



# Scholarly Publishing after the Web

---

Michael Pazzani

Rutgers

Pazzani @ Rutgers.edu



# Outline

---

- Research Impact
- Publication Models and the Internet
- Access to Funded Research
- The internet is more than a way to print a paper stored on a remote computer but a way to analyze that you can't do with paper



# Research Impact

---

As a researcher (and former federal funding official, VP for Research, member of tenure and promotions committee) primarily interested in research impact:

- Create a foundation that others build on
  - Have others read publications
  - Have others cite publications
  - Influence commercial developments
  - Have results reported in lay publications
  - Have findings taught to next generation of researchers
  - Have work summarized in textbooks
- Understand the existing literature
  - New foundations, problems
  - Identify novel combination of existing ideas.



## Increasing the impact of research

---

- Publishing in widely read outlets
- Publishing in prestigious outlets
- Publishing “online” (where readers and search engines can find)
  - Citations are one measure of the impact of a publication.
  - Articles online are cited 2-7x more.
  - S. Lawrence. Online or invisible? *Nature*, 411(687):521, Jan 2001.

# Authors Distributing Publications Before the Internet

## REQUEST-A-PRINT™

3/24/97

Dear Dr. Granger:  
Please send me a copy of  
your article:  
"Distinct memory  
circuits composing the  
hippocampal region"  
published in Hippocampus  
6/6 ( 1996), p.567-578

Lutz Slomianka

  
Thank You

0919  
NO. N. 121  
PAGE PAID

LUTZ BLOMIANKA  
ANATOMY & HUMAN BIOLOGY  
THE UNIVERSITY OF WA  
NEDLANDS 6907  
AUSTRALIA

*ANU. SLOMIANKA*

Requested by:

~~Albert Pines~~ & Human Biology  
The University of WA  
Nedlands 6907  
Australia



# Postscript & email: 80s and early 90s

---

**From:** pazzani@ics.uci.edu

**To:** mooney@cs.utexas.edu

**Subject:** MLC92 Paper

**Date:** Wed, 29 Jan 92 01:07:14 PST

Ray-

Here's a postscript copy of the paper  
you requested...

```
%!PS-Adobe-2.0%%Title: Pazzani-MLC92-  
Average
```

```
/LW{save statusdict/product  
  get(LaserWriter)anchorsearchexch  
pop{dup length 0  
  eq{pop1}{(Plus)eq{2}{3}ifelse}
```



