Enhancing Research Information Systems with Identification of Domain Experts

Gautam Kishore Shahi, Oliver Hummel*

University of Applied Sciences, Mannheim, Germany

Abstract

Research organisations and their research outputs have been growing considerably in the past decades. This large body of knowledge attracts various stakeholders, e.g., for knowledge sharing, technology transfer, or potential collaborations. However, due to the large amount of complex knowledge created, traditional methods of manually curating catalogues are often out of time, imprecise, and cumbersome. Finding domain experts and knowledge within any larger organisation, scientific and also industrial, has thus become a serious challenge. Hence, exploring an institution's domain knowledge and finding its experts can only be solved by an automated solution. This work presents the scheme of an automated approach for identifying (scholarly) experts based on their publications and, prospectively, their teaching materials. Based on a search engine, this approach is currently being implemented for two universities, for which some examples are presented. The proposed system will be helpful for finding peer researchers as well as starting points for knowledge exploitation and technology transfer. As the system is designed in a scalable manner, it can easily include additional institutions and hence provide a broader coverage of research facilities in the future.

Keywords

Research area classification, Scholarly Dataset, Search Engine, Large language model, Domain Experts Search

1. Introduction

In recent years, on the one hand, research institution and their research output have become more visible due to the advancement of scholarly databases, data-sharing policies, and willingness to collaborate amongst research institutions [1]. Nevertheless, most research institutions still follow the traditional approach of merely listing metadata (such as titles and author names) of research results on their websites alongside hand-curated profiles containing usually rather coarse-grained areas of expertise. Hence, these websites usually only provide vague and often outdated information about researchers and especially their specific expertise. This poses a major challenge for stakeholders interested in understanding the research landscape or looking for domain experts or knowledge in, e.g. a nearby institution.

The number of research institutions has roughly doubled in every decade, and the number of researchers has increased in a similar fashion [2]. Most institutions, however, are lagging in updating publication metadata for their researchers, which leads to reduced visibility for researchers as well as existing knowledge and hence limits the value of research institutions as nuclei for innovation, especially for the surrounding regional industrial ecosystem. This situation is especially unpleasant for small and medium-sized companies that cannot afford dedicated scientific staff who are able to screen and penetrate scientific literature. Moreover, current research information management systems (RIMS) are not capable of automatically keeping track of research areas that are tackled by publications. Consider a researcher who started in the field of Natural Language Processing and Information Retrieval and later also started working on chatbots and large language models (LLM) as an example. Due to the involved manual work, RIMSs or websites are often not updated in a timely manner with such evolving research domains. If, in this example, someone would be looking for an expert in artificial intelligence language technology, then a RIMS might not give accurate results due to its rather static

Joint Proceedings of BIR 2024: 14th International Workshop on Bibliometric-enhanced Information Retrieval and IR4U2 2024: 1st Workshop on Information Retrieval for Understudied Users

*Corresponding author.

o.hummel@hs-mannheim.de (O. Hummel)

https://www.informatik.hs-mannheim.de/hummel.html (O. Hummel)

0000-0001-6168-0132 (G. K. Shahi); 0009-0007-3826-9477 (O. Hummel)

content. Moreover, even the use of RIMS might still not be widely adopted, as it is often the case for smaller Universities of Applied Sciences in Germany, where research has only slowly been gaining importance in recent years. Hence, interested stakeholders usually have to rely on open-domain search engines to find helpful experts from nearby research institutions, which in turn have to rely on the rather static web content of these institutions.

