Journal of Machine Learning Research () Submitted ; Published

DataRank: An Online Ranking Algorithm for Ranking Biomedical Datasets

Arya Iranmehr, MS AIRANMEHR@UCSD.EDU
Department of Electrical and Computer Engineering

Huan Wang, MS HUANWNG@UCSD.EDU
Department of Computer Science and Engineering

Xiaoqian Jiang, PhD X1JIANG@UCSD.EDU
Division of Biomedical Informatics
University of California, San Diego
La Jolla, CA 92037, USA

Editor:

Abstract

In this paper, we propose an online ranking algorithm, DataRank for ranking biomedical datasets that are used in the papers index in PubMed Central. DataRank's input is a bipartite citation graph between datasets and articles which each paper is represented by set of corresponding MeSH terms. DataRank works by imputing a set of MeSH terms to each dataset as features, by aggregating MeSH terms from the connected papers in the bipartite graph. For each search query, DataRank first maps the query to set of MeSH terms and present a *offline* ranking of datasets for the MeSH-Query using a Bayesian approach which the likelihood is proportional to Jaccard index and prior is proportional to number of citations of that dataset. DataRank is also extended to a *online* algorithm by incorporating user-feedbacks regarding ranking relevance. The online DataRank again takes an Bayesian approach which uses offline DataRank as its prior and computes its likelihood by estimating the user rating for unknown values using collaborative filtering. A demo web search engine has been developed to rank more than 20,000 dataset that has been discovered in more than 1 million papers.

1. Introduction

Blei (2012)

2. Background

3. Methodology

4. Implementation

5. Experiments

6. Conclusions

References

David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77, April 2012. ISSN 00010782.