# Open archive & anonymous ftp

---

To: mooney@cs.utexas.edu

Subject: MLC94 Paper

Date: Wed, 26 Jan 94 02:19:20 PST

Ray-

My papers are in a ftp archive, here's the  
README file

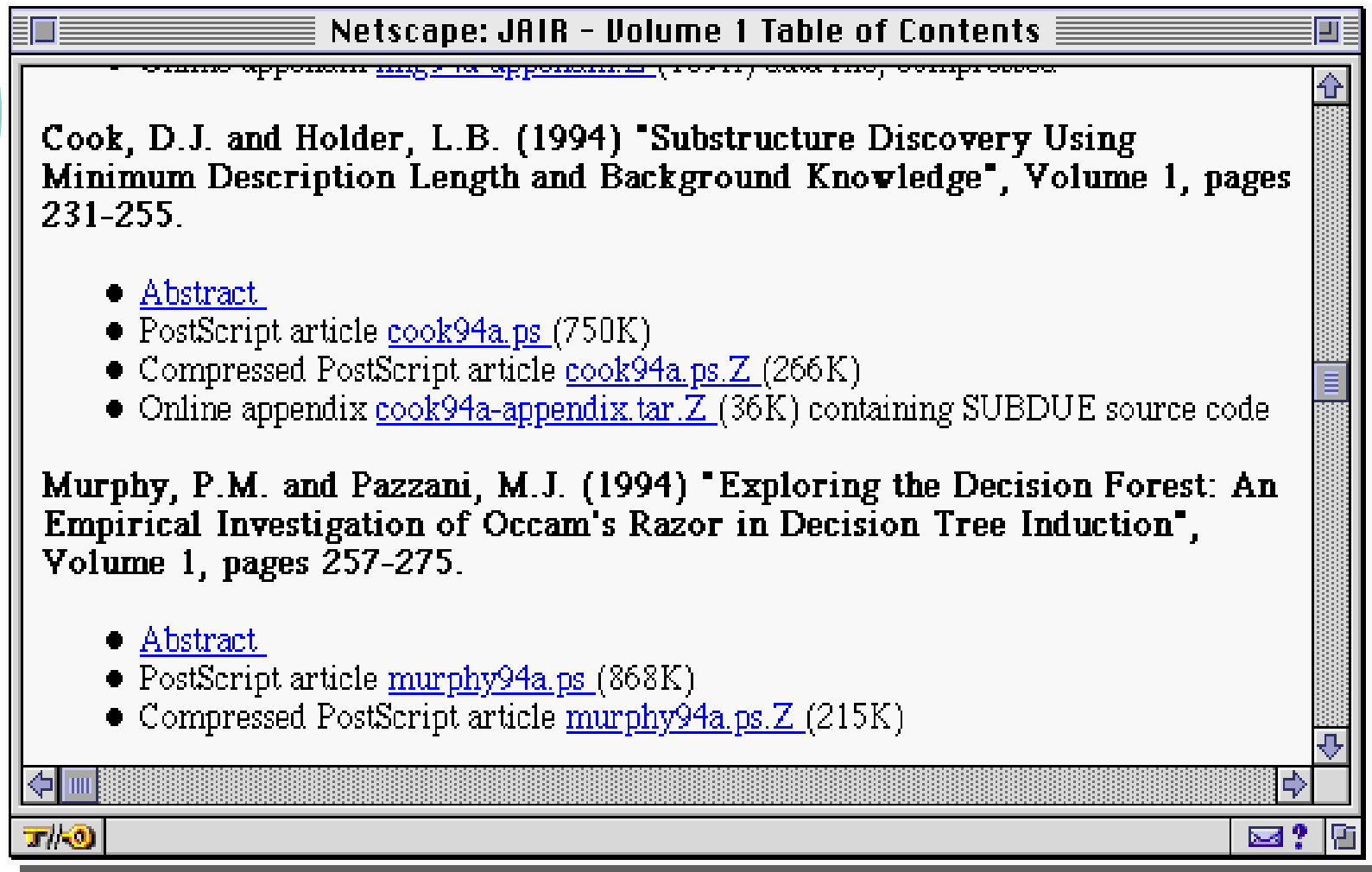
1. ftp to ftp.ics.uci.edu  
Use anonymous as the login name  
and your email as the password
2. cd to /pub/pazzani/papers
3. ls (to see file names)
4. bin (make sure you use binary mode)
5. get paper.ps.Z
6. uncompress paper.ps.Z
7. lpr paper.ps

# Open Archiving & WWW





# Open Access Journals



Netscape: JAIR - Volume 1 Table of Contents

Cook, D.J. and Holder, L.B. (1994) "Substructure Discovery Using Minimum Description Length and Background Knowledge", Volume 1, pages 231-255.

- [Abstract](#)
- PostScript article [cook94a.ps](#) (750K)
- Compressed PostScript article [cook94a.ps.Z](#) (266K)
- Online appendix [cook94a-appendix.tar.Z](#) (36K) containing SUBDUE source code

Murphy, P.M. and Pazzani, M.J. (1994) "Exploring the Decision Forest: An Empirical Investigation of Occam's Razor in Decision Tree Induction", Volume 1, pages 257-275.

