大数据安全高效搜索与隐私保护机制展望

李尚，周志刚，张宏莉，余翔湛
（哈尔滨工业大学计算机科学与技术学院，黑龙江 哈尔滨 150001）

摘　要：随着信息产业的飞速发展，围绕大数据搜索展开的服务已渗透到人们生活的方方面面。相关技术领域也正在经历深刻变革，如数据融合的隐私保护、场景感知的搜索意图理解、统计概率式的搜索模式等。结合国内外最新研究进展，对大数据安全高效搜索与隐私保护问题进行了研究展望：首先，从多源数据发布、用户搜索需求感知及隐私感知的智慧解答 3 个视角凝练了大数据安全搜索与隐私保护的科学问题；其次，提出了面向大数据的信息融合与知识萃取技术、粒度化的知识表示与推演技术、支持平台与用户互动的搜索任务表示模型、基于用户体验驱动的任务管理技术、效用与代价平衡的粒度化搜索技术和基于差分隐私的安全搜索机制等研究内容；最后，对相关的技术路线进行了展望。

关键词：大数据；安全搜索；隐私保护；多源融合
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.2096-109x.2016.00041

Prospect of secure-efficient search and privacy-preserving mechanism on big data

LI Shang, ZHOU Zhi-gang, ZHANG Hong-li, YU Xiang-zhan
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: With rapid development of information industry, services related to big data search have permeated into almost every aspect of human lives. Relevant technologies are undergoing profound changes, such as privacy preservation for data fusion, context-aware search intention comprehending and statistic search pattern. The issues of secure-efficient search and privacy-preserving for big data were studied and looked ahead. Firstly, three chief scientific problems were refined from the aspects of multi-source data publication, awareness of users' search requirements and privacy-aware wise solutions, respectively. Secondly, main research contents were proposed, including big data oriented information fusion and knowledge extraction, granular knowledge representation and inference, interactive representation of search task, user experience driven task management, granular search pattern balancing utility with cost, and secure search mechanism based on differential privacy. Lastly, several promising techniques in future were discussed.

Key words: big data, secure search, privacy preservation, multi-source information fusion

收稿日期：2016-03-07；**修回日期**：2016-04-02。**通信作者**：周志刚，zzgisgod@sina.com
基金项目：国家自然科学基金资助项目（No.61402137）；国家重点基础研究发展计划（“973”计划）基金资助项目（No.2013CB329602）
Foundation Items: The National Natural Science Foundation of China (No.61402137), The National Basic Research Program of China (973 Program) (No.2013CB329602)

1 引言

随着国家“互联网+”行动计划的实施，信息技术与经济社会的交汇融合引发了数据的迅猛增长，大数据已成为国家基础性战略资源，并已渗透到全球人类生产、流通、分配、消费活动的各个环节，正日益对经济运行机制、社会生活方式和国家治理能力产生重要影响。

大数据中蕴含丰富的信息和知识：在政府决策方面，各地区、各行业、各领域的统计数据，蕴含有政策法规成效与影响的宏观态势，经科学利用后可极大提升政府决策效率和风险防范水平；在社会民生方面，交通、气象、旅游等跨部门跨地域的数据，能够协同构成全方位的出行信息，从而方便公众出行、促进旅游事业蓬勃发展；在经济生产方面，研发设计、生产制造、经营管理、市场营销、售后服务等产品全生命周期产生的数据，综合反映出产业链各环节存在的问题，可推动制造模式变革和工业转型升级。然而，这些信息和知识并不会自然呈现，而是需要用智慧的方法将其从海量的数据洪流中萃取出来，根据人们的现实需求进行多维度细粒度的高精度定制及演化，进而服务人们的工作和生活。

从用户的搜索需求看，随着人们生活工作的节奏加快，用户对搜索的期许发生巨大改变，相比枯燥漫长地等待精准解答，人们更期望得到满足精度需求的快速搜索体验。从搜索技术看，传统的搜索引擎主要面向 Web 1.0 静态网页，是基于关键字的“存在性扫描搜索”，不能支持面向 Web 2.0/3.0 应用具有 5V 特性的大数据及其满足用户快速高精度的搜索需求。这些问题催生了研究者对新型大数据搜索技术的探索。本文研究的面向大数据的安全高效搜索技术，是在明确用户搜索需求的基础上，对大数据进行融合、萃取、推演等处理挖掘知识，进而快速地给出满足用户搜索需求及数据隐私保护的智慧解答。大数据搜索具有以下 5 个特点：1) 多源融合，对多通道、多来源的异质、异构数据和信息进行统一表示，在数据层面消除数据缺失、歧义、冗余等现象，在信息层面消除由数据属性关联带来的融合隐私泄露风险；2) 知识综合，对规模庞大的数据属性进行归约，挖掘数据中蕴含的知识，建立统一的知识与关系表示模型，萃取多层次、粒度化且能够随数据的动态更新自我演化的知识聚合体；3) 有限开放，是指针对多源聚合数据在安全需求等级、隐私保护粒度和管理者利益等方面的不同，具备自适应、可重组、动态可调整的搜索信任、安全、隐私保护机制，面向不同的用户呈现差异化的搜索结果；4) 智慧搜索，综合考虑用户搜索情境、历史行为及数据访问权限等要素，整合挖掘出的相关知识聚合体，提供满足用户多元需求的智慧搜索方案；5) 快速响应，是指同时满足用户对搜索时效及搜索精度的诉求，能够有效复用历史同构搜索的粒度化结果视图，给出快速甚至在线解答。

在国家大数据战略和“互联网+”行动计划深入推进的时代背景下，人们期望大数据搜索能够达到“安全搜索随意行、隐私知识两相宜”的境界。为了尽快实现这一目标，需要对大数据搜索相关技术的国内外研究现状、发展态势等进行综合分析，进而凝练科学问题，并提出若干需要突破的关键技术。

2 国内外研究进展

本节从大数据的多源融合发布、用户搜索需求感知及隐私感知的智慧解答这 3 个方面分别介绍国内外最新的研究进展。

2.1 面向大数据的多源融合发布

对于传统的单数据源信息发布的隐私保护研究，如 *k*-匿名[1]（*k*-anonymity）、*l*-多样化[2]（*l*-diversity）、*t*-贴近性[3]（*t*-closeness）已广泛开展，Zhou 等[4]对其进行了综述，此处不做讨论。针对多源融合信息的隐私保护，一个值得讨论的问题是“多源融合信息的隐私保护方法能否继承传统的单数据源信息发布的隐私保护方法”，遗憾的是，答案是否定的。例如，图 1 中的 2 个列表分别是数据源 A 和数据源 B 发布的数据（Name 列为非公开属性），假设敌手已知 Alice 的一些背景信息（年龄 34），易知 2 个数据源发布的数据各自都满足（3-anonymity、2-diversity）的隐私需求，然而当 2 个数据源的数据进行融合时，Alice 患有胃炎的个人隐私信息被暴露。

