Schema Formalism for Semantic Summary Based on Labeled Graph from Heterogeneous Data

Amal Beldi, Salma Sassi, Richard Chbeir, Abderrazak Jemai

Schema formalism for semantic summary based on labeled graph from heterogeneous data

Amal Beldi[1], Salma Sassi[2], Richard Chbeir[2], Abderrazak Jemai[1,3],

[1] Tunis El Manar University, Faculty of Mathematical Physical and Natural Sciences of Tunis, SERCOM Laboratory, 1068 Tunis, Tunisia
amal.beldi@fst.utm.tn
[2] University Pau & Pays Adour, LIUPPA, Anglet, 64600, France
richard.chbeir@univ-pau.fr
[3] Carthage University, Polytechnic School of Tunisia, SERCOM Laboratory, INSAT, 1080, Tunis, Tunisia
abderrazekjemai@yahoo.co.uk

Abstract. Graphs are used in various applications and are used to model real world objects. To understand the underlying characteristics of large graphs, graph summarization becomes a hot topic aiming to facilitate the identification of structure and meaning in data. The problem of graph summarization has been studied in the literature and many approaches for static contexts are proposed to summarize the graph in terms of its communities. These approaches typically produce groupings of nodes which satisfy or approximate some optimization function. Nevertheless, they fail to characterize the subgraphs and do not summarize both the structure and the content in the same approach. Existing approaches are only suitable for a static context, and do not offer direct dynamic counterparts. This means that there is no framework that provides summarization of mixed-source and information with the goal of creating a dynamic, syntactic, and semantic data summary. In this paper,the main contribution relies on summarizing data into a single graph model for heterogeneous sources. It's a schema-driven approach based on labeled graph. Our approach allows also to link the graph model to the relevant domain knowledge to find relevant concepts to provide meaningful and concise summary. After extracting relevant domain, we provide a personalized visualization model capable of summarize graphically both the structure and the content of the data from databases, devices, and sensors to reduce cognitive barriers related to the complexity of the information and its interpretation. We illustrate this approach through a case study on the use of E-health domain.

Keywords: Graph formalism, heterogenous data, real time, interoperability, structure summarization, based content summarization, aggregation, compression

1 Introduction

Data graph management provides better support for highly interconnected datasets [1]. Most big data applications including social networks [2], bioinformatics[3] and astronomy [4] are examples of large-scale interconnected graphs. Such data can be more easily expressed using entities of a graph (nodes and edges). Querying and reasoning about the interconnections between entities in such graph dataset can lead to interesting and deep insights into a variety of phenomena. However, due to sheer volume, complexity, and temporal characteristics, building a concise representation (i.e., summary) helps to understand these datasets as well as to formulate queries in a meaningful way. In this context, graph summarization becomes a hot topic in the database research community in recent years. It facilitates the identification of structure and meaning in data. A summary is a concise representation of the original graph, whose objectives can greatly vary from reducing the number of bits needed for encoding the original graph, to more complex database-style operations that summarize graphs where the resolution could be scaled-up or scaled-down interactively [5]. With the advent of dynamic graphs and streams, there is a demand for analyzing the time-evolving properties of such graphs, and once again graph synopsis construction has found increasing interests[6]. Given its advantages, graph summarization has a wide range of application including interactive and exploratory analysis[7], approximate query processing [8] , visualization [9], data-driven Visual graph query interface construction [10] and distributed Graph Systems[11] among others. The problem of graph summarization has been studied in the fields of graph mining and data management. Many approaches for static contexts such as modularity-based community detection[12],spectral clustering[13], graph-cut algorithms [14] exist

to summarize the graph in terms of its communities, but lack explicit ordering[15]. These approaches typically produce groupings of nodes which only satisfy or approximate some optimization function. They also fail to characterize the subgraphs and do not summarize both the structure and the content in the same approach. Existing approaches also do not offer characterization of the outputs. The lack of explicit ordering in the groupings leaves a user with limited time and no insights on where to begin understanding his data. Furthermore, existing approaches are only suitable for a static context, and do not offer direct dynamic counterparts. Some algorithms like[14] do work in a dynamic setting, but focus only on finding static patterns that appear over multiple time steps. This means that there is no framework that provides summarization of mixed-source and information with the goal of creating a syntactic and semantic data summary. Given the above problems, the proposed paper focus on how we can best describe in one summary both structure and content and thus not just generate succinct summaries for the mixed-sources, but also understand its corresponding interactions and relationships with the past. Thus, towards building a semantic and dynamic summary, the following challenges emerge.

- **Challenge 1**: How to provide multi-sources-based summary, due to multi-modality of data (e.g.,text, video, and image) that can be encoded in different formats ?
- **Challenge 2**: How to provide user oriented semantic based summary, due to the difficulty of retrieving information according to user 'needs ?
- **Challenge 3**: How to incorporate the dynamic nature of real data in computation and perform analysis efficiently ? Our work aims to generate a concise semantic summary of heterogeneous sources to better understand their underlying characteristics.

