Contextualised Browsing in a Digital Library's Living Lab

Zeljko Carevic
GESIS – Leibniz Institute for the Social Sciences
Cologne, Germany
Zeljko.Carevic@gesis.org

Philipp Mayr
GESIS – Leibniz Institute for the Social Sciences
Cologne, Germany
Philipp.Mayr@gesis.org

Sascha Schüller
GESIS – Leibniz Institute for the Social Sciences
Cologne, Germany
Sascha.Schueller@gesis.org

Norbert Fuhr
University of Duisburg-Essen
Duisburg, Germany
fuhr@uni-duisburg.de

ABSTRACT

Contextualisation has proven to be effective in tailoring search results towards the users' information need. While this is true for a basic query search, the usage of contextual session information during exploratory search especially on the level of browsing has so far been underexposed in research. In this paper, we present two approaches that contextualise browsing on the level of structured metadata in a Digital Library (DL), (1) one variant bases on document similarity and (2) one variant utilises implicit session information, such as queries and different document metadata encountered during the session of a users. We evaluate our approaches in a living lab environment using a DL in the social sciences and compare our contextualisation approaches against a non-contextualised approach. For a period of more than three months we analysed 47,444 unique retrieval sessions that contain search activities on the level of browsing. Our results show that a contextualisation of browsing significantly outperforms our baseline in terms of the position of the first clicked item in the result set. The mean rank of the first clicked document (measured as *mean first relevant - MFR*) was 4.52 using a non-contextualised ranking compared to 3.04 when re-ranking the result lists based on similarity to the previously viewed document. Furthermore, we observed that both contextual approaches show a noticeably higher click-through rate. A contextualisation based on document similarity leads to almost twice as many document views compared to the non-contextualised ranking.

1 INTRODUCTION

Exploratory search in a DL, especially on the level of browsing, is a frequent strategy when looking for related content [8], [6], [22]. Due to structured metadata that annotate the content of scholarly DLs, users are able to explore the content based on shared characteristics like keywords, classifications, or author information. These paths to start exploratory search in a DL are referred to as *Stratagems* [3] or search stratagems. Although stratagems are encouraging the user for further exploration, the system support on this level is rather low. Modern DLs support stratagem search as simple boolean filters that disregard information about the present user, his or her information need, and session activities. One way to enhance browsing on the level of stratagems is to integrate the users' search context and thus, tailor search results based on previous search activities [29]. Instead of filtering the documents based on shared characteristic as it is the current state-of-the-art in many DLs, we propose a contextualised stratagem search that extends the basic filtering by re-ranking the results with respect to the users' search context. We develop two short-term contextualisation approaches on the level of stratagems for DLs by utilising implicit user feedback. The goal of our contextualisation approach is to tailor search results on the level of stratagems towards the users' current search task.

Integrating the users' search context to personalise search results has provided large benefits during information seeking. While this is true for ad hoc search, the usage of contextual session information during exploratory search especially on the level of *browsing* has so far been underexposed in research.

Following [13], we understand browsing as a search activity that leads to a pre-defined group of information items sharing a particular metadata, while searching produces ad hoc collections of information that have not been gathered together before.

In this paper, we present two short-term contextualisation approaches that integrate the users' search context in exploratory search. A) A contextual feature that re-ranks documents based on their similarity to the previously viewed document and B) a contextual ranking feature that re-ranks documents based on the users' search activities (e.g. the previously entered query terms and viewed documents) in the current retrieval session.

In this study, we aim to answer the following research question:

- Can we improve the **effectiveness** of exploratory search on the level of browsing by using contextual ranking features in comparison to a non-contextual ranking feature?

One of the major challenges in the evaluation of contextualisation approaches is that large scale log data is usually hard to obtain. For this reason, we evaluate our approaches using the real life DL's living lab for the social sciences *Sowiport* and compare them against a non-contextualised ranking of exploratory search. Using an A/B/C testing, each user is randomly assigned to one approach for the entire session. To evaluate the effectiveness of each approach we measure: a) the position of the first clicked item which we refer to as "mean first relevant" (MFR) b) the click-through rate and c) the usefulness in terms of implicit relevance feedback.

Between September/12/2017 and December/20/2017 we analysed 607,109 sessions of transaction logs provided by our living lab. Our results show that contextualising search activities on the level of browsing significantly improved the retrieval quality in terms of MFR [12].

In modern DLs the most common feature to narrow down search results based on certain document features are *facets* which are widely implemented nowadays. Facets are described as a "set of meaningful labels organized in such a way as to reflect the concepts relevant to a domain" [14]. Although empirical studies identified various beneficial aspects of faceted browsing [10], they usually operate on the level of simple Boolean retrieval. Furthermore, facets require the user to interact with the result lists and to select each filter criterion one after another. In contrast to facets, our contextualisation approach tailors search results on the level of stratagems based on the users' previous interactions without additional effort by the user. In other words: our approach re-ranks documents that result from stratagem search with respect to the users' search context while the advantages of faceted browsing remain unaffected and can still be utilised by the user.

The paper is structured as follows. In the following Section we present the use case of our living lab study. In Section 2 an overview on related work is provided. Our contextualisation approaches and the baseline are presented in Section 3. In Section 4, we describe the metrics used to evaluate the effectiveness of our approach followed by a presentation of the results in Section 5. In Section 6, we discuss the results. Insights into the strengths and weaknesses of our living lab study are presented in Section 7.

Living Lab Use Case

The idea of a living lab is to understand information seeking behaviour in situ involving and integrating users within the research process and providing a context for testing and evaluating of IR models, methods and systems [2]. In general, a living lab is an environment in which researchers are facilitated to test their approaches in real life applications with real life users interacting with the system. One example of a living lab evaluation campaign was initially provided by the CLEF initiative and is now continued as *TREC - Open Search*[1] .

