Checkpoint A: Topic Choice, Initial Schema Definition, and Likely Data Sources

MSDS 459

Technology Sector Team

Nick Butler, Michael Rivera, and Richard Pereira

April 27, 2025

Northwestern University

ABSTRACT

This report outlines the early stages of a research project aimed at developing a knowledge graph and web-crawling framework to support competitive intelligence in the technology sector. The project focuses specifically on tracking emerging competitors within the field of Generative AI. The management problem that motivates this research is the difficulty companies face in identifying and responding to new entrants due to the fragmented nature of data across structured sources like Crunchbase and PitchBook and unstructured sources such as TechCrunch and Wired. To address this challenge, the team is building a graph-based knowledge base that unifies data from financial, technical, and media domains. Work completed to date includes topic selection, schema design, identification of over 20 reputable data sources, and the setup of a focused crawler for web scraping. This foundational work lays the groundwork for building an application that enables efficient information retrieval, media signal analysis, and strategic decision-making.

1. INTRODUCTION

The pace of innovation in the technology sector makes it increasingly difficult for companies to monitor competitors and respond proactively to new market entrants. This is particularly relevant in the field of Generative AI, where startups can gain traction through product innovation, media exposure, and investor support. The core management question guiding this research is as follows: How can a technology company develop actionable intelligence to monitor and respond to emerging competitors in the field of Generative AI?

To help answer this question, the project proposes the development of a knowledge base and corresponding application that will integrate structured and unstructured data into a unified representation of the competitive landscape. The knowledge base will enable users to monitor signals such as patent filings, product announcements, financial performance, and social media activity. The target users of this knowledge base include corporate strategists, market intelligence analysts, product managers, and innovation leaders. These users require reliable, timely insights that can inform investment decisions, product roadmaps, and competitive positioning. The envisioned application will support query-based exploration of competitor intelligence, trend monitoring, and recommendations for strategic action.

2. LITERATURE REVIEW

This research is grounded in the domains of web data mining and knowledge graph construction. Foundational work by Chakrabarti, van den Berg, and Dom (1999) introduced the concept of focused crawling. Focused crawling is a strategic approach to web content discovery that prioritizes topic-relevant material. Their proposed architecture consists of a crawler, distiller (to elevate high-value pages), and a classifier that evaluates content relevance. Together, those components can efficiently navigate the vast and diverse online information space.

Building on this foundation, Chakrabarti (2003) expands on methods for discovering knowledge from hypertext data, highlighting how semantic relationships and link structures can be harnessed to extract meaning from the web. These methodologies have informed our design of both the crawler and the knowledge base schema. Technical resources such as Mitchell (2018), Hajba (2018), and Patel (2020) provide implementation guidance for practical web scraping using scalable Python frameworks

like Scrapy and BeautifulSoup. These tools are essential for automating the retrieval of relevant content from trusted sources.

In the context of information systems and competitive intelligence, knowledge graphs are increasingly used to synthesize heterogeneous data and provide a platform for semantic querying. They are particularly valuable in dynamic environments such as technology and innovation, where structured databases alone fail to capture the complexity and velocity of change.

4. METHODOLOGY

Topic Selection

The research focuses on the technology sector, with a specific interest in capturing competitive and innovation signals within the quickly-evolving domain of Generative AI. This topic was selected due to the high volume of relevant online information and the rapid pace of change in this area. The goal is to support decision-making by enabling structured insights into emerging companies, product developments, and market dynamics.

Schema Design

The knowledge base will follow a graph-relational structure designed for implementation in EdgeDB or a similar platform. The node types include:

- Company: representing technology firms
- Product: representing offerings developed by those firms
- Patent: representing innovation filings tied to companies or products
- News Article: representing unstructured media content
- Financial Event: representing earnings reports, funding rounds, or stock movements
- Trend Indicator: representing metrics from Google Trends or social media APIs

The edges between these nodes will define relationships such as:

- Company to Product (develops)
- Company to Patent (holds or files)
- Company to News Article (mentioned in)
- Company to Financial Event (experiences)
- Company to Trend Indicator (associated with)
- Product to News Article (covered by)

This schema allows for flexible querying and supports the integration of heterogeneous data into a unified analytical model.

