Evaluation of Multi-Terminology Super-Concepts for Information Retrieval

Nicolas GRIFFON[a], Lina F. SOUALMIA[b,c] (PhD), Aurélie NÉVÉOL[d] (PhD), Philippe MASSARI[a] (MD), Benoit THIRION[a,b], Badisse DAHAMNA[a,b] (MSc), Stefan J. DARMONI[a,b,1] (MD, PhD)

[a]*CISMeF, Rouen University Hospital, France*
[b]*TIBS & LITIS EA 4108, Rouen University, France*
[c]*LIM&Bio, University of Paris 13, Sorbonne Paris Cité, France*
[d]*National Center for Biotechnology Information, NLM, Bethesda, MD 20894, USA*

Abstract. Background: Following a recent change in the indexing policy for French quality controlled health gateway CISMeF, multiple terminologies are now being used for indexing in addition to MeSH®. Objective: To evaluate precision and recall of super-concepts for information retrieval in a multi-terminology paradigm compared to MeSH-only. Methods: We evaluate the relevance of resources retrieved by multi-terminology super-concepts and MeSH-only super-concepts queries. Results: Recall was 8-14% higher for multi-terminology super-concepts compared to MeSH only super-concepts. Precision decreased from 0.66 for MeSH only super-concepts to 0.61 for multi-terminology super-concepts. Retrieval performance was found to vary significantly depending on the super-concepts ($p<10^{-4}$) and indexing methods (manual vs automatic; $p<0.004$). Conclusion: A multi-terminology paradigm contributes to increase recall but lowers precision. Automated tools for indexing are not accurate enough to allow a very precise information retrieval.

Keywords. abstracting and indexing; cataloguing; information storage and retrieval; internet; controlled vocabulary

Introduction

The Internet contains a considerable amount of health information that internet users experience difficulties navigating [1]. Several quality-controlled health gateways have been developed to help users find the health information they are looking for. Quality controlled subject gateways were defined by Koch [2] as Internet services which apply a comprehensive set of quality measures to support systematic resource discovery. CISMeF ([French] acronym for Catalogue and Index of Online Health Resources in French) is one such gateway, developed at the Rouen University Hospital. It initially relied on the Medical Subject Headings (MeSH®) thesaurus [3] to manually index the most important sources of institutional health information in French. This thesaurus was chosen because of its granularity (26,142 MeSH keywords describing the biomedical domain in the 2011 version) and the fact that it is well known among

[1] Corresponding Author: Stefan J. Darmoni, CISMeF, Rouen University Hospital, 1 rue de Germont, 76031 Rouen Cedex, France; E-mail: stefan.darmoni@chu-rouen.fr.

medical librarians. Several improvements have been introduced to adapt this scientific publication-oriented indexing vocabulary to internet resources [4].

A notable enhancement was the gathering of MeSH terms under meta-terms. These are super-concepts (SC) which correspond roughly to medical specialties (e.g. surgery), biological sciences (e.g. genetics) or health topics (e.g. diagnosis). MeSH terms were semantically linked to SCs to allow end-users to look for all the resources relevant to one specialty, which is difficult with the MeSH thesaurus, since MeSH terms related to a given specialty are dispersed among the 14 MeSH hierarchies. These semantic links have been hand-crafted by the CISMeF chief medical librarian (BT), based on his expertise. The idea of creating SC came up to maximize information retrieval in CISMeF: a query using the SC corresponds to the union of queries for all the terms semantically linked to it. A comparison of the results of MeSH term-based queries and SC-based queries showed an increased recall with no decrease in precision [5].

The use of multiple terminologies was recommended [6] to increase the number of biomedical concept lexical and graphical forms recognized by the search engine. For this reason, CISMeF evolved recently from a mono-terminology approach using MeSH keywords and qualifiers to a multi-terminology paradigm using, in addition to MeSH: Systematized NOmenclature of MEDicine (SNOMED 3.5), French CCAM for procedures, Foundational Model of Anatomy (FMA), and some classifications from the World Health Organization viz. the 10th revision of the International Classification of Diseases (ICD10), Anatomical Therapeutic Chemical (ATC) Classification for drugs, ICF for handicap, ICPS for patient safety [7]. These terminologies can be used for indexing resources (allowing a more precise indexing) and for querying the Catalogue.

