Term Selection Patterns for Formulating Queries: a User Study Focused on Term Semantics

Anna Mastora
Laboratory on Digital Libraries & Electronic Publishing, Department of Archive & Library Sciences, Ionian University
mastora@ionio.gr

Maria Monopoli
Library Section, Economic Research Department, Bank of Greece
mariamonopoli@hotmail.com

Sarantos Kapidakis
Laboratory on Digital Libraries & Electronic Publishing, Department of Archive & Library Sciences, Ionian University
sarantos@ionio.gr

Abstract

This study aims at investigating query formulation and reformulation patterns within the search procedure focusing mainly on the semantic aspect of submitted queries. Results identified generalisations, specifications, parallel movements and replacements with synonyms; the parallel movement came first in the users' preferences for reformulating a query. Additionally this study revealed that users used terms identified within the retrieved result sets for reformulating their queries. Users also used provided terms for formulating their queries while some terms that users typed in either for formulating or reformulating their queries were repeated even in queries with no strictly related subject of interest. Regarding the length of the queries outcomes showed that users, consistently to previous studies, typed in mostly one word per query. Furthermore, approximately half of the terms users chose for carrying out their searches were actually included in the Eurovoc thesaurus.

1. Introduction

This study aims at investigating common patterns concerning the formulation of users' queries. It is part of a broader research undertaken at the Laboratory on Digital Libraries & Electronic Publishing, Ionian University. Its ultimate goal is to identify users' thinking behaviour while they are formulating their queries within the process of seeking information in a digital library. Such data is useful to designing digital systems as well as data structures according to target's necessities [1].

From a user's perspective, any digital library system is only as good as its retrieval efficiency. In order to achieve efficient and precise retrieval, it is necessary to encode the semantics of a resource [2]. As it is quoted in [3] we must also make available in a machine accessible way the important concepts that are discussed in the documents, the relation of these concepts with those in other documents, relating these concepts to the general background knowledge. This implies having the knowledge of the meaning of terms within the context they are sought by users.

In order to be more precise, this study is primarily focused on the semantic analysis of queries. This analysis attempts to investigate users' thinking behaviour (i.e. whether users narrow or broaden the meaning of predefined search terms or use terms that have somewhat similar meaning). Additionally, it attempts to shed a light on the following issues:

- Whether users use the term provided in the description of the task in order to formulate or reformulate their queries
- Whether users reformulate their queries by using terms included in the retrieved results
- How many terms users type in the searching field
- How many unique terms users actually use and how many of these terms are included in the Eurovoc Thesaurus [4].

2. Related literature

There are two query formulation stages: formulation and reformulation [5]. Previous attempts on tracing the way users retrieve data from the web revealed that approximately half of all Web users reformulate their initial queries – 52% of the users in the 1997 Excite data set and 45% of the users in the 2001 Excite data set [6]. A quite interesting and optimistic conclusion is that users learn how to formulate better queries during their searching process.

In [7] is 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 of searching and have followed different methodologies. Therefore, this comparison may have little meaning. However, it is worth mentioning some beliefs and results. According to [7] 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 [8] indicated that approximately 93% of the Web queries contained between 0 and 4 terms. These findings are, also, consistent with the outcomes of studies [7], [9] based on Web searches which concluded that standard Web search queries tend to be 2-3 terms in length.

It is worth mentioning that in [10] is 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. In [11] is mentioned that the longer the query, the smaller the probability exists that all the terms to appear would be close to each other. Furthermore, when users are not familiar with the subject they are searching for, it would be better to use short queries.

Finally, in [12] 1040 queries are categorised into 11 different transformation types and what is found is that most of the times users simply repeat a query that they have already submitted. A greater attempt on describing the terms users type in the searching field was made in [12].

3. Methodology

This section describes the concepts and methods used for investigating the users' patterns in query formulation and reformulation.