- [Abstract](#)
- PostScript article [murphy94a.ps](#) (868K)
- Compressed PostScript article [murphy94a.ps.Z](#) (215K)

Digitizing doesn't always make  
information easy to find:  
Organization is important





# Internet Publishing Models

---

- Subscription (to user, library, site)
  - Extension of print business model, but less availability
- Open access, peer reviewed journals
  - Ease of access for user, but is there a sustainable business model? will publications be archived
- Self published, open access
  - Not reviewed, quality not assured, but community recommendations?
  - *I could have seen further if it weren't for the giants standing on my shoulders*
  - Hard for new authors, new articles to be found
- Open archive
  - Author publishes in subscription or other venue
  - Author retains the right to
    - Place article in an open archive (government, university)
    - Place article on personal web page



# NIH Publication Policy

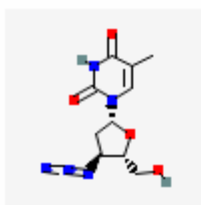
---

- Authors ~~must~~ may put NIH funded publications into PubMed
- Publicly funded research should be free to the public.
  - *But does the public want to see the 30,000 HIV articles published each year?*
- NIH needs to compile these publications into a single archive in order to manage its research portfolio better
- Greater interconnectivity and functional integration between the multiple and large research data bases (e.g., Genbank and PubChem) and an archive of NIH-funded publications has the potential to enhance research in novel ways
  - Manual Integration
  - Link to data (and from data to publications)
  - Automated text analysis

# PubMed PubChem Integration

- 2: **Delavirdine in Combination with Zidovudine in Treatment of Human Immunodeficiency Virus Type 1-Infected Patients: Evaluation of Efficacy and Emergence of Viral Resistance in a Randomized, Comparative Phase III Trial.**
- Joly V, Moroni M, Concia E, Lazzarin A, Hirschel B, Jost S, F, Bentwich Z, Love WC, Hawkins DA, Wilkins EG, Gatell AJ, Vetter N, Greenwald C, Freimuth WW, de Cian W, et al.  
*Antimicrob Agents Chemother.* 2000 Nov; 44(11): 3155-3157
- Links
- ▶ Compound
  - ▶ Substance
  - ▶ PubMed
  - ▶ Cited Articles

- 5: SID: [7980906](#)



CID: [35370](#), zidovudine

Source: [xPharm\(9753\)](#)

IUPAC: 1-[(2R,4S,5S)-4-azido-5-(hydroxymethyl)oxolan-2-yl]-2-methylpyrimidine-2,4-dione

MW: 267.242 | MF: C10H13N5O4

Related Structures Literature

- Literature
- ▶ MeSH Headings
  - ▶ PMC Articles
  - ▶ PubMed via MeSH



# Undiscovered Public Knowledge

---

- Can a M.D. or Ph.D keep up with 30,000 articles a year?
- Swanson: Finding connections between literatures (semi-automated searches)
  - Raynaud's syndrome is related to blood viscosity
  - Blood viscosity related to dietary fish oil
  - Discovery: treating Raynaud's syndrome with dietary fish oil

# Topic, Keyword extraction and analysis

Selection: **Word Vector**

Name: client server

# Topics: 20  All

**Submit Query**

Id	Topic	Probability
97	server_client_servers_...	1
1	planning_plan_plans_...	0
2	hierarchical_clustering...	0
3	cells_neurons_model...	0
4	assumptions_show_i...	0
5	performance_improve...	0
6	traditional_provide_exi...	0
7	recognition_face_hum...	0
8	parameters_values_p...	0
9	product_products_res...	0
10	algebra_algebras_alg...	0
11	beam_scattering_ener...	0
12	map_mapping_maps...	0
13	software_development...	0
14	partial_general_gener...	0
15	se_ar_id_al_tm_	0
16	optimal problem solu...	0

Selection: **Topic**

Name: server\_client\_servers\_clients\_requests\_

Or

Id: 97 Probability: 0.00362118

# Words: 100  All

# Authors: 100  All

**Submit Query**

Id	Word	Probability
24805	server	0.134
4297	client	0.072
24807	servers	0.065
4298	clients	0.04
23284	requests	0.031
133	access	0.03
15745	load	0.025
18263	network	0.024
23280	request	0.02
3274	caching	0.02
24811	service	0.018
21806	proxy	0.018
23412	response	0.014
30147	web	0.009
24285	scalable	0.009
24929	sharing	0.009
2113	bandwidth	0.008


