On the Benefits of Information Retrieval and Information Extraction Techniques Applied to Digital Forensics

David Lillis[1,2] and Mark Scanlon[2]
[1] Beijing-Dublin International College
[2] School of Computer Science, University College Dublin
{david.lillis, mark.scanlon}@ucd.ie

Abstract. Many jurisdictions suffer from lengthy evidence processing backlogs in digital forensics investigations. This has negative consequences for the timely incorporation of digital evidence into criminal investigations, while also affecting the timelines required to bring a case to court. Modern technological advances, in particular the move towards cloud computing, have great potential in expediting the automated processing of digital evidence, thus reducing the manual workload for investigators. It also promises to provide a platform upon which more sophisticated automated techniques may be employed to improve the process further. This paper identifies some research strains from the areas of Information Retrieval and Information Extraction that have the potential to greatly help with the efficiency and effectiveness of digital forensics investigations.

Keywords: Digital Forensics, Information Retrieval, Information Extraction

1 Introduction

Digital forensic investigations remain a labour-intensive manual task. A long backlog has emerged in many jurisdictions whereby investigations may take months or years to yield useful results. This has consequences for the timeline of prosecutions reaching the courtroom, but also has the effect that the timeline involved in a digital forensic investigation is at odds with the normal timeline of investigation, with the digital evidence often being unavailable during the crucial initial stages of a criminal investigation. Clearly, expedited investigations are desirable, and the digital forensics community have been working towards this end.

The primary goal is increased levels of automation so as to reduce the amount of manual work required to conduct an investigation. This can be achieved by improving the speed of evidence processing so as to reduce the time spent waiting for results. If this can be done, there is then scope for more sophisticated automated techniques to be applied to the problem.

Researchers in the areas of Information Retrieval (IR) and Information Extraction (IE) have been developing techniques for decades that help people to sift through large quantities of information as quickly and efficiently as possible. To date, few of these techniques have made the cross-over to the day-to-day lives of digital forensic investigators. This paper seeks to examine the context within which these may be applied to investigations, and the advantages they may bring to this area.

2 A Platform for More Efficient Processing

Traditional digital forensics generally involve the examination of a seized hard drive, using specialist digital forensic software installed on a workstation. Examination of evidence is a lengthy, manual process, which has led to long backlogs in evidence processing for police forces throughout the world.

However, more recent developments in digital forensics technologies, though not yet widely deployed, promise a future forensic investigation platform upon which more sophisticated technologies may be deployed.

The growth of cloud computing in recent years has caused its own challenges for the digital forensics community, as evidence becomes more widely spread across jurisdictional boundaries. However, it also offers great benefits in terms of evidence processing. This type of cloud-based evidence processing has become known as Digital Forensics as a Service (DFaaS) [1] and it has already been deployed with success in the European Cybercrime Centre (EC3), based in the Netherlands [2].

Moving the processing of evidence to the cloud allows for the application of greater computing resources through the use of parallelisation and distribution. Additionally, it allows the introduction of high performance computing techniques, including the application of specialist hardware to particular tasks. It also allows prioritisation of certain investigations as operational requirements demand.

A further advantage of a cloud-based system is that investigations no longer occur in isolation. Many seized hard disks contain large quantities of data in common (e.g. operating system and application files), all of which must be processed anew for each investigation. DFaaS offers a platform for deduplication techniques to be applied [3]. This involves the identification of identical files on different seized hard disks through the use of hashing, which has numerous advantages. Firstly, files that have been previously been analysed in a previous investigation do not require re-examination. This can be used either to eliminate files from consideration or to identify files that have been considered pertinent to a previous investigation (e.g. a shared collection of child pornographic images). An additional speed advantage is that duplicate files do not need to be transferred to the system again, which reduces the time required to capture all available evidence. This is of particular concern when using remote acquisition techniques that transfer evidence through an internet connection while in the field [4]. Reducing the quantity of data to be transferred and analysed is a key step in expediting the investigation process.

3 The Applicability of Information Retrieval

With the development of a high-capacity, cloud-based digital forensic investigation platform comes the opportunity to add more sophisticated automated techniques to the traditional digital forensic process. This has the potential to reduce the time spent by investigators in carrying out their existing tasks, while also potentially empowering them to take a more active role in other aspects of the investigation. One source of such techniques is the field of IR, which is concerned with identifying documents (traditionally text documents, though multimedia IR is also an active research area) that satisfies a user's "information need". Applied in the context of digital forensics investigations, the information need is for anything that is pertinent to the investigation being conducted.

Traditionally, IR effectiveness is evaluated using the oft-conflicting measures of precision and recall. A system with high precision avoids returning documents that are not relevant, whereas a high-recall system aims to ensure that all available relevant documents are returned to the user. While a high level of both is desirable in any situation, different application areas would tend to favour one over the other if this cannot be achieved. Indeed, it is frequently the case whereby improving on one measure involves sacrificing effectiveness according to the other. The classic example of a use case where precision is key is the area of web search. Since potentially millions of relevant documents exist, users do not require access to all of them, and have a strong aversion to spending time on non-relevant documents. In contrast, legal search is frequently proposed as a scenario where recall is of paramount importance. In a situation where a single piece of missing evidence may be crucial to a case, a user will be more tolerant of non-relevant documents if full recall can be provided.

