Research Journal of Applied Sciences 11 (10): 1151-1158, 2016
ISSN: 1815-932X

Distributed Genetic Algorithm for Information Retrieval in Linked Open Data

Asaad Sabah Hadi
Information Technology College, University of Babylon, Hillah, Iraq

Abstract: Information retrieval play an important role in the web. Many research provide many suggestion for retrieve the information to the user in an appropriate way. Linked Open Data (LOD) simplify the information retrieval by focusing on the data rather than the web link and extend the normal web by publishing different data and make an appropriate links between them. The LOD continue growing by adding new web links and that growing make searching the information from it more difficult and need more time. This study proposed an algorithm for using distributed genetic algorithms to improve the information retrieval process in linked open data. Wikipedia is a free online encyclopedia which is created and edited by contributors from many countries around the world. DBpedia which is a web version of DBpedia infer structure information from the wikipedia to make it accessible on the web. In order to facilitate the information retrieval form the DBpedia in the LOD, the suggested algorithm make a clustering for all websites according to their types into seven different clusters and develop a genetic algorithm for each cluster. Steady state Genetic Algorithm with Triple tournament selection is used in each cluster. The replacement strategy was used to improve the total performance of the genetic algorithm by replace the bad individual with a good offspring. Our suggested algorithm give a rise for the importance of using the evolutionary algorithm in linked open data.

Key words: Genetic algorithm, LOD, RDF information retrieval, wikipedia, DBpedia

INTRODUCTION

Genetic algorithm(s) are one of the search algorithms that base on the natural selection and natural genetics. They use thesSurvival of the fittest idea in order to make a stronger population to the solution of the problem at hand (Goldberg, 1989). The most important phases of the genetic algorithm are selection, crossover (reproduction), mutation, fFitness evaluation (Goldberg, 1989; Nowostawski and Poli, 1999). The main idea about reproduction is genetic material combinations between two or more parents in order to obtain two or more offspring. Whereas the mutation is applied to one individual to produce new individual that have a new genetic materials (Goldberg, 1989). The evaluation is the process of giving every individual a fitness value that can be used to select that individual in the next generations (Goldberg, 1989).

There are two main types of genetic algorithms sGA (simple Genetic Algorithm) and ssGA (steady state Genetic Algorithm). There are many difference between sGA and ssGA: first, the main addition in ssGA is replacement strategy in which the worst individual in the population is replaced by the best offspring so that the ssGA population is continue growing and new best individuals will be added in the future. Second the selection in ssGA is mainly tournament selection. Third the mutation probability in ssGA is higher than sGA (Agapie and Wrigh, 2014).

LOD (Linked Opened Data) is a promising technology for storing and providing the structured data. It uses semantic web principles and technology to publish and interlink data, therefore all its entities are referenced by URIs (Universal Resource Identifiers) using the protocol HTTP (Hypertext Transfer Protocol) (Florian and Kaltenbock, 2012).

Evolutionary Algorithm (EA): The main idea about the Evolutionary algorithm is the population of solutions and select among these solutions according to the survival of the fittest mechanism. In the beginning a random population will be created and by using a selection, crossover and mutation operators, a new population will be created (Goldberg, 1989). The fitness function is a heuristic estimation of the solution quality whereas the search process is driven by the selection and variation operators. There are a number of features that poses by the evolutionary algorithm like: EAs are a population based algorithm also it is used recombination in order to mix information for more than one candidate solution to obtain a new solution.

