Artif Intell Law (2010) 18:413–430
DOI 10.1007/s10506-010-9099-3

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

Hans Henseler

Published online: 23 December 2010

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 techniques presented in this article propose 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.

Keywords E-Discovery · Social network analysis · Information retrieval · Email visualisation

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 automated part of this process is called culling and is aimed at selecting and filtering information without manual review. Current E-Discovery products use a variety of culling techniques mostly for

H. Henseler (✉)
Amsterdam University of Applied Sciences, Amsterdam, The Netherlands
e-mail: j.henseler@hva.nl

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 relevant or material, 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. In this paper we propose to introduce statistical analysis and social network analysis based to improve culling of emails.

2 Related work

Research that is related to the ideas and work presented in this article can be found in several sources. The most important one to mention here is probably the Enterprise Track in TREC (Craswell et al. 2005). The goal of this track is specifically to conduct experiments with enterprise data including, but not limited to, email archives. The experiments should reflect the experiences of users in real organisations. To achieve this, the organisers have collected more than 300,000 documents from the World Wide Web Consortium (W3C) containing nearly 200,000 emails posted on list servers. This collection is particularly valuable because the emails have been manually tagged so that the performance of different techniques can be measured. In Weerkamp et al. (2009) this collection is used to improve search in email archives by using contextual information. The authors use thread structure of the emails to get better search results. However, we find that emails in large organizations do not have the thread structure that is typical for internet mailing list servers.

Another source of research is the Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS). At the CEAS 2004 Culotta et al. presented an end-to-end system that extracts a user's social network and contact information about the persons in this network by text mining the emails in the user's inbox. Their research focuses more on text mining than on email search. Also at the CEAS 2004 the Enron email collection was introduced (Klimt and Yang 2004). This collection was made public during the legal investigation concerning the Enron Corporation which is still available for download online. The Enron dataset contains more than 500,000 emails and is representative for corporate email. Today many different approaches can be found on the internet that relate to the application of

social network analysis on the Enron email (see, e.g., Heer 2005). However, few to none focus on using social network analysis to improve email filtering in E-Discovery as it is intended here.

The Association for Computer Machinery (ACM) has several special interest groups that are relevant to the work presented here. In SIGIR (the special interest group on Information Retrieval), Tuulos et al. (2005) have presented a system that combines topic search and social network analysis into a multi-faceted approach for information retrieval in large scale email archives. Although their research is not aimed at filtering as such, their work contains interesting ideas on topic detection that may be used to improve the results presented here. Furthermore, they try to reduce social network complexity by reducing emails to email threads based on email subject matching. The idea is appealing but in E-Discovery typically content-based email threading is preferred since email subject lines are not always reliable.

In the Proceedings of the ACM Special Interest Group on Computer Human Interaction (SIGCHI), Viégas et al. (2006) present 'Themail', a visualization that portrays relationships discovered in email archives. The discovery is based on a combination of textual analysis, social network structure and temporal structure. Research performed earlier (Viégas et al. 2004) revealed that users could effectively map bursts of email exchange to events in their lives without having to rely on the content of the messages. This suggests that network-based filtering combined with temporal analysis may be a successful method for filtering emails in E-Discovery. In the field of Human Computer Interaction more interesting work has been done on information visualization. See for instance the work presented online at Stanford University on visualization of email data (Heer 2005).

As a final source of research we mention the field of Artificial Intelligence and Law where approaches are emerging that are related to analysis and retrieval of electronic stored information in E-Discovery. An overview of these approaches is presented in an article by Ashley and Bridewell (2009). They identify three emerging techniques from this field of which one is related to the application of social network analysis. This application appears to be intended primarily to automatically construct a relational network of persons involved to assist in resolving identities and to provide a basis for litigators who would further augment this network with manual annotations based on organizational structures etc.

3 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 et al. 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 from, to, cc, 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 report results of the experiments that have been performed with the Enron email collection and conclude with recommendations for future research.

4 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 100 Mb 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.

5 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 (arcs) represent emails. A subset of nodes corresponds to the set of custodians. The arcs 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. This resembles closely how the importance of web pages can be rated based on the ratings of pages that refer to it. See for instance the well-known PageRank algorithm used by Google (Brin and Page 1998) and the algorithm suggested by Kleinberg (1999) to find authoritative sources in a hyperlinked environment.

Persons with a high degree centrality are identified as hubs. In E-Discovery custodians typically have a high degree of centrality because the data was collected from their mailboxes. In social network analysis there are different interpretations for nodes with a high degree of centrality. A node can be a “hub”, “authority” (or both) and also brokerage roles are distinguished identifying if a node is a coordinator, liaison etc. (Batagelj and Mrvar 2003). If there is an important node in the email network that is not a custodian, it may be interesting to investigate why the person corresponding to that node is not a custodian. Persons with a high degree of betweenness can be important because they have a ‘liaison’ 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.

6 Experiments

We have conducted a number of experiments using the Enron email collection (Klimt and Yang 2004). The 400 Mb (compressed, without attachments) dataset contains 517,431 emails belonging to 150 custodians that were made public during the legal investigation concerning the Enron Corporation. This email set was deduplicated across all custodians resulting in 252,956 emails.

In order to describe our experiments we will formalize our data structures and algorithms using set logic. This is similar to the representation used in, for instance, Bommarito et al. (2009) where set logic is used to represent citation networks in legal cases. In our experiments the data sets and set operations can be implemented very efficiently using Structured Query Language (SQL). This is convenient because typically commercial E-Discovery software stores information that is extracted during processing in a relational database that can be manipulated using SQL queries.

