ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.5, May 2006, pp.1170−1181
DOI: 10.1360/jos171170

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

SemreX:一种基于语义相似度的 P2P 覆盖网络*

陈汉华, 金 海+, 宁小敏, 袁平鹏, 武 浩, 郭志鑫

(集群与网格计算湖北省重点实验室(华中科技大学),湖北 武汉 430074)

SemreX: A Semantic Similarity Based P2P Overlay Network

CHEN Han-Hua, JIN Hai+, NING Xiao-Min, YUAN Ping-Peng, WU Hao, GUO Zhi-Xin

(Cluster and Grid Computing Key Laboratory of Hubei Province (Huazhong University of Science and Technology), Wuhan 430074, China)

+ Corresponding author: Phn: +86-27-87543529, E-mail: hjin@hust.edu.cn, http://grid.hust.edu.cn/hjin

Chen HH, Jin H, Ning XM, Yuan PP, Wu H, Guo ZX. SemreX: A semantic similarity based P2P overlay network. *Journal of Software*, 2006,17(5):1170−1181. http://www.jos.org.cn/1000-9825/17/1170.htm

Abstract: The decentralized structure together with the self-organization and fault-tolerant features makes P2P network an effective model for information sharing, however, the content location remains a serious challenge of large scale P2P networks. In this paper, the SemreX is introduced, which is a P2P system for literature retrieval. Semantic-similarity-based P2P overlay and routing algorithms are proposed for SemreX. Experimental results show that searching in semantic overlay greatly improves the efficiency of search.
Key words: SemreX; P2P; semantic similarity; ACM topic; semantic overlay networks

摘 要: 对等(peer-to-peer)网络的非集中结构、良好的自治性及容错性等特征,使其可能成为 Internet 上有效的信息共享模型.然而,内容定位问题仍然是大规模 P2P 网络中信息共享所面临的挑战.SemreX 系统是一种 P2P 网络环境下的文献检索系统.针对 SemreX 系统,提出一种基于语义相似度的 P2P 拓扑管理和查询路由算法.仿真实验结果表明,语义拓扑能够有效地提高系统的搜索效率.
关键词: SemreX;P2P;语义相似度;ACM topic;语义覆盖网
中图法分类号: TP311 文献标识码: A

随着网络技术的飞速发展,网络中的信息资源越来越丰富.在人们寻求合理而有效地利用这些信息的途径的过程中,传统的网络计算模型(例如 C/S 结构)潜在的缺陷也越来越明显,对等(peer-to-peer)网络的非集中结构、良好的自治性及容错性等特征,使其可能成为 Internet 上有效的信息共享模型.目前,P2P 技术已经被广泛用于 Internet 环境下的文件共享.这些系统能够提供基于文件名称的数据共享功能,却难以实现类似 Web 搜索引擎提供的信息检索功能.大规模网络环境下的内容定位问题,是现有 P2P 系统深度信息共享面临的首要挑战[1].现有的 P2P 文件共享系统按照结构特征通常被分为 3 类.第 1 类是基于集中索引的 P2P 结构.例如 Napster (http://www.napster.com)系统为 P2P 网络上的资源建立集中索引,所有搜索请求都必须经过集中索引服务器查

* 本文为 2005 年中国计算机大会推荐优秀论文.Supported by the National Grand Fundamental Research 973 Program of China under Grant No.2003CB317003 (国家重点基础研究发展规划(973))
Received 2005-06-15; Accepted 2005-12-16

询后才能定位并访问资源.集中索引简单、实用.但随着 P2P 网络规模的扩大,集中索引服务器必然成为系统性能的瓶颈和单一失效点,从而降低系统的可扩展性和可靠性.第 2 类 P2P 结构基于分布式哈希(distributed hash table,简称 DHT)技术.典型代表是麻省理工学院设计的 Chord[2]和加州伯克利与 AT&T 设计的 CAN (content-addressable network)[3].基于 DHT 的系统具有良好的搜索性能,但在大规模的动态 P2P 环境下,维持系统的结构代价很高[4].另外,基于 DHT 的 P2P 系统只支持精确的对象键值搜索,缺乏模糊搜索能力,更难以有效支持基于内容的定位.第 3 类被称为无结构的 P2P 系统.这类系统对 P2P 网络的拓扑结构不进行限制,主要采用洪泛或随机游走[5]的搜索策略.例如,Gnutella 协议使用洪泛搜索策略定位 P2P 网络上的资源,这种搜索策略使查询消息在 P2P 网络上以指数的方式增长,从而使得 P2P 网络的可扩展性受到限制.对大多数 P2P 网络来说,随机游走是一种比较盲目的路由策略,一般以牺牲查全率为代价降低冗余的消息数量.

本文介绍的 SemreX 系统(http://grid.hust.edu.cn/semrex)是基于 P2P 的参考文献语义检索系统.针对无结构 P2P 系统存在的问题,我们提出基于语义相似度的语义拓扑和查询路由策略.仿真实验结果表明,语义拓扑能有效提高系统查询效率.

本文第 1 节简要介绍 SemreX 的主要功能和系统结构.第 2 节针对内容定位问题,提出 SemreX 语义拓扑和查询路由策略.第 3 节用仿真实验对语义拓扑的搜索性能和 P2P 网络负载进行测试,并分析测试结果.第 4 节介绍相关工作.最后作总结并对未来的工作进行展望.

1 SemreX 系统介绍

为了更好地说明 SemreX 系统的语义拓扑和查询路由策略,我们先对 SemreX 系统的功能和系统结构作简单的介绍.

1.1 系统功能

SemreX 系统为计算机领域研究者提供如下功能:

(1) 基于语义的本地文献信息存储和管理.为了实现 SemreX 网络 Peer 之间在语义上的信息交互,我们设计了文献本体作为所有结点的共享信息模型.本地文件系统中异构格式(目前 SemreX 主要支持 PDF 文档)的文献信息经过信息提取后,按照共享本体模型(如图 1 所示)存储于本地 Sesame 数据库[6]中.Peer 设计可主动检测文献资源的状况,并自动更新存储空间,无须用户干预.在当前版本下,用户可执行本地文献信息的导入、导出和查询等操作.

