(Re-) Discovering Lost Web Pages

Martin Klein
Old Dominion University

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

Follow this and additional works at: http://digitalcommons.odu.edu/computerscience_presentations

Part of the Archival Science Commons, and the Computer Sciences Commons

Recommended Citation

This Book is brought to you for free and open access by the Computer Science at ODU Digital Commons. It has been accepted for inclusion in Computer Science Presentations by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@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)
A Comparison of Queueing, Cluster and Distributed Computing Systems

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

June 1994
Web Infrastructure: Refreshing & Migrating
### Lapsed Website

**WayBack Machine**

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

<table>
<thead>
<tr>
<th>Year</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>0 pages</td>
</tr>
<tr>
<td>1997</td>
<td>0 pages</td>
</tr>
<tr>
<td>1998</td>
<td>0 pages</td>
</tr>
<tr>
<td>1999</td>
<td>1 pages</td>
</tr>
<tr>
<td>2000</td>
<td>12 pages</td>
</tr>
<tr>
<td>2001</td>
<td>11 pages</td>
</tr>
<tr>
<td>2002</td>
<td>5 pages</td>
</tr>
<tr>
<td>2003</td>
<td>13 pages</td>
</tr>
<tr>
<td>2004</td>
<td>11 pages</td>
</tr>
<tr>
<td>2005</td>
<td>0 pages</td>
</tr>
</tbody>
</table>

#### Notes:
- Some duplicates are not shown. [See all](http://web.archive.org/web/*/http://www.dl00.org/)
- * denotes when site was updated.

- **Gambling**
- **Search Engine Portal**
- **ACM DL Conference**
- **Porn**
## URI Content Mapping Problem

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>same URI maps to <strong>same</strong> or very similar content at a later time</td>
<td>2</td>
<td>same URI maps to <strong>different</strong> content at a later time</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>different URI maps to <strong>same</strong> or very similar content at the same or at a later time</td>
<td>4</td>
<td>the content can <strong>not be found at any URI</strong></td>
<td></td>
</tr>
</tbody>
</table>

- **Example Diagrams:**
  - URI 1 (U1) maps to Content 1 (C1) at time A, and maps to the same Content 1 (C1) at time B.
  - URI 1 (U1) maps to Content 2 (C2) at time A, and maps to the same Content 1 (C1) at time B.
  - URI 2 (U2) maps to Content 1 (C1) at time A, and then 404 error at time B.
  - URI 1 (U1) maps to Content 1 (C1) at time A, and then to an unknown URI at time B.
Scenario 1: Same URI, Same Content

JCDL 2008

http://www.jcdl2008.org/

July 2008

http://www.jcdl2008.org/

Today
Scenario 2: Same URI, Different Content

Hypertext 2006

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

http://www.ht06.org/
Today

Hypertext and hypermedia are technologies for supporting structured knowledge work. The Seventh International ACM Conference on Hypertext and Hypermedia (HT 2006) will focus specifically on tools that help us represent, model and interact with social structures, including features, literary, diagramatic, and other types of social structures. Reasort, in those ranging from and acll to large scale hypermedia systems. The conference has been increasingly focused on creating common social phenomena using network tools and social science, as well as social and hypermedia researchers who specialize in building tools to build, manage, and manage structure-intensive systems.

The conference will take place in the Kaiserin SAS H.C. Andersen Hotel in Odense, Denmark. The hotel is located in the center of the central part of Odense, with easy access to several touring attractions and the historic pedestrian walking area of the old city. The conference hotel has extended our attendance a special nightly rate of 880 DKK, which includes breakfast and free wireless internet access. Odense, site of Hypermedia 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 WikSym 2006

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

Co-operation with ACM SIGSOCS

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

Office Administrator
Entrepreneurial, growing and reputable Concrete Construction Company has an opportunity for you. We offer competitive salary and administrative expertise to contribute to the strategic development. Read More
Country: USA, Location: Wisconsin-Northem Waukesha, WI 53186

Informational 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 them. Read More
Country: USA, Location: Wisconsin-Milwaukee, WI 53202

Assistant Controller
Assistant Controller. Reporting to the Senior 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, WI 53223

