Big Data Analytics for Legal Fact-Finding

J.C. Scholtes[1] and H.J. van den Herik[2]

Abstract

Legal fact-finding missions in electronic data sets, also known as e-Discovery, have always been related to *Big Data;* in the last two decades, e-Discovery has developed itself into a major research area in Artificial Intelligence and Law. Over the years, many myths on human versus machine quality have been circulated; at least three of them have now been refuted, but at this moment the end of the list of myths is not in sight. The refuted myths are on review quality (two myths) and on the point of departure. We admit that legal fact-finding missions are in many ways more demanding to search, analytics and machine-learning than any other search-related business application. In particular, defensibility, transparency and the evaluation of quality are a continuous obstacle and therefore deserve our attention. In this contribution we provide an overview of the state of affairs of the three myths by showing the past developments running from 1985 to 2018. We then describe the main applications. At the end we draw our conclusions and we briefly conjecture possible future developments for applications of big-data analytics in e-Discovery.

Keywords: e-Discovery, big-data law, information retrieval, text-mining, text-analysis, machine-learning

1 Introduction

An interesting branch of Artificial Intelligence (AI) is e-Discovery. In the last two decades e-Discovery has developed itself into a major research area. A big stimulus has been received from the advanced use of Big Data. When discussing *big data* in relation to law, we may comfortably state that legal fact-finding missions, also known as e-Discovery, deal with the *biggest* legal data collections of all. Today, an average e-Discovery easily involves several Tera-bytes of electronic data, holding hundreds of millions of documents with highly dynamic and completely unstructured information. These data sets consist of a variety of languages and distributed sources in many different electronic formats and shapes (including legacy and corrupted files); to put it more bluntly: e-Discovery data is truly *dirty* big data.

An e-Discovery is by far the most expensive part of litigation or arbitration. This holds for litigation in the United States, as well as for arbitrations by, for instance, the newly established Netherlands Commercial Court (NCC). The reason is that an expensive attorney manually has to dig through millions of documents to identify *responsive* and *privileged* documents.

[1] Affiliated with the Department of Data Science and Artificial Intelligence, Maastricht University and with ZyLAB Technologies BV, Amsterdam.

[2] Affiliated with e-Law, Faculty of Law; Leiden Centre of Data Science (LCDS), Mathematical Institute (MI), Faculty of Science, and FGGA, all Leiden University.

Where other fields of the law are generally slow technology adaptors, e-Discovery is leading the way in terms of case law, adaptation by law firms and (since a few years) also by corporate legal and government organizations. The course of this article is as follows. After the brief introduction above (Section1), we explain what e-Discovery is in Section 2. Then Section 3 describes the challenges in e-Discovery. Section 4 deals with search in e-Discovery. The main issue is then examined in Section 5: the rise of Technology Assisted Review (TAR). Section 6 provides four examples and Section 7 gives our conclusions. Finally, Section 8 speculates on possible future developments.

2 What is e-Discovery?

e-Discovery (also called electronic discovery) originated in the United States. It refers to the process of discovery facts in legal proceedings. Discoveries are a part of a pre-trial procedure under which one party can request evidence from the opposing party. Discovery is all about fact-finding, and ultimately, truth-finding. In 2006, "e-Discovery" became a definable process under the US Federal Rules of Civil Procedure (FRCP).

Over the years, e-Discovery (in US style) has become the standard approach in Europe for use cases, such as arbitration, answering regulatory requests, (internal-, government- , and criminal) investigations, freedom of information act (FOIA) requests, public records requests, compliance investigation, preparation of mergers and acquisitions (M&A) and recently also *Right to be Forgotten Requests* under the General Data Protection Regulation (GDPR).

In general, e-Discovery handles Electronically Stored Information (ESI), which can be any kind of electronic data, stored anywhere, in any media, format, language or shape. This includes all unstructured data sources such as file shares, email, and SharePoint data, but also more structured data in databases and even mobile - or social media data. Of course, ESI goes well beyond textual information; it also includes audio recordings, images and video.

3 Challenges in e-Discovery

Challenges in e-Discovery are different from challenges in other information quests. e-Discovery challenges may arise from causes that have their own right of existence (their own origin) with a different perspective, viz. *legal, process, data,* or *human*. As a result, methods and technologies used in e-Discovery are more demanding than those in other use cases. Below, we will discuss al four origins.

3.1 A legal related origin

As was clear from the *Zubulake v. UBS Warburg* case [3] in 2003, e-Discovery in human resource disputes can be seen as *asymmetrical warfare.* [4] Let us see for whom the asymmetry will legally be beneficial? Where corporations have to digest a multi tera-bytes collection of information in an e-Discovery, the individual starting the case has to worry only about a small and (probably) cleaned-up mailbox solely containing a few thousand relevant emails and files. This imbalance is causing major headaches to corporate organizations. In some cases, e.g., between organizations, particularly in the US, the e-Discovery process is used to bleed the other party to death (a real risk).

Both US and European legislation are calling for *proportionality*, but the balance between proportionality and incompleteness is not always clear. Contradictions in US e-Discovery laws and European privacy laws only add to the complications: where US courts require large and complete disclosures, European laws prevent sharing personally protected data. Corporations find themselves between a rock and a hard place deciding which regulations to follow.

3.2 A process related origin

Where the use of e-Discovery typically is the exclusive field of law firms, the bill obviously has to be paid by the parties involved in such a litigation: large corporations and governments. Handing over the entire e-Discovery responsibility is no longer something these organizations can afford due to the exploding costs of the e-Discovery process. As a result, organizations have nowadays taken back the control over e-Discovery and brought (parts of) the e-Discovery process in-house. External advisors are used only where their services are essential or required by the law.

