Technology-assisted Investigative Search: A Case Study from an Illicit Domain

Mayank Kejriwal
Information Sciences Institute
USC Viterbi School of
Engineering
Marina Del Rey, CA 90292
kejriwal@isi.edu

Pedro Szekely
Information Sciences Institute
USC Viterbi School of
Engineering
Marina Del Rey, CA 90292
pszekely@isi.edu

CHI'18 Extended Abstracts, April 21–26, 2018, Montreal, QC, Canada

ACM 978-1-4503-5621-3/18/04...$15.00
https://doi.org/10.1145/3170427.3174364

Abstract

For many, search engines like Google and Bing offer excellent facilities for satisfying information needs. However, a class of needs not currently addressed by generic search engines is investigative search, which has massive potential for using adaptive technology for social good. In this case study, we describe the challenges of investigative search in the online sex trafficking domain, along with the insights gained from user feedback in using a real-world investigative search system developed in our group.

Author Keywords

Specialized Information Retrieval; Knowledge Graphs; Investigative Search; Domain-specific Search; Human Trafficking; Illicit Domains.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction

The Web has had a tremendous positive impact on productivity, but has also lowered the barrier for entry for players in illicit domains like identity theft, securities fraud and sex trafficking. For example, ads peddling

Illicit Domains: Some Examples

- Online Sex Trafficking (primary subject of this case study)
- Penny Stock Fraud
- Counterfeit FPGA Manufacturing
- Illicit Mail Shipments
- Narcotics
- Illegal Weapons Sales

sex activities in US territories now number in the tens of millions, based on a recent crawl conducted under the DARPA MEMEX program [16]. Identifying potential victims of trafficking, which can also help identify runaways and missing people, is an important social problem that technology can help address.

The DARPA MEMEX program was established with the intent of building *domain-specific search systems* that could quickly generalize to an arbitrary domain. A domain in this context is a relatively diffuse term, as there may not be an explicit ontology describing it. A class of domains that is of interest to the MEMEX program is *investigative* in nature, whereby consumers of the technology are investigators and field analysts who are investigating *illicit* activity believed to have a significant Web footprint in terms of advertisements, transactions and reviews (or all of them).

In today's digital era, most people rely on search engines like Google and Bing for their needs. Like the smartphone, tablet and desktop, these engines are good examples of technology that are optimized for the general populace, with commercially oriented goals. Historically, and even presently, domains where building specialized technology for search and analytics are necessary tend to have massive commercial (e.g., Business Intelligence [4], [5]) or military implications [6]. Many agencies at the state and local levels, even in a developed country like the United States, cannot afford such proprietary, state-of-the-art technology in the present resource-strapped environment.

The search systems that have been built under MEMEX are designed to adapt in a very general, user-centric way to arbitrary domains by (1) taking sparse specifications of the domain from a user, (2) using these specifications in interactive focused crawling systems to scrape webpages that are likely relevant to fulfilling some subset of the user's informational needs, and (3) extracting useful structured information from the corpus of webpages and building an intuitive search and analytics engine over this constructed 'knowledge base'. The process, expected to be iterative, is refined over time as users invest more effort into the system.

The online sex trafficking domain offers a powerful illustration of the potential of a fully functional domain-specific search pipeline implemented using the high-level workflow above. With this context in place, we use this case study on investigative search in the online sex trafficking domain to achieve several ends. First, we discuss the needs of investigators, and the flavors (and examples) of questions they are looking to answer. Specifically, what is investigative search and why can't generic search engines like Google currently address the challenges of domain-specific search in illicit domains like sex trafficking? In response to these needs, we built (over a multi-year period of research) and made available to investigators, a system called Domain-Specific Insight Graphs (DIG) [7]. In addition to describing the high-level design and workflow in DIG, we also detail the protocol and setup used by NIST to evaluate the system with actual online sex trafficking investigators. We present this case study with the two related aims of shedding light on a new and interesting domain, as well as the design and evaluation of search technology used to assist experts in the domain.

Examples of Lead Investigation Questions

(with artificial identifiers)

Point Fact: List all ads, with social-media-id, containing phone 123-456, optionally with location Key Biscayne, Florida occurring somewhere in the ad text or page.

