10-2-2009

(Re-) Discovering Lost Web Pages

Martin Klein
Old Dominion University

Michael L. Nelson
Old Dominion University, mnelson@odu.edu

(Re-) Discovering Lost Web Pages

Mathematics & Computer Science Seminar
Emory University
October 2, 2009

Martin Klein & Michael L. Nelson
Department of Computer Science
Old Dominion University
Norfolk VA
www.cs.odu.edu/~{mklein,mln}

The Problem

- Web links “break”
  - 404 http status code -- “not found”
  - “soft 404” -- http server returns “200 OK”, but the resource isn’t really there

- Is the content really gone?
  - Did it just move somewhere else in the web?
  - Is there a copy in search engine caches or web archives?

- To find new or different copies, we need to augment *digital preservation* with *information retrieval* techniques

The Actors

Put a human -- lots of humans -- in the loop
for preservation purposes

The Environment

Web Infrastructure (WI) [McCown07]

- Web search engines (Google, Yahoo, MSN Live) and their caches
- Research Projects (CiteSeer, NSDL)
- Web archives (Internet Archive,Web Base)

INTERNET ARCHIVE
WayBackMachine
Google™
NSDL
THE NATIONAL SCIENCE DIGITAL LIBRARY
YAHOO!®
msn™
CiteSeer.IST
Scientific Literature Digital Library

[illegible]
15122
P. 51

NASA Technical Memorandum 109025 (Revision 1)

A Comparison of Queueing, Cluster and Distributed Computing Systems

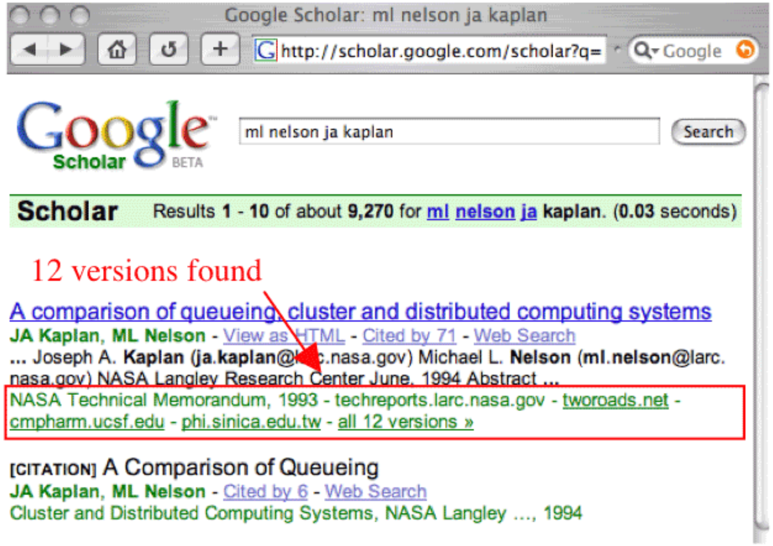
ftp://techreports.larc.nasa.gov/pub/techreports/larc/93/tm109025.ps.Z
http://techreports.larc.nasa.gov/ltrs/PDF/tm109025.pdf

Joseph A. Kaplan and Michael L. Nelson
Langley Research Center, Hampton, Virginia

June 1994

National Aeronautics and
Space Administration
Langley Research Center
Hampton, Virginia 23681-0001

Web Infrastructure: Refreshing & Migrating

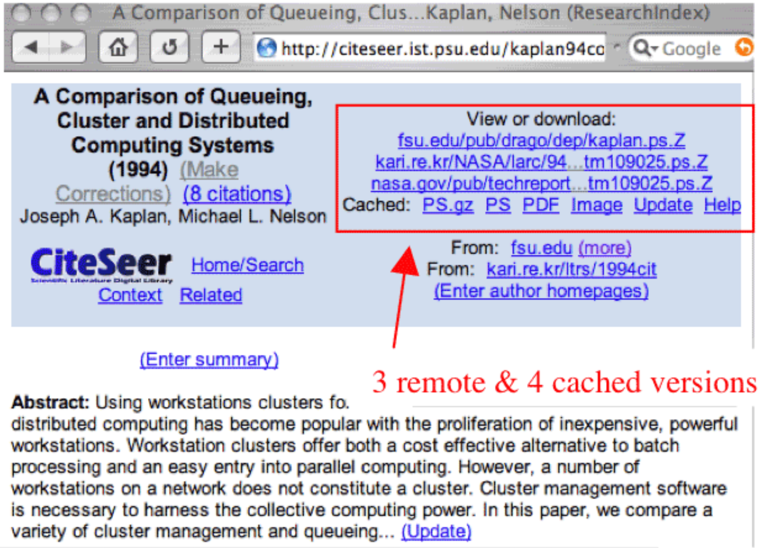

Google Scholar: ml nelson ja kaplan
http://scholar.google.com/scholar?q=
Google
Google Scholar BETA
ml nelson ja kaplan
Search

Scholar Results **1** - **10** of about **9,270** for **ml nelson ja kaplan**. (**0.03** seconds)

12 versions found

A comparison of queueing, cluster and distributed computing systems
JA Kaplan, ML Nelson - View as HTML - Cited by 71 - Web Search
... Joseph A. Kaplan (ja.kaplan@larc.nasa.gov) Michael L. Nelson (ml.nelson@larc.nasa.gov) NASA Langley Research Center June, 1994 Abstract ...
NASA Technical Memorandum, 1993 - techreports.larc.nasa.gov - tworoads.net - cmpharm.ucsf.edu - phi.sinica.edu.tw - all 12 versions »

[CITATION] A Comparison of Queueing
JA Kaplan, ML Nelson - Cited by 6 - Web Search
Cluster and Distributed Computing Systems, NASA Langley ..., 1994


A Comparison of Queueing, Clus...Kaplan, Nelson (ResearchIndex)
http://citeseer.ist.psu.edu/kaplan94co
Google
CiteSeer
Scientific Literature Digital Library

A Comparison of Queueing, Cluster and Distributed Computing Systems (1994) (Make Corrections) (8 citations)
Joseph A. Kaplan, Michael L. Nelson

View or download:
fsu.edu/pub/drago/dep/kaplan.ps.Z
kari.re.kr/NASA/larc/94...tm109025.ps.Z
nasa.gov/pub/techreport...tm109025.ps.Z
Cached: PS.gz PS PDF Image Update Help

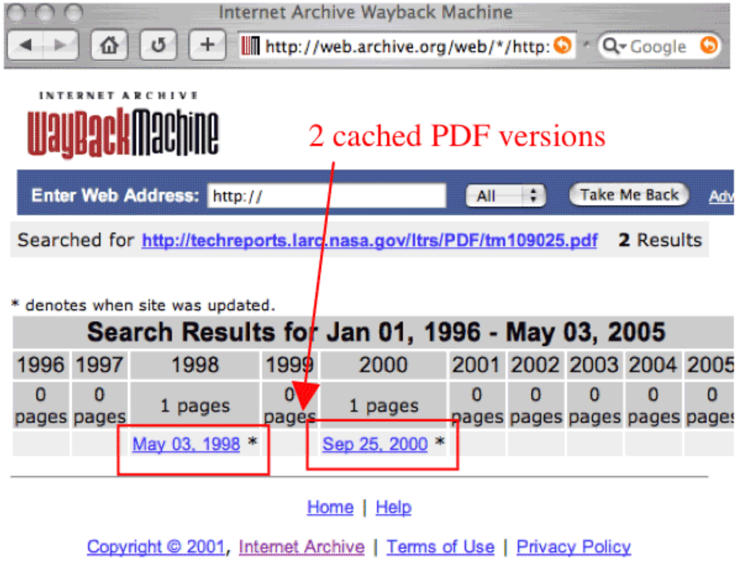
From: fsu.edu (more)
From: kari.re.kr/ltrs/1994cit
(Enter author homepages)

Home/Search
Context Related

(Enter summary)

3 remote & 4 cached versions

Abstract: Using workstations clusters fo. distributed computing has become popular with the proliferation of inexpensive, powerful workstations. Workstation clusters offer both a cost effective alternative to batch processing and an easy entry into parallel computing. However, a number of workstations on a network does not constitute a cluster. Cluster management software is necessary to harness the collective computing power. In this paper, we compare a variety of cluster management and queueing... (Update)



2 cached PDF versions

Searched for http://techreports.larc.nasa.gov/ltrs/PDF/tm109025.pdf **2** Results

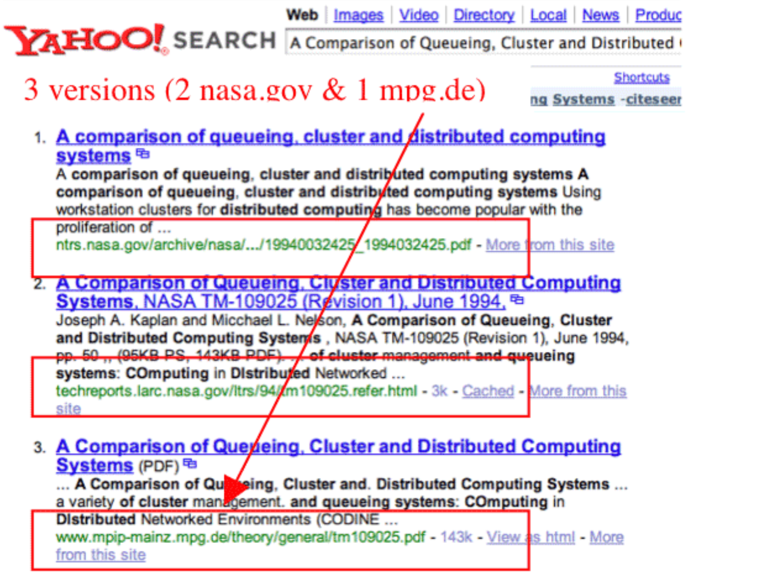
* denotes when site was updated.

Search Results for Jan 01, 1996 - May 03, 2005

1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
0 pages	0 pages	1 pages	0 pages	1 pages	0 pages	0 pages	0 pages	0 pages	0 page
		May 03, 1998 *		Sep 25, 2000 *					

Home | Help

Copyright © 2001, Internet Archive | Terms of Use | Privacy Policy


YAHOO! SEARCH
Web | Images | Video | Directory | Local | News | Produc
A Comparison of Queueing, Cluster and Distributed
Shortcuts
ng Systems -citeseer

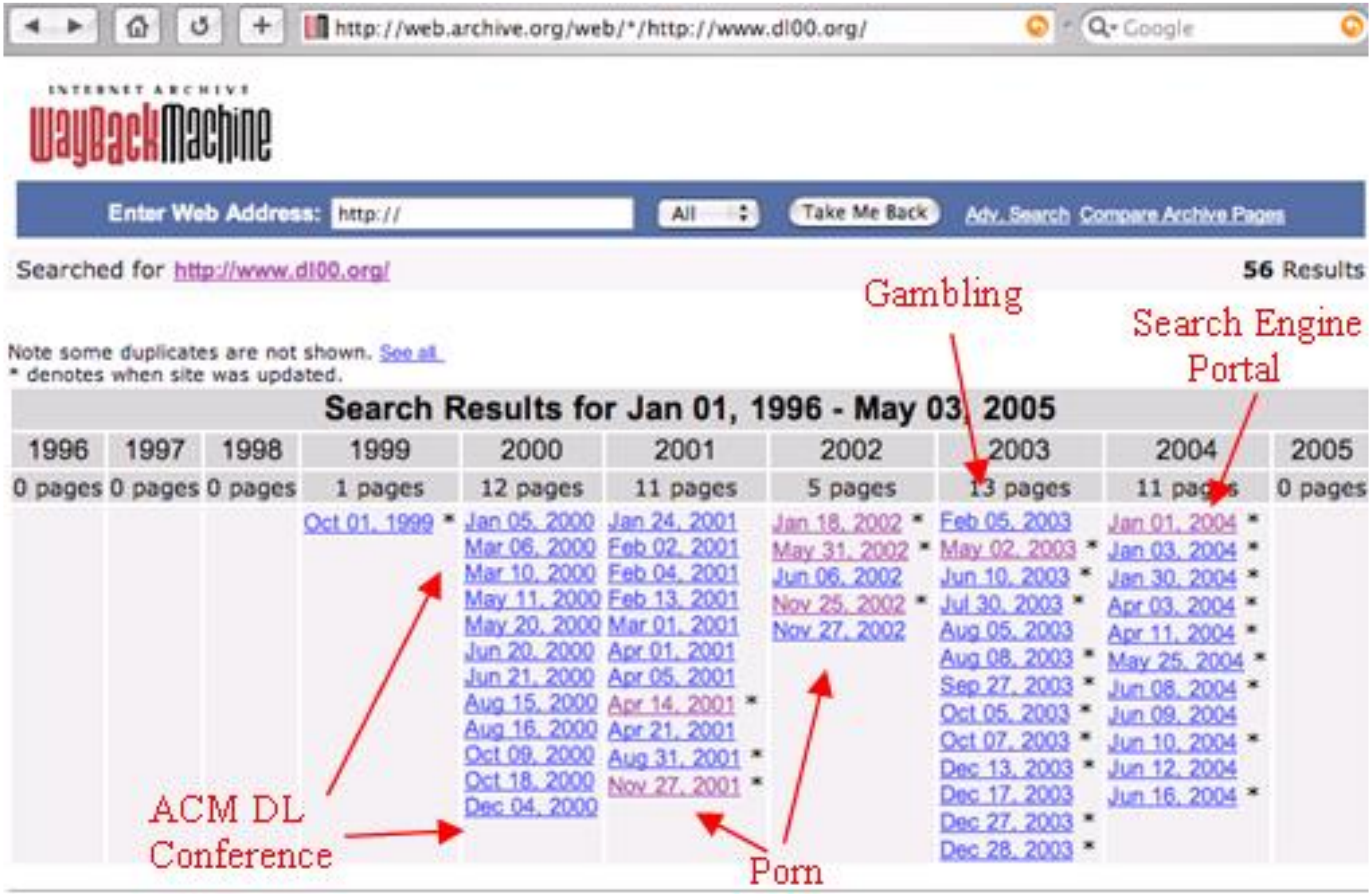
3 versions (2 nasa.gov & 1 mpg.de)

1. A comparison of queueing, cluster and distributed computing systems
A comparison of queueing, cluster and distributed computing systems A comparison of queueing, cluster and distributed computing systems Using workstation clusters for distributed computing has become popular with the proliferation of ...
ntrs.nasa.gov/archive/nasa/.../19940032425_1994032425.pdf - More from this site

2. A Comparison of Queueing, Cluster and Distributed Computing Systems, NASA TM-109025 (Revision 1), June 1994,
Joseph A. Kaplan and Micchael L. Nelson, A Comparison of Queueing, Cluster and Distributed Computing Systems , NASA TM-109025 (Revision 1), June 1994, pp. 50 ,, (95KB PS, 143KB PDF) ... of cluster management and queueing systems: COmputing in DIstributed Networked ...
techreports.larc.nasa.gov/ltrs/94/tm109025.refer.html - 3k - Cached - More from this site

3. A Comparison of Queueing, Cluster and Distributed Computing Systems (PDF)
... A Comparison of Queueing, Cluster and. Distributed Computing Systems ... a variety of cluster management. and queueing systems: COmputing in DIstributed Networked Environments (CODINE ...
www.mpip-mainz.mpg.de/theory/general/tm109025.pdf - 143k - View as html - More from this site

Lapsed Website



Searched for http://www.dl00.org/ **56** Results

Note some duplicates are not shown. See all.
* denotes when site was updated.

Search Results for Jan 01, 1996 - May 03, 2005

1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
0 pages	0 pages	0 pages	1 pages	12 pages	11 pages	5 pages	13 pages	11 pages	0 pages
			Oct 01, 1999 *	Jan 05, 2000	Jan 24, 2001	Jan 18, 2002 *	Feb 05, 2003	Jan 01, 2004 *	
				Mar 06, 2000	Feb 02, 2001	May 31, 2002 *	May 02, 2003 *	Jan 03, 2004 *	
				Mar 10, 2000	Feb 04, 2001	Jun 06, 2002	Jun 10, 2003 *	Jan 30, 2004 *	
				May 11, 2000	Feb 13, 2001	Nov 25, 2002 *	Jul 30, 2003 *	Apr 03, 2004 *	
				May 20, 2000	Mar 01, 2001	Nov 27, 2002	Aug 05, 2003	Apr 11, 2004 *	
				Jun 20, 2000	Apr 01, 2001		Aug 08, 2003 *	May 25, 2004 *	
				Jun 21, 2000	Apr 05, 2001		Sep 27, 2003 *	Jun 08, 2004 *	
				Aug 15, 2000	Apr 14, 2001 *		Oct 05, 2003 *	Jun 09, 2004	
				Aug 16, 2000	Apr 21, 2001		Oct 07, 2003 *	Jun 10, 2004 *	
				Oct 09, 2000	Aug 31, 2001 *		Dec 13, 2003 *	Jun 12, 2004	
				Oct 18, 2000	Nov 27, 2001 *		Dec 17, 2003	Jun 16, 2004 *	
				Dec 04, 2000			Dec 27, 2003 *		
							Dec 28, 2003 *		

Gambling

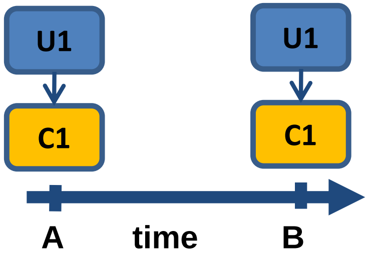
Search Engine Portal

ACM DL Conference

Porn

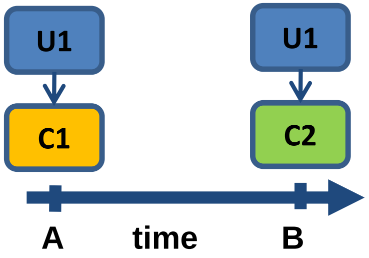
URI Content Mapping Problem

1



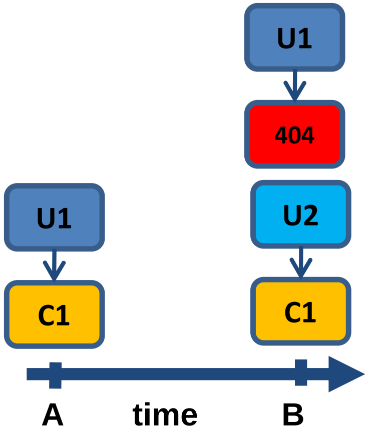
same URI maps to **same** or very similar content at a later time

