A Distributed Information Retrieval Strategy Based on Route Marking

Lijun ZHOU[1,a], Yuan ZHAO[1,b], and Haiyan LV[1,c]

[1]Naval Aeronautical and Astronautical University, YanTai 264001, China

[a]jungle730@163.com, [b]ascendtop@126.com

Keywords: Distributed information retrieval, Flooding algorithm, Route marking

Abstract. By analysing the flooding strategy which is usually used in the distributed information retrieval, we found that the strategy is easy to duplicate retransmission with the same message, which results in a lot of redundant communication, and causes network congestion.The paper proposed a retrieval strategy to improve the flooding algorithm and demonstrated the avalidity of the strategy through theoretical analysis. Finally,it carries on the comparison of two kinds of strategy analysis with examples, and proves the effectiveness of the latter by theoretical derivation.

Introduction

In the open network environment, bandwidth is unevenly distributed, resources are disordered with a high degree of autonomy, and are complex and diverse, traditional centralized resource management mode is difficult to follow, distributed resource management has become a kind of trend, and distributed information retrieval is the basis of building a distributed resource management system.

Compared with the traditional centralized information retrieval, distributed information retrieval is that makes spread geographically distributed and logical link between the number of database through the computer network integration into a whole virtual giant databases provide service for the user. In this case, different database stored information index of different themes, at the time of receiving the user retrieval request, related to the retrieval task was assigned to the corresponding server on the network. Flood information retrieval is usually based on flood mechanism. In order to suppress the serious traffic congestion caused by flooding, domestic scholars have studied a series of improving methods. Chen Duolong etc puts forward a method of restricted flooding based on negative selection[1] . Dou Wen etc proposed a probabilistic broadcast mechanism based on the structure free P2P network [2].

Flooding Strategy

The process of flooding mechanism is realized as follows: any node which receives the searching messages retransmits to neighboring nodes to guarantee the query request be possible to send to the system coverage of every node, but when the same node receives the same message for many times, the number of message retrieval will be exponential amplification, producing a large number of redundant information, and resulting in serious bandwidth consumption. With the increase of nodes, the garbage communication is also increasing, which can cause the network congestion of the whole system and cause the failure of the single point.

1.The retrieval node is a node that transmits request in the retrieval process, we denote it as: n.

2.The degree d of the network node n is the number of neighbor nodes that is connected to the node.

3.The average times per query request be retransmitted in a retrieval process is denoted as the query overhead : f.

Then, we derive the following formula:

$$f = \frac{1}{R}\sum_{i=1}^{N} m_i \tag{1}$$

In the formula (1)，R is the number of nodes that the query request can be reached; m_i is the number of query request be retransmitted by node i; N is the number of nodes that can satisfy the query retransmission in the network. In Figure 1, from the formula (1), the initial node of query request is A, the overhead of a retransmission request is : 2 / 3≈0.6, as shown in Figure 2, the query overhead is 4 / 3 = 1.3.

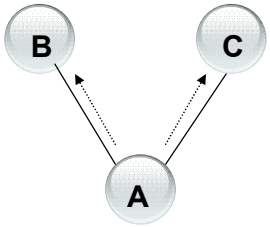

Fig.1 Query overhead: 0.6

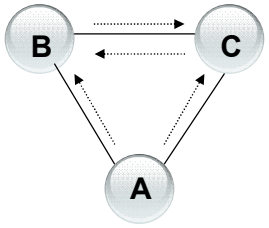

Fig.2 Query overhead: 1.3

4. If the number of nodes that can be satisfied is N, the number of nodes in the query process is R, then the coverage rate C is defined as:

$$C = \frac{R}{N} \tag{2}$$

If the transfer of the message is based on flooding [3], when a node receives a message retrieval for the first time, it will retransmit it to all the neighbor nodes except the source node. Suppose the average degree of node system is $\overline{d}$, by the formula (1),we can get the query overhead as:

$$\frac{1}{N}[1+\sum_{i=1}^{N}(d_i-1)] = \frac{1}{N}+\frac{1}{N}\sum_{i=1}^{N}(d_i-1) \approx \overline{d}-1 \tag{3}$$

In the formula (3), d_i represents the number of neighbor nodes.

