Exploring Query Formulation and Reformulation: A Preliminary Study to Map Users' Search Behaviour

Anna Mastora[1], Maria Monopoli[2], and Sarantos Kapidakis[1]

[1] Laboratory on Digital Libraries & Electronic Publishing, Department of Archive & Library Sciences, Ionian University
72, Ioannou Theotoki Str., Corfu, Greece, 49100
[2] Library Section, Economic Research Department, Bank of Greece
21 Panepistimiou Str., Athens 102 50, Greece
{mastora,sarantos}@ionio.gr, mariamonopoli@hotmail.com

Abstract. This study aims to investigate the query formulation and reformulation patterns such as generalisations, specifications, parallel movements and replacements with synonyms within the search procedure. Results showed that users reformulated their queries by using terms contained in the retrieved results while in the query reformulation process they mainly used terms with parallel meanings. Participants used equally either more specific or more general terms for follow-up queries. Finally, the study revealed that a high proportion of same terms were used instead of unique ones; half of them were included in the Eurovoc thesaurus.

Keywords: Query formulation, Query reformulation, Search behaviour, Search patterns, Query length.

1 Introduction

Kuhlthau [1] has identified six phases of an information search process (ISP), namely initiation, selection, exploration, formulation, collection and presentation; our research is focused on formulation. During this stage users try to formulate a perspective focused on the needed information. This implies the associations which users make and consequently the use of words as query terms.

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

- How users formulate and reformulate their queries
- Whether users use the term provided in the description of the task in order to formulate or reformulate their queries
- Whether users use terms included in the retrieved results for follow-up queries
- How many terms users type in the searching field
- How many unique terms users actually use and how many of them are included in the Eurovoc Thesaurus (http://europa.eu/eurovoc/)

B. Christensen-Dalsgaard et al. (Eds.): ECDL 2008, LNCS 5173, pp. 427–430, 2008.

2 Methodology

We identified two query stages: "formulation" which is the initial stage in which the search strategy is constructed and the following "reformulation" stage in which the initial stage is modified either manually or system-assisted.

For the purpose of this study we used approx. 14,400 bibliographic metadata records, from the "Evonymos Ecological Library" (http://www.evonymos.gr). The system was customised to meet the needs of the experiment, namely we dismissed the Boolean operators, offered only the "Subject" index for submitting queries with the structure "words" and set a "Login area" for all participants to keep track of the log files. There was also a truncation option activated, set as a default "right". These adjustments were considered necessary in terms of simplifying the search task and consequently drawing the participants' attention to select search terms and not to concentrate on the system's functionality.

The participants were 27 undergraduate students who took part under supervision in the Department's laboratories and 21 postgraduate students who participated voluntarily; forty of them were female and eight of them male.

Users had to find relevant documents for each of the following topics: *Migratory birds* (Q1), *Fruit trees* (Q2), *Environmental protection* (Q3), *Greenhouse effect* (Q4) and *Alternative energy sources* (Q5) and fill in the accompanying questionnaire. The participants had to keep track of their queries by filling in some given forms which contained introductory information about the database, guidelines concerning the execution of the task and questions on demographic data. Both the task and the questionnaire were in Greek, thus, for the purpose of this paper, when necessary, we translated some data in English. In order to avoid bias, we also kept transaction log files. However, only data from the questionnaires is currently presented.

In order to motivate participants to put at least some effort into the task, we set a maximum and/or a minimum limit as to how many queries they could submit for each question. We knew in advance which subjects could stand for exhaustive queries and which could not. Additionally, the database was customised by excluding all records contained words in English with a particular concern on excluding records containing Subjects in English in order not to distract the users' attention and allow them to focus on Greek terms. We, also, excluded records on Literature because it would return misleading results. There was also a word limit, i.e. one to three, for the formulation of each term.

We identified few constraints in the task's implementation. The first was associated with the database selection. We thought that a reasonable solution was to use a database which covers a knowledge area that most people are familiar with, like *Environmental issues*. A second difficulty was to customise the database in a way that we could collect valuable information and avoid noise-data. Finally, a third difficulty was to collect the questionnaires given to the postgraduates, even though their motivation in participating was voluntary.

