Domain-specific Insight Graphs (DIG)

Pedro Szekely
USC Information Sciences Institute
Marina Del Rey, California
pszekely@isi.edu

Mayank Kejriwal
USC Information Sciences Institute
Marina Del Rey, California
kejriwal@isi.edu

ABSTRACT

The DARPA Memex program was established with the goal of funding research into building domain-specific search systems that integrated state-of-the-art focused crawling ('domain discovery') information extraction and semantic search, and that could be used by users and domain experts with no programming or technical experience. Domain-specific Insight Graphs (DIG) was proposed and funded under Memex and has led to an end-to-end search system currently being used by over 200 law enforcement for combating human trafficking, by investigators from the Securities and Exchange Commission (SEC) in the US for investigating securities fraud, and for numerous other domains of a difficult, socially consequential (e.g., investigative) and unusual nature.

CCS CONCEPTS

• **Information systems → World Wide Web**; **Users and interactive retrieval**; **Environment-specific retrieval**; *Information systems applications*;

KEYWORDS

Domain-specific search, knowledge graphs, human trafficking, investigative search, dynamic information retrieval

ACM Reference Format:
Pedro Szekely and Mayank Kejriwal. 2018. Domain-specific Insight Graphs (DIG). In *Proceedings of The Web Conference 2018 (WWW 2018).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3184558.3185983

1 INTRODUCTION

Web search today uses a centralized approach that searches the Internet with a one-size-fits-all set of tools for all queries. Despite its extensive commercial success, it does not work well for many use cases, especially in government, defense and also social systems. For example, Web search today still remains a largely manual process that does not save sessions, requires nearly exact input with one-at-a-time entry, and doesn't organize or aggregate results beyond a list of links. Moreover, common search practices miss information in the Deep Web (the parts of the web not indexed by standard commercial search engines), and ignore shared content across pages.

The DARPA Memex program was launched with the goal of advancing online search capabilities far beyond the current state of the art. The goal is to invent better methods for interacting with and sharing information, so users can quickly and thoroughly organize and search subsets of information relevant to their individual interests. The technologies developed in the program would provide the mechanisms for improved content discovery, information extraction, information retrieval, user collaboration and other key search functions.

Table 1: DIG project-level details, including funding and collaborations.

Principal Investigator	Pedro Szekely (USC Information Sciences Institute)
Collaborating Organizations	Columbia University, Inferlink Corp., Next Century Corp.
Duration	3 years (2014-2017)
Project Volume	8.4 million USD
Funding Agency	US Defense Advanced Research Projects Agency (DARPA) under the Memex program and Air Force Research Laboratory
Official Website	http://dig.isi.edu

One of the projects funded under the Memex program is the Domain-specific Insight Graph (DIG) project, led by the USC Information Sciences Institute. To address the challenges posed by Memex, DIG uses a *knowledge graph*-centric approach. In recent years, knowledge graphs have emerged as powerful platforms both in search and the general Artificial Intelligence community (especially, Semantic Web and Natural Language Processing). DIG advances the state-of-the-art both in using and in representing knowledge graphs for search and analytics. Over three years of research, DIG has emerged as a single extensible ecosystem that addresses many of the challenges of Memex, and is now widely in use by many real-world agencies for fighting problems such as human trafficking with the use of technology (Section 2). DIG can be set up and used by people with no programming abilities, an important strength in an era when systems continue to become ever more complex and opaque. DIG is available on GitHub as an open-source project[1] under a permissive MIT license.

Some other project-level details are listed in Table 1. Notably, the project involved both academic and industrial partners, and in addition to the formal collaborations listed in the table, informal collaborations with other teams funded under Memex were also routinely undertaken over the life of the program. For example, the Ache crawler, developed by a research team from New York University, is included in the DIG ecosystem for those who want to truly start from scratch i.e. data collection.

WWW 2018, April 23–27, 2018, Lyons, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3185983

[1] https://github.com/usc-isi-i2/dig-etl-engine

2 SIGNIFICANCE AND IMPACT

Research funded under Memex, including not only DIG but outputs by other performers, has resulted in democratization of search technology, since it allows users with no programming experience to build domain-specific search engines that would normally take significant technical expertise. Users can set up their own information extractors, search engines, focused crawlers, and even customized UIs, all under a unified framework. The software is open-source and has been used to set up the first search engine that is being actively used by law enforcement and other resource-strapped agencies whose funding is constantly under pressure. In the human trafficking domain, Memex tools have resulted in recent prosecutions in the US and are being transitioned permanently to state district attorneys. DIG has also yielded significant research output over the last three years, in addition to research outreach in the form of tutorials, talks and demonstrations at both top-tier academic conferences and industrial venues.

