BUILDING A DIGITAL LIBRARY FROM THE GROUND UP: AN EXAMINATION OF EMERGENT INFORMATION RESOURCES IN THE MACHINE LEARNING COMMUNITY

Sally Jo Cunningham
Department of Computer Science, University of Waikato, Private Bag 3105, Hamilton, New Zealand
sallyjo@cs.waikato.ac.nz

The current crop of digital libraries for the computing community are strongly grounded in the conventional library paradigm: they provide indexes to support searching of collections of research papers. As such, these digital libraries are relatively impoverished; the present computing digital libraries omit many of the documents and resources that are currently available to computing researchers, and offer few browsing structures.

These computing digital libraries were built 'top down': the resources and collection contents are forced to fit an existing digital library architecture. A 'bottom up' approach to digital library development would begin with an investigation of a community's information needs and available documents, and then design a library to organize those documents in such a way as to fulfill the community's needs. The 'home grown', informal information resources developed by and for the machine learning community are examined as a case study, to determine the types of information and document organizations 'native' to this group of researchers. The insights gained in this type of case study can be used to inform construction of a digital library tailored to this community.

1 Introduction

The field of computer science has been well-served by the digital libraries movement since the inception of the digital libraries field. This situation was based on the fact that in the early days of the World Wide Web, only computer science provided a reasonably sized set of digitized documents that were likely to see a significant amount of use by a reasonably sized user base. In the 1980's a convention had grown up that computing research institutions would place their working papers and technical reports in an ftp archive, and that these documents would be held in PostScript format (and likely
other formats as well). It was a natural decision, then, for the computer scientists who constructed the original digital library software to choose computing technical reports as their testbed collection (recall, for example, UCSTRI [11], WATERS [9], DIENST [5], and NZDL (Witten), among others).

These digital libraries have been immensely popular with computing researchers and tertiary computing students — their target user group. The digital libraries provided an invaluable service by enabling the original ftp sites to be

searched as a whole; previously, a document of interest could be located only by laboriously visiting each site in turn, and then examining an index file that was present by convention in most (but not all) sites. The computing digital libraries also offered a powerful new searching capability: keyword searching over the full text of the document. This facility came at a cost, however; sizeable collections could be developed only by 'harvesting' documents from a large number of repositories and WWW sites, but this diversity of sources meant that no bibliographic data could be assumed to exist. The ResearchIndex system (formerly known as CiteSeer) succeeds in offering both full text and bibliographic search in a large (at the moment, 1,000,000 documents) collection by parsing a document to extract the document's bibliographic details and its list of references to other documents [7]. Further, references are automatically linked to the documents they cite, creating a full-fledged citation index. Somewhat disappointingly, however, these computing digital libraries remain strongly rooted in the traditions of conventional libraries and conventional library OPACs. Despite early promises by the digital library community that digital library architectures would support the creation of heterogeneous, multimedia collections, computing digital libraries are restricted to homogeneous, monomedia collections of technical reports, theses, conference papers, and journal articles — the traditional contents for a conventional scientific library. Interactions with the digital library are expected to be based on search, not browsing; the main 'entrance' to the digital library offers a keyword or fielded search as the primary (and in some cases, only) means for exploring the collection's contents.

The difficulty with providing a novel digital library interface or novel contents lies in ensuring that the resulting digital library will be useful to, and usable by, its intended user community. The problem is to discover what documents (construed in the broadest sense) a given community finds useful, what natural organizations of those documents exist, and what vocabulary is used in the community to describe their work. An ethnographic approach seems appropriate for discovering how to tailor the generic digital library architectures to a particular community — that is, to examine the above issues from the target community's point of view, in the community's own words.

