Development of a Scholarly Metadata Knowledge Graph

Rahul Jyoti Nath,[1] and Rebaz Omar[1]

Informatik III, Universität Bonn, Germany
rahuljyotinath@gmail.com, r.omar@hotmail.de

Abstract

While scientific literature has become easily available along with the open access trend in this era, a comprehensive system to explore research metadata is often missing. Enormous number of services have been developed by giants of digital publishing, however their focus is restricted to specific metadata about either authors and publications, or citations or events. A knowledge graph provides more holistic services of metadata integration, management and analysis. OpenResearch.org (OR- a research knowledge graph) is presented in this work as a scholarly metadata management system. By default, OR is a crowd sourcing platform. However, several additional ways of data collection is implemented. In this paper, we particularly introduce one of the data collection ways which is gathering CfPs of scientific events from mailing lists. We also initiate a new definition for "community" and present a use case of such metadata combined with relevant datasets from other sources here DBLP which contains data about publications and authors. We define a set of communities and we use the linkset of metadata integration to detect research communities in the context of OR. Results of such a work flow shows an easy way to run complicated but comprehensive analysis.

1 Introduction

Along with the open access trend in this era, scientific literature has become easily available. However, a time consuming bibliographical analysis is usually performed by scientists before they can really start their research or while searching for related information about upcoming scientific events to which they can submit their research contributions. For the lack of a better support, scholars rely a lot on individual experience, recommendations from colleagues and informal community wisdom, or simple web searches. The following queries are examples of such cases:

- which upcoming and top-ranked scientific events are organized by a group of people who have been active researchers of a specific domain in the last 5 years?
- which countries have the most active research groups in doing research and organizing events in the coming months?
- Which events I should attend or submit publications in to have a higher CV in my graduation period?

Enormous number of services have been developed by giants of digital publishing, however focus of them is restricted to specific metadata about either authors and publications, or citations or events. Research communities use several domain specific ways to disseminate information that might be relevant such as mailing lists or local databases to announce scientific events, services such as ResearchGate[1] to make publications available and etc. However, a comprehensive system to explore research metadata is often missing.

[1] https://www.researchgate.net/

A knowledge graph provides more holistic services of metadata integration, management and analysis. To build such a knowledge graph, we need to build a comprehensive data collection method. Domain specific mailing lists are a medium often used by conference and workshop organizers for posting initial, second and final call for papers as well as deadline extensions. In this paper we focus on events and researchers being active in specific topics. Thus, we present data extraction method from RSS feed of mailing lists.

Furthermore, having access to the networks of a paper's authors and their organizations, and taking into account the events in which people participate enables new indicators for measuring the quality and relevance of research and detecting hidden communities. This paper, we present two main components of OpenResearch.org: 1. a specific way of data gathering, 2. usage of such data to community detection.

A detailed version of work about OR can be found in [6]. In section 2 we represent one way out of several ways of data collection into OR. Section 3 introduces a community detection as a use case of the gathered metadata.

2 Gathering Metadata through CfPs

Announcing CfPs through different mailing lists is a traditional but still most popular way of disseminating information about scientific events. The sheer amount of e-mails containing CfPs are sent through various mailing lists which makes it difficult for an individual to keep track of them. However, it is one of the main and reliable sources for different research communities to share, disseminate and discover upcoming relevant events.

2.1 Methodology

To make usage of critical mass of data being transferred through mailing lists towards creating a research knowledge graph (OR), we process RSS[2] feed of such mailing lists. This is a secondary way of data collection besides OR's crowd sourcing data acquisition process.

In order to do this, we have implemented a specific data extraction method that retrieves metadata from two mailing lists of computer science domains: 1. Public-semantic-web[3] and 2. Linking-open-data[4]). The algorithm starts with an e-mail tracing function to identify call for papers from the other type of e-mails that might be sent through the mailing list e.g. discussions. This tracing function searches the mails which contains "CfP","cfp","Call for paper" and "Call for participation". After detecting all the call for paper e-mails, a list of all relevant information such as author (name and contact information), title and the whole e-mail body gets copied to a document. In the next step of the algorithm regular expressions are used to extract specific information from the main body of the e-mail, e.g. location and important dates. The result is a document with all the important information about the events which then gets imported to OR to create the corresponding event pages. After the event page is created the author of the CfP e-mail gets contacted to verify all the data and fill out missing information.

2.2 CfPs of mailing lists in OpenResearch.org

Data extraction is done for a list of defined metrics that are the required metadata for creating event wiki pages in OR. Figure 1 shows a side by side representation of an email (left) and

[2] https://de.wikipedia.org/wiki/RSS_(Web-Feed)
[3] https://lists.w3.org/Archives/Public/semantic-web/feed.rss
[4] https://lists.w3.org/Archives/Public/public-lod/

metadata that can be extracted and info-box of the extracted and presented metadata in the wiki page of corresponding event.

