




Scenario Generation With Transitive Rules for Counterfactual Event Analysis

Aigerim Mussina¹ ^a, Paulo Trigo² ^b and Sanzhar Aubakirov¹ ^c

¹Department of Computer Science, Al-Farabi Kazakh National University, 71 al-Farabi Ave., Almaty, Kazakhstan

²GulAA, ISEL - Instituto Superior de Engenharia de Lisboa, Lisbon, Portugal

Keywords: Event Detection, Association Rules, What-If Analysis.

Abstract: Event detection on online social networks is one of the comprehensive approaches for analyzing people's discussions. However, it is not enough to detect an event as people often look for ways to influence the course of an event. Often, in the course of a discussion, the introduction of a new topic can shift the focus to another subject and thus move from one event to another. The causal relationship between topics and events can be explored by extracting association rules among the topics covered in each event. The scenario generation based on those causal relationships can support what-if (counterfactual) analysis and explain transitions between events. In this paper our goal is to generate what-if scenarios among topics of detected events. The association rule approach was chosen as a method for its human-readable output that can be transposed into a counterfactual scenario. We propose methods for time-window constrained topic-based what-if scenario generation founded on market-basket analysis.

1 INTRODUCTION

Nowadays people are often immersed in a continuous stream of textual data generated from social networks. A large volume of data usually emerges (and grows) from people's discussions and around aggregating concepts commonly designated as "events". Also, within an event, the people's discussions unfold around certain "topics". There is thus an increase in research effort in the processing of textual data originating from social networks, with the goal of automatically detecting both the events and their topics. Researchers are interested in detecting events-and-topics as it appears to become a powerful method for the follow-up of people's discussions. However, just detecting events-and-topics is not enough.

The follow-up of people's discussions also involves the challenge of trying to predict the cause-and-effect relationship that results from introducing a new topic into a discussion. People usually (and intuitively) seek to understand not only the source-and-flow of a discussion but also the impact that their own participation, via the introduction of a topic, may have


on that same flow.


Therefore, there is a "topic space" where each person searches for a subset of topics through the generation of different scenarios in order to decide how to best intervene in the flow of a discussion. In this context, a scenario is constructed by combining different topics that may have originated from the same event (intra-event scenario) or from different events (inter-event scenario).


The overall process starts with the event detection process being applied in the context of an online social network (OSN). Therefore, an event is described by a set of topics-of-interest (ToI) that, in turn, were addressed by people, in the course of their OSN interactions (discussions) during a given time-window.

We formally define an event as $E(w, T) = \{t_1, t_2, \dots, t_n\}$, where w is a time-window, T is a ToI dictionary and the $\{t_1, t_2, \dots, t_n\} \subseteq T$ is a topic-set; hence, each $t_x \in E(w, T)$ represents the topic, x , as taken from T and addressed, over w , in the detected event E . We point out that the event detection algorithm resorts to a ToI dictionary in order to improve the process accuracy (Mussina et al., 2022).

The counterfactual analysis was chosen as a base approach to the search for "topic space". It aims to explore cause-and-effect relations by searching for statements such as "if A occurred, then B is also

^a  <https://orcid.org/0000-0002-7043-0810>

^b  <https://orcid.org/0000-0001-5850-615X>

^c  <https://orcid.org/0000-0002-8416-527X>

likely to occur” (Menzies and Beebe, 2001). Considering A and B as events and E definition (above), we may rephrase such a statement as “if, within a time-window, w , topics $\{t_{A_1}, t_{A_2}, \dots, t_{A_r}\}$ occurred, then topics $\{t_{B_1}, t_{B_2}, \dots, t_{B_s}\}$ are also likely to occur in w ”; where t_{E_x} , here simply means that topic x was addressed in event E . Following the idea of what-if perspective, we suppose that A ’s topic-set, $\{t_{A_1}, t_{A_2}, \dots, t_{A_r}\}$, includes “topic space” that have been intervened such that discussion goes to B ’s topic-set, $\{t_{B_1}, t_{B_2}, \dots, t_{B_s}\}$.

In this paper, we follow the counterfactual perspective and propose methods for time-window constrained topic-based what-if scenario generation. Generated scenarios have two types: intra-event and inter-event. The inter-event scenario generation includes two approaches: a one-rule-based approach and a two-rules-transitivity-based approach. We followed a market-basket analysis approach and the proposed methods resort to (unsupervised) association rule extraction techniques when applied to detected events (baskets) and the sets of corresponding topics discussed therein (item-sets within baskets).

2 RELATED WORK

An approach to the event analysis process (that follows from event-detection) falls into the broader research field of “event association extraction”. Usually researchers resort to graph-based formulations, where events are usually modelled as nodes (Shahaf et al., 2015) and the value of edges between nodes (events) are computed from the frequency of common words. The “Connecting the dots” concept with graph neural networks outperformed the results of the state-of-the-art methods (Wu et al., 2020). Analyzing event associations one could extract cause-and-effect relationship.

