Geo-Awareness of Learnt Citations Prediction for Scientific Publications (Demo Paper)

Ce Li
Iowa State University
Ames, IA, USA
celi@iastate.edu

Will Postler
Iowa State University
Ames, IA, USA
postler@iastate.edu

Ian Johnson
Iowa State University
Ames, IA, USA
ianjohn@iastate.edu

Paul Brinkmann
Iowa State University
Ames, IA, USA
pbrink21@iastate.edu

Evan Gossling
Iowa State University
Ames, IA, USA
evang@iastate.edu

Bailey Gorlewski
Iowa State University
Ames, IA, USA
bwg@iastate.edu

Goce Trajcevski
Iowa State University
Ames, IA, USA
gocet25@iastate.edu

ABSTRACT

Predicting the citation count of academic/scientific publications has recently spurred a significant amount of research, as a particular variant of the broader cascade prediction for evolving (heterogeneous) networks. However, not much has been done in terms of tying the *geo-social* and *contextual* aspects surrounding the source datasets. Specifically, in complement to determining the trends for the purpose of various mining and prediction tasks, the broader contextual aspects can help in other planning tasks (e.g., teams-forming, allocations of resources, etc.). Given the lack of tools for interactive exploration of the prediction of the models in-concert with (various granularities of) spatial, temporal and other meta-data aspects we took a step towards implementing a prototype system providing such functionalities. In this demonstration paper we present a proof-of-concept implementation of a system that, for a given model for predicting future citations enables: (1) Visual exploration of geo-locations of the institutions with which the co-authors are affiliated, at various granularity; and (2) Access to desired meta-data pertaining to the authors/institutions. We used the open-source data from the APS journal to train the machine learning models to predict the citation count, as well as to enable the (visualization of) other contextual queries. The source code of the implementation of our system is publicly available.

CCS CONCEPTS

• **Information systems** → **Information systems applications**;
• **Computing methodologies** → **Machine learning**.

This work is licensed under a Creative Commons Attribution International 4.0 License.

LocalRec '23, November 13, 2023, Hamburg, Germany

© 2023 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0358-4/23/11.
https://doi.org/10.1145/3615896.3628341

KEYWORDS

citation prediction, academical geo-network, data mining

ACM Reference Format:
Ce Li, Will Postler, Ian Johnson, Paul Brinkmann, Evan Gossling, Bailey Gorlewski, and Goce Trajcevski. 2023. Geo-Awareness of Learnt Citations Prediction for Scientific Publications (Demo Paper). In *7th ACM SIGSPATIAL Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising (LocalRec '23), November 13, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3615896.3628341

1 INTRODUCTION AND MOTIVATION

Predicting the evolution of information cascades has generated a lot of interest among academicians and practitioners [6]. Significant results have been achieved in specialized context such as modeling and learning information diffusion in cellular networks, connections in online social networks, content sharing networks, etc. The term *prediction* may have different meanings in different applications – e.g., predicting the popularity of tweets/hashtags in microblogs, the number of "likes" for a photo/video in Facebook, views of videos in YouTube, etc.

A particular variation is the problem of quantifying and predicting the long-term impact of scientific papers [4]. In addition to the outcome of the model, it may have implicit impacts in various policy decisions like, for example, identifying emerging trends or assessing the merits of proposals for potential funding [5]. There are several literature indexing service providers, such as Google Scholar, dblp and Semantic Scholar providing large data sources that can be used for investigating various impacts of scientific publications [1]. For example, one may be interested in investigating the collaboration among scientists on particular topic, towards which various graph-based representations have been explored [2]. There are also tools that can help visualize the co-authorship networks and search the development of literature topics such as CitNetExplorer[1].

