Analyzing Click-through Data of the UT Search Engine with the aim of Improving Search Engine Rankings

Rob Stortelder
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
rob@stortelder.eu

ABSTRACT
This paper will describe the process of analyzing the ranking of search engines in a federated search engine called Searsia using click-through data. The aim of this analysis is to determine if improvement in search engine rankings in Searsia is possible and how this improvement can be applied. The current ranking is based on sample indexes in combination with response times. This ranking might not return a desired ordering, because this ranking does not involve user or expert feedback. The improved ranking should order search engines with the most relevant search engine first, based on implicit (user) feedback. Click through data was recorded and analyzed with the aim of improving the ordering of search engines in order of relevance to the search query. Data for this research was gathered from the search engine of the University of Twente which is currently used by students, researchers, teachers and other people affiliated with, or interested in the University of Twente. This data showed that the Google search engine almost always showed up high in the rankings and was clicked in more than 80% of the cases. In more than 80% of the time a result at the first position is clicked, making improvements in the search engine limited.

Keywords
Search engine ranking, click data, analysis, federated search engine, web search, implicit feedback

1. INTRODUCTION
Searsia [6] is a search engine, developed at the University of Twente. This search engine is a federated search engine that instead of ranking results, ranks search engines containing results. Only a couple of results with each search engine are shown. Currently, the search engines are displayed in the order of response times in combination with sample indexes instead of relevance to the search query. Analysis of click-through data could supply insight in the possible areas of improvement for the UT search engine. With this insight, possible methods for improvement will be provided.

For this research, click-through data will be recorded and analyzed. Click-through data is data about which links that are clicked in a search engine. The data will be recorded from the UT search engine (currently located at utwente.nl/search). This search engine is an implementation of the Searsia search engine and is developed at the University of Twente. At the point of writing, around 800 queries are executed every day. Researchers, teachers, students and other people affiliated with, or interested in the University of Twente might use this search engine. This search engine is specially configured for the University of Twente, to allow searching multiple resources of the university with one query. The data will be recorded in a way that no explicit feedback from users or manual judgment of experts is required.

Analysis of this data will provide insight in the usage of the UT search engine and the possible areas of improvement. This information can be used by other researchers and developers of the UT search engines to improve the rankings of the search engine.

2. RELATED WORK
T. Joachims describes a method to improve the ranking of search results using click-through data. With machine learning he was able to improve the ranking of search results. Joachims uses a SVM (Support vector machine) to classify results based on click-through data. Click-through data is only valuable if the user deliberately skips results. If only the first result is clicked every time, no improvement is possible. [5]

N. Craswell explains why Click position data is biased. A key source of bias is presentation order, if a relevant document if the most relevant document is located on the tenth page, it might not be found. [3]

Researchers at Microsoft have conducted research to improve ranking for a popular search engine. This research is especially relevant for the amount of data that is recorded. Besides the logging of click-through data, other data about user behavior was logged as well. Examples are 'TimeOnPage' the time the user is on the search page and 'TitleOverlap' are the words shared between the query and the title. The complete list can be found in table 4.1 of this paper [1]

Another Microsoft research project analyses if click through data is adequate for learning search rankings. They describe a method to classify queries using click entropy. Click entropy is based on the percentage of clicks on a document among all clicks in the query. A query with small click entropy generally indicates that it is a navigational query. Users search for ‘Google’ for instance and generally click the first result. Larger click entropy indicates that more documents are clicked with a single query, this is generally related to information search. Several pages
are visited to acquire multiple pieces of information. [4]

2.1 Problem statement
During this study, the search engine of the University of Twente uses around 25 resources, including the University of Twente phone book, University of Twente publications and Google results. When searching for a person one would like the entry in the phone book or publications of this person first. The current result for a person search shows the Google results at a higher rank. By recording click-through data we can analyze if people generally find what they are looking for and if search engine contains this result. With machine learning it might be possible to improve search query results.