For the purpose of our experiments we will represent the deduplicated Enron email set as a set E with emails ε:

$$\begin{aligned} \text{Let } E &= \{\varepsilon_1, \ldots, \varepsilon_{|E|}\} \text{ be the complete set of all Enron email id's;} \\ &\text{The cardinality of } E \text{ id denoted as} |E| \text{ is } 252,956; \\ \text{Let } I &= \{0, 1, 2, \ldots\} \text{ be the set of numbered days starting at } 1/1/1997; \\ \text{Let } E_\tau &= \{\varepsilon_1, \ldots, \varepsilon_{|E_\tau|}\} \text{ be emails sent on } \tau \in I. \end{aligned} \tag{1}$$

In our experiments we will use the addresses of these emails to analyse communication patterns. The direction of communication is also relevant and in addition to the email addresses, we also store the address type, i.e., 'From', 'To' or 'Cc'. The following sets are introduced to describe the experiments that were performed:

$$\begin{aligned} \text{Let } A &= \{\alpha_1, \ldots, \alpha_{|A|}\} \text{ be the set of all occurring email addresses } \alpha; \\ \text{Let } O &= \{\text{to,from,cc}\} \text{ be the set of address types } o; \\ \text{Let } A_{\varepsilon,o} &= \left\{\alpha_1, \ldots, \alpha_{|A_{\varepsilon,o}|}\right\} \text{ be addresses } \alpha \in A \text{ of email } \varepsilon \in E \text{ with type } o \in \text{O}; \\ \text{Let } A_\varepsilon &= \cup_{\forall o \in O} A_{\in,o} \text{ be the union of } A_{\varepsilon,to} A_{\varepsilon,from} \text{ and } A_{\varepsilon,cc}. \end{aligned} \tag{2}$$

The 252,956 Enron emails were parsed using a Perl script resulting in 2 tables. Table Emails representing set $E \times I$ with 252,956 existing (ε, τ) values. Table Relations representing $E \times A \times O$ with 1,635,833 existing (ε, α, o) values. Please note that we have left out the email body and subject in this representation as they are not required for the purpose of the work presented here.

6.1 Custodians

A custodian is defined as a person that is the owner of an email archive on a desktop computer or fileserver or of an mailbox on a server that has been collected as part of E-Discovery. For each custodian we identified one or more email addresses by examining the 'From' field for emails in the 'Sent Items' folder. This typically requires some manual clean-up and this procedure can be represented as follows:

$$\begin{aligned} \text{Let } N &= \{v_1, \ldots, v_{|N|}\} \text{ be the ordered set of } |N| \text{ custodians } v; \\ &\text{Custodian } v_i \text{ can be reference by its } i : 1 \leq i \leq |N|; \\ \text{Let } A_v &= \{\alpha_1, \ldots, \alpha_{|A_v|}\} \text{ be the set of email addresses } \alpha \text{ in 'From' addresses of} \\ &\text{the sent items folder of custodian } v. \end{aligned} \tag{3}$$

6.1.1 Email map

Using a SQL-query a list of a little over 50,000 triplets was constructed containing Custodian, Day and the number of emails where a Custodian is either in the 'From', 'To' or 'Cc' field on that particular day. The computation of the email map M can be represented as a two-dimensional array with elements $M_{v,\tau}$. Let $M_{v,\tau}$ be the number of emails sent or received by custodian v on day τ, then

$$M_{v,\tau} = \sum_{\forall \varepsilon \in E_\tau} |A_v \cap A_\varepsilon| \tag{4}$$

The resulting data set is an array $M_{v,\,\tau}$ denoting the position of a painted rectangle at location $(i,\,\tau)$ with a gray value based on the corresponding $M_{v,\,\tau}$ value and the rank of v. The higher the number of emails on a given day, the darker the painted rectangle. The result is presented in Fig. 1 below.

The email map in Fig. 1 immediately shows that the email traffic differs per custodian both in start date as well as intensity. Small modulations are visible indicating that typically less emails are sent and received during the weekends. Larger modulations reflect season characteristics.

Figure 1 does not only show differences but in some cases there are also interesting similarities. For instance custodians 22, 42, 82, 120, 122 and 146 show very intense email traffic starting after November 2001. Further investigation reveals that these custodians are involved in trading and the final email burst starts around or shortly after the date Enron has filed for bankruptcy on December 2nd, 2001.

From this experiment we can learn that, as could be expected, Enron employees with related work have related email contacts and thus have similar email patterns.

6.2 Search term fingerprint

We can calculate an email map for parts of the email collection by filtering on a particular search term. A search term is represented by a query ψ and we represent the set of emails that are responsive to this query as E_{ψ}:

