Illumine: A Tool to Augment the National Digital Library of India with Full Texts of Research Papers

Debarshi Kumar Sanyal*, Soumya Banerjee*, Gopal Agarwal†,
Samiran Chattopadhyay†, Plaban Kumar Bhowmick*, Partha Pratim Das*
*Indian Institute of Technology Kharagpur, Kharagpur – 721302, India.
†Department of Information Technology, Jadavpur University, Kolkata – 700098, India.
Email (in order): debarshisanyal@gmail.com, soumyabanerjee@outlook.in, gopalagarwal670@gmail.com,
samirancju@gmail.com, plaban@cet.iitkgp.ac.in, ppd@cse.iitkgp.ac.in

***Abstract*—The National Digital Library of India (NDLI) is envisioned as a national educational asset to enable 24x7 learning for learners of all ages and disciplines. It indexes research papers from various publishers, but the full text is often access-restricted and therefore, not freely available to users of NDLI. However, full texts of many papers are available in institutional digital repositories and preprint servers. We have developed a browser extension that allows NDLI users to automatically search the web and retrieve full texts of papers whenever they are available. We describe the design of the tool and report experiments done on a corpus of papers indexed in NDLI. The tool is freely available to all NDLI users.**

***Index Terms*—Open access; E-learning; Academic search; Digital library; Big scholarly data**

I. Introduction

The National Digital Library of India (NDLI) is a very large repository of metadata of educational contents spanning multiple disciplines, languages, and learner requirements. The actual educational content resides on the content provider's site like that of a University or an academic publisher. NDLI simply exposes a common search window a user can submit a query that will then be executed over the hundreds of repositories indexed by NDLI. The user can click on the results and view the metadata and the content. NDLI follows *view-at-source* policy, i.e., the user is redirected to the actual repository that hosts the contents. Therefore, the visibility of the content is limited by the user's access rights with respect to the content *source*. One major constituent of NDLI is its collection of metadata of research papers. It indexes various digital libraries like the ACM DL (`https://dl.acm.org/`), IEEE Xplore (`https://ieeexplore.ieee.org/`), SpringerLink (`https://link.springer.com/`), etc. More often than not, the same document is available in several locations like IEEE Xplore, ACM DL, arXiv (`https://arxiv.org/`), CiteSeerX (`https://citeseerx.ist.psu.edu/`), etc. Sometimes, one of the listed sources is an open access (OA) copy. Here, *OA* means free online access to an article without requiring personal or institutional subscription or other monetary payment from the user. We avoid complex aspects of various OA and sharing policies [1]. It is possible to apply filters on the search results in NDLI to identify OA articles. However, in many cases, research papers are not freely accessible in NDLI. Rather they require a subscription or a one-time access toll to be paid by the user. Given that journal subscriptions and paper download charges are typically high and many engineering colleges in India do not subscribe to international journals, NDLI members from these academic institutions cannot access these papers seamlessly from NDLI.

This paper presents a novel solution to the above access problem. We develop ILLUMINE, an extension for the Chrome browser (`https://www.google.com/chrome/`) to find an OA copy of a research paper present in NDLI. Our experiments show in many cases, ILLUMINE can discover OA copies and thus, make research more freely accessible to researchers. The extension and the experimental results are freely available at `https://github.com/soumyaxyz/illumine/`.

II. System Architecture

A. Overview of the proposed tool

We now introduce ILLUMINE, which is designed as a Chrome extension that finds open access full text for a research paper indexed in NDLI. The extension is visible as a black icon on the browser toolbar. It remains in idle mode until the user navigates to a document in NDLI. If the document is a research paper (as inferred from the resource metadata in NDLI), it searches for its OA full text. If the full text is freely accessible in the current page of NDLI, the button turns green. Otherwise, it attempts to locate an alternative full text from the web. If it succeeds, it turns green; the user can now simply click the extension button to open it in a new tab. If it fails, the extension button turns brown; clicking on it redirects to Google Scholar (`https://scholar.google.com/`) with the paper title (and if necessary, author name) as the query. Figure 1 shows how the ILLUMINE button appears in different cases.

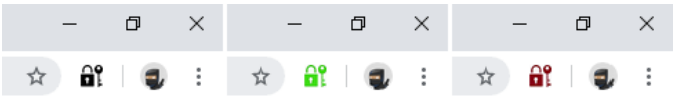

Fig. 1. ILLUMINE button (second from left on second line in each sub-figure) when it is idle, found full text, and could not find full text respectively.

