Original Paper

Finding the best semantic expansion to query PubMed: automatic performance assessment of four search strategies on all MeSH descriptors.

Abstract

Background: With the continuous expansion of available biomedical data, efficient and effective information retrieval has become of utmost importance. Semantic expansions of queries using synonyms may improve information retrieval.
Objective: The aim of this study was to propose an innovative method that could estimate automatically the three main metrics used in information science (precision, recall and F-measure) of four different semantic expansion strategies, assessed on all the descriptors in the MeSH thesaurus (n=28,313).
Methods: Four search strategies were assessed in this study: the standard Automatic Term Mapping (ATM) of PubMed and three strategies that use semantic expansion. ATM queries were of the form: (*"preferred term"[MH] OR "preferred term"[All fields]*). The queries of the other three strategies were of the form: (*"preferred term"[MH] OR "preferred term"[All fields] OR "synonym 1"[All fields] OR "synonym 2"[All fields] OR etc.)*. These three strategies differed by the number and provenance of the synonyms used to build the queries: MeSH synonyms, UMLS mappings and custom mappings (CISMeF). Metrics were assessed by computing (A): the number of all relevant citations, using NLM indexing as gold standard ((*"preferred term"[MH]),* (B): the number of citations retrieved by the added terms, and (C): the number of relevant citations retrieved by the added terms (combining the previous two queries with an "AND" operator). Therefore, it was possible to compute programmatically the metrics for each strategy using every MeSH descriptor as a "preferred term". 239,724 different queries were built and sent to PubMed API. The four search strategies were ranked and compared for each metric.
Results: ATM had the worst performance of the four strategies, for all three metrics. MeSH strategy had the best mean precision (50.93 %, SD = 23 %). UMLS strategy had the best recall and F-measure (40.57 %, SD = 31 % and 35.51 %, SD = 24 %, respectively). CISMeF had the second best recall and F-measure (40.11 %, SD = 31 % and 35.10 %, SD = 24 %, respectively). However, considering a cut-off of 5%, CISMeF had better precision than UMLS for 1180 descriptors, better recall for 793 descriptors and better F-measure for 678 descriptors.
Conclusions: This study highlights the importance of using semantic expansion strategies to improve information retrieval. However, the performances of the strategies were greatly variable depending on the MeSH descriptor, for each metric. These results confirm there is no ideal search strategy for all descriptors. Different semantic expansions should be used depending on the descriptor and the user's objectives. This led our team to develop an interface that allows users to input a descriptor and then proposes the best semantic expansion to maximize the three main metrics, precision, recall or F-measure.

Keywords: Bibliographic database; Information retrieval; Literature searching; Medical Subject Headgins (MeSH); MEDLINE; PubMed; Precision; Recall; Search Strategies; Thesaurus

Introduction

Background

Since the publication of the first model of a modern scientific journal in 1665 ("*Le Journal des Sçavans*") it is estimated that the number of scholarly articles in existence exceeded 50 million in 2009 [1]. Since April 2018, the PubMed® search engine, a service of the U.S National Library of Medicine (NLM®), provides access to over 28 million citations for biomedical literature, of which more than 26 million are from the MEDLINE® database. The number of citations added to MEDLINE each year now exceeds 1 million (1,178,360 citations added in 2016) [2]. With this continuous expansion of available biomedical data, efficient and effective information retrieval has become of utmost importance. PubMed is one of the most used tools to access these data, and its popularity is growing steadily each year: from 2.5 billion searches performed in 2013 to 3.3 billion in 2017 [3].

However, numerous studies have reported that users lack search skills for the effective use of PubMed [4–6]. Although a basic search using PubMed can be relatively straightforward, a deeper understanding of its structure and underlying search algorithm is needed to perform an effective search of the literature. In order to improve the accuracy of information retrieval, MEDLINE citations are indexed in the Medical Subject Headings (MeSH®) thesaurus [7], but most users do not know it well enough and do not commonly use its descriptors to build their queries [8, 9]. Moreover, users rarely employ search tags, therefore not fully exploiting the features of PubMed [10]. Consequently, the NLM has developed an automatic process to modify users' explicit queries: Automatic Term Mapping (ATM). Entry terms are mapped to their corresponding MeSH descriptors and compound words are broken down and combined with Boolean operators "AND" and "OR" and searched with the tag [All fields] [11].

