Building Domain-Specific Search Engines

Subjects: Computer Science, Artificial Intelligence

Contributor: Mayank Kejriwal

With advances in machine learning, knowledge discovery systems have become very complicated to set up, requiring extensive tuning and programming effort. Democratizing such technology so that non-technical domain experts can avail themselves of these advances in an interactive and personalized way is an important problem. myDIG is a highly modular, open source pipeline-construction system that is specifically geared towards investigative users (e.g., law enforcement) with no programming abilities. The myDIG system allows users both to build a knowledge graph of entities, relationships, and attributes for illicit domains from a raw HTML corpus and also to set up a personalized search interface for analyzing the structured knowledge. Both qualitative and quantitative data from five case studies involving investigative experts from illicit domains, such as securities fraud and illegal firearms sales, have been used to illustrate the potential of myDIG.

Knowledge graphs illicit domains human trafficking securities fraud search engines non-technical users

1. Introduction

As research in machine learning and AI continues to progress, front-facing systems have become steadily more complicated, expensive, and limited to specific applications like Business Intelligence[1] (BI). Democratizing technology is a complex issue that involves multiple stakeholders with different sets of abilities. A particular user whose needs are very real, but which are seldom addressed except in BI or military-specific situational awareness situations, is a domain expert with extremely limited technical abilities. In particular, such users do not know how to program, let alone cope with complex machine learning or deep learning algorithms[2], and cannot satisfy their information needs through a simple Google search for several subsequently-described reasons.

A real-world example of such a domain expert that we encountered in the securities fraud domain is an employee from the Securities and Exchange Commission (SEC) who is attempting to identify actionable cases of penny stock fraud. Penny stock offerings in Over-The-Counter (OTC) markets are frequently suspected of being fraudulent, but without specific evidence, usually in the form of a false factual claim (that is admissible as evidence), and their trading cannot be halted. With thousands of penny stock offerings, investigators do not have the resources or time to investigate all of them. One (technical) way to address this problem is to first crawl a corpus of relevant pages from the web describing the domain, a process alternately known in the Information Retrieval (IR) literature as relevance modeling or domain discovery[3]. The latter term is more encompassing, as it involves not just relevance modeling, but the actual crawling of the data.

Once such a corpus is obtained, an expert in information extraction and machine learning would elicit opinions from the users on what fields (e.g., location, company, stock ticker symbol) are important to the user for answering domain-specific questions, along with example extractions per field. This sequence of *Knowledge Graph Construction (KGC)* steps results in a graph-theoretic representation of the data (a Knowledge Graph or KG) where nodes are entities and field values and the (directed, labeled) edges are relationships between entities or assignments of field values to entities. Since the KG is structured, it is amenable to aggregations and to both keyword and structured querying. With a good interface, for example, the domain expert can identify all persons and organizations (usually shell companies) associated with a stock ticker symbol, aggregate prices, or zero in on suspicious activity by searching for hyped-up phrases that indicate fraud.

A Google search is inadequate for the knowledge discovery described above as it does not allow one to specify fields nor to limit the search to a specific corpus. Google also does not offer the specialized aggregation facilities that such domain experts would need. While Google works exceedingly well for document-level information indexing and retrieval, it is not designed to support complex analytical tasks, either with or without training. What is needed is a way of *building domain-specific searching engines* at scale, without enormous technical expertise.

2. Influences

The idea of building domain-specific search engines goes back to a 1945 article written by Vannevar Bush, then Director of the Office of Scientific Research and Development. Bush mentions a futuristic device called Memex that "stores all his books, records, and communications, and which is mechanized" to have speed and flexibility. This device would be so good that it would be like an "enlarged intimate supplement to his memory.” It is important to emphasize that Bush wrote this article in a pre-Internet, pre-Web and pre-smartphone era. We argue that a modern interpretation of Memex is a domain-specific search engine. The Defense Advanced Research Projects Agency (DARPA), which instituted the Memex program a few years ago under which our research was funded for this project, had a similar view. Here, we interpret the "memory" that Bush referred to as the body of subject matter expertise (for example, the body of knowledge that an official from the SEC has acquired over many years of enforcing security laws) and the device is like a search engine specifically designed to help the expert not only retrieve information using a rich set of modalities (such as queries, recommendations, summaries, questions, some of which may not be very well-defined or specified) but to visualize and use that information in a number of different ways.

Another important influence has been the development of knowledge graph technology, especially in the realm of IR. The Google Knowledge Graph, introduced in a blog post back in 2011 as "things, not strings" is a good example. While this knowledge graph was generic and designed for general users, it opened the door for building and using other knowledge graphs for domain-specific IR. Therefore, it significantly influenced the development of our own system, at least in principle.

3. Milestones

In our group, we have developed a system called myDIG[4] (my Domain-specific Insight Graphs) that ingests the output of domain discovery, i.e., a raw corpus of web pages, and presents an intuitive multi-step KGC and search interface to a user that requires no programming, can be refined iteratively, and requires an hour or less of example-based training. An implemented prototype of myDIG has already undergone evaluations by domain experts in five different investigative domains, each of which involves significant technological potential. myDIG's dataflow is modular, since it allows users to construct and search knowledge graphs, and thereby discover new and relevant knowledge by conducting data-driven analysis.

We presented qualitative and quantitative data collected from the five case studies to illustrate the real-world potential of myDIG in making an advanced set of knowledge discovery technologies accessible to non-technical users. Using myDIG, practicing, non-technical domain experts from the five case study domains were able to build a personalized domain-specific search engine over corpora containing more than a million raw web pages in 4--6 working hours. Most notably, myDIG was applied to the domain of human trafficking[2], with promising results.

4. Current Status

The myDIG project is open-source, with a dedicated web page (http://usc-isi-i2.github.io/dig/) and a GitHub (https://github.com/usc-isi-i2/dig-etl-engine) repository associated with it. More details on the project can be found in the open article 'myDIG: Personalized Illicit Domain-Specific Knowledge Discovery with No Programming' published in Future Internet MDPI, 2019. We expect to continue maintaining it and using it for a variety of use cases, including building KGs and search engines for COVID-19, social media and other domains.

References

1. Edge, D., Larson, J., & White, C. (2018, April). Bringing AI to BI: enabling visual analytics of unstructured data in a modern business intelligence platform. In Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-9).
2. Kejriwal, M., & Szekely, P. (2018, April). Technology-assisted investigative search: A case study from an illicit domain. In Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-9).
3. Krishnamurthy, Y., Pham, K., Santos, A., & Freire, J. (2016). Interactive exploration for domain discovery on the web. Proc. of KDD IDEA.
4. Szekely, P., & Kejriwal, M. (2018, April). Domain-specific Insight Graphs (DIG). In Companion Proceedings of the The Web Conference 2018 (pp. 433-434).