Data Sources and Collection

The team has identified 24 trusted sources (see Appendix A) of structured and unstructured data, categorized as follows:

- Financial and investment data: Yahoo Finance, SEC EDGAR, company investor sites

- Company and startup data: Crunchbase, PitchBook, CB Insights
- News and media: TechCrunch, Wired, The Verge, IEEE Spectrum
- Trends and strategy: McKinsey, Statista, Google Trends, OECD STI Outlook
- Patent and government data: Google Patents, WIPO, data.gov
- Social signals: GitHub Trending, Twitter/X API

Initial data collection is being conducted using Python-based tools such as Scrapy and BeautifulSoup. Extracted content will be stored in JSON Lines format, with each entry including metadata such as source URL, publication date, entity mentions, and text body.

User Needs and Questions

The primary users of this knowledge base will be business professionals who are responsible for strategic foresight. Anticipated user questions include:

- Which startups in the AI space are receiving increasing funding or attention?
- How does sentiment toward Company X compare across different media outlets?
- What new patents have been filed in a particular subdomain?
- Which firms are gaining traction in developer communities?

The knowledge base is designed to support these questions through semantic relationships and time-series tracking of key indicators.

Application Usefulness

The envisioned application will offer a suite of capabilities including information retrieval, information extraction, question answering, and personalized recommendations. Users will be able to search for relevant firms or technologies, extract patterns in funding and media activity, and receive alerts about emerging trends or players. This tool will help organizations respond proactively rather than reactively to competitive shifts.

5. RESULTS

At this stage, the team has made progress in several key areas. A set of 24 high-quality data sources has been curated, including structured financial data, company metadata, media articles, and social indicators. Early crawling tests using Yahoo Finance and TechCrunch have confirmed the feasibility of automated data extraction. These tests have successfully captured article metadata and company performance data that will be used to populate the knowledge base.

The team has also completed a preliminary design of the graph schema. The relationships defined in the schema will allow the system to surface connections between firms, innovations, and market responses. This model supports both exploratory analysis and structured querying. The use of JSON Lines format for storing raw content provides flexibility for downstream processing.

Challenges identified during the initial phase include inconsistencies in data formatting, especially in unstructured sources. Media articles vary widely in structure, and some financial sites limit

scraping through rate-limiting or dynamic content loading. In response, the team is evaluating the use of tools such as Selenium or Puppeteer for sites that require browser-based scraping.

6. CONCLUSIONS

This research seeks to enhance the ability of technology companies to monitor emerging competitors by developing a unified, queryable knowledge base that integrates structured and unstructured data. Progress to date demonstrates the feasibility and value of this approach in addressing the underlying management problem. The design of the graph schema and the identification of trusted data sources provide a solid foundation for building a meaningful and dynamic intelligence platform.

While there challenges remain in automating unstructured data extraction and maintaining consistent data quality, these obstacles can be overcome through continued refinement of crawling tools and schema definitions. Additionally, integrating real-time signals from APIs such as Twitter or GitHub may further enhance the system, especially for detecting early innovation trends.

Ultimately, the collected data will enable users to answer strategic questions that are central to maintaining competitive advantage. By surfacing meaningful connections between companies, products, patents, and public sentiment, the knowledge base will support proactive decision-making in an industry defined by its speed and complexity.

7. REFERENCES

Chakrabarti, Soumen. *Mining the Web: Discovering Knowledge from Hypertext Data*. San Francisco: Morgan Kaufmann, 2003.

Chakrabarti, Soumen, Martin van den Berg, and Byron Dom. “Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery.” *Computer Networks* 31, no. 11–16 (1999): 1623–1640. https://www.cse.iitb.ac.in/~soumen/doc/www1999f/pdf/www1999f.pdf.

Hajba, Gábor László. *Website Scraping with Python: Using Beautiful Soup and Scrapy*. New York: Apress, 2018.

