Semantic Data Extraction Towards Relational Databases Using Reinforcement Learning Approach

Sruthimol. A [1] and S.P.Victor [2]
[1]*Research Scholar, Manonmaniam Sundaranar University, Tirunelveli,*
[2]Associate Professor, Department of Computer Science, St.Xaviers College, Tirunelveli.

Abstract
Relational databases are step ahead of normal databases in which the process of accessing the data in particular requirement formats are little bit technically advanced when compare it with the normal procedure from users point of view. The relation databases helps the data viewers and handlers i an easier way of accessing the information's in order to save time and space. The accuracy and clarity the primary constraint in relational databases data extraction procedures. Reinforcement learning is the familiar strategy used in neural networks which can handle the signal and data process based on optimal output tightening by using the repeated training or reinforcing strategies to acquire the expected results. This paper deals with the process of extracting data from relational databases use the reinforcement learning approach for the accuracy and time saves strategies towards optimal output attainments. The mixture of neural reinforcement in database extraction procedures is improving the efficiency in terms of accuracy. In near future we will implement the fuzzy logic combined with neural networks approach in relational databases data extraction methodologies.

Keywords: *data mining, data extraction, relational database, reinforcement learning, optimality.*

I.INTRODUCTION

Relational Database:
A relational database is a type of database that stores and provides access to data points that are related to one another. Relational databases are based on the relational model[1], an intuitive, straightforward way of representing data in tables. In a relational database, each row in the table is a record with a unique ID called the key[3]. The columns of the table hold attributes of the data, and each record usually has a value for each attribute, making it easy to establish the relationships among data points[4]. A relational database is a set of formally described tables from which data can be accessed or reassembled in many different ways without having to reorganize the database tables. The standard user and application programming interface (API) of a relational database is the Structured Query Language (SQL)[7,8].

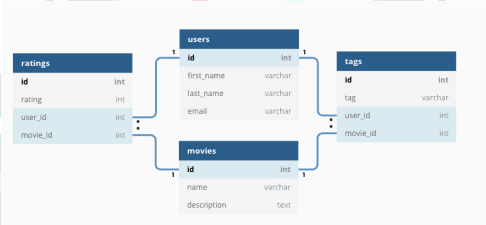

Fig-1: Sample relation Database table

Data Extraction:
Data extraction is a process that involves retrieval of data from various sources. Frequently, companies extract data in order to process it further, migrate the data to a data repository (such as a data warehouse or a data lake) or to further analyze it[2]. It's common to transform the data as a part of this process.The extraction methods in relational databases depend on the source system, performance and business requirements[5].

Reinforcement Learning:
Reinforcement Learning (RL) refers to a kind of Machine Learning method in which the agent receives a delayed reward in the next time step to evaluate its previous action[6]. It was mostly used in games (e.g. Atari, Mario), with performance on par with or even exceeding humans. Recently, as the algorithm evolves with the combination of Neural Networks, it is capable of solving more complex tasks, such as the exact data extraction problem[9].

Fig-2: Data Integrity Features

Then environment refers to the object that the agent is acting on (e.g. the game itself in the Atari game), while the agent represents the RL algorithm. The environment starts by sending a state to the agent, which then based on its knowledge to take an action in response to that state. After that, the environment send a pair of next state and reward back to the agent. The agent will update its knowledge with the reward returned by the environment to evaluate its last action. The loop keeps going on until the environment sends a terminal state, which ends to episode[10].

II.PROPOSED METHODOLOGY

The following figure-3 represents the proposed methodology for reinforcement approach based data extraction from relational database tables. There are two types of policies available in reinforcement learning approaches. They are
a. Online policy-current action derived from current state outputs.

B.Offline policy-current action derived from available existing outputs.

SARSA (State-Action-Reinforcement-State-Action) is an online policy whereas Quality-Learning is an offline policy.

