学校编码：10384　　分类号_____密级_____
学号：200328001　　UDC_____

硕　士　学　位　论　文

基于非结构化的P2P信息检索关键技术研究

Research on Key Technology of Information Retrieval Based on Unstructured P2P Network

曹　阳

指导教师姓名：李 绍 滋　教授
专　业 名 称：计算机应用技术
论文提交日期：2006　年 5　月
论文答辩时间：2006　年　　月
学位授予日期：

答辩委员会主席：_____
评　　阅　　人：_____

2006 年　　月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（　），在　　年解密后适用本授权书。

2、不保密（　）

（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

摘　要

近年来，P2P（peer-to-peer）技术成为人们研究与关注的焦点，以 Napster、MSN、BT 为代表的 P2P 应用软件日渐流行。其中，信息共享是最常见的一种应用。

在P2P共享系统中，每个Peer节点既可以将本地资源贡献出来与其它节点分享，又可以从其它节点获取资源，实现了服务器与客户端的两位一体。然而，所有P2P共享系统都面临一个难题，即如何在缺少集中控制、大规模、分布式的P2P网络中找到并定位信息。遗憾的是，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。

本文在深入研究P2P信息检索技术的基础之上，重点研究基于非结构化P2P网络的信息检索技术，建立了一个新的、能够适应不同粒度的检索要求的非结构化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，通过有针对性的推荐服务，加强Peer节点之间的协作关系，改善P2P检索效率。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、查全率，降低网络负载。

关键词：P2P；信息检索；信息推荐

Abstract

Peer-to-peer (P2P) technology has recently received a lot of attentions, and many P2P platforms have been developed, such as Napster, MSN, BT, etc.

One of the most popular P2P applications is the file-sharing system. In a P2P file-sharing system, each peer performs as both a server and a client, which means they can get resources from remote sites, meanwhile, they are also providing local resources for other peers. However, how to find and locate information in a decentralized and dynamic network is a big problem for all P2P file-sharing systems.

Unfortunately, existing P2P searching mechanisms are usually dissatisfied. For example, structured P2P systems are efficient but lack of actual implements on the Internet, because of their complicated structures. Unstructured P2P systems are inefficient but more popular.

From various perspectives, our work focuses on how to improve retrieval efficiency of unstructured P2P file-sharing systems. In this paper, we present a new approach to P2P information retrieval, using ant colony algorithm and information recommendation services to improve the search efficiency. Ant colony algorithm was used to make routing decisions and it made searches tend to the most favorable direction. Besides, information recommendation services could raise the file-sharing level and reduce blind searches.

In order to evaluate and validate our model, we built a simulated P2P application consist of a network of peer nodes; mobile agent travel through the network, making peer nodes communicate with each other. The results have shown that our searching mechanism has good performances on the search success rate, recall rate, and load balancing.

Key Words: P2P; Information Retrieval; Information Recommendation

目　录

Contents

第一章 绪论

1.1. 背景

1.1.1. P2P的含义

近年来，随着互联网技术的普及和发展，越来越多的机器获得了网络连接。与此同时，计算资源的价格不断下降，性能却在迅速提高。如今，即使是一台普通的个人计算机也具备了相当的服务能力。尽管个人用户无法像专业服务商那样提供大规模的服务能力，但是如果把网络上数量巨大的个人计算机作为一个整体联系起来，就可以提供任何集中式服务器无法比拟的计算资源。正是基于这样的思想，蕴含着巨大的商业和技术价值的P2P（peer-to-peer）技术受到了研究人员越来越多的关注，而以Napster[1]为代表的各种P2P应用软件更是层出不穷，日渐流行。

P2P可以理解为“伙伴对伙伴”的意思，也有些人将其形象的写成“Person-to-Person”。P2P模式使得网络上的主机处于同等地位，每台主机被称为一个“对等点”(peer)。各对等点相互联结，实现了直接的共享和交互而无需经过中间服务器。每个对等点可以随时自由地加入和离开系统，形成一个真正动态的网络环境。根据文献，P2P有两个层面的基本含义[2]：

- P2P通信模式。这种模式区别于传统的客户机/服务器(Client/Sever，C/S)或者主/从（Master/Slave）模式，每个通信方都具有相同的能力，并且每个通信方都可以发起一个通信过程。
- P2P网络。P2P网络是运行在互联网上的动态变化的逻辑网络。这个网络是由一些运行同一个网络程序的客户端彼此互连而构成的，客户端彼此间可以直接访问存储在对方驱动器上的文件。

严格地说，P2P并不是一个全新的概念，P2P是互联网整体架构的基础。互联网最基本的协议TCP/IP并没有客户机和服务器的概念，所有的设备都是通讯的平等的一端。即便是架构在TCP/IP之上的软件的确采用了C/S的结构——浏览

