Author's response to reviews

Title: A MEDLINE Categorization Algorithm

Authors:

Stefan J Darmoni (stefan.darmoni@chu-rouen.fr)
Aurelie Neveol (aurelie.neveol@insa-rouen.fr)
Jean-Marie Renard (renard@univ-lille2.fr)
Jean-Francois Gehanno (jean-francois.gehanno@chu-rouen.fr)
Lina F Soualmia (Lina.Souamia@chu-rouen.fr)
Badisse Dahamna (Badisse.Dahamna@chu-rouen.fr)
Benoit Thirion (benoit.thirion@chu-rouen.fr)

Version: 2 Date: 26 September 2005

Author's response to reviews: see over
Dear Editor,

Please find enclosed the revised manuscript of our article entitled “A MEDLINE Categorization Algorithm” and the answers to the three reviewers. Aurelie Neveol reviewed the English.
Best regards

Reviewer Stuart Nelson

From the first paragraph of Nelson’s review, we have added in the Discussion section, the following sentences:
“As mentioned by one of the reviewers (SN), the indexing of citations is aimed at indicating topics discussed, not at indicating the persons or specialty to whom the citation might be of interest. This information model has the advantage of not requiring that all possible views of why a citation might be of interest be recognized before indexing. The disadvantage is that a search for information of interest in broad categories is not well supported. The use in MeSH of a specialty name as an index term implies that the article cited is about the specialty, not that the article would be of interest to that specialty. For MEDLINE, MeSH is used to index the most specific aspects of citations. In contrast, some other indexing systems intend to capture the broadest aspects of a citation, or possibly even the disciplines involved.”

As proposed by this reviewer, we have moved the first paragraph of the Results section in the Methods section, creating the following paragraph: “3.3 Creation of the MEDLINE Categorization Web site.”

As proposed by this reviewer, we have also modified the Results section, with the following sentences: “An example of the MEDLINE categorization algorithm is displayed in Figure 3 for the 60 articles published in this journal BioMed Central (BMC) Medical Informatics & Decision Making, which are currently indexed in MEDLINE (March 27, 2005). The top 3 medical specialties for this set of articles are: ‘information science’, ‘organization and administration’ and ‘medical informatics’.”

We have also modified the second bullet in the Discussions section as followed: “To categorize the articles of one journal to check if the main fields of coverage are correlated with the top-rated metaterms generated by the algorithm, as shown in Figure 3. Not surprisingly, among the Top 10 medical specialties calculated by the MCA for the journal BMC Medical Informatics & Decision Making, we found medical informatics (ranked second), information science (ranked first, but since it is located (L01) above medical informatics (L01.700) in the MeSH L tree it is inferred by the algorithm at least as many times as medical informatics, epidemiology (ranked fourth, with a score very close to medical
informatics) and statistics (ranked sixth, and which is also very close to medical informatics)."

As proposed by this reviewer, we have added the following sentence in the Discussion section: “In the near future, we will compare the respective precision of the human-driven metaterm algorithm vs. the machine-driven journal descriptor algorithm on a sample of MEDLINE articles.”

**Reviewer Christian Lovis**

We have taking into account all the Minor Essential Revisions proposed by this reviewer.

**Reviewer Alan Aronson**

As proposed by this reviewer, we have largely modified the definition of metaterms and resource types. We have added the following sentences for the metaterms: “In 1997, the primary use of metaterms was to address the relatively restrictive nature of some MeSH terms in information retrieval. The main objective was to improve recall by using metaterms instead of MeSH terms - therefore expanding the queries submitted to the CISMeF health gateway.”

and for the resource types:

“Internet health resources being more heterogeneous than MEDLINE scientific articles, the CISMeF RTs are more diverse than the publication types (PT) of MEDLINE. Specific RTs are dedicated to electronic health resources, such as association, patient information, community networks, or clinical guidelines. For example, in the case of a clinical guideline about carbon monoxide intoxication, ‘carbon monoxide poisoning’ is the MeSH term and ‘clinical guidelines’ is the resource type. CISMeF RTs are organized similarly to MeSH terms and subheadings, in a hierarchical structure with subsumption relationships (allowing the explode property) and a maximum of five-level depth. The Medline publication types are mainly a flat list (see URL: http://www.nlm.nih.gov/mesh/pubtypes2005.html). The controlled list of RTs is available at the following URL: http://www.chu-rouen.fr/documed/typeeng.html. The RT list has been manually built and maintained by the CISMeF team since 1997. Nonetheless, this list is largely driven from the MeSH thesaurus as 187 RTs (76%) are also MeSH terms (e.g. magnetic resonance imaging) and 28 RTs (11%) are also MEDLINE publication types (e.g. technical report).”
As proposed by the reviewer, we have totally rewritten the Abstract section as followed:

“Abstract:

Background: Categorization is designed to enhance resource description by organizing content description so as to enable the reader to grasp quickly and easily what are the main topics discussed in it.

Objective: To propose a categorization algorithm to classify a set of scientific articles indexed with the MeSH thesaurus, and in particular those of the MEDLINE bibliographic database. In a large bibliographic database such as MEDLINE, finding materials of particular interest to a specialty group, or relevant to a particular audience, can be difficult. The categorization refines the retrieval of indexed material. In the CISMeF terminology, metaterms can be considered as super-concepts. They were primarily conceived to improve recall in the CISMeF quality-controlled health gateway.

Methods: The MEDLINE categorization algorithm (MCA) is based on semantic links existing between MeSH terms and metaterms on the one hand and between MeSH subheadings and metaterms on the other hand. These links are used to automatically infer a list of metaterms from any MeSH term/subheading indexing. The semantic links are manually selected by medical librarians.

Results: The MEDLINE categorization algorithm lists the medical specialties relevant to a MEDLINE file by decreasing order of their importance. The MEDLINE categorization algorithm is available on a Web site. It can run on any MEDLINE file in a batch mode. As an example, the top 3 medical specialties for the set of 60 articles published in BioMed Central Medical Informatics & Decision Making, which are currently indexed in MEDLINE are: information science, organization and administration and medical informatics.

Conclusion: We have presented a MEDLINE categorization algorithm in order to classify the medical specialties addressed in any MEDLINE file in the form of a ranked list of relevant specialties. The categorization method introduced in this paper is based on the manual indexing of resources with MeSH (terms/subheadings) pairs by NLM indexers. This algorithm may be used as a new bibliometric tool.”

We have added in the paper (and in the abstract) that the MEDLINE Categorization Algorithm is applied to classify a set of scientific articles.
To answer to the question of the reviewer: "What are the categories for the previous paper? How many are assigned to each resource?"
We have added the following sentences in the 3.2 Categorization algorithm:

The categories of the MEDLINE categorization algorithm are the metaterms.

The number of metaterms to classify one article is not a priori assigned. The number of metaterms is increasing with the number of MeSH terms (and MeSH terms/subheading pairs) and the number of semantic links between MeSH terms, Subheadings and Metaterms.

To answer to the question of the reviewer:
"How are recall and precision determined? How many relevance judgement types are there beyond "fully relevant" and "fairly relevant"?"
We have modified the following sentences in the Introduction section:

"To test the relevance of the CISMeF Categorization Algorithm (CCA), the automatic categorization obtained was compared on 123 randomly picked resources to the classified list of metaterms (or medical specialties) provided by a CISMeF librarian for each resource and defined as the gold standard. This evaluation gave very satisfying results: 81% precision and 93% recall, and 53% of the resources were assigned by the medical librarian as “fully relevant” or “fairly relevant” categorization according to strict standards (as opposed as 20% of the resources as "partially relevant" and 22% of the resources as "non-relevant")"

As proposed by this reviewer, we have also modified the Objective section as followed:
“The objective of this paper is to propose a modified version of this categorization algorithm to classify a set of scientific articles indexed with the MeSH thesaurus and in particular those of the MEDLINE bibliographic database. Categorization allows a more general description with an upper level of granularity than MeSH indexing. In a large bibliographic database such as MEDLINE, finding materials of particular interest to a specialty group, or relevant to a particular audience, can be difficult. The categorization refines the retrieval of indexed material. This algorithm will be able to categorize the scientific production of one or several scientists or an entire research laboratory. It could also be used to categorize a set of articles of one peculiar journal. The categorization of articles from MEDLINE or scientific journals would characterize their contents by bringing out the medical specialties covered by each source.”

