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

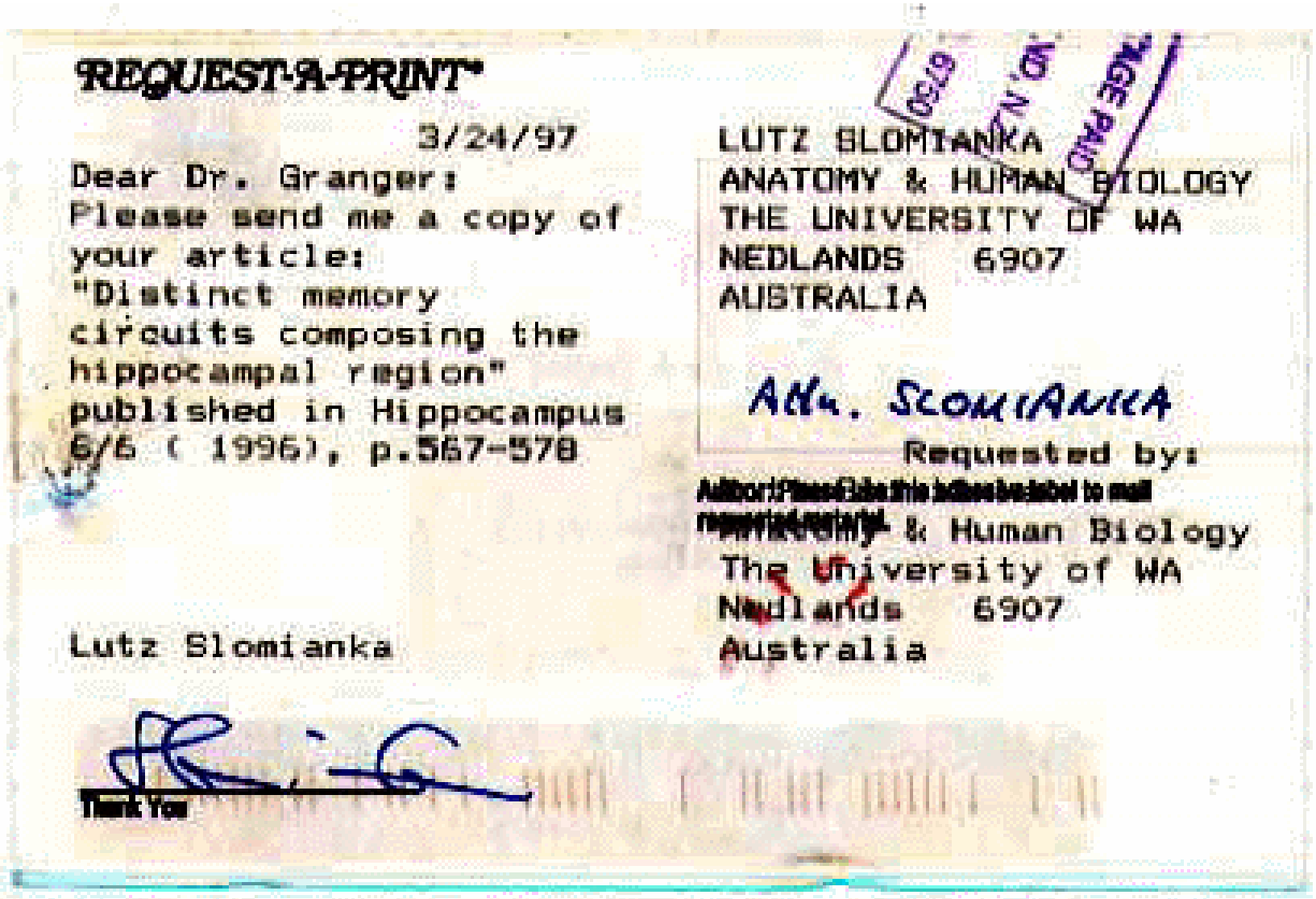
As a researcher (and former federal funding official, VP for Research, member of tenure and promotions committee) primarily interested in research impact:

- Create a foundation that others build on
  - Have others read publications
  - Have other 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

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

Attn. SLOMIANKA

Requested by:
Author: Please fill in this address below to mail reprint directly
Anatomy & Human Biology
The University of WA
Nedlands 6907
Australia