2.1 Research Outputs

DIG has yielded over 15 peer-reviewed publications over the course of three years, some examples being [9], [10], [6], [8], [5], [2], [4], [3], [1] with several more under review, and at least one dataset resource [7]. Two papers have won Best Paper awards [9], [10]. More broadly, our experience has spurred us to write a graduate-level textbook on knowledge graphs, which will be published by MIT Press later this year. DIG has supported the hiring and mentoring of many Master's students, and is largely responsible for the support of at least two Ph.D. students. Moreover, research conducted under DIG is having impact beyond computer science as well. For example, in collaboration with social scientists, we are constructing and visualizing a social network of sex workers in the United States using backpage.com webpages, in order to better understand the domain.

2.2 Outreach Outputs

In addition to research papers and systems, we have also organized workshops[2], and conducted several interdisciplinary tutorials[3] on search and knowledge graphs, including at top-tier conferences such as the International Semantic Web Conference or ISWC (2017), the ACM KDD Conference (2017), the Web Conference (2018), and the AAAI conference (2018). We have also demonstrated the DIG system, both at ISWC 2017 and, more recently, at AAAI 2018, where the system was nominated for a Best Demonstration award. We presented the problem of investigative search as a case study and video at CHI[4] 2018. Finally, we presented DIG as an AI for social good system at industrial venues, including Data Day Texas[5], a 700+ person industrial event held in Austin, Texas in January, 2018.

3 TIMELINE

The DIG graphical user interface (GUI) and early information extraction modules in DIG were already constructed within a few months of program start in 2014, in collaboration with our user interface partner (Next Century Corporation). The second year (2015-2016) was spent on building an advanced search engine for the noisy knowledge graphs ingested by DIG. Throughout this period, the primary domain that DIG was being tested on was the human trafficking domain, due to both the technical difficulties entailed by the domain, as well as its potential for real-world impact. In the final year (2016-2017), the different strands of research in DIG were packaged into a final system called myDIG that allows users to construct domain-specific knowledge graphs in arbitrary domains. myDIG was evaluated by investigative users in multiple domains throughout 2017, and a final 'packathon' testing the capabilities of the system recently concluded, with successful results, in the first week of November 2017 (when the Memex program officially concluded).

4 CONCLUSION AND FUTURE WORK

The Domain-specific Insight Graph (DIG) project was funded under the DARPA Memex program to address important challenges in domain-specific search. DIG successfully showed that good use of Web technology, especially knowledge graph construction from Web data, can be used to address many of the challenges. Recently, DIG has also been used for social science, and is continuing to be expanded in scope and capability through support from other projects. Some use-cases that DIG is now being used to support include causal exploration, search and inference involving time-series, research support for geopolitical forecasting, and multi-modal knowledge graphs. We are continuing to maintain the project, and are committed to keeping it free, open-source and easy to use.

REFERENCES

[1] 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* (2017).

[2] Rahul Kapoor, Mayank Kejriwal, and Pedro Szekely. 2017. Using contexts and constraints for improved geotagging of human trafficking webpages. In *Proceedings of the Fourth International ACM Workshop on Managing and Mining Enriched Geo-Spatial Data*. ACM, 3.

[3] Mayank Kejriwal, Jiayuan Ding, Runqi Shao, Anoop Kumar, and Pedro Szekely. 2017. FlagIt: A System for Minimally Supervised Human Trafficking Indicator Mining. *arXiv preprint arXiv:1712.03086* (2017).

[4] 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.

[5] Mayank Kejriwal and Pedro Szekely. 2017. An Investigative Search Engine for the Human Trafficking Domain. In *International Semantic Web Conference*. Springer, 247–262.

[6] 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* (2017).

[7] Mayank Kejriwal and Pedro Szekely. 2017. Neural embeddings for populated geonames locations. In *International Semantic Web Conference*. Springer, 139–146.

[8] Mayank Kejriwal, Pedro Szekely, and Craig Knoblock. 2018. Investigative Knowledge Discovery for Combating Illicit Activities. *IEEE Intelligent Systems* (2018).

[9] 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.

[10] Linhong Zhu, Majid Ghasemi-Gol, Pedro Szekely, Aram Galstyan, and Craig A Knoblock. 2016. Unsupervised entity resolution on multi-type graphs. In *International Semantic Web Conference*. Springer, 649–667.

[2] http://usc-isi-i2.github.io/home/#workshops
[3] http://usc-isi-i2.github.io/home/#tutorials
[4] https://chi2018.acm.org/authors/case-studies/
[5] http://datadaytexas.com/2018/sessions#kejriwal