Nicholson, S. (2003). Bibliomining for automated collection development in a digital library setting: Using data mining to discover web-based scholarly research works. *Journal of the American Society for Information Science and Technology 54*(12). 1081-1090.

Bibliomining for Automated Collection Development in a Digital Library Setting: Using Data Mining to Discover Web-Based Scholarly Research Works

Scott Nicholson

Syracuse University School of Information Studies

4-127 Center for Science and Technology

Syracuse, NY 13244

Phone: 315-443-1640

Fax: 315-443-5806

http://www.scottnicholson.com

http://www.bibliomining.org

scott@scottnicholson.com

0. ABSTRACT

This research creates an intelligent agent for automated collection development in a digital library setting. It uses a predictive model based on facets of each Web page to select scholarly works. The criteria came from

the academic library selection literature, and a Delphi study was used to refine the list to 41 criteria. A Perl program was designed to analyze a Web page for each criterion and applied to a large collection of scholarly and non-scholarly Web pages. Bibliomining, or data mining for libraries, was then used to create different classification models. Four techniques were used: logistic regression, non-parametric discriminant analysis, classification trees, and neural networks. Accuracy and return were used to judge the effectiveness of each model on test datasets. In addition, a set of problematic pages that were difficult to classify because of their similarity to scholarly research was gathered and classified using the models.

The resulting models could be used in the selection process to automatically create a digital library of Web-based scholarly research works. In addition, the technique can be extended to create a digital library of any type of structured electronic information.

Keywords

Digital Libraries, Collection Development, World Wide Web, Search Engines, Bibliomining, Data Mining, Intelligent Agents

1. INTRODUCTION

> 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 (Nicholson and Stanton, in press and Nicholson, 2002).

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 non-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 institution, or works published in an scholarly peer-reviewed journal. Research, as defined by Dickinson in Science and Scientific Reasoning, is a “systematic investigation towards increasing the sum of knowledge” (1984, pg. 33). This investigation, therefore, may be a literature review, a qualitative or quantitative study, a thinkpiece, 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 / Precision 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 are scholarly research. When applied to the Web as a whole, return can not 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 (as defined above in 1.2.1), 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.

1.3 Research Overview

First, a set of criteria used in academic libraries for print selection is collected from the literature, and a Delphi study was done with a panel of librarians to refine the list. The criteria are then translated into terms appropriate for Web documents, and a Perl program was written that collects aspects of the a Web that correspond to the criteria.

This data collection tool is used to gather information on 5,000 pages with scholarly research works and 5,000 pages without these works. This data set is split, with the majority of the pages used to train the models and the rest used to test the models. The training set is used to create different models using logistic regression, memory-based reasoning (through non-parametric n-nearest neighbor discriminant analysis), decision trees, and neural networks.

Another set of data is used to tweak the models and make them less dependent on the training set. Each model is then applied to the testing set. Accuracy and return is determined for each model, and the best models are identified.

1.4 Literature Review

This section explores closely related literature and the placement of this research in the areas of the selection of quality materials, data mining and similar projects.

1.4.1 Selection of Quality Materials

Should the librarian be a filter for quality? S.D. Neill argues for it in his 1989 piece. He suggests librarians, along with other information professionals, become information analysts. In this article, he suggests that these information analysts sift through scientific articles and remove those that are not internally valid. By looking for those pieces that are “poorly executed, deliberately (or accidentally) cooked, fudged, or falsified”(Neill, 1989, pg. 6), information analysts can help in filtering for quality of print information.

Piontek and Garlock also discuss the role of librarians in selecting Web resources. They argue that collection development librarians are ideal in this role because of “their experience in the areas of collection,

organization, evaluation, and presentation" (1996, pg. 20). Academic librarians have been accepted as quality filters for decades. Therefore, the literature from library and information science will be examined for appropriate examples from print selection and Internet resource selection of criteria for quality.

1.4.1.1 Selection of Print Materials

The basic tenet in selection of materials for a library is to follow the library's policy, which in an academic library is based upon supporting the school's curriculum (Evans, 2000). Because of this, there are not many published sets of generalized selection criteria for academic libraries.

One of the most well-known researchers in this area is S. R. Ranganathan. His five laws of librarianship (as cited in Evans, 2000) are a classical base for many library studies. There are two points he makes in this work that may be applicable here. First, if something is already known about an author and the author is writing the same area, then the same selection decision can be made with some confidence. Second, selection can be made based upon the past selection of works from the same publishing house. The name behind the book may imply quality or a lack thereof, and this can make it easier to make a selection decision.