Cluster Facet: List all ads connected via a *shared* phone number or email to the phone number 987654.

Cluster Identification: List all ads connected via a shared phone number or email to the phone number 987654 that feature a Cuban escort, with location *filtered* on Florida.

Cluster Aggregate: Find *average* height of escorts associated directly or via shared phone/email links with phone number 123-456-7890

Background

Engineering interactive *Search and Analytics* architectures for specialized purposes has been explored in several applications, most notably intelligence and situational awareness dashboards [1], [9] (e.g., in the Humanitarian Assistance and Disaster Relief or HADR domain [15], [10], or like the products developed by Palantir [17]) and Business Intelligence [4], [5]. These domains are well-funded, and of interest to commercial entities or to national and military organizations. In contrast, work on illicit domains has been limited. Taking online sex trafficking as a specific case, many state and city departments do not have a dedicated unit for tackling this problem. At the same time, such domains do have a significant Web presence, which makes them amenable to adaptive analytics at scale.

The AI community has increasingly acknowledged the need for building systems for social good, and for ensuring that algorithms do not incorporate damaging biases. Recently, to that effect, a new conference on AI and Society was jointly established by AAAI and ACM [19]. In addition to offering insight into investigative needs in an actual illicit domain, we also hope that the DIG architecture outlined in this paper serves as a case study of an adaptive AI system that is already being put to good use by over 200 law enforcement agencies in prosecuting human trafficking. DIG is neither commercial, nor developed under commercial funding. The system (but not the data) has been rolled out in an open source release under a permissive license. Technical innovations included in DIG have been described in several specialized papers over the years, representative instances being [15], [18].

Investigative Search

Based on discussions with actual investigators, we can 'break down' investigative search queries into categories: *lead generation* and *lead investigation*. Taking Google as an analogy, lead generation is like clicking on the 'I feel lucky' button, or looking at trending topics. Lead investigation, which can be evaluated in more concrete ways, starts when the user already has a lead in hand. This lead may come from a tip-off on the ground, a suspicious statement or arrest record, or a matter of public record, like the name of a person who has recently gone missing.

Lead investigation questions can be classed into one of four categories (using DARPA and NIST terminology): *point fact, cluster facet, cluster identification* and *cluster aggregate*. Although there is a formal way to define the four classes, we demonstrate via an example of each in the sidebar.

Although not obvious, the lead investigation questions do provide some intuition into why generic search engines like Google are not adequate for satisfying such information needs (at least not without significant engineering). Two problems, *information obfuscation* and *ambiguity*, are respectively related to the quality of the information in the ads and to the difficulty of automating a longstanding Artificial Intelligence (AI) problem called *Information Extraction (IE)* [2], [3]. Concerning the former, online sex ads are written with the intent of being indexed by certain attributes (mainly *location* and the kind of *sex service* being advertised), but other attributes (phone number, social media ids and addresses) are deliberately obfuscated in creative ways so that they are human (but not machine) readable. For example, the following text fragment

AVAILABLE NOW! ?? - (1 two 1) six 5 six - 0 9 one 2 contains a phone number that a human would be able to recognize and dial, but an IE system (also, Google) would have a hard time extracting and indexing in a database, unless it was extensively tuned or engineered. Similarly, ambiguity arises when there is confusion about whether a word (e.g., *Charlotte*) links to the city in North Carolina or a person. Humans use context to make this determination, and while there have been significant developments in machine reading methods, overall accuracy is still quite low, especially for 'difficult' domains like human trafficking [12],[13].

Even without obfuscation and ambiguity, the cluster and aggregate questions illustrate another kind of facility not offered by Google. These questions require a user to retrieve a set of *related* ads e.g. via a shared phone number[1] (for the cluster questions), and also to aggregate quantities across the set. This is a task that involves both search and analytics currently not included in the Google interface.

Unlike lead investigation questions, lead generation questions were significantly more controversial, as users were divided on what makes for a *useful* (or more importantly, a *not harmful*) lead generation question, and even on whether explicitly incorporating lead generation into a domain-specific search system was warranted or the best use of resources. This was an eye-opening insight that significantly changed the way we approached our efforts in the final year of the MEMEX program (2017).