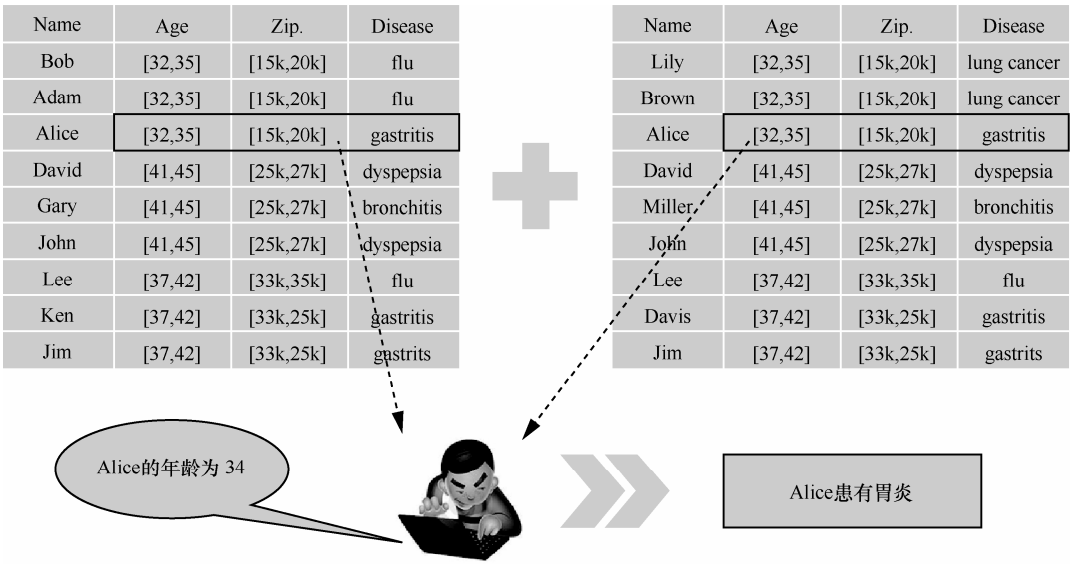

图 1　多源数据融合隐私泄露实例

为此，Xiao 等提出 *m*-不变性[5]（*m*-invariance），*m*-invariance 方案是一种增量式的数据发布方案，要求每次发布的数据与前一次发布的数据相比，满足下列的情况之一：1）新发布的数据与前一次发布的数据属于同一个匿名组；2）为新发布的数据构建新的匿名组；3）当匿名组中的真实数据个数低于预设门限时，将发生匿名组间的融合，但融合后形成的匿名组必须同时满足参与融合的各个匿名组的敏感属性值多样化需求。然而，该方案需要各个数据源的所有者明确知道彼此的真实数据，因此，仅适用于单个数据源多次数据发布的场景。安全多方计算成为一个可行的方案，Clifton 等[6]提出分布式 *k*-匿名算法（D*k*A，distributed *k*-anonymity），该算法假设在垂直划分的数据环境下同一条记录有唯一的全局标识，数据集成的各方都只拥有部分属性的数据，利用可交换加密在通信过程中隐藏原始信息，再构建完整的匿名表判断是否满足匿名门限来实现数据隐私保护。但该算法的时间开销很大，文献[6]中对 defacto benchmark adult 数据集匿名化需要 12 天。文献[7]开发了一个针对关系数据计数、并、交、笛卡尔积 4 种典型操作的安全数据多方数据集成工具。Mohammed 等[8,9]基于分类树结构使用数据泛化技术实现数据集成各方的数据隐私保护，但集成后数据的信息损失较高，具体的信息损失度与数据集相关。文献[10]提出了一种可追责计算框架，该框架可以实现数据集成的各方相互验证。扩展研究[11~13]意在为不同的集成数据挖掘任务设计安全协议，然而这些方法的计算开销过于昂贵，以至于难以在实际场景中应用。

2.2　用户搜索需求感知

用户搜索需求的感知、理解和匹配，是实现智慧和高效大数据搜索的重要前提。文献[14~16]分别根据信息需求类别、意图涵盖范围和查询内容映射提出了用户搜索意图的分类体系，但这些方法均假定用户查询请求的信息表述是全面、准确的，且早期的查询分类较为模糊，经过多次交互过程才能逐渐清晰。文献[17~19]分别研究了基于用户行为分析、伪相关反馈信息和自然语言处理的用户搜索意图理解方法，针对用户搜索请求进行语义识别，但是缺乏必要的结构化知识和深度理解能力，难以准确理解查询中蕴含的真实搜索意图，无法根据用户所处的时空特征进行个性化意图建模。

在用户搜索场景和上下文感知方面，研究人员已提出一些相关技术。Stankovic[20]提出物联网查询机制需要收集和感知大量持续变化的信息，个性化信息能产生可定制的更有效的查询。Maekawa[21]提出一种新颖的 Web 查询方法，根据用户正在从事的日常活动自动检索相关网页，并在附近智能家电上显示。ThingsNavi[22]是一种基于超图的物体搜索方法，该方法通过构建统一的

超图来表示人—物交互中丰富的结构和情景信息，并将相关物体的搜索转换为超图上的排序问题。Ostermaier 等[23]利用实体状态属性的历史数据建立预测模型（APM、SPPM、MPPM），预测可能匹配搜索请求的物理实体的集合，而后访问集合中的物理实体确认是否匹配。Christophe 等[24]通过用户的历史记录预测用户的语境，根据用户语境、物理实体数量和聚类的数量，决定采用哪种搜索算法。SMART[25]的搜索层采用了事件检测和排序检索模型，通过用户的上下文信息（如用户的位置或社会背景），可直接指定或者预测相关查询。现有的研究工作利用预测等模型提高了搜索的效率，但准确度有待进一步提升。在大数据搜索场景下，语境决定了如何跨网，不同网络的异类信息需要不同的融合方法，综合考虑诸如此类动态变化的因素，才能给出高效的智慧解答。