Library Acquisition Policies and Procedures (Futas, 1995) is a collection of selection policies from across the country. By examining these policies from academic institutions, one can find the following criteria for quality works that might be applicable in the Web environment:

- Authenticity
- Scope and depth of coverage
- Currency of date
- Indexed in standard sources
- Favorable reviews
- Reference materials like encyclopedias, handbooks, dictionaries, statistical compendia, standards, style manuals, and bibliographies.

1.4.1.2 Selection of Online and Internet Resources

Before the Internet was a popular medium for information, libraries were faced with electronic database selection. In 1989, a wish list was created for database quality by the Southern California Online Users Group (Basch, 1990). This list had 10 items, some of which were coverage, scope, accuracy, integration, documentation, and value-to-cost ratio.

This same users group discussed quality on the Internet in 1995 (as cited in Hofman and Worsfold, 1999). They noted that Internet resources were different from the databases because those creating the databases were doing so to create a product that would produce direct fiscal gain, while those creating Internet resources, in general, were not looking for this same gain. Because of this fact, they felt that many Internet resource providers did not have the impetus to strive for a higher-quality product.

The library community has produced some articles on selecting Internet resources. Only those criteria dealing with quality that could be automatically judged will be discussed from these studies. The first such published piece, by Cassel in 1995, does not mention the WWW; the most advanced Internet technologies discussed were Gopher and WAIS. She states that Internet resources should be chosen with the same criteria as print resources, such as adherence to the curriculum and supporting faculty research. Other criteria mentioned are the comprehensiveness of the resource, authoritativeness of the creator of the resource, and the systematic updating of the source. However, Cassel feels that unlike in the print world where shelf space is limited, duplication of Internet resources in a collection is not problematic.

A year later, a more formal list of guidelines for selecting Internet resources were published. Created by Pratt, Flannery, and Perkins (1996), this remains one of the most thorough lists of criteria to be published. Some of the criteria they suggest that relate to this problem are:

- Produced by a national or international organization, academic institution, or commercial organization with an established reputation in a topical area
- Indexed or archived electronically when appropriate

2. Linked to by other sites

- Document is reproduced in other formats, but Internet version is most current

- Available on-line when needed
- Does not require a change in existing hardware or software

Another article from 1996 by the creators of the Infofilter project looked at criteria based on content, authority, currency, organization, the existence of a search engine on the site, and accessibility. However, their judging mechanisms for these criteria were based upon subjective human judgments for the most part. Exceptions were learning the institutional affiliation of the author, pointers to new content, and response time for the site.

One new criterion is introduced in a 1998 article about selecting Web-based resources for a science and technology library collection: the stability of the Web server where the document lives. While this does not necessarily represent the quality of the information on the page, it does affect the overall quality of the site. Sites for individuals may not be as acceptable as sites for institutions or companies (McGeachin, 1998).

Three Web sites provide additional appropriate criteria in selecting quality Internet resources. The first is a list of criteria by Alastair Smith in the Victoria University of Wellington LIS program in New Zealand (1997). He looks first at scope by looking for meta information or an introduction discussing the scope of the page. Then, content is judged by looking for original information, political, ideological, or commercial biases on the site, and by looking at the URL for clues about authoritativeness. The final criterion useful for this project is reviews; just as librarians have depended upon reviews for book selection, Web review sites can used to aid in automatic selection.

The second site adopts criteria for selecting reference materials presented in Bopp and Smith's 1991 reference services textbook. Many of the criteria presented have already been discussed in this review, but one new quality-related idea was presented. Discriminating the work of faculty or professionals from the work of students or hobbyists may aid in selecting works that are more accurate and reliable. While this is not always the case, an expert will usually write a better work than a novice (Hinchliffe, 1997).

The final site, that of the DESIRE project, is the most comprehensive piece listed here. The authors (Hofman and Worsfold, 1999) looked at seventeen online sources and five print sources to generate an extensive list of selection criteria to help librarians create pages of links to Internet cites. However, many of the criteria have

either already been discussed here or require a human for subjective judging.

