Knowledge Graph for Discovery and Navigation

Case of Interdisciplinary Ph.D. Program

Stanislava Gardasevic [0000-0002-5758-6968]
University of Hawai'i at Mānoa, Honolulu, HI, USA
gardasev@hawaii.edu

Abstract:

This research is proposing the development of a methodology for eliciting and formalizing relationships that should be organized in a knowledge graph, intended for improved resource discovery and collaboration opportunities in a Ph.D. program. By taking a case of an interdisciplinary Ph.D. program, proposed steps will include participatory design method, text mining, and social network analysis, while reusing available models and vocabularies for the academic domain. The proposed analysis will be based on intellectual outputs, research profiles, information on activities and other relevant data that is produced by the given community. The expected outcome would account for the emphasis of actors' roles in a community, which should result in enhanced opportunities for quality cooperation.

Keywords: Knowledge Graph, Scholarly Data, Social Network Analysis, Topic Modeling, Ph.D. Research, Interdisciplinarity, Knowledge Discovery.

1 Introduction and Relevancy Statement

> One of our professors would often ask us: *What is the best dissertation?*
> And by now, we all know and answer in unison- *A done dissertation!*

The focus of this paper is based on our shared experience; something that Ph.D. students applying for this doctoral consortium are facing now, and something that you evaluating our applications have already gone through (not to say survived). It is about creating a service that would facilitate information discovery and decision making for the Ph.D. students. Considering the growth in the numbers of Ph.D. students around the world [1], this topic is very relevant to the considerably sized population. Not only that, but the proposed tool could provide similar opportunities to students pursuing other university degrees, but also to other actors affiliated with a given program (professors, researchers, alumni, librarians, administration etc.). Therefore, the background theme is: *Navigating academic space, while improving possibilities for quality cooperation, as well as information/knowledge discovery.*

The issue at hand becomes ever more complex in an *interdisciplinary* Ph.D. program, consisted of over 100 alumni, 40 affiliated faculty members, 30 students, 4 schools, and 1 university. This is the case of the program that I am attending as a 2nd-year student. It is called *Interdisciplinary Ph.D. in Communication and Information Sciences* (CIS), at University of Hawai'i at Mānoa. This program is taken as a case for examining, applying different methodologies and developing the "pathfinder" tool. The intention here is to explore the problem of classification, and to create knowledge organization system (KOS) by focusing on utilizing different and interesting *relations* that might be relevant to members of the community. CIS program is taken here as an extreme case because it comprises of 4 disciplines/ 3 schools- School of Communication (COM), Information and Computer Sciences (ICS), Information and Technology Management (School of Business) (ITM), and Library and Information Science (LIS). Not only is this interdisciplinary combination an interesting phenomenon for examining potential intersections in topical and people relations, but the results might be very relevant for the science in general, considering that the research is becoming more and more interdisciplinary [2].

Although my background is in LIS, each discipline of CIS program contributes in its own way to the main goal of my work- facilitating information discovery and improving its relevance. Being exposed to different ideas and paradigms is considered to be a great creativity amplifier, and that creative impulse is what I hope to be the driver of my research, as well as my contribution to the ISWC 2018 Doctoral Consortium.

2 Problem Statement

Although most of the graduate students start with the web search as the first information-seeking activity, doctoral students often consult their faculty advisers, then librarians and peers [3]. People do play a significant role in all phases of Ph.D. research. But how can one find an appropriate person, one you can talk to and hopefully even work with? Collaborators one chooses for their thesis, especially committee chair and members, can be accounted for eventual problems or successes of the thesis research process, something that might influence entire career.

