Domain-Specific Search Engines for Investigating Human Trafficking and Other Illicit Activities

Mayank Kejriwal
University of Southern California, USA

INTRODUCTION

Web advertising related to Human Trafficking (HT) activity has been on the rise in recent years (Szekely et al., 2015). Question answering over crawled sex advertisements to assist investigators in the real world is an important social problem. This problem involves many technical challenges (Kejriwal & Szekely, 2017c). This article will describe the problem of *domain-specific search (DSS)*, a specific set of technologies that can address these challenges. Modern DSS systems for investigative activities draw on cutting-edge techniques developed over three years of DARPA-funded research conducted in collaboratively academic (e.g., the University of Southern California's Information Sciences Institute), government (e.g., NASA's Jet Propulsion Laboratory) and industrial (e.g., Uncharted) settings. Evidence from the HT domain shows that the systems can be used to provide valuable utility to analysts and investigative experts.

In illicit domains such as HT but also others like securities fraud and narcotics, domain-specific search involves a form of *Information Retrieval* (IR) that takes as input a large *domain-specific* corpus of pages crawled from the Web. The system allows investigators to satisfy their information needs by posing *sophisticated queries* to a special-purpose engine. A workflow of this process is shown in Figure 1. Since investigators are largely non-technical, they must be able to issue such queries to (and receive responses from) intuitive, graphical interfaces. A fully functional DSS engine must have some notion of semantics, since sophisticated queries go beyond just keyword specification. This is because investigative queries are more like real-world questions requiring complex operations like aggregations (e.g., *find me all email addresses linked to the phone 123-456*).

A viable solution to the problem has to allow the user to pose queries both intuitively and interactively.

For such a DSS to operate semi-automatically and be useful in the real world, several challenges and desiderata must be fulfilled. Possibly the most important of these is handling the unusual nature of an illicit domain, since investigators who have to use the system have special needs. To understand why this can be challenging, consider the recent advent of technologies like neural networks and deep learning. Pre-trained tools such as word embeddings and Named Entity Recognizers in the natural language processing community have also been released for public use (Pennington, Socher & Manning, 2014). However, many of these tools have been trained on datasets and corpora that are fairly 'regular' i.e. comprise of relatively well-structured text (like news corpora and Wikipedia articles). Consequently, they are not necessarily suitable for language or data acquired in illicit domains. Table 1 illustrates some examples of real text scraped from sex advertisement webpages (but with identifying phone numbers appropriately modified). Acquiring and labeling data from such domains is both expensive and sensitive,

DOI: 10.4018/978-1-5225-9715-5.ch033

Figure 1. A procedural workflow of domain-specific search from the point of view of an investigative user, using the domain-specific insight graph (DIG) DSS for example interfaces