B. Algorithmic aspects

Our search for OA copies of research papers relies on Open Access Button (OA Button) (`https://openaccessbutton.org/`) which is a web application to discover OA research papers. It indexes many sources, especially repositories from universities outside India. By limitation of scope, NDLI is unable to list these sources. Therefore, on detecting a closed access research paper in NDLI, our tool utilizes an Application Programming Interface (API) call to OA Button to search for OA of the paper. For example, for the paper [2], the OA Button query `https://api.openaccessbutton.org/availability?url=The Streaming Capacity of Sparsely Connected P2P Systems With Distributed Control` is triggered. If the returned Uniform Resource Locator (URL) is non-null, the button turns green. See Figure 2 for an example. The button stores the URL

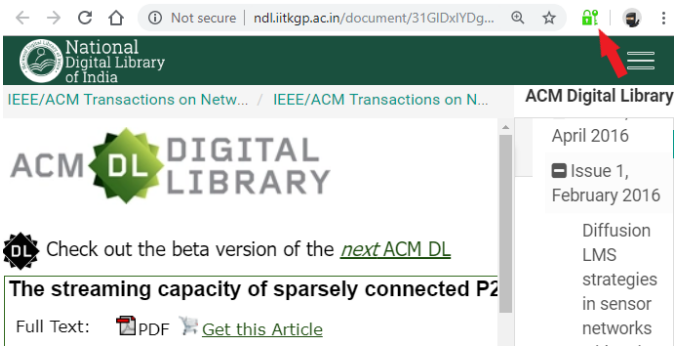

Fig. 2. ILLUMINE button appears in green (shown with the arrow on top-right of the browser window) because an OA full text of the paper is found. The OA full text is unavailable in NDLI but discovered through OA Button. Clicking on ILLUMINE button opens the full text in a new tab.

as *OA_Link*. If the user clicks on the button, the full text opens in a new browser tab. It is important to note that OA Button does not include academic social networks like ResearchGate (`https://www.researchgate.net/`), or personal repositories like Github (`https://github.com/`). If ILLUMINE is unable to locate an alternative OA copy through OA button, it provides the user an option to execute a search in Google Scholar. Google Scholar maintains a much more extensive repository [3], [4]. Unfortunately, its terms and services do not permit frequent script-based use of its query interface. So, ILLUMINE constructs a search string called *Search_URL* using the paper title and if necessary, the first author name and fires a query to Google Scholar allowing the user to further explore the search results for OA copies. For example, for the paper [5] whose OA full text is neither in NDLI nor found by OA Button, ILLUMINE turns brown; when user clicks on it, the query `https://scholar.google.co.in/scholar?hl=en&as_sdt=0%2C5&q="A Scalable and Resilient Layer-2 Network With Ethernet Compatibility"` is fired, and the results are displayed in a new tab for manual inspection. Figure 3 and 4 detail the working of ILLUMINE through self-explanatory state chart and activity diagrams respectively.

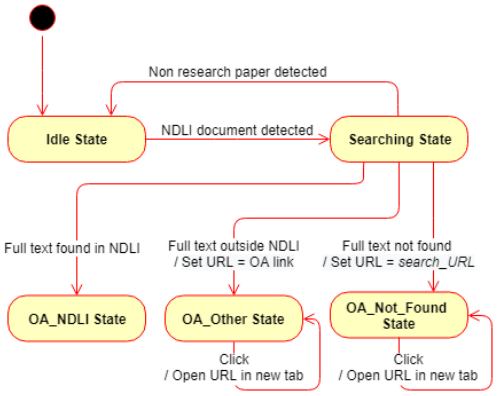

Fig. 3. State chart diagram for ILLUMINE. The `Idle` and `Searching` states show up in black; `OA_NDLI`, and `OA_Other` are both visible as green icons; `OA_Not_Found` appears in brown. The button in `OA_NDLI` state is not clickable.