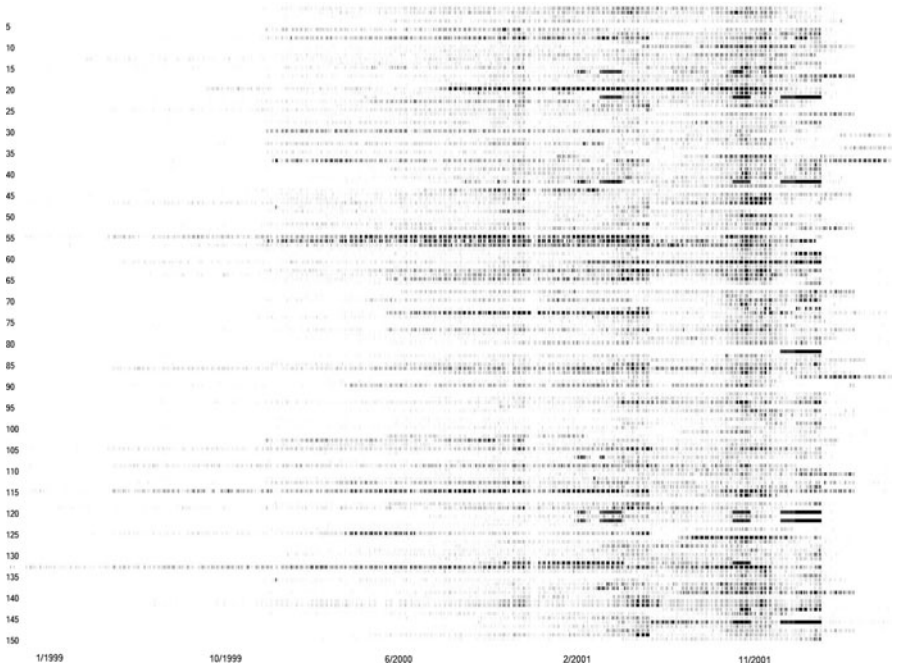

Fig. 1 Map showing the number of emails per custodian (*vertical axis*) per day (*horizontal axis*)

$$\text{Let } E_\psi = \left\{ \varepsilon_1, \ldots, \varepsilon_{|E_\psi|} \right\} \text{ be id's of emails responsive to query } \psi \tag{5}$$

The search term fingerprint $M_{\psi,v,\tau}$ associated with query ψ is an email map representing only E_ψ email messages. Let $M_{\psi,v,\tau}$ be the number of emails sent or received by custodian v on day τ responsive to query ψ:

$$M_{\psi,v,\tau} = \sum_{\forall \varepsilon \in E_\tau \cap \mathrm{E}_\psi} |A_v \cap A_\varepsilon| \tag{6}$$

The email map in Fig. 2 is based on all emails that are responsive to the query "Azurix OR (Wessex Water)". This query was selected because Wessex Water in the UK was acquired by Enron in 1999 and would form the basis for its water subsidiary Azurix. In total 782 emails respond to this query.

The email map in Fig. 3 is based on all emails that are responsive to the query "Blockbuster". This query was selected because in 2000 Enron and Blockbuster announce a important strategic alliance that is called off in March 2001. In total 330 emails respond to this query.

Please note that in Figs. 2 and 3 the timescale on the X-axis is different from the timescale in Fig. 1. For the sake of readability we have magnified the images and consequently have left out dots occurring before July 2000.