What motivates our work is the observation that deep learning based models for predicting scientific impacts (via citations count)

[1] https://www.citnetexplorer.nl/

do not enable explorations of geo-spatial and meta-data contexts along with the temporal evolution (possibly in different granularity). In other words, there are no publicly available tools that can be used to: (a) visualize the future evolutions of citations count in an interactive manner that would allow selecting both the desired future date as well as the frequency of predictions between *now* and the targeted date; (b) and visualize geo-locations of the co-authorship affiliated institutions and providing meta-data for the users. We took a first step in that direction and developed a proof-of-concept (extensible) system that can be used to:

- Select a paper on a particular topic from a given indexed collection.
- Execute a prediction model at a user-specified granularity and target date.
- Visualize the evolution of the number of citations for the selected paper.
- Enable geo-location and meta-data exploration of the institutions of the affiliated authors at various granularities.

In the rest of this paper, we provide a basic background in Section 2, discuss the architecture of our system in Section 3 and describe the steps of a demonstration scenario in Section4. Section5 summarizes and outlines directions for future work.

2 BACKGROUND

When it comes to predicting the impact of scientific publications, the data available from source that index scientific literature typically has a set of *Publications*: $P = \{p_1, p_2, \ldots, p_n\}$. Each p_i (individual publication) in turn, has several attributes such as *paper title*, *authors*, *venue*, *year*, *reference*, etc. We note that, given the data source, different deep learning models have used separate entity classes such as *Authors* and *Venue*. In our prototype system, the attribute *authors* is a composite one, consisting of typical entries for both journals and conferences – e.g., *Name*, *Affiliation*, and *email*.

In our implementation, we use the APS[2] (American Physical Society) dataset which has an academic network containing over 616K papers on 17 APS journals (1893-2017). A heterogeneous bibliographic network is maintained to describe the relationships among author, paper and venue entities. In addition, the network is constantly evolving and will include new published papers as well as emerging researchers. A dynamic heterogeneous graph machine learning model [3] can be deployed here to extract the semantic information of a research paper and get its representation, including the authorship, published venue and keywords. And with the learned representation, a predictor can be trained to make citation prediction for new publications.

3 SYSTEM DESCRIPTION

In most general sense, our prototype system is a website which consists: a *homepage*, *about us* page, *query creation* page, and *visualization page*. The core functionality enables selection of a publication, invocation of the ML model to predict the number of citations for the selected publication for a given choice of parameters (cf. Section4) and displaying the predicted results as well as the geo-locations of the institutions of the respective authors. The basic

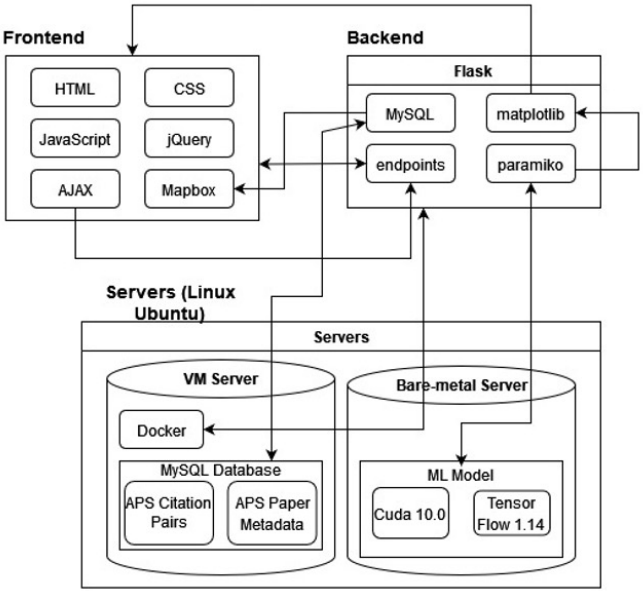

Figure 1: Schema of the infrastructure.

architectural components are illustrated in Figure 1. As shown, our system consists of two basic modules (Frontend and Backend) – however, there is a substantial amount of development in the actual Virtual Machines (VM)/Servers, which is zoomed in at the bottom of Figure 1. Next, we discuss the details of each module.

