

TDK-dolgozat

Szerző: Ficsor Attila

Exploring sustainable development issues: Frequent pattern
mining and hypergraph-based analysis of world news

Szerző: Ficsor Attila

Adattudomány MSc. 1. félév

Témavezetők: Dr. Abonyi János rektor, Cvetkó Tímea PHD hallgató

HUN-REN-PE Complex Systems Monitoring Research Group

Exploring sustainable development issues: Frequent pattern mining and hypergraph-based analysis of world news

*Készítette: Ficsor Attila
Adattudomány mesterképzés*

Témavezető: Dr. Abonyi János, Czvetkó Tímea

Climate change significantly affects the progress toward sustainability goals and the improvement of people's quality of life. This research aims to support a better understanding of global events related to sustainability, offering valuable insights that can assist in informed decision-making by analysing sustainability-related world news using data and network science tools. Additionally, an open-source platform is created to interactively represent and consolidate information from global news, focusing on the interrelationships of events and developments that impact sustainability and relating them to the United Nations Sustainable Development Goals (SDGs). Big data analytics and network analysis are used to automatically collect and organize significant events and insights from a multitude of news articles, directly downloaded from the GDELT project. Data science techniques, including frequent pattern mining and association rule mining, are utilized to identify patterns of events, topics, or countries that frequently co-occur, supporting the understanding of the interrelations of sustainability-related issues. The results have been visualized as networks highlighting the interactions between the elements, while network metrics further allow the identification of central elements. The hypergraph-based visualization provides additional information about group interactions and the multidimensional nature of sustainability issues. This research has the potential to guide more informed public policy decisions and identify effective strategies for sustainable development through the identification of the detailed connections between various real-world occurrences and the benchmarks set for sustainability.

Contents

1	Introduction	3
2	Methodology of hypergraphs-based frequent news pattern analysis	6
2.1	Frequent pattern mining of world news	7
2.2	Association rule mining for identifying interconnections of news domains	9
2.3	Hypergraph representation of frequent itemsets	10
2.4	Summary of the methodology	15
3	Implementation of the proposed methods and the interactive web portal	17
4	Application study on climate change-related news analysis	18
4.1	Interactive web portal	18
4.2	Hypergraph-based analysis of frequent patterns found in world news regarding the SDGs	21
4.3	Association rule mining	27
5	Conclusion	29

1 Introduction

In this study, I utilize the potential of analyzing news through data and network science tools in order to explore sustainable development-related issues and uncover trends, patterns, and connections within news content. Thereby providing a more comprehensive understanding of the multifaced challenges and opportunities associated with sustainability, that contribute to informing targeted strategy-making for sustainable development.

Data science tools play a crucial role in addressing climate change challenges [1], but a comprehensive approach must integrate socio-environmental factors alongside climatic analysis for effective policy-making and strategy development [2]. A possible approach for identifying public awareness and prioritization of sustainability-related issues across countries is news analysis [3]. For this purpose, The GDELT Project (Global Database of Events, Language and Tone) can serve as an effective source of sustainable development-specific world news to be monitored in several languages and considered the latest data available [4]. The GDELT Project is an open platform monitoring the news media of the world in over 100 languages and identifies people, locations, organizations, themes, sources, emotions, counts, quotes, images and events related to the news. Data regarding sustainability-themed news can provide information on the interrelationships of sustainability issues on a global scale. This data can be further refined to provide region or country-specific perspectives, highlighting local sustainability challenges and priorities [5].

Research on Sustainable Development Goals (SDGs)-related news analysis utilizing the potential of the GDELT Project has been made, which proves that the interconnectedness of SDGs can be analyzed through news utilizing network analysis [6], however, the study is limited to a pairwise network representation of sustainability events for a specific time frame, leaving the cooccurrence of news domains and the exploration of potential sequences of sustainability issues unexplored. For analyzing GDELT-based news domains that are frequently observed together and applying association rule mining for identifying possible relations between the domains has been proven to be efficient [7], however, the study only applies this approach for identifying co-location patterns, neglecting the broader potential for uncovering intricate in-

terdependencies among sustainability issues and news domains such as events, people, countries etc..

The representation of the multi-level nature of complex interactions regarding the world news, hypergraphs can provide an efficient tool, and allow e.g. cross-context collaborative associations to support context-aware prediction [8]. Utilizing hypergraphs to represent frequent news patterns can lead to substantial progress in comprehending and effectively addressing the challenges related to sustainable development and news analysis.

Therefore, this study has two major objectives that provide a decision support method and tool utilizing data from the GDELT database.

1) This study provides a method that can support decision-making based on the analysis of complex news patterns, trends, and cross-interactions of specific news domains. I apply frequent pattern mining for identifying news domains, such as themes and countries that are frequently observed together. I further apply association rule mining in order to identify relations between the domains that support a more comprehensive understanding of the intricate interdependencies among sustainability issues and news domains, enabling more informed and strategic decision-making e.g. in the context of Sustainable Development Goals. This information is visualized as hypergraphs, which provide a clear representation of group interactions, allowing for a deeper understanding of the complex relationships and multi-level nature between various news domains. In addition, centrality metrics are applied to identify and analyze key components in the network, shedding light on the most influential factors in the context of sustainability-related news analysis, thereby enhancing the effectiveness of decision-making processes and strategy development.

2) This study also offers an open-source web platform that provides accessibility to sustainability-related world news analysis. The web page has been designed with interactivity in mind, enabling users to explore network-based topical, locational, and choropleth map analyses of news content, while also indicating the tone (positive, negative, neutral) of the news. The web platform, although currently lacking frequent pattern and association rule analysis, holds the potential to significantly bolster decision-making processes. Through this platform, researchers and individuals can access near-real-time demonstrations of world news, allowing for time-based searches and analysis of news events.

This accessibility empowers users to stay informed about global sustainability developments and facilitates data-driven decision-making.

In Section 2, the proposed methodological framework is discussed in detail, including the utilization of frequent pattern mining and association rule mining techniques application for news analysis and the hypergraph-based representation of frequent news patterns. The developed interactive web portal is introduced in Section 3. In Section 4, the applicability of the proposed method is demonstrated through an application study in climate change-related news analysis to identify the interaction and patterns of sustainable development issues.

2 Methodology of hypergraphs-based frequent news pattern analysis

In this section, the proposed method is presented in detail according to the structure of Figure 1, which indicates the proposed methodological steps for the news analysis-based exploration of sustainable development issues.

