

Carstens, Carola; Rittberger, Marc; Wissel, Verena
Information search behaviour in the German Education Index.

formal überarbeitete Version der Originalveröffentlichung in:

International Conference on Digital Library Management (ICDLM). New Delhi : TERI 2011, S. 388-398

urn:nbn:de:0111-dipf-37893

Nutzungsbedingungen

Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Die Nutzung stellt keine Übertragung des Eigentumsrechts an diesem Dokument dar und gilt vorbehaltlich der folgenden Einschränkungen: Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Kontakt:

Deutsches Institut für
Internationale Pädagogische Forschung
Mitglied der Leibniz-Gemeinschaft
Frankfurter Forschungsbibliothek
Schloßstraße 29
D-60486 Frankfurt am Main
publikationen@dipf.de
www.dipf.de/de/bildungsinformation/ffb

Information Search Behaviour in the German Education Index

Carola Carstens*, Marc Rittberger, Verena Wissel

German Institute for International Educational Research, Information Center for Education

Schloßstraße 29, 60486 Frankfurt, Germany

* Corresponding author: Email: carstens@dipf.de

Tel: 0049 (0)69 24708 311, Fax: 0049 (0)69 24708328

Abstract: This paper presents results from the analysis of user interactions in the GEI (German Education Index) specialised search engine. The main interest of this study is getting insights into the GEI users' typical search behaviour to identify optimisation potential for the further development of the GEI search engine. Based on a sample of query logfiles, the study has revealed that the users apply advanced search functionalities and query reformulation tactics adequately, though not frequently, and that more support in their application could possibly increase the effectiveness of the searches in the GEI.

Keywords: Information search behaviour, search tactics, search strategies, logfile analysis, digital libraries

Category: Thesauri and ontologies, semantics, metadata, and retrieval

1 Motivation

In DL (digital library) research, the study of user interactions is a recent trend, as stated by (Xie 2008: 120). Compared to general purpose search engines, digital libraries, as well as specialised search engines, commonly offer more sophisticated advanced search functionalities such as the search in different fields and the support of Boolean search syntax.

While much research on search behaviour in general purpose web search engines exists, this is far less the case for DL and specialised search engines. To contribute to the study of user interactions in DL-like environments, this paper will present results from the analysis of user interactions in the GEI specialised search engine.

2 Related Work

The study of user interactions in retrieval systems can be subsumed under the term information search behaviour, which (Wilson 1999: 263) defines as being concerned with 'the interactions between information user (with or without an intermediary) and computer-based information systems, of which information retrieval systems for textual data may be seen as one type'.

(Bates 1979) and (Harter 1986) enumerate tactics and strategies that can serve to analyse the way users interact with retrieval systems. (Bates 1979) defines a search tactic as 'a move made to further a search', thus serving for the realisation of a certain retrieval goal. She presents a catalogue of 29 search tactics. Amongst others, she distinguishes two kinds of tactics – search formulation tactics and term tactics. The former refer to the design of the query structure. For example, adding a new facet to a query would fall into this category. Term tactics, by contrast, describe the selection of terms in the query formulation. This includes for example the use of related, broader or narrower terms. As defined by (Harter 1986), different tactics may form part of a search strategy, the 'overall plan or approach for a search problem'.

A distinction can be drawn between subject search-strategies and known item search-strategies. While the former serve to query for documents that cover a certain topic, the latter query by already known document facts.

As delineated by (Jansen 2009), query logfile analysis is a well-established method for examining the users' search behaviour, which has often been applied to web search engines. For example (Silverstein et al. 1999) and (Rieh and Xie 2006) have analysed query logfiles of the Excite search engine, while (Spink et al. 2001) have applied this method to gain insights into the Altavista users' search behaviour. The study by (Rieh and Xie 2006) analyses the nature of query reformulations in the Excite search engine in depth. The authors identify parallel movements, such as the use of related terms, as the most common of all reformulation patterns (51.4%). Query specifications are less frequent (29.1%) but still more common than generalisations (15.8%), while the replacement with synonyms occurs seldom (3.7%).