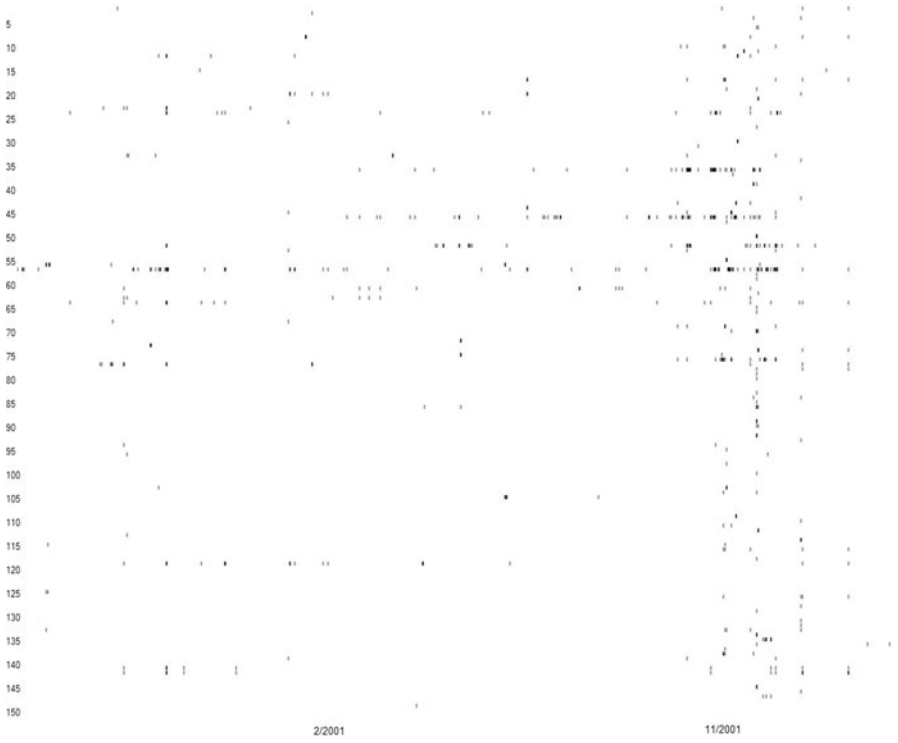

Fig. 2 Map of emails responsive to the 'Azurix' query

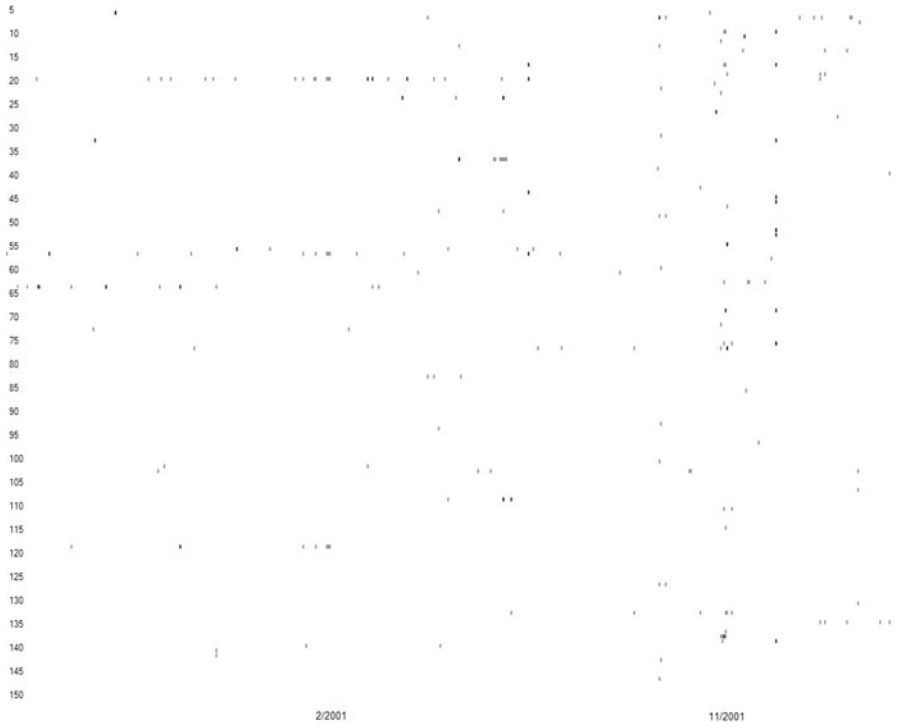

Fig. 3 Map of emails responsive to the 'Blockbuster' query

Typically in E-Discovery projects, during the processing stage, search term statistics are restricted to the number of hits per search term for a given list of search terms. Sometimes these counts are broken down per custodian. By plotting the number of hits for a single search term as a function of custodian and time, a map is produced that reveals when and who was involved on emails related to that search term. Phrases or compound concepts would be handled similarly by replacing search terms with Boolean queries combining multiple search terms.

6.3 Search term timeline

The search term fingerprint (cf. Figs. 2, 3) shows very small details as tiny black dots scattered over the map. When assessing the effectiveness of a search term it can also be useful to ignore the distribution over different custodians and look at the variation over time hoping that we can identify spikes that are related to specific events. We call this the timeline T_τ which represents the number of messages on day τ. Similar to $M_{v,\tau}$ (4) and $M_{\psi,v,\tau}$ (6) we also introduce $T_{\psi,\tau}$ which is the search timeline representing all messages responsive to ψ on day τ.

$$\begin{aligned} \text{Let } T_\tau &= |E_\tau| \text{ be the number of emails on day } \tau. \\ \text{Let } T_{\psi,\tau} &= \left|E_\psi \cap E_\tau\right| \text{ be the number of emails responsive to } \psi \text{ on day } \tau. \end{aligned} \tag{7}$$

The problem with this approach is that the total number of emails per day varies which renders the absolute number of emails responsive to ψ on day τ meaningless. To find out if a number of hits on a certain day is above or below average, we will represent the number of hits for a query on day τ as the percentage of the total number of emails over that same period $\left(\text{i.e.}\frac{100\%T_{\psi,\tau}}{T_\tau}\right)$. This is illustrated in Fig. 4 below for the 'Azurix' query.

Looking at the pattern of spikes in Fig. 4, it is interesting to know that Azurix did an IPO in Spring 1999. However, early 2001 problems arose and Azurix had to sell bonds below their investment grade. This did not help and in August 2001 the CEO of Enron International had to retire from Azurix.

Figure 5 illustrates the timeline for the 'Blockbuster' query. In July 2000 Enron announced a deal with Blockbuster but in March 2001 the companies ended this venture. In August 2001 Enron CEO Jeff Skilling resigned for personal reasons. There must have been many speculations around the problems surrounding Enron with respect to both the Blockbuster deal and the Azurix disaster explaining peaks at August 2001 in both Fig. 4 as well as Fig. 5.

6.4 Search term social network

The email maps give a global overview of email activity that can be fine tuned by filtering on keywords. This is a cheap solution to discover more about the email collection but it lacks a great deal of information that is also available in the email

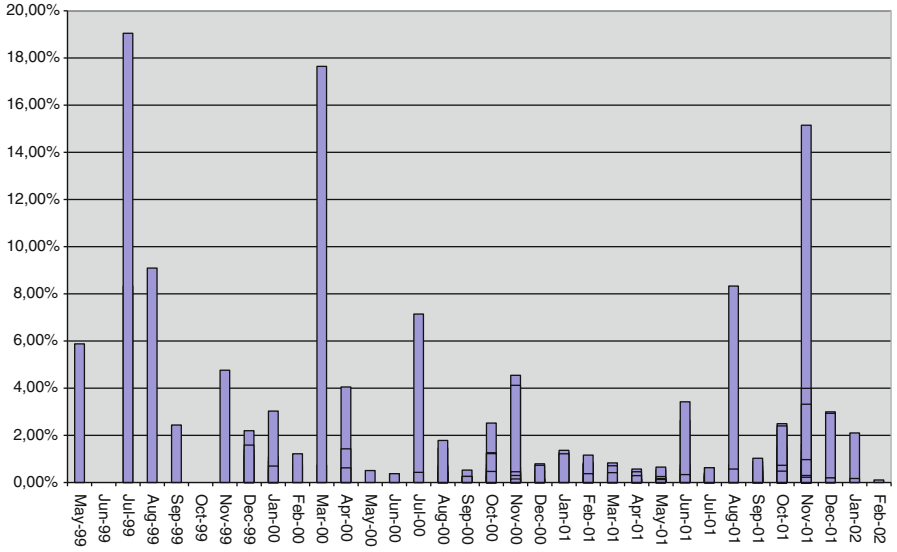

Fig. 4 Timeline for the 'Azurix' query showing number of hits summarized over all custodians per month as a percentage of the total number of emails per month