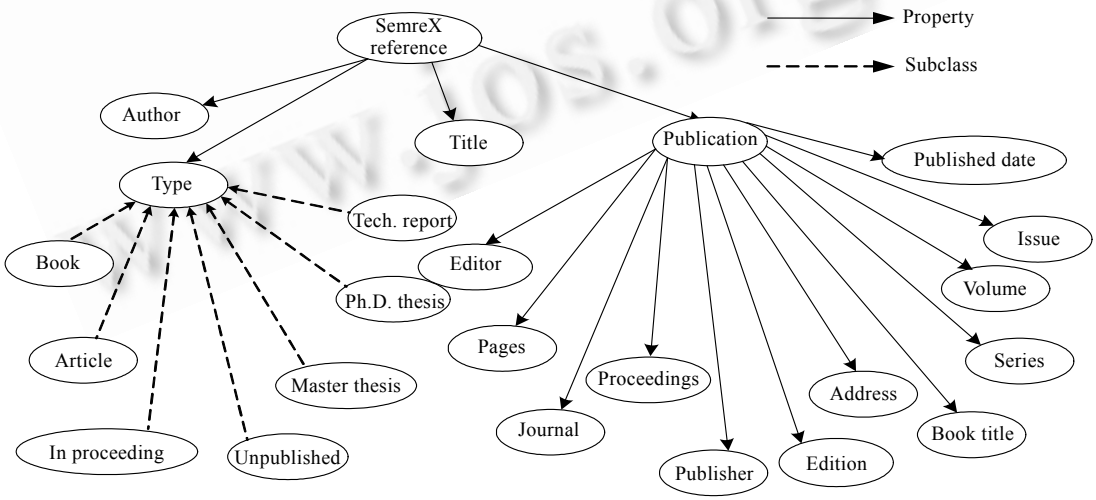


Fig.1 Reference ontology in SemreX

图 1 SemreX 参考文献本体

(2) 基于语义的 P2P 文献搜索.与传统的 P2P 系统基于文件名的搜索功能相比,SemreX 的主要特色是基于语义的文献搜索.为此,SemreX 用两个本体来规范 SemreX 网络环境下文献信息的语义模型:一个是前面提到的参考文献本体;另一个是 ACM Topic(The ACM Topic Hierachy. http://www.acm.org/class/1998),它是计算机研究领域的标准分类模型(如图 2 所示).

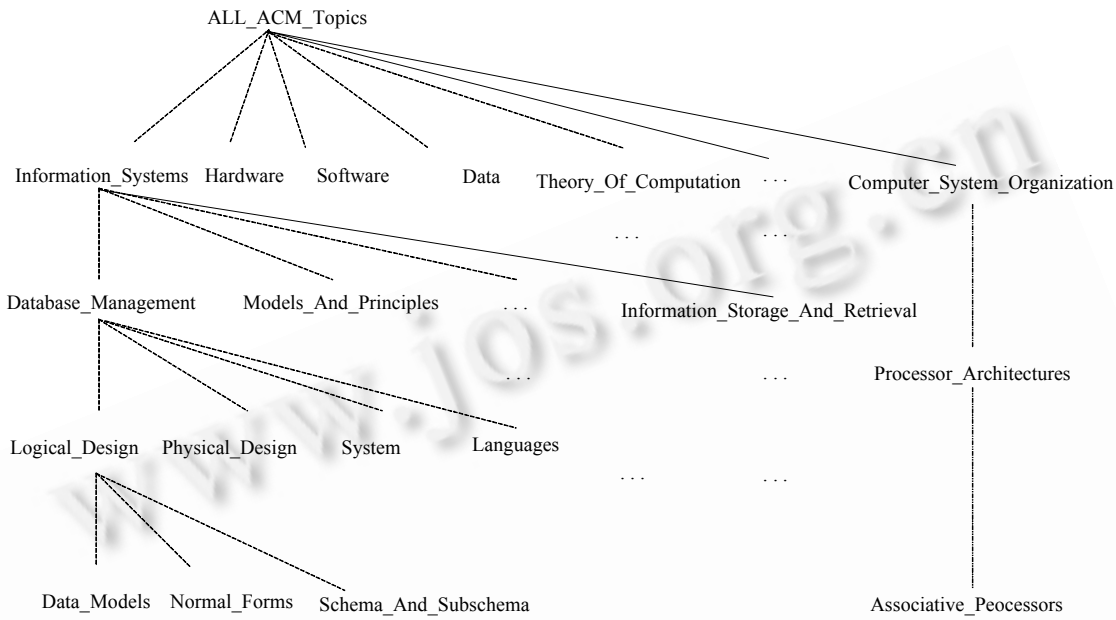


Fig.2 IS-A concept tree of ACM Topic

图 2 ACM Topic 的 IS-A 概念树

1.2 SemreX系统体系结构

SemreX 的结点软件由 5 个模块构成(如图 3 所示),它们分别是:图形用户界面、本地存储管理模块、本地语义库管理模块、系统控制模块和 P2P 网络通信模块.

图形界面是用户使用 SemreX 系统的图形接口,它支持语义查询语句生成和查询结果处理.用户通过图形界面输入查询请求,系统根据用户输入自动生成语义查询语句,并在 P2P 网络上进行搜索,执行查询.用同构本体描述的查询结果,能根据用户需求将信息有效集成起来反馈给用户.

本地存储管理模块主要提供文档分类、异构参考文献信息提取、参考文献信息语义封装和系统监控等功能.文献头部和尾部信息被抽取并进行分类(目前,SemreX 已实现文档尾部参考文献部分信息的抽取),抽取的信息依据 SemreX 的参考文献本体模型进行存储.系统监控模块设计成一个常驻线程,负责监测系统资源状况,在系统 CPU 空闲且无交互操作时,执行主动存储线程(目前版本仅支持用户导入操作).

本地语义库管理模块采用 Sesame 数据库存储本地语义信息,主要包括 SemreX 文献本体、ACM Topic 本体、语义路由表等.

系统控制模块主要负责 SemreX 的逻辑控制,其中包括语义拓扑管理和查询路由控制.本文第 2 节将对此进行详细论述.