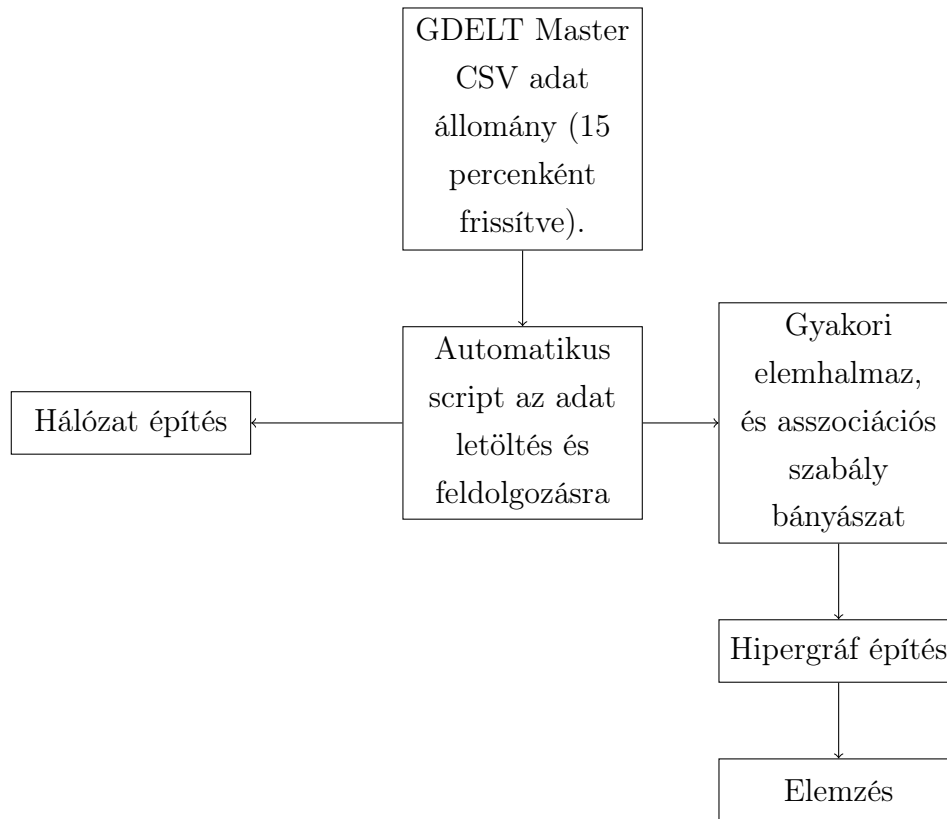


Figure 1: Schematic representation of the proposed methodological steps for new analysis

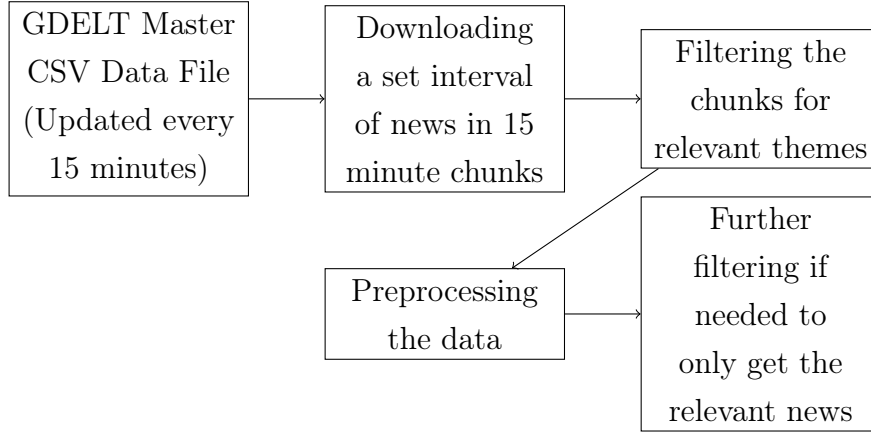


Figure 2: Schematic representation of the proposed methodological steps for new analysis

The news data is derived from GDELT, more specifically the GDELT Global Knowledge Graph [9] which includes several objects of interest such as themes, regions, people, and organizations. These components are crucial for describing the main aspects of a news story. By using this information, world events can be better filtered and analyzed, helping to understand and interpret global news more effectively.

For more specific filtering process, taxonomies can be utilized such as the World Bank Topical Taxonomy that includes a wide range of subjects, which has been linked to the GDELT Global Knowledge Graph [10]. In the presented methodology, data was filtered by specific themes identified by the World Bank Taxonomy codes similar to what can be seen in Table 1, where themes that can be connected to SDG 13 are shown as an example. These themes are from a previous study [6], which links the World Bank taxonomy to Sustainable Development goals. These thematically filtered data are automatically downloaded and preprocessed to allow further analysis.

2.1 Frequent pattern mining of world news

Frequent pattern mining, a key technique in data analysis [11], is centered around identifying recurring patterns within datasets. This approach involves items, that are grouped into transactions, and these transactions are then compiled into a transaction database. In the scope of this study, each news

Table 1: Example of the World Bank themes-based news searching keywords matching with SDG 13 - Climate action.

WB Short Name	WB Full Name
WB_587_	Poverty and Climate Change
WB_1844_	Market-Based Climate Change Mitigation
WB_821_	Disaster Risk Reduction
WB_1787_	Natural Habitats
WB_823_	Post Disaster Recovery and Reconstruction
WB_3138_	Natural Disaster
WB_3358_	Man-Made Disasters
WB_1705_	Disaster Preparedness
...	

article gathered can be viewed as a transaction, the various elements of interest within these articles are the items, and the entire collection of news forms the transaction database.

Let $\mathcal{D}$ represent a database containing N transactions. Within this framework, a transaction $\mathcal{T}$ is defined as a collection of n items. The notation $\mathcal{I}_K$ is used to denote a specific item that can represent a single unique region, theme, organization or person, and $\mathcal{I}$ symbolizes the entire set of items in the database.

$$\begin{aligned}
\mathcal{D} &= \{\mathcal{T}_1, \dots, \mathcal{T}_i, \dots, \mathcal{T}_N\} \\
\mathcal{T}_i &= \{\mathcal{I}_1, \dots, \mathcal{I}_k, \dots, \mathcal{I}_n\} \\
\mathcal{I} &= \{\mathcal{I}_1, \dots, \mathcal{I}_k, \dots, \mathcal{I}_m\}
\end{aligned} \tag{1}$$

The Support of an $A \subseteq \mathcal{I}$ itemset is the number of transactions in which A occurs ($|A|$) normalized between 0 and 1, using the number of all transactions ($|\mathcal{T}|$). This normalized number can be considered as the occurrence probability of A .

$$s_{\mathcal{T}}(A) = \frac{(|A|)}{(|\mathcal{T}|)} = P(A) \tag{2}$$

A closed frequent itemset within this framework is defined as an itemset for which there are no supersets with the same support count. This characterizes

it as a maximal pattern within the transaction database.

$$\mathcal{F} = \{F \subseteq \mathcal{I} \mid \forall G \supseteq F, s_{\mathcal{T}}(G) < s_{\mathcal{T}}(F) \text{ and } s_{\mathcal{T}}(F) \geq \text{minSupport}\} \quad (3)$$

Here, $\mathcal{F}$ is the set of frequent itemsets, and $F \subseteq \mathcal{I}$ represents a single frequent itemset. G represents any superset of F , and minSupport is the predefined minimum support threshold for identifying frequent patterns in the context of the news dataset.

2.2 Association rule mining for identifying interconnections of news domains

To further reveal intricate patterns and relationships within the news data, showing the underlying themes and trends, association rule mining is an efficient technique [12]. In this study, I employ the FP-Growth algorithm to mine these association rules from the dataset.

FP-Growth efficiently generates frequent itemsets using a compact data structure known as the FP-Tree (Frequent Pattern Tree), eliminating the need for candidate generation. The rules derived from FP-Growth are in the form $A \Rightarrow B$, where A and B are subsets of the set of items $\mathcal{I}$. The presence of items in A within a transaction suggests the presence of items in B .

