International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Semantic Representation and Optimized Querying of Cancer Data using Modified Shuffled Frog Leaping Algorithm

Gomathi R,Vidhya N

Abstract*: Cancer registries are most important to predict and treat the cancer disease. Numerous solutions are available in research to analyze the data in cancer registries. However, there is a lack of well defined data model since there is a link to external web pages. In order to overcome this issue a system is proposed to represent the cancer data using a semantic data model. The data model uses a Resource Description Framework (RDF) format to represent the data from the local cancer databases. It also uses an optimized Querying of the semantically represented data using SPARQL query language. The optimization of the queries is done with the Modified shuffled frog leaping algorithm(MSFL). This helps in treatment of cancer patients in an easy way.

Index Terms: Cancer data, Semantic web, biomedical informatics, Resource Description Framework, SPARQL queries, Shuffled frog leaping algorithm

I. INTRODUCTION

Today, most of the data we demand is available to us in the form of websites. Each web page is linked to some other web page which contains related information. Humans find easy to read the information from web pages but the machine finds it difficult to read the sense of the content in web pages. This lead to the materialization of the semantic web. Numerous data models [3] exist to store the semantic web data. One of the primary data models is the Resource Description Framework (RDF). It is a communication methodology for in lieu of semantic web data. The RDF saves the data in the form of a set of triples which contains subject, predicate and object. This Semantic representation provides an easy way to query information from the data registry.

A hospital based cancer registry was implemented by extracting necessary processes and responsibilities. It uses Unified Modeling Language (UML)[1] for representation. The management techniques and registration procedures depends upon the medical facility. There are no adequate systems available for maintenance of electronic health records [2] in healthcare organizations. The problems with electronic record keeping and processing was addressed in literature.

Manuscript published on 30 September 2019
* Correspondence Author
Gomathi R*, Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, India.
Vidhya N, Senior Software developer, The Vanguard group, USA.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

II. MODIFIED SHUFFLED FROG LEAPING ALGORITHM TO OPTIMIZE SPARQL QUERY

A. Query Paths

The semantic web related data stored in RDF arrangement is queried using the most common RDF query language SPARQL protocol and RDF query language (SPARQL).The input SPARQL query can be converted and graphically visualized as a query tree. In the constructed query tree, the leaf nodes of a query tree stands for any one of the triples. The internal nodes are used to join the triples. Different types of representation of query trees are available in research which includes bushy trees, left -deep trees and right -deep trees[5]. The nodes in these query trees can be shuffled in many different ways which produces the same results. After constructing the query tree, it represents the query plan or query path. In this research right-deep trees are used to represent the query plans.

B. Solution Space

Whenever an optimization algorithm is applied to a problem, we need to define a solution space which consists of all possible solutions. Since the problem deals with query plans as solutions, the query tree becomes a solution in the solution space. The size of the plan depends on the kind of tree used for representing the query plan. In general, the number of probable ways to allocate n triples to the trees leaves is n!. These n! ways can be derived by applying the transformation rules.

C. Encoding

After representing the solutions using a format, a suitable encoding mechanism for the solution and a fitness function should be choosen. Since we represent the query plans using right deep trees, the encoding methodology choosen is the ordered list. Solutions in the solution space are denoted using an ordered list [5] of leaves. All the solutions in the solution space is converted into this encoding methodology.

D. Fitness function

To decide RDF query path, any algorithm must choose the fitness function. In this investigation, the fitness function is compared with the cost of the right deep tree. The cost of a right deep tree depends on the selectivity of the triples and cardinality estimation [6]. Consider Ri be the cardinality of the triples and fi,j be the selectivity of the triples.

If pi,j is defined as the join predicate between Ri and Rj,we can define as in Equation(1)

$$fi, j = \frac{| Ri \infty pi, jRj |}{| RiXRj |} \quad (1)$$

For a given join tree T, the ensuing cardinality value |T| can be recursively calculated as in Equation (2) and Equation (3)

|T|= | Ri | if T is a leaf node (2)

$$| T | = (\Pi Ri \in T1, Rj \in T2, fi, j) | T1 \| T2 |$$

if T=T1 ∞ T2 (3)

For a given join tree T, the cost function Cout is defined as in Equation (4) and Equation (5)

Cost = 0

if T is a leaf node (4)

Cost of the tree= Cost of root node cost of subtrees,

if T=T1 ∞ T2 (5)

III. MSFL ALGORITHM

The following are the step by step procedure in the MSFL algorithm. It is applied to solve the problem of query optimization.