There were only a few new criteria appropriate to the research at hand. In looking at the scope of the page, these authors suggest to look for the absence of advertising to help determine quality of the page. Metadata might also provide a clue to the type of the material on the page. In looking at the content of the page, references, a bibliography, or an abstract may indicate an scholarly work. Pages that are merely advertising will probably not be useful to the academic researcher. A page that is inward focused will have more links to pages on its own site than links to other sites, and may be of higher quality. In addition, clear headings can be a judge for a site that is well organized and of higher quality. The authors also suggest looking at factors in the medium used for the information and the system on which the site is located. One new criterion in this area is the durability of the resource; sites that immediately direct the user to another URL may not be as durable sites with a more "permanent" home.

2.1 Data Mining

Once the criteria have been operationalized and collected with the Perl program for a large sample of pages that are linked to academic library Web sites and for another sample of sites that are not scholarly, patterns must be found to help classify a page as scholarly. Data mining will be useful for this, as it is defined as "the basic process employed to analyze patterns in data and extract information" (Trybula ,1997, pg. 199). Data mining is actually the core of a larger process, known as knowledge discovery in databases (KDD). KDD is the process of taking low-level data and turning it into another form that is more useful, such as a summarization or a model (Fayyad, Piatetsky-Shapiro, and Smyth, 1996).

There are a large number of tools available to the data miner, and the tools used must match the task. In the current task, the goal is to look at a database of classified documents, and decide if a new document belongs in an academic library. Therefore, this is a classification problem. According to the Berry and Linoff classic data mining text (1997), some of the tools that may be useful are standard statistics, memory-based reasoning, decision trees, and neural networks. Each will be briefly discussed with this project in mind.

In order to use standard statistics, a technique would be needed that can handle both continuous and

categorical variables and will create a model that will allow the classification of a new observation. According to Sharma (1996), logistic regression would be the technique to use. In this, the best combination of variables is discovered that maximizes the correct predictions for the current set and is used to predict membership of the new observation. This methodology looks for the best combination of variables to produce a prediction. For this project, however, there will be different types of Web pages that are deemed appropriate, and thus it may prove difficult to converge on a single solution using logistic regression.

Memory-based reasoning is where a memory of past situations is used directly to classify a new observation. N-neighbor non-parametric discriminant analysis is one statistical technique used for MBR. This concept was discussed in 1988 by Stanfill and Waltz in The Memory Based Reasoning Paradigm at a DARPA workshop. In MBR, some type of distance function is applied to judge the distance between a new observation and each existing observation, with optional variable weighting. The program then looks at a number of the preclassified neighbors closest to the new observation and makes a decision (Berry and Linoff, 1997).

Decision/Classification trees use a large group of examples to create rules for making decisions. It does this in a method similar to discriminant analysis; it looks for what variable is the best discriminator of the group, and splits the group on that variable. It then looks at each subgroup for the best discriminator and splits the group again. This continues until a set of classified rules is generated. New observations are then easily classified with the rule structure (Johnston and Weckert, 1990).

Neural networks are based on the workings of neurons in the brain, where a neuron takes in input from various sources, processes it, and passes it on to one or more other neurons. The neuron accepts 0-1 measurements of each variable. It then creates a hidden layer of neurons, which weights and combines the variables in various ways. Each neuron is then fed into an output neuron, and the weights and combinations of the neurons are adjusted with each observation in the training set through back-propagation until an optimal combination of weights is found (Hinton, 1992).

Neural networks are very versatile, as they do not look for one optimal combination of variables; instead, several different combinations of variables can produce the same result. They can be used in very complicated domains where rules are not easily discovered. Because of its ability to handle complicated

problems, a neural network may be the best choice for this problem (Berry and Linoff, 1997).

2.1.1 Data Mining in Libraries

Several researchers have discussed the appropriateness of using data mining techniques in libraries. May Chau presents several possible theoretical links between academic librarianship and data mining. She explores Web mining (data mining on the World Wide Web) as a tool to help the user find information. Not only can Web mining be used to create better search tools, but also it can be used to track the searching behavior of users. By tracking this information, librarians could create better Web sites and reference tools (1999).

In addition, Kyle Banerjee explores ways that data mining can help the library. In discussing possible applications, he says "full-text, dynamically changing databases tend to be better suited to data mining technologies" (1998, pg. 31). As the Web is a full-text, dynamically changing database, it is indeed appropriate to use these technologies to analyze it.

A new term to describe the data mining process in libraries is Bibliomining (Nicholson and Stanton, In press). Bibliomining is defined as "the combination of data mining, bibliometrics, statistics, and reporting tools used to extract patterns of behavior-based artifacts from library systems" (Nicholson, 2002). Instead of behavior-based artifacts, however, this project is using bibliomining to discover patterns in artifacts contained in and associated with Web pages. The techniques to discover novel and actionable patterns still apply.