(Spink et al. 2001) further state that many users (48.4%) submit only a single query and consult only few documents. Another finding in this study refers to the use of advanced search functionalities such as Boolean operators, which is reported to be scarce. The authors conclude that 'most people use few search terms, few modified queries, view few Web pages, and rarely use advanced search features'.

In their analysis of search behaviour in a digital library environment, (Jones et al. 1995) report a more frequent use of Boolean operators than (Spink et al. 2001). Nevertheless, they found out that the use of Boolean search was influenced by the default search configuration. During the experimental period, this configuration varied between Boolean and ranked search. In either case, users showed a tendency to maintain the preselected configuration.

(Wildemuth and Moore 1995) have conducted a study on search engine behaviour in a specialised search engine. They analysed 161 searches in the MEDLINE database with a focus on search effectiveness, which is judged by librarians. The authors state that the retrieval effectiveness could be improved by a more frequent use of synonyms, the correct application of Boolean operators and a more frequent consultation of controlled vocabulary resources such as an online thesaurus.

The search behaviour in the GEI specialised search engine has already been explored in a predecessor study by (Carstens et al. 2009). It aimed to assess if search tactics and strategies listed by (Bates 1979) and (Harter 1986) are identifiable in the query logfiles of the GEI. This explorative analysis used a sample of long sessions that were expected to comprise complex searches and focused on the analysis of subject searches. As a result, 19 different tactics and 6 search strategies could be identified in the data set, for which the paper gives illustrative examples.

3 Research Interest

The study at hand builds on the above mentioned results of the study by (Carstens et al. 2009), applies the query logfile analysis to a more comprehensive data set and analyses it more in depth. For example, the number of hits is taken into account in the analysis, which allows to investigate query result-specific user reactions and to draw conclusions regarding the possible effectiveness of the searches. Unlike the previous study, it will also examine characteristics of known item searches.

This way, the study will reveal which tactics and strategies the GEI users typically apply. As stated in the related works section, several studies have shown that users scarcely apply advanced search functionalities through which the effectiveness of their searches could possibly be increased. It will be examined whether this also holds true for the GEI. These findings can serve to derive ideas for the further development of the GEI.

Moreover, the study will deliver insights into user interactions in a DL-like environment which offers sophisticated search functionalities for retrieving metadata documents.

4 The German Education Index

The GEI specialised search engine, a part of the GEP (German Education Portal)¹, comprises more than 700,000 bibliographic references for the domain of educational research, primarily in German language (more than 80%). Its metadata documents consist of fields like index terms, title, author/editor, institutions, abstract and source, while full texts do seldom form part of the GEI.

The GEI is based on the Lucene² open source search engine framework. Its underlying retrieval model is a combination of the Boolean model and the vector space model. The GEI implements both a simple and an advanced search mode, the latter being the default entry point that is interlinked by the GEP. The simple search is implemented as a freetext search over the above listed document fields.

In the advanced search mode, users can define more sophisticated queries than in the simple search (see Figure 1). For example, they can restrict their search to certain fields (1) that can also be combined by the Boolean operators AND, OR and NOT (2). Within each field, query terms are by default connected by the AND operator (3). If desired, this pre-configuration can be changed by the users, thus leading to the combination of query terms by the OR operator. Moreover, a person and an index term register can be consulted to look up query terms for the respective search fields (4, 5). The latter register also serves to identify synonyms of query terms which are used to automatically expand the queries.

Figure 1: Advanced search mode in the GEI

A survey by (Wendt and Patjens 2007) revealed that the majority of the surveyed GEP users has an academic background in education related fields and uses specialised information primarily for research purposes.

5 Research Method

To get a comprehensive insight into the search behaviour of GEI users, the query logfiles of one typical weekday, 15 October 2009, are analysed. For this purpose, the logfile entries are grouped by their anonymised ip addresses, resulting in an amount of 870 ip-specific sessions. As shown in Figure 2, these may comprise several distinct search sessions whose consecutive query formulation steps are both timely and topically related, the latter being verified by a human assessor. This way, an amount of 1823 search sessions is identified.