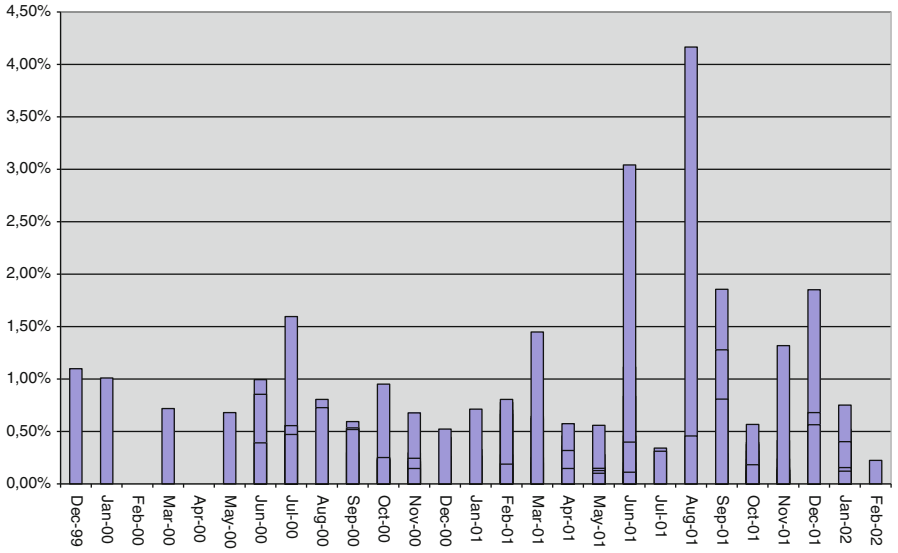

Fig. 5 Timeline for the 'Blockbuster' query showing number of hits summarized over all custodians per month as a percentage of the total number of emails per month

headers. In this section we will show an effective procedure for investigating the social networks that are present in the email collection.

The 'Blockbuster' query resulted in 330 responsive emails. A network for this collection can be constructed as follows. For each responsive email, create pairs of email addresses always starting with the 'From' address and ending with an address from the 'To' and 'Cc' fields. For the 'Blockbuster' query this results in 1,090 directed edges, also known as arcs, with 747 unique email addresses which are interpreted as vertices in the network. More formally the vertices A_ψ of a network for all emails E_ψ that respond to query ψ are represented as follows:

$$A_\psi = \bigcup_{\forall \varepsilon \in E_\psi} A_\varepsilon \tag{8}$$

The communication flow that is represented by the emails that are responsive to query ψ can be represented as subsets of E_ψ. These subsets distinguish between emails sent from and emails copied or sent to an email address @ that that is a vertex in A_ψ:

$$\begin{aligned}
&\text{Let } E_{\psi,@,o} \text{ be the set of emails responsive to query } \psi \text{ where address } @ \in A \\
&\quad \text{occurs as type O} \in o. \\
&\text{Let } E_{\psi,@\rightarrow} = \left\{\varepsilon | \forall \varepsilon \in E_{\psi,@,o} \wedge o = from\right\} \text{ be the set of emails} \\
&\qquad \text{responsive to query } \psi \text{ sent from address } @. \\
&\text{Let } E_{\psi,\rightarrow@} = \left\{\varepsilon | \forall \varepsilon \in E_{\psi,@,o} \wedge (o = to \vee o = cc)\right\} \text{ be the set of emails} \\
&\qquad \text{responsive to query } \psi \text{ sent or copied to address } @. \\
&\text{Let } E_{\psi,@} = E_{\psi,@\rightarrow} \cup E_{\psi,\rightarrow@}
\end{aligned} \tag{9}$$

Based on these definitions we can now construct the directed graph Γ_ψ that represents the directed flow of emails in E_ψ between Email addresses in A_ψ. Let Γ_ψ be the set of triplets (α_1, α_2, ω) where ω messages ε are sent or copied from α_1 to α_2:

$$\Gamma_\psi = \left\{(\alpha_1, \alpha_2, \omega) | \alpha_1 \in A_\psi \wedge \alpha_2 \in A_\psi \wedge \omega = \left|E_{\psi,\alpha_1\rightarrow} \cap E_{\psi,\rightarrow\alpha_2}\right| \wedge \omega > 0\right\} \quad (10)$$

Then Γ_ψ represents a directed graph with its elements representing edges from α_1 to α_2 with weight ω (i.e., the number of emails from α_1 to α_2). In our database we constructed a new table called Edges with records (α_1, α_2, ω) using a single select statement with an inner join to implement the criteria listed in (10).

We load the resulting network in Pajek (Batagelj and Mrvar 2003). Pajek is a program for analysis and visualization of large networks. The Blockbuster network is actually very cluttered making it difficult to visually identify email patterns. Using Pajek, the network can be simplified. The network of 747 nodes (i.e. email addresses) can be decomposed in 56 islands (i.e. disconnected parts of the network). The largest island contains 433 nodes. The network representing this island is still quite complex. In fact, many addresses exist that have no outgoing links (i.e. these addresses do not occur in the 'From' field for any of the selected emails). Using Pajek we can reduce the network by removing nodes that have no outgoing links which results in a network of 65 nodes. By removing these nodes one smaller section and several individual nodes become disconnected leaving the main network at 55 nodes. This network is depicted in Fig. 6 below.

