PoQaa: Text Mining and Knowledge Sharing for Scientific Publications

Keqian Li[1] Ping Zhang[2] Honglei Liu[1] Hanwen Zha[1] Xifeng Yan[1]
University of California, Santa Barbara
[1]{klee, honglei, hwzha, xyan}@cs.ucsb.edu [2]{mr.zhangping}@gmail.com

ABSTRACT

With the arising of popular repositories like arXiv.org, open access publication has become a trend. Publishing becomes easy. While everyone can access research papers from anywhere with a click of a button, new issues emerge. Among the exorbitant number of papers flooding into the Internet everyday, how can we know which one we shall spend time on? For the vast majority of papers that we will never be able to touch upon ourselves, how can we quickly grasp the general idea, in order to keep up with the research trend? When newcomers have questions on a research paper, who is able to help them better understand the work and pinpoint the important follow-up work. We present PoQaa (Paper oriented Question, Answer & Announcement) to address the above challenges. It features the following main functionalities, among others: (1) Feeding popular papers to each reader based on content analysis and user interest (2) Extracting knowledge from massive corpus to provide an up-to-date bird's eye view of the research landscape (3) Mining high-quality concepts and suggesting query terms (4) Hosting a paper centric discussion platform to enable knowledge sharing among readers and authors. We will further integrate new text mining algorithms in PoQaa such as "technology roadmap" and hope it can motivate new development along this direction in order to spread scientific results much faster.

CCS CONCEPTS

• **Information systems → Data mining**;

KEYWORDS

Text Mining; Recommendation; Taxonomy; Collaboration

ACM Reference Format:
Keqian Li[1] Ping Zhang[2] Honglei Liu[1] Hanwen Zha[1] Xifeng Yan[1]. 2018. PoQaa: Text Mining and Knowledge Sharing for Scientific Publications. In *Proceedings of (KDD'18).* ACM, New York, NY, USA, Article 4, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Scientific publications are becoming more accessible than ever. According to [4], in Registry of Open Access Repositories (ROAR), more than 250 subjective-based and 2,200 institutional open access repositories exist. Many prominent ones, such as arXiv, Social Science Research Network (SSRN) and PubMed Central (PMC) have played dominant roles in their respective subject fields. Though the booming of these scientific publication repositories represents the democratization of science, it creates new issues on how to efficiently digest the prohibitively large number of publications whose quality can vary dramatically. First of all, the quantity of papers increases significantly, but the readers' time hasn't. They have to choose what papers to focus on and what papers to avoid, on a day to day basis. It becomes very challenging for beginners. Second, it is not enough to only know a small number of papers one has read. There are much more papers that one needs to be aware of in order to have some breadth. How can a beginner achieve it in a reasonable time? Finally, although many papers are now available as Internet documents, there are very few places for readers to share newly discovered resources, thoughts, doubts or inspirations about the papers. Currently they are spreaded in different websites like Twitter, Reddit and Github, most of which are not paper oriented.

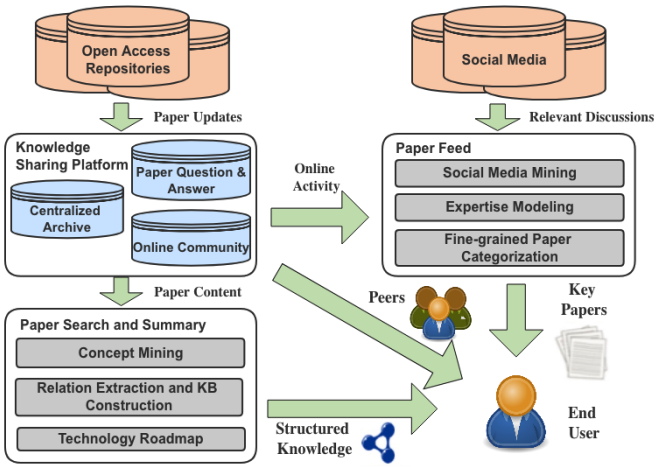

Figure 1: PoQaa Architecture

PoQaa could help people digest the large amount of open access publications based on deep text mining, enable users to share knowledge in a collaborative environment, and keep up with the large amount of research publication by identifying what papers to read and presenting general summary over the research landscape. Shallow information extraction in search engines can hardly meet the demand of deeper and more fine-grained literature mining and retrieval. PoQaa will demonstrate there is such a need.

2 TEXT MINING AND SHARING

In this section we highlight main modules in PoQaa to facilitate advanced text mining and knowledge sharing.

KDD'18, August 19-23, 2018, London, UK

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2.1 Paper Feed

In order to handle an overwhelming number of publications, one has to be very selective on finding important papers. In areas such as deep learning, researchers nowadays have to constantly check new papers on arXiv so that they don't miss newly published algorithms for the problem they're working on. It become critical to develop tools to facilitate easier access to newest research content.