The main contribution of this study relies on summarizing data into a single graph model for heterogeneous sources. It's a schema-driven approach based on labeled graph. Our approach allows also to link the graph model to the relevant domain knowledge to find relevant concepts to provide meaningful and concise summary. So, we propose a summary-driven formalism based on labeled graphs, which provides interesting data summary conserving better data integrity. Also, this formalism allows more structured summary of the data stored within a graph database. The rest of this paper is organized as follows. Section 2 provides a Literature review describing and discussing related works on graph summarization. Section 3 describes the overall of our approach. In section 4, we describe a schema-driven approach for summarization of the LPG graph, based on our proposed formalism. section 5. described the summarization model-based content and . Section 7 provides the implementation of our approach, the conducted experimentation, results, and discussions. Section 8 concludes this study and provides several perspectives.

2 RELATED WORKS

2.1 SUMMARIZATION APPROACHES

- **Static plain graph approach**: most works in static graph summarization focuses on graph structure without side information or labels. At a high level, the problem of summarization, aggregation or coarsening of static is described as simplification-based summarization methods streamline the original graph by removing less "important" nodes or edges, resulting in a sparsified graph[28]. A representative work on node simplification-based summarization techniques is OntoVis[20], representing a visual analytical tool that relies on node filtering for the purpose of understanding large, heterogeneous social networks in which nodes and links. Toivonen et al[16] focus on compressing graphs with edge weights, proposing to merge nodes with similar relationships to other entities (structurally equivalent nodes. SPINE, an alternative to CSI[17], sparsifies social networks to only keep the edges that "explain" the information propagation those that maximize the likelihood of the observed data. In the visualization domain, Dunne and Shneiderman[18] introduce motif simplification to enhance network visualization.
- **Static labled graph approach**: We have reviewed summarization methods that use the structural properties of static graphs without additional information like node and edge attributes. The main challenge in summarizing labeled graphs is the efficient combination of two different types of data: structural connections and attributes[30]. Currently, most existing works focus on node attributes alone, although other types of side information are certainly of interest in summarization. The first and most famous frequent-subgraph-based summarization scheme is SUBDUE [22] employing a greedy beam search to iteratively replace the most frequent subgraph in a labeled graph. The S-Node representation [23] is a novel two-level lossless graph compression scheme optimizing specifically Web graphs. SNAP and k-SNAP are two popular database-style approaches [19]

rely on (attribute and relationship-compatibility), which guarantees that nodes in all groups are homogeneous in terms of attributes, and are also adjacent to nodes in the same groups for all types of relationships. song and al [21] proposes a lossy graph summarization framework as a collection of d-summaries, which intuitively are supergraphs that group similar entities.

- **Dynamic graph approach**: analyzing large and complex data is challenging by itself, so adding the dimension of time makes the analysis even more challenging and time-consuming. For this reason, the temporal graph mining literature is rich, mostly focusing on laws and patterns of graph evolution. Summarization techniques for time-evolving networks have not been studied to the same extent as those for static networks, possibly because of the new challenges introduced by the dimension of time. The methods are sensitive to the choice of time granularity, which is often chosen arbitrarily: depending on the application, granularity can be set to minutes, hours, days, weeks, months, years, or some other unit that makes sense in a given setting. This category's only representative is TCM[26] and TimeCrunch[24], which succinctly describe a large dynamic graph with a set of important temporal structures. (Qu et al. 2014)[31] is a stream of time-ordered interactions, represented as undirected edges between labeled nodes. NetCondense[25] is a node-grouping approach that maintains specific properties of the original time-varying graph, like diffusive properties important in marketing and influence dynamics, governed by its maximum eigen value.

2.2 Discussion and limitations

In order to compare the existing approaches and to overcome the challenges described previously, we define here 7 criteria with respect to the defined challenges:

Challenge 1: How to provide multi-sources-based summary, due to multi-modality of data (e.g.,text, video, and image) that can be encoded in different formats ?

- **Type of input Data (C1)**: this criterion refers to the input data which could be: structured data such as already defined knowledge models include existing ontologies and database schema/graph (ii) Semi-structured data designates the use of some mixed structured data with free text such as Web pages, Wikipedia sources, dictionaries, and XML documents, and (iii) Unstructured data is related to any plain text content , video, signal. etc.
- **Data type (C2)**: this criterion describe the type of data incorporate (text, xml, numeric, video, image) .
- **Representation standard (C3)**: this criterion describes if the approach incorporates standard((i.e. information based standard, document based standard or Hybrid standard) (e.g., Yes or No).

Challenge 2: How to provide user oriented semantic based summary, due to the difficulty of retrieving information according to user 'needs ?

- **Summarization approach (C4)**: this criterion refers to the target of the summarization approach structure or based content,
- **Summarization approach (C5)**: this criterion refers to the objective of the summarization approach query efficiency , compression, influence,
- **Summarization technique (C6)**: this criterion refers to the techniques deployed to summarize ehr which could be: grouping, compression, analysis, pattern- mining, classification, visualization.

