A Proposal for an ACM Computing Portal

SIG Governing Board Portal Committee[1]

September 21, 1999

Abstract

We propose a new ACM service, the *ACM Computing Portal*. This is a web-based repository of bibliographic information of all the computing literature. It will provide pointers from each bibliographic entry to the digitized version of the relevant book, paper, or article, if it resides on the World Wide Web. The ACM Computing Portal is an update and enhancement of the ACM *Guide to Computing Literature*, as well as an entry into the ACM Digital Library. It has as goals qualitatively increasing the effectiveness of scientific research into computing, continuing to place ACM as the premier scientific and educational organization for computing, increasing service of ACM and the SIGs to the computing community, and providing a concrete illustration of the scope of computer science by showing the range of the published literature. The recent availability of underlying technologies, specifically inexpensive scanning and OCR, inexpensive disk space, pervasive access to the world wide web, and inexpensive, high-capacity CD-ROMs, make it possible to consider capturing the entire computer science corpus and making it available for search and retrieval. The catalysts for this endeavor are the ACM Council, the SIG Governing Board, and the ACM Publications Board.

[1] The ACM SIG Governing Board Portal Committee consists of Richard Snodgrass (chair), Department of Computer Science, University of Arizona, Tucson, AZ, rts@cs.arizona.edu, Steven Cunningham, Department of Computer Science, California State University, Stanislaus, Turlock, CA, rsc@castor.csustan.edu, Carol Hutchins, Courant Institute of Mathematical Sciences Library, New York, NY, carol.hutchins@nyu.edu, Robert Krovetz, Computer Science Division, NEC Research Institute, Princeton, NJ, krovetz@research.nj.nec.com, Michael Ley, Informatik, Universität Trier, Germany, ley@uni-trier.de, Andreas Paepcke, Stanford University, Stanford, CA, paepcke@cs.stanford.edu, Kathy Preas, KP Publications on CDROM, Palo Alto, CA, kathy@kppubs.com, Bernard Rous, Deputy Director of Publications and Electronic Publishing Program Director, ACM, New York, NY, rous@acm.org, and Charles Viles, School of Information and Library Science, University of North Carolina, Chapel Hill, NC, viles@cs.unc.edu.

Table of Contents

1 The Vision

The eventual goal, which is admittedly unattainable in the near-, or even medium-, term, is to construct a electronic repository of *all* computing literature, books, journal articles, conference papers, and magazine articles, in digital form. The initial portion, which is possible in the medium term, is a database of bibliographic information, informally termed the "ACM Computing Portal," that will provide a pointer from each bibliographic entry to the digitized version of the relevant book, paper, or article, if it resides on the World Wide Web. Configurations of the database will include the collected works of each author, the collected works of each source (e.g., journal, conference proceedings), the papers cited by a particular paper, and the papers that cite a particular paper.

Four primary objectives motivate this proposal. The first is to qualitatively increase effectiveness of scientific research into computing and increase scholarship among computing scientists. We want to continue to place ACM as the premier scientific and educational organization for computing. Thirdly, the portal will increase service of ACM and the SIGs to the computing community. Finally, the portal will provide an illustration of the scope of computer science by showing the range of the published literature.

The desire is for the Computing Portal to become **the** point of entry to the computing literature stored on the World Wide Web. By providing a comprehensive database, ACM and its SIGs will come to be viewed as the central provider of access to this literature. This initial search should remain free to all users, the majority of whom are not (yet!) ACM members. A fee-based search engine will be used much less than one that is free. The second motivator is to encourage publishers to make their copyrighted material available on the web (whether free or for a fee). As the amount of material easily accessible on the web via a central source increases, material not on the web will be by comparison much more difficult to obtain, and thus less likely to be cited. This will result in increased pressure from the scientific community on those publishers to make their material easily accessible. The third goal derives from ACM's role as a scientific society: to qualitatively improve the effectiveness of scholarship in computing. This also argues for free access by all scholars to the Computing Portal. Finally, people and publishers will be more likely to contribute material to a freely available resource than to one that appears to monetarily benefit the sponsoring organization.

Note: Several have raised the valid point we need to think more about whether the Computing Portal should be free, and also develop a viable business model for its continuing support. Only after such a model exists can the decision about cost of access be made. It is important to ensure a quality product and adequate support for continuing to maintain it.

We envision a four-step process in populating the ACM Computing Portal, hereafter referred to simply as "the portal."

1.1 Step 1 Bibliographic Entries

In this first stage the initial bibliographic database will be collected, of all computing journals, conferences, workshops, technical bulletins, and relevant books and magazine articles. This would cover the entire history of computing, from roughly 1940 to 2000. Extrapolating backwards from the current rate of creation of computing papers, we arrive at a ballpark figure of one million items over this period.[2]

[2] Here are some numbers. Concerning journal papers, there are about 95K pages of ACM journals (transactions, *JACM*, *CACM*) from 1947 to 1990. This translates to very roughly 7K papers. There are 522 individual journals in the Directory of Computer Science Journals. If each journal publishes 40 articles per year, this analysis yields an estimate of 20K journal articles a year. Concerning conference papers, approximately 75K pages of conference articles were published in proceedings from ACM during the period from 1985 to 1990. This translates very roughly to 1K ACM conference papers per year. Concerning

These one million entries would be available for searching without cost to anyone with a World Wide Web connection. Additionally, ACM may issue print or CD-ROM versions, for use during travel or when World Wide Web connections are slow, unavailable, or too costly. (An example is the *SIGMOD Anthology* CD-ROM series, the first volume of which contains some 110,000 bibliographic entries on its initial CD-ROM.) Each located entry would be available in multiple formats, including some combination of html, BiBTeX, refer, Microsoft Word, and endnote.

We hope to have fairly complete coverage of the last two decades of material by December 2000.

1.2 Step 2 Abstracts and Keywords

Either during or after step 1, abstracts and keywords for the entries in the portal will be gathered. Abstracts require care. As they are part of the paper, they are often covered by the article's copyright. Some, but not all, publishers allow abstracts to appear in collections. Some abstracting services, such as INSPEC, employ legions of domain experts to write suitable abstracts. Clearly, ACM cannot take this route, and so we attempt only to provide abstracts that are freely available to ACM.

Keywords also present challenges. Keywords are part of the paper, and so their use may also be restricted by copyright. Even the most unfettered example, that of papers published by ACM, with copyright held by ACM, are problematic, in that the ACM classification scheme has evolved over the years, with the very semantics of some keywords changing over that time. The portal should take a middle ground approach, balancing comprehensiveness and coverage with cost.

1.3 Step 3 Full Text and Bit-mapped Images

When possible, ACM should collect the full text of each paper for use in searching and citation linking (see the next step), and for analyses such as developing lexicons and classification maps possibly for presentation to users. For those papers whose copyright is held by ACM, full text is readily available. For the vast majority of the corpus, for which copyright is held elsewhere, negotiations with the copyright holder may encourage digitization and availability of full text.

It is important to differentiate the Computing Portal, which is comprised of the bibliographic entries, auxiliary information such as URLs and citations, and indexes into those entries created from the full text of the papers, from the papers themselves and from digital versions of those papers that may reside on the World Wide Web in public or proprietary digital libraries. The portal is exactly that, an access point to discover the existence of relevant papers and to locate where on the net those papers reside. Thus, while populating the ACM Digital Library is a useful goal, it is in most regards an activity separate from the portal, except that a populated ACM DL provides highly useful full text from which bibliographic entries, citations, and indexes can be constructed.

---

bibliographic entries, the *ACM Guide to Computing Litera*ture starting with 2K entries in 1976 (just those items reviewed in *ACM Computing Reviews*) but has expanded in an effort to provide a comprehensive bibliography of the research literature. It now adds about 25K entries per year, classified with the ACM Computing Classification System (CCS: www.acm.org/class/) categories and subject descriptors. The coverage is primarily journal and proceedings literature, but the Guide also includes books (16K), dissertations (13K) and technical reports (1K). Extrapolating from a rate of 25K per year over 60 years yields a ballpark figure of 1.5M articles, but of course the rate decreases substantially as one goes back in time. There are 930K entries in the Collection of Computer Science Bibliographies, with many duplicate entries. The DBLP bibliographic database contains 130K unique entries, with much overlap with ACM's bibliographic database.

Ideally, along with the full text, the bit-mapped image of the pages of each paper can be also collected, to retain formatting, equations, tables, and figures. An exact copy of the paper can then be generated. Such images can also provide structure on the full text, such as identifying the sections of the paper (which will be more apparent in the bit-mapped image as a left-justified boldface section title in a larger font). Providing structure can be done manually or semi-automatically, at significant cost per page.

The bit-mapped images require much more disk space (approximately 10–200KB per page) than the bibliographic entry, or even the full text. However, technology continues to reduce this space overhead. Files generated automatically from the formatting source are much smaller than files generated from scanned images. From the bit-mapped images, optical character recognition (OCR) can yield the full-text automatically. Manual retyping or future OCR advances can improve the accuracy of the full text, without rescanning the original page version, if lossless image compression is employed. Note however that OCR still degrades with inferior source. This tends to be the case with older material, particularly publications like conference proceedings where fonts and style were supplied by the author and varied greatly within and between articles.

While publishers may be willing to make the full text of the articles in their journals and conferences available to ACM for use in indexing, they will be justifiably much more protective of formatted (e.g., PDF, postscript, SGML) or bit-mapped images of these papers. And again, we emphasize that the formatted images are distinct from the portal, though certainly should be referenced by the portal.

We hope to have fully populated the ACM Digital Library, with some materials going back to the 1940's, by December 2000.

1.4 Step 4 Citation Linking

From the full text of a paper, or via selective retyping, the bibliography of the paper may be acquired. From this bibliography, an out-linking analysis can identify (or create) the bibliographic entry of each paper referenced by this paper. Following this analysis, the database can be inverted to identify the bibliographic entries of papers referencing this paper. Both lists of references are highly useful to researchers. Additionally, the citation graph can be used for various analyses, such as knowledge diffusion studies. As a simplistic example, an initial analysis of some 30,000 citations in the *SIGMOD Anthology* identified the most-referenced database papers (http://www.acm.org/sigmod/dblp/db/about/top.html). Not surprisingly, the top-scoring one was Codd's *CACM* article introducing the relational model. In fact, the first ten journal/conference papers in this list are found in ACM publications.

