

Network-based filtering for large email collections in E-Discovery

Dr. Hans Henseler¹
j.henseler@hva.nl

*Global E-Discovery / E-Disclosure Workshop DESI III.
June 8th, 2009. Barcelona, Spain*

Abstract

The information overload in E-Discovery proceedings makes reviewing expensive and it increases the risk of failure to produce results on time and consistently. New interactive techniques have been introduced to increase reviewer productivity. In contrast, the research that is discussed in this paper proposes an alternative method that tries to reduce information during culling so that less information needs to be reviewed. The proposed method first focuses on mapping the email collection universe using straightforward statistical methods based on keyword filtering combined with date time and custodian identities. Subsequently, a social network is constructed from the email collection that is analyzed by filtering on date time and keywords. By using the network context we expect to provide a better understanding of the keyword hits and the ability to discard certain parts of the collection.

1. Introduction

E-Discovery is defined as the selection, processing and production of electronic stored information (ESI). This process is illustrated by the E-Discovery Reference Model (EDRM, Socha-Gelbmann, 2006). The well-known EDRM diagram presents an overview of the E-Discovery process and the colored slopes in the background symbolize the transformation of a large volume of general ESI into a small volume of specific and relevant information. The

¹ Lector e-Discovery, Amsterdam University of Applied Sciences, Department of Computer Science and Director Forensic Technology Solutions, PricewaterhouseCoopers Advisory, Netherlands.

automated part of this process is called culling and is aimed at selecting and filtering information. Current E-Discovery products use a variety of culling techniques mostly for removing duplicates, filtering based on file extension, date time and or keywords in order to reduce the volume of ESI. Keyword filtering relies on a set of keywords that is designed by lawyers or investigators based on the context of the investigation, i.e. names of persons, projects, places, companies etc. The remaining documents can then be reviewed by reviewers that have to identify which documents are hot, i.e. documents that should be produced as evidence.

The E-Discovery process has to deal with an explosion of information and the problem of finding relevant information is further amplified as content is more easily generated in a large variety of formats without adding significantly new information (Paul and Baron, 2007). This information overload makes reviewing very expensive and also increases the risk of failure. Already search queries are not ideal and experiments have shown that finding relevant documents using keyword search is far from perfect (Krause, 2009). New techniques are being introduced so that reviewers can review documents faster by using more powerful tools such as conceptual search (see e.g. Chaplin, 2008), detection of near duplicates and visual analysis (see, e.g. Görg and Stasko, 2008). Another approach is to enhance existing culling strategies in order to restrict the volume of information that needs to be reviewed. However, in this paper we propose to introduce statistical analysis and social network analysis based to improve culling of emails.

2. Structured information in emails can improve culling

As the total volume of unstructured ESI increases, the E-Discovery process becomes an increasingly difficult challenge. More advanced techniques are entering the market for E-Discovery products to increase review effectiveness such as detection of near duplicates, email threads and conceptual clusters. The first two techniques enable reviewers to review related documents in one pass improving consistency and review speed. The third technique allows reviewers to discover document categories through concepts and may be useful to exclude irrelevant document categories or prioritize hot categories. These techniques are based on language processing in order to extract meta data from unstructured information allowing reviewers to browse through the data in combination with full-text searches. The quality of full-text retrieval can be improved with language processing to enhance legal discovery (Bobrow, 2007), also in the culling stage. However, language processing can be slow with large search index sizes. Also it is error prone because it is language dependent, for instance, when using synonyms or when extracting named entities. Similarly, unsupervised clustering using extracted

concepts often results in non-relevant document categories that are ignored by reviewers during review.

We suggest using the structured nature of email to enhance legal discovery. In many cases ESI primarily consists of email messages and attachments. These messages are not entirely unstructured. Each email has a header identifying sender, receivers, recipients, date, subject and attachments. The three techniques introduced earlier are general purpose and are based on text mining in unstructured information. Using the structured information of emails, we may be able to further optimize the culling process without having to resort to text mining. The research presented here introduces the application of several statistical analysis and social network analysis techniques to increase the effectiveness of culling emails and their attachments. After a short discussion of these techniques we conclude with recommendations for future research.

