用于对等全文检索的安全覆盖网

霍 林[1,2] 黄保华[2] 鲍 洋[1] 胡和平[1]
(华中科技大学计算机科学与技术学院 武汉 430074)[1] (广西大学计算机与电子信息学院 南宁 530004)[2]

摘 要 对等全文检索充分利用对等节点的资源实现检索,其关键是控制检索请求传播的节点范围。结合全文检索的安全要求提出安全覆盖网(Secure Overlay Network,SON),按安全级支配关系将对等节点组成网络。SON中节点发起的检索请求只能向下传递到安全级受其支配的节点,涉及节点是整个覆盖网中节点的子集,检索结果也是符合安全要求的。给出了SON的定义并分析了其性质,介绍了基于SON的对等全文检索原理和算法并分析了其安全性。实验表明,基于安全覆盖网的对等全文检索具有良好的检索效率。

关键词 安全覆盖网,对等系统,全文检索

Secure Overlay Network for Peer-to-Peer Full Text Search

HUO Lin[1,2] HUANG Bao-hua[2] BAO Yang[1] HU He-ping[1]
(College of Computer Science & Technology, Huazhong University of Science & Technology, Wuhan 430074, China)[1]
(School of Computer, Electronics and Information, Guangxi University, Nanning 530004, China)[2]

Abstract P2P(Peer-to-Peer) full text search utilizes resources of many peers, it is important to control the scope of peers involved in a search request. Considering the requirements of security in full text search, SON(Secure Overlay Network) was proposed for organizing peers into overlay network according the domination of their security level. In SON, a search request of a peer should be transferred down to peers it can dominate. So the peers involved in a search request are only a subset of peers of the overly network, and the search result should accord with the security police certainly. Definition of SON was given and its properties were analyzed. Principle of P2P full text search based on SON and its security problems were discussed. The experiments show that P2P full text search based on SON is efficient.

Keywords Secure overlay network(SON), Peer-to-Peer(P2P), Full text search

1 引言

在网络信息量呈爆炸式增长的今天,全文检索(Full text search)技术不可或缺。如果没有诸如Google等搜索引擎的辅助,我们从Internet获取信息必将非常困难。随着信息化的不断深入,大量具有保密性要求的信息也将存入计算机网络,对这些信息的有效利用就需要安全的全文检索技术。

无论是工业界还是学术界,在全文检索中考虑安全问题的工作并不多。Wen等人提出了一个用于语义搜索的概念模型并在安全访问控制中进行了实现,该模型通过提供访问控制扩展了搜索能力[1]。李新提出的密文全文检索技术在PKI和全文检索技术的基础上,实现了在不解密条件下的数据检索[2]。李瑞轩等提出的基于密文的全文检索系统可以实现密文条件下的全文信息检索,保证了敏感数据的安全性[3]。实际上,被检索数据在检索时不解密只是实现安全全文检索的一个必要方面,检索过程中的访问控制和检索效率也是非常重要的。

索引技术是保证全文检索效率的基础,并行检索则利用多处理机缩短全文检索的响应时间。本文针对涉密信息安全检索的需要,结合对等计算(Peer-to-Peer,P2P)技术提出安全覆盖网(Secure Overlay Network,SON)的概念,并在此基础上构建安全的对等全文检索系统。

对等全文检索利用对等网络中丰富的计算资源和检索结果缓存来提高全文检索的效率,SON则使对等检索只发生在符合安全要求的对等节点(Peer)集中,这不但保证了对等检索的安全性,而且还能够提高对等检索的效率。

2 安全覆盖网

P2P网络中包含许多对等节点,从安全的角度看,这些节点可以分为不同的级别。按照安全级别的不同来聚类对等节点,一方面可使类中的操作满足安全要求,另一方面可限制操作涉及的节点范围,减少无用信息传播和操作,提高整体效率。