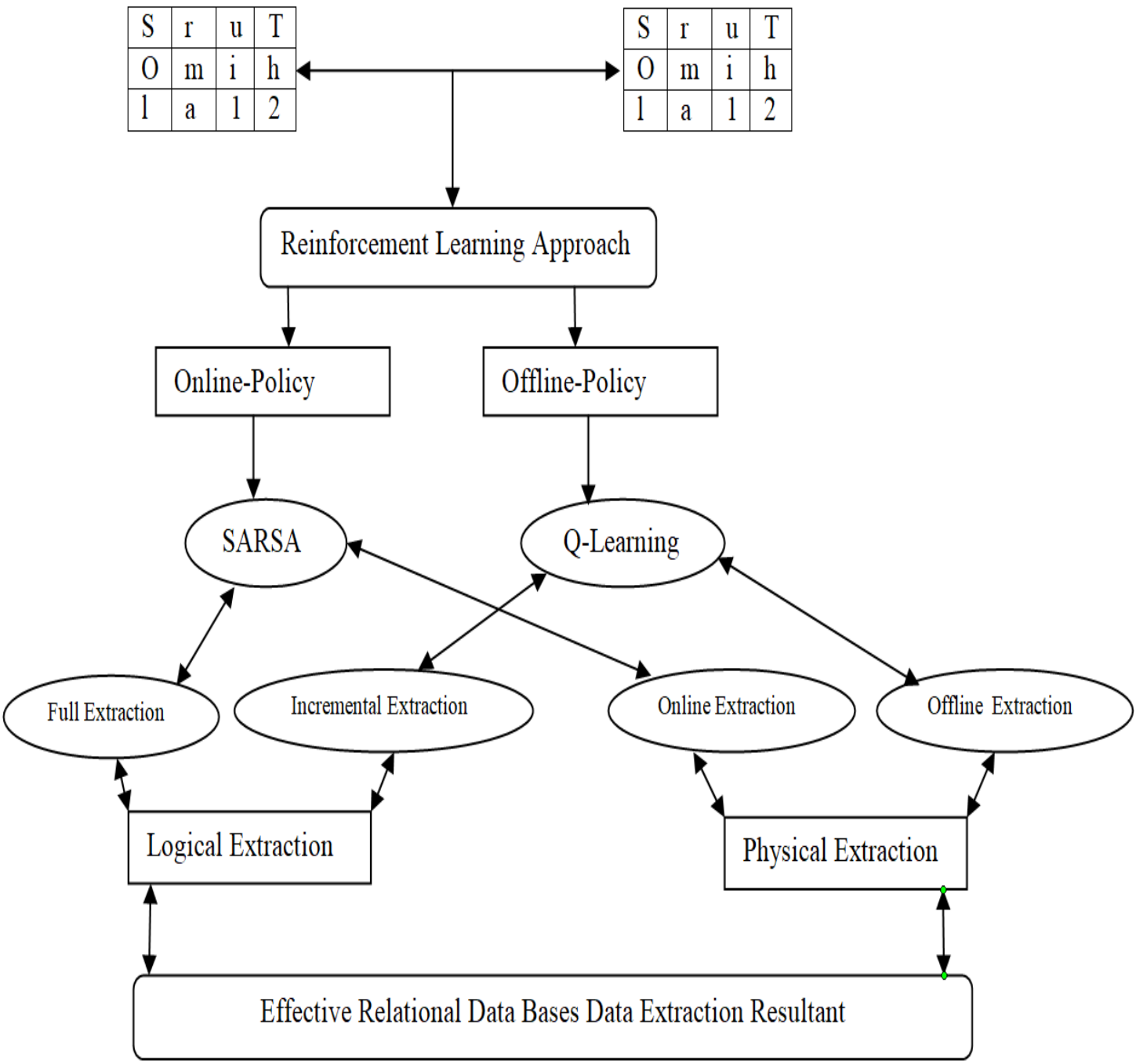

Fig 3: Proposed Data Extraction towards Relational Databases using Reinforcement Learning

There are two primary types of data extraction, they are Logical and Physical extractions such that full and incremental (part by part) belongs to logical along with online and offline belongs to physical extraction respectively.

