Medical Informatics in a United and Healthy Europe
K.-P. Adlassnig et al. (Eds.)
IOS Press, 2009
© 2009 European Federation for Medical Informatics. All rights reserved.
doi:10.3233/978-1-60750-044-5-238

Optimization of the PubMed Automatic Term Mapping

Benoit THIRION [a], Ioana ROBU [b], Stéfan J. DARMONI [a,1]
[a] *CISMeF, Rouen University Hospital & GCSIS, TIBS, LITIS EA 4108, Biomedical Research Institute, Rouen, France*
[b] *Central Library, University of Medicine and Pharmacy, Cluj-Napoca, Romania*

Abstract. PubMed, freely available on the internet, is the best known database for medical information. We propose a method of optimization of the PubMed Automatic Term Mapping (ATM) that includes MeSH terms. This method is evaluated using two queries constructed to emphasize the differences between the PubMed queries as they are at present and also between these queries and the optimized one. The proposed query is significantly more precise than the current PubMed query (54.5% vs. 27%). The optimized query proposed would be easy to implement into PubMed.

Keywords. algorithms, information storage and retrieval, medical subject headings, MEDLINE, PubMed

1. Introduction

One of the most important tools to access scientific information is the Medline bibliographic database, which uses the MeSH thesaurus [1]. Since 1997, Medline has been freely available on the Internet via PubMed. Because one third of Medline queries are performed by members of the general public [2] and furthermore because most health professionals do not know well enough the MeSH thesaurus, the PubMed web site has developed several techniques (*Automatic Term Mapping (ATM)*) to map the end-user query to the MeSH thesaurus, mainly natural language processing (NLP) techniques [3]. The PubMed ATM is also matching the end-user query to other tables (e.g., for Journals and Authors). The objective of this paper is to propose an optimization of PubMed ATM, when the end-user query is composed exclusively of MeSH terms.

2. Material and Methods

When this study was performed (April 2008) and when an end-user performs a query on PubMed using a MeSH term, the PubMed default query was different [4] if the end-user searches with a "preferred term" (e.g., "*liver neoplasms*") or an "entry term" [5] (e.g., "*hepatic cancer*")[2]. (query 1) If an end-user queries for "*liver neoplasms*", which is the preferred

[1] Corresponding Author: Prof. SJ. Darmoni, Head of the CISMeF team, Rouen University Hospital, Normandy, 1 rue de Germont, 76031 Rouen Cedex, France; E-mail: stefan.darmoni@chu-rouen.fr.

[2] "The Main Heading and entry terms are the names of the concepts in a descriptor class (…). An entry term may be a synonym of the descriptor name, or it may be a name of an additional concept in the descriptor" [5].

term in the MeSH thesaurus, the transformed PubMed query is: *"liver neoplasms"[MeSH Terms] OR liver neoplasms[Text Word]*, where "*liver neoplasms*"[MeSH Terms] will retrieve all citations manually indexed with this MeSH term as well as those which include all words and numbers appearing in their title, abstract, MeSH terms, MeSH subheadings, Publication Types, Substance Names, Personal Name as Subject, etc. (query 2) If an end-user queries for "*hepatic cancer*", which is a synonym (or "entry term") of the MeSH term (preferred term) "*liver neoplasms*", the transformed PubMed query is: *("liver neoplasms"[TIAB] NOT Medline[SB]) OR "liver neoplasms"[MeSH Terms] OR hepatic cancer[Text Word].* [TIAB] includes all words and numbers included in the title or the abstract of a citation. NOT Medline[SB] displays citations that have not been indexed yet with MeSH Terms ("in Process Citations" OR "Publisher-Supplied Citations" (formally PreMEDLINE)) [3]. PubMed's in-process records provide basic citation information and abstracts before the citations are indexed with NLM's MeSH Terms and added to MEDLINE. New in-process records are displayed with the tag [PubMed – in process]. Citations received electronically from publishers appear in PubMed with the tag [PubMed – as supplied by publisher] [3]. It is important to note that for the citations which are not yet indexed with MeSH terms, [TIAB] is very similar to [Text Word]. In the April 2008 version of PubMed, the automatically mapped (default) query is different if the end-user enters a "preferred term" (query 1) or enters an "entry term" (query 2). Query 1 is maximizing the recall (with a potential diminution of the precision) when using [Text Word] without restricting it to the citations that have not been indexed yet with MeSH Terms (NOT Medline[SB].). This type of query retrieves "in Process Citations" and "Publisher-Supplied Citations" and also citations that are not indexed voluntarily by indexers with the MeSH term but are present in the title or the abstract (*[Text Word]).* Starting from the assumption that the same concept is sought for by the users, whether entering a preferred MeSH term or a synonym, we propose an optimization of the (default) automatic query mapping in PubMed for any query including MeSH terms (preferred terms or entry terms). Thus, for any user search which includes either preferred term or an entry term the automatically mapped query will be the same, as shown below: (query 3) *"Preferred term"[MeSH Terms] OR (("Preferred term"[Tiab] OR "entry term(s)"[Tiab]) NOT MEDLINE[SB]);* e.g., this proposed default query for "*liver neoplasms*" is: *"liver neoplasms"[MH] OR (("liver neoplasms"[TIAB] OR "cancer of liver"[TIAB] OR "cancer of the liver"[TIAB] OR "hepatic cancer"[TIAB] OR "hepatic cancers"[TIAB] OR "hepatic neoplasm"[TIAB] OR "hepatic neoplasms"[TIAB] OR "liver cancer"[TIAB] OR "liver cancers"[TIAB] OR "liver neoplasm"[TIAB]) NOT MEDLINE[SB]).*