The citations themselves represent a difficult database problem. The 1M papers imply 10M-30M citations. The citations must be captured in standard meta-data formats that enable the look-up, matching and linking process, as well as citation searching mechanisms.

We hope to have the citations from the last ten years (1988-1998) of the ACM Digital Library linked by December 2000. The remaining material in the ACM Digital Library (as well as some other articles via their full text as provided by other publishers specifically for indexing) should be in place by December 2001.

2 Demonstration

As an illustration of how the portal may appear to the user, we show a sample session with the *ACM SIGMOD Anthology*, a project in progress by ACM SIGMOD in cooperation with the DBLP project at the Universität Trier (www.informatik.uni-trier.de/~ley/db/). The *Anthology* is a collection of CD-ROMs, five of which were completed in May 1999, with another set of about five disks to be distributed in February 2000. The *Anthology* contains (a) some 130,000 bibliographic entries of database and logic programming papers, (b) the full set of papers from past proceedings of a dozen-odd of the most prominent database

conferences, (c) some workshop papers, (d) past papers from some prominent database journals, and (e) electronic versions of four classic textbooks, totaling 70,000 pages of some 6,000 digitized papers. The *Anthology* is sent free to all SIGMOD members, and is serving as a wonderful tool for recruiting new members. (The business model for the portal will undoubtedly be different.)

The papers are in PDF (Adobe's portable document format), a file format containing both the bit-mapped image in a lossless compressed TIFF format and the full text obtained via OCR. The full text is indexed, allowing users to search over the entire collection. In Figure 1, we request the papers for which the term "wavelet" appears somewhere in the paper. (The Adobe indexing technology requires the PDF to be on the same CD-ROM disk as the index, thus each CD-ROM has to be searched independently. SIGMOD plans to issue a DVD-ROM version of the *Anthology* in part to alleviate this limitation.)

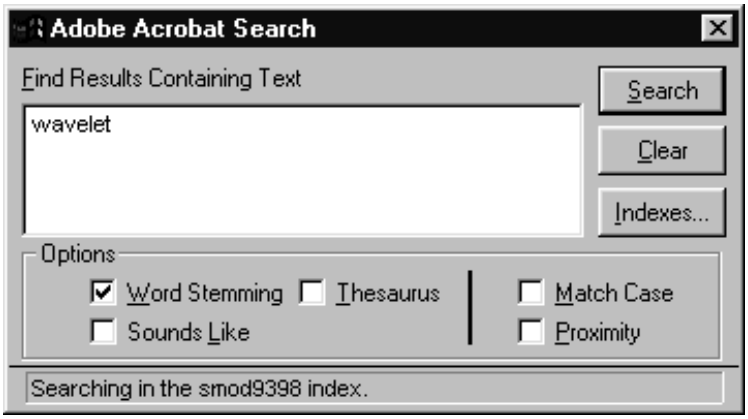


Figure 1 Searching for "wavelet"

This request results in four papers, as shown in Figure 2. The search took only a few seconds, even though it was a full text search over 430 documents (over 500MB of PDF files). The search is over all text in the paper, including title, abstract, bibliography, and even captions and words within figures and tables.

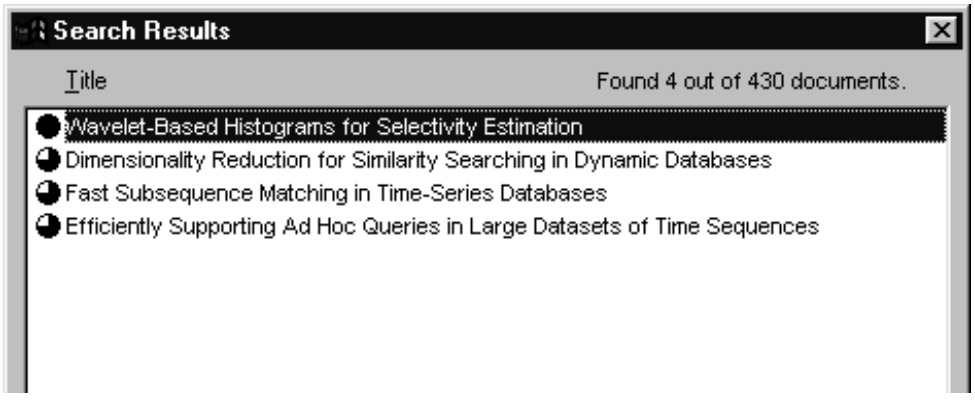
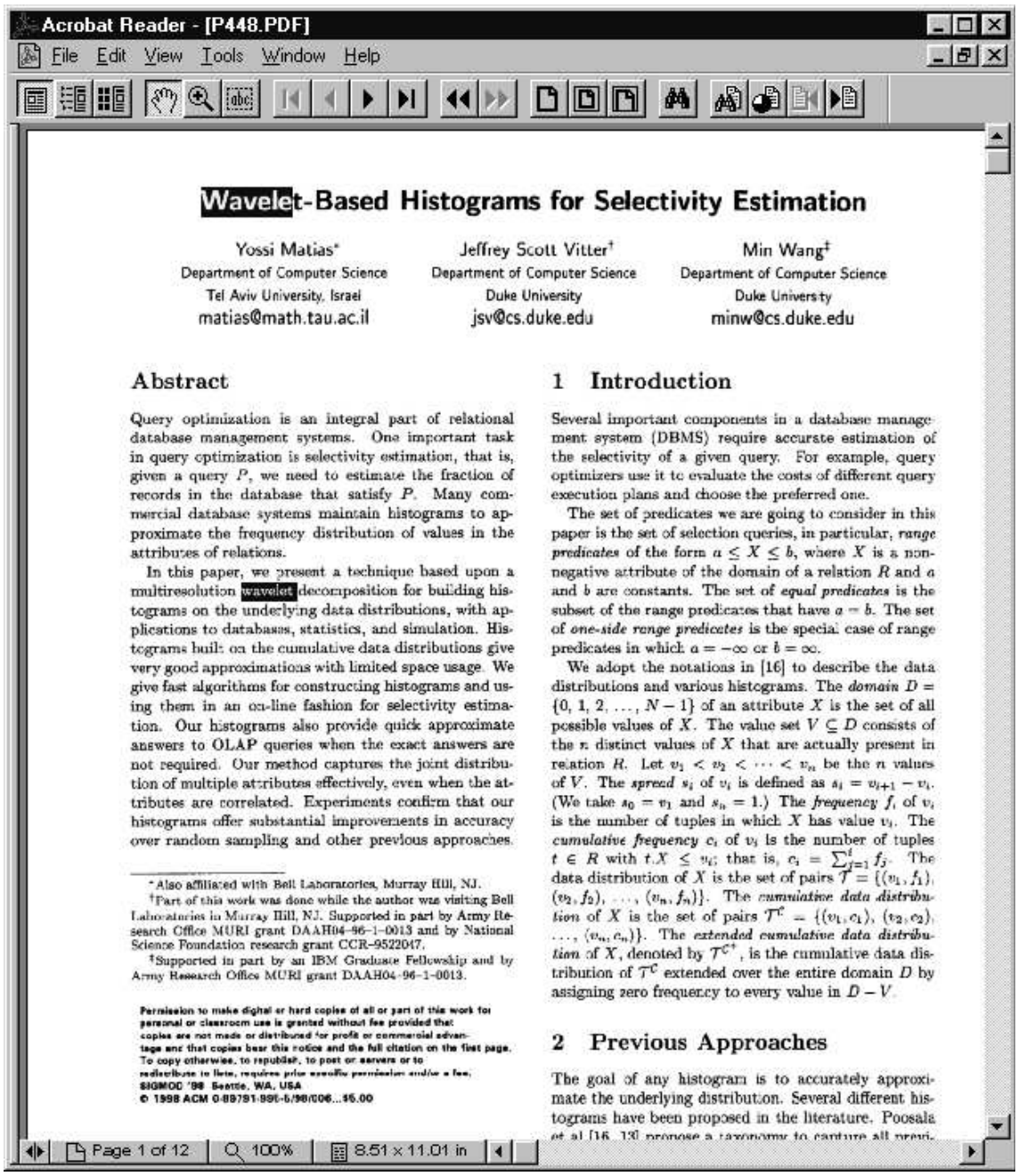


Figure 2 Papers containing the word "wavelet"

Clicking on the first paper listed brings up a bit-map image of that paper, shown in Figure 3. This image is just as the paper appears in the proceedings, with instances of "wavelet" are highlighted. (Due to this image being reduced to fit in this paper, it is rather blurry here. On screen, or printed directly, it is much crisper.)



Wavelet-Based Histograms for Selectivity Estimation

Yossi Matias*
Department of Computer Science
Tel Aviv University, Israel
matias@math.tau.ac.il

Jeffrey Scott Vitter†
Department of Computer Science
Duke University
jsv@cs.duke.edu

Min Wang‡
Department of Computer Science
Duke University
minw@cs.duke.edu

Abstract

Query optimization is an integral part of relational database management systems. One important task in query optimization is selectivity estimation, that is, given a query P, we need to estimate the fraction of records in the database that satisfy P. Many commercial database systems maintain histograms to approximate the frequency distribution of values in the attributes of relations.

In this paper, we present a technique based upon a multiresolution wavelet decomposition for building histograms on the underlying data distributions, with applications to databases, statistics, and simulation. Histograms built on the cumulative data distributions give very good approximations with limited space usage. We give fast algorithms for constructing histograms and using them in an on-line fashion for selectivity estimation. Our histograms also provide quick approximate answers to OLAP queries when the exact answers are not required. Our method captures the joint distribution of multiple attributes effectively, even when the attributes are correlated. Experiments confirm that our histograms offer substantial improvements in accuracy over random sampling and other previous approaches.

*Also affiliated with Bell Laboratories, Murray Hill, NJ.

†Part of this work was done while the author was visiting Bell Laboratories in Murray Hill, NJ. Supported in part by Army Research Office MURI grant DAAH04-96-1-0013 and by National Science Foundation research grant CCR-9522047.

‡Supported in part by an IBM Graduate Fellowship and by Army Research Office MURI grant DAAH04-96-1-0013.

1 Introduction

Several important components in a database management system (DBMS) require accurate estimation of the selectivity of a given query. For example, query optimizers use it to evaluate the costs of different query execution plans and choose the preferred one.

