International Journal of Linguistics and Computational Applications (IJLCA) ISSN 2394-6385 (Print)
Volume 3, Issue 2, April– June 2016 ISSN 2394-6393 (Online)

A Comparative Study of Hierarchical Clustering in Heterogeneous Environment

NehruRevathi[#1], R.Priya [*2]

[1]*Research scholar, Dept of Computer Applications, Vels University, Chennai*
[2]*Assistant Professor, Dept of Computer Applications, Vels University, Chennai*

***Abstract*—** In this project, the Hierarchical clustering is one of the most popular clustering methods that can find hierarchical structure hidden in input data for data analysis. To perform this method, one needs to define the notion of distance between two clusters.

The inter-cluster distance is defined as that the distance between the shortest line connecting two clusters, for connecting single linkage. For the full linkage, the inter-cluster distance is defined as that the longest straight line between the two clusters. For the average linkage, the inter-cluster distance is defined as that the distance between the cluster members. For the centroid linkage, the inter-cluster distance is defined as that the distance between the centers of the two clusters. Maintaining the inter-cluster distance globally, is the challenging one in the wireless network for sharing large data packets.

In a wireless network, the large number of flows can cause congestion. Therefore, a flow control mechanism is required that the amount of data inside the network. In a wireless networks, the bandwidth is a shared resources. The common assumption is that the bandwidth available to a node is shared within the interference range of the node, this affects the available bandwidth inside a network. The estimation of bandwidth in a network results in reducing the overhead of the data transmission. The difficulties in the wireless network is that are Interferences, Absorption and reflection, Multipath fading, Hidden node problem, Resource sharing problem, Capacity, Channel, Network, Safety, etc.

***Keywords*—** Clustering; Network; Bandwidth.

1. Introduction

A wireless network is any type of computer network that uses wireless data connections for connecting network nodes .Wireless networking is a method by which homes, telecommunications networks and enterprise (business) installations avoid the costly process of introducing cables into a building, or as a connection between various equipment locations. Wireless telecommunications networks are generally implemented and administered using radio communication .

The wireless network offers various types are wireless PAN, wireless LAN, wireless mesh network, wireless MAN, wireless WAN, Global area network (GAN), Space area network. For example, Inter-continental network systems, use radio satellites to communicate across the world.

2. Proposed System

In the proposed system, the hierarchical clustering and the neighborhood node techniques are integrated to find the distance between each node from source to destination for the data transformation in a wireless network. The riskless routing protocol is used for the data packets are transferred faster to the destination even the congestion occurs, the data is transferred in a different path when the congestion occurs at different timeslot.

The E-BandEstim method, a novel available bandwidth based flow admission control algorithm for wireless networks. Novel algorithms for estimating intra-flow contention and estimating contention on non relaying nodes, additional MAC layer overhead associated with an increased data traffic load on non relaying nodes, and an algorithm that deals with concurrent admission requests in a first come first serve scheme and also performance analysis can be carried out.

Accurate distance will be calculated from source to destination for data transfer even for the large data packets. If any of the congestion occurs while sending the data in a network, it automatically takes another path to reach the destination but the time varies.

3. Overview

In a wireless network, the large number of flows can cause congestion. Therefore, a flow control mechanism is required that the amount of data inside the network. In a wireless networks, the bandwidth is a shared resources. The common assumption is that the bandwidth available to a node is shared within the interference range of the node, this affects the available bandwidth inside a network. The estimation of bandwidth in a network results in reducing the overhead of the data transmission.The difficulties in

the wireless network is that are Interferences, Absorption and reflection, Multipath fading, Hidden node problem, Resource sharing problem, Capacity, Channel, Network, Safety, etc. The main advantage of using wireless network is that are, Increased mobility and collaboration, Improved responsiveness, Better access to information, Easier network expansion, and enhanced guest access.

