International Conference on Computer Science and Service System (CSSS 2014)

INS: A Novel P2P Semantic Search Approach Using Interest Attenuation Policy

Shuangyuan Yu
School of Computer and Information Technology
Beijing Jiaotong University
Beijing, China
shyyu@bjtu.edu.cn

Zhuoya Ju
School of Computer and Information Technology
Beijing Jiaotong University
Beijing, China
zyju@bjtu.edu.cn

Abstract—Semantic overlay could improve query performance in peer-to-peer (P2P) systems. When peers join or leave frequently, it will lead to network traffic surge, since most semantic search methods maintaining a large number of routing tables. In this paper, we address these problems by proposing Interest Attenuation Search(INS), a novel efficient peer-to-peer semantic search approach based on interest attenuation policy. In INS, the interest attenuation policy is introduced to help peers decide whether to forward messages. Before peer floods the request in semantic overlay network(SON), it will check the history information about the request message, then uses INS to make forwarding decision. Simulation results show that INS significantly improves query performance and reduces the traffic overhead generated by unstable network environment.

Keywords-P2P; semantic search; interest attenuationpolicy; probability

I. INTRODUCTION

P2P systems are gaining popularity quickly due to their scalability, fault-tolerance, distributed control, and self-organizing nature, raising hope for building large-scale information retrieval (IR) systems at low cost[1].

Semantic-based search is a challenging problem in P2P systems.With the advent of the semantic web[2] and semantic web services[3], many researchers and industry developers use these protocols in P2P networks[4]. Reference[5] introduces the definition of Semantic Overlay Network(SON), in which semantically related peers form a SON, and queries are routed to the appropriate SONs, increasingthe chances that matching files will befound quickly, and reducing the search load on peers that have unrelated content.

As social information explosion, there still exists traffic problem in P2P semantic overlay. When the churn rate and routing table update frequency are high (for example, peers join or leave/die frequently), performances of semantic search deteriorate. Heavy network traffic will limit the scalability of P2P semantic overlay, thus we still need to work hard on reducing traffic in P2P semantic overlay.

A number of P2P semantic routing techniques[4-6] havealready been proposed in recent years, but a lot of them still use route table to fulfil the query request.

In this paper, we proposeanapproach based on interest attenuation policy to generate messageforwardingprobability, not just flooding everything, to reduce the traffic. In this way, when peers receive a query request, they will check out how many times this request send to them, when the request generated, where the request came from and other factors. Then peers will use the formulato calculate the probability of forwarding this request and decide whether to forward it.

The interest attenuation policy comes from epidemic algorithms. Reference [7] introduces epidemic algorithms for replicated database maintenance， and its study indicates that interest attenuation policy suits for information dissemination in P2P network. In this paper, we introduce interest attenuation policy to P2P semantic overly, proposal a probability formula for peers forwarding message. Evaluations demonstrate that INS is efficiency in reducing network traffic and enhancing information retrieval.

II. OVERVIEW

A. Concepts

Semantic search is an application of the Semantic Web to search. Semantic Search attempts to augment and improve traditional search results (based on Information Retrieval technology) by usingthe Semantic Webdata[8].

The process of building and using SONsis depicted in Figure 1[5].

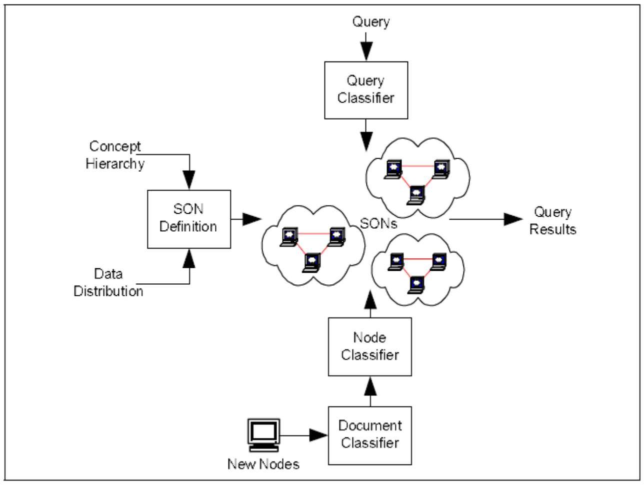

Figure1. Generating Semantic Overlay Networks

Steps of building SONs:

*1)*Classify hierarchies using the actual data distributions in the peers. Thishierarchy will be stored by all of the peers in thesystem and used to define the SONs.

*2)*When newpeer joins the system, itfloods the network with requests for thehierarchy in a Gnutella fashion.

a) Peer runs a documentclassifier based on the hierarchy obtained on all itsdocuments.

b) Peer classifier assigns the peer to specificSONs.

c) The peer joins each SONby finding peers that belong to those SONs. This can bedone again in a Gnutella fashion (flooding the network until peers in that SON are found).

