Efficient Content Location Using Semantic Small World in Peer-to-Peer Networks

Yong Chen[1] , Wei-zhong Xiao[1], Huan-lin Liu[2], Long-zhao Sun[1]
[1]Key Laboratory of Industrial Internet of Things & Network Control, MOE
Chongqing University of Posts and Telecommunications, Chongqing, PR China 400065
[2]Key Laboratory of Optical Fiber Communication Technology, Chongqing, China 400065
e-mail: chauvetzhong@163.com, chenyong@cqupt.edu.cn

Abstrak

Penemuan konten pada jaringan peer-to-peer tak terstruktur adalah masalah yang menantang. Makalah ini menyajikan sebuah mekanisme baru untuk pencarian sumber daya dunia kecil semantik guna menyelesaikan masalah tersebut. Dengan menggunakan model ruang vektor untuk menghitung relevansi semantik dan menerapkan sifat dunia kecil seperti jarak hop rerata rendah dan koefisien pengelompokan tinggi untuk membangun gugus overlay. Dalam sistem dunia kecil semantik, mekanisme pencarian dibagi menjadi dua bagian, pencarian pada gugus dan gugus luar melalui inner-link dan short-link, sehingga dapat menghasilkan penelitian yang berbobot. Ini secara signifikan mengurangi panjang jalur rata-rata dan biaya permintaan. Hasil simulasi menunjukkan bahwa skema dunia kecil semantik yang diusulkan dapat mengungguli skema K-random walk dan skema flooding, dengan laju hit permintaan yang lebih tinggi dan latency permintaan yang lebih rendah.

Kata kunci: *peer-to-peer jaringan, semantik kecil dunia, kecil dunia, gugus*

Abstract

Locating content in unstructured peer-to-peer networks is a challenging problem. This paper presents a novel semantic small world resource search mechanism to address the problem. By using vector space model to compute the semantic relevance and applying small world properties such as low average hop distance and high clustering coefficient to construct a cluster overlay. In semantic small world system, the search mechanism is divided into two parts, searching at cluster and outside cluster through inner link and short link, so that it can achieve the incremental research. It significantly reduces the average path length and query cost. Meanwhile, the simulation results show that semantic small world scheme outperforms K-random walks and flooding scheme than higher query hit rate and lower query latency.

Keywords*: peer-to-peer network, semantic small world, semantic relevance, cluster*

1. Introduction

In recent years, there have been many Peer-to-Peer (P2P) systems deployed in the Internet. The nodes which participate in sharing are increasing quickly and present a massive trend. In essence, a P2P system can be characterized as a distributed network system in which each node has similar functionalities and plays the role of a server and a client at the same time[1]. It is important for users to find relevance resources in the large scale P2P network. Content searching becomes a key step and a complicated problem in such environment.

According to the nodes and resources organization and location method, P2P networks are divided into structured and unstructured. In a structured system, both the network topology and the distribution of keys are predetermined. They adopt DHT technology to indicate resource and queries, such as Chord [2] and Taperstry [3]. The advantage of such a structured system is high query efficiency. But DHT can’t indicate the resources and query semantics, thus it can only be an exact match. On the other hand, unstructured systems like Gnutellla are fault-tolerant and resilient to the fuzzy query. But the search mechanism used in unstructured systems, like flooding, is consuming too much bandwidth and not scalable.

Currently the unstructured peer-to-peer network is still the most widely used. Its basic search method is each node maintains a little neighbor information. Query messages are forwarded between nodes. Then match the query. According to whether there is prompt

Received December 29, 2012; Revised February 25, 2013; Accepted March 11, 2013

information in the search process, the unstructured P2P search can further be classified into blind search and informed search. The common characteristics of blind search are the query will be forwarded to some or all the neighbors without using any prompt information. The search won’t stop until the query meets the stop condition. Modified-BFS[4], Iterative Deepening[5] and random walk[6] are some typical algorithms. Informed routing is proposed aimed at improve search efficiency. The nodes in the network publish and collect resources shared information and use it to guide the forward. Intelligent-BFS and APS[7] mechanisms regard semantics of individual resources of the neighbors as neighbors semantics and as a basis for forwarding the query. [8] presents semantic overlay networks (SON) . Documents in each node are classified and a document hierarchy is spread throughout the P2P network. However, this mechanism needs broadcast or a central directory server to make node join the designated SON. And lookup takes a similar approach with the node join. This algorithm in [10] has a poor scalability and suffers single point failure problem. In this paper, we propose a semantic small world (SSW) mechanism. The search mechanism applies vector space model (VSM) to make semantically related nodes join a semantic cluster, and takes advantages of some specialties of small world to realize incremental search which the node can obtain new related resources when it repeats the search with the same query. According to the experimental results, the method can search out the vast majority of resources users need quickly, and has high efficiency.

