Learning to Love “Ugly”: Making Sense of Archived Email

Douglas W. Oard

College of Information Studies and Institute for Advanced Computer Studies
University of Maryland

November 10, 2008

Smithsonian
Outline

• Thanks!
  – You’ve addressed many important problems
  – Although we have a few more …

• Seeing the forest
• Studying a tree
• Finding that tree
  – Can it be Google?
  – Protecting sensitive content

• techNOLOGy
Email without the “E”

Oey, Jan

From: Jill Schultz [jill.schultz@covance.com]
Sent: Tuesday, October 30, 2001 4:49 PM
To: Holt.Klausvon@pmlint.ch; kuhl.peter@pmlint.ch; rustemeier.Klaus@pmlint.ch; Schepers.Georg@pmlint.ch; Stabbert.Regina@pmlint.ch; tricker.anthony@pmlint.ch; Hans-Juergen.Roethig@pmusa.com; Jan.Oey@pmusa.com; robin.d.kinser@pmusa.com

Subject: Additional glucuronide data as requested:

1) Representative chromatogram of deconjugated nicotine and cotinine glucuronides (chrom.doc).

2) Calibration curves for nicotine and cotinine glucuronides (line1.doc and line2.doc).

3) Freeze/thaw data (3 additional cycles) and room temperature stability (24 hours) was supplied with the original validation data on 18 Oct.

Extract stability of glucuronides was not performed for this study, because the extracted glucuronides are deconjugated during the extraction procedure. Therefore extract stability is covered by the data generated during the aglycone validation study (44/8).

4) Analytical method summary:

Urine samples are thawed at nom. 25xC and spiked with internal standards. The samples are then deconjugated using B-glucuronidase at nom. 37xC for 18-22 hours. The aglycones are extracted from the urine using HCX SPE cartridges and the extracts submitted for analysis by LC-MS/MS.

Best regards,

Jill
Prehistoric Email

• Domain Name System (Jon Postel)
  jack@mail.odin.twenty.net

• ARPAnet “Bang Paths”
  utzoo!decvax!harpo!eagle!mhtsa!ihnss!ihuxp!grg

• CSNET “Care of” addresses
  mary%compsci@udel
Outline

• Thanks!
  – You’ve addressed many important problems
  – Although we have a few more …

➢ Seeing the forest
• Studying a tree
• Finding that tree
  – Can it be Google?
  – Protecting sensitive content
• techNOLOGy
Shneiderman Collection

44,971 email messages
Temporal Rhythm Query
Visualizing Reagan NSC “Email”

Rodman, Peter

Address: NSPWR

Description: An aide to Henry Kissinger ever since undergraduate days at Harvard (which were followed by Oxford and law school), Rodman wrote speeches for Kissinger at the NSC in the early 1970’s, assisted on Kissinger's memoirs, served as Director of Policy Planning at the State Department under George Shultz in the 1980's, and came to the NSC in 1986 as a Special Assistant to the President and Deputy to Poindexter for Foreign Policy (as opposed to Defense Policy etc.). He continued at the NSC as NSC Counselor under Carlucci and Powell, as well as in the Bush administration.

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>Number of messages</th>
<th>Timeline (8/1/85-3/1/87)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cannistraro, Vincent</td>
<td>NSVMC</td>
<td>0 2 2</td>
<td></td>
</tr>
<tr>
<td>Earl, Robert L.</td>
<td>NSRLE</td>
<td>1 0 1</td>
<td></td>
</tr>
<tr>
<td>Fortier, Donald R.</td>
<td>NSDRF</td>
<td>1 0 1</td>
<td></td>
</tr>
<tr>
<td>North, Oliver</td>
<td>NSOLN</td>
<td>1 0 1</td>
<td></td>
</tr>
<tr>
<td>Pearson, W. Robert</td>
<td>NSWRP</td>
<td>0 1 1</td>
<td></td>
</tr>
<tr>
<td>Poindexter, John M.</td>
<td>NSJMP</td>
<td>1 1 2</td>
<td></td>
</tr>
<tr>
<td>Ringdahl, Phil</td>
<td>NSPHR</td>
<td>0 1 1</td>
<td></td>
</tr>
<tr>
<td>Stark, James R.</td>
<td>NSJRS</td>
<td>2 1 3</td>
<td></td>
</tr>
<tr>
<td>Teicher, Howard</td>
<td>NSHRT</td>
<td>1 0 1</td>
<td></td>
</tr>
<tr>
<td>Tillman, Jacqueline</td>
<td>NSJT</td>
<td>0 1 1</td>
<td></td>
</tr>
</tbody>
</table>
People, Clustered by Content