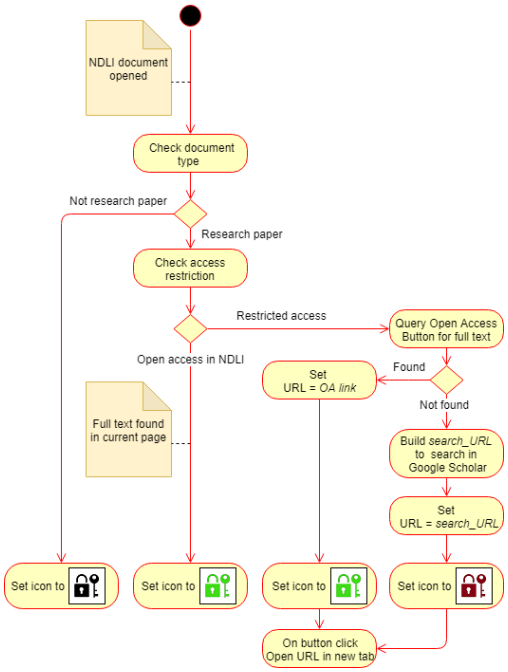

Fig. 4. Activity diagram for ILLUMINE.

C. Search parameters

ILLUMINE, depending on the input case, executes one automated search (through OA Button) and enables one manual search (through Google Scholar). In this section, we discuss the search parameters for these searches.

A research paper is uniquely identified by its Document Object Identifier (DOI). However, DOI is not always available in NDLI. Besides DOI, the title, author list, publication year and other publication metadata can be used to identify a research paper. As DOI references only the primary, publisher-maintained copy, more often than not it does not yield an OA

full text. Publication year and other publication details are not very useful, as this information may vary with the paper version (e.g., OA preprints have a different date compared to the published version). Title and author list are generally uniform across the different available copies of the paper. However, the author list representations vary across sources. It is almost trivial for a human to understand if two author lists are identical or not. However, the sheer variety in names and abbreviation conventions makes its a rather complicated check from an algorithmic perspective. Moreover, the title alone is generally adequate for locating copies of a given paper. Only for concise titles, the title might be inadequate to uniquely identify a paper and the use of author list becomes necessary. One notable issue regarding title and author list is that the latter has a higher rate of noise in NDLI. This noise manifests as an error in author order, quite easy to ignore as a human user but significant issues for an algorithmic approach.

OA Button supports search by DOI, URL, and title. The title is readily available in NDLI metadata, whereas the other two must be fetched from auxiliary sources. Hence, ILLUMINE uses only the paper title for API calls to OA Button. Multiple papers are unlikely to have the exact same title, so this method works well in practice.

Advanced Search in Google Scholar supports title, author names, and some other parameters. To balance computation load (on the machine running ILLUMINE) with detection accuracy, ILLUMINE will normally use only the paper title for building the query string for Google Scholar. In the case of concise titles, in particular, shorter than five words, the first author name is also included in the query string.

III. PERFORMANCE ANALYSIS

We have selected two top-tier journals listed in NDLI: *IEEE/ACM Transactions on Networking (TON)* and *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* for evaluation of ILLUMINE. We randomly selected 25 papers from the latest available volume of each of the above two journals in NDLI. Figures 5 and 6 summarize the results. We observe that around half of the research papers have full text freely available in NDLI. We found OA full text for almost all papers in *TON*: 44% are available in NDLI (through CiteSeerX and arXiv indexed in NDLI), 24% are found through OA Button, and in 28% cases, the query is forwarded to Google Scholar. We have manually checked that there are links to full text for these papers in Google Scholar. In case of one paper, namely [6], the URL returned by OA Button and therefore, ILLUMINE points to an access-restricted link in IEEE Xplore. Since ILLUMINE mistakenly believes the returned URL holds an OA copy, the button turns green but manual verification clubs it in the "Not available" category. In case of *TPAMI*, 52% papers are OA in NDLI, and 16% get OA copies through OA Button. In Google Scholar results, OA full text is not found for only three papers.

