A Semantic Search Engine For Investigating Human Trafficking

Mayank Kejriwal*, Thomas Schellenberg** and Pedro Szekely*

*Information Sciences Institute, University of Southern California
**Next Century Corporation
{kejriwal,szekely}@isi.edu,thomas.schellenberg@nextcentury.com

Abstract. Enabling intelligent search systems that can navigate and facet on entities, classes and relationships, rather than plain text, to answer questions in complex domains is a longstanding aspect of the Semantic Web vision. In this demo, we present an investigative search engine that meets some of these challenges, at scale, for a variety of complex queries in the human trafficking domain. The search engine has been rigorously prototyped as part of the DARPA MEMEX program and has been integrated into the latest version of the Domain-specific Insight Graph (DIG) architecture, currently used by hundreds of US law enforcement agencies for investigating human trafficking. Over a hundred million ads have been indexed. We demonstrate the in-use version of DIG[1], allowing a user to experience the system in real time.

Keywords: Knowledge graphs, Investigative search, Human trafficking, Illicit domains, Knowledge graph construction

Recent studies confirm a formidable reach of illicit players both online and offline. For example, data from the National Human Trafficking Resource Center shows that human trafficking (HT) is not only on the rise in the United States, but is a problem of international proportions [2]. The advent of the Web has made the problem worse [1]. Human trafficking victims are advertised both on the Open and Dark Web, with estimates of the number of (not necessarily unique) published advertisements being hundreds of millions [3].

In recent years, various agencies in the US have turned to technology to assist them in combating this problem through the suggestion of leads, evidence and HT indicators. An important goal is to answer, in *real time*, *entity-centric questions* over noisy Web corpora crawled from a subset of Web domains known for HT-related activity. Entities are typically service providers, such as *escorts*, but could also be latent entities such as *vendors*, who organize the activity.

There are many challenges, both technological and computational, in building and maintaining such a system. First, answering entity-centric questions requires performing robust *information extraction* (IE) over a crawled Web corpora. IE

[1] The prototype described in the in-use paper titled 'An investigative search engine for the human trafficking domain.'

is not a solved problem in either the NLP or Semantic Web community; even on traditional datasets, consistently achieving F-scores above 70% for common attributes like name and location remains challenging. The second challenge is that of offering an intuitive interface to law enforcement so that they are both comfortable using the system, and are also able to make full use of the facilities that such a system offers. Investigative officials, for example, are not adept even at simple query formulation in languages like SPARQL or SQL; even worse, human users are sometimes known to pose a query that is slightly different from what they *intend*. Finally, investigative officials need to know that they can *trust* the system given the sensitive and resource-constrained nature of the human trafficking domain.

In our group, we have developed the Domain-specific Insight Graph (DIG) system for addressing these challenges by semi-automatically constructing and completing, in an offline phase, a *knowledge graph* containing entities, attributes, relationships and clusters over hundreds of millions of otherwise unconnected webpages. During the interactive search phase, DIG uses powerful query reformulation and ranking strategies to handle the noise in the knowledge graph and answer complex queries that are posed by the user by filling out an intuitive form on the GUI. The search engine in DIG has been rigorously evaluated against a set of competitive baselines from academia and industry by the DARPA MEMEX program, and has proven to be both competitive and efficient.

In this demo, we present an in-use prototype of DIG that is already being used by over 200 law enforcement agencies, and is currently being permanently transitioned to the office of the District Attorney of New York. The prototype can answer complex semantic queries over at least a hundred million HT webpages that have crawled and indexed over the last 2-3 years. DIG supports both faceted and entity-centric search, as Figures 1 and 2 illustrate.

Specifically, we will showcase how DIG can be used to jump-start an investigation starting from a vague, under-specified search query. For example, a user searches for a hispanic escort in Chicago who is known to provide certain sex services in her ads. By exploring a ranked list of ads retrieved by the system, along with images, the user is able to drill down, in an entity-centric fashion, on important details like phone numbers and emails that are promising avenues for field investigations. Acquiring such a list without the help of the system would ordinarily have taken months of field level investigations and online searches.

