基于非结构化的 P2P 信息检索关键技术研究

李绍滋,曹 阳,周昌乐
(厦门大学 计算机科学系,福建 厦门 361005)

摘 要:如何在缺少集中控制、大规模、分布式的 P2P(peer-to-peer)网络中找到并定位信息是所有的 P2P 共享系统面临的一个难题.现有的 P2P 信息检索机制存在着种种不足:基于结构化 P2P 网络的检索效率很高,然而由于构造过于严格,难以在 Internet 上普及,而且仅能支持粗粒度的文件共享;非结构化 P2P 网络实现简单,是 P2P 共享系统的主要实现方式,但是由于搜索的盲目性,其检索效率又普遍低下.建立了一个新的非结构化 P2P 共享原型系统.该系统利用改进的蚁群算法进行检索路由,使检索总是倾向于有利的方向.同时,有针对性的推荐服务能够减少盲目搜索,提高文件共享水平.仿真实验的结果表明,该系统所采用的信息检索与信息推荐相结合的策略能够有效地提高 P2P 信息检索的成功率,降低网络负载.

关键词:P2P;信息检索;信息推荐
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-4785(2006)02-0074-05

Research on key techniques about unstructured P2P information retrieval

LI Shao-zi, CAO Yang, ZHOU Chang-le
(Department of Computer Science, Xiamen University, Xiamen 361005, China)

Abstract: How to find and locate information in a decentralized and dynamic network is a big problem for all P2P(peer-to-peer) file-sharing systems. Unfortunately, existing P2P searching mechanisms are usually dissatisfied. For example, structured P2P systems are efficient but lack of actual implementing on the Internet because of their complicated structures. Unstructured P2P systems are inefficient but more popular. In this paper, a new approach to P2P information retrieval based on unstructured P2P systems is presented by using ant colony algorithm and information recommendation services to improve the search efficiency. Ant colony algorithm is used to make routing decisions, which makes the searches tend to the most favorable direction. Besides, information recommendation services can reduce blind searches and raise the file-sharing level. In order to evaluate and validate this model, a simulated P2P application consisted of a network of peer nodes is built. The results show that the searching mechanism has good performances on the search success rate and load balancing.

Keywords: P2P; information retrieval; information recommendation

近年来,P2P(peer-to-peer)技术不仅受到了研究人员越来越多的关注,而且各种 P2P 应用软件也是层出不穷,日渐流行.

信息共享是 P2P 技术流行的重要原因之一,而信息共享的前提是找到并定位信息,即如何检索信息.目前,常见的检索方法包括:以 Napster 为代表集中式的检索机制;以 Gnutella、Freenet[1] 为代表的基于非结构化 P2P 网络的检索;以及以 CAN[2] 和 Chord[3] 为代表基于结构化 P2P 网络的检索.其中,非结构化 P2P 网络中的检索往往由于基于泛洪式的检索方法,检索效率非常低下.文中认为,忽视

收稿日期:2006-09-13.
基金项目:国家自然科学基金资助项目(60373080);福建省自然科学基金资助项目(A0310009);厦门大学 985 二期信息技术创新平台资助项目(2004 - 2007);厦门大学院士启动基金资助项目.

P2P 网络中节点(Peer)的兴趣是造成这一现象的重要原因之一.为此,提出了一种将蚁群算法与推荐服务相结合的 P2P 信息检索方法,使得节点之间针对各自的兴趣互相推荐信息,提高了系统信息共享的程度,进而提高检索效率.

1　相关工作

1.1　蚁群算法思想

仿生学的研究成果表明,蚂蚁个体在搜索食物的运动过程中不但会在其所经过的路径上留下一种被称为信息素的物质,而且还能够感知这种物质的浓度,并以此指导自己的运动方向,即蚂蚁倾向于朝着信息素浓度高的方向移动. 因此,由大量蚂蚁组成的蚁群的集体行为便表现出一种信息正反馈现象:某一路径上走过的蚂蚁越多,后面的蚂蚁选择该路径的概率就越大.受此启发,意大利学者 M. Dorigo[4-6]等人首先提出了通过人工模拟蚂蚁搜索食物的过程对问题进行求解的蚁群算法(ant colony algorithm,ACA).