2



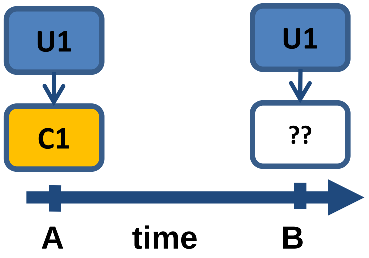
same URI maps to **different** content at a later time

3



different URI maps to **same** or very similar content at the same or at a later time

4

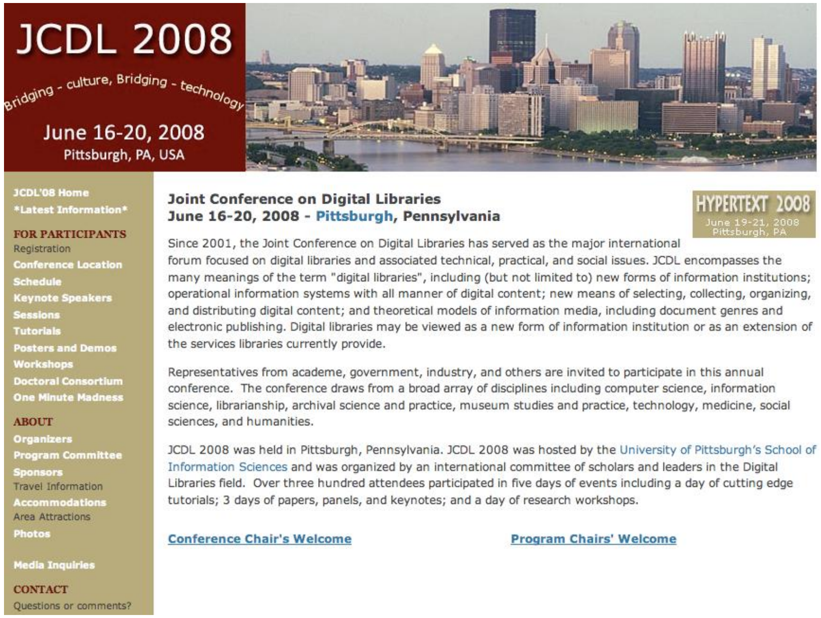


the content can **not be found at any URI**

Scenario 1: Same URI, Same Content

JCDL 2008

http://www.jcdl2008.org/
July 2008



JCDL 2008
Bridging - culture, Bridging - technology
June 16-20, 2008
Pittsburgh, PA, USA

JCDL'08 Home
Latest Information

FOR PARTICIPANTS
Registration
Conference Location
Schedule
Keynote Speakers
Sessions
Tutorials
Posters and Demos
Workshops
Doctoral Consortium
One Minute Madness

ABOUT
Organizers
Program Committee
Sponsors
Travel Information
Accommodations
Area Attractions
Photos

Media Inquiries

CONTACT
Questions or comments?

Joint Conference on Digital Libraries
June 16-20, 2008 - Pittsburgh, Pennsylvania

HYPERTEXT 2008
June 19-21, 2008
Pittsburgh, PA

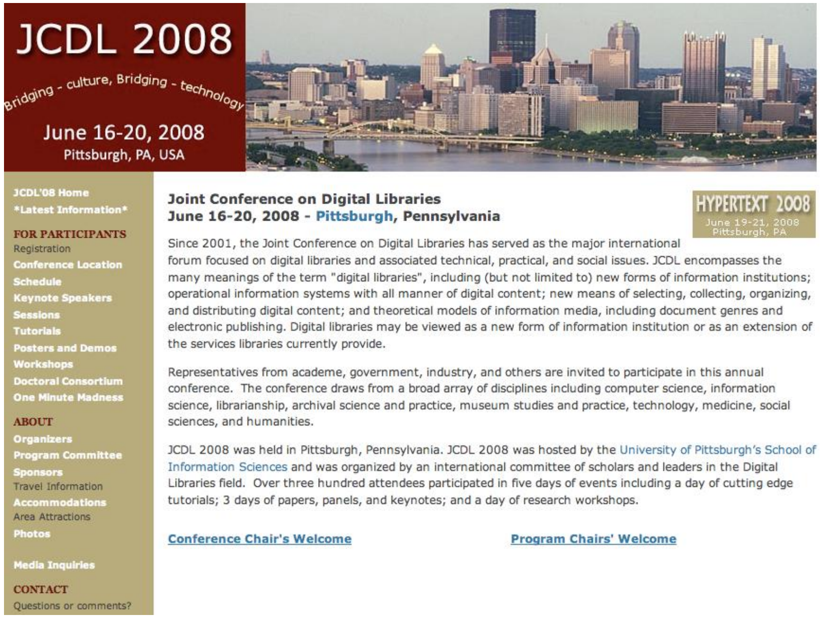
Since 2001, the Joint Conference on Digital Libraries has served as the major international forum focused on digital libraries and associated technical, practical, and social issues. JCDL encompasses the many meanings of the term "digital libraries", including (but not limited to) new forms of information institutions; operational information systems with all manner of digital content; new means of selecting, collecting, organizing, and distributing digital content; and theoretical models of information media, including document genres and electronic publishing. Digital libraries may be viewed as a new form of information institution or as an extension of the services libraries currently provide.

Representatives from academe, government, industry, and others are invited to participate in this annual conference. The conference draws from a broad array of disciplines including computer science, information science, librarianship, archival science and practice, museum studies and practice, technology, medicine, social sciences, and humanities.

JCDL 2008 was held in Pittsburgh, Pennsylvania. JCDL 2008 was hosted by the University of Pittsburgh's School of Information Sciences and was organized by an international committee of scholars and leaders in the Digital Libraries field. Over three hundred attendees participated in five days of events including a day of cutting edge tutorials; 3 days of papers, panels, and keynotes; and a day of research workshops.

Conference Chair's Welcome
Program Chairs' Welcome

http://www.jcdl2008.org/
Today



JCDL 2008
Bridging - culture, Bridging - technology
June 16-20, 2008
Pittsburgh, PA, USA

JCDL'08 Home
Latest Information

FOR PARTICIPANTS
Registration
Conference Location
Schedule
Keynote Speakers
Sessions
Tutorials
Posters and Demos
Workshops
Doctoral Consortium
One Minute Madness

ABOUT
Organizers
Program Committee
Sponsors
Travel Information
Accommodations
Area Attractions
Photos

Media Inquiries

CONTACT
Questions or comments?

Joint Conference on Digital Libraries
June 16-20, 2008 - Pittsburgh, Pennsylvania

HYPERTEXT 2008
June 19-21, 2008
Pittsburgh, PA

Since 2001, the Joint Conference on Digital Libraries has served as the major international forum focused on digital libraries and associated technical, practical, and social issues. JCDL encompasses the many meanings of the term "digital libraries", including (but not limited to) new forms of information institutions; operational information systems with all manner of digital content; new means of selecting, collecting, organizing, and distributing digital content; and theoretical models of information media, including document genres and electronic publishing. Digital libraries may be viewed as a new form of information institution or as an extension of the services libraries currently provide.

Representatives from academe, government, industry, and others are invited to participate in this annual conference. The conference draws from a broad array of disciplines including computer science, information science, librarianship, archival science and practice, museum studies and practice, technology, medicine, social sciences, and humanities.

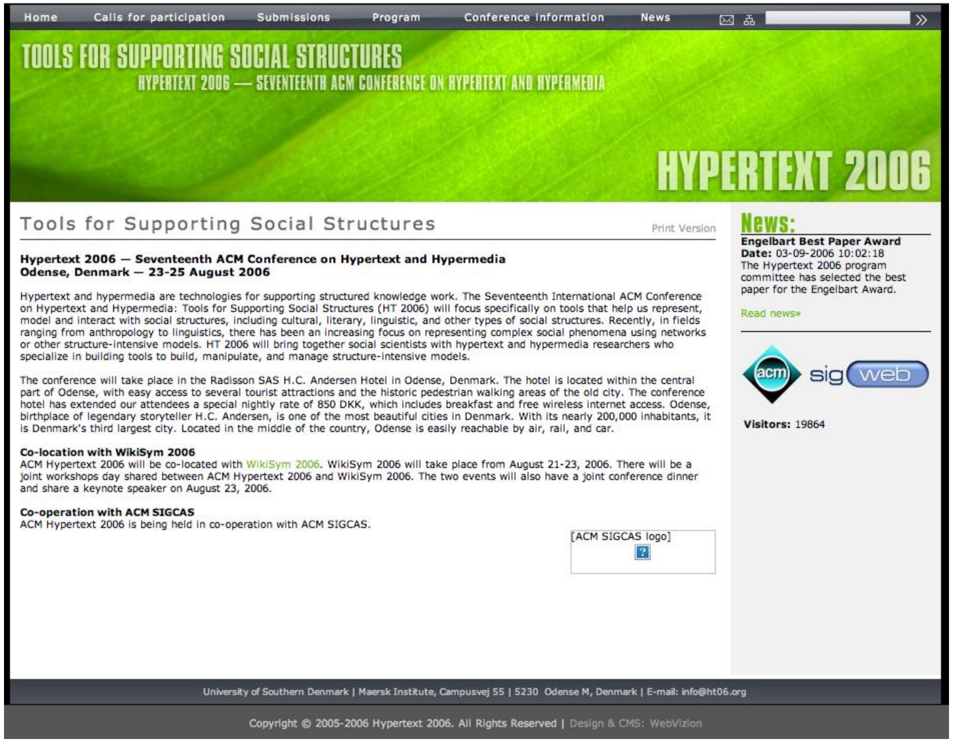
JCDL 2008 was held in Pittsburgh, Pennsylvania. JCDL 2008 was hosted by the University of Pittsburgh's School of Information Sciences and was organized by an international committee of scholars and leaders in the Digital Libraries field. Over three hundred attendees participated in five days of events including a day of cutting edge tutorials; 3 days of papers, panels, and keynotes; and a day of research workshops.

Conference Chair's Welcome
Program Chairs' Welcome

Scenario 2: Same URI, Different Content

Hypertext 2006

http://www.ht06.org/
August 2006



Home | Calls for participation | Submissions | Program | Conference information | News

TOOLS FOR SUPPORTING SOCIAL STRUCTURES
HYPERTEXT 2006 — SEVENTEENTH ACM CONFERENCE ON HYPERTEXT AND HYPERMEDIA
HYPERTEXT 2006

Tools for Supporting Social Structures

Print Version

Hypertext 2006 — Seventeenth ACM Conference on Hypertext and Hypermedia
Odense, Denmark — 23-25 August 2006

Hypertext and hypermedia are technologies for supporting structured knowledge work. The Seventeenth International ACM Conference on Hypertext and Hypermedia: Tools for Supporting Social Structures (HT 2006) will focus specifically on tools that help us represent, model and interact with social structures, including cultural, literary, linguistic, and other types of social structures. Recently, in fields ranging from anthropology to linguistics, there has been an increasing focus on representing complex social phenomena using networks or other structure-intensive models. HT 2006 will bring together social scientists with hypertext and hypermedia researchers who specialize in building tools to build, manipulate, and manage structure-intensive models.

The conference will take place in the Radisson SAS H.C. Andersen Hotel in Odense, Denmark. The hotel is located within the central part of Odense, with easy access to several tourist attractions and the historic pedestrian walking areas of the old city. The conference hotel has extended our attendees a special nightly rate of 850 DKK, which includes breakfast and free wireless internet access. Odense, birthplace of legendary storyteller H.C. Andersen, is one of the most beautiful cities in Denmark. With its nearly 200,000 inhabitants, it is Denmark's third largest city. Located in the middle of the country, Odense is easily reachable by air, rail, and car.

Co-location with WikiSym 2006
ACM Hypertext 2006 will be co-located with WikiSym 2006. WikiSym 2006 will take place from August 21-23, 2006. There will be a joint workshops day shared between ACM Hypertext 2006 and WikiSym 2006. The two events will also have a joint conference dinner and share a keynote speaker on August 23, 2006.

Co-operation with ACM SIGCAS
ACM Hypertext 2006 is being held in co-operation with ACM SIGCAS.

[ACM SIGCAS logo]

News:
Engelbart Best Paper Award
Date: 03-09-2006 10:02:18
The Hypertext 2006 program committee has selected the best paper for the Engelbart Award.

Read news»

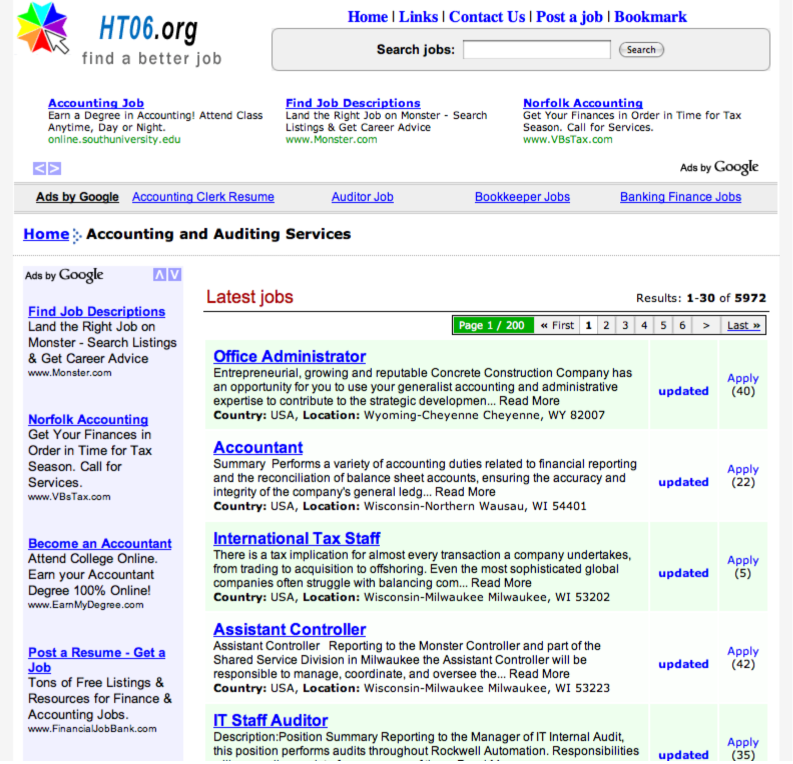
acm sigweb

Visitors: 19864

University of Southern Denmark | Maersk Institute, Campusvej 55 | 5230 Odense M, Denmark | E-mail: info@ht06.org

Copyright © 2005-2006 Hypertext 2006. All Rights Reserved | Design & CMS: WebVizion

http://www.ht06.org/
Today



HT06.org
find a better job

Home | Links | Contact Us | Post a job | Bookmark

Search jobs: Search

Accounting Job
Earn a Degree in Accounting! Attend Class Anytime, Day or Night.
online.southuniversity.edu

Find Job Descriptions
Land the Right Job on Monster - Search Listings & Get Career Advice
www.Monster.com

Norfolk Accounting
Get Your Finances in Order in Time for Tax Season. Call for Services.
www.VBsTax.com

Ads by Google

Ads by Google | Accounting Clerk Resume | Auditor Job | Bookkeeper Jobs | Banking Finance Jobs

Home : **Accounting and Auditing Services**

Ads by Google

Find Job Descriptions
Land the Right Job on Monster - Search Listings & Get Career Advice
www.Monster.com

Norfolk Accounting
Get Your Finances in Order in Time for Tax Season. Call for Services.
www.VBsTax.com

Become an Accountant
Attend College Online. Earn your Accountant Degree 100% Online!
www.EarnMyDegree.com

Post a Resume - Get a Job
Tons of Free Listings & Resources for Finance & Accounting Jobs.
www.FinancialJobBank.com

Latest jobs

Results: **1-30** of **5972**

Page 1 / 200 | « First | 1 | 2 | 3 | 4 | 5 | 6 | > | Last »

Office Administrator Entrepreneurial, growing and reputable Concrete Construction Company has an opportunity for you to use your generalist accounting and administrative expertise to contribute to the strategic developmen... Read More **Country:** USA, **Location:** Wyoming-Cheyenne Cheyenne, WY 82007	updated	Apply (40)
Accountant Summary Performs a variety of accounting duties related to financial reporting and the reconciliation of balance sheet accounts, ensuring the accuracy and integrity of the company's general ledg... Read More **Country:** USA, **Location:** Wisconsin-Northern Wausau, WI 54401	updated	Apply (22)
International Tax Staff There is a tax implication for almost every transaction a company undertakes, from trading to acquisition to offshoring. Even the most sophisticated global companies often struggle with balancing com... Read More **Country:** USA, **Location:** Wisconsin-Milwaukee Milwaukee, WI 53202	updated	Apply (5)
Assistant Controller Assistant Controller Reporting to the Monster Controller and part of the Shared Service Division in Milwaukee the Assistant Controller will be responsible to manage, coordinate, and oversee the... Read More **Country:** USA, **Location:** Wisconsin-Milwaukee Milwaukee, WI 53223	updated	Apply (42)
IT Staff Auditor Description:Position Summary Reporting to the Manager of IT Internal Audit, this position performs audits throughout Rockwell Automation. Responsibilities	updated	Apply (35)

