Name:	Ben Busby
Affiliation:	National Center for Biotechnology Information, Bethesda, MD
Email:	busbybr@ncbi.nlm.nih.gov
Presentation type (s):	**Poster**

PubRunner: A framework to update biomedical text mining analysis with the latest publications

Kishore R. Anekalla[1] J.P. Courneya[2], Nicolas Fiorini[3], Jake Lever[4], Michael Muchow[5] and Ben Busby[6]

[1]Northwestern University, Chicago, IL, USA [2]Health Sciences and Human Services Library, University of Maryland, Baltimore, MD [3]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD [4]Canada's Michael Smith Genome Sciences Centre, University of British Columbia, Vancouver, BC, Canada [5]National Institute of Standards and Technology, Gaithersburg, MD [6]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD

Abstract

Biomedical text mining offers the possibility of extracting large amounts of knowledge from the millions of published articles. These technologies are becoming essential as knowledge is spread across larger numbers of journals and subject domains. However this extracted knowledge is only valuable if it is based on the latest publications and not kept static. Unfortunately many text mining resources are static and after publication these resources are not maintained. The community needs a method to make it easy to keep these analysis up to date. We present PubRunner, a framework that downloads the latest publications from PubMed, runs whichever text mining tools are needed and uploads the data to a publicly available location. We hope to encourage researchers in our community to use this framework to keep text mining results up-to-date so that they are truly valuable to biologists, and extend this service to other corpora used for NLP.

Introduction

All biology researchers face the substantial challenge of keeping abreast of all appropriate research in their field. This problem is exacerbated by the fact that the number of publications in Pubmed, the largest indexed collection of biomedical abstracts, is increasing at an exponential rate[1]. Text mining tools can be used to help researchers in several ways, including extracting knowledge (such as miRNA regulation data[2]), suggesting potential hypothesis (the FACTA+ system[3]) or assisting in improved PubMed search[4]. However text mining tools are only useful if they provide results with the latest papers. Unfortunately some text mining results are static and never updated. The most common reason is that projects are never revisited once published and the technical costs to keep them updated are too high. This necessitates the development of a framework to make updating results easier. Our approach, PubRunner, manages the updates from PubMed, execution of text mining tools and uploads of data to publicly available locations such as FTPs or Zenodo data hosting. We hope this initative will encourage more sharing of tools and up-to-date data in the biomedical text mining community.

Approach

PubRunner consists of four main components shown in Figure 1. The first component manages the download of Pubmed abstracts from the NCBI FTP website. This intelligently checks for new Pubmed files and only downloads what is required. The second component manages the execution of text mining tools on the downloaded Pubmed data. It checks for appropriate exit status. The third component moves the resulting output data to a publicly available space such as a public FTP or Zenodo permanent hosting location. The fourth component then updates the main PubRunner website with information about what tools have been executed and where the data can be found.

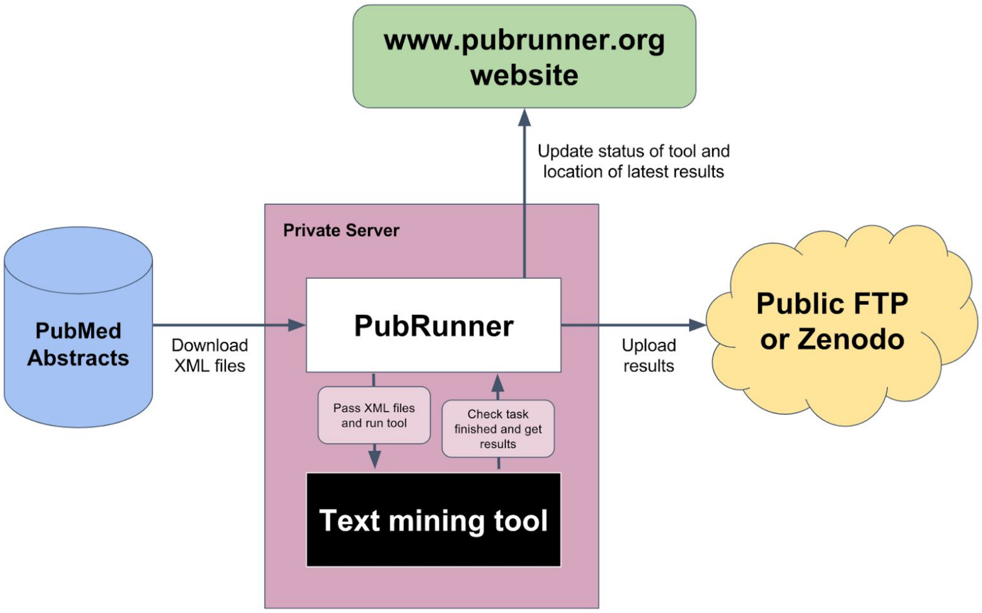
www.pubrunner.org website
Update status of tool and location of latest results
Private Server
PubMed Abstracts
Download XML files
PubRunner
Upload results
Public FTP or Zenodo
Pass XML files and run tool
Check task finished and get results
Text mining tool

Figure 1. Main workflow of PubRunner

The PubRunner framework was tested with several different tools. Three tools were developed as test cases and these tools calculate basic word counts on the abstracts. The main tool tested was word2vec[5] which is a very popular program for calculating word vector representations and is commonly used in text mining analysis. PubRunner is currently run monthly on these four text mining tools and we plan to expand the framework and add more tools.

Conclusion

The text mining community has built a wealth of valuable tools that can extract valuable knowledge for biologists and build useful data for other text mining researchers. However the output of these tools, as presented by the original authors, is often out-dated. PubRunner is the first step towards solving this problem and allowing biologists and text mining researchers to use the latest data in their work.

References

1. Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. Database. 2011 Jan 1;2011.
2. Li G, Ross KE, Arighi CN, Peng Y, Wu CH, Vijay-Shanker K. miRTex: a text mining system for miRNA-gene relation extraction. PLoS computational biology. 2015 Sep 25;11(9):e1004391.
3. Tsuruoka Y, Miwa M, Hamamoto K, Tsujii JI, Ananiadou S. Discovering and visualizing indirect associations between biomedical concepts. Bioinformatics. 2011 Jun 14;27(13):i111-9.
4. Tsai RT, Dai HJ, Lai PT, Huang CH. PubMed-EX: a web browser extension to enhance PubMed search with text mining features. Bioinformatics. 2009 Aug 4;25(22):3031-2.
5. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems 2013 (pp. 3111-3119).