IT Staff Auditor
Description/Position Summary Reporting to the Manager of IT Internal Audit, this position performs audits within Rockwell Automation, Responsibilities...
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 Is 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 2003 Annual Conference Planning Committee
Pieter Bolman (PSP ExCo Chair), Patrick Berndt, Donald Burden, Nigel Fletcher-Jones, Andrew Grabois, Doug LaFrenier, Eric Massant, Ted Nardio, Bill Siewinski AAP: Barbara Meredith, Sara Firestone

Program of Events
MONDAY, FEBRUARY 3, 2003
8:00am-9:00am Continental Breakfast
9:00am-3:00pm Your chance to visit the New Technologies/Services Exhibitors
4:00pm Pre-Conference Session (separate registration fee)
WHERE IS THE USER FOR YOUR SMART CONTENT?
Produced by the AAP/PSP Electronic Information Committee
9-9:00am-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
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”
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>
<table>
<thead>
<tr>
<th>LS</th>
<th>NICNICHOLS NICHOLS NIC STUFF SHOOT COMMAND PENITENTIARY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>NICNICHOLS.COM : DOCUMENTARY TOY CAMERA PHOTOGRAPHY OF NIC NICHOLS : HOLGA, LOMO AND OTHER LO-FI CAMERAS!</td>
</tr>
<tr>
<td>Tags</td>
<td>PHOTOGRAPHY BLOG PHOTOGRAPHER PORTFOLIO PORTFOLIO INSPIRATION PHOTOGRAPHERS</td>
</tr>
<tr>
<td>LNLS</td>
<td>NICNICHOLS PHOTO SPACER VIEW PHIREBRUSH SUBMISSION BOONIKA</td>
</tr>
</tbody>
</table>

Table 1: Data Obtained from www.nicnichols.com
What is a Signature?
(aka “message digest”, examples include “md5” and “sha-1”)
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”
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

<table>
<thead>
<tr>
<th>Rank/Results</th>
<th>URL</th>
<th>LS</th>
</tr>
</thead>
</table>
| 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    | http://www.loc.gov         | library collections congress thomas american  
http://www.google.com/search?q=library+collections+congress+thomas+american |
| (1/174,000 in 01/2008) |                           |                                                                      |
http://www.google.com/search?q=libraries+jcdl+digital+conference+pst |
| (2/77 in 01/2008) |                           |                                                                      |
| 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](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?”

\[
TF_{ij} = \frac{f_{ij}}{m_i}
\]

\[
f_{ij} = \text{freq of } j \text{ in } i
\]

\[
m_i = \text{max freq in } i
\]

\[
IDF_j = \log \left( \frac{N}{n_j} \right) + 1
\]

\[
N = \text{total number of documents}
\]

\[
n_j = \text{number of documents } j \text{ occurs in}
\]
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

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Local Universe Term</th>
<th>TF-IDF</th>
<th>Screen Scraping Term</th>
<th>TF-IDF</th>
<th>N-grams Term</th>
<th>TF-IDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>perfect</td>
<td>7.77</td>
<td>wines</td>
<td>5.97</td>
<td>wines</td>
<td>7.56</td>
</tr>
<tr>
<td>2</td>
<td>wines</td>
<td>6.95</td>
<td>robles</td>
<td>5.3</td>
<td>perfect</td>
<td>7.25</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>6.57</td>
<td>perfect</td>
<td>4.35</td>
<td>robles</td>
<td>7.18</td>
</tr>
<tr>
<td>4</td>
<td>paso</td>
<td>6.29</td>
<td>paso</td>
<td>4.27</td>
<td>paso</td>
<td>6.93</td>
</tr>
<tr>
<td>5</td>
<td>wine</td>
<td>6.18</td>
<td>wine</td>
<td>3.26</td>
<td>wine</td>
<td>4.86</td>
</tr>
<tr>
<td>6</td>
<td>robles</td>
<td>5.4</td>
<td>sauvignon</td>
<td>3.16</td>
<td>10</td>
<td>4.52</td>
</tr>
<tr>
<td>7</td>
<td>sauvignon</td>
<td>3.54</td>
<td>chardonnay</td>
<td>3.15</td>
<td>chardonnay</td>
<td>3.99</td>
</tr>
<tr>
<td>8</td>
<td>cabernet</td>
<td>3.54</td>
<td>robles84</td>
<td>3.11</td>
<td>sauvignon</td>
<td>3.93</td>
</tr>
<tr>
<td>9</td>
<td>monterey</td>
<td>3.36</td>
<td>cabernet</td>
<td>3.09</td>
<td>cabernet</td>
<td>3.89</td>
</tr>
<tr>
<td>10</td>
<td>chardonnay</td>
<td>3.36</td>
<td>enthusiast85</td>
<td>2.91</td>
<td>monterey</td>
<td>3.49</td>
</tr>
</tbody>
</table>
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
  – [https://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html](https://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html)