3.1 Frontend

The Frontend has six main components.
– *HTML*: responsible for displaying the documents and is the basis of our web application. The HTML pages are rendered using Flask via our set endpoints, and any necessary data is passed to the HTML page using Flask. The other Frontend subsystems (described in the sequel) are all integrated or displayed using HTML.
– *JavaScript*: as a programming language usable with HTML, it allows us to bring scripting functionalities and an object-oriented language into our HTML pages. It utilizes many third-party libraries such as jQuery, Ajax, and Mapbox[3].
– *CSS*: This subsystem is used for styling our web applications. It turns what would be a harsh, basic, and inconvenient User Interface into a user-friendly, appealing, and easy-to-use User Interface.
– *jQuery*: is used in the Frontend as a library of JavaScript, for event handling such as select field changes or button on clicks. It also handles some CSS animations and Ajax.
– *Ajax*: is a technique that we use to send POST requests to utilize some of our endpoints in Flask.
– *Mapbox*: is the last JavaScript library used in the Frontend and its purpose is to display pins on a map of the world. The pins correspond to the geo-locations of the institutions for which the authors of a selected paper are affiliated with.

[2]https://journals.aps.org/datasets

[3]https://www.mapbox.com/

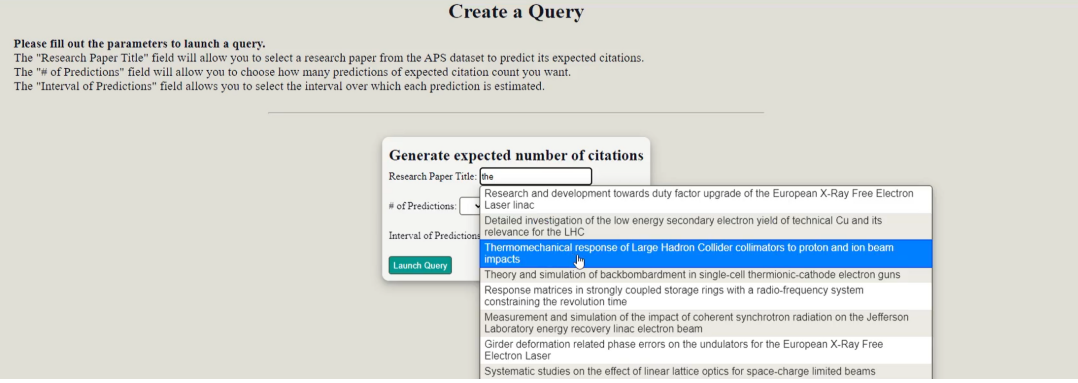

Figure 2: Selecting a publication

3.2 Backend

There are five major parts of the Backend architecture:
– *Flask*: is used as the backbone and foundation of our web application and backend. It enables adding more sophisticated coding into our web app as it allows utilization of all Python libraries. In turn, it enables us to use MySQL, matplotlib, and paramiko. We also use Flask to run our web app and route our backend endpoints. The Flask application is containerized within Docker and then deployed on one of the VMs to host our web application.
– *MySQL*: we use the MySQL library to connect to our MySQL database within one of the VMs. This way, we are able to retrieve data from our database which is necessary for obtaining papers from the APS dataset and the metadata of a selected paper.
– *Matplotlib*: this subsystem is a Python library within the Flask framework and we use it to generate the graphs where the line of interest is plotted using the variables obtained from the ML model.
– *Endpoints*: used in the Backend as a part of Flask and responsible for setting the endpoints used within our web application. These endpoints are set up into two categories: (1) returning and visualizing a HTML page, sometimes with some python logic beforehand to pass data through to the HTML page. This generally involves using other libraries such as MySQL, paramiko, and matplotlib. (2) Used purely for Python logic to supplement our web app.
– *Paramiko*: a Python library that allows us to communicate with our VM servers, used in the backend as a part of Flask. It enabled us to SSH into our server and execute commands on it. We specifically use Paramiko to execute our ML script to obtain the necessary values that allow us to generate the predictive graph with the expected number of citations for a selected publication.