2.1.1 Non-obtrusive data collection
Click-through data will be collected in a non obtrusive way. Users will not be bothered with questions while searching for information. This might introduce privacy issues for search engine users because we keep track of sessions and record clicking data for these sessions. With this information it might be possible to link these sessions to real people. The ethics committee of the University of Twente will be asked for advice on this way of collecting data, ensuring the privacy of search engine users. A different option in collecting this data without a possible invasion of privacy is to ask explicit feedback from users, but experience shows that users are rarely willing to give feedback [5]. Click-through data is currently not recorded in Searsia, a system to record such data needs to be written for this research.

2.1.2 Bias of click-through data
Click-through data is biased [3, 5] which is mainly due to the presentation order. The ordering of the results influences on which results the user clicks. If the third result is clicked, the user deliberately decided not to click on the first two results thus a machine learning system can learn from these deliberately skipped results. A result on the second page might be the best result which might not be seen by the user of the search engine because it is too low in the ranking order. In this case, a machine learning system based on click-through data will not improve this result because no click-through learning data is available for this result.

Another problem with bias of click-through data is that if the first result is clicked, the machine learning system has nothing to improve meaning only results that are deliberately not clicked can be improved. Results that are deliberately not clicked can be interleaved with results that are deliberately clicked. With this system, the better results will occur higher up in the ranking list.

The Searsia search engine shows blocks of results for each search engine. In comparison to conventional search engines, users might scroll to the results of the search engine that seems to have the most relevance to them. If this is the case, this could massively help a machine learning algorithm due to the high amount of deliberately skipped results.

2.2 Research questions
The main research question is defined as follows: How can search engine rankings in the UT Search engine be improved by analyzing click-through data?

To answer this question each subquestion will be answered.

Q1: How can we record click-through data in the UT search engine in a non obtrusive way?

Q2: What information does this data convey about possible improvements for the UT search engine?

Q3: How can this data be used to improve search engine rankings in the UT search engine?

Q4: What other research uses are there for this data set?

3. METHOD OF RESEARCH

3.1 Non-obtrusive data collection
To answer the first question, a method is needed to record click through data unobtrusively. A method described by T. Joachims [5] appears quite similar to the system Google uses for recording click-through data. With this method, each click is sent to a proxy server which records the click event. This proxy forwards the user to the originally clicked page. The fact that Google uses this system can be seen by inspecting the HTML of a Google search results page. The URL directs to Google first, which subsequently forwards you to the original requested page. The difference with the method of T. Joachims is that we do not need information about individual results, but only about which search engine contains this result. Data about which link is clicked inside a search engine will be recorded, but is not directly needed for ordering search engines. A proxy server will be build and implemented to answer this question. This proxy will store at least three pieces of information. The first piece is the query words, the second piece is the order in which the search engines are ordered and the last piece is the list of search engines where the user clicked on a result. This proxy server will make use of sessions to determine if the same user returned to the result page and clicks another result. As described in the problem statement, this will introduce privacy issues. The main issue is that sessions can possibly be linked to users which subsequently links users to search queries. These issues are discussed with the ethical safeguards of the University of Twente. They stated that click-through data may be recorded, but storage of ip-addresses was not allowed. An anonymous session identifier is used to link multiple sub-sequential searches together, this is allowed by the ethics committee. A non-disclosure agreement is signed with the university of Twente for this data. A data management plan is written to state where the data is stored, backed up and how it is managed. This data management plan is handed to the ethics committee.