Thank You

Postscript & email: 80s and early 90s

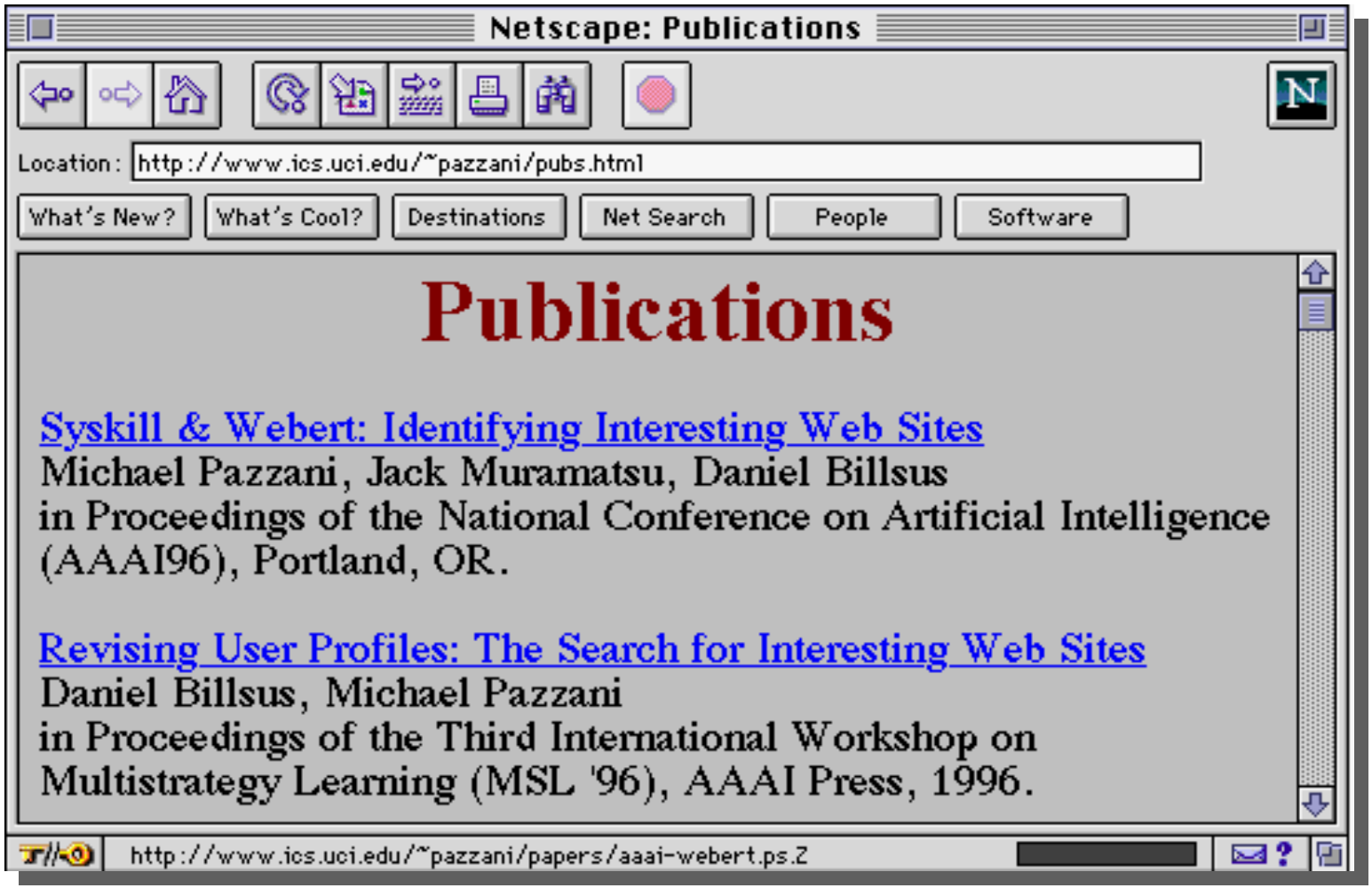
```
From:pazzani@ics.uci.edu
To: mooney@cs.utexas.edu
Subject: MLC92 Paper
Date: Wed, 29 Jan 92 01:07:14 PST
Ray-
Here’s a postscript copy of the paper
  you requested...
%!PS-Adobe-2.0%%Title: Pazzani-MLC92-
  Average
/LW{save statusdict/product
  get(LaserWriter)anchorsearchexch
pop{dup length 0
  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

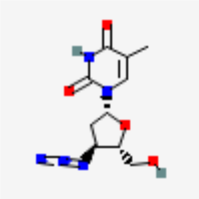
- Authors ~~must~~ may put NIH funded publications into PubMeb
- Publicly funded research should be free to the public.
  - *But does the public want to see the 30,000 HIV articles published each year?*