This query is the same for all entry terms of the MeSH term "*liver neoplasms"* (e.g., "*hepatic cancer"*). This query will systematically include all the entry terms (with an expected increase of the recall). Moreover, we suggest the use of [TIAB] NOT Medline [SB] for the preferred term and entry terms in order to retrieve the "in process" and "as supplied by publisher" citations and at the same time exclude the references deliberately not indexed with the MeSH term composing the query. The choice of *[TIAB]* over *[Text Word]* is explained by an expected higher precision of the results.

To evaluate the respective precision of this new default query, we used the Top Ten MeSH terms from the C tree (Diseases) in terms of frequency use in the MEDLINE bibliographic database (see Table 1). The choice of the C (Diseases) tree from the MeSH thesaurus was driven by its potential impact in daily health care.

To compare the current PubMed default queries (queries 1 & 2) and the query proposed by our team (query 3), we used the following two queries (query 4 & query 5) focusing respectively on the differences between queries 1 & 2 on the one hand and query 3 on the other. For the MeSH term X, we excluded from the queries 1, 2, 3 the common part in these three queries: X*[MeSH Terms]*. Thus, query 1 & 2 was transformed in the following query 4:

(query 4) (*X[Text Word] NOT "X"[MeSH Terms]) AND medline[sb]*; e.g., for "*liver neoplasms*", (query 4) (*liver neoplasms[Text Word] NOT "liver neoplasms"[MeSH Terms]) AND medline[sb]* Query 4 is expected to measure the precision of [*Text Word*], when the respective MeSH term has not been used by the indexers (*NOT "X"[MeSH Terms]*) to describe the content of the indexed article (*Medline[SB]*).

Query 3 was transformed into the following query 5: (query 5) *Synonym(s) of X[TIAB] NOT medline[SB]*; e.g., for "*liver neoplasms*", (query 5) *(("cancer of liver"[TIAB] OR "cancer of the liver"[TIAB] OR "hepatic cancer"[TIAB] OR "hepatic cancers"[TIAB] OR "hepatic neoplasm"[TIAB] OR "hepatic neoplasms"[TIAB] OR "liver cancer"[TIAB] OR "liver cancers"[TIAB] OR "liver neoplasm"[TIAB]) NOT MEDLINE[SB]).* Query 5 is expected to measure the precision of [TIAB] when the citations have not been indexed yet with MeSH Terms (*NOT Medline[SB]*). This query is limited to synonyms because the X[*Text Word*] of the query 1 & 2 is (almost) similar to the *X[TIAB]* of the query 3.