Scenario 3a: Same Content, Different URI

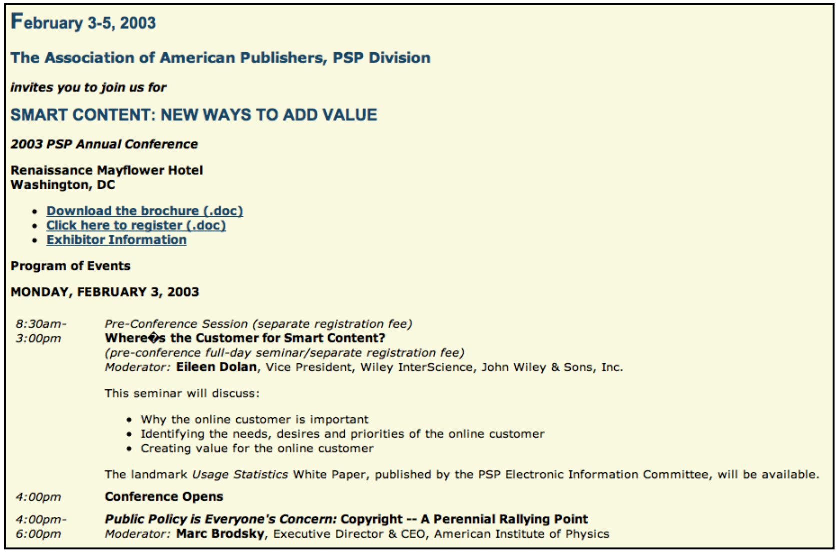
PSP 2003

http://www.pspcentral.org/events/annual_meeting_2003.html

August 2003

http://www.pspcentral.org/events/archive/annual_meeting_2003.html

Today



February 3-5, 2003

The Association of American Publishers, PSP Division

invites you to join us for

SMART CONTENT: NEW WAYS TO ADD VALUE

2003 PSP Annual Conference

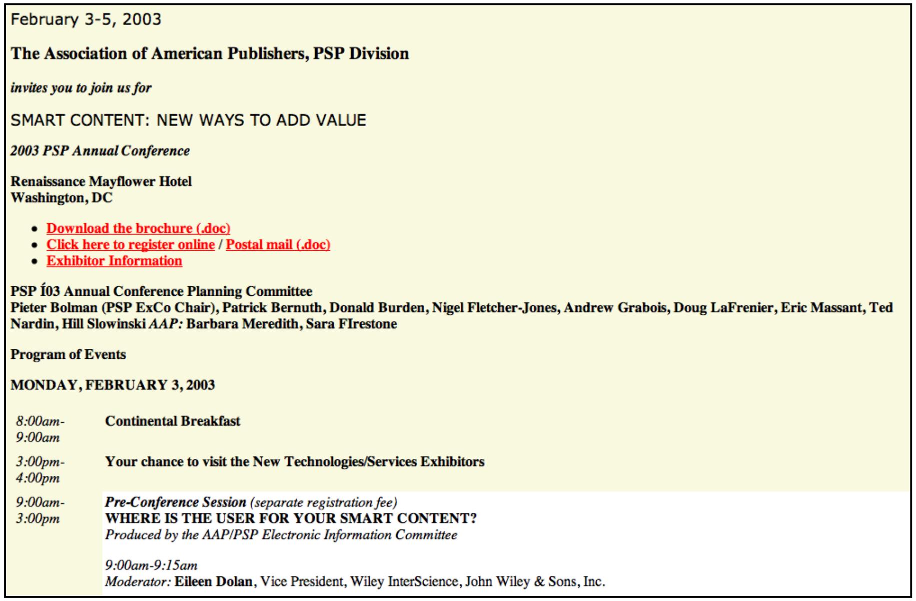
Renaissance Mayflower Hotel
Washington, DC

- Download the brochure (.doc)
- Click here to register (.doc)
- Exhibitor Information

Program of Events

MONDAY, FEBRUARY 3, 2003

8:30am- *3:00pm*	*Pre-Conference Session (separate registration fee)* **Where�s the Customer for Smart Content?** *(pre-conference full-day seminar/separate registration fee)* *Moderator:* **Eileen Dolan**, Vice President, Wiley InterScience, John Wiley & Sons, Inc. This seminar will discuss: • Why the online customer is important • Identifying the needs, desires and priorities of the online customer • Creating value for the online customer The landmark *Usage Statistics* White Paper, published by the PSP Electronic Information Committee, will be available.
4:00pm	**Conference Opens**
4:00pm- *6:00pm*	***Public Policy is Everyone's Concern:* Copyright -- A Perennial Rallying Point** *Moderator:* **Marc Brodsky**, Executive Director & CEO, American Institute of Physics



February 3-5, 2003

The Association of American Publishers, PSP Division

invites you to join us for

SMART CONTENT: NEW WAYS TO ADD VALUE

2003 PSP Annual Conference

Renaissance Mayflower Hotel
Washington, DC

- Download the brochure (.doc)
- Click here to register online / Postal mail (.doc)
- Exhibitor Information

PSP Í03 Annual Conference Planning Committee
Pieter Bolman (PSP ExCo Chair), Patrick Bernuth, Donald Burden, Nigel Fletcher-Jones, Andrew Grabois, Doug LaFrenier, Eric Massant, Ted Nardin, Hill Slowinski *AAP:* Barbara Meredith, Sara FIrestone

Program of Events

MONDAY, FEBRUARY 3, 2003

8:00am- *9:00am*	**Continental Breakfast**
3:00pm- *4:00pm*	**Your chance to visit the New Technologies/Services Exhibitors**
9:00am- *3:00pm*	***Pre-Conference Session*** *(separate registration fee)* **WHERE IS THE USER FOR YOUR SMART CONTENT?** *Produced by the AAP/PSP Electronic Information Committee* *9:00am-9:15am* *Moderator:* **Eileen Dolan**, Vice President, Wiley InterScience, John Wiley & Sons, Inc.

Scenario 3b: Similar Content, Different URI

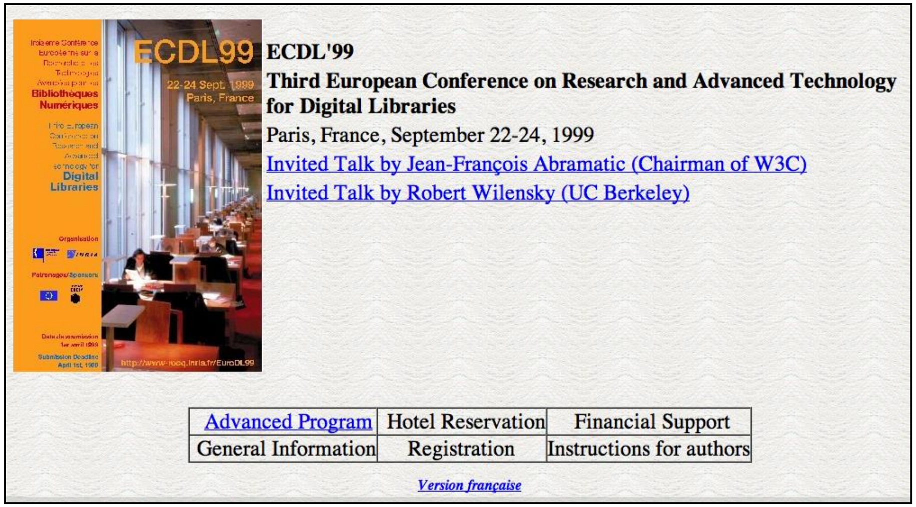
ECDL 1999

http://www-rocq.inria.fr/EuroDL99/

October 1999

http://www.informatik.uni-trier.de/~ley/db/conf/ercimdl/ercimdl99.html

Today



ECDL'99

Third European Conference on Research and Advanced Technology for Digital Libraries

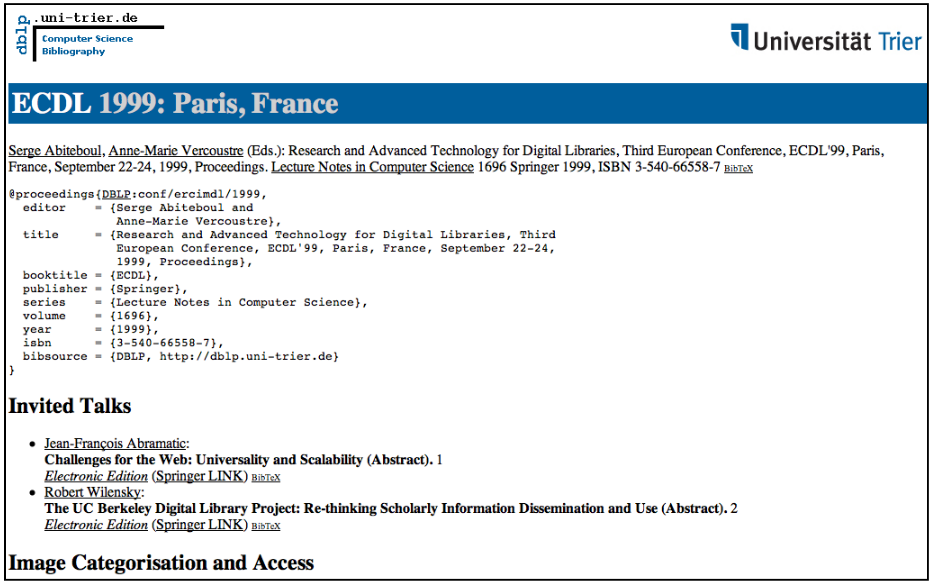
Paris, France, September 22-24, 1999

Invited Talk by Jean-François Abramatic (Chairman of W3C)

Invited Talk by Robert Wilensky (UC Berkeley)

Advanced Program	Hotel Reservation	Financial Support
General Information	Registration	Instructions for authors

Version française



ECDL 1999: Paris, France

Serge Abiteboul, Anne-Marie Vercoustre (Eds.): Research and Advanced Technology for Digital Libraries, Third European Conference, ECDL'99, Paris, France, September 22-24, 1999, Proceedings. Lecture Notes in Computer Science 1696 Springer 1999, ISBN 3-540-66558-7 BibTeX

```
@proceedings{DBLP:conf/ercimdl/1999,
  editor    = {Serge Abiteboul and
               Anne-Marie Vercoustre},
  title     = {Research and Advanced Technology for Digital Libraries, Third
               European Conference, ECDL'99, Paris, France, September 22-24,
               1999, Proceedings},
  booktitle = {ECDL},
  publisher = {Springer},
  series    = {Lecture Notes in Computer Science},
  volume    = {1696},
  year      = {1999},
  isbn      = {3-540-66558-7},
  bibsource = {DBLP, http://dblp.uni-trier.de}
}
```

Invited Talks

- Jean-François Abramatic:
  Challenges for the Web: Universality and Scalability (Abstract). 1
  Electronic Edition (Springer LINK) BibTeX
- Robert Wilensky:
  The UC Berkeley Digital Library Project: Re-thinking Scholarly Information Dissemination and Use (Abstract). 2
  Electronic Edition (Springer LINK) BibTeX

Image Categorisation and Access

Scenario 4: Content Not Findable At Any URI

Greynet 1999

http://www.konbib.nl/infolev/greynet/2.5.htm
1999

?

Today

?

Miller: A lot o' people don't realize what's really going on. They view life as a bunch o' unconnected incidents 'n things. They don't realize that there's this, like, lattice o' coincidence that lays on top o' everything. Give you an example; show you what I mean: suppose you're thinkin' about a plate o' shrimp. Suddenly someone'll say, like, *plate, or shrimp, or plate o' shrimp* out of the blue, no explanation. No point in lookin' for one, either. It's all part of a cosmic unconsciousness.

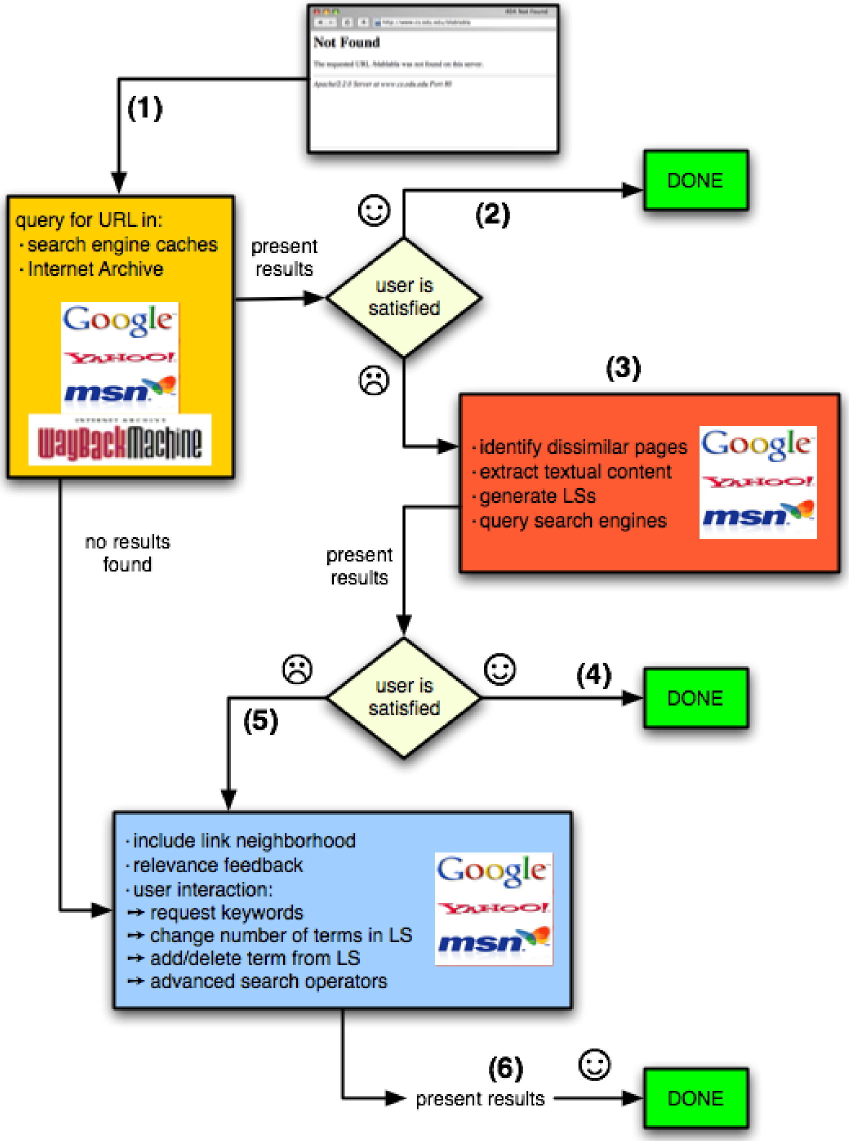
Otto: You eat a lot of acid, Miller, back in the hippie days?

Synchronicity

- Experience of causally unrelated events occurring together in a meaningful manner
- Events reveal underlying pattern, framework bigger than any of the synchronous systems
- Carl Gustav Jung (1875-1961)
    - "meaningful coincidence"

picture from http://www.crystalinks.com/jung.html

Synchronicity Architecture



- Firefox extension catches 404 error (or initiated by user if a “soft” 404 is suspected)

- Discovers copy of missing page in WI (1) and provides to user (2)

- Generates a search engine query based on what the missing page is “about” (3)

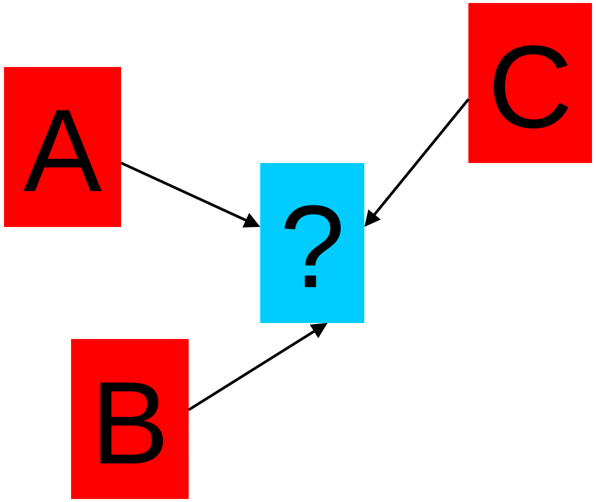
- Finds old content at new URI or provides a “good enough” alternative page (4,5,6)

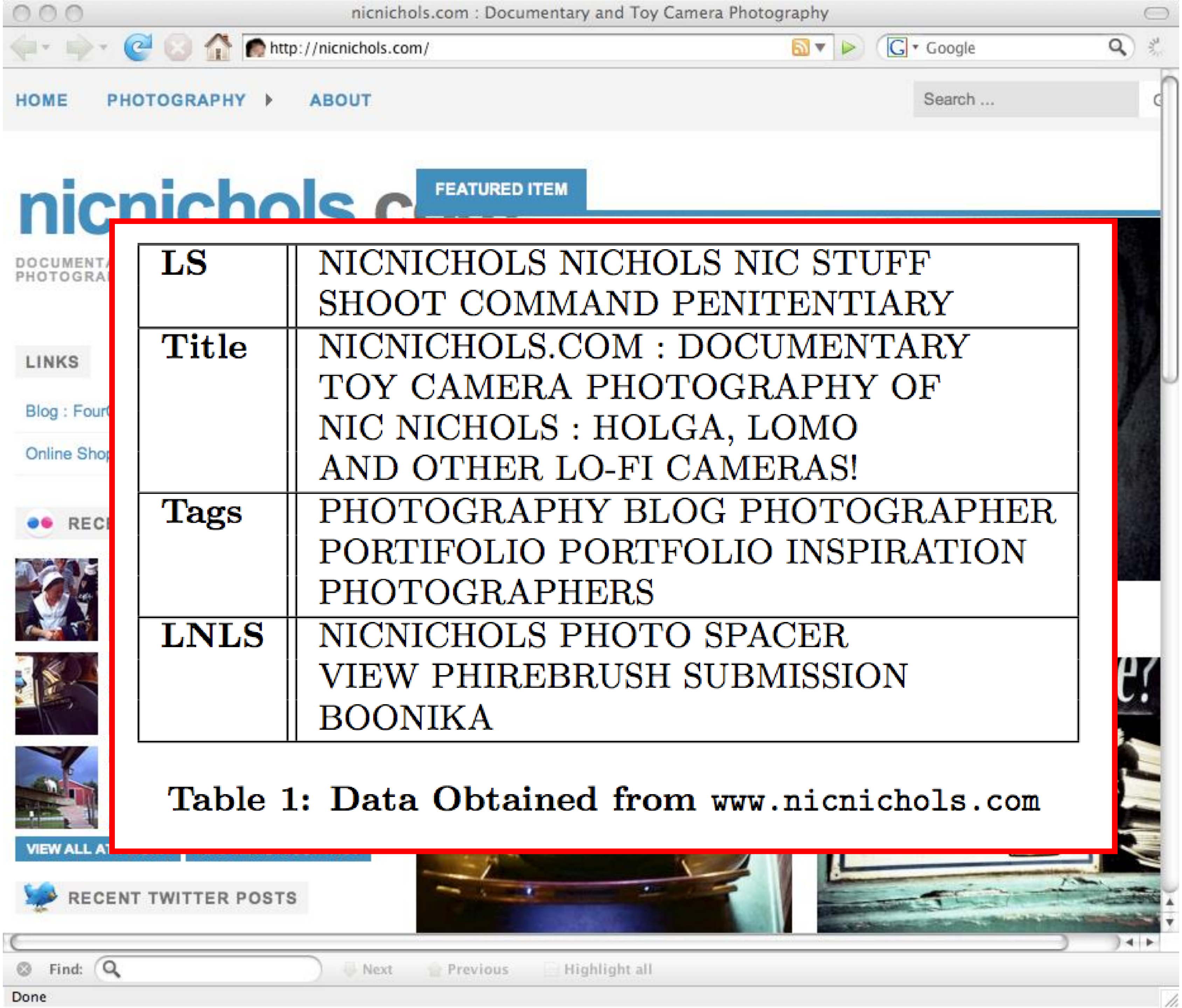
What Was That Web Page About?