With the advance of Generative Artificial Intelligence [3], various AI systems, such as ChatGPT or Gemini, have become publicly available and, according to previous research, might be a promising solution for this challenge [4]. However, after asking ChatGPT the question "*Can you please provide a list of domain experts in the field of big data at Hochschule Mannheim?*", it merely replied the following: "*As of my last update in January 2022, I don't have access to specific lists of domain experts at Hochschule Mannheim (...); I recommend visiting the university's website, department pages, or contacting the relevant faculty or research centres directly. They can provide you with information about faculty members, researchers, and experts in the field of big data at Hochschule Mannheim.*"

This example anecdotally illustrates that even the most advanced open-domain chatbots are currently over-challenged with this specific exercise and merely refer the user to perform a standard web search. To overcome such time-consuming manual searches of domain experts based on static and potentially not updated information, we propose a knowledge-based search engine, which is able to automatically extract the research field(s) of scientists based on their publications and other information published on the Web and, prospectively, also on materials from internal learning management systems, such as Moodle. The key contributions of this paper are as follows, it presents:

- ideas for extracting the field of research from scientific articles.
- a search engine for finding domain experts based on the field of research
- a prototypical version of the proposed system.

In the remainder of this paper, we briefly discuss the state of the art in section 2 and the proposed approach itself in section 3. After that, we discuss its implementation in section 4 and some preliminary results in section 5. Finally, we present important ideas for future work in section 6 and conclude our work with section 7.

2. State of the Art

As the literature illustrates, the topic of classifying scientific articles into their respective field of research is still emerging. Until today, academic institutions have mostly used a manual approach for collecting and analysing scholarly data [5]. For example, while reviewing the research data management at his institution, the author of [6] was confronted with the fact that data is still collected manually to deliver simple services such as a list of publications per researcher. Hence, it is not possible to search for researchers based on a given topic. The research data in the university of the author of [7] was also curated manually; overall this is time-consuming and produces delays in collecting and publishing the data. The same holds true for the University of the authors, where publication lists are still managed with the help of Excel tables and not centrally published at all.

Another interesting study that has been conducted on the information-seeking behaviour of users of 17 search systems for academics has found that these search systems basically use very simple keyword searches and hence bear great potential for improvement through more advanced search functionalities [8]. In the study reported by [9], an exploratory search using semantic technologies is used to provide better access to domain experts. However, it is merely based on the research area provided by the researchers and manually fed into the system. In another study, the author proposes REDI, a Linked Data-powered framework for managing and storing academic data [10]. However, they also still use static data that is provided manually by researchers for this purpose.

Beyond purely theoretical research, now there are also some workshops and challenges emerging, aiming at building and comparing models able to classify research contained in scientific publications, such NSLP 2024[1] for example: This exchange and such challenges are likely to attract different approaches and models for classification in the future and, hence, help with the advancement of the scientific community in this important field.

[1] https://nfdi4ds.github.io/nslp2024

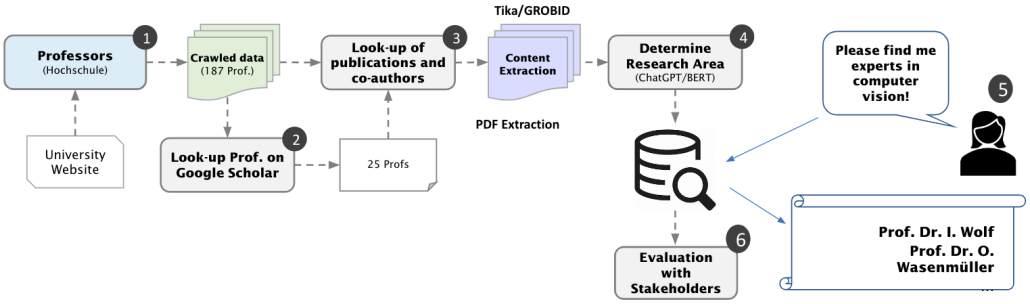

Figure 1: Flow diagram for proposed improved Research Information System.

3. Approach

Given the insights from, e.g., [8], it is clear that a purely search-based solution for research information systems will be as imperfect as other manually curated catalogues, such as in libraries. Since the development of a complex research system is a highly dynamic process that – according to experience from various fields such as software engineering, design thinking, or entrepreneurship – needs to be user-centric, we adopt a highly agile approach with rapid prototyping and brief feedback cycles mainly inspired by the Lean Startup method [11].

