ISSN:0975-9646

Optimal Replication Strategy Using Rough Set Theory for Dynamic Content Replication

Shruthi[#1], Dr. Santosh L Deshpande[*2]

[#1] *Dept of Computer Networks & Engineering, Visvesvaraya Technological University*
Belgavi, India
[* 2]*Department of PG studies, VTU*
Belagavi, Karnataka, India

***Abstract*— Replication strategy is used to minimize the recovery time and data loss at an operational system and it attempts to reduce the system load, access latency and network congestion in a distributed system. Replication improves the reliability and scalability of the services. To increase availability we divide large files into small, medium and large fragments. When there are more number of hits for a data and when there is more delay in accessing a data we need to replicate it more. However these methods lead to high overhead for unnecessary file replication, fragmentation and consistency maintenance. Thus we place data on reliable servers. Any distributed database has three properties, consistency, availability and partition tolerance (CAP). In practice, at most two of these properties can be satisfied for any shared data. Thus a mathematical method called, Rough Set Theory (RST) can be used, that deal with vagueness and uncertainty in data and decision making. When available information in the database is insufficient, lower and upper approximations can be used to determine the exact value of a given set. Rough sets are created based on attributes and allocate memory for the sets. Based on ranking we decide whether it is fixed or variable size set, as we try to maintain time consistent data. The objective of this paper is to determine optimal data replicas using Rough Set Theory methodology, which groups data into sets of different granule size and generates a set of decision rules. This helps in determining which data set should be replicated more in order to provide better availability and consistent data.**

***Keywords*— Replication strategy, rough set theory, consistency of data**

I. INTRODUCTION

A distributed system is a framework in which resources are located on networked computers, communicate and coordinate their activities by passing messages. The components communicate with one another in order to accomplish a common objective. Three important features of distributed frameworks are: lack of a global clock, concurrency of components, and autonomous failure of components. A computer program that keeps running in a distributed framework is known as a scattered program, and dispersed writing of computer programs is the procedure of composing such programs. There are numerous choices for the message passing instrument, including RPC-like connectors and message sequence.

Replication:

Replication is a method used for sharing of information to guarantee consistency among redundant resources, to improve accessibility, reliability, and fault-tolerance. Replication is used to create and manage the versions of a database. Replication duplicates a database, and also synchronizes an arrangement of replicas with the goal that a change made in one replica is circulated to in all other replicas. Replication strategy is used to minimize the data loss and recovery time, at an operational system. The superiority of replication is that, it allows numerous users to work with their individual local replica of a database and yet have the updated database while though they were working on a distinct, centralized database. Replication is the most effective method for database access in database applications where users are geographically broadly distributed. The replication should be transparent. Such that when a failure occurs, a failure of replicas is hidden from the users as much as possible. Replication leads to reduction in the system load, access latency and network congestion. Replication provides high availability of data, consistent information delivery, high performance, easy centralized administration, heterogeneous data source access.

A distributed database has three properties. Consistency, Availability and Partition tolerance (CAP). Consistency is achieved when a read operation is guaranteed to return the most recent write for a given client. Availability is when every request receives a response. Partition tolerance mean that the system will continue to function when there is message loss or failure of a component of the system. CAP theorem states that only two out of three parameters can be satisfied. [6] We have a choice of “two out of three” CAP properties/Parameters which has three design options CA, AP and CP. CA is practically not a meaningful alternative, since a system that is not partition tolerant will be forced to give up consistency or availability during a partition. During a network partitioning a distributed framework must make a choice of either consistency or availability. The CAP theorem presents that the three properties are equally essential. Availability can be measured in spectrum and partition tolerance is measured in binary.

Replication means sharing of data so as to guarantee consistency among repetitive resources, to improve accessibility, reliability, and fault tolerance. Data replication means that the similar data is stored in several storage devices. In distributed systems every computer has a distinct copy of the shared data. Using replication more users can access the data that is shared in distributed environment, i.e., it makes data more available. But replication becomes a burden to make data consistent, i.e., most recent copy of the data should be updated in all the replicas in order to achieve consistency. As the data is replicated in more network partitions, we need to update the data in all the partitions to make data consistent and most recent copy of the data to be available to users.

Consistency models are used as a part of distributed frameworks like distributed shared memory systems. Consistency means concurrent reads and writes on shared replicated state. Consistency models are used to maintain consistency of data. Inconsistency can be handled by characterizing the semantics of data integration system and obtaining consistent database from inconsistent database.

