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

Anna Mastora

Sarantos Kapidakis

Laboratory on Digital Libraries & Electronic Publishing
Department of Archive & Library Sciences
Ionian University (Corfu, Greece)

Presentation Outline

- Aim of the study
- Research background
- Related literature
- Methodology
- Results
- Conclusions
- Future work
- Bibliography

Aim of the study

Improve CQE (Conceptual Query Expansion) effectiveness through better understanding users' interaction with KOSs (Knowledge Organisation Systems)

More specifically:

To what extent do the terms used for query formulation by non domain-experts map to KOSs

Research background

Ongoing research on the field of CQE dealing with:

- The conceptual gap between the queries & the content representation
  - ➢ The user’s perception of knowledge representation => USERS
  - ➢ The knowledge representation itself => KOSs
- The systems’ features => SYSTEMS

✓ The interaction among all the above

Our focus: KOSs and Users

Related literature (i)

- Many studies of the Cognitive IR domain explore QE techniques (Owei & Navathe, 2001; Zazo et al., 2005)
  - Implementations based on lexical (=linguistic) mapping of query terms to KOSs (Shiri & Revie, 2006)
  - Implementations using the hierarchy of KOSs (Tudhope et al., 2006)

Related literature (ii)

- Initial input of the searcher is neither stable nor easily predictable due to searcher's knowledge state, ability to conceptualise and interpret the query, system features… (Wilson, 1997; Spink & Cole, 2006)

- Specifically for the environmental information retrieval it has been stated that KOSs have been introduced to reduce the ambiguity of natural language (Palavitsinis & Manouselis, 2009)

Methodology

We set up an experiment…

- Search Tasks
- Database
- System
- Participants

- Transaction log files (for future use; not currently processed)

We studied:

- First, the lexical (=linguistic) relation between queries and KOSs
- Second, the semantic closeness (=relatedness) between queries and KOSs

Methodology: the Task

- **Participants** were issued with information about the content of the db, the system and the tasks they had to do
- They **had to perform simple searches for given information needs** (we, particularly, encouraged reformulations of the initial term), give demographic data, write down which terms they used, record date & time of logging in and out of the system
- **Queries** only **in Greek** (*no worries! we provide translations for the purpose of this presentation*)

✓ Users did not get involved with the KOSs; the reformulations of initial terms were not assisted

✓ We did the mappings after collecting all the data

Methodology: the Database

- ...of general interest: **Environment**
- ...contained approx. **14400** bibliographic records courtesy of the "**Evonymos Ecological Library**"
  - Temporary access to selected data; the test-collection is not publicly available *(entire db available here: http://www.evonymos.org)*
- ...uses no particular tool for subject indexing
- ...was customised, i.e.:
  - only contained subjects in Greek
  - no literature (*eliminated risk of misleading representation of concepts*)

Methodology: the System

- A minimal interface based on z39.50 search features
  - Search area: **Subject (only!)**
  - Boolean operators: dismissed
  - Structure: “words”
  - Truncation: right

- Why so simple?

- Because users should focus on query terms and not on dealing with system’s features. Plus, their skills as information professionals should not get in the way.

Methodology: the Participants

- Students of the Archive & Library Sciences Department, Ionian University, Corfu (Greece)
  - We could easily locate them to get feedback

- **Undergraduates** (27), mandatory participation, under supervision (lab time)
- **Postgraduates** (21), voluntary participation, without supervision
  - Female (40) & Male (8)

The KOSs (Greek versions)

- EUROVOC thesaurus (***multi-disciplinary***)
  - europa.eu/eurovoc, v. 4.3
  - **6797** descriptors
- GEMET thesaurus (***domain specific***)
  - GEneral Multilingual Environmental Thesaurus
  - eionet.europa.eu/gemet, v. 2.4
  - **5204** descriptors
- WIKIPEDIA (***built by users***)
  - el.wikipedia.org, v. 1.16wmf4(r66620)
  - ~**54500** articles

Results analysis - The topics (i)

- T1: *Mutant products*
- T2: *Genetically modified organisms*

- The description of the information search tasks consisted of related concepts
  - Not only that; many Greek sources use these terms even interchangeably; odd but true...
- We examined each term manually against each of the KOSs, both lexically and semantically

Analysis criteria

- Choice of KOSs: all had **Greek versions**, so location of terms and comparison was safe; no translations were necessary
- We did not take into account:
  - spelling errors
  - singular-plural forms
  - truncated words
- Words were assigned to the relative concept (by two judges, manually)
- We used the article titles for the Wikipedia; no external redirections, no scope notes counted
- **All data was in the users' native language; avoided linguistic barriers and controversial translation of terms**

Facts and figures

- **240 terms examined overall** (not unique)
  - 131 (reformulated) terms for T1
  - 109 (reformulated) terms for T2
- The given term “mutant products” was not included to any of the KOSs; the given term “genetically modified organisms” was not included in the Wikipedia
  - Many users used the terms given within the task description to both formulate and reformulate their queries so occurrences were relatively affected