2.1 安全级

为实现安全的全文检索系统,必须在系统中对检索内容实施访问控制。目前常用的访问控制模型主要有两种,即

到稿日期:2010-03-10 返修日期:2010-05-30 本文受国家自然科学基金(10876012)和广西大学博士启动基金(DD060058)资助。

霍 林(1965—),女,博士生,主要研究方向为全文检索、操作系统等,E-mail:nnxhy@163.com;黄保华(1973—),男,博士,主要研究方向为P2P安全与应用等;鲍 洋(1980—),男,博士生,主要研究方向为对等计算;胡和平(1952—),男,教授,主要研究方向为软件工程、智能决策系统、信息安全等。

BLP 模型和 RBAC 模型[4]。BLP 模型通过二元组 $l=$(密级，范畴)来表示主体和客体的安全级，并通过判定主体和客体的安全级支配情况实施访问控制。RBAC 中角色是权限的集合，在允许角色权限继承的情况下，不同角色间也能形成支配关系。因此，我们可以统一用安全级来描述 BLP 模型中的安全级和 RBAC 模型中的角色层次。

定义 1 安全级指主体和客体的安全性等级，所有安全级的集合是一个偏序集$\langle L;\leqslant\rangle$，对任意 $l_1,l_2\in L$，若 $l_1\leqslant l_2$，则称 l_2 支配 l_1。

在全文检索中，如果主体 s 的安全级 l_s 支配客体 o 的安全级 l_o，即 $l_o\leqslant l_s$，则允许 s 访问 o，否则不允许访问。根据安全级的定义，若主体 $s1$ 的安全级 l_{s1} 支配主体 $s2$ 的安全级 l_{s2}，即 $l_{s2}\leqslant l_{s1}$，则 $s2$ 缓存的检索结果 $s1$ 可以访问。

2.2 安全覆盖网

对等网络中对等节点直接代表使用者，可以看成主体。我们可以通过对等节点间安全级的支配关系将对等节点组织起来，形成安全覆盖网 SON。

定义 2 安全覆盖网是由对等节点按照安全级支配关系连接起来的网络，表示为 SON(P,V)，P 是所有对等节点的集合，V 是对等节点间有向连接的集合，若节点 $p1,p2\in P$，存在 $p1$ 到 $p2$ 的有向连接 $v_{p1->p2}\in V$，则 $p1$ 的安全级 l_{p1} 和 $p2$ 的安全级 l_{p2} 满足 $l_{p2}\leqslant l_{p1}$。

图 1 给出了一个 SON 的示例。从图中可以看出，安全级高的节点直接或间接支配安全级低的节点，安全级越低，可支配的节点数就越少。依据定义 2，参考图 1，结合 P2P 网络的特征，可以得出 SON 的几个重要性质。

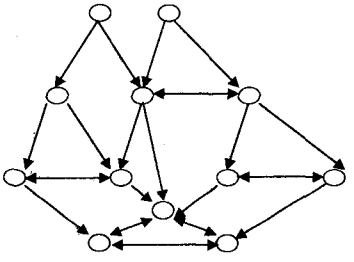


图 1 安全覆盖网示例

性质 1(不完全性) 在 P2P 网络中建立起来的安全覆盖网 SON(P,V)中，V 可能是不完全的，即存在 $p1,p2\in P,l_{p2}\leqslant l_{p1}$，但 V 中不存在从 $p1$ 到 $p2$ 的有向连接。

性质 2(环路存在性) SON 中可能存在从节点 $p1$ 出发，又回到 $p1$ 的有向连接。图 1 中最下层 3 个节点就构成了环路。

性质 3(全支配性) SON 中可能存在安全级别最低的节点 $p1$，对任意节点 px，$l_{p1}\leqslant l_{px}$。图 1 中最下层的 3 个节点安全级就受到 SON 中所有节点的支配。

安全覆盖网的这些性质将直接影响基于 SON 的对等全文检索的性能和安全性，成为安全的对等全文检索要解决的关键问题。

2.3 节点管理算法

从定义 2 可知，P2P 网络中的节点需记录安全级受自己支配的对等节点以形成安全覆盖网。在实际的 P2P 网络中，节点不可能记录全部安全级受自己支配的节点，而只能记录安全级受自己支配的邻居节点。为提高效率，对等节点还可记录安全级来支配自己的邻居节点，以满足安全检查或操作回溯的需要。

