

Exploring Users' Online Search Behaviour: a preliminary study in a library collection

Anna Mastora¹, Sarantos Kapidakis¹, Maria Monopoli²

¹Laboratory on Digital & Electronic Publishing, Archive & Library Sciences Department, Ionian University

²Department of Librarianship & Information Systems, Faculty of Administration & Economics,
Technological Educational Institute (TEI) Athens

{mastora, sarantos }@ionio.gr ; mariamonopoli@hotmail.com

Abstract

The purpose of this study was to investigate the search behaviour of users. Undergraduate and postgraduate users from the Department of Archives and Library Sciences, Ionian University, were invited to indicate the way they formulate and reformulate their queries. Students were asked to formulate queries and complete a questionnaire. Responses from the questionnaires and data collected from log files provided us with valuable information concerning the number of terms users type in the searching field, the type of failed queries users make and whether users reformulate their queries by using terms provided in the retrieved results. Results revealed that users mainly type in the searching field one term per query. Furthermore typographical errors, and specifically the substitution of a character with another, appeared to be the main reason for failed queries. Additionally, the vast majority of respondents declared that they used a term from the retrieved results. And finally, it is worth mentioning that users have equal chances to direct their queries in either a more specified or generalised term, whereas their choice of submitting parallel terms outweighed all other strategies.

Categories and Subject Descriptors

H.1.2.[User/ Machine Systems]; H.3 [Information Storage and Retrieval]; H.3.3 Information Search and Retrieval; H.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation, Evaluation

Keywords

Query formulation, Query reformulation, Query failure, Search behaviour

1 Introduction

This study is part of a broader research in progress at the Department of Archive and Library Sciences, Ionian University regarding the search behaviour of users. Our interest is focused on gaining an understanding of users' thinking behaviour while they are in the process of retrieving information relevant to their information needs. This implies the associations that users make and consequently the use of words as query terms.

In particular, this study aims to shed a light on the following issues:

- How users formulate and reformulate their queries
- How many terms users type in the searching field
- What type of failed queries users make
- Whether users reformulate their queries by using terms provided in the retrieved results

Such results are useful to help library users better achieving their goals and library system developers to better design their systems.

2 Related literature

There are numerous studies on user search behaviour that are focused either on traditional IR systems such as OPACs or on more modern systems such as digital libraries. Concerning the users' search behaviour, there is a large body of literature suggesting that when end-users look for data relevant to their information needs, they prefer to accomplish their goal with the less effort from them (Buzikashvili 2005). Lau and Goh (2006) specified that users are not patient in searching for information and this lack of perseverance leads them to terminate their searches rather than look beyond the first screen of results (hits).

This attitude is more or less the same when users have to deal with failed searches. Jones et al. (2000) defined a failed search as one that matches no documents in the collection. A survey on the OPAC of the Nanyang Technological University (NTU) revealed that a total of 317.840 queries returned zero hits and this figure accounted for 49,5% of the total number of queries submitted. This indicates that users have an almost equal chance of submitting a query that returns no hits (a search failure) or one that returns at least one record (hit) (Lau and Goh 2006).

At this point, it would be useful to provide some reasons that are responsible for users to have searches with zero results. One reason is the 'incorrect use' of searching terms. According to Jansen et al. (2000) an incorrect use is when users do spelling or grammar mistakes, use Boolean operators that are not supported by the IR system or leave spaces between searching terms. Another reason for failed searches is when users make typographical errors. Brown (1988) and Hirst & Budanitsky (2003) have categorised the typographical errors in four (4) groups named substitutions, omissions, insertions and transpositions. A substitution error is when a character of a term is replaced by another incorrect character. An omission error is when one or more characters of a term are missing. An insertion error occurs when one or more characters are added in a term. And a transposition error is when two or more characters of a term are not in the right order. Specifically, research of Lau and Goh (2006) analysed users' searching terms from typographical error point of view and concluded that only 24 out of 1000 failed queries were identified with typographical errors, while substitutions were encountered as frequently as omissions.

However, a quite interesting and optimistic conclusion is that users learn how to formulate better queries during their searching process. Joachims et al. (2007) specified in their article that although users gave a negative feedback about the results, they said that they learned to carry out a better query from these 'not good' results.

Regarding the query length, there is not a clear picture of how many terms users type in the searching field. Jansen et al. (1998) believed that the circumstances and the context between searches done by users of IR systems such as DIALOG and searches done by the general Internet population of the Web are different. Still, studies are concentrated on different aspects related to searching and followed different methodologies. Therefore, this comparison may have little meaning. However, it is worth mentioning some beliefs and results. According to Jansen et al. (1998) Web search queries contain fewer terms than other IR systems queries. Specifically, their study focused on user queries on the Web showed that on average, a query contained 2,35 terms. Similarly, a later study of Jansen (2000) indicated that approximately 93% of the Web queries contained between 0 and 4 terms. These findings are consistent with the outcomes of some other studies based on Web searches (Silverstein et al. 1999; Jansen et al. 2000) which concluded that standard Web search queries tend to be 2-3 terms in length.