Route Marking Strategy

Strategy Thought

In the process of distributed information retrieval, the retransmit node first adds node information in the query request message, when retransmitting the query request, it will check the packet header to make sure whether their neighbor nodes are in the routing label, if it exists, it indicates that it has sent a query request to the neighbor nodes,and it will no longer retransmit to this node, otherwise,it will retransmit the message to the neighbor node. If the routing mark of packet header is modified, the node will immediately notify the neighbors. The algorithm is described as follows:

S_V: represents a collection of target nodes after route marking;

N_V: The neighbor node set of node V;

Routing Label Algorithm(V, message)

if the node V_i receives the message firstly

/* Set neighbor nodes and retransmit information to the target */

{send the message to every neighbor node V_i

and $S_V = S_V \cup N_{V_i}$

}

While $(V_j \in N_{V_i}, j \neq i)$

/* The node which has received the query request will check the routing header, and retransmit information to the nodes which are not in the target node set. */

{ check neighbor $V_j \in S_V$ or not;

$if\,(V_x \in N_{V_j} \wedge (V_x\, is\, not\, in\, S_V))$

{send the message to V_x

and $S_V = S_V \cup N_{V_j}$}

send S_V to neighbor V_j;

} if(all neighbor. V_j is in S_V)

/* When a node's neighbor nodes are in the target set, end the retransmission */

Endwhile;

Endif;

End.

Every host and router in the network has a unique IP address, so we can use the IP information as the node information, and add it to the marked route message packet [5], but the length of packet will increase with the increase of target node set. In order to reduce the extra network overhead caused by packet length, we can use Bloom filter storage algorithm [6] to compress node routing information.

Comparison with flooding

In Figure 3, by the flooding retransmission mechanism, in the first round, node A retransmits queries to its neighbor nodes as follows: A→B, A→C, A→D; in the second round, node B,C,D, which has received the query request respectively retransmits messages to its neighbor nodes as follows: B→C, B→D, B→E, C→D, C→E; D→B, D→C, D→E; in the third round:E→C, E→D.In the whole process , the retransmission number of message is 14.

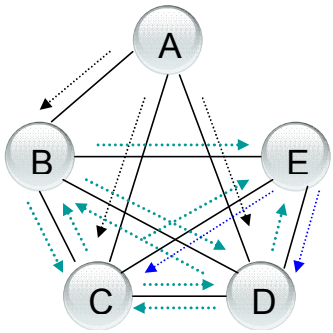

Fig. 3 Mechanism based on flooding retransmission

In Figure 4, based on the routing label retrieval strategy, In the first round, node A retransmits queries to its neighbor nodes as follows: A→B,A→C,A→D, node A and its target nodes form the set {A, B, C, D}, the set is added to the routing message; In the second round, node B, node C and node D first checks the packet marking, they only retransmit to node E,which has not received the queries. At the same time, node E will be added to the target set: B→E, C→E, D→E. Now, the receiving node set is{A, B, C, D,E}. Node E no longer retransmits query request to any node. In the whole process, the retransmission number of message is 6, with filtering redundant information effectivly.