The purpose of ATM is to improve information retrieval, but several studies have proposed alternative processes to enhance users' queries that have yielded better results [12–15]. They consist in performing semantic expansions with synonyms of the entry terms. These processes differ in the Knowledge Organisation System (KOS) they use to perform the expansions. The MeSH thesaurus, developed by the NLM, is a list of descriptors covering the biomedical field. Each descriptor has a preferred term and may have some synonyms. In 2009, Thirion et al. proposed the expansion of queries with MeSH synonyms. This optimization led to a significant improvement in the performances of information retrieval [14]. In 2012, Griffon et al. proposed using the Unified Medical Language System (UMLS®) to perform the expansion [15], leading to a slight increase in recall but a decrease in precision.

MeSH and UMLS expansion strategies were implemented in the InfoRoute tool [16], largely inspired by Cimino's Infobutton [17]. The InfoRoute algorithm allows contextual information retrieval across multiple medical websites in English and French, including PubMed, by generating links-queries.

Prior work

In order to improve information retrieval, our team (physicians and terminology specialists) has developed new mappings between the terms of the 71 KOS available in the HeTOP (Health Terminology/Ontology Portal) cross lingual terminology server, of which 54 are not included in UMLS [18]. These mappings, also known as CISMeF (French acronym, Catalogue et Index des Sites Médicaux de langue Française) mappings, were created by aligning terms describing the same concept, either manually (manual mappings) or automatically using UMLS or Natural Language Processing and then validated by a human (supervised mappings). The performance of this new kind of semantic expansion was manually assessed in a previous study by Massonnaud et al. (not available for citation at the time of publication of this manuscript).

Although these different strategies [14,15] have provided a significant improvement in terms of the effectiveness of information retrieval, there are limitations to the assessment of their performances. Assessments were performed manually, allowing only small samples of descriptors and citations. Moreover, all the studies revealed a great variability of results depending on the descriptor used. This behaviour suggested there was no semantic expansion that would be optimal for all descriptors, but rather that the semantic expansion to be used should be chosen according to the specific descriptor and the user's objective; i.e. when seeking either better precision or recall or a harmonic mean view using F-measure.

Objective

The aim of this study was to propose an innovative method that could estimate automatically the three main metrics used in information science: precision, recall and F-measure, of multiple search strategies. Four different semantic expansions were assessed on all the descriptors (n=28,313) in the MeSH thesaurus, in order to provide users the opportunity to choose the strategy that would best fit their needs.

Methods

Data collection

Four search strategies were assessed in this study. The standard ATM of PubMed and three kinds of queries enhanced with semantic expansions: MeSH, UMLS and CISMeF. Essentially, all these queries are of the form: (*"preferred term"[MH] OR "preferred term"[All fields] OR "synonym 1"[All fields] OR "synonym 2"[All fields] OR etc.).* [MH] being the tag for the MeSH terms field. The objective was to evaluate the

added value of *"preferred term"[All fields]*, in the case of ATM, and the added value of the synonyms, in the case of MeSH, UMLS and CISMeF expansions.

