Vocabulary Density Method for Customized Indexing of MEDLINE Journals

James G. Mork, MSc, Dina Demner-Fushman, MD, PhD, Susan C. Schmidt, MLS, Alan R. Aronson, PhD

Abstract

Automated indexing of MEDLINE citations remains a challenging problem due to the growing volume of citations and the over 27,000 MeSH indexing terms that can be assigned to them. This paper presents a corpus-based approach to improving indexing for specific journals. The Vocabulary Density approach takes into account frequencies of indexing terms previously assigned to a journal when recommending indexing terms for a new citation in that journal. After implementing the approach, we saw a 2.69 (4.44%) improvement in Precision with no loss in Recall.

Introduction

The successes in automatic indexing of MEDLINE® citations using the NLM Medical Text Indexer (MTI)[1] have led to First Line (MTIFL) indexing of several journals. MeSH terms automatically assigned by MTIFL provide the initial indexing to a MEDLINE citation which is then reviewed and completed by an indexer. As we expand the set of journals indexed via MTIFL, we continue seeking improvements to the MTI algorithms. Potential for improvement using journal-specific data was discussed by Tsoumakas et al[2]. To explore whether customizing MTIFL indexing of a journal is worthwhile, we have implemented a simple Vocabulary Density approach for all journals indexed by MTI.

Methods

We used 3,401,111 citations involving 6,606 journals from the 2014 MEDLINE Baseline that have been indexed over the last five years (henceforth referred to as *Corpus*). For each MeSH Heading (MH) used by each journal we captured the number of its occurrences (NOM) and the Number of Articles in the journal (NOA). We then normalized the frequency of each MH in the journal, computing Factor = NOM / NOA. For example, the MH "Swiss 3T3 Cells" occurred four times in the 2,231 articles of the journal "Biochemical Society (Great Britain)" in the *Corpus*. The Factor for this MH is 0.001793 (4 / 2231).

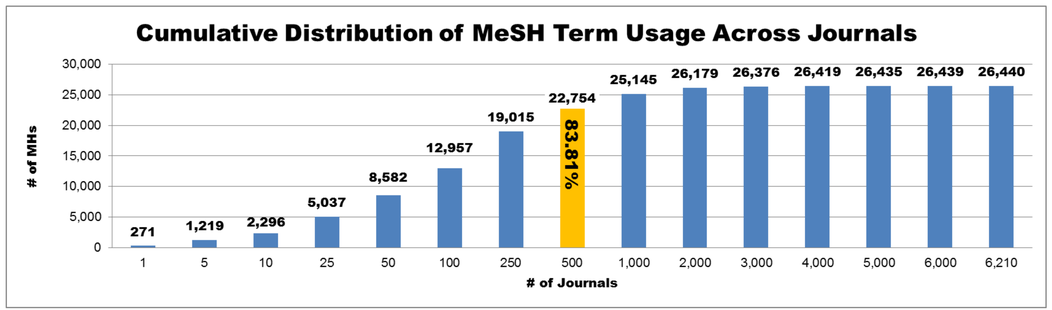
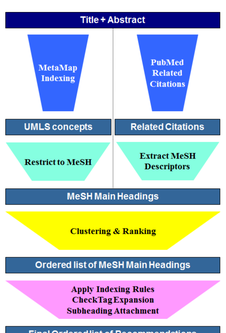


Figure 1. Cumulative Distribution of MeSH Term Usage Across Journals

We applied the Vocabulary Density method to journals that had at least 80 citations in the *Corpus* and to MHs introduced at least a year ago. Given the Vocabulary Density information, MTI does not recommend MHs that are not used for the journal and automatically recommends MHs with a Factor > 0.74. For frequently occurring MHs, e.g., Female or Humans, the threshold for automatically recommending is 1 to reduce incorrect recommendations.



MTI Processing Flow

Results

On average, only 999 unique MHs of the 27,149 available in 2014 MeSH are used per journal in the 6,606 journals in our *Corpus*. 83.81% of the used MHs are found in 500 or fewer journals and 271 MHs are only found in a single journal (see Figure 1). This selective use of MHs confirms the intuition that taking into account journal-specific data can lead to improvements in MTI recommendations. **Furthermore, implementing this simple approach leads to a 2.69 (4.44%) improvement in Precision, 1.36 (2.23%) increase in F_1 score, and a 0.05 (0.08%) increase in Recall.**

Conclusion

The significant improvements in Precision without losses in Recall when taking into account the Vocabulary Density information for a journal shows that exploring other potential approaches to using the journal-specific data is worthwhile.

References

1. James G Mork, Antonio J Jimeno-Yepes, Alan R Aronson. The NLM Medical Text Indexer system for indexing biomedical literature. BioASQ Workshop, Valencia, Spain, September 27, 2013.
2. Tsoumakas G, Laliotis M, Markantonatos N, VlahavasI. Large-scale semantic indexing of biomedical publications at BioASQ. BioASQ Workshop, Valencia, Spain, September 27, 2013.