It is also worth mentioning that Lau and Goh (2006) specified that when query length increases, the number of search terms used in a query could affect the provided results. Specifically, when query length increases there is a higher probability that users would encounter unsuccessful searches. The percentage of successful queries with one search term was double that of failed queries (15,9% versus 7,7% respectively). For queries with two terms, the success rate was 19,7% against a 14,4% failure rate. Barsky and Bar-Ilan (2005) provided some useful advice to users on how to search on

the Web regarding the number of keywords should be included in a query. They mentioned that the longer the query, the less probability that all the words will appear to be close to each other. Furthermore, when users are not familiar with the subject they are searching for, it would be better to use no long queries. Still, users who want to collect as much information as possible are recommended to use single terms and make a variety of similar queries, consisting of all possible synonyms. Users who look for precise answers are advised to create more focused queries.

Regarding the terms that users type in the searching field, Barsky and Bar-Ilan (2005) made a distinction between general queries and complex queries. General queries are consisted of some or all words extracted from the task description, while complex queries are consisted of words from the task description and of other keywords, proposed by users. In both exercises, students showed a preference on both queries; there were 85 general queries and 93 complex queries. They also analysed the use of phrases and concluded that many words within phrases were used in 15,29 percent of the general and 9,67 percent of the complex queries.

Finally, concerning the way that users refine their searches, research has shown so far that relatively few users specialise, generate or reformulate their searches (Lau and Horvitz, 1999). Druza and Dennis (1997) categorised 1040 queries into 11 different transformation types and found that most of the times users simply repeated a query that they have already submitted. A greater attempt on describing the terms that users type in the searching field was made by Suppes and Béziau (2004). According to them the relation between the terms is symmetric.

3 Methodology

3.1 System

For the purpose of this survey we used a z39.50 server for hosting selected bibliographic metadata records, approximately 14,400, from the database of the “Evonymos Ecological Library” (<http://www.evonymos.gr>). The participants used a z39.50 client to query the database. The system was customised to meet the needs of the experiment, namely we dismissed the Boolean operators feature, offered only one search field for submitting queries with the structure “words”, used only the “Subject” index and set a “Login area” for all participants to keep track of the logs. The truncation option was a default “right”.

Using a specific library system and collection, and not a web search engine, had two positive side effects; the system is providing a direct decision on relevance of the documents, by including them or not in the result set, while the result grading methods, like the web search engines, are more difficult to evaluate for relevance. Additionally, we were able to identify both the relevant documents found and those not found during a search, as we had a predefined set of documents.

3.2 Participants

The participants were undergraduate and postgraduate students of the Department of Archive and Library Sciences at the Ionian University in Corfu, Greece. For the 27 undergraduates the task took place under supervision during course attendance while the seven (7) postgraduates participated voluntarily at their own time without supervision. From the total of 34 participants 30 of them were female and four (4) of them were male.

3.3 Task

We provided the participants with a printed form of guidelines and specific tasks to follow in given order. We advised them to formulate queries in order to retrieve the more relevant documents for the given subjects. We set a maximum limit as to how many queries they could submit for each

question along with the term limit we set for each query. Briefly, the instructions we provided the participants with were to fill in the appropriate cells in the given forms by writing down the terms used to formulate a query. This should be done for every single query submitted to the system following the exact order of submission. Both the task and the questionnaire were in Greek language. Thus for the purpose of this paper, when necessary, we translated some data in English.

3.4 Questionnaire

The participants had to keep track of the queries they submitted in the database by filling in the form provided. In this form we included information on the nature of the database, the purpose of the experiment, some guidelines concerning the use of the system and an example of how to complete the form. Users had to complete the questionnaires whose responses we could also verify from the transaction log files, namely time and date of logging in and out of the system and their matriculation number, as well as the terms and queries submitted. The forms ended up with brief questionnaire on demographic data and opinion gathering. In order to avoid biased answers we also used transaction log files.

3.5 Transaction log files

The system kept transaction log files both from the server and the client traffic. Besides the responses we collected from the questionnaires we were additionally able to either confirm or reject certain aspects of them through the data from the log files.

3.6 Difficulties in research

The first difficulty we had to deal with was to decide which database we would use in order to carry out our research. We decided that we should use a database that covers a knowledge area that most people are familiar with and at the same time it offers options for more specific activities in future implementations. Therefore, we concluded that a database covering subjects related to *Ecology* and *Environmental issues* would be a reasonable solution.