FP-Growth constructs the FP-Tree by compressing the database and maintaining the itemset association information. The algorithm then extracts the frequent itemsets from this tree by recursively growing the patterns from each conditional pattern base, which represents a sub-database for each item. The algorithm then identifies rules that meet both a minimum support threshold, denoted as minSupport, and a minimum confidence threshold, referred to as minConfidence. The confidence of an association rule reflects the likelihood of encountering A given that the transaction already includes B , which can be denoted as:

$$c_{\mathcal{T}}(A \Rightarrow B) = \frac{s_{\mathcal{T}}(A \Rightarrow B)}{s_{\mathcal{T}}(A)} = \frac{P(A \cup B)}{P(A)} = P(B|A) \quad (4)$$

The resulting set of association rules includes each rule whose support and confidence exceed these thresholds.

2.3 Hypergraph representation of frequent itemsets

Frequent itemsets can be represented as a hypergraph [13], as an invaluable tool to explore the multilevel connections between the objects that were defined from the news.

A hypergraph is a generalization of a graph in which an edge can join any number of vertices. In contrast, in an ordinary graph, where an edge connects exactly two vertices. A hypergraph can be represented by the incidence matrix $H = (V, E)$. In this structure, $V = \{v_i\}_{i=1}^n$ is a set of vertices, and $E = \{e_j\}_{j=1}^m$ indicates a family of hyperedges where each e_j is a subset of V [14]. The set $\mathcal{I}$, which comprises all individual items from the transactions, provides the vertices of the hypergraph. Specifically, the vertices in the hypergraph H are given by the unique elements of $\mathcal{I}$ that are included in the frequent itemsets $\mathcal{F}$. A hyperedge e_j is formed by a set $F \in \mathcal{F}$, encompassing vertices that share a commonality in the form of co-occurrence within the news items. This co-occurrence is quantified by the support count, adhering to the minimum support threshold. The incidence matrix H defines the relation between vertices and hyperedges, with $H(v_i, e_j) = 1$ if the item v_i is part of the itemset e_j , and $H(v_i, e_j) = 0$ otherwise.

An example incidence matrix of hypergraph H that contains 4 vertices and 3 hyperedges is shown in Table 2 and visualized in Figure 4.

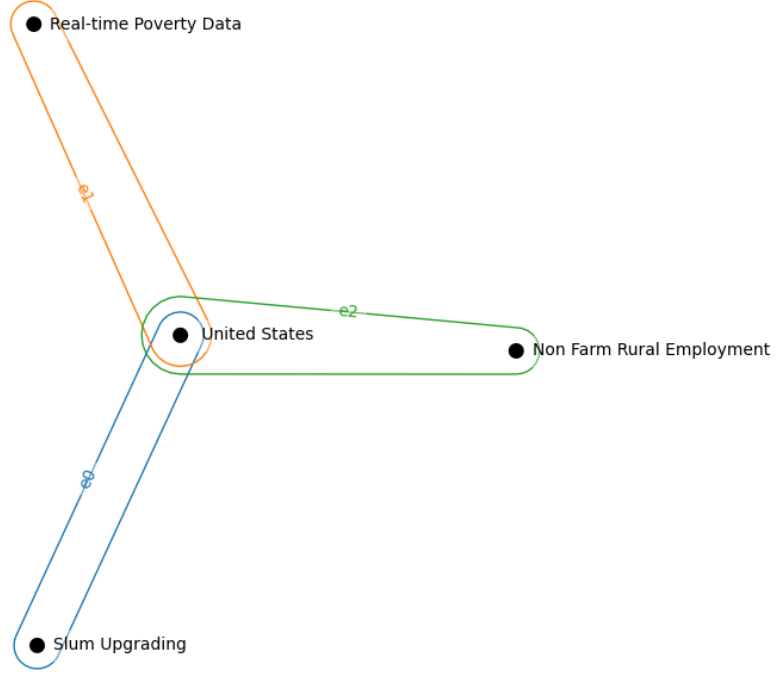


Figure 3: Hypergraph constructed from the frequent patterns found between SDG 1 themed news

Vertex	Label		v_1	v_2	v_3	v_4
v_1	Real-time Poverty Data	$H =$	e_1	e_2	e_3	$\begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$
v_2	Non Farm Rural Employment					
v_3	Slum Upgrading					
v_4	United States of America					

Table 2: Example incidence matrix of hypergraph H .

Weighted hypergraphs extend the standard hypergraph model by incorporating weights into each hyperedge. In this model, represented as $H=(V,E,W)$

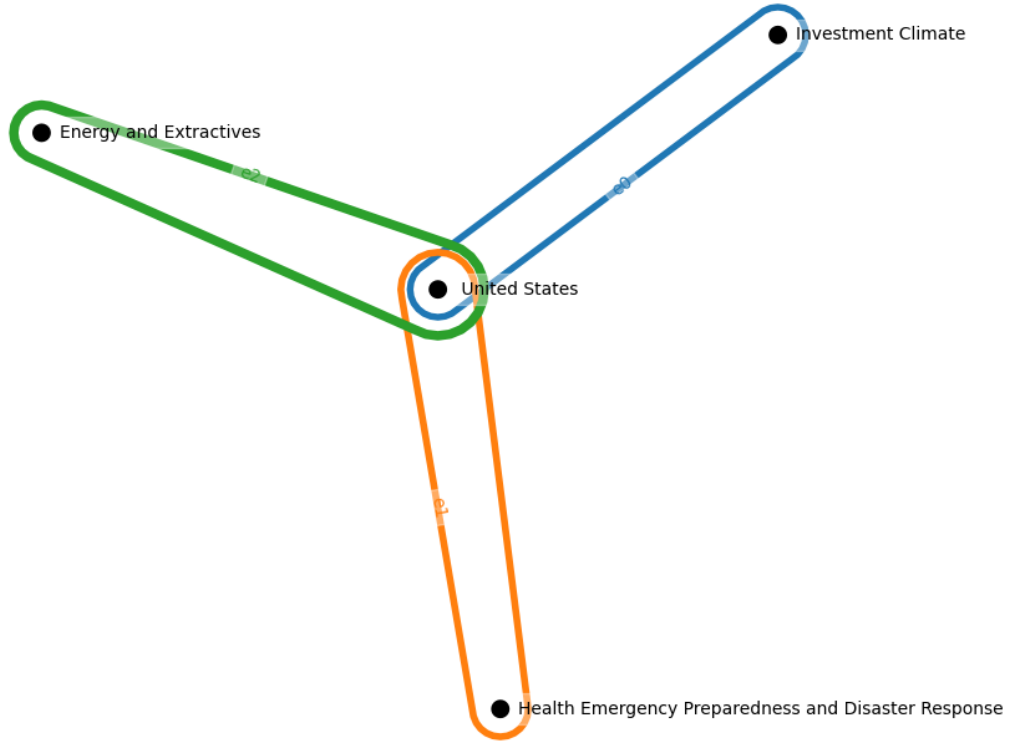


Figure 4: Weighted hypergraph constructed from the frequent patterns found between SDG 13 themed news. The weights correspond to the support of the frequent itemsets, and are visually represented as the edge width

where W is weight, the weight of each hyperedge can signify different attributes, such as the strength or frequency of the connections it represents. These weights can also represent the support count of each itemset. This approach not only reveals the relationships between items (vertices) but also quantifies the importance or frequency of these relationships.