P2P 网络通信模块在整个系统中扮演了“传输层”的角色,它封装了所有底层的 P2P 通信细节,为 SemreX 上层模块提供透明的 Peer 发现、点对点通信、消息封装处理和 Peer 分组管理等功能.

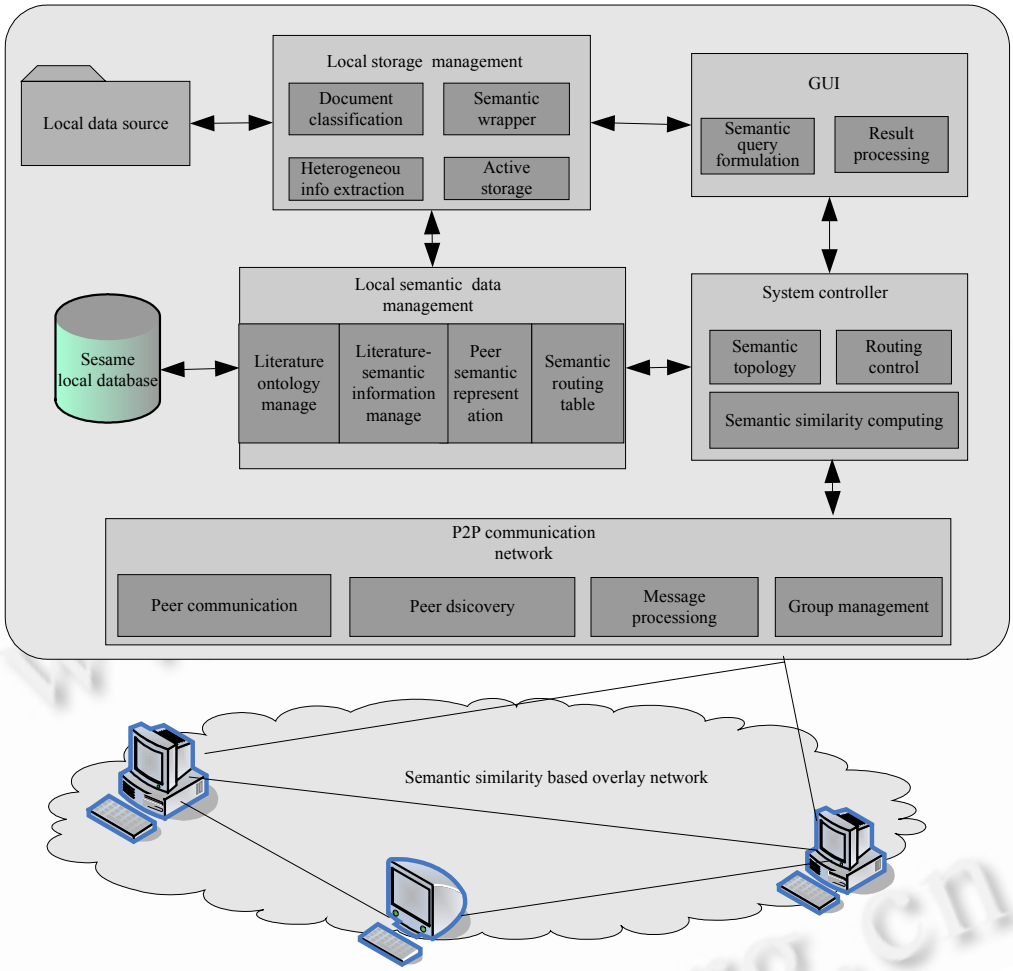


Fig.3 System architecture of SemreX

图 3 SemreX 体系结构

2 SemreX 语义拓扑网络

拓扑管理和查询路由策略是解决 P2P 系统上基于内容定位的关键.SemreX 根据 Peer 之间的语义相似度生成语义拓扑网络.语义拓扑是对人类社会关系网络中兴趣关系网络的模拟[7].

2.1 语义相似度

对象相似性是信息检索领域的一个重要研究课题,按照比较对象的种类,一般被分为两类:同类对象之间的相似性(例如文挡相似性)和不同类对象之间的相似性(例如查询语句和文挡的相似性).基于概念语义相似性度量方面的最新研究成果,我们提出一种用于度量 Peer 之间相似性的统计方法,并以此作为 SemreX 语义拓扑算法的基础.为引入语义拓扑算法,我们先对相似性度量方法做简单的介绍.

近几年,研究者们提出了许多概念相似性的度量方法[8],有些被实践证明能够很好地解决智能计算领域的一些实际应用.总的来说,这些方法可以被分为两类:一类被称为“边计算法”,这类方法将两个概念在语义网络中的最短路径和所处的深度等作为基本度量;另一类被称为“共享信息含量法”,两个概念如果所含的相同信息越多则越相似.

(1) 边计算法.根据两个概念之间的距离来计算它们之间的相似性,是一种直观且容易理解的方法.Rada 在

文献[9]中证明,两个概念在概念层次拓扑上相距的最短路径,是一种有效地比较两个概念相似性的度量.文献[9]同时指出,两个概念在概念树上所处的深度(一组概念在概念树上的深度,可以用它们所有"超类"中最具体的一个概念在概念树上的深度值来度量)也决定了两个概念的相似程度.例如在图 2 中,*Information_System* 与 *Hardware*,以及 *Logical_Design* 与 *Physical_Design* 的最短路径都是 2(即 l=2),但两组概念分别所处的深度是不同的,从而两组概念的相似度也是不同的.Li 在文献[10]中提出了一种有效地度量 IS-A 概念树上两个概念之间相似度的函数,见公式(1).

$$Sim(T_1,T_2)=f_1(l)\cdot f_2(h)=\begin{cases}e^{al}\cdot\dfrac{e^{\beta h}-e^{-\beta h}}{e^{\beta h}+e^{-\beta h}}, & \text{if } (T_1\neq T_2)\\ 1, & \text{if } (T_1=T_2)\end{cases} \tag{1}$$

其中:T_1,T_2 是概念树上的任意两个概念;l 是它们在概念树上的最短路径;h 是它们的深度.公式(1)表明,两个概念的相似度关于 l 单调递减,关于 h 单调递增.α和β用来调整 l 和 h 对概念相似度的影响程度.根据 Li 的测试,α=0.2,β=0.6 是获得最佳度量效果的优化值.