The second difficulty was to customise the database in a way that we could collect valuable information about the searching behaviour of users. We were allowed to have access to the bibliographic metadata records of the Evonymos Ecological Library and make the appropriate customisation. The database was customised regarding the number of records that users had access to, but we extracted all those that could provide data of no use, meaning *noise*.

The third difficulty was to collect the questionnaires given to the postgraduate students. In contrast with the undergraduate students who were obliged to carry out the task and complete the questionnaire during laboratory hours, postgraduate students were volunteers. This means that we should contact them in order to do the experiment and complete the questionnaire.

Finally, the fourth problem was concerned with the transaction log files. Due to a technical problem a log file was corrupted and all information was lost. In this case for the particular respondent we considered only the responses from his questionnaire.

4 Results analysis

In order to analyse our results in terms of identifying the users' thinking behaviour while searching in a digital library we mainly categorised the terms submitted according to the studies of Rieh and Xie (2001; 2006). They examined the facets, subfacets and patterns of query reformulations with a focus on the semantic analysis of queries. The processing of the data was done manually. We present the results related with both the phase of searching and the phase of retrieving results. All

queries were examined manually to identify both the query formulation patterns and the typographical errors.

4.1 Definitions

As stated right above, we used certain subfacets identified at the studies of Rieh and Xie. Below we provide their definitions along with additional definitions that we used in order to categorise our raw data.

4.1.1 Definitions for Subfacets used

Specification: when users specify the meaning of the query by adding more terms or replacing terms with those that have more specific meaning.

Generalisation: when users generalise the meaning of the query by deleting terms or replacing terms with those that have more general meaning.

Replacement with synonyms: when users replace current terms with words that share similar meaning.

Parallel movement: when users do not narrow or broaden previous queries. The previous queries and the follow-up queries have partial overlap in meaning, or two queries are dealing with somewhat different aspects of one concept

Term variation: when users change the format of terms while still searching for the same topic. Variations include spelling out the abbreviation, adding a preposition (e.g., of, to, for), changing from the singular to the plural, or vice versa.

Error correction: when users change query terms in order to correct a typing or spelling error they have made.

4.1.2 Additional definitions

Term provided: a provided term from the description of a task.

Undefined: an inappropriate term for retrieving results for the given task; no apparent connection between the term used and the given task can be identified.

Term: an unbroken string of alphanumeric characters entered by a user (Spink et al. 2000). Terms that were considered as single terms included but not limited to whole words, truncated words and articles.

Unique term: every single term (case non-sensitive) entered by a user either in every session (unique term/ question) or throughout the task (unique term); the varying term could be modifications of any previous term or entirely new term.

Query: a term or a sequence of terms submitted to the system using the “Search” button.

Session: a query or a sequence of queries submitted to the system by a user in order to perform a given task

Failed search: a query that resulted with no hits.

Association: a set of nodes with links between them, meaning the use of notions associated with the input (Suppes & Béziau; 2004).

4.2 Analysis of Questionnaires

4.2.1 General results

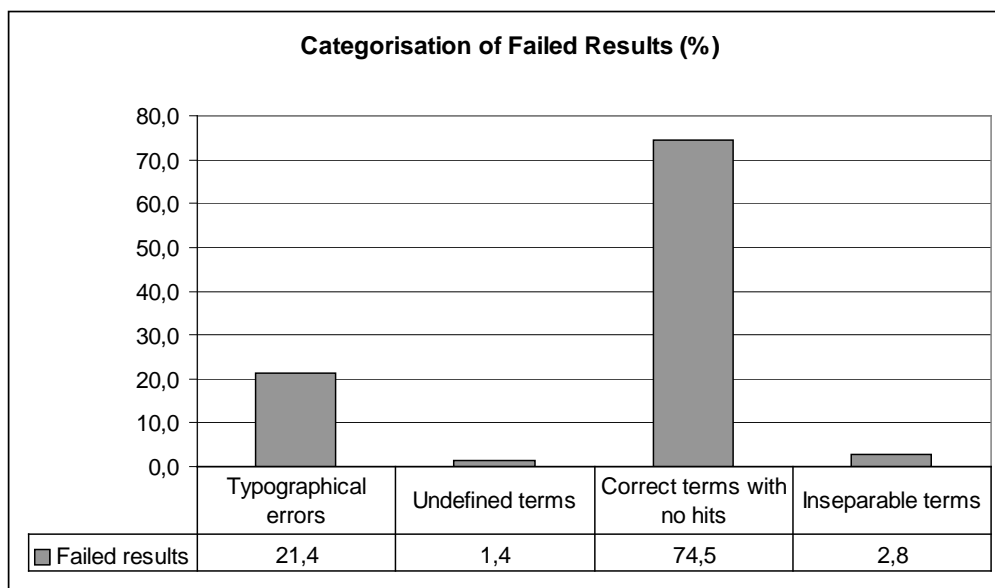
The participants had to find relevant documents for each of the following topics:

- Q 1: migration birds,
- Q 2: fruit-bearing trees,
- Q 3: protection of environment,
- Q 4: greenhouse effect,
- Q 5: alternative energy sources

For the first topic the participants could submit a maximum of two queries. For the second, third and fifth topics they could submit as much as four queries whilst for the “*greenhouse effect*” topic they could go for eight queries. The total number of queries submitted for the completion of the task was 605.