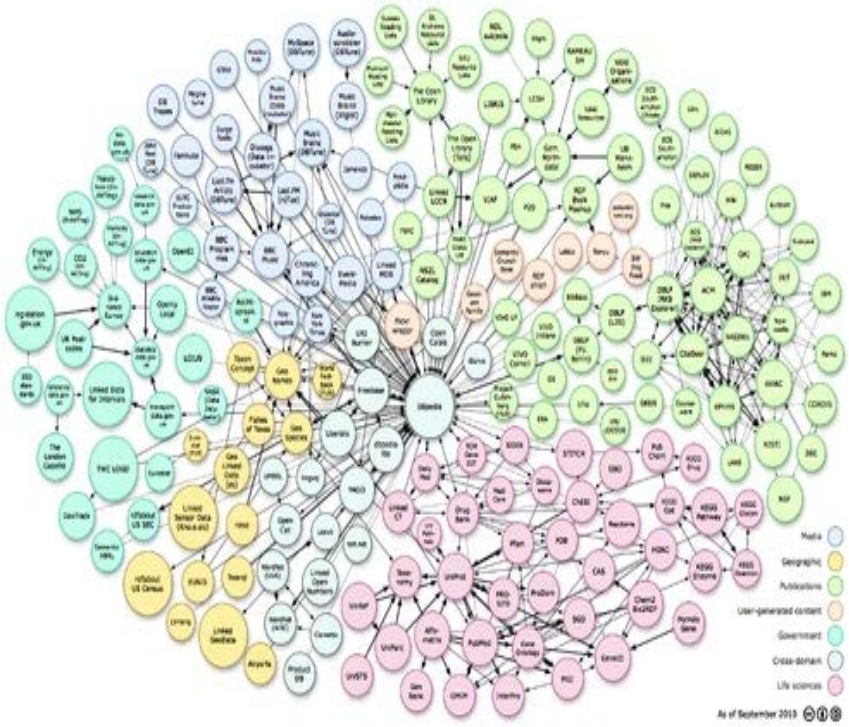

Fig. 1: The LOD at 2012 and it is continue growing (Hogan *et al*., 2012)

Genetic algorithm is one of the evolutionary algorithms. There are two main types of it: sGA (simple Genetic Algorithm), ssGAb (steady state Genetic Algorithm). The main difference between them is: -sGA replace the whole individuals in the current population with a new offspring whereas the ssGA use replacement strategies is which the best offspring is replaced by the worst individual in the current generation in order to produce the next generation (Agapie and Wright, 2014).

There are many research for using multiple genetic algorithm in order to increase the total performance where the main idea for that came from the natural behavior for dividing the various population into groups based on their fitness or distribute the genetic algorithm and use multiple populations (Nowostawski and Poli, 1999).

Linked Opened Data (LOD): Tim Berners-Lee *et al.* (2001) finds the principle of the LOD that overcome some of the limitation in the classical web. The main idea about his work is focusing on the data (which is the important thing in the web). The Linked data is based on many standards for representing, identification and retrieval of data (Agapie and Wright, 2014).

The LOD extend the web by publishing various data sets and put the links between them. The resulting data is named the LOD cloud which gives the key to realize the Semantic web.

In order to fully benefit from opened data, it is necessary to put the information and data into a context that creates new knowledge and enables a powerful applications and services. The LOD is a very important mechanism for information management and integration (Agapie and Wrigh, 2014; Bizer, 2009).

The data may be structured or unstructured. There are several problems when dealing with unstructured data resource like text, query, etc. Some research try to enrich the unstructured data by mapping its contents into a structured knowledge like LOD cloud (Hogan *et al*., 2012). The DBpedia is the major linking hub in the LOD which help the users by linking their unstructured data with it (Hogan *et al*., 2012). There are several advantages of DBpedia over the existing knowledge base (Hogan *et al*., 2012; Bizer *et al*., 2009).

- It represent real community agreement
- It cover many domain
- It truly multilingual, also it is accessible on the web
- It automatically evolves as wikipedia changes

Figure 1 show the LOD at 2012 and it is continue growing (Hogan *et al*., 2012). From Fig. 1, we can see that there are a number of category (Media, b Geagraphic, Puplications, user-generated content, government, cross-domain, life sciences). Each of these category have a number of website that link to it and that number

continue growing with time, also the information inside each website may expand. We will use that categories in our suggested framework architecture later. The principles for publishing data in the LOD is given by Hogan *et al.* (2012).

- Use URIs (Universal Resource Identifiers) as name of things
- Use HTTP URIs in order to simplify the looking up for that names
- Include links to other URIs, so that people can discover new things

There are many steps that be done by the LOD Community (Agapie and Wright, 2014; Hogan *et al.*, 2012):

- W3C has published an open standards for the semantic web build on RDF (Resource Description Framework) which is a standard for representing the meta data and it also used to publish the Wikipedia into DBpedia which is the semantic version of Wikipedia (Agapie and Wright, 2014)
- W3C semantic web standards expect the possibilities to link data sets. If we represent the information in a machine-readable format then these information is mostly the same (or closely) in another resource
- The web applications can use the linked data by standard web services where the semantic web can be used in the most common infrastructure "WWW"

According to the specialist natures of many experts and enterprises, there are many huge information still hidden in the internet and can't be findable or linked to other data. The linked data is a proper solution for that problem.