This paper takes such an approach by examining the WWW-based information resources created by members of a research community, for use by that community. The documents and the organization of information in these resources provide insights into how the current computing digital library architectures can be tailored to meet the information preferences of that community. The community of machine learning researchers is chosen as the focus for this case study. An earlier ethnographic study of computing researchers utilized interviews to explore the participants' information behavior. This work provided interesting insight into this group's preferred information gathering practices, but it proved difficult to turn the focus away from the limitations or facilities of existing resources and onto the potential for innovative new resources [4]. Quan-

titative studies of the transaction logs of computer science digital libraries again can indicate how these researchers use existing digital libraries and their interfaces ([6], [8]), but do not tell us what documents, browsing structures, and searching styles are preferred or more 'native' to these library users. Examination of the 'home-grown' information resources developed within a community will complement these earlier studies by grounding suggestions for further developments of a computing digital library in the authentic behavior of this community.

This paper is organized as follows: the following section lists the community-developed resources that are analyzed in this paper, and briefly describes the criteria for their selection. Section 3 discusses the types of documents stored in or linked to by these resources, and explores the role that each document type can play in supporting the research community. The ways that the community-developed resources group documents to support browsing are discussed in Section 4. Section 5 discusses the implications of these observations to the task of tailoring the generic digital library to this community.

2 Information Resources Analyzed in this Case Study

The following information resources were selected for analysis in this paper:

- Online machine Learning Resources: http:/www.ai.univie.ac.at/oefai/ml/ml-resources.html
- KDnuggets: http://www.kdnuggets.com
- David Aha's Machine Learning Page: http://www.aic.nrl.navy.mil/~aha/research/machine-learning.html
- Mlnet: http://www.mlnet.org
- The Data Mine: http://www.cs.bham.ac.uk/~anp/TheDataMine.html

The primary criterion for selection is that the resource must have been developed and maintained by individuals or groups within the machine learning community; that is, the resource is not part of an 'official' library or corporate information source (KDnuggets, although a .com site, is maintained by a prominent researcher in this field). Further, the resource had to contain sufficient material (links and documents) to be of interest. It was not required that the resource be kept up to date; the interest here is in the documents/information and their organization, and not in the ability of the original resource developer to maintain the currency of that resource.

3 Document Types, and the Information Needs They Support

The information resources listed in Section 2 contain a number of different types of document. This section considers these types, the information that these documents can convey both about the topic of machine learning and the

field of machine learning, and the ways that these documents can support research in machine learning. These documents can be viewed assupporting answers to the questions:

3.1 Who is working in this field?

It can be difficult to for a researcher entering a field to gain an awareness of other researchers in the field—who they are, where they are, and what problems they are working on. In the absence of a strong sense of the extent of a community and the connections between its members, it can be difficult to understand how one fits into this community, and to cultivate the connections that can lead to an 'invisible college' of supporting colleagues.

Collections of homepages of machine learning researchers provide an excellent introduction to the members of that community. The researchers' homepages usually contain a great deal of helpful information for other researchers in the same field:

- contact information, allowing the searcher to easily initiate communication with the target researcher. The contact details nearly always include an email address, with telephone and postal address details also commonly present.
- a list of the researchers' research interests, giving an overview of their personal research programme and an indication of problems they are currently investigating.
- a photo of the researcher, and often a link to more personal information (description of hobbies, pictures from a recent holiday, etc.).
- a list of recent publications, and often a full cv. Frequently the publications list includes links to online versions of the papers.

One conventional information resource that is useful in exploring connections within a community is the citation index, which records the formal reference/citation linkages between specific documents (and by extension, between the sets of documents that form the body of individuals' work). Bibliographic resources allow users to explore connections between authors, in the form of co-authorship linkages. Note that homepage information provides a much richer context for understanding a researcher's place in the community than is available through these conventional resources. Homepage information allows the browser to situate individual documents in the researcher's programme of work, and to evaluate the extent to which that programme has currently been fulfilled. Examination of a researcher's official persona, as presented through the presence — or absence — of photos and personal information, give clues as to the possibility of finding personal, as well as professional, common ground with that researcher.