用户搜索需求还包括对求解精度、执行效率、资源和成本消耗等多维因素的要求。文献[26~28]所实现的各类大数据分析算法（如 top-*k*、skyline、join 等），在查询精度方面取得了令人满意的成果。但这些算法所采用的传统数据处理模式，如构建索引、排序等，从原始的庞大数据集中生成一个新的“大”数据，导致了一个漫长的计算过程，这在大数据搜索场景下是不可取的。为了实现较高的搜索效率，Chaudhuri[29]和 Doucet[30]等采用了一种叫做块状抽样的高效抽样方法，但块状抽样无法保证抽样一致性，在其基础上产生的近似误差取决于样本数据在磁盘上的分布，因而当数据分布集中于个别属性时会导致求解不准确。Hellerstein[31]和 Laptev[32]等分别基于关系数据库和分布式文件系统提出了查询近似结果的早期求解方法，但都没有给出具备普遍性的精度评估策略，从而限制了该方法在通用数据集上的推广能力。文献[33]提出了一种基于语义分析的 MapReduce 程序执行成本优化方法，但该方法依赖于特定的编程语言，在开放式计算平台环境中的适用性较差。文献[34]基于 MapReduce 框架细致地分析了各类平台参数对程序执行效率的影响，进而提出了给定资源下 MapReduce 程序的执行成本（时间）预估以及优化算法，但该工作并没有讨论多任务并发情况下的全局优化策略。通过总结可以发现，已有的研究工作只侧重于搜索精度、效率、能耗中的一个或两个目标，而如何统一多维搜索需求、实现保障全局用户体验的任务管理技术，仍需要进一步地探索和研究。

2.3 隐私感知的智慧解答

从安全搜索的角度看，文献[35]不仅实现了细粒度的访问控制，同时可以抵御如传感器妥协和用户勾结等攻击。文献[36]提出了一种分布式令牌重用检测方案去防止恶意用户对令牌的重用攻击。文献[37]提出了基于行为个性的访问控制机制，利用提取的行为特征进行控制。文献[38]提出了一种 KP-ABE 方法，将访问控制策略嵌入用户的私钥中，实现了细粒度的访问控制。Yang 等[39]提出了一种可撤销的多授权机构的 CP-ABE 结构，有效地解决了属性的撤销问题。文献[40]提出了一种匿名的权限控制方法，在解决数据隐私性的基础上解决了用户身份隐私性问题。文献[41]提出了一种签名方式称为基于属性的签名（ABS）。文献[42]设计了联合控制灵活细粒度的强制访问控制方案。文献[43]提出并实现了基于动态角色的访问控制方案来实施最小特权原则。针对处理隐式访问信息的访问控制问题，Singh 等[44]提出了适用于混杂移动应用的上下文感知权限控制方法。文献[45]开发了被称为 Auto-FBI 的原型系统，实现了敏感数据的自动隔离。文献[46]提出了一种对多人共享的数据的保护方式，设计并实现了一个多机构访问控制策略。现有的访问控制研究大多围绕感知层访问控制、基于属性的访问控制、面向身份隐私保护的访问控制、移动操作系统中的访问控制等技术展开。然而，针对大数据搜索用户的开放性与海量性、节点动态性等特征，海量动态用户访问权限实时更新和撤销的问题有待进一步研究与完善。

从隐私保护的角度看，文献[47]提出了基于差分隐私的查询处理技术。文献[48, 49]将隐私数据拆分成若干个称为“share”的碎片，根据实际的隐私需求，控制获取 share 数目来达到要求的精度。文献[50, 51]将隐私数据通过门限策略映射为 N 个无语义副本，当用户提出数据检索请求时，只需随机地从这 N 个子数据源中任选 K 个副本数据在用户端进行合成。文献[52~54]提出多极的

CP-ABE 策略，将分权机制引入到搜索代理平台，在搜索过程中搜索平台无法准确获知用户的身份信息，进而保护用户的搜索模式。Sankar 等[55]提出了一个抽象模型以支持含有任意数量公共和私有变量的数据库，使用了率失真理论来确定隐私保护和效用之间平衡的基本界限。Guo 等[56]针对社交网络中的用户提出了一个平衡算法来确定用户的隐私设置，以同时满足用户的隐私需求和效用偏好。Gu 等[57]给出了关于隐私泄露和效用损失的概念，并且利用概率论中的散度距离来量化它们。现有的隐私保护研究工作大多围绕某一类特定的场景，根据其隐私需求，提出与之相应的隐私保护方案及其策略，目前，提出的有数据加密、匿名、分割、加噪等代表性的方法。然而，不同的应用环境和情景需要不同的隐私保护方案，随着时间的推移，隐私保护具有复杂的时空场景特征，在这方面的研究还有待开展。

3　科学问题与研究内容

3.1　大数据安全高效搜索与隐私保护科学问题探究

针对大数据安全高效搜索与隐私保护的迫切需求，本文从多源融合数据发布、用户搜索需求感知及隐私感知的智慧解答的视角凝练 3 个具有挑战性的科学问题。

1) 大数据中信息归约与知识发掘

大数据中蕴含了丰富的信息与知识，对这些信息的有效整合并萃取其中的知识是保障搜索质量、提升用户体验的前提和基础。与传统的基于单一静态数据库的搜索不同，大数据具有体量巨大、渠道多源、持续生成等特点，如何整合来自于多源的数据，消除数据间的冲突、过滤数据噪声、消弭由数据融合诱发的隐私问题？如何从规模庞大的众多融合属性中理出其中的复杂关系？如何对复杂的关系进行归约、推演及粒度化建模？挖掘的知识聚合体如何根据数据的持续生成及用户的搜索导向进行自演化？这些都是急需解决的科学问题。

2) 用户搜索需求感知的任务表示与管理

在大数据搜索环境中，由于平台运行状态对用户透明，导致用户快速高精度的搜索需求与平台的任务执行能力产生矛盾。如何将用户搜索需求转化为对平台属性、状态刻面的普适化表示，并以此构建任务计算代价预估模型？如何消除任务所需资源、执行时间、搜索精度 3 个目标刻面的冲突？此外，各搜索任务对最大化占用计算资源的追求导致了对有限资源的竞争，如何设计能够保障全局用户体验的动态优先级刻画方法和任务管理技术？如何实现以任务当前优先级刻面为基准的抢占式资源动态配给机制？如何设计基于供需零和博弈的多目标优化调度策略？这些都是需要解决的科学问题。

