1. Introduction

With the increasing use of the Internet and the World Wide Web, digital libraries have
coloned, and these serve a huge variety of different user audiences. With this expanded view
of libraries, two key insights arise. First, libraries are typically embedded within larger
institutions. Corporate libraries serve their corporations, academic libraries serve their
universities, and public libraries serve taxpaying communities who elect overseeing
representatives. Second, libraries play a pivotal role within their institutions as repositories and
providers of information resources. In the provider role, libraries represent in microcosm the
intellectual and learning activities of the people who comprise the institution. This fact provides
the basis for the strategic importance of library data mining: By ascertaining what users are
seeking, bibliomining can reveal insights that have meaning in the context of the library’s host
institution. Use of data mining to examine library data might be aptly termed bibliomining. With
widespread adoption of computerized catalogs and search facilities over the past quarter century,
library and information scientists have often used bibliometric methods (e.g., the discovery of
patterns in authorship and citation within a field) to explore patterns in bibliographic
information. During the same period, various researchers have developed and tested data mining
techniques—advanced statistical and visualization methods to locate non-trivial patterns in large
data sets. Bibliomining refers to the use of these bibliometric and data mining techniques to
explore the enormous quantities of data generated by the typical automated library. Web sites
contain information that ranges from the highly significant through to the trivial and obscene,
and because there are no quality controls or any guide to quality, it is difficult for searchers to
take information retrieved from the Internet at face value. The Internet will not become a serious
tool for professional searchers until the quality issues are resolved The Quality of Electronic
Information Products and Services, IMO One purpose of the academic library is to provide
access to scholarly research. Librarians select material appropriate for academia by applying a
set of explicit and tacit selection criteria. This manual task has been manageable for the world of
print. However, in order to aid selectors with the rapid proliferation and frequent updating of
Web documents, an automated solution must be found to help searchers find scholarly research
works published on the Web. Bibliomining, a.k.a. data mining for libraries, provides a set of
tools that can be used to discover patterns in large amounts of raw data, and can provide the
patterns needed to create a model for an automated collection development aid. One of the
difficulties in creating this solution is determining the criteria and specifications for the
underlying decision-making model. A librarian makes this decision by examining facets of the
document and determining from those facets if the work is a research work. The librarian is able
to do this because he/she has seen many examples of research works and papers that are not
research works, and recognizes patterns of facets that appear in research works. Therefore, to
create this model, many samples of Web-based scholarly research papers are collected along
with samples of other Web-based material. For each sample, a program in Perl (a pattern-
matching computer language) analyzes the page and determines the value for each criterion.
Different bibliomining techniques are then applied to the data in order to determine the best set
of criteria to discriminate between scholarly research and other works. The best model produced
by each technique is tested with a different set of Web pages. The models are then judged using
measures based the traditional evaluation techniques of precision and recall called accuracy and
return. Finally, the performance of each model is examined with a set of pages that are difficult
to classify. 1.1 Problem Statement Researchers need a digital library consisting of Web-based
scholarly works due to the rapidly growing amount of academic research published on the Web.
The general search tools overwhelm the researcher with no scholarly documents, and the subject-
specific academic search tools may not meet the needs of those in other disciplines. An
automated collection development agent is one way to quickly discover online academic research
works. In order to create a tool for identifying Web-based scholarly research, a decision-making
model for selecting scholarly research must first be designed. Therefore, the goal of the present
study is to develop a decision making model that can be used by a Web search tool to
automatically select Web pages that contain scholarly research works, regardless of discipline.
This tool could then be used as a filter for the pages collected by a traditional Web page spider,
which could aid in the collection development task for a scholarly digital library.
