WATERFORD
TECHNOLOGIES

Highly Efficient eDiscovery Using Adaptive Search Criteria and Successive Tagging

[TREC 2010]

by

Ron S. Gutfinger

12/3/2010

1. Introduction

1.1. Abstract

The most costly component in eDiscovery is the manual review of documents. While this effort can be reduced using sampling, it is still significant and very labor intensive when the data volume is high (e.g., over one million documents). A method is presented here that minimizes the review time needed to achieve high quality eDiscovery, which in turn reduces the overall cost.

Another critical part of eDiscovery is bridging the gap between the investigator (the person defining the searches) and the attorney on the case. While the investigator may have a thorough understanding of the search technology and how to perform effective searches, he lacks the domain and case knowledge the attorney has regarding what deems a document responsive. In addition, the content and language (choice of words) of the document set is unknown upfront to both individuals. The expertise of both people needs to be combined effectively to deliver quality results efficiently.

The algorithm presented here manages and facilitates the discovery process. It integrates effectively document review feedback provided by the attorney in each iteration. Specifically, searches are adapted towards responsive documents. In essence, discovery is influenced and efficiently directed by document review. The resulting algorithm bridges the gap between the investigator and the attorney.

Responsive documents are discovered with an iterative technique using computerized searches in conjunction to quick manual reviews of resulting message titles. Intermediate results are tracked by tagging messages in application UI.
Specifically, in each iteration, a result set is tagged. Then, based on the results review, additional search criteria are defined for false-positives as well as stronger hits. Subsequently, the new criteria are used to search and generate a finer results set. Additional tags are applied to this set. The iterating is repeated until the desired results quality is attained or is limited based on time and resources.

1.2. TREC Background

The Text REtrieval Conference (TREC) facilitates an environment for researching information retrieval methods for real-world, industry scenarios [1]. For this year, 2010, the public Enron messages served as the source of data to analyze. A mock case relating to an oil spill has been defined. Then, using the data the case has been investigated.

1.3. Nomenclature

FP	= False Positive
MS	= Microsoft

1.4. MailMeter and Microsoft SQL Search

MailMeter is an Email archive and search application suite developed by Waterford Technologies. Its eDiscovery application is called Investigate.
MailMeter leverages Microsoft SQL Server full-text search to search documents; specifically, queries using the CONTAINS keyword are performed [2].
For example: Below is a query against a table of message records (MessageDetails), which returns documents containing the phrase ‘oil spill’.

```
SELECT MessageID FROM MessageDetails
        WHERE Contains(BodyText,'"oil spill"')
```

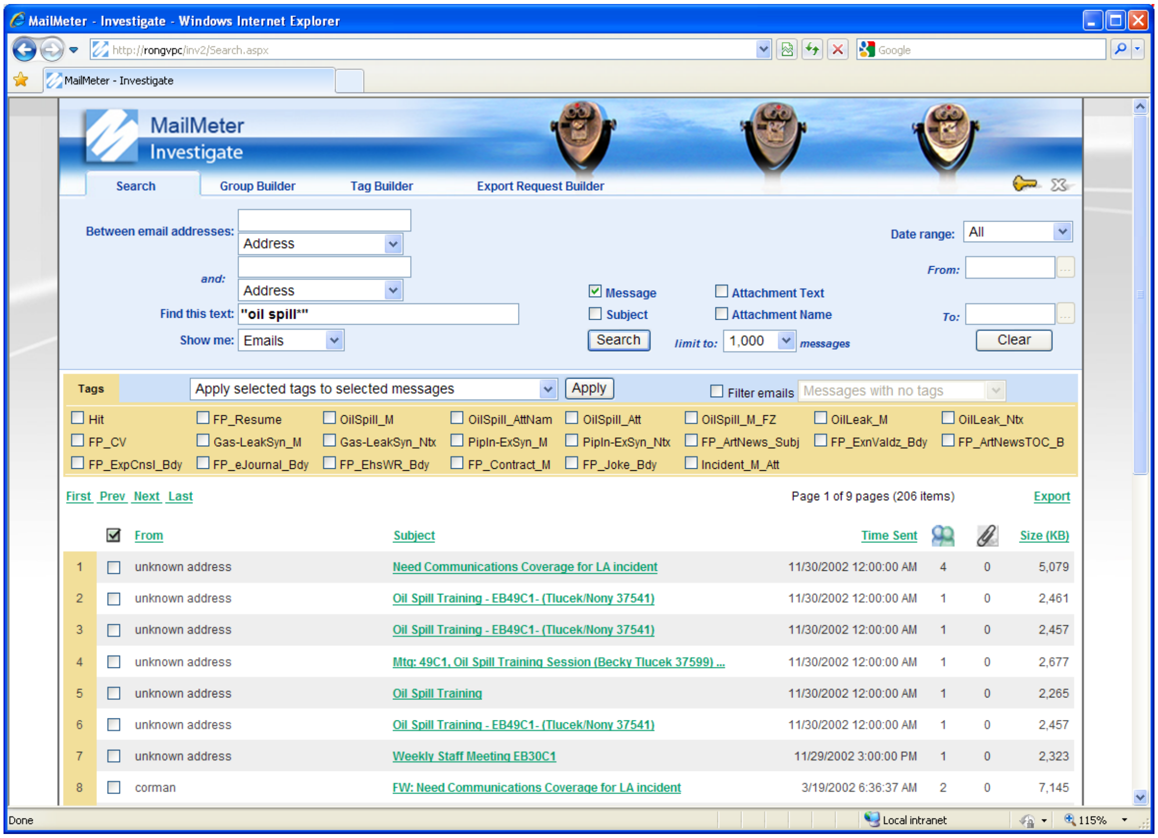
Note: MessageID is the primary key column and BodyText the text column