Consequently, more work is executed internally, starting with the identification, collection, and processing, followed by the initial filtering, culling and a quick, high-level, first-pass review. After these activities, the remaining data is produced to external advisors, who do a more complex review (including the review for privileged data) and finally produce the data for the requesting party. In such an environment, it has to be made clear who is responsible for which activities. Internally, this requires intense cooperation between the *legal* department on the one hand and the *IT* and *business* department on the other hand. Externally, an equally intense cooperation between the corporate legal department and the external counsel is required. When technology is used, it needs to be implemented within the context of such a cooperation and the according responsibilities. In this situation, two facts will reinforce each other, namely (1) legal and IT are different disciplines (with different people and cultures) and (2) the adaptation of e-Discovery by

[3] Zubulake v UBS Warburg, 217 F.R.D. 309, 322 (S.D.N.Y. 2003).

[4] R.A. Satterwhite & M.J. Quatrara, 'Asymmetrical Warfare: The Cost of Electronic Discovery in Employment Litigation', 14 RICH. J.L. & TECH. 9, 2008. Available at: http://law.richmond.edu/jolt/v14i3/article9.pdf.

law firms is traditionally slow. Hence, the two processes (legal and IT) will not lead to beneficial collaborations.

3.3 A data related origin

The main source of challenges in e-Discovery is data related. The reason is straightforward, contrary to other business data, e-Discovery data is *dirty data*. It contains corrupted files, complex embedded objects (emails and attachments, Excel spreadsheets in a Word file), data containers (ZIP, PST), very long files (sometimes 1000's of pages), very short files (emails stating no more than "OK"), PDFs and TIFF without any searchable text, audio, images, video, legacy file formats, proprietary content management systems, encrypted data, social media formats, different languages (even within one document), various data locations, etc. [5] All this makes the application of artificial intelligence techniques harder to use than when used on *clean* data, as explained by Ashley in 2010. [6]

Whatever the case, the biggest obstacles are (1) the sheer volumes and (2) the continuing exponential growth; already in 2007 George Paul and Jason Baron warned for the effects of these two obstacles on a legal system. In their time they spoke of *Information Inflation* and the *Age of Exabytes*, and illustrated the exponential growth by a comparison of the Presidential records and their different administrations. [7] [8]

3.4 A human related origin

All humans are different, therefore it is clear that different humans make different *legal review* decisions. Reviewing enormous amounts of data in an e-Discovery process can be a boring and lonesome task, resulting in errors and low review quality, as explained by Attfield in *the Loneliness of the Long-Distance Reviewer*. [9] Human beings are also adaptive creatures, this means that we learn from our activities. As a result, the same human reviewers make different decisions at different moments in the review process, and this is even worse. The lesson is that in the inconsistency, human beings are inconsistent, leading to a wide variety of different decisions during the review

[5] J.R. Baron & P. Thompson, 'The search problem posed by large heterogeneous data sets in litigation: possible future approaches to research', In: *Proceedings of the 11th international conference on artificial intelligence and law*, pp 141–147, 2007.

[6] K.D. Ashley & W. Bridewell, 'Emerging AI & Law approaches to automating analysis and retrieval of electronically stored information in discovery proceedings', *Artificial Intelligence and Law*, 18(4), 311-320, 2010. doi:10.1007/s10506-010-9098-4.

[7] G.L. Paul & J.R. Baron, 'Information Inflation: Can The Legal System Adapt?', 13 *Rich. J.L. & Tech* 10 (2007). Available at: http://scholarship.richmond.edu/jolt/vol13/iss3/3.

[8] J.R. Baron, J.R. 'Law in the Age of Exabytes: Some Further Thoughts on 'Information Inflation' and Current Issues in E-Discovery Search', XVII *Rich. J.L. & Tech* 9 (2011), Available at: http://jolt.richmond.edu/v17i3/article9.pdf.

[9] S. Attfield, S. de Gabrielle & A. Blandford, 'The loneliness of the long-distance document reviewer: e-Discovery and cognitive ergonomics'. In: DESI III workshop at ICAIL, Barcelona, June 2009.

process. As stated by Grossmann in 2018: "Manual Review is an expensive, burdensome, and error-prone process". [10]Over the years, there have been many studies supporting this view.

To wrap up, we may state that for all the above reasons, e-Discovery calls for more automation and decision support technology, but it is now also clear that e-Discovery will be more demanding with respect to quality, robustness, transparency and defensibility than any other business application.

4 Search in e-Discovery

Searching for a special article, a verdict or an indication is quite difficult. In ancient history, Socrates understood this difficulty very well: "If you search for something you know, then you don't have to search, since you know it already; and if you search for something you don't know, then you also do not have to search since if you stumble into it, then you do not recognize the finding because you did not know what you were searching for."

Below, we will explain searching by indicating how to classify the documents into four/five classes (Subsection 4.1). Then we will describe Boolean search to identify responsive documents in Subsection 4.2. Using Boolean search for finding relevant information is called Assisted Review.

4.1 Search

e-Discovery is all about sorting documents into different categories. Usually there are four main categories: (1) *not-responsive* and (2) *responsive* (often sub-categorized into different issues), (3) *confidential* and (4) *privileged*. Recently, (5) the requirements for *General Data Protection Regulations* have been added to this.

