Chapter XVIII
Query Log Analysis in Biomedicine

Elmer V. Bernstam
University of Texas Health Science Center at Houston, USA

Jorge R. Herskovic
University of Texas Health Science Center at Houston, USA

William R. Hersh
Oregon Health & Science University, USA

ABSTRACT

Clinicians, researchers and members of the general public are increasingly using information technology to cope with the explosion in biomedical knowledge. This chapter describes the purpose of query log analysis in the biomedical domain as well as features of the biomedical domain such as controlled vocabularies (ontologies) and existing infrastructure useful for query log analysis. We focus specifically on MEDLINE, which is the most comprehensive bibliographic database of the world’s biomedical literature, the PubMed interface to MEDLINE, the Medical Subject Headings vocabulary and the Unified Medical Language System. However, the approaches discussed here can also be applied to other query logs. We conclude with a look toward the future of biomedical query log analysis.

INTRODUCTION

Biology and medicine are becoming information disciplines. One of the key drivers of the transformation is development of high-throughput techniques in molecular biology that generate large volumes of data quickly. For example, using a microarray costing a few hundred dollars, a single person can obtain expression data on thousands of genes within a few days. Briefly, a single gene expression microarray experiment can determine whether a large number of genes are more or less active in a particular state (e.g., cancer or some other disease) compared to normal.
These and other advances in biomedicine have led to an information explosion. No human can possibly manipulate data regarding thousands of genes without the aid of information technology. As a result, computers are now essential tools in biomedicine.

The vast amount of knowledge in many areas of biomedicine, and science in general, far exceeds the cognitive capacity of any human. Therefore, the paradigm must shift from knowing all of the answers before being asked the question, to “just in time learning” (i.e., retrieve knowledge after being asked the question). The only practical way to access the published literature in biomedicine is to rely on books, journals and electronic resources rather than human memory (Slawson & Shaughnessy, 2005). Fortunately, biomedicine benefits from relatively well-developed information retrieval resources based on controlled terminologies (ontologies). In this chapter, we will briefly review biomedical information retrieval, focusing on retrieval from the MEDLINE database. We discuss biomedical knowledge sources to which query log analysis has been applied. We then turn to problems and solutions in biomedical query log analysis. Finally, we discuss the contributions that query log analysis can make to the development of biomedical information retrieval systems.

BACKGROUND

The US National Library of Medicine (NLM) of the National Institutes of Health (NIH) developed and maintains many critical resources including databases, knowledge sources and software tools intended to allow access to biomedical information. The NLM “collects materials and provides information and research services in all areas of biomedicine and healthcare” (“About the National Library of Medicine,” 2007). When working with query logs in the biomedical domain, we make extensive use of NLM resources including MEDLINE, a variety of services via the Unified Medical Language System and PubMed, a search interface onto the biomedical literature indexed in MEDLINE.

MEDLINE

MEDLINE is the largest and most comprehensive database of biomedical literature in the world. MEDLINE is maintained by the NLM and is available via multiple interfaces created by a variety of commercial and non-commercial vendors such as Ovid, MD Consult and the NLM itself.

MEDLINE currently indexes over 16 million articles from over 5,000 journals and is growing by over 500,000 articles per year. Since these statistics change continuously, the interested reader should visit http://www.nlm.nih.gov/pubs/factsheets/medline.html for the latest data. For each article, MEDLINE contains bibliographic information including title, authors, journal, publication date and tags from a controlled vocabulary known as the Medical Subject Headings or MeSH (see below).

MEDLINE contains reference information, abstract (if available) and links to full-text articles when available, but does not contain the full-text articles themselves. It is therefore a bibliographic database. In contrast to full-text collections, users identify potentially interesting articles within MEDLINE, but obtain the actual article text elsewhere such as a physical library or journal Website. Biomedical journals increasingly provide online versions of content, therefore medical libraries are shifting from providing primarily print journals and photocopying facilities to providing access to online materials including journals, textbooks, databases and consulting services.

It is hard to overstate the significance of MEDLINE to biomedical researchers, students and clinicians. Indeed, if a biomedical journal is not indexed in MEDLINE, its quality is suspect and authors may think twice before publishing in that journal (“Annals of Family Medicine Selected for Indexing in Index Medicus and MEDLINE,”