Precision was defined as the fraction of relevant citations among the retrieved citations and recall as the fraction of relevant citations retrieved from the total amount of relevant citations. Therefore, in order to estimate automatically these metrics for a given query, it was necessary to identify (A): the set of all relevant citations, (B): the set of retrieved citations and (C): the set of relevant citations retrieved. The set of all relevant citations (A) was defined using NLM's indexing as gold standard. For a query built from a particular MeSH descriptor, a citation was considered relevant if it was indexed with that same descriptor. Therefore, the total number of relevant citations (A) was retrieved via the query: "*preferred term"[MH].* For each search strategy, the number of citations retrieved (B) was computed with the added terms: (*"preferred term"[TIAB])* for ATM, ("*synonym 1"[TIAB] OR "synonym 2"[TIAB] OR etc.)* for the three other strategies. Consequently, the number of relevant citations retrieved (C) could be computed by combining the previous two queries with an "AND" operator as follows: *("preferred term"[MH] AND "preferred term"[TIAB]) for ATM, ("preferred term"[MH] AND* ("*synonym 1"[TIAB] OR "synonym 2"[TIAB] OR etc.))* for the three other strategies. The tag *[All fields]* was replaced with *[TIAB]* since *[All fields]* also searches the indexation field of the citations, therefore conflicting with the *[MH]* tag. Moreover, the scope of the search was reduced to indexed citations by adding the tag *medline[sb]* so that all queries were performed on the same set of manually indexed citations. Table 1 shows a summary of the syntax of the resulting nine different queries.

Table 1. Summary of the syntax of the nine queries used in this study

strategy	relevant citations (A)	retrieved citations (B)	relevant citations retrieved (C)
ATM	"pref. term"[MH]	"pref. term"[TIAB] AND medline[sb]	"pref. term"[MH] AND ("pref. term"[TIAB]) AND medline[sb]
MeSH	"pref. term"[MH]	("MeSH synonym 1"[TIAB] OR "MeSH synonym 2"[TIAB] OR …) AND medline[sb]	"pref. term"[MH] AND ("MeSH synonym 1"[TIAB] OR "MeSH synonym 2"[TIAB] OR …) AND medline[sb]
UMLS	"pref. term"[MH]	("UMLS synonym 1"[TIAB] OR "UMLS synonym 2"[TIAB] OR …) AND medline[sb]	"pref. term"[MH] AND ("UMLS synonym 1"[TIAB] OR "UMLS synonym 2"[TIAB] OR …) AND medline[sb]
CISMeF	"pref. term"[MH]	("CISMeF synonym 1"[TIAB] OR "CISMeF synonym 2"[TIAB] OR …) AND medline[sb]	"pref. term"[MH] AND ("CISMeF synonym 1"[TIAB] OR "CISMeF synonym 2"[TIAB] OR …) AND medline[sb]

The HeTOP terminology server [18] provides relations between multiple KOS. Given a particular concept, it is possible to gather automatically the MeSH preferred term of this concept and its synonyms from the KOS of interest. As the 2018 version of the MeSH was not released at the time of this study, the 2017 version was used, containing 28,313 descriptors, of which 26,636 were used at least once for indexing citations. It was then possible to build programmatically the nine different queries (Table 1) for the 26,636 descriptors; i.e. a total of 239,724 queries. ATM behaviour regarding the split of compound words was reproduced exactly. In order to shorten the length of the queries, the terms were set to lowercase and multiple occurrences of the exact same term were removed. The citation count for each of the 239,724 queries was retrieved via PubMed application programming interface. The process time of these 239,724 queries on a microcomputer was around 3h30. Therefore, it is scalable and can be run frequently.

Statistical Analysis

The mean precision, recall and F-measure were computed for the 26,636 descriptors and for each of the four search strategies. The four strategies were ranked and the number of descriptors for which the CISMeF strategy had better results than each of the three other strategies was computed, considering a difference of at least 5% (arbitrary). The metrics were also computed with stratification on the MeSH category. Statistical analysis was performed using R software (version 3.4.3). As the analysis was performed on the entire set of the MeSH descriptors, confidence intervals and p-values were not needed, and therefore not computed.

Results

Table 2 shows the mean precision, recall and F-measure for each of the four search strategies. ATM had the worst performances of the four strategies, for all three metrics. MeSH had the best mean precision (50.93 %, SD = 23 %). CISMeF and UMLS had identical results for precision (49.20 %, SD = 23 %). UMLS had the best recall and F-measure (40.57 %, SD = 31 % and 35.51 %, SD = 24 %, respectively). CISMeF had the second best recall and F-measure (40.11 %, SD = 31 % and 35.10 %, SD = 24 %, respectively).

