Building A Causality Graph For Strategic Foresight

Jože M. Rožanec
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
joze.rozanec@ijs.si

Beno Šircelj
Jožef Stefan Institute
Ljubljana, Slovenia
beno.sircelj@ijs.si

Peter Nemec
Event Registry d.o.o.
Ljubljana, Slovenia
peter@eventregistry.org

Gregor Leban
Event Registry d.o.o.
Ljubljana, Slovenia
gregor@eventregistry.org

Dunja Mladenić
Jožef Stefan Institute
Ljubljana, Slovenia
dunja.mladenic@ijs.si

ABSTRACT

This paper describes a pipeline built to generate a causality graph for strategic foresight. The pipeline interfaces with a well-known global media retrieval platform, which performs real-time tracking of events reported in the media. The events are retrieved from the media retrieval platform, and content from the media articles is processed with ChatGPT to extract causal relations mentioned in the news article. Multiple post-processing steps are performed to clean the causal relations, removing spurious ones and linking them to ontological concepts where possible. Finally, a sample causality trace is showcased to exemplify the potential of the causality graph created so far.

KEYWORDS

strategic foresight, graph, causality extraction, wikifier, ChatGPT

1 INTRODUCTION

Among the most frequently used strategic foresight methods we find scenario planning [7], that aims to foresee relevant scenarios based on trends and factors of influence. These allow for a better understanding of how actions can influence the future - a key ability in a world full of Turbulence, Unpredictability, Uncertainty, Novelty, and Ambiguity (TUNA) [30]. This ability has fostered an increasing adoption of strategic foresight in the public and private sectors [6, 21].

Domain experts currently plan scenarios by gathering and analyzing the data to determine and report probable, possible, and plausible futures of interest [15]. Nevertheless, the extensive manual work imposes severe scalability limitations and can introduce bias into the assessments [7]. To overcome such limitations, artificial intelligence was proposed to automate information scanning and data analysis [4, 18].

While the value of artificial intelligence for strategic foresight has been recognized, artificial intelligence has not been widely adopted yet [4, 20]. This is also reflected in scientific papers on foresight and artificial intelligence. For example, we queried Google Scholar for "data-supported foresight" and "strategic foresight artificial intelligence" considering the start time is unlimited, and the deadline is September 6th 2023. When analyzing the first 50 search results of each, we got 18% (9/50) and 40% (20/50) relevant hits, respectively. Some approaches described in the literature aim to leverage artificial intelligence to automate time-consuming aspects of strategic foresight, such as performing information scanning and data analysis [4, 18]. Furthermore, text-mining techniques have been used to identify weak signals and trends [10] or extract relevant actions and outcomes that could be mapped to causal decision diagrams [19].

Strategic foresight for environmental purposes has been considered to different degrees by countries and environmental agencies. For example, multiple U.S. Environmental Protection Agency offices began using strategic foresight in the 1980s. Still, they did not do so consistently until 1995, when it began to be institutionalized and connected to the Agency's strategic planning and decision-making, and reinvigorated since 2015 with that purpose [11]. Another example is The Netherlands, where strategic foresight has been encouraged since 1992 to systematically aim to identify critical technologies and scientific possibilities that would allow the fulfillment of environmental policies [29]. Other cases include using strategic foresight to understand how EU-wide policies may affect regions and rural localities [26] or guide decision-making in the face of structural change [2].

Previous work [22, 23] described how artificial intelligence could be used to automate scenario planning. This paper describes a pipeline built to extract and process media news from EventRegistry [16] to create a causality graph. Furthermore, it describes the causality graph created with media news reporting on events related to oil prices, given the abundant research regarding how oil prices impact the environment. Among the benefits of this approach is the ability to extract causal relations with little human intervention and no supervision. The resulting graph enables the creation of link prediction models that can be used to predict future events based on an array of events that have been observed in the past.

This paper is organized as follows. First, section 2 describes how a data extraction pipeline was built, retrieving media events of interest and extracting causal relationships observed in the world and described in them. Section 3 briefly describes some of the results obtained, providing (i) a quantitative assessment of error types and resulting causal relationships after data cleansing procedures and (ii) a qualitative assessment of causality relationships generated through the pipeline. Finally, Section 4 concludes and outlines future work.