Conclusion. Human trafficking is an egregious crime that has been helped, rather than hindered, by the growth of the Web. The DIG search engine attempts to use technologies from various communities, and especially the Semantic Web, in the hopes of significantly raising the barrier-of-entry for would-be traffickers, by providing law enforcement with state-of-the-art tools for expediting prosecution and evidence gathering. DIG, in tandem with other tools developed under the DARPA MEMEX program, has already been used to bring several prosecutions to court in the US[2] and is in the process of being permanently transitioned

[2] A potent example is the recent case described in `http://www.sfgate.com/crime/article/Man-sentenced-to-97-years-in-human-trafficking-7294727.php`

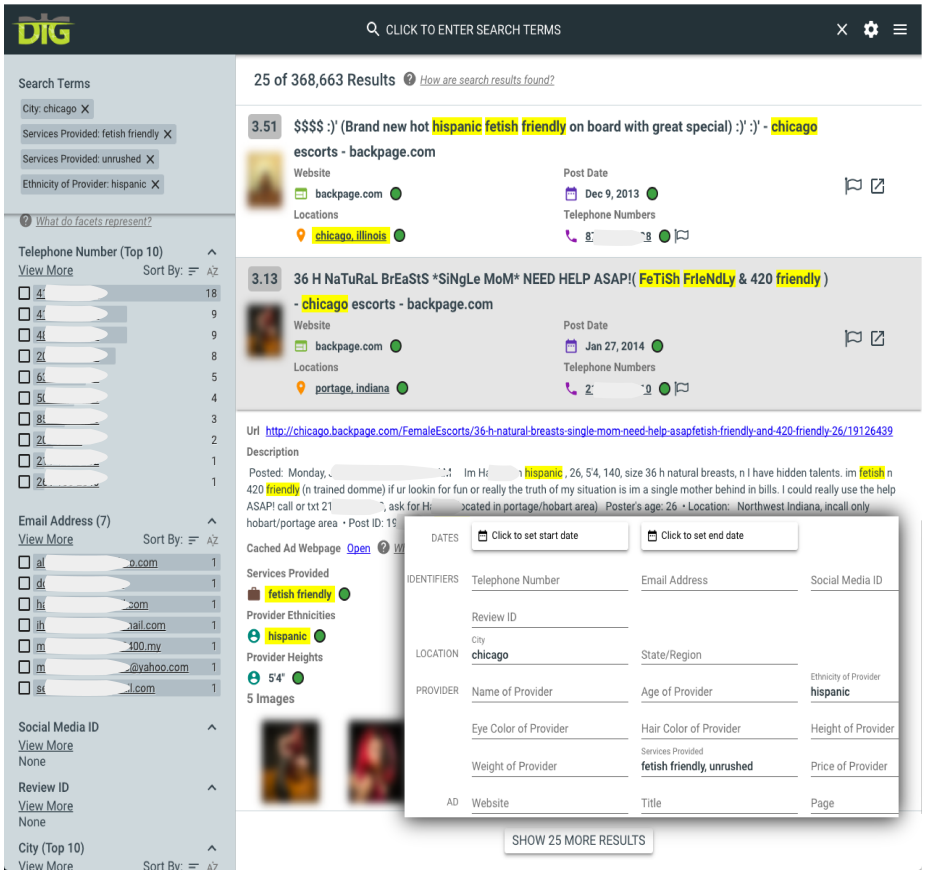

Fig. 1. A case-study illustrating complex investigative search in DIG. Starting from vague details, a user is able to retrieve a set of highly specific ads that help her to narrow down on suspects.

to law enforcement. We will showcase the in-use version of DIG in this demo, with the ultimate hope that our system will serve as a case study of the real-world social impact that Semantic Web research can have.