Database replication can be done using following techniques:

- Snapshot replication: Information on one server is basically duplicated to another server, or to another collection of database on the same server.
- Merging replication: Information from two or more databases can be connected into a single database.
- Transactional replication: Users get complete initial duplicates of the database and consequently receives periodic updates as information changes.
- Mirroring: With mirroring data will be copied in the same system in a back up disk, such that when the primary replica fails, mirrored copy is used. There is a lot of overhead in making multiple mirrored copies of data and managing them.

But these replication technologies have drawbacks in providing correct replica, such as failure of servers, inconsistent data copies, and access latency etc. Thus a new method called rough set theory (RST) can be used. Rough set theory is a new mathematical approach that deals with imperfect knowledge. The rough sets approach use lower and upper approximations and certain possible rule sets concepts to determine the exact value of a given set. Depending upon these approximations we decide which data set should be replicated more. [7]

For defining a set we have to point out its elements. The membership function, distinguishing the belongingness of elements of the universe to the set, can gain one of the two values (0 or 1). Which mean that each and every element of the set is either in or outside of the set under consideration. [9] This description of the membership function doesn’t take into consideration the vagueness or uncertainty of being an element of a given set of elements.

Choosing suitable rough set model for data analysis is a major difficulty in using rough sets. The two rough set models, the probabilistic model and the Pawlak model, gives significance for the decision goals of a user. To define regions in the probabilistic model probabilities are used in these two approaches. These approaches either derive the probability thresholds from the cost associated with making a classification or user-defined parameters. We notice that the accessibility of data with respect to the data analysis is vital for selecting a suitable rough set approach by deciding the ramifications of the outcomes acquired from these models and methodologies. Corresponding to the available data and user needs a list of decision types can be presented. These outcomes may help a user match their decision prerequisites and desires to the model. [8]

A probabilistic methodology has been applied to the theory of rough set in information-theoretic analysis, decision-theoretic analysis, and variable precision analysis [5]. They are characterized depending upon a pair of thresholds presenting the desired levels of accuracy. Bayesian decision-theoretic analysis is followed to give a systematic methodical strategy for deciding the accuracy parameters by utilizing more familiar conception of costs and risks. The rough set approach seems to be of fundamental importance to cognitive sciences and artificial intelligence, particularly in the areas of knowledge acquisition, machine learning, and decision analysis, expert systems, knowledge discovery from databases, pattern recognition and inductive reasoning.

II. Proposed Methodology

Replication gives high availability of data to users by making different copies (replicas) of same data in network, by which network performance can be improved. The architecture of the proposed system consists of several users accessing a network through proxy server which has access to distributed database.

The details of the proposed system as follows;

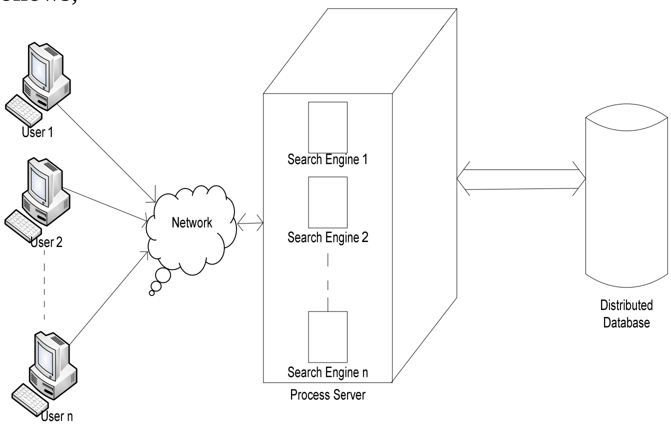

Fig 1 Block diagram of optimal replication strategy using rough set theory for dynamic content replication

Rough set theory is based on the assumption that, with every object in the universe there is some information associated. Objects characterized by the same information have similar view of available information about them. The similarity relation generated in this way has mathematical basis for rough set theory. Elementary set is used to select any set of all similar objects, and form a basic granule of knowledge about the universe. The crisp set is used to select any union of the same elementary sets; otherwise the set is rough (rough sets).

The approximation of lower and upper spaces of a set is the fundamental concept behind rough set theory, the approximation of spaces being the formal classification of knowledge regarding the interest domain. The subset generated by lower approximations is characterized by objects that will definitely form part of an interest subset, whereas the upper approximation is characterized by objects that will possibly form part of an interest subset. Every subset defined through upper and lower approximation is known as Rough Set. It allows reducing original data, i.e. to find minimal sets of data with the same knowledge as in the original data. Below is the state diagram of the proposed system, depicts the processing of a user request for accessing a random number from the distributed database.