¹ http://www.fachportal-paedagogik.de/start_e.html

² <http://lucene.apache.org/>

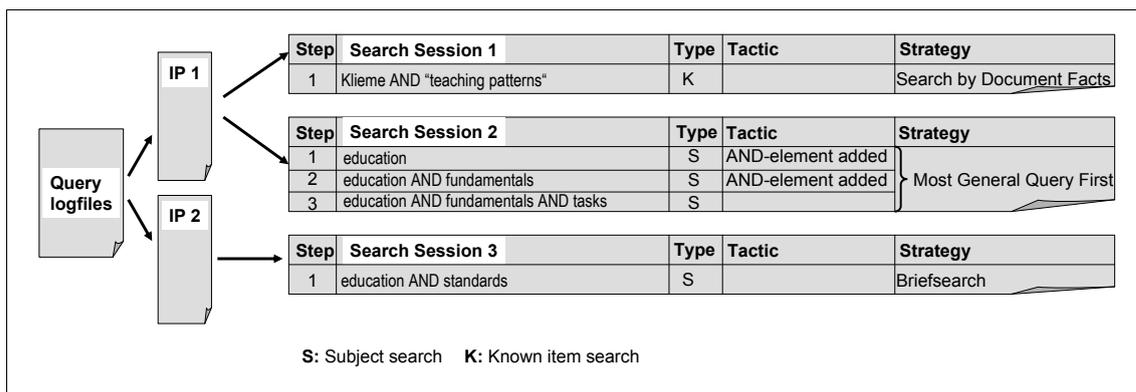


Figure 2: Identification of tactics and strategies in query logfiles

Each search session may comprise one or more distinct search steps, as illustrated in Figure 2. For the purpose of analysis, each search step is classified as either belonging to the subject search- or to the known item search-category. By analysing how queries are reformulated from step to step, search tactics are identified and categorised by a human assessor. A sequence of such distinct tactics can suggest the application of a certain search strategy.

In search session 2 presented in Figure 2, a query is for example specified by applying the tactic of subsequently adding query facets. The searcher may have decided not to include all query facets in the initial query in order to prevent over-specification. New query facets are then consecutively added to the query, following the most general query first strategy.

If the initial query starts with the most specific facet, this is referred to as the most specific concept first strategy by (Harter 1986). Nevertheless, this strategy makes assumptions about the users' expectations on the number of result documents and is therefore not directly identifiable by logfile analysis. But the description of this strategy inspired the definition of two new strategies, referred to as most specific query first and most general query first in (Carstens et al. 2009). Instead of making assumptions about the expected number of result documents, these strategies describe the query structure and the specificity of the query terms. Following the most general query first strategy, the searcher starts with the most general facet(s) and consecutively adds new facets that further specify the search and imply a reduction of the result list. The most specific query first strategy proceeds in the contrary way. It starts with a specific query, often consisting of several facets, which are consecutively deleted or whose terms are generalised. Examples of these strategies are given in Figure 3 that presents an overview of all the subject search strategies that will be analysed in the GEI logfiles.

Subject Search Strategies	Example
Quick Approach	1) pedagogics
Briefsearch	1) deaf and „language support“
Pairwise Facets	1) "intercultural competence" AND pedagogics AND learning 2) "intercultural competence" AND pedagogics 3) "intercultural competence" AND learning
Most General Query First	1) education 2) education AND fundamentals
Most Specific Query First	1) "comparative tests" mathematics 2) tests mathematics
Building Blocks	1) "method competence" AND (library OR "school library")

Figure 3: Examples for subject search strategies

Apart from these sophisticated subject search strategies, the more simple quick approach strategy defined by (Chu 2003) has to be mentioned. It describes a search without using Boolean operators. As query terms are by default combined by the AND operator in the GEI, the quick approach only refers to single term-queries.

To get a first idea of the documents in a retrieval system, the briefsearch strategy is often applied. For this purpose, few query terms are combined with Boolean operators. This basic strategy often serves as an entry point to more complex strategies like the building blocks approach in which distinct query facets are expanded with semantically related terms, as illustrated by search session 3 in Figure 2.

