International Journal of Engineering & Technology, 7 (3.33) (2018) 187-190

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



A novel property graph model for knowledge representation on the Web

Heekyung Moon[1], Zhanfang Zhao[2], Jintak Choi[3], Sungkook Han[1]*

[1]Department of Computer Engimeering, College of Engineering,Wonkwang University, 460 Iksandae-ro, Iksan, Korea
[2]Department of Computer Engimeering, College of Engineering, Hebei GEO University, Huai An Road No. 136, Shijiazhuang, Hebei Provincei, China
[3]Department of Computer Science & Engineering, College of Information Technology, Incheon National University, 119 Academy-ro, Yeonsu-gu, Incheon, Korea
**Corresponding author E-mail:skhan@wku.ac.kr*

Abstract

Graphs provide an effective way to represent information and knowledge of real world domains. Resource Description Framework (RDF) model and Labeled Property Graphs (LPG) model are dominant graph data models widely used in Linked Open Data (LOD) and NoSQL databases. Although these graph models have plentiful data modeling capabilities, they reveal some drawbacks to model the complicated structures. This paper proposes a new property graph model called a universal property graph (UPG) that can embrace the capability of both RDF and LPG. This paper explores the core features of UPG and their functions.

Keywords: *UPG, LPG, RDF, LOD, property graph, key-value pairs, knowledge graph.*

1. Introduction

Resource Description Framework (RDF) and Labeled Property Graphs (LPG) both provide ways to represent knowledge graphically. But they adopt different approaches to model knowledge representation. RDF is a standard data model for opening, sharing and interchanging data on the Web. RDF has been widely used as the core data model for Linked Open Data (LOD) that enables the Web of Data. Several comprehensive LOD products containing plentiful RDF datasets have emerged such as DBpedia and YAGO [1,2]. Besides, a large number of industry RDF datasets have been published by numerous researchers, institutions, and companies, which powerfully contribute to the knowledge sharing on the Web.

Recently, NoSQL databases are attracting increasing attention, since they address the limitation of Relational Databases (RDB) and provide more flexible and available data management for unstructured data [3-5]. Especially, graph databases based on LPG have received significant attention on account of the good performance in dealing with the complex relationships among the data [6]. LPG model owns distinctive features, using any number of key-value pairs to describe the properties of vertices and edges, which makes property graph more expressive and easy to understand for the human being. LPG model shows stronger expressiveness than RDF [7]. In addition, databases based on LPG provide good performance in query and storage for graph data.

The key-value pair properties of LPG model not only have a powerful capability to describe the intrinsic properties of the data objects but also can model complex relationships efficiently. However, the conventional LPG model shows lack of semantic expressiveness to realize semantic interoperability of the data in the open and shared environment of the Web. The data types used in LPG model are also restricted so that it cannot represent the diverse data structures. The core property structures of LPG need to be extended to capture both semantic and structural complexities of the data. This paper proposes a new property graph model called a universal property graph (UPG) that can embrace the capability of both RDF and LPG. This paper also explores the features of UPG model.

The rest of this paper is structured as follows. Section 2 reviews the related work. Section 3 presents the definition of UPG model and analyze its characteristics for data modeling. Section 4 demonstrates the applications of UPG model with the diverse use cases. Section 5 summarizes the contributions and puts forth the prospects for further work.

2. Related Work

Graphs provide an effective way to represent information and knowledge of real world domains. There are two dominant graph data models, RDF and LPG, popular for the applications of big datasets such as social networks and LOD. Although RDF is standard for Web-based knowledge modeling and provides the foundation of LOD, RDF modeling has been widely criticized for its awkward structures and semantic interpretations. Especially, the blank nodes and the reification provoke the serious difficulties in querying and searching [7-9]. Although RDF reification has been withdrawn from the normative sections in the latest RDF Recommendation [10], the expressive capabilities of RDF remain an unresolved problem [11].