For the current early stage of our prototype, we envisage a research interested person searching for domain experts as our central persona and attributed two use cases to it, namely directly finding domain experts based on a classic keyword search and identifying the domains of expertise that are actually represented at an institution. The main idea is to implement a prototypical solution according to the following scheme and, once this is accomplished, evaluate it with researchers and other colleagues involved in research management and technology transfer at our university.

In the approach designed so far (cf. Figure 1 for an illustration), we aim to ingest a given list of researchers from a university or a similar organisation in order to avoid noisy data usually coming from a general web crawl. With an official list provided by the university, one can crawl for publications, e.g., via the university's homepage or scrape it from another source, such as Research Gate or Google Scholar. Another advantage of using an officially provided list is that, in our case (and probably most other cases as well), it contains at least some helpful metadata, such as the department or the broader subject area. For the actual crawling, we apply a heuristic approach, for which we, e.g., take the university name or the email's domain into account to get the best possible matches. Once a researcher's name is identified with a given degree of certainty, we crawl their information, such as citations over past years, co-authors, and lists of publications, and try to find the PDFs of the publications on the Web where possible.

Once the papers are extracted, we extract their content from the PDFs [12]. For the time being, we use ChatGPT's API to identify a research area for the extracted content [13], which is then added to the corresponding researcher's profile. Both the texts of the crawled materials, i.e., the papers, as well as the researcher profile, are then stored in a search engine as described in the section immediately following, where we also describe the other steps involved in this process in more detail. In the section that follows the next one, we elaborate on how we plan to further enhance classic search technology to achieve better results in the quest for domain experts in the future.

4. Implementation

This section discusses a prototypical implementation of our proposed approach with data from Hochschule Mannheim University of Applied Sciences[2]. Currently, the proposed system is implemented based on a list of professors by the university. Below, we highlight the most important insights from implementing the sketch from Figure 1:

[2]https://www.english.hs-mannheim.de/the-university.html

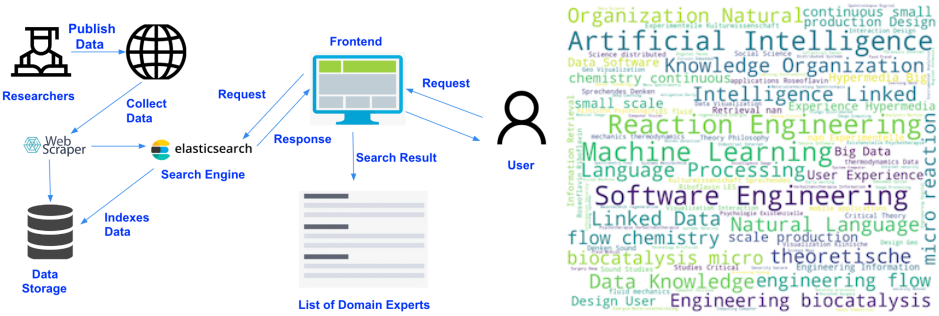

Figure 2: (a) Architecture diagram and data flow for search experience (b) Word cloud generated from the research areas of professors with Google Scholar profiles at Hochschule Mannheim

Gathering Professors. First, we parsed a list of professors from the university website. In total, it contained 188 professors that are listed together with some metadata like department, email, and telephone number.

Crawling Publication Data. Based on this initial list of professor names, we used a crawling script in Python using an open-source[3] library to search for the name of each professor on Google Scholar. If there was a match, we extracted metadata of professors, such as given research areas, citation counts, and list of publications for a total of 28 professors, most of them coming from the departments of computer science and biotechnology. These 28 matches were manually verified for correctness.