3.3 Virtual Machines

– *Docker*: within our VMs it is responsible for containerizing the web application and hosting it on the server. We Dockerize our Flask application and then run the Docker image on our server. This then allows the web application to be hosted on a particular network and be accessible to anyone on that network.
– *MySQL Database*: its purpose is to store the two APS datasets. The first dataset is used to store all of the APS papers, and the second is used to store the metadata of the respective papers. We utilize both datasets to plot the locations affiliated with a specific paper using Mapbox.
– *Machine Learning Models*: This subsystem is used in the Backend and is stored in our physical VM. It is responsible for using the APS data to train a model which will subsequently be used to generate the expected number of citations over time" graph using the input parameters.

We close this section with two important remarks: (1) For testability, we have created a publicly accessible git repository https://github.com/evangossling/sdmay23-35 that can be cloned. (2) The full documentation, describing the design process and ideation, along with more detailed description of the architecture as well as various testing phases (unit, interface, integration, system, regression and acceptance) is publicly available at https://sdmay23-35.sd.ece.iastate.edu/FinalReport.pdf. The document also provides an operations manual with additional details about installation and cloning the git repo (i.e., in addition to the README) as well as setting up the database.

4 DEMONSTRATION SCENARIO

We now describe the main steps of the demonstration of our system that will be presented to the audience.
Step 1: The initial interaction is a basic access to the home page, which provides a concise description of the system and its main functionalities.
Step 2: The very first actual interaction with our demo system consists of selecting a publication for which the user may be interested in: (a) predicting the number of citations; (b) visualizing the geolocations of the affiliated institutions. Figure 2 illustrates the textual field in which, as the user is entering the characters of the title, the system will automatically display the papers the titles of which are matching the partial entry.
Step 3: Upon selecting the desired paper, the user is required to select two additional parameters (partially obscured by the list of papers in Figure 2): – *number of predictions*, which is, how many values for the predictions would the user want to see on a graph; – the *overall interval of interest* for the prediction, showing equally-distributed values of predicted citation for the previously selected number of predictions in the overall interval.

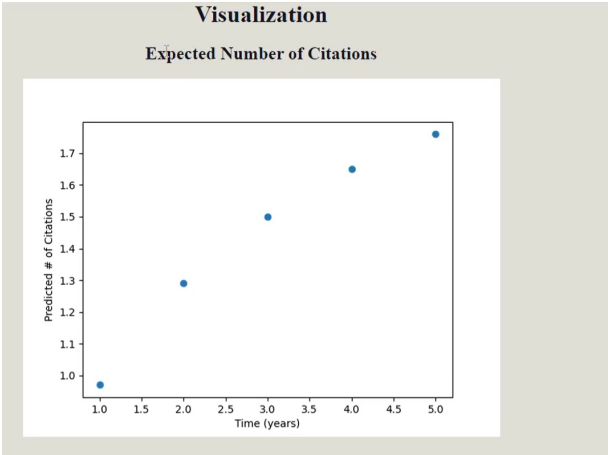

Figure 3: Prediction (interval and values within)

Step 4: Upon completing the entries in the query-window, the user can click the "Launch Query" button which will trigger the execution of the query to generated respective predictions for the selected publication. At this point, the user may experience a slight delay (~10) seconds) due to running the model. The moment the predictions have been calculated, a graph will be automatically generated where the user can see the evolution of the (expected) citations of the selected paper, as shown in Figure 3.
Step 5: Lastly, the user will be able to see the geo-locations of the institutions of all the affiliated authors, as shown in Figure 4.

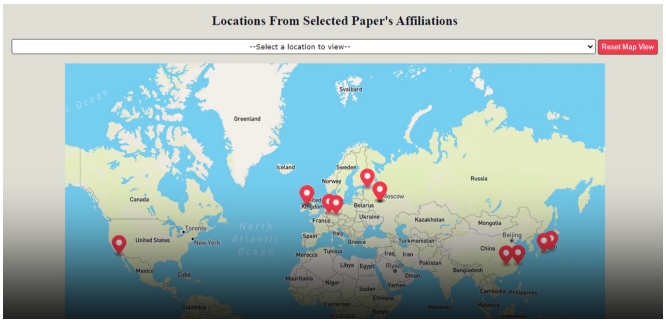

Figure 4: Visualization of affiliated institutions

Step 6: At this point, the user can click on any of the pins on the map, and: (1) the system will automatically zoom in with greater detail on the location of the selected institution, (2) the meta-data about that particular institution (available in the publication) will be displayed for the user, as shown in Figure 5.

