International Journal of Engineering & Technology, *7 (4.19) (2018) 838-842*

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Promote Replica Management based on Data Mining Techniques

Rafah M. Almuttairi*[1], Mahdi S. Almhanna[2], Mohammed Q. Mohammed[3], Saif Q Muhamed[4]

[1]University of Babylon, [2]Information Technology, University of Babylon
[3, 4]University of information Technology and Communications
**Corresponding Author E-mail: rafahmohammed@gmail.com*

Abstract

The data grid technique evolved largely in sharing the data in multiple geographical stations across different sites to improve the data access and increases the speed of transmission data. The performance and the availability of the resources is taken into account, when a total of sites holding a copy of files, there is a considerable benefit in selecting the best set of replica sites to be cooperated for increasing data transfer job. In this paper, new selecrtion strategy is proposed to reduce the total transfer time of required files. Pincer-Search algorithm is used to explore the common characteristics of sites to select uncongested replica sites.

Keywords*: Replica Selection Technique, Replica Optimization , Data Grids, Cloud Computing, Association Rules.*

1. Introduction

Mirroring of data is the solution for many of the network-based applications such as the analysis of climatic data and network physics both of which require a response to the manipulation of the data sets on a large scale. Furthermore, if a multiple of identical copies exist, it would require replica management service to discover the identical copies that are available and select the best replica compatibles with the requirements of user.

The current work is about improving the selection mechanism in management system of data grid architecture. The proposed strategy chooses a group of replica sites. The selected sites feature has a similar characteristic in terms of the stability of the network conditions. At the same time, these sites work all together to send different parts of a large file or a set of different small parts of files. The transferring process could be done using the file transfer service Grid FTP or UDT as shown in Figure 1 below.

Data mining approach was used in our previous work [1, 2, 3, 4, 5, 7] to identify a set of associated sites. In the current work, we have improved the following:

1- Improving Selecting Score. In the current work selecting score has been expanded to enhance the performance of selection of the previous work [1, 2, 3, 4, 5, 7]. Additional factors are imbedded to the selection criteria equation such as: bandwidth, CPU load and availability of memory factors.

2- Minimize search time of selecting best sets of replicas. Pincer-Search algorithm is used instead of a priori algorithm which was used in the previous work. Pincer algorithm consumes less time in the search for a stable set of replica.

2. Traditional Replica Selection System in Data Grids

The computational elements can access all requested files listed in the task immediately in case the file reside in the local site of computing node. If not, requested files should be transferred from a storage element site. Files are replicated in several sites in Data Grids environment so, the need of choosing best site to get the required file from in a big challenge. The best site is the one with lesser transferring time.

Sites in a traditional model get the best site having the file by sending the Logical File Name, LFN to the manager of replica site in Replica Management System, RMS; the manager in turn gets replica Catalog that have required replica to determine all sites store it or part of it. After that physical locations should be determined and saved in Physical File Name, PFN. Network Monitoring Service such as *Ipref* is used to get the current network conditions (number of Hop, Bandwidth, and RoundTrip Time) between computing element node and all replica sites. In the traditional method the replica site which has the maximum bandwidth, or minimum number of hop or the leastround trip time amount is the Best Replica Site, BRS will be chosen to get the required file from. Data Transfer Service is used to transfer the file as shown in Figure1.

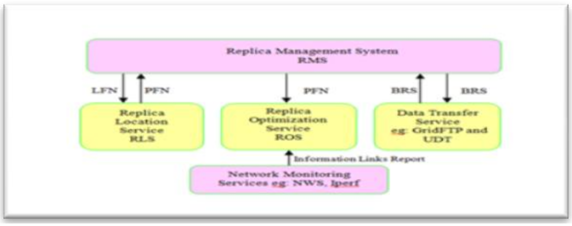

Fig1: Functionality of Replica Management System

3. Proposed Replica Selection Method

3.1 Overview of the Proposed Method

The aim of the proposed method is to enhance the selection of the best replica site using association rules of Data Mining approach.Pincer search is an association rules algorithm. It is used to find the hidden relation in the characteristic of replica sites. The set of replicas called Associated Replica Sites, ARS.
Figure 2 illustrate how replica sites can be grouped depending on its characteristics. In case there are 30 sites having required replica file. After executing the proposed method they have been clustered into three groups.