In the 605 queries, the total number of terms used was 913. An interesting comment from the results analysis is that only 178 of them were unique terms, approximately one fifth of the total terms used. All others were repeated either within the same question or within the whole task. These findings are provided in Figure 4.1.

Figure 4.1: Queries – Terms – Unique terms



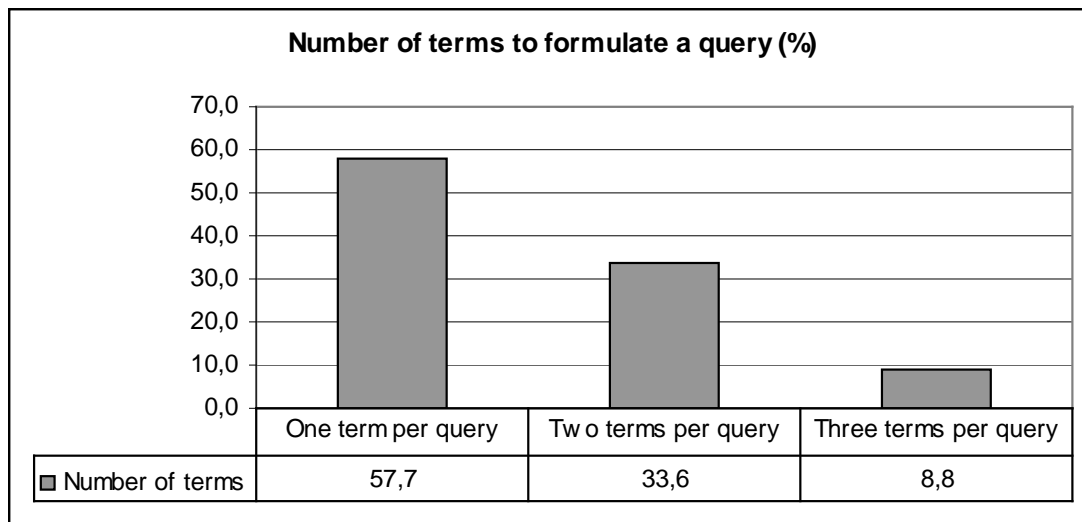
It is worth mentioning that the sum of *Unique terms* which occurs if adding all recorded *unique terms* per each question does not come in accord with the figure given previously for the *Unique terms* identified, namely 178. This is due to the fact that a *term* may be used once within a query but it may also appear as *unique term* within more than one of the given questions (Qs). Consequently, it is worth mentioning that 37 *unique terms* were repeated throughout the whole task. This figure is the occurring difference of the subtraction of the exact number of *Unique terms* identified across all queries (178) from the sum of *Unique terms/ question* (215). The 178 terms were used only once overall.

As we mentioned in the definitions section, the *unique terms* counted were case non-sensitive, meaning that the terms *Environment*, *environment* and *ENVIRONMENT* were counted as one.

Furthermore, since we focused our research on users' associations, truncated terms that were obvious parts of an identified term were also counted as one i.e. *environ*.

The participants could use between one and three terms for each query, not more than three. The majority of queries contained only one term – 57.7% of total queries included one term, while 33.6% of them consisted of two terms. Only 8.8% of total queries had three terms.

Figure 4.2: Number of terms to formulate a query (%)



4.2.2 Queries

Results of the questionnaires showed that queries were divided into seven main categories: specifications, generalisations, replacements with synonym, parallel movements, term variations, use of the term provided and undefined terms. The findings for each category are provided below. Specifications: 15,3% (93) of the total queries were specifications of previous terms or of terms provided for the task.

Generalisations: 15,2% (92) of the total queries were generalisations of previous terms.

Replacements with synonym: 3,14% (19) of the total queries were synonyms of previous terms or of terms provided for the task.

Parallel movements: 47,60% (288) of the total queries were related in a way with previous terms but were neither generalisations nor specifications nor any other of the categories identified. An example is the consecutive use of the words “swallow” and “stork” for the question on *Migratory birds*.

