ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Special Issue , August 2019

International Conference on Recent Advances in Science, Engineering, Technology and Management at Sree Vahini Institute of Science and Technology-Tiruvuru, Krishna Dist, A.P

Content Search Method in Peer-To-Peer Networks

Ch. Ravindra Reddy, Sk. Saida

Assistant Professor, Department Of Computer Science&Engineering, Sree Vahini Institute Of Science & Technology , Tiruvuru, A.P, India

Assistant Professor, Department Of Computer Science&Engineering, Sree Vahini Institute Of Science & Technology , Tiruvuru, A.P, India.

ABSTRACT: Gnutella, A Well-Known P2P System, Uses Resources Inefficiently When Directly Applied To Information Retrieval Problems. In This Paper We Propose An Efficient Search Mechanism That Extends The Standard Gnutella Protocol To Support Content-Based Retrieval In P2P Networks. The Idea Is To Estimate Locally The Relevance Of Peers When They Receive Query Messages. Only Those Peers Esti- Mated As Relevant Will Retrieve The Query And Send Response Messages Back To The Source. Based On A Large Real Testbed Evaluation, We Show That Our Method Improves The Tradeoff Be- Tween The Quality Of Retrieval And Resources Consumed While Preserving Most Advantages Of Standard Gnutella

KEY WORDS: Search, Content-Based, Peer-To-Peer, Retrieval, Computer Networks

I. INTRODUCTION

Peer-To-Peer (P2P) Networks Are A Powerful Architecture For Sharing Computing Resources And Data. In The Strictest Definition, Each Peer Has The Functionality Of Both Server And Client, And Accordingly, Can Both Provide And Request Information. The Decentralized Nature Of P2P Systems Can Be An Advantage Over Client-Server Architectures. First, They Tend To Be More Fault-Tolerant As There Is No Single Point-Of- Failure. Second, Processing, Network Traffic, And Data Stor- Age Can Be Balanced Over All Peers, Which Enables The Network To Scale Well With The Number Of Peers. These Ad- Vantages Come With The Cost Of Requiring A More

Because They May Support A Large Number Of Peers And Have No Central Index, Efficient Search In A P2P System Can Be A Challenge. In General, Peers Can Locate Resources Or Content By Propagating Queries Through The Network And Then Waiting For Results From Peers With Relevant Results. Many Specific Strategies Have Been Proposed. The Simplest Approach Is Taken By The Gnutella [1] Protocol, Which Broadcasts Query Messages To Each Neighbor, Hop-By-Hop Across The Network Within Some Distance From The Source. Although It Is Not Efficient In Terms Of Network Bandwidth, This Technique Is Simple, Robust, And Has A Minimum Re- Quirement On Cooperation And Consistency Among Peers. For Example, It Allows Arbitrary Network Topologies, And Each Peer Stores No Information Regarding The State Of Oth- Ers.

Some Techniques, Such As SETS [3], Try To Reorganize The Topology Of Network So That Topic-Related Peers Are Close To Each Other. By Taking Advantage Of A Rigid Topology, Net- Work Traffic Can Be Reduced. In Other Techniques, Such As The Localized Search Mechanism Proposed By Kalogeraki, *Et Al* [4], Each Node Maintains An Index Or A Profile Of Its Neighbors' Content That Is Used Rank Its Neighbors. Search Is Then Restricted To What Are Believed To Neighbors With Rele- Vant Results. The Cost Of Those Approaches, Like In Dhts Is Increased Coordination Among Peers, Which Must Be Sustained As Peers Join And Leave, And Change Their Content. Moreover, Each Peer Has An Increase Storage Cost To Participate In The System. We Evaluated Our Approach On A Large Testbed With Thou- Sands Of Peers. We Found That The Tradeoff Between The Quality Of Retrieval And Resources Consumed Is Greatly Im- Proved While Most Advantages Of Standard Gnutella Are Preserved. For Example, According To Our Experiments, Gnutella Consumes More Than Three Times The Network Bandwidth Required By Our Approach For The Same Level Of Recall. The Rest Of The Paper Is Organized As Follows: Section 2

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Special Issue , August 2019

International Conference on Recent Advances in Science, Engineering, Technology and Management at Sree Vahini Institute of Science and Technology-Tiruvuru, Krishna Dist, A.P

Describes Related Work. Section 3 Introduces Our Resource- Efficient Algorithm For Content-Based Search In P2P Net- Works. We Present Our Testbed And Evaluation Methodologies In Sections 4 And 5. Experimental Results Are Shown In Section 6. Section 7 Summarizes The Conclu- Sions Of This Paper And Discusses Future Work.

