P2P 网络搜索技术①

王 婕, 王亚美, 廖 婧, 赵婧文

(中国地质大学 软件工程系, 武汉 430074)

摘 要: 随着 P2P 技术的蓬勃发展, 作为 P2P 应用中核心的搜索技术成为研究人员关注的焦点. P2P 网络的搜索技术与其结构有着密切联系, 不同网络体系结构下的搜索技术各不相同. 介绍了 P2P 技术近几年的研究进展,阐述了目前 P2P 系统中不同结构下核心搜索算法, 探讨了 P2P 搜索技术的发展方向.

关键词: P2P; 搜索; 体系结构; 原理

Search Technology of P2P Network Research

WANG Jie, WANG Ya-Mei, LIAO Jing, ZHAO Jing-Wen

(Information Engineering, China University of Geosciences, Wuhan 430074, China)

Abstract: With the rapid development of P2P technology, the P2P Search that is the key technology of P2P applications has become the focus of researches. The P2P search technology its structure and it differs from each other under different network architectures. The research development of P2P was introduced in this paper, and further different search algorithms explored the development and direction of the P2P Search technology.

Key words: P2P; search technology; architectures; principle

1 引言

P2P 又称对等网络, 由一系列地位对等的结点组成, 结点数目可以动态的增加和减少[1]. P2P 网络中结点相互之间直接交换信息和服务, 没有等级、格式、平台的限制. P2P 技术改变了传统的 C/S(客户/服务器)模式, 每一个 P2P 结点既是服务器端, 又是客户端, 被财富杂志列为影响 Internet 未来的四项科技[2]. 在传统的 Web 搜索中, 当用户发出搜索命令后, Web 搜索引擎搜索预先整理好的网页索引数据库, 而在 P2P 网络中, 资源存放在各个结点的 PC 机上, 结点的动态变化给 P2P 网络搜索增加了复杂性.

2 P2P网络体系结构下搜索技术的发展

1998 年, 美国一名大一的新生 Shawn Fanning 为了实现 MP3 音乐共享功能, 编写了一个程序, 这个程序就是后来风靡全球的 Napster, Napster 运用了第一代 P2P 网络—集中式 P2P 网络. Napster 系统采用一个中央的目录服务器, 该服务器不对外提供任何应用服务, 仅存储连接该服务器的各个结点的相关信息, 随着系统的使用, 人们发现集中式 P2P 网络体系的不足, 如果目录服务器瘫痪, 整个系统都会崩溃, 而且当用户数量增加到一定数量后, 系统性能会大大降低, 所以第二代 P2P 网络—全分布式拓扑结构应运而生, 分布式 P2P 网络结构, 它包括两种类型, 一种是全分布式结构化的 P2P 网络, 另一种是全分布式非结构化的 P2P 网络. 分布式结构化的 P2P 网络主要采用分布式哈希表(DHT)技术来组织网络中的结点, 采用完全随机图的组织方式, 分布式非结构化的 P2P 网络最典型的案例是 Gnutella. 随着 P2P 技术的继续发展, 研究者们将集中式 P2P 的快速查找和分布式 P2P 的去中心化优势结合起来, 便形成了一种混合式的 P2P 网络结构, 即第三代 P2P 网络—半分布式网络.

① 收稿时间:2012-06-01;收到修改稿时间:2012-08-08

3 集中式P2P网络搜索技术

3.1 集中式 P2P 网络搜索原理

集中式 P2P 网络搜索方法中, P2P 结点都与已知地址的P2P 目录服务器相连, 服务器负责对P2P 网络中的共享文件进行索引和查询, 服务器集中存放对等节点的地址信息和所保存数据的信息. 当结点资源发生变化时, 比如增加、删除、修改等, P2P 节点服务器会随之更新系统索引表[3], 如图 1.