Steps of query in SONs:

1) Classify request that peer sends out.

2) Peersends requests to the appropriate SONs using Gnutella flooding.

3) Peerswithin the SON findmatched resources by using Gnutella flooding.

According to the above introduction of building and using SONs, we can figure out that Gnutella flooding is used many times in the SONs. Thus when peers join or leave/die frequently, the system will generate a tremendoustraffic.

In order to overcome this problem, we introduce interest attenuation policy into SONs, in whichpeers will lose interestin forwarding when they receive same requests too many times.

B. Design

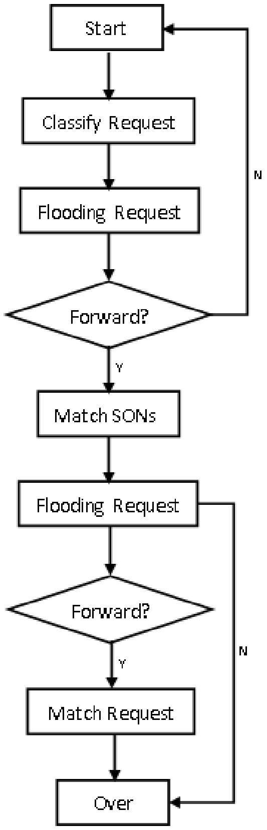

Figure2. Interest attenuation policy in SONs

Rumor mongering[7]: sites are initially “ignorant”; when a sitereceives a new update it becomes a “hot rumor”; while asite holds a hotrumor, it periodically chooses another site atrandom and ensures that the other site has seen the update; when a site has tried to share a hot rumor with too many sites that have already seen it, the site stops treating the rumor as hot and retains the update without propagating it further.

Rumor mongeringmechanism exists in P2P SONs, thus we introduce this concept to SONs.

As is shown in Figure 2, we add “forward decision”modules to SONs,then peers will use the forwarding probability formula to decide whether forwardingthe request or not. The flow of forwarding requestsis as follows:

1) Peers classify requestsbefore sending them out.

2) Peersuse the forwarding probability formula to calculate the forwarding probability.

When the first time peers receive a request, they will forward it to the matched SONs.

If peers receive the same request too many times, they will lose the forwarding interest.

3) Peers utilize Gnutella floodingto send requests to the appropriate SONs.

4) Peers use the forwarding probability formula to calculate the forwarding probability beforethey flood it in the matched SON.

C. Interest Attenuation Policy

Rumor spreading is based on the following scenario[7]: There are n individuals, which are initially inactive (susceptible). We plant a rumor with one person who becomes active (infective), phoning other people at random and sharing the rumor. Every person who hears the rumor also becomes active and likewise shares the rumor. When an active individual makes an unnecessary phone call (the recipient already hears the rumor), the active individual loses interest in sharing the rumor (the individual becomes removed)with probability 1/k.

In line with the epidemiology literature, rumor spreading can be modeled deterministically with a pair of differential equations.We let s, i and r represent the fraction of individuals susceptible, infective and removed respectively, so that $s + i + r = 1$:

$$\frac{\mathrm{ds}}{\mathrm{dt}} = -\mathrm{si} \tag{1}$$

$$\frac{\mathrm{ds}}{\mathrm{dt}} = +\mathrm{si} - \frac{1}{\mathrm{k}}(1 - \mathrm{s})\mathrm{i} \tag{2}$$

The first equation suggests that susceptibles will be infected according to the productsi. The secondequation has an additional term for loss due to individuals making unnecessary phone calls. According to [7] we can solve for *i* asa function of *s* and get a solution:

$$\mathrm{i(s)} = \frac{\mathrm{k+1}}{\mathrm{k}}(1 - \mathrm{s}) + \frac{1}{\mathrm{k}}\log \mathrm{s} \tag{3}$$

We apply this equation into P2P SONs. When $i(s) = 0$we can get the next formula:if an active individual receives same rumor, it loses interest in sharing it with the probability$\frac{1}{k}$[7].

$$\frac{1}{\mathrm{k}} = 1 + \frac{\log \mathrm{s}}{1-\mathrm{s}-\log \mathrm{s}} \qquad (4)$$

Based on the above formulas，we give the message forwarding probability formula as follows:

$$\mathrm{P(n)} = (1-\mathrm{p})^{\mathrm{n}-1}\frac{\mathrm{TTL}}{\mathrm{TTL}_{\max}}\cos\left(\frac{\Delta(\mathrm{t})}{\Delta(\mathrm{t})_{\max}}\frac{\pi}{2}\right) \qquad (5)$$

We denote$p = \frac{1}{k}$, and define $\mathrm{P}(0) = 0$.This means that if a peer never receives a message, it will not forward the message.

n: Peers receive the same message n times.