II. SIGNIFICANCE OF THE SYSTEM

In This Section, We First Define Our Search Problem And De- Tail Our Assumptions. Then, We Introduce Our Efficient Search Algorithm, Which Is An Extension Of The Gnutella Approach.

Problem Definitions And Assumptions

In A P2P Network Each Peer Has The Same Role And The Com- Munication Between Any Two Nodes Is Symmetric. Such A Network Can Be Viewed As An Undirected Graph. Each Node Represents One Peer In The Network. If Peer A Directly Con- Nects To Peer B Then There Is A Logical Edge Between The Two Corresponding Nodes. In That Case, A Is Called The Neighbor Of B And Vice Versa. In This Paper, We Also Assume That The Network Graph Can Be Arbitrary As Long As It Is Fully Con- Nected And We Do Not Modify The Topology.

Each Peer Is Assigned A Local Document Collection On Which Some IR Search Engine Runs. Peers Return Top-Ranked Docu- Ments As The Result For A Given Query. For Simplicity, We Assume Each Search Engine Is Optimal, That Is, It Returns All Of Relevant Documents It Has. The Queries Considered Here Are In Natural Language Style Such As, "Information About What Manatees Eat".

III. LITERATURE SURVEY

Search In Peer-To-Peer Networks Is A Problem Rich In Previ- Ous Work.

As We Stated, Gnutella [5] Takes The Simplest Possible Ap- Proach To Search. A Peer Forwards Query Messages To All Of Its Neighbors Until Some Distance From The Source Is Reached. Since The Query Routing Policy In Gnutella Treats Each Peer Equally Regardless Of Queries, Gnutella Is Very Inefficient In Network Bandwidth.

Other Approaches Improve The Efficiency Of Routing Queries By Storing Information About Other Peers' Content. For Ex- Ample, Kalogeraki, *Et Al* [4] Proposed Storing Profiles Of The Past Query Behavior Of Each Neighbor To Improve The Future Search Efficiency. This Approach Is Not Robust Since It As- Sumes That Users Would Submit Similar Queries. Yang, *Et Al*

[6] Proposed A Technique Where Each Peer Maintains An Index Of Other Peers' Resources Who Are Within Some Num- Ber Of Hops. Maintaining Such An Index Is Costly If Topology Or Membership Changes Are Frequent. One Advantage Of Our Method Is That Each Node Is Fully Decentralized And Does Not Need To Store Other Peers' Information.

Another Idea To Improve Search Efficiency In P2P Networks Is To Reorganize The Topology Of Networks [7] [3]. For In- Stance, Mayank, *Et Al* [3] Propose Maintaining A Topology Where Peers Grouped Together If They Have Libraries Of Simi- Lar Topics; Queries Are Then Routed Only To The Closest Clusters. Although The Search Algorithm Can Take Advantage Of The Reorganized Topology To Improve Efficiency, The Ap- Proach Suggests Some Trade-Offs. First, Placing Peers Together Based On Topic May Degrade Network Performance If Juxtapose Peers Have Poor Bandwidth Between Them. Sec- Ond, Maintenance Costs Does Not Scale Well With Mem- Bership Changes. And Finally, In The Quality Of The Clusters, Which Depends On The Characteristics Of Collections Of Peers, Has A Large Affect On The Success Of The Reorganized Topology [3].