<table>
<thead>
<tr>
<th>Term</th>
<th>All</th>
<th>Buy</th>
<th>Can’t</th>
<th>Is</th>
<th>Love</th>
<th>Me</th>
<th>Need</th>
<th>Please</th>
<th>You</th>
<th>Long</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TC</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td><strong>DF</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

• 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/](http://wacky.sslmit.unibo.it/)
TC Ranks vs DF Ranks Within ukWaC
Rank Correlation Within ukWaC

p-value < 2.2e-16

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

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

for all terms: $|U| = 14$ 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**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>0.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>0.13</td>
<td>0.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td>0.13</td>
<td>0.20</td>
<td>0.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.13</td>
<td>0.33</td>
<td>0.49</td>
<td>0.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>0.20</td>
<td>0.27</td>
<td>0.31</td>
<td>0.46</td>
<td>0.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>0.13</td>
<td>0.33</td>
<td>0.33</td>
<td>0.32</td>
<td>0.48</td>
<td>0.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>0.13</td>
<td>0.13</td>
<td>0.40</td>
<td>0.40</td>
<td>0.47</td>
<td>0.54</td>
<td>0.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>0.13</td>
<td>0.13</td>
<td>0.36</td>
<td>0.35</td>
<td>0.40</td>
<td>0.53</td>
<td>0.60</td>
<td>0.66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>0.13</td>
<td>0.07</td>
<td>0.38</td>
<td>0.37</td>
<td>0.37</td>
<td>0.42</td>
<td>0.50</td>
<td>0.63</td>
<td>0.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>0.13</td>
<td>0.20</td>
<td>0.31</td>
<td>0.35</td>
<td>0.38</td>
<td>0.48</td>
<td>0.51</td>
<td>0.46</td>
<td>0.62</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>0.20</td>
<td>0.20</td>
<td>0.27</td>
<td>0.29</td>
<td>0.37</td>
<td>0.44</td>
<td>0.50</td>
<td>0.37</td>
<td>0.52</td>
<td>0.60</td>
<td>0.90</td>
</tr>
</tbody>
</table>

- 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

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1996-1997</td>
<td>0.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1997-1998</td>
<td>0.40</td>
<td>0.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998-1999</td>
<td>0.73</td>
<td>0.27</td>
<td>0.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999-2000</td>
<td>0.53</td>
<td>0.40</td>
<td>0.49</td>
<td>0.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000-2001</td>
<td>0.47</td>
<td>0.87</td>
<td>0.56</td>
<td>0.62</td>
<td>0.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001-2002</td>
<td>0.53</td>
<td>0.73</td>
<td>0.51</td>
<td>0.52</td>
<td>0.63</td>
<td>0.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2002-2003</td>
<td>0.60</td>
<td>0.73</td>
<td>0.67</td>
<td>0.55</td>
<td>0.67</td>
<td>0.64</td>
<td>0.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003-2004</td>
<td>0.93</td>
<td>0.80</td>
<td>0.76</td>
<td>0.69</td>
<td>0.80</td>
<td>0.83</td>
<td>0.73</td>
<td>0.66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004-2005</td>
<td>0.87</td>
<td>0.80</td>
<td>0.73</td>
<td>0.66</td>
<td>0.82</td>
<td>0.68</td>
<td>0.83</td>
<td>0.74</td>
<td>0.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005-2006</td>
<td>0.93</td>
<td>0.47</td>
<td>0.71</td>
<td>0.72</td>
<td>0.77</td>
<td>0.72</td>
<td>0.84</td>
<td>0.51</td>
<td>0.76</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>2006-2007</td>
<td>0.87</td>
<td>0.53</td>
<td>0.80</td>
<td>0.68</td>
<td>0.83</td>
<td>0.76</td>
<td>0.81</td>
<td>0.49</td>
<td>0.68</td>
<td>0.80</td>
<td>0.90</td>
</tr>
</tbody>
</table>
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