(2) 共享信息含量法.Resnik 提出的共享信息含量法指出,两个概念的相似性决定于它们共享信息的程度[11].他认为,两个概念共享信息含量越大,这两个概念越相似.两个概念的共享信息含量可以用两个概念在概念树上的所有共同"超类"所具有的最大信息含量来表示.根据信息论的基本论点,某个概念 T 的信息含量可以用关于 $p(T)$的单调递减函数–log $p(T)$来表示[12].$p(T)$是指概念 T 在标准文档集中出现的概率,也就是 T 出现的统计次数与所有概念出现的统计次数总和之比,即 $p(T)=freq(T)/N$,且概念树上任何一个"子类"出现的频率应该累加到其父类的出现频率中.由此,一个概念在文集中出现的概率越大,其所具有的信息含量越小.Resnik 的度量模型用以下公式度量 T_1 和 T_2 的相似性:

$$Sim(T_1,T_2)=\max_{T\in S(T_1,T_2)}[-\log p(T)]=-\log p[lso(T_1,T_2)] \tag{2}$$

其中:$S(T_1,T_2)$是 T_1,T_2 的公共超类集合;$lso(T_1,T_2)$是指 T_1 和 T_2"超类"中最为具体的一个概念,也就是 T_1 和 T_2 的所有"超类"中具有最大深度的一个.例如,*Data_Models* 和 *Normal_Forms* 共同的"超类"集为 *S*(*Data_Models*, *Normal_Forms*)={*Logical_Design*,*Database_management*,*Information_Systems*,*ALL_ACM_Topics*},容易求得 *lso*(*Data_Model*,*Normal_Forms*)={*Logical_Design*},故 *Sim*(*Data_Model*,*Normal_Forms*)=–log *p*(*Logical_Design*).

2.2 基于语义相似度的P2P拓扑

SemreX 中,我们对本地存储的文献按照 ACM Topic 进行分类并统计.$P=\{\langle T_i,\lambda_i\rangle, i=1,2,\ldots,m\}$描述 Peer 的语义信息,其中:$T_i$ 是 Peer 中存储的文献被归入的 ACM Topic 类别;λ_i 是被归入 T_i 的文档数量在 Peer 文档集合中的统计权重,λ_i 通过下式计算:

$$\lambda_i=\frac{N_i}{\sum_{j=1}^{j\le|p|}N_j} \tag{3}$$

其中:N_i 是分类到 T_i 的文献在 Peer 中累计出现的次数,也就是 Peer 中以 T_i 为主题的参考文献数量;而 $\sum_{i=1}^{i\le|p|}N_i$ 是 Peer 中文献总数.由此,$P=\{\langle T_i,\lambda_i\rangle,i=1,2,\ldots,m\}$将 Peer 描述为带权 Topic 的集合.联系 SemreX 的应用场景中,P 具有非常直观的意义:T_i 表示 Peer 用户的某个"研究兴趣",而权重 $\lambda_i(0<\lambda_i<1)$可用于度量 Peer 用户对 T_i 感兴趣的程度.值得指出的是,因为在分类时我们将 ACM Topic 概念树展开成平面进行处理,所以这里与第 2.1 节中求 Topic 树结点的信息含量不同,我们不将概念树中 Topic 出现的频率向上累加到其"超类".这样,不会因为某个 Topic 的出现而导致所有的"超类"都成为 Topic 集合中的元素.

我们用公式(4)来度量两个 SemreX 结的语义相似性(如图 4 所示):

$$Sim(P_1,P_2)=\sum_{j=1}^{j\le|P_1|}\sum_{i=1}^{i\le|P_2|}[Sim(T_i,T_j)\times(\lambda_i\times\lambda_j)] \tag{4}$$

其中,函数 $Sim(T_i,T_j)$用于计算 T_i 和 T_j 之间的相似度.可采用第 2.1 节介绍的两种度量相似度的方法.公式(4)的主

要思想是求两个带权概念集合的笛卡儿集中所有概念对之间相似度的加权平均.举例来说明,给定两个 Peer 描述为 $P_1=\{\langle a,0.2\rangle,\langle b,0.8\rangle\}$,$P_2=\{\langle b,0.4\rangle,\langle c,0.6\rangle\}$,同时给定 $Sim(a,b)=0.5$,$Sim(a,c)=0.3$,$Sim(b,c)=0.7$.根据公式(4),P_1 和 P_2 的语义相似度计算如下:$Sim(P_1,P_2)=0.5\times0.2\times0.4+0.3\times0.2\times0.6+1.0\times0.8\times0.4+0.7\times0.8\times0.6=0.772$.

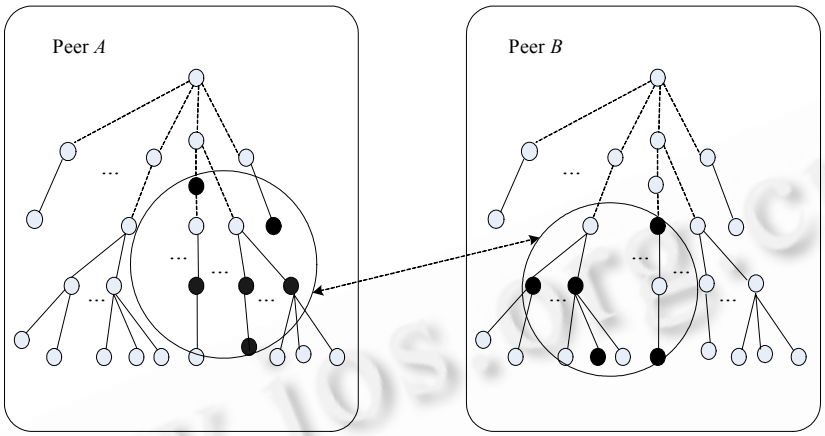


Fig.4 Semantic similarity between peers

图 4 SemreX 中 Peer 的相似性

基于以上 Peer 相似度的度量方法,我们给出 SemreX 语义拓扑算法(见算法 1).算法的基本思想是,当 Peer 加入 SemreX 网络时,将自己的语义描述信息作为广告发送给网络中的某个结点,广告设置一个初始 *TTL*,此结点将广告信息广播给它的所有邻居.所有接收到广告的结点将 *TTL* 减 1,并比较源 Peer 和本地 Peer 的语义相似度:如果语义相似度高于给定的阈值,本地 Peer 将源 Peer 语义信息存入本地路由表;否则,将广告转发给本地 Peer 的所有邻居.语义广告依此在网络上传播,直到 *TTL*=0,从 P2P 网络上删除.