3) 隐私感知的快速高精度智慧解答

从用户和大数据的二元视角进行剖析，一方面，用户渴望快速高精度的搜索体验，另一方面，大数据自身不仅蕴含体量巨大、复杂关联、时间演化的知识，而且对用户的有限开放性，决定了其必须具备“千人千面”的搜索行为识别及控制能力。如何在保障数据隐私的前提下实现快速甚至在线的高精度搜索求解？如何构建精度测量与隐私保护评价统一融合的度量机制？如何在大数据持续量变冲击下定量刻画知识的可复用边界，进而提供知识的柔性复用？如何维护数据更新演化与数据隐私保护粒度的动态平衡模型及隐私保护顽健稳定性边界？这些都是需要解决的挑战性科学问题。

图 2 展示了 3 个科学问题间的关系。如图 2 所示，该研究的核心可概括为三要素：“知识发掘与建模”、“用户需求感知的任务管理”、“隐私感知的智慧解答”。其中，“知识发掘与建模”是从海量数据中萃取知识的关键环节，为大数据搜索提供了粒度化的底层数据支持；“用户需求感知的任务管理”将用户搜索需求与平台运行状态有机融合，是用户搜索请求顺利执行和平台高效运转的有力保障；“隐私感知的智慧解答”是大数据搜索的核心环节，提供效用、代价与隐私保护综合考量的智慧解答。三要素之间相互关联和依存。“知识发掘与建模”与“用户需求感知的任务管理”的协同运作为“隐私感知的智慧解答”提供从底层数据到上层任务的支撑。

3.2　研究内容

针对上述科学问题，可从 6 个方面展开研究，主要研究内容与科学问题的对应关系如图 3 所

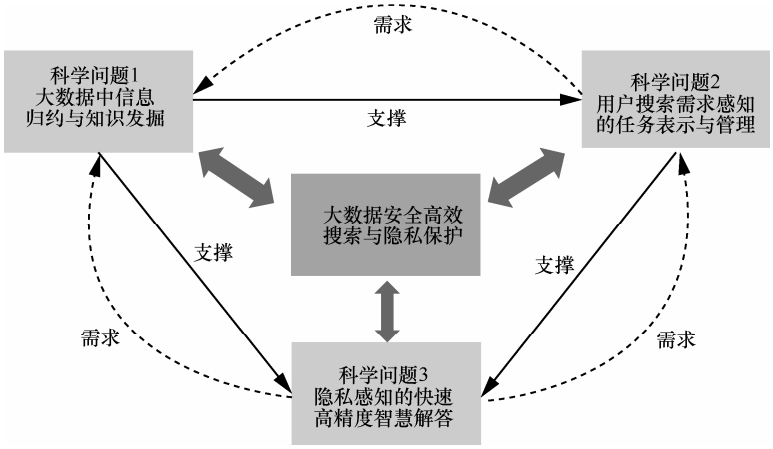

图 2　科学问题及其相互关系

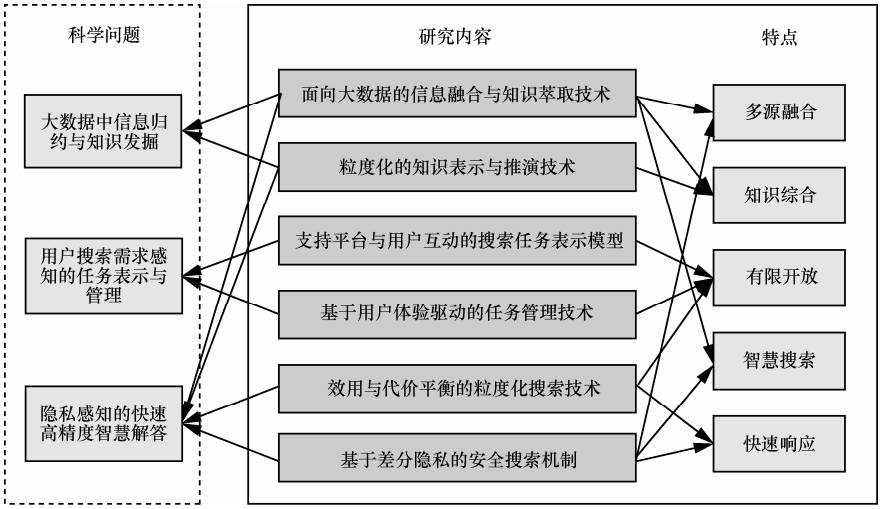

图 3　主要研究内容与科学问题的对应关系

示。主要研究内容具体包括：面向大数据的信息融合与知识萃取技术、粒度化的知识表示与推演技术、支持平台与用户互动的搜索任务表示模型、基于用户体验驱动的任务管理技术、效用与代价平衡的粒度化搜索技术和基于差分隐私的安全搜索机制。

3.2.1　面向大数据的信息融合与知识萃取技术

针对大数据信息具有多维度和多粒度的特点以及用户对大数据搜索需求的多样性（和不可预测性），需要对信息进行预先融合与知识萃取，生成大数据知识聚合体，并进行有效组织，从而支持实时多维度多粒度的搜索请求。具体内容如下。

1) 隐私感知的多源信息融合技术。针对大数据渠道多源化、信息多维度、知识多粒度的特点及大数据搜索服务对不同用户的“有限开放”性，研究粒度化的、支持增量推演的信息融合隐私保护机制，为“千人千面”的搜索结果展现提供基础。

2) 高维属性空间的知识聚合计算技术。针对大数据知识随信息层次、属性维度增加呈指数增长的特点，基于粗糙集、熵计算等知识挖掘理论，研究大数据属性约简方法，并根据各粒度知识间的偏序关系，设计知识聚合计算方法。

3) 支持时空特性的知识更新演化技术。针对大数据自身动态更新及用户搜索需求不断变化的特点，研究知识聚合体的结构化缓存机制，以及知识聚合体自我演化与更新技术。

3.2.2　粒度化的知识表示与推演技术

针对大数据的知识存在高维度、属性依赖、时空演化等特点，构建大数据所蕴含知识的粒度化表示模型，研究面向时空属性的知识推演方法及基于决策依赖的大规模属性推理方法。具体内容如下。

1) 大数据属性间复杂关系的粒度化知识表示方法。针对大数据属性规模大、关系巨复杂、知识粒度化且随时间演化等特点，包括属性间的关联关系、映射关系、偏序关系、基于度的依赖关系等，分析各类关系的不同特点及其随数据更新的演化规律，研究多层次、粒度化的知识表示模型。

2) 面向大数据知识时空属性的推演方法。研究在用户跨平台（子网）数据发布过程中对隐私信息泄露风险的评估预警方法。

3) 基于决策依赖的大规模属性推理方法。针对大数据属性规模大、服务用途多样的特点，研究以用户搜索为导向、“核属性”为主体的属性依赖关系挖掘、冲突消解及知识覆盖、推理技术。