- If an “old” copy can be found:
  - Lexical Signatures
  - <title>…</title>
- If no archived/cached copy:
  - Tags
  - Link Neighborhoods; LSs, anchor tags

```
GET https://user:pass@api.del.icio.us/v1/posts/suggest?url=http://yahoo.com/

<?xml version="1.0" encoding="UTF-8"?>
<suggest>
  <popular>web</popular>
  <popular>tools</popular>
  <popular>searchengines</popular>
  <recommended>yahoo!</recommended>
  <recommended>yahoo</recommended>
  <recommended>web</recommended>
  <recommended>tools</recommended>
  <recommended>search</recommended>
  <recommended>reference</recommended>
  <recommended>portal</recommended>
  <recommended>news</recommended>
</suggest>
```





LS	NICNICHOLS NICHOLS NIC STUFF SHOOT COMMAND PENITENTIARY
Title	NICNICHOLS.COM : DOCUMENTARY TOY CAMERA PHOTOGRAPHY OF NIC NICHOLS : HOLGA, LOMO AND OTHER LO-FI CAMERAS!
Tags	PHOTOGRAPHY BLOG PHOTOGRAPHER PORTIFOLIO PORTFOLIO INSPIRATION PHOTOGRAPHERS
LNLS	NICNICHOLS PHOTO SPACER VIEW PHIREBRUSH SUBMISSION BOONIKA

Table 1: Data Obtained from `www.nicnichols.com`

What is a Signature?

(aka “message digest”, examples include “md5” and “sha-1”)

I LOVE
bhfriszkj2akjw-
igzujk6zjzsg

image from Eddie Kohler http://www.cs.ucla.edu/~kohler/

What is a Lexical Signature?

- First introduced by Phelps and Wilensky [Phelps00]
- Small set of terms capturing the "aboutness" of a document
    - Phelps and Wilensky assumed 5
- "lightweight metadata"

Resource

Removal Policies in Network Caches for World-Wide Web Documents

Stephen Williams, Marc Abrams (Computer Science, Virginia Tech, Blacksburg, VA 24061-0106)

Charles R. Standridge (Industrial Eng., FAMU-FSU College of Eng., Tallahassee, FL 32310)

Ghaleb Abdulla, Edward A. Fox (Computer Science, Virginia Tech, Blacksburg, VA 24061-0106)

{williams,abrams}@cs.vt.edu, stand@eng.fsu.edu, {abdulla,fox}@cs.vt.edu

Revised and Corrected version of Original Publication

Abstract

World-Wide Web proxy servers that cache documents can potentially reduce three quantities: the number of requests that reach popular servers, the volume of network traffic resulting from document requests, and the latency that an end-user experiences in retrieving a document. This paper examines the first two using the measures of cache hit rate (or fraction of client-requested URLs returned by the proxy) and weighted cache hit rate (or fraction of client-requested bytes returned by the proxy). A client request for an uncached document may cause the removal of one or more cached documents. Variable document sizes and types allow a rich variety of policies to select a document for removal, in contrast to policies for CPU caches or demand paging, that manage homogeneous objects. We present a taxonomy of removal policies. Through trace-driven simulation, we determine the maximum possible hit rate and weighted hit rate that a cache could ever achieve, and the removal policy that maximizes hit rate and weighted hit rate. The experiments use five traces of 38 to 190 days of client URL requests. Surprisingly, the criteria used by current proxy-server removal policies (LRU and a proposal by Pitkow and Recker) are among the worst performing criteria in our simulation; instead, replacing documents based on size maximizes hit rate in each of the studied workloads.

1 Introduction

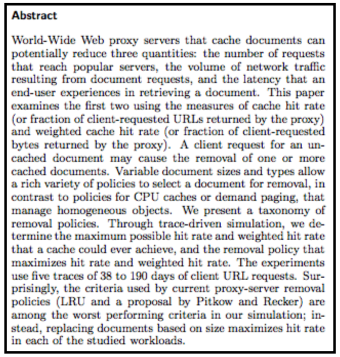
It is ironic that what is arguably the most popular application of the Internet, namely the World-Wide Web (WWW or "Web"), is inherently unscalable. The protocol underlying the Web, HTTP [7], is designed for a setting in which users access a large body of documents, each with a low volume of readership. However the Web today is characterized by high volume of access to popular Web pages. Thus identical copies of many documents pass through the same network links. This has several costs: network administrators see growing utilization that requires bandwidth upgrades, Web site administrators see growing server utilization that requires upgrading or replacing servers, and end users see longer latency for document requests to be satisfied.

These problems can be alleviated by widespread migration of copies of popular documents from servers to points closer to users. Migration can follow a distribution model, in which servers control where document copies are stored, or a cache model, in which copies automatically migrate in response to user requests. Distribution is popular with commercial users that want to protect material with copyrights, and will only trust certain sites to keep copies of such material. However the cache model is the most popular method to date: caching at a server (e.g., [10]), caching at a client (e.g., caches built into Web browsers), and caching in the network itself through so-called proxy servers (e.g., [11]). This paper examines the last of these.

In the following, a *document* is any item retrieved by a Universal Resource Locator (URL), such as a dynamically created page or an audio file. We call a document on a Web server an *original*, and a document on a cache a *copy*. Consider a client configured to use proxy P that requests document D from Web server S. The client sends an HTTP GET request message for URL http://S/D to P. Three cases are now possible: (1) P has a copy of D that it estimates is consistent with the original, (2) P has a copy that is considered inconsistent, and (3) P does not have a copy. (Various algorithms not considered here are used to estimate consistency.) In case (1), P serves the local copy; this is a *hit*. In (2), P sends an HTTP conditional GET message to S containing the `Last-Modified` time of its copy; if the original was modified after that time, S replies to P with the new version (a *miss*). Otherwise, P serves the local copy; this is also a *hit*. In (3), P either forwards the GET message to another proxy server (as in [12]) or to S (also a *miss*).

Proxy caches can dramatically reduce network load. For example, in one of the five Web traffic workload traces used in this paper, representing remote client requests for documents on a network backbone in the Computer Science Department at Virginia Tech, 88% of the bytes transferred in a 37 day measurement period were audio. This surprising result arose because a student created what has been recognized in the WWW as the authoritative Web site for a popular British recording artist. Unfortunately, every request worldwide to listen to a song on this Web site consumes

Abstract



Abstract

World-Wide Web proxy servers that cache documents can potentially reduce three quantities: the number of requests that reach popular servers, the volume of network traffic resulting from document requests, and the latency that an end-user experiences in retrieving a document. This paper examines the first two using the measures of cache hit rate (or fraction of client-requested URLs returned by the proxy) and weighted cache hit rate (or fraction of client-requested bytes returned by the proxy). A client request for an uncached document may cause the removal of one or more cached documents. Variable document sizes and types allow a rich variety of policies to select a document for removal, in contrast to policies for CPU caches or demand paging, that manage homogeneous objects. We present a taxonomy of removal policies. Through trace-driven simulation, we determine the maximum possible hit rate and weighted hit rate that a cache could ever achieve, and the removal policy that maximizes hit rate and weighted hit rate. The experiments use five traces of 38 to 190 days of client URL requests. Surprisingly, the criteria used by current proxy-server removal policies (LRU and a proposal by Pitkow and Recker) are among the worst performing criteria in our simulation; instead, replacing documents based on size maximizes hit rate in each of the studied workloads.

LS

REMOVAL
HIT
RATE
PROXY
CACHE

Query Google

"Removal Policies in Network Caches for World-Wide Web Documents"

LSs as Proposed by Phelps and Wilensky

- “Robust Hyperlink Cost Five Words Each”
- Append LS to URL:

http://www.cs.berkeley.edu/~wilensky/NLP.html
becomes:
http://www.cs.berkeley.edu/~wilensky/NLP.html?lexical-signature=texttiling+wilensky+disambiguation+subtopic+iago

- Limitations:
  1. Applications (browsers) need to be modified to exploit LSs
  2. LSs need to be computed a priori
  3. Works well with most URLs but not with all of them

Lexical Signatures -- Examples

Rank/Results	URL	LS
1/1	**http://www.cs.berkeley.edu/˜wilensky/NLP.html**	**texttiling wilensky disambiguation subtopic iago** http://www.google.com/search?q=texttiling+wilensky+disambiguation+subtopic+iago
1/221,000 (1/174,000 in 01/2008)	**http://www.loc.gov**	**library collections congress thomas american** **http://www.google.com/search?q=library+collections+congress+thomas+american**
1/51 (2/77 in 01/2008)	**http://www.jcdl2008.org**	**libraries jcdl digital conference pst** **http://www.google.com/search?q=libraries+jcdl+digital+conference+pst**
0/10	**http://www.dli2.nsf.gov**	**nsdl multiagency imls testbeds extramural** **http://www.google.com/search?q=nsdl+multiagency+imls+testbeds+extramural**

A “Googlewhack” (http://en.wikipedia.org/wiki/Googlewhack) can be thought of as a two-term LS that produces a 1/1 ranking.

Generating LSs

- Term Frequency (TF)
  - “How often does this term occur in this document?”

- Inverse Document Frequency (IDF)
  - “In how many documents does this term appear?”

$$TF_{ij} \;=\; \frac{f_{ij}}{m_i}$$
$$f_{ij} = freq\ of\ j\ in\ i$$
$$m_i = max\ freq\ in\ i$$

$$IDF_j \;=\; \log\left(\frac{N}{n_j}\right) + 1$$
$$N = total\ number\ of\ documents$$
$$n_j = number\ of\ documents\ j\ occurs\ in$$

Generating LSs

- Park et al. [Park03] investigated performance of various LS generation algorithms

- Evaluated “tunability” of TF and IDF
  - Weight on TF increases recall (completeness, ex. “photography, blog”)
  - Weight on IDF improves precision (exactness, ex. “nicnichols, penitentiary”)

- Computed IDF on closed system (not live web)

- Also assumed “5” to be a good number

- Compared results after 6 months, but did not do an in-depth analysis of LSs over time

Theoretical Underpinnings of Synchronicity

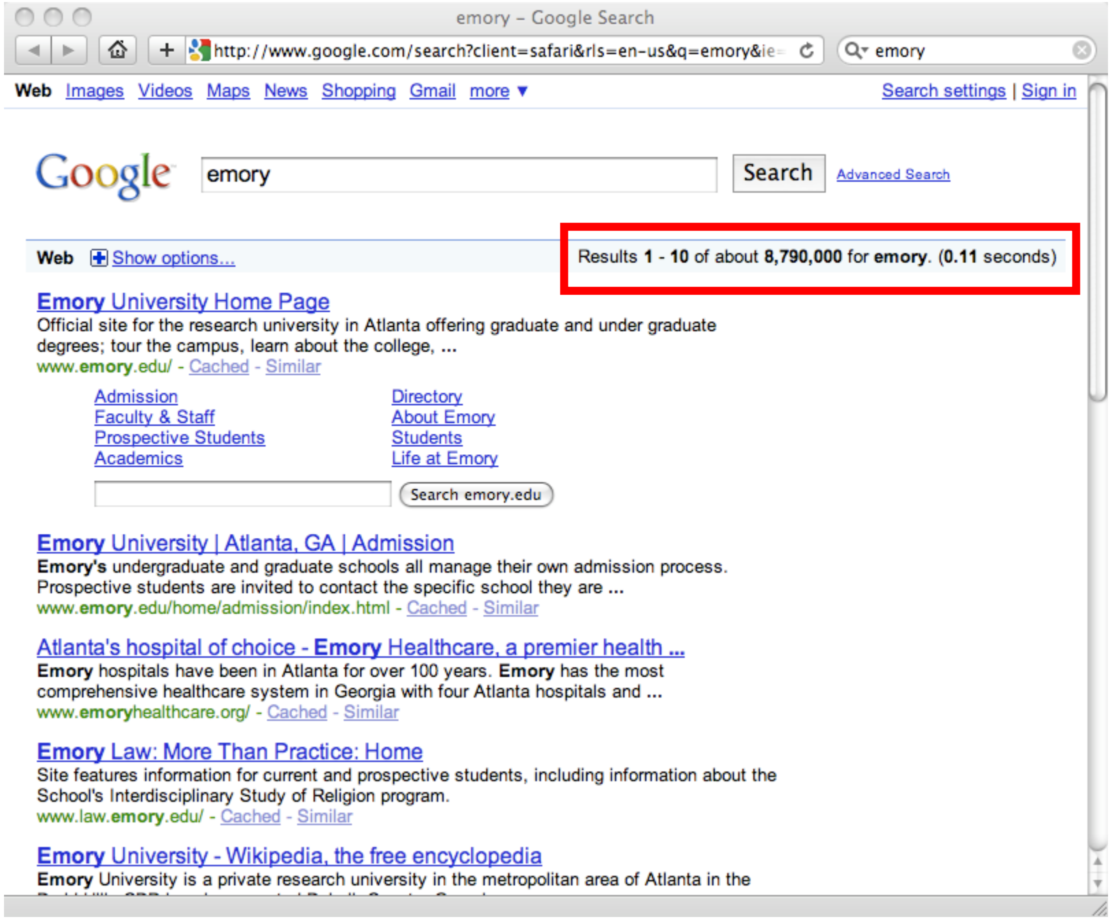
- Estimating IDF values for the Web (WIDM 2008, ECIR 2009)

- Investigated how lexical signatures change over time (ECDL 2008)

- Compared retrieval performance of lexical signatures with titles, tags and lexical signatures generated from link neighborhoods (submitted)

- Investigated how titles change over time (InDP 2009, in preparation)

Hacks for Estimating IDF


emory – Google Search
http://www.google.com/search?client=safari&rls=en-us&q=emory&ie=
emory
Web Images Videos Maps News Shopping Gmail more ▾
Search settings | Sign in
Google
emory
Search
Advanced Search
Web
Show options...

Results **1** - **10** of about **8,790,000** for **emory**. (**0.11** seconds)

Emory University Home Page
Official site for the research university in Atlanta offering graduate and under graduate degrees; tour the campus, learn about the college, ...
www.**emory**.edu/ - Cached - Similar

Admission | Directory
Faculty & Staff | About Emory
Prospective Students | Students
Academics | Life at Emory

Emory University | Atlanta, GA | Admission
Emory's undergraduate and graduate schools all manage their own admission process. Prospective students are invited to contact the specific school they are ...
www.**emory**.edu/home/admission/index.html - Cached - Similar

Atlanta's hospital of choice - **Emory** Healthcare, a premier health ...
Emory hospitals have been in Atlanta for over 100 years. **Emory** has the most comprehensive healthcare system in Georgia with four Atlanta hospitals and ...
www.**emory**healthcare.org/ - Cached - Similar

Emory Law: More Than Practice: Home
Site features information for current and prospective students, including information about the School's Interdisciplinary Study of Religion program.
www.law.**emory**.edu/ - Cached - Similar

Emory University - Wikipedia, the free encyclopedia
Emory University is a private research university in the metropolitan area of Atlanta in the

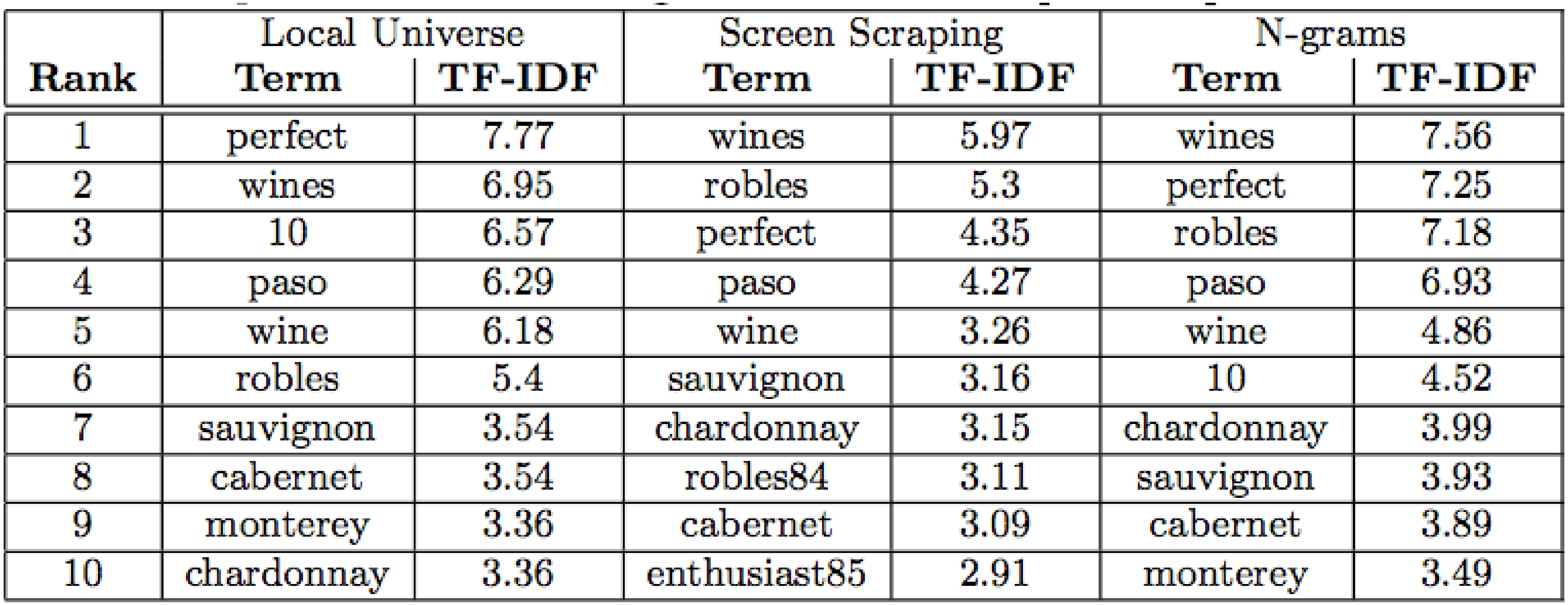
1. everyone knows this value is flaky
2. get N from: http://www.worldwidewebsize.com/

