SetSearch+: Entity-Set-Aware Search and Mining for Scientific Literature

Jiaming Shen1∗, Jinfeng Xiao1∗, Yu Zhang1, Carl Yang1, Jingbo Shang1, Jinda Han1, Saurabh Sinha1, Peipei Ping2, Richard Weinshilboum3, Zhiyong Lu4, Jiawei Han1

1University of Illinois at Urbana-Champaign 2University of California, Los Angeles
3Mayo Clinic 4National Center for Biotechnology Information
1{js2, jxiao13, yuz9, jiyang3, shang7, jhan51, sinhas, hanj}@illinois.edu 2pping@mednet.ucla.edu
3weinshilboum richard@mayo.edu 4zhiyong.lu@nih.gov

ABSTRACT

The ever-increasing volume of scientific literature calls for a better system to help researchers find relevant papers and summarize essential claims. Previous research has shown that a large portion of literature search queries are entity-set queries, that is, queries containing multiple entities of possibly different types. These queries reflect users’ need for finding documents that reveal inter-entity relationships, and pose non-trivial challenges to existing search systems that model each entity independently. In this project, we bring together a team of computing and biomedical experts, and develop SetSearch+, an entity-set-aware search and analysis system for scientific literature. SetSearch+ first leverages a data-driven text mining pipeline to extract typed entities for building entity-enhanced indices. Then, it adopts a novel entity-set-aware ranking model for online document retrieval, which captures entity type information and relations among entity sets. Furthermore, it summarizes top-ranked documents into a concise, interpretable, and interactive concept graph, which enables a user to quickly grasp the gist of all documents and therefore accelerates the knowledge discovery process. Users can interact with the SetSearch+ system conveniently via a web-based interface.

ACM Reference Format:
Jiaming Shen1∗, Jinfeng Xiao1∗, Yu Zhang1, Carl Yang1, Jingbo Shang1, Jinda Han1, Saurabh Sinha1, Peipei Ping2, Richard Weinshilboum3, Zhiyong Lu4, Jiawei Han1. 2018. SetSearch+: Entity-Set-Aware Search and Mining for Scientific Literature. In Proceedings of The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD’18). ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nmnmnn.nmnmnn

1 INTRODUCTION

Literature search helps researchers identify relevant papers and summarize essential claims about a topic. With the fast-growing volume of scientific publications, a good literature search system becomes essential to researchers since few people can master the state-of-the-art comprehensively and in-depth. Previous research [6] has shown that a large set of literature search queries contain multiple entities with possibly different types, which we refer to as entity-set queries. For example, a biologist may want to survey how genes GABP, TERT, and CD11b are associated with cancer and submit a query “GABP TERT CD11b cancer”, which is an entity-set query containing four entities. Entity-set queries reflect users’ need for finding documents that contain multiple entities and reveal inter-entity relationships. Therefore, as in the previous example, returning a paper about only one gene GABP is unsatisfactory. Existing search systems (e.g., Google Scholar and PubMed1) have not yet accommodated well such an information need as they model each entity independently and without types.

Research Project. In this project, we present SetSearch+, an entity-set-aware search and analytics system for scientific literature. SetSearch+ first leverages a data-driven text mining pipeline to extract typed entities from raw text corpus, and builds entity-enhanced document indices. Then, SetSearch+ adopts SetRank, a novel entity-set-aware ranking framework, for online document retrieval. SetRank models an entity-set query as a heterogenous graph and thus explicitly captures the entity type information and inter-entity relations. These techniques enable SetSearch+ to return a high-quality ranked list of documents that are most relevant to the whole entity set. Furthermore, SetSearch+ can summarize a set of top-ranked documents into a concise, interpretable, and interactive concept graph, which enables users to quickly grasp the gist of these documents and possibly discover new entities/relations related to the query primitives. SetSearch+ currently supports literature search in domains including computer science and biomedical science.

Fit with the KDD Ecosystem. This demo is highly relevant to a diverse community of researchers in Data Mining, Information Retrieval, Bioinformatics, and Machine Learning. For more general audience, we believe this demo can also provide a compelling example on how search and mining can be integrated to support and accelerate interactive knowledge discovery. The core methods of system will be open-sourced, and users can customize them in their local environments.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
KDD’18, August 2018, London, UK
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nmnmnn.nmnmnn

1https://www.ncbi.nlm.nih.gov/pubmed
2 MAIN INNOVATIONS

The architecture of SetSearch+ (Figure 1) consists of the following innovative components.

Data-driven entity detection and typing. SetSearch+ is to work on a massive set of “raw” documents without any explicit entity information. To support entity-aware search and analysis functions, we integrate a data-driven text mining pipeline into SetSearch+. We first detect entity mentions using domain-independent phrase mining algorithms [2, 5], and then type extracted mentions using distantly-supervised entity typing techniques [3, 4]. For the resource-rich domains such as biomedical sciences, we further enhance the quality of detected entities using domain-specific tools [8, 9]. Finally, we construct structured data indices, containing document content information, entity information, and document metadata.

Entity-set-aware document retrieval and ranking. A distinctive characteristic of literature search queries is that they reflect users’ need for finding documents containing inter-entity relations. We develop a novel entity-set-aware document ranking model named SetRank [6] to accommodate such an information need. SetRank leverages the above detected entities information to build bag-of-entities document representation. Then, SetRank uses a heterogeneous query graph to capture entity type information and model the inter-entity relations. Finally, the query-document matching process is modeled as a graph covering process. To further enhance the applicability of SetRank, we develop an unsupervised model selection algorithm to automatically choose the parameter settings in SetRank without resorting to a labeled validation set. We integrate SetRank into the SetSearch+ for online real-time document ranking. Experiments [6] show that SetRank helps researchers to identify documents that are most relevant to the whole query entity set.

Concept-graph-based document summarization. Besides showing a ranked list of documents, SetSearch+ also presents concise and interpretable document summarization using a concept graph [1]. The concept graph contains all the entities in the query and its top ranked documents, and the important entity relations inferred by ordinal graphical model [7] from the entity co-occurrence statistics. As shown in Figure 1, by viewing the constructed concept graph, a user can easily understand the interactions among the query entity set, discover “hidden” entities that are closely related to the query set, and form new scientific hypothesis for further investigation.

3 DEMONSTRATION

SetSearch+ can index 27.5 million papers of raw size 53GB with 240 million entity mentions within 6 hours on a single desktop machine. After indices are constructed, SetSearch+ can return online document search results and display the concept graph within a few seconds. We have uploaded a demo video of our system in: http://bit.ly/2slhFae. The final system will be gradually rolled out at: http://han.j.cs.illinois.edu/projs/setsearch.

ACKNOWLEDGEMENTS

This research is sponsored in part by NIH BD2K initiative 1U54GM114838 (NIGMS), ARL-NSCTA (W911NF-09-2-0053), DARPA (W911NF-17-C-0099), NSF IIS 16-18481, IIS 17-04532, IIS-17-41317, DTRA HDTRA11810026, and NIH Intramural Research Program, NLM (ZL).

REFERENCES