3. Statistical analysis of emails

Surprisingly (or may be not surprising at all), very few organizations have centralized email archives that are considered complete enough for discovery purposes. Consequently in E-Discovery projects emails are typically collected from personal email archives found on file servers. Users maintain such personal archives because their mailbox size on the server is limited, varying from 100Mb to a couple of Gb. A backup of this data is available for disaster recovery and a typical organization will have several monthly and yearly backup tape sets available offline. To ensure completeness in the E-Discovery process, all available backups are processed. This results in many duplicates and removing duplicates is an important part of culling.

Knowing that this is the way how emails are collected, it is best practice to assess the completeness of the collected information (after extracting emails and attachments from the archives and removing duplicate items). One approach is to count emails per custodian per time period, e.g. week, month or year. Such an overview can be created by, for instance, running a pivot table on a list of email records containing custodian name and sent date of the email. This table can be compared against information from the personnel department indicating when a person joined or left the organization. The overview may also reveal holiday periods, work patterns etc.

State of the art E-Discovery tools not only extract emails from archives, attachments from emails and store email header information in a database but also create a full-text search index of all email content. The full-text search index is used to filter emails using a keyword list in combination with a filter on the email fields, e.g. a time window. The keyword list is carefully constructed (see for instance Reeves and May, 2008) and typically consists of names of persons, projects, numbers etc. The list is actually perceived as a list of relevant research topics and reviewers are assigned to different topics. To optimize the keyword list, typically a hit count per keyword is generated to identify if keywords are useful.

Another approach might be to create a pivot table on the number of emails per topic per period. This kind of analysis provides a topic map showing the evolution of topics in the course of time. This strategy is also used, for instance, with historical analysis of newspapers looking for trends or specific events.

4. Discovering communication patterns

The structural information of emails (to, from, cc fields) can be presented as a network of communication between persons. In this network nodes (vertices) represent persons and links between the nodes (edges) represent emails. A subset of nodes corresponds to the set of custodians. The edges can be labeled, for instance, with the number of emails. Some E-Discovery tools provide tools to investigate email networks but this is mainly restricted to a manual analysis using a visualization of a cross-section of the network that is typically limited to one person and its direct contacts. If we want to use the network information for culling purposes we need to use more advanced algorithms that provide objective criteria by which emails can be removed before reviewers start with their manual analysis.

Social Network Analysis (SNA) is the mapping and measuring of relationships and flows between people, groups, organizations, computers, web sites, and other information/knowledge processing entities. SNA defines a number of centrality measures to objectively evaluate the role of persons in the network (Scott, 1991). The three most popular centrality measures are degree, closeness and betweenness. Degree centrality is simply calculated as the number of links to or from a person. Closeness centrality is the average number of links it takes a person to reach any other person in the network. Betweenness centrality measures how many times a person is in the shortest path between any two other persons in the network. More advanced measures such as the eigenvector centrality take into account the importance of neighboring persons when

calculating the centrality of a person (cf. like Google rates web pages based on the ratings of pages that refer to the page).

Persons with a high degree centrality are identified as hubs. In E-Discovery custodians typically appear as hubs because the data was collected from their mailboxes. If there is a hub in the email network that is not a custodian it may be interesting to investigate why this person is not a custodian. Similarly persons with a high degree of betweenness can be important because they have a 'broker' role between different networks. Such persons may have access to multiple networks and have valuable other information or access to other unknown networks. These measures of centrality give objective criteria that can be used to evaluate the structure of the email collection and to determine if certain parts of the collection can be discarded or if the collection is insufficient.