Another recent research approach is the “counterfactual event analysis” which is being achieved in the healthcare field where researchers are focused on the prediction of the outcome of treatments (Zou et al., 2022). Research aims not only to predict risks of a treatment, but also to weight the cause-effect relation of alternative interventions (Prosperi et al., 2020). Another approach is related to the “prescriptive analysis” that finds relevant applications in business management processes and decision making (Lepenioti et al., 2020).

The main limitations of modern solutions in prescriptive and counterfactual analysis are both the: a) the generation of models that are complex to explain, and b) focus on narrow areas of application. In this

work we apply a market-basket analysis, which provides, as a result, a human readable model based on association rules. Such a model is used to generate what-if counterfactual scenario. In this context, we could not compare results with other researchers right now.

3 COUNTERFACTUAL EVENT ANALYSIS

The market-basket analysis approach is formalized with the following main concepts: *basket*, *item*, *itemset*, *association rule*, *left-hand-side (LHS)*, *right-hand-side (RHS)*, *support* and *confidence*. In a context of event analysis, the detected event is considered as a *basket* and topic is considered as an *item*. The *itemset* $I = \{t_1, t_2, \dots, t_k\}$ is a, k size, set of *items* where each t_x is a topic (item). The *association rule* is constructed from subitemsets and described as implication of the form $LHS \Rightarrow RHS$, where $LHS, RHS \subset I$ and $LHS \cap RHS = \emptyset$. The *support* and *confidence* are metrics that each association rule satisfies. The *support* is the joint probability, $P(LHS, RHS)$, that items in LHS and RHS occur together. The *confidence* is a conditional probability of the form $P(RHS|LHS)$.

Definition 1. A **what-if scenario** is generated from an itemset I_N as an association rule, $LHS \Rightarrow RHS$, that takes the form:

$$Base_L \cup WI_M \Rightarrow RHS_R \quad (1)$$

where $Base_L \cup WI_M = LHS$ and $Base_L \cap WI_M = \emptyset$, also $LHS \subset I_N$ and $RHS \subset I_N$; each L, M and R subscript is the size of the respective subitemset.

In this way the counterfactual scenario perspective can be read as “if WI_M occurs together with the $Base_L$ then RHS_R is also likely to occur”.

Definition 2. An **intra-event scenario** is taken from a **what-if scenario** where $support(WI_M) = support(Base_L) = support(RHS_R)$

In the **intra-event scenario** the equal support means that subitemsets are likely to appear together in each event. This comes from a small time-window event detection where distinct events describe the same real-world discussions.

Definition 3. A one-rule-based **inter-event scenario** is taken from a **what-if scenario** where:

- $support(WI_M)$ is the **minimum** from all subitemsets of size M ,
- $support(Base_L)$ is the **maximum** from all subitemsets of size L ,
- $support(RHS_R)$ is the **maximum** from all subitemsets of size R .

Events may have common topic-subsets but they actually represent different real-world discussions. In that case, we use the **inter-event** scenario that can show the path from one event to another. The association rules extraction for the **inter-event** scenario is focused on topics that appear in several events. A threshold value, h , is used to control the number of events with common topics.

Definition 4. A two-rules-transitivity-based **inter-event scenario** is taken from two association rules of the form:

$$\begin{aligned} Base_L &\Rightarrow WI_M \\ WI_M &\Rightarrow RHS_R \end{aligned} \tag{2}$$

where $M > L$, $M > R$ and $Base_L, WI_M \subset I_{N1}$, $WI_M, RHS_R \subset I_{N2}$, I_{N1} and I_{N2} are different itemsets.

Analysis of two rules at once is another approach of **inter-event** scenario generation. Here we try to pass from $Base_L$ (left-hand-side) to RHS_R (right-hand-side) through WI_M (transitive subitemset). The key feature is the association rule’s confidence anti-monotone property. It states that confidence is anti-monotone with respect to the number of items on the right-hand-side of the rule. i.e., an increase in right-hand-side dimension decreases or maintains its confidence. In our case, if $M > L$ in $Base_L \Rightarrow WI_M$ then the probability that WI_M will happen when $Base_L$ happened is low. However, if we also have $M > R$ in $WI_M \Rightarrow RHS_R$ then the probability that RHS_R will happen when WI_M happened is high. In this **inter-event** scenario we pass from the low confidence rule to the high confidence rule.