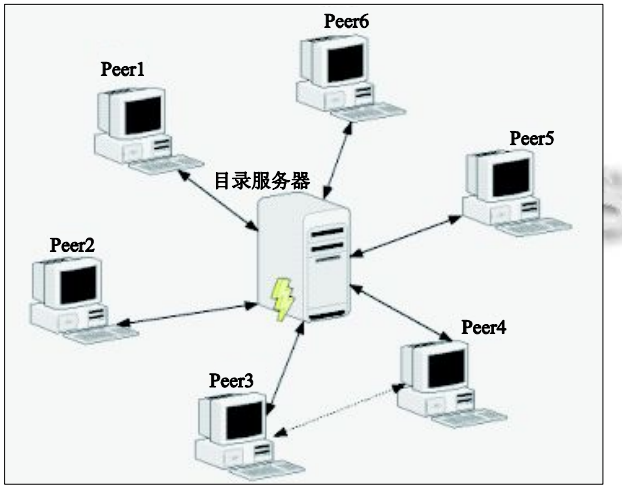

图 1 集中式搜索

3.2 集中式 P2P 网络搜索过程

当查询事件触发时, Peer 结点根据 P2P 目录服务器中的信息进行查询, 通过目录服务器来间接定位其他对等点, 如图 2 中 Peer3 和 Peer4 的通信就是通过目录服务器的媒介作用来完成的. 用于 Mp3 文件共享的 Napster 是集中式 P2P 搜索最具有典型的代表, Napster 系统的目录服务器存储所有该网络的结点的数据信息, 比如结点的 IP 地址, 文件的标题等. 当需要查询某个文件时, 结点向目录服务器发出查询请求, 服务器进行相应的检索和查询, 会返回符合查询条件的结点地址信息列表, 查询发起的结点接收到应答后, 选择最佳的结点与之建立连接, 这样两个结点之间实现文件传输, 完成搜索过程. 集中式网络搜索结构简单, 查询效率高, 速度快, 不足之处在于中央目录服务器负担重, 安全性低.

4 全分布式P2P搜索技术

4.1 全分布式结构化 P2P 网络搜索技术

(1) 全分布式结构化的搜索原理

全分布式结构化基于分布式哈希表(DHT)进行搜索, DHT 中存储形如<关键值, 数值>(< key, value >)的分布式结构, key 代表数据标识, value 代表数据的信息, 比如结点的 IP 地址等[4], 每个结点负责管理一段范围内 keys. 搜索功能主要由 put(key, value)和 get(key)两个函数实现[5], put(key, value)的作用是发布结点信息, get(key)的作用是查询信息, 当需要对 P2P 系统进行文件搜索时, 执行一次 get(key)功能, 便可进行一次搜索. 任何一个关键值 key, 系统中的结点要么拥有 key, 要么能够连接到距离 key 较近的结点.

(2) 全分布式结构化搜索过程

首先定义在分布式哈希表中的一个文件, 名称为 file, 内容为 value, 计算出该文件的 SHA-1 的哈希值, 得到其关键值 key, 执行 put(key, value)操作; 然后在哈希表中找到负责存储关键值 k 的结点, 将(key, value)存储在该结点上; 当其他结点请求 value 时, 系统第二次计算 file 的 key 值, 然后执行 get(key), 发送信息给结构中的任意参与结点, 找到与 key 相关的信息; 最后, 此信息在网络中被传送到负责存储 key 的结点, 此结点收到信息后, 将 value 值传送给请求结点, 完成搜索查询过程. 全分布式结构化搜索优点是结点的自组织能力强, 有良好的可扩展性、鲁棒性, 结点 ID 分配的均匀性, 缺点是服务质量不高, 易拥塞, 安全性低, 不能支持多关键查询, 维护机制复杂.

4.2 全分布式非结构化 P2P 网络搜索技术

(1) 全分布式非结构化搜索原理

在全分布式非结构化网络中, 搜索方法采用泛洪(Flooding)搜索. Flooding 算法首先遍历自己的相邻结点, 然后再层次性的一层层向下遍历, 在遍历过程中, 一个结点向所有邻居结点广播查询消息, 邻居结点再向自己的邻居结点广播, 这个过程不断进行下去. 为了限制搜索的范围, 消息被设置了一个初始的 TTL(Time To Live)值, 消息每经过一个结点, TTL 值减 1, 当 TTL 值为 0 时, 搜索过程结束[6].

(2) 分布式非结构化搜索改进

泛洪算法的算法机制导致了大量冗余消息的存在, 使网络流量增加快速, 从而导致网络中部分低带宽结点失效, 查询结果正确性不高, 所以 P2P 研究者在此搜索算法上进行改进, 产生了 Random Walk[7], 迭代递增搜索, 启发式洪泛搜索等算法.

(3) Random Walk 搜索

Random Walk 搜索也叫随机漫步搜索, 在这个搜

索中，请求者发出 N 个查询请求给随机挑选的 N 个相邻结点，在以后的查询过程中，每个查询信息都直接和请求者保持联系，当得到请求者继续下一步的同意后，又开始进行下一轮的漫步，直到找到要搜索的信息为止，若请求者不同意继续，搜索中止,如图 2.

