Study: eDiscovery + Enterprise Search = Reduced Costs

By Marisa Peacock (@marisacp51) Aug 10, 2010

ZL TECHNOLOGIES INC.

When it comes to collecting and sifting through data for litigation purposes, companies often begin by searching for terms relevant to them at the time, only to return again and again for new relevant keywords. With more and more advanced eDiscovery solutions on the market, many claim to speed up the collection process by searching for the words you didn't know you needed.

But have you ever wondered which approach to culling data for eDiscovery approaches was more efficient? Well, ZL Technologies did and they've got research to back it up.

The report, Comparing Exclusionary and Investigative Approaches for Electronic Discovery using the TREC Enron Corpus, is quite scientific in it's attempt to investigate whether the **data culled from limited document retrieval based on custodian email mailboxes** results in lower recall and produces fewer responsive documents than a **broader, inclusive search process that covers all potential custodians.**

CMS WiRE

Long Story Short

Researchers found **significantly more responsive documents (516% more, to be exact) and initial custodians (1825 % more!) were found when an entire data set was searched**, rather than an approach that relied on custodian-based culling. Ultimately, these results not only show the merit of searching an entire data set at once, but also that companies who employ an “exclusionary, culling-based methodology,” which often requires frequent and subsequent collections, risk not producing enough information despite spending a lot more time and money doing so.

The Long Story: Don't Be Like Enron

While at first glimpse the research results may not be earth shattering, but consider the data set the researchers used to make their point. **Enron.**

That’s right, for their experiment, they prepared two sources of data, an email corpus, containing an Enron data set prepared and distributed by the 2009 TREC Legal Interactive Track, and a custodian list, created from an “ex employee status report”.

To create the data set that would be used in their investigation, the organized it by custodian, yielding 2,965,103 email messages spread across 104 custodians. Then they identified eight distinct dates associated with individual data collection times from August 2000 through March 2002, from which they were able to outline specific complaints filed against Enron, notably:

- **August 2000:** San Diego Gas & Electric Company files a complaint against Enron alleging market manipulation; an event that likely triggered the initial collection.
- **March 2002** is one month after FERC began their investigation into Enron’s involvement in the Western U.S. Energy Crisis.

The researchers created two non-overlapping teams. One team selected a group of custodians to review to simulate the exclusionary approach while the other team performed search and Information Retrieval (IR) on the entire data set to simulate the investigative approach.

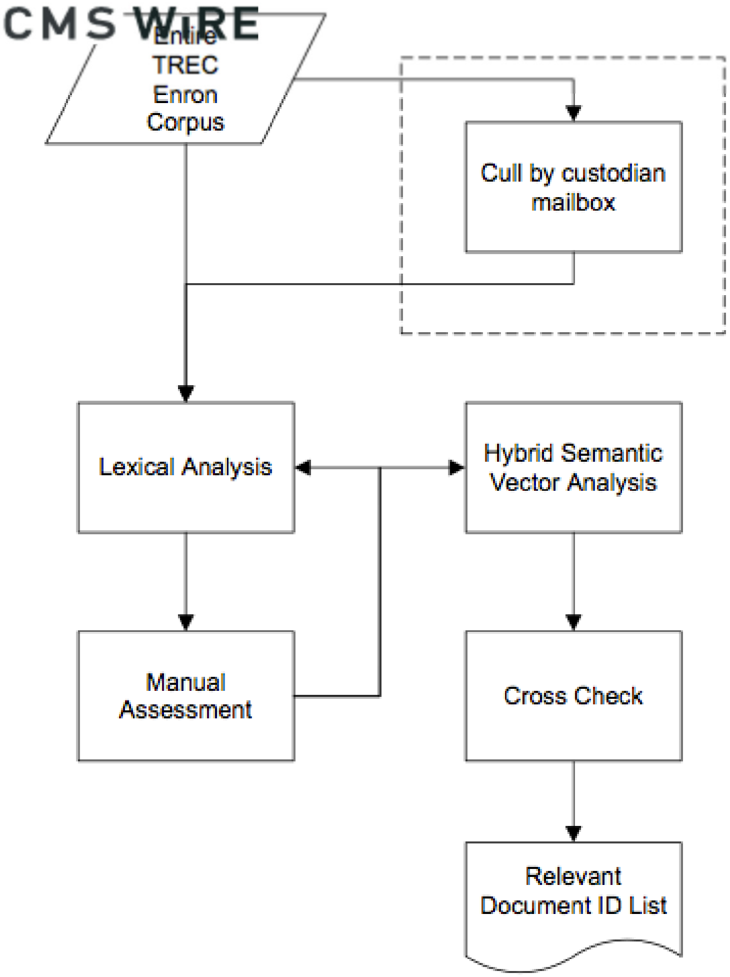
CMS WIRE
Entire TREC Enron Corpus
Cull by custodian mailbox
Lexical Analysis
Hybrid Semantic Vector Analysis
Manual Assessment
Cross Check
Relevant Document ID List

Figure 1. The methodology used by researcher comprised several IR techniques which refined the data set and narrowed down the documents for manual review. The model above outlines the overall process.

The results highly favored the broader approach, so much so that even if the exclusionary team had selected the four custodians with the highest number of responsive documents, the approach would have still overlooked over half of the responsive documents identified by the investigative team.

By cleverly using a company synonymous with audit failure, ZL Technologies and its team of researchers make a convincing plea for organizations to re-evaluate their eDiscovery approach or else risk missing substantial amounts of documents by relying solely on