A Configurable Graph Data Type for Online Exploration of Web Data

Thomas Hornung[1], Wolfgang May[2], and Daniel Schubert[2]

[1] Institut für Informatik, Universität Freiburg,
hornungt@informatik.uni-freiburg.de
[2] Institut für Informatik, Universität Göttingen,
{may,schubert}@informatik.uni-goettingen.de

Abstract. Application domains often include notions that are inherently based on graph structures. In this paper, we propose a comprehensive generic ontology-based datatype for graphs. It focuses on those aspects of graphs that are useful for workflows that require exploration of relevant parts of potentially large graphs by online algorithms. The goal of the ontology is to include as much information as possible supporting the graph exploration process declaratively into the specification of the graph. This allows to separate the (also declarative) specification of the actual exploration process from the maintenance of the graph itself. For concrete applications, the graph specification is given in RDF using this ontology. From the specification, an appropriate instantiation of the abstract datatype is automatically derived which is then used in information workflows.

1 Introduction

A recurring motive when designing informational workflows is the computation of (parts of) transitive closures of graphs. Graph algorithms in general are a traditional research topic; they usually assume a given graph and the focus is on employing additional suitable data structures for efficient algorithms. In the context of the Web, *online algorithms* [1, 4] became more relevant: there, the graph is neither known nor materialized a priori to run algorithms on it, but is *explored* only at runtime, using one or more Web data sources. Often, even the graph data itself is dynamic which does not allow for materialization or caching. These characteristics require completely different algorithms where the exploration and expansion strategy for the graph itself is the central issue. Most algorithms basically follow a *best-first-search* like A^* [15], *breadth-first-search*, or *depth-first-search* for exploration.

In this work, we present the *Configurable Graph DataType* (CGDT) that provides an ontology and an API for *configurable* graphs. The design of CGDT combines generic graph behavior (insertion of edges etc.) with application-specific configurability. CGDT allows to encode the maintenance of the stored graph data inside the graph by (i) assigning properties to vertices, edges, and paths, and (ii) specifying how paths are obtained from existing edges and paths during the

exploration process. This allows to separate the (also declarative) specification of the actual exploration process and the basic acquisition of data from the Web from the maintenance of the graph itself. The CGDT can be embedded in the declarative specification of informational workflows that are specified in RelCCS [6], a process specification language that extends CCS [10] to relational data flow. RelCCS itself is based on the MARS framework [8, 5], an open framework that provides interoperability between nearly arbitrary languages that support relational dataflow. For the actual acquisition of data from the Web that takes place on-demand, MARS enables to embed query languages in RelCCS processes.

In general, Web nodes that implement CGDT can act as Web-wide services for storing, maintaining and querying graph structures not only within MARS-based approaches. The prototype implementation of CGDT uses a relational database for storing the actual contents of the graphs. CGDT calls also set-oriented, i.e., a set of edges (tuples) can be inserted and processed at a time.

Structure of the Paper. In the next section, we describe a concrete use case and analyze the general requirements and concepts for the CGDT ontology. Section 3 introduces the schema part of the ontology. Section 4 adds generic notions to specify how the graph develops during evaluation of an online algorithm. Section 5 gives an overview of related work, and Section 6 concludes.

2 Application Scenario and General Considerations

Consider the problem to find either the cheapest or shortest (in terms of total time spent travelling) route to a given location (e.g., for a conference travel) or a combination of both. Human, manual search usually employs some kind of intuitive strategy. Roughly, the strategy is to start with considering a known set of airports near the hometown and to try to cover as much distance as possible by plane (assuming the distance is above a certain threshold), and then bridge the remaining distance by train or bus; if this fails, do backtracking. This shows that, although human problem solving usually considers one possibility (= tuple) at a time, in this case it is inherently based on a set-oriented model.

With the means of the presented approach, such tasks can be formulated as data workflows. The backtracking is here replaced by a search strategy, where the search space is explored stepwise and pruned based on intermediate results. While for train connections, sources usually are able to return transitive connections, flight portals only return transitive connections over the flights of the same airline. Thus, here an actual graph exploration is required. An typical aspect of this use case (and many other ones) is that the search is subject to additional constraints, like arrival and departure times and required time for changing.