To answer to the question of the reviewer:
3.1 How does the query ‘guidelines in cardiology’ retrieve 11 resources? Using which search engine? This needs to be specified much more carefully. We have modified the following sentences in the 3.1 CISMeF MeSH encapsulated terminology
"To illustrate the difference between MeSH terms and metaterms in terms of information retrieval in the CISMeF health gateway, let us submit the two following sample queries to the Boolean Search of the Doc’CISMeF search engine (URL: http://doccismef.chu-rouen.fr/servlets/Logique): ‘guidelines in cardiology’ or ‘databases in virology’, where ‘guidelines’ and ‘databases’ are CISMeF resource types and ‘cardiology’ and ‘virology’ are viewed alternatively as MeSH terms and CISMeF metaterms. The query ‘guidelines in cardiology’ retrieves 11 resources when ‘cardiology’ is considered as a MeSH term (Boolean query: ‘guidelines.tr AND cardiology.mc’, where tr stands for resource type and mc stands for MeSH term) vs. 143 resources when ‘cardiology’ is considered as a MT (Boolean query: ‘guidelines.tr AND cardiology.mt’, where mt stands for metaterm). The query ‘databases in virology’ retrieves 0 resource when virology is considered as a MeSH term (Boolean query: databases.tr AND virology.mc) vs. 4 resources when virology is considered as a MT (Boolean query: databases.tr AND virology.mt)."

To answer the following question of the reviewer:
"3.1 (and figure 1): The resource types listed in the figure do not seem at all like publication types. If they indeed are resource types, then some explanation needs to be given.", we have added a paragraph explaining the resource types (see above).

To answer to the question of the reviewer: "3.2 How are metaterms automatically inferred? Give an example."
We have added the following sentences in the 3.2 Categorization algorithm section:
"As an example, because the CISMeF librarians have created a semantic link between the MeSH term psychotherapy and the metaterm psychiatry, psychiatry will automatically be inferred for every MEDLINE article indexed with psychotherapy"

As proposed by this reviewer, we have stated what is new in this paper as opposed to the earlier paper in the 3.2 Categorization algorithm as followed:
"There are two main differences between the MEDLINE categorization algorithm and the previous CISMeF categorization algorithm. One concerns the method and one concerns the scope of the categorization:

- Method: The MEDLINE categorization algorithm does not take into account the semantic links between CISMeF metaterms and CISMeF resources types because it will have to restrain to the few resources types which are also MEDLINE publication types (N=28 out of 257, 11%). Using such semantic links would have introduced a major bias. Therefore, we decided not to use them in the MEDLINE categorization algorithm. Furthermore, this strategic choice was driven by the fact that MEDLINE publication types are much less
suitable to categorize medical specialties when compared to CISMeF resources types (e.g. CISMeF RT *lecture* notes for CISMeF MT *medical education* or CISMeF RT *echography* for CISMeF MT *medical imaging*). Finally, only the following six metaterms could have been linked to the CISMeF resources types which are also MEDLINE publication types (N=9 out of 28):

- the CISMeF MT *biostatistics* with CISMeF RT (and MEDLINE PT) *meta-analysis*
- the CISMeF MT *information science* with CISMeF RT (and MEDLINE PT) *database*
- the CISMeF MT *Evidence-Based Medicine* with CISMeF RT (and MEDLINE PT) *consensus development conferences*
- the CISMeF MT *Medical Law* with CISMeF RT (and MEDLINE PT) *Legislation*
- the CISMeF MT *medical education* with CISMeF RTs (and MEDLINE PTs) *examination questions, instruction motion picture & problems and exercises*
- the CISMeF MT *patient* with CISMeF RTs (and MEDLINE PTs) *patient education handout & popular works*

The other 19 shared CISMeF resources types / MEDLINE publication types (e.g. *portrait* or *table*) have no semantic link with any of CISMeF metaterm.