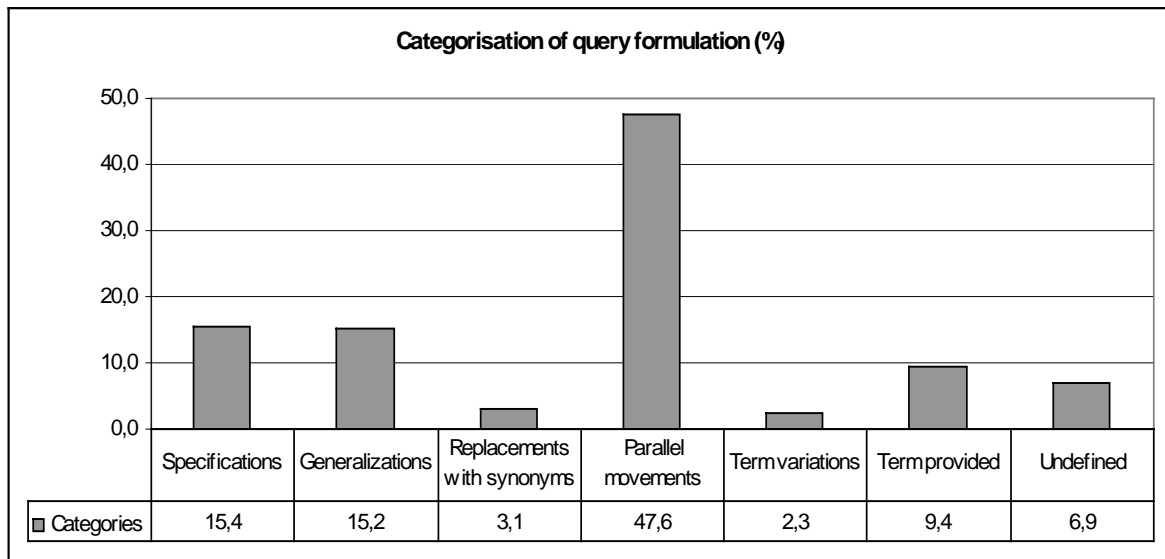
Term variations: 2,31% (14) of the total queries were variations of previous terms or of terms provided for the task.

Term provided: 33,9% (57) of the total queries were queries described in the task.

Undefined: 6,9% (42) of the total queries were queries with very broad meaning. Users were explicitly informed that the content of the database was about ecologic issues. Nevertheless, they used words with very broad meaning, such as *Ecolog*. Also, in this category we included terms which did not have an apparent connection to the given topic, such as *Biology* for Q 4 or terms which were too general in the specific context, such as *Nature* for Q 1.

According to the preceded analysis, the categorisation of queries is presented below in Figure 4.3.

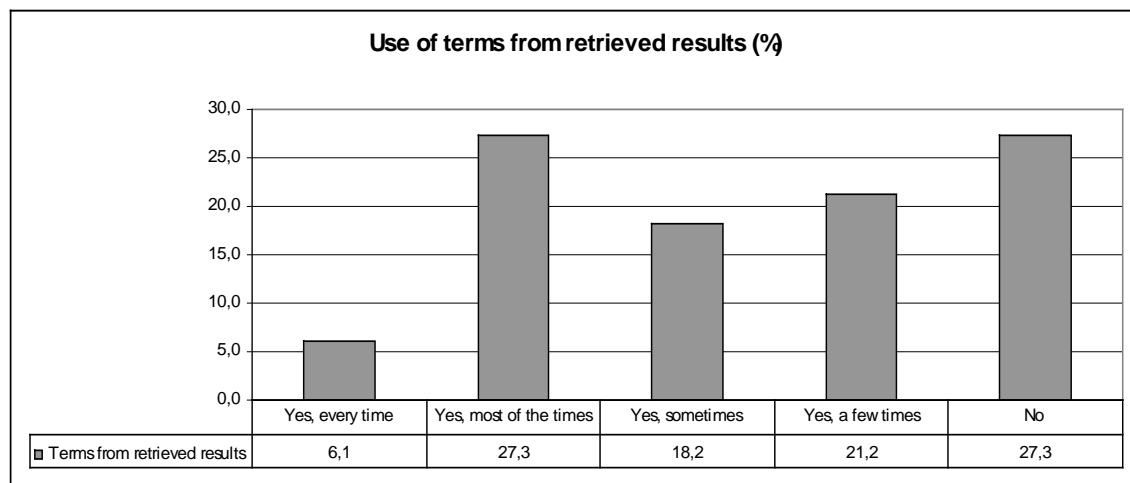
Figure 4.3: Categorisation of query formulation (%)