4. Literature Review

In the Existing system, Hierarchical agglomerative clustering (HAC) is a clustering method widely useful for discovering hierarchical structure embedded in input data. Then proposed a multi-threaded algorithm allocates available threads into two groups, one for managing NN (Neighbor node) chains and the other for updating distance information. In-depth analysis of this approach gives insight into the ideal configuration of threads and theoretical performance bounds.

The evaluation this method by testing it with multiple public datasets and comparing its performance with that of several alternatives. In this test, the proposed method completes hierarchical clustering 3.09-51.79 times faster than the alternatives.

5. Implementation

5.1 Cluster Formation

A. Cluster

Clustering is the task of grouping a set of objects in such a way that objects in the same group called a cluster.

B. Cluster Measurement

For internal : d(I,J) = min D(i,j) where iϵI , jϵJ
For complete: d(I,J) = max D(i,j) where iϵI , jϵJ
For average: d(I,J)= 1/|I| |J| ∑ ∑ D(i,j) where iϵI , jϵJ
For centroid: d(I,J)= D(Ic, Jc)
Where i , j are the two data objects,
D(i,j) = distance between two clusters
I and J are two distinct clusters
d(I,J) = distance between I and J

5.2 Surrogate Path

The alternate path for data transmission through the network when the congestion occurs, based on ip addresses the path will be discovered.

5.3 Riskless Routing Protocol

The integration of Secure Routing Protocol and Nearest Neighbourhood Protocol is said to be Riskless Routing Protocol.

A. Nearest Neighbourhood Protocol

In its original form, the NN-chain algorithm considers only one chain at a time. However, given that maintaining a chain corresponds to searching for clusters that can be merged, there exist ample opportunities for parallelization. Our approach not only exploits these opportunities but further mitigates the impact of starvation and deadlock events which pose a challenge to effective parallelization of the impact of starvation and deadlock events which pose a challenge to effective parallelization of the NN-chain algorithm. In depth analysis of the parallelization scheme reveals its theoretical performance limits and ideal thread configurations for maximum performance.

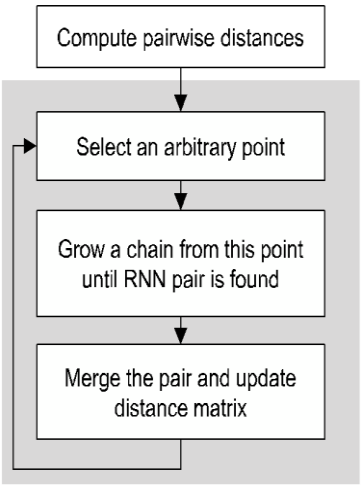

Fig .1: NN chain algorithm

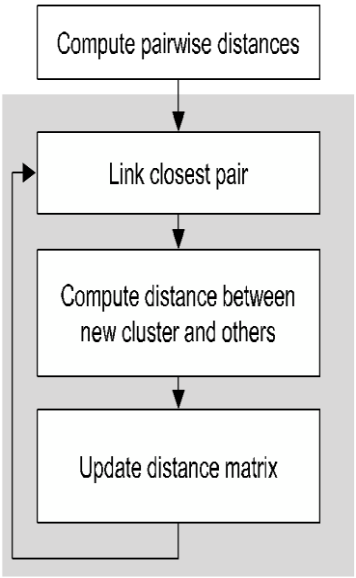

Fig.2: Hierarchical Clustering Algorithm

5.4 Secure Routing Protocol

A. Route Discovery Process

Route discovery process employed by Secure Routing Protocol is almost similar to that of HWMP. The source node S initiates route discovery process for establishing a path to the destination D by broadcasting an RREQ. The RREQ processing rules of SRP are similar to that of HWMP apart from an additional verification required to validate an RREQ. Any intermediate node that receives a broadcasted RREQ verifies the validity of two-hop address present in the RREQ (i.e. node I checks whether the two-hop address received in RREQ is present in its neighbourhood information) along with the usual validation process of HWMP. On validating the RREQ, node I creates a routing entry for the corresponding RREQ-ID, sets its state as transient and rebroadcasts it. Otherwise, it drops the RREQ.