Id	Author	Probability
2675	Bestavros_A	0.012
11807	Eager_D	0.009
6557	Rodriguez_P	0.007
2511	Rexford_J	0.007
8252	Krishnamurthy_B	0.007
29921	Gao_L	0.006
1470	Kaplan_H	0.006
1315	Druschel_P	0.005
10170	Franklin_M	0.005
2231	Dahlin_M	0.005
1600	Chase_J	0.005
7930	Wills_C	0.004
8566	Vernon_M	0.004
2893	Cao_P	0.004
17412	Tewari_R	0.004
1926	Katz_R	0.003
36210	Cherkasova L	0.003

# Recommend publications: cited-by, cites, frequently downloaded, similarity, downloaded together

The screenshot shows a Microsoft Internet Explorer browser window with the title "Michael J. Pazzani: Publications - Microsoft Internet Explorer". The address bar displays "http://agents.ics.uci.edu:9001/~pazzani/Publications/FramedPubs.html". The main content area lists several publications with buttons for "Abstract" and "Postscript". A small cartoon character is visible next to the first publication. Below the list, there is a recommendation interface with a "STOP" sign icon and a "Help" link. At the bottom, a pink box contains a recommendation message and several checkboxes and buttons.

[Suggest](#)  
[Configure Agents](#)  
[Configure Topics](#)

1998

[Start 3D Assistant](#)  
  
[Help](#)

Billsus, Daniel & Pazzani, M. (1997) Learning Probabilistic User Models. in *Workshop Notes of "Machine Learning for User Modeling", Sixth International Conference on User Modeling*, Chia Laguna, Sardinia.

Murphy, C., & Pazzani M. (1997). Combining Neural Network Regression Estimates Using Principal Components. *Preliminary Papers of the 6th International Workshop on Artificial Intelligence and Statistics*.

Domingos, P., & Pazzani, M. (1997). Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. *Machine Learning*, 29, 103-130.

Pazzani, M., See, D., Shroeder, E., & Tilles, J. (1997). Application of an Expert System in the Management of HIV-infected patients. *Journal of AIDS and Human Retrovirology*, 15:356-362.

I recommend: "Learning Probabilistic User Models", because it is the most frequently downloaded article that you have not considered cited by the paper just retrieved.    Suggest after download  Use Cited-by Agent  Machine Learning  Bayesian Classifiers  Intelligent Agents

[A Framework for Collaborative, Content-Based a](#)

Michael J. Pazzani  
Department of Information and Computer  
University of California, Irvine  
Irvine, CA 92697  
pazzani@ics.uci.edu  
phone: (714) 824-5888  
fax (714) 824-4056  
<http://www.ics.uci.edu/~pazzan>

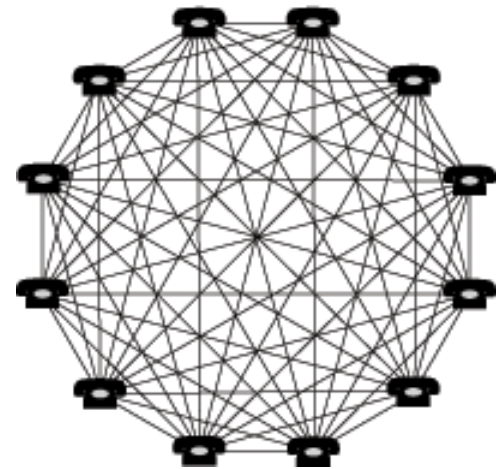
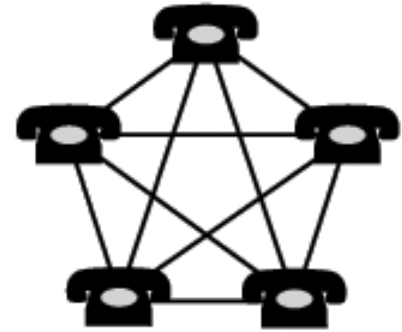
**Abstract**

*We discuss learning a profile of user interests for recommending Web pages or news articles. We describe the types of info whether to recommend a particular page to a particular user content of the page, the ratings of the user on other pages and ratings given to that page by other users and the ratings of user and demographic information about users. We describe how used individually and then discuss an approach to combining sources. We illustrate each approach and the combined recommending restaurants.*



# Metcalfe's law and Corollaries

- The usefulness, or utility, of a network grows with the square of the number of users
- Corollary: Jakob Nielsen
  - The reduced value of partitioning a network into  $N$  isolated components is  $N/N^2$  or  $1/N$  of the value of the original network.
  - Two digital libraries not interconnected have half the value of one



# Citation Links: ACM vs IEEE within CS and also interdisciplinary

and linked references.