3.2 Measuring improvement
We consider a ranking improved if the search engine containing the result is on average higher than without machine learning applied. The method described by T. Joachims in section 5.1 [5] can possibly be applied to our result set. His method merges the results of two ranking strategies alternately. In our case one ranking with machine learning applied and one without machine learning applied. After merging the results of both strategies, the amount of clicks on each strategy is measured. Duplicates are merged to one result in the combined list. The amount of search engines in the UT search engine is quite small (around 25 search engines), result duplication can possibly arise at a high rate if we merge results of multiple strategies. It might be possible that this method does not yield meaningful results due to the high amount of duplicate (thus merged) results. A click on a merged result counts for both strategies. To determine if this is the case for the UT search engine, this needs to be verified by trying this method in practice. Another option to determine if search engine rankings are improved with machine learning is to use one of the two ranking strategies (one with machine
learning and one without) at random with each search query. Measure for both strategies the average rank of the clicked result. The one with the lowest clicked rank on average is the better performing algorithm. The disadvantage is that one person only sees one strategy, and might be searching for something totally different in contrast to someone else. Strategies are not directly compared with the same queries or users.

3.3 Improvement methods

Several papers describe a method using the Kendall rank correlation coefficient [4, 5]. A ranking support vector machine (SVM) can be used to classify documents based on click-through data. The SVM algorithm uses pair-wise ranking to sort documents in relevance to the query. In our case, we can use SVM to sort the search engines in relevance to the query. Due to time restrictions for this research, it was not possible to fully implement a machine learning mechanism for Searsia. The main aim is to provide data analysis and methods that can be used for machine learning, the actual implementation of a machine learning system will be follow-up research.

4. DATA COLLECTION

With each click on a link a query is executed storing the click-through data. The non-obtrusive data collection method that is implemented is similar to the data collection method described in section 3.1. The only difference is that an AJAX call is executed before the user clicks the link. The user is not actually redirected through a click proxy. This method is chosen due to it’s robustness, failing AJAX calls are not noticed by the user of the search engine. The kind of data that is collected can be found in table 1. In a higher percentage of cases in contrast to most other search engines, this shows opportunities for improvement. Strategies are not directly compared with the same queries or users.

5. RESULTS

5.1 Analysis of collected data

5.1.1 Suboptimal rankings

5.1.2 Average rank of clicked search engines

We call a ranking suboptimal if a link at a lower position in the ordering is clicked. This means the rank of the clicked document is either larger than one or is minus one (minus one means that a result in the more results section is clicked). This clicked result could theoretically be improved by moving this clicked result one or more positions higher in the result list. Orderings can only be improved if a result later in the ordering is clicked. If the first result is clicked no improvement is possible, the desired result is already at the best possible position. Results located after the clicked result are quite possibly not seen by the user of the search engine thus nothing can be improved about these results. The ordering of the results influences on which results the user of the search engine clicks. This principle is described in section 2.1.2 and is called click bias.

5.1.3 Average rank of clicked search engines

The first step in analyzing click through data is to determine at what average rank the search engines are clicked. The bar chart in figure 1 shows the average rank of clicked search engines. Note that search engines with less than ten clicks are not displayed in this chart. This clearly shows the ‘googlecustom’ and the ‘aanbevolen’ search engines at higher ranks. (Aanbevolen is Dutch for recommended, some search engines have Dutch names in the UT search engine). Although the recommended result is also just a resource configured in the UT search engine, it is positive to see the recommended result at a relatively high position on average. If a recommended result would turn up low in the rankings, it would not be a recommended result. The Google search engine has an average ranking close to 1. This means that the Google search engine is almost always shown on the top position. ITC and more results are almost always clicked at lower positions in the list. For ‘moreresults’ this is pretty logical, more results are located at the bottom of the page. More interesting is that the ITC search engine is located low on average, but in figure 3 (This figure is further explained in section 5.1.5) illustrated that the ITC search engine is clicked in a higher percentage of cases in contrast to most other search engines, this shows opportunities for improvement.
Note that the ITC search engine is from the newest faculty of the University of Twente. This search engine is also a Google custom implementation, configured specifically for this new faculty. This faculty mainly consists of international students. Language specific research of the UT search engine is one of the further research topics possible with this data.