III. IMPLEMENTATION

Consider the Relational tables for Employees Official and Personal Tables Architecture as follows,

Table-1: Employee Official table

Field Name	Description	Data Type
Emption	Identification Number(Primary Key)	Number
Emp_NAM	Short Name of the Employee	Char(4)
Emp_DEP	Department Name	Char(10)
Emp_DES	Designation	Char(10)
Emp_GEN	Male/Female/Transgender	Char(10)
Emp_DOB	Date of Birth	Date
Emp_SAL	Monthly Salary	Number
Emp_EDN	Education Level-School/UG/PG/PhD	Char(10)
Emp_BGP	Blood Group	Char(5)
Emp_MOB	Mobile Number	Number(10)

The following table represents the employee personal details which can be identified through he primary key as employee ID number.

Table-2: Employee Personal table

Field Name	Description	Data Type
Emp_IDN	Identification Number(Primary Key)	Number
Emp_FNM	First ,Mid and Last Name of the Employee	Varchar2(30)
Emp_ADD	Residential Address	Varchar2(60)
Emp_EDU	Complete Education	Char(10)
Emp_REF	Job reference/Self	Char(10)
Emp_PRM	Physical remarks	Varchar2(20)
Emp_EXP	Total Experience in years	Number
Emp_MST	Marraige Status	Boolean
Emp_CLD	Children Count	Number
Emp_SKL	Any other proficiency-Lang/Skill	Char(20)

The sample data for the employee official database table is represented in table-3.

Table-3: Employee Official table sample data

Emp_IDN	Emp_NAM	Emp_DEP	Emp_DES	Emp_GEN	Emp_DOB	Emp_SAL	Emp_EDN	Emp_BGP	Emp_MOB
1001	STK	Production	Manager	Male	12/10/1978	55000	3	A+	5876543210
1002	NSD	Production	Supervisor	Male	09/06/1975	40000	2	B+	5876543211
1003	PRB	Production	Technician	Female	05/08/1980	25000	1	O-	4758690102
1004	SLM	Sales	Manager	Male	16/12/1982	55000	3	O+	4645484749
1005	ELT	Sales	Supervisor	Male	14/05/1985	40000	2	A+	4142434445
1006	AEC	Accounts	Head	Female	29/11/1981	60000	3	A-	3435363738
1007	RSJ	HR	Manager	Female	17/07/1975	50000	3	B+	2325262829

The sample data with corresponding information stored for the employee personal database table is represented in table-4.

Table-4: Employee Personal table sample data

Emp_IDN	Emp_FNM	Emp_ADD	Emp_EDU	Emp_REF	Emp_PRM	Emp_EXP	Emp_MST	Emp_CLD	Emp_SKL
1001	Satish K	5,Chennai	MSc	Self	Wart on face	18	1	3	Driving
1002	Nirmal Singh D	6.Nellai	BSc	STK	Left hand 6th finger	12	1	2	Hindi
1003	Priya B	7,Palay	Diploma	NSD	Cut on forehead	4	0	0	Nursing
1004	Shakul M	8,Mumbai	MSc	Self	Left leg short	17	1	2	Driving
1005	Edwin Lawrence T	9,Madurai	BSc	MD	Height=6.2" Green eyes	11	0	0	Arts
1006	Abila Elizabeth C	10,Trichy	MSc	MD	Blue eyes short	16	1	1	Computer
1007	Richlin Selene J	11.Salem	MSc	Self	Left hand 4 fingers	20	1	1	Speaking

The reinforcement learning schema act as a situation based approach which can be implemented in the following way of association. There are 4 states of relational database table accessibility nature. They are

a. Initial preloaded state.
b. Online uploading state.
c. Uploaded state.
d. Downloaded state.

The proposed methodology is a mixture of state based action and reinforcement which can be implemented with its maximum level of flexibility to handle the relational database table in any state for processing.