2 DATA EXTRACTION PIPELINE

The data extraction pipeline aims to query relevant media news, process them, and extract causal relationships that can be modeled in a graph. Given the specific interest in modeling causality

Information Society 2023, 9–13 October 2023, Ljubljana, Slovenia

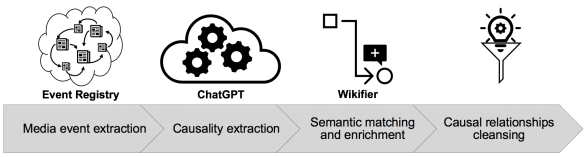

Figure 1: Data extraction pipeline used to retrieve media events and extract causal relationships.

for environmental protection, some research was performed to identify possible topics of interest. Among potential topics, the influence of oil prices on the environment was selected, considering such a topic is frequently covered in the media and was researched to a certain extent. Research has shown that oil price fluctuations (a) affect the consumption of renewable energy sources [1, 28], (b) stimulate green innovation, and that positive shocks in oil prices reduce CO_2 emissions [12], and enhance ecological quality [8, 14].

The data extraction pipeline is summarized in Fig. 1, and each component is briefly described in the following subsections.

2.1 Media Event Extraction

The EventRegistry platform provides real-time insights into media events by sourcing them from the News Feed service [27], processing them and creating media events based on cross-lingual clusters of media news, which are later exposed through an API. The news processing steps require news semantic annotation, extraction of date references, cross-lingual matching, and detection of news duplicates. The cross-lingual clusters denoting a particular media event have a summary describing the media event, information regarding the piece of news considered a centroid to the cluster, and other relevant information.

The first step in the pipeline queries the EventRegistry media event API to extract media events related to a particular concept. This research's query concept was limited to the "Price of Oil". Since EventRegistry has a history of data up to 2014, relevant geopolitical and economic events that influenced oil prices since 2014 were searched. Two events were highlighted by the U.S. Energy Information Administration [1]: (a) the fact that OPEC production quota remained unchanged in the first quarter of 2015 and (b) a reduction in oil demand registered due to the global pandemic in the first quarter of 2020. Furthermore, events between 2022 and 2023 were considered, given the impact of the Russo-Ukrainian War on oil prices [17]. For each event obtained, the centroid media news was queried, its text extracted, wikified, and stored for further processing.

2.2 Causality extraction

To extract causal relations from media events, the OpenAI ChatGPT (gpt-3.5-turbo) was used as a one-shot learning model. To that end, a random media event was sampled, the causality relationships extracted, and both (the text and causal relationships) presented to the model, asking it to recognize causal relationships in the media news. Several iterations of prompt engineering were performed to ensure high-quality results, performing a manual assessment of random results.

The causal relationships persisted in JSON files discriminated the cause, effect, related entities, and locations. In particular, cause, effect, entities, and locations were defined in the following manner:

- **Cause or effect**: contains an entity which is an item, individual, or company that an event happened to;
- **Event**: is an action, development, happening, or state of the entity that is causing or was affected by a cause in the relationship;
- **Location**: geographical location where the event in the cause or effect took place;

Once the causal relationships were extracted, the cause and effect were post-processed, removing adjectives so that only the nouns were left. E.g., *higher diesel prices* was converted to *diesel prices*. The decision was made considering that by doing so, (a) the causes and effects would gain greater support and, therefore, strengthen the information signal in a graph, and (b) that a human expert would be able to determine how a cause and effect may relate given his domain knowledge and a particular context. For example, given the relationship *Inflation → Consumer price index*, the human expert will immediately understand how the consumer price index is affected in a growing or shrinking inflationary context. For each causal relationship, a trace was kept to associate them with the media event from which they were extracted to enable further analysis when required.

2.3 Semantic matching and enrichment

The entire text of the media article was parsed using Wikifier [5]. Data from Wikifier was employed in two distinct ways: firstly, to enrich location data, and secondly, to associate entities to relevant semantic concepts.