5.1.4 Percentage of clicks on first result

The second step is to analyze what percentage of clicks is on the first result. That percentage of clicks cannot be improved with click-through data. Analysis of the recorded data shows that 84.4% of the results that are clicked is on the first result. This means that theoretically 15.6% of the result orderings can be used to learn a machine learning algorithm using click-through data. This resulted in figure 2. This figure shows the average ranks, just like figure 1, but all the clicks on the first ranks are removed. This shows that the Google custom search engine is always located at the second rank if it is not displayed at the first rank. Recommended, Osiris and the phonebook are the next on average lowest ranked and clicked search engines. As expected from the data in figure 1 the more results and ITC search engines are ranked at lower positions on average. The highest values in figure 2 are the search engines that show the most opportunities for improvement.

5.1.5 Clicked search engines

The third step is to analyze the clicked search engines if the engine is not located at the first position. The pie chart in Figure 3 shows the search engines that are clicked if the click rank is larger than one. Every search engine that is
clicked less than 5% is it the other part. This part contains 18 search engines, on average 1.8% per search engine. In most cases, the Google search engine is clicked. The next highest case is the UT news search engine, this is mainly due to the wide range of topics that the UT news search engine covers. This can be seen by looking through the query words associated with the UT news search engine.

The type of url that is clicked also conveys important information about the usage of the UT search engine. This information is important because the 'html_resource_header' leads to the full result list of the search engine. If the 'html_resource_header' is clicked in most of the cases, it means that users did not directly find the result they were looking for in the previewed results. Figure 4 shows a pie chart of the percentage of times a kind of url is clicked. This pie chart clearly shows that an url of the kind 'sub-result_web_full' is clicked in most of the cases. This means that the users that click on links in the UT search engine mostly click on an individual result and do not request the full result list. This is an indication that the preview results are sufficient in most search cases.

5.1.6 Skipped search engines

The last step is to analyze the skipped search engines, these engines can be moved to lower positions in some cases. The pie chart in Figure 5 shows the percentage of times a search engine is skipped if a result at a later position in the result list is clicked. All the skipped search engines are counted in this chart. For example, if the result at the third rank is clicked, both the search engine at the first and second position are counted as skipped. In this case we see that the Google custom search engine is skipped the most times in comparison to the other search engines. This is due to the Google search engine always being on top positions in the current ranking mechanism. If a result is skipped, it is most likely to be the Google search engine due to its high average rankings. The second most skipped result is the recommended result (aanbevolen in dutch). This result should be high in the rankings because it is the recommended result, that this result shows high in the skipped results is only a logical side effect. The aim of this research is not to determine if the recommend results are useful for the users of the search engine, but follow up research in this topic might be useful to reduce the amount of times the recommend result is skipped.

5.1.7 Conclusion of data analysis

As indicated by all the charts, the 'goolecustom' search engine is a popular search engine. The 'goolecustom' search engine is a custom implementation of the Google search engine which specifically targets resources of the University of Twente. This custom Google search engine also indexes all the University of Twente resources, just like Searsia does. The results of all the other search engines in Searsia are also indexed by the Google search engine, meaning that a result that is clicked in the Google search engine might also be available by another resource in Searsia. The Google search engine is always located in the top results because this search engine is very versatile. To illustrate, a result in the Google custom search engine is clicked in 83.24% of the time in the data set. The next largest clicked is the recommended result with 3.14% of the time.

Analysis of the types of clicked links shows that most users click on a previewed result directly instead of clicking the links to all the results of the search engines. This shows that only showing a couple previews is enough information for most users in most cases.

The last step is to analyze the skipped search engines, these engines can be moved to lower positions in some cases. The pie chart in Figure 5 shows the percentage of times a search engine is skipped if a result at a later position in the result list is clicked. All the skipped search engines are counted in this chart. For example, if the result at the third rank is clicked, both the search engine at the first and second position are counted as skipped. In this case we see that the Google custom search engine is skipped the most times in comparison to the other search engines. This is due to the Google search engine always being on top positions in the current ranking mechanism. If a result is skipped, it is most likely to be the Google search engine due to its high average rankings. The second most skipped result is the recommended result (aanbevolen in dutch). This result should be high in the rankings because it is the recommended result, that this result shows high in the skipped results is only a logical side effect. The aim of this research is not to determine if the recommend results are useful for the users of the search engine, but follow up research in this topic might be useful to reduce the amount of times the recommend result is skipped.