A further search strategy is called pairwise facets. It is applicable if the query facets are all considered as equally important. Following the pairwise facets strategy, only two facets at a time are combined and finally, the result sets of all facet combinations are merged.

6 Results

Selected results from the logfile analysis are presented in the following. They refer to characteristics and tactics of both known item and subject searches. Out of the entirety of 3,631 search steps in the analysed search sessions, more than 33% are categorised as known item search steps, while subject search steps make up 67% of the search steps.

6.1 Search Mode

The analysis of the employed search masks demonstrates that most of the search steps are conducted in the advanced search mask, namely 85%. This may be influenced by the fact that the GEI's advanced search mask is directly linked from the German Education Portal entry page.

But although the use of the advanced search mask is high, in 47% of the search steps entered in the advanced search mode, only the freetext field is used. The advanced search mask is thus frequently employed in a simple search mode.

6.2 Strategies of Subject Searches

As can be seen from Figure 4, out of the total amount of subject search-steps, the biggest part (28%) are single-step quick approach searches. In special cases, quick approaches may also comprise more steps, if a single term-query is reformulated, for example by changing the spelling.

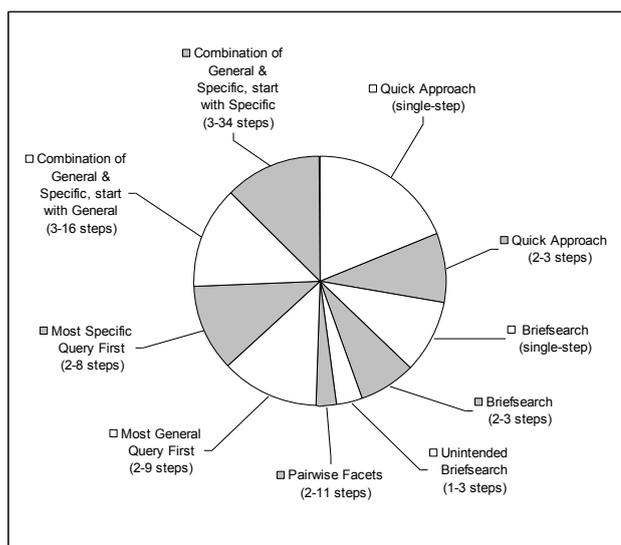


Figure 4: Distribution of search steps that form part of a certain strategy

Briefsearches are also common, making up 17 % of the subject search-steps. These are often single-step search sessions. Sometimes, briefsearches also comprise two or three steps, for example if the query terms are replaced by a spelling variant.

The search steps classified as unintended briefsearches make up 3% of the search steps. These are queries in which phrases or compound terms are entered without phrasing although it would possibly have been adequate. Due to the default configuration in the GEI, they are combined by the AND operator, thus being interpreted as a Boolean combination of terms.

The strategies most general query first and most specific query first consist of several query steps as the queries are consecutively generalised or specialised during the search session. Figure 4 depicts that the length of these strategies ranges from 3 to 9 steps in the GEI data set. Out of the 2,442 subject search-steps, 13% belong to a search sequence following the most general query first strategy, while 11% form part of a most specific query first strategy.

The combination of these two strategies is also common. This is illustrated by the fact that 26% of the subject search steps form part of a combined general & specific search strategy, either starting with a general or a specific query formulation.

The application of the pairwise facets strategy, by contrast, occurs seldom, and the building blocks approach could not be identified at all in this data set.

6.3 Types of Known Item Searches

Figure 5 depicts the distribution of different types of known item search steps. A high percentage (26%) of the total amount of 1,188 known item search steps is precision-oriented where possibly a small set of specific result documents is expected by the users. These are for example searches by title where typically only one document is viewed if result documents are delivered.