Vertex	Label	
v_1	Investment Climate	
v_2	Health Emergency Preparedness and Disaster Response	
v_3	Energy and Extractives	
v_4	United States	

$$H = \begin{matrix} & v_1 & v_2 & v_3 & v_4 \\ \begin{matrix} e_1 \\ e_2 \\ e_3 \end{matrix} & \begin{pmatrix} 0.12 & 0 & 0 & 0.12 \\ 0 & 0.13 & 0 & 0.13 \\ 0 & 0 & 0.18 & 0.18 \end{pmatrix} \end{matrix}$$

Table 3: Example incidence matrix of hypergraph H with corresponding weights.

The hypergraph structure enables us to depict not only pairwise relationships but also complex multilateral associations among the news elements. Such a representation aids in discovering the underlying structure of the data, making enhanced analytical procedures like clustering, anomaly detection, and

trend analysis possible.

Within this structure, the roles of individual nodes and hyperedges become important in understanding the dynamics of the network. Centrality measures such as s-Betweenness and s-Closeness offer valuable insights into the influence and accessibility of elements within the hypergraph.

s-Betweenness centrality reflects the extent to which a node lies on the shortest paths between other nodes, showing how determinative (important) is the element for the hypergraph. This centrality measure captures the influence of a hyperedge in connecting pairs of other hyperedges through the shortest s-walks. It is computed using the formula:

$$BC_s(e) := \sum_{f \neq e \neq g \in E} \frac{\sigma_{fg}^s(e)}{\sigma_{fg}^s} \quad (5)$$

where σ_{fg}^s denotes the number of shortest s-walks between edge f and edge g , and $\sigma_{fg}^s(e)$ denotes those shortest s-walks that include edge e . On Figure 5 a hypergraph can be seen, with the s-Betweenness centrality influencing the node radius, while edge width is influenced by the support of the frequent pattern. For example on Figure 5 the s-Betweenness centrality for the 'United States' node is 0.735, that indicates its importance of linking other items within the dataset.

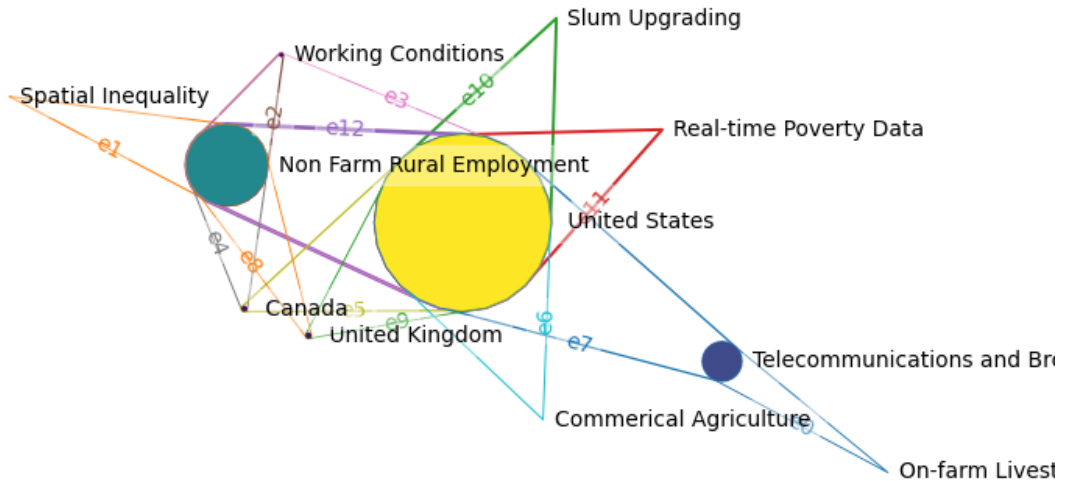


Figure 5: Hypergraph with betweenness centrality influencing the node radius

s-Closeness centrality assesses the degree to which a node is near all other nodes in the network showing how much the element affects other elements. This measure is adapted to reflect how close a hyperedge is to all other hyperedges, as follows:

$$CC_s(e) = \frac{|E_s| - 1}{\sum_{f \in E_s} d(e, f)}. \quad (6)$$

,where E_s is the set of edges such that the cardinality of each edge is at least s , and $d(e, f)$ represents the distance between hyperedges e and f . On Figure 6 a hypergraph can be seen, with the s-closeness centrality influencing the node radius, while edge width is influenced by the support of the frequent pattern. For example on Figure 6 the s-Betweenness centrality for the 'United States' node is 0.489, which indicates its dominant influence on co-occurrence with other elements in the network.

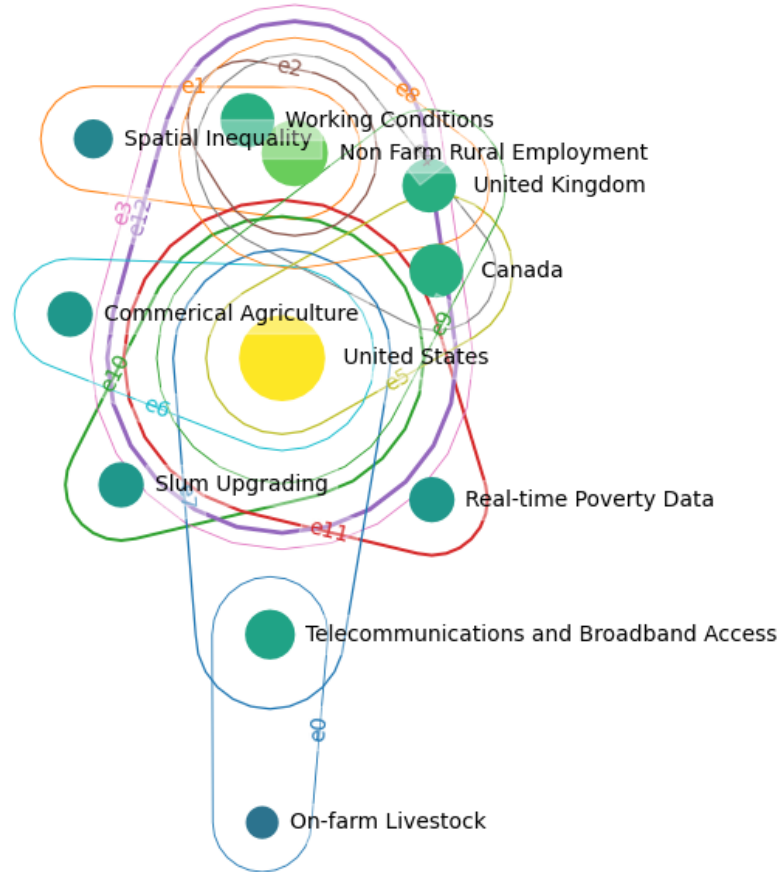


Figure 6: Hypergraph with closeness centrality influencing the node radius

Degree centrality in hypergraphs assesses the significance of a node based on the number of connections it has. In the context of hypergraphs, this measure reflects the extent to which a hyperedge is connected to other hyperedges, highlighting its prominence in the network. If v_i and v_j is a vertex, and $e_k \in E$ is a hyperedge, and w_k is the weight of hyperedge, then the weighted vertex degree centrality for hypergraph to incorporate hyperedge weights [15] is calculated as:

$$DC(v_i) = \sum_{v_i \neq v_j \in V} \sum_{\substack{k=1, \\ \{v_i, v_j\} \subset e_k}}^m w_k \quad (7)$$

2.4 Summary of the methodology

The proposed methodology effectively transforms raw data into informative hypergraphs and insightful association rules through a series of structured steps:

1. **Data acquisition, processing and filtering:** The process begins with the downloading of raw data from GDELT, consisting of news articles. The acquired data is subjected to processing and filtering. This step ensures that only relevant information is retained for analysis.
2. **Transaction formation:** Significant objects of interest are extracted from the filtered news content to form transactions. Each transaction represents a collection of these objects, similar to records in traditional data analysis.
3. **Frequent pattern mining:** FPClose is used to identify recurring patterns, a frequent pattern mining algorithm. This technique helps in uncovering common itemsets that signify trends or patterns in the data.
4. **Association rule mining:** FPGrowth algorithm is used to generate association rules. These rules provide further insights into the relationships between different objects in the data, revealing conditional dependencies.
5. **Hypergraph construction:** Utilizing the frequent itemsets, a hypergraph is constructed. In this representation, vertices correspond to unique

items, while hyperedges are formed based on the identified itemsets, offering a graphical interpretation of the data relationships. Weights can also be assigned based on the support value of the frequent itemsets.

