JCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Volume-5, Issue-11 E-ISSN: 2347-2693

Review Paper on Graph Based Approach for Mining Health Examination Records Using Views

Reshma Ravi[1*], **Remya R**[2]

[1*]Dept. Of CSE and IT, College Of Engineering Perumon, APJ Abdul Kalam Technological University, Kerala ,India
[2]Dept. Of CSE and IT, College Of Engineering Perumon, APJ Abdul Kalam Technological University, Kerala ,India

**Corresponding Author: reshmaravi.cep@gmail.com*

Available online at: www.ijcseonline.org

Received: 23/Oct/2017, Revised: 30/Oct/2017, Accepted: 16/Nov/2017, Published: 30/Nov/2017

***Abstract*—** Answering Queries using Views is proven as an effective technology for querying real life graphs. Real life graphs are really large, so if a query arises from such graph it's a troublesome process. Answering using views is an easy method. When SHG health algorithm is combined with answering queries using views, we can analyze the medical data and based on that data we can predict whether a health examination participant is at risk, if yes what the key associated disease category is. This helps to predict the risks at an early stage. Medical data are usually large and distributed. So we use efficient algorithms like maximally contained rewriting, Minimal containment along with the SHG algorithm to analyze medical data. Semi supervised Heterogeneous algorithm is an efficient algorithm. Maximally contained rewriting algorithm helps to find an approximate answer to the query even if it is not contained in the views.

***Keywords*—**Pattern containment,SHG,Minimal Containment

I. Introduction

Answer Generation using views has been studied for relational, XML and semi structured data. If a query Q and a set of views V= {V_1, V_2,…..,V_n}is given, the idea is to find another query A such that A is equivalent to Q and also A only refers to views in V. If such a query A exists then Q (D) can be answered using A without accessing database. Answering queries using views helps to find answer for real life social graphs easily because they are typically large and distributed. This method helps to query big data, irrespective of the size of underlying data. Graph pattern queries have been widely used in social network analysis. The most important problem related to the real word social graphs are large size and usually they are distributed.

For e.g. Amazon has more than 2 million users with 140n billon links, here the data is distributed to different data centres across the world. The major challenge for the social network analysis is how to cope up with the large size of social graphs. Answering pattern queries using views provided an efficient solution for this problem.

In traditional approach answering pattern queries from graphs takes O ($|Q_S|^2$ + $|Q_S||G|+|G|^2$) time to compute Q_S (G).To identify a match set we have to perform a number of join operations. This increases the computational time .The advantage of our technique is that we only need to visit the views in V (G), without accessing the large graph all the time. The major aspect related to answering queries using views is to decide whether a given pattern query can be answered using a set of views. If a pattern query Q_s and a set of views V= {V_1,……..,V_n} is given, then Q_s can be answered if and only if Q_s is contained in V. For this a notion of pattern containment is proposed instead of the traditional query containment. This help to compute Q_s in O (|Qs||V (G)| +| V (G)|^2) time without accessing G. This helps to find Q_s in minimum computational time than in the traditional approach.

To decide which views in V to use we identified three fundamental problems. They are (1) Containment Problem (to decide whether a given query is contained in views). (2) Minimal containment (to find out a subset of V that minimally contains Q_s). (3) Minimum containment (To find out a minimum subset of V that contains Q_S). Maximally contained rewriting helps to find approximate answer for the query Qs, even if it is not contained in the views. If a query Q_s and a set of views V= {V_1,……,V_n} is given and Q_s is not contained in views. Then we have to find another Query Q_s^1 such that it is a sub query of Q_s and also it is contained in the set of views V. Query processing has two view based approaches. They are query rewriting and query answering. If a query Q and a set of views is given, the idea of query rewriting is to find another Q^1, if the query is not contained in the views. This helps to find approximate answer for the query. Query answering is to find Q_s using another query A which is equivalent to Q.

This Paper helps to predict the risk of medical examination participant at an early stage. It will increase the chance of recovery to a great extend. Usually medical data are really large, so for mining such large data we use Pattern containment algorithms. Maximally contained rewriting algorithm helps to find an appropriate answer to the query.

The work in this paper is organized by describing the related works in section 2. Section 3 describes the methodologies of our paper followed by result in section 4. Section 5 consists about the proposed work and conclusion .

II. RELATED WORK

In [2] studied about answering queries for XML data has been studied .This paper tries to overcome the drawbacks of the previous papers, by setting the inverted list model for evaluating queries on a large persistent XML data. In this approach for materializing the views only those XML tree nodes that occur in the answer to the view is included in the inverted list model. A new time and space efficient algorithm is developed for answering queries. Optimization techniques are proposed to minimize the storage space and also the redundant views can be avoided using Bitmaps.

