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

Proximity Matrix Completion and Ranking Ant Colony Optimization technique in Semantic web

Rubin Thottupurathu Jose, Sojan Lal Poulose

***Abstract*: The semantic web consists of a large number of data that is difficult to retrieve the answer for the user queries. An existing method in the query processing in the semantic web has three main limitations namely, query flexibility, query relevancy or lack of ranking method and high query cost. In this study, Proximity Matrix Completion technique (PMC) is applied to impute the missing data in the dataset that helps to increase the query flexibility and Ranking Ant Colony Optimization (RACO) technique is used to select the relevant features from the dataset and arrange them to increase relevancy. The result shows that the PMC-RACO method has a higher performance compared to the exiting method in semantic web. The mean precision value of the PMC-RACO method in sports data is 87%, while the existing method has the precision value of 83%.**

***Index Terms*: Proximity Matrix Completion technique, query flexibility, query relevancy, Ranking Ant Colony Optimization and Semantic web.**

I. INTRODUCTION

The World Wide Web is the ultimate source of information in the present-day world. The web is highly dynamic and keeps on changing in the information. An increasing number of users in the web makes the search engine are predominant and popular [1]. Semantic web is evolved from the normal web that helps the machine to understand the knowledge of web and can be used for the organization, navigation, integration and task automation [2]. Current techniques are used to represent the semantic data is in the form of RDF Linked. The data in the RDF are growing rapidly and requires an effective technique for search method. There ismore number of data present in the web related to many categories like geographical, life science and government domains [3]. To apply the learning technique in the search process, the factual information is required. There is still the room for improvement in other learning types like analysis, understanding and applications of the learned information [4].

Many existing methods were conducted to increases the performance of the query process with low cost [5, 6]. The three major challenges still need to solved in the semantic web process to increase performance [7, 8]. The flexibility is the first challenge and query cost is the second challenges in the existing method due to large amount of data are present in the RDF graph, which affects the flexibility as well as the query cost [9]. Third, the lack of ranking function in the system that affects performance [10]. In this research, Proximity Matrix Completion (PMC) method is applied to impute the missing data in the dataset. After, the Ranking Ant Colony Optimization (RACO) technique is applied to select the features based on the triple values and the answer to the queries are ranked by the page rank technique. The experimental results show the efficiency of query processing in real-time data.

This paper is organized as follows, Literature Survey given in section II, Proposed method explained in section III, Experimental result is illustrated in section IV. The conclusion of this paper is made in section V.

II. LITERATURE SURVEY

The semantic data are present in the wide range and this is difficult to handle these data and provide the required information to the user. Many existing methods are applied for the semantic web to attain an effective result for the queries.

Acosta, et al. [11] proposed crowdsourcing technique to increase the performance of the completeness of answer in the SPARQL queries. The RDF data are analyzed to measure the completeness of the data. The missing values in the RDF data are imputed based on the crowd answers, which is collected crowd knowledge base. A query engine process the on-the-fly crowd knowledge and measure the RDF model completeness. The method has been analyzed and crowdsourcing technique has a higher performance in data imputation. The query cost is high and optimization technique can be applied to decrease the cost.

Li, et al. [12] developed the fuzzy graph based on the RDF data and extension has been provided to analyze the vagueness in the graph. The pattern matching method is used in the subgraph to retrieve the query. The graph homomorphism is used to denote the graph to reduce the cost of the query. The query cost is much reduced and the method shows the effectiveness in the query processing. The completeness of the answer is not sufficient for the query and normalization technique can be applied to increase the performance.

Jiménez, et al. [13] developed a fuzzy technique for the SPARQL queries to express the fuzzy queries in the RDF data.

Manuscript published on November 30, 2019.
* Correspondence Author
Rubin Thottupurathu Jose*, School of Computer Sciences, M G University, Kottayam, Kerala, India.
Dr Sojan Lal Poulose, Principal, Mar-Baselious Institute of Technology and Science, Kothamangalam, Kerala, India.

The method was made automatic and user-defined enabling wide range of mechanism for categorization and ranking with sentiment analysis and topic detection. The developed method is tested on real time dataset and this shows the higher efficiency than the existing method. The optimization technique can be applied to minimize the query cost and answer completeness is needs to be increased.