6. **Parameter adjustment:** Each step involves careful adjustment of parameters to optimize the performance of the mining algorithms and to ensure the relevance and quality of the results.

This methodology of the study closely follows the Knowledge Discovery in Databases (KDD) process a standard approach in data mining, which can be seen on Figure 7. The KDD process includes the steps of data selection, pre-processing, transformation, mining, and interpretation. The proposed method, starting from gathering data to creating hypergraphs and finding association rules, fits well with these KDD steps.

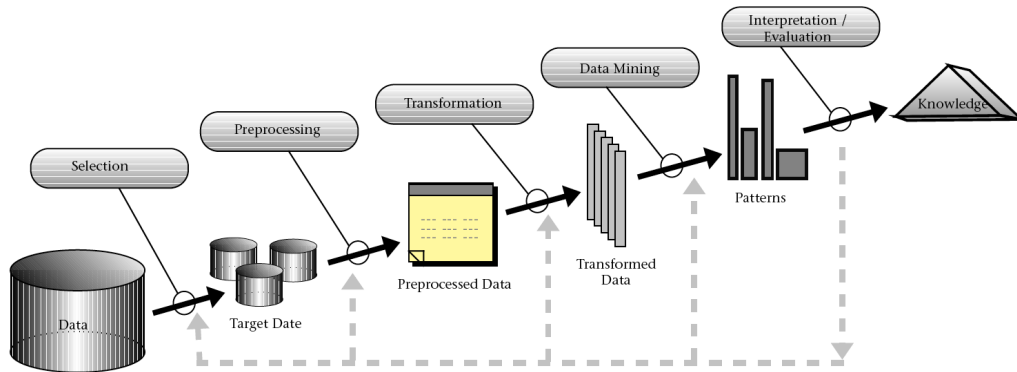


Figure 7: The Knowledge Discovery in Databases (KDD) Process [16]

3 Implementation of the proposed methods and the interactive web portal

The data acquisition and preprocessing for this project involved utilizing a Python script and Dask [17] for efficient memory management and data handling. It allows continuous data acquisition, setting a foundation for the following stages of data analysis and interpretation. In this project, I used a 24-hour interval due to processing power limitations however, a more extensive interval could be set to acquire a more comprehensive analysis.

To start, the script deletes old data files, ensuring the workspace remains manageable and only retains the most relevant and recent files. The script then proceeds to download 15-minute interval data files within a specified duration from the official website of GDELT, ensuring access to the most up-to-date data for analysis. The downloaded files, in a zipped format, are then extracted and loaded into Dask DataFrames, offering a Pandas-like interface while enabling operations on larger-than-memory datasets. The records undergo a filtering process where only the entries with non-null locations and theme values are retained. Furthermore, the themes in each record are screened against the predefined list to ensure relevance and focus in the context of the study.

Based on the dataset obtained, I have developed a web platform using Plotly Dash [18] that visualizes complex interconnections between news topics, sustainability-related World Bank themes, and SDGs, specifically focusing on SDG 13 which targets climate action. The construction of the visual graphs was carried out employing the methodologies outlined in Ref. [6].

The foundation of the platform is a PostgreSQL database, which is responsible for storing and managing all of the data efficiently. To ensure smooth operation and easy deployment, the web platform operates within a Docker container. This setup is hosted on Google Cloud Run for now, providing a scalable and secure environment for the application. This combination of technologies ensures both reliability and ease of maintenance for the platform.

The data mining process utilizes the SPMF Java library, accessed through a Python interface for integration and execution.

4 Application study on climate change-related news analysis

The presented methods are applied for analyzing 24-hour sustainable development goals-related news. For data downloading I utilized the climate change-related keywords based on the World Bank taxonomy identified in Ref. [6]. Table 4 includes the data specifications.

Table 4: News data analysis overview (2023-11-02)

Metric	Count
Total number of news items (SDG-related)	96,663
Total number of objects (columns)	1,363
Total number of regions	243
Total number of topics	1,120
News related to SDG 13	59,131

4.1 Interactive web portal

The developed web portal offers an interactive view of the intertwining of world news by providing different graphs, in order to analyze the graphs such as:

- **Theme-based network graphs:**

Theme-based network graphs present networks where nodes represent news articles aligned with World Bank themes related to SDG 13. Edges in the graph are colored and weighted to signify the overall sentiment or tone of the connected articles, and the frequency of their occurrence respectively. This visual aid serves to identify how the discourse on climate action is shaped by the tone of the news, providing insights into global narratives. An example theme-based network graph visualization is shown in Figure 8.

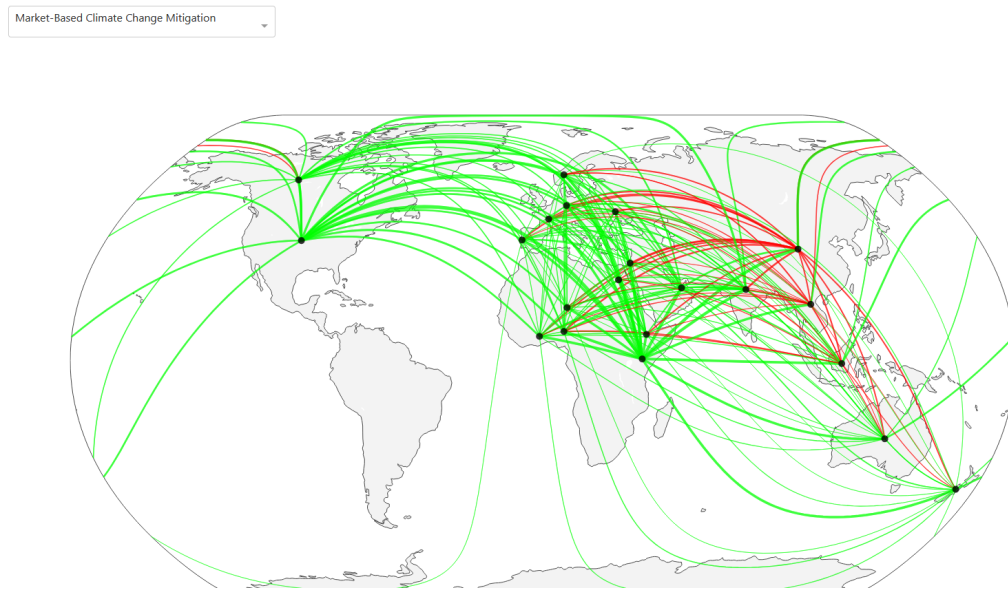


Figure 9: Graph constructed from the data showing the connections between countries based on a selected "Market-based climate mitigation" theme.