4 RESULTS

An overall view of our work is depicted in figure 1. The left side of the figure represents our previous work on the event detection where we explored two input data sources – corpora from Twitter “Events2012” and Topics (ToI) (McMinn et al., 2013) – and implemented the SEDTWik (Morabia et al., 2019) algorithm for event detection replacing Wikipedia with ToI dictionaries (Mussina et al., 2022). Since event detection method uses frequency counting to determine the burst in the use of certain words, it was necessary to remove the noise and reduce each word to the same form. Tweets were cleaned of stop words and converted to their stemmed form. The NLTK library was used in the process of cleaning tweets from stop words and stemming (with PorterStemmer) (NLTK, 2022). The detected events were combined into one file for each ToI dictionary, where each line corresponds to an event.

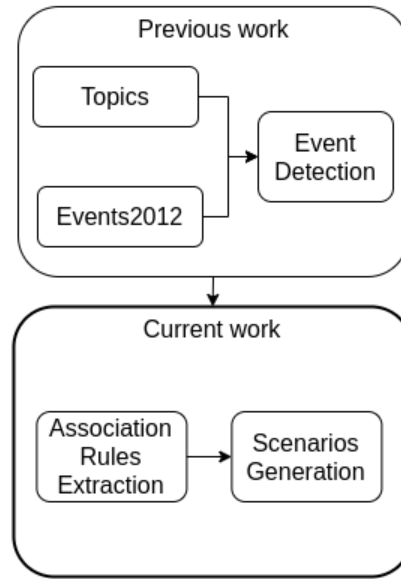


Figure 1: Overall view of the work.

The right side of the figure 1 represents this paper’s current work. The association rules extraction resorts to the market-basket approach being applied to the detected events. Given that the approach may identify a large number of rules, it was limited by the number of rules and minimum support and confidence. Therefore, the support-and-confidence space was explored to extract groups with at most 10 rules. From those extracted rules we generate the scenarios. As it was mentioned above the scenarios could be intra-event and inter-event. Below we describe the scenarios’ generation details along with preliminary results.

4.1 Intra-Event Associations

Since detected events have a small number of topics and the number of words in a “topic space” is enormous, we faced many rules being extracted from one event. One such rule, where the support of each subitemset is about 0.029, is presented in the table 1. From that rule, the following counterfactual scenario was generated: “if ‘*shrine*’ occurs together with the ‘*nearbi*’, ‘*flood*’ then ‘*sanctuari*’, ‘*evacu*’ are also likely to occur”. This scenario can be interpreted as “if a flood occurs near the shrine, then people from the sanctuary should be evacuated”.

4.2 Inter-Event Associations

The first tests showed that the size of topic-sets should be increased to extract the inter-event association rules. The subitemsets support was lower than 0.1,

Table 1: Association rules for one real-world event.

	$Base_L$	WI_M	RHS_R
itemssets	'nearbi', 'flood'	'shrine'	'sanctu- ari', 'evacu'
support	0.0294	0.0294	0.0294

which means that the probability of topics appearing together is tiny. The average topic-set size was 9, which is not enough to search for common topics in a certain number of events. For the inter-event scenario generation, the topic-set expanded by tweets containing topics, which construct event clusters in event detection. The added tweets were cleared from non-ToI dictionary topics. As a result, the average topic-set size increased to 453.

In the context of an inter-event scenario generation, we use topics that appear in several events. If the number of events containing a topic is greater than the threshold value, h , then this topic will be included in the itemset, I . We intuitively set the threshold value as $h = 10$.

4.2.1 One-Rule-Based Approach

This subsection provides two examples. The first example is from the dictionary "Armed Conflicts and Attacks", and the second is from "Arts, Culture and Entertainment". For the first experiments, the following sizes of subitemsets of the what-if scenario were intuitively chosen: $L = 2$, $M = 1$, and $R = 2$. The extracted association rules corresponding to the inter-event scenario definition are presented in the table 2.

Table 2: Association rules support - example 1.

	$Base_L$	WI_M	RHS_R
itemssets	'news', 'kill'	'car'	'least', 'bomb'
support	0.6712	0.4521	0.4246

Table 3: Association rules support - example 2.

	$Base_L$	WI_M	RHS_R
itemssets	'news', 'kill'	'car'	'least', 'bomb'
support	0.6712	0.4521	0.4246

It is necessary to identify events from the extracted rules in order to better understand the results. We found which detected events contain the WI_M subitemset. Detected events are listed in the table 4. More detailed and readable descriptions of the real-world situations are shown in the table 5. Descriptions of the events taken from the open-source "Wikipedia

Current Events Portal" (Wikipedia, 2022).

Table 4: Detected events topics.

	Detected Event A	Detected Event B
Ex. 1	'beirut', 'car', 'eight', 'central', 'explod'	'kill', 'bomb', 'syria', 'turkey', 'car', 'attack', 'suicid', 'central', 'wound'
Ex. 2	'pope', 'coptic', 'egypt', 'christian', 'bishop', 'chosen', 'chosen', 'select'	'egypt', 'christian', 'chosen', 'blindfold', 'crystal', 'copt'

Table 5: Real-world events description.

