this case this resulted in an even worse-case scenario than originally intended.

Because the experiment portraits a worst-case scenario, the order in which each resource will be simulated to respond is from least relevant to most relevant. This means that the most relevant results will always be received last. This will then cause some display strategies to perform more badly in a worst-case scenario, and some display strategies to thrive. Whether this is the case will then be tested by monitoring which links the test participant clicks.

4.2 User testing

To test each strategy, a number of human test participants were be asked to perform a number of searches. For each of these searches, a different search strategy will be selected, making it possible for the test participant to compare their experiences with the different strategies.

For each experiment, each test participant was asked to perform a search task to find an answer to a question, using a predetermined query and set of results, which were stripped from information regarding their origin (e.g., Google or Bing) in order to prevent search engine bias (Section 2.4). An example task would be: *What family of animals does the jaguar belong to?*. The subject would then proceed to use the simulated search engine as they normally would, except only being allowed to view one single result at any given time, to answer the query. After completing and answering the search task, the participant was asked to complete a questionnaire containing questions on the display strategy that was just experienced (Section 4.5). The answers to this questionnaire were closed questions on a Likert scale to allow the ranking of the display strategies.

For this research 23 participants were found willing to participate in the experiment, experiencing each of the display strategies. Each of the experiments were supervised, such that the aim of the experiment was explained prior to the execution of the experiment and any questions that arose could be answered. Though not required, supervision also allowed exclusion of participants that tried to manipulate results or who do not perform the given task (e.g., directly answer the search task without using the search engine or using another search

engine to answer the task).

n the future, it will be useful to employ eye tracking equipment to accurately measure where the participant looks and what they see. This will give further insight into whether the participant has really seen relevant user interface artifacts and to witness the way they interact with these artifacts. It may also provide a basis for future improvements of the tested strategies.

4.3 Time to first click

By recording the amount of time between starting the search and the moment the user clicks a search result, it is possible to measure the time a user needs to click a relevant result. This provides insight into the responsiveness of each individual display strategy and may provide a maximum amount of time a user is willing to wait. It may also provide insight if the unexpected moving of results (Section 3.3.3) confuses the user, causing them to wait for all items to appear, before examining the reported results.

4.4 Time to last click

For every experiment, the last link clicked is considered to have lead the test participant to the page where they have found the answer to their search task. In order to test the speed of each proposed display strategy, the time each participant needs to reach the last click is also recorded. This metric is somewhat skewed because it also includes the time spent on previous clicked links, but will overall still provide insight into the time a test participant requires to find their answer.

4.5 User questions

After completing each search experiment — answering the search task asked in the user test — the user will be asked some general questions on the search experience. Some questions are specific to a display strategy and will only be asked if that strategy was used in the experiment:

1. What is the answer to the question asked? (open question, used to determine whether the test participant truthfully performed their search task);
2. I was easily able to find an answer (1=I do not agree; 7=I strongly agree);
3. How fast did the search engine feel? (1=very slow; 7=very fast);
4. I would use this search engine more often (1=I do not agree; 7=I strongly agree);
5. The "found x more results" notice helped me finding more results (1=I do not agree; 7=I strongly agree);
6. I noticed that the search results would move around (1=I do not agree; 7=I strongly agree);

7. I noticed the "found x more results" notice being displayed (1=I do not agree; 7=I strongly agree);
8. Without the progress indicator, I would not have waited for the search engine to finish (1=I do not agree; 7=I strongly agree);
9. Describe in no more than three sentences how you think this search engine works or what just happened (open question).

4.6 Statistical significance for Likert questions

In case of the answers to the Likert questions (section 5.2), we need to compare the dependent variable *user satisfaction* (a continuous variable as expressed in answers on a Likert scale) with the independent variable *the display strategy used* (a categorical variable). For this, the Wilcoxon signed ranks test is prescribed. This test makes a number of assumptions [27]:

- that the scale of measurement for XA and XB has the properties of an equal-interval scale;
- that the differences between the paired values of XA and XB have been randomly drawn from the source population;
- that the source population from which these differences have been drawn can be reasonably supposed to have a normal distribution.

Using the Wilcoxon signed ranks test, the difference between two sets of measurements (the participants' answers) can be calculated. $W+$ denotes *the sum of positive ranks* (A is higher than B, or baseline score), $W-$ *the sum of negative ranks* (B is higher than A, or alternative score). The test also yields a Z-score, from which p can be calculated. Any $p \leq 0.05$ is considered to be significant. Equal pairs of recorded scores are ignored. This analysis is performed for all common Likert questions, in sections 5.2.1, 5.2.2 and 5.2.3.

Chapter 5

Results

This chapter presents the results of the research, set up as described in chapters 3 and 4. First the test population is described in section 5.1. In section 5.2 the answers given to the user questions are presented and discussed. Section 5.3 then presents and discusses measured metrics. Section 5.4 discusses the observed phenomena of domain bias, where users tend to click results based on the familiarity of the result's domain, rather than position within the result set. Section 5.5 will answer the research question as set in section 1.1. Finally, section 5.6 will summarize this chapter. For readability, each set of histograms has an equal vertical scale.

5.1 Test population

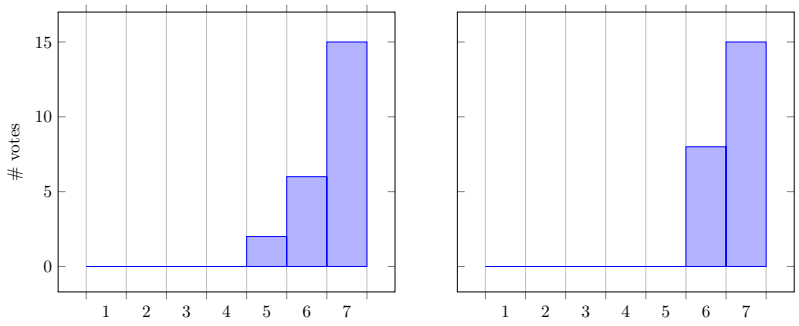
The proposed display strategies were tested using a group of 23 test participants of various backgrounds, ranging from computer science to medicine and from physical therapy to education. Most (20) of these subjects were or are being higher educated, the rest (3) being in a middle educational level. 16 of the subjects were male, 7 were female, and most (22) people were aged in the range of 20–31 years. Most claimed to be experienced using search engines and all used search engines very often, as depicted in figures 5.1a and 5.1b.

5.2 User questions

Directly following each experiment, the test participant was asked a number of questions regarding their experience with the tested display strategy. The answers to the general questions (i.e., the questions following each display strategy) are discussed first, followed by section 5.2.4 where the display strategy specific questions are discussed.

5.2.1 I was easily able to find an answer

The first question to be asked was *i was easily able to find an answer*. Here, the answer refers to the task of finding an answer to the search task for the specific experiment. An example would be the task *in billions of years, how old is the*



(a) experience operating search engines (1=none; 7=a lot)

(b) use of search engines (1=never; 7=daily)

Figure 5.1: Search engine usage by test participants

sun, for which the answer is about 4.567. The test participant is asked this, as it provides an indication whether the display strategy provided results in an order that helped the participant finding the answer to the search task.

Figure 5.2 depicts a series of histograms that represent the participant answers to this question. When subjecting the answers to the Wilcoxon signed ranks test, and comparing each display strategy with *Wait for all results to be retrieved*, we get the results in figure 5.3.