3.1. System

For the purpose of this study we used a z39.50 server to host selected bibliographic metadata records, approximately 14,400, from the database of the "Evonymos Ecological Library" (http://www.evonymos.gr). The system was customised with the intent to simplify the search process drawing, thus, the participants' attention away from the system's functionality and allowing them to concentrate on the choice of terms. In particular, we offered only one searchable field, the "Subject" field, and dismissed the Boolean operators feature trying to keep the search process as simple as possible. There was also a right truncation feature and we finally, set a "Login area" to keep track of the log files.

3.2. Participants

The participants were 27 undergraduate and 21 postgraduate students of the Department of Archive and Library Science. From the total of 48 participants 40 of them were female and eight (8) of them were male. We selected these participants, because we thought that it would be more likely to find and persuade them to participate in future phases of the experiment (e.g. interviews) if necessary.

3.3. Task

We provided the participants with written guidelines and specific search tasks to carry out in a predefined order. For the 27 undergraduates the tasks took place under supervision at the Department of Archive and Library Science laboratories. The 21 graduate students participated voluntarily.

The participants 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).

In order to further motivate the students concerning the number of queries they would submit, we set both a maximum and a minimum limit as to how many queries they could submit for each question. These limits were set firstly according to the number of queries which could actually be produced concerning the data collection and secondly, according to the quality of results they would return. We knew in advance which subjects could stand for exhaustive queries and which could not. The database was customised regarding the characteristics of records that users had access to. Namely we excluded all records containing words in English with a particular concern in eliminating records containing Subjects in English. We, also, excluded Literature content because it would return misleading results concerning the semantics of the selected terms. There was also a word limit, i.e. one to three, for the formulation of each term. Participants were informed in advance that it would take approximately 20 minutes to complete the tasks nevertheless they could exceed this time limitation.

The participants had to keep track of the queries they submitted in the database by filling in the given form. In this form we included some introductory information about the 'Evonymos' database and the purpose of the experiment, some guidelines

concerning the use of the system and an example of how to complete the form. Also, the form included a brief questionnaire asking from participants to specify some demographic data and give their opinion on certain issues. Both the task and the questionnaire were in the users' native language, Greek. Thus, for the purpose of this paper, when necessary, we translated some data in English. In order to avoid biased answers we also used transaction log files which will be the focus of a future study.

4. Results analysis

In order to analyse our results in terms of identifying the users' thinking behaviour while formulating and reformulating queries we mainly categorised the terms submitted according to Rieh and Xie [15]. They examined the facets, subfacets and patterns of query reformulations with a focus on the semantic analysis of queries. In the current study we used their identified subfacets of the "Content" facet. All queries were examined manually to identify both the query formulation and reformulation patterns.

The characterization of the terms was made according to the Eurovoc thesaurus (version 4.2), which was used as the reference concept hierarchy in our approach. According to the concept of the question, we used the related section of the thesaurus, e.g. for *Alternative energy sources* (Q5) we used the "Energy" section. Each word or sequence of words was treated in the semantic context it belonged and not as a general term. For example, in the context of the question about the *Greenhouse effect* (Q4) a search for *greenhouse,* according to the applying definition, was characterised as a *generalisation* and not as an irrelevant term as it could be within a different context. The terms that did not exist in the thesaurus were characterised according to the judgement of the authors.

4.1. Definitions

Right below we provide the definitions of Rieh and Xie for query formulation patterns as identified in their study [15].

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

Generalisation: users 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: 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

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 the Dictionary of Modern Greek language by George Babiniotis, the 1998 edition [15].

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

4.2. Categorisation of term patterns for query formulation

Formulation is the initial stage in which the search strategy is structured [5]. We categorized the results of users' formulations into seven main categories: generalisations, specifications, parallel movements, replacements with synonym, errors, undefined terms and use of the term(s) provided. The findings for each category are presented below and a summary is available in Figure 1.

Generalisations: 18.1% (43) of first queries were generalisations of the provided term.

Specifications: 25.2% (60) of first queries were specifications of the provided term.

Parallel movements: 8.4% (20) of first queries were related in a way with the given term but were neither generalisations nor specifications nor any other of the categories identified. An example is the use of terms *environmental problems* in a first query for the question on *Environmental protection.*

Replacements with synonym: 5.9% (14) of first queries were synonyms of the term provided in the description of the task.