$(1-p)^{n-1}$indicates that ifa peer receives same message too many times, it will lose the interest to forward the message.

TTL: Time to live (TTL) is a mechanism that limits the lifespan or lifetime of data in a network. When$TTL = 0$, peers will stop forwarding message.

$TTL_{\max}$: The maximum TTL in current P2P SONs.

$\frac{TTL}{TTL_{max}}$ indicates that the TTL gets bigger, the forwarding probability will be higher.

$$\Delta(\mathrm{t}) = \mathrm{t} - \mathrm{TimeStamp} \qquad (6)$$

$t$presents current time.

$TimeStamp$: A timestamp is a sequence of characters when a certain event occurred, usually giving accurate to a small fraction of a second.

$\Delta(t)_{max}$ presents the maximum time delay. If$\Delta(t) = \Delta(t)_{max}$, then $P(n) = 0$.This indicates that if delay exceeds users-endurance, peers won’t forward the message.

$\cos(\frac{\Delta(t)}{\Delta(t)_{max}}\frac{\pi}{2})$indicates that as time goes on, the message forwarding probability gets down. This is consistent with the actual situationwhere people are not interested in old news.

From the above analysis we can infer that the forwardingprobability formula is in line with the interest attenuation policy.

Algorithm is as follows:

(1) Search initiator*I* sends message *M* to its neighbors，then it turns to be silent status *S*.

(2)To any remaining peer*v*，when it receives messageM, it will usethe forwarding probability formula to decide whether to forward it. If*v*decides to send *M* to its neighbors, it will turn statusfrom *S*to transfer status *T*, then forwards M and turns tostatus S again; otherwise *v* will keep silent status *S*.

(3)A silent status peer will keep silent if it doesn’t receive any message.

(4)When there isn’t any transfer status *T*, the algorithm ends.

In this algorithm, the search initiator*I*is equivalent tothe infective in rumor mongering; peers forwarding messages is equivalent to individuals susceptible in rumor mongering.

III. IMPLANTATION & EVALUATION

To evaluate the effectiveness of interest attenuation policy, we first generatenetwork topologies. Based on generated networks, wesimulate P2P flooding search, peers joining/leaving behavior.

A. Simulation Setup

Two types of topologies, physical topology and logicaltopology, are generated in our simulation. Physical topology presents the shape of the cabling layout used to link devices;logical topology is the way that the signals act on the network media, or the way that the data passes through the network from one device to the next without regard to the physical interconnection of the devices.

Our logical topology represents the P2P overlaytopology built on top of the physical topology.

GT-ITM [9] is a collection of routines to generate and analyze graphs using a wide variety of models for network topology. The graphs are generated in Don Knuth's SGB format. Using GT-ITM, wegenerate 3 physical topologies each with 5,000 nodes (peers).

GnutellaSim [10] is a scalable packet-level Gnutella simulator that enables the complete evaluation of the Gnutella system with a detailed network model. GnutellaSim is based on a framework for packet-level peer-to-peer system simulation; the framework is designed to be extensible to incorporate different implementation alternatives for a specific peer-to-peer system and is portable to different network simulators.

Using GnutellaSim, wegenerate the logical topologies with the number of nodes ranging from 1,000 to 3,000. The averagenumber of neighbors of each node ranges from 4 to 10. We attach leaf physical nodes to the stub router nodes generated by GT-ITM, and only run Gnutella peers on those leaf nodes.

B. Performance Metrics

The simulation parameters and their default values aregiven in Table 3. Our work focuses onQuery Throughput and Receive Rate.

TABLE I. EVALUATE INDICATORS

Item	Description
Query Success Rate	the probability for a Query to succeed
QueryThroughput	number of Queries (including copies) delivered to peers
Query Response Time	the time it takes for the first QueryHit to get to the Query initiator
Available Peers	number of online peers
Connectivity	degree of connectivity
Message Losses	message losses in Gnutella happens during the forwarding process on peers
Receive Rate	number of messages received per second while online

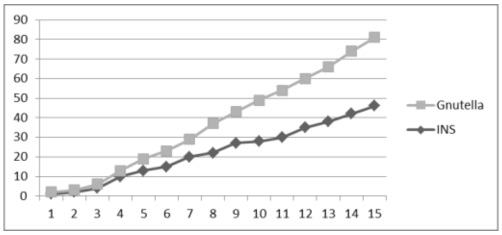

Figure 3.Number of Queries (including copies) delivered to peers

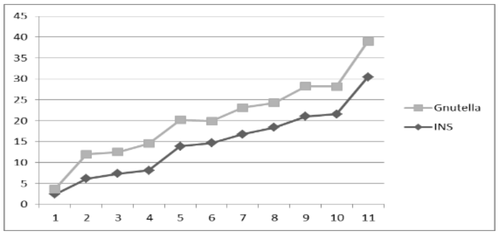

Figure 4.Number of Queries received

C. *Performance Evaluation*

Lots of simulations have been performed to evaluate the performance of INS.For comparisons, we simulated our searchapproach in conjunction with a randomlyconnected Gnutella approach which is simplicity and prevalence, a widely used benchmark approach for many researches.