5.2 Analyzing the Google search engine

The analysis of the data shows that improvement in the Searsia search engine is possible, but not to a very large extent. The Google search engine is by far the most popular search engine which is clicked in 83.24% of the time. 84.4% of the clicks are on the first result. If a click is not on the first result, the ITC search engine is clicked in 7.5% of the time, but is almost always located at a low position in the result list. Improvement for this search engine is certainly possible but not to a large extent, the Google search engine outruns all the other search engines by far.

Figure 6. Percentage of times a path of an url at the utwente.nl site is clicked in the custom Google search engine. (everything below 5% is in the other section, the other part contains 97 sub domains)
are clicked. The pie chart in figure 6 shows the first part of the path and in which percentage of times this path is clicked. Paths clicked in the English or German parts of the site are indicated with /en/* and /de/* . The other part in the chart consists of 97 different paths, on average 0.6% per path.

The chart in figure 6 clearly shows why the Google search engine is popular. Information contained by the Google search engine is simply not available in other configured search engines. The information in the Google search engine is the only search engine that completely spans the whole University of Twente site. No other search engine is aimed at resources in the 'www.utwente.nl/*' domain. This directly explains why the ITC search engine is relatively popular as described in 5.1, this searches the new ITC faculty of the University no other search engines searches faculties of the University of Twente. This is the only faculty with a dedicated search engine that is configured as a resource in the UT search engine. If there are more faculty specific search engines available, it would be recommend to configure these in the UT search engine. It will quite possibly turn out that these search engines are also relatively popular.

5.3 Improving the UT search engine

Although improvement in the UT search engine is limited (section 5.1.7), it is possible to improve the UT search engine rankings to some extent. Due to time limitations of this study, it will not be possible to actually improve search engine rankings.

Joachims describes an Support Vector Machine (SVM) algorithm for learning ranking functions [5]. This SVM algorithm has shown to learn a highly effective retrieval function for the ordering of search results. This retrieval function combines current ranking functions to create a highly effective retrieval function.

To determine if his method can be applied to the UTWente search engine, the current retrieval functions need to be analyzed. Analysis of the source code of Searsia shows that two methods are currently used for ranking documents. The first method is a Lucene score that is calculated for every result. This is software created by the apache foundation for calculating scores based on full text search. [2]. The second method is a prior that is manually set for each search engine. This prior is used to determine the relevance of a search engine, even before the search query is actually analyzed. This prior in combination with a bias creates a value indicating the general relevance of the search engine. The current value of the bias is set to 1, meaning it does not influence the prior values that are manually set. These two methods are combined in a single value indicating the relevance of the search engine to the search query.

It is certainly possible to apply the SVM algorithm described by Joachims, the only data that is not available for the learning mechanism is the Lucene score itself. This way, the SVM can use the Lucene function and the prior data to learn between both ranking functions with each logged click. For better machine learning options, it might be helpful to provide tf.idf or bm25 scores. These are numerical statistics for word relevance in a document. With multiple scores provided, the SVM can learn between multiple ranking functions, in stead of only the Lucene and prior values. Before implementing a machine learning mechanism based on click-through data, the Lucene scores (and possible scores from other ranking mechanisms) need to be stored in the data set. With these values available the SVM machine learning algorithm can be implemented in the UT search engine.

6. CONCLUSION

The University of Twente recently switched to Searsia as the primary search engine of the University instead of the custom Google search engine that was used before Searsia. The Google search engine was added as a resource to Searsia and is clicked in 83.42% of the cases in the data set. The Google search engine is always displayed at the first or second position in the rankings. This could possibly be an indication of click bias because users of a search engine tend to click the top results.