算法 **1**. 语义拓扑生成算法.

输入:当前 Peer 的语义描述 $P^j=\{\langle T_i^j,\lambda_i^j\rangle,i=1,2,\ldots,m^j\}$.

输出:当前 Peer 的语义路由表 $Neighbor_{semantic}(P^j)$.

算法描述:

```
1.  while(true) do
2.      监听网络上的广告信息;
3.      当监听到来自P^k的广告信息后;
4.      设置P^k.TTL=P^k.TTL−1;
5.      if (Sim(P^j,P^k)>SIMILARITY_THRESHOLD)
6.          将P^k的语义描述加入P^j的邻居路由表:Neighbor_semantic(P^j)=Neighbor_semantic(P^j)∪{P^k};
7.          通知P^k,并将自己的语义广告发送给P^k;
8.      end if
9.      if (P^k.TTL>0)
10.         将广告信息({⟨T_i^k,λ_i^k⟩,i=1,2,…,m^k} and P^k.TTL)转发给P^j的邻居结点
11.     end if
12. end do
```

2.3 基于语义相似度的**P2P**查询路由算法

查询路由策略直接影响 P2P 系统的可扩展性和查询效率.不同于传统 P2P 路由策略,SemreX 系统通过比较查询请求和目标结点的语义相似度,选择最有可能返回查询结果的 Peer 转发查询请求.SemreX 采用 Sesame 数

据库支持的语义查询语言 SeRQL 来描述一个搜索请求.图 5 给出了一个查询语句示例及其对应的图形化描述.这个语句描述的语义是“查询 2005 年发表的所有以操作系统为主题的文章及其作者”,其中 */ALL_ACM_Topics/Software/Operting_Systems* 指定了文献被分类到的 Topic.

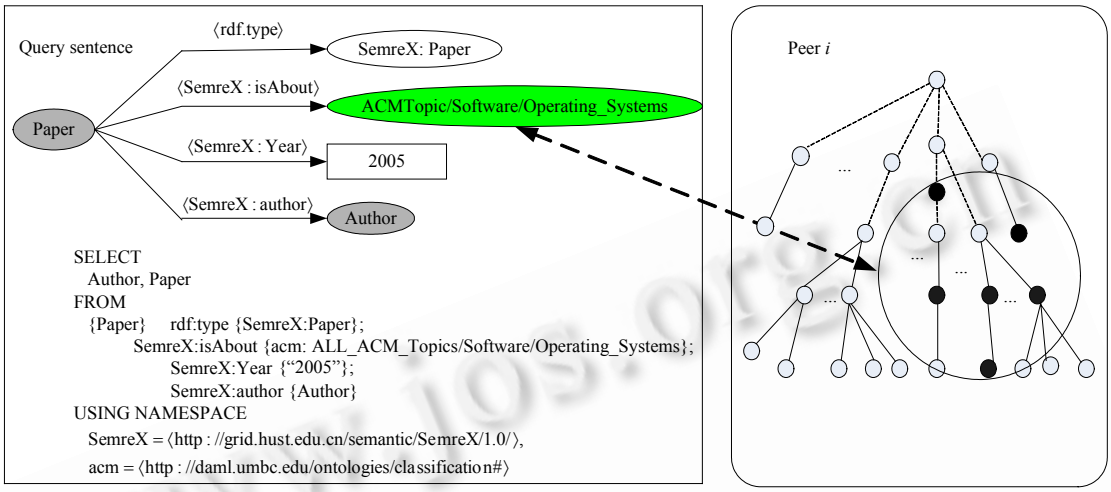


Fig.5 Semantic similarity between a query and a peer

图 5 SemreX 中查询语句与 Peer 的语义相似性

如何度量查询语句和 Peer 的语义相似性是 SemreX 路由策略的关键.我们采用的主要策略是比较查询语句中的 Topic 和路由表 Peer 中带权 Topic 的语义相似度,并找到其中的最大值

$$Sim(T_Q, P) = \underset{T_i \in P}{\mathrm{Max}}\{Sim(T_Q, T_i) \times \lambda_i\} \tag{5}$$

由公式(5)可知,查询语句和 Peer 的语义相似度决定于查询请求和目标结点 Topic 的相似度和被度量的 Topic 在 Peer 中权重的乘积.基于查询请求和 Peer 的相似性度量方法,算法 2 给出 SemreX 的查询路由策略的详细描述.其主要思想是,源 Peer 向 SemreX 语义拓扑网络中发送具有初始 *TTL* 值的语义查询消息(SeRQL 语句),当某个 Peer 收到该消息时,将搜索语句中的 Topic 与路由表中的邻居 Peer 进行语义相似度比较,并将查询请求转发给相似度高于给定阈值的 Peer.特别地,如果本地 Peer 与查询消息具有较高的相似性,则在本地执行查询请求,并向源 Peer 返回查询结果.

算法 **2**. SemreX 查询路由算法.

假设 P^j 是当前 Peer.