Fig. 5. Availability of OA full text of papers published in *IEEE/ACM Transactions on Networking, Vol. 24, 2016.*

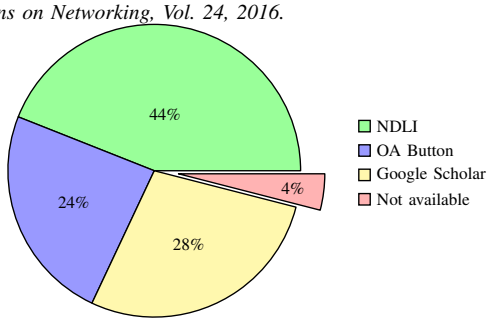

Fig. 6. Availability of OA full text of papers published in *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 38, 2016.*

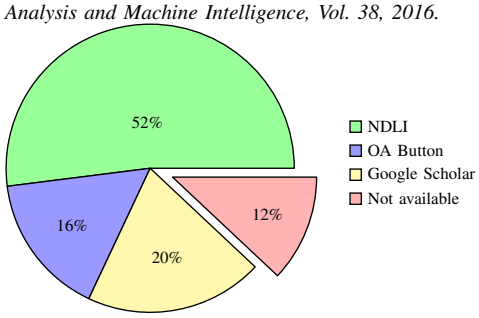

Figure 7 highlights the advantages of the ILLUMINE tool integrated with NDLI. Suppose the user navigates to a chosen search result from the main page of NDLI. Thereafter, of the 50 sampled papers, the user can find OA full text for 34 papers in a single click (i.e., directly clicking an *open* button in NDLI or the green ILLUMINE button). Another 12 can be reached through ILLUMINE followed by Google Scholar, and therefore, through two clicks.

Fig. 7. Number of clicks to access full text in our sample dataset.

Journal	One click	Two clicks	Not available
IEEE/ACM Transactions on Networking	17	7	1
IEEE Transactions on Pattern Analysis and Machine Intelligence	17	5	3

IV. CONCLUSION AND FUTURE SCOPE

In this paper, we presented the tool ILLUMINE to retrieve OA copies of access-restricted research papers in NDLI. We detailed the workings and advantages of the same. Its limitations include its upstream dependency on OA Button and Google Scholar. We have tested it only for papers in Computer Science, and it remains to be seen whether its performance will generalize well to other disciplines. We proposed the concept of surrogates of research publications in [7], [8]. In the future, we plan to integrate the concept of surrogacy into the presented tool so that wherever OA copies are not found, OA copies of suitable surrogates are returned.

ACKNOWLEDGMENT

This work is supported by *National Digital Library of India* Project sponsored by Ministry of Human Resource Development, Government of India at IIT Kharagpur.

REFERENCES

[1] H. Piwowar, J. Priem, V. Larivière, J. P. Alperin, L. Matthias, B. Norlander, A. Farley, J. West, and S. Haustein, "The state of OA: a large-scale analysis of the prevalence and impact of open access articles," *PeerJ*, vol. 6, p. e4375, 2018.

[2] C. Zhao, J. Zhao, X. Lin, and C. Wu, "Capacity of P2P on-demand streaming with simple, robust, and decentralized control," *IEEE/ACM Transactions on Networking (TON)*, vol. 24, no. 5, pp. 2607–2620, 2016.

[3] A. Martín-Martín, R. Costas, T. van Leeuwen, and E. D. López-Cózar, "Evidence of open access of scientific publications in Google Scholar: a large-scale analysis," *Journal of Informetrics*, vol. 12, no. 3, pp. 819–841, 2018.

[4] M. Gusenbauer, "Google Scholar to overshadow them all? comparing the sizes of 12 academic search engines and bibliographic databases," *Scientometrics*, vol. 118, no. 1, pp. 177–214, 2019.

[5] C. Qian and S. S. Lam, "A scalable and resilient layer-2 network with ethernet compatibility," *IEEE/ACM Transactions on Networking (TON)*, vol. 24, no. 1, pp. 231–244, 2016.

[6] H. Gao, V. Yegneswaran, J. Jiang, Y. Chen, P. Porras, S. Ghosh, and H. Duan, "Reexamining DNS from a global recursive resolver perspective," *IEEE/ACM Transactions on Networking (TON)*, vol. 24, no. 1, pp. 43–57, 2016.

[7] T. Santosh, D. K. Sanyal, P. K. Bhowmick, and P. P. Das, "Surrogator: A tool to enrich a digital library with open access surrogate resources," in *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. ACM, 2018, pp. 379–380.

[8] D. K. Sanyal, P. K. Bhowmick, P. P. Das, S. Chattopadhyay, and T. Santosh, "Enhancing access to scholarly publications with surrogate resources," *Scientometrics*, 2019, (To be published).