- **Choropleth map:**

The choropleth map illustrates the average sentiment of each country on selected themes, offering a geographical perspective on the global sentiment. It is interactive and theme-adjustable. To demonstrate the capabilities of the platform, images of the graphs and choropleth map for specific themes have been provided. An example choropleth map visualization is shown in Figure 10.

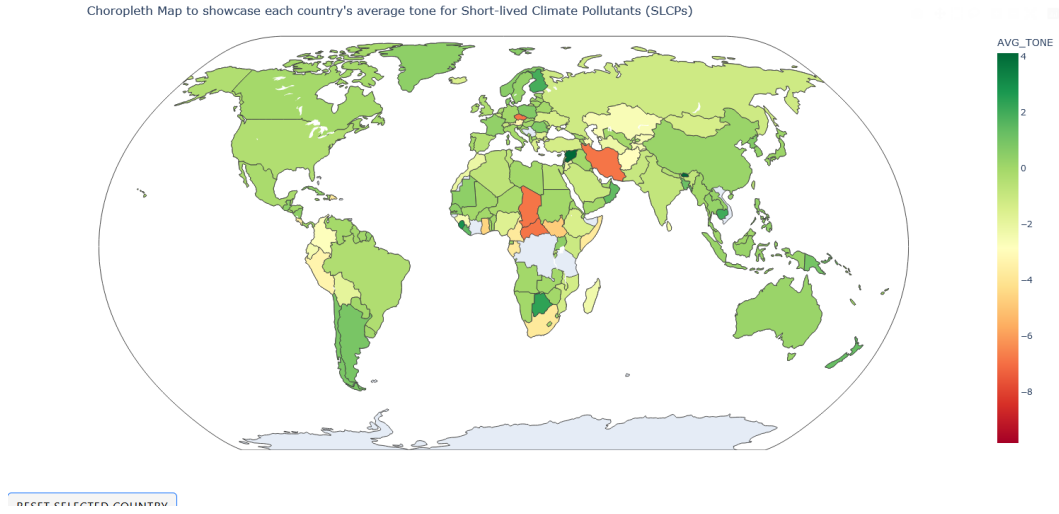


Figure 10: Choropleth map showing the average tone of countries towards the "Short-lived climate pollutants" theme.

4.2 Hypergraph-based analysis of frequent patterns found in world news regarding the SDGs

To showcase the effectiveness of the method presented multiple hypergraphs were constructed from the news using different subsets of the full transaction database. The hypergraph presented in Figure 11 illustrates the web of regional interconnections found in 24 hours of news coverage. Nodes, representing different countries, vary in size indicating the degree centrality within the network, with the United States in a central position, denoting its frequent co-occurrence with a multitude of other regions. Edges, marked with varying thickness, symbolize the strength of association between regions, corresponding to the support of frequent patterns in the news data. The graph shows how international news stories from the given time period focus on the United States and the Middle East, especially Egypt, Israel, and the Gaza Strip, highlighting these areas' importance in current geopolitical media coverage.

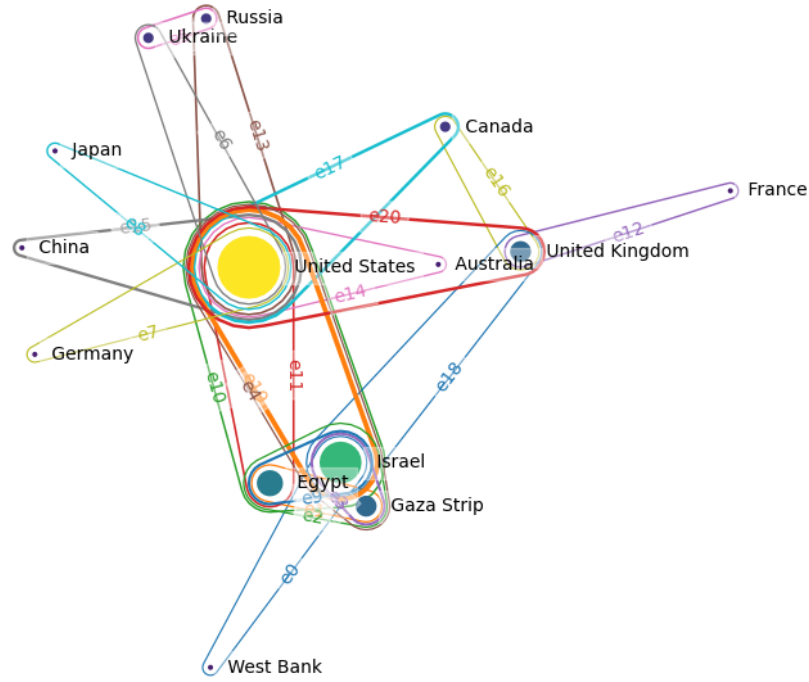


Figure 11: Hypergraph constructed from the frequent patterns found between the regions with degree centrality influencing the node radius and support influencing the edge widths

The method was also tested on three specific countries, namely the United States of America, United Kingdom, and the Republic of Kenya, in order to uncover patterns and gather new insights.

The hypergraph depicted in Figure 12 illustrates the most important themes in news narratives involving the United States for the given time period. It also highlights certain pairs of themes that frequently appear together. Notably, pairs such as 'Justice' and 'Public Sector Management', 'Fragility, Conflict and Violence' alongside 'Conflict and Violence', and 'Information and Communication Technologies' in conjunction with 'Digital Government' are recurrent. The co-occurrence of 'Fragility, Conflict and Violence' with 'Conflict and Violence' is a natural pairing, given the subset relationship of these themes. In contrast, the combinations of 'Justice' and 'Public Sector Management', and 'Information and Communication Technologies' with 'Digital Government' suggest more complex associations. These are not straightforward antecedents but

rather indicative of intertwined narratives, possibly reflecting extensive media coverage of specific events or legislative actions.

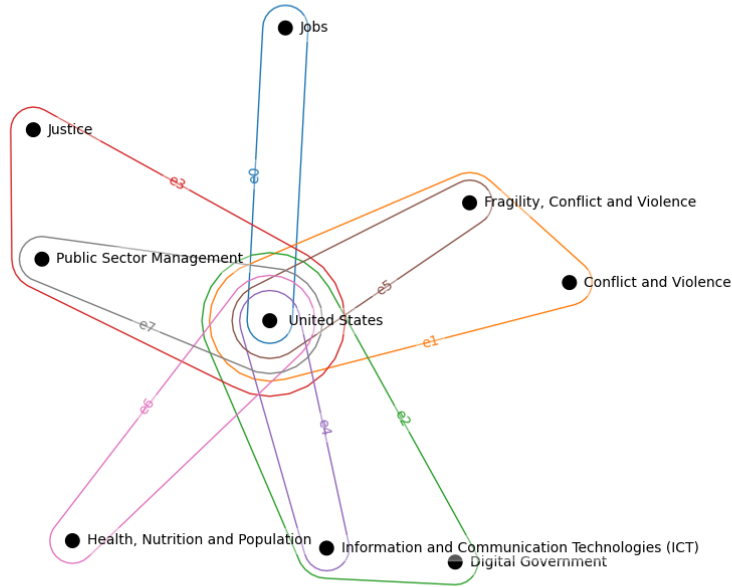


Figure 12: Hypergraph constructed from the frequent patterns mined from the news regarding the USA

The hypergraph in Figure 13 shows the key themes in news narratives about the United Kingdom for the same time period. Like the United States, the UK has similar recurring theme pairs in its news coverage. These include 'Fragility, Conflict and Violence' with 'Conflict and Violence', and 'Information and Communication Technologies (ICT)' together with 'Digital Government'. Another interesting observation is the frequent mention of the United States alongside the United Kingdom, indicating a close tie in global news narratives. This similarity in the themes of the UK and US reflects shared areas of focus in international media coverage.

