Semantic Search in E-Discovery

Research on the application of text mining and information retrieval for fact finding in regulatory investigations

David van Dijk, Hans Henseler
Amsterdam University of Applied Sciences
CREATE-IT Applied Research
Amsterdam, the Netherlands
d.van.dijk@hva.nl, j.henseler@hva.nl

Maarten de Rijke
University of Amsterdam
Intelligent Systems Lab Amsterdam
Amsterdam, the Netherlands
derijke@uva.nl

***Abstract*— For forensic accountants and lawyers, E-discovery is essential to support findings in order to prosecute organizations that are in violation with US, EU or national regulations. For instance, the EU aims to reduce all forms of corruption at every level, in all EU countries and institutions and even outside the EU. It also aims to prevent fraud by setting up EU anti fraud offices and actively investigates and prosecutes violations of competition regulations. This position paper proposes to address the application of intelligent language processing to the field of e-discovery to improve the quality of review and discovery. The focus will be on semantic search, combining data-driven search technology with explicit structured knowledge through the extraction of aspects, topics, entities, events and relationships from unstructured information based on email messages and postings on discussion forum.**

Keywords: E-Discovery, Semantic Search, Information Retrieval, Entity Extraction, Fact Extraction, EDRM

I. Introduction

Since the ICT revolution took off around 50 years ago the storage of digital data has grown exponentially and is expected to double every 18 months [16]. Digital data became of crucial importance for the management of organizations. This data also turned out to be of significant value within the justice system. Nowadays digital forensic evidence is increasingly being used in court. The Socha-Gelbmann Report from 2006 shows a usage of this kind of evidence in 60% of the court cases [31].

The process of retrieving and securing digital forensic evidence is called electronic data discovery (E-Discovery). The E-Discovery Reference Model [8] gives an overview of the steps in the e-discovery process. The retrieval of information from large amount of digital data is an important part of this process. Currently this step still involves a large amount of manual work done by experts, e.g. a number of lawyers searching for evidence in all e-mails of a company which may include millions of documents [30]. This makes the retrieval of digital forensic evidence a very expensive and inefficient endeavor [24].

Digital data in E-Discovery processes can be either structured or unstructured. Structured data is typically stored in a relational database and unstructured data in text documents, emails or multimedia files. Corporate Counsel [6] indicates that at least 50% of the material of contemporary electronic discovery environment is in the form of e-mail or forum and collaboration platforms. Finding evidence in unstructured information is difficult, particularly when one does not exactly know what exactly to look for.

The need for better search tools and methods within the area is reflected in the rapid growth of the E-Discovery market [32,10], as well as in the growing research interest [34,15,29]. This paper positions the research that is carried out through joined work from CREATE-IT Applied Research at the Amsterdam University of Applied Sciences [7] and the Intelligent Systems Lab Amsterdam at the University of Amsterdam [17]. It focuses on the application of text mining and information retrieval to E-Discovery problems.

II. Text Mining and Information Retrieval

Information retrieval (IR) can be defined as the application of computer technology to acquire, organize, store, retrieve and distribute information [19]. Manning defines IR as finding material (usually documents) of unstructured nature (usually text) from large collections (usually stored on computers) that provides in an information need [23]. Text mining (TM), also called text analytics, is used to extract information from data through identification and exploration of interesting patterns [9]. In TM, the emphasis lies on recognizing patterns. TM and IR have a considerable overlap, and both make use of knowledge from fields such as Machine Learning, Natural Language Processing and Computational Linguistics.

Both TM and IR provide techniques useful in finding digital forensic evidence in large amounts of unstructured data in an automated way. The techniques can be used for example to extract entities, uncover aspects of and relationships between entities, and discover events related to these entities. The extracted information can be used as metadata to provide additional guidance in the processing and review steps in E-

Discovery. Without such guidance, plain full-text search in large volumes of data becomes useless without proper relevance ranking. Metadata can be used to support interactive drill down search that is more suited for discovering new facts.