Search is key to all five tasks. As more than 60 years of research in *Information Retrieval* have shown, there are many different approaches for searching and finding relevant data. It usually starts with a straightforward *Boolean* search, followed by introducing mechanisms for *relevance ranking* and *relevance feedback* to overcome some of the limitations of Boolean search. Further improvements come from using *semantic information extraction* and *knowledge-based approaches* to identify and highlight relevant aspects of the documents. Recently, the success of *machine learning* has also boosted retrieval quality by teaching a retrieval algorithm what is relevant and what not.[11]

[10] M.R. Grossman & G.V. Cormack, 'Quantifying Success: Using Data Science to Measure the Accuracy of Technology-Assisted Review in Electronic Discovery in Data-Driven Law: Data Analytics and the New Legal Services', 1st Edition, Edward J. Walters, 2018.

[11] An excellent overview is given by C.D. Manning, P. Raghavan & H. Schütze, *Introduction to information retrieval.* New York: Cambridge University Press, 2009.

4.2 Boolean search to identify responsive documents

The first intuition is to do a straightforward search for the relevant words. However, there are many reasons why keyword searches are inadequate. We mention four of them. (1) If you have not written the text yourself, how will you know what to look for? It is difficult to define the right query. (2) Keyword search is based on Boolean logic with AND, OR and NOT operators. These are hard to understand for lawyers as queries quickly become complex, and almost look like programming. (3) Even if done well, the results are never satisfactory. It is feast or famine, with an AND operator narrowing your search, leading to fewer results, and an OR operator broadening it, leading to more results but a large amount of noise too. (4) Besides, how do you know if you have found all the relevant documents, when your Boolean search has completed? You simply do not know.

The shortcomings of using keyword search to carry out discovery were already exposed as far back as in 1985. In a groundbreaking study [12], a group of lawyers was given access to a document retrieval system and asked to continue searching by using Boolean operators until they felt that they had found 75% to 51 discovery requests related to a train accident. It turned out the lawyers massively overestimated the recall they achieved key word searches: confident that they had hit 75%, the average recall was only 20% of relevant documents.

The 1985 *Blair and Maron* paper had – and still has – a huge and rather perverse impact. First, it did *not* lead to a professional understanding in the sense that lawyers turned their attention to more advanced machine-based document classification technologies (despite the fact that the advanced machines were at that time in a state of being developed). Second, lawyers used – and some are still using – the low recall procedure as found in *Blair and Maron* to justify a full manual document review. Third, several new studies were ignored by the lawyers. We admit that the studies used other setups and other information retrieval methods, but they once more confirmed the low performance by humans compared to machines in the e-Discovery information retrieval tasks. [13]

Let us now return to the lawyers, and particularly to the time when lawyers continued to review large document sets manually (1985). From that moment onwards, the technologies and institutions around Assisted Review were advancing in leaps and bounds. First, they started to use Boolean search and thereafter early versions of what was called *concept search* came to the forefront. In the third stage it included semantic and clustering approaches, also known as *topic modeling.*

[12] D.C. Blair & M.E. Maron, 'An evaluation of retrieval effectiveness for a full-text document-retrieval system', *Communications of the ACM*, 28(3), 289-299, 1985. doi:10.1145/3166.3197.

[13] H.L. Roitblat, A. Kershaw & P. Oot, 'Document categorization in legal electronic discovery: computer classification vs. manual review', *Journal of the American Society for Information Science and Technology*, 61(1), 70-80, 2009. doi:10.1002/asi.21233.

5 The Rise of Technology Assisted Review

In the information retrieval community, an alternative approach to identify responsive documents was adopted from the field of text-classification. The idea was "selecting responsive documents for production requests" could also be considered a text-classification problem. The task now was simply: classify documents as either relevant or not-relevant. Such classifiers were in the beginning rule-based composed by hand-crafted rules. Later the idea came up to train a classifier with groups of relevant and non-relevant documents. The field started to evolve during the 3rd Message Understanding Conference (MUC-3) in 1991 (see Lewis et al., 1992). [14]

The first myth

Barnett and Cormack were among first to describe "using machine learning as a search method for e-Discovery". [15] [16] Initially Roitblat [17] defended the myth that manual review is the gold standard – that is, if you have unlimited time, unlimited money and an unlimited pool of alert lawyers, the result of a manual review would always be more defensible than the result of 'computer-assisted categorization' (as their paper named it). In 2011 Grossman and Cormack strongly invalidated this myth with an extensive study comparing (1) rule-based text-classification with machine learning to (2) text-classification with a human as classifier; [18]a new research field named Technology Assisted Review (or TAR) was born.

It took only a year for any US court to acknowledge this new research finding, and to incorporate the technology assisted review into US case law applications. In 2012, in the case *Da Silva Moore et al. v. Publicis Groupe & MSL Group*, in which five woman sued the advertising giant for sex discrimination, magistrate Judge Andrew Peck of the US District Court in the Southern District of New York ruled that the defendants could use 'computer-assisted review' to search 3 million electronic documents as part of the parties' e-Discovery protocol. It is worth quoting part of Judge Peck's opinion: "While some lawyers still consider manual review to be the 'gold standard', that is a myth, as statistics clearly show that computerized searches are at least as accurate, if not more

[14] An insightful overview of principles and methods for text-classification can be found in F. Sebastiani, 'Machine learning in automated text categorization', *ACM Computing Surveys*, 34(1), 1-47,2002. doi:10.1145/505282.505283 and in C.D. Manning, P. Raghavan & H. Schütze, *Introduction to information retrieval*. New York: Cambridge University Press, 2009.

[15] T. Barnett, S. Godjevac, C. Privault , J.M. Renders & R. Wickstrom, 'Machine learning classification for document review' In: DESI III Workshop at ICAIL, Barcelona, June 2009.