1.5. MailMeter Investigate UI and Tagging

A tag is an attribute or property that a user can apply to a mail message to identify it (e.g., Confidential, Internal).
Below is a screenshot of the MailMeter Investigate application Search screen. It is composed of three sections:

- Search criteria
  - The search phrase used is typed in the text box following “Find this text” and is searched in the selected components (e.g., Message and Attachment Text checkboxes).
- Tags
- Result set

The user can apply any of the selected tags (in Tags section midway through the screen) to messages in the result set. In addition, the user can select tags to filter the results (see checkbox to right of Apply button in center of image).



Between email addresses: Address

and: Address

Find this text: "oil spill*"

Show me: Emails

☑ Message ☐ Subject ☐ Attachment Text ☐ Attachment Name

limit to: 1,000 messages

Date range: All

From:

To:

Tags: Apply selected tags to selected messages

☐ Filter emails: Messages with no tags

☐ Hit ☐ FP_Resume ☐ OilSpill_M ☐ OilSpill_AttNam ☐ OilSpill_Att ☐ OilSpill_M_FZ ☐ OilLeak_M ☐ OilLeak_Ntx
☐ FP_CV ☐ Gas-LeakSyn_M ☐ Gas-LeakSyn_Ntx ☐ PipIn-ExSyn_M ☐ PipIn-ExSyn_Ntx ☐ FP_ArtNews_Subj ☐ FP_ExnValdz_Bdy ☐ FP_ArtNewsTOC_B
☐ FP_ExpCnsl_Bdy ☐ FP_eJournal_Bdy ☐ FP_EhsWR_Bdy ☐ FP_Contract_M ☐ FP_Joke_Bdy ☐ Incident_M_Att

Page 1 of 9 pages (206 items)

	☑	From	Subject	Time Sent	👥	📎	Size (KB)
1	☐	unknown address	Need Communications Coverage for LA incident	11/30/2002 12:00:00 AM	4	0	5,079
2	☐	unknown address	Oil Spill Training - EB49C1- (Tlucek/Nony 37541)	11/30/2002 12:00:00 AM	1	0	2,461
3	☐	unknown address	Oil Spill Training - EB49C1- (Tlucek/Nony 37541)	11/30/2002 12:00:00 AM	1	0	2,457
4	☐	unknown address	Mtg: 49C1, Oil Spill Training Session (Becky Tlucek 37599) ...	11/30/2002 12:00:00 AM	1	0	2,677
5	☐	unknown address	Oil Spill Training	11/30/2002 12:00:00 AM	1	0	2,265
6	☐	unknown address	Oil Spill Training - EB49C1- (Tlucek/Nony 37541)	11/30/2002 12:00:00 AM	1	0	2,457
7	☐	unknown address	Weekly Staff Meeting EB30C1	11/29/2002 3:00:00 PM	1	0	2,323
8	☐	corman	FW: Need Communications Coverage for LA incident	3/19/2002 6:36:37 AM	2	0	7,145

Figure 1 MailMeter Investigate search page

2. Algorithm

The eDiscovery algorithm discussed here is depicted in the diagram below. Each step is detailed on the subsequent sections.