在 ACA 中,每个蚂蚁根据一定的状态转移规则确定转移方向,t 时刻蚂蚁 k 由位置 i 转移到 j 的概率 P_{ij}^k 为

$$P_{ij}^k=\begin{cases}\dfrac{\tau_{ij}^{\alpha}(t)\eta_{ij}^{\beta}(t)}{\sum_{s\in allowed_k}\tau_{is}^{\alpha}(t)\eta_{is}^{\beta}(t)}, & j\in allowed_k,\\ 0. & \text{otherwise.}\end{cases}\tag{1}$$

式中:$\tau_{ij}(t)$ 表示 t 时刻路径 ij 上残留的信息素强度;$\eta_{ij}(t)$ 表示由 i 到 j 的期望程度;$allowed_k$ 表示蚂蚁 k 下一个可转移位置的集合;α 表示在路径 ij 上残留信息素的重要程度;β 表示启发式因子在蚂蚁选择路径中所起的作用.

随着时间的推移,以前经过的蚂蚁留在路径上的信息素逐渐减弱,而后经过的蚂蚁会留下新的信息素.如果用参数 $1-\rho$ 表示信息素衰减程度,$\Delta\tau_{ij}$ 表示路径 ij 上信息素的增量,则经过 n 个时刻后,各路径上信息素强度根据式(2)调整

$$\tau_{ij}(t+n)=\rho\cdot\tau_{ij}(t)+\Delta\tau_{ij},\rho\in(0,1).\tag{2}$$

近年来的研究结果表明:一方面蚁群算法用于组合优化具有很强的发现较好解的能力,另一方面也存在收敛速度慢、易于停滞的缺点.为此,许多学者提出过改进方法,其中相当一部分是针对信息素策略的改进,如:T. Stutzled[7]等提出了 MMAS(max-min ant system)算法,通过对路径上的信息素进行限制克服停滞问题;黄国锐等[8]提出了基于信息素扩散的蚁群算法加强蚂蚁之间的协作,信息素不仅会被留在蚂蚁经过的路径上,而且还会扩散到一定半径距离内的其他路径上.

1.2　Anthill

Anthill 是由 Montresor A.[9-10]等人提出的以蚁群网络模型为基础的 P2P 应用框架.类似真实蚁群系统,Anthill 中有 2 个比较重要的概念——"巢"和"蚂蚁".巢是提供计算或存储资源的 Peer 节点,互相连接的巢组成了 P2P 网络;蚂蚁是能够在巢之间移动的自主 Agent,并能根据巢提供的资源,蚂蚁可以在巢内执行各种操作,如运算、查询文件、或如同真实的蚂蚁留下信息素一样散发关于其他巢或文件的信息等.

在 Anthill 中,不同的 P2P 应用需要采用不同的蚂蚁算法.由于 Anthill 的目的是为研究人员提供一个设计与分析 P2P 算法的框架,所以并没有规定具体的应用领域,不过为了与现有的 P2P 系统进行比较,A. Montresor 等人利用 Anthill 建立了一个将 Gnutella 和 Freenet 的优点相结合的文件共享系统 Gnutant[10].

2　基于蚁群算法的 P2P 信息检索

目前,P2P 技术应用最多的形式之一是文件共享系统,如何在 P2P 网络中检索信息是这些系统面临的关键问题.蚁群算法为发展 P2P 检索技术提供了一种新的思路(如基于蚁群思想的 Gnutant).受此启发,文中提出了一种新的基于蚁群算法的 P2P 信息检索方法.

2.1　Peer 结构

同以往的 P2P 共享系统相比,该系统中由于引入了信息推荐服务,Peer 节点不仅能够对来自本地和其他节点的查询请求进行信息检索、查询转发,还能利用检索历史预测其他节点的兴趣,从而主动推荐信息. Peer 节点具体的组成模块见图 1.

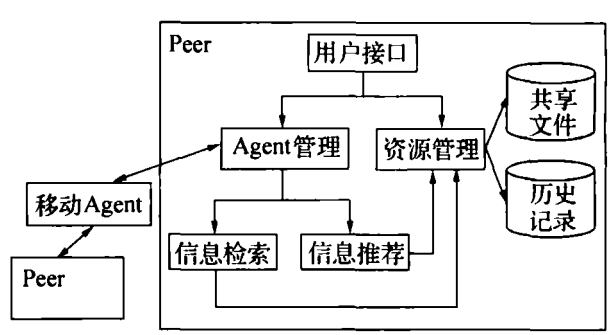

图 1　Peer 节点结构

Fig. 1　Peer node

1)用户接口:帮助用户与 Peer 节点内部各功能模块交互的中间部件.接收本地用户的查询请求,根据用户需求设定参数(如查询关键字的权重、返回结果的最多数目等),将本地或其他节点返回的查询和推荐结果提交用户.