Challenge 3: How to incorporate the dynamic nature of real data in computation and perform analysis efficiently ?

- **Output type (C7):**this criterion concerns the type of displayed summarized data which is a combination of: numerical data, textual data, document, graph.
- **Context-aware criterion (C8):** defines two types of context-aware:(i) Partial, used to demonstrate if an existing system uses concepts about the deployed context of the devices (e.g., time, location, and trajectory) or concepts about the static data and (ii) Total, used to determine if an existing system uses both of deployed context of devices and other static data context.
- **User oriented summarization (C9):**this criterion represent that the approach oriented user (e.g., yes or No).

Our comparison highlights that the evolution of the summary is still an open challenge. So, we observe that most of existing studies[20] [16] [17] [18] do not consider real data in their analysis and do not consider the context on creating the summary and they rely only on the time property.Thus, existing systems[27][26][25] are still unable to contextually interpret and reason on the transferred knowledge among real data, and consequently cannot synthetize data in order to provide accurate desired results. All existing systems focus on one objective, while none of them provide in the same framework various

functionalities despite its importance in supporting users' preferences to find the data according to various needs. All objectives should be an integral part of a summarization-based system. Most of the above studies [29][30] [20][19] can only satisfy a certain aspect of users'needs. Finally, another important part of this study is the output type of summarized data. They do not propose dedicated tools that make the summary accessible to the user nor provide them with appropriate perceptions of their needs. Users are more and more concerned about security, confidentiality, understanding their data, and the accuracy and completeness of their data. In this study , the main approach that we address the aforementioned problems by proposing an appropriate approach able to model heterogeneous sources based in a single graph based on a schema-driven approach, providing a personalized summary model capable of synthesize graphically the content based and finally summarizing the structure of the graph in order to reduce its size and minimize its complexity and keep the important nodes and relations.

Table 1. Qualitative Comparison of static ,static labeled dynamic plain Graph Summarization Approach

	Challenge 1			**Challenge 2**				**Challenge 3**	
Existing study/Criterion	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**
Category 1:Dynamic graph									
(Adhikariet al 2017)	Structured	Weighted, Directed, Unidrected	Yes	Structure	Influence	Grouping	Supergraph	Partial Time	Yes
(Tan et al.,2016)	Structured	Weighted, Directed, Undirected	Yes	structure	Query efficiency	Grouping	Supergraph	Partial time	Yes
(Shah et al.,2015)	Structured	Unweighted, Directed, Undirected	Yes	Structure	Visualization	Compression	List of temporal structure	Partial time	Yes
(Qu et al.,2014)	Structured	Unweighted, Undirected	Yes	Structure	Influence	Influence	Subgraph	Partial time	Yes
Category 2: Static graph									
(Maccioni et al.,2016)	Structured	Unweigted, Directd, Undirected	No	Structure	Query efficiency	Grouping	Sparsified graph	No	No
(Dunne et al, 2013)	Structured	Unweighted, Undirected	No	Structure	Visualization	Grouping	Supergraph	No	No
(Toivonen et al 2011)	Structured	Weighted, Directed, Undirected	No	Structure	Compression	Grouping	Supergraph	No	No
(Mathioudakis et al ,2011)	Strctured	Weighted, Directed	No	Structure	Influence	Influence	Sparsified graph	No	No
Category 3: Static labeled graph									
(Song et al 2016)	Structured	Unweighted, Undirected	No	Structure	Query efficiency	Grouping	Supergraph	No	No
(Mehmood et al 2013)	Structured	Unweighted, Directed	No	Structure	Influence	Influence	Supergraph	No	No
(Toiven et al 2011)	Structured	Weighted, Undirected, Directed	No	Structure	Grouping	Compression	Super graph	No	No
(Shein et al 2006)	Structured	Unweighted, Undirected	No	Structure	Simplification	Visualization	Sparsified graph	No	No
Proposed approach	Structured, Unstructured	Unweighted, Undirected	Yes	Structure, Content	Summarization	Aggregation, Mathematics operations	Graph summary	Yes	Yes

3 Contribution

The proposed approach aims to summarize data into a single graph model for heterogonous sources. It's a a schema-driven approach based on labeled graph. It allows also to link the graph model to the relevant domain knowledge to find relevant concepts in order to provide meaningful and concise summary. Last but not least, it provides a personalized visualization model capable to summarize graphically both the structure and the content of the data from databases, devices and sensors to reduce cognitive barriers related to the complexity of the information and its interpretation. To achieve this goal, our framework architecture is composed of four main modules as shown in 1:

A) **Data Pre-Processing module**: consists of processing and indexing data in order to summarize them. Every incoming data is processed and transformed according to two-steps: data cleaning and data semantization. This module is composed of:
   a) **Data Cleaning**: consists of preprocessing data and involves transforming raw data into an understandable format. It consists of data extraction from multiple and heterogeneous sources. Then, data cleaning is applied which is the most important task in building any analysis model. This includes outliers quantizing and handling missing values.
   b) **Data Semantization**: integrates semantics into preprocessed data by normalizing them based on an existing domain knowledge. Based on the heterogeneity characteristics of data, we propose an integrated framework composed of three processes:
      i) **Modeling unstructured data**: we use here NLP tools to identify data and to convert them into their appropriate data types.
      ii) **Mapping data with domain knowledge**: we combine here structured data with the previous process (unstructured data) output before mapping them with knowledge domain metadata for better normalization
      iii) **Integrating Data**: we integrate the different normalized data into a generic framework that supports direct generation of the data in a common format.