- NIH needs to compile these publications into a single archive in order to manage its research portfolio 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 Tre**
Human Immunodeficiency Virus Type 1-Infected P
Evaluation of Efficacy and Emergence of Viral Res
in a Randomized, Comparative Phase III Trial.
Joly V, Moroni M, Concia E, Lazzarin A, Hirschel B, Jost
F, Bentwich Z, Love WC, Hawkins DA, Wilkins EG, Gatell AJ,
Vetter N, Greenwald C, Freimuth WW, de Cian W,.
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]-5-methyl-pyrimidine-2,4-dione
MW: 267.242 | MF: C10H13N5O4

Related Structures Literature

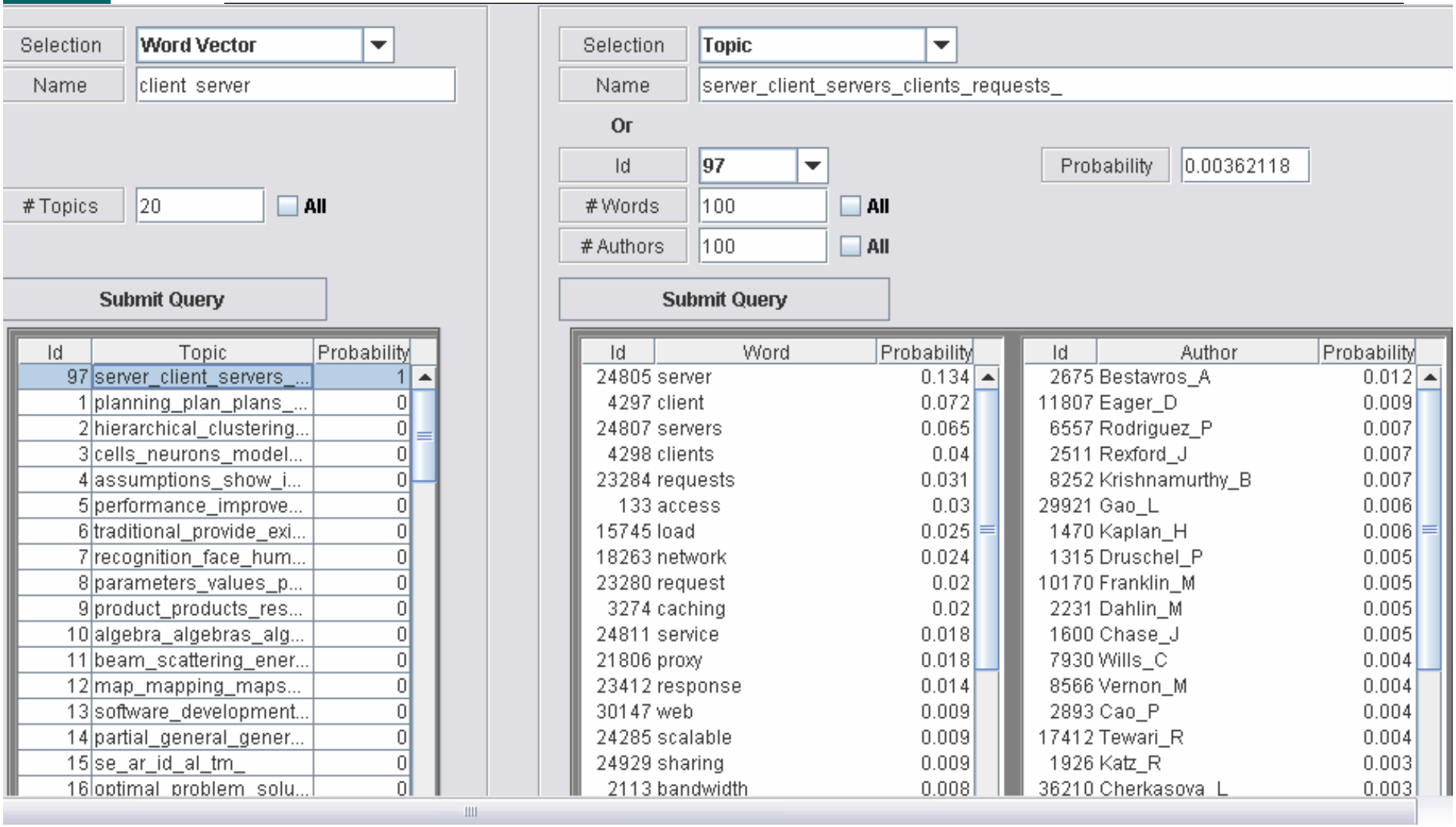
Literature
- MeSH Headings
- PMC Articles
- PubMed via MeSH

Undiscovered Public Knowledge

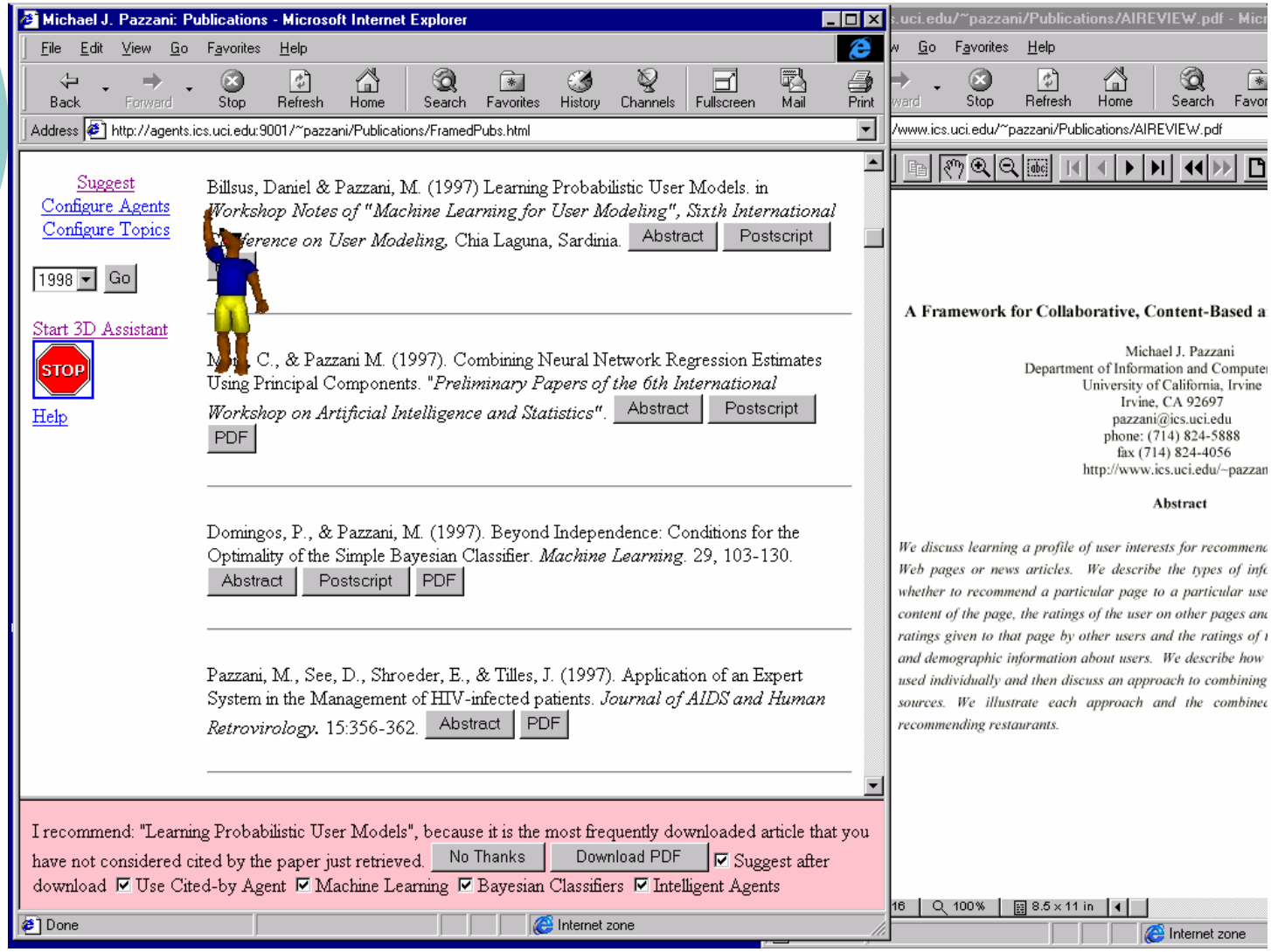