The goal of this study is to assess the effect of multi-terminology SC (MT-SC) definition compared to MeSH-only SC (MeSH-SC) definition on information retrieval performance in CISMeF.

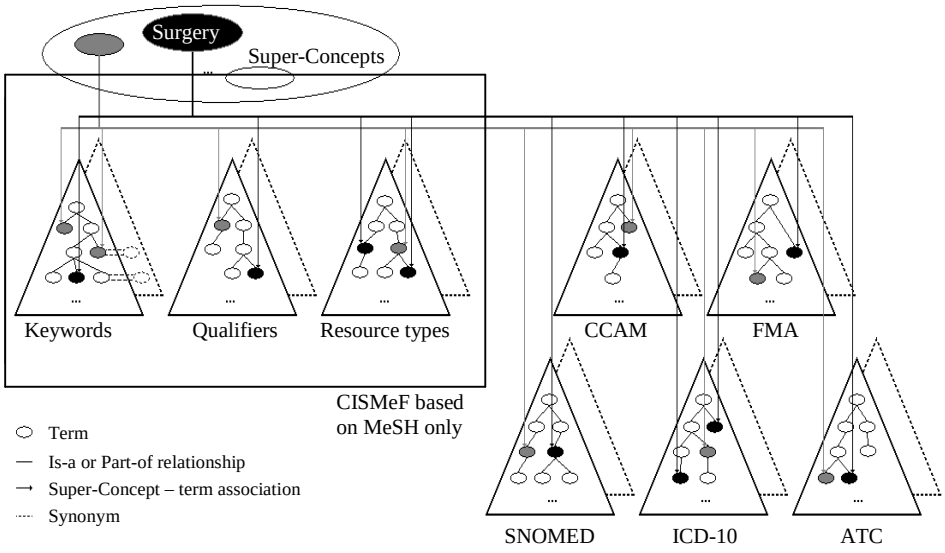

Figure 1. Semantic links between CISMeF Super-Concept and terminologies terms and resource types. Terminology terms describe the subject matter of the resources, resource type categorize the nature or genre of the resource content.

1. Material and Methods

1.1. CISMeF Information Model

The addition of multiple terminologies to CISMeF did not induce modifications in the tasks performed for using, maintaining and updating the catalogue: manual resource indexing, automatic resource categorization, visualizing and navigating through the concept hierarchies in the CISMeF Health Multi-Terminology Portal (URL: http://www.pts.chu-rouen.fr) and information retrieval using the Doc'CISMeF search tool. Nevertheless, the tools used for indexing and retrieving information needed important modifications [7].

As shown in figure 1, new terminologies have been linked to SCs manually by experts: one physician (PM) for ICD10 and CCAM, one pharmacist-librarian for ATC, and one resident (NG) for FMA. For instance, SC "cardiology" was initially linked to MeSH keywords such as "cardiology", "stents", and their descendants. With the integration of new terminologies, additional links completed the definition of SC “cardiology”: links to "cardiovascular system", "Antithrombotic agents" and others from ATC, links to "Cardiomyopathy", "Heart" and their descendants from ICD10 and so on. These mappings are available at: http://pts.chu-rouen.fr.

1.2. Information Retrieval Queries

Our aim is to compare the precision and recall of MT-SC compared to MeSH-SC queries in CISMeF. As MT-SC are based on MeSH-SC plus semantic link to some terms in other terminologies, the query results for MeSH-SC are all included in the query results for MT-SC, which became the gold standard for recall. So, we have to evaluate the precision of the query retrieving resources indexed by a term linked to MeSH-SC (MeSH-SC query), on the one hand, and by a term linked to MT-SC and not to MeSH-SC (Delta query) on the other hand.