PoQaa's Paper Feed will provide user with the trending papers in their area of interest, where it (1) constantly mines major social media like Twitter and detects relevant and popular paper mentions (2) identify the research area of each paper by leveraging paper categorization technique [3]. It is able to provide users with the newest papers every day, ranked by their popularity and relevance to users' interest.

2.2 Paper Summary and Search

In order to help users quickly gain insight from the large body of scientific publications, PoQaa provides Paper Summary functionality. Specifically, it extracts high quality concepts from corpus and based on that, produces a "concept level representation" for each document, where concepts like "convolutional network" are assembled into atomic units. We can leverage the distributed semantics of these concepts [3] and map them to a concept similarity graph, where concepts with similar meaning and similar occurring context becomes neighbors. One immediate application of the concept graph is query suggestion: when a user inputs a search query, such as "generalization error," the system will automatically find similar concepts in the graph such as "leave one out error," "empirical risk minimization," and help user better explore the scientific literature. It is a function very needed by beginners as they have hard time to identify right terms.

2.3 Structured Knowledge

By leveraging the relationship between concepts, we can further group them together into a hierarchy [1] and construct a taxonomy. When time information is incorporated, one can study the evolution of concepts: For each concept (e.g., algorithm), we will link its successors, predecessor, competitors, and based on that, form a "technology roadmap" for each subarea. For example, to better understand all the follow-up work of "generative adversarial network (GAN)," we arrange variants like "conditional generative adversarial network (CGAN)" and "Cycle-Consistent Adversarial Networks (CycleGAN)" in a directed graph (technology roadmap). Readers can quickly grasp the evolution of these algorithms and find which variant is the most important one and which one is popularly used. These kinds of structured knowledge will save outsiders massive amount of time for retrieving technology families and allow them to have an overview of a fast-moving area.

2.4 Knowledge Sharing

PoQaa will also let readers and authors directly share their own knowledge, similar to Quora and StackOverflow, but in a *paper centric* fashion: All of the discussions, questions/answers and data/codes are associated with the specific papers they are originated from.

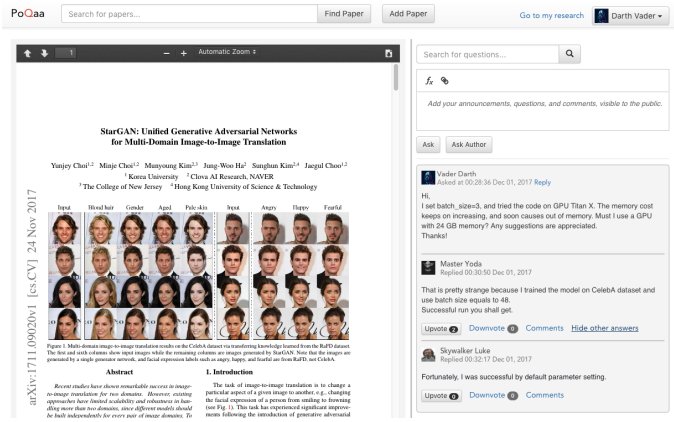

Figure 2: PoQaa Interface

Instead of being a general question answering platform, it caters to questions and thoughts one would encounter when reading a specific paper. By hosting open discussion forums, closed discussion groups, and effectively connecting users based on their interest and social interaction [2], we hope to utilize the wisdom of the crowds to help each individual gain better understanding of scientific publications and extend the outreach of scientific research to a wider audience.

3 DEMONSTRATION

The PoQaa system is available at http://www.poqaa.com/. Figure 2 is a snapshot of the PoQaa interface. We are integrating algorithms developed in our previous work into it. It has been used by about 40 students in a graduate deep learning course for question/answer sharing in the computer science department at UCSB.

In this showcase, we present PoQaa that provides paper question/answering, feed, trending and deep content mining to users so that all the after-publication knowledge can be recorded and shared in a single place. PoQaa's paper recommendation and text mining will help new researchers quickly get an overview of a field, select top papers to read from, surf the content more intelligently, and keep up with the latest related research. PoQaa is our effort to democratize scientific research and break its barrier so that more people can access first-hand scientific knowledge as soon as possible.

REFERENCES

[1] Keqian Li, Yeye He, and Kris Ganjam. 2017. Discovering Enterprise Concepts Using Spreadsheet Tables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1873–1882.

[2] Keqian Li, Wei Lu, Smriti Bhagat, Laks V.S. Lakshmanan, and Cong Yu. 2014. On social event organization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1206–1215.

[3] Keqian Li, Hanwen Zha, Yu Su, and Xifeng Yan. 2018. Unsupervised Neural Categorization for Scientific Publications. In *Proceedings of the 2018 SIAM International Conference on Data Mining*. SIAM, 37–45.

[4] Mark Ware and Michael Mabe. 2015. *The STM report: An overview of scientific and scholarly journal publishing.*