---

[1]http://trec-open-search.org/

We evaluate our contextual ranking features in *Sowiport*, a Digital Library for the social sciences [16] that is based on VuFind 2 with an Apache Solr 5.3 index. For the present study we utilised Sowiport as a Living Lab environment in which we injected our re-ranking based on contextualisation. Sowiport comprises about 9.7 mio. literature references from 23 different databases covering topics from social and political sciences. On a weekly basis, Sowiport reaches around 20,000 unique users. A document in Sowiport usually covers information about the author(s), keywords, classifications, source information (journal or conference proceedings). All these document properties are implemented as hyperlinks and when employed, a result list of documents that share that particular metadata is generated.
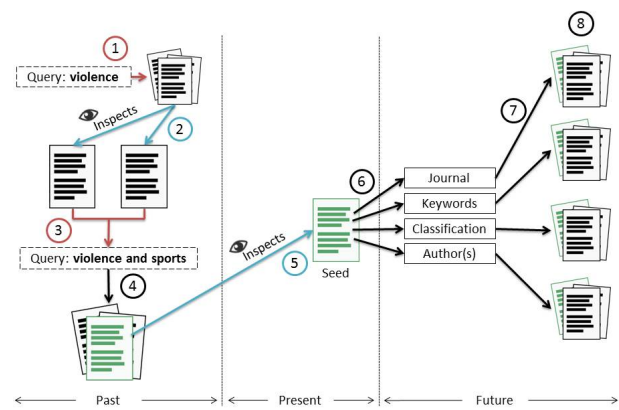
Figure 1: Schematic visualisation of contextualised stratagem searching.

A schematic visualisation of our contextualisation approach can be found in Figure 1. In this example, we have a user who is seeking information on the topic *violence and sports*. After entering a query (1), inspecting two documents (2) and refining the query (3), the user has found a document of interest in a result set (4: highlighted green in the figure) that he inspects in detail (5 and 6). To seek further related content the user could for instance look at the journal the document was published in or select a certain keyword that is contained in the current document. Each of these interactions (7) leads to a new result list containing documents that shares the same attribute with the seed document which is also part of the result list (8). Our approach is to re-rank these result lists based on contextual information about the users search sessions.

A more practical example is displayed in Figure 2a and 2b. In this example a document with the title "*Football in Southeastern Europe: ...*" has been retrieved via the query for "violence and sports". To perform a stratagem search a user could select a keyword, the name of the author, or the journal (*Südosteuropäische Hefte – Southeastern Europe Magazine*). Each of these interactions would lead to a result list of documents containing that particular filter criterion. Such a result list is displayed in Figure 2b which was generated by selecting the keyword *sport* from the seed document. In this example, the

Figure 2: Example of a keyword search in Sowiport.

(a) Example document from Sowiport retrieved via a query search for "violence and sports".

(b) Example result list after a stratagem search for the keyword "sports".

ranking has been contextualised towards the similarity of each document in the result set to the seed document. The displayed result list contains various characteristics worth noticing: One can see that the top ranked document is highly related to football in the former Republic of Yugoslavia which shows a topical relatedness to the seed document. Furthermore, one can see that the result list contains documents from different languages. In this case the top ranked document is in German. A click on a document in the result list leads to a detailed view of that particular document. It is important to notice that we did not modify the interface of the result list but rather only re-rank the results towards their similarity

to the seed document. Thus, the user is still able to use features like facets or to narrow down the search results based on the year of publication. This is displayed in the left box of Figure 2b.

For each document a bookmarking feature, a lookup functionality in Google Books and Google Scholar, and a check for the availability in the local library is placed next to each record as well as a link to the full text of that record, if available. This is displayed in the right box of Figure 2b.

2 RELATED WORK

In the early 2000s, Lagoze et al. [19] and Smeaton and Callan [29] made a first move to put the concept of *context* into the perspective of DL research. They postulate that a DL "should be contextual, expressing the expanding web of inter-relationships and layers of knowledge" [19]. In this sense, they called for *proactive* [29] or *collaborative* [25] or *adaptive* [11] DLs. Since that time, state-of-the-art DLs have not changed much and remain often in the "simple catalog model" [19]. Personalisation and recommender systems have evolved and found their way into modern DLs, but deep contextual system support and user guidance are often still missing.

Exploratory search tasks usually comprise search activities on the level of learning and investigating that go beyond simple look up tasks such as known item search [20]. Due to the complexity of exploratory search tasks, various search activities on the level of moves, tactics, and stratagems are typically involved [6]. To date numerous studies have been conducted that aim to understand users' search behaviour and search activities during exploratory search. Ellis [9] studied the search behaviour of social scientists and identified six generic features: *Starting* (e.g. to identify a paper to start with), *Chaining* (e.g. to follow references in a certain paper), *Browsing* (e.g. to browse all papers by a certain author), *Differentiating* (e.g. to judge a source based on their nature), *Monitoring* (e.g. to subscribe to an alerting service) and *Extracting* (e.g. to identify material in a well-known journal).

Contextualisation (in web search more commonly referred to as personalisation) has drawn a lot of attention in web search, e.g. ([30], [21], [27]) and is distinguished between short-term (the current session) and long-term (over many user sessions) personalisation. While short-term personalisation provides better results for recent search interests, it lacks support of the users' long-term interests. Long-term personalisation on the other hand is suitable to capture the users' long-term interests but may not represent recent shifts in search interests [5]. In this paper, we focus on short-term personalisation which we refer to as contextualisation.