Results analysis (i)

Lexical mapping of concepts

	Terms lexically mapped		
	T1	***T2***	***MA***
EUROVOC	53.4%	59.6%	**56.5%**
GEMET	26.7%	48.6%	**37.6%**
WIKIPEDIA	38.2%	40.4%	**39.3%**

- **Terms treated as *bags-of-words***; e.g. the user searched for “metallaxis” which stands for *mutation*. The word was detected in all KOSs, so, it is a “lexical match”
- *EUROVOC* offered more lexically mapped terms than the other KOSs; *GEMET* offered the fewest
- *T2* offered more lexically mapped terms than *T1* in all cases

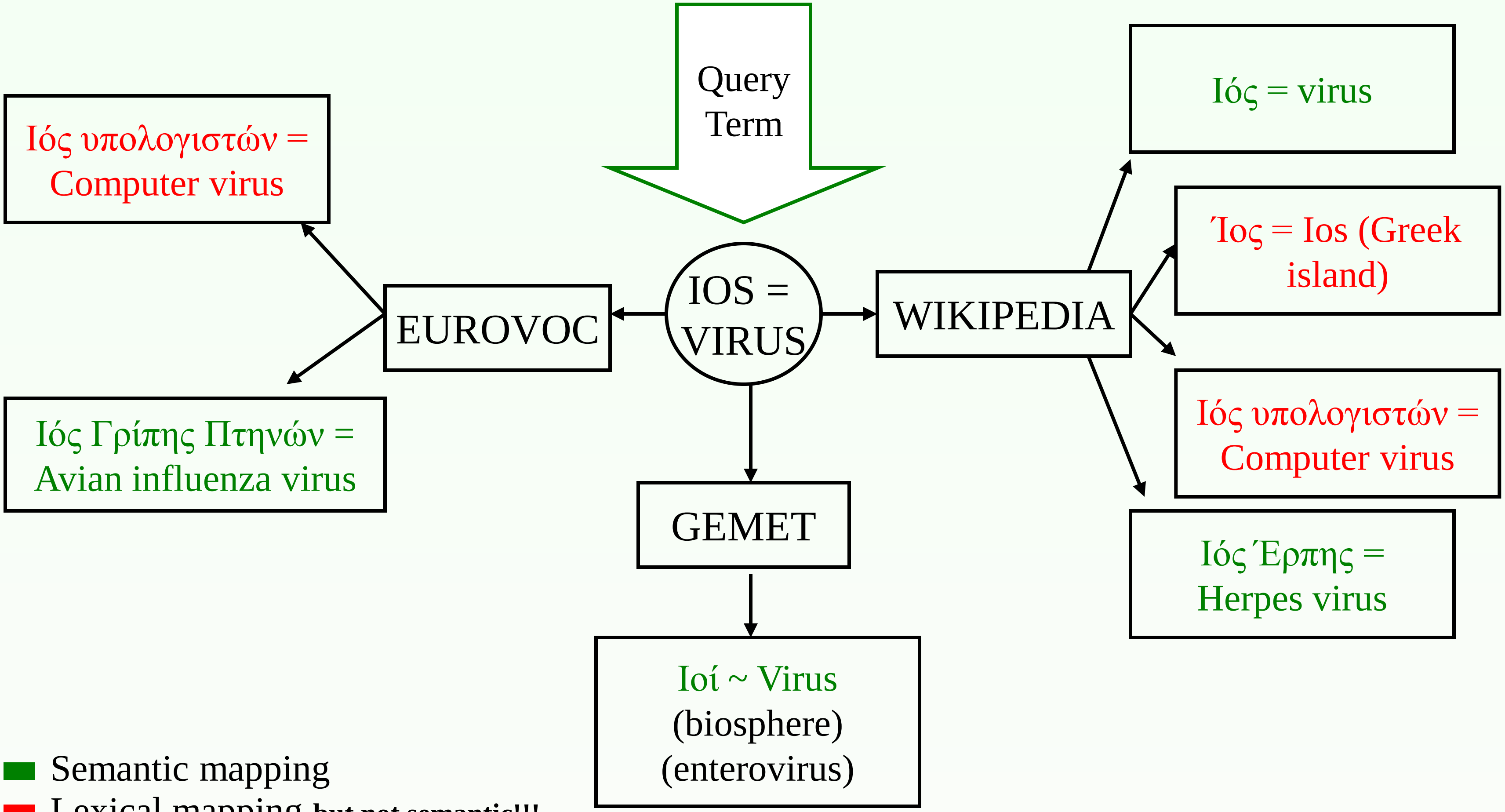
Results analysis (ii)

Semantic mapping of concepts

	Terms semantically mapped		
	T1	*T2*	*MA*
EUROVOC	33.6%	55.0%	**44.3%**
GEMET	16.0%	14.7%	**15.4%**
WIKIPEDIA	14.5%	34.9%	**24.7%**