The centrality measures give an overview of general statistics and their use is limited. The specific email pattern of a user can also reveal aliases of a user. Custodians may use more than one email address to communicate. This might be intentionally to hide another identity but more often this is due to a change of email server or the use of different email address formats, e.g. internal versus external address format. Social network similarity may be used to identify if two different email addresses might belong to the same person. Resolving aliases is an important step in reducing the complexity of the network and increasing review consistency.

Social network analysis can also be performed on cross sections of the email collection by using keyword filtering combined with date-time intervals. By analyzing the email network in different time slices, the centrality measures mentioned above can be compared over time. Similar to the topic map introduced earlier, it might be useful to restrict the email network to certain topics and then study the effect this has on the role of different persons. It could be that a person is a hub on a specific topic while he or she has a more standard role on another topic. With a full-text filter it is possible to decompose a complicated network in underlying smaller and less complicated topic-based networks.

In case specific events are being investigated, e.g. the merger of two companies or the negotiation of a commercial contract, combining a time window with a full-text filter can be helpful to discover which persons have received emails that are related to the event (based on a

full-text search filter). Once such a network is identified, the date time restriction may be changed to discover other topics that are discussed within this particular network.

5. Conclusions and recommendations

A new method has been proposed to improve culling in E-Discovery by taking advantage of structured information in email messages. First a statistical analysis should be performed by creating a pivot table from structured email information. Such tables form two-dimensional maps that help identify the completeness of data that has been collected from many different sources. By adding date time, custodians and keyword filter results in these pivots the universe of available information is put on a map. Second part of the method introduces concepts from social network analysis that can be applied to a network of persons in which two persons are linked if they have communicated by email. Objective measures for centrality can be calculated to determine “hubs” and “brokers” in the network. Traditional date time and full-text filtering can be used to decompose complex networks into less-complex sub networks that are focused on particular events and/or topics. By presenting full-text filtering results in social network context culling becomes more objective.

For future research we recommend to investigate how the discussed techniques can be integrated in the culling process with objective parameters that minimize manual involvement. Also, with the fast development of technology for entity extraction we recommend that the proposed techniques are not only applied on structured email information but also on social networks that have been obtained through text mining on unstructured information, e.g. email bodies and word processor documents.

6. Literature

Bobrow, Daniel G., King, Tracy H., and Lee, Lawrence C. (2007). Enhancing Legal Discovery with Linguistic Processing. DESI I. Second International Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings.<http://www.umiacs.umd.edu/~oard/desi-ws/papers/bobrow.pdf>.

Chaplin, David T. (2008). Conceptual Search – ESI, Litigation and the issue of Language. DESI II. Second International Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings.
<http://www.cs.ucl.ac.uk/staff/S.Attfield/desi/9.%20Chaplin.pdf>.

Görg, Carsten and Stasko, John (2008). Jigsaw: Investigative Analysis on Text Document Collections through Visualization. DESI II. Second International Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings.

<http://www.cs.ucl.ac.uk/staff/S.Attfield/desi/7.%20Gorg.pdf>.

Krause, Jason (2009). In Search of the Perfect Search. ABA Journal.

http://www.abajournal.com/magazine/in_search_of_the_perfect_search.

Paul, George L. and Baron, Jason R. (2007). Information Inflation: Can the Legal System Adapt? Richmond Journal of Law & Technology, Volume XIII, Issue 3.

<http://law.richmond.edu/jolt/v13i3/article10.pdf>

Gloor, Peter A., Laubacher Rob, Zhao, Yan and Dynes, Scott (). Temporal Visualization and Analysis of Social Networks. NAACSSOS Conference, June 27-29, Pittsburg PA.

<http://www.ickn.org/documents/CKN4NAACSSOS.pdf>

Reeves, A. and May, C. (2008). Term testing: A Case Study. DESI II. Second International Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings. <http://www.cs.ucl.ac.uk/staff/S.Attfield/desi/4.%20May.pdf>.

Scott, John (1991). Social Network Analysis. London: Sage.

Socha-Gelbmann. (2006). EDRM E-Discovery Reference Model. <http://www.edrm.net>.