Acknowledgements. We gratefully acknowledge our collaborators and all (former and current) members of our team who contributed their efforts and expertise to DIG, particularly during the dry and final evaluation runs: Amandeep Singh, Linhong Zhu, Lingzhe Teng, Nimesh Jain, Rahul Kapoor, Muthu Rajendran R. Gurumoorthy, Sanjay Singh, Majid Ghasemi Gol, Brian Amanatullah, Craig Knoblock 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

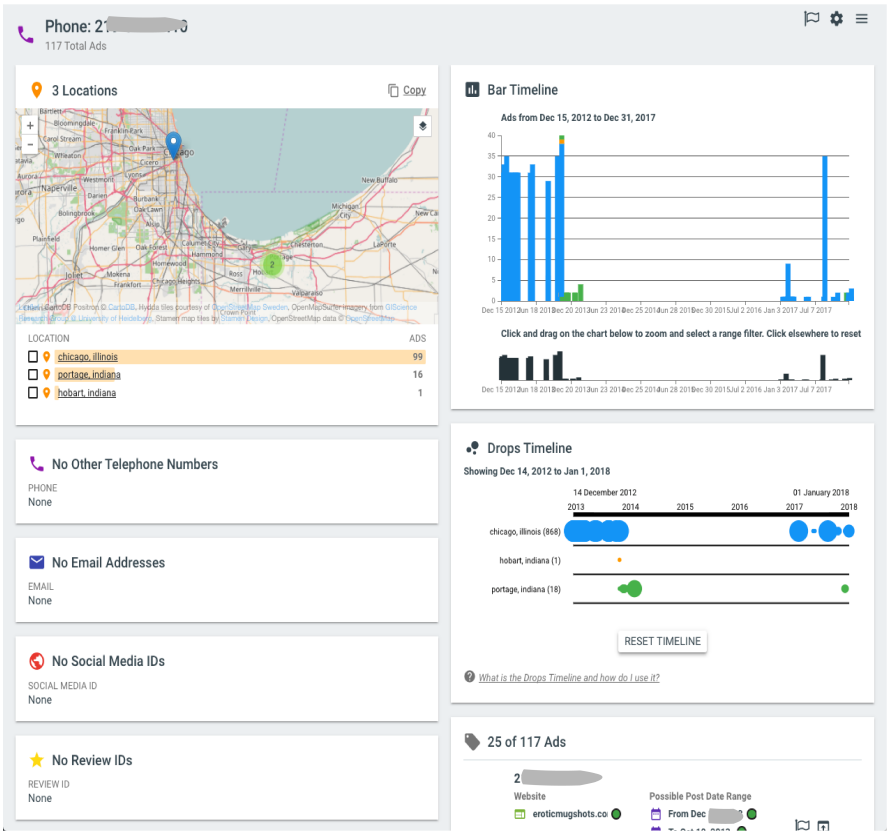

Fig. 2. Continuing from Figure 1, the user is able to use entity-centric search facilities in DIG to collect a list of 117 ads, with details such as timelines and locations. These numbers can be used to conduct an investigation on the ground.

as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, AFRL, or the U.S. Government.

References

1. V. Greiman and C. Bain. The emergence of cyber activity as a gateway to human trafficking. In *Proceedings of the 8th International Conference on Information Warfare and Security: ICIW 2013*, page 90. Academic Conferences Limited, 2013.
2. S. Harrendorf, M. Heiskanen, and S. Malby. *International statistics on crime and justice*. European Institute for Crime Prevention and Control, affiliated with the United Nations (HEUNI, 2010.
3. P. Szekely, C. A. Knoblock, J. Slepicka, A. Philpot, A. Singh, C. Yin, D. Kapoor, P. Natarajan, D. Marcu, K. Knight, et al. Building and using a knowledge graph to combat human trafficking. In *International Semantic Web Conference*, pages 205–221. Springer, 2015.