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Recently, LPG model is very popular in an emerged NoSQL database paradigm that is non-relational, non-ACID, schema-less database systems to handle a huge amount of diverse data generated on the Internet. The conventional relational databases have been confronted with tough difficulties in the advent of Big Data, especially, query processing time drastically increases due to a number of complex JOIN-style operations [3]. LPG provides more compact, expressive representation of graph data modeling and efficiently store the key-value pairs with index-free adjacency that can allow for fast querying. In short, LPG is basically about storage and querying. However, LPG has also revealed some drawbacks although many enhanced functionalities have been proposed for LPG.

Since RDF and LPG are dominant graph modeling that has similar objectives, some studies have been accomplished to map from one model to another [10,12]. The conceptual comparisons of two models have also been studied [13,14]. The unified model that can harmonize the distinguishable features of two graph modeling approaches should be investigated to make graph modeling more powerful and practical.

3. Universal Property Graph Model for Knowledge Representation

Graphs are flexible and intuitive for modeling information resources, their relationships and the conceptual structure of their domain. In addition to the expressiveness, graphs can be stored efficiently and processed consistent with the well-known algorithms. This section explores the formal, conceptual properties of UPG model.

3.1. Definition of UPG model

The UPG model is based on LPG. However, it embraces the open-world features of RDF and the unified structures. Since UPG is a kind of the reshaped LPG, UPG provides more compact, expressive representation of graph data modeling and efficiently store the key-value pairs with index-free adjacency that can allow for fast querying. In addition, UPG can supersede RDF with more powerful expressiveness in the open world.

Definition (Key-Value Pair): A key-value pair (KVP) is an abstract data type consisting of a set of two linked data items: a key, which is a unique identifier for some item of data, and the value, which is literal, a set or an identifier of that data. The key plays a role of metadata of the data. The ontological vocabulary is usually used as the key. The value can be literal such as string, number, and date, a structured value such as array and list, or identifier of the key-value pairs. The key-value pair is also called the property in general.

Definition (PPI): The Public Property Identifiers (PPI) serves as unique public identifiers that can be used to identify any key-value structures in the open world. The concept of PPI is very similar to the Internationalized Resource Identifiers (IRIs) of RDF 1.1. However, the main aim of PPI is to identify the key-value structures, not Web resources. The PPIs are usually defined in the form of IRIs.

Definition (Universal Property Graph): A UPG G is a directed, labeled, attributed, multi-relational graph consisting of G = (V, E, K, P, L, □□□□□□□□), where V is a set of vertices, E is a set of edges, K is a set of the key-value pairs, P is a set of PPIs denoting the key-value pair in K, L is a set of labels, □□is the property identification function K □ P, □ is the labeling function mapping (V □□E) □ L, and □ is the identification assignment function mapping (V □□E) □ P.

In UPG data model, entities or resources are represented as vertices and relationships as edges. Both vertices and edges are labeled with their roles and can have only one PPI. The edges are directed. There can be multiple-edges between any two vertices.

An important aspect of UPG is that both vertices and edges can have labels and only one PPI. The multiple labels are essentially useful and flexible for providing the diverse, informative metadata related to resources and relationships. The PPI plays a role of the unique identifier for a vertex and an edge. This also provides reusable property constructs. So UPG data model can be ground, universal model to generate various type of graph-based data model by simply adding or abandoning specific constraints on UPG. Fig. 1 is an example of a typical UPG.

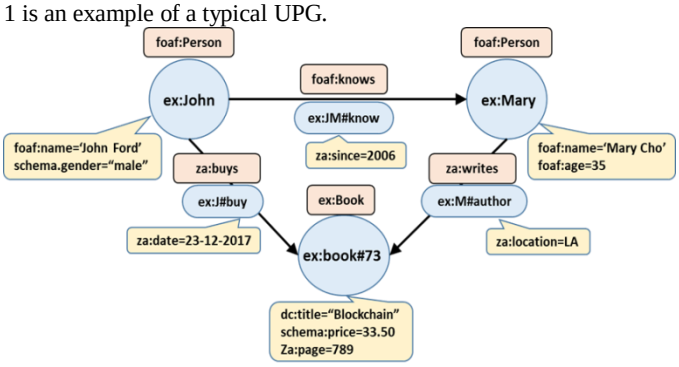


Fig. 1: Example of a typical UPG

Although UPG model is conceptually similar to the conventional LPG, it has its distinct features that are more uniform and powerful capability. These features can seamlessly embrace the idiosyncratic functions of both RDF and LPG. UPG can provide uniform modeling of data in the diverse environment.

