MedKit: A Helper Toolkit for Automatic Mining of MEDLINE/PubMed Citations

Jing Ding

Department of Electrical and Computer Engineering,
Iowa State University, Ames, IA 50011, USA

Daniel Berleant[*]

Department of Electrical and Computer Engineering,
Iowa State University, Ames, IA 50011, USA

Running head: MEDLINE/PubMed Toolkit

[*] To whom correspondence should be addressed

ABSTRACT

Summary: MEDLINE/PubMed is one of the most important information sources for bioinformatics text mining. However, there remain limitations in working with MEDLINE/PubMed citations. For example, PubMed imposes an upper limit of 10,000 for downloading PMID list or citations; and MEDLINE files are too large for most off-the-shelf XML parsers. We developed a Java package, MedKit, to work-around the limitations, as well as provide other useful functionalities, e.g. random sampling. Its four modules (querier, sampler, fetcher and parser) can work independently, or be pipelined in various combinations. It can be used as a stand-alone GUI application, or integrated into other text mining systems. Text mining researchers and others may download and use the toolkit free for non-commercial purposes.

Availability: http://metnetdb.gdcb.iastate.edu/medkit

Contact: berleant@iastate.edu, dingjing@iastate.edu

INTRODUCTION

MEDLINE (National Library of Medicine, 2004b) is a standard literature database for bioinformatics text mining. It can be accessed through annual releases in XML format (with weekly updates) or through its web interface, PubMed (National Library of Medicine, 2004c). Both access methods have limitations, however. For example, the MEDLINE release files are intended for automatic processing. Although the XML format is easy for navigation and manipulation within a file, the files are too large (average size: ~100 MB) for most off-the-shelf XML parsers. Despite the popularity of XML, some text-mining libraries, e.g. the “Bow” toolkit (McCallum, 1996), still take plain text files as input. It is also difficult to generate a subset of citations directly from MEDLINE release files in response to a user query, which causes users to turn to PubMed. PubMed is more focused on human users. Its query system is designed to return a manageable set of relevant documents. Its upper limit for downloading query results (currently 10,000 hits) can be a hindrance to automated text-mining systems. In addition, the MEDLINE XML format and the PubMed XML format are not identical.

To work around these limitations, as well as add other useful functionalities (e.g. to randomly sample a subset of MEDLINE/PubMed abstracts), we developed a Java package, MedKit. It can be used as a stand-alone GUI application, or its modules may be integrated into other automated MEDLINE/PubMed mining systems.

PROGRAM OVERVIEW

MedKit integrates four modules (Java classes), a *querier*, a *sampler*, a *fetcher*, and a *parser*. They may be used together through the interface of Figure 1, or incorporated in any combinations into other programs. The querier takes a query (keyword terms plus other conditions, such as publication dates, fields, etc.) as input and returns a list of PMIDs. The number of returned PMIDs has no upper limit. The sampler draws random samples from a list of PMIDs. The fetcher retrieves citations from PubMed given a list of PMIDs. Like the querier, the number of retrieved citations has no upper limit. The parser can parse very large MEDLINE/PubMed XML files, split them into small ones, extract PMIDs, and/or extract abstracts into plain text files (not limited by the size of input files, but by available disk space for output).

Figure 1

Common mining-related tasks in working with MEDLINE/PubMed citations can be disassembled into atomic steps, and carried out by individual modules in a step-by-step style, saving the output of each step as the input to the next. Alternatively, the modules can be pipelined in various combinations to carry out more sophisticated tasks in a single run. For example, the task of retrieving three random samples of 50 citations each of PubMed abstracts in compressed XML format mentioning “*red blood cell*” in MeSH terms published in the last 5 years, can be accomplished by a pipeline of “querier → sampler → fetcher” (Fig. 1). Other valid pipeline workflows, their inputs and outputs are shown in Table 1.

Table 1

The querier and fetcher take advantage of NCBI Entrez Utilities’ ESearch and EFetch services, respectively (National Library of Medicine, 2004a). In other words, MedKit simply passes users’ queries to PubMed. Therefore, any legal PubMed boolean queries can be used, for example, *red blood cell[text word] AND review[publication type]*. It is, however, not our intention to build another MEDLINE interface to compete with PubMed. On the contrary, MedKit is designed to compliment PubMed. PubMed also provides other facilities (i.e. Limits, History and Clipboard) to enhance query capability and efficiency. The results of direct PubMed queries (PMID lists and/or XML citations) can be saved locally, and then used as input to MedKit for further processing, e.g. parsing and/or sampling.

