Multimodal Data Knowledge Representation for Situational Awareness: Domain-agnostic Approach With Evolving Knowledge Graphs

Introduction

Streaming data comes in different forms (video, text, image and other modalities); usually each data modality is represented in its own unique way which permits to analyze and extract patterns and knowledge from these modalities separately but does not allow to discover latent interconnections *across* the modalities and between their entities and knowledge objects. In order to capture semantic similarities between the multimodal data, as well as to predict new connections between the knowledge objects already available in store (historically accumulated knowledge and data at rest) and the new data that is streaming in (real-time data), we propose a framework which extends ontological knowledge graph with temporal and spatial relationships, i.e. a knowledge representation with axes that describe time and location on top of semantical connection between entities.

In this work, the term Knowledge Graph stands for a general concept of a graph-like structure that models the semantics between nodes. Other popular synonyms in literature include RDF, Ontologies, Knowledge Bases, Semantic Web and Linked Open Data. When properly designed and amply filled, KGs are capable to heterogeneously combine data sources across diverse domains and reason on highly complex relational data. Thus, KGs are a popular representation of the knowledge and they facilitate inter-connectivity of unimodal datasets by finding latent connections and discovering hidden links between them (e.g. DBPedia, Wikidata).

Since KG can be designed in different ways, there is no one size fits all structure. Depending on the domain and the purpose of the KG, it can be modified and expanded with additional artifices, such as second-order entities, events, locations, and so on.

Human domain expert involvement is crucial during the design phase since it is important to extract the meaningful entities and relations. In case KG are incorporated into a recommendation system and suggest knowledge that the user is interested in, it is needed to determine key user roles with their initial queries in order to resolve the cold start problem (problem, where the system is not able to recommend items to users because nothing has been recommended and no items were requested by anybody yet). In the context of search and recommender systems, knowledge graphs extended with temporal axis can be used to predict new suggestions that the user is interested in based on his previous searches.

Knowledge Representation for Multimodal Data

KGs in their basic form can be modeled as triples in the form <Subject, Predicate, Object> where two entities (subject and object, or subject and attribute, sometimes also called literal) are in a relationship described by a semantic predicate (e.g. Donald Trump, presidentOF, USA).

In a typical approach to knowledge graph construction, the KG is static, which means that it can grow more nodes and links as new data becomes available but it is ultimately considered time and place invariant. Sometimes, different time periods related to the entity and/or related locations are added as separate attributes of that entity, but this structure makes the graph bulky and connection discovery problematical.

To represent such knowledge graph in the mathematical model that allows fast computations and finding dependencies, it is transformed into a tensor, which is essentially a generalization of matrix into more than two dimensions (tensor orders). Given that KG statements consist of three elements, we can use a third-order tensor to map them: two orders for entities and another one for relationship encoding. The intersection of the three axis is a point for a specific node in the KG.

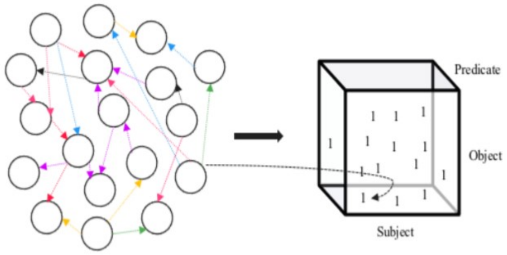


Figure 1. KG as a Tensor [1]

It is easy to see that when building a KG especially with streaming and real-time data, the triples may not be time-invariant. In the example above, for the time frame 2008-2016, the actual triple would be “Obama, presidentOf, USA”. Thus, knowledge representation can be encoded as a triple (Object, Relation, Object) in time and space: this is the reason we extend the knowledge graph with additional temporal and spatial axis to encode the timestamp and location (both or just one of them; depending on what is applicable for a particular triple) of the event. So, the first three dimensions encode objects and the relationship between them and the additional dimensions encode time and space.

For example, if KG contains Person and Movie entities with relationships between them (e.g. “likes”, “watched”, “directed”, “starred”, “produced”, etc), we can encode a record “John likes Jurassic Park” in the 3-dimensional space where one axis will list objects Person class, one axis will list objects in Movie class and the third axis will encode all possible relationship between 1st and 2nd entities. Thus, in Figure 2, “John likes Jurassic_Park” will be represented by a point s_2 where coordinates for entity axes are “John” and “Jurassic Park” and the coordinate on the relationship axis corresponds to “like”.