The set of predicates we are going to consider in this paper is the set of selection queries, in particular, *range predicates* of the form $a \leq X \leq b$, where X is a non-negative attribute of the domain of a relation R and a and b are constants. The set of *equal predicates* is the subset of the range predicates that have $a = b$. The set of *one-side range predicates* is the special case of range predicates in which $a = -\infty$ or $b = \infty$.

We adopt the notations in [16] to describe the data distributions and various histograms. The *domain* $D = \{0, 1, 2, \ldots, N-1\}$ of an attribute X is the set of all possible values of X. The value set $V \subseteq D$ consists of the n distinct values of X that are actually present in relation R. Let $v_1 < v_2 < \cdots < v_n$ be the n values of V. The *spread* s_i of v_i is defined as $s_i = v_{i+1} - v_i$. (We take $s_0 = v_1$ and $s_n = 1$.) The *frequency* f_i of v_i is the number of tuples in which X has value v_i. The *cumulative frequency* c_i of v_i is the number of tuples $t \in R$ with $t.X \leq v_i$; that is, $c_i = \sum_{j=1}^{i} f_j$. The data distribution of X is the set of pairs $\mathcal{T} = \{(v_1, f_1), (v_2, f_2), \ldots, (v_n, f_n)\}$. The *cumulative data distribution* of X is the set of pairs $\mathcal{T}^{\mathcal{C}} = \{(v_1, c_1), (v_2, c_2), \ldots, (v_n, c_n)\}$. The *extended cumulative data distribution* of X, denoted by $\mathcal{T}^{\mathcal{C}^+}$, is the cumulative data distribution of $\mathcal{T}^{\mathcal{C}}$ extended over the entire domain D by assigning zero frequency to every value in $D - V$.

2 Previous Approaches

The goal of any histogram is to accurately approximate the underlying distribution. Several different histograms have been proposed in the literature. Poosala et al. [16, 15] propose a taxonomy to capture all previ-

Figure 3 Selecting the first paper containing the word "wavelet"

Figures and tables and mathematics are rendered just as the author specified, such as that shown in Figure 4. Words used in figures (such as “Error” and “tree”) are included in the full-text version, and in the index.

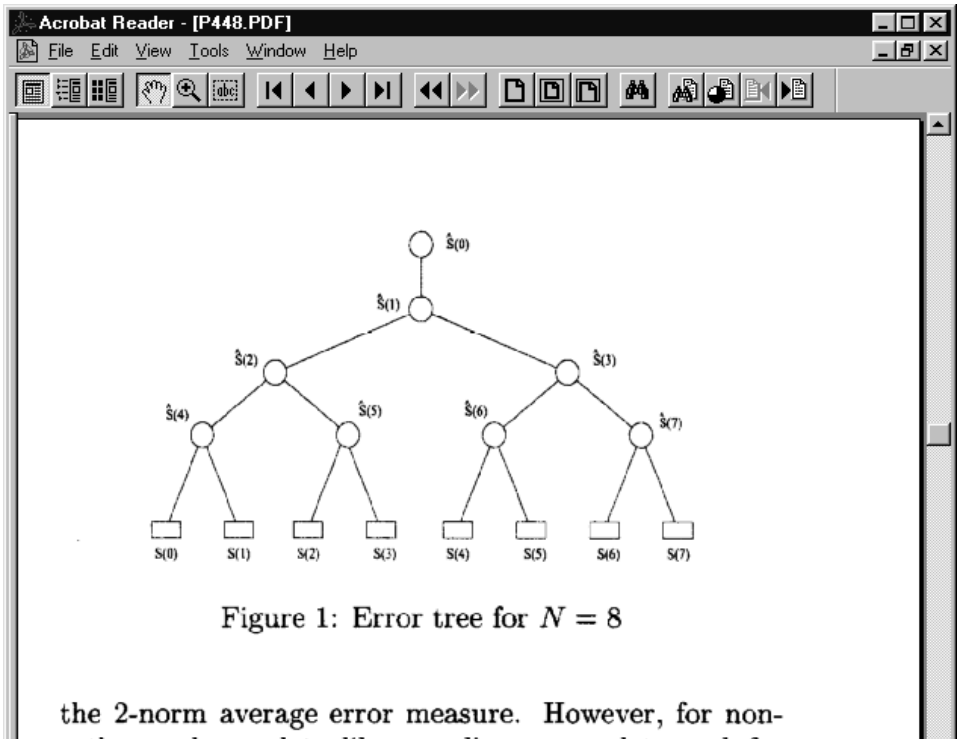


Figure 4 A later page in that paper

Going to the citation page for this paper (Figure 5), we see the bibliographic information on the paper, along with the abstract, and (further down this page), the BiBTeX entry, entries in the bibliography of the paper, and papers referencing this paper.

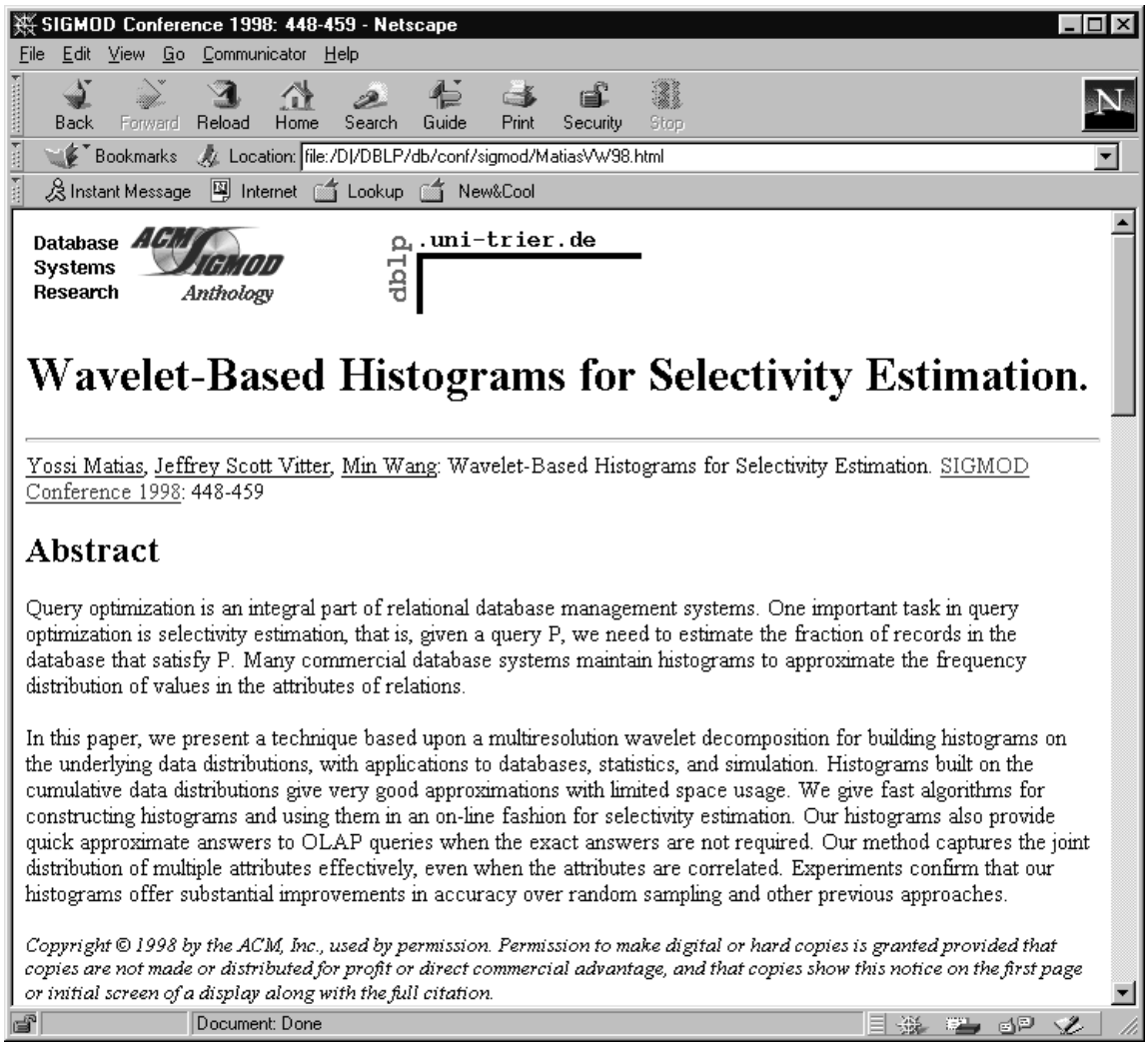


Figure 5 The citation page for the Matias paper

Scrolling down (Figure 6) displays the references in the paper's bibliography. All references shown are pointers into the bibliographic database, and are hyperlinks. Those not shown (e.g., [1] and [2]) are not yet in the bibliography database (they should still be shown, without being hyperlinked.**)**