Table 2. Mean performances of the four search strategies for the 26,636 MeSH descriptors

KOS	Precision, % (sd)	Recall, % (sd)	F-measure, % (sd)
ATM	44.24 (± 24)	31.12 (± 29)	28.41 (± 23)
MeSH	50.93 (± 23)	38.01 (± 31)	34.59 (± 24)
CISMeF	49.20 (± 23)	40.11 (± 31)	35.10 (± 24)
UMLS	49.20 (± 23)	40.57 (± 31)	35.51 (± 24)

CISMeF and UMLS had identical precision for 1007 descriptors, identical recall for 2397 descriptors and identical F-measure for 2721 descriptors. CISMeF and MeSH had identical precision for 2476 descriptors, identical recall for 2 descriptors and identical F-measure for 1876 descriptors. CISMeF and ATM had identical precision for 622 descriptors, identical recall for 8 descriptors and identical F-measure for 1054 descriptors. MeSH, CISMeF and UMLS had identical precision for 8986 descriptors, identical recall for 8714 descriptors and identical F-measure for 9732 descriptors. Table 3 shows the number of descriptors for which CISMeF had better results than each of the three other strategies and vice-versa, by at least 5%.

Table 3**.** Number of descriptors for which CISMeF strategy had better results than each of the three other strategies and vice-versa, by at least 5%

	precision	recall	F-measure
CISMeF better than UMLS	1180	793	678
UMLS better than CISMeF	1017	1262	1140
CISMeF better than MeSH	170	2372	1403
MeSH better than CISMeF	2088	9	669
CISMeF better than ATM	9112	9724	8895
ATM better than CISMeF	4557	2949	2628

The analysis stratified on the category (tree) of the Mesh descriptor revealed the same trends for all three metrics. The best precision was obtained with category C (diseases) by the MeSH strategy (58.16%). MeSH had the best precision for all categories expect category B (organisms) for which ATM had the best precision (data not shown). The best recall was obtained with category B by UMLS (64.28%), which had the best results for 11 out of the 15 categories. CISMeF had the best recall for the remaining four categories: H (Disciplines and Occupations), K (Humanities), L (Information Science) and N (Health Care). ATM had the worst recall for all categories (data not shown). Figure 1 shows the F-measure scores for each of the four strategies depending on the MeSH category of the descriptor. The best F-measure was obtained with category B by UMLS (47.55%). UMLS had the best F-measure for all categories except category H, for which CISMeF had the best score (19.82%), and category I (Anthropology, Education, Sociology and Social Phenomena), for which MeSH had the best F-measure (22.38%).