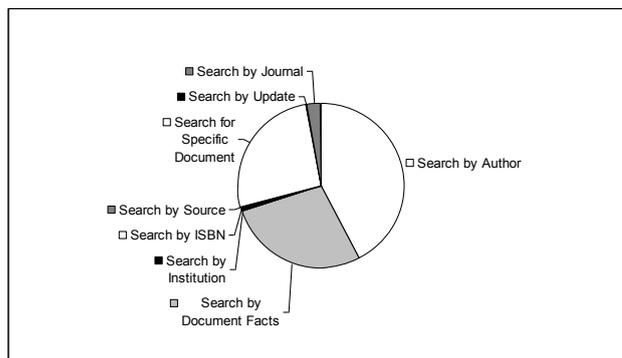


Figure 5: Types of known item search steps

Search steps classified as searches by document facts, by contrast, do not necessarily query for one specific document. Instead, already known facts are used to specify search criteria like the combination of author and year or topic and year. 28% of the known item search steps fall into this category.

Searching only by author name is a very common strategy. It makes up the majority (42%) of the known item search steps. About half of these search steps (54%) form part of longer search sessions, while the remaining 46% are made up by single-step search sessions. Although in 74% of the single-step sessions by author name, result documents exist, in the majority of these search sessions no document is viewed (61%). This may indicate that either the GEI is used to merely assess if a certain author is listed in the database or to check if new publications by an author are listed.

But the users often seem to be unsure about the adequate query formulation strategy for author names. The 501 search steps that query by author name comprise 92 two- or more-step sequences where author names are varied, e.g. by spelling variants, changing the order of first and last names or consulting the person register. These reformulations thus make up a large part of the search steps by author and show

the need for user support. In the current GEI implementation, this is offered by the person register which is used in about one quarter of the author searches.

6.4 Search Fields of Known Item Searches

In the cases where the users do not make use of the person register, they most frequently employ the designated author search field to define searches by author. But with nearly equal frequency, the freetext field is used for this purpose, as shown in Figure 6.

If users search for a specific document by the title or parts of the title, these queries are most frequently entered into the freetext search field while the title field is used in 19% of the cases, as illustrated in Figure 7. The author field is often used to specify these queries for specific documents. In 18% of the searches for specific documents, the author field is combined with the title field, and in 12% of these searches, it is employed in combination with the freetext field.

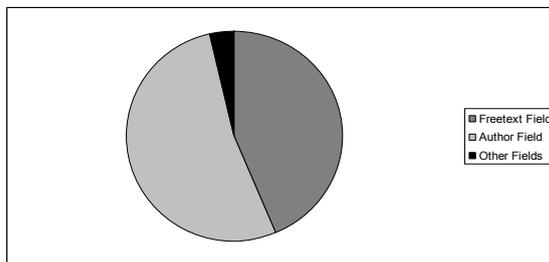


Figure 6: Fields used for searches by author name

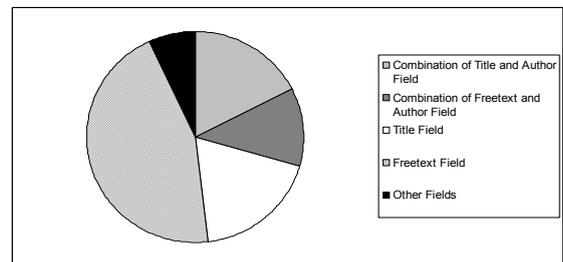


Figure 7: Fields used for searches for a specific document

In the queries for title searches, the use of phrasing would usually be appropriate. Nevertheless, this is employed in only 4% of the searches for specific documents. While the title searches without phrasing probably lead to the expected result documents if the entered title is both long and specific, large and probably unsatisfying result lists are delivered for short and general title queries, especially if the search is entered into the freetext field.

6.5 Search Tactics

Figure 8 shows that in the entirety of the 3,631 analyzed search steps, most types of search tactics occur only seldom.

While (Bates 1979) makes a distinction between term tactics (e.g. the use of a synonym) and search formulation tactics (e.g. the use of an AND-operator), the results of our predecessor study (Carstens et al. 2009) have inspired us to specify these tactics by analysing term and query structure characteristics in combination. For example, a certain term type like a synonym can either be added to a query, it can replace another term or it can be deleted from the query. Consequently, these variants are recorded as distinct tactics, as shown in Figure 8.

register terms from the keyword register or replace a term by a broader term. The users thus apply adequate tactics to overcome zero result-queries.