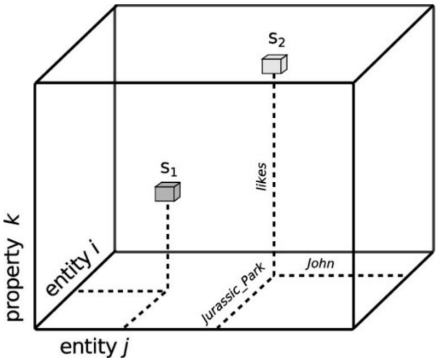


Figure 2. Representing Knowledge Graph triples in a 3-dimensional tensor [2]

However, for knowledge that can evolve in time and space, 3rd order tensor representation is not enough. For this reason, we extend knowledge space with two additional axes for time and location (Figure 3).

Without loss of generality, we assume that new information arrives in the form of events: for a given time step on the temporal axis, the event creates a link between specific nodes from the semantical axis with a point in time and space.

Thus, the incoming event can be described as a dot in 5-dimensional space where the first dimensions are reserved for representing the semantics of the event, e.g. the knowledge graph triple (Subject, Relation, Object) and other dimensions describe time and space.

Consequently, the event is encoded in the form of a quintuple with values (subject, relation, object, time, location).

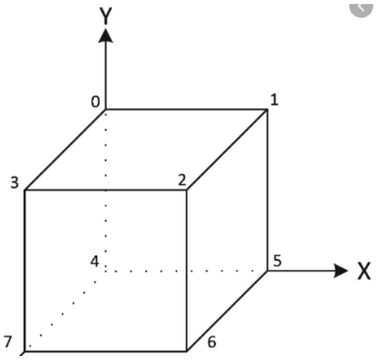


Figure 3. Additional temporal and spatial axes for 3rd order tensor with KG triples results in quintuples that encode objects, relationship between them, time and location. While 5-dimensional object is hard to picture on the plot, it is useful to think about it as a collection of points where each point has 5 coordinates and can be mapped to 5 axes.

Note that this quintuple can contain information extracted from *any* modality that we have access to thus making this approach domain-agnostic. The value of the point can be binary (0/1 or true/false) when it identifies the existence of the relationship between entities in this point in time or a real value when it reflects the measurement for the certain relationship (e.g. distance between objects captures with LIDAR) in specific time and location.

Therefore, an additional perquisite of our approach is its capability to evolve over time as well as to expand with additional data modalities. Adding a new modality is reduced to adding a set of points to a 5-dimensional space, where the first three coordinates describe the semantics, while the additional two coordinates capture the spatial and temporal point of the measurement. For instance, in case of a sensor data it can be the temperature or the distance between two objects (e.g. with LIDAR sensor). Thus, the subsegments of the multidimensional tensor representing a fused knowledge object from multiple modalities can be identified and extracted.

For example, from police dispatch reports the extracted event can be encoded in the form (“Saturn 1994”, “stolen from”, “PersonA”, March 31 2020, West Lafayette), from the video data (“Saturn 1994”, “parked by”, “Chipotle State Street”, April 1 2020, West Lafayette), additional quintuple can be extracted from the image data, etc.

Once these events are normalized in the 5-dimensional space with regard to their ontologies, time and location, it becomes possible to predict missed connections and future events.

In the example with the police dispatch report and video camera footage given above, we have an Object on the semantic axis and Location on the spatial axis that share the same value, “Saturn 1994” and West Lafayette respectively, therefore two quintuples that essentially describe different relations and nodes in the Knowledge graph may be connected to the same fused knowledge object: now it describes a stolen Saturn from PersonA parked by Chipotle.

The approach described above allows to order the incoming events with regard to their semantics, chronological and spatial attributes, and establish new connections between historical events contained in the graph and incoming events with link prediction algorithms.

Semantic connections are established if the entity nodes coincide or lie within certain vicinity. Semantic similarity can be calculated through similarity distance measurements, for example, cosine similarity.

Temporal connections are established if the time of the new event is known and it coincides with the time of other events or even if only certain recurring attributes of the timestamp coincide, such as day of the week or month. This can predict connections between repeated events, such as if PersonA is seen by the temple periodically, we can predict that this person will attend church next Sunday.

Spatial connections are established if the location/vicinity of the new event is known.

Once the events are connected through one or more axes, they become candidates for a fused knowledge object which describes one event or a combined scene with more details.