Distributed IR Research Assumes A Central Sever-Client Ar- Chitecture Where The Central Server Has Direct Access To The Indices Of All Collections In Clients. One Of Distributed IR Problems Most Related To This Paper Is Resource Selection, That Is, How To Pick The Most Relevant Collections. The CORI Resource Selection Algorithm [9] Uses A Bayesian Inference Network Model In The INQUERY System To Rank Collections. Although It Is Stable And Effective, It Is Hard To Integrate This Method To Search Engines Other Than INQUERY. Xu Et Al[10] Proposed A Method Where Collec- Tions Are Ranked By Kullback-Leibler Divergence Between Query Language Model And Collection Model. In [11] Luo Si Et Al Used The Very Similar Approach For Resource Selection. Both Methods Estimate Language Models By Word Fre- Quency. The Only Differences Are In Details On How To Estimate The

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 6, Special Issue , August 2019
International Conference on Recent Advances in Science, Engineering, Technology and Management at Sree Vahini Institute of Science and Technology-Tiruvuru, Krishna Dist, A.P

Language Models. In Our Experiments We Use The Same Approach To Estimate The Relevance Of A Collection Given A Query.

IV. METHODOLOGY

The Efficient Search Mechanism: Extending The Gnutella Protocol

A Naive Application Of Gnutella To Content-Based Retrieval Would Flood Query Messages Within Some Predefined Search Depth Limit And All Peers Receiving The Query Message Do Retrieval And Reply To The Source. By Doing So, A Lot Of Net- Work Bandwidth Is Wasted Since Only A Few Peers Have Documents Relevant To A Given Query.
Our Goal Is To Achieve A Desired Level Of Recall As Efficiently As Possible While Preserving The Advantages Of Gnutella, Which Includes The Lack Of Coordination Among Nodes. In A Random Graph, Peers That Have Relevant Documents Are Ran- Domly Distributed Across The Network. Thus, On Average, The More Peers That Are Visited By A Query, The More Relevant Documents The Source Peer May Receive; And Consequently, The More Bandwidth That Is Consumed. We Cannot Achieve A Sufficient Recall Level By Only Visiting A Small Number Of Nodes In Our Defined Problem. There Is An Unavoidable Trade-Off Between The Quality Of Retrieval And The Network Bandwidth. We Are Interested In How To Minimize This Trade-Off. Since A Peer Is Visited Only When It Receives The Query Message, It Is Difficult To Reduce The Number Of Query Messages If A Reasonable Recall Level Is Required. However**,** The Response Message, Which Is Much Larger Than The Query Message, Can Be Used More Efficiently.
We Modify Gnutella Such That, When Receiving Query Mes- Sages, Only Peers Estimated As Relevant Reply To The Source. No Matter Estimated As Relevant Or Not, All Peers Receiving Query Messages Still Forward Query Messages To All Of Their Neighbors Until A Distance From The Source Is Reached. At Each Peer Receiving The Query Message, We Calculate *P(Q|C),* The Probability Of Generating Query, *Q,* From The Collection, *C,* Of A Peer. Then We Set A Threshold. Peers With *P(Q|C)* Above The Threshold Are Regarded As Relevant. Only Relevant Peers Will Retrieve The Query Against Their Collec- Tions And Send Back The Source Retrieval Results.
Even Though Our Approach Floods Query Messages, It Is Still Resource-Efficient. The Reasons For This Are Two-Fold. First, The Size Of Query Message Is Much Smaller Than The Size Of Response Message That Contains Retrieval Results. Response Messages Are Well Utilized In Our Algorithm. Secondly, In Traditional Server-Client Architecture Or Hybrid P2P Architec- Ture, A Central Server Or A Super Peer Is Responsible For Calculating The Relevance Of A Larger Number Of Neighboring Peers In Order To Rank Them. In Our Approach Such A Re- Source-Consuming Computation Is Divided Over Individual Peers, Which Is Desirable Because Of The Decentralized Na- Ture Of P2P Networks.