B. Route Selection Process

The primary goal of SRP route selection process is to select congestion free paths. Nodes monitor the received RREQ's for a necessary and sufficient condition to classify a path to be free from congestions. Once an RREQ is verified to be congestion free, the corresponding routing entry is elevated to stable state from transient state. The route selection process is shown in Algorithm 5.1.

Consequent upon receiving a new RREQ (RREQN) , the intermediate node I processes it to take an appropriate action. Initially, the intermediate node I verifies if a transient routing entry corresponding to the RREQ-ID and sequence number received in RREQN exists. Multiple transient routing entries may exist for the same RREQ-ID that are received through a unique two-hop node. Node I then compares the two-hop address present in RREQN with the two-hop addresses of a set of existing routing entries represented by {RREQO }. If it matches with any of the existing routing entries RREQO , it is updated with RREQN provided that it offers better metric. In case of no matching address, node I further compares the three and four-hop addresses present in,

Frame Control	Duration	Addr1	Addr2	Addr3	Sequence Control	HT Control	Flag	Neighbor Address	1 CS

Algorithm: Route selection process. On receiving $RREQ_N$ by an intermediate node I

1: if no routing entry exists for S then
2: if (2Hop is valid) then
create corresponding RENTRY
state ← transient broadcast RREQN
3: else
drop RREQN
4: end if
5: else
6: if (RREQID , SEQ− No, 2Hop are valid) then
7: for(all routing table entries)
8: if (RREQN _2Hop ==RENTRY2Hop) then
9: if (RREQN _Metric < RENTRYMetric) then
update(RENTRY ← RREQN)
10: else

11: end if
12: else
drop RREQN
13: if (RREQN _2Hop ==(RENTRY3Hop | RENTRY4Hop)) then
update(RENTRY ← RREQN)
state ← stable
14: else
update(RENTRY ← better(RREQN , RENTRY))
end if
16: else
← stable
create routing entry for RREQN
state ← transient
broadcast RREQN
17: end if
18: else
drop RREQN
19: end if
20: end if

in RREQN with the two- hop addresses of routing entries represented by RREQO or vice versa. If any one of the two addresses match (two- hop address in RREQN with three-/four-hop addresses of RREQO or vice versa), provided the 2HA of RREQN does not match with transmitter address of RREQO or vice versa, an optimal of the two RREQ's (RREQO or RREQN) is selected and state of the routing entry is set to stable. If none of the comparisons match, a new transient routing entry is created for the corresponding RREQ-ID.

This matching of addresses is carried out to select an optimal congestion free path. The necessary and sufficient condition for detecting a congestion free path. Finally, if an intermediate node I receives an RREQN when it already has a stable routing entry to a destination D, I processes the RREQN only if the new route request offers a better metric than the existing route. SRP creates a separate routing entry for

RREQN and updates the existing stable entry with RREQN , only after it has been verified to be free from congestions. This process assures the selection of a congestion free path.

Route reply process: Like any intermediate node I, the destination node D processes multiple RREQs before selecting an optimal congestion-free path, satisfying the route selection criteria. It unicasts an RREP through which a stable the RREQ has been received. Subsequently, intermediate nodes propagate the RREP through congestion free routers.

Route maintenance: Route maintenance in SRP is similar to that of HWMP. Whenever a node I discovers a link failure, it initiates the route maintenance process by transmitting a RERR message addressed to the source. Node I can optionally initiate route discovery process on behalf of the source to reduce the route selection latency. Intermediary nodes on receiving a RERR message mark the corresponding routing entry and propagate the RERR message towards the source S. The source S on receiving the RERR message can reinitiate the route discovery process by broadcasting an RREQ.