2.2 Similar Projects

There are many manually-collected digital libraries of scholarly research works, two of the largest are Infomine(http://infomine.ucr.edu) in the United States and BUBL(http://bubl.ac.uk) in the United Kingdom. Other projects are European Research Papers Archive, which indexes pages from online series of papers (Nentwich, 1999), and Argos and Noesis by the Internet Applications Laboratory, which index papers in ancient studies and philosophy (Beavers,1998). However, being created by hand, these libraries can not keep up with the rapid production of Web pages.

There are currently several projects that automatically gather scholarly Web pages. Lawrence, Giles, and

Bollacker have created CiteSeer (now called ResearchIndex), which is based around citations and link analysis. In order to verify that the page is a research article, the tool looks to see if there is a works cited section (Lawrence, Giles, and Bollacker, 1999). Another project to identify scholarly research works is CORA. This tool selects scholarly Web pages in the computer science domain by visiting computer science department Web sites and examining all of the Postscript and PDF documents, keeping those which have sections commonly found in a research paper(McCallum, Nigam, Rennie, and Seymore, 1999). Both ResearchIndex and CORA might benefit from an expansion of their inclusion criteria using the models presented in this paper.

In addition, Yulan and Cheung (2000) created PubSearch. This tool creates customizes searches for a user by taking a selection of articles and searching for related articles through citation and author analysis. This tool, therefore, is useful for users who have already done research in an area and would like to discover similar research. This research could provide a filter for PubSearch to use in order to go beyond the user's specified Web sites.

3. Methodology

3.1 Criteria Creation

A list of criteria used to select academic research was gathered from a literature review of criteria used in selecting print and electronic documents for academic libraries (Nicholson, 2000). This list was presented to a panel of 42 librarians. The criteria were ranked and the librarians were allowed to suggest new criteria. The list was then changed to remove low-ranking criteria and add new suggested criteria. This process was repeated until consensus was reached. A summary of the final list of criteria follows.

Summary of Selection Criteria for Web Pages

Author Criteria

Author has written before

Experience of the author

Authenticity of author

Content Criteria

Work is supported by other literature

Scope and depth of coverage

Work is a reference work

Page is only an advertisement

Pages are inward focused

Writing level of the page

Existence of advertising on the site

Original material, not links or abstracts

Organizational Criteria

Appropriate indexing and description

There is an abstract for the work

Pages are well-organized

Currency of date/ Systematically updated

Producer/Medium Criteria

Document is reproduced in other forms

Available on-line when needed

Does not require new hardware or software

Past success/failure of the publishing house

Produced by a reputable provider

Unbiased material

Stability of the Web server

Response time for the site

Site is durable

External Criteria

Indexed in standard sources

Favorable reviews

Linked to by other sites

3.2 Design

A Perl program was then created that would retrieve a Web page and analyze it in regard to each criterion. The part of the program to analyze each criterion was developed and tested before being integrated into the entire program. Once the program was complete, it was tested on other pages to ensure that the program was working correctly.

3.3 Page Collection Techniques

In order to collect pages containing scholarly research works, several techniques were employed. Requests were posted to scholarly discussion lists, online journals and conference proceedings were explored, and search tools were utilized. Only Web pages that were free to access, written by someone in academic or a non-profit research institution or published in an scholarly peer-reviewed journal, were in HTML or text, and contained the full text of the research report on a single Web page were accepted. As some sites had many scholarly works, no more than 50 different works were taken from a single site. After 4,500 documents were collected for the model creation sets, another 500 were collected for the test set. Care was taken to ensure that none of the documents in the test set came from the same Web site as any other document in the model or test set.

In order to create models that can discriminate between pages with scholarly works and those without, a set of pages not containing scholarly works was gathered. Since this agent was designed to work with the pages captured by a typical Web spider, the non-scholarly pages for model-building were taken from the Web search tools. The first step in selecting random pages was to use Unfiltered MetaSpy (http://www.metaspy.com). MetaSpy presents the text of the last 12 searches done in MetaCrawler. These queries were extracted from the MetaSpy page and duplicates were removed.