Constructing the user context can be done in various ways like as per explicit (a user explicitly providing feedback information) or implicit feedback which utilises all the information available about the current user, like for instance the viewed/skipped documents or the set of query (re-)formulations. A comparison between implicit and explicit feedback is presented in [34]. In total, 16 participants were asked to solve 16 tasks that were taken from the TREC-10 interactive track. In their experiment two interfaces were developed that connect to the Google search engine: one that uses implicit and one that uses explicit relevance feedback. The results showed no significant difference in terms of search effectiveness between the usage of implicit and explicit relevance feedback with regards to viewed result pages, task completion, and task time.

In the *TREC session track* [18] which ran from 2010 to 2014, the main goal was to consider the session history for (re-)ranking the output for subsequent queries. The TREC session track combines a given corpus, topics, queries and relevance judgments with retrieved results, click data and dwell times from crowd workers conducting searches within a session. Similar is the *INEX Interactive Track* [24] where different tasks of real users are conducted (see an overview of different interactive IR data sets in [4]). In contrast to the TREC evaluation setup, we only consider specific queries in the form of contextualised browsing, and we evaluate our approach in a living lab.

Implicit relevance feedback

In [31] various implicit feedback models are evaluated each aiming to enhance the representation of a user's information need. Each model is constructed by gathering relevance information from the user's exploration of the particular IR system. The different variations are evaluated using searcher simulations. The authors show that each implicit feedback model increased search effectiveness through query expansion. In [26] implicit relevance feedback is incorporated for query expansion and to select new terms that are added to the user's query. The authors conducted five experiments on incorporating behavioural information into the process of relevance feedback with 30 participants and 6 tasks. They observed the overall search behaviour, search effectiveness and the subjects perception of the different systems. The results indicate that search behaviour can be utilised in the process of relevance feedback. In [28] the authors studied the effect of implicit feedback, including query history and click-through history to improve information retrieval performance. They developed four context-sensitive language models that use statistical language models which incorporate context information into a basic retrieval model. Their experiments showed that the retrieval performance could be substantially improved without requiring any user effort. In [27] a long-term personalisation approach is developed that incorporated click behaviour into the document ranking. The approach is evaluated with data from the *Personalized Web Search Challenge* organised by Yandex and Kagell where researchers are able to provide their own solution to a personalisation task on Yandex log data. White et al. [32] developed a task-based personalisation approach that utilises session history from other users with similar tasks referred to as *groupisation*. A contextualisation approach aiming to support browsing in a local web site is presented in [1]. Based on the implicit feedback of groups of users, a document summarisation for browsing is developed. By matching all queries from a group of users a document summarisation was developed that performs on two levels: single document summarisation and a multi document summarisation, i.e. a summary that is generated from a collection of related documents.

Previous Work

The present paper is part of a large scale investigation on the usage of stratagems in exploratory search. In a position paper, we discussed the idea of contextual exploratory search [7]. We presented a concept for bibliometric enhanced stratagem searches that contextualises search activities and integrates further re-ranking features like co-citation analysis and author centrality. In [8] we presented a first approach on gathering a deeper understanding on the usage of stratagems by conducting an online survey with 128 respondents from twelve different fields of research. Our survey showed a general need for a contextual ranking in exploratory search which we tested using a journal run scenario in which the respondents

were asked to arrange the content of a journal run based on two contextual features and four non-contextual features like date or title. The results of the survey showed that the respondents assess the ranking features based on contextualisation noticeably higher than the four non-contextual features. In a recent study [6] on exploratory search in a DL, we showed that the majority of search activities were performed on the level of browsing in comparison to queries and the usage of recommendations. A similar observation can be found in [33] in which the authors showed that less than one third of the interactions belonged to search result pages while the remainder belongs to pages that lie on the hyperlink trail from a search result page.

3 CONTEXTUALISED RE-RANKING APPROACH

We compare three methods for the ranking of exploratory search results. The first approach is the default ranking of Sowiport while the other two approaches are developed to support contextual browsing. We apply contextualised browsing on the level of stratagems and, thus, need a seed document to start with. From a set of documents $D = \{d_1, ..., d_n\}$ the seed document d_s is defined as the document which is currently opened and examined by a user and contains a set of possible browsing activities (*BrowsAct*) which are defined as follows:

$$BrowsAct(d_s) = \{Keywords_{d_s}, Authors_{d_s}, Classifications_{d_s},$$
$$Journal_{d_s}\}$$

where $Keywords_{d_s} = \{k_1, .., k_j\}$ is a set of keywords, $Authors_{d_s} = \{a_1, .., a_k\}$ is a set of authors, $Classifications_{d_s} = \{c_1, .., c_l\}$ is a set of classifications in d_s and $Journal_{d_s}$ is the journal the seed document d_s was published in. For all three approaches we excluded the seed document from the result set ($d \neq d_s$) to prevent a potential click bias towards the previously inspected document[2].

3.1 Default Ranking

In this section we describe the default ranking of Sowiport which we use as baseline. The default ranking is based on SOLR and can be formalised as follows:

$$DR : Q \times D \rightarrow \mathbb{R} \tag{1}$$

A filter query q (selected by the user) is submitted to an expansion function which takes the filter query and expands it with synonyms and corresponding translations (see conceptual model in [23]). If, for instance, a user clicks on the keyword "violence", this keyword is first expanded with synonyms and different translations. Furthermore, the filter type "keyword" is expanded to related metadata fields. The actual ranking is then provided by boosting the results based on the contained metadata field type. The boosting order is predefined by the DL which assumes that different metadata fields are of higher value than others (e.g. keywords > free keywords). The resulting list of ranked documents in this approach has no connection to the previous search activities of the user and is basically a simple Boolean filter which is extended by a boosting on particular metadata fields. We decided to utilise the described

Sowiport default ranking as a baseline although an out-of-the-box VuFind-Solr configuration which performs a simple Boolean filter would decrease the complexity and allow a better reproducibility of our results. The main reason for this decision is that we want to compare our contextualisation approach to a realistic real life DL ranking which also provides us with a strong and established baseline.