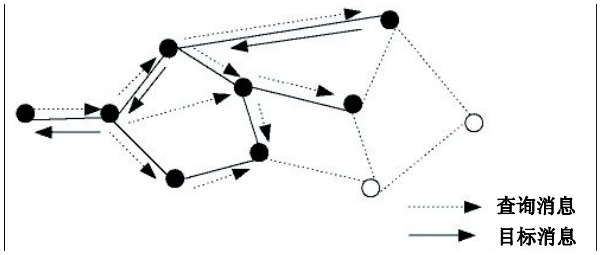

图 2 Random Walk 搜索

与前面的 Flooding 搜索相比，Random Walk 搜索对结点信息的搜索范围有更强的控制性，搜索范围的灵活性也增加了. 全分布式非结构化查询容错性好，支持复杂查询,受结点的动态变化影响小，但是查询速度慢，结果可靠性不高，带宽消耗大，可扩展性不好.

5 半分布式P2P网络搜索技术

5.1 半分布式 P2P 网络搜索原理

在半分布式 P2P 网络搜索是指在搜索过程中，运用了两种或两种以上的搜索技术进行混合搜索的方法，这种网络结构中包含两类结点，一类是搜索结点，另一类是普通结点，搜索结点和其临近的普通结点之间形成一个集中目录式的结构体，如图 3.

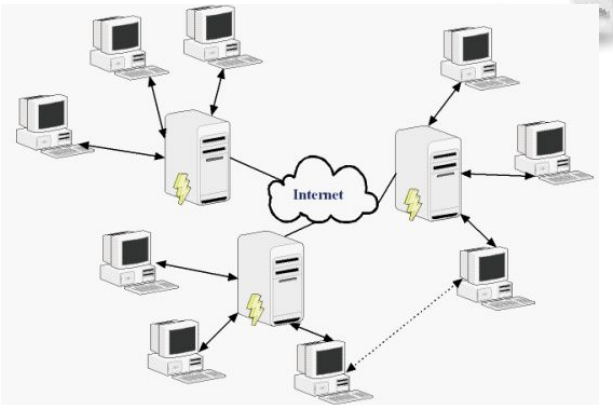

图 3 半分布式 P2P 搜索

4.2 Geutella2 的搜索算法

Geutella2 是半分布式 P2P 网络搜索的代表，该网络结构的搜索结点中存储中与之临近的普通结点的信息，同时搜索结点之间相互连通. 当普通结点需要查询文件时，首先从与它连接的搜索的索引中寻找，如果找到文件，则直接和具有该文件的结点建立连接，否则搜索结点把该查询请求发给与它连接的其他搜索结点，直到搜索成功. 半分布式 P2P 网络搜索消除了网络阻塞，搜索效率低等问题，提高了网络的负载均衡性，但是对搜索结点依赖性大，易于受到集中攻击，容错性不好.

6 总结

本文针对不同 P2P 网络结构的搜索技术进行总结分析，得出以下结论，如表 1，从表中可以看出，集中式网络结构的的可维护性，搜索效率是最好的，全分布式结构化总体的性能较高，全分布式非结构化的优势在于可扩展性，支持复杂查询，半分布式网络结构虽然混合使用了集中式搜索和分布式搜索，但是性能总体是中.

表 1 P2P 网络结构综合性能对比表

比较标准/网络结构	集中式网络结构	全分布式结构化网络结构	全分布式非结构化网络结构	半分布式网络结构
可扩展性	差	好	差	中
可靠性	差	好	好	中
可维护性	最好	好	最好	中
发现算法的效率	最高	高	中	中
复杂查询	支持	不支持	支持	支持

7 P2P网络搜索展望

如今 P2P 的搜索技术研究不仅仅处于可行性研究阶段，而是以提高搜索成功率，缩短搜索时间为目标，综合带宽节约、负载均衡等性能要求，研究出更专业化、个性化、智能化的搜索算法. 所以未来的 P2P 搜索研究，可以从以下方面进行考虑:

在全分布式网络结构下，如何实现多条件的复杂查询;

在全分布式非结构化网络结构中，用什么网络模型来改进算法;

在半分布式网络结构中提高混合后的算法效率;

研究兴趣网络，探究搜索优化算法[8].