B) **Data Graph generation module**: it consists of transforming input data and generating aggregated items into a graph-based data model. We introduce here a new Data Graph Model (DGM) representing important structured and unstructured data in a domain. We define the DGM graph to efficiently represent a domain data and the relationship between them. the DGM model will be detailed in section 4.

C) **Data Summarization module**: defines the data summarization model-based graph, which is the core module of our framework. It allows to transform input data and generates the summary. So, it aims, firstly, at modeling through a data graph schema the most appropriate data that must be summarized. Secondly, summarizing data using a driven schema approach based on structure and content. The data summarization model-based graph will be detailed in next section. This module is composed by two sub-module:
   a) **Based content module** This sub-module provides a user-centered summarization model depending on the user preferences. Our goal behind this proposed graph summarization-based content is to provide data model adjusted based user preferences and needs. For this end, we define, a new node to allow users personalizing the content according to the analysis needs and preferences. We allow creating a calcul to one or many Data Nodes from the graph data GD to calculate a mathematic function from any number of incoming numeric values. Then the resulting score is placed in a new Data Node. The node result is related to the data nodes sources through a calculation Node, a condition Node, and a logic Node. The proposed graph summarization-based content will be detailed in section 5.
   b) **Based structure module** Summarization model based structure consists of summarizing the graph in terms of its topology in order to reduce the size and minimize the complexity of the graph and keep the important nodes and relations. It is called structural summarization. In order to summarize the graph structurally, we define new "Super-DataNode" and "Super-Edge" nodes. Our goal is to generate a summary network by grouping similar data nodes and finding hierarchical “Super-DataNode” (representing collection of data nodes) and “Super-Edges” (representing similarities between groups of Data Nodes). Once the graph GD constructed, the summarization process begins. The goal is to generate a smaller network: GDs = (DNs, DRs, Ls, Ts, As, Vs) from the original network GD = (DN, DR, L, T, A, V) such that data nodes representing similar/relevant nodes in GD are grouped into a single node (a "super-DataNode") in GDs. We call GDs = (DNs, DRs, Ls, Ts, As, Vs) a "summary network," where super-DataNodes (SuperDNs) are the groups of related data nodes and super-Edges

(SuperE) represent the average similarity between group of data nodes represented by the two end points. We obtain GDs via a series of "Assign" relations. The Assign Relation assigns data nodes to their super-nodes. This relation partitions the original network DG and groups each partition to form a Super-DataNode in the summary network GDS. We define three types of assign relation:

i) **Data Nodes aggregation relation** consists of summarizing the graph by aggregating the same data node types into super nodes.
ii) **Relation aggregation relation** consists of summarizing the graph by aggregating the same relationship types into super relations.
iii) **Compression relation** consists of defining graph summary from the input. by Minimizing the number of bits needed to describe the input graph via its summary

D) **Data Post-Processing module**: is responsible of the visual representation of data. It provides visual and interactive communication and includes the techniques to graphically present data so as to summarize and understand the meaning of data. Also, it allows to rapidly find insights in data.

The domain description module provides users (and mainly experts) the domain knowledge description to enable meaningful domain interoperability. This description provides the organizational and functional interoperability, the semantic interoperability using domain terminology, and the device data annotation using an existing IoT ontology.