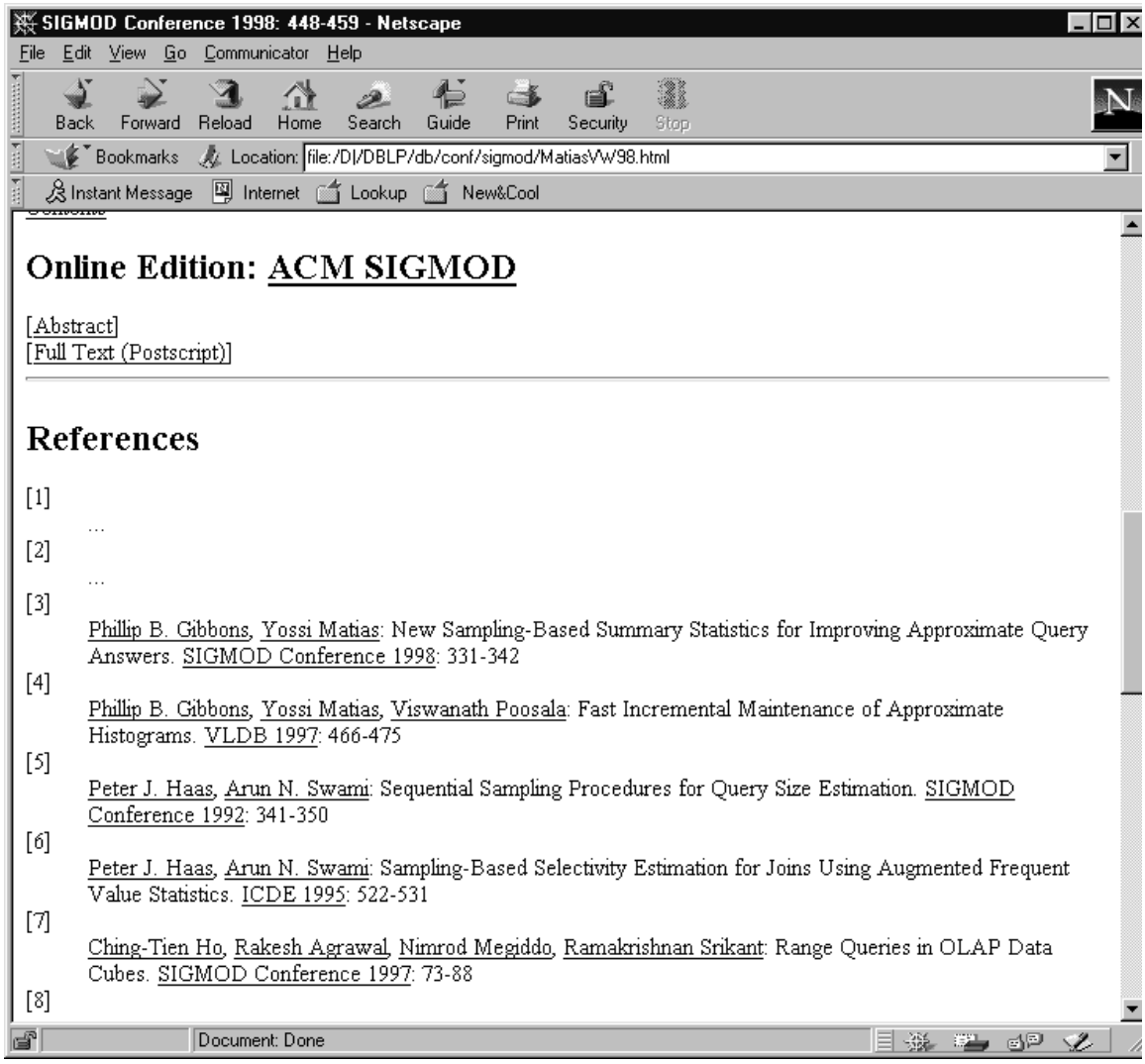


Figure 6 Scrolling down the Matias citation page, to the references

We then scroll down further in the bibliography, arriving at Figure 7.

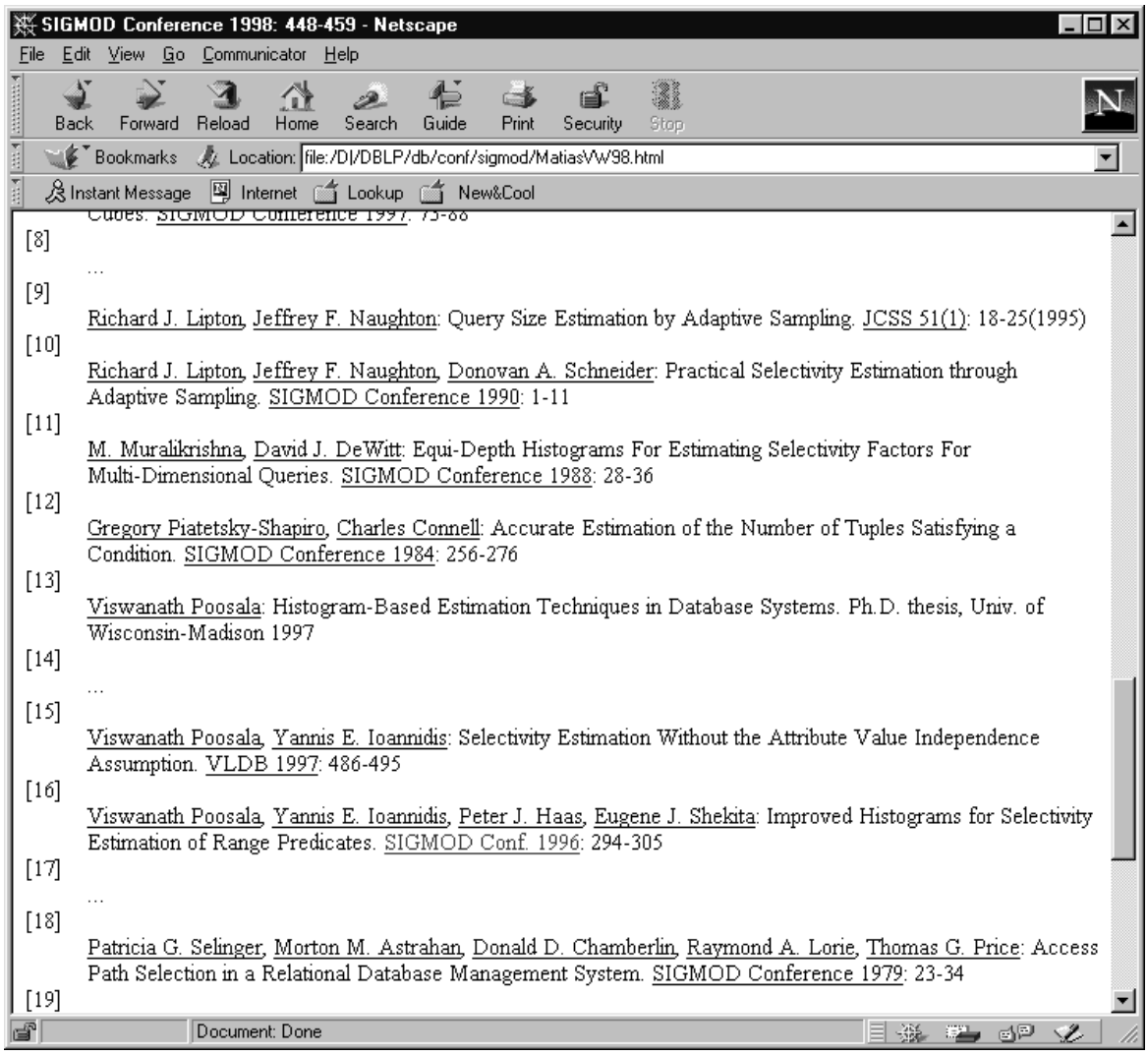


Figure 7 Scrolling further down the Matias citation page

Clicking on reference [16] results in the citation page in Figure 8, for a paper by Viswanath Poosala, et al., which is referenced by the Matias paper.

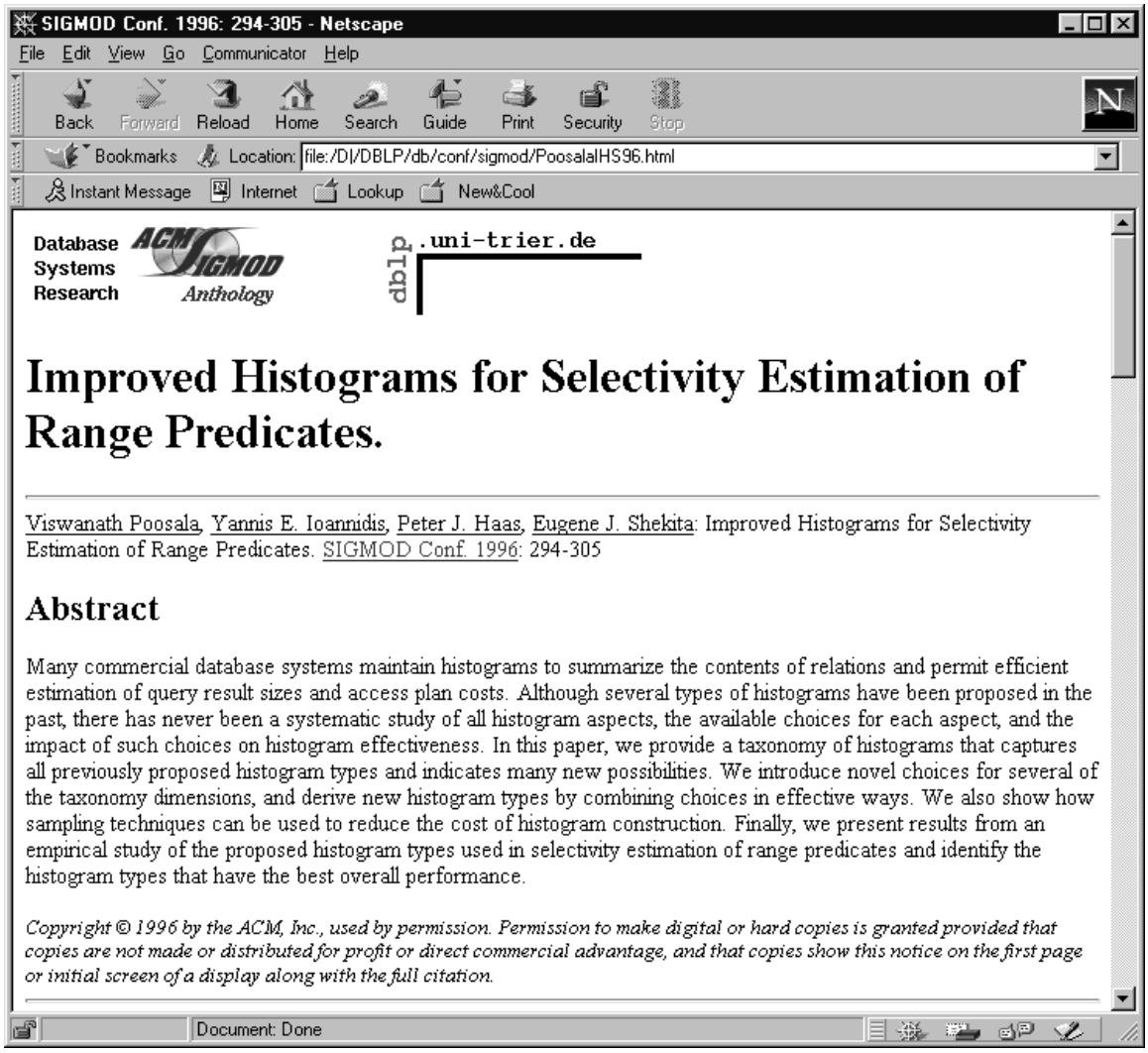


Figure 8 A paper referenced by the Matias paper

Scrolling down this citation page to the referenced-by section (Figure 9) shows that the original paper (fifth in this list) indeed references this paper. Such forward references are quite valuable to researchers to identify work building on a particular paper. They also can indicate a pattern of propagation of knowledge. Even though it appeared only three years ago, the Poosala paper has been referenced by quite a few papers, suggesting that it has been quite influential.