For this purpose, we build Boolean queries using SC themselves: for the "surgery" SC, MeSH-SC query was "surgery[MeSH-SC]" and Delta query was "surgery[MT-SC] NOT surgery[MeSH-SC]". Retrieved resources returned were assessed for relevance according to a three modality scale used in other standard Information Retrieval test sets [8]: irrelevant (0), partly relevant (1) or fully relevant (2). A medical resident (NG) manually assigned relevance scores to the top 20 resources returned for each SC query in our study (see Table 1). We chose to assign relevance scores to the top 20 resources returned because 95% of the end-users do not go beyond this limit when using a general search engine [9], and 80% when using a biomedical search engine [10].

Weighted precisions for MeSH-SC queries and for Delta queries were computed given the level of relevance considered and compared using χ^2 test. Indexing methods and SC were compared too. Relative recall for MeSH-SC queries were computed given the level of relevance considered.

2. Results

For the purpose of assessing SCs for Information Retrieval, we have developed a test collection comprising relevance judgments for the top 20 resources returned for a

Table 1. Relevance of resources retrieved by 20 Super-Concept queries

Super-Concept query	Number of resources retrieved		Relevance of top 20 retrieved resources					
	MeSH-SC Query	Delta Query	MeSH-SC Query*			Delta Query*		
			0	1	2	0	1	2
Diagnosis	13,132	350	0	2	15	14	1	5
Toxicology	11,980	482	0	0	20	16	1	3
Neurology	9,325	2,168	8	4	8	11	5	4
Infectious diseases medicine	6,557	2,573	0	0	20	3	16	1
Paediatrics	7,560	251	4	4	12	2	4	13
Cardiology	5,288	2,388	1	0	18	4	10	6
Oncology	5,626	1,063	0	1	18	2	14	4
Surgery	5,504	320	17	0	3	5	0	15
Rheumatology	4,408	856	3	8	9	11	5	4
Gastroenterology	4,069	1,106	0	0	20	8	11	1
Study of allergies and immunology	4,598	573	1	17	2	2	17	1
Metabolism	3,797	849	14	2	4	0	2	18
Dermatology	3,196	1,427	7	0	13	0	4	16
Nutrition	3,455	1,027	0	1	19	0	9	11
Pneumology	3,466	584	0	7	12	0	14	6
Gynaecology	3,186	850	6	1	12	0	1	19
Haematology	2,906	1,075	13	2	5	7	10	3
Endocrinology	3,168	666	15	1	4	0	9	11
Obstetrics	3,063	316	5	1	12	20	0	0
Virology	3,122	257	1	11	6	0	20	0
Total	**107,406**	**19,181**	**95**	**62**	**232**	**105**	**153**	**141**

*: Due to dead link, some queries had less than 20 resources evaluated

selection of 20 SC queries. This collection is made available to the research community. Table 1 shows that the queries yielded 126,587 resources (59224 unique), of which 788 (754 unique) were assessed for relevance (0.6%).

The mean weighted precision of Delta queries was 0.33 and 0.76 for, respectively, full and partial relevance. The mean precision of MeSH-SC queries was 0.66 and 0.80 for, respectively, full and partial relevance. The difference between MeSH-SC and MT-SC was significant for full relevance (0.66 vs 0.61; $p<10^{-4}$, χ^2) but not for partial relevance (both 0.80; p=0.3, χ^2). The mean recall of MeSH-SC queries was 0.92 and 0.86 for, respectively, full and partial relevance. Table 2 shows that, whatever the relevance considered was, results varied significantly according to the indexing method: manual (precision of 0.50 and 0.81 for, respectively, full and partial relevance) perform better than automatic (precision of 0.38 and 0.48 for, respectively, full and partial relevance), and to the SC studied.