Furthermore, information about entities and aspects makes it possible to retrieve facts about a person as to what kind of position he currently holds, what positions he has previously had and what is important about him. Information about relationships can be used to identify persons closely connected with each other, but also to identify what persons are strongly connected to specific locations or (trans)actions. And events related to the entity can help one to extract temporal patterns.

III. Applications

The above techniques can be useful in many areas, both within and outside the domain of E-Discovery. Opportunities can be found in the areas of fraud, crime detection, sentiment mining (e.g., marketing), business intelligence, compliance, bankruptcies and, as one of the largest areas, e-discovery [27,12]. Large regulatory compliance investigations in the areas of anti-corruption and anti-trust offer excellent opportunities for text mining and information retrieval. Known techniques can be optimized and further developed to extract facts related to corruption and competition and to identify privileged and private information that should be excluded from the investigation.

For the detection of competition law infringements one can look at how prices develop [4]. For finding corruption one could search for suspicious patterns in transactions between entities, e.g., clients and business partners. In determining confidential data one can think of social security numbers, correspondence between client and attorney, medical records, confidential business information, etc. But often it is not clear beforehand what is sought, and therefore techniques are of interest that make the information accessible and provide insights so that a user can easily interact with it.

The entities and relations retrieved by the aforementioned techniques can be made accessible to the user in various ways. Either as additional metadata to documents to be combined with full-text search or as relational data in a separate system which can process questions in natural language (Question Answering System). The former gives a list of documents in response, the second can answer in natural language.

IV. Objective

Our research will focus on review and in particular on the search process. Generic search technology is not the answer. It has its focus on high precision results, where the top-ranked elements are of prime importance, whereas in forensic analysis and reconstruction all relevant traces should be found. In e-discovery, both recall and precision must be simultaneously optimized [26]. As a consequence, in e-discovery, the search process is typically iterative: queries are refined through multiple interactions with a search engine after inspection of intermediate results [5].

Analysts often formulate fairly specific theories about the documents that would be relevant and they express those criteria in terms of more-or-less specific hypotheses about who communicated what to whom, where, and, to the extent possible, why [2]. Representing, either implicitly or explicitly, knowledge associated with analysts' relevance hypotheses so that an automated system can use it, is of primary importance in addressing the key issues in e-discovery of how to identify relevant material [14]. Our research is aimed at providing analysts with more expressive tools for formulating exactly what they are looking for.

In particular, our research questions are as follows:

RQ1: At least 50% of the material in today's e-discovery environment is in the form of e-mail or forum and collaboration platforms [6]. How can the context (such as thread structure or the participant's history) of email messages and forum postings be captured and effectively used for culling entire sets of messages and postings (as they do not answer the question posed)?

RQ2: How can the diversity of issues that relate to the question posed be captured in a data-driven manner and presented to analysts so as to enable them to focus on specific aspects of the question?

RQ3: Social networks, graphs representing probable interactions and relations among a group of people, can enable analysts to infer which individuals most likely communicated information or had knowledge relevant to a query [28,13]. How can we effectively extract entities from e-mail messages and forum postings to automatically generate networks that help analysts identify key individuals?

RQ4: How can we semi-automatically identify the principal issues around the question posed? Creating an "information map" in the form of a domain-specific and context-specific lexicon will help improve the effectiveness of the iterative nature of the e-discovery process [36].

Based on typical user needs encountered in E-Discovery best practices, these research questions are situated at the interface of information retrieval and language technology. Answering them requires a combination of theoretical work (mainly algorithm development), experimental work (aimed at assessing the effectiveness of the algorithms developed) and applications (implementations of the algorithms will be released as open source).

V. Innovation

In recent years the field of information retrieval has diversified, bringing new challenges beyond the traditional

text-based search problem. Among these new paradigms is the field of semantic search, in which structured knowledge is used as a complement to text retrieval [25]. We intend to start a research project which pursues semantic search along two subprojects:

Subproject 1: integrating structured knowledge (discussion structure, topical structure as well as entities and relations) into information retrieval models;

Subproject 2: extracting structured knowledge from user generated content: entities, relations and lexical information.