In the initial phase of the program, one of the task challenges that technical performers were called upon to participate in was that of predicting whether a given set of scraped sex activity ads (each set representing a 'case') *indicated* some level of trafficking. It was believed that this was a useful problem, and that an imputed 'risk score' (similar to a credit score) for such a set could be used to generate leads [11]. Many investigators who were told about this task challenge felt that the problem was misguided, if not impossible to solve, for several reasons. One important reason is that the ad often does not contain enough information for such predictions, and algorithms may end up being biased in favor of certain ad characteristics (e.g., racy words in the text). Some investigators also feel that they already have enough leads to investigate, and that they are not looking to generate more leads. Others felt that they could not *trust* the outputs of a machine learning system without some *explanation*. In response to these concerns, participants in the program investigated novel methods and investigations into the emerging topic of dataset and algorithmic bias, and explainable machine learning. The open-source ELI5 package[2], which supports off-the-shelf explainable machine learning for text data [8], was implemented by a MEMEX participant and was a direct output of this exercise.

Domain-specific Insight Graphs (DIG)

Our group at the USC Information Sciences Institute developed a system called DIG [7], [14] to facilitate some of the desired search capabilities described previously.

[1] Two phone numbers are shared either when they are extracted from the same ad or when the same person has been linked to both phone numbers through a sequence of ads.

[2] Explain it like I'm 5.

Technologies

DIG is a complex ensemble of various adaptive technologies that is ultimately designed to help users perform investigative search. The system described in this case study was specifically optimized for investigative querying in the online sex trafficking domain, although it is currently also being extended to address search in arbitrary domains. DIG includes:

- A suite of IE technologies [12], [13] that use rules, heuristics and advanced machine learning to automatically extract a variety of relevant structured attributes from webpages, including phones, emails, names, ages, physical attributes, sex services, and locations. We refer to this collective set of interlinked extractions as a *knowledge graph.*
- Indexing techniques for ensuring fast retrieval over both the knowledge graph and the text in the webpages [14].
- A cached copy of every processed webpage, in case the webpage is taken offline on the live Web (as it often is), and also to foster trust in, or verify, IE.
- Provenance data about which algorithm led to which extraction in the knowledge graph.
- Images and image similarity search based on deep neural network-based computer vision.
- Big Data architecture to support streaming ingestion of new ads.