3. Discussion & Conclusion

This study evaluates the precision and the recall of the MT-SC compared to MeSH-SC queries in information retrieval in quality controlled subject gateway CISMeF. For full relevant resources, the precision decreases with the shift from MeSH-SC to MT-SC (from 0.66 to 0.61) for an 8% improvement in recall. For partial relevance, the increase of recall with multiple terminology is even higher (14%) at no cost in terms of precision (0.80)

Table 2. Determinants of relevance

Variable	Full relevance	Partial relevance
Specific query$	$p < 10^{-4}$*	p = 0.3*
Indexing method	p = 0.004*	$p < 10^{-4}$*
Super-concept	$p < 10^{-4}$*	$p < 10^{-4}$*

$: MeSH-SC vs MT-SC ; *: χ^2 test

Because of the significant difference in relevance between MT-SC and MeSH-SC queries, MT-SC queries will be best used when the MeSH-SC result set is small. In this case, MT-SC queries can offer a larger result set with good partial relevance.

A limitation of this study is that only the top 20 results are assessed for relevance. This possibly induced bias, because resources are sorted by a relevance algorithm, but we think this method reflects the real life, since most users usually do not look at results beyond the first page, i.e. the top 20 documents returned [9, 10].

This analysis underlines that the performance of the automatic indexing algorithm is lacking and needs to be improved significantly. However, even resources indexed manually (thus having higher quality indexing) were less relevant for MT-SC than for MeSH-SC. We have possible explanations: (1) some hand-crafted links between descriptors and SC have been found to be erroneous and will be corrected soon, (2) the shift to multi-terminology occurs recently and concern only new resources that are different from old MeSH only indexed resources. These 2 sets of resources are not comparable (e.g. some of these new resources, providing very standardized and precise information, need new indexing strategy to avoid them inducing noise).

Overall, the multi-terminology paradigm for super concepts definition was found to increase the recall but lower the relevance of retrieved resources. Automated tools for indexing are not accurate enough to allow a very precise information retrieval.

References

[1] Keselman A, Browne AC, Kaufman DR. Consumer health information seeking as hypothesis testing. J Am Med Inform Assoc. 2008 Jul-Aug;15(4):484-95. doi: 10.1197/jamia.M2449

[2] Koch T. Quality-controlled subject gateways: definitions, typologies, empirical overview, subject gateways. Online Information Review. 2000;24(1):24-34. doi: 10.1108/14684520010320040

[3] Nelson SJ, Johnson WD, Humphreys BL. Relationships in Medical Subject Heading. In: Relationships in the organization of knowledge. Bean CA, Green R. Kluwer Academic Publishers, 2001:171-84

[4] Douyère M, Soualmia LF, Névéol A, Rogozan A, Dahamna B, Leroy JP, Thirion B, Darmoni SJ. Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. Health Info Libr J 2004 Dec;21(4):253-261. doi: 10.1111/j.1471-1842.2004.00526.x

[5] Gehanno JF, Thirion B, Darmoni SJ. Evaluation of meta-concepts for information retrieval in a quality-controlled health gateway. AMIA Annu Symp Proc. 2007;269-73

[6] Wagner MM. An automatic indexing method for medical documents. Proc Annu Symp Comput Appl Med Care. 1991;1011-7.

[7] Darmoni SJ, Pereira S, Sakji S, Merabti T, Prieur E, Joubert M, Thirion B. Multiple terminologies in a health portal: automatic indexing and information retrieval. In: Conference on Artificial Intelligence in Medicine. 2009;255-9. doi: 10.1007/978-3-642-02976-9_37

[8] Hersh W. Buckley C, Leone TJ, Hickam D. OHSUmed: An interactive retrieval evaluation and new large test collection for research. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval (1994), pp. 192-201.

[9] Spink A, Jansen BJ. Web search: Public searching on the web. Kluwer Academic Publishers, 2004;199.

[10] Islamaj Doğan R, Murray GC, Névéol A, Lu Z. Understanding PubMed user search behavior through log analysis. Database. 2009 ;bap018. doi: 10.1093/database/bap018.