Research on the data distribution mechanism of MP2P based on the semantic routing

Li Lu

Defence Industry Secrecy Examination and Certification Center, Laboratory, China

kevin_psguy@163.com

Keywords: MP2P, semantic routing, data distribution

Abstract. The routing information is easy to fail as the disturbance of MP2P terminals. The channel would be occupied by the traditional flooding mode to ensure the success rate of information exchange. The interconnection of nodes computed by the semantic similarity, the choice of routing nodes and routing path is researched in the paper. Meanwhile, the relevant data is pushed to the potential users in order to improve the resource utilization in the transmission. The experiment shows that when the nodes move in a certain speed, if the data is scheduled by semantic routing, the recall ratio and query efficiency of information would be improved. The resource download success rate and bandwidth efficiency would be increased if the data is pushed to the chosen potential users, and the data distribution efficiency of MP2P network would be improved.

Introduction

MP2P (Mobile Peer-to-Peer) is the evolution product of P2P technology [1] and mobile network, inherits the advantage of resource load balance, no centralized network framework and network extensibility in the P2P network [2]. The MP2P network as the mobile Ad Hot network [3] is easy to build the information interaction platform. The temporary and flexible characteristics of MP2P can reduce personnel directly participation in the complex and dangerous environment such as battlefield and disaster area. Therefore, The MP2P without fixed base station network has practical value and broad market.

But MP2P network has the following problems with the resource location and transmission because of the wireless and multi-hop [4]: bad link, limit resource, low transmission efficiency.

In order to optimize the MP2P network routing strategy and network content distribution, the opportunistic routing strategy based on the semantic similarity is proposed. In the strategy, the potential users are discovered. The resource distribution and bandwidth usage efficiency is enhanced by the active and passive hybrid routing mechanism.

The MP2P data distribution mechanism based on the semantic routing strategy

The nodes that have the similar requirement are chosen as the relay nodes in the transmission. In this way, the relevance among the transmission nodes is strengthened and the potential users that have the interest in the similar resources are discovered.

MP2P network model

The MP2P network is divided into several clusters. In every cluster, the super node O is the center node and the length of radius is r. The nodes belong to different clusters are the opportunistic nodes that can exchange information among clusters. The formula of the cluster is shown as (1):

$$B(O,R)=\{i\in V \mid \rho(O,i)<(1+\frac{\Delta n_T}{N_T})\cdot r\} \tag{1}$$

The super node is the center of the sphere with the radius of R. The radius of the cluster changes periodically as the number of nodes in the same cluster. V is the node set, $\rho: V\times V\to R$, Δn_T is the number of nodes that changes in the period T. N_T is the number of nodes that is indexed in the period.

node information module

The nodes query the information by access to the information table. The information table structure of node is shown as table 1:

Table 1. information table structure of node

NodeID	Cluster	State	Interesting	Rarity

NodeID variable is the unique identification of node.

Cluster variable identifies the cluster that the node belongs to.

State variable is the state of the nodes. The node has three states in the network: super node that is in charge of the cluster, opportunistic node that transmits information among nodes and common node that contains the rest nodes.

The nodes stay more time in the network, then the nodes are more stable and valuable and the information is stored more. The residence time (T) is an important factor that is used to measure the node performance. The nodes of MP2P frequently join and leave the network and in this way the data distributed efficiency is low. The churn is related to the move speed (v) of nodes. The MP2P network is also affected by the computing power (C) and storage space (M). The combined weight of nodes is shown as formula (2): $W_n = \lambda_1 M + \lambda_2 C + \lambda_3 T + \lambda_4 v$ (2)

In the formula, λ is the factor of weight, and $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$

It is assumed that the threshold is β, the node that the weight is greater than the threshold is chosen as super node and the node information of the cluster is stored in the super node. The super node set is defined as formula (3): $P_w = \{(O_w, \beta) \mid W_n > \beta\}$ (3)

The interest preference is stored in the interesting variable. The history of mutual information among nodes is recorded as vector. The history information contains query records, the resource access frequency, the similarity between resource and node and the similarity between different nodes. So it is convenient to build the information transmission path by the nodes that are interested to the resource.