Now Data extraction using logical extraction schema:
The customized query is as follows,

List out the employees with PG qualification and able to drive the vehicle.

a. If Relational Databases are in the form of initial preloaded state, then Full extraction schema will yield the optimal results (SARSA)
Full extraction:

SELECT EMP_IDN FROM EMPOFFICE WHERE EMP_EDN=3 INTERSECT
SELECT EMP_IDN FROM EMPPERS WHERE EMP_SKL='Driving';

The resultant record is: 1001
1004

b. If Relational Databases are in the form of online uploading state, then Incremental extraction schema will yield the optimal results (Quality learning)

Incremental Extraction:
SELECT EMPOFFICE.EMP_IDN,EMPPERS.EMP_FNM FROM EMPOFFICE
INNER JOIN EMPPERS
ON EMPOFFICE.EMP_IDN=EMPPERS.EMP_IDN WHERE EMPPERS.EMP_SKL='Driving';

Table-5: Query resultant data-1

Emp_IDN	Emp_FNM	Emp_EDU	Emp_SKL
1001	Satish K	MSc	Driving
1004	Shakul M	MSc	Driving
1006	Abila Elizabeth C	MSc	Computer
1007	Richlin Selene J	MSc	Speaking

Table-5: Query resultant data-2

Emp_IDN	Emp_FNM	*Emp_EDU*	*Emp_SKL*
1001	Satish K	*MSc*	*Driving*
1004	Shakul M	*MSc*	*Driving*

c. If Relational Databases are in the form of online web uploaded state, then online extraction schema will yield the optimal results (SARSA)

Online Extraction:

SELECT EMPOFFICE.EMP_IDN,EMPPERS.EMP_FNM FROM EMPOFFICE
FULL JOIN EMPPERS
ON EMPOFFICE.EMP_IDN = EMPPERS.EMP_IDN WHERE EMPPERS.EMP_SKL='Driving';

Table-5: Compact Query resultant data

Emp_IDN	Emp_FNM
1001	Satish K
1004	Shakul M

d. If Relational Databases are in the form of downloaded state, then Offline extraction schema will yield the optimal results (Quality learning).

Offline Extraction:
SELECT EMP_FNM FROM EMPPERS WHERE EMPPERS.EMP_IDN= (SELECT EMP_IDN FROM EMPOFFICE WHERE EMP_EDN=3) AND = (SELECT EMP_IDN FROM EMPPERS WHERE EMP_SKL='Driving');

Table-6: Final Query resultant data

Emp_IDN	Emp_FNM
1001	Satish K
----	-------
-----	-------
1004	Shakul M

The proposed methodology with all of its 4 level of implications produced the accurate results with optimal way of expected outputs. The Resultant data produces the dealing strategy for huge set of relational table queries can be implemented in a flexible way if it is analyzed with the proper reinforcement schema for case analysis method. The final results are optimized with the execution of case based analysis system.

IV.RESULTS AND DISCUSSION

The proposed methodology focused with 20 different relational table-sets, in which the customized query from the user's side execution yields different possible combination of accurate results in a flexible way. The online policy deals with current state reinforcement learning schema whereas the offline policy deals with the existing updated policy for current state reinforcement learning. Both methodologies are optimal with their own way of diversifications along with the expected query results. The following table illustrates the success rate for each of our reinforcement learning approaches for 25 sample relational table sets with 50 customized queries.

Table-7: Samples Experiment resultant data

Reinforcement Schema	Strategy wise Solution	Queries
Full Extraction	14	20
Incremental Extraction	8	10
Online Extraction	7	10
Offline Extraction	15	20
Total success (44/50=88%)	44	50

The following diagram illustrates the success rate for each of our reinforcement learning approaches for 25 sample relational table sets with 50 customized queries.

