Methods for Generation, Recommendation, Exploration and Analysis of Scholarly Publications

Gianmaria Silvello[1*], Oscar Corcho[2] and Paolo Manghi[3]

[1*]Department of Information Engineering, University of Padua, Via Gradenigo, 6/b, Padova, Italy.
[2]Departmento de Inteligencia Artificial, Universidad Politécnica de Madrid, Boadilla del Monte, Madrid, Spain.
[3]Institute of Information Science and Technologies, National Research Council, Via Giuseppe Moruzzi, 1, Pisa, Italy.

*Corresponding author(s). E-mail(s): gianmaria.silvello@unipd.it;
Contributing authors: ocorcho@fi.upm.es; paolo.manghi@isti.cnr.it;

Abstract

In the shifting landscape of sharing knowledge, it is no longer only about writing papers. After a paper is written, what comes next is an integral part of the process. This special issue delves into the transformative landscape of scholarly communication, exploring novel methodologies and technologies reshaping how scholarly content is generated, recommended, explored and analysed.
Indeed, the contemporary perspective on scholarly publication recognizes the centrality of post-publication activities. The criticality of refining and scrutinizing manuscripts has gained prominence, surpassing the act of dissemination. The emphasis has shifted from publication to ensuring visibility and comprehension of the conveyed content.
The papers compiled in this special issue scrutinize these evolving dynamics. They delve into the intricacies of post-processing and close examination of manuscripts, acknowledging the impact of these aspects. The overarching objective is to stimulate scholarly discussions on the evolving nature of communication in academia.

Keywords: Scholarly Document Processing, Text Analytics, Citation Behaviours, Author Name Disambiguation, Recommendation System, Open Access, Web Archiving

Preface

The first paper [2], "Deep Author Name Disambiguation using DBLP Data," by Zeyd Boukhers and Nagaraj Bahubali Asundi, confronts the pervasive challenge of Author Name Ambiguity in the dynamic and expansive academic landscape. As the number of scientists and authors with identical names continues to rise, the imperative for precise Author Name Disambiguation becomes increasingly pronounced. Leveraging the extensive dataset from the DBLP repository, this paper introduces an innovative approach that harnesses information about co-authors and research domains to establish definitive links between author names and real-world entities. The efficacy of the proposed neural network model is

underscored by its significant effectiveness, validated through exhaustive experiments conducted on a large dataset. This paper is the revised and extended version initially published in TPDL 2022 [1].

The second paper [4] delves into the intricacies of referencing behaviors across disciplines and focuses on how scholars define bibliographic references. The paper "Referencing Behaviours Across Disciplines: Publication Types and Common Metadata for Defining Bibliographic References" by Erika Alves dos Santos, Silvio Peroni, and Marcos Luiz Mucheroni, comprehensively analyzes citation practices across diverse scholarly disciplines. The study unveils distinct citing behaviors and structures that exhibit variation across disciplines by scrutinizing the metadata employed in bibliographic references across subject areas. This paper is the revised and extended version of the paper initially published in TPDL 2022 [3].

The third paper [8], "DETEXA: Declarative Extensible Text Exploration and Analysis through SQL," by Yannis Foufoulas, Eleni Zacharia, Harry Dimitropoulos, Natalia Manola, and Yannis Ioannidis, introduces a text analysis framework implemented in extended SQL, addressing the significant task of metadata enrichment in digital libraries. This framework facilitates end-to-end text mining pipelines, leveraging the scalability features of modern database management systems. The declarative nature of SQL not only allows for swift experimentation but enables domain experts to edit text-mining workflows through user-friendly graphical interfaces effortlessly. This paper is the revised and extended version published initially in TPDL 2022 [7].

The fourth paper [16], "The Digitization of Historical Astrophysical Literature with Highly-Localized Figures and Figure Captions," by Jill P. Naiman, Peter K. G. Williams, and Alyssa Goodman, delves into the challenges of digitizing scientific articles from the pre-digital age. Focusing on historical astrophysical literature, the paper introduces a YOLO-based method for extracting figures and captions from scanned pages through Optical Character Recognition (OCR). Applying this method to NASA Astrophysics Data System (ADS) holdings demonstrates a substantial improvement in F1 scores compared to other state-of-the-art methods. This paper is the revised and extended version of the paper originally published in TPDL 2022 [15].

These four papers illuminate the evolving scholarly landscape beyond the initial manuscript creation phase. They provide valuable insights into innovative approaches for author disambiguation, citation practices, text exploration, and figure extraction from historical literature. The exploration of post-processing and analysis techniques emerges as a pivotal frontier in enhancing the accessibility, discoverability, and usability of scholarly content.