[16] G.V. Cormack & M. Mojdeh, 'Machine learning for information retrieval: TREC 2009 Web, relevance feedback and legal tracks', In: *The eighteenth Text REtrieval Conference proceedings* (TREC 2009), Gaithersburg, MD, Nov 2009. National Institute of Standards and Technology (NIST), USA.

[17] H. L. Roitblat, A. Kershaw & P. Oot, P. Document categorization in legal electronic discovery: computer classification vs. manual review. Journal of the American Society for Information Science and Technology, 61(1), 70-80, 2009. doi:10.1002/asi.21233.

[18] M.R. Grossman & G.V. Cormack, 'Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review', 17 *Rich. J.L. & Tech* 11 (2011). Available at: http://scholarship.richmond.edu/jolt/vol17/iss3/5.

so, than manual review (…) While this Court recognizes that computer-assisted review is not perfect, the Federal Rules of Civil Procedure do not require perfection."

5.1 Continuous active learning

The TAR method as approved by Judge Peck was still based on first manually reviewing a random sample.

The second myth
Where Boolean search was the standard to assist reviewers in eDiscovery, the combination of an interactive sequential learning approach as introduced by David Lewis in 1994, [19] was more efficient. Grossman and Cormack named their machine-learning protocol *Continuous Active Learning* (CAL). [20] The main point is that human involvement could be better replaced by intelligent programs involvement. Still the scientific question remained: do we really need Boolean search? The answer is: no, machine learning algorithms can do a better job. [21] It is the end of the second myth.

The third myth
The TAR method as supported by Judge Peck used the human approved random sample to start the machine learning in order to avoid any bias. The rise of the Monte-Carlo Tree Search that was started by Coulom [22] in the area of computer chess paved the way for the idea that machine-learning with a random selection is the best starting point. Even more so, if a proper balance could be found between *exploration* and *exploitation*. So, information-retrieval efforts in e-Discovery focused from then on finding *responsive* documents with random starts. In 2017, Ben Ruijl showed that MCTS could be better replaced by a local search process with an advanced learning algorithm and with hand-picked training examples. [23] In the earlier mentioned publication from Grossman and Cormack from 2018, the authors show that the same principle also applies to Assisted Review in e-Discovery: starting with non-random methods such as a Boolean search in combination with an interactive sequential learning approach is better than starting with a random sample. This is the end of the third myth.

[19] D.D. Lewis & M.R. Tong, 'Text filtering', In MUC-3 and MUC-4, 51-66, 1992. 10.1145/1072064.1072069.

[20] G.V. Cormack & M.R. Grossman, 'Evaluation of machine-learning protocols for technology-assisted review in electronic discovery' *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* - SIGIR '14, 2014. doi:10.1145/2600428.2609601.

[21] An excellent overview is provided in M.R. Grossman & G.V. Cormack, 'Quantifying Success: Using Data Science to Measure the Accuracy of Technology-Assisted Review in Electronic Discovery', In *Data-Driven Law: Data Analytics and the New Legal Services*, 1st Edition, Edward J. Walters, 2018.

[22] R. Coulom, 'Efficient Selectivity and Backup Operators in Monte-Carlo Tree Search', *Computers and Games*, 5th International Conference, CG 2006, 2006.

[23] B.J.G. Ruijl, *Advances in computational methods for QFT calculations.* PhD thesis, Leiden University, Leiden, the Netherlands, 2017.

5.2 How about review for privileged information?

The majority of machine-learning and information-retrieval efforts in e-Discovery focuses on finding *responsive* documents. In this area, the current TAR protocols work rather well. But when reviewing documents, lawyers also should look for *privileged* documents. For a proper understanding it is good to know the difference between *confidential* documents and *privileged* documents. Confidential documents should be handed over to the opposing party in an e-Discovery request. Privileged documents may contain (for instance) information on an agreement between the public prosecutor and the suspect on the diminishing of sentences after the suspect's providing of secret information. So, privileged documents need not be handed over in an e-Discovery. However, so far AI techniques are not able to distinguish between confidential and privileged documents. Hence, all relevant documents have to be checked manually. Disclosing only *one* privileged document can seriously damage one's interest, as seen in a recent Oracle-Google dispute. [24] Therefore, extensive (manual) efforts are implemented to review responsive documents for privilege. [25]

A privilege review highly depends on (1) the names of the organization and persons mentioned and (2) expressing certain statements, as well as on (3) the context of the case. In addition, a long document can contain just a few paragraphs or even sentences, which contain expressions protected by client-attorney privilege. It is not possible to find such text snippets by using TAR. [26]

Keyword search for names of attorneys, law firms (including email domains), and specific context-related keywords allow us to find *potentially* privileged documents, but here human review is still essential. [27] Better quality privilege review will remain a topic of research for the upcoming years.

5.3 Redactions for privileged data and data protection

Identifying, creating and reviewing redactions (also known as anonymizations or blacklining) are labor intensive, boring and error-prone activities. The recent effectiveness of the General Data Protection Regulation (GDPR) and the new California privacy regulations add to the existing

[24] P. Favro, 'Perspective: Oracle v. Google Trial Highlights Importance of Privilege Reviews', In *Bloomber law: Big Law Business*, 2016. https://biglawbusiness.com/perspective-oracle-vs-google-trial-highlights-importance-of-privilege-reviews/

[25] E.S. Epstein, The Attorney-Client Privilege and the Work-Product Doctrine. ABA 2001.

[26] M. Gabriel, C. Paskach & D. Sharpe, 'The Challenge and Promise of Predictive Coding for Privilege', ICAIL 2013 DESI V Workshop, 2013, Rome, Italy.