3.2.3　支持平台与用户互动的搜索任务表示模型

针对现有平台无法有效感知用户个性化需求，且平台特征参数及运行状态对用户透明的现实，以探求可行的用户搜索请求为目标，构建与平台无关的任务表示模型，研究基于平台刻面的任务计算代价预估模型，化解用户搜索需求中的冲突因子。具体内容如下。

1) 跨平台自适应的（与平台无关的）任务模型表示方法：针对用户搜索需求多样及数据存储平台异质的现实，通过对用户搜索需求的抽象化表示及对平台属性、状态刻面的泛化表示，构建搜索需求与平台刻面统一融合的表示计算模型，为任务间定量可比的预估及计算提供基础。

2) 基于平台刻面的任务预估与冲突消解模型：针对平台特征参数及运行状态对用户透明的现实及用户不具备任务预估专业知识的既定假设，以任务需求与平台刻面统一融合的表示模型为基础，研究以平台刻面为主体，任务所需资源、执行时间、搜索精度 3 个目标预估及优化模型，以及面向多目标刻面的冲突消解方法。

3.2.4　基于用户体验驱动的任务管理技术

针对多用户搜索任务的数据分布特征、对资源的时变需求以及任务间的覆盖关系，研究任务类型感知的多队列任务动态管理技术及基于气球驱动机制的抢占式资源分配技术，实现用户与平台双赢的资源动态管理机制。具体内容如下。

1) 任务需求感知的多队列任务动态管理技术。针对多用户搜索任务对平台优先调度的普遍需求，通过分析搜索任务间覆盖关系及数据分布特征，以最大化保障全局用户体验为愿景，研究优先级刻画方法与多队列的任务管理技术。

2) 基于气球驱动机制的抢占式资源分配技术。针对多用户搜索任务对资源的时变需求，通过构建平台资源的全局视图，以任务当前优先级刻面为基准，研究基于气球驱动机制的抢占式动态资源分配技术。

3) 用户与平台双赢的资源动态管理技术。针对同时满足用户搜索体验与平台资源利用率最大化的双边应用需求，以多队列任务动态管理技术及抢占式资源分配技术为基础，研究基于供需零和博弈的多目标优化调度技术。

3.2.5　效用与代价平衡的粒度化搜索技术

针对用户对大数据搜索高精度、低成本的利益需求，基于统计抽样理论和方法，实现满足用户求解精度需求的快速搜索技术，同时研究粒度化历史结果复用机制以及与数据更新联动的粒度化视图维护机制。具体内容如下。

1) 支持快速求解的高精度搜索技术。针对用户对大数据搜索结果的高精度、快响应的现实需求，以伯努利、重采样等统计抽样理论为依据，构建样本容量与响应时间、求解精度联动的可量化模型，实现同时满足效用与代价约束的搜索技术。

2) 基于粒度化视图的结果复用机制。针对用户搜索请求普遍存在覆盖、偏序、同构的特征，研究粒度化历史结果视图的缓存与复用机制，以进一步减小计算代价、提高响应速度为目标，设计以历史结果视图为基础的增量求解模型。

3) 支持数据更新的粒度化视图维护机制。针对大数据基数庞大、量变有限的特点，探求粒度化结果视图的可复用边界，研究具有强顽健性的粒度化视图维护机制，从而为历史结果的柔性复用机制提供理论支撑。

3.2.6　基于差分隐私的安全搜索机制

针对大数据搜索生命周期的各个阶段，包括数据发布、信息融合、数据检索、数据更新与维护，着力研究面向多源隐私信息的融合理论，设计面向快速高精度搜索的柔性差分隐私保护机制，实现基于差异化隐私预算的访问控制策略及

支持数据增量发布的差分隐私预算分配机制。

1) 面向快速高精度搜索的差分隐私保护机制。针对差分隐私保护机制的严苛约束与快速高精度的搜索需求的不匹配问题，研究精度测量与隐私保护评价统一融合的度量机制，设计面向高精度快速搜索的柔性差分隐私保护策略。

2) 基于差异化隐私预算的访问控制策略。针对大数据“有限开放”的搜索机制及“千人千面”视图展现的现实需求，为防范以大数据“微”查询为代表的个体隐私泄露攻击，研究基于交互式差异化隐私预算动态配给的访问控制策略。

3) 数据增量发布感知的差分隐私预算分配机制。针对大数据持续生成、频繁更新对既有数据隐私保护程度造成的冲击，研究数据增量与数据隐私保护粒度的动态平衡模型及隐私保护顽健稳定性边界，设计支持数据增量发布的差分隐私预算分配机制。

4 技术路线展望

4.1 融合统一的隐私度量标准

与传统的基于关键字的 Web 搜索不同，由于大数据具有体量巨大、内涵知识丰富且有限开放的特点，导致用户对大数据搜索的需求不再是单一维度的存在性解答，而是从搜索结果的精度、可用性以及搜索的时效性等多个维度进行评价，使传统针对单一维度的搜索评价指标不再适用于对大数据搜索的评判。此外，从大数据搜索的生命周期看，期间包括多源数据融合隐私保护、用户搜索行为隐私保护、搜索过程中数据拥有者对不同用户设定的细粒度访问控制等，这些隐私保护技术虽然都有各自的评价指标，但这些指标难以相互量化转化。因此，需要构建一套集成多维隐私保护、搜索精度、时效三位一体的度量体系。

首先，统一各隐私保护策略的度量指标。目前，主流的隐私保护方案可以分为两类：基于 k-匿名及其变种的隐私保护方案和基于差分隐私的保护方案，可通过对这两类隐私保护方案对数据可用性的度量来对其进行融合。即基于 k-匿名及其变种的隐私保护方案通过加噪、泛化等手段保护隐私，可以通过噪声比，信息泛化树来量化表示隐私保护粒度与信息可用性；而差分隐私保护方法可以直接通过隐私预算来确定加噪比，因此，可以构建 k-匿名类隐私保护算法与差分隐私保护算法在信息可用性度量上的映射。

其次，统一隐私保护度量与搜索精度的度量指标，拟通过扩展差分隐私的定义来实现。差分隐私假设这样一个场景：设有 2 个几乎完全相同的数据集（两者的差别仅在于一个记录不同），分别对这 2 个数据集进行查询访问，同一查询在 2 个数据集上产生同一结果的概率比值接近于 1。而基于抽样理论，搜索精度天然地可用(ε, δ)-原语来量化，其中，ε 表示误差边界，δ 表示抽样误差大于 ε 的概率。这里，可对差分隐私的定义增加一个松弛因子 δ，这样不但增强了差分隐私保护方法在现实应用领域的适用性，而且与搜索精度的量化表示建立了一一映射。