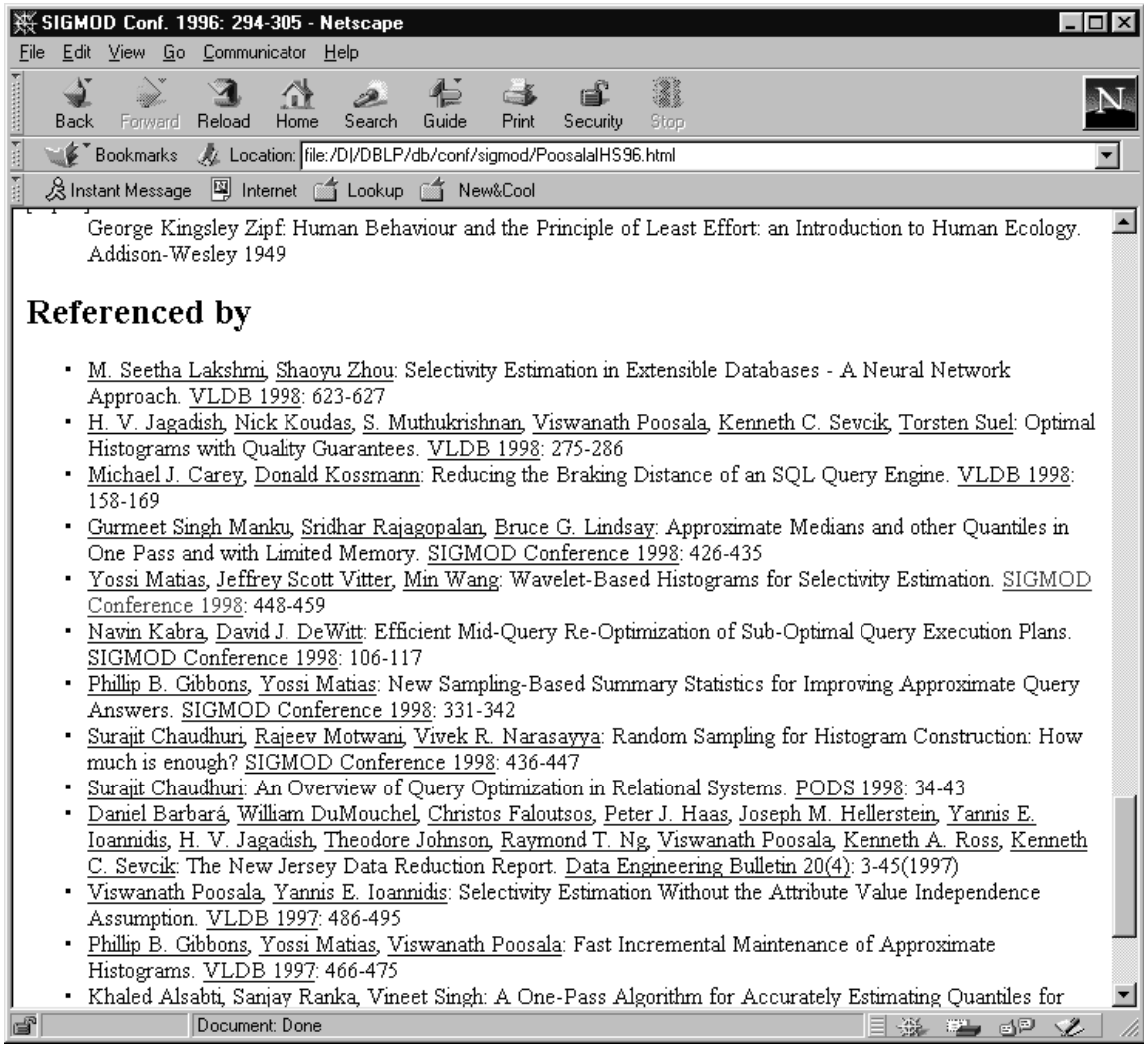


Figure 9 The referenced-by portion of the citation page for the Poosala paper

Clicking on the author of a paper brings up the author page; Figure 10 shows that for Jeffrey Vitter, who co-authored the aforementioned paper with Yossi Matias and Min Wang. The author page provides a link to the home page for that person (this link is currently generated manually, with maintenance an obvious concern); the author page is being augmented to include contact information such as the author's email address and phone number. It also provides a chronological list of the author's papers. The "EE" appearing to the left of an entry indicates that an "electronic edition" is available for that paper. Generally that edition is a citation page that includes a link to the bit-mapped contents of the paper. For the rest, we hope to provide links into publishers' digital libraries, if the paper is available on-line.

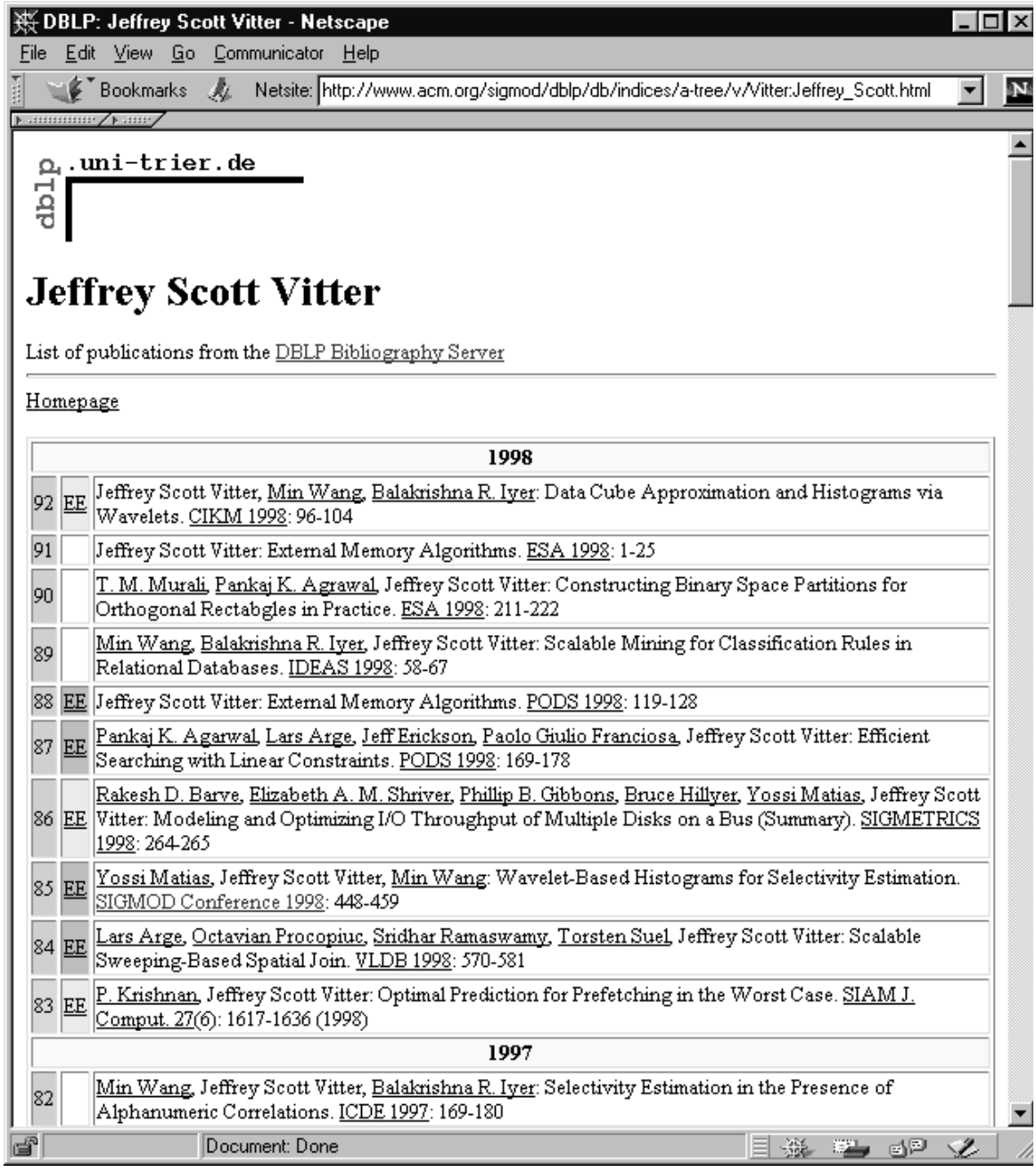


Figure 10 Jeffrey Vitter's author's page

Anecdotal evidence shows that many in the database community use author pages extensively to locate people, to check on a person's recent or long-term productivity prior to naming them to program committees or editorial boards, and to request a review for a journal submission.

As a comparison, INSPEC's fee-based bibliographic search allows one to search on author, title, journal or conference title, subject (terms in the INSPEC Thesaurus), keyword (words in titles, subjects and conference titles), or INSPEC classification code. Searching for "Vitter" as an author yields 90 bibliographic entries on 8 pages, one of which is shown in Figure 11.