[27] G.L. Fordham, 'Using Keyword Search Terms in e-Discovery and How They Relate to Issues of Responsiveness, Privilege, Evidence Standards, and Rube Goldberg', 15 RICH. J.L. & TECH. 8, 2009. Available at: http://law.richmond.edu/jolt/v15i3/article8.pdf.

redactions the need to redact and pseudonymize personal data. Automatic methods to identify and apply redactions have been around for some time now. These are primarily based on using principles from the field of text-mining [28] to identify named entities and potentially personal-information. [29] However, identifying more complex information, such as indirect identifications (identifying an individual from the context) or larger textual sections, is still just as challenging as detecting privileged information. This will also continue to be a topic of research in the upcoming years. [30]

5.4 Analytics for early case assessment: strategic decision support

In parallel, text-analytics have been used for the more strategic application of what is called *Early Case Assessment* (ECA). When an organization is confronted with litigation, a regulatory request, or an (internal) investigation, the initial e-Discovery can generate terabytes of electronic data. It is not easy to start comprehending what a case is about, let alone making well-informed strategic decisions. This is where ECA can help. ECA is an umbrella term for many different methods to understand the structure and content of large unstructured data sets in order to make better decisions in an early phase of the e-Discovery without the need to have to review all documents in great detail in advance.

Depending on the type of e-Discovery case, there are different dimensions that may be interesting for an early case assessment: custodians, data volumes, location, time series, events, modus operandi, motivations, etc. As described by Attfield and Blandford in 2010, [31] traditional investigation methods can provide guidance for the relevant dimensions of such assessments: *Who, Where, When, Why, What, How, and How Much* are the basic elements for analysis. *Who, Where and When* can be determined by Named Entity Recognition methods. *Why* is harder, but personal experience of the first author in law enforcement investigations shows that data locations with high emotion and sentiment values also provides a good indication of the motivation or insights into the *modus operandi*.

[28] R.Feldman, & J.Sanger, *The text mining handbook: Advanced approaches in analyzing unstructured data*. New York, NY, 2007.

[29] C.D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard & D. McClosky. 'The Stanford CoreNLP Natural Language Processing Toolkit', In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60, 2014.

[30] C. Cardellino, M. Teruel, L.A. Alemany & S. Villata, 'A low-cost, high-coverage legal named entity recognizer, classifier and linker', *Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law* - ICAIL '17, 2017. doi:10.1145/3086512.3086514.

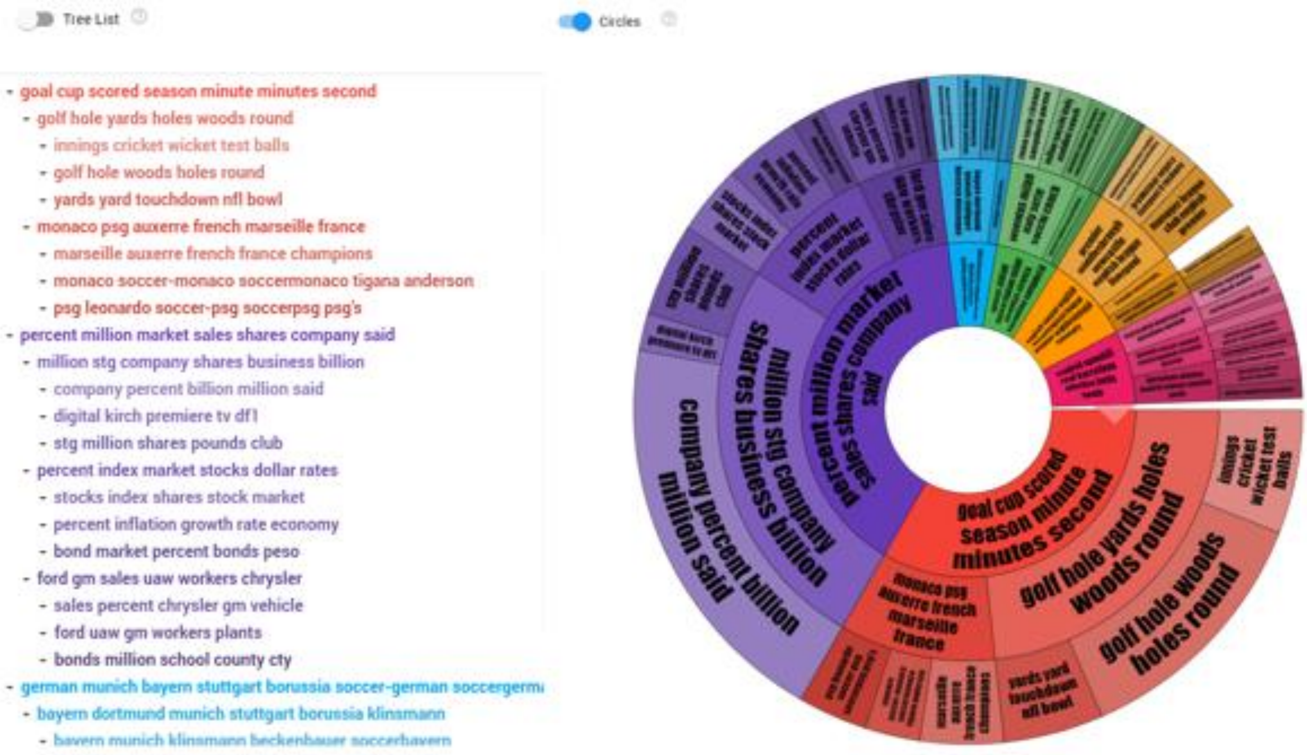
[31] S. Attfield, & A. Blandford, 'Discovery-led refinement in e-Discovery investigations: sensemaking, cognitive ergonomics and system design', *Artificial Intelligence and Law*, 18(4), 387-412, 2010. doi:10.1007/s10506-010-9091-y.