A simplified example query for the default ranking is displayed in Listing 1. Again, we utilise our fictitious example of a user who is looking at a document about "Football in Southeastern Europe". In this example, the user selected the keyword "violence". One can see that the baseline approach not only performs a Boolean query on the keyword level but extends the query to related fields like in this case the metadata field "keyword_free" which is an alternative and less formal descriptor for keywords. Furthermore, one can see that the term violence is translated into the German word "Gewalt". The ranking of the retrieved documents is based on TF*IDF whereby the weighting of the fields is taken into account. In our example it can be seen that the metadata field "keyword" is boosted by a higher factor than the "keywords_free" field.

Listing 1: Example query for the default ranking

```
1  q => keyword:((violence OR "Gewalt")^400 OR
2  keyword_free:((violence OR "Gewalt"))^250
```

In the following two sections, we describe our contextualisation approach which is then compared against the baseline.

3.2 Re-Ranking based on Document Similarity

In this approach we perform a re-ranking of the result list based on the similarity of each document in the result set compared to the seed document.

This is described formally as follows:

$$SR : Q \times D \times D_s \rightarrow \mathbb{R} \tag{2}$$

The ranking function SR is an extension to the default ranking DR. The ranking function takes the filter query and the seed document (D_s) as input to produce a ranked list of documents. In a first step, the documents are filtered correspondingly to the default ranking. Instead of only boosting the documents based on field types containing the filter query, the ranking is extended by a similarity score of each document compared to the seed document. To estimate the similarity of each document compared to the seed document we utilise the *MoreLikeThis (MLT)* query parser[3] built in SOLR which is usually employed to provide related documents to a given seed document. An example query for the re-ranking based on document similarity is provided in Listing 2. One can see that the query is based on the default ranking described in Section 3.1 but uses the *MLT* query parser. The seed document is specified by the *DocID* parameter. To compute the similarity of all documents to the seed document we use the keywords, journal information, the abstract (in different languages if available), and the author names of the seed document. These are specified in the *qf* parameter in Listing 2.

---

[2]During a pretest, we observed an extraordinary high click rate on the duplicate entry of the result list.

[3]http://archive.apache.org/dist/lucene/solr/ref-guide/apache-solr-ref-guide-5.3.pdf

Listing 2: Example query for the re-ranking based on document similarity

```
1  q => {!mlt}DocID
2  keyword:((violence OR "Gewalt")^400 OR
3  keyword_free:((violence OR "Gewalt"))^250
4  qf=authors,keywords,journal,abstract
```

3.3 Re-Ranking based on Session Details

In this approach, the re-ranking is performed based on the user's session context which is derived from a set of actions the user has performed during the current search session.

This is described formally as follows:

$$CR : Q \times D \times U_c \to \mathbb{R} \qquad (3)$$

where

$$U_c = \{Keywords(U), Categories(U), Queries(U)\}$$

The ranking function CR is again an extension to the default ranking DR. The function takes the filter query and the user's search context (U_c) as input to produce a ranked list of documents. Again, the first step is to filter all documents corresponding to the default ranking. The re-ranking based on the user's search activities is then provided by boosting the filtered documents using the user's session context (U_c) which consists of three features: keywords, categories and the queries a user U has submitted.

We create the keyword and category features from two sources: a) each keyword/category is considered that was contained in the list of documents the user has seen during the session, b) each keyword/category is considered that was contained in documents within a result set during the session. We then sort the list of keywords and categories based on the number of their occurrences within the session. If a certain keyword has appeared five times in the list of viewed documents and four times in documents within the result list, the score for that particular keyword is nine. To reduce noise, we limit the number of keywords and categories to the top three for each feature and normalise the count to range between zero and one. Regarding the queries, we did not create a threshold as we assume that all queries are equally important to describe the user's information need and do not contain any noise. An example session context could be described as shown in Listing 3:

Listing 3: Example session context

```
1  {"query":"violence sports","rank":1},
2  {"keyword":"Football","rank":1},
3  {"keyword":"Radicalism","rank":0.5},
4  {"keyword":"Ethnic Conflict","rank":0.5},
5  {"category":"Political Sociology","rank":1},
6  {"category":"Decision Making","rank":0.66},
7  {"category":"Sociology","rank":0.66}}
```

In this example, the user has submitted the query *violence sports*. From the corresponding result set and the viewed document, the categories and keywords are derived with a rank that represents the normalised frequency of that term in the result list and the viewed document. It may happen that we have an overall occurrence count of 1 resulting in a rank of 1 for each keyword and classification,

which is the case when a user enters our DL from a web search engine and goes straight to the detailed view of a document. In this case, we use each keyword and classification from the seed document as session context.

The actual boosting is performed in the following order: by the entered query terms, the keywords, and the classifications each multiplied by their normalised rank (see Listing 3).

An example query for the re-ranking based on session details is displayed in Listing 4. The session context in this example is derived from the example session context in Listing 3. The first part of the query is based on the default ranking which is extended by a boosting parameter *bq*. In line 5, it can be seen that we look for the previously entered query term *violence sports* in the title field which is also the metadata field with the highest boosting value. In the lines 6 to 8 we boost documents that contain the most frequent keywords in the session context and boost each keyword with a decreasing factor depending on the keywords' frequency in the session context. In line 9 to 11 we boost documents that contain the category terms from the session context.