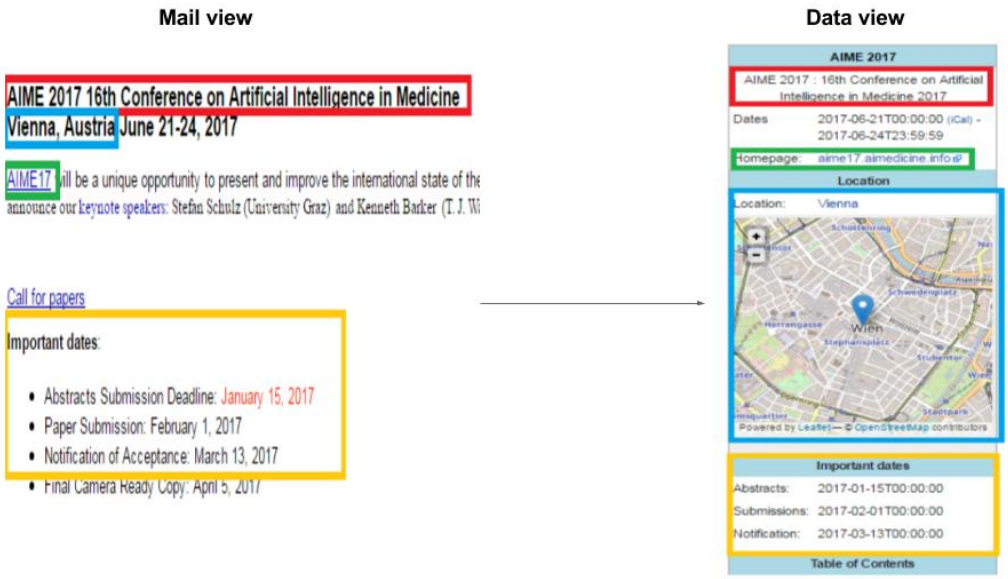

Figure 1: Metadata within a mail and the corresponding OR view

This metadata contains information about event and event organizers, their affiliation and location of events that will be used for community detection together with other relevant datasets.

3 Metadata usage: Community Detection

The objective of this attempt is to detect communities in research scholar metadata. We are using information about authors and publication gathered from the DBLP Data dump[2] combined with the metadasets gathered from CfPs.

3.1 Research Community

The term "community" is used as a broad concept in the meaning "a sizeable social unit that shares common values"[5]. It can point to online, physical or virtual groups. **Community** in the context of OpenResearch.org is defined as: "groups of scientists that shares common values". However, the values we are going to define can point to a set of metrics relevant to an element on the scientific communication. This can be a combination of all research entities (authors, journals, conferences etc.) which belong to a specific group of bodies with similar properties (co-authorship, papers published in the same conferences etc.) from our point of interest. Researchers can search for information about relevant publications, co-authors of an author collaborated with, conferences or journals in the research domains of their interest. A sample set of three communities and their overlaps is shown in 3.1.

[5] `http://semanticscience.org/resource/SIO_001064`

These can the cases when researchers change their affiliations and yet carry the research and their networks on.

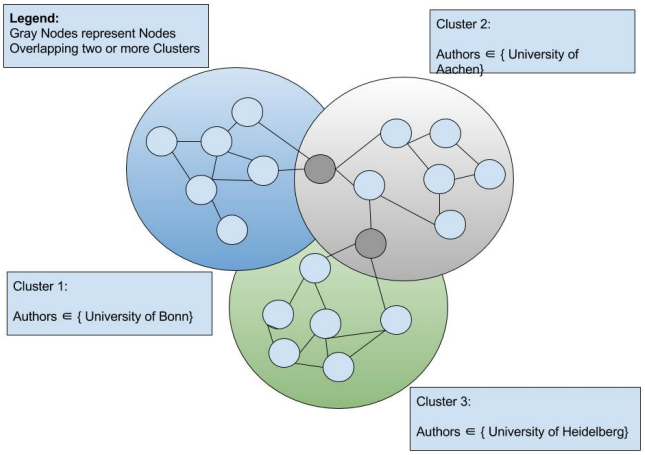

Figure 2: Sample Clustering of Authors from different Universities

3.2 Graph Clustering Tools

In order to do the graph clustering, we explored the already existing tools The following three tools have been selecting as the candidate ones for this research.

METIS is a set of serial programs for partitioning graphs to partition based on sparse matrices. The algorithms implemented in METIS are based on the multilevel recursive-bisection, multilevel k-way, and multi-constraint partitioning schemes [3]. When given the number of desired communities, it finds the maximum number of communities in such a way that the communication inside the community is maximized and connections between the communities is minimized.

semEP semEP is an edge partitioning approach that combines a data mining framework for link prediction, semantic knowledge (similarities) from ontologies, and an algorithmic approach to partition the edges of a heterogeneous graph.[1]

Currently, the the community detection is under development using METIS and as future work we are planning to do the same using KOI and finally come up with a performance evaluation of the three tools.