One method to examine nodes in this network is to calculate the *degree centrality* which measures the importance of a node directly by the number of neighbors it has. This is similar to, for instance, Mazzega et al. (2009) who use centrality to measure the importance of legal codes in a citation graph representing the network of French legal codes.

Assuming that our network has non-weighted and undirected links, then the degree centrality $C_{\psi,@}$ for vertex @ in Γ_ψ can be represented as follows:

$$\begin{aligned}
&\text{Let } A_{\psi,@\rightarrow} = \left\{\alpha | \forall\alpha \in A_\psi \wedge \left(E_{\psi,@\rightarrow} \cap E_{\psi,\rightarrow\alpha} \neq \phi\right)\right\} \text{ be all addresses with}\\
&\quad \text{have received emails responsive to query } \psi \text{ sent from address } @.\\
&\text{Let } A_{\psi,\rightarrow@} = \left\{\alpha | \forall\alpha \in A_\psi \wedge \left(E_{\psi,\rightarrow@} \cap E_{\psi,@\rightarrow} \neq \phi\right)\right\} \text{ be all the addresses with}\\
&\quad \text{have sent or copied emails responsive to query } \psi \text{ to address } @.\\
&\text{Let } A_{\psi,@} = A_{\psi,@\rightarrow} \cup A_{\psi,\rightarrow@}\\
&\text{Let } C_{\psi,@} \text{ denote the degree centrality of vertex } @ \in A_{\psi,@}, \text{then}: C_{\psi,@} = \left|A_{\psi,@}\right|
\end{aligned} \quad (11)$$

The top three addresses for this network are Jeff Dasovich (12.0), Steven Kean (10.0) and Karen Denne (9.0). These names are well known in the Enron investigation. Jeff Dasovich was government Relations Executive. Steven Kean was Vice President and Chief of Staff and Karen Denne was VP of Public Relations for Enron.

A more sophisticated measure is *eigenvector centrality* which measures the importance of a node by incorporating the importance of linked nodes. The importance of a node is obtained by calculating the eigenvector of the corresponding connectivity matrix that has the largest eigenvalue. Alternatively, the eigenvector centrality can also be simulated using the Google Page Rank algorithm (Brin and Page 1998).

In short the Page Rank algorithm calculates the importance of a vertex @ as a minimum value (typically 0.15) plus a constant (typically 0.85) times the sum of the

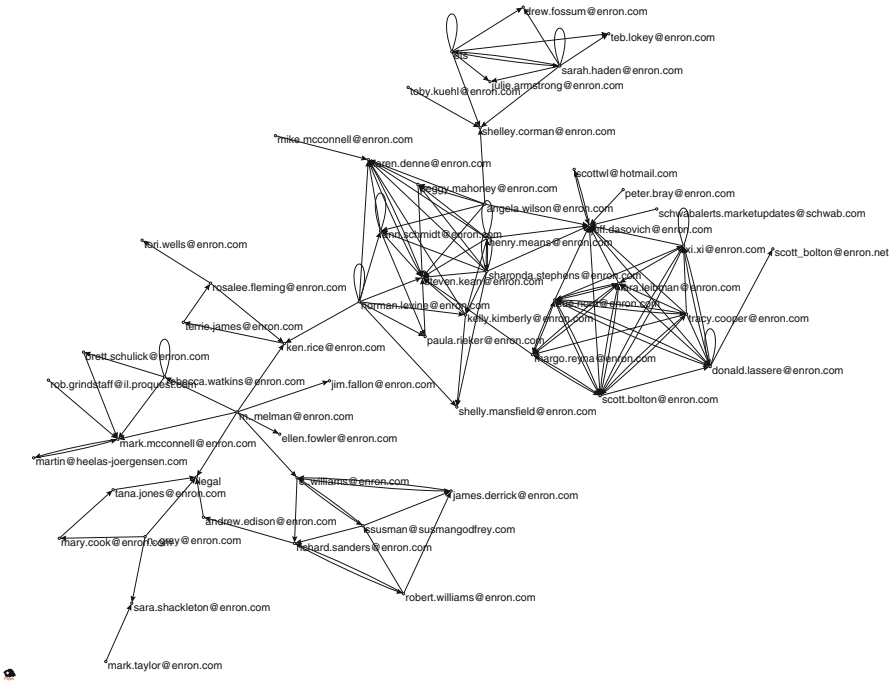

Fig. 6 Reduced directed network for the 'Bockbuster' query containing 55 nodes

rated importance of neighbors that have an incoming link to @. The rated importance of a vertex is calculated by dividing its importance by the number of outgoing links. Initially the importance of vertices is set to a real random value in [0, 1]. The Page Rank algorithm typically converges quite fast and the experiments here show little change after 10 iterations.

We will demonstrate how the Page Rank algorithm can be represented using our email representation. Page Rank calculation requires that we pre-compute the number of outgoing links for each vertex. In the directed graph Γ_ψ the *indegree* $C_{\psi,\to@}$ and *outdegree* $C_{\psi,@\to}$ centrality, can be calculated as follows:

$$\begin{aligned} C_{\psi,\to@} &= \left|A_{\psi,\to@}\right| \\ C_{\psi,@\to} &= \left|A_{\psi,@\to}\right| \end{aligned} \tag{12}$$

Let $P_{\psi,@} \in \Re$ denote the Page Rank value for vertex @ in graph A_ψ which can be calculated using the Page Rank algorithm:

$$\begin{aligned} \forall @ \in A_\psi : P_{\psi,@} &:= \text{random}[0,1] \\ &\quad \text{loop 10 times} \\ \forall @ \in A_\psi : P_{\psi,@} &:= 0.15 + 0.85 \sum_{\alpha \in A_{\psi,\to@}} \frac{P_{\psi,\alpha}}{C_{\psi,\alpha\to}} \end{aligned} \tag{13}$$

The update rule can be implemented using one SQL update query. This query updates a new table called Vertices with records (α, P_α, $C_{\alpha\rightarrow}$) and joins with the Edges table (cf. 10) to select incoming edges.