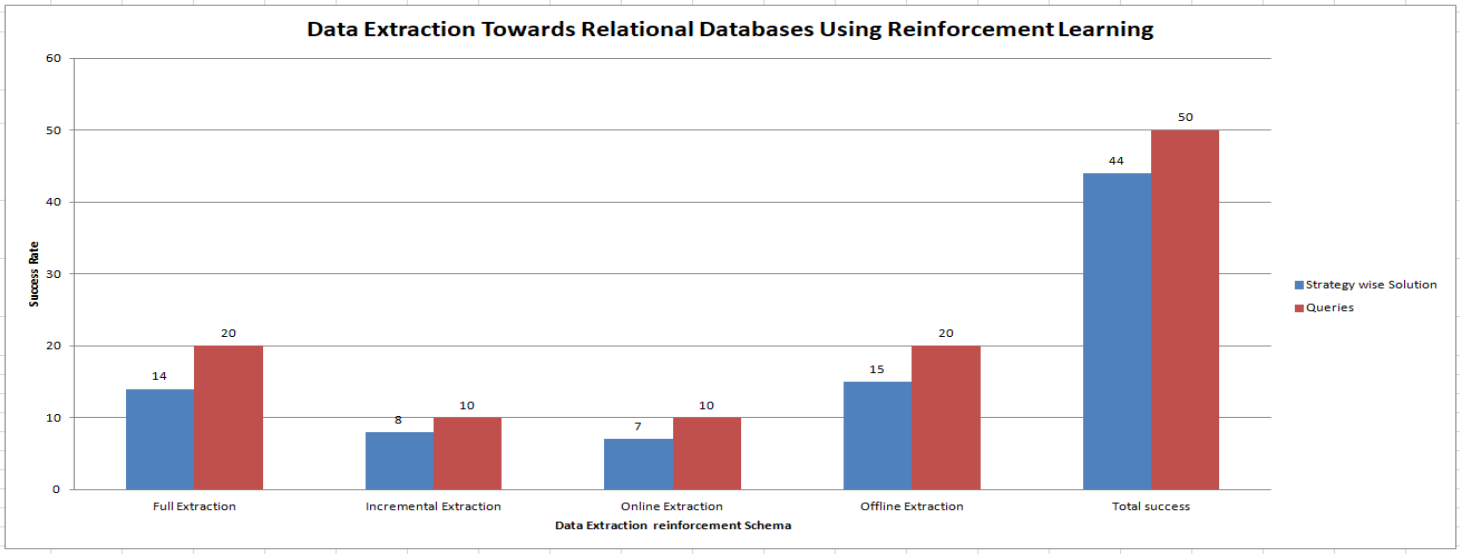

Fig-4: Proposed methodology Efficiency Rate

V.CONCLUSION

The data extraction from relational database tables is acting in a primary role due to its heavy impact in the online web databases. Our proposed methodology applies the reinforcement learning approaches with online and offline policy structures to classify the position to attack the database to retrieve the information's in a proper and accurate way .the consistency in results for all the individual approaches only gives the entire reliability for the proposed approach. This paper deals with the entire retrieval or part by part retrieval based on its state of online or offline with preloaded content, on loading content, uploaded content and downloaded content form a relation database table set. Our proposed approach yields 88% efficiency with the sample set of 25 relational table sets along with the 50 customized query states with 4 levels of implications. In near future we will implement the Fuzzy based neural networks approach for Data extraction schema in relational database tables.

References

[1] C. R. Aberger, S. Tu, K. Olukotun, and C. Ré. EmptyHeaded: A relational engine for graph processing. In SIGMOD, 2016.
[2] S. Abiteboul, R. Hull, and V. Vianu. Foundations of databases, volume 8. Addison-Wesley Reading, 1995.
[3] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. Reviews of modern physics, 74(1), 2002.
[4] R. Angles and C. Gutierrez. Survey of graph database models. ACM Computing Surveys (CSUR), 40(1):1, 2008.
[5] A. Apostolico and G. Drovandi. Graph compression by BFS. Algorithms, 2(3):1031–1044, 2009.
[6] Y. Asano, Y. Miyawaki, and T. Nishizeki. Efficient compression of web graphs. In International Computing and Combinatorics Conference, pages 1–11, 2008.
[7] P. Atzeni, P. Cappellari, R. Torlone, P. A. Bernstein, and G. Gianforme. Model-independent schema translation.VLDB Journal, 17(6):1347–1370, 2008.
[8] P. Boldi and S. Vigna. The web graph framework I: compression techniques. In WWW, pages 595–602, 2004.
[9] N. R. Brisaboa, S. Ladra, and G. Navarro. K2-trees for compact web graph representation. In Intl. Symposium on String Processing and Information Retrieval, 2009.
[10] Y. Bu, V. Borkar, J. Jia, M. J. Carey, and T. Condie. Pregelix: Big (ger) graph analytics on a dataflow engine. PVLDB, 2014.