Listing 4: Example query for the re-ranking based on session details

```
1   q =>
2   keyword:((violence OR "Gewalt")^400 OR
3   keyword_free:((violence OR "Gewalt"))^250
4   [bq]=
5   (title:violence sports^1700) OR
6   (keyword:Football^1200 OR
7    keyword:Radicalism^1080 OR
8    keyword:Ethnic Conflict^1080) OR
9   (category:Political Sociology^800 OR
10   category:Decision Making^560 OR
11   category:Sociology^560 )
```

4 METHODOLOGY

At the beginning of a session each user is assigned one ranking method for the entire duration of a session. As most users click only on one suggested document, we use the rank position of the first clicked document as quality criterion. The most obvious metric to be used for this case would be mean reciprocal rank, but it has been argued in [12] that this metric is not on an interval scale, and thus the mean cannot be applied. Instead, we used the proposed alternative *mean first relevant (MFR)* which takes the rank position of the first clicked document in a result set and computes the arithmetic mean for all result sets that were generated using a stratagem search. This measure is proportional to the effort a user has to invest in finding the first relevant document in a result list: An MFR value of x represents the x-fold effort in comparison to the ideal value of 1. As the number of documents in a result set varies, it may happen that a result set only contains a small number of documents. One typical example is the search for other documents of a certain author or the search for a highly specific keyword. In this case, the MFR is usually rather low and might bias our results. Therefore, we additionally measured the MFR≥20 for all result sets that contain at least 20 documents which is the default number of documents on the first result page in Sowiport. Due to potential outliers that

may distort our results, we furthermore disregard document views that were performed on the third page of a result set (first relevant > 40) when measuring MFR.

Besides the MFR values, we measure other session-related features like session length, number of document views, session duration, and usefulness of stratagems. To measure the usefulness of stratagems, we apply a measure previously described in [15] in which different interaction signals within a session can be used to estimate the success of a search session. For instance, if a user bookmarks a certain document one can consider this document as relevant. The list of implicit relevance signals considered to measure the usefulness is displayed in Table 1.

Short	Description
Add to favourites	Bookmark a single or multiple records to favourites
Goto Google Scholar	Search a record in Google Scholar
Goto Google Books	Search a record in Google Books
Goto Fulltext	View the full text of the record
Goto local availability	Check availability in the local library
Export record	Export the record in different citation styles or via e-mail

Table 1: List of implicit relevance feedback signals

To estimate if and how useful a stratagem was for the outcome of a search session, we measure the usefulness by observing the number of positive signals in the log file after the usage of a stratagem on two levels:

- Local usefulness
  The local usefulness counts the total number of implicit relevance signals on a document that was contained in the result set immediately after a stratagem run. For a visual example of implicit relevance feedback on a result set see the right box in Figure 2b.
- Global usefulness
  For each session, we count the total number of implicit relevance signals contained in the entire session after the first usage of a stratagem.

Using the local usefulness we determine the "immediate" relevance of a document in a result set after a stratagem search. The global usefulness on the other hand measures the usefulness of a stratagem search for the entire session.

5 RESULTS

In total, we analysed 607,109 unique sessions during our living lab study.

Descriptive statistics on the overall usage of stratagems can be found in Table 2. During our experiment, all three approaches were nearly equally distributed among the users (see column 2). Users applied a stratagem search 77,036 times from which 5,839 documents were clicked from the result lists. In column 5, we measured the mean dwell time starting from the first usage of a stratagem until the end of the session. When looking at the dwell time, we identified

numerous outliers. We therefore removed all sessions that exceeded a dwell time of more than 20 minutes when measuring the average dwell time. On average, the users continued their search for 2.1 minutes after the first stratagem usage. All three approaches show a similar dwell time. Users that were assigned approach A or C continued their search for 123 seconds while approach B was continued for 133 seconds. In column 4, the mean number of interactions for all sessions containing a stratagem search is displayed. It can be seen that the number of interactions do not differ between the three approaches. In column 3, the number of document views from stratagem usage per approach is displayed. Both contextualised approaches have a considerably higher number of document views from stratagem search in comparison to the non-contextualised approach. Using the baseline, only 1,985 documents were viewed while the contextualisation based on similarity received more than 3,200 document views from stratagem searches. The contextualisation based on the session context also clearly outperforms the baseline with 2,627 document views. This tendency can also be seen in the click-through rates which are discussed later on in this section.

To estimate the effectiveness of the contextual ranking features in comparison to the non-contextualised baseline, we utilised the mean first relevant metric (MFR) as described in Section 4: the idea of the MFR metric is to take the position of the first clicked document in the result set. In Table 3, the MFR values for all sessions are displayed. The contextualisation based on similarity to the seed document performed best with an MFR of 3.10. The contextualisation based on session details received an MFR of 3.62 while the non-contextualised ranking performed the worst with an MFR of 4.66[4]. Besides the better results in terms of MFR, we can observe a higher click-through rate (denoted as N) for both contextualised approaches. For the contextualisation based on similarity, we observed nearly twice as many first relevant clicks (N=1999) compared to the baseline (N=1078).

The contextualisation based on the session context also clearly outperforms the baseline with 1,571 first relevant clicks.

Due to highly skewed data, we utilised a non-parametric Mann-Whitney U-Test to seek for significant differences in the results with a Bonferroni corrected $p^* = 0.016$. We see that both contextualised ranking features significantly outperform the non-contextualised ranking with A > B, p^*=0.001 r=0.13 and A > C, p^*=0.014 r=0.04. Furthermore, we see that the contextualisation based on similarity significantly outperforms the overall session context with (B > C, p^*=0.001 r=0.04).

As we may have different result set sizes, we furthermore measure the MFR$\geq$20 for all result sets that contain at least 20 documents. The MFR$\geq$20 values are displayed in column 4 of Table 3. The MFR$\geq$20 values for all three approaches increased with a stronger impact on approach A and C. We can observe that the MFR for the non-contextualised approach increased from 4.66 to 6.47. However, we can not observe a substantial difference when

---

[4]In addition to MFR, we calculated mean reciprocal rank (MRR) values for all three approaches. MRR showed the same tendency as MFR, so we decided not to report MRR separately in this paper.