4.2.3 Query reformulation based on retrieved results

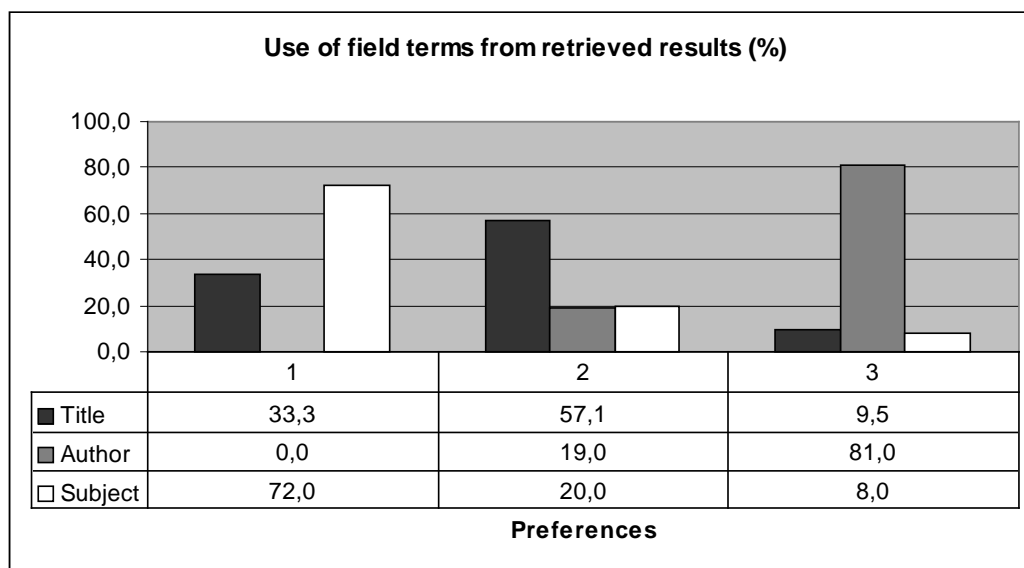
After the completion of the task we asked the participants to provide information as to whether they used or not terms identified within the retrieved result set in order to reformulate their queries. We received 33 responses which are distributed as shown in Figure 4.4.

Figure 4.4: Distribution of terms used from retrieved results (%)



From the participants' responses we, also, observed that the most commonly used field to provide ideas for query reformulation was the field of *Subject*. This was not only the most commonly used field (72%) but also the one that participants described as the most important to use by giving it the relative value (1 in Figure 4.5) in their responses. The fields of *Title* and *Author* were used equally in order to derive information for the formulation of a new query, although *Title* was chosen as second preference (2 in Figure 4.5) against the *Author* field which most of the participants ranked as their third choice (3 in Figure 4.5). A summary of the preferences of users is shown in Figure 4.5.

Figure 4.5: Distribution of terms used from retrieved results (%)



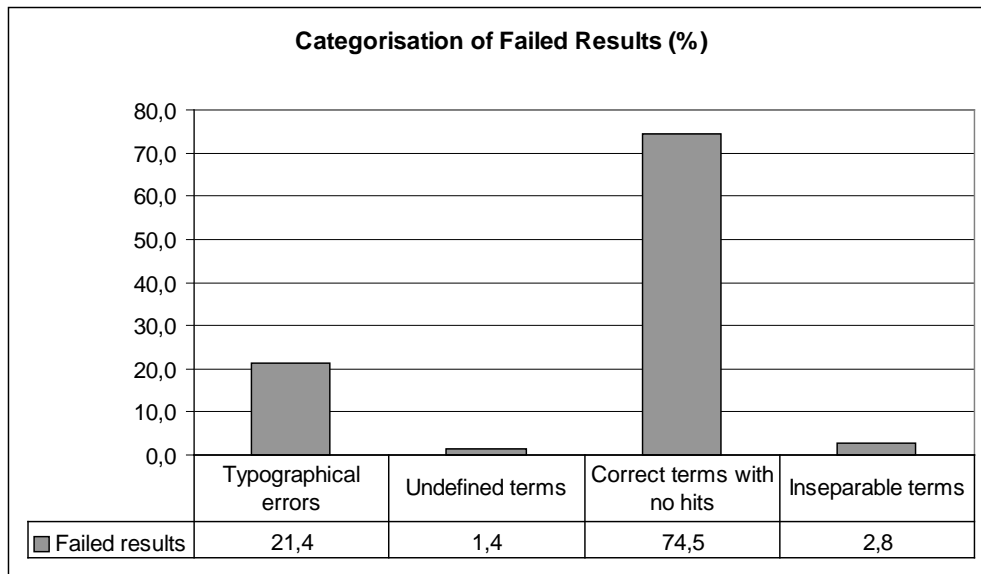
4.3 Analysis of Log files

Transaction logs described the use of the “Evonymos Ecological Library” online database by 33 users. These users made 548 queries in total, while 26,4% of these queries had zero results (failed queries). This means that due to the fact that the information system did not retrieve any relevant records, if users wanted to continue to use the system they should started their search from the beginning.

