一种基于非结构化对等网络的改进搜索算法①

An Improved Search Algorithm Based on Unstrntured P2P

张 伟 欧阳松 (中南大学 信息科学与工程学院 湖南 长沙 410083)

摘 要: 目前非结构化的 P2P 网络系统中，一般以广播方式作为其搜索的基本策略，引发较大的网络流量。因此，提出改进的搜索策略，根据历史查询记录，为每个节点建立朋友节点，同时又在搜索过程中把节点划分为超级节点和普通节点。实验表明改进算法提高了搜索效率，同时减少了网络信息流量。

关键词: P2P 非结构化 搜索

1 引言

计算机对等网络(Peer-to-Peer network，P2P)技术是目前流行于国际计算机网络技术研究领域的一个热点。P2P 网络模型是一种新型的体现结构模型[1]，其许多优势有待于进一步挖掘。P2P 被认为是未来重构分布式体现结构的关键技术，已经引起了越来越多网络用户的关注。而这种系统的一个核心技术就是搜索[2]算法。

目前的 P2P 网络按照搜索机制可分为结构化 P2P 网络和非结构化 P2P 网络两大类。结构化 P2P 网络对节点进行了有效地组织，能得到较高的搜索效率，但是因为结构化 P2P 网络对结构要求过于严格，更新代价过大，仅适用于小规模的 P2P 应用，难以在 Internet 上普及。而非结构化 P2P 网络因其简单和健壮性而获得广泛应用，引发了许多致力于非结构化 P2P 网络搜索算法改进的研究。

2 相关工作

非结构化 P2P 网络模型按照拓扑结构可以分为三类：集中式的(以 Napster[3]为代表)、完全分布式的(以 Gnutella[4]为代表)、混合式的(以 Kazaa[5]为代表)。集中模型中的搜索系统使用特定的节点（目录节点）来存储 P2P 系统中资源的目录。用户进行搜索时，会通过查询目录节点来找出拥有目标资源的节点地址信息。集中式搜索机制的优点是查找效率高，开销小；但是存在单点故障和瓶颈问题，因此可靠性差。完全分布式模型中没有目录节点，所有的节点都是平等的，查询在节点间广播直到匹配的资源被找到。虽然这种搜索机制简单而且具有鲁棒性，但是每产生一次查询，都会造成巨大的网络开销。混合模型中选择少数节点作为超节点，由超节点作为局部的集中式服务器来保存部分 P2P 节点的共享资源信息，超级节点构成完全分布式模型。混合模型综合了集中模型和分散模型的特点，吸取了两者的优点。但是要构建高效的分散模型，需要考虑网络的规模和超级节点的选取。

在 Gnutella 搜索机制的基础上提出一种新的搜索算法，利用节点自身兴趣[6]，逐步“聚集”具有相似兴趣的朋友节点，同时又能利用文档资源在节点间的分布信息，动态建立“超级节点”，在大幅度降低网络带宽消耗的同时保证了查全率，从而提高了系统的搜索效率。

3 改进的搜索算法

经大量统计发现，节点存储的信息资源基本上能反映节点的兴趣，同时兴趣相似的节点更有可能存储相似的信息资源。如果它们能进行直接的连接查询，不仅查找效率高，而且信息流通量小，网络带宽消耗小。

根据这个思想，提出一个建立在无结构对等网络之上、基于兴趣节点的网络搜索模型。在此网络模型中，查询节点可以依据各自的兴趣建立与其它节点之

① 收稿时间:2008-05-31

间的链接，因此为每个节点建立朋友节点。若通过朋友节点查询的资源不能满足需求时，查询转发到有较大可能性应答查询请求的节点。在这里引入超级节点的概念。如果在一个节点上成功查询的次数超过了一个设定的阈值，这个节点就被定义为超级节点。超级节点一般来说是数据资源丰富的节点，它是在搜索过程中动态建立的。

为了叙述方便，这里引进几个概念，首先假设系统中包括 N 个节点，每个节点维护几张路由表。

①本地信息表，保存本地共享信息列表。由（fag，num）组成，其中 fag 是超级节点的标志，num 代表成功查询的次数。

②朋友节点表，保存朋友节点的有关兴趣资源信息列表。由（listlen，numf，address）组成，其中，listlen 代表朋友节点表长度，numf 代表朋友节点的数量，address 代表朋友节点的地址。