Mitchell, Ryan. *Web Scraping with Python: Collecting More Data from the Modern Web*. 2nd ed. Sebastopol, CA: O’Reilly Media, 2018.

Miller, Thomas W. *Web and Network Data Science: Modeling Techniques in Predictive Analytics*. Upper Saddle River, NJ: Pearson FT Press, 2015.

Nair, Vineeth G. *Getting Started with Beautiful Soup: Build Your Own Web Scraper and Learn All About Web Scraping with Beautiful Soup*. Birmingham, UK: Packt Publishing, 2014.

Olston, Christopher, and Marc Najork. “Web Crawling.” *Foundations and Trends in Information Retrieval* 4, no. 3 (2010): 175–246.

Patel, Jay M. *Getting Structured Data from the Internet: Running Web Crawlers/Scrapers on a Big Data Production Scale*. New York: Apress, 2020.

Smith, Vincent. *Go Web Scraping Quick Start Guide: Implement the Power of Go to Scrape and Crawl Data from the Web*. Birmingham, UK: Packt Publishing, 2019.

APPENDIX A – Data Sources Table

Content	Type	Details	URL
Yahoo Finance	Financial Data	Stock ticker and financial performance data	https://finance.yahoo.com/
Crunchbase	Company Data	Company and startup data	https://www.crunchbase.com/
TechCrunch	News & Blogs	Tech news blog	https://techcrunch.com/
CB Insights Tech Trends 2024 Report	Industry Trends	Industry articles and thought leadership	https://www.cbinsights.com/research/report/top-tech-trends-2024/
McKinsey Technology Trends Outlook 2024	Industry Trends	Industry articles and thought leadership	https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-top-trends-in-tech
Apple IR	Financial Data	Apple Investor Relations site	https://www.apple.com/investor/
Microsoft IR	Financial Data	Microsoft Investor Relations site	https://www.microsoft.com/en-us/investor

NVIDIA IR	Financial Data	NVIDIA Investor Relations site	https://investor.nvidia.com/
SEC EDGAR Site	Financial Data	SEC filings for publicly traded companies	https://www.sec.gov/search-filings
Google Trends	Industry Trends Data	Google Search trend data	https://trends.google.com/trends/
The Verge	News & Blogs	Tech blog	https://www.theverge.com/tech
US Government Open Data	Government & Open Data	US government data portal covering tech, workforce, etc.	https://data.gov/
Google Patents	Government & Open Data	Company patent filing information	https://patents.google.com/
Twitter/X API	Social API	Twitter API	https://docs.x.com/home
Alphabet Investor Relations	Financial Data	Financial reports and filings for Google's parent company	https://abc.xyz/investor/
Meta Investor Relations	Financial Data	Investor disclosures for Meta (Facebook, Instagram, WhatsApp)	https://investor.fb.com/home/default.aspx

Amazon Investor Relations	Financial Data	Financials, press releases, and shareholder info for Amazon	https://ir.aboutamazon.com/
Tesla Investor Relations	Financial Data	Official source for Tesla's financial data and presentations	https://ir.tesla.com/
Intel Investor Relations	Financial Data	Financials and strategic updates for Intel	https://www.intc.com/
Wired – Business & Tech	News & Blogs	Covers business, innovation, and technology culture	https://www.wired.com/category/business/
IEEE Spectrum	News & Blogs	Engineering-focused tech news and innovation stories	https://spectrum.ieee.org/
OECD STI Outlook	Government & Open Data	Trends and outlooks in global science, tech, and innovation policy	https://www.oecd.org/sti/
GitHub Trending	Government & Open Data	Live view of trending software projects and repos by language	https://github.com/trending
Statista – Technology & Telecommunications Reports	Government & Open Data	Quantitative data, forecasts, and trend analysis in tech	https://www.statista.com/markets/424/topic/482/technology-telecommunications/
WIPO – IP Statistics Data Center	Government & Open Data	International patent data and innovation metrics	https://www3.wipo.int/ipstats/

PitchBook – Emerging Technology Research (Free Reports)	Company Data	Free briefs on verticals like AI and fintech; useful for trend discovery	https://pitchbook.com/news/reports