2) Agent 管理模块:本系统利用移动 Agent 技术实现 Peer 节点之间的交互,完成检索和推荐任务.这些 Agent 按照功能分有 3 种类型:用以处理查询请求的检索 Agent,用以返回查询结果的反馈 Agent,以及用以向其他 Peer 节点推荐本地资源的推荐 Agent.

3)信息检索模块:检索本地共享文件以响应查询请求;过滤历史纪录以帮助查询调度模块确定路由选择;对检索结果进行排序.

4)信息推荐模块:根据检索历史,信息检索模块周期性地将一些本地热点资源推荐给其他可能感兴趣的节点,加强 Peer 节点之间的合作,

5)资源管理模块:通过建立字典对存储在 Peer 节点上的两类资源——共享文件和历史记录进行管理.

2.2 基于蚁群算法的 P2P 信息检索

在共享系统中,Peer 节点收到用户提出的查询请求后,首先检查本地文件是否满足查询;如果本地没有适合的文件,Peer 节点将启动网络检索在整个网络中搜索信息.

本地检索时,为了支持语义丰富的信息检索,同时也为了方便对检索结果排序,使用了向量空间模型(vector space model, VSM)描述文件和查询请求,并利用二者之间的相似度判断文件判断是否满足查询.文件 d 表示为 $(w_{1,d}, w_{2,d}, \cdots, w_{id}, \cdots, w_{t,d})$, $w_{i,d}$表示第 i 个关键字 key_i 在文件 d 中的权重.查询请求 q 表示为 $(w_{1,q}, w_{2,q}, \cdots, w_{t,q})$, $w_{i,q}$表示关键字 key_i 在 q 中的权重.那么 d 与 q 的余弦相似度为

$$\mathrm{sim}(d,q) = \frac{\sum_{i=1}^{t} w_{i,d} \cdot w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,d}^2} \cdot \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}. \quad (3)$$

检索本地文件时,如果 $\mathrm{sim}(d, q)$ 大于设定的阈值,则认为 q 满足 d.

网络检索时,利用蚁群算法进行路由选择,检索历史记录就相当于信息素. Peer 节点通过分析检索历史概率确定转移方向,使检索方向倾向于对查询内容感兴趣的节点.

用 $\tau_{ij}(t)$ 表示 t 时刻路径 ij 上与查询请求 q 同类的路由信息强度,$\eta_{ij}(t)$ 表示表示由 i 向 j 转发 q 的期望程度,计算公式分别为

$$\tau_{ij}(t) = \mathrm{Result}_{ij}(q) + \mathrm{Ad}_{ij}(q), \quad (4)$$

$$\eta_{ij}(t) = \frac{C}{d_{ij}}. \quad (5)$$

式中:$\mathrm{Result}_{ij}(q)$ 表示当前有效期内 i 收到来自 j 的与 q 同类的查询结果数量, d_{ij} 表示 i 到 j 的网络延时, $Ad_{ij}(q)$ 表示当前有效期内 i 收到来自 j 的与 q 同类的查询结果数量.网络检索任务由检索 Agent 完成,其算法如下:

1)调用 Peer 节点的信息检索模块对历史记录进行过滤,使用式(3)计算以前收到的查询结果、推荐内容与查询 q 的相似性,确定与 q 相关的邻居节点;

2)利用式(4、5)计算本地节点到各邻居节点的路由信息强度、期望程度,并确定 N 个最近邻居;

3)利用式(1)计算向各邻居节点转移的概率 P_{ij}^{q},确定转移方向,并移动到目标节点;

4)在当前节点进行本地检索,检索完成后,如果发现满足查询请求的文件,则生成反馈 Agent 将结果立即返回初始节点,并通过用户接口将检索结果提交给用户;否则,重复步骤 1)~4)直到跳数达到存活时间(time to live, TTL).

需要指出的是,推荐结果与推荐内容的不同之处在于:推荐结果是对一个独立的查询请求的返回结果,而推荐内容是某一节点对收到的多个查询请求分析后针对某些节点的兴趣主动推荐的信息.

2.3 信息推荐

该系统的一个特色是将信息推荐服务引入非结构化 P2P 共享系统中.这样做的最大好处是,由于信息持有者针对其他节点的兴趣主动推荐相关信息,以后遇到类似的查询请求时,节点就可以直接向相关信息持有者转发查询请求,快速查找到需要的资源.由于加快了路由信息的更新速度,整个系统无需较长的时间就可获得比较满意的检索效率.遗憾的是,包括 Gnutant 在内,现有的 P2P 共享系统都没有这样做.