Typical Search and Analytics Workflow

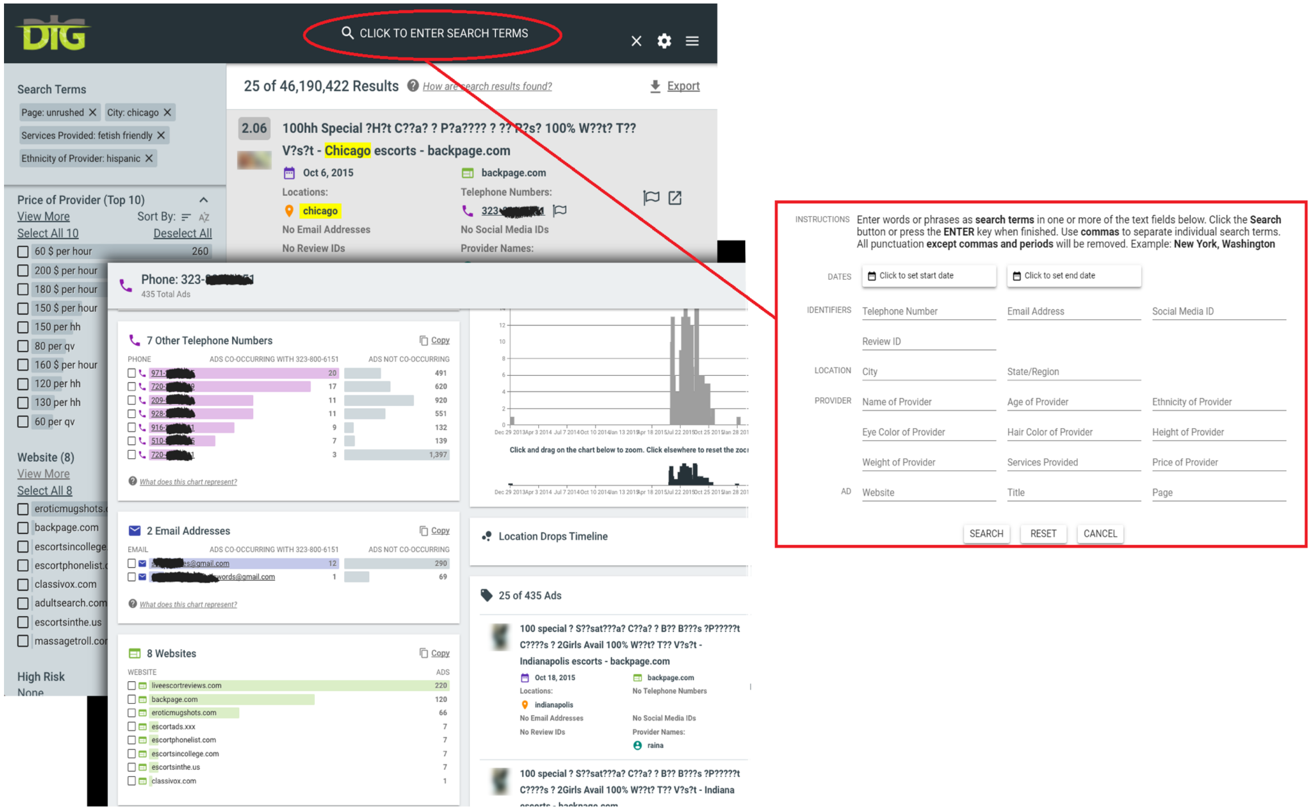
Search in DIG is supported in a variety of ways. Users start by filling out some fields (or just free text) on a search form (Figure 1). As DIG retrieves results, *facets* on the side get populated so that users get a sense of how the data is distributed (e.g., if results indicate that many hits seem to be spatially localized). Users can browse any document in the ranked list of documents, obtain provenance for the extractions, open the cached webpage associated with each document, and click on the document to see temporal and geospatial information. Users can continue exploring by directly manipulating facets (by clicking on the tick boxes next to a facet item) and thereby making the search narrower or updating the original search form. We received direct feedback from users that having flexible exploratory capabilities did much to foster trust in DIG, and the underlying AI components. DIG does not require installation as it is hosted in the cloud and is accessible (using a password) through a browser.

NIST Evaluations

Given the unusual nature of investigative search in illicit domains, we believe the evaluation and user study protocol decided upon by DARPA and NIST is an important aspect of this case study.

A total of 16 questions (6 lead generation, 6 lead investigation and 4 operationally relevant i.e. generated in conjunction with the office of a state District Attorney based on its actual investigative needs during that time) were used to evaluate the assistive potential of DIG. A second system, TellFinder, which has a similar philosophy to DIG and has also been independently funded and developed under MEMEX, was also evaluated. These questions (with annotated answers) were independently derived by subject matter experts (SMEs) from a private research organization that had no part in building DIG, and that has been actively involved in training investigators in using search systems developed under MEMEX.

Facets on the side (in underlaid figure, facets for the attributes *Price of Provider* and *Website* are shown) summarize the distribution of values for hits and allow the user to refine search further by checking the tickbox(es) next to an item. In this case, we can see that the price of *60 $ per hour*, and the website *eroticmugshots* are overrepresented among the user's hits. All identifying information has been deliberately obfuscated.