(下转第 47 页)

地址; 流通环节监控数据包括: 产品条件参数(入库人、出库人、保质期、温度、湿度、干燥度、入库时间)和实时监控数值(入库人、出库人、保质期、温度、湿度、干燥度、记录时间); 仓储厂家信息包括: 厂家名称、联系人、联系方式、网址、地址. 以种子溯源防伪为例: 首先每包种子从生产厂家一出厂, 便给每一小包装贴上二维码防伪标签, 给大箱包装贴上 RFID 标签, 同时种子从运输到销售的整个过程中, 通过各种传感器来实时监测流通环节监控数据, 以此来跟踪种子的生产、物流、仓储、批发及零售等各个环节. 具体来说, 本平台以 RFID 为主要信息载体, 依托网络通信、系统集成及数据库应用等技术, 建立了农资产品安全信息追溯平台, 实现了农资产品从生产到销售等环节于一体的信息化平台, 确保农资产品“来可追溯、去可跟踪、信息可保存、责任可追查、产品可召回”.

2) 农资调度

农资调度子系统是为实现农资的合理化调度而开发的信息平台, 主要是根据对全国各地农资的需求而实行最优路径规划的调度配送, 同时还提供了对农资商品运输过程的实时追踪功能, 即通过 GPS 定位, 在 GIS 地图上实时显示, 并通过车载摄像头提供实时运输视频. 首先平台集成了全国各地农业信息数据, 包括来自全国各地的农资生产厂商、农资仓储中心、农资分销商和农民合作社的数据信息, 以便在调度过程中根据需求的不同选择路程最短、时间最优、费用最低和综合最优等路径规划算法时, 对数据进行有效的计算和存储, 来达到对农资智能调度的目的. 最终实现了农资流通的有效利用与共享, 并可达到最优服务的目的.

3) 农资知识服务

利用现有的农业知识数据并对数据进行异构融合与挖掘, 为广大农民、农业合作社及农资企业等提供农资产品供求智能对接、广告精准投放与个性化农资技术知识推送等全方位的农资增值信息服务.

4 结语

农资产品是农业生产的主要投入品, 是农业生产的重要物质基础, 其质量优劣直接关系着农业生产和农产品质量安全. 因此, 对农资产品进行溯源防伪是农业发展的保障. 本文对应用于农资产品溯源服务的物联网技术及其体系架构进行了分析, 展示了具体的应用实例. 通过分析可知物联网在农资方面的推广应用将会给农业生产领域带来一次全新的改革, 对增强食品安全, 提高农民收入具有重要的现实意义.

参考文献

1 International Telecommunication Union UIT. ITU Internet Reports 2005: The Internet of Things.2005.
2 Weber RH. Internet of things – Need for a new legal environment. Computer Law and Security Report, 2009, 25(6): 522–524.
3 施亮,傅泽田,张领先.基于 RFID 技术的肉牛养殖安全可追溯系统研究.计算机应用与软件,2010,27(1):40–43.
4 祝胜林,黄显会,张守全.大型猪场基于 RFID 的信息平台研究与应用.广东农业科学,2007,(3):63–64.
5 谢菊芳,陆昌华,李保明.基于.NET 架构的安全猪肉全程可追溯系统实现.农业工程学报,2006,22(5):218–220.
6 阎敏杰,夏宁,万忠,段洪洋.物联网在现代农业中的应用.中国农学通报,2011,27(8):464–467.
7 王保云.物联网技术研究综述.电子测量与仪器学报,2009,23(12):1–7.
8 张捍东,朱林.物联网中的 RFID 技术及物联网的构建.计算机技术与发展,2011,21(5):56–59.

(上接第 15 页)

参考文献

1 幸冬梅,朱洪.P2P 结构与搜索机制研究.计算机工程与科学,2007,29(10):108–110.
2 熊仕勇.基于 P2P 网络的搜索算法研究.科技创新导报,2010,27:35.
3 韩运宝,戚建勋.P2P 网络搜索技术的研究现状.计算机与信息技术,2007,16:316.
4 刘维光,陈立伟.一种基于 DHT 的 P2P 搜索方法.网络与通信,2006,22(3):131–133.
5 管磊,等.P2P 技术揭秘.北京:清华大学出版社,2011.
6 林鹏程,李文正.基于混合式 P2P 架构的资源搜索机制研究.科技咨询导报,2007,10:39–43.
7 欧阳柏成.非结构化 P2P 中搜索算法的性能分析.计算机工程与科学,2009,31(6):67–70.
8 吴思,欧阳松.基于兴趣相关度的 P2P 网络搜索优化算法.计算机工程,2008(6):102–107.