输入:语义查询消息 Q=*Query*(*Expression*(T_Q), P^k.*id*, *TTL*)

当前 Peer 的语义描述 P^j={⟨ T_i^j , λ_i^j ⟩,i=1,2,…,m^j}

当前 Peer 的邻居路由表 $Neighbor_{semantic}(P^j)$= {P_1^j,...,P_m^j}

输出:返回查询结果 R^j

转发查询消息的目的 Peer 列表 $P_{sim}(Q)$={P_1,P_2,...,P_n}

算法描述:

1. 初始化转发列表$P_{sim}(Q)$=∅;
2. while (true) do
3. 监听P2P网络上的查询消息
4. 当监听到来自P^k的查询消息Q=*Query*(*Exprssion*(T_Q),P^k.*id*,Q.*TTL*)时;
5. 设置Q.*TTL*=Q.*TTL*−1;
6. if (*Sim*(T_Q,P^j)>*High_SIMILARITY_THRESHOLD*)

7. 执行本地查询,返回查询结果R^j;
8. 将结果返回给P^k;
9. end if
10. if ($Q.TTL$>0)
11. for 每个 $P_i^j \in Neighbor_{semantic}(P^j)$ $1 \le i \le m$
12. if ($sim(T_Q, P_i^j)$>$Route\_SIMILARITY\_THRESHOLD$)
13. 将 P_i^j 加入转发列表$P_{sim}(Q)=P_{sim}(Q) \cup \{P_i^j\}$;
14. end if
15. end for
16. 将查询请求转发给转发列表$P_{sim}(Q)$中的每个Peer;
17. end if
18. end do

3 仿真实验及性能评价

为了测试 SemreX 基于语义相似度的 P2P 拓扑的有效性,我们将语义拓扑与 Gnutella 网络进行了对比测试.测试的性能指标是信息查全率和 P2P 网络中每查询平均需要处理的消息数.

信息查全率是衡量基于内容的检索的重要指标之一,它反映检索到的文档占所有相关文档的比例.在我们的模拟测试中,语义查询中指定相关的 Topic 和若干个关键字.查全率定义见公式(6),其中 $Doc_{relevant}$ 是针对某个查询,P2P 网络上分布的所有相关文档的集合,$Doc_{retrieved}$ 是实际搜索到的文档集合.

$$Recall = \frac{\left|Doc_{relevant} \cap Doc_{retrieved}\right|}{\left|Doc_{relevant}\right|} \tag{6}$$

3.1 实验设置

在仿真测试中,我们用 Pajek 网络分析软件[13]生成大致符合α=3.0,结点平均度为 2.8~2.9 的 Power law 网络[14]来模拟 Gnutella,节点规模为 1 000~5 000.我们按照参数为α_{zipf_doc}=1.2,n=MAX_*DOCUMENTS*=500 的 Zipf 分布为每个节点分布文档,每个节点发出的查询数符合参数为α_{zipf_doc}=1.0,n=MAX_*QUERIES*=200 的 Zipf 分布.虽然 ACM Topic 包含计算机的 1 287 个研究领域,但我们发现每个研究者感兴趣的领域一般只有几个,而且往往有一个集中研究的领域,在测试中我们假设每个 Peer 上有 4 个 Topic,并保证其中一个 Topic 的权重较高(80%).为了模拟语义拓扑,我们让底层 Power law 网络上的每个结点向网络上广播 *TTL_overlay*=2 的语义广告,结点间按照 *SIMILARITY_THRESHOLD*=0.5 的阈值生成语义链接.具体的参数设置见表 1.

Table 1 Configurations of the simulation

表 **1** 仿真实验的参数设置

Parameters	Description	Value
NO_PEERS	Number of nodes in the network	1k~5k
A	Exponent α of power law	3.0
AVG_DEGREE	Average degree	2.8~2.9
NO_TOPICS	Number of topics	30
MAX_*DOCUMENTS*	Number of documents each peer	1~500
SIMILARITY_THRESHOLD	Threshold of similarity	0.5
MAX_*QUERIES*	Number of queries by each peer	1~200
MAX_*CONTENTS*	Number of keywords each document	20
TTL_overlay	TTL for clustering	2
TTL	TTL for search	2~4
α_{zipf_doc}	Exponent α of document distribution	1.2
α_{zipf_query}	Exponent α of query distribution	1.0

Topic 语义相似度是语义拓扑的基础.在实验中,我们采用边计算方法对 Topic 的相似度进行了测试.由于

ACM Topic 的概念模型已知,我们在测试中用树型数据结构来表示.任意给定树上的两个结点 a,b,设其深度分别为 $A=depth(a)$和 $B=depth(b)$,首先在树上找到这两个结点的共同(祖)父结点中深度最大的结点 c,设其深度为 $C=depth(c)$,那么,这两个结点的最短路径可快速计算为$(A+B)–2C$.因此,依照公式(1)的度量模型,它们之间的相似度容易测得.表 2 列举了部分 Topic 的相似度的测试值.

Table 2 Similarity testing between topics

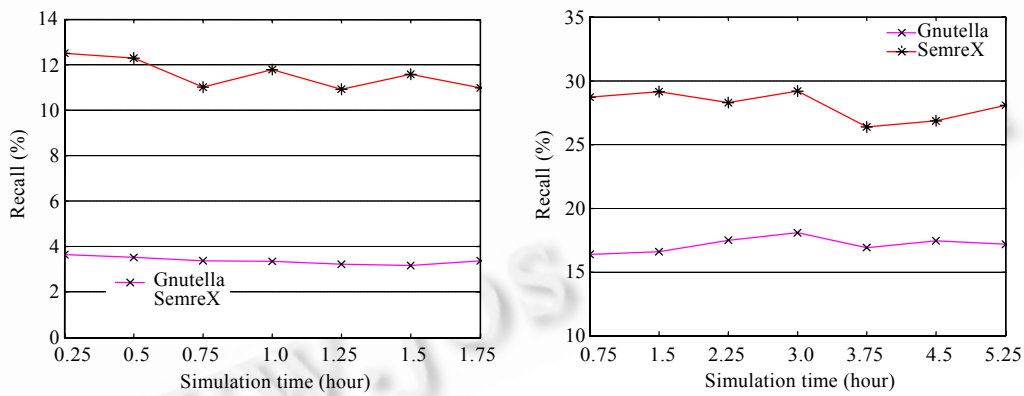
表 2 Topic 相似度测试

Topic 1	Topic 2	Similarity
Distributed systems	High-Speed networks	0.519
Client/server	Network operating systems	0.659
Distributed database	Data models	0.132
Routers	Super computers	0.251
Digital library	Information search and retrieval	0.634
Linguistic processing	Associative processors	0.132
Formal languages	Pattern matching	0.306
Data models	Information search and retrieval	0.306
Algebraic language theory	Formal languages	0.775
Content analysis and indexing	Retrieval models	0.519
Computer system organizations	Theory of computing	0.359

这里需要指出的是,基于边计算的方法是一种相对静态的方法,没有考虑测试文档集合的统计信息,而基于信息含量的相似度适合在集中文集中进行测试.本文主要针对无集中服务器的语义拓扑模型,因此只对基于边计算的方法进行了测试.