Table.2: Notations and their Meaning in Securing Routing Protocol

Notation	**Meaning**
RREQ	Route request
RREQID	Route request identity
SEQ− No	Route request sequence number
RREQN	Newly received route request
RREQO	Existing routing table entry
2Hop	Two-hop address in route request
RREQN_iHop	ith hop address in RREQN
RENTRY	Entry in routing table
RENTRYiHop	ith hop address of routing entry
RREQN_Metric	Routing metric in RREQN
RENTRYMetric	Routing metric of a particular routing entry

8. Conclusion

Addressing congestion in a network is a crucial issue to ensure secure data transmission. In this paper, proposed a Riskless Routing Protocol (RRP) that relies on shorter alternate paths to detect a congestion in the path. During route discovery, SRP monitors for alternate paths for a cached RREQ and quarantines such RREQ that fails to meet the necessary and sufficient condition. And also, proposed an approach to parallelization of hierarchical clustering based on the nearest-neighbor chain algorithm. The algorithm grows multiple chains simultaneously by partitioning available threads into two groups, one for growing chains and the other for updating the distance matrix. The proposed method will be helpful for extending the applicability of hierarchical clustering to large-scale data that otherwise cannot be analyzed due to the limited scalability of current approaches.

Summary and Future Work

In this paper the implementation for the data transmission in congestion free path without any lose of data and also, briefly explained about the distance calculation and path discovery and its process.

In future, the project may be extended to estimate the bandwidth by measuring the bandwidth using its parameters (are Time to Live TTL , Round Time Trip RTT , File size) and the performance analysis can be carried out

References

[1] Yongkweon Jeon, Student Member, IEEE and Sungroh Yoon, Senior Member, (2015) introduced "Multi-Threaded Hierarchical Clustering by Parallel Nearest-Neighbor Chaining" IEEE transactions on parallel and distributed systems.

[2] Han.J , Kamber.M , and Pei.J , (2006) introduced "Data Mining : Concepts and Techniques", San Mateo, CA, USA : Morgan Kaufmann, 2006.

[3] Frank.E and Witten.I. H , (2005) introduced "Data Mining : Practical Machine Learning Tools and Techniques," San Mateo, CA, USA : Morgan Kaufmann, 2005.

[4] Day.W and Edelsbrunner.H , (1984) introduced "Efficient algorithms for agglomerative Hierarchical clustering methods," J. Classification, vol. 1, no. 1, pp. 7–24, 1984.

[5] Defays.D. , (1977) introduced "An efficient algorithm for a complete link method," Comput. J., vol. 20, no. 4, pp. 364–366, 1977.

[6] Khabbazian.M, Mercier.H , Bhargava.V.K , (2009) introduced "Severity analysis and countermeasure for the wormhole attack in wireless ad hoc networks". IEEE Trans. Wireless Commun. 8(2), pp.736-745.

[7] Choi.S , Kim.D.Y , Lee.D.H , Jung.J.I , (2008) introduced "WAP : Wormhole attack prevention algorithm in mobile ad hoc networks" IEEE International Conference on Sensor Networks, Ubiquitous and Trustworthy Computing. (IEEE, Taichung, 11–13 June 2008), pp. 343–348

[8] Wang.X , Wong.J , (2007) introduced "An end-to-end detection of wormhole attack in wireless adhoc networks" Thirty-First Annual International Computer Software and Applications Conference. (IEEE, Beijing, 24–27July'07).

[9] Tun, AH Maw.Z , "Wormhole attack detection in wireless sensor networks". J. World Acad. Sci. Eng. Technol. 46(2), 545–550 (2008)

[10] Rakesh Matam , and Somanath Tripathy , (2013) introduced "WRSR : wormhole-resistant secure routing for wireless mesh networks", Springer 2013.

[11] A.S.Aneeshkumar and Dr. C.Jothi Venkateswaran, "A novel approach for Liver disorder Classification using Data Mining Techniques", Engineering and Scientific International Journal, Volume 2, Issue 1, January - March 2015, pp.15-18.