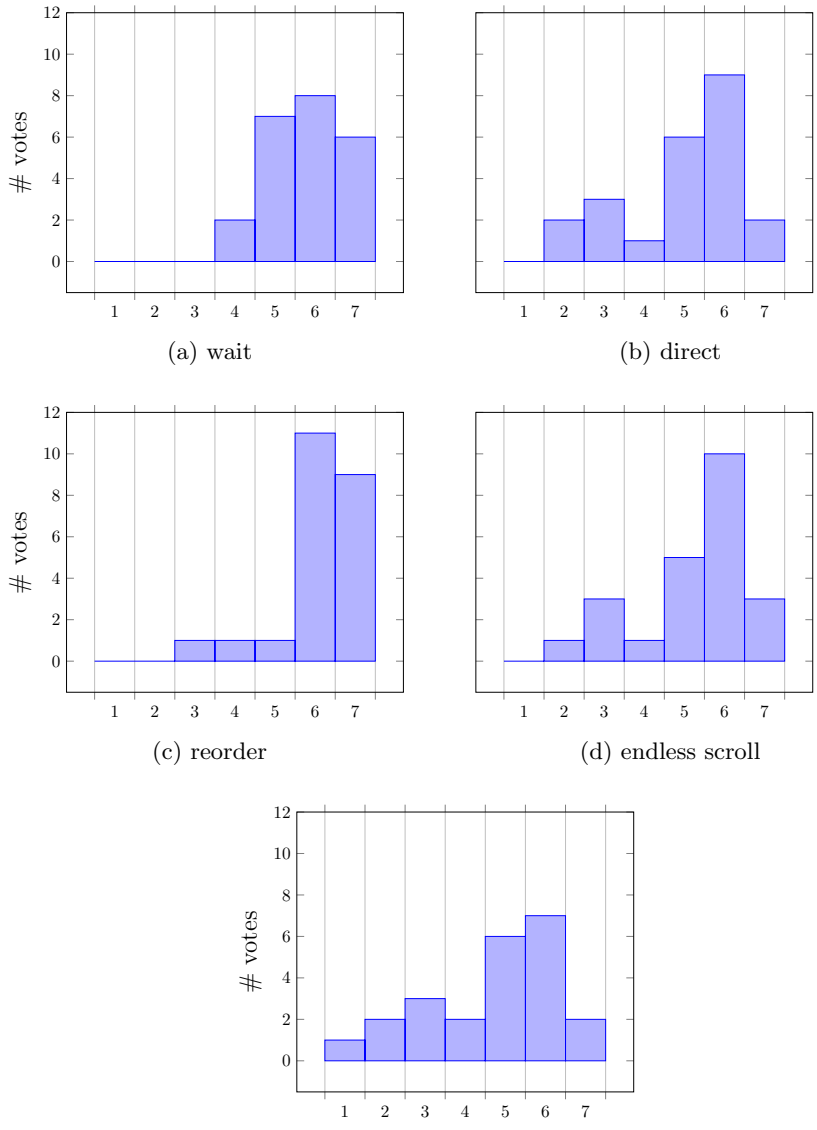
Since the results for the *Re-order displayed search results when results are retrieved* and *Screen fill with "endless" scrolling* display strategies are not significant, we can not successfully reject the null hypothesis. Since the results for *Directly display result sets when retrieved* and *Visual indication of more relevant results* are significant, we will only discuss these results.

Looking at the histogram and $W+$ and $W-$ values for *Directly display result sets when retrieved*, we see that it scores less than the baseline system. This is likely caused by the relevant results being more evenly distributed throughout the search results, rather than being clumped at the top of the results, like the baseline system. *Visual indication of more relevant results* did score similarly, with the baseline getting better marks. This is likely caused by the problems the test participants had operating that display strategy. Problems focused around the notification of more search results, which participants found confusing. As such, they indicated they were less easily able to find their answer than when using the baseline system.

When answering the question *i was easily able to find an answer*, none of the proposed systems – of which significant data was collected – performed better than the baseline.

5.2.2 How fast did the search engine feel?

The second question the test participants were asked, was *how fast did the search engine feel?* Here, the task was to rank the display strategy's speed subjectively on a Likert scale. The results are listed in figure 5.4. When subjecting the answers to the Wilcoxon signed ranks test, and comparing each display strategy with *Wait for all results to be retrieved*, we get the results in figure 5.5.



(a) wait

(b) direct

(c) reorder

(d) endless scroll

(e) visual

Figure 5.2: I was easily able to find an answer; X-axis: Likert score (1–7), Y-axis: number of votes

Strategy (S)	$W+$ $(A > S)$	$W-$ $(S > A)$	p	Significant? $(p \leq .05)$
direct (b)	100.5	19.5	0.02144	yes
reorder (c)	29.5	75.5	0.14986	no
endless scroll (d)	69	22	0.101	no
visual (e)	144	27	0.01078	yes

Figure 5.3: Wilcoxon signed ranks test of *i was easily able to find an answer*; A=baseline

For each of the alternative display strategy, the answers to this question were significant. As such, each will be shortly discussed. What quickly becomes apparent is that the baseline system performed worse than all alternative systems. This is likely due to the fact that it is the only display strategy that waits for the last results to be received, before rendering results. When using the baseline system, participants were required to wait for eight seconds before search results were shown. Even though a progress bar was displayed, informing participants on simulated progress, the system felt slow for most participants.

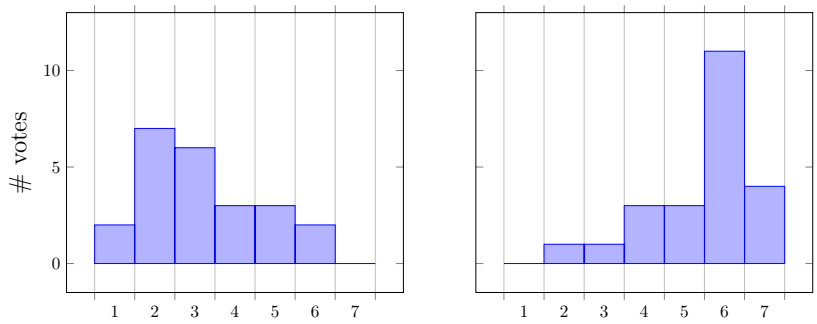
The least improvement over the baseline system was by *Visual indication of more relevant results*. This will likely be caused by the notification of more search results. Participants tended to be confused by this notification and first try alternative results, not covered by the notification. Then, if no answer was found, they resorted to clicking the notification. This may have have caused perception of slowness of the display strategy. It is an effect that might be lessened by continuous use over a longer period of time.

Directly display result sets when retrieved was also ranked to feel faster than the baseline. This display strategy displays results at the precise moment they are received, and as such was the fastest to render the first result. *Re-order displayed search results when results are retrieved* and *Screen fill with "endless" scrolling* also display results at the moment they are received, but handle subsequent results differently. The first has moving and animated results while the second tries to re-order search results in a more transparent method. Not having its results moving may have improved the perception of *being done* operating, thus giving the perception of working faster.