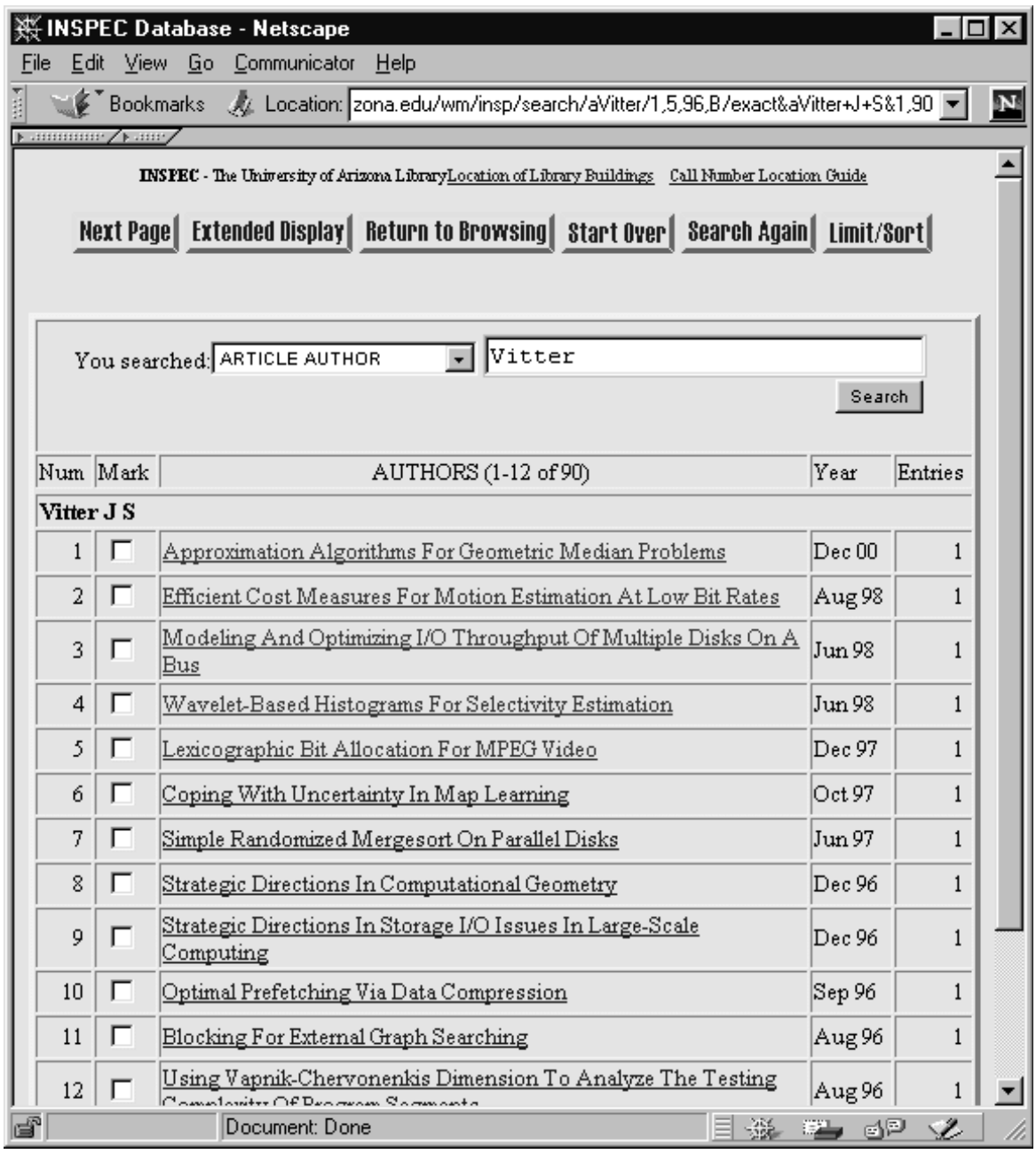


Figure 11 Searching INSPEC for author Vitter

INSPEC lists just the title of each paper, which can be clicked to get more bibliographic information. This initial listing of titles is less useful than the *Anthology* page, which is why many rarely use INSPEC, even though it is provided free to some through their home institution. Some information, such as contact information and the electronic edition, are not provided by INSPEC (instead, INSPEC indicates which library in the University of Arizona system, where INSPEC was run, has the paper, or helpfully suggests using interlibrary loan to get a copy of the paper).

Interestingly, INSPEC lists only 3 papers for Jeff Vitter in 1998, whereas the *Anthology*, a free service, lists 10 papers for that. Hitting "Next Page" on the INSPEC page four times yields the page shown in Figure 12.

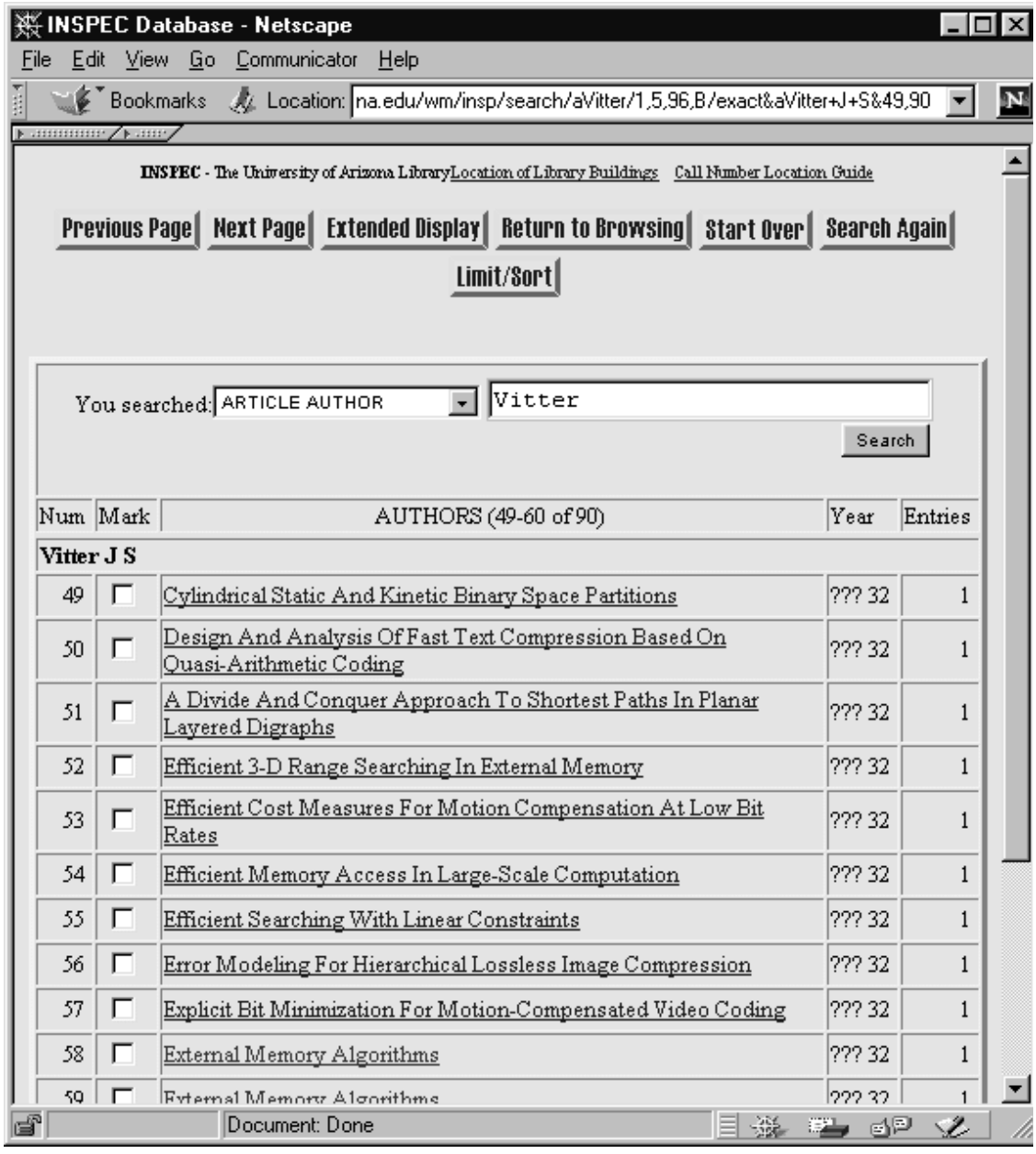


Figure 12 A further page from INSPEC on author Vitter

Interestingly, the year field of all of these entries is obviously erroneous. Comparing this with Jeffrey Vitter's author page in the *Anthology* (Figure 10), we see that item 58 from the INSPEC page ("External Memory Algorithms") appears in the correct place (item 88, in 1998) in the *Anthology* page. Both searches, the *Anthology* and INSPEC, turned up about 90 papers for Jeffrey Vitter, though the coverage overlaps only partially.

Our conclusion is that a portal derived from the model of the *SIGMOD Anthology* may indeed be as comprehensive in its coverage, more convenient in its presentation, richer in some of its content (e.g., forward citation links) while not providing other content (e.g., a restricted searching vocabulary, professional abstracting), and much more available, by virtue of it being free, than the fee-based INSPEC service.

Again, we have to be careful about this business model. The fact that INSPEC costs a lot should serve as a flag to us that even a cost-recovery model may dictate that some fee be charged. On the other hand, as we'll see below, the costs to create the portal will be far lower than those for INSPEC.

3 Realizing the Computing Portal

The portal will be a challenging project for ACM. It will work only if the various components, the SIGs, the Publications Board, the Publications and Information Systems departments at ACM, and ACM Council all contribute their resources and expertise.

While the process is articulated as a sequential series of four steps, in reality these steps should occur in parallel, with the results of a step (e.g., extraction of citations from full text) feeding another step (populating the digital bibliography).

3.1 Scope

An important initial question is, what should the portal include? The easy answer is that the portal should include all relevant computing literature. This raises other questions. Should this be restricted to computer science, or enlarged to include all aspects of information technology? Should it be restricted to core research literature, or be extended to include substantial practitioner literature? Should it focus only on current literature, or attempt to be comprehensive, including all literature back 60 years? Which journals should be included (as food for thought, should we consider the *Journal of Electronic Materials*, the *Journal of Information Science*, the *Journal of Computational Physics*, the *Journal of New Music Research*, the *Journal of Multivariate Analysis*, the *Information Resources Management Journal*, *Science*)?

We recommend that pragmatics dictate the answers. So our initial take is the following.

- The interests of the SIGs should ultimately determine the scope. Each SIG should identify the journals, conferences, and other resources most relevant to that SIG, with the portal comprised of the combination of all of these resources. This approach essentially defines "computing" to be the union of the interests of the SIGs.
- Gathering material from the last ten years (1990-1999) is the most important. As the rate of production has been increasing, this material will also be the most voluminous. Gathering material from the prior fifteen years (1975-1989) is also rather important; there is a good deal that was produced during that time. The prior thirty-five years (1940-1974) is less important, and also represents a smaller amount of material. Nevertheless, many classic papers in programming languages, compilers, and theory came out of that time, and should be included. We rely on citation linking and the SIGs to identify what is important.
- Certainly archival journal papers are the most important publications to include in the portal. Conference proceedings are also very important, as many of the ideas appeared there, and never made

it into journals. Workshops tend to be trendier, and so are relatively less important. Technical newsletters also fall into that category. Research books are much less numerous, and so at least the bibliographic information on them should be relatively easy to capture. We would like to include relevant theses and dissertations. Trade press books are more problematic. While some are useful, the contents of many don't hold up well over time, and the quality varies highly. These reservations apply even more to trade magazines and newspapers. Software goes out of date quickly and is notoriously difficult to index, and so should not be included.