The Wikifier tool marks which words in the wikified text correspond to certain semantic concepts. Such annotations were matched to the entities extracted by ChatGPT as part of the causal relationships. To successfully match strings to semantic concepts, some preprocessing was required. First, the non-letter symbols and stopwords were removed, followed by the stemming of each word. It was considered a match if at least one identical string between the text related to marked concepts and the causal relationship. Not all of the semantic concepts listed by the Wikifier were considered: (a) the concepts were required to have a PageRank higher than 0.0001; (b) for location data, only the concepts categorized as *"place"* were considered, and (c) when substituting the original entity by the associated semantic concept, the semantic concept with the highest cosine similarity between the article it's corresponding Wikipedia page was considered.

2.4 Cleansing causal relations

After extracting causal relations, we focused on analyzing the data and cleansing to ensure only relevant relations were considered and used to build a causality graph. Subsequent random sampling iterations were performed, extracting 300 causal relationships in each iteration, which were then analyzed. In each iteration, the causal relations were assessed to determine whether they were meaningful to the topic under consideration, to identify common errors, and to propose mitigation strategies that could amend such errors or filter useless causal relations. We typified six such cases, five originating from ChatGPT and one when semantically post-processing the causal relations with concepts obtained from the Wikifier:

- **repeated entity**: [ChatGPT] the same entity is registered for cause and effect. E.g., *Oil price → Oil price*.

[1]The events were highlighted in the following report, last accessed on August 25th 2023: https://www.eia.gov/finance/markets/crudeoil/spot_prices.php.

- **empty entity**: [ChatGPT] an entity is missing as cause or effect. E.g., → *Oil price*.
- **missing entity**: [ChatGPT] ChatGPT omits the actual entity but could be inferred from the text by the human reader. E.g., *S&P 500 capital expenditures* → *growth*, *energy policy* → *defiance*, or *survey* → *Nasdaq 100*.
- **time entity**: [ChatGPT] some time-period is considered an entity. E.g., *drilling activity* → *2016*, or *(US) shale oil supply* → *end of the year*.
- **non-entity**: [ChatGPT] words marked as entities don't mean anything coherent. E.g., *retail sales* → *risk appetite*.
- **wrong conversion**: [Wikifier] the entity was changed to something unrelated to the one stated in the text. E.g., *Australian government* > *Australian dollar*, or *political tensions* > *Breakup of Yugoslavia*.

While the mitigation strategy for most of the abovementioned errors is to remove the causal relationship, for *missing entity*, a follow-up question will be provided to ChatGPT to get a more concrete answer. This last mitigation strategy has not been implemented yet. Furthermore, a list of concept mappings will be considered to reduce clutter. For example, *Wage Growth* or *1980s Oil Glut* should be replaced by *Wage* or *Oil Glut*, respectively. *Breakup of Yugoslavia* could be replaced by *Country Breakup*. Finally, a more thorough linking to semantic concepts and ontologies is required (e.g., *Jerome Powell* could be linked to *Central Bank*).

After the abovementioned cleansing, the strings were turned into lowercase and trimmed, and most non-alphabetical characters were removed. Further sampling and entity evaluation were performed, creating a dictionary to match string occurrences to a particular concept. It must be noted that the dictionaries do not provide an exhaustive mapping and that ongoing work is being done to further refine and complete the mapping phase. Such dictionaries were created to provide ground for future ontological mapping based on existing ontologies and ontologies that will be developed for this purpose. Finally, all the relations that, after the described process, were extracted from only one media event were discarded, given they are very likely to introduce noise.

2.5 Creating a causality graph

Once causal relationships were extracted, a causality graph was created by matching $cause \rightarrow effect$. Furthermore, some metrics were computed to assess the graph characteristics. The graph can be sampled and visualized with the NetworkX [2] library, which creates a dynamic HTML interface to view it. For each cause and all the possible effects following it, probabilities of each effect occurring were computed based on the ratios present in the data.