1.2 Definitions
1.2.1 Scholarly Research Works
To specify the types of resources that this predictive model will identify, the term “scholarly
research works” must be defined. For this study, scholarly research is limited to research written
by students or faculty of an academic institution, works produced by a non-profit research
instituion, or works published in a scholarly peer-reviewed journal. Research, as defined by
Dickinson in Science and Scientific Reasoning, is a “systematic investigation towards increasing the sum of knowledge”. This investigation, therefore, may be a literature review, a qualitative or quantitative study, a think piece, or another type of scholarly exploration. A research work is defined as a Web page (a single HTML or text file) that contains the full text of a research report. As the Web page has become the standard unit for indexing and reference by search tools and style manuals, the Web page is used here as the information container. 1.2.2 Accuracy revision and Return / Recall the models are judged using measures named accuracy and return; these are based off the traditional IR measures of precision and recall. Accuracy (precision) and return (recall) are both defined in their classical information retrieval sense, as first defined by Cleverdon (1962). Accuracy is measured by dividing the number of pages that are correctly identified as scholarly research by the total number of pages identified as scholarly research by the model. Return is determined by dividing the number of pages correctly identified as scholarly research by the total number of pages in the test set that is scholarly research. When applied to the Web as a whole, return cannot be easily defined. However, a higher return in the test environment may indicate which tool will be able to discover more scholarly research published on the Web. 1.2.3 Problematic Pages Problematic pages are Web pages that might appear to this agent to be scholarly research works but are not. Categories of problematic pages are author biographies, syllabi, vitae, abstracts, corporate research, research that is in languages other than English, and pages containing only part of a research work. Future researchers will want to incorporate some of these categories into digital library tools and this level of failure analysis will assist those researchers in adjusting the models presented in this research. For several decades, library and information services in corporations, schools, universities, and communities have had the ability to capture information about their users, circulation history, resources in the collection, and search patterns (Koenig, 1985). Collectively, these data can provide library managers more information about common patterns of user behavior to aid in decision-making processes. Unfortunately, few libraries have taken advantage of these data as a way to improve customer service, manage acquisition budgets, or influence strategic decision-making about uses of information in their organizations. The application of advanced statistical and data mining techniques to these kinds of data may provide useful ways of supporting decision-making at every library where user, cataloging, searching, and circulation interfaces are automated.[16] Use of data mining to examine library data records might be aptly termed bibliomining. With
widespread adoption of computerized catalogs and search facilities over the past quarter century, library and information scientists have often used bibliometric methods (e.g. the discovery of patterns in authorship and citation within a field) to explore patterns in bibliographic information. During the same period, various researchers have developed and tested data mining techniques -- advanced statistical and visualization methods to locate non-trivial patterns in large data sets. Bibliomining refers to the use of these techniques to plumb the enormous quantities of data generated by the typical automated library. Forward-thinking authors in the field of library science began to explore sophisticated uses of library data some years before the concept of data mining became popularized. Nutter (1987) explored library data sources to support decision making, but lamented that “the ability to collect, organize, and manipulate data far outstrips the ability to interpret and to apply them”. Johnston and Weckert (1990) developed a data-driven expert system to help select library materials and Vizine-Goetz, Weibel, & Oskins (1990) developed a system for automated cataloging based on book titles (also see Morris, 1992; Aluri and Riggs, 1990). A special section of Library Administration and Management (“Mining your automated system”) included articles on extracting data to support system management decisions (Mancini, 1996), extracting frequencies to assist in collection decision-making (Atkins, 1996), and examining transaction logs to support collection management (Peters, 1996). More recently, Banerjee (1998) focused on describing how data mining works and ways of using it to provide better access to the collection. Guenther (2000) discussed data sources and bibliomining applications, but focused on the problems with heterogeneous data formats. Doszkocs (2000) discussed the potential for applying neural networks to library data to uncover possible associations between documents, indexing terms, classification codes, and queries. Liddy (2000) combined natural language processing with text mining to discover information in “digital library” collections. Lawrence, Giles, and Bollacker (1999) created a system to retrieve and index citations from works in digital libraries. Gutwin, Paynter, Witten, Nevill-Manning, and Frank (1999) used text mining to support resource discovery. These projects all shared a common focus on improving and automating two of the core functions of a library – acquisitions and collection management. What these projects did not discuss was the use of library data to support strategic management decisions for libraries and their host institutions. A few authors have recently begun to address this need by focusing on understanding library users: Schulman (1998) discussed using data mining to examine changing trends in library user behavior; Sallis,
Hill, Jance, Lovetter, and Masi (1999) created a neural network that clusters digital library users; and Chau (2000) discussed the application of Web mining to personalize services in electronic reference. We extend these efforts by taking a more global view of the data generated in libraries and the variety of decisions that those data can inform. Thus, the focus of this chapter is on describing ways in which library and information managers can use data mining to understand patterns of behavior among library users and staff and patterns of information resource use throughout the institution. The chapter will examine data sources and possible applications of data mining techniques as well as explore the legal and ethical implications of bibliomining.