In our attempt to categorise these failed queries we organised them into four (4) main groups. The first category referred to the queries with typographical errors namely, substitutions, omissions, insertions and transpositions. These subcategories are described extensively in the articles of Brown (1988) and Hirst and Budanitsky (2003). The second category included the queries whose terms were inseparable. For example, a user typed in the searching field the terms *environment greenhouseeffect*, instead of *environment greenhouse effect*. In the third category, we grouped together all the undefined terms. This means that we included the terms that they did not have any particular meaning or were inexistent. And in the forth category, we included the queries whose terms did not have any spelling or grammar mistake nevertheless users did not retrieve any record.

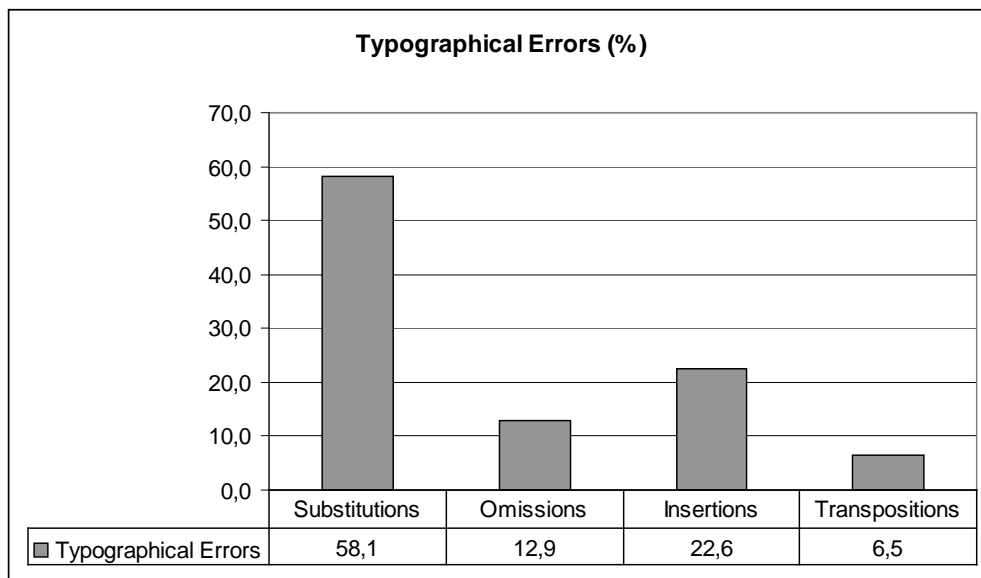
Proportionally most queries (74,5%) belonged to the last group, which implies that users typed one or more meaningful terms, but still they did not find any relevant record. The second most famous category was the typographical errors where 31 queries were typed incorrectly - 21,4% of the total number of failed queries (Figure 4.6).

Figure 4.6: Categorisation of Failed Results (%)



Regarding the typographical errors, results showed that the most regular error for users was to replace a character with another character (substitutions); 58,1% of the total number of typographical errors belonged to this subcategory. For example users typed *cleening*, instead of *cleaning*. The next most typical error was to insert additional character(s) to a term (22,6%), while the least famous error was to type two or more characters of a term not in the right order -only 6,5% of the failed queries included in this group (Figure 4.7).

Figure 4.7: Typographical Errors (%)



5 Conclusions – Future work

Previous literature has shown that it is difficult to measure users' search behaviour because it varies according to certain circumstances. Thus, it is not easy to setup appropriate experiments covering and controlling - if not all - at least an adequate proportion of factors affecting the users' search behaviour. Based on relative studies and former experience on the field we attempted to put into scheme some of these aspects. In our study we attempted to retrieve some initial results on user search behaviour. The results are summarised below.

In terms of query length, although we set limitations to the participants regarding the number of terms they could use to formulate a query, our results seem to be consistent with the outcomes of previous studies. According to these findings, users mainly type in the searching field one term per query. Our study showed that 57,7% of the total queries contained only one term. Two-term queries appeared in 33,6% of the cases and only a limited 8,8% of queries contained three terms.

An important observation is the number of *Unique terms* that respondents used. The number of total terms used for all queries was 913. Among them only 178 i.e. one fifth were actually *unique*. This fact is encouraging for our research because it shows that users' queries can actually be modeled, at least under certain circumstances, and lead to specific patterns.

Furthermore, we observed that users have equal chances to direct their queries in either a more specified (15,4%) or generalised (15,2%) term, whereas their choice of submitting parallel terms outweighed all other strategies (47,6%). This is comparably very close to the results presented in the study of Rieh and Xie (2006) who specified that 51.4% of movements turned out to be parallel movements.