Collecting Publications. For each professor, we scraped a list of publications using Beautifulsoup [14] to gather additional publication information like title, author, and link to the PDF of the paper. Currently, we have links for 420 publications and were able to download 268 of them for our analysis. The remaining papers were not avaiblable to us, mainly because they were behind a paywall or otherwise not accessible. Once the PDFs were downloaded, we used another Python script to parse the PDF using TIKA[4] and GROBIRD[5] to extract the textual content of the paper, excluding references as this might add unnecessary noise to searches later. After cleaning out further unwanted information, like email ID or URLs, we indexed the texts in our search engine.

Identifying Research Areas. To identify the research area of each professor, we are currently evaluating three approaches. First, we used the metadata from the university homepage; second, we scraped the data entered by the researchers themselves in Google Scholar. Third, we aim to extend these by extracting more fine-granular information from each downloaded paper with the help of a Large Language Model (LLM), as indicated before. Currently, we do this by simply calling the ChatGPT API, providing it with the research area classification from the Library of Congress [15], and asking it to classify the field of research for each given paper accordingly. For each author, we merged the research areas delivered by ChatGPT for his papers with those provided by the university, as well as with those retrieved from Google Scholar. The result gives a relatively broad overview of the research expertise of each professor.

Illustrating Research Areas. From the union of all extracted research areas, we derived a word cloud using bi-gram tokens. An example is shown in Figure 2 (b).

This word cloud is intended to get stakeholders interested in a university an up-to-date overview of research topics that are currently addressed at an institution.

4.1. Search Engine

We are currently using Elasticsearch 8.7.0[6] as our core search engine since it provides out of the box text search functionality as well as advanced text analysis features for the data we collect. The overall architecture, data model, and search & browse interface are discussed in the following.

[3]https://pypi.org/project/scholarly/
[4]https://pypi.org/project/tika/
[5]https://github.com/kermitt2/grobid
[6]https://www.elastic.co/guide/en/elasticsearch/reference/current/release-notes-8.7.0.html

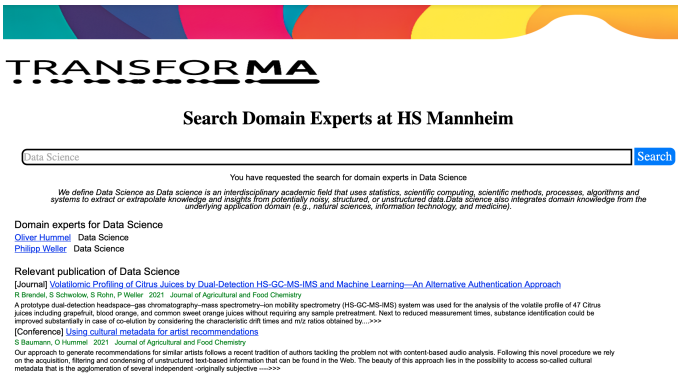

Figure 3: Exemplary Search Results with domain experts and relevant publications.

Architecture. We propose a proof of concept for our system based on a classic client-server architecture. The backend consists of a Flask application [16], which operates as a server and is connected to an Elasticsearch index containing all collected data. The front end is a client-side single-page application also based on Flask with a "thin server" architecture. I.e., most business logic is moved from the server to the client that requests data when needed, thereby allowing for a seamless user experience. This architecture is shown in Figure 2 (a), which also explains the flow of data triggered by a user request.

Data Model. The data model of the application is based on the information required from a researcher. Hence, it consists of the data extracted from the university website, Google Scholar, and the research areas extracted from scholarly publications. The data model can later be discretionarily extended to support further entities and their associated data.

Search & Browse. The homepage of the search engine initially shows the word cloud of the research areas available in the institution to provide an overview of the fields of expertise that are present there. An information-seeking stakeholder can then start looking for the desired experts by entering a search term (i.e., a desired research area) in the text box, or they can browse a sortable list of research fields that serve as a starting point to get to a domain expert. Once the search button is clicked, the user gets a basic definition of the search term extracted from Wikipedia, as well as a list of available domain experts. The user can further click on the domain expert to get more detailed profile information, such as on the research area, publications, and potential links to other bibliographic sources. A glimpse of the search interface with the interface is shown in Figure 3.