We should gather all the materials that make sense into the portal, but we should not go to extraordinary lengths to get materials that are difficult.

3.2 Step 1 Bibliographic Entries

One approach is to have each SIG be responsible for collecting the relevant entries and for ensuring the accuracy of the entries it contributes to the portal. The SIG Portal Committee would coordinate, to reduce overlap between the SIGs. The SIG Governing Board should provide or fund software for data entry, validation, conversion, and duplicate detection, for use by all the SIGs.

Another approach requires less of each SIG, requesting primarily funds for the majority of the effort to be done in a centralized manner.

Clearly, it is desirable to utilize the expertise and knowledge of the field resident in each SIG. For example, people in an area can much more easily disambiguate author names, because they often know the authors personally.

Starting with an estimate of 1 million computing entries, each SIG would be responsible for very roughly 30,000 entries, a manageable number.

There are many resources for collecting bibliography. The table of contents of conference proceedings and journals could be typed in and converted into the proper format. Clerical staff could do much of that work. The *SIGMOD Anthology* already includes some 130K entries; the *ACM Guide to Computing Literature* contains 300K entries. The bibliographies of the 30K-odd papers currently residing in the ACM Digital Library contain perhaps a million citations (see step 4), with many to the same influential articles. (As the ACM Digital Library expands (see step 3), this resource of citations will also grow. Note however that citations are often incomplete. Nonetheless, it may be possible to harvest bibliographic entries from the bibliography of the full-text version of articles, when available.) The Collection of Computer Science Bibliographies (http://liinwww.ira.uka.de/bibliography/index.html) contains some 930K entries (again, with much overlap), which could be used to check the accuracy of the Portal entries. There are many other smaller bibliographic collections (http://www.library.cmu.edu/bySubject/CS+ECE/bibs.html provides an impressive list of such resources). Individual people have bibliographies that they may be willing to give to ACM for conversion and consolidation. Indeed, whenever possible the portal should utilize existing volunteer efforts. Such an approach would significantly benefit from a freely accessible portal; people are much less likely to contribute to a portal that they themselves or the community with which they identify would have to pay to use.

The bibliographic entry should provide all the standard fields (author(s), title, etc.) These formats should exploit the careful work done by librarians, for example in developing the MARC format. Several major commercial and society publishers have agreed upon a prototype XML DTD for the exchange of bibliographic metadata for journal articles.

The bibliographic entry should also include an internal unique identifier, for use by within the bibliographic software. It would be useful to provide an externally visible unique identifier, such as a *Unique Reference Identifier* (URI) or *Digital Object Identifier* (DOI), but only if available. ACM should not attempt to impose a unique identifier on all the computing literature, but should adhere to any standards that arise in the future.

Bibliographic entries should be stored in a database, to enable easy manipulation. Now that some DBMSs support XML, it might make sense to store them as XML records. The cost of acquiring and maintaining the DBMS is an issue, especially if the backend of the portal is distributed, with portions provided by different groups. The database should record the provenance of each field of each record, to track changes to these records.

The portal should support multiple output formats for bibliographic entries. Relevant output formats include BiBTeX, refer, endnote, MS Word, HTML and XML. XSL might be used to generate various formats from XML data. Ideally the portal can provide tools for popular word processing systems to import entries via the World Wide Web. Imagine including in a paper you are writing citations identifying entries in the portal, and having the tool automatically download and format those entries into a bibliography.

This information can be presented in a variety of ways. Various search tools can provide lists of relevant entries, sorted by some relevance measure. (There is the concern that in searching several hundred thousand full-text articles, an unhelpfully large number of articles may be returned from an unfocused search.) From individual entries found on these lists, the user can request related entries, again, ranked in some way. There should also be hierarchical searching through keywords or the ACM CCS, as well as search over identified bibliographic fields (e.g., author, title, abstract, year). A variety of recomputed lists can be provided, including the following.

- Articles by each author. ACM should be able to provide every author with a virtual and permanent bibliographic home page, and allow the author to provide corrections to these entries, as well as links to his/her current site. An example of such a bibliographic home page may be found in the *SIGMOD Anthology* at www.acm.org/sigmod/dblp/db/indices/a-tree/v/Vitter:Jeffrey_Scott.html (Figure 10).

- Articles in a collection, such as a proceedings or book or conference series.

- Articles by journal and year, by journal and volume, by journal issue.

- Highly rated articles, such as the frequently cited database articles (www.acm.org/sigmod/dblp/db/about/top.html) and the *ACM SIGMOD Digital Review* (data.cs.washington.edu/digrev/) which reviews and recommends articles appearing in the Computing Research Repository, CoRR.

- SIG-specific portals over a relevant subset of the entries.

If the article is available on the web, the bibliographic entry should provide a URL. This URL should refer the specific work; thus cooperation with publishers to ensure correct URLs is important. Such a facility provides customers for cost-based digital libraries; the ACM should use this to negotiate good terms with publishers owning these digital libraries. The hope is that with a complete repository of bibliographic entries for the entire corpus, and that users will routinely go to the portal first to find material, even if that material is distributed across many sites and proprietary digital libraries.

3.3 Step 2 Keywords and Abstracts

It makes sense for the SIGs to also collect these. We may need to acquire copyright permission, to be negotiated by ACM HQ. As one source, the Collection of Computer Science Bibliographies has 100K abstracts.

The challenge here will be to keep the collection cost down, while maintaining acceptable accuracy. A primary expense of commercial abstracting and indexing services is paying professionals to write abstracts and index papers, which is very expensive. We should distinguish between author-supplied keywords, which are readily available, and professional indexing terms, which are extremely costly to generate. We

recommend that ACM not attempt to develop a precise classification scheme from a controlled vocabulary or produce abstracts in-house, and instead rely on terms and abstracts provided by the authors. The ACM Computing Classification Scheme (There are now three versions, specified in 1964, 1991, and 1998.) is very useful, and should definitely be used with the 300K entries now present in the *ACM Guide to Computing Literature*. It might make sense to apply this scheme over time to the 700K expected additional entries, but such classification should be attempted only if possible in a cost-efficient manner. (One additional problem is that the CCS has evolved over time.) Full text searching partially alleviates the need for a controlled vocabulary of index terms. Also, it might be possible to automatically extract index terms from the full text, either from the beginning of the article, where they are provided by the authors, or by simply extracting terms identified in some manner as useful from the body of the article.

3.4 Step 3 Full Text and Bit-mapped Images

We propose that the SIGs fund populating the ACM Digital Library, creating a PDF file for each journal, conference, and newsletter article. (As in Section 1.3, we emphasize that the documents in the ACM Digital Library are logically separate from the portal, which consists of bibliographic entries that include URLs into the ACM Digital Library, as well as into the libraries of other publishers.) Each page will be scanned, to create a TIFF image, which can be OCRed to produce full text, at 95-99% accuracy, termed a *PDF-wrapped TIFF with hidden OCR*. This will cost about $1.00 per page. (Going to SGML or XML and increasing the accuracy to 99.9% would cost $8-$10 per page, which is prohibitive.) There are about 125K pages of ACM journals, 275K pages of conference and workshop proceedings, and 120K pages of newsletters, for a total of about 550K pages at a digitization cost of $600K, or an average of $20K per SIG. Of course, the coverage of SIGs varies dramatically, as does the available funds from the various SIGs, so a fair allocation will need to be determined. The SIG Opportunity Fund can also be used.

The PDF files already encapsulate compressed bit-mapped images of the pages of the article. The OCRed version will certainly contain errors; the images will contain smudges and blurred portions. These annoyances can be partially corrected, but manual correction is extremely costly, at $3-$10, or more, per page. The best bet seems to be to await advances in OCR and automatic image tuning technology, and then apply these corrections when these technologies mature, to clean the stored images.

ACM should as a general policy not pay for digitization of material whose copyright is owned by someone else. To digitize the entire computing corpus would cost approximately $15M, which clearly exceeds available funds. Rather, the portal should encourage other copyright owners to digitize their material, and to provide full text to the portal for indexing. In this way, the portal via an investment of perhaps $1M can eventually effect digitization of the full corpus. Additionally, ACM may wish to negotiate reciprocal agreements with other societies, providing their members with access to the ACM DL in exchange to access by ACM members to other digital libraries.

The ACM Digital Library should clearly delineate the source of the article. While ACM has control over the content of its publications, it does not over those of other publishers. It is to everyone's benefit that the source of the article be obvious, which will be a challenge when articles from many publishers are but a click away.

The ACM Digital Library when populated will comprise some 550K pages, or about 40K papers. This represents very roughly 5% of the total of one million papers (the *SIGMOD Anthology* experience is that the ACM papers constitute many of the highest quality papers). For the remaining vast majority of papers, ACM should negotiate with publishers to provide customers for their digital libraries by utilizing such URLs in the portal, in exchange for the full text of the papers for use in indexing and extracting bibliographic information. As the SIGs are the primary beneficiaries, it seems reasonable that they pay for the costs of integrating acquired full text into the portal (which will cost on the order of 25 to 50 cents a page, over the cost of scanning).

Again, the *SIGMOD Anthology* experience is instructive. Publishers were not contacted directly by SIGMOD. Instead, the relevant scientific body (conference steering committee or journal editorial board)

was contacted, in the spirit of joint cooperation. That scientific body was almost always extremely enthusiastic about being involved. Conversely, the reaction of the publisher was almost always initially negative (to its credit, Morgan Kaufmann Publishers has been supportive from the beginning). Only through dedicated lobbying by the scientific body have several forward-looking publishers, including the IEEE Computer Society and Springer Verlag, agreed to allow their copyrighted work to appear in the *Anthology*.

A second factor was the extremely favorable terms offered by SIGMOD. Anyone providing significant content was given the right to purchase copies of the *Anthology* at cost, and to sell it for whatever amount that organization desired. In many cases, the publisher didn't want to bother. For some conferences, the conference decided to give the *Anthology* away to the conference attendees, who were, after all, paying for digitizing the proceedings of that conference. The IEEE Technical Committee on Data Engineering is distributing the *Anthology* free of charge to all of its members (specifically, those who won't already be getting it through their SIGMOD membership). SIGMOD's view was that the more copies distributed, by whomever, the better, for everyone concerned.