6 Four Examples of Technology Assisted Review

Below we provide four examples of TAR with the help of visualizations. In Subsection 6.1 we deal with the task to find an answer to the *What* question by means of NMF modeling[32]. In subsection 6.2 we focus on community detection through Dynamic topic modeling. In Subsection 6.3 we search for priorities given to news items and events. In Subsection 6.4 we search for the communication of emotions.

6.1 NMF modeling

An example of the visualization of the *What* question, can be found in Figure 1. The NMF topic modeling is combined with clustering and a basic visualization. This visualization allows users to dynamically browse e-Discovery document sets based on the automatically derived topic hierarchy.


Tree List
Circles
percent million market sales shares company said
million stg company shares business billion
company percent billion million said
percent index market stocks dollar rates
goal cup scored season minute minutes second
golf hole yards holes woods round
golf hole woods holes round
innings cricket wicket test balls
monaco psg auxerre french marseille france
yards yard touchdown nfl bowl

- goal cup scored season minute minutes second
  - golf hole yards holes woods round
    - innings cricket wicket test balls
    - golf hole woods holes round
    - yards yard touchdown nfl bowl
  - monaco psg auxerre french marseille france
    - marseille auxerre french france champions
    - monaco soccer-monaco soccermonaco tigana anderson
    - psg leonardo soccer-psg soccerpsg psg's
- percent million market sales shares company said
  - million stg company shares business billion
    - company percent billion million said
    - digital kirch premiere tv df1
    - stg million shares pounds club
  - percent index market stocks dollar rates
    - stocks index shares stock market
    - percent inflation growth rate economy
    - bond market percent bonds peso
  - ford gm sales uaw workers chrysler
    - sales percent chrysler gm vehicle
    - ford uaw gm workers plants
    - bonds million school county cty
- german munich bayern stuttgart borussia soccer-german soccergerm
  - bayern dortmund munich stuttgart borussia klinsmann
    - bayern munich klinsmann beckenbauer soccerbayern

Figure 1: Visualizing the *What* question in e-Discovery.

[32] NMF stands for Non-Negative Matrix Factorization.

The figure represents two adjacent visualizations of the *What* question: on the left side a traditional text hierarchical tree view and on the right side a so-called Word-Wheel representation. Both can be navigated interactively. For the text it happens by clicking on a line and for the graph by clicking on an area in the graph, one can either enlarge it, make it smaller, navigate to the documents holding a specific topic that is most dominant, or start the training of a classifier to find similar documents for Assisted Review. An example of text clicking is as follows. Clicking on the red entry on the left side “golf hole woods hole round” will show the documents describing Tiger Woods successes in the 1996 World Golf tournaments. All these topics and corresponding labels have been recognized by the topic modeling algorithm using unsupervised machine learning, the algorithm does not need any labeled or other initial information to build such topic models. They are also language and domain independent.

6.2 Community detection

Once the *Who’s* are identified by using *Named-Entity Recognition*, methods from social network analysis can be used to identify relevant groups and communities, allowing the reviewers to prioritize the review better and identify information which can be used in negotiations to obtain a more favorable settlement. An example of such a community detection on correspondence of the Museum of Modern Art in Amsterdam will led us to the automatic derivation of communities. [33] See Figure 2.