器和Web服务器，邮件客户端和邮件服务器，对于服务器来说，它们之间仍然是对等联网的，比如，邮件服务器相互协作把电子邮件传送到相应的服务器上去。P2P的意义在于把这种“对等联网”理念拓展到整个互联网范围，改变目前互联网以大网站为中心的格局、重返“非中心化”，把权力交还给用户。

1.1.2. P2P的优势

P2P非中心化的基本特点使其具备了以下优点：

- 可扩展性好。在P2P系统中，每当新用户加入，新的资源也随之加入，系统整体的资源和服务能力也就同步地扩充，从而始终能较容易地满足用户的需要。
- 容错性好。由于P2P系统是分散化的，单个节点故障对整个系统的影响很小。并且，P2P网络一般在部分节点失效时自动调整整体拓扑，保持其它节点的连通性。这就消除了C/S系统由于单一访问点造成的瓶颈和单点故障问题。
- 成本低、充分利用分布资源。P2P系统汇聚了来自因特网边缘的大量低廉设备的空闲资源，不仅达到高性能计算和海量存储的目的，而且减少系统的成本。
- 提高匿名性，保护用户隐私。在P2P系统中，信息的传输分散在各节点之间而无需经过某个集中环节，用户的隐私信息被窃听和泄漏的可能性大大缩小。此外，由于所有参与者都可以提供中继转发的功能，因此可以将通信的参与者隐藏在众多的网络实体之中，大大提高了匿名通讯的灵活性和可靠性。

1.1.3. P2P的应用

目前，P2P技术主要应用在以下领域：

- 信息资源共享——P2P技术最典型的应用。利用基于分布式计算的P2P技术，可以方便地组织和存储信息资源，对等节点通过不同的查询机制

定位含有所需资源的对等节点后，直接与其建立连接获取所需要的信息资源。典型的系统包括Napster、Gnutella[3]、Freenet[4]、Free Haven[5]、Ohaha[6]等。

- 普适计算（Pervasive Computing）——充分利用网络中各种的计算单元来共同完成大规模的计算任务。P2P普适计算整合了互联网中闲散的计算资源，利用整个网络上可得的CPU周期来完成超计算密集型应用。典型的系统如SETI@home [7]、Distributed.net[8]等。
- 协同工作——多个用户之间利用网络中的协同计算平台互相协同来共同完成计算任务，共享信息资源等。较之传统的基于C/S和Web的协作方式，因为采用文件直接共享的方式，P2P协作平台既保证了协作的实时性，也节省了对单点服务器存储以及性能的要求。典型的系统如Groove [9]。
- 实时通信技术。以ICQ、MSN、Yahoo Messenger等为代表的实时通信产品已经被广大用户所接受和使用。这些产品使网络中的用户可以通过直接连接来实现文字、语音和视频聊天。
- 搜索引擎。P2P技术使用户能够深度搜索文档，而且这种搜索无需通过Web服务器，也可以不受信息文档格式和宿主设备的限制，可达到传统目录式搜索引擎无可比拟的深度。

1.1.4. P2P技术面临的问题

P2P系统本质上也是一个分布式系统，但是较之传统分布式系统更强调自组织、对等、动态。P2P系统的这些特性使其面临以下的技术问题[10，11] 。

- 网络拓扑结构

拓扑结构是指分布式系统中各个计算单元之间的物理或逻辑的互联关系。节点之间的拓扑结构一直是确定系统类型的重要依据。目前互连网络中广泛使用集中式、层次式等拓扑结构。Internet本身是世界上最大的非集中式的互联网络，但是九十年代所建立的一些网络应用系统却是完全集中式的系统，很多Web

应用都是运行在集中式的服务器系统上。集中式拓扑结构系统目前面临着过量存储负载、Dos攻击等一些难以解决的问题。层次式拓扑结构是一种应用比较广泛的分布式拓扑结构，DNS系统是其最典型的应用。P2P系统一般要构造一个非集中式的拓扑结构，在构造过程中需要解决系统中所包含的大量节点如何命名、组织以及确定节点的加入、离开方式、出错恢复等问题。

- 资源定位和路由机制

在典型的P2P网络中数据资源分布在各个独立的节点上，如何高效地索引、查找、定位以及访问这些数据信息资源是另一个需要关注的重要问题，在分布式系统中这些问题同样也是正在研究的热点问题。URL是目前在Web上使用最普遍的信息定位策略，DNS则提供了一套层次式的查找机制。一般来说，在P2P共享应用中所采用的检索方式是利用关键字来查询自己所需的信息资源，同时人们也期望能够将数据资源的索引信息存放在系统中的每一个节点上。路由机制是指节点之间通信的消息传递路径，合适的路由机制可以充分的利用网络带宽资源并使系统具有很好的容错性、可扩展性。目前，很多系统中的路由机制都是和这些系统的逻辑拓扑结构紧密相关的。在数据的访问过程中则期望能够采用流水、并行或者选择传输路径的方式来加快数据的访问速度。