5.2.3 I would use this search engine more often

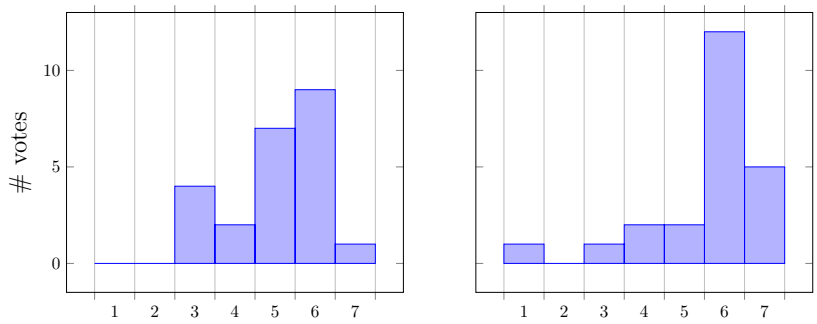
The final general question asked is *i would use this search engine more often*. This question is asked to measure whether the test participant likes the display strategy. Even when a system provides favorable results, the possibility exists that the participant will not use the proposed system when more pleasant alternatives exist. This question is meant to measure this response. Figure 5.6 shows the participant's answers to this question. When subjecting the answers to the Wilcoxon signed ranks test, and comparing each display strategy with *Wait for all results to be retrieved*, we get the results in figure 5.7.

Here we see that, although *Directly display result sets when retrieved* got a favorable $W-$ score, its result is not significant and can as such not be further discussed. *Re-order displayed search results when results are retrieved* got simi-



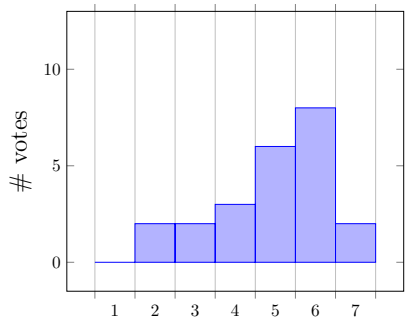
(a) wait

(b) direct



(c) reorder

(d) endless scroll



(e) visual

Figure 5.4: How fast did the search engine feel; X-axis: Likert score (1–7), Y-axis: number of votes

Strategy (S)	$W+$ $(A > S)$	$W-$ $(S > A)$	p	Significant? $(p \leq .05)$
direct (b)	2	208	0.00012	yes
reorder (c)	18.5	212.5	0.00076	yes
endless scroll (d)	7.5	223.5	0.00018	yes
visual (e)	16.5	173.5	0.00158	yes

Figure 5.5: Wilcoxon signed ranks test of *how fast did the search engine feel?*; A=baseline

lar $W+$ and $W-$ scores, but was also not significant. An even better score was for *Visual indication of more relevant results*, but the results for this display strategy could also not be considered as it wasn't significant either. *Screen fill with "endless" scrolling* was significant and had a higher $W-$ score than the baseline system, and can be considered to have scored better than said baseline. For this user question it is the only alternative that can be said to perform better than the baseline system.

Here, the difference between $W+$ and $W-$ is higher for *Screen fill with "endless" scrolling* than it is for *Visual indication of more relevant results*. This indicates that participants favored the former. When subjecting these two systems to a separate Wilcoxon signed ranks test, p was 0.28 and thus the test was deemed not to be significant, though the suspicion remains that *Screen fill with "endless" scrolling* performed best, because it performed the most like a regular search engine such as Google or Microsoft Bing.

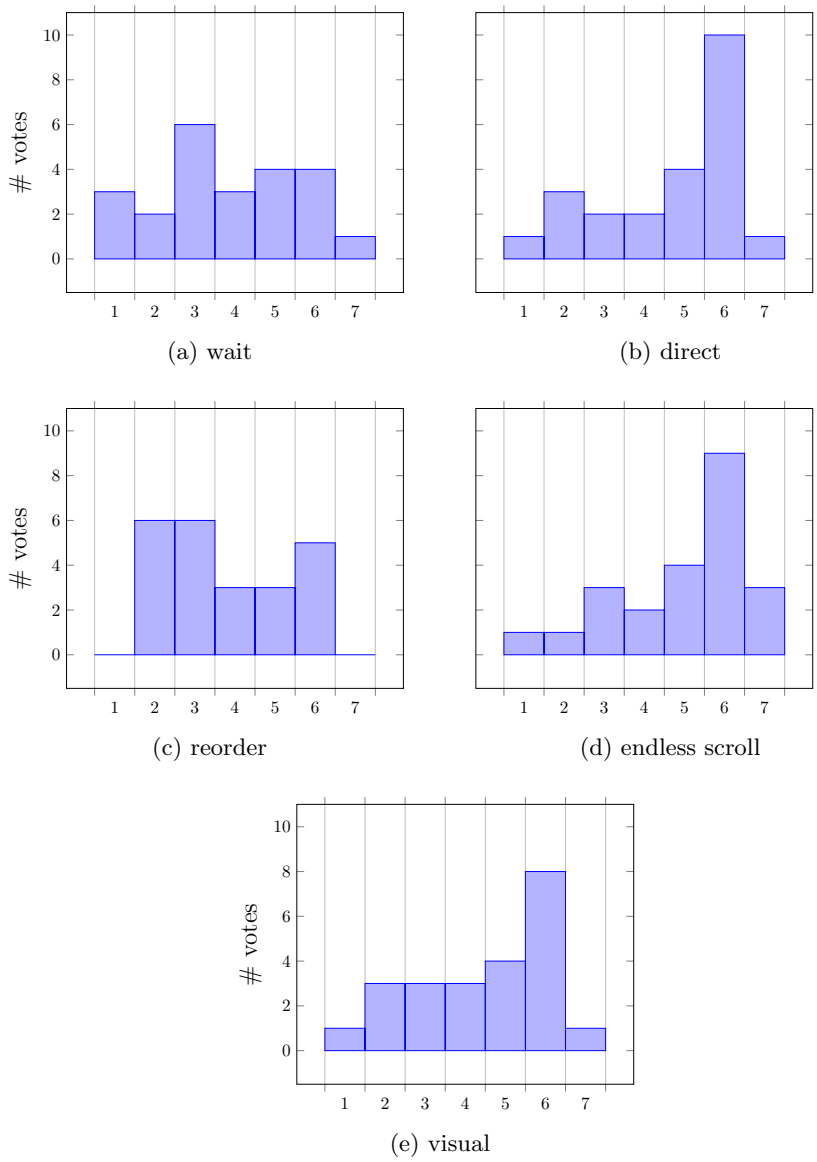
5.2.4 Display strategy specific questions

In addition to questions asked for every display strategy, there were some questions specific to some specific display strategies. Though the answers to these questions can not be compared to the baseline, they may be informative of the participant's experience. This section discusses the answers to these questions.

Progress indicator

The first display strategy specific question was *Without the progress indicator, I would not have waited for the search engine to finish.* This question was asked for the baseline system in order to determine whether the progress bar that was displayed attributed to a better user experience with this display strategy. The results are depicted in figure 5.8

Here, we see a high number of participants indicating they would consider not waiting for the search engine to finish if no progress indicator would be visible. When asked, alternatives ranged from restarting the search by re-pressing the *search* button, to using an alternative search engine to find their answer, if such a search engine were available. This result confirms the view that web users are not willing to wait [5], even when an indication of progress is given.