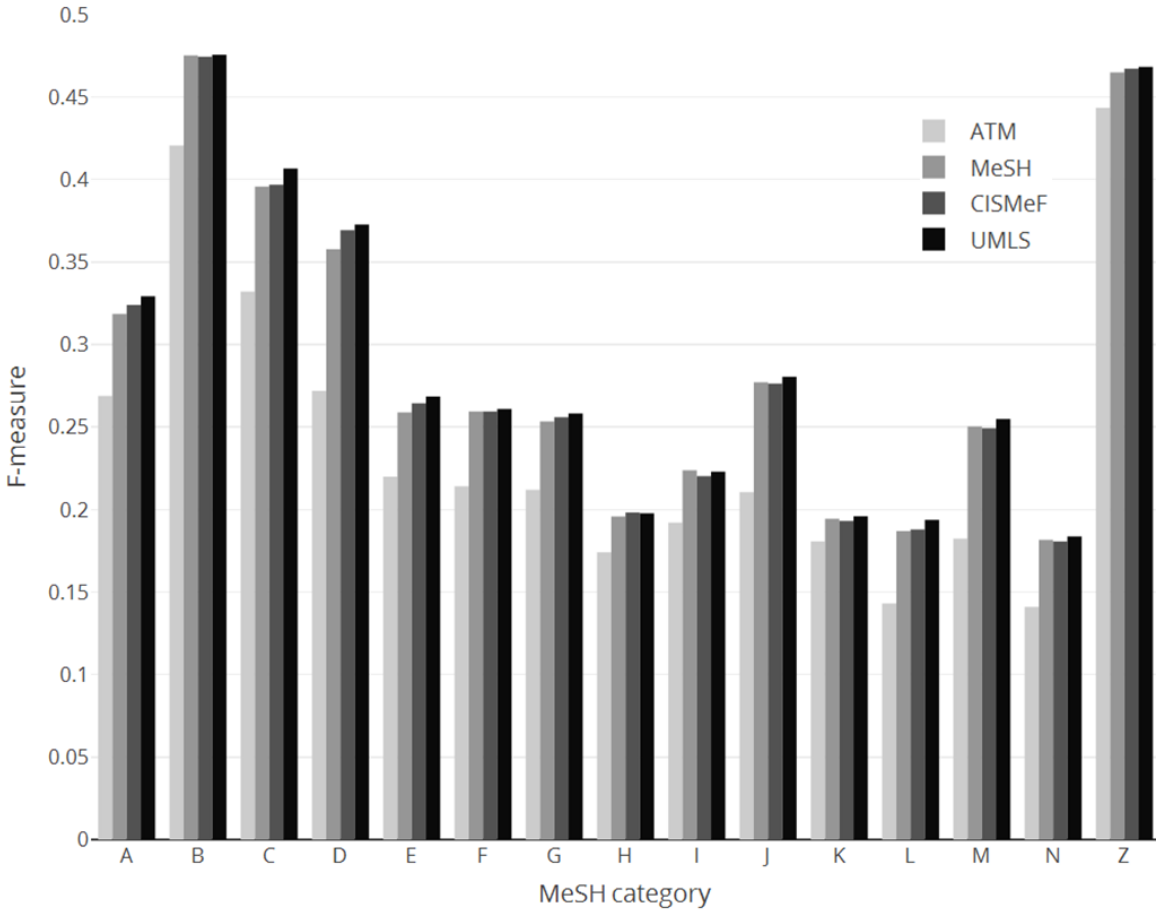

Figure 1. F-measure scores of the four search strategies depending on the MeSH category of the descriptor.

Discussion

Principal Results

Of the four strategies assessed in this study, PubMed's standard ATM had the worst mean performances for the three metrics measured (i.e. precision, recall and F-measure). These results are consistent with previous studies [14, 15]. The best precision was obtained with the MeSH strategy (50.93 %). The mean values of precision for both CISMeF and UMLS strategies were identical (49.20 %). For recall and F-measure, the best performance was obtained by the UMLS strategy, then by CISMeF and then by MeSH.

Even though the differences between the mean performances of the three enhanced strategies (i.e. MeSH, CISMeF and UMLS) were small, with no difference at all for numerous descriptors, this did not reflect the huge variability of the results. Indeed, for each metric, the ranking of the four strategies was greatly dependant on the descriptor. For instance, although UMLS and CISMeF had identical performances for mean precision, CISMeF had better precision than UMLS for 1180 descriptors.

Likewise, the CISMeF strategy had better recall for 793 descriptors and better F-measure for 678 descriptors. Even the ATM strategy, which had significantly lower mean results for all three metrics, was ranked first for several descriptors (data not shown).

The huge variability of the performances found here is consistent with previous studies [14, 15]. This variability was an important limiting factor for these studies since the assessments were performed manually, therefore on a restricted set of descriptors. Consequently, the interpretation of the results was difficult and had important limitations. The objective of this study was to implement and evaluate an original approach for the automatic assessment of the three main metrics of information science. This innovative method allowed us to test the four different strategies of semantic expansion on the entire set of MeSH preferred descriptors (n=28,313), rather than on a small subset. The results found with this new method conform to those of previous studies, with a similar ranking of the four search strategies. These results also confirm the great variability of performances depending on the descriptor. The analysis of different possible confounding factors did not reveal any specific pattern that could explain the variability. The ranking of the four strategies was similar after stratification on the descriptor's category in the MeSH tree. The exact same evaluation was performed with different tags in the queries [19]. The tag **[majr]* was tested instead of **[mh]*, and all strategies were tested with and without the explosion behaviour, which is activated by default [20]. The assessment was also performed over different time intervals. All these different analyses revealed similar ranking of the four strategies and similar variability (data not shown).