2. Proposed SSW Scheme

2.1. The Basic Idea of SSW Algorithm

The search algorithm of SSW is a mix of semantic relevance and small world. The query messages are forwarded to new relevant node, and achieve incremental search.

(1) Apply VSM to compute semantic relevance, and then make the relevant nodes join a semantic cluster.
(2) Each node, on one hand, is connected to some semantically relevant neighboring nodes in semantic cluster, on the other hand, keeps a small number of links to some randomly chosen distant nodes. Links to relevant neighboring nodes are called Inner links while links to irrelevant distant nodes are called short links.
(3) Source node forwards the query messages to not only relevant nodes through inner links but also irrelevant nodes through short links. Thus realize the incremental research.

An example of SSW topology is illustrated in Figure 1.

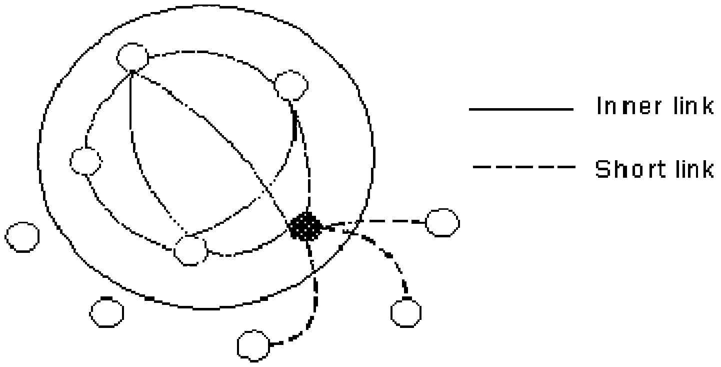

Figure 1. A Topology Graph of SSW

2.2. Related Concepts

SSW scheme aims at locating content efficiently. It contains two concepts. One is semantic cluster, the other is small world.

2.2.1. Semantic Cluster

In this paper, a cluster which is formed by some semantic relevant nodes is called semantic cluster. The semantic relevance has many kinds of computation models. Vector Space

Model (VSM)[9] is one of the most effective semantic representation models. The node vector is a centroid vector of all documents in a node. The steps we calculate node vectors are as follows. First, a term vector is derived to represent a document, in which each term's weight is allotted by its frequency in the document. Second, all temporary term vectors of X's documents are summed up, thus we gain a new vector in which each term element *T* has a weight w_t. The value is calculated by formula TF-IDF:

$$W_t = TF_t \times IDF_t \tag{1}$$

Where TF_t is frequency of term Tin document X. and IDF_t is inverse frequency. We re-weigh IDF_t using formula

$$W_t = TF_t \times \log(\frac{N}{n} + 0.01) \tag{2}$$

Where N is the number of documents on X. n is the numbers of feature terms.

Third, we normalize the weighed term vector to unit length. Given two nodes of documents (node X and node Y), their relevance score is the cosine similarity of their normalized node vectors listed below:

$$\mathrm{Re}levance(X,Y) = \sum_{t \in X,Y} W_{X,t} \cdot W_{Y,t} \tag{3}$$

The formula can be written as follows:

$$\mathrm{Re}levance(X,Y) = Cos(X,Y) = \frac{\sum_1^n W_{X,t} \times W_{Y,t}}{\sqrt{\sum_1^n W_{X,t}{}^2} \times \sqrt{\sum_1^n W_{Y,t}{}^2}} \tag{4}$$

In this formula, t is term existing in both node X vector and node Y vector. $W_{X,t}$ is the weight of term t in X, $W_{Y,t}$ is the weight of term t in Y. If the relevance score is less than a certain threshold, these two nodes are deemed to be irrelevant. Otherwise, they are considered relevant.

Nodes vectors are also used to calculate the relevance of a node X and a query Q according to (3), the following formula applies:

$$\mathrm{Re}levance(X,Q) = \sum_{t \in X,Q} W_{X,t} \cdot W_{Q,t} \tag{5}$$

2.2.2. Small World

Small world phenomenon is found by psychologists Milgram[10], also known as "six degrees of separation" theory. Milgram's experiment shows that the average distance between any two individuals in the social network is six. It has been observed that the small world phenomenon is pervasive in a wide range of settings such as social communities, biological environments, and communication networks. Recent studies have shown that peer-to-peer networks such as Freenet may exhibit small world properties. The average shortest distance between two randomly chosen nodes is approximately six hops. This property implies that one node can locate information stored at any random node of a small world network by a small number of links.