3.2 实验结果与分析

为了测试语义拓扑中查询的效率,我们以 NO_PEERS=1000 的规模为例对语义拓扑和 Gnutella 网络进行了对比测试,结果显示(如图 6 所示),当 TTL=2 和 TTL=3 时,语义拓扑能有效提高查全率.



(a) TTL=2 (b) TTL=3

Fig.6 Comparison of recall in SemreX and Gnutella

图 6 SemreX 和 Gnutella 查全率对比

我们发现,当 $TTL\geq4$ 时,两种方式的查全率都非常高而且相当接近,甚至某些时候,Gnutella 方式具有更好的效果.我们分析了这个结果,认为这是由于 Gnutella 网络的在规模相对较小的情况下,当 $TTL\geq4$ 时已经基本上遍历了整个网络.我们又分别对 NO_PEERS=2000~5000 的网络规模进行了查全率测试.

在图 7 中,我们将语义拓扑和 Gnutella 的网络路由 TTL 固定为 3,改变网络规模,测试结果显示,语义拓扑的查全率与 Gnutella 方式相比,在大部分测试的网络规模下都能提高 100%甚至更多.而且,当网络规模增加时,语义拓扑查全率的优势更为明显.

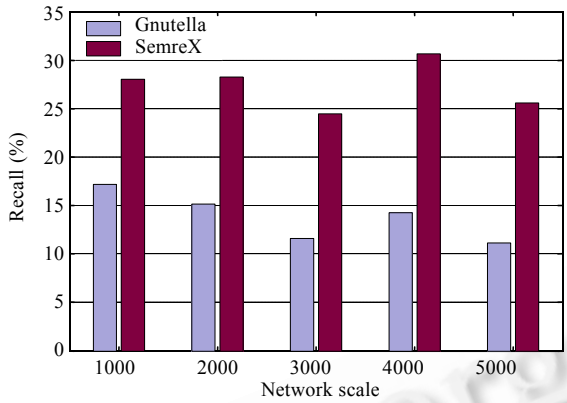
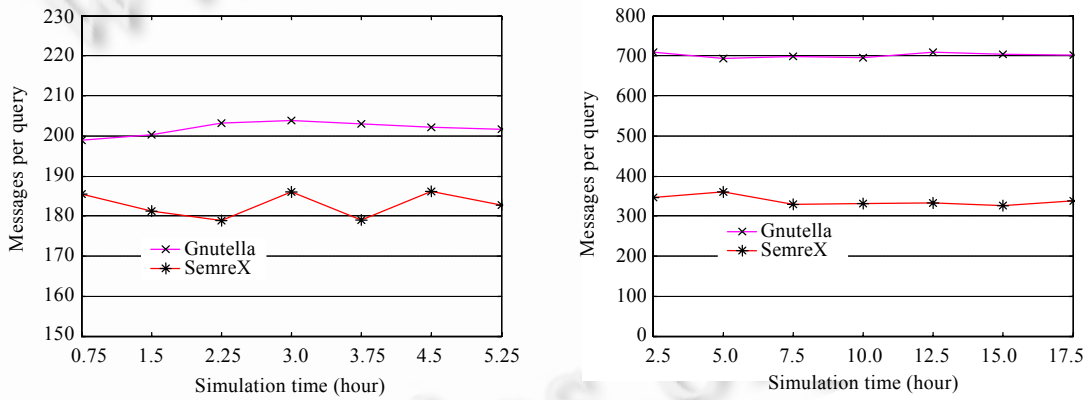


Fig.7 Comparison of recall of different network Scales in SemreX and Gnutella

图 7 SemreX 和 Gnutella 在不同网络规模下的查全率对比

为了更有效地分析语义拓扑的特征,在分析查全率的同时,我们分别在两种拓扑下进行了 Flooding 测试,对比两种网络产生的负载情况.为此,我们测试了查询的平均消息数 *AVG_MSGS_Per_QUERY*,它是查询请求在 P2P 网络上传播时平均产生的消息数.为测试该值,我们设置网络规模 *NO_PEERS*=1000,*TTL*=3 和 *TTL*=4,并比较在两种拓扑的消息数.图 8 显示,即使同样使用 flooding 策略,语义拓扑也能有效降低 P2P 网络的消息负载.

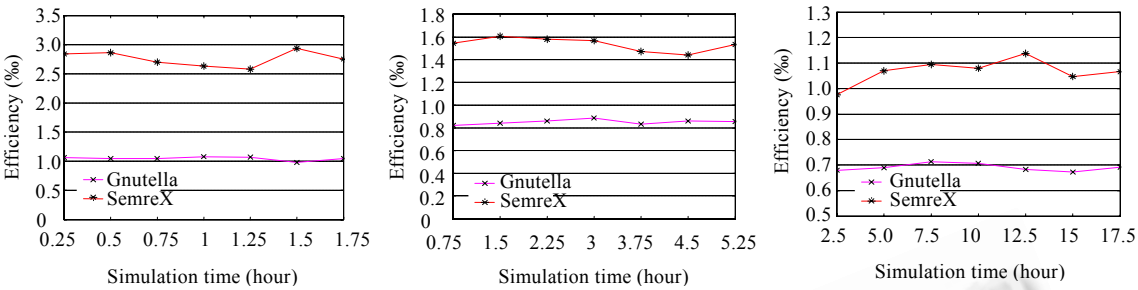


(a) *TTL*=3 (b) *TTL*=4

Fig.8 Comparison of messages in flooding test of SemreX and Gnutella

图 8 SemreX 和 Gnutella 在 flooding 测试中的消息数对比

为了综合考虑查全率和消息负载,查询效率定义为 *Efficiency=Recall/AVG_MSGS_Per_QUERY*,表示单位消息贡献的查全率.图 9 显示了语义拓扑和 Gnutella 在 flooding 情况下的查询效率.图 10 将 SemreX 语义拓扑和 Gnutella 网络在不同网络规模下的查询效率进行了对比.实验结果显示,语义拓扑能有效提高 P2P 系统的查询效率.