Errors: no error was identified in the selection of terms for the first queries of the task.

Undefined terms: 0.8% (2) of first queries was queries which could be considered irrelevant to the specific task. Terms in English were counted in this category as well, since, according to the description of the task, participants were asked to use terms in Greek only.

Term(s) provided: 41.6% (99) of first queries contained the term(s) provided in the description of the task.

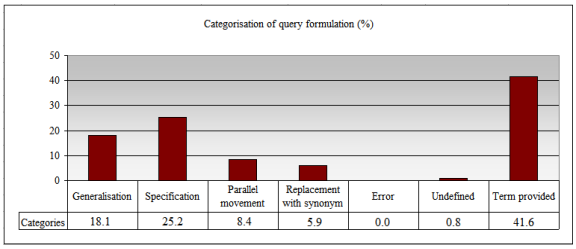

Fig 1. This figure shows the categorisation of query formulation patterns, i.e. the characterisation of all first queries submitted for each question (Q) of the task (%).

4.3. Categorisation of term patterns for query reformulation

Reformulation is the stage during which the initial stage is adjusted manually or with the assistance of a system [5]. In our case, outcomes of the manual adjustment of queries provided the following data concerning the query reformulation term patterns, a summary of which is presented in Figure 2.

Generalisations: 20.0% (121) of the reformulated queries were generalisations of the previous term.

Specifications: 20.3% (123) of the reformulated queries were specifications of the previous term.

Parallel movements: 47.6% (288) of the reformulated queries were related in a way with the previous term but were neither generalisations nor specifications nor any other of the categories identified.

Replacements with synonym: 5.3% (32) of the reformulated queries were synonyms of the term provided in the description of the task.

Errors: 0.8% (5) of the reformulated queries was errors, meaning that inexistent terms were used.

Undefined terms: 6.0% (36) of the reformulated queries were queries which could be considered irrelevant to the specific task. Terms in English were counted in this category as well, since, according to the description of the task, participants were asked to use terms in Greek only. Another example of such characterisation is the use of the word *insects* for the query on *Migratory birds.*

In cases where the characterisation of a term was *Error* or *Undefined*, the comparison was done with the first valid preceded term.

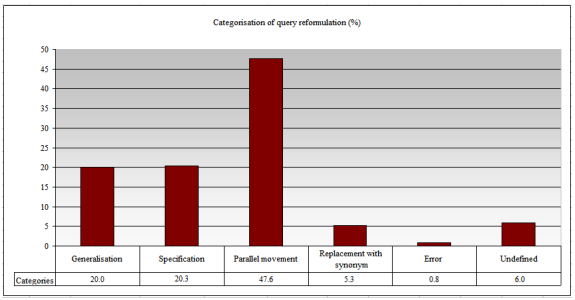

Fig. 2. This figure shows the categorisation of query reformulation patterns, i.e. the characterisation of all queries submitted for each question (Q) of the task except from the first ones (%).

4.4. Additional findings of queries patterns

In addition to the results referring to the way that users formulate and reformulate their queries, this research provided some valuable data about the selection and use of terms by users.

Use of Term(s) Provided. Regarding the use of the *term(s) provided* for formulating their queries, results showed that 41.6% of the terms were either the actual terms given in the task description or the terms with the use of the truncation feature. Users preferred to use mostly other terms and not the term provided although this does not exclude their returning to the use of this term for a query reformulation later in the search process.

Concerning the use of the *term(s) provided* for reformulating a query, we examined its frequency separately because this term could, also, belong to one of the other identified categories for query reformulation, for example it could also be either a specification or a generalisation. Results showed that 5.3% of subsequent queries consisted of the *term(s) provided*.

Examining the selection of terms during all queries, 15.5% of the terms used were *term(s) provided* in the task description.

Use of Terms from Retrieved Results. After the completion of the tasks we asked participants to provide information as to whether they had used or not terms identified within the retrieved result sets for reformulating their queries. Their responses are summarised as follows: 4.3% of the respondents specified that they chose a term from the retrieved result sets *every time*, 27.7% *most of the times*, 25.5% *sometimes*, 21.3% *a few times* whilst 21.3% stated that they used no terms from the retrieved result sets. These results will be further checked using the transaction log files.