3.1 超级节点的建立

系统初始时，每个节点都定义为普通节点。当节点 A 刚加入 Gnutella 网络时，根据本地的共享资源建立本地信息表，并在搜索过程中由共享资源的变化而进行更新。当节点 A 第一次发出查询时，采用 Gnutella 的泛洪搜索机制在网络中传播，同时返回一系列拥有该资源的朋友节点，只有当节点 A 成功地从节点 B 中下载资源时，节点 B 的本地信息表中 num 加 1，同时节点 A 将节点 B 加入到朋友节点表中，节点 A 的朋友节点表中 numf 加 1；当节点 C 的查询继续在节点 B 上搜索成功并下载完成时，节点 B 的本地信息表中 num 继续加 1，节点 C 也把节点 B 加入到朋友节点列表中，同时节点 B 的朋友节点表中 numf 加 1。多次查询之后，num 达到某个阈值，节点 B 的本地信息表中 fag 设置为 1，节点 B 定义为“超级节点”。

同时超级节点还可以通过朋友节点建立。如节点 A 发出查询，首先查询朋友节点 B，如果查询成功，则节点 B 的本地信息列表中 num 继续加 1，如果 num 达到系统开始时设定的阈值，则节点 B 的本地信息表的 fag 设置为 1，节点 B 定义为“超级节点”；否则，num 减 1，如果 num 的值小于设定的阈值，则节点 B 的 fag 设置为 0，节点 B 变为普通节点。

建立超级节点的算法描述如下：

算法 1

名称：建立超级节点

功能：节点发出查询请求，根据该节点成功查询的次数判断是否定义为超级节点。

输入：查询

输出：超级节点

步骤：

①如果 A 是新加入节点，转②，否则转④；

②采用 Gnutella 的泛洪算法，找到一些朋友节点，修改查询节点的信息表；

③判断朋友节点的本地信息表中 num（查询成功的次数）的值是否大于设定的阈值，若是则本地信息表的 fag 设置为 1，该朋友节点定义为“超级节点”；否则，修改查询节点信息表及该朋友节点的 num 值；

④在朋友节点的基础上表进行新的搜索，根据查询成功或失败对朋友节点的 num 值进行调整。

3.2 朋友节点的建立

我们将每个节点看作一个语义向量，为了描述节点之间的相似性，对节点的共享资源用向量模型（VSM）表示。设 d1,d2,…, di,…,dn 为同属一类的一组对象，每个文档 d 表示一个范化特征向量 $V(d)=(t_1,w_1(d);t_i,w_i(d);\cdots,t_n,w_n(d))$，其中 t_i 为特征词，$w_i(d)$为 t_i 在 d 中的权值，一般被定义为 t_i 在 d 中出现频率 $tf_i(d)$的函数。其中，n 为所有文档的数目，n_i 为含有特征词 t_i 的文档数。节点加人到网络的同时，将自己的兴趣特征向量发送到邻居节点，该节点接受到这个信息后，与自己的兴趣特征向量进行相似度计算，计算公式：

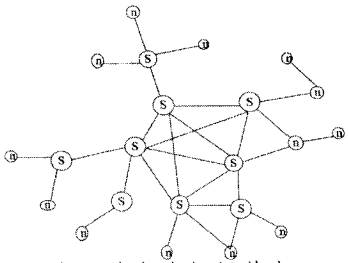
$$S=\sum_{k=1}^{n} w_i(d_i)\times w_i(d_j)\Big/\sqrt{(\sum_{k=1}^{n} w_{i2}(d_i))(\sum_{k=1}^{n} w_{i2}(d_j))}$$

判断两个对象的语义相似度。若节点 d1 向节点 d2 发出查询请求，计算 S 的数值。S 越大，表示节点 di 和 dj 的兴趣相关程度越高，反之表示兴趣相关程度低。预先设定一个阈值，阈值的设定方法可以根据应用环境的不同而改变。若 S 的值大于这个阈值，节点 d_i 将节点 d_j 的信息添加到自己的朋友节点表当中，同时对应的朋友信息列表中 numf 加 1。否则，不添加。列表长度设定为某一个值，多次查询之后，朋友节点表填满。当新的朋友节点要加入朋友节点表时，淘汰朋友节点列表中最近最少使用的朋友节点。系统通过节点成功地提供资源的历史记录来评价其应答能力。