Arnaout and Elbassuoni, [14] developed the general framework to increase the efficiency of the search in the RDF knowledge graph. The search was conducted based on the knowledge graph and triple score value, which allows the user to analyze the large data. The results are ranked in this method based on the statistical machine translation and query relaxation were performed automatically to increase the recall measure. The ranking method increases the query efficiency and the recall value is higher for the method. Analyzing a large number of queries increases the query cost and optimization technique can be applied to reduce the cost of queries.

Halvorsen and Stolpe, [15] defined row-reduction and column-reduction technique operation, if the SPARQL resultset is viewed as the table. The single operator is used to combine row and column operator that can be applied in the semantic query. The query cost is low and this method can process number of data. The effectiveness of the query is low and the method needs to analyze the large data of the RDF. The flexible representation technique can be applied to increase efficiency without affecting the performance.

A. Problem definition and solution

From the analysis of the recent methods in the semantic query processing, some of the limitations are observed. The proposed PMC-RACO method is applied for the multi-objective optimization and the proposed method consider the three problem definition. The current challenges in the semantic web query processing are explained as follows

1. **Query Flexibility:** There is a large number of data present in the semantic web and continuous to increases. Existing representation method involves the missing data and affects the flexibility of the method to search the information in the system. Solution: The proposed PMC-RACO method involves in using the predictive model for impute the missing data and supports to analyze the relationship between the data by the search method.
2. **Query cost:** A large number of data present in the system and analyzing the data depends on the feature value. The irrelevant feature selection consumes more time to process the data and this is denoted as query cost. Existing search techniques are having a high query cost. The effective representation of data and select the suitable features in the system helps to decreases the query cost. Solution: The RACO method helps to effectively analysis the feature in the RDF data. This method helps to reduce the query cost of the system.
3. **Result Ranking:** Most of the existing method was not using the ranking technique to order the result based on its importance and this affects the relevance retrieval of the system (Recall value). The recall evaluation parameter can be used to measure the relevance retrieval of the method. The recall value of the existing method is low and can be increased by the ranking method. Solution: The proposed PMC-RACO method of optimization technique applies the page rank technique to rank the query based on its importance. Hence, the proposed method has a higher recall value.

III. PROPOSED METHOD

This research aims to solve the three major problems in the semantic web query processing namely, flexibility, query cost, and lack of ranking. The PMC is applied to impute the missing data in the semantic dataset and this method helps to solve the flexibility problem. The RACO method applies in the features of the semantic data to reduce the cost value and page rank technique is applied to improve the efficiency of the method. This section gives a detailed explanation about the proposed method. The architecture of the proposed PMC-RACO method is shown in Fig. 1. The RDF graph is plotted from the semantic web with vertices and edges and the triple values are extracted as same as in the research [16].

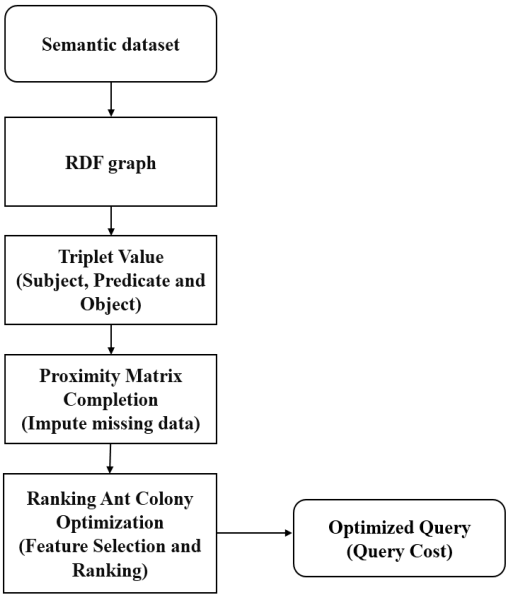

Fig. 1. The architecture of PMC-RACO method in semantic web data

A. Proximity Matrix Completion

Assume n data points of the semantic web data $\{x_i : i = 1, 2, ...n\}$ and distance metrics is measured between the two data are $d_{ij} = dis\,\text{tance}(x_i, x_j)$, results are shown in the proximity matrix [17], as shown in Eq. (1).

$$D = \begin{pmatrix} d_{11} & \dots & d_{1n} \\ \vdots & \ddots & \cdots \\ d_{n1} & \dots & d_{nn} \end{pmatrix} \quad (1)$$

Once the distance metrics are measured for the whole data and then reduces the missing data in the dataset. Hence, the clustering techniques of RACO can be directly applied to the whole data including imputed data. The cluster data points are generated based on the existing cluster methods.

The clustering technique includes single-line and complete-link algorithms.