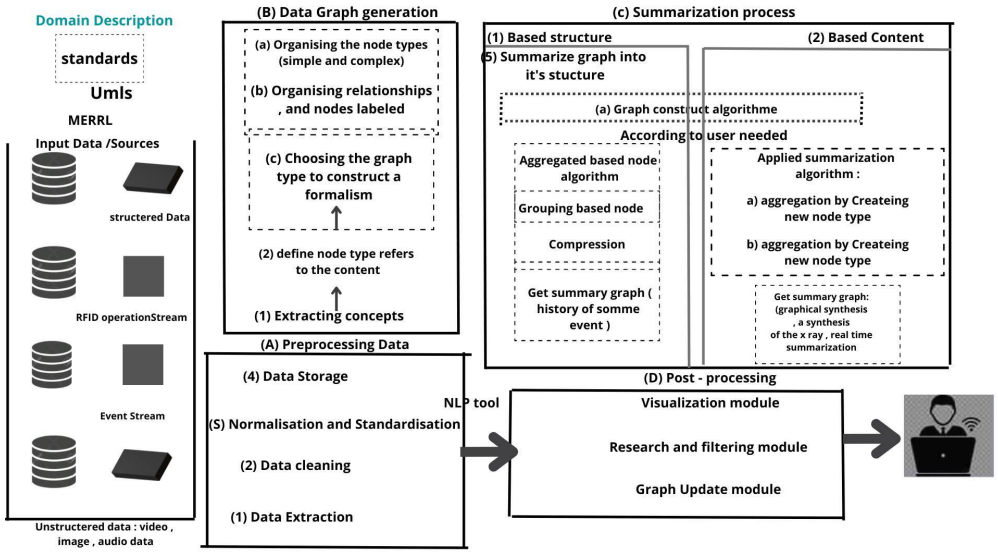

Fig. 1. Architecture of our proposed system

4 A Graph Data Model

An aggregation process is performed on the transformed input data and generates the aggregated value. Indeed, given a set of data items, a graph-based data model of aggregated items is iteratively built. The graph root is an aggregated item that represents the whole data set. Each aggregated item consists of one or more children which can be the original data items (leaves) or aggregated items (nodes). We introduce here a new Data Graph Model (DGM). The main goal behind DGM is to build and manipulate a common synthesis of a large amount of data to facilitate and perform the summarization process.

4.1 Data Graph Formalism

The Data Graph Model (DGM) represents important structured and unstructured data in a domain. We define the DGM graph to efficiently represent a domain data and the relationship between them.
Definition 1: Data Node A Data Node (DN) represents the information contained in a data structure. The data node contains a value of structured or unstructured data. Nodes are represented by a

single parent node. A Data Node (DN) is described by an identification, a name, and a type. We define here a data node which can be simple node or complex node.A data node is defined as follow:

DN: (IdN, NameN, Val, ValType, TypeN)

where

- IdN: is the node identifier
- NameN:is the node name
- Val: is the a value for each node attribute
- ValType : is the type of the value of each node attribute
- TypeN : is the type of node which can be simple Node or complex node We distinguish two types of Data Node:
  - Simple Data Node (SDN) is the most elementary unit. It can only be of the following ones: textual node, numerical node, Boolean node, image node, video node.
  - Complex Node (CDN) is composed of one or many simple and/or complex nodes.

Definition 2: Data Relationship: A Data Relationship (DR) connects two or many data nodes in the graph G. It is a directed edge consisting of an ordered pair of data nodes. It is characterized by a set of attributes, a role, a Data Node Source, a Data Node Destination. A relationship is defined as follow:

DR (IdR, NameR, nS, nD, Label)

Where

- IdDR: is the relation identifier
- NameDR is the relation's name
- dnS: is the node source of the relation ri
- dnD: is the node destination of the relation ri Label: is a word or a set of words used to describe the relation

Definition 3: Data Graph Model: We denote a graph GD as (DN, DR, L, T, A, V, ft, fa, fv, fr,) where N is the set of nodes, and R is a set of relationships.

Each Ri N * N representing the set of edges of a particular type. Nodes in a graph have a set of associated attributes, which is denoted as A Each node has a value for each attribute.These attributes are used to describe the features of the objects that the nodes represent.

DG = (DN, DR, L, T, A, V, ft, fa, fv, fr)

where :

- DN: is a set of (nodes) ni, denoting model Nodes Simple nodes and Complex Nodes.
- DR: is a set of relationships,
- L: is a set of edge labels li designating each a node or a relationship
- T: is a set of types of ti
- A: is a set of attributes ai
- V: is a set of values vi
- ft: N→ T is a function associating each node ni to its type (ft(ni))
- fa: N→A is a function associating each node ni to a set of attributes (fa(ni))
- fv: A→V is a function associating each type of attribute ai A to a possible value (fv(ai))
- fR: is a function defined on R, assigning a label from L to each edge in R
- fSr : R→ N is a function associating for each relation ri to its node source ((fSr(ri)))
- fD : R→ N is a function associating for each relation ri to its node destination ((fD (ri)))

Algorithm 1: Data Graph Formalism

Input: heterogeneous data: image, document, text, numeric
Output: DG

```
begin;
Construct data node;
Construct simple node and complex node:DN;
Construct relation types:DR;
Attribute the properties for node types:A;
Value of nodes: is a set of values vi ;
Attribute label for nodes (L);
for i ← 0 to the number of different data type do if (the content of data is one type) then
  typeT = simple data node : SDN
  else
  complex data node : CDN
  endif
  endfor
  For j ← 0 to the number of nodes N do fl : Defined label L ft : Defined type T fv : Defined value, V
  Endfor
  Return DG
  end
```

The algorithm Graph Formalism8 describes how to build our Data Graph DG into it simple nodes , complex nodes and relations.

5 Summarization-based content model

We Provide a user-centered summarization model depending on the user preferences. Our goal behind this proposed graph summarization-based content is to provide data model adjusted based user preferences and needs. For this end, we define, a new node to allow users personalizing the content according to the analysis needs and preferences.We can create a calcul to one or many Data Nodes from the graph data GD to calculate a mathematic function from any number of incoming numeric values. The node result is related to the data nodes sources through a calculation Node.