3 RESULTS

A total of 2,503 media events were extracted from EventRegistry. When processed with ChatGPT, 12,290 unique causal relationships were extracted, totaling 14,226 unique entities. Those were processed to remove possible errors. Considering *repeated entity* and *empty entity* errors, 253 causal relations were removed. After applying wikification, 9,726 unique causal relations remained, totaling 7,723 entities. 845 causal relations were removed, considering *repeated entity* and *empty entity* errors. Table 1 shows the number of causal relations affected by a particular error type, considering a random sample of 300 causal relations.

[2]The library is documented at the following website: https://networkx.org/

Error type	Count	Percentage
Wrong conversion	17	5.7%
Missing entity	15	5.0%
Non-entity	9	3.0%
Time entity	3	1.0%

Table 1: Statistics for typified errors based on a random sample of 300 causal relationships.

After performing the abovementioned cleansing and dictionary-based mappings, 7,723 nodes and 9,726 edges were obtained. Removing causal relationships reported only in a single media event reduced the graph size to 489 nodes and 877 edges.

3.1 Causality graph and causality chain analysis

Causal chains were created by linking causes and effects extracted from media events. While these are not always completely accurate, they help to identify sequences of events that may take place. Furthermore, while currently not implemented, graph link prediction could be used to predict future event sequences based on patterns observed in the past.

This section provides an example regarding a causality chain of interest retrieved from the causality graph. The causality chain is briefly analyzed to demonstrate how it captures relevant knowledge. In particular, many causality chains displayed the following pattern: *Pandemic* → *Currency* → *Price of Oil* → *Economic Growth* → *Oil Glut* → *Inflation* → *Central Bank* → *Stock Market* → *Investment*.

The complete causality chain summarized above was: *Pandemic* → *Currency* → *Price of Oil* → *Crude Oil Futures* → *Fuel Pricing* → *Economic Growth* → *Petroleum* → *Oil Glut* → *Consumer Price Index* → *Monetary Policy* → *Inflation* → *Central Bank* → *Stock Market* → *Investment* → *Bond*.

To validate the causality chain, scientific literature and events from the past few years were reviewed to find research and examples to validate the causal relationships. For the causality chain described above, we found that the *Pandemic* influenced *Currency*: countries experiencing a sharp daily rise in COVID-19 deaths usually saw their currencies weaken [13]. Causality between exchange rates (*Currency*) and *Price of Oil* has been reported by the European Central Bank [9]. In particular, it has been noticed that the exchange rates can affect oil prices through financial markets, financial assets, portfolio rebalancing, and heading practices. It has also been noted that given the oil prices are expressed in US dollars, the oil futures can be used to hedge against an expected depreciation in US dollars - something that explains the causal relationship between *Price of Oil* and *Crude Oil Futures*. Furthermore, a relationship exists between futures and spot prices (futures prices tend to converge upon spot prices [3] and between oil prices and fuel prices[4], validating the causal relationship between *Crude Oil Futures* and *Fuel Pricing*.

[3]See *"Futures Prices Converge Upon Spot Prices"*, last accessed at https://www.investopedia.com/ask/answers/06/futuresconvergespot.asp in September 7th 2023.
[4]See *"Gasoline explained: Factors affecting gasoline prices"*, last accessed at https://www.eia.gov/energyexplained/gasoline/factors-affecting-gasoline-prices.php in September 7th 2023.

When considering the relationship *Fuel Pricing* and *Economic Growth*, we found that the relationship is validated with energy prices [3], e.g., with gas prices: higher gas prices negatively impact the economy[5]. Economic growth can affect the petroleum market and, in particular, lead to an oil glut (a significant surplus of crude oil caused by falling demand) as it happened at the beginning of the COVID-19 pandemic[6]. Furthermore, oil pricing can have direct or indirect effects on *Inflation* [24], which is reflected in the *Consumer Price Index*, and which can trigger a particular *Monetary Policy* from the *Central Bank* in response to it. Finally, monetary policies affect the stock market and investments [25].

While the causality chain displayed in this case is mostly clean, some improvements are required to make it neater. For example, based on domain knowledge, and depending on the context, the *Consumer Price Index* and *Inflation* could be merged into a single concept, and *Monetary Policy* and *Central Bank* could be considered as one.