5. Preliminary Results and Lessons Learned

We have implemented the approach as described before, and since search has become a fairly well-understood topic in recent years, we have experienced no unexpected issues with its basic functionality. What we have learned so far is that manually attributed research areas are typically more coarse-grained than those that are extracted by ChatGPT, so it seems like a good complement at first glance. Even after a closer look, the research areas extracted from ChatGPT make sense and nicely illustrate how this approach is able to highlight even recent trends in a personal publication history. Consider the following research areas extracted for one professor at our university:

- **University Website**: Big Data, Data Science, Information Retrieval, Software Engineering
- **Google Scholar**: Big Data, Software Engineering, Information Retrieval
- **Extracted from publications by ChatGPT**: Cognitive Neuroscience, Software Design Patterns, Object-Oriented Programming

The ChatGPT data is apparently fine-grained and also mirrors a very recent collaborative work of this colleague in an unusual field. However, it is, of course, reasonable to question whether one joint publication in an area like "Cognitive Neuroscience" turns a computer scientist into an expert in psychology. Thus, it might make sense to consider further metrics, such as the number of papers in an area or something similar, to obtain even better results in the future.

Another lesson learned is that although a word cloud seems to be a nice visual aid at first glance, our current implementation reveals some weaknesses at second glance. First and foremost, it is visible that not all research areas are well represented by bi-grams. Moreover, a mixture of languages (such as English and German in our case) in non-English speaking countries might be somewhat confusing for prospective users and needs to be fixed in future versions.

6. Future Work

We plan to extend the search database to a partner university to add more researchers and demonstrate that our approach is generalisable to multiple institutions in the future. We have also planned to integrate additional data sources, like other academic search engines or platforms such as Research Gate or DBLP, to increase the coverage of our approach. Moreover, it is necessary to improve the quality of the word cloud since, obviously, not all research areas are well represented by bi-grams. One way to handle this better might be to use a positive list of research areas as a filter for retrieved results. As mentioned before, it also seems necessary to add some basic language detection and translation capabilities so that a word cloud does not contain a mixture of various languages, such as English and German, as in our example. In the short term, we are also aiming to improve the design of the search page by adding more functionality, such as a chatbot for answering questions about the domain and providing contact details.

In the spirit of the Lean Startup approach, we also plan to gather feedback from potential users of our system to make sure that we are actually developing a useful piece of technology. We also plan to use semantic web technologies for the mapping of research areas within ontologies [17, 18].

Another possible more long-term extension for our work is to test different (and locally hosted) LLM implementations for the extraction of subject areas from papers and to evaluate the results obtained from them. As the preliminary results from ChatGPT illustrated, it still seems necessary to better understand the accuracy of the search results in general and the applicability of LLMs for such tasks in particular.

The knowledge embodied in research publications is certainly important and valuable; however, it is probably only one side of the coin as it mostly covers the latest research results. As most scientists also have teaching obligations, it is probably safe to assume that a large part of their more fundamental knowledge is embodied in teaching materials. This is obviously less interesting for research transfer, but nevertheless, it might be interesting for finding potential teachers for advanced training, e.g., technical domains, and, of course, for broadening and sharpening the recognised knowledge areas of domain experts.

However, teaching materials are usually not published and, hence, not freely available. With access to the course management platform (CMP) of an institution, which should be possible for our system once it becomes officially used there, downloading these materials is probably less a technical issue and more a question of Copyright and willingness of the affected colleagues to at least share their materials for analysis with our system, if they do not want their scripts and slide decks to become publicly available. Hence, integrating a "stealth" mode for files that should be analysed but not indexed, as well as a crawler for our institution's CMP (which is Moodle) into our system, is another future task we are about to tackle soon.

