Crowd-PANDA: Using Crowdsourcing Method for Academic Knowledge Acquisition

Zhaoan Dong[1], Jiaheng Lu[1,2](✉), and Tok Wang Ling[3]

[1] DEKE, MOE and School of Information, Renmin University of China,
Beijing, China
jiahenglu@gmail.com
[2] Department of Computer Science, University of Helsinki, Helsinki, Finland
[3] School of Computing, National University of Singapore, Singapore, Singapore

Abstract. Crowdsourcing is currently being used and explored in a number of areas. Since some automatic algorithms are extremely hard to handle diverse academic documents with different quality and layouts, we present Crowd-PANDA, a **Crowd**sourcing sub-module in the **P**latform for **A**cademic k**N**owledge **D**iscovery and **A**cquisition, which combines algorithms and human workers to identify and extract meaningful information objects within academic documents named Knowledge Cells (e.g. *Figures*, *Tables*, *Definitions*, etc.) as well as their relevant key information and relationships. The extracted Knowledge Cells and their relationships can be used to build an **Academic Knowledge Graph**, which could provide a fine-grained search and deep-level information explore over academic literature.

Keywords: Crowdsourcing · Academic knowledge acquisition

1 Introduction

With an exponential growth of the volume of scientific publications, tremendous interests have been spent on extraction and management of research data within scientific literature. One example is **Digital Curation** (***DC***) [1] which indicates the activities including *selection*, *preservation*, *maintenance*, *collection* and *archiving* of digital assets and the process of *extraction* of important information from scientific literature. Another example is **Deep Indexing** (***DI***) [2] by which *ProQuest* [3] indexes the research data within scholarly articles that are often invisible to traditional bibliographic searches. Similar capabilities are available in *CiteSeerX* [4] and *ScienceDirect* [5] et al. However, they only focus on *Figures* and *Tables*, none of them have pay attention to other kind of objects like *Algorithms*, *Theroms*,*Lemmas*, etc., and the relationships among them. In PandaSearch [6–8], the information units within academic literature including all mentioned above are defined as **Knowledge Cells**. Knowledge Cells and

This work is partially supported by NSFC (No. 61472427), NSSFC (No. 12&ZD220) and RUC Fund (No. 11XNJ003).

B. Cui et al. (Eds.): WAIM 2016, Part II, LNCS 9659, pp. 531–533, 2016.
DOI: 10.1007/978-3-319-39958-4

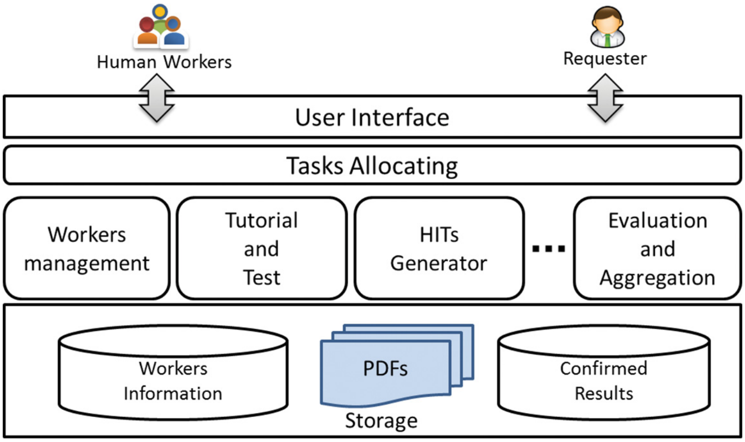

Fig. 1. The system architecture of Crowd-PANDA.

their relationships can be used to build an**Academic Knowledge Graph** for a fine-grained search and deep-level information explore over academic literature.

However, the most important prerequisite is to correctly identify and extract the Knowledge Cells and their relationships from huge amount of documents [8]. Existing automatic computer algorithms can extremely hard to handle diverse PDF documents that published in different years, conferences or journals with different quality and layouts.

As *crowdsourcing* has become a powerful paradigm for large scale problem-solving especially for those tasks that are difficult to computers but easy to human [9, 10], we make use of crowdsourcing to identify and extract those Knowledge Cells as well as their relevant key information and relationships from huge amount of PDFs. It is notable that during the process of identification and extraction some activities can generally be broken into small tasks which are often repetitive and do not require any specific expertise. For example, a human worker can firstly locate the content of a **Figure** by browsing the PDF pages and then crop the content only by "*drag and draw*".

Obviously, if all the documents are crowdsourced to anonymous workers, the crowdsourcing cost will be extremely high. Therefore, the natural alternative is to combine automatic algorithms with crowdsourcing, which is the most challenging work mentioned in [8]. In this demo, we only focus on the crowdsourcing sub-module for identifying and extracting Knowledge Cells.

2 System Design and Implementation

2.1 System Architecture

The architecture of Crowd-PANDA is illustrated in Fig. 1. The system is implemented in Python and use PostgreSQL to store the data.

2.2 Crowdsourcing Tasks

There are two basic kinds of Human Intelligence Tasks (HITs): *identifying* and *review*. **Identifying tasks** ask workers to identify the knowledge cells from those PDF pages that are difficult for automatic extraction algorithms [8]. **Review tasks** ask workers to check the answers of other workers for the sake of quality control. To guarantee the effectiveness and the accuracy, human workers have to accomplish the tutorial tasks and some tests to learn how to perform the tasks. Additionally, each HITs should be assigned a time limit.

2.3 Results Aggregations

Majority Vote is the common-used strategy for results evaluation and aggregation where an answer will be accepted if it is confirmed by most of the reviewers. For more complex tasks, we can devise a weighted voting strategy based on the confidences and historical performances of human workers.

3 Demonstration Scenarios

We plan to demonstrate the Crowd-PANDA system with the following scenarios:

Issuing Crowdsourcing Tasks: The requester can set some parameters of the crowdsourcing tasks before publishing them including the type of extracted Knowledge Cells, time limit and the number of tasks, etc.

Identifying Tasks: Human workers undertake the identifying tasks by browsing the PDF pages and crop the contents of the Knowledge Cells such as Figures, Definitions, Tables, Algorithms, etc.

Reviewing Tasks: Human workers are asked to confirm or reject the answers from other workers.

User Performance: Human workers can view the performance of their completed tasks and see their performance rank among all the workers.

References

1. Digital Curation. http://en.wikipedia.org/wiki/Digital_curation
2. Illustrata Deep Indexing. http://proquest.libguides.com/deepindexing
3. ProQuest. http://search.proquest.com/
4. Citeseer. http://citeseer.ist.psu.edu/
5. ScienceDirect. http://www.sciencedirect.com/
6. PandaSearch. http://pandasearch.ruc.edu.cn/
7. Huang, F., Li, J., Lu, J., Ling, T.W., Dong, Z.: PandaSearch: a fine-grained academic search engine for research documents. In: 2015 IEEE 31st International Conference on Data Engineering (ICDE 2015), pp. 1408–1411. IEEE Press (2015)
8. Dong, Z., Lu, J., Ling, T.W.: PANDA: a platform for academic knowledge discovery and acquisition. In: 2016 3rd International Conference on Big Data and Smart Computing (BigComp 2016). IEEE Press (2016)
9. Howe, J.: The rise of crowdsourcing. Wired Mag. **14**, 1–4 (2006)
10. Luz, N., Silva, N., Novais, P.: A survey of task-oriented crowdsourcing. Artif. Intell. Rev. **44**, 1–27 (2014)