Semantic Search in E-Discovery: An Interdisciplinary Approach

David Graus[1], **Zhaochun Ren**[1], **Maarten de Rijke**[1]
David van Dijk[2], **Hans Henseler**[2]
Nina van der Knaap[3]

{d.p.graus, z.ren, derijke}@uva.nl
{d.v.van.dijk, j.henseler}@hva.nl
n.van.der.knaap@law.leidenuniv.nl

[1] ISLA, University of Amsterdam
[2] Lectoraat E-Discovery, Amsterdam University of Applied Sciences
[3] eLaw Group, Leiden University

Abstract

We propose an interdisciplinary approach to applying and evaluating semantic search in the e-discovery setting. By combining expertise from the fields of law and criminology with that of information retrieval and extraction, we move beyond "algorithm-centric" evaluation, towards evaluating the impact of semantic search in real search settings. We will approach this by collaboration in an interdisciplinary group of four PhD candidates, applying an iterative two-phase work cycle to four subprojects that run in parallel. The first phase we work individually. We determine the use and needs of search in e-discovery (subproject 1), and simultaneously explore and develop state-of-the-art semantic search approaches (subprojects 2–4). In the second phase we collaborate, designing user experiments to evaluate how and where semantic search can support the analysts' search process. By repeating this cycle multiple times we gain specific and in-depth knowledge and propose solutions to specific challenges in search in e-discovery.

1 Introduction

At its heart, e-discovery is the practice of sensemaking in textual corpora. Most of the time it is not exactly clear beforehand what is sought in the e-discovery setting, therefore search often starts exploratory. Moreover, forensic analysts typically refine their line of enquiry by discoveries in the data (Attfield and Blandford, 2010).

Forensic analysts are facing a large increase in the amount of digital information that needs to be processed as part of their investigations, where time and resources are limited. To facilitate exploratory search and to provide insights in large text corpora to forensic analysts. In this setting generic search, which typically focusses on high precision over recall, is not the answer.

In our work, which follows up on previous work by van Dijk et al. (2011), we study how semantic search technologies can be developed and implemented in a search engine, to support forensic analysts in their broad line of work.

To apply and evaluate semantic search in the e-discovery setting, we will combine expertise from the fields of criminology and law with that of information retrieval (IR) and information extraction (IE). In doing so, we move beyond "algorithm-centric" evaluation, towards evaluating the impact of semantic search in real search tasks.

1.1 Approach

We approach this by collaboration in an interdisciplinary group of four PhD candidates, where we apply an iterative two-phase work cycle to four subprojects that run in parallel, see Fig 2. In the first phase, we work individually. We determine the use and needs of search in e-discovery (subproject 1), and simultaneously explore and develop state-of-the-art semantic search approaches (subprojects 2–4; *Semantic Analysis*). In the second phase, we collaborate and design user experiments to be able to evaluate more precisely how and where semantic search can support the analyst's search process while at the same time gaining new insights into this process.