We have requested funding for two PhD students, one for each of the two subprojects. Subproject 1 will primarily address RQ1 and RQ2. Subproject 2 will focus on RQ3 and RQ4.

Work on RQ1 will start from earlier work at ISLA [35] and extend the models there with ranking principles based on thread structure and (language) models of the experience of participants in email exchanges and collaborative discussions.

Work on RQ2 will take the query-specific diversity ranking method of [11], adapt them to (noisy) social media and complement them with labels to make the aspects identified interpretable for human consumption and usable for iterative query formulation.

Work on RQ3 will focus on normalization, anchoring entities and relations to real-world counterparts as captured in structured information sources. This has proven to be a surprisingly hard problem [20]. So far, mostly rule-based approaches have been used in this setting; the project will break down the problem in a cascade of more fine-grained steps, some of which will be dealt with in a data- driven manner, and some in a rule-based step, following the methodology laid down in [1].

Finally, in work on RQ4, principal issues in result sets of documents will be identified through semi-automatic lexicon creation based on bootstrapping, using the initial queries as seeds [21].

For all the questions described above we plan to conduct experiments in which we will implement our newly designed techniques and evaluate them by measuring commonly used metrics like precision and recall. By experimenting with different designs and evaluating them we expect to reach the desired level of quality expected from these techniques. Evaluation will take place by participating in benchmarking events like TREC [33], CLEF [3] and INEX [18] and by cooperating with business organizations within the stated areas.

As the aim of the TREC Legal Track [34] is to evaluate search tools and methods as they are used in the context of e-discovery, participating in this track seems to be an attractive way to start of our project. We will join the 2011 track with our first implementation for which we will use the Lemur Language Modeling Toolkit [22], complemented with implementations of the lessons learned at ISLA in the work referenced above. The track will provide us with workable data, focus, a deadline and it will provide us with a first evaluation of our work.

VI. Relevance for quality in E-Discovery

This research derives its relevance for quality in E-Discovery from three factors:

First, the research connects with the present (and growing) need of trained E-Discovery practitioners. Both national and international regulators and prosecutors are facing a large increase in the amount of digital information that needs to be processed as part of their investigations.

Second, the research is relevant for legal processes, as it directly addresses evidential search. The proceedings of their investigations impact in-house and outside legal counsel who are acting on behalf of companies that are under investigation. Intelligent language processing techniques can be a solution to effectively discover relevant information and to filter legal privileged information at the same time. This is not only a Dutch problem but also extends to international cases with US en EU regulators.

Third, the research will result in (open source) web services that can be exploited in E-Discovery settings. For testing and development purposes, open sources and/or existing data sets are available.

These factors and the active involvement of E-Discovery practitioners will be realized through their involvement in use case development, data selection and evaluation. We expect that this combination will increase the effectiveness and the quality of E-Discovery while information volumes will continue to explode.

REFERENCES

[1] Ahn, D., van Rantwijk, J., de Rijke, M. (2007) A Cascaded Machine Learning Approach to Interpreting Temporal Expressions. In: Proceedings NAACL-HLT 2007.

[2] Ashley K.D., Bridewell, W. (2010) Emerging AI & Law approaches to automating analysis and retrieval of electronically stored information in discovery proceedings. Artificial Intelligence and Law, 18(4):311-320.

[3] CLEF: The Cross-Language Evaluation Forum, http://www.clef-campaign.org/

[4] Connor, John M., (2004). How high do cartels raise prices? Implications for reform ofthe antitrust sentencing guidelines, American Antitrust Institute, Working Paper.

[5] Conrad J., (2010). E-Discovery revisited: the need for artificial intelligence beyond information retrieval. Artificial Intelligence and Law, 18(4): 321-345.

[6] Corporate Counsel, (2006). The American Bar Association (ABA), section of litigation, committee on Corporate Counsel. http://www.abanet.org/litigation/committees/corporate/

[7] CREATE-IT applied research - Onderzoek/lectoren, http://www.create-it.hva.nl/content/create-it-applied-research/onderzoeksprogrammas/

[8] EDRM: Electronic Discovery Reference model, http://www.edrm.net/

[9] Feldman, R., and Sanger, J. (2006). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press.