3.2 Which people are working together in this field?

Formally established research groups are represented with a project or group homepage that serves much the same purpose as the individual researcher's homepage: the group homepage projects the group's persona to the world, describes the group's research programme, and present the results achieved to date. These results can include documents such as formal publications, working papers, manuals, and reports to various governing or funding bodies. Results can also include online demonstrations and fully functioning software.

Conventional information resources such as bibliographic and citation databases provide a much more limited picture of research groups; they are only able to show which researchers have co-authored or have referenced each other in the past. Group homepages give the 'big picture' of how a number of individual research agendas fit together, and what potential exists for future collaboration between individuals within the group.

3.3 What events are occurring, …

The events that are important to a field can include conferences, workshops, courses, funding opportunities, and seminars. In machine learning, data mining competitions are also significant events — occasions when individuals and groups compete to see who can provide the 'best' model of a dataset. Events are generally listed in reverse chronological order, with future
events preceding past events. Sometimes the listing of an event is limited to a simple text notice of its dates, topic, and associated deadlines; more often, the event's description includes a link to the event's webpage or website.
Generally notices of seminars, courses, and funding opportunities are deleted after their deadlines pass. These notices are useful only when they are timely — when planning attendance at a seminar, or when attempting to gain research funding. After the fact, the information in the notices are mildly interesting as historical artifacts for the field, but for the most part are not deliberately searched for or consulted.

3.4 …and what events have occurred in the past?

By contrast past conference notices, when accompanied by a link to a conference website, are often stored long after the conference itself has ended. The conference website contains a number of items that remain highly relevant for researchers: the names and institutional affiliation of invited speakers and other presenters, which can be useful in locating active researchers in the field; groupings of accepted papers into sessions, useful in determining the research topics that were represented in that conference — and the number of papers associated with each topic indicate how relatively 'hot' that topic was; and finally, most websites contain either an abstract for each paper, or an online version of the paper.

Conference websites, then, can provide both useful documents and a sense of context for those documents. This sense of context is significant; while the papers themselves may be available elsewhere, the grouping of the papers into the conference provides a valuable snapshot of the state of that field at that particular time.

Links to data mining competition websites are also frequently maintained long after the competition has ended. The competition websites describe the winners and their successful techniques for constructing a data model; past competition websites thus provide a record of comparative analysis at a degree and level of detail that is difficult to find in more formal publications.

3.5 What are people in the field talking about?

Conferences and journals are the media through which formal scientific 'conversations' occur: the arguments, shifts in topic, and achievement of consensus occur through the byplay of reference and citation, replication and refutation.

In many fields informal scientific conversations are conducted through mailing lists; in the case of machine learning, the mailing lists of note are KDnuggets News and ML-List. Archives exist for both lists, and the KDnuggets archives are searchable. Other mailing lists exist for sub-fields (for example, Machine Learning for User Modeling) or related fields (for example, AI & Statistics). This present case study will focus on KDnuggets and ML-List, as they are the oldest lists and have the broadest focus.

Both mailing lists are moderated, and are mailed out at regular intervals to subscribers. In format, then, the lists are closer to electronic newsletters than to freewheeling discussion groups. This more formal, limited type of conversation appears to be preferred by the machine learning community, presumably because the signal to noise ratio is greater on a more tightly controlled discussion forum.

What types of postings are made to the lists? Event notifications such as conference calls for papers or participation, job openings, and funding application deadlines are staple topics. Members post announcements of their publications that they feel are particularly significant, or of interest to, the community. Technical questions are posed, and answered; opportunities for collaboration are described; refereeing practices are questioned; and so forth. The mailing lists, though formally structured, still offer a forum for community building and for development of an individual research agenda.

3.6 Where can scientific 'results' be found?

The conventional scientific result is, of course, a formally refereed and published scientific article — and research papers (or rather, links to research papers) are included in the machine learning resource websites. Additionally,

many of the other sites linked to (homepages, research group pages, and conference websites) also contain machine learning research articles.