Therefore, the problem this research is going to address is facilitating the discovery of relevant *resources* that are considered as necessary for the success of a Ph.D. student- e.g. finding an appropriate supervisor, thesis committee members, collaborators, courses, projects, information on conferences, seminars, etc. Relevant information can come from many sources. This research is aiming to develop technology based on a *knowledge graph*, envisioned to help with connecting people; pointing us to those around, who can potentially provide us with valuable pieces of information, thus help us with the decision-making process. The research will attempt to address the problem of establishing a methodology for a knowledge graph creation, based on combining methods already implemented in other solutions, and applying it to the case of CIS program and its pertinent domains. Still, the intention is for the methodology to be reusable in any given academic program. The research will address the issues of i) choosing methods that could be applied to extract interesting relations from data produced by a community, ii) order in which they should be applied, and iii) discovering interesting relations that should be included in the graph/KOS by means of data science. Through addressing these issues and checking them with the participants from the community, the results will be used in developing a particular application, and then tested.

3 Related Work

Academia and scholarly communication is an interesting area for developing recommendation-based systems. Due to its rather complex, yet relatively structured and well documented body of knowledge, it is offering a great testbed for developments in different domains such as: information retrieval (IR), LIS (sciento/bibliometrics, KOS & classification, academic librarianship), social network analysis (SNA) and visualization, Semantic Web, including numerous ontologies developed for this purpose, and many other. This research is intended to re-use and mix relevant solutions, methodologies and paradigms from these different domains, that are already validated.

For example, one of the rather comprehensive schemas in this area is VIVO Ontology for Research Discovery[1]. This model comes as a part of a Semantic Web *OpenVIVO* platform that is freely available for use and upload of data, whether by an institution or an individual researcher [4]. Not only is the VIVO ontology well elaborated, but the mentioned implementation allows for different explorations/navigations of data- e.g. author-topic connections through the Capability Map[2], co-authorship network, etc. Another application, Rexplore [5], develops the possibility of scientific data exploration even further. The proposed solution, is combining many functionalities in facilitating expert search on a fine-grained level by treating research areas as semantic concepts, rather than syntactic (keywords- usually utilized in IR systems). Furthermore, the system offers an interesting exploration through the graph view, that can be interactively navigated based on different relations between authors, but also ranked based on various metrics, and filtered with respect to years, topics, venues of publishing activity.

These are examples of good practice in facilitating the scientific information discovery- VIVO with focus on open and reusable scientific data, and Repox aiming at the eventual business processes and usages. Still, both cover the research data on the global scale. Contrarily, the research presented in this paper is more locally focused. Being strongly grounded in a particular *place* (geolocation, implying an organization end even more precise, unit(s) within), entails having local norms and requirements related to the research topic and practices. These norms will be taken as paramount of the *Topic* class modeling effort, hopefully resulting in the increased relevance and usage in that community. Furthermore, the graph view is intended to be used beyond the visualization (sensemaking) purpose, but also for interactive navigation of the knowledge base.

Research on expert profiling and recommendation has been popular lately. One of such endeavors has elicited a methodology that might be potentially reused here. STEP methodology [6] is incorporating extraction of concepts based on domain ontologies, and their consolidation by annotating lexically different but semantically similar entities, in order to create the automatic and time-depended expert profiles. That methodology was further extended with statistical methods: Topic modeling and N-Gram modeling in an attempt to improve results.

Still, in cases where no semantic reasoning is applied as a method, and only probabilistic methods- such as topic modeling [7] or author-topic modeling [8] were used, a network science-based methodology presented by Paranyushkin [9] could be utilized,

[1] https://bioportal.bioontology.org/ontologies/VIVO
[2] http://openvivo.org/vis/capabilitymap#

by which one might run particular document or a subset of the corpus assigned to a particular topic, in order to validate results and/or name topics more adequately.