(a) wait

(b) direct

(c) reorder

(d) endless scroll

(e) visual

Figure 5.6: I would use this search engine more often; X-axis: Likert score (1–7), Y-axis: number of votes

Strategy (S)	$W+$ $(A > S)$	$W-$ $(S > A)$	p	Significant? $(p \leq .05)$
direct (b)	57.5	152.5	0.7672	no
reorder (c)	95.5	94.5	0.98404	no
endless scroll (d)	45	165	0.0251	yes
visual (e)	82.5	148.5	0.25014	no

Figure 5.7: Wilcoxon signed ranks test of *i would use this search engine more often*; A=baseline

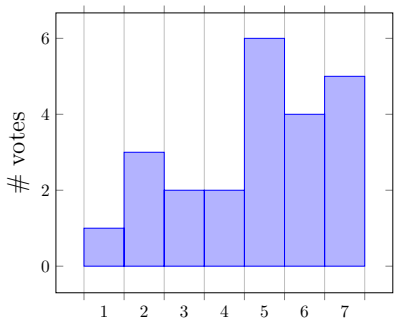


Figure 5.8: Without the progress indicator, I would not have waited for the search engine to finish

Moving results

For the *Re-order displayed search results when results are retrieved* display strategy, the question *I noticed that the search results would move around* was asked. Figure 5.9 shows the answers to this question. This question was asked in order to determine whether the animation speeds were sufficiently large for the effect to be noticeable, which ensures the user has knowledge that more relevant results have appeared above now less relevant results. An answer to this question complements that to the common questions described in the previous sections. Should the animation be too fast, participants would likely not notice the animations and may miss any new results. If animations are too slow, they will hamper the participant's ability to read result titles and snippets while they move. This was generally considered to be annoying.

Here, we see that the majority of participants took notice of the moving results, though a minority (5 out of 23 participants) entered values < 4. This indicates that the animations generally have the correct timings, though for some users they should last longer. Perhaps the length of the animation should depend on slowness of the queried resource; results from a fast resource should result in a quick animation, results that come in after a longer amount of time should perhaps animate more slowly.

Indication of more results

For the *Visual indication of more relevant results* display strategy, two questions were asked: *I noticed the "found x more results" notice being displayed* and *The*

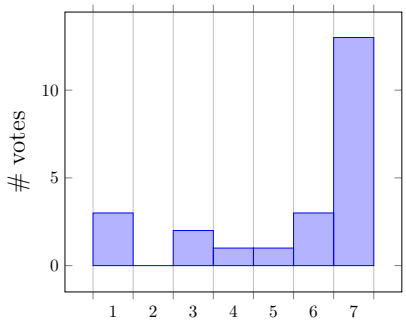
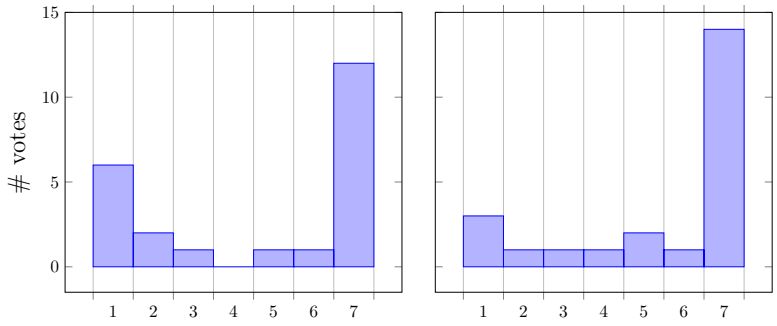


Figure 5.9: I noticed that the search results would move around

"found x more results" notice helped me finding more results. Participants were instructed to enter a average score (i.e., 4) for the latter question when they had answered not to have notices the notification. Figure 5.10 shows their answers.



(a) i noticed the notification (b) the notification helped me

Figure 5.10: Questions for the *Visual indication of more relevant results* display strategy

Here, we witness a divide. The majority of people said they noticed the notification of more results. At the same time, 9 participants said they did not notice the notification. As some of the search tasks allowed finding an answer without clicking the notification of more results, the results in figure 5.10b yields slightly better results. Some more participants reported on their initial reaction. Some, who only saw the notification after some time reported a low score for the question in figure 5.10a, while once seeing the notification, they were helped finding their answer.

5.2.5 What was your favorite search engine

One of the final questions asked was to try and describe the search engine that the participant found most pleasant to use. This question was set up in this way to prevent the test participant from picking an engine from a list. When answering however, participants would generally be unable to give a factually correct description of their favorite search engine (i.e., they described a non-

existing search engine), were ambiguous or could not describe their favorite at all. Therefor, the answers to this question are not considered in this thesis.

5.3 Measured metrics

In addition to user questions, a number of metrics were recorded during user testing. This section details these metrics.

One of the test participants recorded two experiments without registered clicks. This is most likely caused by middle-clicking results, rather than with the left mouse. This event is not captured by client side JavaScript and as such not recorded. Because of this, none of the metrics for this participant are used in this section.

5.3.1 Click statistics

One of the metrics recorded for the user experiment was the time when participant clicked a link. Here, the last link clicked holds special meaning, as the assumption is made that the last link clicked lead the participant to a page finding their answer. Figure 5.11 depicts the time needed for a user to click the last link and finding an answer for their search task.

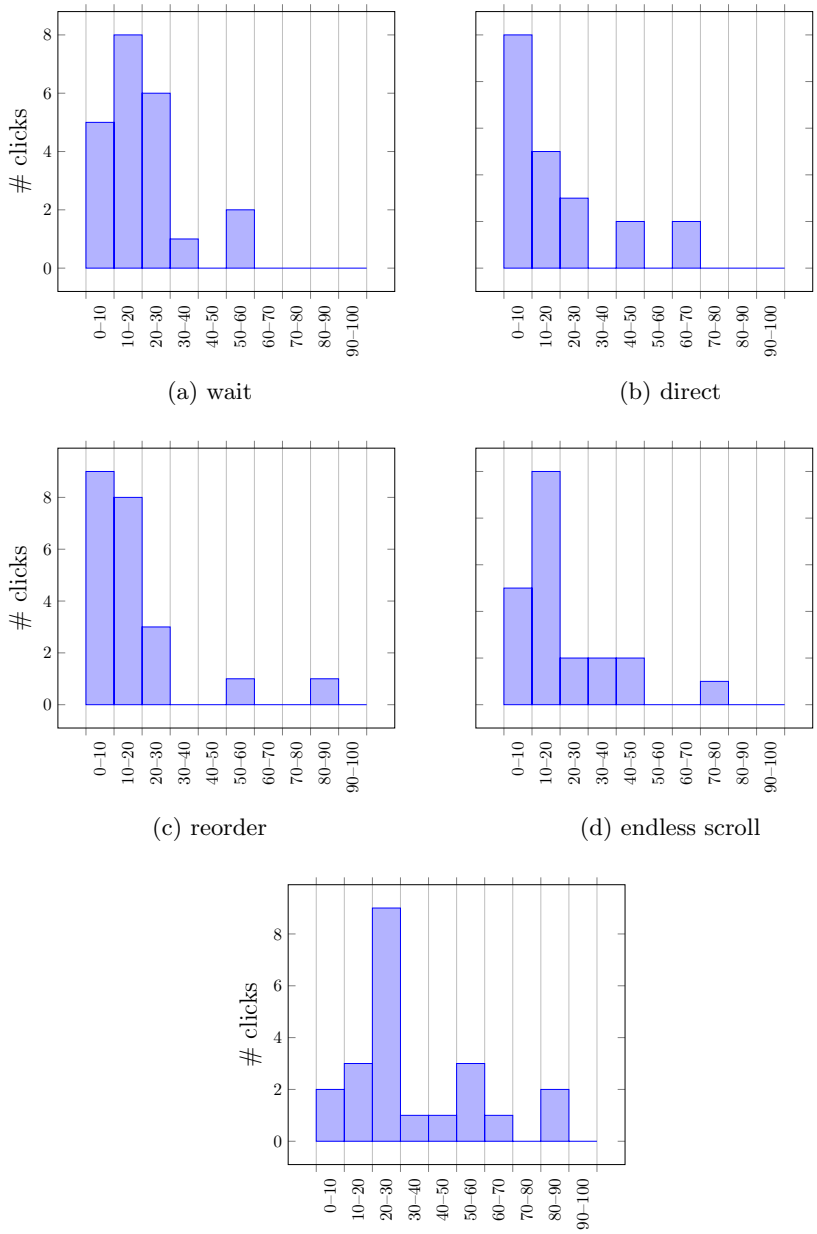
The values recorded were subjected to a t-test in order to determine whether any of the proposed systems performs better than the baseline. The results of this t-test are depicted in figure 5.12. The t-test shows most of the results not being significant, the exception being the *Visual indication of more relevant results* display strategy. This display strategy did perform noticeably worse than the baseline system. This is probably due to test participant confusion as to what the notification (see section 6.2) meant, how to use it and whether it was allowed to be clicked.