When calculating eigenvector centrality for the network in Fig. 4, number 1 is again Jeff Dasovich (0.331), number 2 is Sharonda Stephens (0.284) and number 3 is Steven Kean (0.275). Sharonda Stephens is not a custodian and she has primarily been sending emails into the network. According to the eigenvector centrality, she has been sending and receiving emails to and from relatively more important persons then Karen Denne who has a higher degree centrality but a lower eigenvector centrality (0.246).

A similar analysis can be performed for the 'Azurix' query (cf. Fig. 2). Persons that appear with high degree centrality are again Steve Kean and Karen Denne who also appeared important in the Blockbuster network. In this network also other important custodians show up such as CFO and Treasurer Rod Hayslett, VP of Regulatory Affairs Richard Shapiro and Chairman for Azurix Michael Anderson.

The network in Fig. 6 can be detailed further by taking into account the number of messages sent between addresses. The arcs in the network are not only directed but then also have weights. This is visualized in Fig. 7 below.

In the email network the weight of a link is equal to the number of messages from α_1 to α_2 (cf. 10). This change is reflected in the calculation of the eigenvector centrality by redefining the *indegree* and *outdegree* introduced in (12) as the number of incoming or outgoing emails instead of neighboring addresses:

$$\begin{aligned} C'_{\psi,\rightarrow@} &= \left|E_{\psi,\rightarrow@}\right| \\ C'_{\psi,@\rightarrow} &= \left|E_{\psi,@\rightarrow}\right| \end{aligned} \tag{14}$$

If we continue the analysis of nodes in this network we can elaborate on the eigenvector centrality and identify two types of important vertices: hubs and authorities. A vertex is a good hub, if it points to many good authorities, and it is a good authority, if it is pointed to by many good hubs. The eigenvector centrality introduced earlier will find authoritative addresses in our network. If we calculate the eigenvector centrality of the inverse network, then we find the hub addresses (Kleinberg 1999). For the sake of illustration, we will use the standard analysis function in Pajek to identify 5 hubs and 5 authorities leading to the classification in Table 1 below.

Depending on the nature of the E-Discovery task, emails can be included or excluded in a filter depending on the importance of the address that they were sent from or were sent to. When an address has a high degree centrality for outbound links in a relatively small network such as the 'Blockbuster' query network, it probably corresponds to an address that has sent out emails to large distribution lists. For instance, a quick investigation of Sharonda Stephens indicates that she has sent 37 emails to an impressive total of 1,449 addresses with subjects such as "Enron Mentions" which typically contain press headlines that mention Enron. When email duplicates are removed within custodians a duplicate will remain after deduplication for every custodian that received this email. This is not the case in the experiment described here since we removed duplicates across custodians. Removing such non-specific emails during culling can reduce the review load. In contrast, Karen

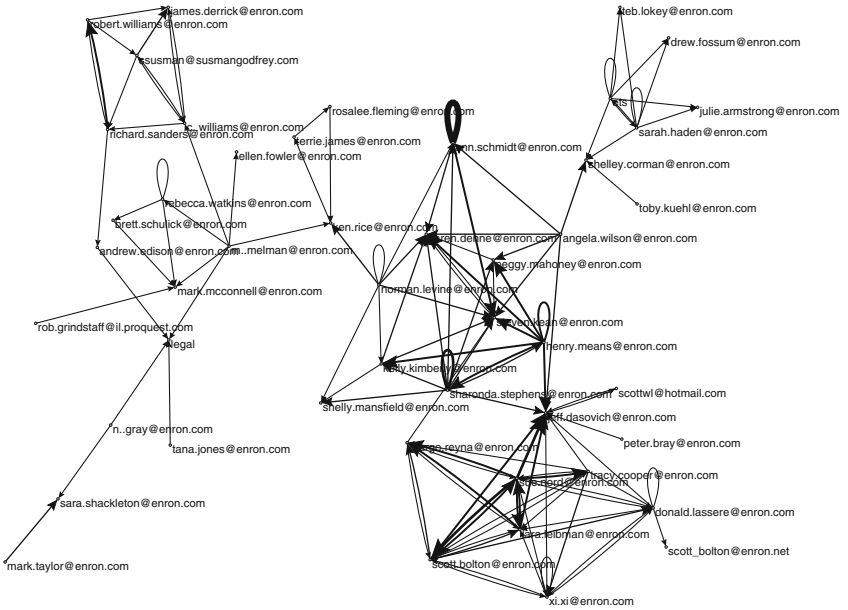

Fig. 7 Reduced network for the 'Blockbuster' query with *width of lines* indicating number of emails between nodes in the network

Table 1 Identification of five hubs and five authorities in the network presented in Fig. 7

Email address	Type
sue.nord@enron.com	Hub
henry.means@enron.com	Hub
angela.wilson@enron.com	Hub
sharonda.stephens@enron.com	Hub/Authority
ann.schmidt@enron.com	Hub/Authority
jeff.dasovich@enron.com	Authority
steven.kean@enron.com	Authority
karen.denne@enron.com	Authority

Denne has sent 337 emails to 2,589 email addresses, so an average of 7.7 addresses per email versus an average of 39.2 addresses per email sent by Sharonda Stephens.