Search Terms

Page: unrushed ✕ City: chicago ✕ Services Provided: fetish friendly ✕ Ethnicity of Provider: hispanic ✕

Price of Provider (Top 10)
View More Sort By:
Select All 10 Deselect All

- □ 60 $ per hour 260
- □ 200 $ per hour
- □ 180 $ per hour
- □ 150 $ per hour
- □ 150 per hh
- □ 80 per qv
- □ 160 $ per hour
- □ 120 per hh
- □ 130 per hh
- □ 60 per qv

Website (8)
View More
Select All 8

- □ eroticmugshots.c
- □ backpage.com
- □ escortsincollege.
- □ escortphonelist.c
- □ classivox.com
- □ adultsearch.com
- □ escortsinthe.us
- □ massagetroll.cor

High Risk
None

25 of 46,190,422 Results How are search results found? Export

2.06 100hh Special ?H?t C??a? ? P?a???? ? ?? P?s? 100% W??t? T?? V?s?t - Chicago escorts - backpage.com

Oct 6, 2015 backpage.com
Locations: chicago
Telephone Numbers: 323[redacted]
No Email Addresses
No Social Media IDs
No Review IDs
Provider Names:

Phone: 323-[redacted]
435 Total Ads

7 Other Telephone Numbers Copy

PHONE	ADS CO-OCCURRING WITH 323-800-6151	ADS NOT CO-OCCURRING
□ 971-[redacted]	20	491
□ 720-[redacted]	17	620
□ 209-[redacted]	11	920
□ 928-[redacted]	11	551
□ 916-[redacted]	9	132
□ 510-[redacted]	7	139
□ 720-[redacted]	3	1,397

What does this chart represent?

2 Email Addresses Copy

EMAIL	ADS CO-OCCURRING WITH 323-800-6151	ADS NOT CO-OCCURRING
□ [redacted]es@gmail.com	12	290
□ [redacted]words@gmail.com	1	69

What does this chart represent?

8 Websites Copy

WEBSITE	ADS
□ liveescortreviews.com	220
□ backpage.com	120
□ eroticmugshots.com	66
□ escortads.xxx	7
□ escortphonelist.com	7
□ escortsincollege.com	7
□ escortsinthe.us	7
□ classivox.com	1

Click and drag on the chart below to zoom. Click elsewhere to reset the zoo

Location Drops Timeline

25 of 435 Ads

100 special ? S??sat???a? C??a? ? B?? B???s ?P?????t C????s ? 2Girls Avail 100% W??t? T?? V?s?t - Indianapolis escorts - backpage.com

Oct 18, 2015 backpage.com
Locations: indianapolis
No Telephone Numbers
No Email Addresses
No Social Media IDs
No Review IDs
Provider Names: raina

100 special ? S??sat???a? C??a? ? B?? B???s ?P?????t C????s ? 2Girls Avail 100% W??t? T?? V?s?t - Indiana escorts - backpage.com

INSTRUCTIONS Enter words or phrases as **search terms** in one or more of the text fields below. Click the **Search** button or press the **ENTER** key when finished. Use **commas** to separate individual search terms. All punctuation **except commas and periods** will be removed. Example: **New York, Washington**

DATES	Click to set start date	Click to set end date	
IDENTIFIERS	Telephone Number	Email Address	Social Media ID
	Review ID		
LOCATION	City	State/Region	
PROVIDER	Name of Provider	Age of Provider	Ethnicity of Provider
	Eye Color of Provider	Hair Color of Provider	Height of Provider
	Weight of Provider	Services Provided	Price of Provider
AD	Website	Title	Page

Figure 1. The main search page in DIG (underlaid) and an *entity* page (overlaid) describing a complete profile of a phone number, including co-occurrence with other phone numbers and email addresses, and the timestamps associated with ads from which that phone was extracted. Ads from which that phone were extracted are also suggested in the lower right pane. The search form is illustrated on the right.

At the time of evaluation, DIG had indexed more than one hundred million sex ads collected over a period of more than two years through focused crawling systems also developed under MEMEX (by different teams). The purpose of the evaluation was not to collect a detailed set of quantitative results that could be put through statistical significance testing, but to understand the nature of investigative search itself, and the role that assistive systems like DIG and TellFinder could play, both through their exploratory UX facilities as well as

Accuracy of answers provided to each question was evaluated by NIST on a scale of 0-3 per question using the following guidelines:

0: No answer given

1: Answers had no overlap with SME answers

2: Answers had some (but not significant) overlap with SME answers

3: Answers had significant overlap with SME answers

the advanced AIs that are used to populate the knowledge base at scale. Fourteen users from multiple state and federal agencies volunteered their time to participate in the study[3]. The user study itself was delegated to NIST by DARPA, with the stated evaluation goals being *usefulness, usability* and *accuracy* (compared to the SME-annotated answers), along with subsequent data collection and analysis.