5.3.2 Position of clicked results

Another metric was the position of the clicked results at the time of the user clicking them. Figure 5.13 shows a box plot of the positions of the *last clicked* results. These results are considered to be the results that lead the participants to the answer to their search task. What becomes immediately apparent is that the *Directly display result sets when retrieved* display strategy produces the greatest spread of search result positions. This is not really surprising, given how the display strategy operates. Every incoming search result is appended to the result list, regardless of the relevant of the new results. The other display strategies in contrast, perform much better as they actively try to position or reposition search results, either by waiting for all items to be retrieved or actively trying to order displayed results.

5.4 Domain bias

The test users exhibited a strong bias towards familiar websites. This phenomena is called domain bias [13] and causes users to tend to click on links that lead to familiar web sites, even when those links do not belong to the



(a) wait

(b) direct

(c) reorder

(d) endless scroll

(e) visual

Figure 5.11: Time to last click; X-axis: time in seconds; Y-axis: number of clicks

Strategy	T-value	p	Significant? ($p \leq .05$)
direct	-0.2527	0.8030	no
reorder	-0.5904	0.5613	no
endless scroll	0.1038	0.9183	no
visual	2.6156	0.0162	yes

Figure 5.12: T-test results for time to last click; X-axis: time in seconds to last click; Y-axis: number of clicks

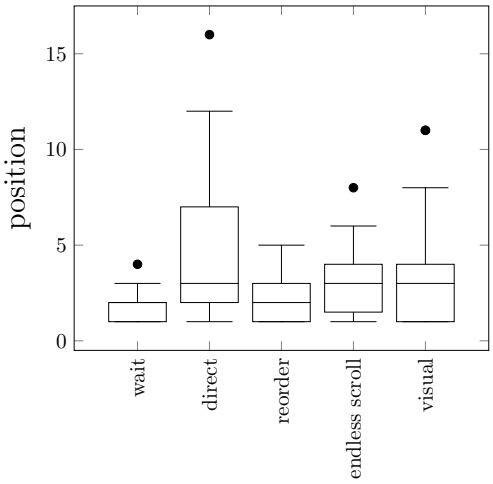


Figure 5.13: Positions of all clicked links

first results of the search engine, which one would expect to lead to the best results (section 2.4). Figure 5.15 enumerates the domains which were visited by the test participants, per search task. It shows a clear preference towards `wikipedia.org`, even though the first search results would not always lead to Wikipedia. It should be noted that, when given multiple results leading towards the same domain – as observed in the South African independence search task – resulted in users clicking the first result, even though the result title clearly indicated that the page would not answer the task. In follow-up research this can be prevented by not displaying the links with the search results, and by obscuring the links by implementing a redirect link from the research system to the targeted result. It may also prove fruitful to strip any domain information from search result titles, such as those for Wikipedia. Figure 5.14 depicts these problems. This figure will not contain the clicked links for the search task *When did South Africa gain its independence from the United Kingdom*, as for this task erroneously not enough relevant resources were selected.

5.5 Answer the to research question

Answering the question which single display strategy is the best is a hard thing to do. This is because of the differences between the answers to each question. As

Figure 5.14: Search result

Search task	Domain	# of clicks
How many people were President of the United States in the 19th century?	en.wikipedia.org	17
	whitehouse.gov	9
	facebook.com	1
How many seasons of the television show "ER" have aired?	nbc.com	13
	imdb.com	7
	tv.com	1
In billions of years, how old is the sun?	en.wikipedia.org	20
	books.google.com	2
What is the currency used in Afghanistan?	en.wikipedia.org	12
	nl.wikipedia.org	8
	state.gov	6
	britannica.com	1
	uni-muenchen.de	1
	nytimes.com	1
When was the album "The wall" by Pink Floyd released	ask.com	10
	nl.wikipedia.org	8
	en.wikipedia.org	1
	amazon.com	1

Figure 5.15: Domain of all clicked links per answer

such, for each of the following three metrics, which relate to the user questions, a winner is determined, which then will be combined into one display strategy that will be elected the best.

If the best display strategy is the one operating the fastest, than for the answers to the user question it is a tie between *Directly display result sets when retrieved*, *Re-order displayed search results when results are retrieved* and *Screen fill with "endless" scrolling*. Each of these display strategies had similar times to last click (though not significant), with the latter scoring the highest when asking participants how fast they felt the search engine responded. When looking at the times to last clicks, then these three display strategies performed similarly to the baseline system. In this case, *Screen fill with "endless" scrolling* is elected to be the winner.

If the best display strategy is the one where users most easily find an answer to their search tasks, then both *Directly display result sets when retrieved* and *Visual indication of more relevant results* got good marks. Both results were significant, with *Directly display result sets when retrieved* scoring the highest $W+$ rank. For results to be easily findable, the clicked results should appear on the top of the result page. All display strategies did this, with the exception of *Directly display result sets when retrieved*, which distributes relevant results

throughout the result page. As such, we have no clear winner.

If the best display strategy is the one which users would use again, then only *Re-order displayed search results when results are retrieved* scored similarly to the baseline system. Though, only *Screen fill with "endless" scrolling* had significant results and thus would win for this question.

Combining these three sub scores, *Screen fill with "endless" scrolling* is mentioned positively the most often. As such, within the scope of this research, this display strategy is the recommended one. It should be noted though, that when looking for more specific conditions, another display strategy might perform better.

5.6 Summary

In this chapter, first the test population was described, followed by an analysis of the answers to the user questions. Then the measured metrics were discussed. Then the observed domain bias was shortly described.