Approach	Total stratagem usage	Document views from stratagem search	Mean interactions per session	Mean dwell time (s)
(A) Baseline	25,426	1,985	7.61	123.79
(B) Similarity	25,475	3,212	7.91	134.98
(C) Session Context	26,135	2,627	7.76	123.27

Table 2: Descriptive statistics for the period of the study

Approach	MFR	SD	MFR ≥20	SD
(A) Baseline	4.66 (N=1078)	6.45	6.47 (N=607)	7.74
(B) Similarity	3.10* (N=1999)	4.41	3.39* (N=1528)	4.81
(C) Session context	3.62* (N=1571)	4.74	4.30* (N=1097)	5.33

Table 3: Mean first relevant

comparing MFR and MFR>20. The contextualisation based on document similarity still performs best while both contextualisation approaches still outperform the baseline significantly.

The efficiency of approach C highly depends on how rich the interaction of the user was before performing a stratagem search. If a user just entered the DL without any previous interaction we have a cold start problem. Therefore, we compared the MFR for different history sizes (defined as the number of interactions prior to the stratagem usage) to evaluate whether the usage of a stratagem later on in a session has any influence on the performance of the approaches. In Table 4, the MFR values for 3 different history sizes are displayed.

The results in Table 4 show that with a growing history size the MFR values of all three approaches increase. Looking at the non-contextualised approach the MFR at early stages of the session is 4.52 while increasing to 5.06 after a history size > 5. If we compare sessions with a history size ≥ 10, we can observe that the MFR for all approaches further increases (A=5.49, B=3.59, C=3.80). The differences between the MFR values get even more evident if we look at the percentage increase between the different history sizes. For a history size of H ∈ [11, ∞] the MFR values of the baseline increased by 21.46% compared to a history size of H ∈ [2, 5] while the contextualisation based on session context only increased by 6.14%. However, the number of sessions with a history size ≥ 10 is comparably low with only 487 sessions. Having a larger sample size would improve the reliability of these observations.

To measure the usefulness of the contextualisation in comparison to the baseline, we utilise implicit relevance signals which can be found in Table 1. In general, we measure the usefulness by observing the number of positive signals in the log file after the first usage of a stratagem on two levels: a) the local usefulness b) the global usefulness.

To exclude outliers from this experiment, we only considered sessions with a total of implicit relevance signals ≤ 10 . The results of this experiment are displayed in Table 5.

Approach	Local usefulness	Global usefulness
(A) Baseline	232	5,385
(B) Similarity	628	5,684
(C) Session Context	334	5,294

Table 5: Usefulness of stratagem browsing per session

Regarding the number of local usefulness signals displayed in column 2, we can observe that the contextual ranking (B=628,C=334) again outperforms the non-contextualised baseline (A=232). Regarding the global session usefulness, the results show only marginal differences between the contextual approaches and the non-contextual baseline. The similarity approach again performs best with 5,684 implicit relevance signals in total after the first usage of a stratagem search. However, the results for the global usefulness show only marginal differences between the contextual approaches and the non-contextual baseline.

6 DISCUSSION

By measuring the position of the first relevant document in the result set (MFR), we showed that both contextualisation approaches lead to significantly better results in comparison to the baseline. Besides the better results of the MFR we see a considerably higher click-through-rate of the contextualisation (B=1999, C=1571) compared to the baseline (A=1078). Both, the MFR and the click-through rates are strong indicators for a better effectiveness of the contextualised approaches in comparison to the non-contextualised baseline.

When limiting the MFR measure to all result sets that contain at least 20 documents we observed an increase of the MFR for all three approaches. The non-contextualised baseline for instance increased from an MFR of 4.66 to 6.47.

The results for MFR and the higher click-through rate do not include information about the relevance of the clicked documents but rather about a topical relatedness to the users search interests.

To overcome this problem and to get an insight into the perceived relevance, we additionally utilised the usefulness metric that takes implicit relevance signals into account. This analysis was carried out on two levels: a) the local usefulness of a search result after a stratagem run and b) the global usefulness which measures the implicit relevance signals for the entire session. The results for the local usefulness are in line with the results of the MFR. Both

Approach	MFR H $\in [2, 5]$	MFR H $\in [6, 10]$	MFR H $\in [11, \infty]$
(A) Baseline	4.52 (N=802)	5.06 (+11.94%, N=276)	5.49 (+21.46%, N=112)
(B) Similarity	3.04 (N=1491)	3.29 (+8.22%, N=508)	3.59 (+18.09%, N=215)
(C) Session context	3.58 (N=1201)	3.77 (+5.30%, N=370)	3.80 (+6.14%, N=160)

Table 4: Mean first relevant for different history sizes (H)

contextualised approaches outperform the baseline. The contextualisation based on document similarity gathered more than twice as many direct relevance signals as the baseline while the contextualisation based on the session context was also considerably higher (A=232,B=628,C=334). This underpins the assumption that the contextualised approaches provide a more effective topical relatedness to the users search interests. By utilising the global usefulness, we measured the number of implicit relevance signals after the first stratagem usage for the entire session. For the global usefulness, we identified only minor differences in the overall success of the sessions. This may be an indicator for a rather equal performance of the three approaches in terms of satisfying the users information need.

For the session context approach we identified a typical cold start problem. A majority of users visit Sowiport via a web search engine which indexes the detail view of a document. Users employing a stratagem for further exploration are very common in the transaction logs. However, in this case we have no information about the users information need and therefore, only little information can be used for contextualisation. The more interactions a user performs the more detailed the session context of the user can be modeled. This was also indicated by the results of our segmentation into different sets of window sizes in Table 4. We observed an increase in terms of MFR for larger history sizes. We assume that the performance of the session context depends on the complexity of the search task. If a session contains various interactions this is also an indication for the complexity of a search task. We assume that sessions with a higher number of interactions also have greater demands on the quality of a result list as all three approaches performed worse when the number of interactions increased.