3.3 搜索算法描述

在改进搜索算法中搜索过程描述如下：

如图 1 所示，如果节点 A 是超级节点，发出查询后，首先把自己的搜索请求向自己的朋友节点转发，收到消息的朋友节点将在本地搜索并将搜索请求转发给自己的朋友节点，搜索成功后将搜索结果返回给查询节点 A，同时更新本地信息表以及朋友节点表。如果朋友节点没有获得足够的资源，更新本地信息列表，如果本地信息类列表中 num 数值小于设定的阈值，则超级节点转变为普通节点，同时将查询消息直接通过泛洪的方式广播。



s 表示节点为超级节点

n 表示节点为普通节点

图 1　转发查询消息

如果节点 A 是普通节点，按朋友节点的建立过程以一定的 TTL 值广播查询消息。若搜索成功，则将搜索结果返回给查询节点 A，同时更新本地信息表，如果 num 的值大于设定的阈值，则节点 A 转变为超级节点；搜索不成功，则继续广播查询，直到 TTL 减为 0。转发算法描述如下：

算法 2

名称：改进的搜索查询算法

功能：根据查询历史有选择地转发查询。

输入：查询

输出：符合转发条件的节点

步骤：

①节点 A 发出查询，首先向所有的朋友节点转发查询消息，如果成功则返回查询结果，并修改相应的朋友列表信息，本次查询结束；否则转②；

②若节点 A 是超级节点，搜索本地资源并转发给自己的邻居超级节点，但是不转发到将请求信息转发给自己的邻居超级节点，如果成功则本次查询结束；否则转③；

③若节点 A 是普通节点，则仅按泛洪搜索机制在网络中传播，不进行资源搜索，直到遇到超级节点，转②。

4　仿真及结果分析

本模拟实验是在一台 PC 上完成的，网络拓扑结构由 PLOD 算法产生。构造了一个具有 1000 个节点和 l00000 份文档资源的 P2P 网络，文档在节点间采用 80:20 分布。为了评价改进算法是否有效，文档的稀有程度设置各不相同，热门文档可能超过 200 份，而稀有文档可能不足 10 份，实验中设置节点的朋友节点数目取值为 l0。仿真实验主要从以下两个方面分析比较了 Gnutella 环境下的洪泛算法和本文构造的改进搜索算法。

查全率：由图 2 可知，刚开始时两者的查全率是基本一致的，但是随着朋友节点的出现以及超级节点聚集的信息逐步增加，改进搜索算法的查全率明显高于 Gnutella。

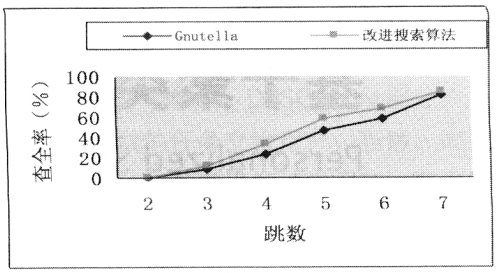


图 2　查全率

网络信息流量：随着查询次数的增加，朋友节点也越来越多，改进搜索算法中转发消息的节点明显减少，而且没有影响到搜索结果，因此减少了网络带宽的占用，提高了带宽利用率。

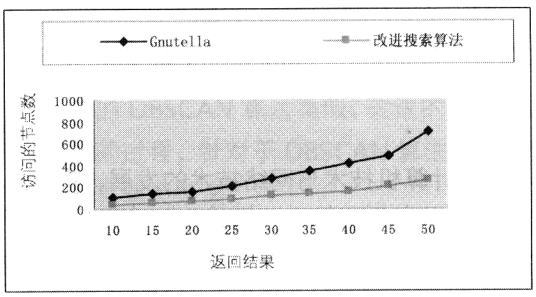


图 3　网络信息量

5　总结

本文提出的改进搜索算法，利用节点的自身兴趣，建立朋友节点，同时又能利用文档资源在节点间的分布信息，动态建立“ 超级节点 ”。搜索请求首先在朋友节点上查询，如果查询结果不能满足需求，则继续在超级节点和普通节点转发。模拟实验结果显示，改进的搜索算法中查询请求在小范围节点上传播就可以获得足够的信息资源。大幅度降低了网络带宽消耗，同时也保证了一定的查全率，从而提高了系统的搜索效率。

参考文献

1 张亮,邹福泰,马范援.对等网信息检索的研究现状与展望.计算机学报,2004,31(4):74–78.

2 凌波,陆志国,黄维维,钱卫宁,周敖英.基于 Peer-to-Peer 的信息检索系统.软件学报，2004,15(9):1375–1384.

3 Napster Home Page. http://www.napster.com/

4 Gnutella Home Page. http://www.gnutella.com/

5 KaZaA File Sharing Network.KaZaA website.http://www.kazaa.com/, 2003-10-12.

6 吴连龙.基于特别兴趣组的 P2P 网络搜索算法.计算机应用,2007,27(8):1871–1876.