The developers of DIG and TellFinder each had about an hour to brief the domain experts in small batches, illustrate key uses of the tool, and answer any questions they had. Following this training phase, each user was given the option of working either remotely or onsite. A facilitator from NIST took notes in the background without assisting or interfering in any way. A user was given a limit of 30 minutes to generate answers for each of the 4 operationally relevant questions, and 15 minutes for each other question. Except the facilitator, no one was present in the room during tool use. The order of assigned questions per user was random, and each question was exposed in turn i.e. right before the search session for that question commenced. Posthoc, the accuracy of users' answers was scored on a scale of 0-3, with the scale described in the sidebar. We do not comment on other metrics in this case study, but the operational impact of DIG and TellFinder (subsequently described) indicates significant uptake of both systems in the real world.

Findings

Accuracy results on DIG and TellFinder, using the scale in the sidebar, are illustrated in Figure 2. The results indicate that, with few exceptions, users could find answers that overlapped to some degree with the SME answers. However, the high values (for both DIG and TellFinder) for the value of 1 on the scale also indicates an interesting problem with conducting accuracy evaluations in a dynamic Web domain. There is good *reason* to believe, based on conversations that took place after the fact between NIST, the users and the MEMEX participants, that many of the answers found by users that were non-overlapping may have been relevant to answering the question. This raises a troublesome question about evaluating accuracy. One solution is to ask the users to provide relevance or 'satisfaction' scores for their own answers, but this would be subject to other kinds of biases.

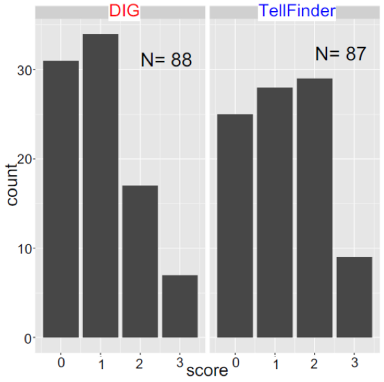


Figure 2. Accuracy results of DIG and TellFinder on 88 and 87 user-question data points respectively, using the scale described in the sidebar.

A second bias that may have crept into the evaluation, but was recognized after the fact as well, was that the SMEs who generated many of the evaluation questions

[3] Questions contain identifying information, or the affiliations of the users, cannot be presently exposed.

were *not* field operators, despite their otherwise broad knowledge about the domain. In contrast, many of the investigative users were exposed in their professional lives to law enforcement, judges, district attorneys, and both sex trafficking victims and perpetrators, but were not necessarily academic students of the subject matter. While both kinds of experts are domain experts, the nature of the expertise is varied enough that it may have caused the evaluations to be biased in favor of how the academic SMEs perceive the domain.

The operational impact of both DIG and TellFinder has been widespread. Both are currently being used, often in complementary ways, by more than 200 law enforcement agencies in the US, specifically to combat sex trafficking. Outputs (particularly cached webpages, later taken offline) from these tools have led to evidence formally presented in court to arraign a sex trafficker. As the MEMEX program is approaching the end, both tools are currently in the process of being transitioned permanently to the office of the District Attorney of New York.

Concluding Notes

Illicit domains have a significant presence on the Web, and domain experts looking to investigate leads and profile activity in such domains have specific technological needs that current Internet-based search technology is not designed to address. Using online sex trafficking as a case study, we described the nature of an illicit domain, the kinds of needs users typically have, and the experiences and insights we gained from building and evaluating a real-world system, DIG, for assisting field investigators across the US. We hope that this case study provides a framework for the HCI community to study the role and impact of interactive technology-assisted investigative search in illicit domains. While it remains an open question whether the findings in this case study extend in a reasonable way to other illicit domains, we believe that this question is worth tackling, especially given the evidence we have been receiving on the operational impact such systems can have on resource-strapped state departments looking to significantly curtail illicit activity.