There are a number of steps that can be done to publish the data in LOD (Alden and Wrightm, 2014: Hogan *et al.*, 2012):

- Analyze your data
- Clean your data
- Model your data
- Choose appropriate vocabularies
- Specify license
- Convert data to RDF
- Link your RDF data into another data
- Publish and promote your LOD

METERIALS AND METHODS

Resourse Description Framework (RDF): XML is designed for producing an interchanged format for weakly structured data by defining the data model in a schema format. XML is not convenient for defining the meta data (Wilkinson *et al.*, 2003). RDF was proposed to fill this gap where it is a model for representing meta data as a triples "subject, predicate, object".

The general keyword scanning can't give a specific information because the machine can't easily understand the semantic of the sentences. RDF provides a good mechanism for recording the resource statements in order to simplify the machine understanding and interpretation of the statements (Powers, 2003).

RDF is based on the Domain-neutral model where it allow one set of statements to connect with another statements even they are dramatically differs (Powers, 2003).

Rather than generating one XML file in XML vocabulary for many different applications, one RDF file can be generated to contain all these information. Any application can find everything they need and the information is continue growing (McGlothlin and Khan, 2009) .

In order to fully define a knowledge we need a three piece of information, the first piece is the subject such as name, book, car, etc. The second part in the RDF is the property type (or predicate). There are many facts that are related to the subject like car color, college degree and so on. The intersection between the subject and the predicate is the third part of the triples which is the object (or value) for example, I (subject) have a name(predicate) which is Asaad (Value) (McGlothlin and Khan, 2009).

In RDF the subject is the thing being describes (in RDF terms) which is identified by a URI (Universal Resource Identifier). The predicate is the property of the resource (such as relationship, characteristic). The object is equivalent to the value of the resource property type for a specific subject (Powers, 2003). There are three important fact about RDF (Powers, 2003; Miller, 1998):

- Each RDF triples represent a complete and unique fact
- Each RDF triples is made up by :subject, predicate, object
- Each RDF triple can be joined with another RDF triple

The RDF directed graph consists of a set of nodes connected by arcs where the node come in three varieties uriref, blank node, literal. A uriref node consists of a URI (Uniform Resource Identifier) references that give a specific identifier that unique the node and it is better to point to a specific location on the web. The blank node is the node without URI. Within the directed graph the uriref

Fig. 2: RDF Directed Graph for a specific statement

```
_:j0 <http://www.w3.org/1999/02/22-rdf-syntax-ns#subject> <http://www.webreference.com/
dhtml/hiermenus> .

_:j0 <http://www.w3.org/1999/02/22-rdf-syntax-ns#predicate> <http://burningbird.net/
schema/Contains> .

_:j0 <http://www.w3.org/1999/02/22-rdf-syntax-ns#object>
"Tutorials and source code about creating hierarchical menus in DHTML" .

_:j0 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#Statement> .

_:j0 <http://burningbird.net/schema/recommendedBy> "Shelley Powers" .
```

Fig. 3: RDF N-Triples with generated blank node identifier

```
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:pstcn="http://burningbird.net/postcon/elements/1.0/">
<rdf:Description rdf:about="http://burningbird.net/articles/monsters3.htm">
<pstcn:author>Shelley Powers</pstcn:author>
<pstcn:title>Architeuthis Dux</pstcn:title>
</rdf:Description>
</rdf:RDF>
```

Fig. 4: RDF/XML for an article

resource node are drawn within Eclips and URI within circle and blank nod ein blank circle. RDF literals are drawn in rectangle. Figure 2 show a directed graph for a simple statement (Powers, 2003; Faye *et al*., 2012). There are many different kinds of file formats that can be used to serialize RDF like: RDF-XML, RDF N-Triples, N3.

The RDF N-Triples format break the RDF into separate triples one in each line. N-triples is based on another notation called N3 or Notation 3. The basic structure of N3 is: "subject predicate object" separated by space and ending with". "Each line in the N-Triples consists of either a comment or a triple but not both. Figure 3 show an example of the RDF N-Triple.

We can merge two or more RDF graph into one RDF according to the similarities between them. The Entailment of the RDF graph is the similarity between two graph in all aspect. For example forthat, the sub-graph lemma states that a graph entails all of its sub-graph because any assertion that can be made in the whole graph will also be made to the sub-graph. The monotonicity lemma in a graph state that if the sub-graph of the graph entails another graph, then the original graph can also entail that second graph (Powers, 2003).