7 Conclusions and recommendations

Several new techniques and existing techniques from related fields in computer science have been introduced to improve culling in E-Discovery by taking

advantage of structured information in email messages. The experiments reported here have been described in set logic. Using a matching relational database model, the set operations can be transformed into SQL statements. Since most E-Discovery systems store their information in relational databases, the application to real cases is straightforward. Moreover, using SQL enables researchers to leverage the computational power of corporate database servers that can handle very large collections of data.

Two different types of analysis and visualization were investigated. First a statistical analysis is performed resulting in an email map showing number of emails per day per custodian. Such two-dimensional maps help identify the completeness of data that has been collected from many different sources and to some extent it allows general correlation of custodians. Furthermore, creating such a map based on filtered results from a specified search term creates a of fingerprint of this search term that provides more detail than the standard hit statistics that are typically provided when negotiating search terms. By reducing the map to a timeline showing a graph of filtered email intensity as a percentage of the unfiltered email intensity, periods can be identified where the filtered email appears to be more important.

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, authorities and other brokerage roles 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 based on additional parameters such as addresses and time becomes more objective. In particular analyzing a social email network as a temporal network could be a very powerful solution to understand the structure of the network.

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. Using query expansion to improve search term filtering so that it becomes a better topic filter (cf. Tuulos et al. 2005) will upgrade the search term timeline to a more precise topic time line and may improve culling effectiveness in E-Discovery. Related to this, the work from Viégas et al. (2006) appears to have some interesting techniques for identifying distinct and frequently used words that characterize communication between persons and changes in communication over time. In the area of social network analysis further investigation on the different types of brokerage roles and their importance to E-Discovery given the nature of the underlying investigation is expected to be useful.

Finally, with the rapid advancement of language technology we recommend that the proposed techniques are not only applied on structured email information but also on social networks and semantic networks (see for instance Bommarito et al. 2009) that have been obtained through text mining in unstructured information, e.g. email bodies, attachments and electronic documents.

References

Ashley KD, Bridewell W (2009) Emerging AI+law approaches to automating analysis and retrieval of ESI in discovery proceedings. DESI III Global E-Discovery/E-Disclosure workshop, Barcelona. http://www.law.pitt.edu/DESI3_Workshop/Papers/DESI_III.KAshley.pdf

Batagelj V, Mrvar A (2003) Pajek—analysis and visualization of large networks. In: Jünger M, Mutzel P (eds) Graph drawing software. Springer, New York, pp 77–103

Bobrow D, King T, Lee L (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

Bommarito II MJ, Katz D, Zelner J (2009) Law as a seamless web? Comparison of various network representations of the United States supreme court corpus (1791–2005). In: Proceedings of the 12th international conference on artificial intelligence and law, pp 234–235

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst Arch 30(1–7):107–117. http://infolab.stanford.edu/~backrub/google.html

Chaplin D (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

Craswell N, de Vries A, Soboroff I (2005) Overview of the TREC-2005 enterprise track. In: The fourteenth text retrieval conference proceedings (TREC 2005). http://trec.nist.gov/pubs/trec14/papers/ENTERPRISE.OVERVIEW.pdf

Culotta A, Bekkerman R, Mccallum A (2004) Extracting social networks and contact information from email and the web. In CEAS-1. http://www.ceas.cc/papers-2004/176.pdf

Görg C, Stasko J (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

Heer J (2005) Exploring Enron: visual data mining of email. Available online at http://jheer.org/enron/

Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5): 604–632. http://www.cs.cornell.edu/home/kleinber/auth.pdf

Klimt B, Yang Y (2004) Introducing the Enron corpus. In: Proceedings of the collaboration, electronic messaging, anti-abuse and spam conference. http://www.ceas.cc/papers-2004/168.pdf

Krause J (2009) In search of the perfect search. ABA J. http://www.abajournal.com/magazine/in_search_of_the_perfect_search

Mazzega P, Bourcier D, Boulet R (2009) The network of French legal codes. In: Proceedings of the 12th international conference on artificial intelligence and law, pp 236–237

Paul G, Baron J (2007) Information inflation: can the legal system adapt? Richmond J Law Technol XIII(3). http://law.richmond.edu/jolt/v13i3/article10.pdf

Reeves A, 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 J (1991) Social network analysis. Sage, London

Socha-Gelbmann (2006) EDRM E-discovery reference model. http://www.edrm.net

Tuulos VH, Perkiö J, Tirri H (2005) Multi-faceted information retrieval system for large scale email archives. In: SIGIR '05, pp 683–683. http://cosco.hiit.fi/Articles/wi05-mail.pdf

Viégas F, Boyd D, Nguyen D, Potter J, Donath J (2004) Digital artifacts for remembering and storytelling: post history and social network fragments. In: HICSS-37. http://alumni.media.mit.edu/~fviegas/papers/posthistory_snf.pdf

Viégas F, Golder S, Donath J (2006) Visualizing email content: portraying relationships from conversational histories. In: Proceedings of ACM CHI 2006, pp 979–988. http://www.research.ibm.com/visual/papers/themail_chi_paper.pdf

Weerkamp W, Balog K, de Rijke M (2009) Using contextual information to improve search in email archives. In: 31st European conference on information retrieval conference (ECIR 2009), LNCS 5478, pp 400–411. http://staff.science.uva.nl/~mdr/Publications/Files/ecir2009-discsearch.pdf