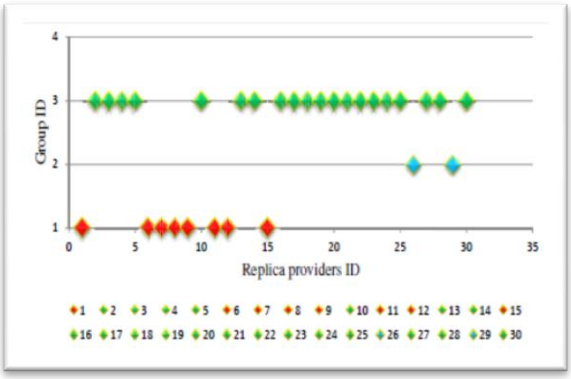

Fig. 2: Associated Replica Sites (ARS)

3.2 Function of the Proposed Method

The following functions are used in the proposed model to obtain Associated Replica Sites
Function 1: it is used to extract the round trip time between computing site and replica sites. It is done by obtaining the RTTs from network monitoring service report and save it in the text file, which is called as Network History File (NHF).
Function 2: **Mapping Function,** it is used to convert (NHF) into (LHF), The values of the table of the Logical History File are either 0 or 1. Mapping algorithm is applied on the (RTT) between replicas and computing site in a duration of time $[t_0, t_z]$.
Function 3: Efficient Set Technique (EST) function [1], discovers the Associated Replica Sites, ARS.

3.3 Simulation Tools

the following tools have been used in the practical part.
1) Ipref: PRAGMA Network is a real network environment provides Network History File "NHF" via network monitoring services called Ipref. NHF is used to obtain the Round Trip Time between replica sites.

Current work is to study the transfer of data in only one direction i.e, from the server to the client, hence, it would depend just on a Single Trip Time STT, Which is almost equal to*STT=RTT/2.*
2) Data Network Services: "OptorSim" helps to test and evaluate improvement strategies in the network. OptorSim proposed model simulator used for comparison between the selections methods.

4. Improve Selecting Score of the Replica Selection Technique

In the current work, the score of selecting best set of replicas has been improved. It updates the formula used to determine and calculate the weight. The score depends on the situations in the various devices of the servers that had been collected. Bandwidth and latency of the links are the most important factors that directly affect the data transfer process. Other two factors CPU and I/O slightly effect the performance of data transfer.
In our work, in accordance with the first three factors, the function of the scope is defined as in the following formula;

$$Scorei = pi^{BW} \times wi^{BW} + pi^{CPU} \times wi^{CPU} + pi^{I/O} \times wi^{I/O} \quad (1)$$

$$\text{And } Wi^{BW} + Wi^{CPU} + Wi^{I/O} = 1 \quad (2)$$

Where
Score i: The cost of selection model corresponding to the server iwhere $1 \leq i \leq n$
pi^{BW}: Representing the proportion of available bandwidth from the server i to the computing site; the current bandwidth is divided by the highest Theoretical Bandwidth
wi^{BW}: Network to Bandwidth ratio of replica of server i which has been determined by the Director of the Organization
pi^{CPU}: The percentage inertia of the CPU of the replica Server i.
wi^{CPU}: The CPU load percentage determined by the Director of the Organization.
$pi^{I/O}$: percentage of free memory space of server i
$wi^{I/O}$: The proportion of free space of memory defined by an official of the Organization of the data grid network
Factors affecting represent the wi^{BW}, wi^{CPU}, and $wi^{I/O}$ which represent the weight of all of the bandwidth and CPU load and I/O memory as shown in equation 2.
These weights can be determined by the Director of the Organization of the data network using Ipref service or Network Monitoring Service of Data Grids. According to the different properties of storage systems for the nodes in the data grid network.

5. Minimize Search Time of Selecting Replica Sites using Pincer Search Algorithm

Pincer Search Algorithm generates sets of correlating replicas. It discovers the hidden feature of link of connected replica at the execution time such as, latency; those sets of links have a similar characteristic (uncongested links). A frequent item set is another name of correlating replicas, it uses *NHF.*

Algorithm1: The proposed Model

- **Input:**
  NHF (Network History File)
- **Output:**
  LHF (Logical History File), MFS (Maximum Frequent Set)

- **Begin**

➢ **Step 1:** Calculate ($MSTT_{i;j}$) "the Mean of $STT_{k,j}$ "

$$MSTT_{i;j} = \frac{\sum_{k=i}^{(l-1)+j} (STT\ k,j)}{l}$$

➢ **Step 2:** Calculate the Standard deviation of STTi;j

$$STDEV_{i,j} = \sqrt{\frac{\sum_{k=i}^{(l-1)+j} (STTj - MSTTi,j)^2}{l}}$$

➢ **Step 3:**Calculate $Q_{i,j} = \frac{STDEVi,j}{MSTTi,j} * 100$

➢ **Step 4:** Find a Coefficient Variation of Replica provider, j using:

$AV_i = \frac{\sum_{j=1}^{m} (Qi,j)}{M}$, where M is the number of replicas, AVi represents average of the variation of $Q_{i;j}$

➢ **Step 5:** Classify links into stable and unstable using the condition.