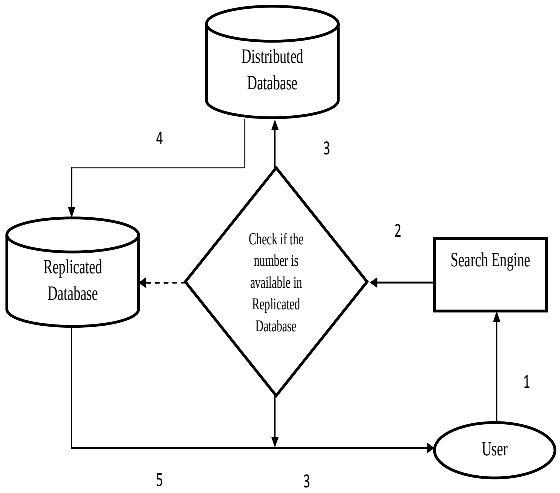

Fig 2 State diagram

STEPS:

1. User requests for a random number.
2. Check if the requested number is available in the replicated database.
3. If YES then replicate the number to the user. If NO then search in the main database.
4. Replicate the number from the main database to the replicated database.
5. Replicate the requested number from replicated database to the user.

The structure of the database (sets database) in the proposed system is shown below. Data is stored in the form of sets. Here the sets also contain overlapped segments, which have similar data between two sets. With this data access for users can be made easy.

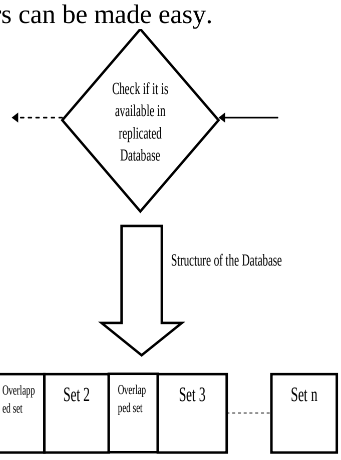
Check if it is available in replicated Database
Structure of the Database
Set 1
Overlapped set
Set 2
Overlapped set
Set 3
Set n

Fig 3 Structure of the sets database

The concept of rough set theory can be understood with the following illustration; several parameters of a network (network system) can be considered, and grade the parameters as high, average and low depending on importance of the parameter. By taking into account the similarity of the grades for individual parameter, one can group the parameters and reduce it to have a common set, which is called as rough set.

The data set can be considered as Information System (IS),

IS= {U, A}

Where, universe U and attributes A correspond to a set of objects and to set of variables respectively.

U= {x1, x2, x3, x4, x5, x6… x10}

A= {size, hits, number of users, paths, latency, failure of servers, power and bandwidth}

The domains of the particular attributes are:

V= {replication, availability, consistency}

Grades are gives as follows:

High=3

Average=2

Low=1

The information function for this system is presented in Table I.

Let us group all objects based on the grades for three variables considered. The results are presented in Table II.

Lower and upper approximation for the system is as follows:

Lower approximation is given by BX= {x2, x3, x5}

Upper approximation is given by BX= {x1, x2, x3, x4, x5, x6, x7}

Boundary of X in U={x1, x2, x3, x4, x5, x6, x7}-{x2, x3, x5}

={x1, x4, x6, x7}

Cardinality ratio of the upper and lower approximation is as follows:

Card (BX) = 3

Card (BX) = 7

Accuracy set of X is given by,

X= µ(X) = 3/7

= µ(X) <1

TABLE I
ELEMENTARY SETS OF UNIVERSE U IN SPACE A

Parameters(A/X)	Replication	Availability	Consistency
Size(x1)	3	3	2
Hits(x2)	3	2	1←
Number of parameters(x3)	3	2	1←
Paths(x4)	3	3	2
Latency(x5)	3	2	1←
Failure of servers(x6)	2	1	1
Power and bandwidth(x7)	2	3	2

TABLE IIi
ELEMENTARY SETS OF UNIVERSE U IN SPACE B

Parameters(B/X)	Replication	Availability	Consistency
(x2, x3, x5)	3	2	1
(x1, x4)	3	3	2
(x6)	2	1	1
(x7)	2	3	2