最后，统一搜索精度与搜索时效的度量指标。在现实的搜索应用场景中，由于搜索对象（数据集）、可用资源（内存、带宽等）受限，根据中心极限定理，当搜索精度确定时，可以确定抽样样本个数，因此，抽样与求解的时间可以预估，从而将搜索精度与搜索时效建立映射。

4.2 支持多源信息增量融合的隐私反推演技术

针对面向多源信息发布下的隐私泄漏预判方法，可采用基于冲突图的超图消解算法，构造相容性规则，对已发布的数据进行等价类划分。通过分析属性间的覆盖、偏序等关系，构建信息属性辨识矩阵，基于粗糙集、熵计算等知识挖掘理论，在进行属性约简、构建粒度化的知识聚合体的同时，从中得出信息属性与敏感属性间的关联冲突规则集，然后将每一个属性视为超图的一个顶点，每一个冲突规则视为超图的一条超边。最后通过最优超边消解原则，开发基于贪心策略的超边消解启发式算法，获得可能引发隐私泄露的最小属性集。

针对多源信息融合过程的隐私保护，可采用基于轮询迭代的加细信息匿名隐私保护策略，主要思想是：数据融合的各方就自己所拥有的本地数据计算各属性的信息熵并公布最大的熵值进行比较，各方选出本轮全局熵值最大的属性。该属性的所有者基于上一轮的数据划分结果对其进行

加细划分，若划分结果不违背数据匿名约束，则公布划分结果，否则直接进行下一轮，直至没有属性能在满足匿名约束的前提下对数据加细划分产生贡献。数据的匿名性是由算法本身保证的。基于轮询迭代的加细信息匿名隐私保护策略对数据实施自顶向下逐步细化，在交互过程中，每轮具有全局信息增益最大属性的用户严格遵照匿名门限细化数据集，第 q 轮细化的最终结果就是融合数据 T，且易知前 $q-1$ 轮的细化结果集都比 T 粗糙，即租户在交互过程中不可能学到比集成数据表 T 更多的知识。

针对增量数据发布可能诱发的隐私泄露风险，可通过整合 4.1 节提出的隐私保护强度与数据精度融合统一的度量机制，分析增量数据对既有数据隐私保护程度造成的冲击，量化隐私保护顽健稳定性边界，基于差分隐私策略，设计一套隐私预算增量分配机制，以实现在满足数据增量发布隐私保护的前提下最大化数据可用性的现实需求。

4.3　支持时空特性的多维细粒度访问控制机制

大数据搜索中用户搜索需求多元化且动态多变，针对传统的基于用户静态属性的访问控制模型（如 CP-ABE）扩展性较差、难以继续适用的问题，可采用如下方案进行解决：1) 采用层次聚类法自动挖掘用户静态角色属性，通过分析角色属性构造概念格从而引导权限集的聚合，再引入基于场景驱动的方法分析具体业务场景，将需求信息以语义或者启发式规则的方式对自底向上得到的结构进行优化，增强属性挖掘结果的准确性；2) 基于属性挖掘的结果，结合用户搜索精度、时效、成本等动态需求，构建用户的动态“画像”并给出形式化描述；3) 设计给定计算资源场景下搜索精度与时效的预估算法，基于计算资源与用户“画像”的联动模型，粒度化地控制呈现给用户的搜索结果视图，从而建立适用场景广、动态可扩展的多维细粒度访问控制模型。

大数据多源融合的特性，引入了多个授权机构的局部数据访问控制规则，对于各访问控制规则间的冲突消解机制，可按照以下方案展开研究：1) 建立各授权机构访问控制规则中数据、授权用户、访问权限等属性的统一形式化表示模型，量化研究不同粒度访问控制的安全风险与数据效用；2) 在访问规则属性统一表示模型的基础上，构建访问控制属性辨识矩阵，从中得出访问控制冲突属性集；3) 采用基于冲突图的超图消解算法处理访问控制冲突属性集，从而统一融合各授权机构差异化的访问控制规则。

对支持时空特性的访问控制演化机制的研究，可从以下方面展开：1) 归纳和分析大数据搜索场景中各类具有时空特性的访问控制要素，如数据拥有者的授权体系、用户的动态访问权限及数据隐私保护需求等，抽象出时空特性在访问控制模型中的表达方法；2) 基于对用户搜索意图的分析，研究授权及权限更新等操作对访问者及数据等所带来的风险，分析风险间的关联关系，基于风险评估研究授权与权限的实时更新方法；3) 在粒度化的访问控制视图的基础上，进行时空相关的属性约束扩展，包括角色时间属性、用户位置属性、上下文时间属性等；4) 整合搜索周期中涉及的信任、安全、隐私保护机制，设计支持时空特性的风险计算方法及权限更新策略，进而实现自适应、可重组、动态可调整的访问控制机制。

4.4　用户搜索意图保护的安全搜索技术

在大数据搜索中，对用户搜索意图的准确理解是提供高质量搜索结果的保障，但用户表达其真实意图的同时可能会暴露位置、行为等隐私。

对于搜索意图的表示方法与隐私检测机理，可从以下 2 个方面展开研究：1) 以个性化辨识属性集为目标对用户意图进行建模，结合用户的行为和语境信息，构建搜索目标“与/或树”，围绕这个主结构伸展意图，实现用户搜索意图的显式量化表示方法；2) 从现实的大数据搜索应用出发，基于抽取用户搜索意图中包含的各类实体，建立搜索意图及位置、行为等隐私属性的二元结构与用户实体的映射关系，通过频繁模式挖掘对映射关系进行关联分析，从而检测出隐含在用户搜索意图中的隐私信息，实时提供行为隐私、身份隐私及位置隐私的泄露预警。

针对用户对搜索质量和隐私保护的二元需求，可按照以下方案研究用户搜索意图保护策略：1) 基于语义树设计搜索意图的模糊及泛化方法，并量化分析该方法对搜索质量的影响；2) 基于扩展的差分隐私定义，根据搜索质量与隐私保护预算的映射关系图，得到对应于给定搜索质量的隐私保护预算，并使用基于高斯分布的加噪机制替代差分隐私保护原有的拉普拉斯加噪方案；3) 从搜索意图中抽取出行为、身份、位置等隐私属性，基于行为隐私、身份隐私及位置隐私联动的保护机理，实现用户与其搜索意图解耦合的安全搜索模型。