Small world network is a network between random network and regular network. A node in small world network maintains many connections to the close distance nodes and several connections to the long distance nodes. There are many attractive properties in small world such as low average path length and high clustering coefficient. In order to better explain the two specialties. We define undirected graph G as the connection graph of small world network, where v represent a node. The total number of nodes is N.

Definition.1. Clustering Coefficient (CC). Suppose the out degree of v is Kv, and the actual number of edges connected to the node is Ev. Because the most number of edges will be Kv=kv(kv-1)/2 in this graph, the clustering coefficient of node v is Cv=Ev/Kv .the clustering coefficient C is defined to be the average of all node clustering coefficient, as equation (4).

$$C=\frac{1}{N}\sum_{v=1}^{N}C_V \tag{6}$$

CC represents a network tightness, and it can be expressed as the closeness between people in a social network.

Definiton.2. Average Path Length (APL). Let D(i,j) denote the shortest path length between the node I and node j. APL can be defined as the average L of the shortest path length between all pairs of two nodes in G.

$$L(G)=\frac{1}{N(N-1)/2}\sum_{i\neq j}D(i,j) \tag{7}$$

The CC and APL of random network are both small, to the contrary, they are both large in regular network. However, in small world network, CC is large and APL is small. The two good properties help locate content quickly.

There is other good articles study the small world network. Manfredi [11] analyzes how the small world occurs in network communication. Inaltekin [12] compares the small world with the other technologies on the average path length through experimental test. [13-14] apply small world to peer-to-peer network. In [15], small world is applied to wireless P2P network. However, all the above methods which construct small world are based on physical distance. Our SSW mechanism considers users' indeed need by using semantic cluster as distance metric. Otherwise, our method studies small world on the top of unstructured P2P network instead of structured P2P.

2.3. The Construction and Adaptation Algorithm of SSW

The task is to organize relevant nodes into semantic cluster through inner links and to facilitate semantic cluster discovery by maintain some short links. Each node periodically performs neighbor discovery and adaptation to adjust its neighbor links thereby construct the small world overlay topology.

In neighbor discovery part, each node maintains two types of neighbors: random neighbors and semantic neighbors. Random neighbors are the semantically irrelevant nodes, whose relevance score is lower than REL_THRESHOLD, they are connected by random links, in this paper, we define random links as short links. Semantic neighbors are the semantically relevant node, whose relevance score is higher than or equal to REL_THRESHOLD. They are connected by inner links. The source node issues a random walk query message which contains the source node's node vector, a relevance threshold REL_THRESHOLD, the maximum number of responses MAX_RESPONSES, and TTL (time-to-live). When reaching a node, it will compute the relevance score between the node and the source node. Note that the relevant score is computed using formula (3) in section 2.2.1. Then the value of TTL minus one and forwards the query message randomly. The random walk returns a set of nodes until TTL equal to zero. According to their relevance scores, the returned nodes are added to the query source node's two neighbor candidate caches: random neighbor cache and semantic neighbor cache. The nodes with relevance score lower than REL_THRESHOLD will be added to random neighbor cache and those with relevance score higher than or equal to REL_THRESHOLD will be added to semantic neighbor cache, shown as figure 1.

In adaptation part, the purpose includes selecting new neighbors from the neighbor candidate caches and following existing neighbors. Each node periodically updates its semantic neighbors by choosing new neighbors from the semantic neighbors. At the same time, each node also periodically updates its random neighbors by selecting new neighbors from the random neighbor cache.

2.4. Search Procedure

Given a query, first, we search the local peer. If there are enough resources meet the MAX_RESPONSES which users need, the procedure terminates. Otherwise turn into the next step. In the second stage, query messages will be forwarded through inner links and short links separately. It means the query, on one side, is forwarded to semantic relevant neighbor in cluster. On the other side, is forwarded to irrelevant neighbor outside the cluster. In cluster search, we use k-Iteration Deeping mechanism. At the same time, through short links, the query adopts random walks. The search won't stop until MAX_RESPONSES=0 or TTL=0.