44. Burghardt, Raymond
56. Burns, William J.
52. Cannistraro, Vincent
18. Carlucci, Frank C.
53. Cockell, William A.
23. Covey, James P.
6. Coy, Craig P.
31. Degraffenreid, Kenneth E.
13. Djerejian, Ed
24. Douglass, John
47. Earl, Robert L.
10. Fortier, Donald R.
14. Gantt, Florence
15. Hall, Fawn
55. Hall, Wilma
17. Hughes, Phillip
2. Keel, Alton G.
16. Lavin, Alton L.
8. Major, David
54. Martin, William F.
41. McDaniel, Rodney B.
42. McFarlane, Robert
29. Miller, Johnathan
12. Mitsler, Elaine M.
11. Morton, Elaine
22. Murphy, Jaqueline A.
62. NSBEB

73. NSCLM
72. NSDBK
64. NSDFP
71. NSFWE
68. NSGMM
67. NSJL
66. NSKMH
65. NSKS
SPWH
61. NSRTC
70. NSSRS
36. North, Oliver
1. Pearson, W. Robert
27. Poindexter, John M.
5. Powell, Colin L.
38. Ringdahl, Phil
40. Rodman, Peter
7. Ross, Dennis
46. Sable, Ronald K.
28. Stark, James R.
48. Steiner, Steve
39. Stevens, Paul S.
21. Teicher, Howard
37. Thompson, Paul
30. Tillman, Jacqueline
Content, Clustered by People

1. Lib
2. Libya Broadcasting
3. no subject
4. BENNETT DEBRIEF
5. More Libyan Fighting
6. BUSTO VISIT
7. Lib
8. BENNETT DEBRIEF
9. Free help
10. Free help
11. RR Mtg w/Oakley
12. Egypt status
13. Libya next steps
14. Israeli Spy Case
15. CRAZY GOING ONS IN COLOMBIA
16. Iraq Financial Situation
17. Libyan Terrorism
18. Lib
19. PRIVATE BLANK CHECK
20. no subject
21. PRIVATE BLANK CHECK
22. no subject
23. Further On Qadhafi's Travel
24. no subject
25. INTERESTING GOINGONS
26. Nicaragua Finding
27. no subject
28. no subject
29. Libyan terrorism
30. Anything New??
31. Anything New??
32. no subject
33. Special M
34. no subject
35. CONTRA PROJECT
36. no subject
37. CONTRA PROJECT
38. no subject
39. Iran
40. no subject
41. Public Affairs Campaign Central Ar
42. no subject
43. no subject
44. no subject
45. no subject
46. no subject
47. no subject
48. How's It Going?
49. no subject
50. IRAN AND TERRORISM
Outline

• Thanks!
  – You’ve addressed many important problems
  – Although we have a few more …

• Seeing the forest
  ➢ Studying a tree

• Finding that tree
  – Can it be Google?
  – Protecting sensitive content

• techNOLOGy
Visualizing Asynchronous Conversations

Perer and Shneiderman, InforVis 2005
Nobody Writes Letters Anymore: Helping people make sense of historically significant email collections

Tamer Elsayed and Douglas W. Oard

College of Information Studies and
UMIACS Laboratory for Computational Linguistics and Information Processing
Did Sheila want Scott to participate? Looks like the call will be too late for him.
**55 Sheila’s!!**

