Users and KOSs: When Can We Trust Those Two Together for Conceptual Query Expansion?

Anna Mastora and Sarantos Kapidakis
Laboratory on Digital Libraries & Electronic Publishing, Archive & Library Sciences Department, Ionian University, 72 Ioannou Theotoki Str., GR-49100, Corfu, Greece
{mastora, sarantos }@ionio.gr

1 Background & Aim

This study is part of our ongoing research on the field of conceptual query expansion, currently focusing on a user-centred approach relating to whether non-expert users could directly interact with various types of KOSs[1]. In previous phases, we studied the initial query formulation for given information seeking tasks. We mapped, both lexically and semantically, the terms used to formulate a query to terms included in certain KOSs, namely the EUROVOC[2] and GEMET[3] thesauri and WIKIPEDIA[4]. Lexical mapping of terms, which is under constant research [1], gave poor results, meaning that, in average, the terms were matched only for 49.7%, 41.4% and 27.4% of the cases respectively. We concluded that, concerning the initial query formulation, it is highly unlikely for non-expert users to use terms mapped to a domain-specific thesaurus such as GEMET. Considering, however, the high performance of its semantic relatedness (57.7%, 62.2% and 14.6% respectively for the three KOSs) we were led to believe that a domain-specific thesaurus is a good choice for bridging the semantic gap between the users' input and the terms used to represent certain content.

Building on the initial query formulation results, we explored the reformulations of the initial query, too. The hypothesis was that the outcome of this study would elaborate on the question whether non-expert users would eventually use terms included in the KOSs and are, thus, familiar with them constituting the KOSs candidate tools for conceptual query expansion in direct interaction with the users.

2 Method

For this purpose, we setup a small-scale experiment. Users (48 undergraduate and graduate students of Archive & Library Sciences Department) were initially assigned fifteen information seeking tasks, two of which are the subject of the analysis in hand: *T1* was about "Mutant products" and *T2* was about "Genetically modified organisms". Both of these represent the same concept and we, deliberately, provided different terms in the description of the task. For *T1* participants were asked to write down potential queries before logging in the system, while *T2* was the twelfth task in the overall process. For both tasks we processed the first three reformulations of the initial term. The tasks were in Greek and the users had to conduct only queries in their native language, i.e. Greek. For the mapping of terms we used the Greek versions of the previously mentioned KOSs.

3 Results

For *T1* we recorded a total of 131 terms occurring from the reformulation process while for *T2* we recorded 109 terms; not all users took advantage of the full potential of reformulations. Concerning the lexical mapping of terms we observe that there is an increase of mapped terms from *T1* to *T2* in all three KOSs. As shown in results of previous phases of our study, all three KOSs perform rather

[1] Related to workshop theme No6*"User-centred issues relating to the above, or (e.g.) user behaviour studies, user-centred design methodologies, user-centred evaluation relating to KOS"*

[2] http://europa.eu/eurovoc/. Version 4.3. Last accessed 20 June 2010.

[3] http://eionet.europa.eu/gemet. GEMET - Themes, version 2.4, 2010-01-13. Last accessed 20 June 2010.

[4] http://el.wikipedia.org/. MediaWiki version 1.16wmf4 (r66620). Last accessed 20 June 2010.

poorly concerning the lexically matched terms barely reaching half of the terms used in the best of cases and, again, the GEMET thesaurus holds the lowest percentage in average. For the semantic mapping of terms we isolated, from the already lexically mapped terms, and not from the total number of used terms, the ones actually relevant to the initial query. This means that we matched those terms which, if used in a potential thesaurus-based query expansion, would either return directly the term used in a relevant-to-the-initial-query context or following the hierarchy of the KOS would result to the same outcome. Again, the performance of GEMET thesaurus is rather limited in contrast to the less strict thesaurus, namely EUROVOC. The case of WIKIPEDIA is quite an intriguing surprise because one would expect it to be closer to non-expert users' conceptualisation of recorded information. Table 1 summarises the findings of our metrics.

Table 1. Lexical and semantic mapping of reformulation terms to KOSs

	Terms lexically mapped			**Terms semantically mapped**		
	T1	***T2***	***MA***	***T1***	***T2***	***MA***
EUROVOC	53.4%	59.6%	56.5%	62.9%	92.3%	77.6%
GEMET	26.7%	48.6%	37,6%	60.0%	30.2%	45.1%
WIKIPEDIA	38.2%	40.4%	39.3%	38.0%	86.4%	62.2%

Another observation upon the results is that *T2* gives both lexically and semantically more matched terms in all cases except in the case of the GEMET thesaurus. In our former analysis of the initial query formulation we yielded this behaviour in the users' learning-as-searching process. If this was the case, what happened in the case of GEMET? The study of raw data revealed an additional factor. As users reformulated their queries, they developed the tendency to use either more general, but still domain-specific, terms or more specific terms, but not domain-specific. In the first case, users used terms more likely to appear in the higher levels of the hierarchy which represent more basic concepts and, thus, more likely for them to be included in any KOS concerning a certain domain knowledge. For example, for *T2* we counted two occurrences of the term *biology*, during the initial formulation of queries but nine occurrences during the reformulation process. In the second case, users used more specific terms but not the terms to be included in a domain-specific thesaurus. For example, users searched for *Darwin* which in the case of WIKIPEDIA led to relevant results, but the term was not included in either of the thesauri.

4 Conclusions

Evidence so far shows that, to a great extent, users tend to search with terms not included in KOSs. Approximately half of the terms used can be matched to terms from KOSs. Concerning potential use of KOSs for query expansion we have to take into consideration that non-expert users are highly unpredictable as to what terms they use for initial query formulation, as well as for subsequent reformulations. They seem to better match to less strict structures of knowledge representation having the disadvantage that they lack in semantic relatedness. We conclude that users cannot directly interact with all kinds of KOSs during a query expansion process, therefore, the use of an intermediate between the users and a domain-specific knowledge structure would be appropriate to deal with the term mismatch problem, which is very well summarised in [2].

5 References

1. Fang, H. (2008). A Re-examination of Query Expansion Using Lexical Resources. In: *Proceedings of ACL-08: HLT*. Columbus, Ohio, USA: Association for Computational Linguistics. p139–147.
2. Gray, A.J.G., Gray, N., Hall, C.W. and Ounis, I. (2010). Finding the right term: Retrieving and exploring semantic concepts in astronomical vocabularies. *Information Processing and Management*. 46, p470–478.