Definition 1: Calculation Node A calculation Node performs calculations on a single value. The following Calculation Nodes represent basic mathematical calculation: add, subtract, multiply, divide, exponent, remainder,average, count, last, Max, Min, Sum. A Calculation Node is defined as follows:

CalculNodeV (FirstV, CalculSymb, SecondVal, DisplayList)

Where:

- FirstVal: is the first input value that will be used in the calculation
- CalculSymb: the symbol that corresponds with the mathematical calculation that is performed by the node.
- SecondVal: is the second input value that will be used in the calculation
- DisplayList: determines the label that appears on the node in the policy model

A Data node result is generated contain respectively, The max value of diabetes for the first collection and the max value diabetes for the second collection. After That, we apply an Average Node to calculate the average value of the max values. Also we can integrate other nodes type proposed such as

- **Definition2: Logic node** And and Or that you can use in a policy model to specify whether or not policy execution should continue based on the results of the incoming logic paths. Specifically: The And node evaluates whether or not all incoming logic paths result in a value of true passed to the And node. The Or node evaluates whether or not at least one incoming logic path results in a value of true passed to the Or node.
- **Definition3: Conditional node** We define this type to apply a variety of conditions, calculations, and logic to the values represented by data nodes.
- **Definition4:Comparison node** To compare two values using the comparison operator that corresponds to the name of the node. The following comparison nodes are available: Equal, Not Equal, Greater Than, Greater Than or Equal, Less Than. In our work the based content summarization process focused in the numeric node , formalized by these different steps.

Algorithm 2: Based content summarization

Input: Data Graph
Output: summary result: GDs, supernodes , graphical, list
based content summarization applied;
For i = 1 to n ;
if typeDN=(numeric node);
applied operation node according to user needs;
case 1 : CalculNodeV (FirstV, CalculSymb, SecondVal, Displayresult);
case 2: ComparisonNode (FirstV, operatorSymb, SecondVal, Displayresult);
$\forall DGi, DGsj \in G, i = j, DGi \cap DGsj \neq \varnothing$
$Returnresult$
end

6 Summarization based structure model

In this model is to summarize the graph in terms of its topology, the objective of which is to reduce the size and minimize the complexity of the graph and keep the important nodes and relations. We note it structural summarization. Graph structure is prominently used in summarization technique (compression, grouping, simplification, visualization). In order to summarize the graph structurally, we define new Super-Data-Node and Super Edge. Our goal is to generate a summary network by grouping similar data nodes and finding hierarchical "Super-Data-Node" and "Super-Edges" . Once the graph GD constructed, the summarization process begins. The goal is to generate a smaller graph GDs = (DNs, DRs, Ls, Ts, As, Vs) from the original graph GD = (DN, DR, L, T, A, V) such that data nodes representing similar/relevant nodes in GD are grouped into a single node (a "super- node") in GDs.

Algorithm 3: Structured summarization

Input: Data Graph
Output: Data graph summary DGs
Structured summarization applied;
For i= 1 to n;
regrouping nodes that have same types;
if DN=textual node, DN=numeric node or DN=image node;
Regrouping nodes in the supernode;
DGs = DNs1, DNs2,..DNsk;
if $\forall DGi \in DG, DNs(DGi) \cap DNs(DGs) and DGi \neq \emptyset$
$\forall DGi, DGsj \in DG, i = j, DGi \cap DGsj \neq \varnothing$
$Returnsupernodeof SDNtypesregrouping$
end

7 Experimentation and results

In this section we provide an experimental study of our proposed approach and analysis of the proposed algorithms and operational nodes. We have developed a prototype of our methodology implemented in NEO4J visualization and Python database. Graphs GD are stored using the following formalism proposed and described in 4.1.

7.1 Datasets and Experimental Setup

Our scenario is described by various essential steps for our approach, the first step consists of loading heterogeneous information containing a patient's medical file in several formats (Word doc, PDF doc, xlsx, image, video, audio) that represented our heterogeneous database. The second step is the building of a data graph that describes how to model heterogeneous data into graph: each node follows the formalism proposed (data node, data relationship, calculation node, logic node, condition node) and each node type has its own color (exp image node refers the red color shows in fig2textual node refers the green color3. The summarization process, in this context we have proposed two types of tha graph summary. The first one a based content which is interested in summarizing the digital measurements

(temperature, blood measure, glucose level) coming from the 3 sensors. The result of this type of synthesis shows us either the maximum value, the minimum value, the average of these measurements during the period mentioned (1 month)5 or a curve which interprets the variations of the measurements6. The second type of summarization is interested in complex nodes, we start with an extraction and a cleaning is explained that a PDF document will be extracted in cleaned pages (texts apart and images apart) afterwards each text will be accompanied by an annotation (node annotation) which comes from mapping with the UMLS dictionary to facilitate the diagnosis and decision-making of doctor.
For structural summarization it's an aggregation operation and compression algorithm, node based attribute or relationship .The result of summary graph visualization containing only linked nodes (an attribute that expresses user need such as a particular disease or to Summarize medical prescriptions of X-ray interpretation 3 figured in textual node (in green color) or image node in (red color) to summarize only X-ray. Or an aggregation of images (ultrasounds by date) presented with pink color Or aggregation to display a summary graph mentioned the last nodes visualised (patient history)4.