In the context of RDF/XML, the schema or vocabulary is a rule-based dictionary that define the elements of importance to a specific domain, then describes how to relate each element with the other. Many type of these vocabulary like: -rdf: type, pstcn:bio. The RDF Schema is a domain-neutral way for describing the metadata that can later be used to describe the data for a domain-specific vocabulary (Powers, 2003). The RDF schema elements always marked by a specific namespace like:

Xmlns: rdfs = "http://www.w3.org/2000/01/rdf-schema#"

Within that schema specification, there is a core group of properties and classes that can be used to describe a domain-specific RDF elements. There are few RDF schema classes (Powers, 2003):

- Rdfs: Resource, all the RDF resources are member in this class
- Rdfs: Class, it explain the category of the resource
- Rdfs: Literal, it explain the literals such as text string
- Rdfs: XMLLiteral , it explain the literal that use XML syntax
- Rdfs: Container, it is a superclass for all container class
- Rdfs: Container MembershipProperty, it explain the members of containers
- Rdfs: Datatype, it explain the information of the data typing
- Explain a simple RDF/XML document that describe an article

The triples for that RDF/XML is shown in Fig. 4 and 5 whereas the graph for that RDF is shown in Fig. 6.

Number	Subject	Predicate	Object
1	http://burningbird.net/articles/monsters3.htm	http://burningbird.net/postcon/elements/1.0/author	"Shelley Powers"
2	http://burningbird.net/articles/monsters3.htm	http://burningbird.net/postcon/elements/1.0/title	"Architeuthis Dux"

Fig. 5: The Triples for RDF/XML

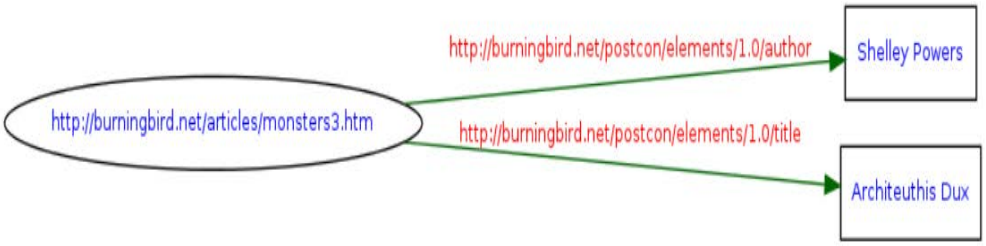

Fig. 6: Graph representing the RDF/XML

RESULTS AND DISCUSSION

Cluster analysis (clustering): Cluster analysis or simply clustering is the process for partitioning a set of observations into many subsets. Each of these subset called" a cluster and all its elements similar to each other (Han *et al*., 2012).

Cluster analysis has been widely used in many applications like: web search, security, image recognition, business intelligence. Clustering can also be called "data segmentation" in some application because it partitions a large data sets into groups according to their similarities. There are many classified into the following categories with its main characteristics (Han *et al*., 2012)

- Partitioning methods: it has the following general characteristics
  - Find mutually exclusive clusters of spherical shape
  - Distance-base clustering
  - May use mean or medoid to represent the center of the cluster
- Hierarchical methods: it has the following general characteristics
  - The clustering is done in multiple levels
  - Cannot correct erroneous merge or splits
  - May incorporate other techniques like microclustering
  - Density-based methods: it has the following general characteristics
  - It has the ability to find arbitrary shaped clusters
  - Each point must have a minimum number of points within its "neighborhood"
  - It has the ability to filter out of the outliers
  - Grid-based methods: it has the following general characteristics
  - Use a multi resolution grid data structure
  - Fast processing time where it typically independent of the number of data objects

K-means: centroid-based technique (Han *et al*., 2012): It is one of the Partitioning Methods that distribute the objects (n) in the data set (D) into clusters (k). The objective function is used to assess the quality of the partitions in the form that the objects within a cluster are similar to one another but dissimilar to objects in other clusters.