The first user question, *i was easily able to find an answer*, only got significant results for the *Directly display result sets when retrieved* and *Visual indication of more relevant results* display techniques. To this question, users answered positively for both alternatives, with the latter receiving more $W+$ points than the former. This was explained by the former display technique displaying search result subsets one after another, and the latter display technique reordering the search results in order to display the most relevant results at the top of the list, making it easier to find an answer to the user's assignment.

The second user question, *how fast did the search engine feel?* had significant results for every display technique. Each proposed alternative ranked better than the baseline system, with the highest $W-$ score for the *Screen fill with "endless" scrolling* display technique. This is likely because of the baseline system waiting for all results to be displayed before displaying the results, taking the longest time before any results are shown. Of all display techniques, *Visual indication of more relevant results* scored the lowest $W-$ value. This is likely because even though the display technique provided search results quickly, the notification made some participants feel slowed down.

The last question, *i would use this search engine more often* had only one significant result: *Screen fill with "endless" scrolling*. This alternative scored better than the baseline system, most likely because it visually operates much like regular display techniques (i.e., Google, Bing, Yahoo), which participants were used to using. Even though not significant, the same is seen for *Directly display result sets when retrieved*, which also operates in a similar manner as regular search engines.

When asked, participants indicated they would not be willing to wait for the baseline system to complete operation if no progress indicator would be displayed. The baseline waits for about eight seconds before displaying search results, which is likely too long to operate without visual indication of progress.

Participants generally noticed the elements moving around for the *Re-order displayed search results when results are retrieved* display technique. This indicates that the animations last long enough to be noticed. It does, however, not indicate whether the animation is fast enough to provide pleasant operation or

whether the animations should be faster in order for the display technique to be more usable. This is something that can be further explored in future work (section 7.2).

Most participants took notice of the visual indication of more results for the *Visual indication of more relevant results* display technique. Though not everyone answered positively, most participants found the notification useful once they noticed the message. This display technique might have been hindered by the red background color of the notifications, a color which is usually associated with errors or warning messages. Further research could show if different colors could provide better usability scores for this display technique.

When looking to the time to last click, which is an indication of how quickly the participant found a result answering their assignment, most alternatives do not provide significant differences. The only alternative with significant results is *Visual indication of more relevant results*, which performed worse than the baseline system. This is probable due to the participants first exhausting every other result before resorting to clicking the notification of more results. Once they did click the notification, they usually quickly found an answer to the assignment.

The positions of clicked results vary only slightly for each display technique, with the bulk of clicks at the top of the result page. The only exception is with the *Directly display result sets when retrieved* display technique. This can be explained by it being the only display technique that does not re-order its received results in some manner. When results come in they are simply appended to the results that are already being displayed. In the simulated worst-case scenario, this means that the least relevant results appear at the top of the result page, only then followed by the most relevant results.

A domain bias by the participants was observed as they evaluated each search result. Participants tended to prefer links to familiar websites, even when these results were not displayed at the top of the results. For future work, it may be fruitful to not display links to search results and to strip the result titles from any clues as to the domain of the search result.

Finally, *Screen fill with "endless" scrolling* is elected to be the best alternative for the baseline system, due to the fact that in two of the three metrics, it scored better than the other alternatives. However, it should be noted that the definition of the best alternative might vary when emphasis is placed on different criteria.

Chapter 6

Discussion

During user testing a number of observations were made of things that seemed to affect the tests or their outcome. This chapter will shortly discuss these influences and proposes methods to prevent or reduce these influences in the future. It will start with a discussion on the influence of the demo, which test participants used to get familiar with the testing process. Then is discussed how the color of the notifications for the *Visual indication of more relevant results* display strategy might have influenced participant's reactions. Finally is a section that touches on the worst-case nature of the test, and how real life will likely be more average cased.

6.1 Influence of demo on test

Giving test users a short demo before starting the test, may have influenced their responses. Even though the demo setup did use a different display strategy for each demo, the test participants could sometimes see in advance how a display strategy would function. An example is the *Visual indication of more relevant results* display strategy that shows a notification that more results were found. This notification then had to be clicked in order to see the extra search results. Test participants that had seen this notification during the demo knew they could click the notification and more results would show up. Subjects that had not yet seen this display strategy seemed to hesitate longer before clicking.

This problem may be solved in future work by implementing a search engine specifically for demo purposes, with animations and notifications that test participants may be pointed to when preparing them for having to answer a question of how they think the search engine works.

6.2 Color of notification

Performance of the *Visual indication of more relevant results* display strategy may have been hampered by the fact that the notification had a red background color. Originally this was thought to grab the user's attention, but during testing it appeared to do the exact opposite. Subjects simply did not read the text of what they perceived to be an error message, or if reading the text some

participants actually requested permission to click, even though the notification (figure 6.1) stated it had to be clicked in order for more results to be displayed.

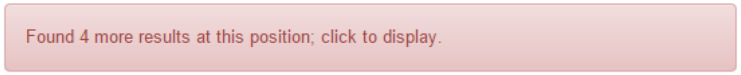


Figure 6.1: Notification of extra search results

In most cases participants first scanned the entire result set, even clicking some non-relevant results and only after exhausting every available option, they clicked the notification that led them to finding an answer to their search task. This could have had any number of reasons. For example, red is usually associated with errors or warning messages. Users tend not to read these. It may also have been the case that the users were suspicious of being tested, not wanting to do anything that would later be revealed to be inappropriate. Users who had seen this display strategy in the demo (see section 6.1) did not exhibit this behaviour and clicked the notification more quickly.

Perhaps this conclusion can lead to more research where differing colors of the notification can be tested. Further variations should be tested to be fully able to dismiss this display strategy. Variations could be different colored background (e.g., green or blue), where it should be noted that certain colored notifications will likely not be read. Examples are red, which often denotes warnings or errors, and green, which often denotes success messages. Use of eye tracking may help determining when a user does or doesn't see such a notification.

6.3 Worst case scenario

The simulated display strategies all performed under the assumption of a worst-case scenario (section 3.6). In a worst-case scenario, resources produce results in reverse order of relevance. That is, the resource to respond first produces the least relevant results whereas the slowest resource will produce the most relevant results. In a best-case scenario, the proposed systems will perform similar, if not exactly, to the baseline system which waits for all resources to respond before displaying the combined results. Though required for testing the effectiveness of the proposed display strategies, this worst-case assumption is not entirely realistic. In reality, a more average-case performance is to be expected. This will mean that sometimes the search engine will perform like in a best-case scenario, sometimes like a worst-case scenario and most times somewhere in between. In order to best test user satisfaction for the most promising proposed alternatives, it is suggested to test each alternative in a real-life situation, serving real responses to real queries. In some cases this will mean being unable to observe improvements, but in the long run, it will likely show which alternative provides the best user satisfaction in the average case situation.

The worst-case nature of this research will also cast some shadow on the user responses, as they are colored by this scenario. Users will tend to be used to a more average-case scenario or, in the case of well known display strategies (e.g.,

Google, Bing, Yahoo) to a best-case scenario where the search engine always responds in a very short time.

Worst-case scenarios could in their entirely be avoided when only relevant resources are queried. This will likely eliminate non-relevant results end likewise eliminate the need for results to be re-arranged. After all, no non-relevant results are displayed and one could argue that any relevancy score of the displayed items is close enough to not warrant further ordering as any item worth displaying is most likely already visible and users tend to scan at least the first few items [9, 10, 11] (section 2.1).