The goal of the interest attenuation policy is to reduce traffic cost asmuch as possible while retaining the Query Success Rate.Oursimulation results are consistent on overlay networks of 1,000 nodes, 2,000 nodes, and 3,000 nodes.

Figure3 compares the traffic cost incurred by the originalGnutella-like system and by the system after adding “forward decision” modules.The number ofqueries delivered to peers increases as time goes on, but the INS’growth rate is lower. It means INS cuts down traffic overhead.

Figure4 compares the “Receive Rate”between two models. It indicates when network sizeincreases, the received queriesimprove and INS’sarealmost as well as Gnutella.

IV. RELATED WORK

Many recent semantic search techniques relate to ourresearch. We list only some ofthe most relevant ones due to space limitation:

Reference[11]uses constrained flooding routingalgorithm to reduce traffic in unstructured P2P system.

Reference [4]prunes search trees’ branches which have no chance to proceed to a response, in orderto limit the size of the indexes.

Reference[12]discusses lexical-based ontology to provide foundation for indexing in structuredP2P system.

Reference[13] proposes an ontology-based scheme which could measureinterestsimilarity between peers. Thus peer floodsqueries either within local peer groups or towards remote groups sharing similar interest.

Reference [14]introduces small-world characteristic to structure P2P network to decrease the maintenance cost.

V. CONCLUSION

In this paper we presented an efficient model for P2P semantic search. We introduce interest attenuation policyinto SONs and add “forward decision” modules to limit network traffic.

This system has been evaluated by a group of simulations, which show that using the proposed INS willreduce overhead significantly. We will scale up semanticinformation to further enhance information retrieval success rate.

REFERENCES

[1] C. Tang, Z. Xu, and S. Dwarkadas, "Peer-to-peer information retrieval using self-organizing semantic overlay networks," in Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications. ACM. 2003, pp. 175-186.

[2] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," Scientific american, vol. 284, 2001: pp. 28-37.

[3] S. A. McIlraith, T. C. Son, and H. Zeng, "Semantic web services," Intelligent Systems, IEEE, vol. 16, 2001: pp. 46-53.

[4] H. Rostami, J. Habibi, and E. Livani, "Semantic routing of search queries in P2P networks," Journal of Parallel and Distributed Computing, vol. 68, 2008: pp. 1590-1602.

[5] A. Crespo and H. Garcia-Molina, "Semantic overlay networks for p2p systems," in Agents and Peer-to-Peer Computing. 2005, Springer. p. 1-13.

[6] L. Juan and V. Son, "OntSum: A Semantic Query Routing Scheme in P2P Networks Based on Concise Ontology Indexing," in Advanced Information Networking and Applications, 2007. AINA '07. 21st International Conference on. 21-23 May 2007. 2007, pp. 94-101, doi: 10.1109/AINA.2007.104.

[7] A. Demers, D. Greene, C. Hauser, W. Irish, J. Larson, S. Shenker, H. Sturgis, D. Swinehart, and D. Terry, "Epidemic algorithms for replicated database maintenance," in Proceedings of the sixth annual ACM Symposium on Principles of distributed computing. ACM. 1987, pp. 1-12.

[8] R. Guha, R. McCool, and E. Miller, "Semantic search," in Proceedings of the 12th international conference on World Wide Web. ACM. 2003, pp. 700-709.

[9] http://www.cc.gatech.edu/fac/Ellen.Zegura/graphs.html

[10]http://www.cc.gatech.edu/computing/compass/gnutella

[11]S. Vuong and J. Li, "Efa: an efficient content routing algorithm in large peer-to-peer overlay networks," in Peer-to-Peer Computing, 2003.(P2P 2003). Proceedings. Third International Conference on. IEEE. 2003, pp. 216-217.

[12]C. Sangpachatanaruk and T. Znati, "Semantic driven hashing (sdh): an ontology-based search scheme for the semantic aware network (sa net)," in Peer-to-Peer Computing, 2004. Proceedings. Proceedings. Fourth International Conference on. IEEE. 2004, pp. 270-271.

[13]Q. Wang, R. Li, L. Chen, J. Lian, and M. T. Özsu, "Speed up semantic search in P2P networks," in Proceedings of the 17th ACM conference on Information and knowledge management. ACM. 2008, pp. 1341-1342.

[14]W. Xianghui and G. Guochang, "Low maintenance cost small-world P2P network," in Electronic and Mechanical Engineering and Information Technology (EMEIT), 2011 International Conference on. IEEE. 2011, pp. 3618-3621.