在大数据搜索中集成面向数据和用户意图的隐私保护技术，同时保证用户对搜索精度、时效的需求，以实现高可用的大数据安全搜索架构。这部分的研究可从如下方面展开：1) 基于隐私保护需求对搜索结果精度的刻画机制，以伯努利、重采样等统计抽样理论为依据实现近似求解，通过构建样本容量与精度、时效联动的可量化模型，实现同时满足隐私、精度、时效约束的搜索模型；2) 针对存在覆盖、偏序、同构关系的用户搜索请求，设计粒度化历史结果视图的缓存与复用机制，建立以历史结果视图为基础的增量求解模型，实现支持数据更新的历史结果视图维护机制；3) 分析多用户搜索任务的数据分布特征、对资源的时变需求以及任务间的覆盖关系，实现优先级刻画方法与多队列的任务管理技术，通过构建平台资源的全局视图，以任务当前优先级刻面为基准，基于气球驱动机制实现抢占式的动态资源配给技术。

5 结束语

大数据搜索以数据体量巨大、知识蕴含丰富为“形”，多粒度、高效智慧地满足用户需求体验为“神”，这一理念的兴起革新了以往对安全高效搜索与隐私保护的求解思路。一方面，用户在精度、时效、能耗等方面对大数据搜索提出新的期许；另一方面，数据拥有者急需适用于大数据环境下的访问控制机制及数据隐私保护方案。本文从大数据多源融合发布、用户搜索需求感知及隐私感知的智慧解答 3 个方面综述了国内外最新的研究进展，并以此为基础，从大数据安全高效搜索与隐私保护全局一体化的视角凝练了亟待解决的科学问题，提出了 6 个富有挑战性的研究议题，最后展望了相关的技术路线，以期为该领域的研究提供借鉴价值。

参考文献：

[1] SWEENEY L. *K*-anonymity: a model for protecting privacy[J]. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.

[2] MACHANAVAJJHALA A, KIFER D, GEHRKE J, et al. *L*-diversity: privacy beyond k-anonymity[J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 1-47.

[3] LI N, LI T, VENKATASUBRAMANIAN S. Closeness: a new privacy measure for data publishing[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(7): 943-956.

[4] 周水庚，李丰，陶宇飞，等．面向数据库应用的隐私保护研究综述[J]．计算机学报, 2009, 32(5): 847-861.
ZHOU S G, LI F, TAO Y F, et al. Privacy preservation in database applications: a survey[J]. Chinese Journal of Computers, 2009, 32(5): 847-861.

[5] XIAO X K, TAO Y F. *M*-invariance: towards privacy preserving republication of dynamic datasets[C]//International Conference on Sigmod, New York. c2007: 689-700.

[6] JIANG W, CLIFTON C. A secure distributed framework for achieving anonymity[J]. The International Journal on Very Large Data Bases, 2006, 15(4): 316-333.

[7] CLIFTON C, KANTARCIOGLU, VAIDYA J. Tools for privacy preserving distributed data mining[J]. ACM Sigkdd Explorations Newsletter, 2010, 4(2): 1-7.

[8] MOHAMMED N, FUNG B C M, DEBBABI M. Anonymity meets game theory: secure data integration with malicious participants[J]. VLDB Journal, 2011, 20(4): 567-588.

[9] MOHAMMED N, FUNG B C M, et al. Centralized and distributed anonymization for high-dimensional healthcare data[J]. ACM Transactions on Knowledge Discovery from Data, 2010, 4(4):885-900.

[10] JIANG W, CLIFTON C, KANTARCIOGLU M. Transforming semi-honest protocols to ensure accountability[C]// IEEE International Conference on Data Mining Workshops (ICDM). c2006: 524-529.

[11] DU W, HAN Y S, CHEN S. Privacy-preserving multivariate statistical analysis: linear regression and classification[C]//The SIAM International Conference on Data Mining, Florida. c2004:222-223.

[12] PINKAS B. Cryptographic techniques for privacy-preserving data mining[J]. ACM Sigkdd Explorations Newsletter, 2003, 4(2): 12-19.

[13] VAIDYA J, CLIFTON C. Privacy-preserving *k*-means clustering over vertically partitioned data[C]//ACM Sigkdd International Conference on Knowledge Discovery & Data Mining. c2003: 206-215.

[14] BRODER A Z, FONTOURA M, GABRILOVICH E, et al. Robust classification of rare queries using web knowledge[C]//The 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. c2007: 231-238.

[15] KOHLSCHÜTTER C, CHIRITA P A, NEJDL W. Using link analysis to identify aspects in faceted web search[J]. Sigir Faceted Search Workshop. 2006.

[16] SHEN D, PAN R, SUN J T, et al. Query enrichment for web-query classification[J]. ACM Transactions on Information Systems, 2006, 24(3): 320-352.

[17] TEEVAN J, KARLSON A, AMINI S, et al. Understanding the importance of location, time, and people in mobile local search behavior[C]//International Conference on Human Computer Interaction with Mobile Devices & Services. c2011:77-80.

[18] BRODER A, FONTOURA M, JOSIFOVSKI V, et al. A semantic approach to contextual advertising[C]//The International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, the Netherlands. c2007:559-566.

[19] SARKAS N, PAPARIZOS S, TSAPARAS P. Structured annotations of Web queries[C]//ACM SIGMOD International Conference on Management of Data. c2010:771-782.

[20] STANKOVIC J A. Research directions for the internet of things[J]. Internet of Things Journal, 2014, 1(1): 3-9.

[21] MAEKAWA T, YANAGISAWA Y, SAKURAI Y, et al. Context-aware Web search in ubiquitous sensor environments[J]. ACM Transactions on Internet Technology, 2012, 11(3): 1-23.

[22] YAO L, SHENG Q Z, FALKNER N J G, et al. ThingsNavi: finding most-related things via multi-dimensional modeling of human-thing interactions[C]//The 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services. c2014: 20-29.

[23] OSTERMAIER B, ROMER K, MATTERN F, et al. A real-time search engine for the web of things[C]//2010 Internet of Things (IOT). c2010: 1-8.

[24] CHRISTOPHE B, VERDOT V, TOUBIANA V. Searching the 'web of things'[C]//The 5th International Conference on Semantic Computing. c2011: 308-315.

[25] Smart. Search engine for multimedia environment generated content[EB/OL]. http://www.smartfp7.eu/.

[26] HAN X, LIU X, LI J, GAO H. TDEP: efficiently processing top-*k* dominating query on massive data[J]. Knowledge & Information Systems, 2014, 43(3): 689-718.