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

- 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)

<table>
<thead>
<tr>
<th></th>
<th>Google</th>
<th></th>
<th>Yahoo</th>
<th></th>
<th>MSN</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top</td>
<td>Top10</td>
<td>Top100</td>
<td>Undis</td>
<td>Top</td>
<td>Top10</td>
</tr>
<tr>
<td>LS5</td>
<td>50.8</td>
<td>12.6</td>
<td>4.2</td>
<td>32.4</td>
<td><strong>67.6</strong></td>
<td>7.8</td>
</tr>
<tr>
<td>LS7</td>
<td>57.3</td>
<td>9.1</td>
<td>2.6</td>
<td>31.1</td>
<td><strong>66.7</strong></td>
<td>4.5</td>
</tr>
<tr>
<td>TI</td>
<td><strong>69.3</strong></td>
<td>8.1</td>
<td>2.9</td>
<td>19.7</td>
<td>63.8</td>
<td>8.1</td>
</tr>
<tr>
<td>TA</td>
<td>2.1</td>
<td>10.6</td>
<td>12.8</td>
<td>75.5</td>
<td><strong>6.4</strong></td>
<td>17.0</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Google</th>
<th></th>
<th>Yahoo</th>
<th></th>
<th>MSN</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top</td>
<td>T10</td>
<td>T100</td>
<td>Undis</td>
<td>Top</td>
<td>T10</td>
</tr>
<tr>
<td>LS5-TI</td>
<td>65.0</td>
<td>15.2</td>
<td>6.1</td>
<td>13.6</td>
<td><strong>73.8</strong></td>
<td>10.0</td>
</tr>
<tr>
<td>LS7-TI</td>
<td>70.9</td>
<td>11.7</td>
<td>4.2</td>
<td>13.3</td>
<td><strong>75.7</strong></td>
<td>7.4</td>
</tr>
<tr>
<td>TI-LS5</td>
<td>73.5</td>
<td>9.1</td>
<td>3.9</td>
<td>13.6</td>
<td><strong>75.7</strong></td>
<td>9.1</td>
</tr>
<tr>
<td>TI-LS7</td>
<td>74.1</td>
<td>9.4</td>
<td>3.2</td>
<td>13.3</td>
<td><strong>75.1</strong></td>
<td>8.7</td>
</tr>
<tr>
<td>LS5-TI-LS7</td>
<td>65.4</td>
<td>15.2</td>
<td>6.5</td>
<td>12.9</td>
<td><strong>73.8</strong></td>
<td>10.0</td>
</tr>
<tr>
<td>LS7-TI-LS5</td>
<td>71.2</td>
<td>11.7</td>
<td>4.2</td>
<td>12.9</td>
<td><strong>76.4</strong></td>
<td>7.8</td>
</tr>
<tr>
<td>TI-LS5-LS7</td>
<td>73.8</td>
<td>9.1</td>
<td>4.2</td>
<td>12.9</td>
<td><strong>75.7</strong></td>
<td>9.1</td>
</tr>
<tr>
<td>TI-LS7-LS5</td>
<td>74.4</td>
<td>9.4</td>
<td>3.2</td>
<td>12.9</td>
<td><strong>75.7</strong></td>
<td>9.1</td>
</tr>
<tr>
<td>LS5-LS7</td>
<td>52.8</td>
<td>12.9</td>
<td>6.5</td>
<td>27.8</td>
<td><strong>68.0</strong></td>
<td>7.8</td>
</tr>
<tr>
<td>LS7-LS5</td>
<td>59.9</td>
<td>9.7</td>
<td>2.6</td>
<td>27.8</td>
<td><strong>71.5</strong></td>
<td>4.9</td>
</tr>
</tbody>
</table>

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® lighted keyboard DataCity 800-326-5051 Service Disabled Veteran Owned Business SDVOB
Content Change vs. Title Change

![Graph showing content change vs. title change](image)
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