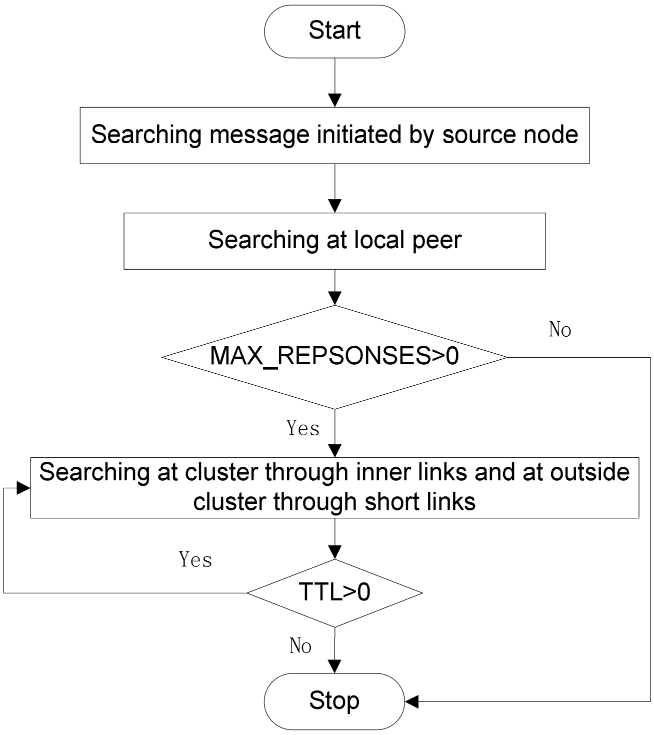
Start
Searching message initiated by source node
Searching at local peer
MAX_REPSONSES>0
No
Yes
Searching at cluster through inner links and at outside cluster through short links
Yes
TTL>0
No
Stop

Figure 2. The process of resource search

3. Simulation and Results

This paper's experiment is on top of network simulation platform PeerSim, which is an open source network simulation platform written in Java. In order to judge which scheme is better for the P2P search, we take two realistic meaningful evaluation criteria: average path length and query cost.

Compare to flood algorithm and K-random walks.(K=2), we can see the high performance of SSW. Experiment configuration is as follows: network size is 1000 nodes; the maximum inner links are 100; the network topology is a random network; the REL_THRESHOLD=0.8; the iteration depth k is set 4 and we set TTL=k; the MAX_RESPONSES we set 100.

Figure 3 shows the results of the average path length of the three algorithms. Flooding algorithm maintains at about 24 hops, while the K-random walks algorithm maintains at about 16 or so., SSW algorithm finally maintains at about 11 hops. Contrast to flooding, K-random and SSW have obvious improvement; and compared with K-random, SSW has better improved. Figure 4 shows the comparison of query cost. Query cost can be defined as the number of visited nodes. According to make the semantic relevant nodes join a cluster, when the number of returned resources is equal; SSW algorithm visits the minimum nodes. Along with the number of returned resources is increasing, the gap becomes bigger. The result shows that SSW mechanism has lower query cost comparing to flooding and K-random walks.