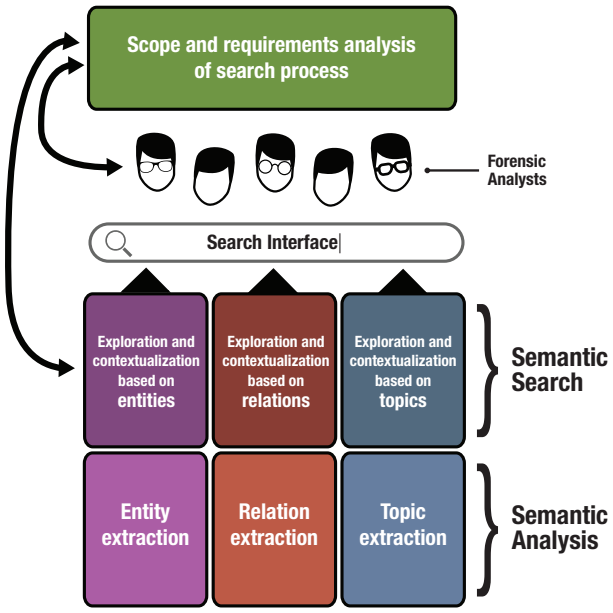

Figure 1: Schematic representation of subprojects

1.2 Challenges

In e-discovery, retrieving all relevant traces is important. In the search context, this means that the focus should be on high recall, in contrast to traditional (web) search (Oard and Webber, 2013). Furthermore, data used in e-discovery is typically on a case-by-case basis, it can be noisy and is diverse in nature and origin. The amount of digital information to process in investigations is continually growing. Time and resources are limited in investigations, so one cannot rely on vast amounts of manual annotations, as is common in widely used supervised machine learning approaches. We take these intuitions and observations as a starting point, and expect to gain more specific knowledge on challenges in the field from subproject 1.

1.3 Semantic Search

Semantic Search is a paradigm in Information Retrieval (IR) which applies structured knowledge, e.g. discussion structure, topical structure or entities and relations, as a complement to text retrieval (Pound et al., 2010). In this work, we apply semantic search in the e-discovery search setting in two ways:

1. To complement (traditional) retrieval tasks, e.g. document classification (relevance/non-relevance or privileged/non-privileged) and document similarity metrics.
2. To provide guidance in analysts' search or sense making process.

We believe it is important to both understand the intricacies and specificities of forensic analysts' search process, as the available and suitable state-of-the-art in semantic search technologies, in order to effectively determine how semantic search fits in this search process. We approach this task with an interdisciplinary team, combining domain-specific expertise in criminology and law[1] with expertise in semantic search[2].

The rest of this paper is organized as follows; in section Section 2 we explain subproject 1: analysis and review of the search process, in Section 3 we describe sub projects 2 through 4, in Section 4 we describe the collaborative approach, and finally in Section 5 we describe the contribution and novelty of our interdisciplinary approach.

2 Subproject 1: Analyzing the search process

The first subproject focusses on the human aspect of e-discovery in investigations. Here, we establish the use and needs of the field of e-discovery, because it can provide information about the use of such methods, the need for improvement and where these needs lay. Furthermore, we can observe the search process of professionals to see where enhancements can be made.

We will first conduct an extensive literature study in a multidisciplinary fashion. This will consist of literature found in three important fields of study, namely; information technology research, mostly on the topic of e-discovery and digital forensics in al ways, shapes and forms (Attfield and Blandford, 2010; Oard and Webber, 2013; Biros et al., 2007; Casey, 2011; Garfinkel, 2010), combined with criminological research and law research (van Wilsem, 2011; Carrier, 2002). Together this will provide a balanced overview of e-discovery in investigations.

Technological, legal and criminological literature will be gathered and used to provide a strong framework upon which we can build by interviewing pro-

[1] http://law.leiden.edu/organisation/metajuridica/elaw/
[2] http://ilps.science.uva.nl