By construction, queries 4 & 5 have no intersection: (no citations in common). We have manually evaluated the precision of the Top 20 answers of queries 4 & 5. This manual evaluation was performed by consensus of two authors. The assessment of the indexing quality was performed based on Title and Abstract only, not in the full text of the article. To obtain a rough estimation of the respective recall of queries 4 & 5, we have applied two methods and used two extrapolations. We assumed that the precision found in the top 20 retrieved citations is the same for the overall citation retrieved, which is a rough extrapolation. *Method 1*: if we take for example "asthma", we can compute the respective number of relevant citations as: 0.25*15,102 = 3,775 for query 4 and 0.60*416= 250 for query 5 and then make the sum for the ten MeSH terms to obtain 10,713 and 51,662 respectively for the queries 4 & 5. *Method 2*: we have extrapolated the total number of relevant citations by multiplying the total number of retrieved citations for the ten MeSH terms and for the two queries 4 & 5 by the mean percentage of relevant citation and obtained 38,506 (0.270*142,615) and 43,946 (0.545*80,635) respectively.

3. Results

The results of this evaluation are summarized in Table 2. The current PubMed query for MeSH terms (query 1) provides slightly more overall results than the proposed optimization (query 3) for the Top ten MeSH terms of the Disease tree: 2,959,484 vs. 2,857,971 (+3.55%). In four cases out of ten, the proposed optimization (query 3) has provided more results. In two cases, the current PubMed query (query 1) has provided largely more results than query 2: for the MeSH terms "asthma" (99,539 vs. 84,345, + 18.02%) and "hypertension" (269,983 vs. 174,979, + 54.29%). The precision for queries 4 & 5 are varying from one disease to another (min-max respectively: 0 (hypertension) – 55 (neoplasms) for query 4 and 25 (hypertension) – 85 (HIV infections) for query 5). The precision between query 4 & 5 has a significant difference in only one case (HIV infections) (p=0.003; Exact Fisher test) when applying the Bonferroni correction, the precision with query 5 being higher in this case (see Table 1). There is a significant difference between the mean percentage of relevant answers for queries 4 & 5 (27.0 vs. 54.5) (p < 0.0001; Fisher exact test), the percentage being higher with query 5 (see Table 2).

The estimation of the respective recall of queries 4 & 5 are the following (see Table 2): 17.2% vs. 82.8% (method 1) and 46.7% vs. 53.3% (method 2). We did not apply statistical test for the recall because it was extrapolated from the precision. The optimization of the default query proposed in this paper is already accessible from the CISMeF catalogue [6] ([French] acronym for Catalog and Index of French Language Health Resources on the Internet [7]) and from the French MeSH browser [8, 9]. From the CISMeF catalogue and from the French MeSH Browser, the French end-users have already accessed with one click (cross-lingual Infobutton) to the optimized PubMed query (query 3).

Table 1. Current and optimized PubMed queries

	Query 1 (N)*	Query 3 (N) **	Query 4 (%) $	Query 5 (%) $	P (Fisher exact test)
Asthma	99,539	84,345	25 (15102)	60 (416)	0.054
Breast neoplasms	148,438	152,103	25 (24)*	65 (6011)	0.025
Neoplasms	1,925,740	1,944,867	55 (3576)	65 (67542)	0.748
Hypertension	269,983	174,979	0 (94612)	25 (255)	0.047
Coronary disease	168,928	167,335	35 (2190)	30 (1297)	1.000
Lung neoplasms	120,199	122,704	30 (38)	55 (3056)	0.200
Myocardial infarction	144,552	118,509	15 (26366)	50 (536)	0.041
HIV infections	166,490	166,983	35 (573)	85 (1093)	0.003 $$
Liver neoplasms	91,666	91,317	20 (77)	65 (484)	0.009
Skin neoplasms	71,926	71,277	30 (33)	45 (395)	0.514
Total	2,959,484	2,857,971	(142,615)	(80,635)	

* Current PubMed query (number of answers in the PubMed database); ** Optimized query (number of answers in the PubMed database); $ Precision among the Top 20 answers (Number of answers); $$ Fisher exact test: Significant difference applying the Bonferroni correction

Table 2. Overall results of queries 4 & 5

	Query 4	Query 5
Top 20 answers		
Relevant answers	54	109
Total of answers	200	200
Precision	0.270 *	0.545 *
Overall answers		
Relevant answers (method 1)	10,713	51,662
Relevant answers (method 2)	38,506	43,946
Total of answers	142,615	80,635
Recall (method 1)	0.172	0.828
Recall (method 2)	0.467	0.533