- Can a M.D. or Ph.D keep up with 30,000 articles a year?

- 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

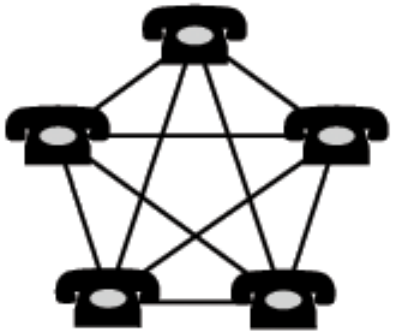


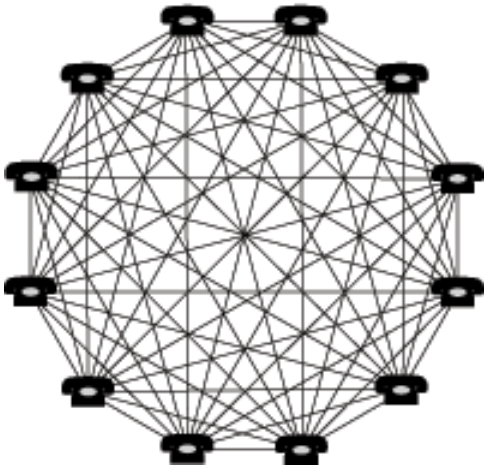
Recommend publications:
cited-by, cites, frequently downloaded, similarity, downloaded together



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 internatic 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 Comput 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 We 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 or 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 c 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 internatio 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