[10] Gartner, (2009).MarketScope for E-Discovery Software Product Vendors report 2009, http://www.gartner.com/DisplayDocument?id=1262421

[11] He, J., Meij, E., de Rijke, M. (2011) Result Diversification Based on Query-Specific Cluster Ranking. Journal of the American Society for Information Science and Technology, to appear.

[12] Henseler, J.,(2010A). Openbare les E-Discovery: Op zoek naar de digitale waarheid. Amsterdam University of Applied Sciences.

[13] Henseler, J.,(2010B). Network-based filtering for large email collections in E-Discovery. Journal Artificial Intelligence and Law,Volume 18, Number 4, p.413-430

[14] Hogan C, Bauer R, Brassil D (2010) Automation of legal sensemaking in e-discovery. In: Artificial Intelligence and Law, 18(4):321-345

[15] ICAIL 2011: The Thirteenth International Conference on Artificial Intelligence and Law, http://www.law.pitt.edu/events/2011/06/icail-2011-the-thirteenth-international-conference-on-artificial-intelligence-and-law/

[16] IDC 2007: Research report on the Information Explosion.

[17] ISLA: Intelligent Systems Lab Amsterdam, http://isla.science.uva.nl/

[18] INEX: Initiative for the Evaluation of XML Retrieval, http://www.inex.otago.ac.nz/

[19] Jackson P., Moulinier I., (2002). Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization. Amsterdam: John Benjamins Publishing Company.

[20] Jijkoun, V., Khalid, M., Marx, M. de Rijke, M. (2008) Named Entity Normalization in User Generated Content. In: Proceedings of the second workshop on Analytics for noisy unstructured text data (AND 2008), pages 23-30, ACM.

[21] Jijkoun, V., de Rijke, M., Weerkamp, W. (2010) Generating Focused Topic-specific Sentiment Lexicons. In: 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010).

[22] LEMUR: Language Modeling Toolkit http://www.lemurproject.org/lemur/

[23] Manning, C. D., Raghavan, Prabhakar, and Schütze, Hinrich (2008). Introduction to Information Retrieval. Cambridge University Press.

[24] Oard, D.W., Baron, J.R., Hedin, B., Lewis, D.D., Tomlinson, S.,(2010). Evaluation of information retrieval for E-discovery. Journal Artificial Intelligence and Law,Volume 18, Number 4, p.347-386

[25] Pound, J., Mika, P., Zaragoza, H. (2010) Ad-hoc object retrieval in the web of data. In WWW 2010, pp 771-780.

[26] Rosenfeld L., Morville, P. (2002) Information architecture for the World Wide Web, 2nd edn. O'Reilly Media, Sebastopol.

[27] Scholtes, J. C., (2009). Text mining: de volgende stap in zoektechnologie. Inauguratie. Maastricht University

[28] Schwartz, M.F., Wood, D.C.M. (1993) Discovering shared interests using graph analysis. Communications of the ACM 36:78–89

[29] Sedona: The Sedona conference, http://www.thesedonaconference.org/

[30] The Sedona Conference® Best Practices Commentary on Search & Retrieval Methods (2007). http://www.thesedonaconference.org/publications_html

[31] The 2006 Socha-Gelbmann Electronic Discovery Survey Report, (2006). http://www.sochaconsulting.com/2006survey.htm/

[32] The 2008 Socha-Gelbmann Electronic Survey Report, (2008). http://www.sochaconsulting.com/2008survey.php/

[33] TREC: Text REtrieval Conference, http://trec.nist.gov/

[34] TREC Legal, http://trec-legal.umiacs.umd.edu/

[35] 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, pages 400-411

[36] Zhao F.C., Oard, D.W., Baron, J.R. (2009) Improving search effectiveness in the legal e-discovery process using relevance feedback. In: Proceedings of the global E-Discovery/E-Disclosure workshop on electronically stored information in discovery at the 12th international conference on artificial intelligence and law (ICAIL09 DESI Workshop). DESI Press, Barcelona