[27] HAN X, LI J, GAO H, YANG C. SEPT: an efficient skyline join algorithm on massive data[J]. Knowledge & Information Systems, 2015, 43(2): 355-388.

[28] HAN X, LI J, YANG D. PI-join: efficiently processing join queries on massive data[J]. Knowledge & Information Systems, 2012, 32(3): 527-557.

[29] CHAUDHURI S, DAS G, SRIVASTAVA U. Effective use of block-level sampling in statistics estimation [C]//ACM SIGMOD International Conference on Management of Data, Paris. c2004: 287-298.

[30] DOUCET A, BRIERS M, SNCAL S. Efficient block sampling strategies for sequential monte carlo methods[J].Journal of Computational and Graphical Statistics, 2006,15(3):693-711.

[31] CONDIE T, CONWAY N, ALVARO P, et al. Online aggregation and continuous query support in MapReduce[C]//ACM SIGMOD International Conference on Management of Data. c2010: 1115-1118.

[32] LAPTEV N, ZENG K and ZANIOLO C. Early accurate results for advanced analytics on MapReduce[J]. The VLDB Endowment, 2012, 5(10): 1028-1039.

[33] HERODOTOU H, DONG F, and BABU S. MapReduce programming and cost-based optimization? crossing this chasm with starfish[J]. The VLDB Endowment, 2011(4):1446-1449.

[34] HERODOTOU H, BABU S. Profiling, what-if analysis, and cost-based optimization of mapreduce programs[J].The VLDB Endowment, 2011(4): 111-1122.

[35] YU S, REN K, LOU W. FDAC: toward fine-grained distributed data access control in wireless sensor networks[J]. IEEE Transactions on Parallel and Distributed Systems, 2011, 22(4): 673-686.

[36] ZHANG R, ZHANG Y, et al. Distributed privacy-preserving access control in sensor networks[J]. IEEE Transactions on Parallel and Distributed Systems, 2012, 23(8): 1427-1438.

[37] FRIAS-MARTINEZ V, SHERRICK J, STOLFO S J, et al. A network access control mechanism based on behavior profiles[C]// Annual Computer Security Applications Conference. c2009: 3-12.

[38] GOYAL V, PANDEY O, SAHAI A, et al. Attribute-based encryption for fine-grained access control of encrypted data[C]//The 13th ACM CCS. c2006: 89-98.

[39] YANG K, JIA X. Expressive, efficient, and revocable data access control for multi-authority cloud storage[J]. IEEE Transactions on Parallel and Distributed Systems, 2014, 25(7): 1735-1744.

[40] CAMENISCH J, LEHMANN A, NEVEN G, et al. Privacy-preserving auditing for attribute-based credentials[C]// Computer Security-ESORICS. c2014: 109-127.

[41] MAJI H. Attribute-based signatures: achieving attribute-privacy and collusion-resistance[J]. IACR Cryptology ePrint Archive, 2008, (4): 1-23.

[42] BUGIEL S, HEUSER S, SIT F. Flexible and fine-grained mandatory access control on android for diverse security and privacy policies[C]//Usenix Conference on Security. c2013:131-46.

[43] ROHRER F, ZHANG Y, CHITKUSHEV L, et al. DR BACA: dynamic role based access control for android[C]//The 29th Annual Computer Security Applications Conference. c2013: 299-308.

[44] SINGH K. Practical context-aware permission control for hybrid mobile applications[M]//Research in Attacks, Intrusions, and Defenses. Berlin Heidelberg: Springer, 2013: 307-327.

[45] AVE S. Auto-FBI: a user-friendly approach for secure access to sensitive content on the Web[C]//The 29th Annual Computer Security Applications Conference. c2015: 349-358.

[46] HU H, AHN G, et al. Multiparty access control for online social networks: model and mechanisms[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(7): 1614-1627.

[47] DWORK C, NAOR M, VADHAN S. The privacy of the analyst and the power of the state[C]//IEEE 53rd Annual Symposium on Foundations of Computer Science. c2012: 400-409.

[48] MARIAS G F, DELAKOURIDIS C, KAZATZOPOULOS L, et al. Location privacy through secret sharing techniques[C]//The 6th IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks. c2005: 614-620.

[49] WERNKE M, DURR F. PShare: position sharing for location privacy based on multi-secret sharing[C]//IEEE International Conference on Pervasive Computing and Communications. c2012: 153-161.

[50] SANTOS V, BAIÃO F, TANA A. An architecture to support information sources discovery through semantic search[C]//IEEE International Conference on Information Reuse&Integration. c2011: 276-282.

[51] ZHANG B, ROSS B, TRIPATHI S, et al. Network-aware data caching and prefetching for cloud-hosted metadata retrieval[C]// The Third International Workshop on Network-aware Data Management. c2013: 1-10.

[52] SABRINA D C D V, FORESTI S, JAJODIA S, et al. On information leakage by indexes over data fragments[C]//The 29th International Conference on Data Engineering Workshops, IEEE. c2013: 94-98.

[53] BETHENCOURT J, SAHAI A and WATERS B. Ciphertext-policy attribute-based encryption[C]//IEEE Symposium on Security and Privacy. c2007: 321-334.

[54] YANG K, JIA X, REN K, ZHANG B. DAC-MACS: Effective data access control for multi-authority cloud storage systems[J].IEEE Transactions on Information Forensics & Security, 2014, 8(11): 2895-2903.

[55] SANKAR L, RAJAGOPALAN S R, POOR H V. A theory of utility and privacy of data sources[J]. IEEE International Symposium on Information Theory, 2010, 41(3): 2642-2646.

[56] GUO S, CHEN K. Mining privacy settings to find optimal privacy-utility tradeoffs for social network services[C]//International Conference on Privacy, Security, Risk and Trust .c2012: 656-665.

[57] GU Y, WU W. A quantifying method for trade-off between privacy and utility[C]//The International Conference on Information and Communications Technologies. c2013:270-273.

作者简介：

李尚（1989-），男，山东济宁人，哈尔滨工业大学博士生，主要研究方向为大数据安全搜索。

周志刚（1986-），男，山西太原人，哈尔滨工业大学博士生，主要研究方向为云安全和隐私计算。

张宏莉（1973-），女，吉林榆树人，博士，哈尔滨工业大学教授、博士生导师，主要研究方向为网络与信息安全、网络测量与建模、网络计算、并行处理等。

余翔湛（1973-），男，黑龙江哈尔滨人，博士，哈尔滨工业大学研究员、博士生导师，主要研究方向为网络容灾、信息安全、物联网安全等。