<table>
<thead>
<tr>
<th>Name</th>
<th>Name</th>
<th>Name</th>
<th>Name</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>weisman</td>
<td>maynes</td>
<td>jarnot</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pardo</td>
<td>nacey</td>
<td>kirby</td>
<td></td>
<td></td>
</tr>
<tr>
<td>glover</td>
<td>ferrarini</td>
<td>knudsen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rich</td>
<td>dey</td>
<td>boehringer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>jones</td>
<td>macleod</td>
<td>lutz</td>
<td></td>
<td></td>
</tr>
<tr>
<td>breeden</td>
<td>howard</td>
<td>glover</td>
<td></td>
<td></td>
</tr>
<tr>
<td>huckaby</td>
<td>darling</td>
<td>wollam</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tweed</td>
<td>watson</td>
<td>jortner</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mcintyre</td>
<td>perlick</td>
<td>neylon</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chadwick</td>
<td>advani</td>
<td>whanger</td>
<td></td>
<td></td>
</tr>
<tr>
<td>birmingham</td>
<td>hester</td>
<td>nagel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>kahanek</td>
<td>kenner</td>
<td>weights</td>
<td></td>
<td></td>
</tr>
<tr>
<td>foraker</td>
<td>lewis</td>
<td>gra make</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tasman</td>
<td>whitman</td>
<td>mclaughlin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fisher</td>
<td>berggren</td>
<td>venville</td>
<td></td>
<td></td>
</tr>
<tr>
<td>petitt</td>
<td>osowski</td>
<td>rappazzo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dombo</td>
<td>kelly</td>
<td>miller</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Robbins</td>
<td></td>
<td>swatek</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chang</td>
<td></td>
<td>hollis</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hi Shari,

Hope all is well. Count me in for the group present. See ya next week if not earlier.

Liza

Elizabeth Sager
713-853-6349

-----Original Message-----
From: SStack@reliant.com@ENRON
Sent: Monday, July 30, 2001 2:24 PM
To: Sager, Elizabeth; Murphy, Harlan; jcrespo@hess.com; whenze@jonesday.com
Cc: ntillett@reliant.com
Subject: Shhhh.... it's a SURPRISE!

Please call me (713) 207-5233
Thanks!

Shari

---Elsayed and Oard, CEAS 2006---
Hi Sheila,

For the new account funds will have to be in the account, before any trades being put on the account.

Sheung
Shane,
Laurel and I met with Sara Shackleton, Legal, yesterday.
Sara provided the following updates for Australia ISDA Agreements:
West Pac: Our end done, sent to them. Susan Flynn to follow-up to id where in process currently.
ANZ Bank: We need to know the exact name of the entity we want the agreement with so we can apply to the credit group to approve getting an agreement in place.

thanks, sheila

Enron Broadband Services, Inc.
1400 Smith, Suite EB-4573A
Houston, TX 77002
Shane,
Laurel and I met with Sara Shackleton, Legal, yesterday.
Sara provided the following updates for Australia ISDA Agreements:
West Pac: Our end done, sent to them. Susan Flynn to follow-up to id where in process currently.
ANZ Bank: We need to know the exact name of the entity we want the agreement with so we can apply to the credit group to approve getting an agreement in place.

thanks, sheila
Representational Model of Identity

77,240 non-trivial identity models

Elsayed and Oard, CEAS 2006
Judgment Process

Correct
Very Informative
Informative

Incorrect

Judged Associations

Wrong
kmpresto@msn.com <-> "home email"

Correct (but Useless)
june-rick@reliantenergy.com <-> “june rick”

OK
terriecovarru@hotmail.com <-> "terrie covarru"

Great
lemelpe@nu.com <-> "phyllis"
Did Sheila want Scott to participate? Looks like the call will be too late for him.
Great news. Lets get this moving along. Sheila, can you work out GE letter? Vince, I am in London Monday/Tuesday, back Weds late. I'll ask Sheila to fix this for you and if you need me call me on my cell phone.