The ingestion pipeline requires further work to enhance the concept mappings. We envision that the dictionaries will be further evolved and linked to specific ontologies that could be used to assign semantic meaning and, e.g., contract links in a chain with the same semantic ancestor.

4 CONCLUSIONS

This research has described a pipeline created for causality extraction from media news and aimed toward a strategic foresight tool, and currently focused on events affecting oil prices. Particular errors in the causality extraction were identified and typified, and mitigation measures were implemented. Nevertheless, further work is required to improve the pipeline. Future work will consider three directions: (a) string to ontologies mapping to ensure the captured causes and effects can be tied to particular semantic knowledge and exploit it, (b) generate richer cause and effect representations so that based on encoded metadata, better causality patterns can be elucidated, and (c) create a link prediction model based on the causality graph.

ACKNOWLEDGMENTS

The Slovenian Research Agency supported this work. This research was developed as part of the Graph-Massivizer project funded under the Horizon Europe research and innovation program of the European Union under grant agreement 101093202.

REFERENCES

[1] Nicholas Apergis and James E Payne. 2015. Renewable energy, output, carbon dioxide emissions, and oil prices: evidence from South America. *Energy Sources, Part B: Economics, Planning, and Policy* 10, 3 (2015), 281–287.
[2] M Bruce Beck. 2005. Environmental foresight and structural change. *Environmental Modelling & Software* 20, 6 (2005), 651–670.
[3] Istemi Berk and Hakan Yetkiner. 2014. Energy prices and economic growth in the long run: Theory and evidence. *Renewable and Sustainable Energy Reviews* 36 (2014), 228–235.
[4] Patrick Brandtner and Marius Mates. 2021. Artificial Intelligence in Strategic Foresight–Current Practices and Future Application Potentials: Current Practices and Future Application Potentials. In *The 2021 12th International Conference on E-business, Management and Economics*. 75–81.
[5] Janez Brank, Gregor Leban, and Marko Grobelnik. 2017. Annotating documents with relevant wikipedia concepts. *Proceedings of SiKDD* 472 (2017).
[6] George Burt and Anup Karath Nair. 2020. Rigidities of imagination in scenario planning: Strategic foresight through 'Unlearning'. *Technological Forecasting and Social Change* 153 (2020), 119927.
[7] Ashkan Ebadi, Alain Auger, and Yvan Gauthier. 2022. Detecting emerging technologies and their evolution using deep learning and weak signal analysis. *Journal of Informetrics* 16, 4 (2022), 101344.
[8] Ali Ebaid, Hooi Hooi Lean, and Usama Al-Mulali. 2022. Do oil price shocks matter for environmental degradation? Evidence of the environmental Kuznets curve in GCC countries. *Frontiers in Environmental Science* 10 (2022), 860942.
[9] Marcel Fratzscher, Daniel Schneider, and Ine Van Robays. 2014. Oil prices, exchange rates and asset prices. (2014).
[10] Amber Geurts, Ralph Gutknecht, Philine Warnke, Arjen Goetheer, Elna Schirrmeister, Babette Bakker, and Svetlana Meissner. 2022. New perspectives for data-supported foresight: The hybrid AI-expert approach. *Futures & Foresight Science* 4, 1 (2022), e99.
[11] Joseph M Greenblott, Thomas O'Farrell, Robert Olson, and Beth Burchard. 2019. Strategic foresight in the federal government: a survey of methods, resources, and institutional arrangements. *World futures review* 11, 3 (2019), 245–266.
[12] Jinyan Hu, Kai-Hua Wang, Chi Wei Su, and Muhammad Umar. 2022. Oil price, green innovation and institutional pressure: A China's perspective. *Resources Policy* 78 (2022), 102788.