A graph-theoretic formulation of data points is used as the imputation method and mainly used to distinguish between the two categories of missing not-at-random values in the proximity matrix. The data points are denoted as $V=\{1,2,....n\}$, which will interpret as vertices in a graph $G=(V,E)$, where the edge $e=(i,j)\in E$ denotes the similarity between the corresponding data points (edges are in the missing value in). A clique $C\subseteq V$ in G is the subset of vertices such that each pair of distinct vertices is connected by an edge. A clique denotes the data points with the complete distance information. A maximum clique is a clique with the property that if one more vertex is added and the subset of vertex is no longer a clique. The detailed description the method of PMC and vertices measure is given in the research [17].

B. Raking Ant Colony Optimization technique

The ACO is a heuristic method that is inspired by the behavior of ant nature [18]. The ant can find the shortest path to the food source based on depositing pheromone information without any visual cues. The ACO solves the TRP problem effectively and the ACO is applied to solve many hard problems [19]. This has been applied to the various combinatorial problem due to its efficiency [18 – 19]. The description of the ACO in the TRP is discussed as an example due to this it has effective performance in that manner. The objectives in TRP is to find the shortest path that transverses all the cities once. Consider n cities and ants as m, the cities are fully connected with the edges E^n. The optimization steps are explained as follows:

1. **Initialization: Place** the m ants in the n cities randomly, and pheromone value is assigned as small positive variable.
2. **Path construction:** Each ant chooses the next city to visit based on the transition probability as defined as in Eq. (2).

$$p(i,j)=\begin{cases}\dfrac{[\tau(i,j)]\alpha[\eta(i,j)\beta]}{\sum_{u\in j}[\tau(i,u)]\alpha[\eta(i,u)]\beta} & if\ j\in J\\ 0 & otherwise\end{cases} \quad (2)$$

Where ant k is in the city i and the next city is j,τ is the amount of pheromone on the edge (i,j), the heuristic edge value is η_{ij}, the two weighting factors that controls the importance between the pheromone and heuristic information is α and β, and set of cites J is not visited.

3. If all the ants have the complete traverses path to all the cities, then go to step 4. Otherwise, go to step 2.
4. Pheromone is updated using the Eq.

$$\tau(i,j)\leftarrow(1-\rho).\tau(i,j)+\rho.\Delta\tau(i,j) \quad (3)$$

Where $\rho\in(0,1)$ is the evaporation rate and the value $\Delta\tau(i,j)$ is related to the fitness function value.

5. If the stopping criteria is achieved, then stop the process, otherwise randomly place the ants and go to step 2, and the best solution is obtained [19].

Algorithm 1: Ant colony optimization.

```
Initialize pheromone values.;
while termination criteria not met do
    ConstructSolutions()
    ApplyLocalSearch()      % optional
    UpdateTrails()
end
```

1) PageRank Method

The solution provided by the ACO algorithms are ranked based on the vertices and edges from the RDF graph. The activities are observed to solve the rank sink problem and phenomenon is found that the all users is not follows the existing links but directly go to page b, not directly link to page a. For this purpose, the users just type the URL of page b into the URL text field and directly jumps to page b, as in Eq. (4). In this case, the page b is not affected even when the page a is directly connected. Therefore, no absolute rank sink is provided [20].

$$PR(u)=(1-d)+d\sum_{v\in B(u)}\frac{PR(v)}{N_v} \quad (4)$$

Where damping factor is provided as $d=0.85$. We could think of user's probability d following links and regarded as $(1-d)$ for the PageRank distribution from non-directly linked pages.

IV. EXPERIMENTAL RESULT

The proposed PMC-RACO is evaluated with two different datasets such as DBpedia and IMDB dataset to measure the flexibility, efficiency and query cost of the method. The precision metric is used to measure the flexibility of the method and efficiency can be measured by the recall values.

A. Dataset

DBpedia: DBpedia dataset (2014) consists of the semantic web queries that is used to measure the performance of the method. The 50 benchmark queries are used to analyze the triple pattern in the dataset and investigate the performance of the methods.

IMDB: IMDB dataset consists of the data related to the movies like actor, director etc.., and consist of the 17.2k vertices. The instance in the dataset is high and this is very convenient for measuring the query cost of the method.

These two datasets are used to measure the performance of the method and analyze the impact of the semantic web challenges.

B. Experiment Setup

The experiment was conducted on the system that has the specification of the Intel i7 processor, 8GB of RAM, and 500GB hard disk. The method is implemented using the Java eclipse 8.2, Apache jena 1.8 and jdk.