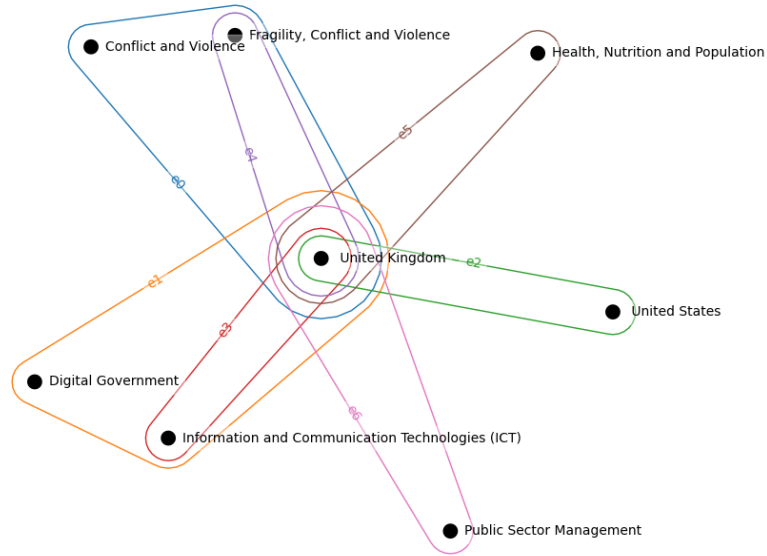


Figure 13: Hypergraph constructed from the frequent patterns mined from the news regarding the UK

The hypergraph in Figure 14 shows the main themes in news stories about the Republic of Kenya. Like the US and UK, the news involving Kenya often mentions 'Fragility, Conflict and Violence' along with 'Conflict and Violence'. Also, news stories that talk about Kenya frequently also mention the United States and the United Kingdom. A big difference in the news involving Kenya is the focus on 'Education'. This includes all kinds of education topics, from making sure everyone can go to school ('Education for All') to teaching skills that help people get jobs ('Education, Skills Development and Labor Market'). This focus shows that education is a very important topic in Kenya. It reflects the commitment of the country to improving educational standards and accessibility, as well as the recognition of education as a key driver for economic and social development.

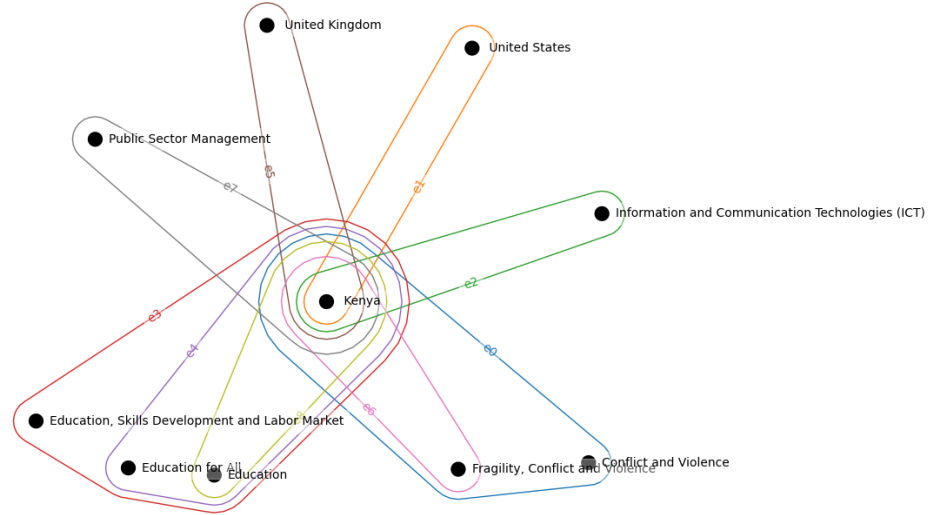


Figure 14: Hypergraph constructed from the frequent patterns mined from the news regarding the Republic of Kenya

One important similarity in the hypergraphs created from news data about a specific country is that the country itself is always at the center and is always the most significant element. This is expected, but to gain a deeper insight into the themes and how they are connected in relation to a country, a different approach can be tried. By initially focusing on news related to a particular country, and then intentionally removing that country from the data, a clearer view can be obtained of how various themes are interlinked. Figure 15 illustrates this idea using the United States as an example. It shows the connections between different themes in news stories about the USA, but with the USA itself removed. This helps us see how these themes are intertwined and related to each other in the context of news about the USA. here the closeness centrality gives us more information about the connections:

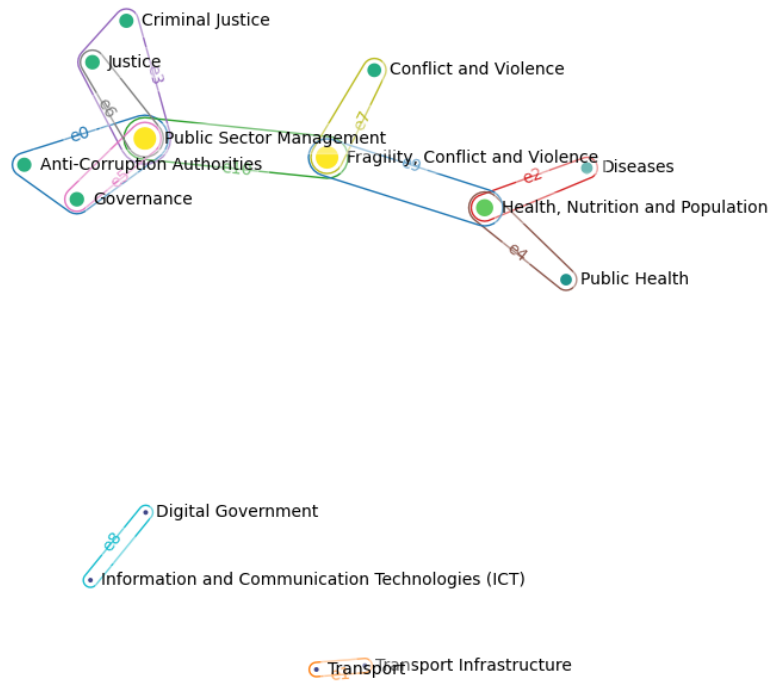


Figure 15: Hypergraph constructed from the frequent patterns mined from the news regarding the USA, with closeness centrality influencing the node radius.

In Figure 15, the hypergraph displays news themes linked to the USA, with the USA itself removed to better show how these themes are connected. The graph uses closeness centrality, shown by the size and color of the nodes, to show how important each theme is. Themes like 'Public Sector Management' and 'Fragility, Conflict and Violence' have a high closeness centrality of 0.264, which means they're more connected to other themes in the news. 'Public Sector Management' is often found with themes like 'Governance', 'Justice', and 'Criminal Justice', showing its role in various governance-related news stories. Also, the combination of 'Health, Nutrition and Population' with 'Diseases' and 'Public Health' points to a strong focus on health in the news. Besides, the graph highlights pairs like 'Transport' with 'Transport Infrastructure', and

'Information and Communication Technologies (ICT)' with 'Digital Government'. These pairs show themes that are closely related in meaning.