For LS purposes, it doesn’t matter much…

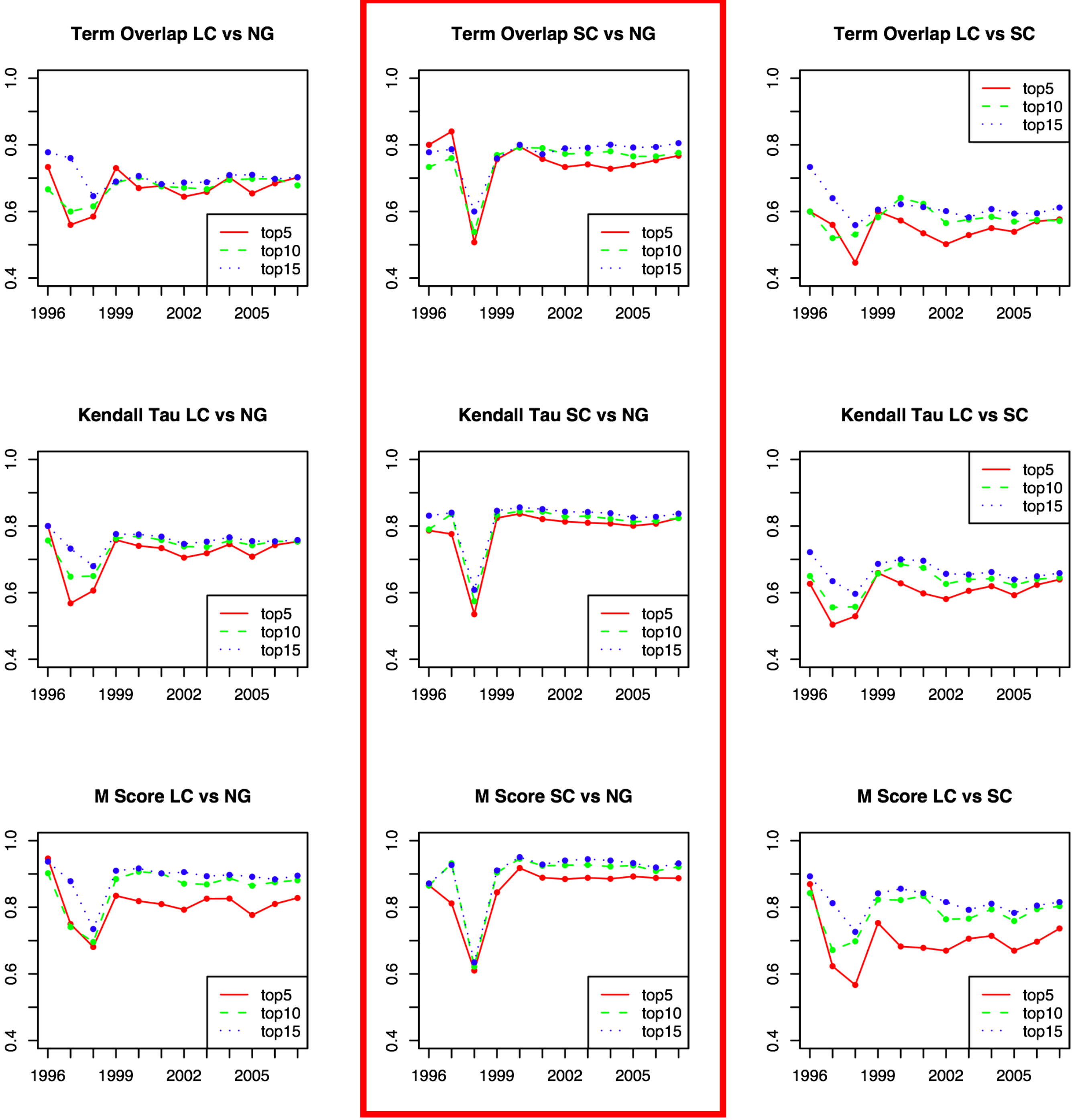
URL: http://www.perfect10wines.com

Year: 2007

Union: 12 unique terms



Rank	Local Universe		Screen Scraping		N-grams	
	Term	**TF-IDF**	**Term**	**TF-IDF**	**Term**	**TF-IDF**
1	perfect	7.77	wines	5.97	wines	7.56
2	wines	6.95	robles	5.3	perfect	7.25
3	10	6.57	perfect	4.35	robles	7.18
4	paso	6.29	paso	4.27	paso	6.93
5	wine	6.18	wine	3.26	wine	4.86
6	robles	5.4	sauvignon	3.16	10	4.52
7	sauvignon	3.54	chardonnay	3.15	chardonnay	3.99
8	cabernet	3.54	robles84	3.11	sauvignon	3.93
9	monterey	3.36	cabernet	3.09	cabernet	3.89
10	chardonnay	3.36	enthusiast85	2.91	monterey	3.49



Comparing LSs

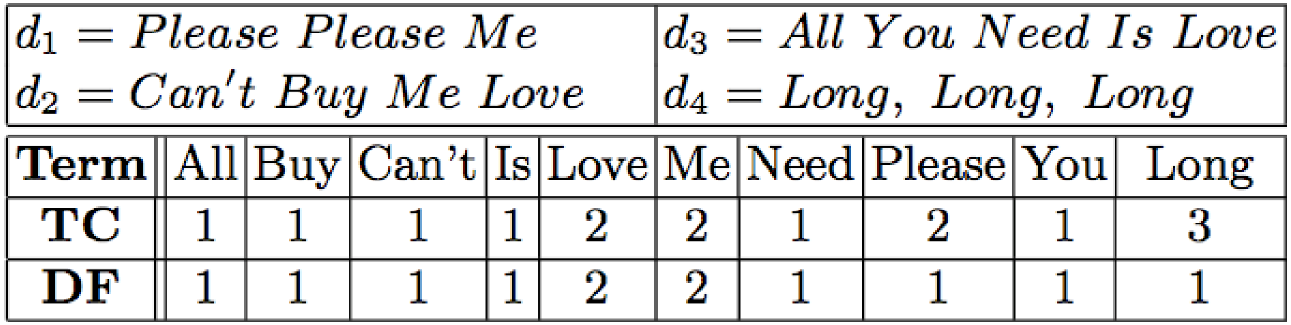
Top 5, 10 and 15 terms

LC – local universe
SC – screen scraping
NG – N-Grams

~4 of 5 LS terms are the same

How Does Google N-grams TC Relate to DF?

- Google N-grams has only Term Count (TC), not Document Frequency
  - where TC >= DF
  - http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html



$d_1 = Please\ Please\ Me$

$d_2 = Can't\ Buy\ Me\ Love$

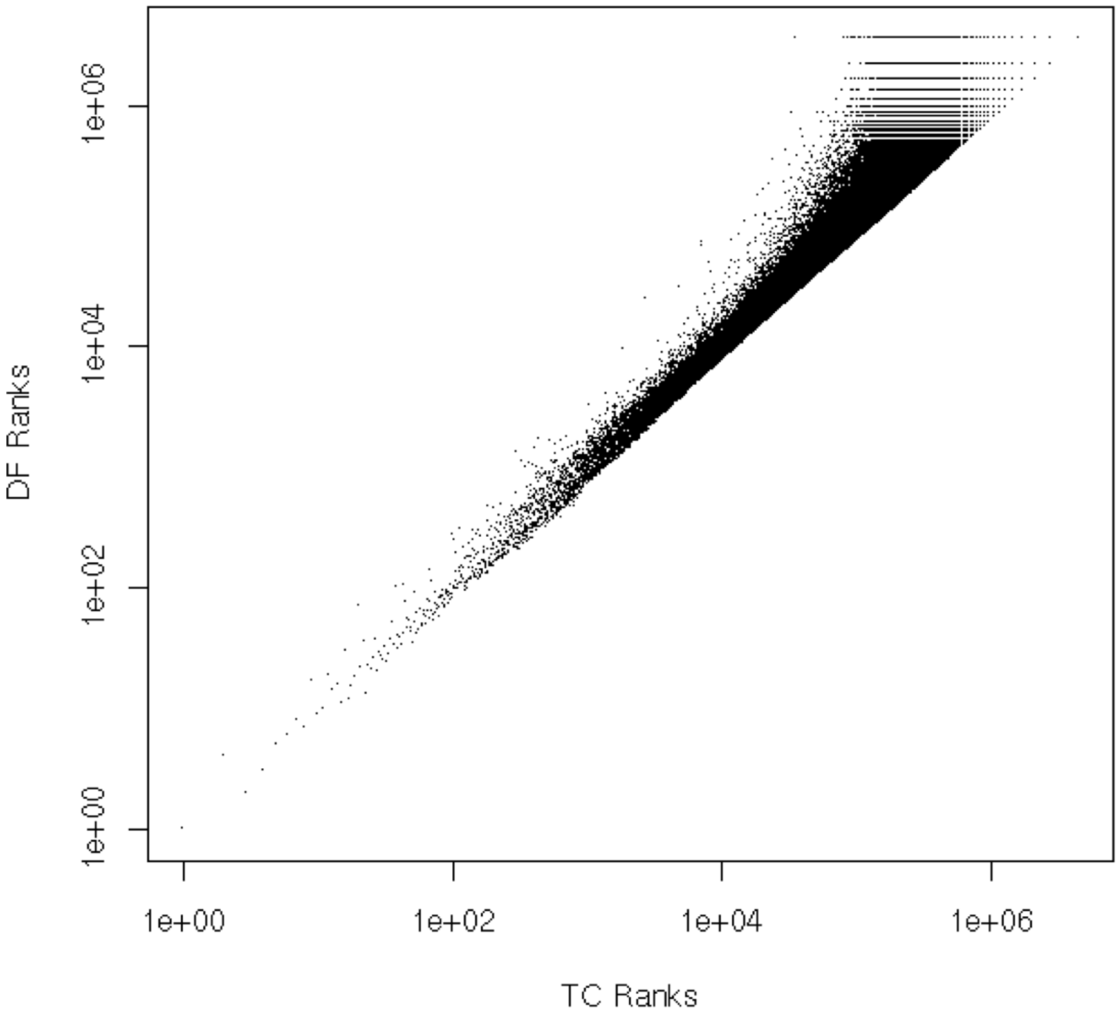
$d_3 = All\ You\ Need\ Is\ Love$

$d_4 = Long,\ Long,\ Long$

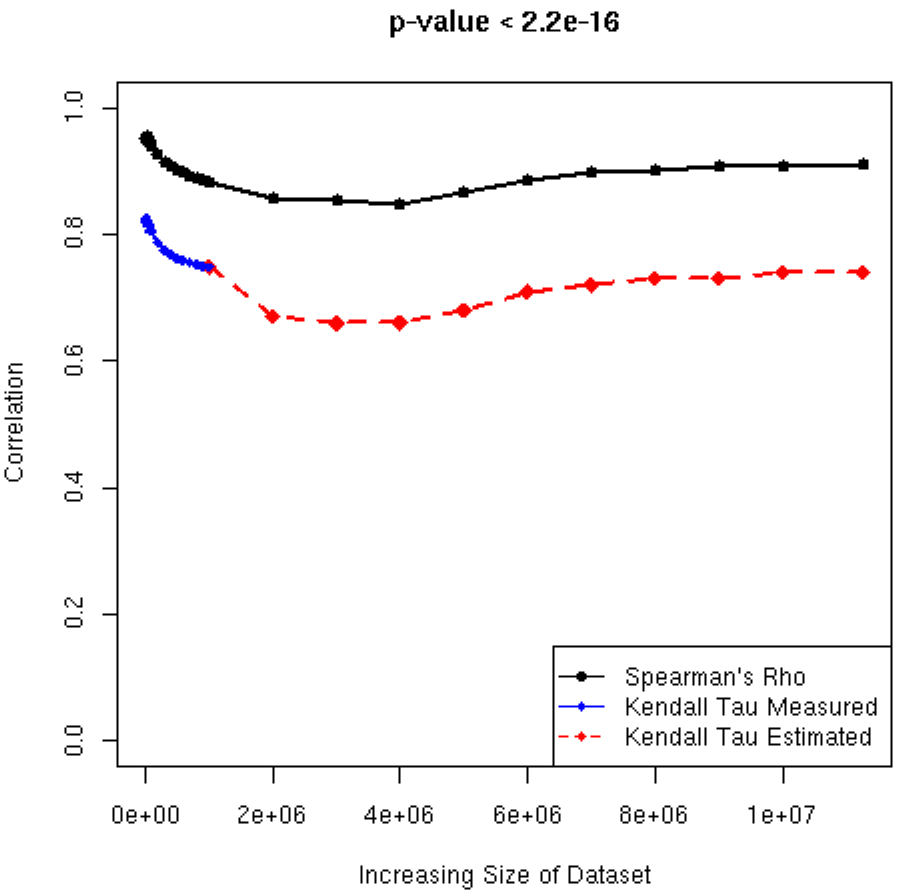
Term	All	Buy	Can't	Is	Love	Me	Need	Please	You	Long
TC	1	1	1	1	2	2	1	2	1	3
DF	1	1	1	1	2	2	1	1	1	1

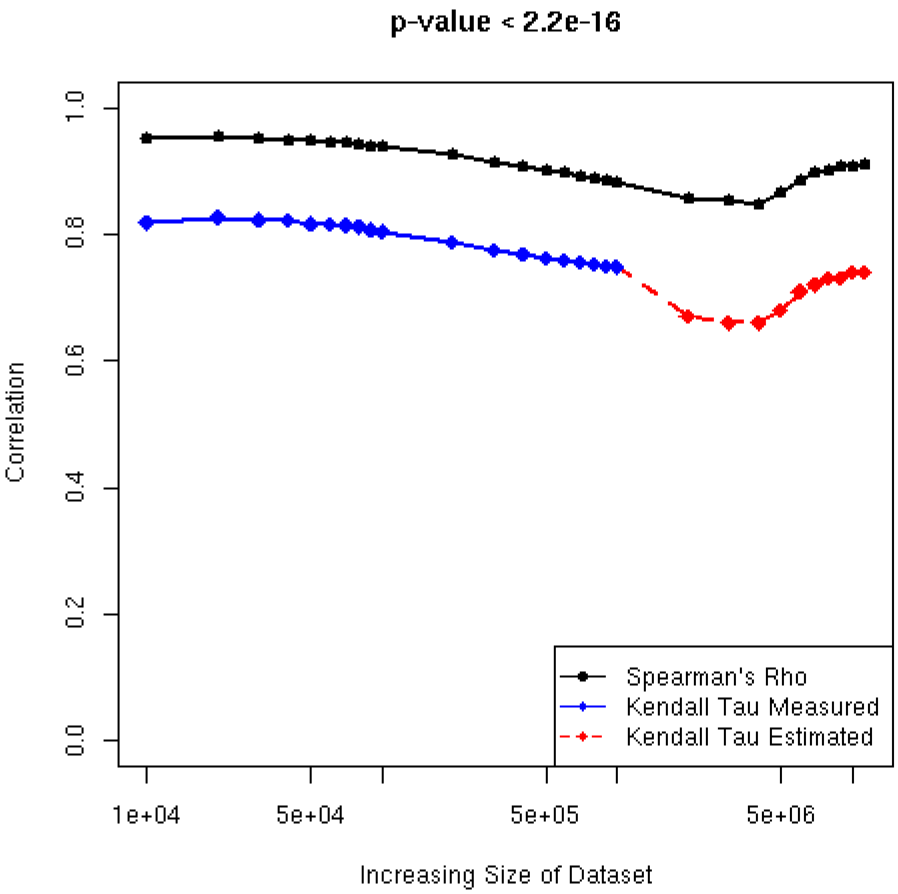
- Idea: compare TC & DF in a known collection, then compare that collection's TC to the Google N-grams TC
  - we used ukWaC, from WaCKy: http://wacky.sslmit.unibo.it/