In the machine learning world, another possible type of 'result', or end product of research, is software implementing the ideas embodied in a paper — for example, an implementation of a new machine learning technique. The software is useful to machine learning practitioners, who build on existing implementations or run comparative trials of algorithms; practitioners, who use the software in evaluating real world databases; and novices to the field, who examine the software to learn about the algorithms.

3.7 Where can 'raw material' for research be located?

When a new machine learning algorithm is developed, the algorithm's performance must be compared to that of other algorithms. Generally, this is accomplished by comparative trials over several datasets. A single algorithm may also be used to process a variety of datasets of known characteristics, to evaluate that algorithm's effectiveness and efficiency in the face of those characteristics.

Several collections of test datasets exist, with the repository maintained at The University of California, Irvine being by far the most comprehensive [1]. The repository includes an informal cataloging of the datasets that, while not optimal [3], is useful in locating suitable datasets for an algorithm trial.

3.8 How do sub-fields, and the field as a whole, define themselves?

A frequently encountered type of information is a tutorial or extended discussion of a specific topic in machine learning — for example, link analysis, machine learning applications in games, or minimum message length encoding. This type of document defines the sub-field, and frequently presents links to associated documents describing that community or sub-field (homepages, software, research articles, etc.).

These documents are of particular interest when a sub-field is in its infancy; they delimit the boundaries of the emerging sub-field, pointing out its distinguishing concerns or methods. This type of document can form a manifesto for a fledgling discipline, providing an identity and a focus for an emerging group of researchers.

But machine learning as a whole is also an emerging discipline, and its boundaries are also in flux. What is the relationship between machine learning, data mining, and knowledge discovery in databases — or are they the same field? Is computational learning theory (COLT) a sub-field of machine learning, or is COLT a separate, theoretical discipline distinct from an application-focused machine learning field? Each of the resources analyzed in this case study implicitly defines its own view of the boundaries of machine learning through the topics of the documents and links that it includes in the website, and through groupings of links into the field proper and into 'related field' categories.

4 Organization of Documents for Browsing

Browsing, rather than searching, is intended as the primary interaction technique for the websites analyzed in this paper. Only one of the machine learning resource websites supported searching over the entire resource (KDnuggets), and even at that site a browsing hierarchy is more prominently displayed than is the search facility. Remembering that these websites are, by definition, created and maintained by members of the machine learning community, it seems likely that this preference for browsing is at least partly based on the pragmatics of web resource construction and maintenance — the site developers likely did not have a search engine available to them, or did not wish to invest the time to learn to create and manage an index. The techniques used to organize documents for browsing are still of interest, however, in that they illustrate the 'natural' groupings that the resource developers see in the collection.

The machine learning webpage resources offer two fundamental types of document organization: clustering of documents by type (that is, grouping homepages together, research articles together, etc.), and clustering by topic (for example, grouping all documents describing Computational Learning Theory together). The previous section discussed the different types of documents, and the information needs that could be supported by those document types. Here, the focus will be on support provided for browsing by topic.

The most striking observation about the topic-based groupings found in the resources developed by machine learning researchers is that this community recognizes and uses a far richer and finer-grained categorization of this field than do the formal computing resources. For example, the 1998 version of the ACM Computing Classification System (http://www.acm.org/class/1998/) does indeed contain a Learning classification (I.2.6), but this category contains a relatively sparse set of subject descriptors (Analogies, Concept learning, Connectionism and neural nets, Induction, Knowledge acquisition, Language Acquisition, and Parameter learning). Far more, and far more finely grained, topics appear in the machine learning websites: for example, David Aha's Machine Learning Page contains links to topics such as bias shift, learning classifier systems, context sensitive learning, and decision trees. Moreover, the ACM's Learning classification is not an exact fit to the evolving description of the field that is emerging from its research community: I.2.6 includes subject descriptors that much of the machine learning community would recognize as being a part of the field (induction, for example), omits other subjects (for example, time series analysis), and includes subjects that most of the machine learning community would see as outside of the field proper (for example, Connectionism and neural nets).