Digital forensics is typically seen as the latter situation (e.g. in [5, 6]). Whereas high recall inevitably leads to a higher rate of false positives, this is tolerated due to the requirement to find all available evidence. Certainly for a case to be presented in court, all available evidence must be identified, which requires recall to be maximised. Absent evidence clearly has the potential to undermine both criminal and civil cases. The principal advantage of using IR techniques is that it helps to achieve this. Once the initial processing stage has been completed, queries can be run extremely quickly. Indeed, achieving high recall likely requires many queries to be run. As noted in [7], there is a large degree of variation in the vocabulary that searchers use to describe their information needs. Less than 20% of users use the same keywords for topics they are interested in. The use of standard IR techniques such as synonym matching and query expansion can aid with the process of improving recall, without requiring investigators to enter every possible query manually.

One example of this is [8], where query expansion and query reduction techniques were applied to the popular Enron email dataset to improve retrieval performance. WordNet was used for domain-independent query expansion, with a Latent Semantic Indexing approach to query reduction being applied afterwards.

However, this focus on recall arguably has negative consequences at earlier stages in the investigation. As noted in [2], because the results of digital forensics investigation are not typically available within the first few days of an investigation, they are rarely taken into account during the crucial initial stages where hypotheses are being formed and leads being investigated. This leads to the observation that while high recall is necessary for court, the manual work it requires is too time- consuming for the early stages of the investigation.

Figure 1 offers a simple illustration of the timeline of an investigation. At the end of the investigation, evidence is required to be complete and court-admissible in order to prove a case on court. This requires high recall as in classic legal search scenarios. However, towards the initial stage of an investigation, precision is of far greater importance. Investigators will require relevant documents quickly. This allows the most relevant evidence to be used during the early stages of the investigation process as quickly as possible. Additionally, in cases where large numbers of devices have been seized (e.g. from an office or data centre), a triage stage is required to identify hard disks that may contain pertinent information. Achieving high precision at early stages has the potential to very quickly identify those machines that merit further investigation.

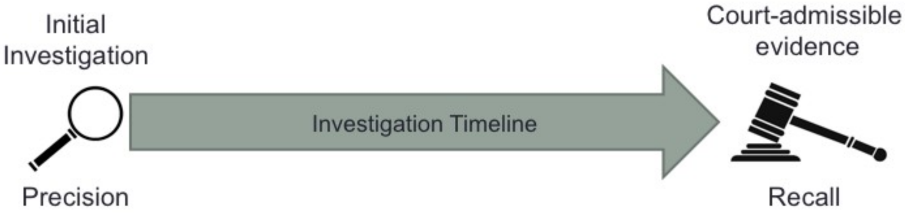

Fig. 1. Investigation Timeline

As the investigation progresses, there is a growing requirement for higher recall, leading to the classic legal search by the end. However, it is clear that investigative requirements shift during the process of an investigation, and a configurable approach to IR is likely to very advantageous in both increasing the quality and the efficiency of the process.

Additional features of IR research that have been applied to digital forensics include visualisation approaches such as ranking [9] and clustering [10]. While these have become commonplace in many search implementations, they have not yet reached ubiquity in popular forensic software. Innovations such as these have the potential to reduce the manual burden on investigators, even when the late recall-oriented stage of the investigation is reached.

4 The Applicability of Information Extraction

In addition to IR, which seeks only to identify documents containing relevant information, IE further attempts to extract meaningful structured information from unstructured files. This also has the potential to improve the efficiency of investigations by automatically discovering and extracting evidence on an investigator's behalf.

Initial efforts have already begun in applying IE to forensics investigations. For example, [11] uses two-phase IE for forensic investigation (evaluated using the Enron dataset). The two phases employed were named entity recognition and association rule mining. The initial phase sought to identify entities in emails (people, places, organisations, etc.) with the second attempting to identify how entities are related. This type of system has great potential in aiding investigators. This work also had a strong emphasis on the visualisation aspects of the system, whereby named entities were highlighted in text, named entities were displayed in word clouds and the nature of mined entity relationships were presented also. Another application in the context of email can be found in [12], while named entity extraction has also previously been applied to police reports [13]. All of these approaches help to lower the cognitive load placed on an investigator while examining large quantities of digital evidence.

One other long-standing aspect of IE that has potential for digital forensics is text summarisation [14]. This is most commonly seen in news reports and web search results where relevant snippets of a document are presented to a user to help decide whether a document is relevant without requiring the entire document to be read. By showing the context in which key words are displayed, time can be saved in identifying relevant documents. While this has not yet reached mainstream forensic investigation interfaces, it has the potential to expedite the process if incorporated in the future.