TC Ranks vs DF Ranks Within ukWaC



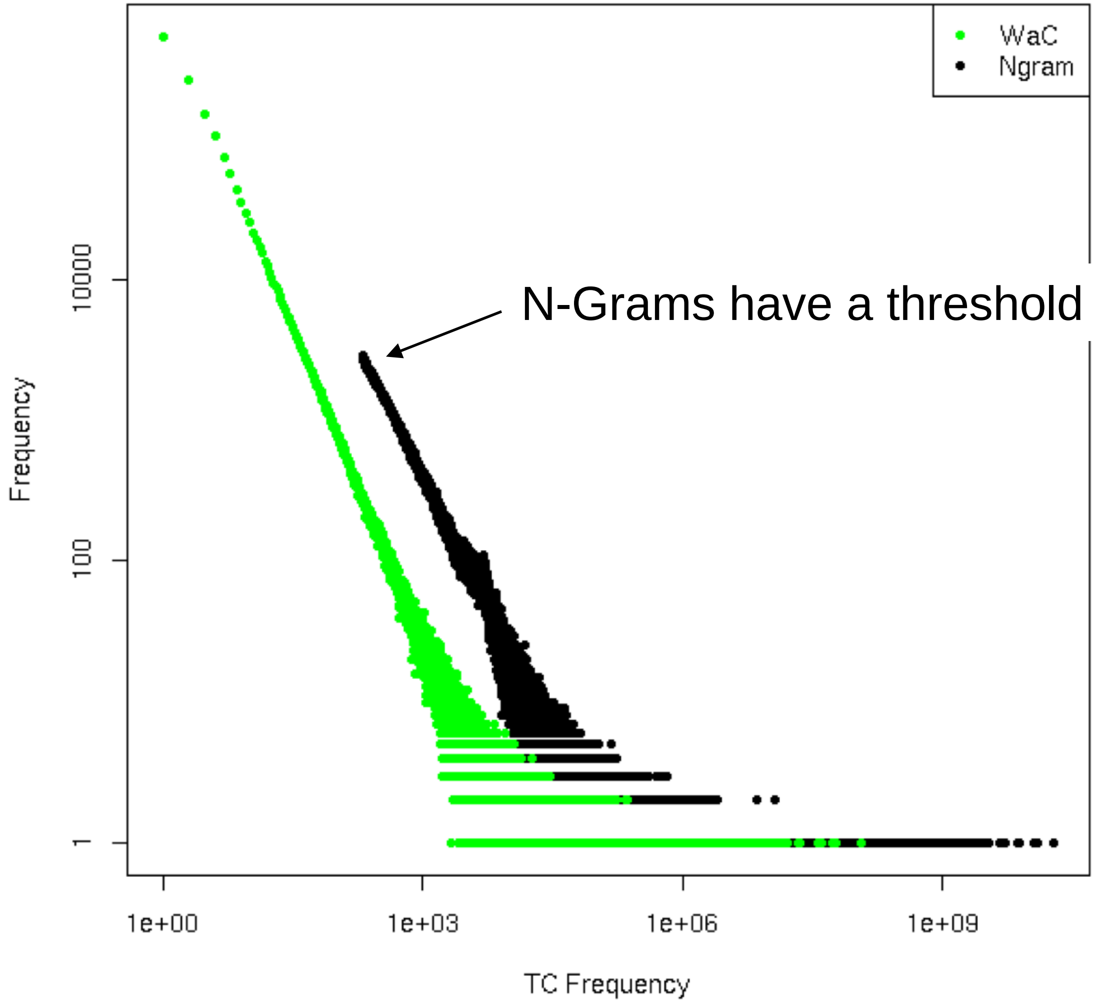
Rank Correlation Within ukWaC





semi-log scale

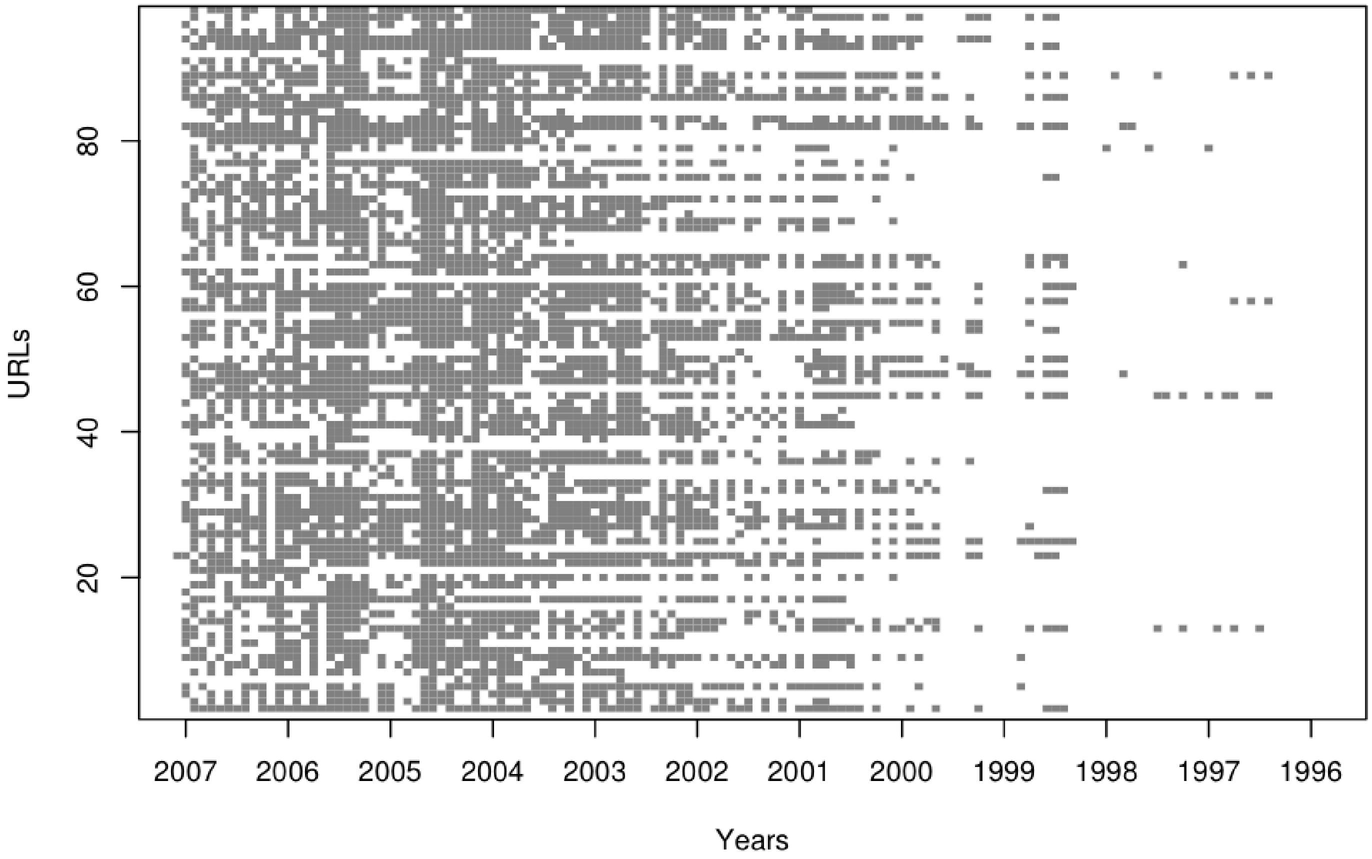
TC Frequencies in ukWaC and N-Grams



N-Grams have a threshold of 200

LS Evolution Over Time

Copies of web pages from the IA (1996-2007)

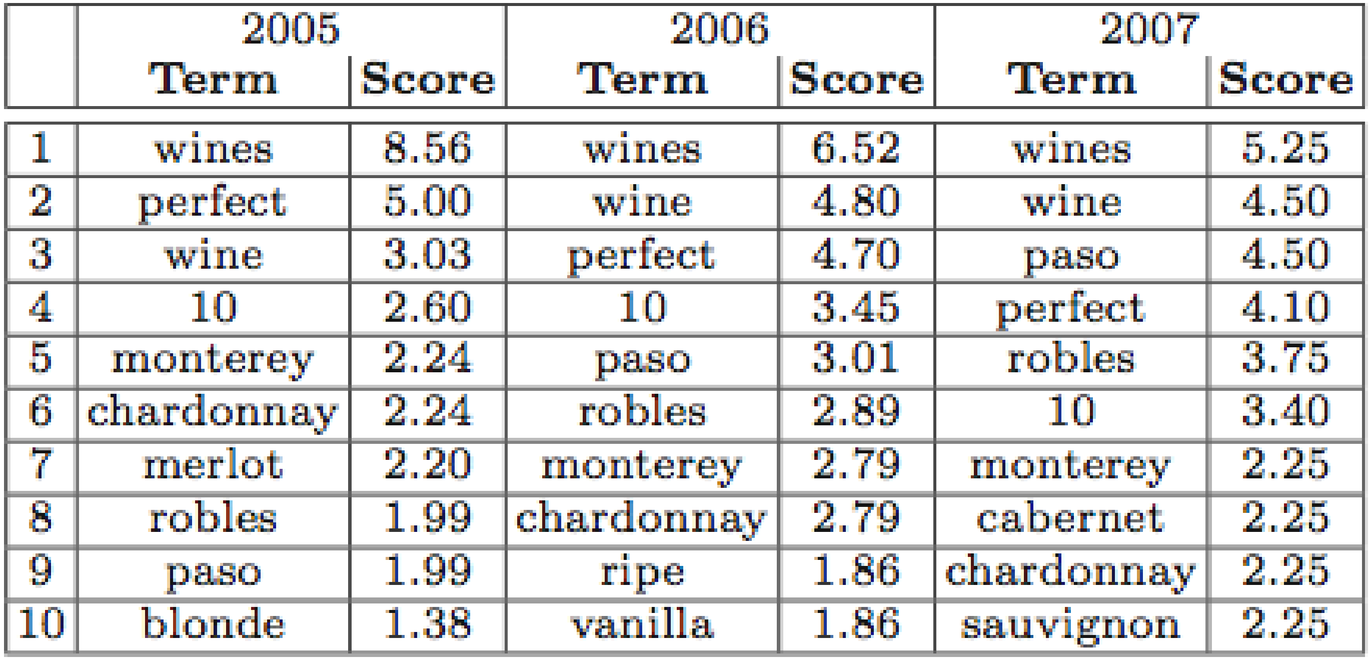


300 Random URLs, winnowed to 98, 10493 observations over 12 years

Evolution Over Time -- Example

10-term LSs generated for

http://www.perfect10wines.com



	2005		2006		2007	
	Term	Score	Term	Score	Term	Score
1	wines	8.56	wines	6.52	wines	5.25
2	perfect	5.00	wine	4.80	wine	4.50
3	wine	3.03	perfect	4.70	paso	4.50
4	10	2.60	10	3.45	perfect	4.10
5	monterey	2.24	paso	3.01	robles	3.75
6	chardonnay	2.24	robles	2.89	10	3.40
7	merlot	2.20	monterey	2.79	monterey	2.25
8	robles	1.99	chardonnay	2.79	cabernet	2.25
9	paso	1.99	ripe	1.86	chardonnay	2.25
10	blonde	1.38	vanilla	1.86	sauvignon	2.25

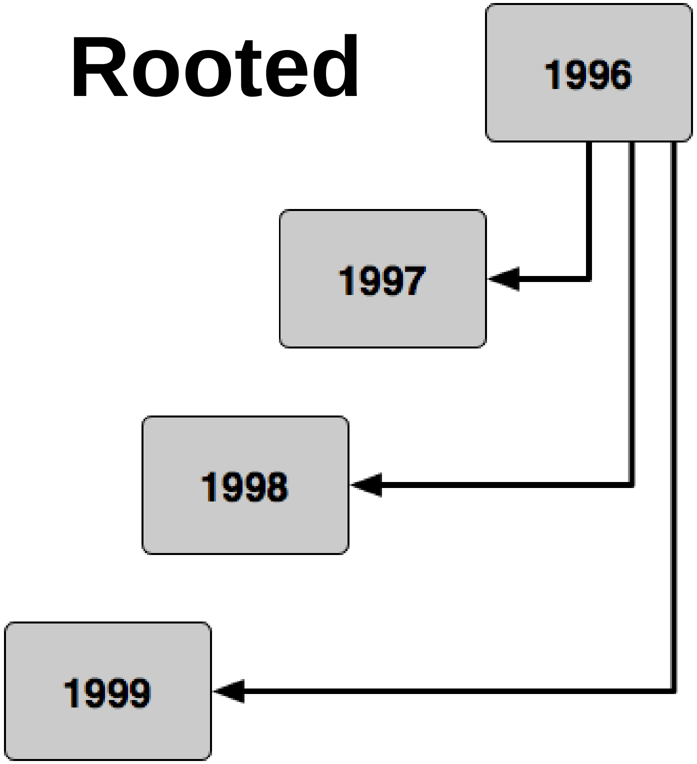
$$for\ all\ terms: \ |\cup| = 14\ and\ |\cap| = 8$$

Two Methods for Measuring Evolution

Idea

- Generate LSs from copies of URLs
- Conduct overlap analysis

Rooted



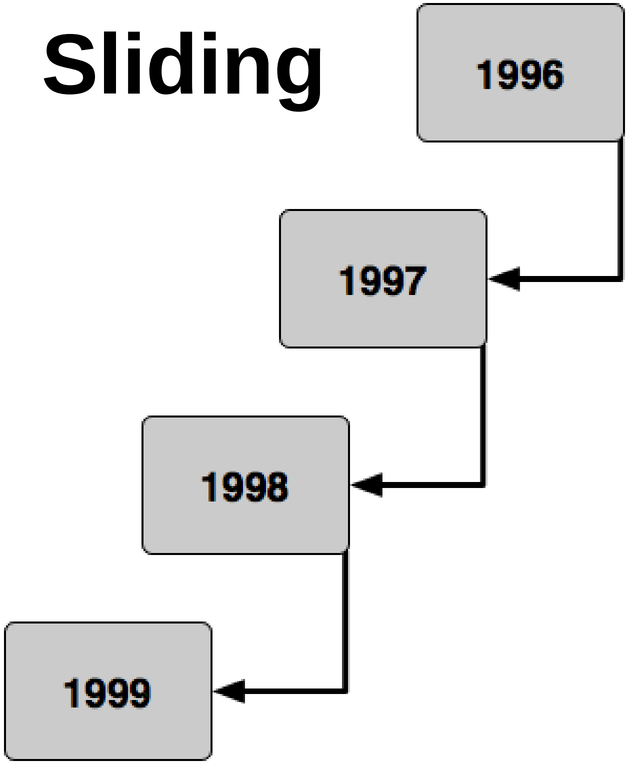
1996

1997

1998

1999

Sliding



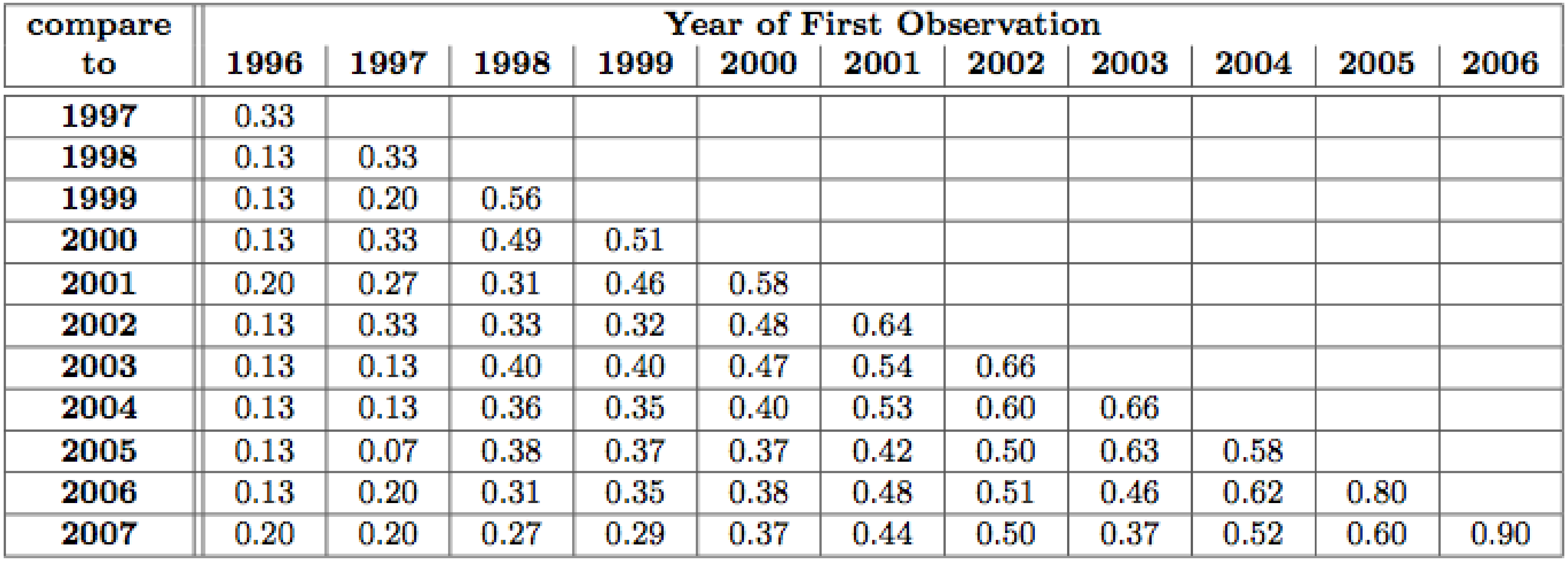
1996

1997

1998

1999

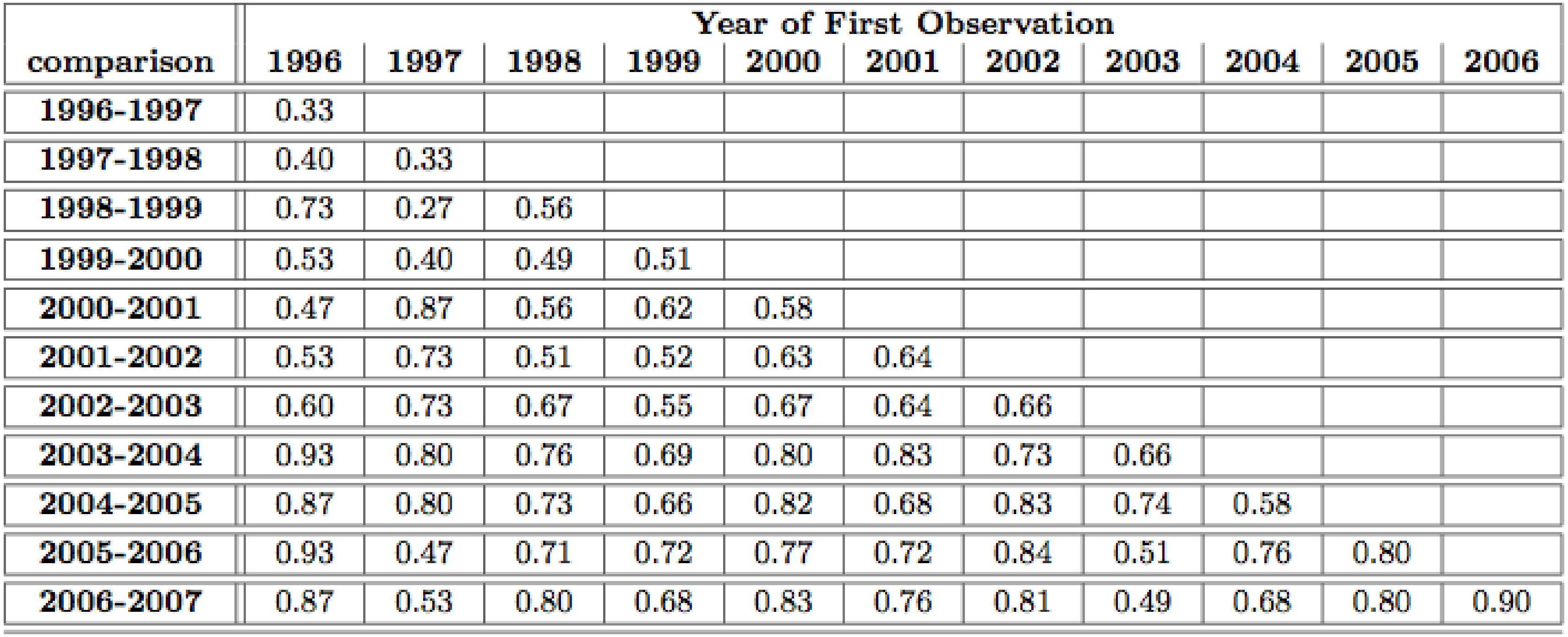
Evolution Over Time - Rooted



compare to	Year of First Observation										
	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
1997	0.33										
1998	0.13	0.33									
1999	0.13	0.20	0.56								
2000	0.13	0.33	0.49	0.51							
2001	0.20	0.27	0.31	0.46	0.58						
2002	0.13	0.33	0.33	0.32	0.48	0.64					
2003	0.13	0.13	0.40	0.40	0.47	0.54	0.66				
2004	0.13	0.13	0.36	0.35	0.40	0.53	0.60	0.66			
2005	0.13	0.07	0.38	0.37	0.37	0.42	0.50	0.63	0.58		
2006	0.13	0.20	0.31	0.35	0.38	0.48	0.51	0.46	0.62	0.80	
2007	0.20	0.20	0.27	0.29	0.37	0.44	0.50	0.37	0.52	0.60	0.90