A centroid-based partitioning technique uses the Centroid of a cluster, C_i to represent that cluster. The difference between an object p C_i and c_i, the representative of the cluster is measured by dist (p, c_i), where dist (x, y) is the Euclidean distance between the point x and the point y. The quality of the Cluster C_i can be measured by the following equation:

$$E = \sum_{i=1}^{k} \sum_{p \in Ci} dist(p, ci)^2$$

Where:

E = The sum of the squared error for all the objects in the data set

p = The point that represent a give object

c_i = The centroid of the cluster C_i

This objective function tries to make the resulting k clusters as separate as possible.

K-means partitioning algorithms

Algorithm: k-means:
Input:
K: the number of clusters
D: a data set containing n objects.
Output: A set of k clusters.
Method:
Arbitrary choose k objects from the data set D as ainitial cluster centers;
Repeat
Assign (re-assign) each object to a specific cluster that the objects is the most similar, based on the mean value;
Update the cluster means by calculating the mean values of the objects for each cluster;
Until no change;

Framework architecture and implementation: The main idea of our framework can be explained in the following phases.

Phase-1: This is a preprocessing phase. We design an algorithm for reading the content for each website in RDF-XML Format then convert it into RDF-NTriples format in the form of (subject-predicate-object), then check the triples that share the same predicate which help us in specifying the type of that website according to the seven cluster. Also, save the voting value for each website that will be used in the next phases.

Phase-2: Clustering of all the category according to its type. We use K-means clustering where K = 7. The seven categories are:

- Media
- Geagraphic
- Puplications
- User-generated content
- Government
- Cross-domain
- Life sciences

After completing this phase, Fig. 7 can be seen as illustrated in Fig. 8. We use a sample of Fig. 7 in our work and make an estimation.

Phase-3: Calculate the number of node in each clusters which means the number of website in each cluster. Then put that number in a link between the DBpedia and that cluster as label in Fig. 8.

Phase-4: Use the idea of distributed genetic algorithm in each cluster in order to find the best way for retrieving the information in a minimum cost. The proposed steady state Genetic Algorithm (ssGA) given by the following steps:

Step 1: Generate the population according to the voting for each website where the chromosome form is given in Fig. 8. The chromosome length and population size is different for each cluster and (m) is the chromlength. Each part of the chromosome contain (the website-name, Website Vote). The voting for the website will be an integer value between (0,100). The Value '0' means No voting for that website.

Step 2: Calculate the fitness for each individual. The problem here is maximization. That means the maximum fitness is better than the minimum one in survival of the fittest selection. We propose new fitness function in which we divide the whole population size by 2, then we give the first part a positive Voting and give the second part a negative voting, then the fitness function is given by:

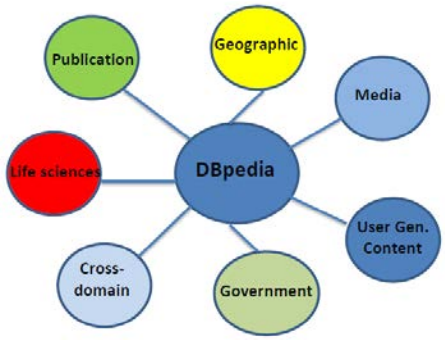

Fig. 7: The Clusters result from phase 2

Website1	Vote-1	Website2	Vote-2	…	Website-m	Vote-m

Fig. 8: Chromosome design for each cluster

Web-1	10	Web-2	15	Web-2	30	Web-4	45	Web-5	1	Web-6	0

Fig. 9: Chromosome

$$\sum_{i=1}^{(\text{chrom.Length}/2)} \text{Voting(chromosome[i])} - \sum_{j=i+1}^{(\text{Chrom\_Length})} \text{Voting(chromosome[j]}$$

Which means Fitness = Positive Voting+Negative Voting, then the positive fitness mean that the positive voting is larger than the negative voting. Ex. suppose we have the following chromosome:

Fitness = first part+(-second part) = first part-second part = (10+15+30)-(45+1+0) = 9

Step 3: Select the individual from the population. We use triple tournament selection. We choose three random individuals then choose the Best one from them and repeat that to take another individual. From this step we have two selected individuals.

Step 4: Crossover between the selected parent from Step 3. Because the values in the chromosome are integer we can't use the normal crossover algorithm, then we will use one of the following algorithm that are suitable for integer Encoding with probability within [0.5, 0.9]: partially Matched Crossover (PMX), Cycle Crossover (CX) order crossover (OX).

Step 5: The mutation mechanism is done by randomly choose two position in a chromosome and exchange its value. The mutation probability is taken between [0.1, 0.5], hopefully we can make some correction for the bad individuals.