Chapter 7

Future work

This chapter presents a number of methods to improve the research done for this thesis. It starts by proposing an experiment set in real-life, rather than in a simulated environment. Then a method is proposed to test small variations on (some of) the proposed display strategies. An example might be the background color of notifications or the way displaying new results are animated. Finally, it proposes the use of eye tracking equipment to be able to more precisely be able to see what the user sees, without the need to rely on their spoken recollection of events.

7.1 Real-life experiment

One area in which future work can be conducted is bringing (a subset of) the display strategies into real life. Searches would then be performed using real-life resources, rather than a predefined database of search results (section 3.8). This will allow users to define their own queries and use their own wording in order to answer questions they themselves have. It also shifts the focus from a worst-case situation to a more average situation. This will either require an algorithm for determining the best resources to query, or a manageable set of resources, for which it can be expected they can be queried in acceptable time.

Testing systems in this manner will probably require a different approach on which display strategy to test. Rotating the selected engine for each query defeats the purpose by introducing a variable in the sense that the user will never be able to, in advance, expect to know how the search engine works. It is probably better to allow the user to either select which display strategy to use, or by switching the display strategies only after some time. It also is not feasible to ask the user about their experience after each query. They will either grow tired of giving the same answer over and over, or skip the questionnaire entirely. This approach will likely only work by recording metrics such as time to last click (section 5.3.1), time spent on the search result page and the number of repeat clicks.

7.2 Variations on proposed display strategies

One other thing that can be tested is how well the proposed display strategies perform when variations are applied. This practice is usually called A/B testing [28, 20]. It works best when a search engine has a large user base and is used to perform many searches each day. Small variables, such as color of clickable links or line spacing, are being changed from time to time, and metrics are recorded in order to determine differences in behaviour and inferred user satisfaction.

The *Re-order displayed search results when results are retrieved* display strategy can be varied on in the way items are displayed. User satisfaction can change when the length of animations are varied, and when the animation used is changed. Perhaps items should move from the bottom all the way up to their intended location, or maybe animations should last longer.

The *Visual indication of more relevant results* display strategy currently has a notification with a red background color. This color is usually associated to danger, warning messages or error messages. This may explain why a large percentage of the test participants did not notice the notification. It may be that they did see a red color but chose to ignore it, as users often do when error messages are displayed. Changing the color to green may also not improve user satisfaction as this color is usually associated to success messages. Users tend to ignore these messages as they expect websites to function properly. The author suggest a blue color, and to try and use an icon indicating the notification can be *expanded* to reveal additional results.

Lastly, the *Screen fill with "endless" scrolling* and *Directly display result sets when retrieved* display strategies may be extended to wait a short time before displaying initial results. During this time, any results that are received may be ordered from relevant to non-relevant in order to prevent that the least relevant results immediately reach the top of the result page. How long users accept waiting for initial results may also be an aspect of this future research.

7.3 Eye tracking

Finally, future work can consist of applying eye tracking to the proposed display strategies. Eye tracking can provide insight into whether test participants actually see visual cues such as notices or effects. It can also provide insight into which color works best for notifications and what the user uses do determine whether or not to click a link. It can also be used to determine if domain bias (section 5.4) really has an effect on which link users click when operating an experimental search engine. It may be that they want to cling onto the one thing familiar to them (i.e., having search results displayed using the style of a familiar search engine), or that they do not notice it at all, as they are preoccupied trying to see how the search engine in front of them works.

The advantage of using eye tracking for this kind of research is that observers can precisely see what the user sees and where the user looks, rather than having to ask the user afterwards, when they may feel pressured to give the *right answer*, even when there is none. The disadvantage of eye tracking is that equipment is still rather expensive and not available for every researcher.

An alternative to eye tracking may be mouse tracking [29]. This method blurs the entire web page, except for a small area around the mouse. This forces

test participants to point their mouse at the regions they want to examine or read. The advantage of this method is that every computer is equipped with a mouse, making this a cheap option. The downside is that not every user may move their mouse as quickly as their eyes shift focus. This may mean that an animation may be finished by the time their mouse reaches the area of effect. Another downside is that the provided software currently only works when providing text to a service, and the software can not be dropped in place on, for example, a search engine. Because of this, the software should likely be re-engineered, which costs time.

Chapter 8

Conclusion

In this thesis the topic of search result presentation in distributed search engines was discussed. The research question was *on the subject of distributed search (e.g., peer-to-peer or federated search), do methods exist to present intermediate search results that allow early access to relevant results?* This question was divided into:

- Whether there is a method of displaying results that provides greater user satisfaction than waiting for all results to be retrieved;
- Whether the user prefers more rapidly displayed search results, even if it means presenting a number of less relevant results first;
- Whether the proposed strategies improve relevant result distribution.

In order to answer this research question, a number of display strategies were implemented and tested against a baseline system. These strategies are: (*i*) *Visual indication of more relevant results*, which notifies users of new search results at the position where these new results will be displayed. These results are only displayed after clicking the notification; (*ii*) *Re-order displayed search results when results are retrieved*, which actively reorders results when new results are received; (*iii*) *Directly display result sets when retrieved*, which appends newly received results onto the list of already rendered results; (*iv*) *Screen fill with "endless" scrolling*, which renders results until the screen is filled whilst continuing to receive results. When the user starts scrolling, just enough newly received results are appended to allow the user to keep scrolling and to allow the search engine to receive more results. The performance of these techniques are compared to that of the baseline: *Wait for all results to be retrieved*, which waits for all resources to produce their individual results, then combines these results and orders them by relevance and displays them all at once.

Implementing each display technique in a real world situation would result in a system that is hard to test due to real life variables such as network delay and resource delay. This would have introduced additional variables, making it harder to test. To test whether the proposed alternatives improve upon the baseline system, a simulation of a distributed search engine was implemented, operating in a worst-case scenario. Testing the alternatives in a best-case scenario would obfuscate any improvements by the alternatives, as they would essentially operate the same as the baseline system.

Answering the first question, *Whether there is a method of displaying results that provides greater user satisfaction than waiting for all results to be retrieved*, is *yes*, there are display techniques that perform better in this regard than the baseline system. Although it is a question that can not be easily answered — there are simply multiple definitions of *better* — *Screen fill with "endless" scrolling* has the best results overall. It provided the best user satisfaction when the user was asked how fast they felt the display technique to be and whether they would use the system more often. When looking at the metric of *time to last click*, it also performed quite good, although the results were not significant. The positions of clicked links were slightly more distributed throughout the results, but this came with the benefit of rapidly displayed results and a display technique that feels very similar to regular search engines (e.g., Google, Bing, Yahoo).

The second question, *Whether the user prefers more rapidly displayed search results, even if it means presenting a number of less relevant results first* is harder to infer. When asked, participants indicated they felt every display techniques to be fast more often than the baseline system. When being asked whether they would use the display technique more often, the only significant answer was for the *Screen fill with "endless" scrolling* technique, with a favorable score. From this, it is inferred that users indeed prefer more rapidly displayed search results, even if it means having to face a number of less relevant results first.