Finally, there has not been much application of deeper SNA methods in the KOS design, beyond visualizing collaboration networks [4, 5], and recommendation systems based on similarity of user profiles [10]. The proposed research tends to explore this frontier further. Research presented by Kadriu [11] shows exactly how the network science metrics (in this case centrality metrics- degree, closeness, betweenness, and PageRank) can bring valuable insights of the state of topical expertise in an institution. Including such information in KOS could be a valuable asset, since it could inform on the roles that certain people might play in the community (e.g. high betweenness centrality would point out people who would be best to spread information, as they connect those in disparate parts of the network). Except for the centrality and the degree of separation (pointing out the connecting nodes), my plan is to apply (overlapping) community detection algorithms, assortativity, affiliation and other SNA algorithms for further analysis [12].

4 Research Questions

Tentative research questions behind this proposal are:

RQ 1 What are the information needs of a Ph.D. community?

- What information is deemed as relevant for successfully fulfilling a program requirements?
- What type of social support aspects are lacking in current tools?
- How can people use novel technology to navigate the academic information space?

RQ2 How do you organize the domain information in a coherent way, by means of creating and navigating knowledge graph?

RQ3 What are the more appropriate methods for knowledge discovery- the created knowledge graph or the existing ones?

- In which extent is new KOS improving the information discovery experience/ fulfillment of information needs for this community?

5 Hypotheses

Considering that the proposed research is harnessing methodologies from different disciplines, including social sciences (participatory design) and IS (design science), it is not possible to answer to all of the proposed questions by means of quantitative research methods. Still, several hypotheses could be posted in order to answer to the RQ3.

- ❖ The created knowledge graph is a more appropriate (faster, relevant) method for information discovery than the already existing means (e.g. CIS website).
- ❖ The created knowledge graph has positive impact on the lives of CIS students (e.g. it helps in finding more relevant courses, projects, mentors, etc.).

- ❖ Overall satisfaction with information discovery possibilities is higher when using the created knowledge graph, then the already available means.

6 Approach

When designing an information tool, it is considered as a good practice to go the community this tool should serve. Research has shown that through participatory design approach, research participants become responsible agents, deemed as partners rather than subjects of a research [13]. Not only that, but such agency can potentially make the underlying values more visible, and thus facilitate establishing a more comprehensive rationale for cooperation. For that reason, participatory design methodology is considered as the most appropriate for i) answering to the RQ1, ii) informing the design of the knowledge graph, as well as iii) evaluating and improving it. The community that would be involved in this research are my CIS peers (group of about 30 students) and CIS committee (the core of 5 professors included in the decision-making processes). Several workshop sessions will be organized in order to elicit the valuable group experience. Also, an online questioner will be conducted in in the same community, in order to capture the information that will be included in the social graph, i.e. important years in the program, classes/directed readings taken, committee members, topics of interests, estimated relationship with other students, research methodologies used, and other variables that might be interesting for the purpose of analysis, visualization and/or recommendation.

Participatory design is in sync with yet another interesting theoretical approach to knowledge organization called- domain analysis. By looking at *discourse communities* [14], study of knowledge domains should be taking in consideration factors such as the structure of a knowledge organization, as well as its cooperation patterns, language and communication forms, information systems and other relevance criteria. This approach will inform the data collection and analysis, with the intention to inform the RQ2. Except for the *intellectual outputs* (publications, posters, research data, thesis, course material), it should include the community members' *activities* (research, projects, teaching, supervising etc.), *meta-information* (research profiles) and other, often tacit and implied information that might be deemed relevant, still not equally available to all members of the community. The collected data will be analyzed by different means utilized in data science, including IR/text mining methods (LDA and author-topic modeling), SNA metrics and other methods that should allow for formalization of certain information that is pertinent, yet not apparent. Throughout this research, we will try to re-use the proven methods for the analysis and combine them with other methods that are not so frequently used for this purpose. This should result in a novel approach to the stated problem.

6.1. Modeling

Much of the modeling efforts in this research will rely on already established schemas, e.g. *People* class will be in much informed by the VIVO ontology one, with the focus on people's *activities*- such as publishing, co-authorship, mentorship, courses teaching/attending, projects, labs involvement, etc.