V. EXPERIMENTAL RESULTS

We Group Documents In WT10G Into More Than Ten Thou- Sand Collections According To Their IP Addresses. Each IP Address Corresponds To One Peer. Table 1 Shows The Number Of Peers, And The Number Of Documents Per Peer, Where The Number Of Documents Is Binned Into Four Ranges. For Ex- Ample, There Are 6,188 Peers With A Collection Size Of Between 5 And 29 Documents.

N(The # Of Docs)	[5,29]	[30,59]	[60,99]	More Than 100
The #Of Peers	6188	1951	1068	2305

Table 1. Document-Peer Distribution

In P2P Networks, We Want To Minimize Resources Consumed For A Fixed Level Of Retrieval Quality. In This Paper, Resources We Are Interested In Are Network Bandwidth And Query Proc- Essing Cost. We Measure Retrieval Quality

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Special Issue , August 2019

International Conference on Recent Advances in Science, Engineering, Technology and Management at Sree Vahini Institute of Science and Technology-Tiruvuru, Krishna Dist, A.P

With Two Metrics: Recall And A Modified Version Of *Mean Reciprocal Rank* (MRR). We Evaluate Our Approach In The Following Three Ways: Query-Processing Efficiency, Network Efficiency, And A Modified Version Of MRR.

Query-Processing Efficiency

Here The Query-Processing Cost Is Referred To As The Cost Consumed By The Search Engine On Individual Peers. As- Suming Each Peer Uses The Same Search Engine, Query- Processing Cost Can Be Measured By The Number Of Peers Required To Do Retrieval Given A Query. We Are Interested In This Measure Because Many Effective Retrieval Algorithms Are Both Time And Space Consuming And Peers May Receive A Lot Of Queries At The Same Time. On The Other Hand, In A Real P2P System, Given A Query, Many Peers Do Not Contain Any Relevant Documents. In Order To Utilize The Computing Re- Sources Of Each Peer As Efficiently As Possible, We Introduce Query-Processing Efficiency. Given A Query, If We Assume *N* Peers Are Searched And *M* Out Of *N* Peers Need To Perform Retrieval On Their Collections, Then The Recall Level, *R*, Is The Number Of Relevant Documents In Those *M* Peers Over The Total Number Of Relevant Documents In The Network. The Query Processing Efficiency Is Measured By The Average Ratio Of (*R/M)*.

Network Bandwidth Efficiency

The Efficiency Of Network Bandwidth Is Measured By The Average Bandwidth Consumed At A Certain Recall Level. Given A Query, We Assume The Cumulative Recall After N Peers Are Searched Is *R*, And *M* Out Of *N* Peers Need To Reply The Source, The Bandwidth Consumed At Recall Level R Is Calculated As:

Bandwidth= *N*S1+M*S2*,

Where *S1* Denotes The Size Of The Query Message And Is Set To 100 Bytes In Our Experiments; *S2* Denotes The Size Of The Response Message. We Assume Each Response Message Con- Tains 10 Documents With The Size Of 1,000 Bytes And The Response Message Header With The Size Of 100 Bytes. So *S2* Is Set To 10,100 Bytes In Our Experiments.

Results For Query-Processing Efficiency

We First Investigate The Effect The Smoothing Parameter Lambda Has On The Performance. We Use Centralized Archi- Tecture And Standard Gnutella As Our Baseline. In Figure 1, The Line Labeled "Central Mode" Denotes The Accumulative Recall When All Peers Are Directly Connected To The Central Server And The Server Ranks Peers According To *P(Q|C)* De- Fined In Section 3.2 And Retrieves Peers In That Order. "Gnutella" Denotes The Naive Application Of Gnutella To Content-Based Retrieval Where Each Peer Receiving The Query Message Will Reply The Source. Gnutella Can Be Seen As The Special Case Of Our Approach When Every Node Is Considered