	Real-world Event A	Real-world Event B
Ex. 1	"A car bomb explodes at Sassine Square in the Lebanese capital of Beirut, killing and injuring 12..."; at least eight people ..."	"A car bomb detonates in Semdinli, Turkey, killing 1 and injuring 12..."; "A suicide car bomber detonates a bomb in the Hama province of Syria ..."
Ex. 2	"Bishop Richard Williamson is expelled from the Society of Saint Pius X (SSPX) ..."	"A shortlist of successors to the Coptic Pope is drawn up; a blindfolded child is then expected to pick from a list of three. (BBC)"

From the obtained results, we could derive the following scenarios:

- "if 'car' occurs together with the 'news', 'kill' then 'bomb', 'least' are also likely to occur". If there is a news about murder and additional information as a car occurred, then we can suggest that a bomb in a car exploded.
- "if 'choos' occurs together with the 'pope', 'egypt' then 'christian', 'coptic' are also likely to occur". Even though events in real-world are not strictly connected, it is interesting that Event A was about exclusion of bishop, Event B was about including new members in the Coptic Pope. Events connected by topic 'choos'. This topic is a stemmed version of word choose.

4.2.2 Two-Rules-Transitivity-Based Approach

In this subsection we present an example of a two-rules-transitivity-based inter-event scenario generation. The example is presented in the table 6.

Table 6: Two-rules-transitivity-based approach example.

	<i>LHS</i>	<i>RHS</i>
Rule 1	'least', 'syria'	'news', 'car', 'kill', 'bomb'
Rule 2	'news', 'car', 'kill', 'bomb'	'soldier', 'sever'

Rule 1, from the table 6, describes an event with description: "A car bomb explodes at Sassine Square in the Lebanese capital of Beirut, killing at least eight people and wounding up to 78 others. (BBC)". Rule 2 describes an event with description: "Syrian civil war: A Jordanian soldier dies during a gunfight between Jordanian troops and Islamic militants attempting to cross the border into Syria. (CTV News)".

From that rules, the following counterfactual scenario was generated: "if 'news', 'car', 'kill', 'bomb' occurs together with the 'least', 'syria' then 'soldier', 'sever' are also likely to occur".

5 CONCLUSION

In this paper, we proposed methods for time-window constrained topic-based what-if scenario generation, in the counterfactual perspective, founded on market-basket analysis and association rules extraction. Definitions of counterfactual scenarios, both intra-event and inter-event, are given. Preliminary results illustrate the extraction of coherent causal effects and require more analysis and controlled experiments. Future work will apply the methods to events detected using other ToI dictionaries. We will also include evaluating the proposed methods in the context of the usefulness of association rules and scenarios.

ACKNOWLEDGEMENTS

This work was supported by the grant of the Ministry of Science and Higher Education of the Republic of Kazakhstan, project BR10965311 "Development of intelligent information and telecommunication systems for urban infrastructure: transport, ecology, energy, and data-analytics in the Smart City concept".

REFERENCES

- Lepenioti, K., Bousdekis, A., Apostolou, D., and Mentzas, G. (2020). Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, 50:57–70.
- McMinn, A. J., Moshfeghi, Y., and Jose, J. M. (2013). Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 409–418.
- Menzies, P. and Beebee, H. (2001). Counterfactual theories of causation.
- Morabia, K., Murthy, N. L. B., Malapati, A., and Samant, S. (2019). Sedtwik: segmentation-based event detection from tweets using wikipedia. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 77–85.
- Mussina, A., Aubakirov, S., and Trigo, P. (2022). Parametrized event analysis from social networks. *Scientific Journal of Astana IT University*, 10(10).
- NLTK (2022). NLTK :: sample usage for stem. <https://www.nltk.org/howto/stem.html>. [Online; accessed 09-November-2022].
- Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., Rich, S., Wang, M., Buchan, I. E., and Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2(7):369–375.
- Shahaf, D., Guestrin, C., Horvitz, E., and Leskovec, J. (2015). Information cartography. *Communications of the ACM*, 58(11):62–73.
- Wikipedia (2022). Wikipedia Current Events Portal. https://en.wikipedia.org/wiki/Portal:Current_events/. [Online; accessed 15-December-2022].
- Wu, Z., Pan, S., Long, G., Jiang, J., Chang, X., and Zhang, C. (2020). Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 753–763.
- Zou, H., Li, B., Han, J., Chen, S., Ding, X., and Cui, P. (2022). Counterfactual prediction for outcome-oriented treatments. In *International Conference on Machine Learning*, pages 27693–27706. PMLR.