3.3 Implementation and Workflow of Community Detection Service

The overall community detection workflow is shown in 3.3 shows the which we plan to implement. The data imported from DBLP[6] imported into OpenResearch knowledge graph will be used as an input for the detector. A threshold function for edges in the graph was calculated for preparing the graph. Metis is currently used for clustering the graph 3.1.

[6] `http://dblp.uni-trier.de/`

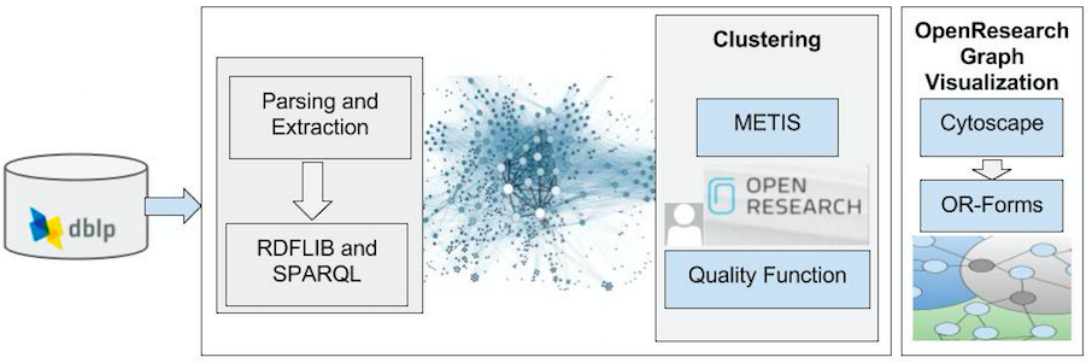

Figure 3: White Box View of Community Detection

The detector itself is composed of three sub-components: Parser, clustering and visualizer. Rdflib parser [4] used to parse the DBLP data dump in n-triple file format, and later sparql query filters the data according to requirements considering the type of community under consideration. A sample use case could be "Clustering of Authors from Netherlands who published in the 2012 Conference of ISCW on the topic of semantic web." The threshold function prepares the edges in the graph, and a wrapper function prepares the graph file suitable for the clustering Tool metis. A wrapper was created to convert the output cluster file from metis into a suitable Excel Graph file format acceptable as input for the visualizer. For visualization, cytoscape [5] was used which plots the graph showing the clusters in different clusters. It is an open source software platform for visualizing of graphs with annotations.

4 Conclusion

We represented the current enhancing plans of OpenResearch platform. An extra data collecting way that leads to an easy and multiple way of building the research knowledge graph. Collection of data from CfPs can increase visibility of such data by potential target users. CfPs are offered in a more flexible and queriable way through our crowd sourcing platform. This enables communities to contribute to the event call for papers by adding or editing and easily disseminating them as well as performing complicated queries on top of them. With the help current available clustering tools, we are working on graph partitioning. We show sample use case of communities which can be an interest of researchers, however due to lack of support not enabled to be detected. Detection of hidden communities in the scholarly communication is the main purpose of this work.

5 Future Work

In future, we envision to intensify data flows and service integration between OpenResearch and other open scholarly services. Currently, metadata extraction from mailing lists are done for two mailing lists of semantic web community but we plan to increase this for other lists. The process of data gathering from mailing list will be integrated to OR in a way that automatic emails will be sent to organizers pointing to the wiki page of their event on OR and asking

them to disseminate the event using OR.

A systematic list of potential communities that can be an interest fr several research domain will be developed by interviewing experts of the domains. Each of this communities will be implemented as an stand alone criterion in the detector. A more user friendly way of searching for communities will be implemented on top OR. We plan to implement the community detection using the other tools, KOI and semEP and an evaluation of the three tools will be a future work.

6 Acknowledgement

Authors of this work are students doing this research for their theses and would like to thank supervisors: Prof. Maria-Esther vidal, Dr. Christoph Lange, M.Sc. Sahar Vahdati.

References

[1] Kevin Bleakley and Yoshihiro Yamanishi. Supervised prediction of drugtarget interactions using bipartite local models. *Bioinformatics*, 25(18):2397, 2009.

[2] Olaf Görlitz and Steffen Staab. Federated data management and query optimization for linked open data. In *New Directions in Web Data Management 1*, pages 109–137. Springer, 2011.

[3] George Karypis and Vipin Kumar. Metis – unstructured graph partitioning and sparse matrix ordering system, version 2.0. Technical report, 1995.

[4] D Krech et al. Rdflib python library. Technical report, Tech. rep, 2002.

[5] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[6] Sahar Vahdati, Natanael Arndt, Sören Auer, and Christoph Lange. Openresearch: Collaborative management of scholarly communication metadata. In *EKAW*. Springer, 2016.