These results suggest that the choice of the semantic expansion strategy used to build the query must be made according to the descriptor. Since the automatic assessment tested here allowed assessment of all the MeSH preferred descriptors, it is now possible to choose which semantic expansion strategy to use to build a query for a given descriptor, according to the performances of the three metrics (precision, recall and F-measure). As the process time of this automatic assessment is quite low, it can be updated frequently (each day, each week or each month). Technically, the assessment could also be performed in real time, although this does not seem necessary since the results should not vary greatly during short periods of time.

The availability of quantitative measurements of the performances of different strategies now allows users to decide which semantic expansion to use given a particular MeSH descriptor. Depending on their specific needs, users could either choose the strategy providing best precision, best recall or best F-measure, since these performances could be accomplished by different strategies. These considerations led our team to develop an interface that allows users to input their MeSH preferred descriptor and that displays the performances of the different strategies for all three metrics. Thus, users can choose which strategy to adopt depending on their needs, with the possibility to build and send the query

automatically to PubMed search engine. A perspective could be to go even further in the customization of the queries, with the possibility to add successively each synonym, each time assessing in real time the performances of this custom query.

Limitations

This study has some limitations. First, in order to assess the metrics in an automatic manner, the scope of the search had to be restricted to the indexed citations of the MEDLINE® database. The assessment of recent, non-indexed, citations could only be performed manually, with all the limiting factors previously described in the literature. However, it is legitimate to assume that the different semantic expansions would perform in the same way on the entire database since there is no reason to think that the indexing paradigms would shift suddenly for a given descriptor. Moreover, the results presented here are consistent with manual evaluations of previous studies, suggesting there is no major bias in this new methodology. Secondly, the queries built were simple queries based on only one MeSH preferred term. It would be necessary to evaluate the performances of these different semantic expansions with more complex queries, associating multiple MeSH preferred terms. However, the behaviour of such queries would be identical because the semantic expansion of each term would be treated independently, and then recombined with Boolean operators, which is the default behaviour of PubMed's ATM.

Conclusions

In this study we present an innovative method to automatically compute, on PubMed citations, the three main metrics used in information science. This new method allowed us to compare four semantic expansion strategies to query PubMed on all MeSH descriptors. These results confirm the great variability depending on the descriptor. Hence, the need to propose to users the semantic expansion that best fits their specific objectives. Thanks to the possibility to update regularly the performances of these search strategies on all the MeSH descriptors, our team has developed an interface that allows users to input a descriptor and then proposes the best semantic expansion to maximize either precision, recall or F-measure.

Acknowledgements

The authors are grateful to Nikki Sabourin-Gibbs for her help in editing the manuscript.

Conflicts of Interest

The authors declare that there is no conflict of interest.

Abbreviations

ATM: Automatic term mapping
CISMeF: Catalogue et index des sites médicaux de langue française
HeTOP: Health Terminology/Ontology Portal
KOS: Knowledge organization system
MeSH: Medical subject headings