The scarcity of resources is indicated with the rarity variable. The entire resource is divided into multiple data blocks in the network. The resource information in the cluster is maintained by the super node. The scarcity information of data blocks is recorded in the super node too and the data blocks are ranked by priority according to the scarcity of resources. Therefore, the probability of getting rare data blocks is improved. The priority of resource is defined as follows:

$$Rarity \propto \frac{n}{s^f \cdot hop_i} \quad (4)$$

In the formula (4), hop is defined as hops of node i. The more hops the resource takes, the more possible the data blocks are reserved. On the contrary, the less hops the resource takes, the less possible the data block are reserved. In the formula, n is the query number, s is the storage space of data blocks and f is the calculating parameter.

routing message mechanism

(1) The recognition of opportunistic nodes

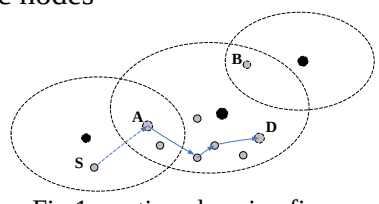


Fig.1. routing choosing figure

As the dynamic characteristic of MP2P network, the network is divided into several clusters that are managed separately. The information among different clusters is transmitted by the opportunistic nodes. The information is transmitted by the nodes in the overlap area in the opportunistic routing strategy. There are two requirements for the information transmitting among clusters. The first is that the intersection of different clusters comes out in the dynamic moving process. The second is that the nodes that are named opportunistic nodes exist in the intersection of different clusters. If the information does not exist in the cluster, the super node would query the information storage table

and wait for the appearance of the opportunistic nodes, then the query would go on in another cluster. Every common node has its own metric space. The other nodes in the metric space are considered as neighbor nodes and the information is transmitted by the query method of heartbeat. Meanwhile, the node will check whether its super node ID and its neighbor super node ID are the same. If they are different, it can be inferred that the node is opportunistic nodes that are in the cross area.

The opportunistic node set is defined in formula (5): $D_i = \{(U, S_n) | \sum S_n \neq 1, n = 1, 2, 3.....\}$ (5)

(U,Sn) is two-tuples. U is the cluster that node i belongs to last period. Sn is the number of super nodes that node i have queried. If the node could query more than one super node in a period, it is the opportunistic node.

Compare to the wired network, the invalid link and the bad link are more possible to appear in the wireless network. Therefore, the routing strategy should be refined based on the opportunistic nodes. The routing path is shown as figure 1, and the node A and node B are opportunistic nodes.

(2) Preference relationship based on semantic similarity

The query node calculates the similarity between the request resource and preference of neighbor nodes and then chooses the node that is most interested in the resource as the next hop. In this way, the routing is available.

The preference of nodes is related to the query history and resource access frequency. The tendency to the resource depends on the query history. The nodes that have high response probability to the transmitting resource are the potential customer and are valuable for resource push.

There are 4 steps to choose the routing nodes based on the weighted semantic similarity:

Step 1: feature words are represented as vector.

The preference of nodes include the query history $I_x(D)$ and keyword access frequency $I_k(D)$. The keyword vector is shown as formula (6): $I_x(D) = (x_1, x_2, x_3.....x_n)$ (6)

The preference to the keyword is affected by its access frequency in the node. The keyword frequency vector is show as formula (7): $I_k(D) = (k_1, k_2, k_3..... k_n)^{-1}$ (7)

and $K_n = check(n_i) / check(N)$

Check(N) is the sum of resource query number from node N; check(n_i) is the sum of query number for a certain resource (keyword) ni.

The preference to the resource is defined as Interesting(i): $Interesting(i) = I_x(D) \cdot I_k(D)$.