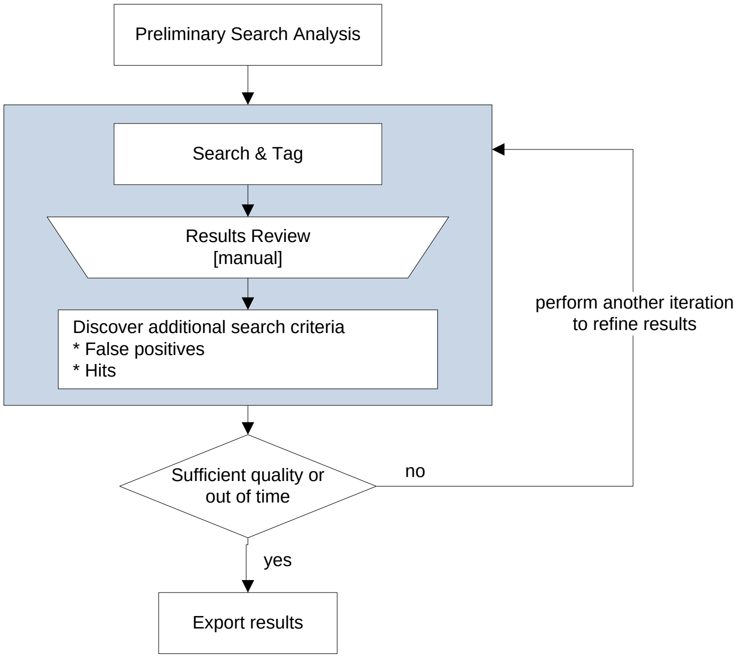


Figure 2 eDiscovery algorithm

2.1. Preliminary Search Analysis

2.1.1. Candidate Search Criteria

Using legal case description, candidate search criteria are derived.

- E.g., “oil” near “spill” in message or attachment
- Phrases
  - E.g., “oil spill”
- Generalization of phrase to expand result set [3]
  - Use * (e.g., “oil spill*” – expanded hit list)
  - Use synonyms (e.g., “oil spill*” or “oil leak*”)
  - The word ordering could be reversed (e.g., “leaking oil”).

2.1.2. Exploring Search Criteria

Searches are run using the candidate criteria and based on the respective result sets it is decided which criteria are of interest. These searches are exploratory in nature.

The emphasis here is on ‘coverage’ (recall) rather than ‘accuracy’ (precision). It is more important to get all relevant documents than getting a low percentage of irrelevant documents. This is because any documents that are not returned by the search at this point would be missed.

- For example, while searching for a particular phrase (e.g., “oil spill”) in attachment name may yield a high success rate, it is likely to exclude most of the responsive documents. Specifically, the phrase is expected to be found in attachments text but unlikely to be in each respective attachment’s name.

The goal is to expand the results set but with reason.

- For example, if a phrase or criterion expands the results set three-fold but all the sample documents reviewed (e.g., 3% of the set) are found to be irrelevant, then this criterion should not be employed.
    - Using the *near* construct (e.g., “oil” near “blowout”) was found to be too broad a criterion and of limited value because MS-SQL server does not provide proximity control and may return document with distant words.
        - Example: *Near* yielded documents having distant search terms such as the one below:
            - e.g., *“A classic example of 'buy the rumor, sell the fact' was seen Friday when Crude **__Oil__** futures sold off on confirmation of OPEC production cuts.”... ....HOLIDAY SWEEPSTAKES -- LAST CALL -- FREE ENTRY -- WIN A $500 PERSONAL SPENDING SPREE **__BLOWOUT__** !!!”*
                - This document satisfies the search criteria but has no relevance to the case at hand.

2.1.3. Finalizing Search Criteria

After the exploratory phase, the final search criteria are identified.