安全级支配节点 p 的节点可记录在 $UTab_p$ 中，安全全级受节点 p 支配的节点记录在 $DTab_p$ 中。$UTab_p$ 和 $DTab_p$ 位于节点 p 上，由节点 p 自己维护。

节点 p 加入 SON 的算法步骤如下：

- 节点 p 在 P2P 网络中广播加入 SON 的消息；
- 收到 p 广播的节点，安全级支配 p 的节点集 Pu 应答 u 并将 p 加入到自己的 $DTab$ 中，安全级受 p 支配的节点集合 Pd 应答 d 并将 p 加入到自己的 $UTab$ 中；
- 节点 p 将节点集 Pu 和 Pd 分别加入到 $UTab_p$ 和 $DTab_p$ 中。

在节点加入算法中，需要进行安全级的判断，这可以采用许多方式，比如 PMI 服务、属性证书、Kerberos、P2P 信任评价等，可根据具体应用环境进行选择。另外，如果要限制 $UTab$ 和 $DTab$ 的大小，则可采取适当的置换策略，如 LRU，LFU 等。

节点 p 退出 SON 的算法步骤如下：

- 节点 p 给邻居节点发送退出 SON 的消息；
- 收到 p 退出消息的节点从自己的 DTab 和 UTab 中删除 p 的记录。

节点退出除由节点主动请求外，SON 中节点还需增加主动检测机制，按计划检测邻居节点的状态，当检测到某节点不在线时，主动从自己的 DTab 和 UTab 中删除该节点。

3 基于安全覆盖网的对等全文检索

全文检索是计算密集型过程，需要大量计算资源，而 P2P 网络中富含存储、信息、计算等资源。通过对等全文检索技术，合理利用这些资源可提高全文检索的效率。对等全文检索中要解决的一个关键问题是限制查询请求传播的节点范围，安全覆盖网不但可以解决这个问题，而且还满足了检索过程安全性的要求。

3.1 系统框架与检索原理

一般全文检索服务由一台服务器或多台服务器组成的服务器集群提供，我们把它叫做集中式检索(Centralized Search，CSearch)。基于安全覆盖网的全文检索(Search based on SON，SSearch)在这种模式基础上增加对等检索(P2P Search，PSearch)功能，系统原理如图 2 所示。

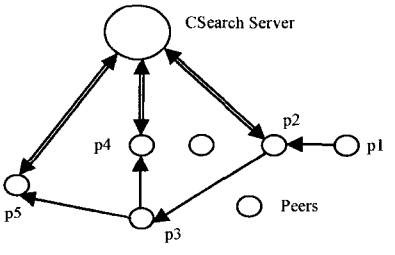


图 2 基于安全覆盖网的全文检索原理

从图 2 中可以看出，基于安全覆盖网的全文检索 SSearch 由集中式检索 CSearch 和基于安全覆盖网的对等检索 PSearch 两部分组成。CSearch 由节点直接向检索服务器发起，检索的结果由节点缓存，作为节点提供检索服务的数据。PSearch 由对等节点在安全覆盖网中发起并执行，能够检索的对象就是各节点缓存的内容。

节点 p 执行检索请求 q 的 SSearch 执行步骤如下：

- 节点 p 自己执行 q；
- 节点 p 将 q 发送给所有 $DTab_p$ 中的节点；
- 接收到 q 的节点自己执行 q；
- 接收到 q 的节点将 q 发给其 $DTab$ 中的所有节点；
- 节点 p 将 PSearch 检索结果呈现给用户，如果用户满意，则结束执行；
- 节点 p 将 q 发给 CSearch Server，缓存检索结果并呈现给用户。

安全覆盖网的性质 2 指出其具有环路特征，节点在转发请求时可在查询中增加自己为转发节点，这样收到转发的节点就可以判断出转发的路径，如果发现环路，可将请求丢去。

图 2 中节点 $p2$ 将请求发给 $p3$，由 $p3$ 发给 $p4$ 和 $p5$，检索结果不满足要求，然后 $p2$ 直接向 CSearch Server 请求检索。节点 $p1$ 将请求发给 $p2$，然后发给 $p3$ 并发给 $p4$ 和 $p5$，检索结果满足要求，因此不向 CServer Server 发起检索。