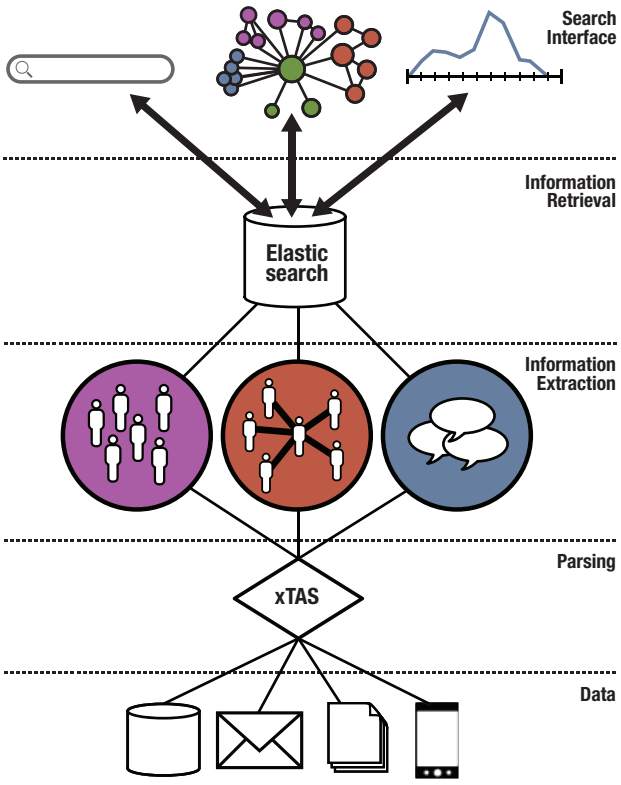

Figure 2: Information Extraction system layout

fessionals and practitioners who work with digital evidence. These interviews provide an overview of the uses of e-discovery and the desired situation according to professionals and practitioners. By gaining insight in their search process, uses, needs and limitations of current search, we identify areas where we can effectively improve search systems.

3 Subprojects 2–4: Information Extraction for Semantic Search in E-Discovery

The initial focus for projects 2 through 4 is on developing approaches for Information Extraction (IE), the step that precedes the integration of structured knowledge into retrieval systems for semantic search: Information Retrieval (IR). In three subprojects, we focus on entity extraction, relation extraction and topic extraction In the initial phase, we evaluate these subprojects using standard test collections. This allows us to study where and how semantic search technologies can improve traditional retrieval tasks, e.g. classification.

1. **Entity extraction** addresses the task of identifying and resolving entities in documents. Extracting entities can provide building blocks for more elaborate information extraction tasks such as issue mapping, event extraction, or 'cold start knowledge base creation' (van Dijk et al., 2011)

2. **Relation extraction** entails identifying entities in a specified semantic relation, which can for example be combined to help analysts identifying key individuals, by generating networks from E-Discovery data sources.

3. **Topic extraction** refers to the task of detection and tracking of topics within document collections. In E-Discovery, topic extraction could help analysts understand what is happening by supporting in discovering hidden events.

As we described in Section 1, an important task in e-discovery is to provide insights from large corpora. These three focal points provide building blocks for more elaborate information extraction tasks such as identifying events, interactions between individuals and evolution or change of topics in conversation logs, email databases or collaboration platforms (van Dijk et al., 2011).

Considering the e-discovery-specific constraint of diverse, noisy and "case-by-case" nature of data and its increasingly large scale, we cannot rely on extensive manual annotations. We thus restrict ourselves to approaches based on (i) semi-supervised learning, where a small amount of seed-data is used as a starting point for recognizing patterns, and/or (ii) unsupervised approaches, where the learning process doesn't rely on any annotations. Furthermore, since availability of experts can be assumed – forensic analysts who search for relevant documents – we also consider interactive machine learning methods, such as active learning (Settles, 2009), where experts improve algorithms by labeling 'border-cases'.

4 Collaboration

In the second phase we collaborate and join findings from subproject 1 with the algorithms developed in subprojects 2–4.

By collaboratively designing user experiments that leverage both findings of analysts' search with

application of developed information extraction algorithms, we can both gain more insights into the search process of analysts, and allow us to measure whether and how semantic search supports analysts in their search process.

A typical approach would be to compare two groups of analysts performing a common e-discovery task, each using a different search system: one representing the current practice, and another representing our "semantic search-enabled" system.

To study the exploitation of extracted structured information, a suitable search interface should allow the end-users to intuitively and flexibly interact with documents and available information in increasingly large data collections. Because of the size and focus on high recall, methods of efficiently presenting extracted information and allowing users to interact with it is an important subtask. E.g. possibilities include visualizing identified entities, their relationships or interaction, while at the same time allowing analysts to interact with the temporal dimension of the data. Expertise in effectively designing user interfaces for specific tasks and domains builds on previous work (Bron et al., 2012; de Rooij et al., 2013).