[13] Aamir Jamal and Mudaser Ahad Bhat. 2022. COVID-19 pandemic and the exchange rate movements: evidence from six major COVID-19 hot spots. *Future Business Journal* 8, 1 (2022), 17.
[14] Foday Joof, Ahmed Samour, Mumtaz Ali, Turgut Tursoy, Mohammad Haseeb, Md Emran Hossain, and Mustafa Kamal. 2023. Symmetric and asymmetric effects of gold, and oil price on environment: The role of clean energy in China. *Resources Policy* 81 (2023), 103443.
[15] Kevin Kohler. 2021. Strategic Foresight: Knowledge, Tools, and Methods for the Future. *CSS Risk and Resilience Reports* (2021).
[16] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*. 107–110.
[17] Gaye-Del Lo, Isaac Marcelin, Théophile Bassène, and Babacar Sène. 2022. The Russo-Ukrainian war and financial markets: the role of dependence on Russian commodities. *Finance Research Letters* 50 (2022), 103194.
[18] Nathan H Parrish, Anna L Buczak, Jared T Zook, James P Howard, Brian J Ellison, and Benjamin D Baugher. 2019. Crystal cube: Multidisciplinary approach to disruptive events prediction. In *Advances in Human Factors, Business Management and Society: Proceedings of the AHFE 2018 International Conference on Human Factors, Business Management and Society, July 21-25, 2018, Loews Sapphire Falls Resort at Universal Studios, Orlando, Florida, USA 9*. Springer, 571–581.
[19] Lorien Pratt, Christophe Bisson, and Thierry Warin. 2023. Bringing advanced technology to strategic decision-making: The Decision Intelligence/Data Science (DI/DS) Integration framework. *Futures* 152 (2023), 103217.
[20] Norbert Reez. 2020. Foresight-Based Leadership. Decision-Making in a Growing AI Environment. In *International Security Management: New Solutions to Complexity*. Springer, 323–341.
[21] Aaron B Rosa, Niklas Gudowsky, and Petteri Repo. 2021. Sensemaking and lens-shaping: Identifying citizen contributions to foresight through comparative topic modelling. *Futures* 129 (2021), 102733.
[22] Joze Rozanec, Peter Nemec, Gregor Leban, and Marko Grobelnik. 2023. AI, What Does the Future Hold for Us? Automating Strategic Foresight. In *Companion of the 2023 ACM/SPEC International Conference on Performance Engineering*. 247–248.
[23] Jože M Rožanec, Radu Prodan, Dumitru Roman, Gregor Leban, and Marko Grobelnik. 2023. AI-based Strategic Foresight for Environment Protection. In *Symposium on AI, Data and Digitalization (SAIDD 2023)*. 7.
[24] Siok Kun Sek, Xue Qi Teo, and Yen Nee Wong. 2015. A comparative study on the effects of oil price changes on inflation. *Procedia Economics and Finance* 26 (2015), 630–636.
[25] Peter Sellin. 2001. Monetary policy and the stock market: theory and empirical evidence. *Journal of economic surveys* 15, 4 (2001), 491–541.
[26] Anastasia Stratigea and Maria Giaoutzi. 2012. Linking global to regional scenarios in foresight. *Futures* 44, 10 (2012), 847–859.
[27] Mitja Trampuš and Blaz Novak. 2012. Internals of an aggregated web news feed. In *Proceedings of 15th Multiconference on Information Society*. 221–224.
[28] Victor Troster, Muhammad Shahbaz, and Gazi Salah Uddin. 2018. Renewable energy, oil prices, and economic activity: A Granger-causality in quantiles analysis. *Energy Economics* 70 (2018), 440–452.
[29] Barend Van der Meulen. 1999. The impact of foresight on environmental science and technology policy in the Netherlands. *Futures* 31, 1 (1999), 7–23.
[30] Angela Wilkinson. 2017. Strategic foresight primer. *European Political Strategy Centre* (2017).

[5] See *"How Gas Prices Affect the Economy"*, last accessed at https://www.investopedia.com/financial-edge/0511/how-gas-prices-affect-the-economy.aspx in September 7th 2023.

[6] See *"Oil glut means there's little hope for oil price recovery until 2021"*, last accessed at https://www.conference-board.org/topics/natural-disasters-pandemics/COVID-19-oil-glut in August 30th 2023.