4.3 Association rule mining

Table 5 presents association rules derived from the analysis, with the support values indicating the count of occurrences as provided by SPMF.

The association rules indicate clear patterns in how topics related to SDGs are interconnected in media coverage. For instance, a strong and frequent connection exists between governance, justice, and anti-corruption themes, especially related to the United States, which often tie in with discussions about public sector management and anti-corruption investigations, as shown in rules r1 to r3. Similarly, health-related issues are frequently linked to pharmaceuticals and emergency response, as shown in rules r4 and r6, highlighting a media emphasis on the impacts and management of health crises. While these insights are based on a 24-hour news dataset and might not capture more complex patterns, the methodology proves effective and suitable for analyzing such associations in news reporting.

Table 5: Association rules mined from the news data with support and confidence

Rule ID	Antecedent	Consequent	Support	Confidence
r1	United States, Governance, Justice, Anti-Corruption Authorities	Public Sector Management, Criminal Justice, Anti-Corruption Investigation	5338	0.97
r2	Governance, Justice, Criminal Justice, Anti-Corruption Authorities	Public Sector Management, Anti-Corruption Investigation	8975	0.98
r3	Governance, Public Sector Management, Justice, Anti-Corruption Authorities	Criminal Justice	9171	0.99
r4	Health, Nutrition and Population, Health Technologies, Fragility, Conflict and Violence, Conflict and Violence	Pharmaceuticals	6397	0.67
r5	Governance, Public Sector Management, Criminal Justice	United States, Justice	5819	0.60
r6	Health, Nutrition and Population, Public Health, Health Emergencies	Health Emergency Preparedness and Disaster Response	6028	0.77
r7	Public Sector Management, Peace Operations and Conflict Management, Fragility, Conflict and Violence	Justice	5213	0.75

5 Conclusion

The presented method is a robust tool for dissecting complex news patterns, trends, and the interplay of various news domains, which in turn enables support for informed decision-making. Through the application of frequent pattern mining, commonly occurring themes and countries have been successfully discovered. Furthermore, association rule mining has been instrumental in uncovering relationships between these domains, thereby improving the comprehension of the intricate connections between sustainability issues and news domains.

One of the key highlights of the proposed techniques is the visualization of this complex information through hypergraphs. These hypergraphs not only provide clarity in representing group interactions but also allow for an in-depth exploration of the relationships between diverse news domains. The application of centrality metrics further adds value by pinpointing and analyzing key components within the network. This analysis sheds light on the most influential factors in sustainability-related news, thereby augmenting the effectiveness of decision-making processes and strategy development.

In parallel, the study introduces an innovative web platform that offers access to sustainability-related world news analysis. This platform is a gateway for researchers and individuals to engage with near-real-time world news, facilitating a data-driven approach to understanding global sustainability trends and events.

The application study showcases the effectiveness of the proposed methodology in analyzing news data, leading to the extraction of not only interesting results but also meaningful knowledge, thereby demonstrating the practical utility and success of these developed techniques.

This study opens the door to many exciting future projects. For example, hypergraph analysis could be integrated into the developed web portal, making the data more interactive and accessible. Exploring the use of machine learning in hypergraph networks is particularly intriguing, as it's a relatively new and uncharted area of research, offering the potential for groundbreaking developments in data analysis methods. The foundation for a comprehensive targeted analysis is already established as illustrated in the application study.

Further expansion and integration of news tone analysis could significantly enhance the depth of examination for any chosen organization, region, or subject of study. This addition would not only enrich the understanding of prevailing themes but also provide a detailed perspective on the sentiment portrayed in the media. Such an enriched analysis could offer a more complete and nuanced view, crucial for in-depth understanding and interpretation.

References

- [1] James H Faghmous and Vipin Kumar. A big data guide to understanding climate change: The case for theory-guided data science. *Big data*, 2(3):155–163, 2014.
- [2] Viktor Sebestyén, Tímea Czvetkó, and János Abonyi. The applicability of big data in climate change research: The importance of system of systems thinking. *Frontiers in Environmental Science*, 9:70, 2021.
- [3] Ralf Barkemeyer, Frank Figge, and Diane Holt. Sustainability-related media coverage and socioeconomic development: A regional and north-south perspective. *Environment and Planning C: Government and Policy*, 31(4):716–740, 2013.
- [4] GDELT. The gdelt project - global database of events, language and tone. <https://www.gdeltproject.org/> **last accessed:** [2023-10-31].
- [5] Xiaohui Lu, Faith Ka Shun Chan, Nan Li, Chuke Chen, Wei-Qiang Chen, and Hing Kai Chan. Improving urban flood resilience via gdelt gkg analyses in china’s sponge cities. *Scientific Reports*, 12(1):20317, 2022.
- [6] Tímea Czvetko, Gergely Honti, Viktor Sebestyen, and Janos Abonyi. The intertwining of world news with sustainable development goals: An effective monitoring tool. *Heliyon*, 7(2), 2021.
- [7] A Garaeva, F Makhmutova, I Anikin, and Kai-Uwe Sattler. A framework for co-location patterns mining in big spatial data. In *2017 XX IEEE International Conference on Soft Computing and Measurements (SCM)*, pages 477–480. IEEE, 2017.
- [8] Yunshan Ma, Chencheng Ye, Zijian Wu, Xiang Wang, Yixin Cao, and Tat-Seng Chua. Context-aware event forecasting via graph disentanglement. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1643–1652, 2023.
- [9] GDELT. Querying, analyzing and downloading. <https://www.gdeltproject.org/data.html> **last accessed:** [2023-10-31].

- [10] GDELT. World bank group topical taxonomy now in gkg. <https://blog.gdeltproject.org/world-bank-group-topical-taxonomy-now-in-gkg/> last accessed: [2023-10-31].
- [11] Toon Calders. Introduction to pattern mining. *Business Intelligence: Third European Summer School, eBISS 2013, Dagstuhl Castle, Germany, July 7-12, 2013, Tutorial Lectures 3*, pages 1–32, 2014.
- [12] Mohammed Alghobiri. Exploring the attributes of influential users in social networks using association rule mining. *Social Network Analysis and Mining*, 13(1):1–13, 2023.
- [13] Rumeng Chen, Feng Hu, Feng Wang, and Libing Bai. Hypergraph-clustering method based on an improved apriori algorithm. *Applied Sciences*, 13(19):10577, 2023.
- [14] Alain Bretto. Hypergraph theory. *An introduction. Mathematical Engineering. Cham: Springer*, 1, 2013.
- [15] Komal Kapoor, Dhruv Sharma, and Jaideep Srivastava. Weighted node degree centrality for hypergraphs. In *2013 IEEE 2nd network science workshop (NSW)*, pages 152–155. IEEE, 2013.
- [16] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37, 1996.
- [17] Matthew Rocklin. Dask: Parallel computation with blocked algorithms and task scheduling. In *Proceedings of the 14th python in science conference*, number 130-136. Citeseer, 2015.
- [18] Plotly Technologies Inc. Collaborative data science. <https://plot.ly>, 2015.