- 1 [Krishna Bharat , Monika R. Henzinger, Improved algorithms for topic distillation in a hyperlinked environment, Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, p.104-111, August 24-28, 1998, Melbourne, Australia](#)
- 2 [Monica Bianchini , Stefano Fanelli , Marco Gori, Optimal Algorithms for Well-Conditioned Nonlinear Systems of Equations, IEEE Transactions on Computers, v.50 n.7, p.689-698, July 2001](#)
- 3 Björck, A. 1996. Numerical Methods for Least Squares Problems. Society for Industrial and Applied Mathematics.
- 4 [Immanuel M. Bomze , Walter Gutjahr, The dynamics of self-evaluation, Applied Mathematics and Computation, v.64 n.1, p.47-63, Aug. 1994](#)
- 5 Bomze, I. and Gutjahr, W. 1995. Estimating qualifications in a self-evaluating group. Qual. Quant. 29, 241--250.
- 6 [Allan Borodin , Gareth O. Roberts , Jeffrey S. Rosenthal , Panayiotis Tsaparas, Finding authorities and hubs from link structures on the World Wide Web, Proceedings of the 10th international conference on World Wide Web, p.415-429, May 01-05, 2001, Hong Kong, Hong Kong](#)
- 7 Brin, S., Motwani, R., Page, L., and Winograd, T. 1998. What can you do with a web in your pocket? IEEE Bulle. Techn. Comm. Data Eng., IEEE Comp Soc. 21, 2, 37--47.
- 8 [Sergey Brin , Lawrence Page, The anatomy of a large-scale hypertextual Web search engine, Proceedings of the seventh international conference on World Wide Web 7, p.107-117, April 1998, Brisbane, Australia](#)
- 9 Brin, S., Page, L., Motwani, R., and Winograd, T. 1999. The PageRank citation ranking: Bringing order to the Web. Tech. Rep. 1999-66, Stanford University. Available on the Internet at <http://dbpubs.stanford.edu:8090/pub/1999-66>.
- 10 [David Cohn , Huan Chang, Learning to Probabilistically Identify Authoritative Documents, Proceedings of the Seventeenth International Conference on Machine Learning, p.167-174, June 29-July 02, 2000](#)
- 11 Cohn, D. and Hofmann, T. 2001. The missing link---A probabilistic model of document content and hypertext connectivity. In Neural Inf. Proc. Syst.
- 12 [Michelangelo Diligenti , Marco Gori , Marco Maggini, Web page scoring systems for horizontal and vertical search, Proceedings of the 11th international conference on World Wide Web, May 07-11, 2002, Honolulu, Hawaii, USA](#)
- 13 Golub, G. H. and Van Loan, C. F. 1993. Matrix computation. The Johns Hopkins University Press.
- 14 Haveliwala, T. H. 1999. Efficient computation of pagerank. Tech. Rep. 1999-66, Stanford University. Available on the Internet at <http://dbpubs.stanford.edu:8090/pub/1999-66>.
- 15 [Taher H. Haveliwala, Topic-sensitive PageRank, Proceedings of the 11th international conference on World Wide Web, May 07-11, 2002, Honolulu, Hawaii, USA](#)



# What can CRA do

---

- NSF and CRA: encourage publication models that support access to publications by people **and machines.**
- NSF
  - Fund research into sustainable, archived publishing models with internet availability.
  - Create a new publication outlet
    - Give access to final reports of grants
      - Synthesis of research project vs. collection of publications
      - Report negative results, false starts