A Task-Focused View of the Community Digital Library

Kevin Ros and ChengXiang Zhai
Department of Computer Science, The Grainger College of Engineering, University of Illinois Urbana-Champaign
Email: kjros2@illinois.edu

INTRODUCTION

The Community Digital Library (CDL) is an online [1] and open-source [2] social bookmarking platform that supports task-focused search and recommendation.

The CDL consists of a stand-alone website and a Chrome browser extension. A user of the CDL can form communities, and they can describe and save webpages to these communities. If a user is a member of a community, then the user can search over all submissions made by all users to that community.

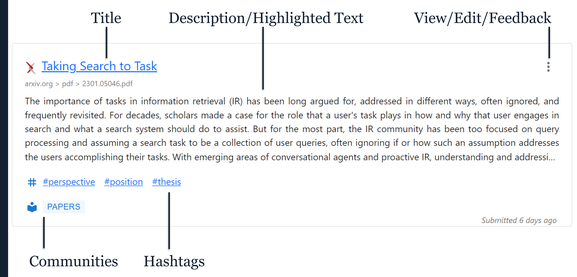

Figure 1: An example CDL submission

[1] https://textdata.org/

[2] https://github.com/thecommunitydigitallibrary/cdl-platform

MOTIVATION AND BACKGROUND

Why did we build the CDL?
- We wanted to investigate ways to support collaborative information seeking during knowledge-intensive tasks (especially for students in courses).
- There is a lack of open-source, free-to-use social bookmarking platforms, and access to the system and the data are critical for testing new search and recommendation techniques.

How is the CDL built?
- The website and extension are built using React and Next.js. The backend is written in Python, and the API endpoints are exposed through Flask.
- The data is stored in MongoDB and OpenSearch
- There is a fully local (offline) and cloud-based version.

How is the CDL related to other platforms?
See *Kevin Ros and ChengXiang Zhai. 2023. The CDL: An Online Platform for Creating Community-based Digital Libraries. In Computer Supported Cooperative Work and Social Computing (CSCW '23 Companion), October 14–18, 2023, Minneapolis, MN, USA. ACM, New York, NY, USA, 4 pages.*

COURSE-BASED TASK SUPPORT

The Structures of Student Information Needs
Students in the same upper-level college course face unique challenges in information seeking and discovery:

- Students have similar or share specific goals, such as understanding material, completing assignments, and attaining certain grades.
- Students often come from similar backgrounds, have completed the same pre-requisites, attend the same institution, and are a similar age.
- Students tend to complete coursework at similar times, as due dates and course milestones are shared.

As a consequence, students tend to have similar information needs at similar times. However, much of the information seeking is done independently, which results in wasted time and effort:

- Each student turns to online search engines, finding and evaluating online information independently from the other students.
- Instructors do not receive feedback on the students' understanding of the course material.

How can we leverage these implicit structures of a course to support students for more effective information seeking?

Enriching Course Content
We focus on the common task of needing additional information while watching lectures or completing assignments.

During the spring of 2023, in a large class at the University of Illinois, Urbana-Champaign, each student was assigned to submit, per lecture, a webpage that they found helpful for understanding the lecture's content. Submissions were made to a course-specific CDL community, and each submission was accessible by every student of the course.

The title and description of the submission were to include why the corresponding webpage was helpful for the lecture, and the hashtag was to indicate the corresponding lecture.

After filtering via student consent, there were ~1000 submissions made by ~40 students over ~20 lectures. The submissions covered a wide range of topics and were often diverse with respect to the domain and URL. The collected webpages helped other students to find lecture-related content and helped instructors better understand the topic focal points of lectures.

TOWARDS A GENERAL TASK FRAMEWORK

The data curation process from the course in the spring of 2023 required manual effort from the students. Ideally, we can design a system and framework which:
- Requires little to no effort by automatically curating, organizing, and recommending content.
- Leverages the implicit and explicit effort of past students to benefit future students of the same course, and across different courses.

Figure 2 describes how the CDL can be used to learn associations among webpages for contextual search and recommendation.

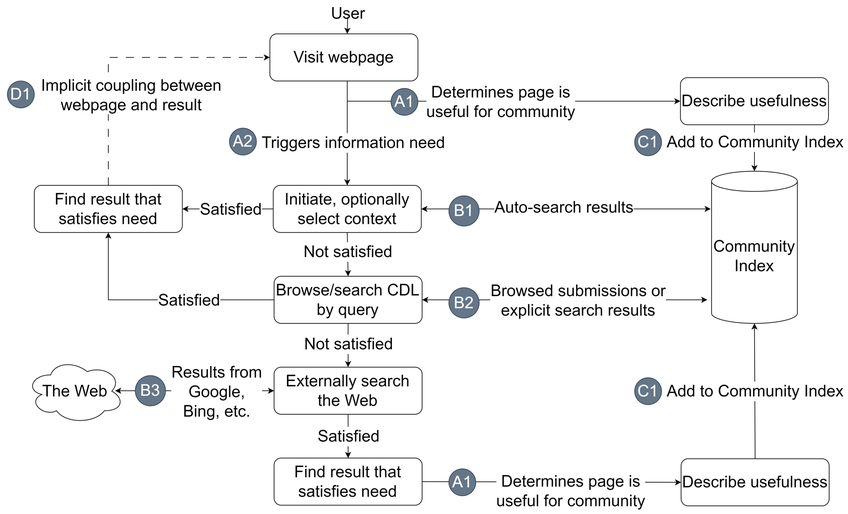

Figure 2: A framework for indexing and coupling webpages

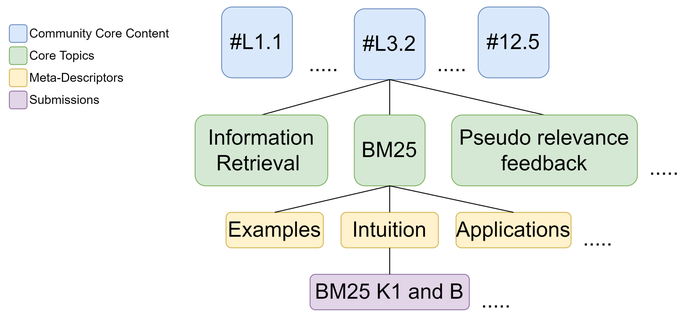

Figure 3: Organizing submissions around community core content

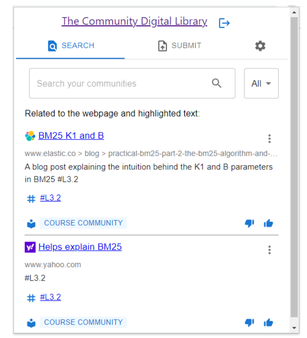

Figure 4: Contextual recommendations using the CDL browser extension

FUTURE WORK

Integrating Generative AI into the CDL
We plan to incorporate RAG-based result summarizations and multi-turn interactions through conversational agents. Additionally, we are investigating how webpages can be augmented with generative content to connect new information to a user's known information.

Data Organization and Model Improvement
The webpage submission process can be used to label data for downstream model training and for visualizing structured data representations (e.g., Figure 3).

Ideas for collaboration? Email kjros2@illinois.edu

ACKNOWLEDGEMENTS

We would like to thank Dhyey Pandya, Ashwin Patil, Kedar Takwane, and Rakshana Jayaprakash for their help in developing the CDL. This work is in part supported by the IBM-Illinois Discovery Accelerator Institute (IIDAI), by the NSF under grand number 1801652, and by the Jump ARCHES grant.

ILLINOIS