These queries were then put into several major search tools. The first ten URLs were extracted from the

resulting page and one was selected at random and verified to make sure the page was functioning through a Perl program. Each page was then manually checked to ensure that it did not contain scholarly research. The next query from Search Voyeur was then used to perform another search. This process continued until 4,500 URLs were gathered for the model building sets. The same technique was used for the test set with another search tool providing the pages.

Each of the 10,000 URLs was then given to the Perl program to process. For each page, the HTML was collected and analyzed, and the URL submitted to four different Web search tools and analysis tools in order to collect values for some of the criteria. After this, the datasets were cleaned by manually examining them for missing data, indicators the page was down, or other problems.

After the data were cleaned, the datasets were prepared for model development and testing. One set of 8,500 document surrogates was created for model creation, and a second set of 500 document surrogates was created for tweaking the models. The third dataset consisted of the 1,000 documents selected for testing. Each of these sets had equal numbers of documents with and without scholarly research works. Finally, the dataset of surrogates for the problems pages was prepared.

3.4 Analysis of Web Page Surrogates through Data Mining

Four models were then created and tested using different data mining techniques. In SAS 6.12, logistic regression and n-nearest neighbor nonparametric discriminant analysis were used to create models. Clementine 5.0 was used to create a classification tree and a neural network for prediction. Each model was created with the large dataset and tested against the tweaking dataset. If there were settings available, these were adjusted until the model produced the best results with the tweaking dataset. Once settings were finalized, the testing dataset were run through the model. The actual group membership was compared to the predicted group membership in order to determine accuracy and return for each model.

4. Model Exploration and Performance Discussion

4.1 Logistic Regression

4.1.1 Model Description

Stepwise logistic regression selects a subset of the variables to create a useful, yet parsimonious, model. In this case, SAS selected 21 criteria for inclusion in the model.

The R^2 for this regression was .6973. On the model-building dataset, the model was 99.3% accurate. All of the criteria used to start a stepwise regression, and the ones that remained in this model were:

- Clearly stated authorship at the top of the page
- Number of age warnings and adult-content keywords
- Statement of funding or support at the bottom of page
- Number of times a traditional heading appeared on the page (such as Abstract, Findings, Discussion, etc.)
- Presence of labeled bibliography
- Presence of a banner ad from one of the top banner ad companies
- Existence of reference to “Table 1” or “Figure 1”
- Existence of phrase “presented at”
- Academic URL
- Organizational URL
- Existence of a link in Yahoo!
- Number of full citations to other works
- Existence of meta tags
- Number of words in the meta keyword and dc.subject meta tags
- Average sentence length
- Average word length

- Total number of sentences in document
- Average number of sentences per paragraph
- Ratio of total size of images on page to total size of page
- Number of misspelled words according to Dr. HTML
- Average length of misspelled words.

4.1.2 Performance on Test Dataset

The model created by logistic regression correctly classified 463 scholarly works and 473 randomly chosen pages. Therefore, it has a accuracy of 94.5% and a return of 92.6%. It had problems with non-scholarly pages that were in the .edu domain, that contained a large amount of text, or that contained very few external links. In addition, it had problems identifying scholarly pages that were in the .com domain, that did not use traditional headings or a labeled bibliography, or that contained large or numerous graphics.

4.1.3 Performance on Problematic Dataset

This model misclassified 30% of the documents in te problematic dataset. It had the most difficulty with non-annotated bibliographies, vitae, and research proposals; however, it correctly classified all of the non-scholarly articles.

4.2 Discriminant Analysis (Memory-Based Reasoning)

4.2.1 Model description

This technique does memory-based reasoning by using all of the variables to plot a point for each identified page. New pages are plotted in the space, and the model looks at the nine nearest neighbors. The classification of the majority of those neighbors is assigned to the new page. There is no way to tell which variables are most useful in the model. This model correctly identified the items in the model dataset 97.74% of the time.

4.2.2 Performance on Test Dataset

This model classified 475 non-scholarly works and 438 scholarly works correctly. Therefore, it had a accuracy of 94.6% and return of 87.6%. It had many of the same problems as the logistic regression model. Long textual pages, pages with few graphics, and pages in the .edu domain were common features of misclassified non-scholarly pages. Scholarly pages that were misclassified usually had two of the following features: many graphics, no labeled bibliography, unusual formatting such as forced line and paragraph breaks or many tables, no traditional headings, or from a commercial domain. In addition, any page on one of the free home page servers (Geocities, Xoom) was deemed as non-scholarly. This criterion was removed and the model was generated again to see if there was some underlying problem, but the performance was worse without that criterion.