- Little overlap between the early years and more recent ones
- Highest overlap in the first 1-2 years after creation of the LSs
- Rarely peaks after that – once terms are gone they do not return

Evolution Over Time - Sliding



comparison	Year of First Observation										
	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
1996-1997	0.33										
1997-1998	0.40	0.33									
1998-1999	0.73	0.27	0.56								
1999-2000	0.53	0.40	0.49	0.51							
2000-2001	0.47	0.87	0.56	0.62	0.58						
2001-2002	0.53	0.73	0.51	0.52	0.63	0.64					
2002-2003	0.60	0.73	0.67	0.55	0.67	0.64	0.66				
2003-2004	0.93	0.80	0.76	0.69	0.80	0.83	0.73	0.66			
2004-2005	0.87	0.80	0.73	0.66	0.82	0.68	0.83	0.74	0.58		
2005-2006	0.93	0.47	0.71	0.72	0.77	0.72	0.84	0.51	0.76	0.80	
2006-2007	0.87	0.53	0.80	0.68	0.83	0.76	0.81	0.49	0.68	0.80	0.90

- Overlap increases over time
- Seem to reach steady state around 2003

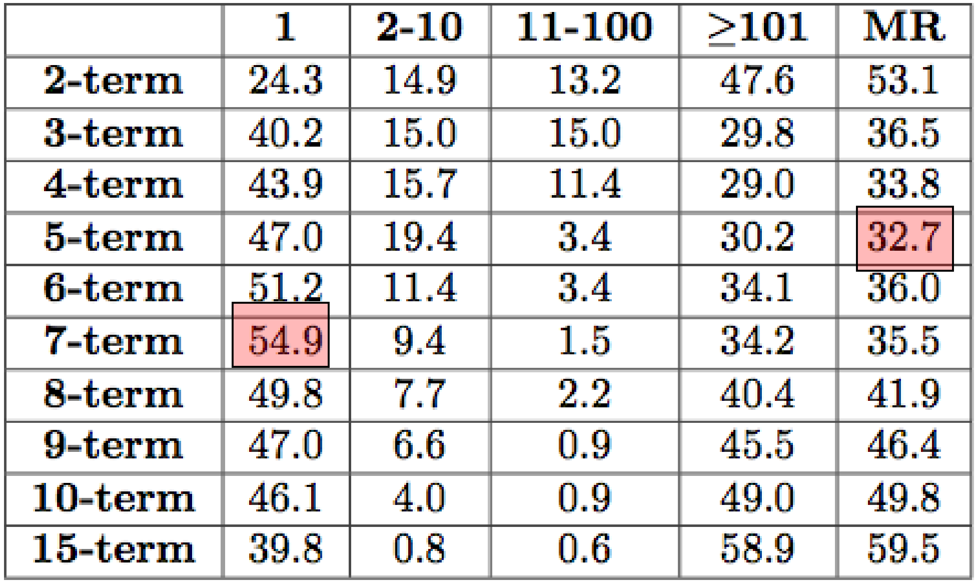
Performance of LSs

Idea

- Measure performance in respect to age of LS and number of terms it contains
- Query Google search API with LSs
- Identify URL in result set:

1. Top ranked
2. Ranked between 2-10
3. Ranked between 11-100
4. Ranked beyond 100 (considered undiscovered)

Performance – Number of Terms



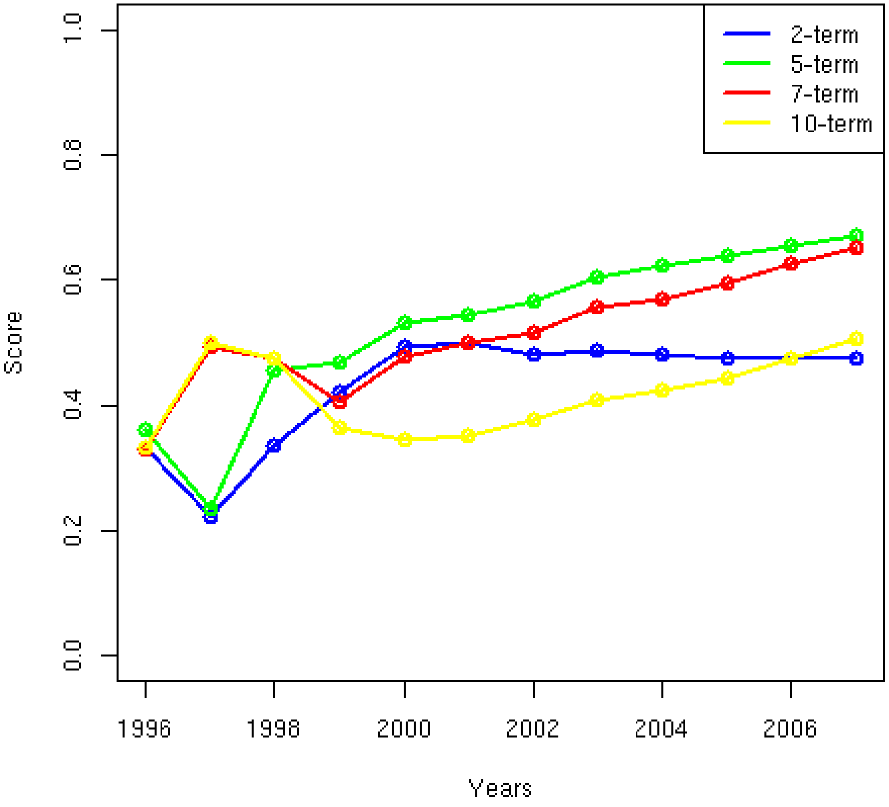
	1	2-10	11-100	≥101	MR
2-term	24.3	14.9	13.2	47.6	53.1
3-term	40.2	15.0	15.0	29.8	36.5
4-term	43.9	15.7	11.4	29.0	33.8
5-term	47.0	19.4	3.4	30.2	32.7
6-term	51.2	11.4	3.4	34.1	36.0
7-term	54.9	9.4	1.5	34.2	35.5
8-term	49.8	7.7	2.2	40.4	41.9
9-term	47.0	6.6	0.9	45.5	46.4
10-term	46.1	4.0	0.9	49.0	49.8
15-term	39.8	0.8	0.6	58.9	59.5

- 2-, 3- and 4-term LSs perform poorly
- 5-, 6- and 7-term LSs seem best
    - Top mean rank (MR) value with 5 terms
    - Most top ranked with 7 terms
    - Binary pattern: either top 10 or undiscovered
- 8+ terms -- decreased performance

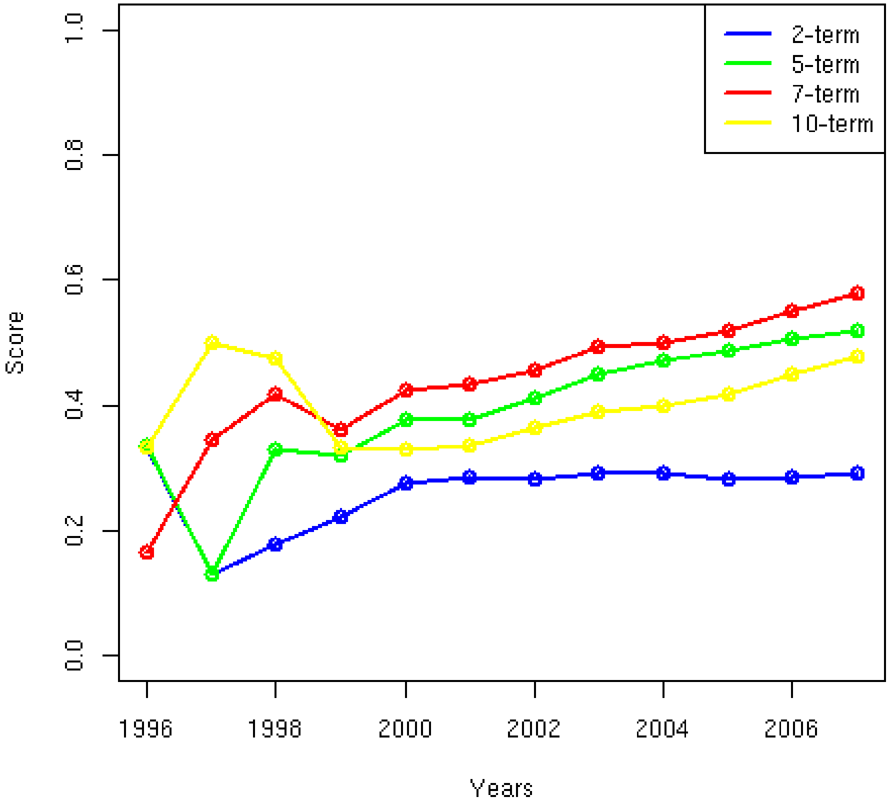
Performance – Age

Score of LSs consisting of 2, 5, 7 and 10 terms

Fair



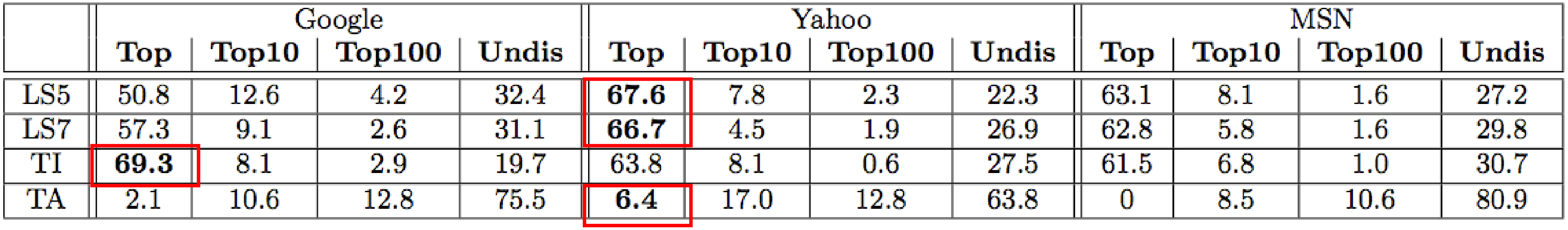
Optimistic



•Example, scores for the position of an URL in a list of 10:

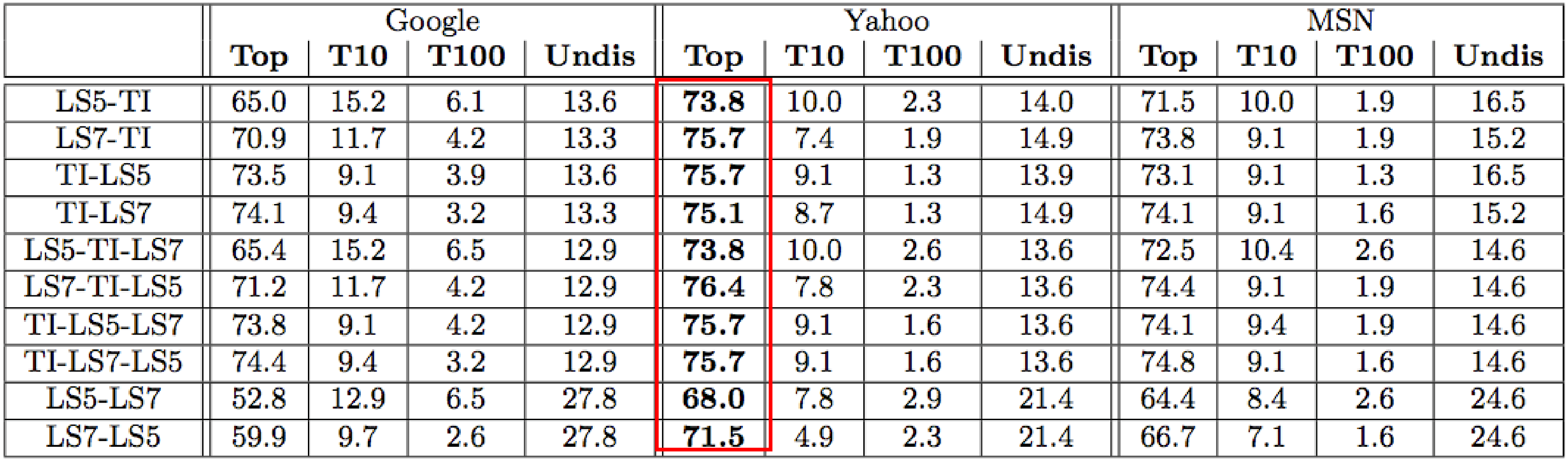
- fair: 10/10, 9/10, 8/10 ... 1/10, 0
- optimistic: 1/1, 1/2, 1/3 ... 1/10, 0

Titles (TI), 5- & 7-term Lexical Signatures (LS5, LS7), Tags (TA)



	Google				Yahoo				MSN			
	Top	Top10	Top100	Undis	Top	Top10	Top100	Undis	Top	Top10	Top100	Undis
LS5	50.8	12.6	4.2	32.4	**67.6**	7.8	2.3	22.3	63.1	8.1	1.6	27.2
LS7	57.3	9.1	2.6	31.1	**66.7**	4.5	1.9	26.9	62.8	5.8	1.6	29.8
TI	**69.3**	8.1	2.9	19.7	63.8	8.1	0.6	27.5	61.5	6.8	1.0	30.7
TA	2.1	10.6	12.8	75.5	**6.4**	17.0	12.8	63.8	0	8.5	10.6	80.9

Table 2: Relative Number of URLs Retrieved with one Single Method from Google, Yahoo and MSN



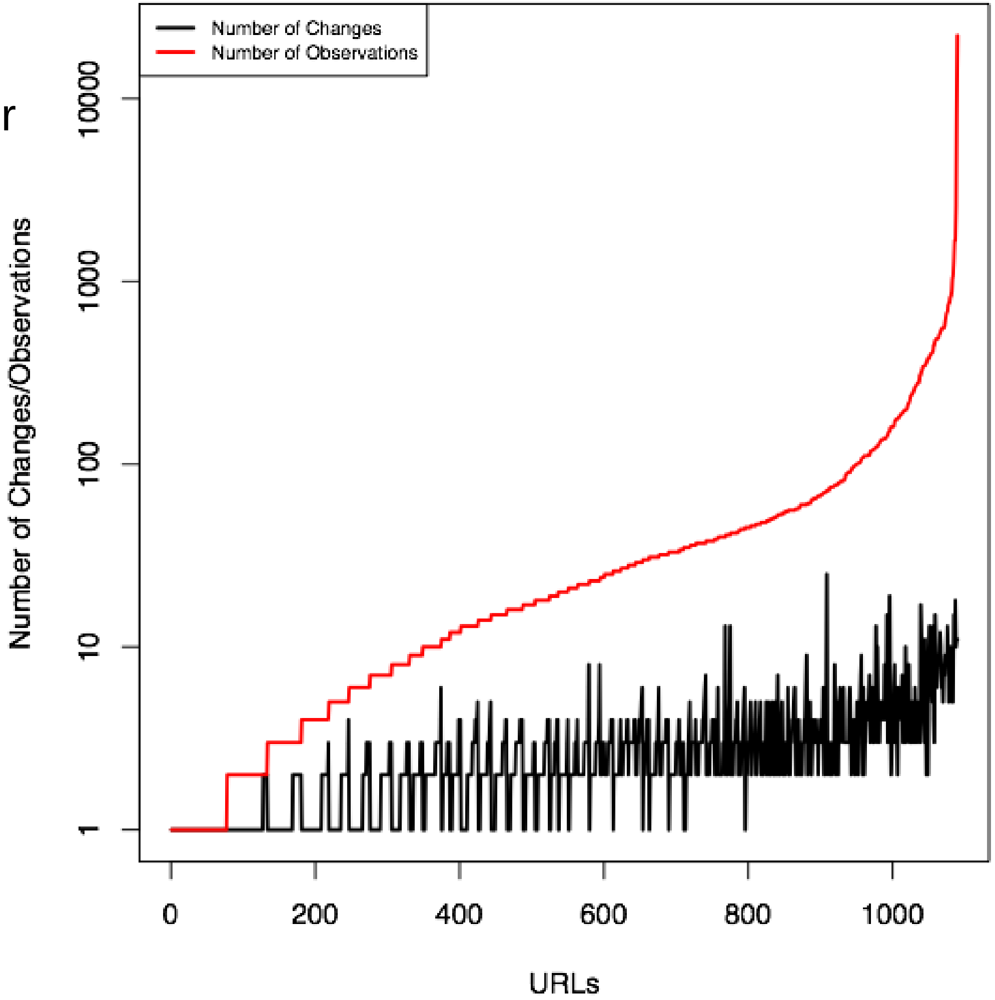
	Google				Yahoo				MSN			
	Top	T10	T100	Undis	Top	T10	T100	Undis	Top	T10	T100	Undis
LS5-TI	65.0	15.2	6.1	13.6	**73.8**	10.0	2.3	14.0	71.5	10.0	1.9	16.5
LS7-TI	70.9	11.7	4.2	13.3	**75.7**	7.4	1.9	14.9	73.8	9.1	1.9	15.2
TI-LS5	73.5	9.1	3.9	13.6	**75.7**	9.1	1.3	13.9	73.1	9.1	1.3	16.5
TI-LS7	74.1	9.4	3.2	13.3	**75.1**	8.7	1.3	14.9	74.1	9.1	1.6	15.2
LS5-TI-LS7	65.4	15.2	6.5	12.9	**73.8**	10.0	2.6	13.6	72.5	10.4	2.6	14.6
LS7-TI-LS5	71.2	11.7	4.2	12.9	**76.4**	7.8	2.3	13.6	74.4	9.1	1.9	14.6
TI-LS5-LS7	73.8	9.1	4.2	12.9	**75.7**	9.1	1.6	13.6	74.1	9.4	1.9	14.6
TI-LS7-LS5	74.4	9.4	3.2	12.9	**75.7**	9.1	1.6	13.6	74.8	9.1	1.6	14.6
LS5-LS7	52.8	12.9	6.5	27.8	**68.0**	7.8	2.9	21.4	64.4	8.4	2.6	24.6
LS7-LS5	59.9	9.7	2.6	27.8	**71.5**	4.9	2.3	21.4	66.7	7.1	1.6	24.6

Table 3: Relative Number of URLs Retrieved with Two or More Methods Combined

500 random URLs from dmoz.org winnowed to 309 (only 47 of 309 had tags in delicious.com).
Due to query restrictions, link neighborhood only run on Yahoo -- results were similar to tags.