Did Sheila want Scott to participate? Looks like the call will be too late for him.
Re: GE Conference Call has been rescheduled

Did Sheila want Scott to participate? Looks like the call will be too late for him.

ESA Option Execution

Can you initial the ESA assignment and assumption agreement or should I ask Sheila Tweed to do it? I believe she is currently en route from Portland.

Thanks,

Rebecca
Contextual Resolution

Elsayed, Oard and Namantha
ACL/HLT 2008

**Mathematical Expression**

\[
p(c \mid m, X(m)) = \frac{p(c, m, X(m))}{p(m, X(m))}
\]
## Test Collections

<table>
<thead>
<tr>
<th>Collection</th>
<th>Emails</th>
<th>Identities</th>
<th>Mention Queries</th>
<th>Min.</th>
<th>Avg.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sager</td>
<td>1,628</td>
<td>627</td>
<td>51</td>
<td>1</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>Shapiro</td>
<td>974</td>
<td>855</td>
<td>49</td>
<td>1</td>
<td>8</td>
<td>21</td>
</tr>
<tr>
<td>Enron-subset</td>
<td>54,018</td>
<td>27,340</td>
<td>78</td>
<td>1</td>
<td>152</td>
<td>489</td>
</tr>
<tr>
<td>Enron-all</td>
<td>248,451</td>
<td>123,783</td>
<td>78</td>
<td>3</td>
<td>518</td>
<td>1785</td>
</tr>
</tbody>
</table>
Which Context is the best?

- Social
- Topical
- Conversational
- Baseline
- Local

MRR

Sager Shapiro Enron-sub Enron-all
Outline

• Thanks!
  – You’ve addressed many important problems
  – Although we have a few more …

• Seeing the forest

• Studying a tree
  ➢ Finding that tree
    – Can it be Google?
    – Protecting sensitive content

• techNOLOGy
Text Retrieval Conference (TREC)

• Goals
  – Foster development of research communities
  – Create “benchmark” evaluation resources
  – Establish “baseline” results

• History
  – Sponsored by NIST since 1992
  – “Legal Track” started in 2006 with an E-Discovery focus
Federal Rules of Civil Procedure
(as amended 12/1/06)

Rule 26(f)

At the parties’ planning meeting, issues expected to be discussed include:

– “Any issues relating to disclosure or discovery of electronically stored information, including the form or forms in which it should be produced”

– “Any issues relating to preserving discoverable information”
From: Lisa Fislla

Subject: CIGNA Well-Being Newsletter - Future Strategy During our last CIGNA Aetna Plan meeting, the issue of whether to stop publishing articles and discontinue mailing CIGNA Well-Being newsletter to employees was under discussion. Several employees wanted to keep the newsletter and suggested alternative solutions for future communication. I believe everyone's input is valuable, and would appreciate hearing from each of you on whether you support my recommendation.

I have done some research and wanted to provide you with my findings and preliminary recommendations for PM's strategy involving CIGNA Well-Being newsletter. I am aware of the need to improve our communications and would appreciate hearing from each of you on whether you concur with my recommendations...

From: Lisa Fislla

Subject: CIGNA Well-Being Newsletter - Future Strategy During our last CIGNA Aetna Plan meeting, the issue of whether to stop publishing articles and discontinue mailing CIGNA Well-Being newsletter to employees was under discussion. Several employees wanted to keep the newsletter and suggested alternative solutions for future communication. I believe everyone's input is valuable, and would appreciate hearing from each of you on whether you support my recommendation.

I have done some research and wanted to provide you with my findings and preliminary recommendations for PM's strategy involving CIGNA Well-Being newsletter. I am aware of the need to improve our communications and would appreciate hearing from each of you on whether you concur with my recommendations...