TABLE III
EXPERIMENTAL RESULTS

Data Set	Granule Size	Number of Level 2 Threads	Number of Level 3 Threads	Number of Requests from 3rd Level	Number of Requests that went to Main Thread	Number of Hits	Hit Rate %	Miss Rate %
1000	5	3	100	4200	127	4093	97%	3%
1000	5	3	200	11391	170	11271	99%	1%
1000	10	3	100	5149	169	4980	97%	3%
1000	10	3	200	10731	164	10567	98%	2%
1000	20	3	100	5502	160	5342	97%	3%
1000	20	3	200	9690	162	9528	98%	2%
1000	50	3	100	5962	181	5781	97%	3%
1000	50	3	200	9310	143	9167	98%	2%
1000	100	3	100	3900	178	3722	96%	4%
1000	100	3	200	7890	109	7781	99%	1%
1000	500	3	100	3259	101	3158	97%	3%
1000	500	3	200	8972	137	8835	98%	2%

It means the set is rough set; otherwise it is crisp set where no elements are member of the set. By applying this method to network parameters which are of greater importance, the parameters can be grouped into one set, which have similar grades. Depending upon these grades one can work for betterment of the network performance.

This is proved using a simple program, which generates random numbers between 0 to 999 and it is considered as main set. In the second level these random numbers are equally shared among three sets (taken as threads). And then these numbers are grouped into segments of different granule size, in which some segments consist of same numbers. These segments are considered as overlapped segments which may contain data from any of the two or three sets. And when a random number is generated, it can be replicated to user from one of the three sets at the second level or from overlapped segments. Otherwise the number is replicated to the user from the main set. This experiment shows that grouping of the data gives better availability of the data and the system performance can be improved by using rough sets.

III. EXPERIMENTAL ANALYSIS

Table III shows experimental results.

Correlation between number of requests to number of hits= 0.99

Correlation between number of requests to number of misses= 0.35

IV. OUTCOMES

The table gives the results requests from various numbers of users requesting for a random number between 0 to 999.

Better availability and increased hit rate by grouping the data into segments of different granule size and by overlapping of segments. CAP parameters are satisfied. And the results show that, better availability and consistency can be achieved by partitioning of the data into different segments. The system performance is improved.

V. CONCLUSION AND FEATURE WORK

Replication of the data gives high availability, reliability and better performance of the system. The experimental result shows that CAP theorem parameters can be satisfied. And the system performance can be improved using the Rough Set Theory, where several parameters of a network (network system) are considered. And grade these parameters as high, average and low depending on importance of the parameter. By taking into account the similarity of the grades for individual parameter, one can group the parameters and reduce it to have a common set, which is called as rough set. It provides efficient methods, algorithms and tools for finding hidden patterns in data. It allows reducing original data i.e. to find minimal sets of data with the same knowledge as in the original data. It is best suited for concurrent/distributed processing.

The rough set approach seems to be of fundamental importance to cognitive sciences and artificial intelligence (AI), particularly in the areas of knowledge acquisition, machine learning, and decision analysis, expert systems, knowledge discovery from databases, pattern recognition and inductive reasoning.

REFERENCES

[1] R. Jimenez-Peris, M. Patino-Martinez, G. Alonso, B. Kemme "How to select a replication protocol according to availability, communication overhead and scalability" Reliable Distributed systems, May 2001.

[2] Chun-chen Hsu, Chein-min Wang, Pangfeng Liu "optimal replication transition strategy in distributed hierarchical systems" Parallel and Distributed Processing, April 2008.

[3] Xin Sun, Jun Zheng, Quiongxin Liu, Yushu Liu "Dynamic data replication based on access cost in distributed systems" Computer Sciences and Convergence Information Technology, November 2009.

[4] Didier Dubois, Henric Prade "Rough fuzzy sets and Fuzzy rough sets" Transactions on Fuzzy Systems, April 2007.

[5] Eric C. C. Tsang, Degang Chen, Daniel S. Yeung, Xi-zhao wang, John W. T. Lee "Attributes reduction using fuzzy rough sets" IEEE Transactions on Fuzzy Systems, October 2008.

[6] Seth Gilbert, Nancy A. Lynch "Perspectives on the CAP Theorem" 2012.

[7] Thabet Slimani "Application of Rough Set Theory in Data Mining" November 2013.

[8] Joseph P. Herbert, JingTao Yao "Criteria for choosing a rough set model" Computers and Mathematics with Applications, 2008.

[9] B. Walczak, D.L. Massart "Tutorial Rough set theory" 1999.