



Beyond PubMed and Google Scholar: Using Vertical Digital Library Portals to Enhance Exploratory Biomedical Literature Mining

Adam D. Troy¹, Guo-Qiang Zhang^{2*}

¹ Case Center for Proteomics, School of Medicine

² Department of Electrical Engineering and Computer Science
Case Western Reserve University, Cleveland, Ohio 44106, USA
{adam.troy, gq}@case.edu

Abstract. The scientific literature is a vast mine of past and current knowledge painstakingly constructed by the research community, and yet our means to access it is constrained by the current query interfaces and search engines, despite tremendous progresses made in its online availability. The problem is two-fold: one is to get to the optimal set of literatures in response to an information need, and the other is to digest the contents of the retrieved literature. We propose a vertical digital library portal (VDL) approach to address the first issue, particularly aimed at providing biomedical investigators and interested users an enhanced access interface with interactive, multifaceted presentation of results and focused content coverage. The VDL framework consists of four interrelated components: (a) newly enabled methodological techniques for information organization and retrieval; (b) a general-purpose system architecture; (c) a systematic and semi-automated process for constructing new verticals; (d) a feasible content harvesting, maintenance and update procedure. The focus of this paper will be on components (a) and (b), after the rationale for VDL is presented. New techniques include: chronological term rank, query expansion using term association, concept-guided content organization, cross-metadata search, and collaboration network analysis. The VDL system architecture features algorithms based on these techniques and supports dynamic, multifaceted organization of search results beyond paginated linear lists of items in the standard catalog-

* Corresponding Author. Email: gq@case.edu.

style. We also discuss our experiences with a publicly accessible experimental prototype system to demonstrate feasibility.

Keywords: vertical digital libraries, literature mining, social network, search-result organization.

1 Introduction

Biomedical researchers access the scientific literature of their discipline on a regular basis in order to integrate their studies with past and current research. They also search the literature for information about the latest trends, related techniques or new approaches, and similar efforts in comparable or distant fields. For these purposes, tools such as Google Scholar and PubMed have become indispensable for disseminating biological knowledge. A recent study [10] of PubMed's usage profile involved a typical day's query log in 2005, with around 2.7 million queries issued by over 600,000 users. The NCBI's published statistics [27] on PubMed indicates accelerated increase in usage, with monthly queries having quadrupled since 2000.

This trend is expected to continue as the ability to effectively search the scientific literature for biological information has become increasingly important with the rapid expansion of biomedical research [6] and its increased information dependence.

Access to scientific literature consists of two basic modes [4]. One is **targeted** (also called navigational, a somewhat misleading term), where a user – knowing what to look for – comes with specific pieces of information about the title or author, or other contextual data and tries to retrieve a corresponding set of abstracts or full articles from a digital library.

The second is **explorative**, also referred to as informational, where the goal is to explore and mine the literature when queries often cannot be easily or precisely formulated, either due to user's limited knowledge of the subject area or the use of synonyms and ambiguous terms in the subject area; all aspects of the well-known vocabulary problem [7]. This can result in a user getting few or too many hits. For example, a recent study [10] estimates that almost 75% of the queries to PubMed are exploratory, consistent with the profile of web searches [17]. This is an inevitable trend because the ever-increasing complexity of science and engineering requires many important research problems to be addressed by collaborative, multidisciplinary teams. To facilitate collaborative and interdisciplinary work, scientists and engineers face the growing need to access the literature beyond their immediate domain of expertise. The NLM estimated in 2002 [13] that one third of PubMed's users were members of the general public, while the remaining two thirds were healthcare professionals and researchers within or outside the medical domain. Using this as an estimate and assuming equal distribution of queries from the two populations, one estimates that 50% of all queries to PubMed have been exploratory ones submitted by researchers and healthcare professionals. Therefore, advances in supporting exploratory literature mining will be crucial for the effective dissemination of knowledge in science, engineering and medicine.