具体的信息推荐任务由信息推荐模块完成,其算法如下:

1)确定被推荐文件.提供推荐服务的节点周期性的检查本地共享文件,如果上一周期有新文件到来或某一旧文件被频繁检索,则可将该文件选为待推荐文件.

2)确定推荐目标.

①根据被推荐文件的关键字查找相关的历史记录,过滤出对被推荐文件未知的节点(未向其发送过相同的推荐内容或检索结果).

②利用式(6)计算各条路径上推荐强度,由此确定 M 个最有可能对被推荐文件感兴趣的邻居节点.

3)生成推荐 Agent,向推荐目标发送被推荐文件的信息.

用被推荐文件与节点之间的相似度预测节点对该文件的兴趣度:

$$\mathrm{sim}(d,p)=\frac{\sum_{i=1}^{t} w_i \cdot m_i}{\sqrt{\sum_{i=1}^{t} w_i^2}\cdot\sqrt{\sum_{i=1}^{t} m_i^2}}. \quad (6)$$

式中:w_i 表示关键字 key_i 在被推荐文件 d 中的权重,m_i 表示节点 p 向当前节点发送的包含关键字 key_i 的查询请求的次数.

由于根据检索历史进行推荐,一方面可以将推荐服务看成是对以往查询结果的补充.另一方面,由于推荐信息影响了节点检索路由强度的计算方法(见式(4)),这种推荐服务也可以看成是蚁群算法在信息素更新规则方面的改进.

3　实验结果与分析

为了验证文中提出的 P2P 信息检索和信息推荐算法的有效性,通过仿真实验重点对系统的检索效率进行测试,并与 Gnutant 进行对比.

3.1　评价指标

主要采用以下 2 个指标对检索效率进行评价:

1)检索成功率:指 Peer 节点发出的查询消息数和收到的查询成功的消息数之比.成功率越高证明检索机制越有效.

2)带宽利用效率:用满足每个查询请求的平均消息量表示,计算方法见式(7).

$$\text{带宽利用效率}=\frac{\text{一定周期内网络上的消息总量}}{\text{得到满足的查询数目}}. \quad (7)$$

3.2　实验结果与分析

为了使算法分析结果更为准确、客观,进行了多次实验,每一次实验在不同的资源分布和查询请求条件下进行,并根据模拟过程中收集的相关数据计算平均值.实验的数据集采用 A. Montresor 等人从 Gnutella 网络中收集的真实的 10 000 条查询请求.具体的实验参数如下:Peer 节点的总数为 1 000,每个节点的连接度为 6,每个节点初始拥有 5 个资源,TTL 为 10,实验重复进行 10 次,每隔 1 000 条查询请求采样一次.

初始条件下,各节点随机选择 N 个邻居.对于同类别的查询请求,系统经过多次检索后,Gnutant(不使用推荐服务)和本系统(使用推荐服务)对比实验的结果见图 2、图 3.