IF ($AV_i \leq Q_{i,j}$) then LV_i= 0 Otherwise
LV_i= 1 /* LV represents a mapped Boolean Value of each STT*/

➢ **Step 6:** Calculate $Score_i$ usingEquation 1

➢ **Step 7:** Calculate the summation of all scores

$Sum_k = \sum_{k=1}^{M}$ Score k

- **Step 8:** Calculate Average of Scores , Avg= Sum_k/M
- **Step 9:** Construct Logical Score Value Table (LSV_i)

LSV_i= 1If $Score_k$>= Avg Otherwise LSV_i =0

- **Step 10:**Use **AND** Logical Gate to get Logical History File (LHF)

LHF=LVi ^ LSV_i , where i=1, M

- **Step 11:** Apply Pincer Search Algorithm

Input : Logical History File (LHF), Support (S)

Output: Maximum Frequent Set (MFS).

- **Step 12**: Use MFS sites to download required files via GridFTP.
- End.

6. Implementation

6.1 Beyond Network History File

To get Network History File (NHF),Ping report is a real Network History File of the given replica providers, it can be taken from Network Monitoring Service of Data Grid environment.

Table 2: Network History File

Checkpoint	Result	min. rtt	avg. rtt	max. rtt	IP
Australia - Perth (auper01)	OK	260.959	261.050	261.259	107.180.41.150
Australia - Brisbane (aubne02)	OK	225.185	225.308	225.570	107.180.41.150
Argentina - Buenos Aires (arbue01)	OK	197.326	200.739	202.621	107.180.41.150
Australia - Sydney (ausyd04)	OK	226.663	226.690	226.730	107.180.41.150
United States - Atlanta (usatl02)	OK	15.648	15.688	15.756	107.180.41.150
Australia - Sydney (ausyd03)	OK	213.467	213.573	213.675	107.180.41.150
Brazil - Sao Paulo (brsao04)	OK	172.639	175.201	177.051	107.180.41.150
Brazil - Porto Alegre (brpoa01)	OK	152.549	152.620	152.721	107.180.41.150
Canada - Vancouver (cavan03)	OK	89.505	89.561	89.731	107.180.41.150
Belgium - Antwerp (beanr03)	OK	128.776	132.492	134.371	107.180.41.150
Bulgaria - Sofia (bgsof02)	OK	107.502	107.563	107.662	107.180.41.150
India - Bangalore (inblr01)	OK	240.875	241.048	241.265	107.180.41.150
United States - Boulder (uswbu01)	OK	41.988	42.047	42.201	107.180.41.150

Table2 shows the hourly data of round trip time from Iraq site to all sites around the world at Feb 6, 2018. In Figure3 some links of sites are stable and others are not. Link of a site might be in few hours of the day stable and for few hours could be congested. Several sites enjoy together good latency of Round Trip Times as shown in Figure4. For this reason, a new method is used to discover the hidden relationship between the links that connect the replica sites. Pincer Search Algorithm is used to discover the hidden relations of replica links.