Experimental Settings: The initial population of the ACO method is set as 50, maximum generation = 50, decaying rate is set as 0.1 and evaporation rate is set as 0.1. The maximum generation is set as 50 and generation about 50 provides the low increase in performance.

The decaying rate and evaporation rate above 0.1 doesn't consider the important features.

C. Query Flexibility Analysis

The PMC is applied to impute the missing data in the system and this provides the relationship between the data. This technique helps in increases the flexibility of the method by impute and analyze the relationship between the data. The ranking technique helps to provide effective answering in the data. The DBpedia dataset are used to measure the performance of the proposed PMC-RACO method. The precision metric is used to measure the effectiveness of the method and precision metric in the different dataset provides the flexibility of the method. The precision value for the different data is shown in Table 1.

Table I. The precision value of query processing

Query	Sports			Music			Life sciences			Movies			History		
	HARE-BL [11]	HARE [11]	PMC-RACO	HARE-BL [11]	HARE [11]	PMC-RACO	HARE-BL [11]	HARE [11]	PMC-RACO	HARE-BL [11]	HARE [11]	PMC-RACO	HARE-BL [11]	HARE [11]	PMC-RACO
Q1	1	1	1	1	1	1	1	0.5	0.63	0.34	1	1	N/A	1	1
Q2	1	1	1	1	1	1	1	1	1	0.64	0.96	0.97	1	1	1
Q3	0.33	1	1	1	1	1	1	1	1	0.53	1	1	0.75	0.75	0.78
Q4	0.13	0.55	0.62	0.5	0.5	0.62	0.5	1	1	1	1	1	0.63	0.77	0.79
Q5	0.8	1	1	N/A	0.57	0.65	0.18	1	1	0.5	0.8	0.85	0.77	0.95	0.97
Q6	0.6	0.69	0.73	0.5	0.6	0.62	1	1	1	1	1	1	0.78	0.93	0.96
Q7	0.67	1	1	N/A	0.48	0.53	0.54	0.75	0.77	0.89	1	1	0.71	0.63	0.67
Q8	0.5	0.92	0.94	0.43	0.39	0.43	0.71	0.87	0.88	0.87	1	1	0.33	0.93	0.95
Q9	0.3	0.5	0.51	0.92	0.36	0.41	0.54	1	1	0.58	1	1	0.72	0.54	0.62
Q10	0.4	0.91	0.93	0.39	0.52	0.57	0.7	1	1	1	1	1	0.48	0.95	0.97

The precision value is measured for the DBpedia dataset and also for the different data as shown in the Table. (1). This shows that the precision value of the PMC-RACO method is high compared to the other existing methods in the query processing technique. The high efficiency is due to the ranking technique in the feature selection technique that ranks the solution of the ACO and provides the answer to the queries. The movie type dataset has a higher value compared to the other dataset and the proposed PMC-RACO method has a higher precision value. The overall performance of the proposed method is high in the different types of data and the proposed PMC-RACO method has higher performance. This shows that the proposed PMC-RACO method is more flexible compared to the other techniques.

D. Query Ranking Analysis

The ranking method provides a more relevant answer to the user queries based on optimization technique. Recall is the measure of relevant instance retrieved from the dataset based on the user queries and this provides the relevance measure to the answers. The recall value can be used to measure of the relevance value of the proposed PMC-RACO method. The recall value is measured in the DBpedia dataset and this is shown in the Table. (2).

Table II. Recall value of query processing

Query	Sports			Music			Life Sciences			Movies			History		
	HARE-BL [11]	HARE [11]	PMC-RACO	HARE-BL [11]	HARE [11]	PMC-RACO	HARE-BL [11]	HARE [11]	PMC-RACO	HARE-BL [11]	HARE [11]	PMC-RACO	HARE-BL [11]	HARE [11]	PMC-RACO
Q1	1	1	1	1	1	1	1	1	1	0.55	0.41	0.57	0	1	1
Q2	1	1	1	1	1	1	1	1	1	0.7	1	1	1	1	1
Q3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Q4	0.14	0.86	0.92	1	1	1	0.2	1	1	1	1	1	0.28	0.94	0.95
Q5	0.8	1	1	0	0.8	0.85	0.33	1	1	1	1	1	0.94	1	1
Q6	0.67	1	1	0.25	0.75	0.77	1	1	1	0.16	1	1	0.27	0.96	0.97
Q7	1	1	1	0	0.92	0.95	0.78	1	1	0.89	1	1	0.24	0.95	0.97
Q8	0.55	1	1	0.43	1	1	0.38	1	1	0.87	1	1	0.07	1	1
Q9	0.5	1	1	0.35	1	1	0.58	1	1	0.7	1	1	0.84	1	1
Q10	0.6	1	1	0.2	0.91	0.95	0.54	1	1	0.88	1	1	0.98	1	1