Step2: calculate the similarity. $sim(\mathrm{Re}sources, Interestng) = (C(R) \cdot C(I)) / (|C(R) \cdot C(I)|)$

In the formula, resources are the keywords of query. Interesting is the vector of node preference. C(R) is the keywords vector of resource. C(I) is the node preference vector. |C(R)|is the norm of C(R) and |C（I）| is the norm of C(I). If the cosine value of vector C(R) and C(I) is 1, it means that the preference of node matches exactly with the resource. If the cosine value of vector C(R) and C(I) is 0, it means that the resource is not needed by the node and the node would not be chosen as routing node.

The value interval of similarity is: 0<=sim(Resources,Interesting)<1.

By the similarity calculation, the node that has the highest semantic similarity is chosen as the routing node. The node that is interested in the resource is chosen as the next hop. In this way, the resource is pushed to the potential customers and the resource and bandwidth utilization level is improved.

Step3: calculate the node characteristic. In the step 2, the nodes that are interested in the resource are chosen by similarity calculating, but the routing cannot traverse all the nodes that are interested in the resource. In that way, the efficiency of distribution to the potential customers can be improved, but a large of network resource would be occupied if there are too many relay nodes and the transmission delay increases. The resource download success would be reduced because of the instability of nodes. The routing condition can also be affected by the node itself, therefore, the node condition should be considered into the selection condition based on the weighted similarity. The state attribution of node in the network can be shown in formula (8):

In the formula (8), $\alpha_{(i,j)}$ is the probability of node i and j in the same cluster, T is the period length, $\zeta(x_i, y_j)$ is the common online time of query node and routing nodes.

$$STATE(D_i) = \alpha_{(i,j)} \cdot \frac{\zeta(x_i, y_i)}{n \cdot T} \tag{8}$$

Step 4: The routing nodes are determined by their state attributes and preference to the resource, just as formula (9): $Q_{rout} = STATE(Di) \cdot Sim(\text{Resources,Interesting})$ (9)

The active distribution strategy with the information pushing can avoid plenty of duplication of data to fill the channel because of leaflets type transmission and reduce the waste of the network resource. The resource transmits through some nodes that are interested in it. In this way, the resource spreads in a limited area and the resource download rate is enhanced.

(3) The routing query algorithm based on the semantic similarity

The data transmission is divided into intra-cluster and inner-cluster by the cluster management in the MP2P network. The node information table is managed by the super node (Oi). The opportunistic nodes (node.oppo) that are in the crossing area of clusters can transmit query information and resource among multiple different clusters. When the node queries the resource, firstly, it would visit the super node in the same cluster to locate the resource. If the retrieval information of request resource is in the super node of the same cluster, the resource transmission would go on in the cluster. Otherwise, if the opportunistic nodes come out, the query information would spread.

The routing query algorithm 1 based on the semantic similarity:

```
1. Select()
2. {  select_routenode();
3.     Query(Ri)→node.route;
4.     If (node.route has query)
5.        Addquery();}
6. Query(Ri)→Oi  //visit the super node to query the relative nodes
7. If(Oi.content != null) && (Ri is alive)
8.     Select();
9. else if (TTL!=null)
10. {  If  num(node.oppo) == 1
11.         query(Ri)→node.oppo;
12.         Select();
13.    else if num(node.oppo) !=null
14.        Select();}
```

Routing overhead

The MP2P network strategy based on the cluster partition management is proposed by the paper. The resource is transmitted by the opportunistic nodes among different clusters. The overhead used to locate the resource in MP2P network is shown as formula: $\cos t = y_0 \cdot Spend_{sub} + (1 - y_0) \cdot Spend_{\text{inter}-sub}$ (10)

In the formula (10), the overhead of MP2P network can be sectioned into two parts: the overhead of the request for resource in the cluster and among the clusters. The overhead used in the cluster is shown as formula (11): $Spend_{subnet} = \min \sum_{i,o \in V} \sum_{k \in K} [d_{ik} \cdot z_{ok} \cdot x_{io} \cdot c_{io} + {}_{ik}\ x_{io}(1 - z_{ok})(c_{io} + c_{os})]$ (11)