Another significant feature of the topic groupings present in the machine learning resources is that they are not hierarchically organized. In keeping with

the fluid perception of the boundaries of the field, these websites make little or no attempt to define the relationships between topics. Further, the topic-based browsing structures are wholly pragmatic in nature — topic listings are restricted to those topics for which the resource actually stores documents or links to documents, and there is no attempt to construct a comprehensive ontology for the discipline as a whole.

5 Conclusions

This paper examines a set of WWW-based information resources created by members of the machine learning community, to support the needs of that community. These information resources cover a far broader range of document types than do the present digital libraries designed for computer science researchers. These latter digital libraries limit their collections to 'official' research documents: technical reports, conference papers, journal articles, and theses. The additional types of documents and information provided through the machine learning resources can provide a richer support for some of the tasks that a conventional collection also supports — for example, exploring the relationships between researchers within that field. The 'results' of research are construed more broadly in the machine learning resources than they are in the computer science digital libraries; the resources, for example, include links to algorithm implementations. The resources also provide a broader support to the different phases of a machine learning research project — for example, by providing links to test datasets.

The practitioner-based collections can be browsed by document type, or by subject/topic grouping. These subject titles and the relationships between them are not static, and are not fitted into an over-arching definition of the field. It has long been recognized that a mismatch frequently exists between the 'official' terminology used by classification systems or thesauri to describe a field, and the language used by the practitioners in the field (see, for example, [2]). One approach to providing a subject-related browsing structure that is tailored to a discipline's own description of itself is to use the terminology extant in a collection's document as the basis for browsing. Paynter, et al. [10], for example, present a set of techniques for extracting the 'significant' phrases from a document collection; these phrases then form the basis for a digital library interface that supports searching and browsing by phrase.

References

1. Blake, C.L. & Merz, C.J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.
2. Carlyle, A. Matching LCSH and user vocabulary in the library catalog. Cataloging & Classification Quarterly 10:1/2 (1989), 38-42.
3. Cunningham, S.J. (1997) Dataset cataloging metadata for machine learning applications and rsearch. Proceedings of the Sixth International Workshop on AI and Statistics '97 (Fort Lauderdale, FL, Jan.).
4. Cunningham, S.J., and Connaway, L.S. Information searching preferences and practices of computer science researchers. Proceedings of OZCHI '96 (Hamilton, New Zealand) (1996) 294-299.
5. Davis, J., and Lagoze, C. 'Drop-in' publishing with the World Wide Web. Proceedings of the Second International WWW Conference (Chicago, 1994). http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/Pub/davis/davis-lagoze.html
6. Jones, S., Cunningham, S.J., McNab, R.J., and Boddie, S. A transaction log analysis of a digital library. International Journal on Digital Libraries 3(2) (2000) 152-169.
7. Lawrence, S., Giles, L.C., and Bollacker, K. Digital libraries and autonomous citation indexing. IEEE Computer, 32: 6 (1999) 67-71.
8. Mahoui, M., and Cunningham, S.J. A Comparative Transaction Log Analysis of Two Computing Collections. Research and Advanced Technology for Digital Libraries: Proceedings of the 4th European Conference, ECDL (Lisbon, Portugal, Sept.) (2000) 418-423.
9. Maly, K., Fox, E.A>, French, J.C., and Selman, A.L. Wide Area Technical Report Server. Technical Report, Dept. of Computer Science, Old Dominion University. http://www.cs.odu.edu/WATERS/WATERS-paper.ps.
10. Paynter, G.W., Witten, I.H., Cunningham, S.J., and Buchanan, G. Scalable browsing for large collections: a case study. Proceedings of Digital Libraries 2000 (San Antonio, Texas, June) (2000).
11. Van Heyningen, M. The Univied Computer Science Technical Report Inex: lesions in indexing diverse resources.
12. Proceedings of the Second International WWW Conference (Chicago, 1994).
13. http://www.cs.indiana.edu/ucstri/paper/paper.html#ref-odlyzko