We observed a general tendency for an improved performance of the session context with growing history sizes. The MFR values for the contextualisation based on similarity increased by 18.09% for larger history sizes in comparison to a history size of $H \in [2.5]$. The contextualisation based on session details only increased by 6.14%. The non-contextualised approach had the largest increase with 21.46%. Unfortunately, the number of sessions with a history size ≥ 10 is comparably low with only 487 sessions. Having a larger sample size would improve the reliability of these observations.

To overcome the cold start problem a hybrid approach could be implemented that contextualises browsing based on document similarity at the beginning of a session while longer sessions are contextualised based on session details. As the majority of search sessions in *Sowiport* comprises (1 to 5) interactions we can not generalise the results of our session context approach (C). We assume that for highly interactive search sessions the results of the re-ranking will differ and lead to a better performance.

Surprisingly, other session related measures like the dwell time or the number of interactions did not differ substantially between the approaches.

7 STRENGTHS AND LIMITATIONS OF THE STUDY

The results of the transaction log study provide insights towards a general usefulness of contextualised ranking features in exploratory search. Both, short-term and long-term ranking features outperform the baseline. While the baseline describes the general behaviour of many DLs it may not be suitable for comparison with more sophisticated approaches like the ones that are presented in this paper.

One downside of the present paper lies in the nature of transaction log studies. Our results indicate a significantly better ranking in terms of MFR on the level of contextual ranking features. However, we are not able to interpret our results in the light of quality and overall relevance. A similar argumentation is provided in [17] where a presentation bias in click-through data is observed. To gain a more qualitative view on the usage of contextual ranking features a user study could be conducted that gives insights into the relevance of documents and particular reasons for regarding or disregarding certain documents.

We used implicit relevance ratings to measure the overall success of each session. Although the implicit relevance ratings had previously been used to measure the usefulness of DL features [15], they are not evaluated entirely. Thus, the results for this observation are influenced by our subjective relevance ratings of search activities. However, they provide a suitable measure to estimate the usefulness of search results.

One downside of the present approach is that the user is not aware of the contextual ranking features. In a first attempt, we provided the users with an information box which aimed to give some information on the ranking mechanism and the ability to disable the contextualised ranking. After running a pretest we decided to remove all information about the ranking features on the user interface as this led to confusion and distorted results.

By assigning each user one approach for the entire session, we kept the overall usage balanced. In this way, we eliminated seasonal effects like the start of a university semester or weekends. On the other hand, we ignored potential shifts in the users' search interest during the session. A more sophisticated approach would be to identify session boundaries.

Another advantage of the present study is that it can be easily reproduced by other researchers and DL developers. As we relied on the MLT query parser which is a standard similarity function build in SOLR one can reproduce our approach in its own environment. The re-ranking based on the session context uses SOLR boost

functions which also can be reproduced as long as a suitable user model can be derived.

8 CONCLUSION

In this paper, we have introduced two contextualisation approaches on the level of browsing in a DL: a) a contextualisation based on document similarity and b) a contextualisation based on implicit feedback about the users' search activities. Using a living lab environment, we evaluated our experiment on 47,444 sessions that contained browsing activities. Our results show that a contextualisation of browsing in DLs significantly outperforms the baseline in terms of MFR. Furthermore, we observed an increase in the click-through rate for our contextualisation approaches in comparison to the baseline. Other session related factors like dwell time or the total number of interactions per session did not differ between the approaches. As we utilised basic SOLR functions like the *More Like This* query parser to contextualise stratagem searches, our approach can easily be reproduced by other researchers and DL developers.

Although our contextualisation approach outperforms the baseline, a further investigation is necessary as our living lab study can not give insights into the overall relevance of the viewed documents. To tackle this problem we measured the local and global usefulness which take implicit relevance feedback into account. The results of our local usefulness analysis show that the contextualisation had a strong impact in terms of topical relatedness when selecting a document from the result set. Again, both contextualised approaches outperform the non-contextualised baseline. The global usefulness for the entire session, however, did not show significant differences between the contextualised and the non-contextualised re-ranking. Thus, we can state that the contextualised ranking features performed significantly better in terms of MFR, click-through rate and local usefulness. However, insights on the usefulness of our feature with regards to the users information need can not be derived from our living lab study. The increase in terms of the click-through rate and the significantly better MFR values, however, provide us with a strong indication that users could benefit from contextualised ranking features on the level of stratagems and that further research in this area could be of benefit for the DL community. Future work will be to conduct a user study that provides us with qualitative feedback on the effectiveness of our contextualisation approaches.

REFERENCES

[1] Azhar Alhindi, Udo Kruschwitz, Chris Fox, M Albakour, et al. 2015. Profile-based summarisation for web site navigation. *ACM Transactions on Information Systems (TOIS)* 33, 1 (2015), 4.

[2] Leif Azzopardi and Krisztian Balog. 2011. Towards a living lab for information retrieval research and development. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 26–37.

[3] Marcia J Bates. 1990. Where should the person stop and the information search interface start? *Information Processing & Management* 26, 5 (1990), 575–591.

[4] Nicholas J. Belkin, Daniel Hienert, Philipp Mayr, and Chirag Shah. 2017. Data Requirements for Evaluation of Personalization of Information Retrieval - A Position Paper. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum*. CEUR-WS.org, Dublin, Ireland. http://ceur-ws.org/Vol-1866/paper_193.pdf

[5] Paul N Bennett, Ryen W White, Wei Chu, Susan T Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. 2012. Modeling the impact of short-and long-term behavior on search personalization. In *35th SIGIR conference*. ACM, 185–194.