Number of Title Changes and Observations in the IA

ordered in increasing order by:
1) observations
2) changes

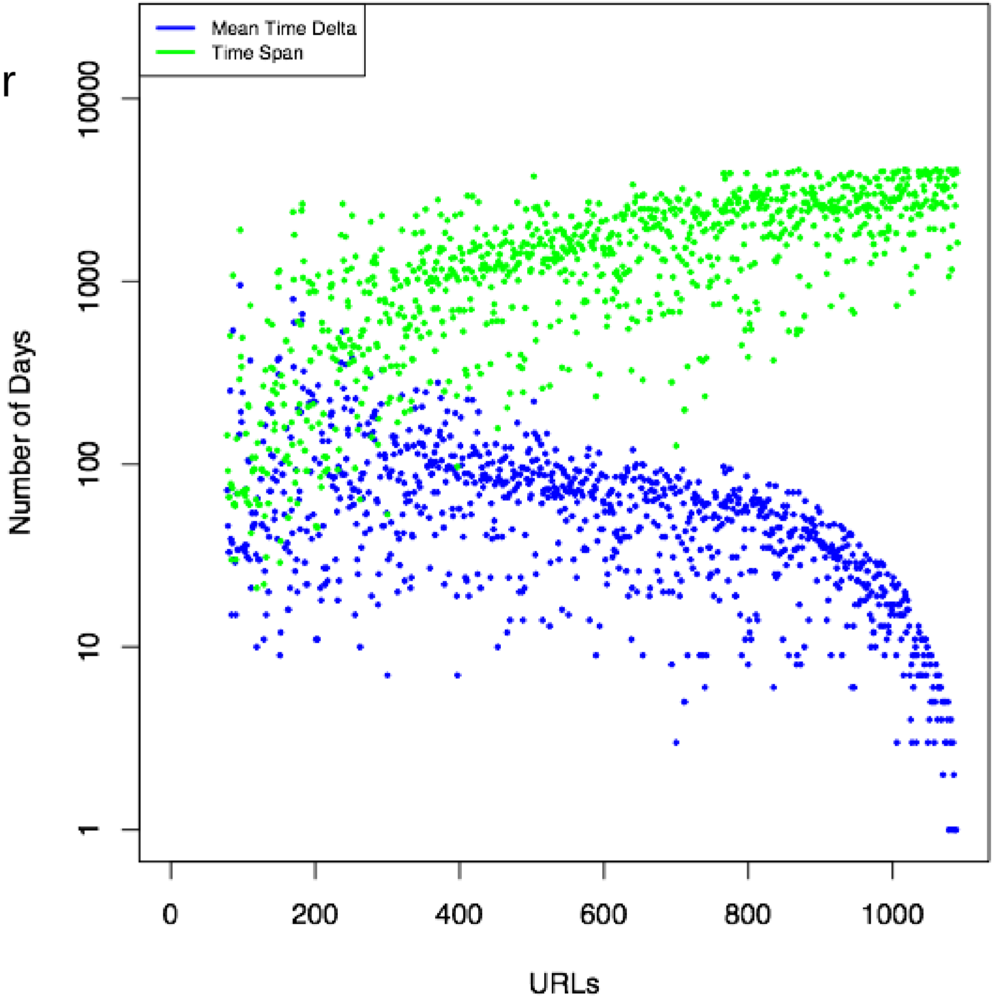


- generally low number of changes
- max changes: 25
- number of observations does not impact the number of changes

6000 random URLs from dmoz.org, winnowed to 1090 URLs and 100k+ observations

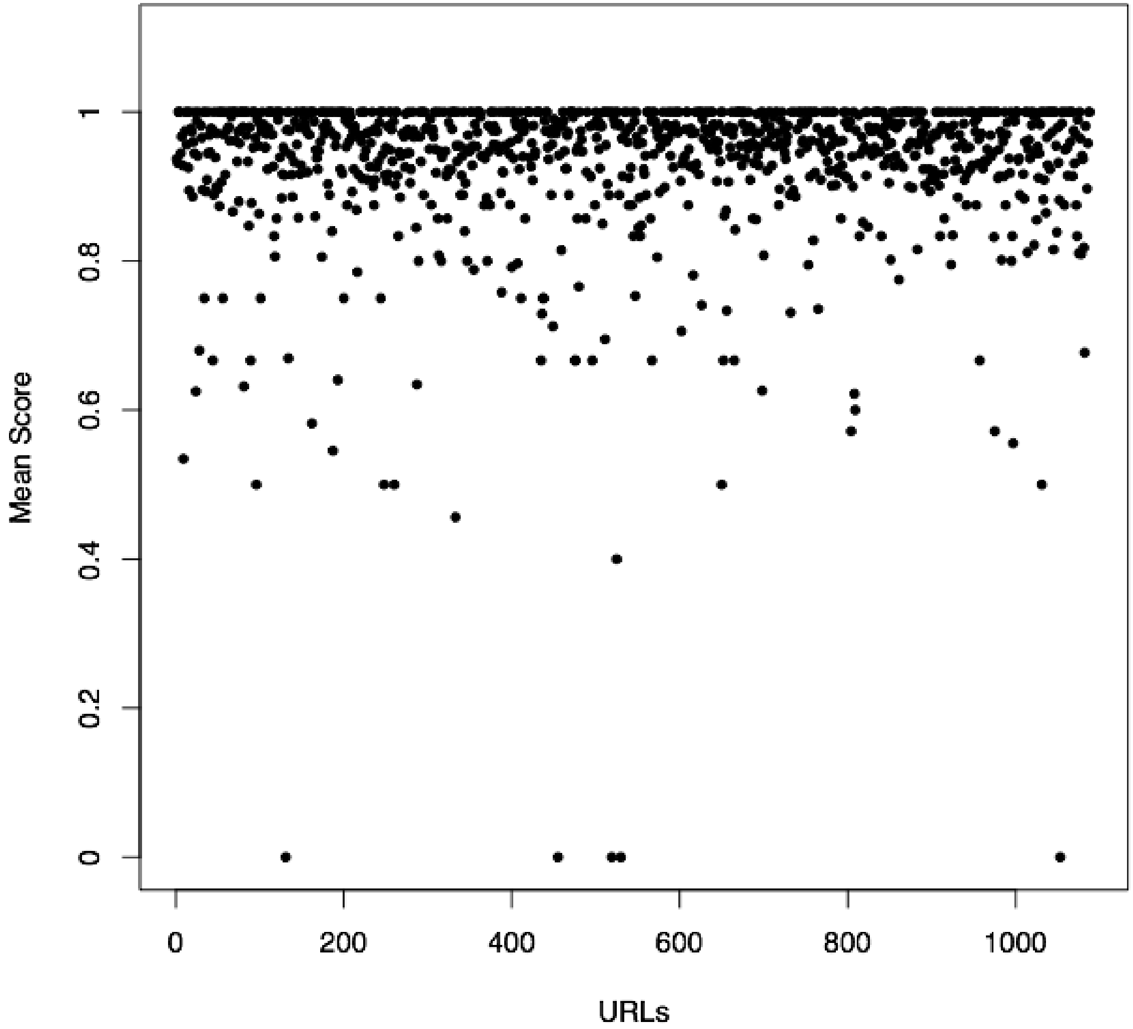
Mean Time Delta Between Changes Time Span Between First and Last Observation in the IA

ordered in increasing order by:
1) observations
2) changes



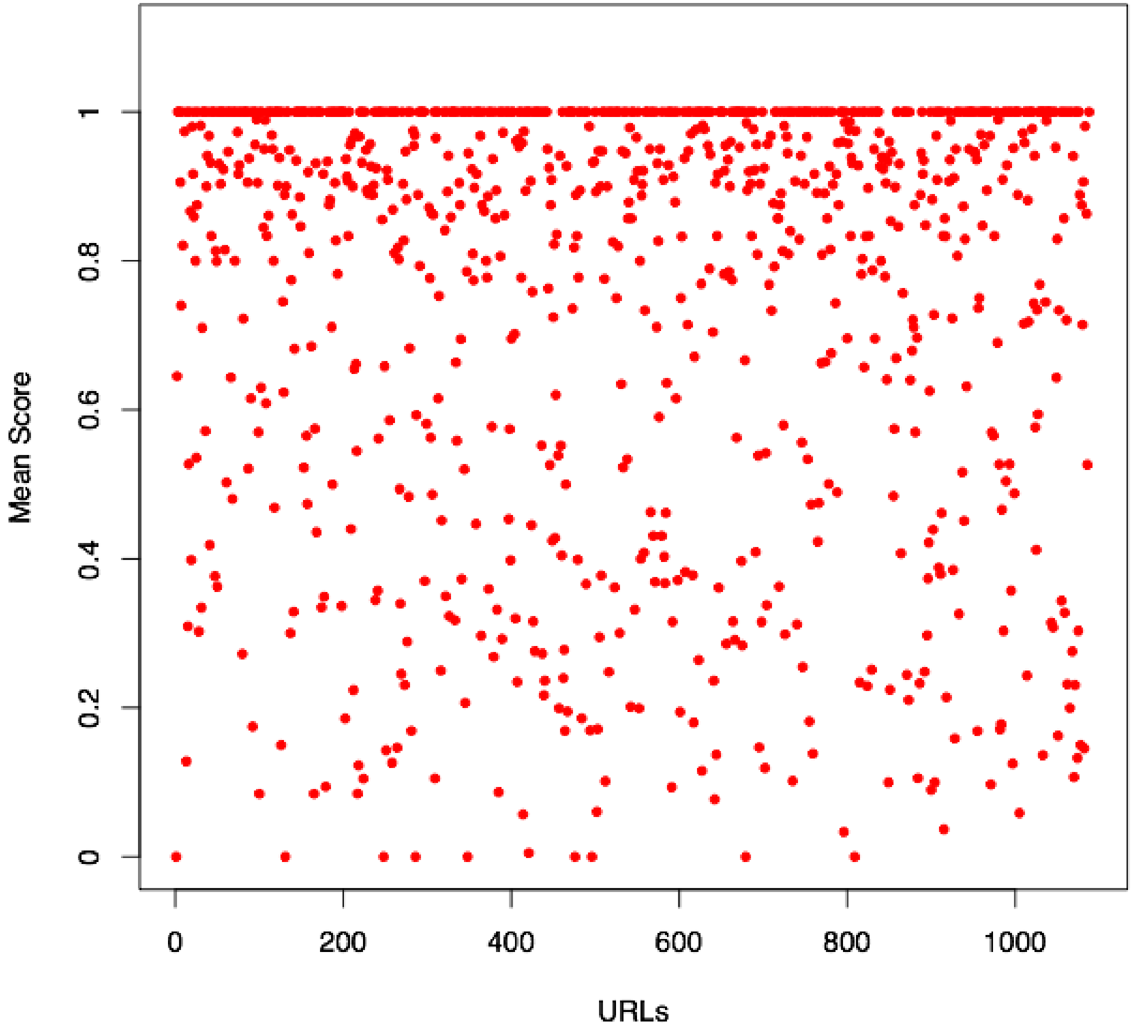
- time span between observations decreases with increasing number of observations
- overall time span just slightly increases
- URLs with many observations are being crawled frequently in a short period of time

Mean Levenshtein Scores of all Titles - Sliding



- 5 URLs with score = 0
- 85% of URLs with score >=0.8
- titles rarely change drastically

Mean Levenshtein Scores of all Titles - Rooted



- 9 URLs with score = 0
- 56% of URLs with score >=0.8
- titles more likely to change compared to their first observation

http://www.sun.com/solutions
mean Levenshtein score
sliding: 0.84 rooted: 0.29

1998-01-27
Sun Software Products Selector Guides -Solutions Tree

1999-02-20
Sun Software Solutions

2002-02-01
Sun Microsystems Products

2002-06-01
Sun Microsystems - Business & Industry Solutions

2003-08-01
Sun Microsystems - Industry & Infrastructure Solutions

2004-02-02
Sun Microsystems - Solutions

2004-06-10
Gateway Page - Sun Solutions

2006-01-09
Sun Microsystems Solutions & Services

2007-01-03
Services & Solutions

2007-02-07
Sun Services & Solutions

2008-01-19
Sun Solutions

http://www.datacity.com/mainf.html
mean Levenshtein score
sliding: 0.68 rooted: 0.15

2000-06-19
DataCity of Manassas Park Main Page

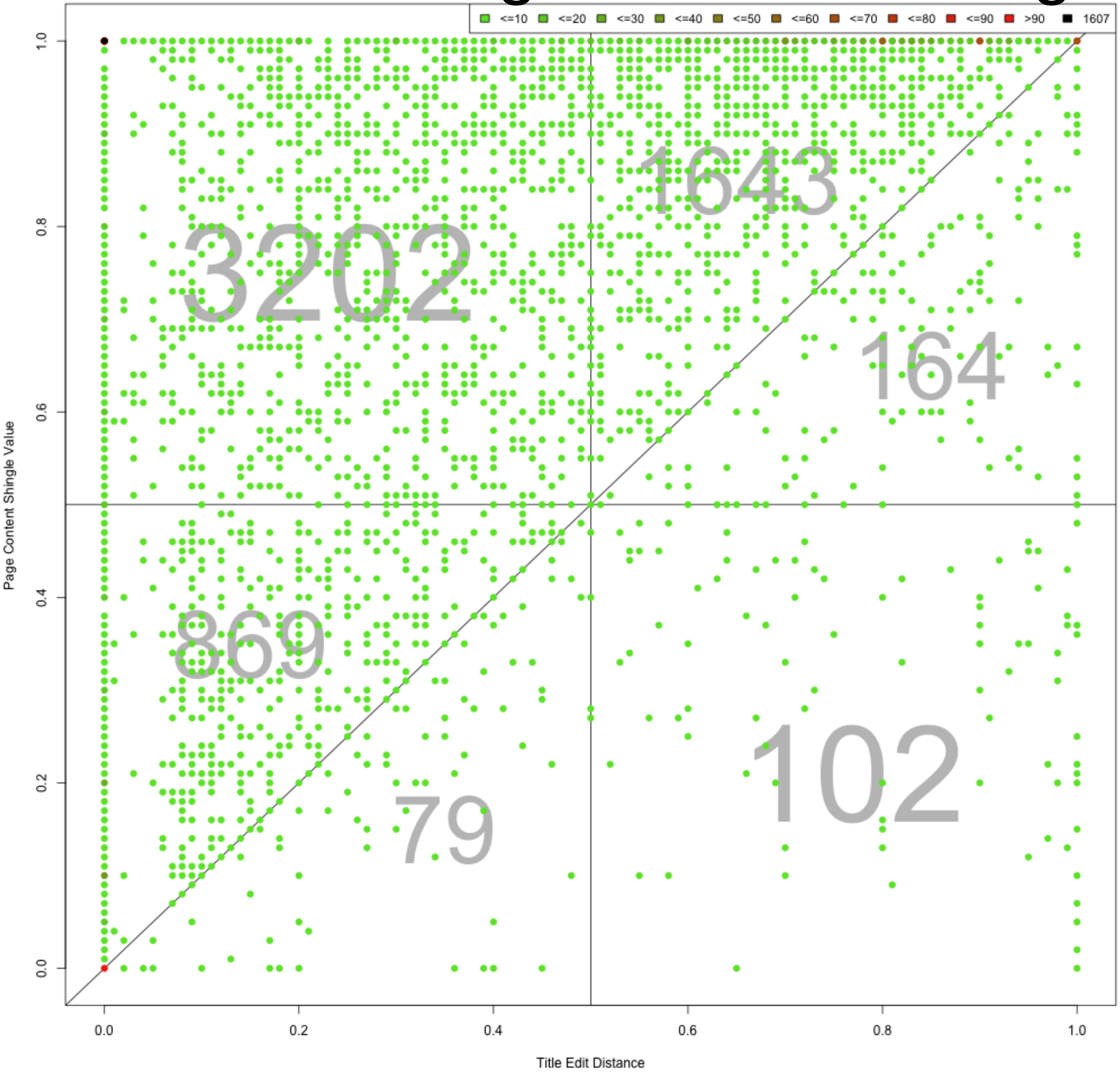
2000-10-12
DataCity of Manassas Park sells Custom Built Computers & Removable Hard Drives

2001-08-21
DataCity a computer company in Manassas Park sells Custom Built Computers & Removable Hard Drives

2002-10-16
computer company in Manassas Virginia sells Custom Built Computers with Removable Hard Drives Kits and Iomega 2GB Jaz Drives (jazz drives) October 2002 DataCity 800-326-5051 toll free

2006-03-14
Est 1989 Computer company in Stafford Virginia sells Custom Built Secure Computers with DoD 5200.1-R Approved Removable Hard Drives, Hard Drive Kits and Iomega 2GB Jaz Drives (jazz drives), introduces the IllumiNite&reg; lighted keyboard DataCity 800-326-5051 Service Disabled Veteran Owned Business SDVOB

Content Change vs. Title Change



Conclusions & Future Work

- LSs decay over time, Titles decay less
    - Rooted: quickly after generation
    - Sliding: seem to stabilize
- Titles give comparable performance to LSs
- Titles + LSs give better performance

- Future work:
    - can we know in advance if a title is “good”? (i.e., not “welcome to my home page”)
    - can we use tags to augment titles / LS?
    - how big should a link neighborhood be?

- Contact us to get a beta version of the Firefox extension (real soon now!)

Necronomicon