- The word “metallaxis” was detected in the KOSs, but in the case of EUROVOC it represented the concept of “social reform”; thus, it is not a “semantic match”
- *EUROVOC* offered more semantically mapped terms than the other KOSs; *GEMET* offered the fewest
- *T2* offered more semantically mapped terms than *T1* except in the case of *GEMET*

Lost in… concepts: an example



Query Term

IOS = VIRUS

EUROVOC

Ιός υπολογιστών = Computer virus

Ιός Γρίπης Πτηνών = Avian influenza virus

GEMET

Ιοί ~ Virus (biosphere) (enterovirus)

WIKIPEDIA

Ιός = virus

Ίος = Ios (Greek island)

Ιός υπολογιστών = Computer virus

Ιός Έρπης = Herpes virus

🟩 Semantic mapping

🟥 Lexical mapping **but not semantic!!!**

Discussion (i)

	Terms lexically mapped			Terms semantically mapped		
	T1	*T2*	*MA*	*T1*	*T2*	*MA*
EUROVOC	53.4%	59.6%	**56.5%**	33.6%	55.0%	**44.3%**
GEMET	26.7%	48.6%	**37.6%**	16.0%	14.7%	**15.4%**
WIKIPEDIA	38.2%	40.4%	**39.3%**	14.5%	34.9%	**24.7%**

- Lexical mapping of terms used reaches a certain percentage; Semantic mapping reduces this percentage in all cases. The loss in mapped terms is the following: EUROVOC: -12.2%, GEMET: -22.2%, WIKIPEDIA: -14.6%
- The domain-specific thesaurus, GEMET, seems not to contain terms likely to be used by non-expert users
- Wikipedia was expected to be closer to users' conceptualisation of terms; the final outcome, though, could be due to the specific subjects used and/or the structure of the tool

Discussion (ii)

- During reformulation of queries users tend to search using terms either
  - ➢ more general, but domain-specific, i.e. biology
    - we counted 2 occurrences of “biology” during initial query formulation but 9 during reformulations
  - ➢ or, more specific, but not domain-specific, i.e. *Darwin*
    - in the case of Wikipedia “Darwin” was detected and even led to relative results; “Darwin” did not appear in the two thesauri giving negative match in the metrics

✓ *More general terms are more likely to appear in any kind of KOSs as Top Terms, so it is more likely to give positive matches if used in queries*

Further observations

	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%**

- Only terms that were lexically mapped were, then, used for this additional semantic mapping computation
- Terms that were not lexically mapped during the first phase were not included in this metric

Conclusions

- Users use terms lexically mapped to KOSs in ~37-56% of the cases
- Users use terms semantically mapped to KOSs in ~15-44% of the cases

The reduced ratio between lexical and semantic mapping is an issue we have to overcome

- Lexical mapping is a good starting point but not safe for implementing CQE mechanisms
- Non-expert users need less strict structures of KOSs

Future work

- Would the use of more KOSs solve the problem of conceptual gap between queries and knowledge representation?
- What happens with the terms that were not mapped to the thesauri?
  - Mapping of KOSs creating mediators between users and KOSs?
  - Move to query clustering by identifying the query intent?

Bibliography

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.

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.

Owei, V. and Navathe, S. B. 2001. Enriching the conceptual basis for query formulation through relationship semantics in databases. *Inform. Sys.* 26, 6 (Sep. 2001), 445-475. DOI= 10.1016/S0306-4379(01)00029-1.

Palavitsinis, N. and Manouselis, N. 2009. A survey of knowledge organization systems in environmental sciences. In *Information Technologies in Environmental Engineering*, I. N. Athanasiadis et al., Ed. Environmental Science and Engineering. Springer-Verlag, Berlin, Heidelberg. DOI= 10.1007/978-3-540-88351-7.

Shiri, A. and Revie, C. 2006. Query expansion behavior within a thesaurus-enhanced search environment: A user-centered evaluation. *J. Am. Soc. Inf. Sci. Technol.*57, 4 (Feb. 2006), 462-478. DOI= 10.1002/asi.v57:4.

Spink, A. and Cole, C. 2006. Human information behavior: Integrating diverse approaches and information use. *J. Am. Soc. Inf. Sci. Technol.* 57, 1 (Jan. 2006), 25-35. DOI= 10.1002/asi.v57:1.

Tudhope, D., et al. 2006. Query expansion via conceptual distance in thesaurus indexed collections. *J. Doc.* 62, 4, 509-533. DOI= 10.1108/00220410610673873.

Wilson, T. D. 1997. Information behaviour: An interdisciplinary perspective. *Inf. Process. Manage.* 33, 4 (Jul. 1997), 551-572.

Zazo, A. F. et al. 2005. Reformulation of queries using similarity thesauri. *Inf. Process. Manage.* 41, 5 (Sep. 2005), 1163-1173. DOI= 10.1016/j.ipm.2004.05.006.

That’s all!

Thank you for your attention!

Any questions?

Laboratory on Digital Libraries & Electronic Publishing
Department of Archive & Library Sciences
Ionian University (Corfu, Greece)
Email (Anna Mastora): mastora@ionio.gr