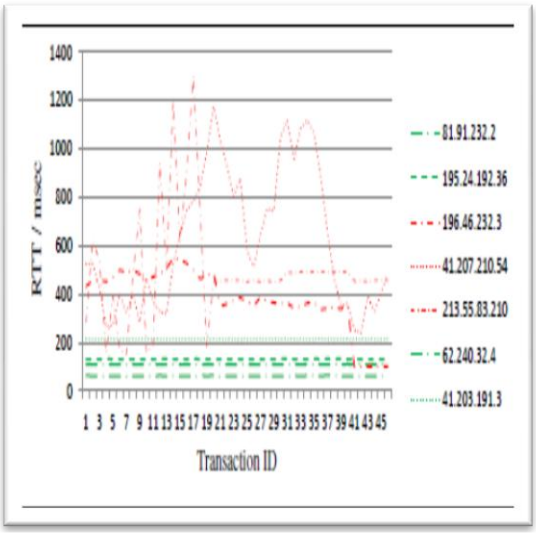

Fig. 3: Network History of seven data providers

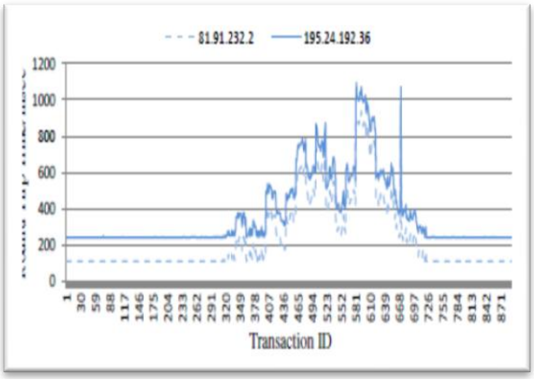

Fig. 4: RTT between cern.ch and waib.gouv.bj

Figure4 illustrates the link among waib.gouv.bj (Government of the Republic of Benin) and CERN site (CERN, the European Organization for Nuclear Research), with IP (81.91.232.2). The stability of the link differs from an hour to another. Thus, the link among the sites in the beginning of February 6, 2018 was stable, then in the mid-day became unstable, and then again became stable. The site /www.uobabylon.edu.iq/ with IP (5.77.37.165) has a stable link whereas the site www.netfind.com/Search with IP (80.249.75.2) has an unstable link as shown in Figure 5.

When the user transmits a request early in the day, both of the sites /www.uobabylon.edu.iq/ and www.netfind.com/Search have a stable link they will show in Associated Replica Sites "ARS" after applying the proposed strategy, whereas www.netfind.com/Search will be discarded at the mid of the day because at that time it is unstable [6].

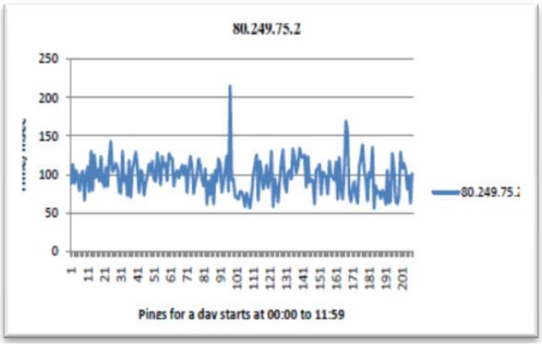

Fig. 5: RTT of univ-sba.dz

6.2 Study Case

In case there are eight replica sites distributed in different places. As illustrate in Table3. First the STT of eight servers. S_i={S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8} has to be calculated. Which is almost equal to *STT= RTT/2 as shown in Table 3 below.*Then, for the purpose of converting data of STT from a numerical number to logical number, algorithm 1 is executed on Table 3.

Table 3: Transactions Table of STTS Values (NHF)

STT	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
1	60	176	202	336	78.	7 256	76.4	213
2	65	211	209	34	376.5	256	86.8	202
3	303	175	298	338	64.8	258	57.2	205
4	305	213	203	273	92.6	255	85.1	202
5	300	210	223	334	95	292	55	212
6	313	176	207	335	66.7	298	55.3	212
7	310	175	298	273	94.2	255	85.8	202
8	307	216	260	271	94.4	256	60	212
9	310	217	260	342	95.2	289	90.1	212
10	310	211	224	339	66.3	257	90.3	212
11	310	175	204	272	92.8	262	91.4	202
12	307	176	205	271	69	256	88.7	210
13	310	211	227	344	92.5	299	64.3	212
14	50	175	202	270	216	66.3	299	90.5

15	316	214	260	336	63.4	296	57.8	224
16	74	209	206	341	94.3	287	56.4	222

Step 1-5:
Logical Values $LV_i$of Table 3 can be obtained after executing first five steps of Algorithm 1.