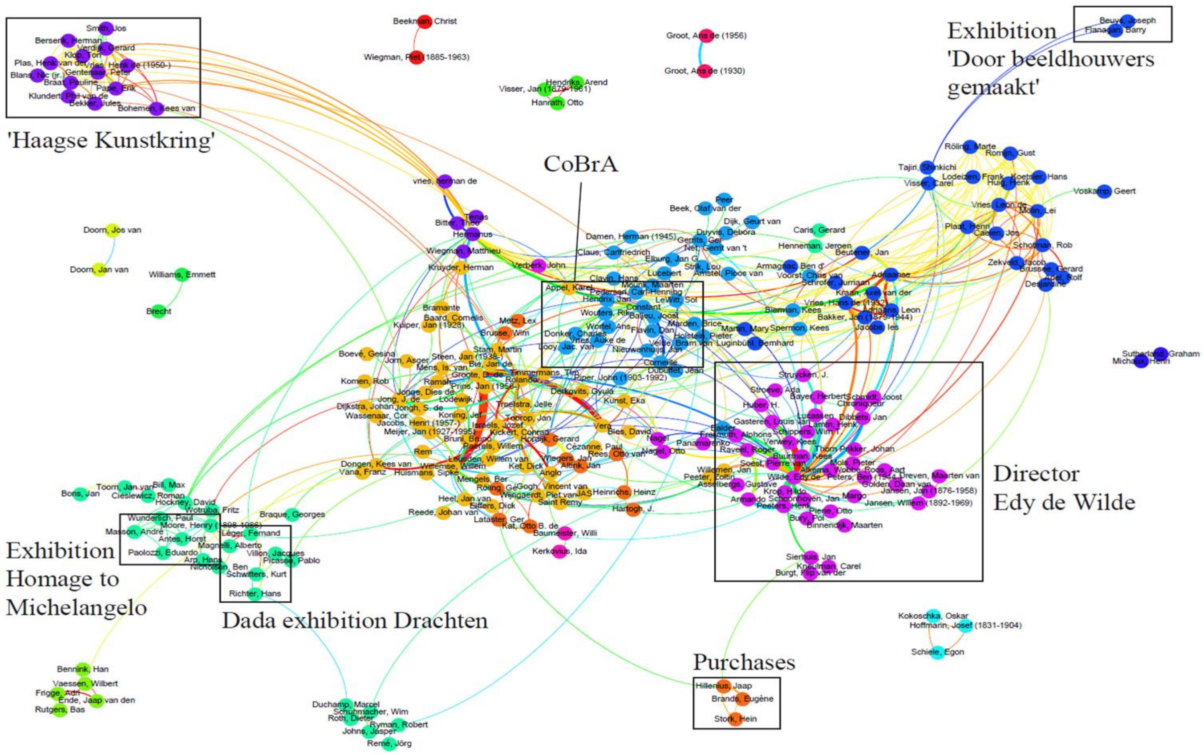


Figure 2: Community Detection on Communication from the *Stedelijk Museum of Modern Art* Amsterdam, the Netherlands.

[33] J. Smeets, J. C.Scholtes, C. Rasterfoff & M. Schravemaker, ‘SMTP: Stedelijk Museum Text Mining Project’. Digital Humanities Benelux (DHBenelux), Luxemburg, June 2016.

Here we remark that basic dimensions of the ECA can also be combined in more complex overviews such as *What-When*, a form of dynamic topic modeling, also referred to as *Topic Rivers*. [34]

6.3 Topic rivers

Figure 3 displays the visualization of so-called Topic Rivers on 8 months of Reuters news from 2014. For each week, the system determines (in this case) the 20 most dominant topics. Next, for each period, the number of new, growing or declining topics is determined and connected to corresponding topics in the previous and next period. In the resulting graph, the invasion in Ukraine can clearly be observed in March 2014, pushing away all other news. Other topics, such as the Israel-Palestine conflict can be seen present in the news for the entire year. Another anomaly is the blue one on the bottom of the graph, representing the news when Malaysia Airlines Flight MH17 was shot down over Ukraine.

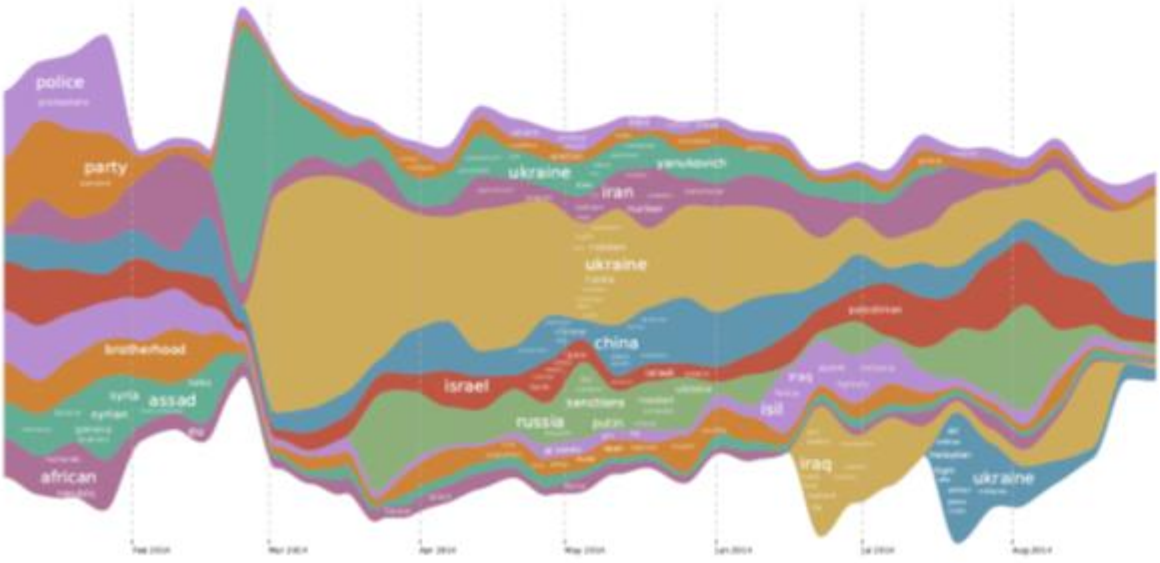


Figure 3: Answering the *What-When* question: Topic Rivers on 8 months of Reuters News from 2014

6.4 Emotion mining from song lyrics

Emotions and sentiments often lead to interesting emails in e-Discovery. Both can be measured. Combining emotions with custodians (persons) can lead to the discovery of relevant issues that could be the starting point of the Assisted Review.

For this reason, part of the *Why* question can often be identified by looking at the communication with the highest levels of emotions of negative sentiments. By identifying these and linking them to the persons expressing them, one can obtain an answer to the *Why-Who* question.

In Figure 4, we can observe the analysis of the text from 220.000 pop song lyrics for the basic emotions: *Trust, Anticipation, Joy, Anger, Fear and Sadness.* The name of pop artist is displayed on the lines connecting the most dominant emotions in their songs. As we can observe, rappers are in the left bottom corner around Anger and Fear. Elvis, the Beatles, and David Bowie are more in the top right corner around

[34] M. Tannenbaum, A. Fischer & J.C. Scholtes, 'Dynamic Topic Detection and Tracking using Non-Negative Matric Factorization', *Benelux Artificial Intelligence Conference* (BNAIC), Hasselt, Belgium, November 5-6, 2015.

Joy, Trust and Anticipation. Similar analysis have been made on movies, books and other content, leading to similar satisfying results. For court cases these techniques have also been used.