Query length, Unique terms and Terms from Eurovoc. The total number of queries (formulations and reformulations) submitted for the completion of the tasks was 843. In most cases (57.7%) users typed in only one (1) term in the search area before submitting their query. In 33.6% of the cases users typed in two (2) terms while only 8.8% of the queries contained three (3) terms. The system performs a default “and” for input terms so we considered setting this limitation to prevent failed queries deriving from multiple term combination during searching. Within the 843 queries, the total number of terms used was 1372. An interesting comment

from the results analysis is that only 205 of them were unique terms.

We, finally, measured how many of the query terms exist in the Eurovoc thesaurus. Our findings showed that considering the total number (205) of *Unique terms* used, 124 of them are actually included in the thesaurus, i.e. more than half of them.

5. Conclusions

In our study we attempted to examine the users' term selection and to analyse their search behaviour during query formulation and reformulation. The results are summarised below.

In our study in 41.6% of the cases, users used the *term provided* in the task description in order to start their search tasks. Furthermore, we observed that users had somewhat equal chances to direct the formulation of a query in to either a more *specified* (25.2%) or *generalised* (18.1%) term, whereas their choice of submitting *parallel terms* was less frequent (8.4%). Even less users, 5.9%, preferred to use a *synonym* of the term provided in order to formulate their first query.

The occurring results of the query reformulation patterns showed that users preferred *parallel movements* for the reformulation of their queries, i.e. 47.6% of total reformulations belonged in this category. This is close enough to the results presented in [12] in which is specified that 51.4% of movements turned out to be parallel movements. Following, 20.0% and 20.3% of the queries were *generalisations* and *specifications* of previous terms, respectively. As well as in the query formulation, participants also made little use (5.3%) of *replacing a term with a synonym* to reformulate a query meaning that they most certainly did not retrieve valuable information due to probable lack of either domain or language knowledge or even both.

We expected that users would find either helpful or easier to start their search strategy by submitting the term provided for at least formulating their first query. However, although we observed that users are not much creative in selecting original terms for formulating their queries, less than half of them proceeded as expected. Compared to the excess use of the *term(s) provided* to formulate a query, on the other hand, in query reformulations only a percentage of 5.3% of queries contained the *term(s) provided* or a slightly altered form of it. If taking into account the overall process, though, the use of the *term(s) provided* were 15.5% of all queries submitted.

The responses we received concerning the use of terms from the retrieved data sets, at least in the context of searching within known items, is a strong element as to what users would find helpful in the process of searching. A percentage of 74.5% admitted that they used a term from the retrieved results. Variations constitute the frequency with which they recorded this use, namely *for every query*, *for most queries*, *for some queries* or *for few*.

In terms of query length and since using terms from retrieved results is such a common strategy, the users' tendency of using mostly one word per term could, in some cases, lead to retrieving misleading or even irrelevant results. Our study showed that 48.9% of total queries contained only one term. Two-term queries appeared in 39.6% of the cases and only a limited 11.5% of queries contained three terms. The results concerning the query length have considered all the queries submitted during the tasks. Further analysis of query length has shown that the tendency is, also, confirmed if formulation and reformulation patterns are considered separately. To be more specific, concerning formulation query length 42.9% of the queries contained one term, 42.0% contained two terms and 15.1% contained three terms. Finally, concerning reformulation query length 51.2% of the queries contained one term, 38.7% contained two terms and only 10.1% contained three terms. Furthermore, the limitations we had set to the participants regarding the number of the terms they could use to formulate a query, does not seem to lead to biased answers since the outcome of our study is consistent with the outcomes of previous studies as it is mentioned previously.

Concerning the findings of *Unique terms* we found that approximately one seventh (205) of total terms (1372) used across all queries were actually unique terms. All others were repeated either within the same question or within all tasks. The subjects that users had to retrieve data for were not so strictly related to expect this outcome. This, probably, shows a tendency of users' preference in common terms for formulating queries within a subject domain regardless of the specificity of the subject they are looking for. According to this finding we can suggest that such users would be more easily and adequately satisfied by a recommendation system compared to more demanding users.