In an era marked by an unprecedented influx of scholarly submissions to top-tier conferences and journals, the fifth paper [12], "Towards Automated Meta Review Generation via an NLP/ML Pipeline in Different Stages of the Scholarly Peer Review Process," by Asheesh Kumar, Tirthankar Ghosal, Saprativa Bhattacharjee, and Asif Ekbal, unveils a groundbreaking approach to alleviate the daunting task of identifying proficient reviewers and meta-reviewers. The relentless growth in submissions poses unique challenges, including the intricate process of crafting meaningful meta-reviews. The paper introduces an innovative natural language processing (NLP) and machine learning (ML) pipeline that predicts review recommendation and confidence scores and employs a transformer-based seq2seq architecture to generate decision-aware meta-reviews. This methodology represents a paradigm shift, significantly surpassing standard summarization baselines and prior works in the domain. This paper is the revised and extended version of the paper originally published in TPDL 2022 [11].

Shifting focus to the intricate process of selecting open-access journals for scholarly publication, the sixth paper [6], "Comparing different search methods for the open access journal recommendation tool B!SON," by Elias Entrup, Anita Eppelin, Ralph Ewerth, Josephine Hartwig, Marco Tullney, Michael Wohlgemuth, and Anett Hoppe, introduces the B!SON web-based journal recommendation system. Navigating the complexities of journal selection amidst a growing array of options, B!SON utilizes a systematic requirements analysis, drawing insights from a survey to enhance its functionality. Recommendations are driven by user-provided content, leveraging open data, and maintaining publisher independence. Incorporating a pre-trained transformer model amplifies the

tool's recommendation quality, underscored by positive feedback from the academic community and robust performance in test sessions. This paper is the revised and extended version of the paper originally published in TPDL 2022 [5].

In the realm of scholarly publications and knowledge graph integration, the seventh paper [14], "RDFtex: Knowledge Exchange between LATEX-based Research Publications and Scientific Knowledge Graphs," addresses the limitations posed by traditional research papers in the context of Scientific Knowledge Graphs (SciKGs), by proposing RDFtex, the paper pioneers a framework that facilitates bidirectional knowledge exchange. It allows the import of contributions from SciKGs into LATEX-based research publications, easing the preparation process. It enables the export of original contributions from papers to SciKGs, enhancing their integration. This innovative bridge between traditional publications and knowledge graphs promises to reshape how scholarly content contributes to the collective scientific understanding. This paper is the revised and extended version of the paper originally published in TPDL 2022 [13].

The final paper [10], "Robots Still Outnumber Humans in Web Archives in 2019, But Less Than in 2015 and 2012," by Himarsha Ransirini Jayanetti, Kritika Garg, Sawood Alam, Michele Weigle, and Michael Nelson, delves into the intricate dynamics of user interactions with web archive content over the years. Leveraging access logs from Internet Archive and Arquivo.pt, the research unveils the shifting patterns of robot prevalence and user access behaviors across three years. This exploration into the evolving landscape of web archives sheds light on the significant role of both human and robot users, emphasizing the importance of understanding their distinct needs and access patterns. This paper is the revised and extended version of the paper originally published in TPDL 2022 [9].

These last four papers collectively underscore the multifaceted transformations underway in scholarly communication. They provide a rich tapestry of innovative methodologies and technologies that address contemporary challenges in content generation, recommendation, and exploration and point toward the future trajectory of scholarly endeavors.

Overall, these eight papers represent a revised and extended version of the original papers published in the Proceedings of the 26^{th} International Conference on Theory and Practice of Digital Libraries (TPDL 2022) held in Padua, Italy, from 20 to 23 September 2022. The papers were selected among the best full papers published in TPDL 2022.

References

[1] Boukhers, Z. and Asundi, N. B. (2022). Whois? deep author name disambiguation using bibliographic data. In Silvello, G., Corcho, O., Manghi, P., Di Nunzio, G. M., Golub, K., Ferro, N., and Poggi, A., editors, *"Linking Theory and Practice of Digital Libraries", Proceedings of the 26th International Conference on Theory and Practice of Digital Libraries, TPDL 2022, Padua, Italy, September 20-23, 2022*, pages 201–215, Cham. Springer International Publishing.

[2] Boukhers, Z. and Asundi, N. B. (2023). Deep author name disambiguation using dblp data. *International Journal on Digital Libraries*.

[3] dos Santos, E. A., Peroni, S., and Mucheroni, M. L. (2022). The way we cite: Common metadata used across disciplines for defining bibliographic references. In Silvello, G., Corcho, O., Manghi, P., Di Nunzio, G. M., Golub, K., Ferro, N., and Poggi, A., editors, *"Linking Theory and Practice of Digital Libraries", Proceedings of the 26th International Conference on Theory and Practice of Digital Libraries, TPDL 2022, Padua, Italy, September 20-23, 2022*, pages 120–132, Cham. Springer International Publishing.

[4] dos Santos, E. A., Peroni, S., and Mucheroni, M. L. (2023). Referencing behaviours across disciplines: publication types and common metadata for defining bibliographic references. *International Journal on Digital Libraries*.