To enlarge result sets with few results, the most common tactics are the replacement of a query term by a related term and the deletion of a facet.

If a query delivers only few results, the search can either be interpreted as successful in the case of a precision-oriented goal, or as the contrary, in the case of a recall-oriented goal. In the logfiles, the possible goal can be inferred from the number of inspected documents. If none of the few results is inspected, the query can be classified as unsuccessful. If at least one of the few results is viewed, this raises the probability of a successful precision-oriented search. In the GEI logfiles, search sessions with an initial query with few results (1-3 hits) are immediately quit in 44% of the cases. But more than half of these searches can be interpreted as possibly successful precision-oriented searches as at least one document is viewed. Nevertheless, query logfiles do not allow to make an assumption about the relevance of the viewed documents. The inspection of a document is only an indicator of possible relevance.

6.7 How Users React to Large Result Sets in Subject Searches

If users receive large result sets (more than 100 hits) in the initial subject search-query of a search session, more than half of these search sessions (65%) are immediately ended, either without having viewed any document (30%) or after having viewed at least one result document (35%).

In the search sessions that are not ended, different tactics are employed to reduce the result set in the first reformulation step. The most common tactic for reducing result lists is the addition of a new search facet. Furthermore, the replacement of terms by related terms is common, and narrower terms are also employed for this purpose, although more seldom. Another frequent tactic is the change of search fields, e.g. from the freetext to the keyword field. But still, in many of these search sessions where initial queries with large result sets are reformulated (53%), no document is viewed.

7 Interpretation of Results

Overall, it can be stated that the majority of the search steps in the GEI is made up by subject searches. But known item searches are also frequent, a high fraction of which are searches for specific documents or searches by author name (6.3). This can be explained by the GEI's primary role as a bibliographic reference database. It is obviously often consulted to look up specific documents in order to assess their availability.

The study has shown that the advanced search mask is generally preferred by the users (6.1). Nevertheless, it is often used in a simple search mode, many searches in the advanced search mask only employing the freetext search field. This is also frequently used for known item searches although users also partly employ specific search fields like the author and title fields adequately for these purposes (6.3).

But in general, the GEI users do not frequently employ further advanced search functionalities like phrasing (for example for title searches (6.4)), truncations or register terms (6.5). This behaviour supports the findings by (Wildemuth and Moore 1995) who state that the retrieval effectiveness could be improved by fostering these functionalities. Even in DL-like environments, the users thus do not seem to fully exploit the potential of advanced search functionalities, which has already been noted by (Spink et al. 2001) with reference to a web search engine.

The fact that the advanced search mask is employed for a high fraction of the queries may be due to the GEP's pre-configuration which offers the advanced mask as the default option. This assumption would be in line with the findings by (Jones 1998) who states that users tend to maintain search pre-configurations.

This could also explain why the pre-configuration of Boolean operators is seldom changed in the advanced search mask. The analysis of term tactics has revealed that query terms are hardly ever combined by the OR-operator. Instead, queries are mostly reformulated by replacing terms, most frequently by related terms, which have also been identified as the most common term tactic by (Rieh and Xie 2006). Nevertheless, combining original query terms with semantically related terms by the OR-operator would possibly often be more effective than replacing the former. Especially in queries that comprise several facets, the expansion of a facet with a related term may be more effective for increasing recall than its replacement with a related term. But currently, the expansion of facets is difficult to define in the GEI, which may explain their scarce occurrence in the logfiles, as well as the non-application of the building blocks strategy in this data set.

Sophisticated search strategies are thus seldom applied and quick approaches make up for the biggest part of the search steps (6.2). These single-term queries often deliver a high number of results, which have been shown to frequently lead to immediate endings of the search sessions (6.6). The contrary situation of empty result sets has incurred similar reactions (6.7).

Although the users employ adequate tactics for specifying and generalising their queries (6.6, 6.7), the analysis has shown that the use of narrower terms (for specifying) and broader terms and synonyms (for generalising) is seldom (6.5). Instead, users tend to replace query terms by related terms or vary the number of query facets.

