Extracting Conversation Timelines from Email

by Philip Greenspun, March 5, 2012, updated April 24, 2013

This is a specification for a software product that would be useful in litigation, e.g., in handling discovery requests for email conversations, and that also would be useful in business and personal situations when one's goal is simply to review what has been previously discussed.

Big Picture

Given a body of email, it would be nice to construct a timeline of conversations relevant to a particular topic. Each entry in the timeline would be a hyperlink to the actual conversation. This may be very useful when an organization is trying to figure out how a decision was made or in litigation.

The big challenges here are making it convenient to group conversations by subtopic and making it easy to exclude irrelevant or trivial conversations, e.g., ones that are purely regarding the logistics of scheduling a meeting.

Example Output (organized by topic and then by conversation)

Note that the organization by topic is done by the human user, who picks what topics are of interest and classifies conversations into "Topic A", "Topic B", and "Irrelevant" (the irrelevant ones, e.g., trying to schedule a lunch, don't get into the report).

Contract Negotiations

- 3/15/2009: request for quotation (link to Gmail-style conversation)
- 4/19/2009: proposal accepted (hyperlink to conversation from this and all subsequent)
- 4/25/2009: first draft of contract
- 5/2/2009: comments on first draft (hyperlink to Gmail-style conversation that may include responses at later dates than 5/2)
- 5/7/2009: second draft of contract
- 5/12/2009: exchange of signed copies

Corrosion Resistance

- 2/2/2009: question regarding corrosion of fasteners (link to Gmail-style conversation including the answer to the question)
- 3/18/2009: question regarding warranty in the event of corrosion
- 7/16/2009: complaint regarding corrosion around screws

Example Output (individual emails)

Contract Negotiations

- 3/15/2009, 9:30 am: request for quotation (link to one email)
- 3/17/2009, 4:15 pm: quote (link to one email)
- ...

Display Options

Options include at least the following:

- show names of correspondents
- show email addresses (from/to) on each line
- show times as well as dates

Workflow

Step 1: Define all of the emails that might become part of the timeline. This could include expressions such as "all emails from philg@mit.edu" or "all emails from philg@mit.edu to billg@microsoft.com" or "all emails containing 'project athena' in the subject or body".

Step 2: Categorize/exclude conversations. Selecting a conversation should also highlight ones with a similar subject line or similar words in the body (perhaps the best approach may be to use the subject of the selected email as a search string into the bodies of other emails). The system would then offer the opportunity to categorize all of those selected into a subtopic, possibly a new one, or to exclude them from the timeline (can be considered the "irrelevant" category).

Step 3: Review the timeline. Similar format to the final report, but editable to recategorize a conversation and with the option to show excluded emails in gray (for consideration to hoist them back up into a category other than "irrelevant").

Print Options

The above specification is for display on a Web page. If a print version is desired it would be ideal to have conversations printed more or less as they are currently done by the Gmail system. A new page would be started for every new conversation. For discovery in litigation lawyers

would prefer to have the output be PDF but plain text would also be acceptable and then the files could be turned into PDF and Bates-numbered by Adobe Acrobat Pro.