Also, an interesting finding is the limited use of *Term variations* as a search strategy, only 2,3% of the queries involved use of this pattern. It seems that users tended to alter significantly the terms used rather than try submitting a different form of them, such as the same term truncated or the changing between singular and plural. The participants also made little use (3,1%) of *Replacing with synonym* term in order to reformulate a query.

Indicative of both the need to easily retrieve the required information and the associations the users make was the monitored use of the *Term provided* in the task description. In 9,4% of the queries the participants used the given term or a slightly altered form of it. Specifically, the majority of such instances occurred in the very first query.

The responses we received concerning the use of terms from the retrieved data sets is a strong element as to what they would find helpful in the process of searching. A percentage of 72,8% declared that respondents used a term from the retrieved results. Variations include the frequency with which they recorded this use, namely others for every query, for most of the queries, for some queries or for few. Furthermore, what occurred from the participants' responses was that the most favoured field for providing new ideas for queries was the *Subject* field with a percentage of 72%.

Regarding the type of failed queries, results showed that typographical errors, and specifically the substitution of a character with another appeared to be the main reason. The cases where the search term was not contained within the searchable data were excluded from this remark.

In the context of further confirmation of the results we have already planned to repeat the task adding more queries and collecting data from additional users.

Furthermore, in our plans is to verify certain factors which we have evidence that influence the search behaviour of users, such as using different terms for describing the same subject and asking the participants to retrieve relevant results. Evidence collected for the *Term provided* category shows that users would not make the same selection of words if they were provided with different initial terms.

References

1. Barsky E. and Bar-Ilan J. 2005. From the search problem through query formulation to results on the web. In *Online Information Review* 29(1): 75-89.

2. Brown A. (1988). A Singaporean corpus of misspellings: analysis and implications. In *Journal of the Simplified Spelling Society* 3. Available at: <http://www.spellingsociety.org/journals/j9/singapore.php>, (Accessed 23 October 2007).
3. Buzikashvili N. (2005). Information searching behavior: between two principles. In: F. Crestani and I. Ruthven (eds), *CoLIS 2005* 3507: 79-95.
4. Druza P. D. and Dennis S. 1997. Query reformulation on the Internet: empirical data and the Hyperindex search engine. In *5th RIAO Conference*. Available at: <http://citeseer.ist.psu.edu/cache/papers/cs/2109/http:zSzzSzwww.dstc.edu.auzSzRDUzSzreportszSzraio97.pdf/pd97query.pdf>, (Accessed 15 September 2007).
5. Hirst G. and Budanitsky A. 2003. Correcting real-world spelling errors by restoring lexical cohesion. Available at: <http://ftp.cs.toronto.edu/pub/gh/Hirst+Budanitsky-2005ms.pdf>, (Accessed 30 October 2007).
6. Jansen B. J. 2000. An investigation into the use of simple queries on Web IR systems. In *Information Research: An Electronic Journal* 6(1). Available at: <http://jimjansen.tripod.com/academic/pubs/ir2000/ir2000.html>, (Accessed 15 September 2007).
7. Jansen B. J., Spink A. and Saracevic T. 2000. Real life, real users, and real needs: a study and analysis of user queries on the Web. In *Information Processing and Management* 36: 207-227.
8. Jansen B. J., Spink A., Bateman J. and Saracevic T. 1998. Real life information retrieval: a study of user queries on the web. In *SIGIR Forum* 32(1): 5-17.
9. Joachims T., Granka L., Pan B., Hembrooke H., Radlinski F. and Gay G. 2007. Evaluating the accuracy of implicit feedback from clicks a query reformulations in web search. In *ACM Transactions on Information Systems* 25(2): 1-27.
10. Jones S., Cunningham S., McNab R. and Boddie S. 2000. A transaction log analysis of a digital library. In *International Journal Digital Libraries* 3(2): 152-169.
11. Lau E. P. and Goh D. H.-L. 2006. In search of query patterns: a case study of a university OPAC. In *Information Processing and Management* 42: 1316-1329.
12. Lau T. and Horvitz E. 1999. Patterns of search: analyzing and modelling web query refinement. In *Proceedings of the 7th International Conference on User Modeling* : 119-128.
13. Rieh S. Y. and Xie H. 2001. Patterns and sequences of multiple query reformulations in web searching: a preliminary study. In *Proceedings of the 64th ASIST annual meeting* 38: 246-255.
14. Rieh S. Y. and Xie H. 2006. Analysis of multiple query reformulations on the web: the interactive information retrieval context. In *Information Processing and Management* 42: 751-768.
15. Suppes P. and Beziau J.-Y. 2004. Semantic computations of truth based on associations already learned. In *Journal of Applied Logic* 2: 457-467.
16. Silverstein C., Henzinger M., Marais H. and Moricz M. 1999. Analysis of a very large AltaVista query log. In *SIGIR Forum* 33(1): 6-12.