3.2 查准率与查全率

查准率(Precision)和查全率(Recall)是衡量全文检索系统检索效果的重要指标。设检索 q 相关的文档数为 r，检出的相关和不相关文档数量分别为 c 和 e，则 $Precision=c/(c+e)$，反映拒绝非相关信息的能力；$Recall=c/r$，反映检出相关信息的能力。

在基于安全覆盖网的对等检索中，设 q 被转发到的节点集合为 Ps，则：

$$Precison_{SON}=\sum_{p\in Ps}c_p/\sum_{p\in Ps}(c_p+e_p) \quad (1)$$

$$Recall_{SON}=\sum_{p\in Ps}c_p/r \quad (2)$$

设各节点采用相同的查询算法和软件，对任意节点 p，查准率可看成常数 $Precision$，则 $c_p=Precision\times(c_p+e_p)$，结合式(1)可得：

$$Precison_{SON}=Precision \quad (3)$$

从式(3)可以看出，查准率是由各个节点执行 q 的查准率决定的，与安全覆盖网无关。

对于查全率，从式(2)可以看出，与查询 q 被转发到的节点集 Ps 有关，自然就与安全覆盖网有关系。安全覆盖网的性质 1 指出，其具有不完全性，q 可能没有被转发到所有包含 q 检索结果缓存的节点，导致查全率降低。

改变安全覆盖网的不完全性是困难的，但可以改善 q 转发的覆盖范围。在节点的 UTab 中，增加记录节点安全级的列，并要求 q 中附带请求节点的安全级。这样，节点在转发 q 时，除选择 DTab 中的节点外，还选择 UTab 中安全级受 q 安全级支配的节点。

3.3 安全性

基于安全覆盖网的对等全文检索的安全性主要涉及 3 个方面：(1)节点安全级确认的安全性；(2)对等节点缓存的检索结果的控制是自主的访问控制；(3)查询请求泄密。

准确确定邻居接点的安全级是安全覆盖网得以构建并安全工作的基础。一般处理涉密信息的网络中都会部署身份认证和权限管理的服务，比如 PKI/PMI，Kerberos 等，可以保证安全级确认的安全性。在一些安全要求不高的场合，也可以采用信任级表示安全级，信任的管理有许多方案。

在节点要缓存检索结果时采用对等的全文检索，节点拥有了这些内容，就可以按照自己的意愿允许其它节点访问这些内容。这种访问控制是自主的，可能和原内容要求的强制访问控制策略相违背。可以通过两方面措施来缓解这一安全问题。一方面通过加密等措施，让缓存的内容只能由对等检索系统识别和处理，而节点用户不能直接得到；另一方面通过软件完整性技术实现对等检索系统的完整性，保证系统正在实施规定的安全策略。

查询请求 q 由高安全级节点向低安全级节点传递，而 q 可能反映了发起查询(高全级)节点的意图，即 q 携带了信息。由于安全覆盖网的全支配特征(性质 3)，恶意节点就可以以最低安全级加入安全覆盖网，网络中所有查询请求就可以到达它，而它不用执行查询，也不用返回结果或只返回没有所查询内容的应答。可以通过限制 q 到达节点的最低安全级来解决这一问题。

4 实验研究

为检验基于安全覆盖网的对等全文检索的效率，以开放源代码的全文检索软件 Lucene[5] 和对等计算平台软件 JXTA[6] 为基础设计实现了一个对等全文检索系统(如图 3 所示)。系统中采用 JXTA PeerGroup 来组织节点的邻居节点，每一节点创建自己的 UGroup 和 DGroup 两个 PeerGroup，分别用于管理安全级来支配该节点和受该节点支配的节点。