2.2. 1st Iteration

2.2.1. Search & Tag

In this step, a search is run for each phrase, and the respective results are tagged.
The tags used need to be meaningful enough to describe the reasons the document was selected. The naming convention used is as follows:

- Tag_Msg - includes results based on searching message, subject, and header
    - E.g., “OilLeak_syn_Msg”
- Tag_Ntx - includes results based on searching attachment name and text
    - E.g., “OilLeak_syn_Ntx”

2.2.2. Efficient Review of Results [Subject/Doc-80/5]

The results are evaluated by reviewing a sample of documents using the Investigate application UI. This is sufficient to gauge quality [4].

For efficiency purposes, documents (message bodies and attachments) are not reviewed initially. Instead only the subject is reviewed first. Then, based on the subject the document may be reviewed as well; even so, it may be reviewed only in part to save time.

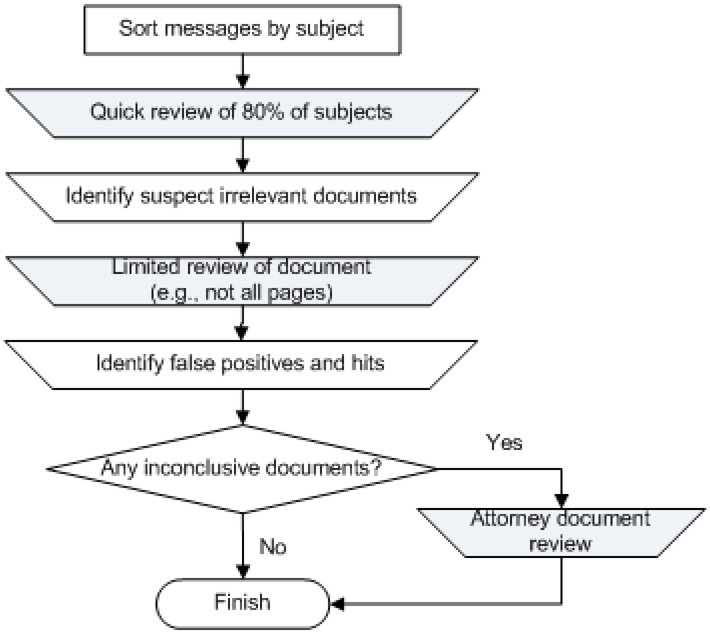

Sort messages by subject
Quick review of 80% of subjects
Identify suspect irrelevant documents
Limited review of document (e.g., not all pages)
Identify false positives and hits
Any inconclusive documents?
Yes
Attorney document review
No
Finish

Figure 3 Efficient review of results

As shown in the figure above, the review process is comprised of these steps:

- Before starting the review, sort the messages by subject.
    - This brings efficiency as it makes duplicates and overlaps more visible.
- Quick subject review
    - Review quickly 80% of subjects (a fraction of a second per subject)
        - Identify suspect irrelevant documents
- Document review
    - Review 5% of suspect documents (message body or attachment body)
        - To minimize review cost, start by reviewing only part of a document (e.g., only selected paragraphs or pages).
        - Identify false positives
        - Identify sure hits
    - If unsure of any document relevance, confirm with attorney on case (Topic Authority)

Note: The above review is defined here as a *subject/doc-80/5* review.

2.2.3. Discover Next Iteration Search Criteria

- Using false positives (FP) found above, identify second set of criteria on tagged documents that deem document as not responsive.
    - Each criterion, may involve keywords and/or phrases
- Using sure hits found above, identify second set of criteria on tagged documents, not having false positive phrases; deeming documents as more likely to be responsive (a stronger hit).
    - Each criterion, may involve keywords and/or phrases

2.3. 2nd Iteration

2.3.1. Search & Tag

- Search documents tagged from Iteration1 with FP phrases, and tag as FP (e.g., FP_*iteration_phrase*)
- Search documents tagged from Iteration1 with Hit phrases and no FP tags, and tag as Hit (e.g., Hit_*iteration_phrase*).
  Note: The selected documents are still open to elimination in subsequent iterations as search criteria evolve.

The result set is now reduced and is more likely to be responsive due to added criteria. Its search criteria *adapted* to the review findings.