- 异构网络环境的互操作性和扩展性

P2P网络连接了各种自治资源和系统，它需要考虑如何屏蔽操作系统、网络协议的异构性和复杂性，使分布在网络上的不同机器能够相互传递消息协同工作。P2P网络形成的初期，计算规模较小。随着大量计算单元的不断加入，系统的资源规模也随之扩大。因此需要考虑在资源规模不断扩大、应用不断增长的情况下系统的可扩展性，不降低网络的整体性能。

- 元数据组织与表示

P2P网络面向的是异构网络与操作系统，需要在这些系统之间交换数据资源，但是因为这些系统的数据表示并不都是完全相同的，这就需要一个能够在多个系统之间确定一个通用的元数据表示方案。关于元数据的组织包括数据资源的表示、消息通信协议等，很多系统都支持SOAP或者XML-RPC等协议。

- P2P网络的支撑技术

Internet 技术的发展使得连入互联网络中的设备不再局限于计算机，在P2P的计算环境中要求任何设备都可以在任何地点很容易的加入到这个环境中。所谓的计算设备既包括有线设备也包括无线设备，这样就需要很多很多网络传输的支撑技术来支持各种不同设备连入整个P2P网络。

- 安全问题

安全问题是一直伴随着互联网发展的重要问题，安全问题包括很多相关的问题，比如应该防止他人控制整个系统、增加恶意信息等，同时系统应能够保证系统中信息资源的正确性。在P2P系统中，系统安全同样面临着巨大的挑战。P2P系统需要在没有中心节点的情况下，提供身份的认证、授权以及数据信息的安全存储、数字签名、加密、安全传输等工具，同时P2P系统要有能力抵抗过量存储负载、Dos攻击等攻击行为。

1.2. P2P信息检索的优势

信息共享是P2P技术流行的最重要的原因之一，然而信息共享的前提是找到并定位信息。目前人们在网络中搜索信息的主要工具是搜索引擎。但是，常见的搜索引擎如Google、Yahoo、baidu等都是集中式的搜索引擎，远远无法涵盖所有互联网上的共享内容。较之集中式的搜索引擎，P2P信息检索在检索的即时性、检索深度等方面具有明显的优势。

- 保证搜索信息的即时性。传统的搜索引擎，搜索的并非实际的内容，而是预先通过“网络蜘蛛（Spider）”或者其他工具形成的索引。一方面，由于网络规模的庞大，基于网络蜘蛛的传统搜索引擎的索引刷新周期较长；另一方面，如今互联网上的信息具有动态变化、周期短的特征。这就导致传统的集中式引擎无法胜任这种实时性强的海量信息检索。P2P信息检索则不同，当请求消息到来时，将在本地机器上搜索共享的内容，并返回查询的结果，因此，返回的信息都是即时的。
- 弥补传统搜索引擎无力深度挖掘网站信息的弱点。传统的搜索引擎，主要检索的是网络中的静态信息，HTML页面以及静态的主页等。而网站

数据库中相当大部分的信息是以动态网页的形式来提供的，因此搜索引擎仅能检索到网络上很小一部分的信息。P2P检索提供了一种不同的方式：各个信息提供者作为一个节点加入P2P共享网络，各个节点分别对自己本机上存储的信息制作索引，所有的信息提供者一起构成一个庞大的分布式数据库以供检索。P2P技术使用户能够深度搜索文档，而且这种搜索无需通过Web服务器，也可以不受信息文档格式和宿主设备的限制，可达到传统目录式搜索引擎无可比拟的深度。

- 挖掘移动终端的信息。随着3G的到来，智能手机、智能终端的功能不断加强。这些移动终端存储的数据具有分布面广、地域性强、存储信息和用户终端密切相关的特点，这些特点在互联网的P2P共享网络中也同样存在，可以看成是P2P共享网络的一个自然延伸。充分挖掘这些分布的信息，并使得信息在各个用户群体内部流通顺畅，具有相当大的实用价值和研究意义。P2P信息检索十分适合于解决这些实际的问题。
- 构建人性化的信息终端。在P2P网络中分布于各个终端的数据非常直接而深刻地反映了用户的兴趣。在对这些兴趣进行挖掘的基础上，可以很方便地对产品和业务进行个性化推送，组织协作，研究和开发更加富有人性化的信息终端。这无论对提升企业内部信息流通的质量还是对改善公众网络信息服务都有十分重大的意义。

1.3. P2P信息检索的难点

在P2P网络环境中检索信息面临着种种难题。

首先，P2P环境下文档的分布和节点的分布往往不一致，资源定位尤其是稀疏资源的定位非常困难。在传统P2P网络中(如Gnutella)，由于资源随机分布在网络中的节点上，检索过程往往需要遍历比较多的节点才能获得较高的查全率，导致网络承受很大的带宽压力。因此，保证高查全率的前提下，要尽量把检索请求路由到所有可能的答案提供节点，而尽可能不把查询转发给不相干的节点。

其次，如何均衡负载是另一个难题。在网络中存储热点内容的节点将会被