Discussion and Ongoing Work

The suggested approach for knowledge representation can be efficiently used for multimodal data where there is a combination of historical data at rest and incoming streaming data. Challenges in working with multimodal data for situational awareness are as follows:

1. Mismatch in different modalities. Data coming from different modalities may be mismatched which makes it difficult to see the latent relations across these modalities or it can be linked indirectly or not linked at all as opposed to the case when different modalities describe exactly the same problem. For example, in the dataset with twitter videos and the corresponding tweets captures, the same event is described both in video and user’s comment in text. However, in our use case, we may have video and text data that come from unrelated sources and it is part of the problem solution to identify the relationship between those sources.
2. The data comes in large quantities and quickly accumulates in giant graphs which can grow to have millions of nodes and connections. Methods for storage, growth and link prediction must be robust, reliable and cost-efficient.
3. Missing information. The data might not describe the scene in its entirety and in certain cases it is not possible to form fully developed quintuples that have all five coordinates.

Ways to approach the challenges listed above:

1. Accurate alignment of data from different modalities. Tensor with 5 dimensions that represents the semantic triple with its temporal and spatial coordinates for both text and video modalities will allow to align these modalities whenever it is possible. Entity alignment is essential when combining data from heterogeneous sources into one knowledge object.

2. Since knowledge graph quintuple can be encoded in a 5 order tensor, we can apply tensor decomposition and factorization methods to make pattern discovery and data prediction fast and robust. These data that comes from heterogeneous sources might be sparse, and we can use data mining methods developed for sparse matrices and tensors for computational efficiency.
3. Default value is allocated for missing coordinates. Similarity methods can be used for completion of the missing data.

Benefits of our approach:

1. Enables both contextual and temporal reasoning about real-world and/or domain-specific events and allows predictions based on the recorded data and history.
2. Additional modality can be added, as long as there are methods to extract a quintuple with meaningful information from that modality, such as timestamp, location, sensor values, meaning of the values, etc. Correlations with the modalities already in place can be found retrospectively.
3. Tensor representation allows to use a number of methods for link prediction, such as techniques developed for tensor completion. This information can be used for situational awareness and decision support.
4. The design allows for the system to be scalable and distributed across multiple sites if needed to decrease computational costs.
5. The design allows to extend the knowledge representation with reasoning mechanism and causality diagrams (separate diagrams for different domains created by experts) to infer not only connections but causes and reasons for the events. If conditional dependencies are identified, probabilistic model can be built to derive consequence of action and analyze causality of the events.

The principal dimensions for knowledge representation are narrowed down to 3 main axes and are recapped as follows:

1. Semantic axis (responsible for knowledge graph triples that connect object via a relationship between them): Contains topics and objects learned from text and videos in the form of triples (PersonA standing with Person B in the video, ObjectA stolen from Person B in the police dispatch report)

2. Temporal Axis: Timestamps of posts in social networks, timestamps of video record, timestamp of sensor measurements.

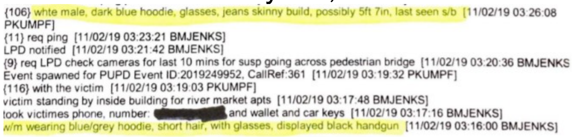
3. Spatial Axis: Location of a dashcam, geocoordinates of a signal device, identified geographical location of a user who made a post on twitter.

To summarize, each event or knowledge object which is derived from an accessible data source is analyzed for the presence of the entities with relationship between them, timestamp and geocoordinates. Based on these values a unique point for the event in the 5-dimensional space can be assigned.

Experiments and Demos (work in progress):

Real-life scenario with the data provided by WLPD:

Police provided us with the dispatch reports that describe the incidents; this data represent our text modality. We have obtained access to several public surveillance cameras located on the streets of West Lafayette; this data stands for our text modality.



```
{106} whte male, dark blue hoodie, glasses, jeans skinny build, possibly 5ft 7in, last seen s/b  [11/02/19 03:26:08
PKUMPF]
{11} req ping  [11/02/19 03:23:21 BMJENKS]
LPD notified  [11/02/19 03:21:42 BMJENKS]
{9} req LPD check cameras for last 10 mins for susp going across pedestrian bridge  [11/02/19 03:20:36 BMJENKS
Event spawned for PUPD Event ID:2019249952, CallRef:361  [11/02/19 03:19:32 PKUMPF]
{116} with the victim  [11/02/19 03:19:03 PKUMPF]
victim standing by inside building for river market apts  [11/02/19 03:17:48 BMJENKS]
took victimes phone, number: [redacted] and wallet and car keys  [11/02/19 03:17:16 BMJENKS]
w/m wearing blue/grey hoodie, short hair, with glasses, displayed black handgun  [11/02/19 03:16:00 BMJENKS]
```

Figure 4. Excerpt from Police Event Report describing a robbery. Key features are highlighted.