All tags are applied in addition to previously applied tags. The collection of tags on a given document illustrate its discovery history. At a later time, when the document is revisited, the investigator can quickly piece together the reasons for the document's relevance or irrelevance as well as get a summary of its content pertaining to the case. The figure below presents a message with three tags portraying the history of findings in three iterations.

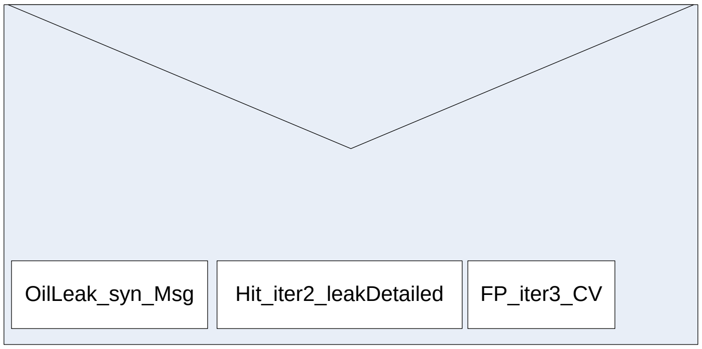


Figure 4 Example message with three tags (iter=iteration)

2.3.2. Review Results – Documents Excluded

- Use same subject/doc-80/5 review strategy
  - Assess each FP criterion accuracy and potentially:
    - Retain FP criterion
      - i.e., criterion validated over a large document set
    - Refine FP criterion
      - i.e., to apply to a large document set, criterion may need additional conditions
    - Withdraw FP criterion
      - i.e., criterion is invalid or inapplicable to a large set of documents

Note: Since the document set coming into this review is smaller than the one in the previous iteration, the sample needing manual review is proportionally smaller. Therefore, this review will be less time consuming and less costly.

2.3.3. Review Results – Documents Selected

- Same review as above with Hit phrases

2.4. Export or Continue Iterating

At this point, if result quality is sufficient or there is no time remaining, the final result set is exported in PST format which is then forwarded to the requesting attorney. Conversely, if quality is insufficient, additional iterations are performed.

Although, the results quality is unknown since a very small set of messages is reviewed, it can be discerned based on the impact of the search criteria on the results set. The investigator can get a sense of convergence as the iteration's impact is reduced.

For example: if at a given iteration, the FP criteria yield a 30% reduction of the result set, further iterating should be time-worthy. However, if the FP criteria yield a 0.5% reduction of the result set, additional iterating may be of little value.

3. Process Applied to Case

For TREC2010, the public Enron messages served as the source of data to analyze. The algorithm was used to investigate Topic 302, which is focused on oil spill related documents.

3.1. Search Criteria

- The search criteria used are these exact phrases and their synonyms (appearing in message or attachment)
    - "gas leak*" or "gas spill*" or "gas blowout*" or "gas release*"
    - “oil leak*” or “oil spill*”
    - "pipeline erupt*" or "pipeline rupture*" or "pipeline explod*" or "pipeline explos*"
        - This criterion includes “pipeline erupted”, “pipeline eruption”, “pipeline exploded”, and “pipeline explosion”.

3.2. 1st Iteration

The searches using the criteria above yielded 1,374 messages of the 1.16 million Enron messages (exact total count was 1,161,516). This is 0.118% of the messages.

3.3. 2nd Iteration

3.3.1. FPs

Here are some key false positives' criteria identified in the reviews for the case at hand.

Many messages that contained the phrase oil spill were resume submissions by job applicants. The criteria to identify these false positives were:

- Body criterion: "CV" and "attach*" and not "oil spill"
  AND
- Attachment criterion: "experience" or "qualification*"

Some of the messages containing the phrase oil spill were news articles. The criteria to identify them were:

- Subject: "article*" or "news*" or “CSIS watch” or "Petroleumworld weekly review" or "e-journal"
  Note: “news*” covers ‘newsletter’.

The searches using all the FP criteria reduced the message count to 942 (of 1,374).
This is 68.6%; i.e., over 30% reduction.

3.3.2. Hits

Below are observations per the reviews of responsive documents.