It is worth mentioning that if adding all recorded unique terms per each question, the occurring sum is greater than 205, actually 251. 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). An additional noticeable remark is that only 46 *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 (205) from the sum of *unique terms/ question* (251).

Concluding on the results of the present study, we mapped 124 of the 205 *unique terms* used to formulate and reformulate queries to a term in the Eurovoc Thesaurus. The remaining half were terms

not included in the thesaurus, thus, leading us to the suggestion that using the thesaurus for semantically relating the terms of the users' queries could be a good starting point, though it should, also, be enriched and supported by other means, too, perhaps such as an ontology.

6. Future work

Our future concern is to further monitor the behaviour of users in terms of query formulation and identify the factors which affect the users' choice of terms. The significant use of terms which were identified in the Eurovoc thesaurus could be considered as a motivation for investigating whether such tools the purpose of which is to organise Knowledge, e.g. thesauri and ontologies, could effectively support the query formulation process by offering semantically related reformulation suggestions and to which extent.

Furthermore, we will investigate the behaviour of users during the reformulation of queries associated to both successful and unsuccessful queries.

Finally, we will add findings from additional results and verify aspects of others using the transaction log files which have not been processed yet.

7. References

[1] S.M. Ferreira and D.N. Pithan, "Usability of digital libraries: a study based on the areas of information science and human-computer-interaction", *OCLC Systems and Services*, 2005, vol. 21(4), pp. 311-323.

[2] A.R.D. Prasad and D.P. Madalli, "Faceted infrastructure for semantic digital libraries", *Library Review*, 2008, vol. 57(3), pp. 225-234.

[3] J. Krause, "Semantic heterogeneity: comparing new semantic web approaches with those of digital libraries", *Library Review*, 2008, vol. 57(3), pp.235-248.

[4] Eurovoc thesaurus, Edition 4.2, ISSN: 1725-4299, http://europa.eu/eurovoc/

[5] E.N. Efthimiadis, "Query expansion", *ARIST*, 1996, vol. 31, pp. 121-187.

[6] A. Spink, B.J. Jansen, D. Wolfram and T. Sarasevic, "From E-sex to E-commerce: Web search changes", *IEEE Computer*, 2002, vol. 35(3), pp. 133-135.

[7] B.J. Jansen, A. Spink, J. Bateman and T. Saracevic, "Real life information retrieval: a study of user queries on the web", *SIGIR Forum*, 1998, vol. 32(1), pp. 5-17.

[8] B.J. Jansen, "An investigation into the use of simple queries on Web IR systems", *Information Research*: *An Electronic Journal*, 2000, vol. 6(1).

[9] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz, "Analysis of a very large AltaVista query log", *SIGIR Forum*, 1999, vol. 33(1), pp. 6-12.

[10] E.P. Lau and D.H.-L. Goh, "In search of query patterns: a case study of a university OPAC" *Information Processing and Management*, 2006, vol. 42, pp. 1316-1329.

[11] E. Barsky and J. Bar-Ilan, "From the search problem through query formulation to results on the web" *Online Information Review*, 2005, vol. 29(1), pp. 75-89.

[12] P.D. Druza and S. Dennis, "Query reformulation on the Internet: empirical data and the Hyperindex search engine", *Proceedings of the RIAO 97 Conference*, 1997, Last accessed on 9th July 2008 through http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.44.1331.

[13] P. Suppes and J.-Y. Beziau, "Semantic computations of truth based on associations already learned", *Journal of Applied Logic*, 2004, vol. 2, pp. 457-467.

[14] S.Y. Rieh and H. Xie, "Analysis of multiple query reformulations on the web: the interactive information retrieval context", *Information Processing and Management*, 2006, vol. 42, pp. 751-768.

[15] Babiniotis, G. *Dictionary of Modern Greek language: with comments for the correct usage of words*. Lexicology Centre, Athens, 1998 (In Greek).