Table 4: Logical Values LV_i

S1	S2	S3	S4	S5	S6	S7	S8
1	0	0	1	1	1	1	1
1	0	1	1	0	0	1	1
1	1	0	0	1	1	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
1	1	0	1	1	1	1	1
1	1	0	0	1	1	1	1
1	1	0	1	1	1	1	1
1	1	0	1	1	1	1	1
1	1	1	1	1	1	1	1
1	1	0	1	1	1	1	1
1	1	0	1	1	1	1	1
1	1	1	1	1	1	1	1
0	0	0	1	1	0	0	0
1	1	1	1	1	1	1	1
1	0	1	1	1	1	1	1

Step 6: Calculate Score for S_i as shown in Table 5

Table 5: Scores values ofS_i= {S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8}

Score (S_1)	Score (S_2)	Score (S_3)	Score (S_4)	Score (S_5)	Score (S_6)	Score (S_7)	Score (S_8)
200	160	190	185	144	250	99	80

Step 7-8:Calculate Summationand Average of Scores

Table 6: Summation and Average of $Score_i$

Sum of $Score_i$	Avg. of $Score_i$
1308	163.5

Step 9: Construct Logical Score Values (LSV_i) to get Table 7 below

Table 7: Logical Score Values (LSV_i)

Score (S_1)	Score (S_2)	Score (S_3)	Score (S_4)	Score (S_5)	Score (S_6)	Score (S_7)	Score (S_8)
1	0	1	1	0	1	0	0

Step 10: Construct Logical History File (LHF)

Table 8: Logical History File Table (LHF)

S1	S2	S3	S4	S5	S6	S7	S8
1	0	0	1	0	1	0	0
1	0	1	1	0	0	0	0
1	0	0	0	0	1	0	0
1	0	1	1	0	1	0	0
1	0	1	1	0	1	0	0
1	0	0	1	0	1	0	0
1	0	0	0	0	1	0	0
1	0	0	1	0	1	0	0
1	0	0	1	0	1	0	0
1	0	1	1	0	1	0	0
1	0	0	1	0	1	0	0
1	0	0	1	0	1	0	0
1	0	1	1	0	1	0	0
0	0	0	1	0	0	0	0
1	0	1	1	0	1	0	0
1	0	1	1	0	1	0	0

Step 11: Apply Pincer Search Algorithm for Support (100% & 90%)
Input : LHF (Table 8) , Support (S = 100%) .
Output : MFS = {S_1}
Input : LHF (Table 8) , Support (S = 90%) .
Output : MFS = {S_1,S_6}
Step 12: Requested files will be concurrently transferred from S_1, S_6 Sites in case more than one site are required. Otherwise file can be transferred form S_1 only.
To explain the effect of using score of *Equation 1*. Next steps a Pincer search algorithm is applied on Table 4, where logical values had not reflect score effects. Step 11 will be repeated with new inputs
Step 11: Apply Pincer Search Algorithm for Support (100% & 90%)
Input : LHF (Table 4) , Support (S = 100%).
Output : MFS = {S_1,S_7,S_8}
Input : LHF (Table 8) , Support (S = 90%) .
Output : MFS = {S_1,S_6,S_7,S_8}
Analyzing Results:
To explain the effect of using score in our proposed method, the selected sites in MFS with score (Table 8) and without Score (Table 4) were compared.
MFS (Score) when the support is 100% was{S_1}, whereas MFS(without Score) is MFS = { S_1,S_7,S_8}. The network conditions of selected sites {S_7,S_8} are not fairly good to concurrently work and share transferring requested files at specified time. It might be better to get the file from a single site instead of getting it from more than one but, congested links As shown in Figure 6.
MFS(Score) when the support is 90% was{S_1,S_6}, whereas MFS(without Score) is MFS= {S_1,S_6,S_7,S_8}. The network conditions of selected sites { S_7,S_8} are not fairly good to concurrently work and share transferring requested files at specified time. It might be better to get the file from {S_1,S_6} sites, as they are uncongested links.The congested sites{S_7,S_8} has been neglected using score criteria as it is shown in the list of sites of MFS (Score) As shown in Figure 6.