5 Contribution

By our interdisciplinary approach, spanning the fields of criminology, law, Information Retrieval and Natural Language Processing, we position ourselves between strictly empirical/field work of Attfield and Blandford (2010) and Computer Science perspective efforts such as work by Oard and Webber (2013).

By starting work individually, and collaborating only after this first phase, we ensure an efficient workflow, and minimize in collaborative effort. Furthermore, this iterative workflow allows us to gain detailed insights into a specific subset of tasks in the general e-discovery search process.

The use and practice of e-discovery on a small scale, particularly because of the feedback-loop by means of user studies, we can get more specific and in-depth understanding of subtasks in the search process.

The field of IR will benefit from new insights into a sub-domain with specific and well-understood characteristics which is relatively unexplored – our findings could prove useful in other domains with similar constraints, e.g. in exploratory search for historians and similar tasks in the field of digital humanities, where current natural language processing and retrieval models prove insufficient due to mismatches between available training data and real data.

And finally, the fields of law and E-Discovery practitioners will benefit from improved tooling and understanding of the search process.

6 Acknowledgements

This research was partially supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements nr 258191 (PROMISE Network of Excellence) and 288024 (LiMoSINe project), the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.004.802, 727.011.005, 612.001.116, HOR-11-10, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP and BILAND projects funded by the CLARIN-nl program, the Dutch national program COMMIT, by the ESF Research Network Program ELIAS, Elite Network Shifts project funded by the Royal Dutch Academy of Sciences, the eLaw group of Leiden University's Institute for the Interdisciplinary Study of the Law, and Fox-IT.

References

[Attfield and Blandford2010] Simon Attfield and Ann Blandford. 2010. Discovery-led refinement in e-discovery investigations: sensemaking, cognitive ergonomics and system design. *Artificial Intelligence and Law*, 18(4):387–412.

[Biros et al.2007] David P Biros, Mark Weiser, and John Witfield. 2007. Managing digital forensic knowledge an applied approach. In *Australian Digital Forensics Conference*, page 11.

[Bron et al.2012] M. Bron, J. van Gorp, F. Nack, M. de Rijke, and S. de Leeuw. 2012. A subjunctive exploratory search interface to support media studies

researchers. In *SIGIR '12: 35th international ACM SIGIR conference on Research and development in information retrieval,*, pages 425–434, Portland, Oregon. ACM, ACM.

[Carrier2002] Brian Carrier. 2002. Open source digital forensics tools: The legal argument.

[Casey2011] Eoghan Casey. 2011. *Digital evidence and computer crime: Forensic science, computers, and the internet*. Academic press.

[de Rooij et al.2013] O. de Rooij, D. Odijk, and M. de Rijke. 2013. Themestreams: Visualizing the stream of themes discussed in politics. In *SIGIR'13: 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, ACM.

[Garfinkel2010] Simson L Garfinkel. 2010. Digital forensics research: The next 10 years. *Digital Investigation*, 7:S64–S73.

[Oard and Webber2013] Doug Oard and William Webber. 2013. Information retrieval for e-discovery. *To Appear*.

[Pound et al.2010] Jeffrey Pound, Peter Mika, and Hugo Zaragoza. 2010. Ad-hoc object retrieval in the web of data. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 771–780, New York, NY, USA. ACM.

[Settles2009] Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

[van Dijk et al.2011] David van Dijk, Hans Henseler, and Maarten de Rijke. 2011. Semantic search in e-discovery. In *DESI IV Workshop on Setting Standards for Searching Electronically Stored Information In Discovery Proceedings*, pages 109–112, Pittsburgh PA, June. University of Pittsburgh School of Law.

[van Wilsem2011] Johan van Wilsem. 2011. bought it, but never got itassessing risk factors for online consumer fraud victimization. *European Sociological Review*.