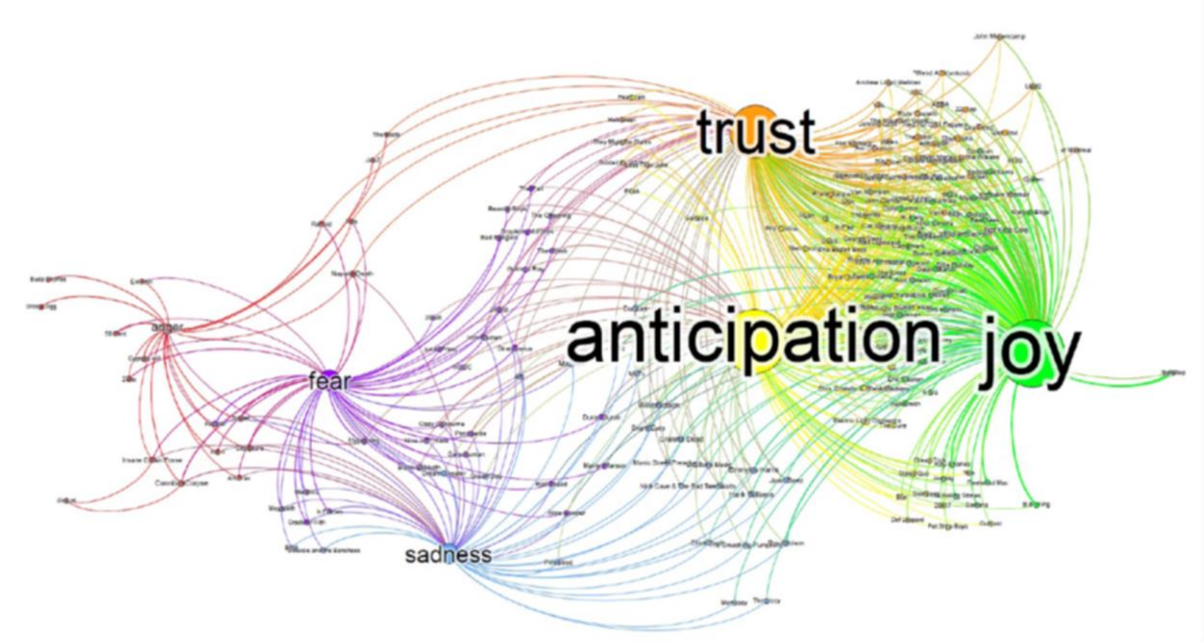


Figure 4: Answering the *Who-Why* question: Emotion mining on 220.000 song lyrics and the corresponding artists

Many other combinations, analysis, clustering and visualization methods can be created, and we will most likely see more of these in future research.

7 Conclusions

Starting in 1985 with Blair and Maron, who showed that the recall estimate of manual reviewers was highly exaggerated, many myths in e-Discovery have been invalidated by scientific research, contrary to the original beliefs. It is now accepted that: (1) *computer algorithms provide better review quality than human review*, (2) *machine-learning algorithms provide better review quality than Boolean search*, and (3) *machine learning starting with a random selection is not better than starting with hand-picked training documents.* Undoubtedly, more myths will perish in the near future.

For many years, knowledge-engineering methods prevailed over the empirical machine-learning methods in Artificial Intelligence and Law, this trend has been changed after e-Discovery had entered the field and even more after it had taken the lead. This change was not only due to scientific research, but it can be credited to judges such as Judge Peck explained in his verdict for Da Silva Moore v. Publicis Groupe, 2012.[35] They are enthusiastic evangelists promoting and even mandating the use of machine-learning based review methods in their court rooms. Thank you, for such a cooperation.

[35] Da Silva Moore v. Publicis Groupe, Case No. 11 Civ. 1279 (ALC) (AJP), 2012 WL 607412 (S.D.N.Y. Feb. 24, 2012).

8 Possible future developments

The behavior of Judge Pack et al. is not a surprise, because they were the first to observe the trend that the high cost of e-Discovery were obstructing the righteousness of their justice system. This made them understand the necessity of technology. But they also emphasized the defensibility and transparency of such methods. Yet, both concepts are still the Achilles heel of the more empirical artificial intelligence. As it will never be possible to explain a laymen jury or a judge how the complex mathematical models and high dimensional feature spaces of such models work, we have to develop evaluation methods around legal applications of artificial intelligence. These methods should continuously monitor quality and verify decision making in the same way as other complex techniques, such as DNA-evidence, have found their way into court. Recently, David Gunning from DARPA started a dedicated research program exactly on this topic. [36]

Other concerns in Big Data Law are of a more ethical nature: bias and prejudice are high risk factors, also in e-Discovery. As in other fields of Artificial Intelligence, we have to fence algorithmic freedom to prevent the development of too aggressive strategies or morally unacceptable behavior. All these issues are on the radar of today's e-Discovery researchers. [37]

The first generation of e-Discovery technology taught us how to deal with *big data*, the next generation will teach us how we can learn from big data and train our algorithms to provide better decision support and ultimately, on the condition it is properly fenced, defensible and understood, make better decisions by itself.

[36] D. Gunning, 'Explainable Artificial Intelligence (XAI)'. https://www.darpa.mil/program/explainable-artificial-intelligence.

[37] H.J. van den Herik & T.A.M. de Laat, 'The Future of Ethical Decisions Made by Computers', In: L.A.W. Janssens (red.), In: *The Art of ethics in the information Society*. pp. 49-54, 2016. Amsterdam: Amsterdam University Press.