* $p < 0.0001$ Fisher exact test

4. Discussion

The default query proposed in this article is significantly more precise that the default query used in PubMed during this study (54.5 vs. 27%)[3]. Furthermore, the estimated recall of the current query is also lower than the estimated recall of the optimized query by two different methods (17.2% vs. 82.8 – method 1 – & 46.7% vs. 53.3% – method 2 –). The optimized version of the query (query 2) provides brand-new citations that are not retrieved by the current query (query 1) because query 1 does not retrieve all the

[3] Since May 2008, the PubMed default query has evolved (NLM Technical Bulletin http://www.nlm.nih.gov/pubs/techbull/mj08/mj08_pubmed_atm_cite_sensor.html). This new Automatic Term Mapping (ATM) is even more noisy (e.g., Old ATM: “gene therapy”[MeSH Terms] OR gene therapy[Text Word] New ATM: “gene therapy”[MeSH Terms] OR (“gene”[All Fields] AND “therapy”[All Fields]) OR “gene therapy”[All Fields].

entry terms. These citations will appear among the first ones that are provided by the PubMed web site. The citations that are retrieved by the current PubMed query and not retrieved by the optimized query are already indexed by NLM indexers. As mentioned in the evaluation section, assessment of indexing quality was performed based on Title and Abstract only, not in the full text of the article. This could be considered to be a bias of this study. However, this is the way that end-users predominantly use PubMed, which contains only titles and abstracts.

In PubMed, this proposed optimization could have some interest for the end-user and could be very easily implemented by the NCBI. The systematic use of synonyms for the [TIAB] and the fact that our default query rejects the references that are not indexed with the MeSH term present in the query is, according to us, more robust conceptually than the current default query. Furthermore, our default query is the same if the MeSH term is a preferred term or not. Finally, it adds more precision because the proposed new default query also uses the synonyms of the MeSH term (entry terms) or the synonyms of the MeSH supplementary concepts.

If a complete consensus among indexers existed (NLM indexers vs. the two indexers of this study), the global precision of query 4 should be 0. The global precision of query 4 reaches 27%. In these cases, the two authors would have indexed with the MeSH term that was not used by US NLM indexers. However, inter-expert variability among indexers is known to be quite low [10]. We are planning to evaluate in the near future another default query where [TIAB] will be replaced by [Title]. This will certainly optimize the precision but decrease the recall.

5. Conclusion

The optimized query proposed in this study would be easy to implement into PubMed. It is more conceptually robust and would increase the precision of the searches, thus helping the increasing number of users of PubMed not familiar with the MeSH thesaurus.

References

[1] Nelson, S.J., Johnson, W.D., Humphreys, B.L. (2001) Relationships in medical subject headings. In Bean, C.A., Green, R. (Eds.) *Relationships in the Organization of Knowledge*. Kluwer Academic Publishers, New York, 171–184.

[2] Herskovic, J.R., Tanaka, L.Y., Hersh, W., Bernstam, E.V. (2007) A day in the life of PubMed: Analysis of a typical day's query log. *Journal of the American Medical Informatics Association* 14(2):212–220.

[3] http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helppubmed.section.pubmedhelp.Displaying_the_Search.

[4] http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html.

[5] http://www.nlm.nih.gov/mesh/meshrels.html.

[6] Douyère, M., Soualmia, L.F., Névéol, A., Rogozan, A., Dahamna, B., Leroy, J.P., Thirion, B., Darmoni, S.J. (2004) Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. *Health Information and Libraries Journal* 21(4):253–261.

[7] http://www.cismef.org & http://www.chu-rouen.fr/cismef.

[8] http://www.chu-rouen.fr/terminologiecismef/.

[9] Thirion, B., Pereira, S., Névéol, A., Dahamna, B., Darmoni, S.J. (2007) French MeSH browser: A cross-language tool to access MEDLINE/PubMed. *AMIA Annual Symposium Proceedings 2007*, 1132.

[10] Funk, M.E., Reid, C.A., McGoogan, L.S. (1983) Indexing consistency in MEDLINE. *Bulletin of the Medical Library Association* 2(71):176–183.