In [3] studied about doable and undoable for distributed graph simulation. This approach provides algorithm whose response time and data shipment are not a function of G. Experimental studies shows that these algorithms scale well with the large real world graphs. Also studies show that the Distributed simulation is Partition Bounded. This paper studied about the fundamental problems of the graph simulation. Proposed possibility and impossibility theorem. Scalability and efficiency of the algorithm is checked.

In [4] a new algorithm is proposed for rewriting the semi structured queries (Q) that access the semi structured views V and are equivalent to Q. In the first step the content mappings (which are used to produce the candidate rewriting) are used. And checks whether the composition is equivalent to the original query or not. This technique is complicated due to the lack of the schema and of structuring capabilities of TSL views. Our algorithm uses Containment mappings, the Chase and the query composition for efficient query answering.

In [5] the research has been done to revisit the Tree Pattern [TP] queries .The answer which is obtained by evaluating the annotated views is similar to the answer over the original View Analysis. The study proposed an algorithm for identifying the redundant view answers. In this approach V is divided into a finite set of sub views and finding MCR s of Q using each of the Sub views. In this approach a new technique is proposed for identifying the redundant view answers, which can be ignored while evaluating the maximally contained Rewriting.

In [6] a new method is proposed for incremental solutions for graph pattern matching based on simulation sub graph isomerism and bounded simulation. This approach developed incremental algorithms for batch updates and patterns. Incremental graph pattern matching helps to identify the answers to the queries even though the views are updated. This technique helps to apply graph simulation independent of the graph size. Incremental algorithms help to calculate changes in the matches when the graph is updated this helps to minimize unnecessary recommendation.

III. METHODOLOGY

Query answering and rewriting. There are two view-based approaches for query processing: query rewriting and query answering. Given a query Q and a set V of views, (1) query rewriting is to reformulate Q into an equivalent query Q0 in a fixed language such that for all D, Q(D) = Q0(D), and moreover, Q0 refers only to V; and (2) query answering is to compute Q(D) by evaluating a query A equivalent to Q, while A refers only to V and its extensions V(D). While the former requires that Q0 is in a fixed language, the latter imposes no constraint on A.

Answering pattern queries using views is an effective technique for querying large graphs. In this paper query answering using views has been done with the help of graph simulation. A notion of pattern containment is used instead of Query containment to characterize the pattern query problems. It is sufficient to determine whether a query Q_s is contained in the views. Algorithm Matchjoin is to find a match set between the data graph (G) and the pattern query (Q_s).

Approximation algorithms like contain, minimal and minimum are used to find out whether a query is contained in the view or not. Minimum containment helps to effectively reduce the redundant views. When the query is not contained in the views our approximation algorithm helps to find the answer at reasonable accuracy. Optimization strategy in this paper makes the view base matching up to 1.66 times faster. Minimal algorithm is defined as follows; the algorithm returns either a non-empty subset V1 of V that minimally contains Q_s or 0 to indicate that the Q_s is not contained in the views. Minimum algorithm identifies a subset V1 of V such that (1) Qs is contained in V1 if Q_s is contained in V and card (v1) ≤log (|Ep|), card $V_{opt.}$

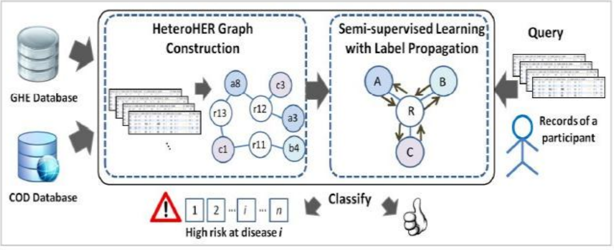

Fig.1 SHG-Health Algorithm for risk predication

Our SHG-Health algorithm takes health examination data (GHE) and the linked cause of death (COD) labels as inputs. Its key components are a process of Heterogeneous Health Examination Record (HeteroHER) graph construction and a semi-supervised learning mechanism with label propagation for model training. Given the records of a participant as a query, SHG-Health predicts whether patient falls into any of the high-risk disease categories or "unknown" class whose instances do not share the key traits of the known instances belonging to a high-risk disease class.

SHG HEALTH ALGORITHM

Input: select of health examination records
1. Heterogeneous graph construction
2. Normalised weight calculation
3. Update weight for new records
4. Semi supervised learning

Output: optimized computed soft labels (risk prediction)
A graph representation allows us to model data that is sparse. To capture the heterogeneity naturally found in health examination items, we constructed a graph called HeteroHER consisting of multi-type nodes based on health examination records. The process of HeteroHER graph construction includes the following steps:

Step 1. Binarization: As a preparatory step, all the record values are first discretized and converted into a 0/1 binary representation, which serves as a vector of indicators for the absence/presence of a discretized value.

Step 2.Node Insertion: Every element in the binary representation obtained in Step 1 with a value "1" is modeled as a node in our HeteroHER graph, except that only the abnormal results are modelled for examination items (both physical and mental).