3.1.1. Features of Vertices

Vertices or nodes denote entities or resources of the domain. Vertices contain only one PPI consisting of any number of the key-value pair. The PPI is the identifier of a vertex but plays a role of the placeholder for the key-value pairs. Although there are no restrictions to give the key-value pairs to vertices, the properties of a vertex are usually the intentional properties representing conceptual attributes inherent in the entity. Vertices can have one or more labels. Vertex labels can play a vital part in specifying the roles of vertices. This makes it possible to form conceptual schema or hierarchy of a certain concept efficiently.

3.1.2. Features of Edges

An edge, also known as a link, arc or relation, represents a relationship between two connected vertices to establish a conceptual context for each vertex. Edges have a direction to connect two vertices. Even though edges can be self-referencing or looping, they can never be dangling. As the mandatory feature of UPG model, similar to LPG, every edge must have one and only one label to represent the edge uniquely. The edge label represents the relationships between two vertices while vertex labels represent the roles or categories of the vertex.

Much like vertices, edges also have only one PPI to hold their properties. The key-value properties usually describe the circumstantial or contextual attributes when the relationship is built between two vertices such as time, location and modality.

3.1.3. Features of the key-value pairs and PPIs

The property is the foundational mechanism of UPG model to describe the attributes of vertices (entities) and edges (relationships). Since the attributes are the intentional characteristics of an entity, object or relation, the property usually represents intrinsic or conceptual features such as color, weight, and size for vertices, time and location for edges. Since the property is essential for expressing non-relational data, it should be distinguished from the

associative features that are generally represented by the edges. The metadata or ontological vocabularies can be used for the key of the property.
The UPG makes a set of the key-value pairs an object that can be identified by PPI. However, the PPI is not used as the common identifier but plays a role of the placeholder that contains a set of the key-value pairs. In some sense, the PPI is a container of the key-value pairs.

3.1.4. Features of Labels

Labels are one of the foundational elements of UPG data model. Both vertices and edges have labels, however, their applications are different. The vertices labels are a way to assign the roles to vertices and to categorize vertices by means of their semantic features. The vertex labels are similar to rdf:type of RDF, but more efficient and powerful. The vertex labels can be used for many different purposes such as sub-graph creation, efficient UPG data store, and schema generation. On the other hand, every edge has mandatorily one and only one label that represents the relationship between two connected vertices. The edge label plays a role of the unique identifier of the edge.
The features of the major elements of UPG can be observed in Fig. 1. The UPG is more intuitive and precise than LPG and RDF in creating knowledge graph of a certain domain.

3.2. Advanced Features of UPG Model

The UPG is a pure graph model intrinsically based on the key-value pairs and their PPIs. This approach supports several useful capabilities essential for the complex knowledge modeling. These features can resolve some cumbersome problems such as the reification and expand the expressive power so that UPG represents the complicated relationships

3.2.1. Vertex as a Resource

The vertices of LPG that are the placeholders for the data properties consisting of the key-value pairs play a similar role of the resources of RDF. However, this nature of the vertices in UPG is different from those of the resources of RDF uniquely identifiable by IRI. In UPG, labels used to indicate the roles and categorize the vertices are essentially important than the vertex identifiers. The vertex identifier is mainly used to define the edge relationships. Some NoSQL systems based on LPG internally assign the unique identifier to each vertex for efficient management of graph operations.
For the identification of vertices, UPG can use two ways. One is to use PPI to access the key-value pairs directly although this seems to be impractical. The other is to use the properties to access semantically by means of a conceptual naming, for example, Student@[name='John'].
The data objects should be semantically complete and uniquely identifiable to be published and shared in the open world of the Web. In general, the vertices of UPG have resource properties of RDF so that they can provide the useful information when they are accessed.

3.2.2. Ontology Vocabularies and Namespaces

The conventional LPGs use localized vocabularies within a specific system for the labels and the keys of the property. It is strongly required to use ontology vocabularies and namespace so that UPGs can be shared in the open environment like RDF graphs.
The ontology vocabularies and namespaces used in the labels can generate a conceptual schema of the domains. The conceptual schema makes UPG an abstract knowledge model and provides the substantial basis for high-level knowledge processing. And besides, the property also uses ontology vocabularies with namespaces. The property efficiently consists of the key-value pairs similar to tagging data values. The keys can be regarded as metadata for vertices and edges and feasibly represented by ontology vocabularies with the namespaces. Ontology vocabularies with the namespace for the key can provide the commonly shared vocabularies and the coherent semantic interoperability to UPG as RDF graphs.

3.2.3. Nested Property

The original definition of key-value does not give a strict constraint for the datatype of the value. So the value of the key-value pair is usually an opaque string of bytes of arbitrary length. However, different systems expand the datatype for the convenient data modeling and graph traversal, for example, lists for heterogeneous ordered collections of values and maps for heterogeneous, unordered collections of the key-value pairs. UPG needs adequate value types than such the expansion for conceptual modelling.

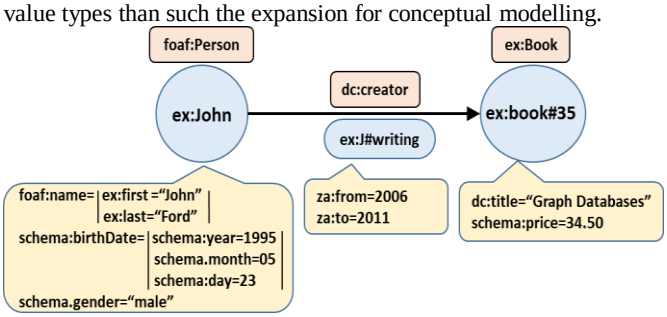


Fig. 2: Example of the nested key-value data type

The key-value datatype has been widely used in feature-based systems to provide more concise and understandable conceptualization for compound attributes. The nested key-value datatype also has complete theoretic basis and application use cases. As an example of the key-value datatype shown in Fig. 2, it provides a preferable conceptualization of resources and relationships.

3.2.4. Formalized Data Type Definition

Many graph database systems based on LPG use the diverse datatypes for the effective modeling and management. However, there are no common specifications to define the datatypes consistently. Moreover, the specifications of the aggregation datatypes such as containers and collections, especially, consisting of vertices or resources as the primitive elements have not been definitely addressed yet. This paper uses the key-value pair specification to define the complex datatypes involving vertices. This approach can keep methodical consistency in LPG modeling based on the key-value property.

The predefined key vocabularies are used to specify the datatypes of data aggregation related to vertices as follows:

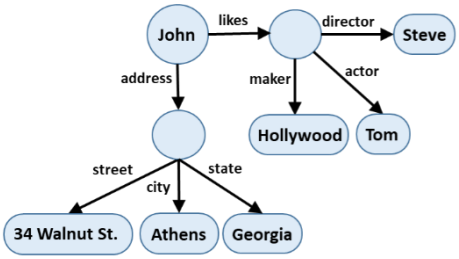
- za:Construct declare the aggregated data structure.
- za:tag gives the data aggregation a literal name that can be used as the reference.
- za:datatype specifies the type of the aggregated structure such as rdf:Bag and rdf:List.
- za:order represents a sequential number of the resource in the aggregated structure.

Assuming that the interpretation of the za:datatype can be accomplished in the given system, this method can specify the diverse datatypes of arbitrary data structures consisting of vertices. Note that PPIs can be used as the values of the keys in UPG.

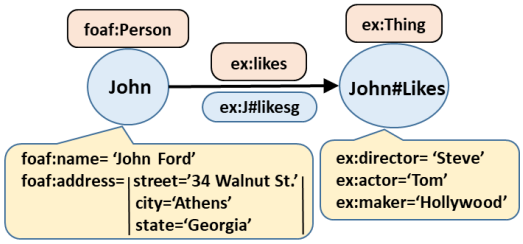
4. Applications of UPG

UPG has powerful modeling capability and shows explicit expressiveness. So UPG can tackle not only the cumbersome problems of RDF and LPG but also the general issues related to knowledge

representation. Fig. 3 is a typical example that UPG addresses the blank nodes of RDF model.



(a) RDF graph with blank nodes



(b) UPG graph of (a)

Fig. 3: UPG modeling of RDF graph with blank nodes

5. Conclusion

There are two dominant graph data models, RDF and LPG, popular for the applications of LOD and NoSQL databases. Although these graph models have plentiful data modeling capabilities, a new graph model that can embrace the capability of both RDF and LPG. This paper proposes a new property graph model called a universal property graph (UPG) with some unique modeling capabilities. UPG can construct the conceptual resources by means of PPI that can be used to identify any key-value structures in the open world. This capability provides a unified view of the resources or entities. UPG expands its representative capability by allowing the nested property. The nested property also offers compact and comprehensive data modeling and solves some structural problems such as blank nodes. Since UPG model uses namespaces and ontological vocabularies, UPG models can be applied in the open world like RDF when appropriate access methods are supports.

This paper presents the conceptual view of UPG. The sound implementation is an important research project. In addition, the associated developments such as graph traversal and serialization should also be investigated.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and future Planning (NRF-2017R1A2B4005185).

References

[1] J. Hoffart, F. M. Suchanek, K. Berberich & G. Weikum (2013), "YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia", *Artificial Intelligence*, Vol. 194, pp. 28-61.

[2] T. Heath & C. Bizer (2011), "Linked Data: Evolving the web into a global data space", *Synthesis lectures on the semantic web: theory and technology*, Vol. 1, No. 1, pp. 1-136.

[3] M. Stonebraker (2010), "SQL databases v. NoSQL databases", *Communications of the ACM*, Vol. 53, No. 4, pp. 10-11.

[4] R. Angles & C. Gutierrez (2008), "Survey of graph database models", *ACM Computing Surveys (CSUR)*, Vol. 40, No. 1, pp. 1-39.

[5] J. Han, H. E, F. Le & J. Du (2011), "Survey on NoSQL database", *Processing of the 6th international conference on Pervasive computing and applications (ICPCA)*, pp. 363-366.

[6] M. A. Rodriguez & P. Neubauer (2011), "Constructions from dots and lines", *Bulletin of the Association for Information Science and Technology*, Vol. 36, No. 6, pp.1-11.

[7] A. C. Kanmani, T. Chockalingam & N. Guruprasad (2016), "RDF data model and its multi reification approaches: A comprehensive comparative analysis", *Processing of the International Conference Inventive Computation Technologies (ICICT)*, pp. 1-4.

[8] D. Hernández, A. Hogan & M. Krötzsch (2015), "Reifying RDF: What works well with wikidata?", *Processing of the 11th International Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS)*, pp. 32-47.

[9] O. Hartig & B. Thompson (2014), "Foundations of an alternative approach to reification in RDF", *arXiv preprint arXiv:1406.3399*, pp. 1-14.

[10] T. Dominik (2016), "RDF data in property graph model", *Processing of the 10th International Conference on Metadata and Semantics Research (MTSR)*. pp. 104-115.

[11] W3C, RDF 1.1 concepts and abstract syntax. W3C Recommendation (2014), available online https://www.w3.org/TR/rdf11-concepts/

[12] O. Hartig (2014), "Reconciliation of RDF* and property graphs", *arXiv preprint arXiv:1409.3288*, pp. 1-14.

[13] S. Das, J. Srinivasan, M. Perry, E. I. Chong and J. Banerjee (2014), "A Tale of Two Graphs: Property Graphs as RDF in Oracle", *Processing of the 17th International Conference on Extending Database Technology (EDBT)*, pp. 762-773.

[14] M. Margitus, G. Tauer & M. Sudit (2015), "RDF versus attributed graphs: The war for the best graph representation", *Processing of the 18th International Conference on Information Fusion*, pp. 200-206.