However, the modeling of the *Topic* class is going to be tackled in a slightly different way. It should include not only the *topic of interest*, but also notions such as *application area*, *methodology* used, as well as *domains of expertise*- both sought and obtained (possibly indicated by *courses* thought and/or taken). Finally, epistemological studies are considered as the crucial part in domain analysis approach [15], therefore different traditions that are due to *epistemological schools* should be part of the modelling effort. Such approach is intended to support the needs of a particular local community, since this level of granularity is usually not available in present discovery tools. Also, wherever possible, existing vocabularies will be reused (e.g. subset of the FAST thesaurus[3], for the broader research domain), while keeping in consideration local trends.

6.2. Building the Graph

The data (relations) deemed relevant should be stored and organized in a graph database system Neo4J[4], set up for this purpose. This noSQL database is considered as appropriate for capturing relations between entities, serving recommendations, as well as allowing for more dynamic knowledge representation and data update. This database was successfully applied in the scholarship domain in a Research Graph[5] project [16].

6.3. Maintaining and Updating the Graph

Methodology for the development of this tool will accommodate for the future maintenance and update of the graph, so the process is mostly automated. Considering that the topic modelling is dependent on the most recent publications, the actual implementation of such tool would imply the stricter compliance with the institutional policies- such as uploading the publications to the institutional repository (in this case ScholarSpace[6]), but also updating researcher profiles in the departmental website. While data for the paper co-authorship graphs can be automatically harvested and injected from DBLP[7], the metadata on thesis and project would need manual curation (by a program Teaching Assistant or a designated librarian). Furthermore, the design of tool will aim to support interoperability with other systems, thus use of APIs for automatic ingest and update of data, as shown in the case of *OpenVivo* [4].

6.4 Evaluation Plan

In order to make sure the final product is indeed a useful tool, one or more means of evaluation will be utilized. As previously mentioned, participatory design approach will be used to indicate whether the attempt is going in the good direction, to advise the design, and possible functionalities of the tool. Also, the same group can be used in order to answer to the RQ3. However, possibly the more appropriate way to test the hypotheses, posted in section 5, would be by using the quasi-experimental method.

[3] https://www.oclc.org/research/themes/data-science/fast.html
[4] https://neo4j.com/
[5] http://researchgraph.org/
[6] https://scholarspace.manoa.hawaii.edu/
[7] http://dblp.uni-trier.de/

Students, and preferably newly admitted students, would be asked to perform a set of tasks, both by using the new tool and the already existing one- the CIS website. Metrics that could be used to measure the eventual improvement in this case are: speed of discovery, relevance of results and overall satisfaction level with the new tool.

7 Pilot Project and Preliminary Results

The pilot project was set up for the purpose of testing different methods of data analysis and establishing procedures and software solutions that will be used on the full dataset. The pilot project dataset comprises of 95 publications, out of which 20 are the theses produced by CIS Ph.D. candidates, 74 papers, and a single book. The chosen publications are the most recent full text available and are including intellectual outputs by 30 professors (in average 3 papers per professor), 20 alumni, and 3 current students.

Topic modeling approach- LDA was performed on the corpus, by using R language and *tm* and *topicmodel* libraries. The number of 45 topics was chosen as appropriate for this purpose, by using *ldatuning* library [17]. Interestingly, in 90% cases, theses can indeed be considered as interdisciplinary, because of their assignment to a disparate topic from the ones that were assigned to professors. Results have shown that most of the recent research in this community is related to civic activity in social networks.

The result from this analysis will be further examined in order to inform the *Topic* class modeling effort (some of the attempts can be seen in obtained visualizations done by Gephi[8] software, e.g. the those showing words' assignment to topics[9]).