- These words indicate internal correspondence - not some external report/newsletter
  - Confidential, Litigation, Memorandum or Memo, "Enron Litigation Unit", "Litigation Unit"
- These words indicate an issue or a problem
  - Emergency, aftermath, “urgent*”, “alert*”, crisis, “issue*”, “incident*”, “accident*”, “concern*”, “risk*”, liability, “claim*”, compliance, “damage*”, “cause*”, disaster.
- These words imply activity relative to a leak or problem
  - Fixed or "fix*" (e.g., oil leak fixed), "repair*", follow-up, Operations (they usually get involved), “shut*” (e.g., shut down), "response plan", “clean*” (e.g., cleanup), remediation, "contingency plan", recovery.
- These words describe symptoms of the problem
  - “barrel*” (describes size of leak), “rainbow*” (patterns), "discharge volume".

Due to time constraints, the above criteria were simplified into this single criterion:

- Body or Attachment contain: ("emergency" or "urgent" or "aftermath" or “crisis" or "issue" or "concern" or "incident" or "risk" or "liability" or "operations" or "confidential" or "litigat*" or "memo*" or "claim*" or "damage*" or "cause*" or "fix*" or "repair*" or "clean*" or "follow-up" or "shut*" or "remediation" or "disaster" or "contingency plan" or "recovery")

The searches using all the Hit criteria reduced the message count to 896 (of 942).
This is 95.1%; i.e., about 5% reduction.

3.4. Summary Graph

The results are depicted in the graph below. It displays how the message count decreases at the different steps of the process.

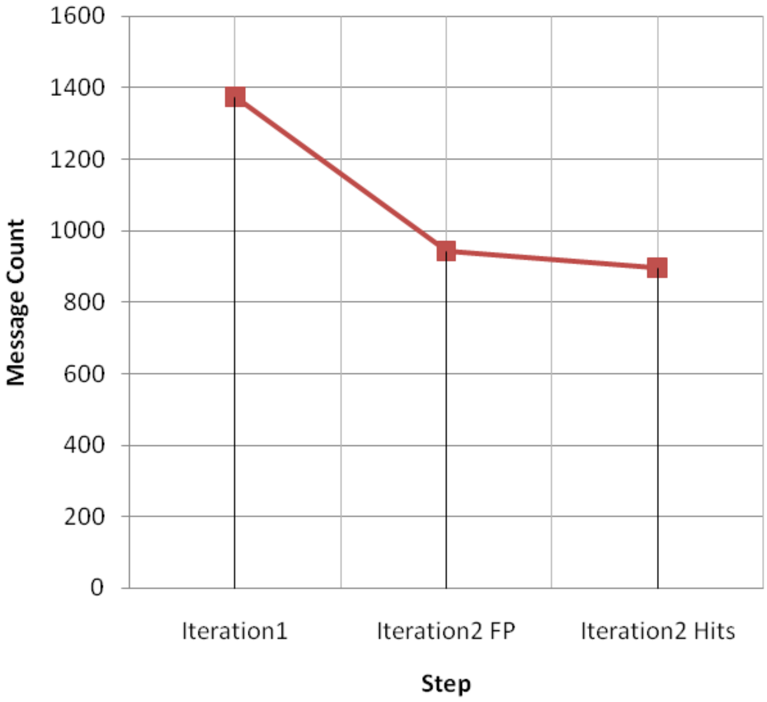


Figure 5 MessageCount vs. Step

4. Conclusions

A method has been presented to efficiently discover responsive documents.
The algorithm facilitates the discovery process and integrates effectively document review feedback provided by the attorney. In addition, it minimizes the review time, which reduces cost. Lastly, the technique is flexible enough to allow for replacing its core piece of search engine and allows for iterating per allocated time.

5. Acknowledgments

A special thanks to Lorcan Kennedy (CTO of Waterford Technology) for managing the MailMeter import process, which involved 1.2 million messages.

6. References

[1] TREC web site: http://trec.nist.gov/overview.html.

[2] "T-SQL Reference," Sql Server 2008 Books Online, Microsoft Corporation, 2008.

[3] Baron, Jason R., Braman, Richard G., and Withers, Kenneth J., "Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery," The Sedona Conference Journal, Volume 8, Fall 2007, pp. 189-223.

[4] Baron, Jason R. and Burke, A. Macyl, "Commentary on Achieving Quality in the E-Discovery Process", The Sedona Conference, May 2009.