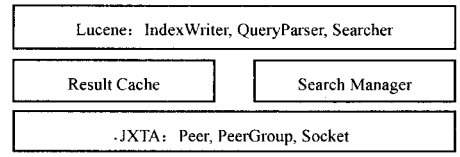


图 3 基于安全覆盖网的对等检索系统结构

测试在一个由 16 个对等节点和 1 个服务器组成的局域网上进行。网络为 100Mbps 快速交换以太网，对等节点配置为 2.0G CPU / 1GB RAM / Red Hat Linux 9，服务器为 2×2.6G CPU / 4GB RAM / Red Hat Enterprise Linux 4。服务器用于提供集中式检索服务，软件基于 Lucene 编码实现。节点被分成绝密、机密、秘密、公开 4 个安全级，对应节点数分别为 2,4,8,2。测试结果如图 4 所示。

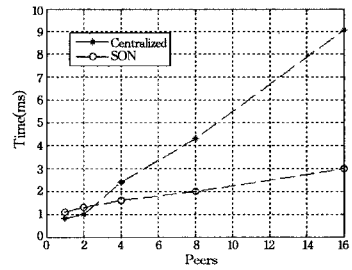
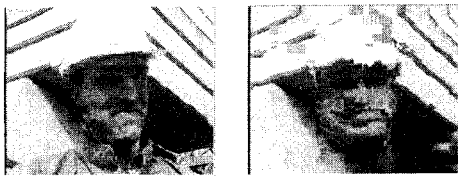


图 4 集中式和基于安全覆盖网的检索响应时间

从图 4 中可以看出，集中式全文检索服务的响应时间随并发节点数的增加快速增长，与并发节点数成正比关系，这是由所有并发请求都要由单一的服务器处理造成的。基于安全覆盖网的对等全文检索方式响应时间虽然也随并发节点数的增加而增加，但相对于集中式检索要少很多，这是由于并发请求在 SON 中得到了并发处理。以上实验表明，基于安全覆盖网的对等全文检索系统能够利用对等节点的检索结果缓存和计算能力，充分发挥对等网络中大量对等节点的功能，显著提

(下转第 139 页)



(a) 50ms (b) 10ms

图 10 不同时延门限下的视频回放快照

结束语 本文提出的 OMNET++与 H.264 视频编解码器联合仿真的方法，能够很好地评估在视频传输等多媒体实时业务的要求下 Ad hoc 网络中协议算法的综合性能，并为网络协议算法的进一步开发与改进提供了有效的检验平台。经过以上的性能分析，可以看出 IEEE802.11 协议并不能很好地支持视频在 Ad hoc 网络中的传输，尤其是网络负载较大、视频业务对时延有较高要求时性能比较恶劣。为了支持视频在 Ad hoc 网络中的实时传输，保证视频业务的服务质量，相应的网络协议算法还有很大的改进空间。

参 考 文 献

[1] Frodigh M, Johansson P, Larsson P. Wireless ad hoc networking-The art of networking without a network[J]. Ericsson Review, 2000(4): 248-263

[2] 徐雷鸣，庞博，赵耀. NS 与网络模拟[M]. 北京：人民邮电出版社，2003

[3] 陈敏. OPNET 网络仿真[M]. 北京：清华大学出版社，2006

[4] Varga A. OMNET++-Discrete Event Simulation System Version 4.0 User Manual[EB/OL]. http://www.omnetpp.org/, 2003

[5] Joint Video Team(JVT) of ITU-T VECG and ISO/IEC MPEG. Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification(H.264/AVC)[S]. JVT, 2003

[6] IEEE. Wireless LAN Medium Access Control(MAC) and Physical Layer(PHY) Specification[S]. IEEE 802.11, 1999

[7] 余兆明，查日勇，黄磊，等. 图像编码标准 H.264 技术[M]. 北京：人民邮电出版社，2006

[8] Qiao D, Choi S, Shin K G. Goodput Analysis and Link Adaptation for IEEE802.11a Wireless LANs[J]. IEEE Transactions on Mobile Computing, 2002, 1(4): 278-291

(上接第 106 页)

高全文检索的效率。

5 相关工作比较

索引技术作为全文检索的核心技术之一得到了广泛研究，成果包括倒排文件、签名文件、后缀树等索引结构[7]。并行全文检索很早就受到重视[8]，直到今天仍然是重要的热点研究领域。近年来随着对等网络的发展，基于对等网络的信息检索得到大量研究[9]。