Data Mining in Library and Information Services One of the challenges put forth to the library profession by Michael Buckland (2003) is to gain a better understanding of library user communities. Most current library evaluation Techniques focus on frequencies and aggregate measures; these statistics hide underlying Patterns. Discovering these patterns key in understanding the communities that use library Services. In order to tailor services to meet the needs of different user groups, library Decision-makers can use the bibliomining process to uncover patterns of data-based artifacts of use. The bibliomining process, which consists of data warehousing and data mining, the term “bibliomining” was first used by Nicholson and Stanton (2003) in discussing data mining for libraries. In the research literature, most works that contain the terms “library” and “data mining” are not talking about traditional library data, but rather using library in the context of software libraries, as data mining is the application of techniques from a large library of tools. In order to make it more conducive for those concerned with data mining in a library setting to locate other works and other researchers, the term “bibliomining” was created.[9]The term pays homage to bibliometrics, which is the science of pattern discovery in scientific communication. Bibliomining is the application of statistical and pattern-recognition tools to large amounts of data associated with library systems in order to aid decision-making or justify services. The bibliomining process consists of:
· determining areas of focus;
· identifying internal and external data sources;
· collecting, cleaning, and anonymizing the data into a data warehouse;
· selecting appropriate analysis tools;
· Discovery of patterns through data mining and creation of reports with traditional Analytical tools
Analyzing and implementing the results.

The process is cyclical in nature: as patterns are discovered, more questions will be raised which will start the process again? As additional areas of the library are explored, the data Warehouse will become more complete, which will make the exploration of other issues much Easier. There are many benefits to using bibliometrics to study the growth of the CI Literature. The purpose of bibliometrics “is to shed light on the process of written Communications and the nature and course of a discipline, by means of counting and Analyzing the various facets of written communication” (Pritchard, 1969). Librarians and Information professionals benefit most from the practical application of bibliometric data, especially since this information is useful in bibliographic control, database evaluation, and collection development. By determining core authors, core journals and CI literature Growth and size, criteria are established on which to base decisions on database Evaluation and journal selection, retention or cancellation. Examining the bibliometric Characteristics of CI literature will reveal the structure and impact of CI and will clarify Bibliometric studies are like a mirror: they give researchers a chance to glimpse the reflection of many facets of a research area (Sellen 1993). The scarcity of bibliometric research pertaining to the production and distribution of CI-related literature coupled with the increased attention to competitive information gathering signals the need and importance of conducting a bibliometric analysis on this topic. With the everincreasing interest in competitive intelligence, and the clear relationship CI has to the field of library and information science (LIS), it is important that a bibliometric study be conducted to more fully understand the development of the CI discipline and the related implications on the business and library science bodies of work. Although business journals and scholarly articles frequently cover CI topics, indepth authorship patterns surrounding this body of work has not been explored. Furthermore, no study has explored the CI discipline and the bibliometric characteristics related to LIS, further necessitating research on library literature as it relates to CI. Inresearching the bibliometric characteristics of competitive intelligence literature in business and in answering this question, prolific authors, core journals, and overlap between databases will be examined to understand fundamental differences.