Step 6: Use replacement strategy in which the best chromosome in a population have a chance to stay in the next generation whereas the bad individuals will replaced with a good offspring. Binary tournament replacement used in our study with probability within [0.5, 0.9].

Step 7: Repeat the previous steps (step-3 to 6) until we reach the population size.

Step 8: Check the stopping condition. The stopping criteria for our suggested algorithm is an on-line performance technique in which we reach the convergence when the average fitness of the population in current generation is near enough to the average fitness of the previous generation by selecting an appropriate value for that difference. At the end of this step we sort the individual in each population in decreasing order.

Phase-5: After completing phase-4, we have seven different size population that have the best websites for each cluster. According to the information user requirements from the DBpedia which is the core of the LOD in Fig. 1, we will take the next action: if the requirement need information from one cluster then the algorithm directly return the best website for that cluster according to the population, otherwise the algorithm take the minimum path to retrieve the information according to the values shown in Fig. 8 which represent the number of website in each cluster. We make an Estimation for the number of the website in each Cluster as follows:

No. of Website: cluster-1 (Value-1) = 10, cluster-2 (Value-2) = 30, cluster-3 (Value-3) = 50, cluster-4 (Value-4) = 25, cluster-5 (Value-5) = 40, cluster-6 (Value-6) = 75, cluster-7 (Value-7) = 60

All genetic algorithms for all clusters stop in an appropriate number of generation where the maximum generation is (200) for cluster-6 and the minimum number of generation is (97) for cluster-4.

CONCLUSION

As a conclusion for our work, we seen that the use of distributed genetic algorithm help end user by suggest the appropriate website for his information request.

For the future work we suggest to design an intelligent agent for navigating the website for each cluster and automatically update all the information. Also, design an algorithm for using evolutionary algorithm to find minimum path between the user request and the websites.

REFERENCES

Agapie, A. and A.H. Wright, 2014. Theoretical analysis of steady state genetic algorithms. Appl. Math., 59: 509-525.

Berners-Lee, T., J. Hendler and O. Lassila, 2001. The semantic web. Sci. Am., 284: 34-43.

Bizer, C., 2009. The emerging web of linked data. IEEE. Intell. Syst., 24: 87-92.

Bizer, C., J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak and S. Hellmann, 2009. DBpedia-A crystallization point for the web of data. Web Semant.: Sci. Serv. Agents World Wide Web, 7: 154-165.

Faye, D.C., O. Cure and G. Blin, 2012. A survey of RDF storage approaches. Arima J., 15: 11-35.

Florian, B. and M. Kaltenbock, 2012. Linked Open Data: The Essentials a Quick Start Guide for Decision Makers. Thomas Thurner Publisher, New York, USA.,.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning. 1st Edn., Addison-Wesley Publishing Company, New York, USA., ISBN: 0201157675, pp: 36-90.

Han, J.W., M. Kamber and J. Pei, 2012. Data Mining Concepts and Techniques. 3rd Edn., China Machine Press, Beijing, China, pp: 327-390.

Hogan, A., J. Umbrich, A. Harth, R. Cyganiak and A. Polleres *et al.*, 2012. An empirical survey of linked data conformance. Web Semantics Sci. Serv. Agents World Wide Web, 14: 14-44.

McGlothlin, J.A.M.E.S. and L. Khan, 2009. RDF Join: A scalable data model for persistence and efficient querying of RDF datasets. M.sc Thesis, The University of Texas at Dallas, Richardson, Texas.

Miller, E., 1998. An introduction to the resource description framework. Bulletin Am. Soc. Inf. Sci. Technol., 25: 15-19.

Nowostawski, M. and R. Poli, 1999. Parallel genetic algorithm taxonomy. Proceedings of the Third International Conference on Knowledge-Based Intelligent Information Engineering Systems, Audust 31- September 1, 1999, IEEE, Dunedin, New Zealand, ISBN: 0-7803-5578-4, pp: 88-92.

Powers, S., 2003. Practical RDF. O'Reilly Media Inc, Sebastopol, California, ISBN:0-596-00263-7, Pages: 53.

Wilkinson, K., C. Sayers, H. Kuno and D. Reynolds, 2003. Efficient RDF storage and retrieval in Jena2. Proceedings of the First International Conference on Semantic Web and Databases, September 7-7, 2003, ACM, Berlin, Germany, pp: 120-139.