The recall value of the proposed PMC-RACO method is high compared to the other existing methods and this is due to the ranking technique and answer completeness. The answer completeness is achieved by the data imputation technique based on PMC. The PageRank method gives a more relevant answer in order and this increases the recall value. The overall recall value for the proposed PMC-RACO method is high. The mean recall value in the movie data is 95%, while the existing method HARE-BL has the recall value of 70%. Hence, the proposed PMC-RACO method solves the problem of the low recall value. The average execution time of the existing and proposed method is 0.01 s in the DBpedia dataset.

E. Query Cost Analysis

The proposed PMC-RACO methods are analyzed in the IMDB dataset for the query cost. The execution time of the proposed PMC-RACO method is measured and compared with the executing method [12] for the query cost. The IMDB dataset has more data to process and this increases the execution time and this dataset is selected for the measure of query cost. The vertices and edges are varied to analyze the performance of the proposed PMC-RACO method and compared with existing methods such as NAGA [12], PQ [12], and Backtracking [12].

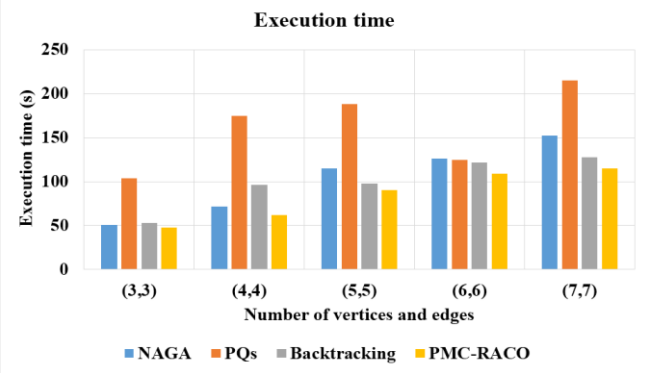

Fig. 2. The Execution time of the proposed PMC-RACO method in IMDB dataset

The execution time of the different method is measured for varying the vertices and edges in the graph. The execution time of the different method is compared with the existing method in Fig. (2). The execution time of the proposed PMC-RACO method is low compared with other existing methods. Increases in the vertices and edges also increase the execution time of the method. The execution time of the proposed PMC-RACO method is low due to the optimal selection of the features by the ACO and rank them to PageRank method. The average precision value of the bracktracking [12] method is 60.17%, while proposed method has the average precision value in query processing is 65.25%.

The investigation of the proposed PMC-RACO method for the different metric and compared with existing method shows that the proposed PMC-RACO method solves the problem of the Query flexibility, Low efficiency and query cost.

V. CONCLUSION

Semantic web has a wide range of data and processing the data to find the relevant information is an important task. There are some limitations in the semantic web query processing namely, query flexibility, query efficiency and query cost. This research aims to improve the performance of query processing based on the imputation and optimization method. The PMC-RACO method imputes the missing data in the database and RACO method finds the relevant features from the dataset. The PMC-RACO method is analyzed for efficiency in the two datasets for the query flexibility and query cost. The experimental result shows that the performance of the proposed PMC-RACO method is high compared to other techniques. The experimental investigation shows that the PMC-RACO method solves the problem of query flexibility, query efficiency and query cost. The mean precision value of the PMC-RACO method for sports data is 87%, while the existing method has the precision value of 83%. In future work, the proposed PMC-RACO method can be applied for the top k queries in the semantic web to increase the efficiency.