7. Conclusion

Finding domain experts at research institutions is a challenging task that current search- or even catalogue-based approaches are not able to achieve. Hence, in this work, we presented the core of a modern research information system that is able to extract the research field from scientific publications and can be searched using a web interface. Built upon an open-source search engine, it can provide a list of domain experts for a given topic as well as links to their profile page or personal website for users in need of further information. Hence, it will be useful for stakeholders to easily identify available expertise at an institution, e.g., to initiate collaborations, research transfer, or advanced training.

One of the current limitations of our approach is that not all researchers have a profile on a scholarly database like Google Scholar, which might make it difficult to retrieve their publications and derive the research areas in which they are active. Another limitation is that getting a PDF version of a publication

is often not possible due to the paywalls used by many publishers. Hence, in conclusion, although our system already delivers promising results in its early stages of development, there is plenty of room and need for future work.

ACKNOWLEDGEMENT

The work has been carried out under the TransforMA project. Authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article. This project has received funding from the federal-state initiative "Innovative Hochschule" of the Federal Ministry of Education and Research (BMBF) in Germany.

Declaration on Generative AI

The present study does not use any AI tool for text generation or rephrasing.

References

[1] R. Farooq, Knowledge management and performance: a bibliometric analysis based on scopus and wos data (1988–2021), Journal of Knowledge Management 27 (2023) 1948–1991.
[2] Y. Dong, H. Ma, Z. Shen, K. Wang, A century of science: Globalization of scientific collaborations, citations, and innovations, in: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017, pp. 1437–1446.
[3] G. Strobel, L. Banh, F. Möller, T. Schoormann, Exploring generative artificial intelligence: A taxonomy and types (2024).
[4] A. Askari, M. Aliannejadi, E. Kanoulas, S. Verberne, A test collection of synthetic documents for training rankers: Chatgpt vs. human experts, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023, pp. 5311–5315.
[5] G. Guest, E. E. Namey, M. L. Mitchell, Collecting qualitative data: A field manual for applied research, Sage, 2013.
[6] L. Perrier, E. Blondal, A. P. Ayala, D. Dearborn, T. Kenny, D. Lightfoot, R. Reka, M. Thuna, L. Trimble, H. MacDonald, Research data management in academic institutions: A scoping review, PLoS One 12 (2017) e0178261.
[7] F. Schuetzenmeister, University research management: An exploratory literature review (2010).
[8] Y. R. Nedumov, S. D. Kuznetsov, Exploratory search for scientific articles, Programming and Computer Software 45 (2019) 405–416.
[9] T. Schopf, N. Machner, F. Matthes, A knowledge graph approach for exploratory search in research institutions., in: KMIS, 2023, pp. 265–270.
[10] J. Ortiz Vivar, J. Segarra, B. Villazón-Terrazas, V. Saquicela, Redi: Towards knowledge graph-powered scholarly information management and research networking, Journal of Information Science 48 (2022) 167–181.
[11] E. Ries, The lean startup: How today's entrepreneurs use continuous innovation to create radically successful businesses, Currency, 2011.
[12] D. Lin, Revolutionizing retrieval-augmented generation with enhanced pdf structure recognition, arXiv preprint arXiv:2401.12599 (2024).
[13] O. D. Okey, E. U. Udo, R. L. Rosa, D. Z. Rodríguez, J. H. Kleinschmidt, Investigating chatgpt and cybersecurity: A perspective on topic modeling and sentiment analysis, Computers & Security 135 (2023) 103476.
[14] L. Richardson, Beautiful soup documentation, 2007.
[15] A. Salaba, L. M. Chan, Cataloging and classification: an introduction, Rowman & Littlefield, 2023.
[16] G. K. Shahi, W. Kana Tsoplefack, Mitigating harmful content on social media using an interactive user interface, in: International Conference on Social Informatics, Springer, 2022, pp. 490–505.

[17] D. Nandini, G. K. Shahi, An ontology for transportation system, Kalpa Publications in Computing 10 (2019) 32–37.
[18] B. Dutta, D. Nandini, G. K. Shahi, Mod: metadata for ontology description and publication, in: International Conference on Dublin Core and Metadata Applications, 2015, pp. 1–9.