Lastly, we look at the distribution of relevant results. When considering the inner workings of each proposed technique, no technique can perform better than the baseline system as it is able to order all results by relevance before displaying. The *Screen fill with "endless" scrolling* and *Visual indication of more relevant results* techniques order their results similarly, though each uses a distinct method of informing the user of new results. *Visual indication of more relevant results* performs worst, as it orders the results by time received, not by relevance. *Screen fill with "endless" scrolling* initially also orders its displayed results by time received, then transitioning to displaying all new results by relevance, if possible. To answer this question, another metric needs to be considered. Section 5.3.2 shows that the positions of clicked results vary for each display technique. For the baseline system, the clicks occurred mostly at the top of the result page. Most of the alternatives come close to the baseline system, but here too, no alternative performs better than the baseline system. To this question the answer must therefor be *no*, none of the proposed strategies improve relevant result distribution.

In order to answer the main research question (i.e., *on the subject of distributed search (e.g., peer-to-peer or federated search), do methods exist to present intermediate search results that allow early access to relevant results?*), the above results must be combined such as to create one winning contender. As described in section 5.6, *Screen fill with "endless" scrolling* performs better than the other strategies in two of the three metrics (i.e., test participant questions, combined with click and result position metrics). This earns it the title of best performing display technique, though it should be noted that the definition of *best* may vary when more emphasis is placed on differing criteria. This answers the main research question with: *yes*, although not literally, as the chosen display method does not guarantee relevant results to be displayed first. Rather, it provides a

compromise and tries to deliver relevant results *as soon as possible*.

Appendices

Appendix A

Resources per participant assignment

Participant assignment	Resource
How many people were President of the United States in the 19th century?	Babylon Gigablast Mamma.com Ask Bing LinkedIn Jobs Paypal Starbucks MPRA Volvo Cars
How many seasons of the television show "ER" have aired?	Babylon Bing AOL Ask Yahoo TechRepublic Tweedehands.net Tweepz MPRA University of Twente Publications
When did South Africa gain its independence from the United Kingdom?	Wikipedia Technet Techrepublic The Register USAJobs Volvo Cars

Participant assignment	Resource
In billions of years, how old is the sun?	Mamma.com Yahoo! Image Google Books SpringerLink Picasa MPRA USAJobs University of Twente Publications Volvo cars Yahoo Answers
What is the currency used in Afghanistan?	Encyclopedia Britannica AOL Babylon Mamma.com Yahoo Tweedehands.net UAB Digital Repository MPRA USAJobs University of Twente Publications
When was the album "The wall" by Pink Floyd released?	Ask Babylon Amazon AOL Mamma.com Volvo Cars Wiktionary WordPress Yahoo Answers Yahoo! Image

Bibliography

[1] Milad Shokouhi and Luo Si. Federated search. *Foundations and Trends in Information Retrieval*, 5(1):1–102, 2011.

[2] Adele E Howe and Daniel Dreilinger. Savvysearch: A metasearch engine that learns which search engines to query. *AI Magazine*, 18(2):19, 1997.

[3] Antonio Gulli and Alessio Signorini. Building an open source meta-search engine. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 1004–1005. ACM, 2005.

[4] Jinyang Li, Boon Thau Loo, Joseph M Hellerstein, M Frans Kaashoek, David R Karger, and Robert Morris. On the feasibility of peer-to-peer web indexing and search. In *Peer-to-Peer Systems II*, pages 207–215. Springer, 2003.

[5] Fiona Fui-Hoon Nah. A study on tolerable waiting time: how long are web users willing to wait? *Behaviour & Information Technology*, 23(3):153–163, 2004.

[6] Pratibha A Dabholkar and Xiaojing Sheng. Perceptions of download delays: relation to actual waits, web site abandoning, and stage of delay. *The Service Industries Journal*, 28(10):1415–1429, 2008.

[7] Laura A Granka, Thorsten Joachims, and Geri Gay. Eye-tracking analysis of user behavior in www search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 478–479. ACM, 2004.

[8] Stuart K Card, Peter Pirolli, Mija Van Der Wege, Julie B Morrison, Robert W Reeder, Pamela K Schraedley, and Jenea Boshart. Information scent as a driver of web behavior graphs: results of a protocol analysis method for web usability. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 498–505. ACM, 2001.

[9] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161. ACM, 2005.

[10] Edward Cutrell and Zhiwei Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 407–416. ACM, 2007.

[11] Lori Lorigo, Bing Pan, Helene Hembrooke, Thorsten Joachims, Laura Granka, and Geri Gay. The influence of task and gender on search and evaluation behavior using google. *Information Processing & Management*, 42(4):1123–1131, 2006.

[12] Kerstin Klöckner, Nadine Wirschum, and Anthony Jameson. Depth-and breadth-first processing of search result lists. In *CHI'04 extended abstracts on Human factors in computing systems*, pages 1539–1539. ACM, 2004.

[13] Samuel Ieong, Nina Mishra, Eldar Sadikov, and Li Zhang. Domain bias in web search. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 413–422. ACM, 2012.

[14] Eric Goldman. Search engine bias and the demise of search engine utopianism. *Yale Journal of Law & Technology*, pages 06–08, 2006.

[15] German court orders google to block max mosley sex pictures. *Reuters*. `http://www.reuters.com/article/2014/01/24/us-google-germany-court-idUSBREA0N0Y420140124` (accessed February 26, 2014).

[16] Clive Thompson. Googles china problem (and chinas google problem). *The New York Times*, 23, 2006. `http://www.nytimes.com/2006/04/23/magazine/23google.html` (accessed February 26, 2014).

[17] Lucas D Introna and Helen Nissenbaum. Shaping the web: Why the politics of search engines matters. *The information society*, 16(3):169–185, 2000.

[18] Séamus Byrne. Stop worrying and learn to love the google-bomb. *Fibreculture Journal*, (3), 2004. `http://three.fibreculturejournal.org/fcj-015-stop-worrying-and-learn-to-love-the-google-bomb/` (accessed February 26, 2014).

[19] Tom McNichol. *The New York Times*, 2004. `http://www.nytimes.com/2004/01/22/technology/circuits/22goog.html` (accessed February 26, 2014).

[20] Jake Brutlag. Speed matters for google web search. *Google. June*, 2009.

[21] Andrei Broder. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM, 2002.

[22] Daniel E Rose and Danny Levinson. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, pages 13–19. ACM, 2004.

[23] Krishna Bharat and Andrei Broder. A technique for measuring the relative size and overlap of public web search engines. *Computer Networks and ISDN Systems*, 30(1):379–388, 1998.

[24] Antonio Gulli and Alessio Signorini. The indexable web is more than 11.5 billion pages. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 902–903. ACM, 2005.

[25] Jesse Alpert and Nissan Hajaj. We knew the web was big. *The Official Google Blog*, 21, 2008. `http://googleblog.blogspot.nl/2008/07/we-knew-web-was-big.html` (accessed February 26, 2014).

[26] Dong Nguyen, Thomas Demeester, Dolf Trieschnigg, and Djoerd Hiemstra. Federated search in the wild: the combined power of over a hundred search engines. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1874–1878. ACM, 2012.

[27] Richard Lowry. The wilcoxon signed-rank test. *htt p://faculty. vassar. edu/lowry/ch12a. html*, 2010.

[28] Ron Kohavi, Alex Deng, Roger Longbotham, and Ya Xu. Seven rules of thumb for web site experimenters. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14)*. ACM, 2014.

[29] Dmitry Lagun and Eugene Agichtein. Viewser: Enabling large-scale remote user studies of web search examination and interaction. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 365–374. ACM, 2011.