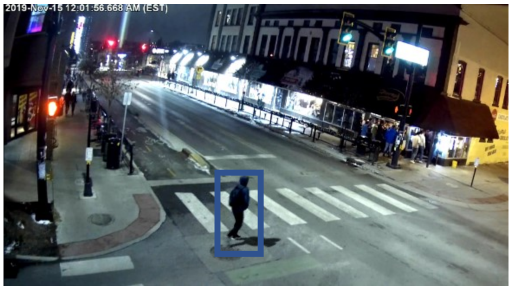


Figure 5. Surveillance camera footage. Detected objects: person. Extracted features: gender and color of jacket. Timestamp and location are available with camera metadata.

Graphical schema for representation of the text and video data knowledge object has the following format:

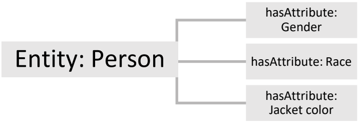


Figure 6. Schema to describe detected knowledge objects with attributes. These objects with detected features can be stored as triples in the 5-dimensional space. When combined with timestamp and geolocation, the triple is transformed into quintuple.

The schema is grown depending on the entities and their attributes that come from the multimodal data sources. E.g., entity "University" can be added and connected with the "Person" entity for college towns.

The protocol below describes in detail how the proposed approach allows to identify a connection between a filed incident and a video of the suspicious person taken by the camera with aligning objects along the semantic, temporal and spatial axes.

1. Police dispatch reports and video of West Lafayette are then processed for objects, features and relationship extraction:

    1.1 Text data extracted (with the help of named entities recognition methods), relationships extracted from the event report (Figure 4):
    Entity: "Person"
    Attributes: "White", "Male", "Blue jacket"

    1.2 Video object detection and recognition is performed (Figure 5):
    Entity: "Person", "Male"

2. Locating the extracted semantical triples on the temporal and spatial axis and populating the knowledge tensor with the extracted text and video data features on the semantical axes and timestamps and geographical coordinates on the temporal and spatial axes. The table below lists selected quintuples from a 5-dimensional tensor.

	Semantic axes values			Temporal axis values	Spatial axis values	
	Object	Relationship	Object/Attribute	Timestamp	Location	Value
Video data organized in 5-dimensional space	(Person,	hasGender,	Male,	T_1,	L_1)	2
	(Person,	hasGender,	Male,	T_1,	L_1)	2
	(Person,	hasJacket,	Blue,	T_1,	L_1)	1
	(Person,	hasRace,	White,	T_1,	L_1)	1
Text data organized in the same space	(Person,	hasGender,	Male,	T_2,	L_2)	1
	(Person,	hasJacket,	Blue,	T_2,	L_2)	1
	The first three coordinates that encode the semantics of the event coincide for data from both dimensions			The temporal and spatial coordinates are different but within the suggested threshold from each other		Denotes number of objects that share the same coordinates

Next steps and ongoing research:

3. Applying machine learning and pattern detection methods to find latent connections and make predictions (e.g. finding objects with the same attributes, time-based or geographical closeness in the 5-dimensional tensor)
4. Recording the user interests from his queries.
5. Streaming new multimodal data into the knowledge representation framework and expanding it with new quintuples.
6. Establishing the connections with the existing data and triggering notification to the user if the new data matches his interests for situational awareness.

References:

[1] Tay, Yi, et al. "Random semantic tensor ensemble for scalable knowledge graph link prediction." *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 2017

[2] Wilcke, Xander, Peter Bloem, and Victor De Boer. "The knowledge graph as the default data model for learning on heterogeneous knowledge." *Data Science* 1.1-2 (2017): 39-57.

[3] Palacios, Servio, K. M. A. Solaiman, Pelin Angin, Alina Nesen, Bharat Bhargava, Zachary Collins, Aaron Sipser, Michael Stonebraker, and James Macdonald. "WIP-SKOD: A Framework for Situational Knowledge on Demand." In Heterogeneous Data Management, Polystores, and Analytics for Healthcare, pp. 154-166. Springer, Cham, 2019.