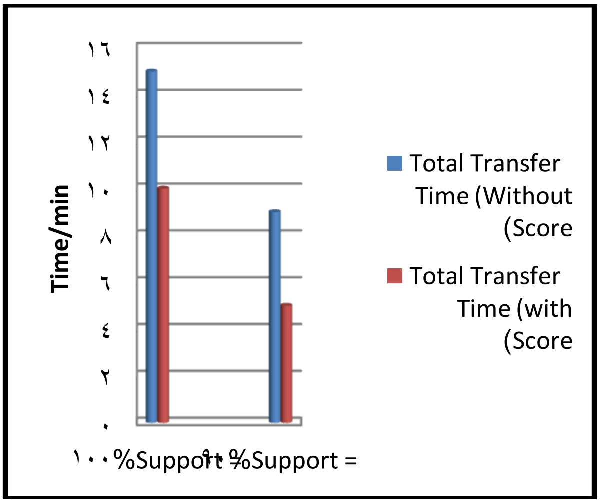

Fig. 6: Total Time for transfering required datausing GridFTB

7. Discussion and Conclusion

Cuurrent work is to enhanced replica selection strategy in data grid management system using data mining approach.The proposed strategy chooses best set of replicas dynamically in Data Grids environment. It has been compared with other methods to serve same job of getting replica files from several replica sites. Observations can be listed below:
Scalability. It is a scalable technique, because the output of the proposed strategy specifies a set of associated replica providers. The set of replica sites can be used as scalable nodes to divide the files among them.
Availability. The selection of most previous strategies depended on the previous requests, which is may or may not have the required files. Whereas, our approach has no possibility of non-

existence of files, it counts on the current roster of processors with the aid of the replica location service.

Transfer speed. All previous strategies need to re-run for number of times whenever data provider has to work concurrently for dealing parts of files. The selected sites in all other methods have no relation among them. However, some of them might be congested and others might be not. If the links are unstable, the percentage of lost packets becomes more and therefore, retransmission is needed which increase the total time of transfer files. In the proposed strategy, set of sits with a hidden relation can be chosen, the providers are correlated and all have stable links and can cooperate together to send the required file(s).

Easy to deployment. The proposed strategy spreads the solution very easily. It enables the organization to progress productivity and reduces transferring costs within Data Grid environments. The output of “Promote Replica Management Based on Data Mining Techniques” relies on the core services that are already available in the data network, such as the replica Location service and the Network Monitoring service.

Reference

[1] Rafah M. Almuttairi, R. Wankar, A. Negi, C.R. Rao, A. Agrawal, R. Buyya, A two phased service oriented broker for replica selection in data grids (2SOB), Future Generation Computer Systems (2012), doi:10.1016/j.future.2012.09.007.

[2] Rafah M. Almuttairi , Praveen Ganghishetti, Rajeev Wankar, C. Raghavendra Rao., “Rough Set Based Quality of Service Design for Service Provisioning in Clouds”,In Pro. of 6th International Conference on Rough Sets and Knowledge Technology (RSKT 2011) pp. 268-273, Lecture Notes in Computer Science 6954, Springer, 2011, ISBN 978-3-642-24424-7, October 2011, Canada.

[3] Rafah M. Almuttairi, R. Wankar, A. Negi, and C. R. Rao., “Replica selection in data grids using preconditioning of decision attributes by k-means clustering”. In Proceedings of the Vaagdevi International Conference on Information Technology for Real World Problems, pages 18-23, Los Alamitos, CA, USA, 2010. IEEE Computer Society.

[4] Rafah M. Almuttairi; Rajeev Wankar; Atul Negi; Raghavendra Rao Chillarige; Mahdi S. Almahna 2010 “New replica selection technique for binding replica sites in Data Grids”. 1st International Conference on Energy, Power and Control (EPC-IQ) ,Pages: 187 – 194,Year: 2010

[5] Mahdi S. Almahanna ; “A Dynamic Strategy to Transmit Files in Data Grids DSTF” / Journal of Babylon University/Pure and Applied Sciences/ No.(2)/ Vol.(25): 2017, DOI:10.1109/NTICT.2017.7976134, http://ieeexplore.ieee.org/abstract/document/7976134/

[6] Haider J. ABD,Mahdi S. Almahanna, “ Suppression of a Nonlinear Effect for High Data Transmission Rate Systems with a Wavelength Division Multiplier Using WITH A WAVELENGTH DIVISION MUL The Optimization of Fiber Properties.” ,UKRAINIAN JOURNAL OF PHYSICS, ISSN 2071-0186. Ukr. J. Phys. 2017. Vol. 62, No. 7. doi: 10.15407/ujpe62.07.0583

[7] Mahdi S. Almhanna, “Minimizing replica idle time”Annual Conference on New Trends in Information & Communications Technology Applications (NTICT), 2017, Conference Paper · March 2017, DOI: 10.1109/NTICT.2017.7976134

[8] Dao-I Lin; Z. M. Kedem, Pincer-search: an efficient algorithm for discovering the maximum frequent set , IEEE Transactions on Knowledge and Data Engineering , Year: 2002, Volume: 14, Issue: 3 , Pages: 553 – 566.