Using Topic Models to Interpret MEDLINE's Medical Subject Headings

David Newman[1,2], Sarvnaz Karimi[1], and Lawrence Cavedon[1]

[1] NICTA and The University of Melbourne, Victoria, Australia
[2] University of California, Irvine, USA
{david.newman,sarvnaz.karimi,lawrence.cavedon}@nicta.com.au

Abstract. We consider the task of interpreting and understanding a taxonomy of classification terms applied to documents in a collection. In particular, we show how unsupervised topic models are useful for interpreting and understanding MeSH, the Medical Subject Headings applied to articles in MEDLINE. We introduce the resampled author model, which captures some of the advantages of both the topic model and the author-topic model. We demonstrate how topic models complement and add to the information conveyed in a traditional listing and description of a subject heading hierarchy.

1 Introduction

Topic modeling is an unsupervised learning method to automatically discover semantic topics in a collection of documents and allocate a small number of topics to each individual document. But in many collections, documents are already hand-categorised using a human-constructed taxonomy of classification terms or subject headings. We report on a number of experiments that use topic modeling to interpret the meaning of categories, and explain subtle distinctions between related categories, by analysing their use over a document collection. These experiments are performed in the context of the Medical Subject Headings (MeSH) taxonomy.

MeSH are the subject headings used for tagging articles in MEDLINE, the largest biomedical literature database in the world. PubMed – the interface for searching MEDLINE – extensively uses these MeSH headings. Most PubMed queries are mapped to queries that involve MeSH headings, e.g. the query "teen drug use" is mapped to a longer query that searches for the MeSH headings "Adolescent" and "Substance-Related Disorders" (this mapping is explained in [1]). Therefore, it is critical for researchers and health-care professionals using PubMed to understand what is meant by these MeSH headings, since MeSH headings have a direct effect on search results.

One possible approach would be to attempt to understand MeSH headings by analysing how MeSH headings are applied to documents. However, MeSH tagging is a complex procedure performed by a team of expert catalogers at the National Library of Medicine in the US[1]. These catalogers use a range of

[1] MeSH tagging is described in detail at http://ii.nlm.nih.gov/mti.shtml

A. Nicholson and X. Li (Eds.): AI 2009, LNAI 5866, pp. 270–279, 2009.

Table 1. Most frequent MeSH headings, major MeSH headings, major qualifiers and MeSH-qualifier combinations in articles published since 2000

MeSH heading	Major MeSH heading	Major qualifier	MeSH-qualifier combination
Humans	Brain	metabolism	Signal Transduction (physiology)
Female	Breast Neoplasms	physiology	Antineoplastic Combined Chemotherapy Protocols (therapeutic use)
Male	Neoplasms	genetics	Magnetic Resonance Imaging (methods)
Animals	Apoptosis	methods	Apoptosis (drug effects)
Adult	HIV Infections	chemistry	Neurons (physiology)
Middle Aged	Neurons	pharmacology	DNA-Binding Proteins (metabolism)
Aged	Signal Transduction	therapeutic use	Transcription Factors (metabolism)
Adolescent	Antineoplastic Agents	pathology	Antineoplastic Agents (therapeutic use)
Mice	Magnetic Resonance Imaging	immunology	Anti-Bacterial Agents (pharmacology)
Child	Anti-Bacterial Agents	diagnosis	Brain (metabolism)

techniques, leveraging various biomedical resources and ontologies, and applying machine learning tools that score and suggest MeSH categories for a given document.

We take a statistical approach to this analysis, using topic models of large sets of search results over MEDLINE, to provide a semantic interpretation of MeSH headings. By analyzing large scale patterns of MeSH tagging, and patterns of co-occurring words in titles and abstracts, we independently learn the meaning of MeSH terms in a data-driven way. We argue that this leads to an understanding of the way MeSH headings have been applied to the MEDLINE collection, providing insight into distinctions between headings, and suggesting MeSH terms that can be useful in document search. While this paper focuses on MEDLINE and MeSH, the approach is more broadly useful for any collection of text documents that is tagged with subject headings.

Background on MeSH Headings: MeSH headings are arranged in a large, complex and continually evolving hierarchy. Currently there are over 25,000 MeSH terms arranged in a directed acyclic graph, which includes a root and 11 levels. On average there are 16 MeSH headings attached to a MEDLINE article. All MeSH tags on a given article have an additional attribute MajorTopicYN which can take on the value Y or N, indicating whether the MeSH tag is the primary focus of the article. Furthermore, each application of a MeSH tag on an article may be qualified using zero, one, or more qualifiers, e.g. one could qualify the MeSH tag *Methadone* with the qualifier *therapeutic use*. There are over 80 qualifiers, but only a specific subset of qualifiers may be used with each MeSH heading. Qualifiers applied to articles also always have the attribute MajorTopicYN.

To gain some familiarity with the usage of MeSH headings and qualifiers, we provide lists of most frequent terms in Table 1. Rows in the table do not correspond – the four columns are separate. The first column shows the most frequent MeSH headings, irrespective of MajorTopicYN. We see headings that act as "check tags" (e.g. Human), used to restrict search results to certain classes of interest. The second column shows the most common *major* MeSH headings, where the heading or one of its qualifiers has MajorTopicYN=Y. Here we see a broad range of topics, covering both conditions/diseases (Neoplasms, HIV) and