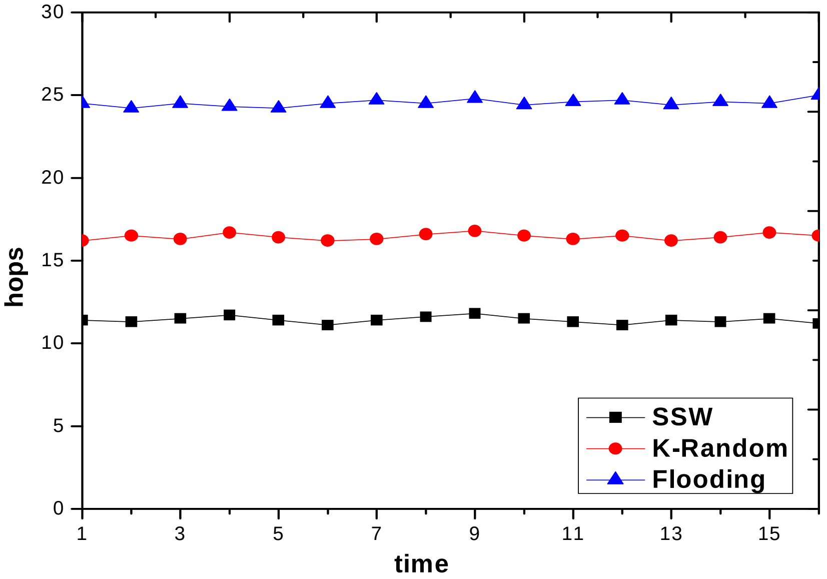

Figure 3. average path length

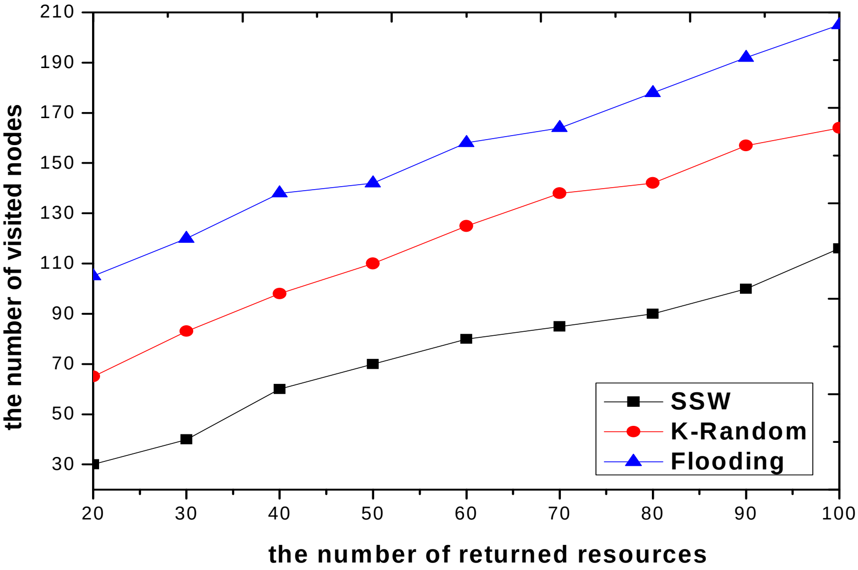

Figure 4. Query cost

Figure 5 illustrates the results of query hit rate of the three algorithms. Because query messages forwards to relevance nodes through inner links, along time is increasing, the hit rate of SSW is improving, and is better than the other methods. Figure 6 shows the results of query latency. SSW algorithm finally maintains at about 1, K-random walks appears small amplitude fluctuations, and maintains at 6.4 or so. The latency of flooding algorithm is greatest. These results show that SSW is better.