4.2.3 Performance on Problematic Dataset

This tool classified almost every item in the problematic dataset as scholarly. It classified only 17 out of the 200 as being non-scholarly; thus it was incorrect 91.5% of the time on these difficult pages. It performed the best with abstracts, only misclassifying about half of them.

4.3 Classification Tree

4.3.1 Model Description

The classification tree creates a series of IF-THEN statements based upon certain values of criteria. Three options were selected in C5.0: simple method, no boosting, and accuracy favored over generality. This tree used 13 criteria and was 98.09% accurate on the model dataset. All of the criteria were available to the algorithm, and the criteria selected were:

- Number of references in the text
- Average word length
- Existence of reference to “Table 1” or “Figure 1”
- Number of times a traditional heading appeared on the page (such as Abstract, Findings, Discussion,

etc.)

- Number of times phrases such as “published in,” “reprinted in,” etc. appear
- Academic URL
- Ratio of total size of images on page to total size of page
- Number of misspelled words according to Dr. HTML
- Number of words in the meta keyword and dc.subject meta tags
- Average number of punctuation marks per sentence
- Average sentence length
- Number of sentences in the document
- Commercial URL.

4.3.2 Performance on Test Dataset

The classification tree correctly classified 478 scholarly pages and 480 non-scholarly pages. This gives it a accuracy of 96% and a return of 95.6%. This tool misclassified many non-scholarly pages that were at an educational domain, contained links to educational sites, or that were long textual documents with few graphics. Common features in misclassified scholarly documents were a commercial URL, a lack of traditional headings, and large graphics on the page.

4.3.3 Performance on Problematic Dataset

This tool misclassified 32.5% of the pages in the problematic dataset. It did the worst with research proposals and abstracts, but classified most of the syllabi correctly.

4.4 Neural Networks

4.4.1 Model Description

Neural networks combine nodes holding values for criteria in iterations until there is just one node left. This

neural network started with 41 nodes and was processed through one hidden layer of three nodes, which were then combined for the decision node. The multiple training method was used with the "prevent overtraining" option selected. This model correctly classified the model dataset 97.12% of the time. Although the neural network uses all of the criteria, the ten most important criteria are:

- Number of sentences
- Average word length
- Number of times a traditional heading appeared on the page (such as Abstract, Findings, Discussion, etc.)
- Number of times Dr., PhD, Professor, and similar academic titles are used
- Number of misspelled words according to Dr. HTML
- Number of times "journal," "conference," or "proceedings" appear
- Presence of labeled bibliography
- Existence of reference to "Table 1" or "Figure 1"
- Number of references in the text
- Average paragraph length.

4.4.2 Performance on Test Dataset

The neural network classified 469 non-scholarly pages and 465 scholarly pages correctly. This gives it a accuracy of 93.75% and a return of 93%. It had a problem with non-scholarly pages that were long textual documents with few graphics. Conversely, scholarly pieces that were shorter, contained no labeled bibliography, and did not use traditional headings caused problems for this model.

4.4.3 Performance on Problematic Dataset

The neural network misclassified 31% of the problematic dataset. Just like logistic regression, this tool had problems with non-annotated bibliographies and research proposals. It correctly classified all of the

non-scholarly articles and did well with syllabi, book reviews, and research in a foreign language.

4.5 Model Comparison

The classification tree had the highest accuracy and return, although the accuracy for all tools was quite close (93.75% to 96%). The return was spread out between 87.6% and 95.6%, with discriminant analysis performing the worst. The difference in proportions between the highest and lowest performing model is statistically significant at the 95% level for accuracy and at the 99% level for return, however, the differences between the middle-ranked models and the extreme performers are not statistically significant at a reasonable level.

Table 1. Accuracy and Return of Models

	Accuracy	Return
Logistic Regression	94.5%	92.6%
Discriminant Analysis	94.6%	87.6%
Classification Tree	96%	95.6%
Neural Network	93.75%	93%

Even the worst model here would perform well in powering an Web search tool. The classification tree uses only twelve easily attained criteria and an easily programmable if-then structure to make rapid classification decisions.

All of the models used criteria based on the existence of a labeled bibliography and/or number of references, the reading level of the text (word length, sentence length, etc.), and the structure of the document (use of traditional headings, table references, etc.). This suggests that in order for a future automated classification to be successful, suggested guidelines or even standards for electronic scholarly publishing are needed.