The overhead of the request for resource used among cluster is shown as formula (12):

$$Spend_{subnet} = \min \sum_{i,o \in V} \sum_{k \in K} [d_{ik} \cdot z_{ok} \cdot x_{io} \cdot c_{io} + d_{ik} x_{io}(1 - z_{ok})(c_{io} + c_{os})] \tag{12}$$

In the formula (10)(11)(12), the meaning of variables are shown as follows:

d_{ik}: the probability of node i requests for resource k;

c_{ij}: the overhead of establishing connection link(i, j) between the node i and j;

p: the probability of nodes meet(opportunistic nodes);

$$x_{io} = \begin{cases} 1, \text{the node i is in the metric space} \\ \text{of cluster o;} \\ 0, \text{otherwise;} \end{cases} \qquad z_{ok} = \begin{cases} 1, \text{the resource k is in the cluster} \\ \text{head o;} \\ 0, \text{otherwise;} \end{cases}$$

$$y_o = \begin{cases} 1, \text{the resource k is in the cluster;} \\ 0, \text{otherwise;} \end{cases}$$

In the formula (11)(12), $d_{ik} \cdot z_{ok} \cdot x_{io} \cdot c_{io}$ represents the overhead that node i get the resource from node o in the cluster, $d_{ik} x_{io} (1 - z_{ok})(c_{io} + c_{os})$ represents the overhead that node i get the resource from node in the cluster by access to the resource routing table of super node, $d_{ik}(1 - z_{ok}) \cdot p \cdot (c_{io} + c_{oa} + c_{bs} + (1 - z_{s'k})c_{ss'})$ represents the overhead that node i get the resource from other cluster by opportunistic nodes.

The experiment results and analysis

The environment configuration

The Inet+Oversim+OMNeT++ is chosen as the simulation platform [5] in the paper. Table 3 shows the primary parameter configuration in the simulation platform.

Table 2. The primary parameter configuration of experiment

Parameter	Value
Topology	Cluster
MP2P size	3000m×3000m
Number of nodes N	2000
Vmax	20m/s
Vmin	0m/s
Block	10kb
Bandwidth	250kbs
Number of documents each peer	200
number of keywords each document	10
TTL	6

data analysis

The data transmission in the MP2P network is improved by the opportunistic routing and node preference strategy in the paper. Then the strategy would be measured from transmission delay, bandwidth utilization, resource recall ratio and resource query efficiency.

(1) Transmission delay

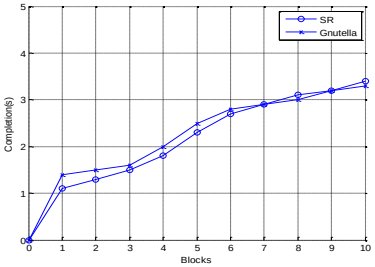
Gnutella that transmits information by flooding is adopted as the unstructured network framework in some mobile network. But it is easy to cause overflow and retransmission because the Gnutella transmits all the information to the neighbor nodes without choice..

The resource transmission is divided into three sections, the first is that the query is issued by the node, the second is that the query reaches the node that contains the resource, and the third is that the request resource is transmitted to the request node. The whole file is split into lots of data blocks, so it is needed that all the data blocks should be got together.

In the figure 3, when the nodes in the network is sufficient, as the number of blocks increases, the delay time of semantic routing (SR) and Gnutella would grow. When there are less than seven blocks, the delay time of semantic routing is less than Gnutella. When there are more than seven blocks, the semantic routing does not have more advantage than the Gnutella, even has more delay time.

(2) Bandwidth utilization

The load balance is one of the factors that are used to measure the bandwidth utilization. It is shown in figure 4, when the same file is queried for several times, the semantic routing has lower bandwidth occupation rate than the Gnutella. The semantic routing strategy has choice on the forwarding nodes when the query is transmitted. Therefore, when the target resource is sent back to the request node, it would pass through the nodes that have common preference. In this way, the utilization of resource is improved, the retransmission probability of resource is reduced and the bandwidth utilization is enhanced.