NLM: National Library of Medicine
UMLS: Unified medical language system

References

1. Jinha AE. Article 50 million: an estimate of the number of scholarly articles in existence. Learn Publ. 2010 Jul 1;23(3):258–63. doi: 10.1087/20100308
2. Yearly Citation Totals from 2017 MEDLINE/PubMed Baseline: 26,759,399 Citations Found [Internet]. Available from: https://www.nlm.nih.gov/bsd/licensee/2017_stats/2017_Totals.html. Archived at: http://www.webcitation.org/73Woae0hy
3. Key MEDLINE® Indicators [Internet]. Available from: https://www.nlm.nih.gov/bsd/bsd_key.html. Archived at: http://www.webcitation.org/73WolvSJp
4. van Dijk N, Hooft L, Wieringa-de Waard M. What are the barriers to residents' practicing evidence-based medicine? A systematic review. Acad Med J Assoc Am Med Coll. 2010 Jul;85(7):1163–70. PMID: 20186032
5. Zwolsman S, te Pas E, Hooft L, Wieringa-de Waard M, van Dijk N. Barriers to GPs' use of evidence-based medicine: a systematic review. Br J Gen Pract J R Coll Gen Pract. 2012 Jul;62(600):e511-521. PMID: 22781999
6. Majid S, Foo S, Luyt B, Zhang X, Theng Y-L, Chang Y-K, et al. Adopting evidence-based practice in clinical decision making: nurses' perceptions, knowledge, and barriers. J Med Libr Assoc JMLA. 2011 Jul;99(3):229–36. PMID: 21753915
7. Nelson S.J., Johnston W.D., Humphreys B.L. (2001) Relationships in Medical Subject Headings (MeSH). In: Bean C.A., Green R. (eds) Relationships in the Organization of Knowledge. Information Science and Knowledge Management, vol 2. Springer, Dordrecht. ISBN:978-90-481-5652-8
8. Herskovic JR, Tanaka LY, Hersh W, Bernstam EV. A day in the life of PubMed: analysis of a typical day's query log. J Am Med Inform Assoc JAMIA. 2007 Apr;14(2):212–20. PMID: 17213501
9. Hoogendam A, Stalenhoef AFH, Robbé PF de V, Overbeke AJPM. Analysis of queries sent to PubMed at the point of care: observation of search behaviour in a medical teaching hospital. BMC Med Inform Decis Mak. 2008 Sep 24;8:42. PMID: 18816391
10. Mosa ASM, Yoo I. A study on PubMed search tag usage pattern: association rule mining of a full-day PubMed query log. BMC Med Inform Decis Mak. 2013 Jan 9;13:8. PMID: 23302604
11. How PubMed works: Automatic Term Mapping [Internet]. Available from: https://www.nlm.nih.gov/bsd/disted/pubmedtutorial/020_040.html. Archived at: http://www.webcitation.org/73WpHYH2O
12. Aronson AR, Rindflesch TC. Query expansion using the UMLS Metathesaurus. Proc Conf Am Med Inform Assoc AMIA Fall Symp. 1997;485–9. PMID: 9357673
13. Hersh W, Price S, Donohoe L. Assessing thesaurus-based query expansion using the UMLS Metathesaurus. Proc AMIA Symp. 2000;344–8. PMID: 11079902

14. Thirion B, Ioana R, J DS. Optimization of the PubMed Automatic Term Mapping. Stud Health Technol Inform. 2009;238–242. PMID: 19745304
15. Griffon N, Chebil W, Rollin L, Kerdelhue G, Thirion B, Gehanno J-F, et al. Performance evaluation of Unified Medical Language System®'s synonyms expansion to query PubMed. BMC Med Inform Decis Mak. 2012 Feb 29;12:12. PMID: 22376010
16. Merabti T, Lelong R, Darmoni S. InfoRoute: the CISMeF Context-specific Search Algorithm. Stud Health Technol Inform. 2015;216:544–8. PMID: 26262110
17. Cimino JJ, Li J, Bakken S, Patel VL. Theoretical, empirical and practical approaches to resolving the unmet information needs of clinical information system users. Proc AMIA Symp. 2002;170–4. PMID: 12463809
18. Grosjean J, Merabti T, Dahamna B, Kergourlay I, Thirion B, Soualmia Lina F, et al. Health Multi-Terminology Portal: A Semantic Added-value for Patient Safety. Stud Health Technol Inform. 2011;129–138. PMID: 21685618
19. PubMed help [Internet]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK3827/. Archived at: http://www.webcitation.org/73WpbIhZ1
20. PubMed tutorial [Internet]. Available from: https://www.nlm.nih.gov/bsd/disted/pubmedtutorial/020_055.html. Archived at: http://www.webcitation.org/73WpgjMtg