对等网络能够汇集网络边缘的存储、计算资源以及网络带宽，具有高可靠性和可扩展性[10]，为信息检索提供了有利条件。对等网络环境下信息检索的关键在于将请求发往恰当的节点执行。在结构化对等网络中一般采用 DHT[11] 技术，内容所处的位置由内容经 Hash 得到，即内容的位置由内容决定，因此可以高效率地找到查询节点。结构化对等网络中一般只能进行精确匹配，不能实现全文检索，文献[12]提出的质心法解决了这个问题。

非结构化网络中最简单的方式是洪泛，这种方式效率很低。提高非结构化对等网络中信息检索性能的最主要的方法是节点聚类，即按照某种方法将对等节点分类，检索就可以被控制在一类或少数几类节点中进行。文献[13]基于 Small World 原理，采用自适应线型空间算法将语义向量映射在一维小世界模型中以聚类节点。文献[14]定义了反映用户偏好的用户模式树，通过计算用户模式树间的相似性，将用户分类为不同的社区。

本文利用主体的安全级来分类对等节点，不但在检索效率方面是对已有工作的重要补充，而且考虑了安全性，能够实现安全的对等全文检索。

结束语 基于安全覆盖网的对等全文检索充分发挥了对等系统对数据和计算资源、网络带宽的汇聚作用，并通过安全覆盖网将检索参与节点限制在一个小而满足安全要求的范围内，不但提高了检索效率，而且能够满足涉密信息检索的安全需求。实验结果表明，安全覆盖网对检索效率的提高具有明显作用。下一步将在实际政务办公自动化系统中部署基于安全覆盖网的对等全文检索系统，研究和检验其效果并为改进提供依据。

参 考 文 献

[1] Wen K, Lu Z, Li R, et al. A Semantic Search Conceptual Model and Application in Security Access Control[C]// Proceedings of the First Asian Semantic Web Conference. Beijing, China, 2006

[2] 李新. 密文全文检索技术[P]. 200410070113.2004

[3] 李瑞轩，卢正鼎，宋伟，等. 基于密文的全文检索系统[P]. ZL200610124691.1.2006

[4] Sandhu R, Ferraioloy D, Kuhny R. The NIST Model for Role Based Access Control Towards A United Standard[C]// Proceedings of the Fifth ACM Workshop on Role-based Access Control. Berlin, Germany, 2000

[5] Welcome to Lucene[OL]. http://lucene.apache.org/. January 10, 2010

[6] JXTA Community Projects[OL]. https://jxta.dev.java.net/. January 10, 2010

[7] 刘小珠，彭智勇. 全文索引技术时空效率分析[J]. 软件学报，2009, 20(7): 1768-1784

[8] Salton G, Buckley C. Parallel Text Search Methods[J]. Communications of the ACM, 1988, 31(2): 202-215

[9] Jiang Q, Guan J. A Peer-to-Peer Based Text Sharing and Retrieval System[C]// Proceedings of the Future Generation Communication and Networking 2007. Jeju Island, Korea, 2007

[10] Milojicic D S, Kalogeraki V, Lukose R. Peer-to-Peer Computing [Z]. Hewlett-Packard Company, 2002

[11] Kaashoek M F, Karger D R. Koorde: A Simple Degree-optimal Hash Table[C]// Proceedings of the 2nd International Workshop on Peer-to-Peer Systems. Berkeley, CA, USA, 2003

[12] 程学旗，吕建明，周昭涛. 基于对等网络的全文信息检索[J]. 计算机研究与发展，2004, 41(12): 2148-2155

[13] Li M, Lee W-C, Sivasubramaniam A, et al. A Small World Overlay Network for Semantic Based Search in P2P Systems[C]// Proceedings of the 2nd Workshop on Semantics in Peer to Peer and Grid Computing. New York, USA, 2004

[14] 张亮，邹福泰，张文举，等. 基于社区的对等网络信息检索[J]. 上海交通大学学报，2006, 40(5): 767-770