4.5.1 Discussion of Problematic Pages

Analysis of the problematic pages suggest a number of new criteria that could be introduced into the next iteration of these models. In order to create these criteria, a set of pages that fall into the category can be examined for common facets. Out of all of the common facets, those which do not apply to scholarly research work can be introduced as new criteria. For example, in order to discover non-annotated bibliographies, a criterion of the percentage of the entire work dedicated to the bibliography could be introduced. This would aid in reducing misclassifications.

All of the models had trouble distinguishing research proposals from scholarly research works. This suggests that the definition used in this work for scholarly research works may be too limiting, and needs to include research proposals. The table below summarized the number of pages misclassified by each tool in each area.

Table 2. Number of Pages Misclassified (out of 20).

Category	Logit.	Discrim.	Class.	NN
Non-annotated bibliographies	10	15	6	13
Syllabi	2	20	1	2
Vitae	11	19	5	7
Book Reviews	2	20	5	2
Non-scholarly Articles	0	20	2	0
Research Written in Foreign Language	3	18	3	3
Partial Research	4	20	6	5
Corporate Research	7	20	10	8
Research Proposals	16	20	17	15
Abstracts	5	9	10	7
Total Missed	60	183	65	62

5. Conclusions and Discussion

This study created an information agent for collection development, in the guise of an automated filter, for a class of documents. One of the requirements for this type of agent to function is that the document be in an electronic form. When electronic publishing is accepted and all documents are produced in an electronic form, information filtering agents will be a useful and necessary tool in dealing with the rapid production and dissemination of information.

The four-step technique developed in the research can be used to create these filters for other groups of structured documents. First, criteria are selected that may discriminate between the desired type of documents and other documents. Second, the criteria are operationalized and programmed into a computer program. Third, both documents that are desired and that are not desired are gathered. Finally, data mining is used to create a parsimonious model that can discriminate between documents.

One problem with this methodology is that the data set modeled on is not representative of the real Web, as the percentage of pages containing academic research on the Web is much lower than 50% (Nicholson, 2000). Due to the small size of the dataset used (10,000 pages), the 50%/50% split was used in order to have a more robust failure analysis. Due to this unrealistic split, the retrieval will err on the side of return; however, this can be easily compensated in implementation by making it easy for users to report a non-academic page. Others using these techniques for similar explorations will need to use non-representative samples or oversampling techniques in order to create a rich data set for bibliomining.

5.1 Future Research

The next step in this research is to revise the list of criteria to take advantage of common facets in misclassified and problematic pages. By analyzing each area of failure for commonalities, new criteria could be produced for new models. In addition, by applying multiple techniques to create an overall model, the accuracy of these models can be improved. After adding new criteria, the data set should be modified to more accurately represent the real Web world and therefore create more generalizable models.

Another step in this research is to remove some of the restrictions placed upon the definition of scholarly

research works. Future researchers could see if this technique can be applied to documents that are broken up over several Web pages. As many collections of documents require submission to be in LaTex, PDF, or Postscript files, as compared to HTML or plain text, moving this research beyond just analyzing HTML and plain text documents may be the next step most needed to continue this line of research.

This technique can also be applied to different types of research. By adding foreign-language terms for some of the criteria to the Perl program, this technique might be able to be used to not only collect research in other languages, but identify the language used as well.

In conclusion, the application of data mining and agent techniques to the World Wide Web for information retrieval is a new and open research area, but it may prove to be one of the best ways to organize the chaotic and expanding Web.

6. References

Banerjee, K. (1998). Is data mining right for your library? *Computers in Libraries, 18*(10), 28-31.

Basch, R. (1990). Databank software for the 1990s and beyond. *Online, 14* (2),17-24.

Beaver, A. (1998, December). Evaluating search engine models for scholarly purposes. *D-Lib Magazine*. Retrieved November 23, 2002, from http://www.dlib.org/dlib/december98/12beavers.html.

Berry, M. J. and Linoff, G. (1997). *Data Mining Techniques*. New York: Wiley Computer Publishing.

Cassel, R. (1995). Selection criteria for Internet resources. *C&RL News 56*(2), 92-93.

Chau, M. (1999). Web mining technology and academic librarianship: Human-machine connections for the twenty-first century. *First Monday 4*(6) Retrieved November 23, 2002 from http://www.firstmonday.dk/issues/issue4_6/chau.

Cleverdon, C. (1962). *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing System*. Cranfield, U.K.: College of Aeronautics.