(a) *TTL*=2 (b) *TTL*=3 (c) *TTL*=4

Fig.9 Comparison of efficiency in flooding test of SemreX and Gnutella

图 9 SemreX 和 Gnutella 在 flooding 测试中的有效性比较

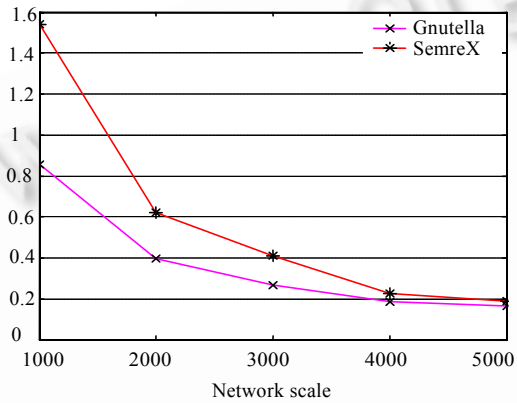


Fig.10 Comparison of search efficiency of different network scales in SemreX and Gnutella

图 10 SemreX 和 Gnutella 在不同网络规模下的查询效率对比

4 相关工作

在 P2P 拓扑和路由方面,许多现有的研究工作集中在提高查询效率上,例如随机游走、基于 DHT 的结构化 P2P.语义拓扑方面的研究工作现在还相当少.Bibster[15]是德国 Karlsruhe 大学组织开发的一个共享 BibTeX 格式参考文献信息的 P2P 系统.网络中具有相同 Topic 的结点建立起连接.与 Bibster 系统不同,SemreX 针对异构类型的文档信息进行提取,SemreX 语义拓扑和路由算法建立在基于统计的相似性度量方法上.

5 总结和未来工作

针对 SemreX 系统 P2P 网络的内容定位问题,我们设计了基于语义相似度的 P2P 拓扑管理算法和查询路由算法.仿真实验结果表明,语义拓扑能有效提高系统查询效率.

下一步,我们将在大规模网络下模拟测试语义路由算法,并进一步分析语义拓扑在大规模网络环境下的网络特性.在前一阶段的测试中,我们发现参数的设置对查询效率影响较大.鉴于单次测试时间开销非常大,且有的参数处在连续空间上,在后面的工作中,我们将使用启发式算法对本文算法中的参数进行优化.

References:

[1] Shen HT, Shu Y, Yu B. Efficient semantic-based content search in P2P network. IEEE Trans. on Knowledge and Data Engineering, 2004,16(7):813−826.

[2] Stoica I, Morris R, Karger D, Kaashoek MF, Balakrishnan H. Chord: A scalable peer-to-peer lookup service for Internet applications. In: Govindan, ed. Proc. of the ACM SIGCOMM 2001. ACM Press, 2001. 149−160.

[3] Ratnasamy S, Francis P, Handley M, Karp R, Shenker S. A scalable content-addressable network. In: Govindan, ed. Proc. of the ACM SIGCOMM 2001. ACM Press, 2001. 162−172.

[4] Castro M, Costa M, Rowstron A. Debunking some myths about structured and unstructured overlays. In: Proc. of the 2nd USENIX Symp. on Networked Systems Design and Implementation (NSDI 2005). USENIX, 2005.

[5] Gkantsidis C, Mihail M, Saberi A. Random walks in peer-to-peer networks. In: Proc. of the IEEE INFOCOM 2004. New York: IEEE Press, 2004. 120−130.

[6] Broekstra J, Kampman A, Harmelen FV. Sesame: A generic architecture for storing and querying RDF and RDF schema. In: Horrocks I, Hendler JA, eds. Proc. of the ISWC 2002. Berlin: Springer-Verlag, 2002. 54−68.

[7] Milojicic DS, Kalogeraki V, Lukose R, Nagaraja K, Pruyne J, Richard B, Rollings S, Xu Z. Peer-to-Peer Computing. Palo Alto: HP Laboratories: Hewlett-Packard Company, 2002. 1−52.

[8] Budanitsky A, Hirst G. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In: Proc. of the Workshop on WordNet and other Lexical Resources. 2001.

[9] Rada R, Mili H, Bicknell E, Blettner M. Development and application of a metric on semantic nets. IEEE Trans. on System, Man, and Cybernetics, 1989,19(1):17−30.

[10] Yuhua L, Bandar ZA, McLean D. An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans. on Knowledge and Data Engineering, 2003,15(4):871−882.

[11] Resnik P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research, 1999,11:95−130.

[12] Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proc. of the Int'l Conf. Research on Computational Linguistics (ROCLING X). 1997.

[13] BatagelJ V, Mrvar A. Pajek—Analysis and visualization of large networks. In: Mutzel P, Jünger M, Leipert S, eds. Proc. of the 9th Int'l Symp. on Graph Drawing. Berlin: Springer-Verlag, 2001. 477−478.

[14] Matei R, Lamnitchi A, Foster I. Mapping the Gnutella network. IEEE Internet Computing, 2002,6(1):50−57.

[15] Haase P, Broekstra J, Ehrig M, Menken M, Mika P, Plechawski M, Pyszlak P, Schnizler B, Siebes R, Staab S, Tempich C. Bibster—A semantic-based bibliographic peer-to-peer system. In: McIlraith SA, Plexousakis D, Harmelen FV, eds. Proc. of the ISWC 2004. Berlin: Springer-Verlag, 2004. 122−136.

陈汉华(1978－),男,湖北武汉人,博士生,主要研究领域为网格计算,对等计算,语义 Web.

袁平鹏(1972－),男,博士,副教授,主要研究领域为 CSCW,网格计算,语义 Web.

金海(1966－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为计算机体系结构,集群计算与网格计算,并行与分布式计算,对等计算,高性能网络存储和并行 I/O,Web 和网络安全,多媒体技术,移动计算,普适计算.

武浩(1979－),男,博士生,主要研究领域为网格计算,对等计算,语义 Web.

宁小敏(1978－),男,博士生,主要研究领域为网格计算,对等计算,语义 Web.

郭志鑫(1977－),男,博士生,讲师,主要研究领域为网格计算,语义 Web.