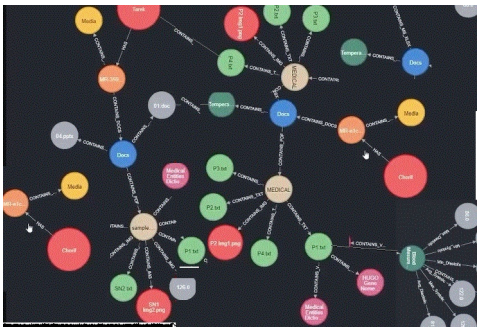

Fig. 2. Electronic health record model represented by data graph

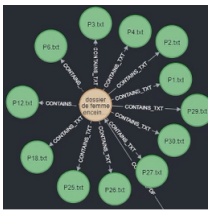

Fig. 3. structured summarization (based on aggregation node(same type of node))

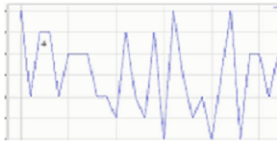

Fig. 5. based content summarization visualized in graphical node

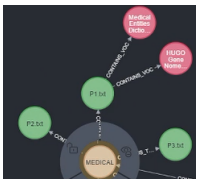

Fig. 4. structured summarization (based on aggregation by attribute

Fig. 6. based content summarization using calculation node(max,min,avg)

7.2 Evaluation

For our benchmarking, we used formalism graph data(GD) for medical database. Moreover, we known that for graph summarization based on content there are not yet evaluation metrics to use in the literature. We have proposed two metrics the running time and the loss of information to evaluate our approach . First step for the content summarization we associate this evaluation: In fig7, we depict

the impact of the variation of the number of node and relationships on the summarizing algorithms. We compared our algorithm in term based content summarization to other ones based structure in the literature. We choice the approach proposed in the literature[19] using execution time metric .We considered the execution time of our algorithm always remains low and this guarantees its applicability to large graph (nodes, relationships) and that shows good performance for our approach. In fig 8we compare our algorithms from aggregation nodes ,aggregation relationships with algorithm of approach[19] using a loss of information metric. we noted that our approach is more efficient(yellow curve and orange curve) keep a large percentage of content graph (nodes and relationships).

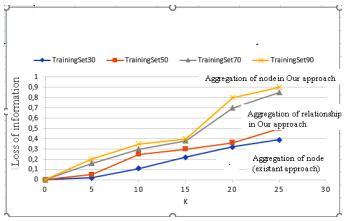

Fig. 7. Relative improvement on runtime between based content summary(our approach) vs structured summarization approach

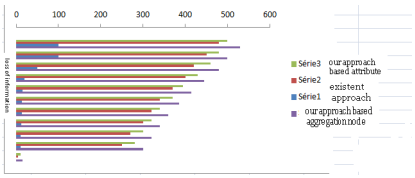

Fig. 8. Relative improvement on loss of information between graph summary aggregated (node , relationship) in our approach vs based ksnap graph summary

8 Conclusion and Future work

In this work, we study utility-driven graph summarization in-depth and made several novel contributions. We present a new, lossless graph summary, the first one structured based and the second one content based. Moreover we introduced our approach by the formalism proposed of data graph into heterogeneous data in the input. We proposed four main operations to the summarization process. We design a scalable, lossy summarization algorithm in our experimentation, based on two principal metrics the running time and the non-loss of information. Finally, the problem of graph summarization has been extensively addressed for existing graph data models, such as static, labeled, and weighted graphs that mentioned in our related works. Furthermore, our interesting future direction would be to investigate quality metrics for summaries and evaluation benchmarks for structured graph summary and based content. Also to ameliorate summarization process with integration of all operation node types proposed.

References

1. Boldi, P., Rosa, M., Santini, M., & Vigna, S. (2011, March). Layered label propagation: A multiresolution coordinate-free ordering for compressing social networks. In Proceedings of the 20th international conference on World Wide Web (pp. 587-596)
2. Cudré-Mauroux, P., & Elnikety, S. (2011). Graph data management systems for new application domains. Proceedings of the VLDB Endowment, 4(12), 1510-1511.
3. Hooper, S. D., & Bork, P. (2005). Medusa: a simple tool for interaction graph analysis. Bioinformatics, 21(24), 4432-4433.
4. Barceló, P., Pérez, J., & Reutter, J. L. (2012). Relative Expressiveness of Nested Regular Expressions. AMW, 12, 180-195.
5. Tian, Y., Hankins, R. A., & Patel, J. M. (2008, June). Efficient aggregation for graph summarization. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 567-580).
6. Tang, N., Chen, Q., & Mitra, P. (2016, June). Graph stream summarization: From big bang to big crunch. In Proceedings of the 2016 International Conference on Management of Data (pp. 1481-1496).
7. Fan, W., Li, J., Wang, X., & Wu, Y. (2012, May). Query preserving graph compression. In Proceedings of the 2012 ACM SIGMOD international conference on management of data (pp. 157-168).
8. Feigenbaum, J., Kannan, S., McGregor, A., Suri, S., & Zhang, J. (2009). Graph distances in the data-stream model. SIAM Journal on Computing, 38(5), 1709-1727.