The parser works around the file size limitation without sacrificing performance (parse medline04n001.xml.gz containing 30,000 citations in 82 sec on a Pentium II 500Hz machine with 384 MB memory running Windows 2000 and Sun’s JRE 1.4.2). It is done by combining a regular Java file reader with open source XML libraries, dom4j (MetaStuff Ltd., 2004) and Piccolo (Yuval Oren, 2004). A MEDLINE/PubMed XML file is first opened as a plain text file by the regular file reader, and read into memory in small chunks. A chunk of text, containing exactly one citation unit from start tag to end tag, is then passed to the XML parser. After the citation is processed, the next chunk is passed to the parser, and the previous one is discarded. Thus, the MedKit parser is able to process very large files in a stream-like fashion while retaining the convenience and flexibility of XML within a citation unit. This design is based on the observation that most MEDLINE/PubMed text mining systems focus on the information contained within single citations; cross talk among citations is rare.

The sampler's random sampling algorithm is backed by the Colt distribution, open source libraries for high performance scientific and technical computing in Java (European Organization for Nuclear Research (CERN), 2004).

The package is freely available at: http://metnetdb.gdcb.iastate.edu/medkit.

References

European Organization for Nuclear Research (CERN) (2004) The Colt Distribution: Open Source Libraries for High Performance Scientific and Technical Computing in Java. http://hoschek.home.cern.ch/hoschek/colt/

McCallum, A.K. (1996) Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering. http://www-2.cs.cmu.edu/~mccallum/bow/

MetaStuff Ltd. (2004) dom4j. http://www.dom4j.org

National Library of Medicine (2004a) Entrez Utilities. http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html

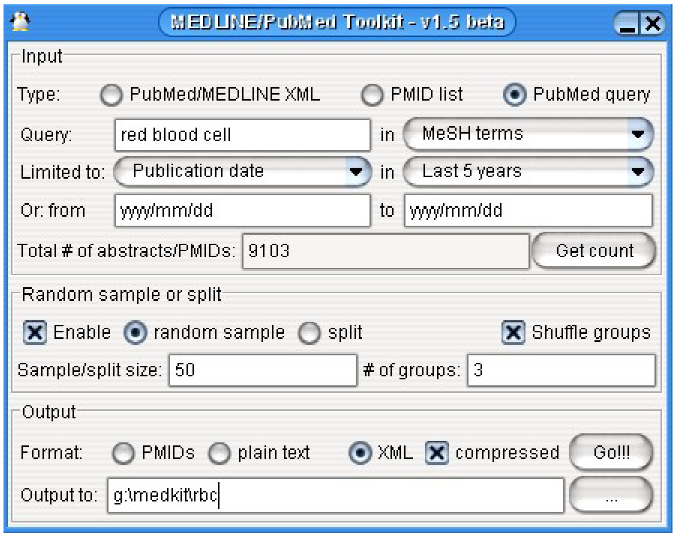
National Library of Medicine (2004b) MEDLINE. http://www.nlm.nih.gov/pubs/factsheets/medline.html

National Library of Medicine (2004c) PUBMED. http://www.ncbi.nih.gov/entrez/query.fcgi

Yuval Oren (2004) Piccolo XML parser. http://piccolo.sourceforge.net/

Table 1. Valid workflows of MedKit. (Q: querier; S: sampler; F: fetcher; P: parser; r: required; o: optional)

Input	Modules				Output
	Q	S	F	P	
Query terms	r	o			PMID list(s)
Query terms	r	o	r	o	XML abstracts
Query terms	r	o	r	r	Plain text abstracts
PMID list		o			PMID list(s)
PMID list		o	r	o	XML abstracts
PMID list		o	r	r	Plain text abstracts
XML abstracts		o		r	XML abstracts
XML abstracts		o		r	PMID list(s)
XML abstracts		o		r	Plain text abstracts



MEDLINE/PubMed Toolkit - v1.5 beta

Input

Type: ○ PubMed/MEDLINE XML ○ PMID list ● PubMed query

Query: red blood cell in MeSH terms

Limited to: Publication date in Last 5 years

Or: from yyyy/mm/dd to yyyy/mm/dd

Total # of abstracts/PMIDs: 9103

Random sample or split

☒ Enable ● random sample ○ split ☒ Shuffle groups

Sample/split size: 50 # of groups: 3

Output

Format: ○ PMIDs ○ plain text ● XML ☒ compressed

Output to: g:\medkit\rbc

Figure 1. A screenshot of MedKit GUI.