7 Conclusion

Based on the above stated results, two main areas for further developments in the GEI can be identified: the prevention of empty result sets, as well as of immediate endings of possibly unsuccessful searches. While the adaptation of the retrieval algorithm and the implementation of further automatic query expansion mechanisms may serve the former purpose, support in the application of search tactics and strategies may help to overcome unsuccessful query formulations. A more supportive term suggestion functionality than the currently implemented and scarcely used registers may serve this purpose. Based on these findings, it can for example be hypothesised that the implementation of query expansion mechanisms may lead to an increase in retrieval effectiveness, which is currently investigated by (Carstens 2009).

Moreover, the study has revealed search characteristics which are due to the GEI's role as a bibliographic reference database where known item searches make up for a big part of the queries. The offered advanced search functionalities are employed adequately by the users, though seldom. While these results are specific to the GEI, they are comparable to studies in other DL environments that may also based on the Lucene search engine, apply similar search forms, support Boolean search syntax and offer advanced search functionalities.

References

- Bates, M. 1979. **Information Search Tactics**. *Journal of the American Society for Information Science* 30(4): 205–214.
- Carstens, C. 2009. **Effects of Using a Research Context Ontology for Query Expansion**, pp 919-923. In *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*, edited by Aroyo, L. et al. Berlin et al.: Springer.
- Carstens, C., Rittberger, M., and Wissel, V. (2009). **How Users Search in the German Education Index – Tactics and Strategies**, pp 76-83. In *Proceedings of the Workshop Information Retrieval (WIR 2009)*, edited by Mandl, T., Frommholz, I. Darmstadt: TU.
- Chu, H. 2003. *Information Representation and Retrieval in the Digital Age*. Medford: Information Today.

- Harter, S.P. 1986. **Search Strategies and Heuristics**. In *Online Information Retrieval. Concepts, Principles and Techniques*, pp 170-204. Orlando: Academic Press.
- Jansen, B.J. 2000. **The Methodology of Search Log Analysis**. In *Handbook of Research on Web Log Analysis*, pp 100–123, edited by Jansen, B.B., Spink, A., Taksa, I. Hershey: Information Science Reference.
- Jones, S., Cunningham, S.J., and McNab, R. 1998. **Usage Analysis of a Digital Library**, pp 293-294. In *Proceedings of the International Conference on Digital Libraries 1998*, edited by Witten, I., Akscyn, R., Shipman, F.M. New York: ACM.
- Rieh, S.Y., Xie, H. 2006. **Analysis of Multiple Query Reformulations on the Web: The interactive Information Retrieval Context**. *Information Processing and Management* 42(3): 751-768.
- Silverstein, C., Marais, H, Henzinger, and M, Moricz, M. 1981. **Analysis of a Very Large Web Search Engine Query Log**. *SIGIR Forum* 33(1): 6-12. New York: ACM.
- Spink, A., Wolfram, D., Jansen, B.J., and Saracevic, T. 2001. **Searching the Web: The Public and Their Queries**. *Journal of the American Society for Information Science and Technology* 52(3): 226–234.
- Wendt, J., Patjens, S. 2007. *Auswertung zur Online-Umfrage unter Nutzern und Nichtnutzern des Fachportals Pädagogik*. Project report. Retrieved 13 October 2010 from http://evalinfo.dipf.de/evalinfo/upload/Fachportal_Paedagogik_2007_Nutzer-Nichtnutzerbefragung.pdf
- Wilson, T.D. 1999. **Models in Information Behaviour Research**. *Journal of Documentation* 55(3): 249-270.
- Xie, I. 2008. **Interactive IR in Digital Library Environments**. In *Interactive Information Retrieval in Digital Environments*, pp 116-152. Hershey: IGI Publishing.
- Wildemuth, B.M., Moore, M.E. 1995. **End-user Search Behaviors and their Relationship to search Effectiveness**. *Bulletin of the Medical Library Association* 83(3): 294-304.