To answer the following question of the reviewer:
3.2 In the text beginning "Assume there are: ..." what is the scope of the computations?
We have modified the following sentences in the 3.2 *Categorization algorithm section*
"Assume there are; in a set of MEDLINE articles to be categorized:"

As proposed by this reviewer, we have clearly stated in the 3.2 Section that:
"The categories of the MEDLINE categorization algorithm are the metaterms."

As proposed by this reviewer, we have added the following paragraph to explain what is the purpose of semantic link in the 3.1 section:
"A semantic link between a CISMeF term (MeSH terms, MeSH subheadings, or CISMeF resource types) and a CISMeF metaterm means that the CISMeF term is related to the concept denoted by the CISMeF metaterm. Therefore, an article indexed with the CISMeF term can be categorized by the corresponding metaterm: e.g. the MeSH term *psychiatric somatic therapies* is linked to the CISMeF metaterm *psychiatry* (see Table 1). There is a 0 to many relation between CISMeF terms and CISMeF metaterms: e.g. the MeSH term ‘Paris’ has no semantic link with any of the CISMeF metaterms. On the contrary, the MeSH term *acquired*
immunodeficiency syndrome is linked to two different metaterms: `virology' and `allergy and immunology'"

To answer the following question of the reviewer:
3.2 Clarify the example on the 'Method:' subsection involving "biostatistics" and "meta-analysis". It is not at all clear.

We have added the following sentences in the 3.2 Categorization algorithm section

- **Method:** The MEDLINE categorization algorithm does not take into account the semantic links between CISMeF metaterms and CISMeF resources types because it will have to restrain to the few resources types which are also MEDLINE publication types (N=28 out of 257, 11%). Using such semantic links would have introduced a major bias. Therefore, we decided not to use them in the MEDLINE categorization algorithm. Furthermore, this strategic choice was driven by the fact that MEDLINE publication types are much less suitable to categorize medical specialties when compared to CISMeF resources types (e.g. CISMeF RT lecture notes for CISMeF MT medical education or CISMeF RT echography for CISMeF MT medical imaging). Finally, only the following six metaterms could have been linked to the CISMeF resources types which are also MEDLINE publication types (N=9 out of 28):
  - the CISMeF MT biostatistics with CISMeF RT (and MEDLINE PT) meta-analysis
  - the CISMeF MT information science with CISMeF RT (and MEDLINE PT) database
  - the CISMeF MT Evidence-Based Medicine with CISMeF RT (and MEDLINE PT) consensus development conferences
  - the CISMeF MT Medical Law with CISMeF RT (and MEDLINE PT) Legislation
  - the CISMeF MT medical education with CISMeF RTs (and MEDLINE PTs) examination questions, instruction motion picture & problems and exercises
  - the CISMeF MT patient with CISMeF RTs (and MEDLINE PTs) patient education handout & popular works

The other 19 shared CISMeF resources types / MEDLINE publication types (e.g. portrait or table) have no semantic link with any of CISMeF metaterm.

As proposed by this reviewer, we have added the following paragraph to discuss Susanne Humphrey's work on journal descriptor indexing:

"CISMeF MTs can be viewed as quite similar to JDs (Journal Descriptors) used for indexing journals per se (see URL: [www.nlm.nih.gov/tsd/serials/jdi.html](http://www.nlm.nih.gov/tsd/serials/jdi.html); List of journals indexed for Medline, 2005). Humphrey has developed the JDI (Journal Descriptor Indexing) system based
on the statistical associations between JDs and text words or starred (major) MeSH terms. In
the near future, we will compare the respective precision of the human-driven metaterm
algorithm vs. the machine-driven journal descriptor algorithm on a sample of MEDLINE
articles. After this study, we will collaborate with Humphrey to improve semantic links
between MTs and MeSH terms, using statistical associations between JDs and MeSH terms.
We also plan to use statistical associations between JDs and MeSH terms to improve
information retrieval in the CISMeF search engine, limiting the scope of a query by proposing
the most frequently associated JDs (e.g. for the query ‘asthma’, the system will suggest
restricting its scope to the JDs ‘Critical care’ or ‘Pediatrics’).

Finally, we have taking into account all the Minor Essential Revisions proposed by this
reviewer.