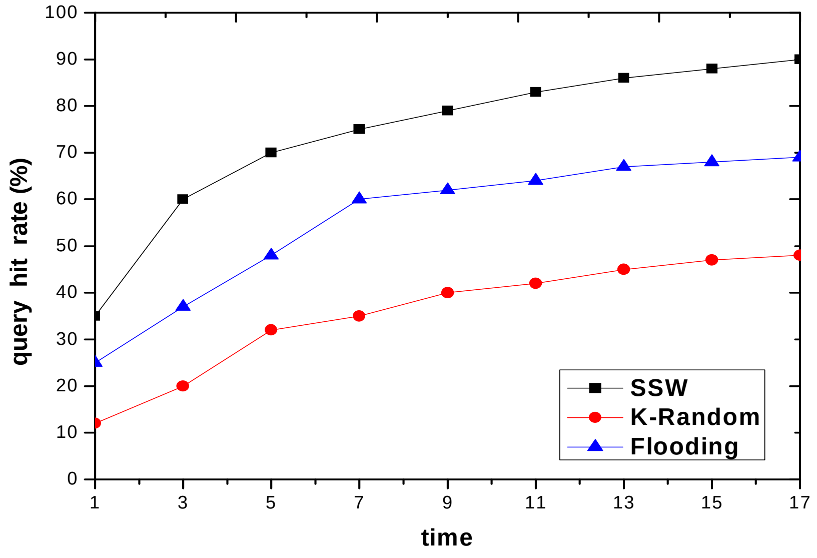

Figure 5. query hit rate

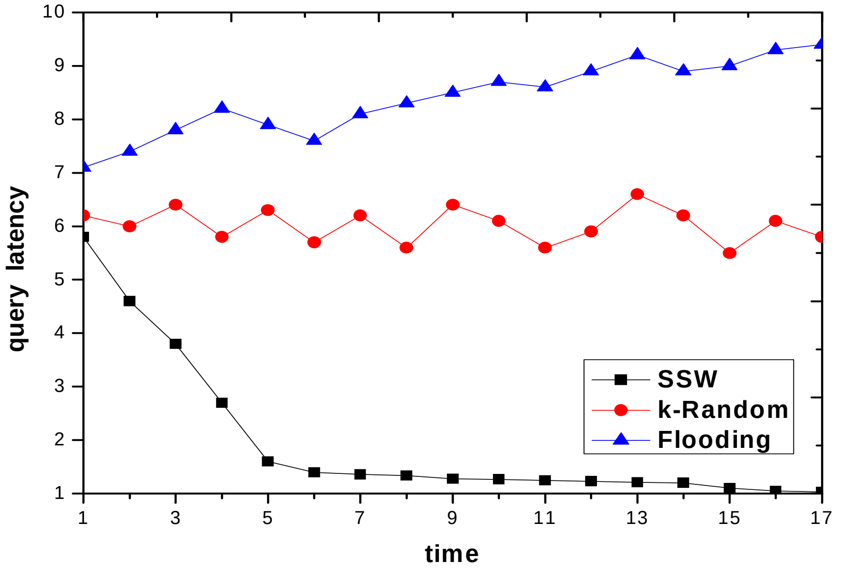

Figure 6. Query latency

4. Conclusion

Using the users' semantic, and following the small world rule, we propose a semantic-based small world search mechanism. The SSW algorithm applies small world properties to peer-to-peer network, and applies VSM to compute semantic relevance. During the search, query messages are forwarded through not only inner cluster links but also short links, thus guarantee the node get more new relevant resources. The simulation shows SSW method has smaller average path length, lower search cost, higher hit rate and lower latency than flooding and K-random walks. In the future, we will consider the heterogeneous topology, which the nodes in the networks have different capacities.

References

[1] Xuemin Shen, Heather Yu, John Buford Mursalin Akon. Handbook of Peer to Peer Networking. Springer. New York. 2010.

[2] Stocia I, Morris R, Karger D R, et al. Chord: A scalable peer-to-peer lookup service for Internet applications. *IEEE/ACM Transactions on Networking*, 2003; 11(1): 17-32.

[3] Zhao B Y, Kubiatowicz J D, Joseph A D. Tapestry: A fault-tolerant wide-area application infrastructure. *Computer Communication Review*. 2002; 32(1): 81.

[4] Kalogeraki V, Gunopulos D, Zeinalipour-Yazti D. *A local search mechanism for peer-to-peer networks*. Proc of the 11th Int Conf on Information and Knowledge Management. New Yord. 2002: 300-307.

[5] Yang B, Garcia Molina H. *Improving search in peer-to-peer networks*. Proc of the 22nd Int Conf on Distributed Computing Systems. IEEE Computer Society. Washington. 2002: 5-14.

[6] Lv Q, Cao P, Cohen E, et al. *Search and replication in unstructured peer-to-peer networks*. Proc of the 16th Int Conf on Supercomputing. New York. 2002: 84-95.

[7] Tsoumakos D, Roussopoulos N. *Adaptive probabilistic search for peer-to-peer Networks*. Proc of the 3rd IEEE Int Conf on P2P computing. IEEE Computer Society. Washington. 2003: 102-109.

[8] Crespo A, Garcia-Molina H. *Semantic overlay networks for P2P system*. Proc of the 3rd Int Workshop on Agents and Peer-to-Peer Computing. Springer. Berlin. 2005:1-14.

[9] M. W. Berry, Z. Drmac, E. R. Jessup. *Matrices, vector spaces, and information retrieval*. SIAM Review. 1999; 41(2): 335-362.

[10] S. Milgram. The small world problem. *Psychology Today*. 1967; 2: 60-67.

[11] Manfredi, S. di Bernardo, M. Garofalo F. *Small world effects in networks: an enginerring interpretation*. Proceedings of the 2004 International Symposium Circuits and Systems (ISCAS'04). 2004; 4:23-26.

[12] Inaltekin H, Mung Chiang, Poor H.V. Average Message Delivery Time for Small-world Networks in the Continuum Limit. *IEEE Transactions on Information Theory*. 2010; 56(9): 4447-4470.

[13] Bader, D.A, Madduri K. SNAP. *Small-world Network Analysis and Partitioning: An open-source parallel graph framework for the exploration of large-scale network*s. IEEE International Symposium Parallel and Distributed Processing. 2008: 1-12.

[14] Yan Ma, Bin Gong, Lida Zou. *Resource Discovery Algorithm Based on Small-world Cluster in Hierachical Grid Computing Environment*. GCC 7th International Conference on Grid and Cooperative Computing. 2008: 110-116.

[15] ChangJie Jiang, Chien Chen, JeWei Chang, RongHong Jan, Tsun Chieh Chiang. *Construct Small Worlds in Wireless Networks Using Data Mules*. IEEE 8th Int Conf on Trustworthy Computing of Sensor Networks. 2008: 28-35.