The user can also achieve the same effect by selecting from a drop-down menu, containing the list of the institutions (not shown). The expected duration of the 6 steps of the demo is ~4 minutes, at which point a completely new run can be started.

We note that, for convenience, a video illustrating the functionality of our system (as well as highlighting some of the steps related to installation/initialization) is publicly available at https://sdmay23-35.sd.ece.iastate.edu/FinalDemoVideo.mp4.

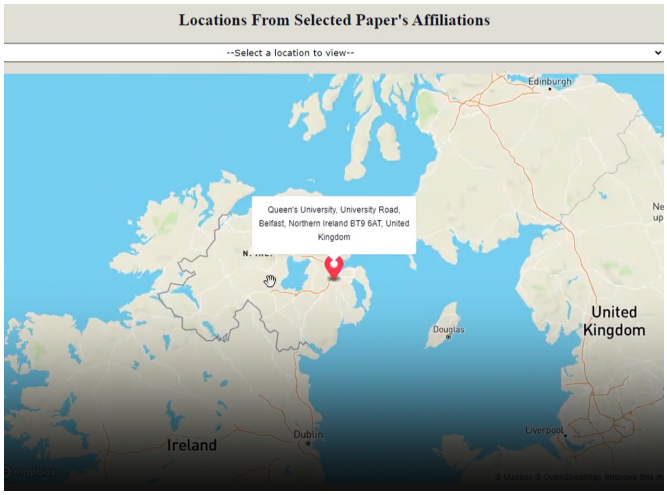

Figure 5: Selected affiliated institution

5 CONCLUDING REMARKS

We presented a system that enables interactive visualization of the geo-location of co-authorship affiliated institutions for predicted number of citations for a scientific paper for a desired future-time (and with user-specified temporal granularity). In addition, the system enables zooming in on the institutions and obtaining the corresponding meta-data. The architecture of the system makes it quite extensible and as part of the future work, we will try to enhance its capabilities by: (1) providing additional sources of bibliographical data; (2) enabling use of multiple prediction models; and (3) extending the functionality to enable comparison of predictions for multiple papers and institutions.

ACKNOWLEDGMENTS

We thank the Electronics Technology Group at the Department of Electrical and Computer Engineering at Iowa State University for providing the computing facilities throughout the development of the prototype system. Research in part supported by the NSF SWIFT grant 2030249.

REFERENCES

[1] Yuxiao Dong, Reid A Johnson, and Nitesh V Chawla. 2015. Will this paper increase your h-index? Scientific impact prediction. In *Proceedings of the eighth ACM international conference on web search and data mining*. 149–158.
[2] Jared David Tadeo Guerrero-Sosa, Víctor Hugo Menéndez-Domínguez, María-Enriqueta Castellanos-Bolaños, and Luis Fernando Curi Quintal. 2019. Use of Graph Theory for the Representation of Scientific Collaboration. In *ICCCI*. 543–554.
[3] Song Jiang, Bernard Koch, and Yizhou Sun. 2021. HINTS: Citation time series prediction for new publications via dynamic heterogeneous information network embedding. In *WWW*. 3158–3167.
[4] Dashun Wang, Chaoming Song, and Albert-László Barabási. 2013. Quantifying long-term scientific impact. *Science* 342, 6154 (2013), 127–132.
[5] Xovee Xu, Ting Zhong, Ce Li, Goce Trajcevski, and Fan Zhou. 2022. Heterogeneous dynamical academic network for learning scientific impact propagation. *Knowl. Based Syst.* 238 (2022), 107839.
[6] Fan Zhou, Xovee Xu, Goce Trajcevski, and Kunpeng Zhang. 2022. A Survey of Information Cascade Analysis: Models, Predictions, and Recent Advances. *ACM Comput. Surv.* 54, 2 (2022), 27:1–27:36.