Title: CIGNA WELL-BEING NEWSLETTER - FUTURE STRATEGY

Organization Authors: PMUSA, PHILIP MORRIS USA

Person Authors: HALLE, L

Document Date: 19970530

Document Type: MEMO, MEMORANDUM

Bates Number: 2078039376/9377

Page Count: 2

Collection: Philip Morris
A “Production Request”

RequestNumber: 52

RequestText: Please produce any and all documents that discuss the use or introduction of high-phosphate fertilizers (HPF) for the specific purpose of boosting crop yield in commercial agriculture.

Proposal: "high-phosphate fertilizer!" AND (boost! w/5 "crop yield") AND (commercial w/5 agricultur!)

Rejoinder: (phosphat! OR hpf OR phosphorus OR fertiliz!) AND (yield! OR output OR produc! OR crop OR crops)

FinalQuery: ("high-phosphat! fertiliz!" OR hpf) OR ((phosphat! OR phosphorus) w/15 (fertiliz! OR soil))) AND (boost! OR increas! OR rais! OR augment! OR affect! OR effect! OR multipl! OR doubl! OR tripl! OR high! OR greater) AND (yield! OR output OR produc! OR crop OR crops)

B: 3078
2006/07/08 Research Teams

Clearwell
Carnegie Mellon U
Dartmouth College
H5
Long Island U
Sabir Research, Inc.
U Buffalo
U Iowa (2 teams)
U Massachusetts
U Maryland
U Missouri, Kansas City

U Pittsburgh
U Washington
Ursinus College
CNIPA (IT)
Fudan U (CN)
National U of Singapore (SG)
Open Text Corporation (CA)
RMIT (AU)
U Amsterdam (NL)
U Waterloo (CA)
Deconstructing “Concept Search”

“Features” — “Method” — “Specification” — “Result”
Representing “Documents”

**Content**
- Count the words
- Weight the words
- Ascribe meaning to words

**Context**
- Who said this?
- When was it said?
- Who did they say it to?

**Description**
- What was said about it?

**Behavior**
- What was done with it?
Controlling the Search System

- **Proactive**
  - “Keyword” query
  - “Boolean” query
  - Query by example

- **Reactive**
  - Ranked list selection
  - “More like this” query
  - Category exploration
  - Social network exploration

- **Iterative**
  - Query refinement
  - “Search within” query
Generating Results

- Logic
- Similarity
- Probability

**Features**

**Method**

**Specification**

- Result set
- Ranked list
- Classification
- Clustering
- Visualization
2006/07 “Relevancy” Assessors

Bank of America
Department of Justice
FTI Consulting
H5 Technologies Inc.
NARA
Lewis & Roca LLP
New Mexico Attorney General
Preston Gates LLP
Reasonable Discovery LLC
SAIC
Private individuals (CA, UK)

Law Schools
Boston University
Case Western Reserve
George Mason
George Washington
Loyola-Los Angeles
Loyola-New Orleans
U Dayton
U Indiana-Indianapolis
U Maryland
U Texas
Boolean Run Estimated Recall

Mean EstR@B: 0.22
- Boolean run missed 78% of the relevant documents (on average per topic)

Topic 84 (1960’s films)  
EstR@B=100%

Topic 77 (smoke NOT tobacco)  
EstR@B= 0%
How to retrieve corrupted documents?

Expand query with author and recipient names
Protecting Sensitive Content

• Pre-review becoming increasingly impractical
  – End-user review is hard to enforce
  – Volumes growing too large for post hoc review

• No automated system will be good enough
  – Precision-recall tradeoff

• On-demand review may be viable alternative
  – Significant implications for workflow and tools
Outline

• Thanks!
  – You’ve addressed many important problems
  – Although we have a few more …

• Seeing the forest

• Studying a tree

• Finding that tree
  – Can it be Google?
  – Protecting sensitive content

➢ techNOLOGy
TechNOLOGY

• Changing technology landscape

• Many cognate professions
  – Electronic records
  – E-discovery / e-disclosure
  – Law librarians
  – Forensic analysis (e.g., for law enforcement)

• “iSchools”