Collins, B. (1996). Beyond cruising: Reviewing. *Library Journal, 121*(3), 122-124.

Dickinson, J. (1984). *Science and Scientific Researchers in Modern Society*. (2nd ed.). Paris: Unesco.

Evans, G. E. (2000). *Developing Library and Information Center Collections*. (4th ed.). Englewood, CO: Libraries Unlimited.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37-54.

Futas, E., (Ed.). (1995). *Library Acquisition Policies and Procedures.* (3rd ed.). Phoenix: Oryx Press.

Information Market Observatory (IMO). (1995). *The Quality of Electronic Information Products and Services*. Retrieved November 23, 2002 from http://www.midas.gr/info2000/market/fn954ww.zip.

Hinchliffe, L. J. (1997). *Evaluation of Information*. Retrieved November 23, 2002 from http://alexia.lis.uiuc.edu/~janicke/Eval.html.

Hinton, G. (1992). How neural networks learn from experience. *Scientific American, 267*(3), 145-151.

Hofman, P., and Worsfold, E. (1999). A list for quality selection criteria: A reference tool for Internet subject gateways. *Selection Criteria for Quality Controlled Information Gateways.* Retrieved November 23, 2002 from http://www.ukoln.ac.uk/metadata/desire/quality/report-2.html.

Johnston, M. & Weckert, J. (1990). Selection Advisor: An expert system for collection development. *Information Technology and Libraries, 9*(3), 219-225.

Lawrence, S., and Giles, C. (1999). Accessibility of information on the Web. *Nature, 400,* 107-109.

Lawrence, S., Giles, C., and Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *IEEE Computer, 32*(6), 67-71.

McCallum, A., Nigam, K., Rennie, J., and Seymore, K. (1999). Building domain-specific search engines with machine learning techniques. In *Proceedings of the AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*. Retrieved November 23, 2002 from: http://citeseer.nj.nec.com/cache/papers/cs/1785/http:zSzzSzwww.cs.cmu.eduzSz~mccallumzSzpaperszSzcora-aaaiss98.pdf/mccallum99building.pdf.

McGeachin, R. B. (1998). Selection criteria for Web-based resources in a science and technology library collection. *Issues in Science and Technology Librarianship, 18.* Retrieved November 23, 2002 from: http://www.istl.org/98-spring/article2.html.

Neill, S. D. (1989). The information analyst as a quality filter in the scientific communication process. *Journal of Information Science, 15,* 3-12.

Nentwich, M. (1999). Quality filters in electronic publishing. *The Journal of Electronic Publishing, 5*(1). Retrieved November 23, 2002 from: http://www.press.umich.edu/jep/05-01/nentwich.html.

Nicholson, S, and Stanton, J. (in press). Gaining strategic advantage through Bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries. In H. Nemati & C. Barko (Eds.), *Organizational Data Mining: Leveraging Enterprise Data Resources For Optimal Performance.* Hershey, PA : Idea Group Publishing.

Nicholson, S. 2000. *Creating an Information Agent through Data Mining: Automatic Indexing of Academic*

Research on the World Wide Web. Unpublished doctoral dissertation., University of North Texas, Denton. Retrieved November 23, 2002 from: http://www.scottnicholson.com/ scholastic/finaldiss.doc.

Nicholson, S. (2002). *Bibliomining: Data Mining for Libraries*. Retrieved November 22, 2002, from http://www.bibliomining.org.

Piontek, S. and Garlock, K. (1996). Creating a World Wide Web resource collection. *Internet Research: Electronic Networking Applications and Policy, 6*(4):20-26.

Pratt, G.F., Flannery, P., and Perkins, C. L. D. (1996). Guidelines for Internet resource selection. *C&RL News, 57*(3), 134-135.

Sharma, S. (1996). *Applied Multivariate Techniques*. New York: John Wiley & Sons.

Smith, A. (1997). *Criteria for Evaluation of Internet Information Resources*. Retrieved November 23, 2002 from: http://www.vuw.ac.nz/~agsmith/evaln.

Trybula, W. J. (1997). Data mining and knowledge discovery. In M. E. Williams (Ed.) *Annual Review of Information Science and Technology, 32*, 196-229. Medford, NJ: Information Today.

Yulan, H. and Cheung, H. (2000). Mining citation database for the retrieval of scientific publications over the WWW. *Proceedings of Conference on Intelligent Information Processing*, 64-72. Publishing House of Electron. Ind; Bejing, China.