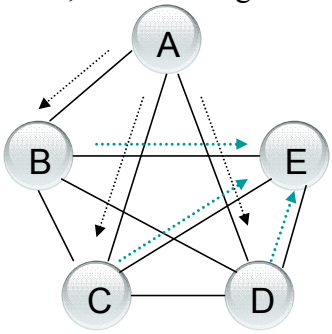

Fig.4 Retransmission mechanism based on route marking

Results and Discussion

If the number of retransmission query request nodes is N, the number of sides is E, the average degree of nodes is $\bar{d} = 2E/N$. Faloutsos has led in the concept of the degree index [7], he presented and verified that, for a route chart in the network , if the number of sides is E , the number of nodes is N , and the degree index is R, then:

$$E = \frac{1}{2(R+1)}(1-\frac{1}{N^{R+1}})N \tag{4}$$

While $\bar{d} = 2E/N$, substitutes the formula (4) for E ,then:

$$\bar{d} = \frac{1}{R+1}(1-\frac{1}{N^{R+1}}) \to \frac{1}{R+1}(N \to \infty) \tag{5}$$

In summary, there is no direct relationship between $\bar{d}$ and N.

The number of queries retransmited in flood mechanism is:

$$Number1 = \sum_{i=1}^{N}(d_i - 1) = N(\bar{d} - 1) \tag{6}$$

In the retrieval strategy based on route marking, every time, the new nodes be added to the target collection will be deleted in the next round, and the number of query requests in the searching period is:

$$Number2 = \sum_{i=1}^{N}(d_i - v_i) = N\bar{d} - \sum_{i=1}^{N} v_i \tag{7}$$

Under the same condition, which means the same number of nodes, the same topology, comparing the route marking strategy with the flooding strategy, for the former, the reduction number of transmitted packet is:

$$Number1 - Number2 = \sum_{i=1}^{N} v_i - N \tag{8}$$

In communication, there is a proportional relationship between the generation of redundant messages and nodes' degree d, from the formula (7), we can see, with the increasing network node connectivity, d becomes larger, and V_i will inevitably increase, in a network which N is fixed, the value of $\sum_{i=1}^{N} v_i - N$ increases. So the route marking strategy can effectively reduce the redundant messages, the performance is better than flooding strategy.

Summary

In the process of distributed information retrieval, in order to reduce the number of redundant messages, the paper adds node information in message, and uses a retrieval strategy based on routing label information. Through the analysis, this method can effectively reduce the redundancy forwarding overhead caused by flooding mechanism, and can save the network bandwidth. But at present, it is only from theoretical data that the effectiveness of the strategy is demonstrated.

Acknowledgement

The authors wish to thank the helpful comments and suggestions from my teachers and colleagues in NAAU. And also thank National University of Def ense Technology to provide software test datas.This work is supported by the study fund of HYJC at YanTai (No.HYJC2015024).

References

[1] Chen Duo-Long, Meng Xiang-ru, "A Constrained Network Flooding Algorithm Based on Negative Selection",Micro Electronics&Computer, 2013,vol.30,No.9, pp.53-57.

[2]DOU Wen,WANG Huai Min, " A Rumor Spreading Analog on Unstructured P2P Broadcast Mechanism", Journal of Computer Research and Development. 2004,vol.41,No.9,pp.1460-1465.

[3]LIU Hong, LIU Xi-yu, "A Heuristic Search Approach of Web Information",MINI-MICRO SYSTEMS, 2003,vol. 24,No.3, pp.427-429.

[4]ZHANG Long, LI Wei, "Distributed resource discovery model based on modified DHT", Application Research of Computers 2007,vol.24, No.12, pp.313-316.

[5]Matei Ripeanu, "Peer to Peer Architecture Case Study", Gnutella Network. University of Chicago Technical Report TR-2001-26, pp.223-231.

[6]XIE Kun, ZHANG Da-Fang. "A Trace Label Based Consistency Maintenance Algorithm in Unstructured P2P Systems", Journal of Software, 2007,vol.18, No.1, pp.105-116.

[7]Michael Mitzenmacher, Member, IEEE. Compressed Bloom Filters. IEEE/ACM TRANSACTIONS ON NETWORKING. 2002,vol.10,No.5, pp.110-121.

[8]FALOUTSOS M, FALOUTSOS P, EALOUTSOS C. "Power-law relationships of the internet topology. SIGCOMM ,1999, pp.251-262.