Start
Generate a population of possible P query plans
For each individual in population P, calculate the fitness value of each query plan
Sort the query plans in ascending order of cost
Determine the cost of global best query plan
Divide P into m query plans
For each plan m
Determine the cost of local best and worst solution
Improve the position of worst solution with respect to local best solution
Find out new position and calculate cost
Check if cost improves, combine the evolved plans
Classify the population in ascending order of cost until termination condition becomes true
End

The traditional Shuffled frog leaping algorithm uses shuffling the memeplexes in the decreasing order of performance. As a variation, our proposed algorithm follows the shuffling process with increasing order of cost.

IV. EXPERIMENTAL RESULTS

The cancer registry was created from local databases which includes 5,00,000 cases diagnosed within last three years. The patients were grouped according to gender and age.These databases are first represented using RDF format and queried using SPARQL queries which goes through an optimization phase. Implementation is executed in a Microsoft windows environment on an Intel Pentium 4 machine with 4GB RAM. Tests are implemented by considering 500,000 triples from local cancer database which is represented using RDF. To measure the concert of the proposed technique, several experiments are conducted with queries of unstable number of predicates. The optimization algorithm is iterated for hundred times to increase the correctness of the results. The SPARQL query produces results about the patients which includes the type of cancer, type of therapy: Surgery, chemotherapy, radiotherapy, hormonal therapy and so on. Based on the queried results weights are assigned to the type of cancer and the type of therapy based on the age limit of the patients.

Since the body condition of the patients affects the type of therapy they can undergo, such kind of weights is assigned. The weight value is the probability with which the patient can undergo a particular treatment A sample set of results obtained by optimized querying is given in Table I. This kind of diagnosis will help the health professionals to treat cancer disease in an efficient manner.

Table I Sample results of querying cancer datasets

Patient	Age	Type of cancer	Type of therapy	Weight
A(Male)	63	Lung	Surgical	0.33
			Chemotherapy	0.42
			Radiotherapy	0.12
B(emaleF)	48	Breast	Surgical	0.34
			Chemotherapy	0.62
			Radiotherapy	0.15

V. CONCLUSION

In this work the MSFL algorithm is applied to optimize biomedical queries of cancer data. The algorithm proceeds with a solution space having all possible query plans. The cost of the all query plans is based on the arrangement of joins occurring in the query. In this research work the application of the algorithm is tested using cancer datasets and queries are presented. The extent to which the algorithm work efficiently is measured in terms of probability of treatment of the diseases.

REFERENCES

1. Shiki, N., Ohno, Y., Fujii, A., Murata, T., & Matsumura, Y. (2008). Unified Modeling Language (UML) for hospital-based cancer registration processes. Asian Pac J Cancer Prev, 9(4), 789-96.
2. Shortliffe, E. H. (1999). The evolution of electronic medical records. ACADEMIC MEDICINE-PHILADELPHIA-, 74, 414-419.
3. Liu C, Wang H, Yu Y, Xu L. Towards efficient SPARQL query processing on RDF data. Tsinghua Science & Technology. 2010 Dec 1;15(6):613-22.
4. Eusuff, M., Lansey, K., & Pasha, F. (2006). Shuffled frog-leaping algorithm: a memetic meta-heuristic for discrete optimization. Engineering optimization, 38(2), 129-154.
5. Steinbrunn M, Moerkotte G, Kemper A. Heuristic and randomized optimization for the join ordering problem. The VLDB Journal—The International Journal on Very Large Data Bases. 1997 Aug 8;6(3):191-208.
6. Gomathi R, Sharmila D. A novel adaptive cuckoo search for optimal query plan generation. The Scientific World Journal. 2014;2014.

AUTHORS PROFILE

Dr.R.Gomathi completed her Doctorate in Information and Communication Engineering during April 2016.She is presently working as an Associate Professor in the Department of Computer Science and Engineering,Bannari Amman Institute of Technology. She is having nearly 16 years of teaching experience and have published nearly 15 papers in reputed Journals and nearly 12 papers in Conferences. She has received the "Young Faculty Achiever Award 2018" by Engineering Professional Society.

Ms.Vidhya.N completed her graduate in Engineering in 2003 and presently working as a Senior Software Developer in the Vanguard Group, USA.She has nearly 15 years of working experience in IT companies.