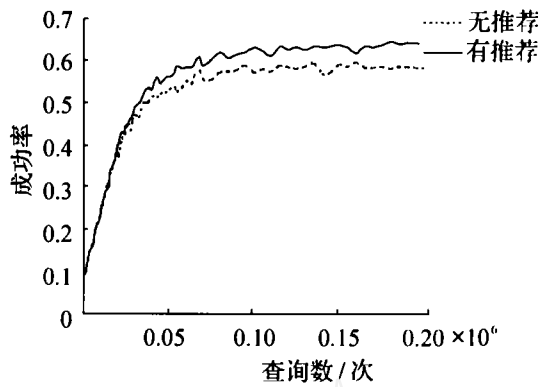

图 2　检索成功率

Fig. 2　Success rate

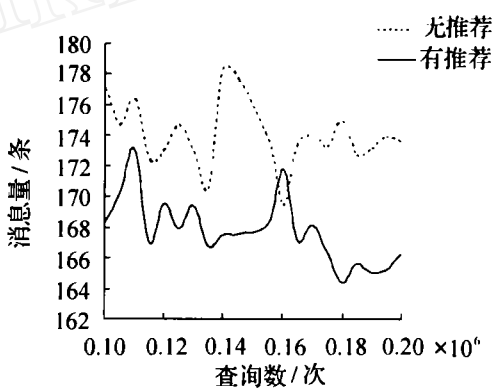

图 3　带宽利用效率

Fig. 3　Average number of message

上述实验结果表明:除了初始阶段外,本系统在成功率方面始终高于 Gnutant.随着查询请求数的增加,成功率逐渐趋于稳定.无推荐服务的 Gnutant 系统的成功率稳定在 58 %左右,本系统采用推荐服务后使得成功率提高到 63 %左右.在带宽利用效率方面,由于信息推荐服务引发的消息量非常少(一次推荐的消息量远小于一次检索过程的消息量),因此并不会增加网络负载.相反,随着系统趋于稳定,满足每个查询请求的平均消息量反而有所降低,如图 3 中 100 000 条查询后,Gnutant 的平均消息量在 174 左右,本系统为 168 左右,这也意味着要达到同样的成功率,本系统总的消息量比 Gnutant 减少 4 %左右.

不难看出,由于引入推荐服务,本系统在检索成

功率和带宽利用效率2个指标上表现均好于Gnutant,检索效率有了明显的改善.

4 结束语

针对P2P文件共享系统中的定位问题,提出了一种基于蚁群算法的P2P检索方法,并将信息推荐引入P2P系统以改善检索效率.仿真结果表明,本文提出的P2P检索方法可以明显改善检索效率.鉴于对P2P信息检索技术的研究还处于初级阶段,而仿真实验毕竟只是对现实的模拟,预设了诸多的限定条件,尚有许多问题需要解决.如何在真实的网络环境中部署、实现本文提出的检索策略,如何实现较好的可扩展性,如何均衡负载、避免热点资源的提供者被过度访问,以及安全性等方面问题将是下一步研究的重点.

参考文献:

[1]CLARKE I, SANDBERG O, WILEY B, et al. Freenet: a distributed anonymous information storage and retrieval system [A]. ICSI Workshop on Design Issues in Anonymity and Unobservability[C]. [S.l.],2000.

[2]RATNASAMY S, FRANCIS P, HANDLEY M, et al. A scalable content-addressable network [A]. In ACM SIGCOMM'01[C]. [S.l.],2001.

[3]DABEK F, et al. Building Peer-to-Peer systems with chord, a distributed lookup service [A]. In: Proc of the 8th Workshop on Hot Topics in Operating Systems (HotOS)[C]. Schloss Elmau, Germany, 2001.

[4]COLORNI A, DRIGO M, MANIEZZO V. Distributed optimization by ant colonies [A]. In: Proc 1st European Conf Artificial Life[C]. Pans, France: Elsevier, 1991.

[5]COLORNI A, DORIGO M, MANIEZZO V. An investigation of some properties of an ant algorithm[A]. In: Proc PPSN'92[C]. London, 1992.

[6]DORIGO M, GAMBARDELLA LM. Ant colony system: a cooperative learning approach to the traveling salesman problem [J]. IEEE Transactions on Evolutionary Computation, 1997, 1(1):53-66.

[7]STUTZLE T, HOOS H H. MAX-MIN ant system and local search for the traveling salesman problem[A]. In: IEEE Int'l Conf on Evolutionary Computation[C]. Indianapolis: IEEE Press, 1997.

[8]黄国锐,曹先彬,王煦法.基于信息素扩散的蚁群算法[J].电子学报,2004,32(5),865-868.
HUANG G R, CAO X B, WANG X F. An ANT colony optimization algorithm based on pheromone diffusion [J]. acta Electronica Sinica, 2004, 32(5):865-868.

[9]MONTRESOR A. Anthill: a framework for the design and analysis of Peer-to-Peer systems [A]. In: Proc 4th European Research Seminar on Advances in Distributed Systems[C]. Bertinoro, Italy, 2001.

[10]BABAOGLU O, MELING H, MONTRESOR A. Anthill: a framework for the development of agent-based peer-to-peer system [A]. In: Proc 22nd ICDCS'02[C]. Vienna, Austria,2002.

作者简介:

李绍滋,男,1963年生,教授,博士生导师,原机械工业部跨世纪学术骨干、河南省首批优秀中青年骨干教师和厦门大学中青年骨干教师,主要研究方向为人工智能与智能信息检索、网络多媒体与CSCW. E-mail: szlig@xmu.edu.cn

曹 阳,女,1980年生,厦门大学计算机科学系硕士研究生,主要研究方向为人工智能、P2P信息检索.

周昌乐,男,1959年生,教授,博士生导师,厦门大学信息科学与技术学院院长.长期从事人工智能及其应用技术、文理交叉学科领域的研究工作,主要研究方向为计算语言学、理论脑科学、智能中医学等.