9. Han, W., Miao, Y., Li, K., Wu, M., Yang, F., Zhou, L., ... & Chen, E. (2014, April). Chronos: a graph engine for temporal graph analysis. In Proceedings of the Ninth European Conference on Computer Systems (pp. 1-14).
10. Kang, U., & Faloutsos, C. (2011, December). Beyond'caveman communities': Hubs and spokes for graph compression and mining. In 2011 IEEE 11th international conference on data mining (pp. 300-309). IEEE.
11. Kang, U., Tong, H., Sun, J., Lin, C. Y., & Faloutsos, C. (2011, August). Gbase: a scalable and general graph management system. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1091-1099).
12. Huang, J., Abadi, D. J., & Ren, K. (2011). Scalable SPARQL querying of large RDF graphs. Proceedings of the VLDB Endowment, 4(11), 1123-1134.
13. Khan, K. U., Nawaz, W., & Lee, Y. K. (2017). Set-based unified approach for summarization of a multi-attributed graph. World Wide Web, 20(3), 543-570.
14. Shah, N., Koutra, D., Zou, T., Gallagher, B., & Faloutsos, C. (2015, August). Timecrunch: Interpretable dynamic graph summarization. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1055-1064).
15. Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: a survey. Data mining and knowledge discovery, 29(3), 626-688.
16. Toivonen, H., Zhou, F., Hartikainen, A., & Hinkka, A. (2011, August). Compression of weighted graphs. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 965-973).
17. Mathioudakis, M., Bonchi, F., Castillo, C., Gionis, A., & Ukkonen, A. (2011, August). Sparsification of influence networks. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 529-537).
18. Dunne, C., & Shneiderman, B. (2013, April). Motif simplification: improving network visualization readability with fan, connector, and clique glyphs. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 3247-3256).
19. Tian, Y., & Patel, J. M. (2008, April). Tale: A tool for approximate large graph matching. In 2008 IEEE 24th International Conference on Data Engineering (pp. 963-972). IEEE
20. Shen, Z., Ma, K. L., & Eliassi-Rad, T. (2006). Visual analysis of large heterogeneous social networks by semantic and structural abstraction. IEEE transactions on visualization and computer graphics, 12(6), 1427-1439.
21. Lebanoff, L., Song, K., & Liu, F. (2018). Adapting the neural encoder-decoder framework from single to multi-document summarization. arXiv preprint arXiv:1808.06218.
22. Cook, D. J., & Holder, L. B. (2000). Graph-based data mining. IEEE Intelligent Systems and Their Applications, 15(2), 32-41.
23. Raghavan, S., & Garcia-Molina, H. (2003, March). Representing web graphs. In Proceedings 19th International Conference on Data Engineering (Cat. No. 03CH37405) (pp. 405-416). IEEE.
24. Zhang, N., Tian, Y., & Patel, J. M. (2010, March). Discovery-driven graph summarization. In 2010 IEEE 26th international conference on data engineering (ICDE 2010) (pp. 880-891). IEEE.
25. Adhikari, B., Zhang, Y., Amiri, S. E., Bharadwaj, A., & Prakash, B. A. (2017). Propagation-based temporal network summarization. IEEE Transactions on Knowledge and Data Engineering, 30(4), 729-742.
26. Tang, N., Chen, Q., & Mitra, P. (2016, June). Graph stream summarization: From big bang to big crunch. In Proceedings of the 2016 International Conference on Management of Data (pp. 1481-1496).
27. Tan, J., Wan, X., & Xiao, J. (2017, July). Abstractive document summarization with a graph-based attentional neural model. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1171-1181).
28. Maccioni, A., & Abadi, D. J. (2016, August). Scalable pattern matching over compressed graphs via dedensification. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
29. Shi, L., Tong, H., Tang, J., & Lin, C. (2015). Vegas: Visual influence graph summarization on citation networks. IEEE Transactions on Knowledge and Data Engineering, 27(12), 3417-3431.
30. Fan, W., Li, J., Wang, X., & Wu, Y. (2012, May). Query preserving graph compression. In Proceedings of the 2012 ACM SIGMOD international conference on management of data (pp. 157-168).
31. Qu, Q., Liu, S., Jensen, C. S., Zhu, F., & Faloutsos, C. (2014, September). Interestingness-driven diffusion process summarization in dynamic networks. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases.