The Value of AI Tools: Some Lessons Learned

Steven N. Minton

■ *We are in the midst of an AI spring, and it's an exciting time for the AI community. AI is poised to change the world. Nevertheless, for many new AI technologies, it is still unclear how these technologies will be successfully productized, and which types of companies will be winners and losers. In this column, I reflect on some of the thorny issues that can arise when commercializing AI technology, based on my personal experience developing information-extraction software.*

Artificial intelligence is a hot commodity, and for the first time in decades we are seeing the emergence of companies whose primary business purports to be AI. I'm curious to see whether these companies can sustain themselves in the long run, especially since AI has traditionally delivered most of its value under the hood. Today, AI plays an important role behind the scenes in industries from finance to manufacturing, but it's only recently that there has been enough interest to raise its visibility to the forefront of a company's value proposition.

One area with a history of commercial interest is web-based information extraction. The advent of the World Wide Web in the 1990s created a new set of opportunities for applied AI, and for many people, Google's highly public focus on AI and machine learning helped validate the growing importance of these technologies for business. In fact, there have been many other companies, big and small, that have also worked on technologies for understanding and extracting web data. In this column I will comment on some of the thorny business issues, as well as opportunities, that I have seen in this arena, which continues to be a strong area of interest for both AI researchers as well as practitioners.

Personally, I became involved in research on web data extraction at the University of Southern California's Information Sciences Institute in the mid-1990s, when I worked on a series of research projects for extracting data from semistructured sites, with Craig Knoblock and other colleagues. Late in 1999, at the height of the Internet boom, we launched Fetch Technologies, Inc., to commercialize this work. At the time, companies with less capable technology were selling for many hundreds of millions of dollars, and we were looking forward to becoming very wealthy. Unfortunately, just as we brought our first product to market — a machine-learning tool for extracting data — the Internet bust caught up with us and we ran out of cash.

Rather than shut down the company completely, we kept a few people, and with money from federal research grants, we were able to limp along. After the bust, there wasn't much interest in our initial product, but we were able to continue baking the technology and after several years of hard work produced a very solid product. With just a few examples, the Fetch Agent Platform could be taught how to navigate through a website and extract semistructured data (for example, data on web pages). So, for instance, one could easily show the system how to extract from a telephone directory site, including filling out the relevant forms on the site, and the system would effectively reverse-engineer the site, extracting a database of harvested data. To achieve this, we developed a series of increasingly sophisticated machine-learning methods for data extraction (Gazen and Minton 2005; Knoblock et al. 1998; Lerman, Minton, and Knoblock 2003; Minton, Ticrea, and Beach 2003; Muslean, Minton, and Knoblock 2006) as well as a data flow architecture that could harvest and extract data at scale (Barish et al. 2000).

The Fetch Agent Platform was a very impressive AI system, one that I'm still proud of, but our business never quite hit that inflection point we were looking for. On the plus side, we eventually did have some important successes. For instance, Fetch was used by many background-checking companies to harvest criminal records from county and state court sites; and the company was acquired by Connotate, which still uses the technology today. Nevertheless, Fetch was never the big commercial success we had hoped for.

One reason is that we never found a vertical market that was big enough. Instead, each different business case had different requirements, and often required us to change or customize our sales pitch, our services, and sometimes even our technology, to fit the use case. Some use cases involved sites with complex forms to navigate, others did not. Some use cases involved deep scraping of very structured sites, others involved many unstructured pages on many sites. Some needed images, others needed PDF extraction, and so forth.

As one example, we had a very successful engagement with Dow Jones Factiva, where we helped them harvest "all the news on the planet." This service was able to monitor and extract articles (title, text, authors, publication date, and other information) from tens of thousands of news sites. But it also required us to customize the technology, such as adding semisupervised classifiers to recognize authors. While we did find one or two other customers for this specialized news harvester, the overall effort barely paid for itself.

We ran into other business problems as well. For instance, web data is inherently free, and this often affected the perceived value of the system. Also, it's not too hard for programmers to scrape specific sites, so unless a company wanted to harvest from a large number of sites, they could often do it themselves. In fact, while other companies had competing products, in sales situations our biggest competitor was almost always the customer's internal developers, who typically wanted to do the job themselves.

Fetch was a good example of how hard it is to make money off a horizontal technology that has many different use cases. Indeed, there have been many companies that have similarly developed tools for extracting data from semistructured sources, as well as unstructured text, and none of them have been big winners, to my knowledge. Of course, there are technologies that have succeeded horizontally —— databases being a good example. Perhaps the difference is that every midsize to large business needs a database. But not every company needs a generic semistructured information extractor.

I believe that the easiest way to build a business based on extraction technology is to focus on building a specific data product that has a clear value. For example, HG Data, an early Fetch customer, was a startup that now has a rapidly growing business selling data about the adoption of software and hardware, which is useful for market intelligence. I don't know what technologies HG Data is using today underneath the hood, but I know that their data has clear value. Another company, Cytenna (which I cofounded), aggregates data about cyber vulnerabilities. Cytenna uses AI underneath the hood, but

Cytenna's products are all about the data, not the technology.

I still work on generic tools for information extraction. I think it's an exciting and important research area, which continues to progress. But I also try to maintain realistic expectations about the commercial prospects for these tools.

References

Barish G.; DiPasquo, D.; Knoblock, C. A.; and Minton, S. 2000. Dataflow Plan Execution for Software Agents. In *Proceedings of the Fourth International Conference on Autonomous Agents* (Agents-2000). New York: Association for Computing Machinery. doi.org/10.1145/336595.337087

Gazen, B., and Minton, S. 2005. AutoFeed: An Unsupervised Learning System for Generating Webfeeds. In *Proceedings of the Third International Conference on Knowledge Capture*. New York: Association for Computing Machinery. doi.org/10.1145/1088622.1088625

Knoblock, C. A.; Minton, S.; Ambite, J. L.; Ashish, N.; Modi, P. J.; Muslea, I.; Philpot, A. G.; and Tejada, S. 1998. Modeling Web Sources for Information Integration. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 211–218. Palo Alto, CA: AAAI Press.

Lerman, K.; Minton, S.; and Knoblock, C. A. 2003. Wrapper Maintenance: A Machine Learning Approach. *Journal of Artificial Intelligence Research* 18: 149–181.

Minton S.; Ticrea S.; and Beach J. 2003. Trainability: Developing a Responsive Learning System. Paper presented at the 2003 IJCAI Workshop on Information Integration on the Web, Acapulco, Mexico, August 9–10.

Muslea, I.; Minton, S.; and Knoblock, C. A. 2006. Active Learning with Multiple Views. *Journal of Artificial Intelligence Research* 27.

Steven N. Minton is president of InferLink Corporation, a research and development company that develops technologies for aggregating, integrating, and analyzing information from multiple sources. In addition to InferLink, Minton has helped launch several AI companies, including Fetch Technologies, Cytenna, Innotrieve, GeoSemble, and the nonprofit AI Access Foundation. Minton was also the founder and first executive editor of the *Journal of Artificial Intelligence Research* (JAIR). He is a AAAI fellow and his awards include AAAI's Classic Paper award (2008), the AAAI Robert S. Englemore Memorial Lecture award (2012), and Best Paper at AAAI-88.