A further consideration is that event timeline reconstruction is extremely important in criminal investigations [15]. Investigators desire to construct a chain of events in temporal order. Existing automated approaches to this task are typically done at a low level (e.g. filesystem and logging events) [16], with events such as connecting a USB stick being considered as “high-level” in that context.

Related to this, efforts have been ongoing to extract temporal information from unstructured text [17]. While the tasks are not identical, both the Text REtrieval Conference (TREC)[3] and the Text Analysis Conference (TAC)[4] feature a temporal track. In the TAC Event Track, participants are asked to extract information about events so that a knowledge base can be populated. This is then used to identify when mentions of events in text related to the same event, as well as to link mentions in terms of the role they play in an event and their timing within it. The TREC Temporal Summarization Track has the goal of developing systems that can efficiently monitor event details over time. These are research areas with the potential to be of great benefit to digital forensics investigators.

5 Conclusions

Research from the text retrieval and analysis communities have great potential for reducing the manual workload on digital forensics investigators. This in turn has the potential to help clear the evidence backlog, and so allowing for digital evidence to be meaningfully included at an earlier stage in investigations. The traditional focus on recall for IR in forensics is not necessarily appropriate for all stages of an investigation, with precision being arguably more appropriate at earlier stages. A movement of forensics towards cloud computing and related technologies will provide a platform upon which cutting-edge techniques such as named entity extraction, association rule mining, temporal information extraction and others to be incorporated into the investigation process in the future.

[3] http://trec.nist.gov

[4] http://www.nist.gov/tac

References

1. Lee, J., Un, S.: Digital forensics as a service: A case study of forensic indexed search. In: ICT Convergence (ICTC), 2012 International Conference on. (2012) 499–503
2. Van Baar, R.B., van Beek, H.M.A., van Eijk, E.J.: Digital Forensics as a Service: A game changer. Digital Investigation 11 (2014) S54–S62
3. Watkins, K., McWhorte, M., Long, J., Hill, B.: Teleporter: An analytically and forensically sound duplicate transfer system. Digital Investigation 6(SUPPL.) (2009) 43–47
4. Scanlon, M., Kechadi, M.T.: Online Acquisition of Digital Forensic Evidence. In Goel, S., ed.: Digital Forensics and Cyber Crime: First International ICST Con- ference, ICDF2C 2009, Albany, NY, USA, September 30-October 2, 2009, Revised Selected Papers, Berlin, Heidelberg, Springer Berlin Heidelberg (2010) 122–131
5. Beebe, N.L., Clark, J.G.: Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results. Digital Investigation 4(SUPPL.) (2007) 49–54
6. Beebe, N.: Digital forensic research: The good, the bad and the unaddressed. In: Advances in digital forensics V. Springer (2009) 17–36
7. Furnas, G.W., Deerwester, S., Dumais, S.T., Landauer, T.K., Harshman, R.A., Streeter, L.A., Lochbaum, K.E.: Information retrieval using a singular value decomposition model of latent semantic structure. In: SIGIR ’88: Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA (1988) 465–480
8. Du, L., Jin, H., de Vel, O., Liu, N.: A latent semantic indexing and WordNet based information retrieval model for digital forensics. In: 2008 IEEE International Conference on Intelligence and Security Informatics, IEEE (2008) 70–75
9. Beebe, N.L., Liu, L.: Ranking algorithms for digital forensic string search hits. Digital Investigation 11(SUPPL. 2) (2014) 314–322
10. Beebe, N.L., Clark, J.G., Dietrich, G.B., Ko, M.S., Ko, D.: Post-retrieval search hit clustering to improve information retrieval effectiveness: Two digital forensics case studies. Decision Support Systems 51(4) (2011) 732–744
11. Yang, M., Chow, K.P.: An Information Extraction Framework for Digital Forensic Investigations. In: Advances in Digital Forensics XI. Springer (2015) 61–76
12. De Vel, O., Anderson, A., Corney, M., Mohay, G.: Mining e-mail content for author identification forensics. ACM Sigmod Record 30(4) (2001) 55–64
13. Chau, M., Xu, J.J., Chen, H.: Extracting meaningful entities from police narrative reports. In: Proceedings of the 2002 annual national conference on Digital government research, Digital Government Society of North America (2002) 1–5
14. Salton, G., Singhal, A., Mitra, M., Buckley, C.: Automatic text structuring and summarization. Information Processing & Management 33(2) (1997) 193–207
15. Chabot, Y., Bertaux, A., Kechadi, M.T., Nicolle, C.: Event Reconstruction: A State of the Art. In: Handbook of Research on Digital Crime, Cyberspace Security and Information Assurance. IGI Global (2014) 231–245
16. Hargreaves, C., Patterson, J.: An automated timeline reconstruction approach for digital forensic investigations. Digital Investigation 9 (2012) S69–S79
17. Campos, R., Dias, G., Jorge, A.M., Jatowt, A.: Survey of Temporal Information Retrieval and Related Applications. ACM Computing Surveys 47(2) (2014) 1–41