Figure 1: Query-Processing Efficiency

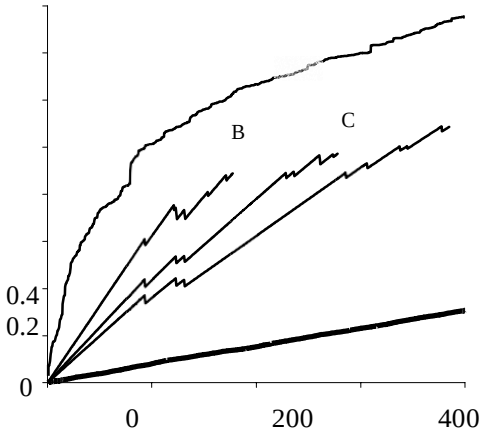

Fig 1: The Number Of Nodes Retrieved

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Special Issue , August 2019

International Conference on Recent Advances in Science, Engineering, Technology and Management at Sree Vahini Institute of Science and Technology-Tiruvuru, Krishna Dist, A.P

Figure 1 Shows That Our Approach Is Significantly Better Than Gnutella No Matter How Lambda Is Set. This Is Because Our Approach Can Intelligently Estimate The Relevance Of Each Peer. Secondly, A Higher Lambda Is Helpful For Improv- Ing Query-Processing Efficiency. Lastly, A Higher Lambda Has The Risk That Some Of The Queries May Not Have A High Recall Even When The Whole Network Is Searched.

VI. CONCLUSION AND FUTURE WORK

We Proposed An Efficient Search Algorithm By Extending The Gnutella Protocol. The Idea Is To Estimate The Relevance Of Peers Locally When Receiving Query Messages. Only Those Peers Deemed As Relevant Will Retrieve The Query And Send Response Messages Back To Source. Based On Evalua- Tion Of A Large Real Testbed, We Show That The Tradeoff Between The Quality Of Retrieval And The Resources Con- Sumed Is Greatly Improved Over Gnutella While Simplicity, Robustness And The Autonomy Of Peers Are Maintained. We Also Notice That The Collection Language Model P(Q|C) Is Far From Being The Perfect Indicator Of The Relevance Of Peers. For The Future Work, P(Q|C) May Be Improved By Trying Different Smoothing Techniques Such As Dirichlet Smoothing. We Are Also Interested In Other Methods That May Help To Predict The Relevance Well In P2P Systems

REFERENCES

[1]The Gnutella Protocol Specification V4.0 Http://Www9.Limewire.Com/Developer/Gnutella_Protocol_V4

.0.Pdf

[2]S. Ratnasamy,Et Al. A Scalable Content-Addressable Net- Work In *ACM SIGCOMM '01,Auguest 2001*

[3] M. Bawa, G. S. Manku, P. Raghavan. SETS: Search Enhanced By Topic Segmentation. In Proceedings Of The 26th Annual International ACM SIGIR Conference, 2003

4] H. Zhang, W.B.Croft And B.N. Levine. Efficient To- Pologies And Search Algorithms For Peer-To-Peer Content Sharing. In *Univ. Of Massachusetts,Amherst, CIIR Technical Report IR-314,August,2003*

5J. Lu, J. Callan. Content-Based Retrieval In Hybrid Peer- To-Peer Networks" In *Proceedings Of The ACM CIKM 03 Conference ,2003*

[6] J. Callan "Distributed Information Retrieval" In *W.B.Croft ,Editor,Advances In Information Retrieval . Klu- Wer Academic Publishers. (Pp.127-150)*

[7] J. Xu, W.B. Croft. Cluster-Based Language Models For Distributed Retrieval In *Proceedings Of The 22th Annual In- Ternational ACM SIGIR Conference, 1999*

8 .L. Si, R. Jin, J. Callan, P. Ogilvie. Language Modeling Framework For Resource Selection And Results Merging In *Proceedings Of The ACM CIKM 02 Conference ,2002*

10 M. Faloutos, P.Faloutsos, C. Faloutos On Power-Law Relationships Of The Internet Topology. In *FIGCOMM,1999*

11.C. Palmer And J. Steffan Generating Network Topolo- Gies That Obey Power Law. In *Proceedings Of GLOBECOM '2000.*

12. J. Lin And B. Katz. Question Answering Techniques For The World Wide Web. In *Tutorial Presentation At EACL,2003*

13. Http://Trec.Nist.Gov/Data/Topics_Eng/Topics.501-550.Txt