[5] Entrup, E., Eppelin, A., Ewerth, R., Hartwig, J., Tullney, M., Wohlgemuth, M., and Hoppe, A. (2022). B!son: A tool for open access journal recommendation. In Silvello, G., Corcho, O., Manghi, P., Di Nunzio, G. M., Golub, K., Ferro,

N., and Poggi, A., editors, *"Linking Theory and Practice of Digital Libraries", Proceedings of the 26th International Conference on Theory and Practice of Digital Libraries, TPDL 2022, Padua, Italy, September 20-23, 2022*, pages 357–364, Cham. Springer International Publishing.

[6] Entrup, E., Eppelin, A., Ewerth, R., Hartwig, J., Tullney, M., Wohlgemuth, M., and Hoppe, A. (2023). Comparing different search methods for the open access journal recommendation tool B!SON. *International Journal on Digital Libraries*.

[7] Foufoulas, Y., Zacharia, E., Dimitropoulos, H., Manola, N., and Ioannidis, Y. (2022). Detexa: Declarative extensible text exploration and analysis. In Silvello, G., Corcho, O., Manghi, P., Di Nunzio, G. M., Golub, K., Ferro, N., and Poggi, A., editors, *"Linking Theory and Practice of Digital Libraries", Proceedings of the 26th International Conference on Theory and Practice of Digital Libraries, TPDL 2022, Padua, Italy, September 20-23, 2022*, pages 107–119, Cham. Springer International Publishing.

[8] Foufoulas, Y., Zacharia, E., Dimitropoulos, H., Manola, N., and Ioannidis, Y. (2023). Detexa: declarative extensible text exploration and analysis through sql. *International Journal on Digital Libraries*.

[9] Jayanetti, H. R., Garg, K., Alam, S., Nelson, M. L., and Weigle, M. C. (2022). Robots still outnumber humans in web archives, but less than before. In Silvello, G., Corcho, O., Manghi, P., Di Nunzio, G. M., Golub, K., Ferro, N., and Poggi, A., editors, *"Linking Theory and Practice of Digital Libraries", Proceedings of the 26th International Conference on Theory and Practice of Digital Libraries, TPDL 2022, Padua, Italy, September 20-23, 2022*, pages 245–259, Cham. Springer International Publishing.

[10] Jayanetti, H. R., Garg, K., Alam, S., Nelson, M. L., and Weigle, M. C. (2023). Robots still outnumber humans in web archives in 2019, but less than in 2015 and 2012. *International Journal on Digital Libraries*.

[11] Kumar, A., Ghosal, T., Bhattacharjee, S., and Ekbal, A. (2022). Investigations on meta review generation from peer review texts leveraging relevant sub-tasks in the peer review pipeline. In Silvello, G., Corcho, O., Manghi, P., Di Nunzio, G. M., Golub, K., Ferro, N., and Poggi, A., editors, *"Linking Theory and Practice of Digital Libraries", Proceedings of the 26th International Conference on Theory and Practice of Digital Libraries, TPDL 2022, Padua, Italy, September 20-23, 2022*, pages 216–229, Cham. Springer International Publishing.

[12] Kumar, A., Ghosal, T., Bhattacharjee, S., and Ekbal, A. (2023). Towards automated meta-review generation via an nlp/ml pipeline in different stages of the scholarly peer review process. *International Journal on Digital Libraries*.

[13] Martin, L. and Henrich, A. (2022). Rdftex: Knowledge exchange between latex-based research publications and scientific knowledge graphs. In Silvello, G., Corcho, O., Manghi, P., Di Nunzio, G. M., Golub, K., Ferro, N., and Poggi, A., editors, *"Linking Theory and Practice of Digital Libraries", Proceedings of the 26th International Conference on Theory and Practice of Digital Libraries, TPDL 2022, Padua, Italy, September 20-23, 2022*, pages 26–38, Cham. Springer International Publishing.

[14] Martin, L. and Henrich, A. (2023). Rdftex in-depth: knowledge exchange between latex-based research publications and scientific knowledge graphs. *International Journal on Digital Libraries*.

[15] Naiman, J. P., Williams, P. K. G., and Goodman, A. (2022). Figure and figure caption extraction for mixed raster and vector pdfs: Digitization of astronomical literature with ocr features. In Silvello, G., Corcho, O., Manghi, P., Di Nunzio, G. M., Golub, K., Ferro, N., and Poggi, A., editors, *"Linking Theory and Practice of Digital Libraries", Proceedings of the 26th International Conference on Theory and Practice of Digital Libraries, TPDL 2022,*

Padua, Italy, September 20-23, 2022, pages 52–67, Cham. Springer International Publishing.

[16] Naiman, J. P., Williams, P. K. G., and Goodman, A. (2023). The digitization of historical astrophysical literature with highly localized figures and figure captions. *International Journal on Digital Libraries*.