A constructive momentum thereby developed, so that once a critical mass of participating organizations and publications was gathered, later publications began to feel, rightly so, that they would be marginalized if they didn't participate. Grabbing an article off of an *Anthology* CD-ROM is so much easier than buying an individual proceedings, or going to the library, or waiting a week or more for interlibrary loan to procure the article. (Currently, 156 volumes of conference proceedings and journals are included in the *Anthology*, equivalent to an acquisition cost to individuals of many thousands of dollars.) The *Anthology* is to the point where conference organizers regularly contact SIGMOD for information on how to be included.

We conclude from this experience that the scientific community will enthusiastically support this venture. Publishers will be initially highly reticent, and will have to be cajoled into participating by that community. Economic incentives, such as the portal delivering customers to proprietary digital libraries, should be emphasized, so that it is in the best interest of the publishers to participate. Care should to be taken to configure the project so that everyone, including researchers, ACM, publishers, and the scientific societies, benefit. The sequencing of participants is important, with momentum critical to get everyone on board.

The issue of which format to use is an important one. We should differentiate here between storage format and distribution format. Ideally, a single storage format should be utilized, with multiple distribution formats provided. The full text will be stored in proprietary indexes used by the search engine, but will not otherwise be available directly to users. Bitmapped images of the articles that have been scanned in should be stored in PDF as wrapped TIFFs with hidden OCR full text, which is easy to generate from scanned images. When the document's source is available, as is the case with many recent conferences, native PDF should be used. That format requires much less space, and is of higher quality.[3] SGML is difficult and costly to produce from scanned images, and so is not appropriate as a storage format. Similar considerations eliminate XML. And the gif and jpeg formats are not suitable either.

Output formats should be platform independent and not represent a significant cost to the user. PDF and postscript are currently the most prevalent means of distributing articles on the web. Readers for these formats are freely available on major platforms. Postscript can be converted to ASCII using Prescript, though it is not clear the portal should distribute ASCII versions of papers. DjVu (www.djvu.att.com) appears to require less disk space and is more amenable to transfer across slow internet connections. However, there are concerns about the availability of free readers, of long-term support, and of the feasibility of converting to DjVu from the single storage format. It may be possible to convert PDF to DjVu

[3] Steve Cunningham provided some interesting figures. He just did a proceedings in native PDF with many images and a lot of color, and the entire proceedings at 200 pages requires only 2.9MB (15KB per page). A 10-page monochrome scanned article from the UIST'97 proceedings (p189-dethick.pdf) requires 1.5MB (150KB per page) at lower quality. Some care has to be taken with images, though. A 12-page native PDF article from Interactions (V5N1-p25-friedland.pdf) is 2.2MB (183KB per page, of mixed text and color), while a native page from the SIGGRAPH'96 proceedings (page 128) is only 126K, even when made into a one-page document that carries the entire document overhead.

to support the latter as an additional output format. It might also be useful to support jpeg as an output format, though the utility of doing so given the prevalence of PDF viewers is uncertain.

The PDF files require several seconds per page to download over a phone-line connection. SIGs may want to provide CD-ROMs of relevant portions to members; such disks are especially useful for international members, who often have slow or costly connections to the web, and for those traveling or working from home. At some point, there will be uniformly fast access to the web from everywhere, but that eventuality is at least a decade off.

Due to the large size of bit-mapped images, numerous CD-ROMs are required to capture significant portions of the computing corpus, even specialized portions. As an example, the conference proceedings related to the field of databases require an estimated 15 CD-ROMs, totaling about 10 GB. Fortunately, DVD is becoming prevalent; all that SIGMOD-specific material will fit on a single DVD (two-layer, double-sided).

3.5 Step 4 Citation Linking

The 30,000 citations in the *SIGMOD Anthology* (now up to 55,000) were done manually, at a cost of 20 cents a citation. Most of this was done starting with the bibliography in the paper's full-text; some was done by manually typing in the bibliography. There has been some research on fully automatic citation linking, which would cost much less, though it is not known the degree to which accuracy would suffer. Perhaps a combination of automatic linking with manual scanning and correction would achieve an appropriate balance of low cost and high accuracy. Eventually the portal will contain tens of millions of citations, providing an important resource for searching and for analyses of propagation of information. At the same time, the cost of these citations must remain low, only a few cents per citation, for them to be cost-effective to collect on a large scale.

The bibliographic entries and citations for the entire computing corpus could fit on a single DVD. This one resource alone would be highly valuable to researchers and libraries, independent of its availability on the web. Producing this DVD disk in quantity would cost perhaps $5–$8, and could easily be sold to departments and libraries for $100, thereby providing a partial revenue stream.

3.6 Time Frame

The time frame will depend heavily on who does the work and how and when funding is available. Our hope is that populating the ACM Digital Library could be completed fairly quickly, as this is probably the most straightforward aspect. Setting up the infrastructure to support a million or more bibliographic entries and tens of millions of citations will certainly be challenging, and will take some time to do it right.

3.7 Maintenance

The ACM already has a mechanism in place for collecting bibliographic entries for its *Guide to the Computing Literature*. These new entries can be funneled into the portal to keep it up to date. Arrangements should be negotiated with publishers for them to provide accurate bibliographic entries, as well as full-text for indexing and URLs into their digital libraries. It would be useful to reach an agreement about the minimal bibliographic information provided by digital libraries. For example, author names should not be abbreviated. Editors, exact titles, and ISBN should be available at the volume level (this information is missing in the ACM DL…). Session titles are listed in the IEEE DL, but not in the ACM DL, and conversely, page numbers are available in the ACM DL, but not the IEEE DL.

Additionally, conferences should be encouraged to provide native PDF files of accepted papers, as that is the most space-efficient and highest quality format.

It may be possible to exploit the parallelism of the web by enabling user correction of entries via a web interface and/or an email interface, while not requiring human intervention at the backend by having an automated process that could update the entry. Of course, this would have to be designed carefully so that it avoids corruption of previously collected information.

4 Open Architecture

It is critical that searching the bibliographic data and the full text of the papers be available to scholars, whether ACM members or not, via a web interface and at the central http://acm.org site. Additionally, specialized searching should be provided at individual SIG portals, to provide visibility and a sense of community for the SIGs that contribute to the portal. We should also attempt to make the portal available for mirroring, both on a geographical basis, for those areas from which access to the http://acm.org site is too slow, and on an institutional basis, say an individual library or even a specific department within a company or academic institution mirroring the portal on its intranet. At the same time, it is important that ACM and perhaps the sponsoring SIGs be prevalent on the mirrored pages, and that a link (and perhaps a version number or the mirror date) be present back to the definitive location of each page.

The bibliographic data and the full text should also be made available, perhaps under stated constraints, to those doing research in information retrieval, knowledge propagation, and other disciplines. The bibliographic data will be somewhat easier to make available, as ACM will have more control over it. Distribution in any way of full text acquired from publishers and other sources is restricted by the highly disparate copyright policies of these various publishers. In negotiations with these publishers, ACM should push for allowing distribution of full text for research purposes.

The bibliographic and full text data can be helpful in evaluating new algorithms and in providing insights into the scientific process, at least as exercised in the field of computer science. The resulting algorithms and insights can be fed back into the portal to increase its effectiveness. As a concrete example, say a researcher develops a new information clustering approach. It can be applied to the full text in the portal, and evaluated using metrics designed by the researcher. If found effective, the results from that analysis could be made available by the portal, thereby helping to guide the use of the portal by subsequent researchers. By allowing others to develop better user interfaces to the information stored in the portal, wider usage and dissemination of that information will result.

Defining appropriate constraints and their enforcement mechanisms will need careful thought. For example, it is not clear whether enforcement will be by technology, by policy, or by license. One approach would be to take the courageous step to establish a culture of free high quality bibliographic data. We recommend that ACM develop a copyright for such data that is in the tradition of the GNU public license, including a detailed list that clearly states what is permitted and what is forbidden. A truly comprehensive portal is only possible if ACM cooperates with other learned societies and profit-oriented publishers. This will work only if the rights are symmetrical.

5 Previous Efforts

Several SIGs have over the last few years pioneered the distribution of electronic versions of their literature.

- SIGDA organized an ambitious effort to capture all the literature concerning design automation. Papers were retyped and converted into SGML, for high accuracy. This 9 CD-ROM publication cost $1.5M. Unfortunately, due to its high cost, $200 per copy, less than 1000 copies have been sold.

- Many conferences distribute a CD-ROM of the papers for that conference. SIGOPS does so for its SOSP conference, SIGGRAPH for its annual conference.

- SIGPlan distributes as a member benefit a CD-ROM containing 10 years of POPL, as PDF files.

- SIGMOD, as previously mentioned, has distributed volume 1 of the *Anthology* (5 CD-ROMs) and volume 1 of the annual *DiSC* (2 CD-ROMs), with additional volumes of each comprising a similar number of CD-ROMs planned for Spring 2000. The first volume of these two publications cost about $120K, and have been distributed to over 3000 individuals and libraries.

6 Summary

The *ACM Computing Portal* represents a qualitative change in the searching and retrieval of computing literature resulting from a confluence of technology, necessity, and catalyst. The newly available underlying technologies include efficient and effective OCR algorithms, inexpensive, high-resolution scanning devices, inexpensive disks, fast and pervasive World Wide Web access, high capacity, inexpensive CD-ROMs, and prevalent CD-ROM readers. The research community is abundantly receptive to services that provide responsive access to the technical literature, which is growing at an exponential rate and is already overwhelming. ACM, through its highly regarded publishing arm, its three dozen high-caliber Special Interests Groups, and its long-standing commitment to furthering scholarly activity, is an effective catalyst. It is impossible to fully gauge the impact of a freely searchable database of the complete computing corpus, but that impact will assuredly be enormous.

7 The Next Step

Carla Ellis, chair of the Executive Committee of the SIG Governing Board, in spring 1999 appointed the SGB Portal Committee and charged it with making technical and logistical recommendations to the SIGs on the implementation of the portal. This proposal is an initial product of the discussions of that committee.

There are many challenges ahead. This white paper has focussed on populating the ACM Digital Library and the portal, and has thus focused on the existing corpus. While the ACM has a plan for integrating new material into the Digital Library, procedures are still needed to maintain the portal. The role of the individual SIGs vis-à-vis the SGB and ACM staff still must be worked out. Will the SIGs primarily supply resources (specifically, funds), or expertise, or some combination of both? What is the appropriate interaction between the portal and the Computer Research Repository (CoRR)? What are the implications of the U.S. Copyright Law, which has evolved over the forty-year history of computer science?

The next year will indeed be exciting, as these issues are worked through, and the ACM Computing Portal realized.