3 Results Analysis

We mainly categorised the terms submitted according to Rieh and Xie [2]. All queries were examined manually to identify both the query formulation and reformulation patterns. The attribution of characterisations to the terms was made according to the Eurovoc thesaurus (version 4.2). Few terms that did not exist in the thesaurus were characterised according to the judgement of the authors.

Right below we provide the definitions of the study of Rieh and Xie [2] concerning the identified query patterns.

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

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

Replacement with synonyms: replace current terms with terms that share similar meaning

Parallel movement: 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

We additionally used our own definitions in order to meet the needs of our study as follows:

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

Error: an inexistent term according to Babiniotis Dictionary [3]

Undefined: an inappropriate term for describing 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

Query: a term or a sequence of terms submitted to the system

3.1 Analysis and Conclusions

Concerning the users' behaviour for query formulation we observed that users in 41.6% of the cases used the *term provided* in order to start their search task. Furthermore, we observed that users had somewhat equal chances to submit an either more *specified* (25.2%) or *generalised* (18.1%) term, whereas using a *parallel term* fell to 8.4%. Even lesser users (5.9%) preferred the use of a *synonym* term for formulating their first query.

When it comes to reformulation, users showed a preference on *parallel movements*, i.e. 47.6% of total reformulations belonged in this category. Their second choice was either *generalisations* or *specifications* of terms with 20.0% and 20.3%, respectively. Similar to the query formulation, participants made little use (5.3%) of *replacing a term with a synonym* to reformulate a query. Compared to the excess use of the *term provided* to formulate a query, in query reformulations only 5.3% of queries contained the term provided. If taking into account all the tasks, though, the use of the *term provided* represented the 15.5% of all submitted queries.

Regarding the use of terms from the retrieved results, outcomes showed that 74.5% of users admitted that they used a term from the retrieved results. This result is a valuable element to what users would find helpful in the process of searching.

In terms of query length, we set limitations to the participants regarding the number of terms they could use to formulate a query. The system performs a default "and" for input terms and we considered that submitting more terms in the search field would probably return zero results. Our study showed that 48.9% of total queries contained only one term, 39.6% of the cases had two-term queries and only 11.5% of queries contained three terms.

Concerning the findings of the *Unique terms* we found that approximately one seventh (205) of total terms (1372) used in all queries, were actually unique terms. All others were repeated either within the same question or within the whole task, although the given subjects were not strictly related in meaning. It is worth mentioning that if we add all recorded unique terms per question, the occurring sum is greater than 205; it is actually 251. This is due to the fact that a term was used once within a query but it also appeared as *unique term* within more than one of the given questions. An additional noticeable remark is that only 46 *unique terms* were repeated throughout the whole task. This number is the occurring difference of the subtraction of the exact number of *unique terms* identified across all queries (205) from the sum of *unique terms/ question* (251).

Concluding on the results of this study, we mapped 124 of the 205 *unique terms* used to terms of the Eurovoc Thesaurus.

4 Future Work

Our future concern is to further monitor the users' behaviour regarding the terms they selected for formulating their queries and identify possible factors that affect these selections. In this context it would, also, be interesting to associate search success with certain reformulation patterns. Additionally, the significant use of terms from the Eurovoc thesaurus could be considered as a starting point to deal with semantic heterogeneity. Finally, we will add findings from additional results and verify aspects of others using the transaction log files which have not been processed yet.

References

1. Kuhlthau, C.: Inside the search process: information seeking from the user's perspective. JASIS 42(5), 361–371 (1991)
2. Rieh, S.Y., Xie, H.: Analysis of multiple query reformulations on the web: the interactive information retrieval context. Information Processing and Management 42, 751–768 (2006)
3. Babiniotis, G.: Dictionary of Modern Greek language: with comments for the correct usage of words. Lexicology Centre, Athens (1998) (in Greek)