REFERENCES

1. G. Deepak, and J. S. Priyadarshini, (2018). Personalized and Enhanced Hybridized Semantic Algorithm for web image retrieval incorporating ontology classification, strategic query expansion, and content-based analysis. *Computers & Electrical Engineering, 72*, pp. 14-25.
2. B. Fazzinga, G. Gianforme, G. Gottlob, and T. Lukasiewicz. (2011). Semantic Web search based on ontological conjunctive queries. *Journal of Web Semantics, 9(4)*, pp.453-473.
3. A. Poulovassilis, P. Selmer, and P. T. Wood. (2016). Approximation and relaxation of semantic web path queries. *Journal of Web Semantics, 40*, pp.1-21.
4. T. Yilmaz, R. Ozcan, I. S. Altingovde, and Ö. Ulusoy. (2019). Improving educational web search for question-like queries through subject classification. *Information Processing & Management, 56(1)*, pp.228-246.
5. C. Huang, H. Xu, L. Xie, J. Zhu, C. Xu, and Y. Tang. (2018). Large-scale semantic web image retrieval using bimodal deep learning techniques. *Information Sciences, 430*, pp.331-348.
6. G. Zenz, X. Zhou, E. Minack, W. Siberski, and W. Nejdl. (2009). From keywords to semantic queries—Incremental query construction on the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web, 7(3)*, pp.166-176.
7. O. Hartig, and J. Pérez. (2016). Ldql: A query language for the web of linked data. *Journal of Web Semantics, 41*, pp.9-29.
8. F. Cai, and M. de Rijke. (2016). Learning from homologous queries and semantically related terms for query auto completion. *Information Processing & Management, 52(4)*, pp.628-643.
9. J. Singh, and A. Sharan. (2018). Rank fusion and semantic genetic notion based automatic query expansion model. *Swarm and Evolutionary Computation, 38*, pp.295-308.
10. W. Van Woensel, and S. Casteleyn. (2016). A mobile query service for integrated access to large numbers of online semantic web data sources. *Web Semantics: Science, Services and Agents on the World Wide Web, 36*, pp.58-76.
11. M. Acosta, E. Simperl, F. Flöck, and M. E. Vidal. (2017). Enhancing answer completeness of SPARQL queries via crowdsourcing. *Journal of Web Semantics, 45*, pp.41-62.
12. G. Li, L. Yan, and Z. Ma. (2019). Pattern match query over fuzzy RDF graph. *Knowledge-Based Systems, 165*, pp.460-473.
13. J. M. Almendros-Jimenez, A. Becerra-Terón, and G. Moreno. (2018). Fuzzy queries of social networks with FSA-SPARQL. *Expert Systems with Applications, 113*, pp.128-146.
14. H. Arnaout, and S. Elbassuoni. (2018). Effective searching of RDF knowledge graphs. *Journal of Web Semantics, 48*, pp.66-84.
15. J. Halvorsen, and A. Stolpe. (2018). On the size of intermediate results in the federated processing of SPARQL BGPs. *Journal of Web Semantics, 51*, pp.20-38.
16. E. G. Kalayci, T. E. Kalayci, and D. Birant. (2015). An ant colony optimisation approach for optimising SPARQL queries by reordering triple patterns. *Information Systems, 50*, pp. 51-68.
17. S. Karimzadeh, and S. Olafsson. (2019). Data clustering using proximity matrices with missing values. *Expert Systems with Applications, 126*, pp.265-276.
18. Y. Wu, M. Gong, W. Ma, and S. Wang. (2019). High-order graph matching based on ant colony optimization. *Neurocomputing, 328*, pp.97-104.
19. E. Papenhausen, and K. Mueller. (2018). Coding Ants: Optimization of GPU code using ant colony optimization. *Computer Languages, Systems & Structures, 54*, pp. 119-138.
20. W. Xing, and A. Ghorbani. (2004). Weighted pagerank algorithm. In Proceedings. *IEEE Second Annual Conference on Communication Networks and Services Research*, pp. 305-314.

AUTHORS PROFILE

Rubin T Jose, MCA, M Tech. Scholar in the area of Semantic web and Ontology Engineering, already published 3 Journal Papers and 8 Conference publications in the area. Attended short course in the Protégé tool in Stanford University, USA. He has got a total of 15 years of teaching experience in the field of Computer Science and Engineering.

Dr P. Sojan Lal has more than 30 years of blended experiences, with major international petroleum companies in Middle East and premier educational institutions in India. He has authored 6 technical books, two of them published in Germany and other books published in India. He is an approved research supervisor of Mahatma Gandhi University, Kottayam; University of Petroleum and Energy Studies, Dehradun; and APJ Abdul Kalam Technological University, Kerala, India.
Dr. Sojan Lal has authored 65 National and International Journal and Conference papers and guided 4 PhD scholars. He has the world record for the highest number of publications within shortest period in 2014. He has been listed in “Marquis Who's Who in the World” since 2009, representing the world's most accomplished individuals.