Furthermore, the same dataset was used for the purpose of creating networks for the analysis- the *co-authorship network* and the *thesis-mentorship* network, latter made of 20 most recent CIS theses. The co-authorship network shows cliques of co-authors, usually based on a departmental/school setting; but also, cooperation between departments and with students (see visualization[10]). Because of the small sample, not all of the relations are apparent, yet the method shows promising results for exploring community/interdisciplinary overlaps. Also, the *thesis-mentorship network* shows who are the important actors when it comes to chairing or participating in dissertation committee (see visualizations[11]). These visualizations make roles of individuals in the community more apparent than it is currently possible to notice on the CIS website.

8 Reflections

The created knowledge graph could be utilized for the purpose of visual navigation and discovery of information, recommendations including, where different dimension/granularity levels of the data can be explored in an interesting and intuitive way. It should allow the possibility to navigate through the graph in various directions. For example, starting from a particular topic of interest, one might get to a professor who is working on it, see her collaborators (co-authors and potential committee members),

[8] https://gephi.org/

[9] https://stasha.net/vizualizations-topic-modeling-lda-results-2/

[10] https://stasha.net/visualizations-co-authorship-network/

[11] https://stasha.net/visualization-thesis-mentorship-graph/

her students (one might ask for advice), activities (classes and projects), even publications. All of this can be done in a seamless way and without visiting several different web pages, which is the case now. Also, the highest-level graph view would show the gestalt view of the community and activities in it.

Furthermore, one of the main issues when developing novel technologies is related to attracting the critical mass to it. Most of the state of the art applications in this area, although having the impressive technology and complex modeling, are arguably scarcely used by researchers (with except for the Google Scholar and other corporate solutions). This opacity, beyond the circle of scientists that are interested in developing similar technology, is partly because the service is buried in the Open Web among many other similar tools. Although one might argue that the science knows no geographical boundaries and the wider coverage of the data is always better, the strategy of focusing on much smaller scale (in this case CIS program, or any given academic department) and offering such service on the website of the program or library where each student is bound to access it, would influence in much the actual usability of the technology, but also facilitate the discovery of relevant knowledge that might be directly obtained from the senior researcher in one's vicinity.

Finally, this research is having a sociotechnical focus, aiming to influence the changes and behavior that should be beneficial for academia by i) supporting the open access movement by enforcing the policies about submission of research papers and thesis into the institutional repository; ii) encouraging the university librarians to be better connected with faculty and their activities; iii) facilitating the valuable connections within the community by promoting collaboration opportunities, and thus facilitating education process and its quality; iv) promoting visibility of the people's activity, therefore encouraging other to be more active in collaboration/mentoring. The last point is a rather interesting one, since visibility often calls for accountability [18], and this feature can have the twofold benefit. Firstly, in the case of the thesis-mentorship graph, professors can see their peers as prominent nodes, which might inspire them to take on more of mentoring themselves. Secondly, student mentoring activity can be a base for an alternative metrics that might help administrators evaluate the impact of a professor, rather than using only publications-based ones. Professors who are taking part in many students' researches are investing their time and expertise in molding the future scientists, educators, and science as such. Thus, this metric should be more prominent in the current education system.

Acknowledgments. I would like to thank Dr. Lipyeow Lim for his advising and support of my efforts in concurring new frontiers (IR), and for asking important and interesting questions that helped shape this proposal and envision the technology that should come out of it. Also, I would like to thank Dr. Dan Suthers for his fervor in teaching Network Science and his dedication to each individual student project, mine including. Finally, thanks to CIS program for allowing me the opportunity to peruse my dreams and supporting me to attend ISWC 2018 Doctoral Consortium. I hope this research will contribute to making the program even better for the students to come.