There are areas of improvement for the UT search engine but these areas are limited due to the high amount of clicks on the Google custom search engine. Analysis of the Google search engine showed that the Google custom search engine is the only search engine that searched inside the www.utwente.nl/* domain. The ITC search engine is relatively popular in contrast to other search engines but is ranked low on average. Ranking improvements for this search engine is certainly possible. It could also be helpful to configure more faculty specific search engines if these are available, just like the ITC search engine.

Analysis of this data also showed that users mostly click preview links directly leading to results in stead of clicking the search engine to show the full result list. In most cases only the top 3 results are shown as previews, this data shows that in most cases these results are enough for users of the search engine.

The current ranking methods of the UT Search engine could be improved with tf.idf or bm25 scores for example. A machine learning system using click-through data could use these values to learn between ranking functions. It is certainly possible to implement the SVM algorithm as described by Joachims [5], only the current ranking scores need to be logged in order to apply a learning algorithm. This SVM might perform better with multiple ranking functions available. At this point, only two ranking functions are combined. One of these ranking functions uses static variables to indicate the general relevance of the engine to any search query.

This is the first research in click data of the UT search engine. The click logging data mechanism that is created for this research will continue to expand the click logging database. This opens a lot of opportunities for follow up research in this search engine. Not only research in the field of improvement of the search engine rankings is possible, also research in the users and usage of the search engine are possible. Further research topics are described in section 7.

7. FURTHER RESEARCH

This data set can be used for further research purposes concerning the UT search engine. Some topic require some additional data but others have all the data already available in this data set.

7.1 Internal and external use

Due to privacy concerns it was not allowed to store ip-addresses in combination with search data. Although ip-addresses may not be stored, it is possible to store if the request was made from the campus (130.89.*) or if the request was made from outside the campus. This data in combination with the data already available could contain information about the users of the search engine. For
machine learning it was not required to store where the request was made from the campus or not thus for this reason the data is currently not stored. This information could be used to improve the search engine to more specific needs of different user groups.

7.2 Research current ranking mechanisms
This research mainly focused on click-through data in combination with search engine rankings. Research in the improvement of the current ranking mechanisms and how they perform can be desired to check if improvement in these fields are possible before implementing the learning mechanism with click through data. Data from this research can be used in how the current ranking results in practice. This research might turn out that the current rankings are almost always the same because they are heavily reliant on the prior, which is a static value indicating the possible relevance to any query.

7.3 Language use
Research in language usage can be used to improve search engine rankings. Research in what type of languages are used in queries can be used for this research. With this data it might be possible to determine if result are better found in one language in contrast another language. Currently no language is specified for the search engine and the results contain both dutch and English documents. Automatic translation might help to improve search engine usability.

7.4 Navigational or informational usage
For this research only click data is stored. With small adjustments it is possible to store all the queries that are entered. This way, information can be retrieved about how much queries are needed before the desired result is found. Data about how much clicks that are needed before the desired result can already be derived from this data set. With this information, assumptions can be made of the intentions of the users of the search engine. As described by [4] in section 5.3, search queries can be navigational or informational. Navigational means that in the case of the UT search engine the query ‘blackboard’ is typed in to find the blackboard site. Informational is looking for specific information. The informational queries have generally more clicked links (click entropy), because more clicks are generally needed to find all the information needed. This gives insight in the kind of users that use the search engine. This data can also be used to improve the search engine rankings.

7.5 Compare Searsia with Google
The most clicked search engine in Searsia is the Google search engine. The Searsia search engine is a replacement of the Google search engine, but still uses the Google search engine as a resource. This raises the question that it might be better to reinstate the Google search engine as the default search engine for the University of Twente. Research in user experiences in both search engines could determine if the Searsia search engine is experienced as more usable in comparison to the Google search engine. This result could be used for deciding to choose one of the two search engines.

7.6 Recommend results
Click-through data also contains information about clicks on the recommend result. This improvement can easily be measured by counting the percentage of clicks and skips of the recommend result.

8. REFERENCES