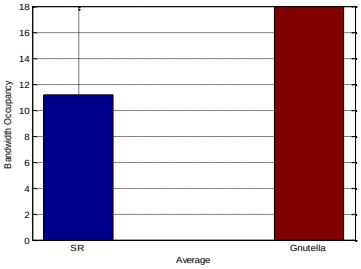


Fig.2. the comparison of network transmission delay

Fig.3. the bandwidth utilization comparison between semantic routing and Gnutella

(3) Recall rate

The recall rate is one of the factors that are used to measure the retrieval ability. From the experimental results, when the TTL is greater than six, the semantic routing and Gnutella can get high recall rate in the mobile network. As in a small network with high TTL, the query can traverse the entire network. The recall rate is shown as formula (13):

$$\mathrm{Re}call = \frac{\sum_{got} File}{\sum_{related} File} \tag{13}$$

(4) Query efficiency

In order to compare the query efficiency of semantic routing and Gnutella, the number of nodes and TTL in the network is fixed. As the coverage area of network increases, the query efficiency of semantic routing and Gnutella change. The query efficiency is shown as formula (14):

$$Check\_Efficiency = Right\_\mathrm{Re}call / Check\_Query \tag{14}$$

From the experiment, it is shown that the query efficiency of semantic routing and Gnutella is reduced as the coverage area of network increases. Meanwhile, the query efficiency of semantic routing is better than Gnutella in the same condition. As the information in the Gnutella is spread by broadcast, there are a lot of nodes involved in the query process. Nevertheless the semantic routing strategy transmits information by nodes that have preference on it. As the coverage area of network increase while the number of nodes remains unchanged as before, the semantic routing strategy would have little advantage, but Gnutella is little affected by the change. The experiment result is shown as figure 6.

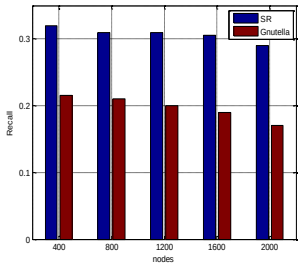


Fig.4. the recall rate comparison between semantic routing and Gnutella

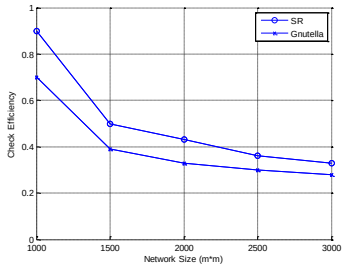


Fig5. The comparison of query efficiency of semantic routing and Gnutella in different network scale

Summary

The semantic routing strategy is proposed by the paper to calculate the relevancy degree as the routing of data transmission is effected by the dynamic characteristic of MP2P network. The routing is optimized according to the preference on transmission resource. In this way, the potential users would get the resource interested and the resource retransmission would be reduced. It is shown in the experiment that when the nodes move at a certain speed, the recall ratio, the query efficiency, the resource download success rate and the bandwidth utilization are improved by the strategy. Pushing data to potential users can strengthen the cooperation of nodes and the data distribution is improved.

References

[1] Xinxin Fan, Mingchu Li, Hui Zhao, et al. Peer cluster: a maximum flow-based trust mechanism in P2P file sharing networks[J]. Security and Communication Networks, 2013, 6(9):1126-1142.

[2] Zhuo Chen, Gang Feng, et al. Game Theoretical Bandwidth Request Allocation Policy in P2P Streaming Network[J]. Journal of Electronics and Information Technology. 2013, 35(7): 1725-1731.

[3] Dong Ping,Qian, Huanyan,Lan, Shaohua. Opportunistic multipath routing protocol in mobile Ad hoc networks[J]. Journal of Nanjing University of Science and Technology, 2013, 37(3):337-343.

[4] Chuanchi Lai, Chuanming Liu. Approaches for Data Synchronization on Mobile Peer-to-Peer Networks[J]. Advances in Intelligent Systems and Applications, 2013, 2(1):599-608.

[5] Privalov, A.Yu, Tsarev, A. Analysis and simulation of WAN traffic by self-similar traffic

model with OMNET, Wireless Communications and Mobile Computing Conference (IWCMC), 2014 International, 2014: 629-634.