Step 3. Node Typing: Every node is typed according to the examination category that its original value belongs to, for example, the Physical tests (A), Mental tests (B), and Profile (C).

Step 4. Link Insertion: Every attribute (non-Record) type node is linked to a Record type node representing the record that the observation was originally from. The weight of the links is calculated.

A simple function g(t) can be defined as:

$$g(t) = (t - s + 1)/l \quad (1)$$

where t is the time of current record, l is the time window of interest, and s is the starting time of the time window.

IV . RESULTS AND DISCUSSION

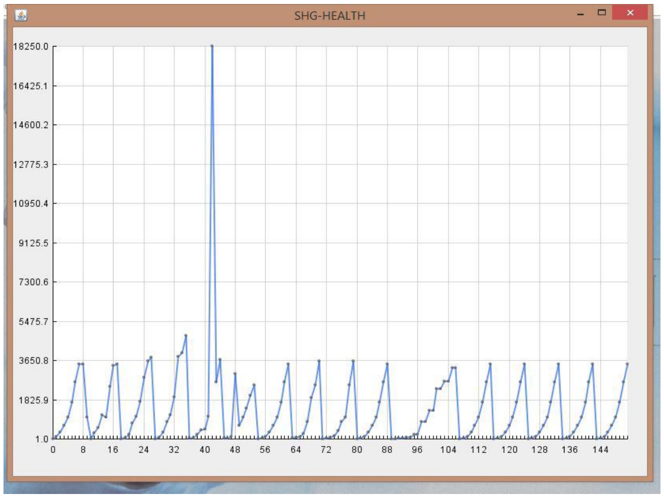

Fig.2. Result

SHG health algorithm helps to predict the risk of health patients. In this first stage of the two-stage evaluation, we compared algorithms based on their abilities to identify high-risk cases regardless of what disease category they belonged to. In the second stage, we further evaluated the algorithms' conditional performance on multi-class classification. Only the cases that were predicted into one of the disease classes were considered.

V. CONCLUSION

Answering pattern queries using views helps to find out answer from real life graphs. Mainly three problems are addressed; they are pattern containment problem, minimal containment problem and the maximally contained rewriting. SHG algorithm is used for early risk prediction .A graph G and a set of views is given, then a query can be answered using views if the query is contained in the views. Maximally contained rewriting helps to find the answer if the query is not contained in the views. Minimal containment problem is used to find out the minimum set of views that can be answered using the views. Semi supervised Heterogeneous Graph based algorithm is used to predict the risk of a participant. First a heterogeneous graph is constructed based on the dataset, and then using semi supervised learning method risk prediction is analyzed. This paper helps to predict the risk of health patient by comparing big set of data. As future work new algorithms will be designed to increase the efficiency of our proposed system.

REFERENCES

[1] A.Y. Halevy's, "*Answering queries using views: A Survey*, "VLDBJ, vol.10, no.4, pp-270-294 2001.

[2] X. Wu, Theodoratos and W .H. Wang, "*Answering XML queries using materialized views revisited*", in proc.18th ACM conf .inf. Knowl. Manage. , 2009, pp.475-484.

[3] W. Fan, X. Wang, and Y. Wu, "*Distributed graph simulation: impossibility and possibility"*, proc. VLDB Endowment, vol.7, no.12, pp.1083-1094, 2014.

[4] Y. Papakonstantinou and V. Vassalos , "*Query rewriting for the semi structured data,"* in the ACM SIGMOD int. conf. Manag. Data, 1999, pp. 455-466.

[5] J. Wang, J, J, X. Yu' and J.Li's paper, "*Answering tree pattern queries using views: A revisit,"* in the 14 th Int. conf. Extending Database Technol., 2011, pp. 153-164.

[6] W. Fan, X. Wang, and Y. Wu, "*Incremental graph pattern matchin*g ," ACM Trans. Database syst., vol.38, no. 3,2013.

[7] W. Fan, J. Li, X. Wang, and Y. Wu, "*Query preserving graph compression,*" in proc. ACM SIGMOD Int. conf. Manag . Data, 2012, pp. 157-168.

[8] R . Pottinger and A. Y. Levy, "*A scalable algorithm for the answering using views,"* Very Large Data Bases, of 2000, pp. 485.

[9] D. Calvanese, G. D. Giacomo, , M. Y. Vardi and M. Lenzerini, : "*View based query processing and the constraint satisfaction,"* in Proc. 15th Annu . IEEE Symp. Logic Comput. Sci., 2000, pp. 475-484.

[10] Y. Zhuge and H. Garcia-Molina , *" Graph structured views and their Incremental maintenance algorithm ,"* in 14th Int .Conf, held at 1998, pp. 116-125.

[11] M. Muralidharan, V.Valli Mayil, "*A Study of Natural Language Processing Procedures*", International Journal of Computer Sciences and Engineering, Vol.5, Issue.6, pp.300-304, 2017.