The expected answer is the *set* of k best alternatives (wrt. a weighted function of price and duration), where each solution contains the actual connection data (flight and train numbers, departure/arrival times). Furthermore, it should in general be possible to extend the process specification in such a way that the best available one is actually booked automatically.

Pitfalls. Experiences with conference travels showed that real travel agencies are often challenged with finding the potential nearest airports to less standard destinations (e.g., St. Malo/France as for ICLP 2004), and are rather weak in finding non-direct flight connections using different airlines (e.g. Lufthansa + AirFrance) or via unexpected intermediate airports (via London Stansted to reach Dinard/France), or surprising connections (fly to Jersey Island and take the ferry to St. Malo) – actually, ferries are often contained in the railway portals, so we do not consider these separately. The latter example shows also that it would not be advantageous to try to save time by predefining the set of destination airports by the user, but to use a fully algorithmic search that is not biased.

Comparison to Classical Graph Algorithms. On first sight, the problem looks like an application for classical "shortest path" graph algorithms like Prim [11, 2] or Kruskal [7, 2]. A more detailed analysis shows that even under some optimistic assumptions, this would not be an appropriate solution:

- Dynamics: the complete graph is not available, and is continuously changing: the availability of flights and their prices changes every moment.
- Constraints: the paths are further constrained by the requirement that the departure time must be after the arrival time at intermediate airports.
- Completeness: the airline connections' graph (which is, neglecting the availability issue, of a size that could efficiently be processed by breadth-first or A^* search), is not sufficient. Additionally, the connections between airports and the final destination must be considered. Thus, finding the solution in the graph depends on further information since below a certain remaining distance the process continues outside the main graph.

Generalization. The above considerations show that in such cases, a large search space has to be explored, and application-specific properties of the paths, like price and duration, have to be maintained *incrementally*. The stepwise exploration corresponds to *inductive* characterizations of these properties that are in fact common to the idea of properties of paths in a graph. The CGDT ontology provides generic notions to specify how this information is combined from the actual input (i.e., information about edges obtained from Web sources). The relevant features are mapped and expressed in terms of the generic ontology.
CGDT supports the following generic functionality:

- materializing the relevant graph fragment (including the inductively defined properties) based on the explored edges,
- creating paths according to specified criteria,
- accessing the vertices that should be explored next according to BFS or A^*,
- querying the result graph.

The design of the data workflow can then be separated into three issues:

- describe the domain-specific characteristics of the graph in terms of the CGDT ontology. This consists of the basic schema of the graph, the *constructive specification* how to extend the graph with relevant newly obtained

knowledge, and *constraints* when newly obtained knowledge is relevant to expand the graph;
- actual acquisition of the data from the Web (including Deep Web sources). This means to identify appropriate data sources and to encode the access to them. Potentially, for each step also two or more sources must be accessed – for instance one to identify potential edges (in our example: which airports can be reached from a given one), and the second to query for the actual existence of the edges (in our example: actual availability and departure/arrival times of that connection for a given date);
- fill in a common breadth-first-search or A^*-search workflow pattern as a RelCCS process with case splits and Web queries.

After configuring the graph once during the initialization, the process will only submit edges to the graph, and query it for the vertices where the exploration should be continued. The compilation of the information in the graph itself, and the choice of the vertices for the next step is done automatically by the graph.

3 An Ontology for Graphs in Online Algorithms

The basic notions of any graph ontology are vertices, edges, and paths. In the following, we consider directed, labeled graphs of the form $G := (V, E, P)$, where V is the set of vertices, $E \subseteq V \times V$ is the set of directed edges between these vertices, P is a set of paths. While in the usual notion of graphs, the set of paths is defined as the transitive closure of edges (i.e., the set of paths is $\{(v_1, \ldots, v_n) \mid (v_1, v_2), \ldots, (v_{n-1}, v_n) \in E\}$), the set P of relevant paths in a configurable graph is a certain *subset* of all existing paths in the graph that satisfy additional constraints. Nevertheless, each path $p \in P$ is a path in the traditional sense which consists of multiple connected edges. A path p that ends in a vertex x can be extended by an edge (x, y), denoted by $p \circ (x, y)$. The set P will contain paths that are obtained by such extension steps according to configurable criteria.

3.1 Properties

A central feature of CGDT is that vertices, edges and paths can be adorned with sets VP, EP, and PP of (typed) properties. Each property is associated with a literal type, taken from the XML Schema datatypes [18].

The properties can optionally be specified in terms of view definitions over other properties, or by external queries. For instance, given a vertex with its airport code, the timezone can be obtained by a suitable Web query. The distance of a flight from A to B is the geographical distance between A and B's coordinates, and the price of a path is the sum of the prices of its edges.

For properties of vertices and edges where no definition is given, the value must be given when adding the edge to the graph. Often, vertices are added only with their key (when found by exploring edges), and their additional properties are obtained by external queries that are automatically executed upon insertion

of the vertex. As paths are not inserted manually, but automatically by extending an existing path with an edge, all path properties must be derived properties. Here, often an inductive definition over the length of the paths is used.

3.2 Signature and Operations

The operations of CGDT are divided into a *Data Definition Language (DDL)* where the properties and the constraints are *defined*, and a *Data Manipulation Language (DML)* that provides *generic* operations for updating and querying the graph which are used during the actual process of exploration.

3.2.1 The DDL

While in SQL and related languages, the DDL has an own syntax, the DDL of CGDT is actually the *ontology language* RDF [12] that *declaratively* specifies which properties exist, together with their definitions, and with the constraints how to expand the graph.

In contrast to SQL, where the main notion of the schema is the *table*, the CGDT is based on three subschemas, i.e., a VertexSchema, an EdgeSchema, and a PathSchema. Each of them defines some properties (i.e., VertexProperties, EdgeProperties, and PathProperties) and optionally some constraints (to be discussed in Section 4) that guide the exploration process. Each of the subschemas can be regarded (and stored) as a table. The notions of the generic graph ontology itself (i.e., the DDL notions) are depicted in UML in Figure 1; an excerpt of the RDF Schema [13] definition can be found in the long version of this paper[1].

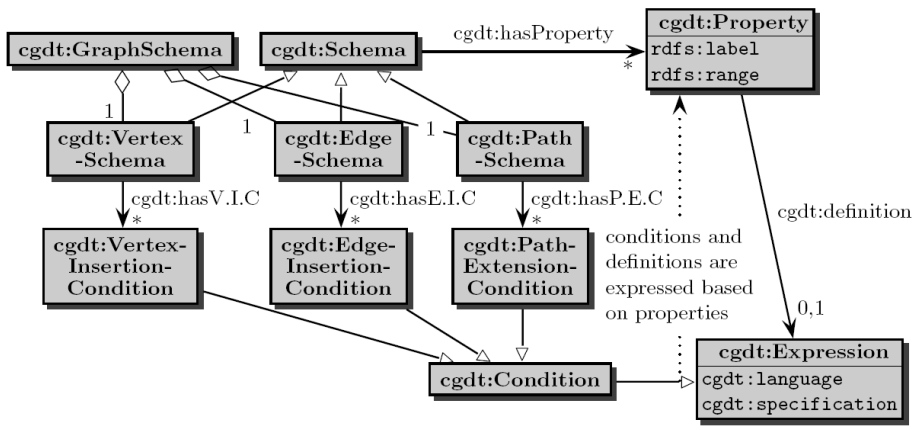

Fig. 1. Basic notions of the CGDT ontology

The three subschemas contain some mandatory, built-in properties:

– vertex schema: id serves as key,

[1] Available at `http://www.dbis.informatik.uni-goettingen.de/Publics/`

- edge schema: id (key, internally generated and used), from and to, referring to vertices x and y for an edge (x, y). Note that from and to are not key to allow different edges between the same vertices (e.g., several flights at the same day),
- path schema: id (key, internally generated and used), from, to, front and last, referring to a, y, p and (x, y) (the latter two referring to ids of an edge and a path, respectively) for the path $p \circ (x, y)$ where a is the first vertex of p.

A concrete application-specific CGDT specification then defines

- the names and datatypes of the additional application-specific properties of each subschema,
- the definitions of the derived properties,
- conditions to configure the exploration process (to be discussed in Section 4).

Derived Properties. Since derived properties are *views* over other properties or Web queries, they can be expressed by query languages. For accessing the Web, external queries can be embedded using the language management of the MARS framework. Throughout this paper, we use pseudocode expressions. Properties of paths are often defined inductively. For these, the specification of the base case (which is an edge, and thus builds upon the edge's properties) and of the inductive step (potentially using the path and the extending edge) have to be given. Instead of giving an inductive definition, path properties can also be specified to be SumProperties, CountProperties, or {Min|Max}Properties, which are defined as the aggregation of the values of a specified edge property.

Example 1 *In our running example, the concrete instantiation of CGDT is tailored to the travel application scenario and rooted shortest path search.*

The vertices (which are the train stations and airports) have two properties, i.e., the id *(which is e.g. the airport code) and the timezone. The timezone is defined by a Web query (against a wrapped Web source)*

timezone = getTimezone(<http://www.theairdb.com>, code) .

Edges, which are the direct connections, e.g., FRA-CDG (Frankfurt to Paris Charles de Gaulle), have domain-specific properties code *(the flight number),* dept, arr *(departure and arrival time wrt. the local timezone) and* price. *The* duration *is a derived property:*

duration = arr - dept + from.timezone - to.timezone.

The properties of the paths, from, to, dept, arr, price *and* duration *are defined inductively. For the base case where a path is just a single edge, they have the same values as for the edge. For paths of length* > 1*, they are defined as follows:*

from = front.from (built-in), dept = front.dept,
to = last.to (built-in), arr = last.arr,
price = front.price + last.price or equivalently as a SumProperty
= sum[e:edge](e.price) (= sum of prices of all edges of the path)
duration= front.duration + last.duration + last.dept - front.arr
which equals last.arr - front.dept + from.timezone - to.timezone.

The Constructor. The constructor *gid* ← getGraph(*rdf-spec*) initializes a new CGDT instance with a given specification *rdf-spec* (which is an RDF specification of the desired instance) and returns a unique graph id.

3.2.2 The DML

The DML is also independent from the actual application domain. The modifiers allow to add items to the graph:

- addVertex(*id*, *vertex-property-name-value-pairs*) adds a new vertex *id* with the given vertex property values,
- addEdge(*from*, *to*, *edge-property-name-value-pairs*) adds a new edge (*from, to*) with the given property values (and adds the target vertex if not yet present).

In the pseudocode, we use a slot-based notation, e.g. addEdge("FRA", "CDG", [dept ← "10:30", arr ← "11:50", code ← "LH123", price ← 185.00]).
The accessors include the following:

- *var* ← getNewVerticesBFS() supports breadth-first-exploration and binds *var* to the ids of each of the new vertices that have been added since the previous call of getNewVerticesBFS(),
- *var* ← getNextVertexAStar() binds *var* to the id of the next vertex that has to be extended according to A^* best-first-search (and a given valuation function, cf. Sec. 4.4),
- $(v_1, \ldots, v_n)$ ← getResultPaths(v_1 ← *attr*$_1$, …, v_n ← *attr*$_n$) returns a binding for variables $(v_1, \ldots, v_n)$ to the corresponding attributes of each path that is considered as a result. In Section 4.4 we will discuss how the intended result paths are specified in the ontology.

4 Configurability of the Exploration Process

Although breadth-first-search, best-first-search and depth-first-search proceed different in the large, the *configuration* of the exploration process can be specified by the same notions. Thus, we exemplify it for the use in *breadth-first-search*, which shows the set-oriented features best by doing the expansion in parallel.

4.1 Breadth-First Search

The underlying principle of breadth-first-search is simple and makes the strategy well-suited for graph exploration in online algorithms: Starting with a set of one or more known vertices (e.g., the nearest airports to the starting place), consider all edges from these vertices to any other (known or yet unknown) vertex. These edges are added to the graph, and (i) can be used to extend existing paths, and (ii) result in newly known vertices that can be used in the next step.

The configuration of the behavior of the graph consists of conditions that specify the following:

1. when a new edge is found, add it to the graph or discard it (e.g., when certain airlines or intermediate airports should be excluded),

2. when a new edge is inserted: under which conditions can it be used to extend an existing path p (e.g., its departure time must obviously be later than the arrival time of p),
3. under which conditions should a vertex be considered for the next exploration step?

By this, CGDT separates the acquisition of edges (that must be programmed explicitly in the processs) from the actual handling of their contributions to the graph (that is configured into the graph).

4.2 Insertion Conditions

For vertices and edges, conditions can be stated that need to be satisfied for insertion of the item into the graph. *Vertex insertion conditions* are *only* concerned with properties of the vertex itself (e.g., the exclusion of flights via London Heathrow (LHR) because of luggage handling problems can be expressed as $\mathsf{id} \neq$ "LHR"). *Edge insertion conditions* are *only* concerned with properties of the edge itself (e.g. $\mathsf{duration} <$ "10:00"), its start and end vertices, and with general properties of a graph (e.g., forbid to make the graph cyclic). An edge is also not inserted if one of its vertices does not satisfy the insertion conditions.

4.3 Path Extension Conditions

Path Extension Conditions allow to state application-specific constraints whether a new edge (x, y) can be *used* for extending a path p that ends in x to $p \circ (x, y)$. They are formulated in terms of the properties of the edge and of the path.

Example 2 *In our example, for a path* $((s, \ldots, x), [\mathsf{arr} = t_1])$ *and a new edge* $(x, y, [\mathsf{dept} = t_2])$*, the new path* $((s, \ldots, x, y), [\ldots])$ *is only added if* t_2–t_1> *"01:00".*

Consider an invocation of $\mathsf{addEdge}(x, y, [...])$ *(i.e. a direct connection). If the destination airport* y *is not yet contained in the graph, it is added as a vertex (automatically retrieving its* timezone *property from the Web). The connection itself is added as an edge with its properties, and for all paths* $p = (s, \ldots, x)$*, the path* $p' = (s, \ldots, x, y)$ *is a candidate for insertion. If the new edge's departure is more than one hour later than* p*'s arrival,* p' *is actually inserted with the appropriately computed property values.*

If for such newly added paths, edges (y, z) are already stored, the respective extended paths $(s, \ldots, x, y, z)$ are also candidates for insertion, and so on. Note that edges (like in the example $(x, y, [\mathsf{dept} = t_3])$ with $t_3 < t_1$) that cannot yet be used for extending an (already known) path can possibly be used later for extending other paths that reach y with an earlier arrival time than p. For that, path extension conditions are usually stricter than edge insertion conditions. Vertices are only considered as "new" to be extended in the next step if they became actually newly reachable by a path.

4.4 Specification of Desired Result Paths and Termination

The above conditions control *how* the internal information of the CGDT instance is extended when adding edges. Additionally, it must be specified, when the process ends, and preferably already during the process, only new vertices that are promising to continue the search should be selected for the next step.

The *Result Specification* is expressed via a filtering condition which paths can qualify as intended results (in the example, those that end in the final destination), and optionally a valuation function on paths (that can be seen as a cost measure and that must be strictly monotonic wrt. path extension) and an integer k, how many results should be finally returned.

Example 3 *In our example, the valuation function is* duration (in hours) + price / 100 *(means, for 100€ saved, one accepts one more hour to travel). The filter condition is* to = finalDestination, *and* $k = 5$.

When breadth-first-search is applied, paths that are "above" (i.e. more expensive) the limit of the best k results so far are not further extended, and vertices that are only reachable by such paths are not expanded. This prunes the search space as soon as k paths have been found that satisfy the filter condition, and guarantees termination. In case of A^* search, the valuation function is used to choose the next vertex to be extended.

5 Related Work

The notions of *online algorithms* [1] in general and *dynamic graph algorithms* [4] cover a broad spectrum of aspects. This includes scenarios where the *current* situation is completely known, but *changes*, as well as situations where the underlying situation is actually static, but is not completely known and is processed *incrementally*, like dynamic search algorithms. CGDT is tailored to the special, but still very common case where exploration is dynamic, but monotonic: vertices and edges once added to the graph will remain unchanged forever. The underlying graph is also dynamic, but every run is based on a (non-transactional) snapshot that is *explored* dynamically.

Online algorithms over unknown graphs are investigated by many authors under different aspects (total exploration [3], search etc.). For path search, breadth-first-search and best-first-search by A^* (see e.g. [15] for an overview) are the most prominent ones. Also, research on composition of Web Services like [14, 9] is a related area, but in general deals with a higher level of abstraction where the concrete modeling and algorithmic handling of the data is not described. Such approaches can be complemented with the use of CGDT, since it declaratively covers the data-oriented aspects.

Most works on graph schemas have a different goal, namely to describe a *graph-based data model* in the sense of semistructured data like RDF on the schema level by the labels of its vertices and edges. In these languages the graph is not part of the domain and it is not used for applying graph algorithms, but

the domain is modeled as a graph which is updated and queried. Some works in the area of graph transformations, e.g. PROGRES [16] allow –like CGDT– to assign (optionally derived) attributes not only to vertices, but also to edges and paths. Paths are seen as derived edges that are declared in a rule-based way.

6 Conclusion

We presented an ontology for a configurable graph datatype CGDT that supports explorative online algorithms using Web information sources. CGDT allows to declaratively specify and encapsulate the handling of the collected graph data, and to separate it from the data acquisition and process control.

A prototype of the implementation has been completed. An online prototype for MARS and RelCCS, with further documentation and the above process can be found at http://www.semwebtech.org/mars/frontend/ → run CCS Process.

References

1. S. Albers. Online Algorithms: A Survey. *Math. Prog.*, 97(1-2):3–26, 2003.
2. T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms*.
3. X. Deng and C. H. Papadimitriou. Exploring an Unknown Graph. In *FOCS*, pages 355-361, 1990.
4. D. Eppstein, Z. Galil, and G. F. Italiano. Dynamic graph algorithms. In *Algorithms and Theory of Computation Handbook*. CRC Press, 1999.
5. O. Fritzen, W. May, and F. Schenk. Markup and Component Interoperability for Active Rules. In *Web Reasoning and Rule Systems (RR)*, Springer LNCS 5341, pages 197–204, 2008.
6. T. Hornung, W. May, and G. Lausen. Process algebra-based query workflows. In *CAiSE*, Springer LNCS 5565, pp. 440-454, 2009.
7. J. Kruskal. On the shortest spanning subtree and the traveling salesman problem. *Proceedings of the American Mathematical Society*, (7):48–50, 1956.
8. W. May, J. J. Alferes, and R. Amador. Active rules in the Semantic Web: Dealing with language heterogeneity. In *RuleML*, Springer LNCS 3791, pages 30–44, 2005.
9. S. A. McIlraith and T. C. Son. Adapting GOLOG for composition of Semantic Web Services. In *KR*, pages 482–496. Morgan Kaufmann, 2002.
10. R. Milner. Calculi for synchrony and asynchrony. *Theoretical Computer Science*, pages 267–310, 1983.
11. R. C. Prim. Shortest connection networks and some generalisations. *Bell System Technical Journal*, (36):1389–1401, 1957.
12. Resource Description Framework (RDF). `http://www.w3.org/RDF`, 2000.
13. Resource Description Framework (RDF) Schema specification. `http://www.w3.org/TR/rdf-schema/`, 2000.
14. D. Roman and M. Kifer. Reasoning about the Behavior of Semantic Web Services with Concurrent Transaction Logic. In *VLDB*, pages 627–638. 2007.
15. S. Russell and P. Norvig. Artificial Intelligence: A Modern Approach, 2003.
16. A. Schurr, A. J. Winter, and A. Zündorf. The Progres approach: language and environment. In *Handbook on Graph Grammars and Computing by Graph Transformation: Applications, Languages, and Tools*. World Scientific, 1999.
17. Turtle - Terse RDF Triple Language. `http://www.dajobe.org/2004/01/turtle/`.
18. XML Schema part 2: Datatypes. `http://www.w3.org/TR/xmlschema-2`, 1999.