[6] Zeljko Carevic, Maria Lusky, Wilko van Hoek, and Philipp Mayr. 2017. Investigating exploratory search activities based on the stratagem level in digital libraries. *International Journal on Digital Libraries* (2017), 1–21. http://link.springer.com/10.1007/s00799-017-0226-6

[7] Zeljko Carevic and Philipp Mayr. 2015. Extending search facilities via bibliometric-enhanced stratagems. (2015), 40–46. http://ceur-ws.org/Vol-1344/paper5.pdf

[8] Zeljko Carevic and Philipp Mayr. 2016. Survey on High-level Search Activities based on the Stratagem Level in Digital Libraries. In *International Conference on Theory and Practice of Digital Libraries*. Springer, 54–66. http://link.springer.com/10.1007/978-3-319-43997-6_5

[9] David Ellis. 1989. A behavioural approach to information retrieval system design. *Journal of Documentation* 45, 3 (1989), 171–212.

[10] Jody Condit Fagan. 2010. Usability studies of faceted browsing: A literature review. *Information Technology and Libraries* 29, 2 (2010), 58.

[11] Enrique Frias-Martinez, Sherry Y. Chen, and Xiaohui Liu. 2009. Evaluation of a Personalized Digital Library Based on Cognitive Styles: Adaptivity vs. Adaptability. *Int. J. Inf. Manag.* 29, 1 (Feb. 2009), 48–56.

[12] Norbert Fuhr. 2017. *Some Common Mistakes In IR Evaluation, And How They Can Be Avoided.* Technical Report. University of Duisburg-Essen, Germany.

[13] Marti Hearst. 2009. *Search user interfaces.* Cambridge University Press.

[14] Marti A Hearst. 2006. Clustering versus faceted categories for information exploration. *Commun. ACM* 49, 4 (2006), 59–61.

[15] Daniel Hienert and Peter Mutschke. 2016. A usefulness-based approach for measuring the local and global effect of IIR services. In *Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval*. ACM, 153–162.

[16] Daniel Hienert, Frank Sawitzki, and Philipp Mayr. 2015. Digital library research in action: supporting information retrieval in Sowiport. *D-Lib Magazine* 21, 3 (2015), 8. https://doi.org/10.1045/march2015-hienert

[17] Thorsten Joachims et al. 2003. Evaluating Retrieval Performance Using Click-through Data. (2003).

[18] Evangelos Kanoulas, Ben Carterette, Mark Hall, Paul Clough, and Mark Sanderson. 2011. Overview of the TREC 2011 Session Track. In *TREC 2011*. NIST.

[19] Carl Lagoze, Dean B. Kraft, Sandy Payette, and Susan Jesuroga. 2005. What Is a Digital Library Anyway?: Beyond Search and Access in the NSDL. *D-Lib Magazine* 11, 11 (Nov. 2005). https://doi.org/10.1045/november2005-lagoze

[20] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.

[21] Nicolaas Matthijs and Filip Radlinski. 2011. Personalizing web search using long term browsing history. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 25–34.

[22] Philipp Mayr and Ameni Kacem. 2017. A Complete Year of User Retrieval Sessions in a Social Sciences Academic Search Engine. (2017), 560–565. https://doi.org/10.1007/978-3-319-67008-9_46

[23] Philipp Mayr and Vivien Petras. 2008. Cross-concordances: terminology mapping and its effectiveness for information retrieval. In *74th IFLA World Library and Information Congress*. IFLA, Quebec, Canada. http://www.ifla.org/IV/ifla74/papers/129-Mayr_Petras-en.pdf

[24] Nils Pharo, Thomas Beckers, Ragnar Nordlie, and Norbert Fuhr. 2011. Overview of the INEX 2010 Interactive Track. 227–235.

[25] M.Elena Renda and Umberto Straccia. 2005. A personalized collaborative Digital Library environment: a model and an application. *Information Processing & Management* 41, 1 (Jan. 2005), 5–21. https://doi.org/10.1016/j.ipm.2004.04.007

[26] Ian Ruthven, Mounia Lalmas, and Keith Van Rijsbergen. 2003. Incorporating user search behavior into relevance feedback. *Journal of the Association for Information Science and Technology* 54, 6 (2003), 529–549.

[27] Anna Sepliarskaia, Filip Radlinski, and Maarten de Rijke. 2017. Simple Personalized Search Based on Long-Term Behavioral Signals. In *European Conference on Information Retrieval*. Springer, 95–107.

[28] Xuehua Shen, Bin Tan, and ChengXiang Zhai. 2005. Context-sensitive information retrieval using implicit feedback. In *28th SIGIR conference*. ACM, 43–50.

[29] Alan F. Smeaton and Jamie Callan. 2005. Personalisation and recommender systems in digital libraries. *International Journal on Digital Libraries* 5, 4 (01 Aug 2005), 299–308. https://doi.org/10.1007/s00799-004-0100-1

[30] Jaime Teevan, Daniel J Liebling, and Gayathri Ravichandran Geetha. 2011. Understanding and predicting personal navigation. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 85–94.

[31] Ryen White, Joemon Jose, C van Rijsbergen, and Ian Ruthven. 2004. A simulated study of implicit feedback models. *Advances in Information Retrieval* (2004), 311–326.

[32] Ryen W White, Wei Chu, Ahmed Hassan, Xiaodong He, Yang Song, and Hongning Wang. 2013. Enhancing personalized search by mining and modeling task behavior. In *22nd international conference on World Wide Web*. ACM, 1411–1420.

[33] Ryen W White and Steven M Drucker. 2007. Investigating behavioral variability in web search. In *Proceedings of the 16th international conference on World Wide Web*. ACM, 21–30.

[34] Ryen W White, Joemon M Jose, and Ian Ruthven. 2002. Comparing explicit and implicit feedback techniques for web retrieval: Trec-10 interactive track report. In *Proceedings of the Tenth Text Retrieval Conference (TREC-10)*. 534–538.