Extensions and Future Steps

DIG is currently being extended to handle arbitrary domains, and has been evaluated on a variety of other illicit domains. Additionally, the technology in DIG is constantly being updated and refined using state-of-the-art Artificial Intelligence tools.

Acknowledgements

We gratefully acknowledge our collaborators and all (former and current) members of our team who contributed their efforts to DIG, particularly during the evaluation runs. We especially thank Amandeep Singh, Majid Ghasemi Gol, Brian Amanatullah, and Steve Minton. This research is supported by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under contract number FA8750-14-C-0240. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, AFRL, or the U.S. Government.

References

1. Marilyn Jager Adams, Yvette J Tenney, and Richard W Pew. 1995. Situation awareness and the

cognitive management of complex systems. Human factors 37, 1 (1995), 85–104.

2. Charu C Aggarwal and ChengXiang Zhai. 2012. Mining text data. Springer Science & Business Media.
3. Chia-Hui Chang, Mohammed Kayed, Moheb R Girgis, and Khaled F Shaalan. 2006. A survey of web information extraction systems. IEEE transactions on knowledge and data engineering 18, 10 (2006), 1411–1428.
4. Surajit Chaudhuri, Umeshwar Dayal, and Vivek Narasayya. 2011. An overview of business intelligence technology. Commun. ACM 54, 8 (2011), 88–98.
5. Hsinchun Chen, Roger HL Chiang, and Veda C Storey. 2012. Business intelligence and analytics: From big data to big impact. MIS quarterly 36, 4 (2012).
6. Scott M Diamond and Marion G Ceruti. 2007. Application of wireless sensor network to military information integration. In Industrial Informatics, 2007 5th IEEE International Conference on, Vol. 1. IEEE, 317–322.
7. DIG. Domain-specific Insight Graphs (DIG). http://usc-isi-i2.github.io/dig/. (2017). Accessed: 2017-10-10.
8. ELI5. Explain it Like I'm 5. https://pypi.python.org/pypi/eli5. (2017). Accessed: 2017-10-10.
9. Mica R Endsley. 2016. Designing for situation awareness: An approach to user-centered design. CRC press.
10. Qunying Huang and Yu Xiao. 2015. Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery. ISPRS International Journal of Geo-Information 4, 3 (2015), 1549–1568.
11. Kyle Hundman, Thamme Gowda, Mayank Kejriwal, and Benedikt Boecking. 2017. Always Lurking: Understanding and Mitigating Bias in Online Human Trafficking Detection. arXiv preprint arXiv:1712.00846.
12. Rahul Kapoor, Mayank Kejriwal and Pedro Szekely. 2017. Using contexts and constraints for improved geotagging of human trafficking webpages. Proceedings of the Fourth International ACM Workshop on Managing and Mining Enriched Geo-Spatial Data (p. 3). ACM.
13. Mayank Kejriwal and Pedro Szekely. 2017. Information extraction in illicit web domains. In Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 997–1006.
14. Mayank Kejriwal and Pedro Szekely. 2017. Knowledge graphs for social good: an entity-centric search engine for the human trafficking domain. IEEE Transactions on Big Data.
15. Shamanth Kumar, Geoffrey Barbier, Mohammad Ali Abbasi, and Huan Liu. 2011. TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief. In ICWSM.
16. MEMEX. DARPA MEMEX. https://www.darpa.mil/program/memex. (2017). Accessed: 2017-10-10.
17. Palantir. Palantir. https://www.palantir.com/. (2017). Accessed: 2017-10-10.
18. Pedro Szekely, Craig A Knoblock, Jason Slepicka, Andrew Philpot, Amandeep Singh, Chengye Yin, Dipsy Kapoor, Prem Natarajan, Daniel Marcu, Kevin Knight, and others. 2015. Building and using a knowledge graph to combat human trafficking. In International Semantic Web Conference. Springer, 205–221.
19. AI and Society. AAAI/ACM Conference on AI, Ethics and Society. http://www.aies-conference.com/. (2017). Accessed: 2017-10-10.