References:

1. Cyranoski, D., Gilbert, N., Ledford, H., Nayar, A., Yahia, M.: The PhD Factory, Nature **472**, pp. 276–279 (2011). https://doi:10.1038/472276a
2. Porter, A. L., Rafols, I.: Is science becoming more interdisciplinary? Measuring and mapping six research fields over time, Scientometrics, **81**(3), pp. 719–745 (2009). https://doi.org/10.1007/s11192-008-2197-2
3. Catalano, A.: Patterns of graduate students' information seeking behavior: a meta-synthesis of the literature, Jour. of Doc., **69**(2), pp. 243–274 (2013). https://doi.org/10.1108/00220411311300066
4. Ilik, V., Conlon, M., Triggs, G., Haendel, M. A., Holmes, K.: OpenVIVO: Transparency in Scholarship. Front. Res. Met. Ana., **2** (2017). https://doi.org/10.3389/frma.2017.00012
5. Osborne, F., Motta, E., Mulholland, P.: Exploring Scholarly Data with Rexplore. In In: Alani H. et al. (eds) The Sem. Web – ISWC 2013. Lecture Notes in Com. Sci., Springer, Berlin, Heidelberg **8218**, (2013). https://doi.org/10.1007/978-3-642-41335-3_29
6. Ziaimatin, H., Groza, T., Hunter, J.: Semantic and Time-Dependent Expertise Profiling Models in Community-Driven Knowledge Curation Platforms. Fut. Int., **5**(4), 490–514 (2013). https://doi.org/10.3390/fi5040490
7. Blei, D. M.: Probabilistic topic models. Com. of ACM, **55**(4), pp. 77–84 (2012). https://doi.org/10.1145/2133806.2133826
8. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th Conference on Uncertainty in artificial intelligence, pp. 487–494, AUAI Press, Baniff (2004). https://mimno.infosci.cornell.edu/info6150/readings/398.pdf
9. Paranyushkin, D.: Identifying the pathways for meaning circulation using text network analysis. Berlin: Nodus Labs. (2011). https://noduslabs.com/research/pathways-meaning-circulation-text-network-analysis/ on
10. Thiagarajan, R., Manjunath, G., Stumptner, M.: Finding Experts by Semantic Matching of User Profiles, In: 3rd Expert Finder Workshop on Personal Identification and Collaborations: Knowledge Mediation and Extraction (PICKME 2008) Innsbruck, Austria, (2008). http://ceur-ws.org/Vol-403/paper1.pdf
11. Kadriu, A.: Discovering value in academic social networks: A case study in ResearchGate. In: ITI 2013, pp. 57–62. IEEE Press, Cavtat/Dubrovnik (2013). doi:10.2498/iyi.2013.0566
12. Barabasi, A. L.: Network Science. Cambridge University Press, Cambridge UK (2016). http://barabasi.com/networksciencebook/
13. Carroll, J. M., Rosson, M. B.: Wild at Home: The Neighborhood as a Living Laboratory for HCI. ACM Tran. Com.-Hum. Int., **20**(3), pp. 1–28 (2013). https://doi.org/10.1145/2491500.2491504
14. Hjørland, B., Albrechtsen, H.: Toward a new horizon in information science: Domain-analysis. Jour. Ame, Soc. Inf. Sci. **46**(6), pp. 400–425 (1995). https://doi.org/10.1002/(SICI)1097-4571(199507)46:6<400::AID-ASI2>3.0.CO;2-Y

15. Hjørland, B.: Domain analysis in information science: Eleven approaches – traditional as well as innovative. *Journal of Documentation*, ***58***(4), pp. 422–462 (2002). https://doi.org/10.1108/00220410210431136
16. Aryani, A., Wang, J., Zhang, H., Xiang, A., Zhou, Z., Wang, K.: Visualising Research Graph using Neo4j and Gephi, In: *Open Repositories Conference*, Brisbane (2017). doi:10.4225/03/58c8e8cc8a1ec
17. Nikita, M.: Select number of topics for LDA model (2016). https://cran.r-project.org/web/packages/ldatuning/vignettes/topics.html
18. Erickson, T.: 'Social' systems: designing digital systems that support social intelligence. Ai & Soc, **23**(2), 147–166. (2009). https://doi.org/10.1007/s00146-007-0140-3