doi:10.11959/j.issn.1000-436x.2015311

智慧搜索中的实体与关联关系建模与挖掘

王晓阳，郑骁庆，肖仰华
（复旦大学 计算机科学技术学院 上海市数据科学重点实验室，上海 201203）

摘 要：随着网络搜索空间从互联网扩展到人、机、物互联的泛在网络空间，以及大数据时代的到来，传统的搜索引擎已经不能满足时代的需求，新时代的搜索引擎技术——大搜索（或称智慧搜索）概念应运而生。因此，讨论实现大搜索所需关键技术之一的实体与关联关系建模与挖掘，以及相关的设计思想和实现技术。
关键词：大搜索；实体与关系建模；知识图谱；知识仓库
中图分类号：TP393 **文献标识码**：A

Entity-relation modeling and discovery for smart search

WANG X Sean, ZHENG Xiao-qing, XIAO Yang-hua
(Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai 201203, China)

Abstract: Nowadays, by connecting the mobile networks, Internet of Things and the sensor networks to the Internet, the cyberspace has expanded to a ubiquitous space of human beings, machines and things. Combining with the technology of big data, the traditional search engines are evolving into their next generation—big search (or smart search). Entity-relation modeling and discovery are the key techniques to fulfill the vision of smart search. Approaches to model the entities and their relations in large scale by knowledge graph and knowledge warehouse, and ways to discovery new entities and the relations between them in the cyberspace are discussed.
Key words: big search; entity-relation modeling; knowledge graph; knowledge warehouse

1 引言

自万维网（World Wide Web）诞生以来，经历半个多世纪的迅速发展与演化，其形式内容与应用模式都发生了显著的变化。网络应用模式从由专业人员开发、以高访问量为目标的综合门户网站为主导的 Web1.0 时代，发展至众人皆可参与、高度交互的社交媒体 Web 2.0 时期。万维网正在向更高级的、以语义和智能技术应用为代表的 Web3.0 发展，更加强调通过综合多源异质信息，以提供个性化的智能解答与服务。

与此同时，大数据概念及技术迅速渗入社会各层面。大数据的目标是从存在“噪声”的海量多源异质异构数据中，自动高效地发掘有价值的信息。将大数据分析中技术共性部分抽取出来，加以扩展，开发新一代面向网络空间的搜索引擎，推进搜索引擎向对象多元化、数据多样化、信息融合化、解答智能化的方向发展，从而能够提供契合用户搜索意图的智慧解决方案——“大搜索”的概念也应运而生[1]。

大搜索或称“智慧搜索”，指的是根据搜索请求，在网络空间中进行搜索，形成相应的智慧解决方案，最后返回以解决方案为搜索结果的过程。它与传统搜索最大的不同在于：它的搜索内容和对象由传统的文本信息扩展到了物体、信息和人物，以及他们之间的关联关系；它要求从网络空间中获取智能解答方案而非简单的返回相关网页。

实现智慧搜索面临以下挑战。1）网络空间的数据获取与组织。当前网络空间中所描述的实体对

收稿日期：2015-10-09；**修回日期**：2015-12-10
基金项目：国家自然科学基金资助项目（61370080）；上海市自然科学基金资助项目（13ZR1403800）
Foundation Items: The National Natural Science Foundation of China (61370080); Shanghai Municipal Natural Science Foundation (13ZR1403800)

象（如人、物、概念、事件等）及关联关系（如朋友、购买、参与等）的数量巨大、种类繁多。数据来源可包括互联网、社交网络、时空数据、企业、运营商等。智慧搜索需要融合多渠道、多模式的各种类型数据，挖掘和发现其中潜在的、有价值的信息，并且形成相应的知识框架及索引体系，以便于搜索、查询与利用。2）用户意图的准确理解。用户查询输入方式多样，充满了语义方面的歧义。这需要智慧搜索能够洞察与理解用户真实的搜索意图，在海量、多源、异构、多态的数据中，利用他们之间语义关联关系，实现实体对象及其关联关系相关信息的有效搜索，提供最贴合用户需求的搜索结果。3）满足用户查询需求的智慧方案形成。传统搜索引擎一般只能为用户提供符合搜索要求的存在性信息（相关的网页），而用户的意图具有多样化、个性化等特点，需要根据其意图形成一系列可供选择的智慧解决方案。这需要实现搜索解答方案的智慧化，为用户求解出智慧答案。因而如何根据用户的搜索意图，基于知识仓库对有关知识进行求解，通过推理演算形成若干综合的智慧解决方案则成为智慧搜索技术的关键所在。

应对上述智慧搜索技术的挑战，一个重要的任务就是对实体对象及关联关系进行建模，将网络空间包含的各类实体关联知识用有效的组织方式存储，以支持智慧搜索。这里，“实体对象”或简称“实体”应被理解为广义的对象，包含世界中客观存在的事物以及人类思维空间中的概念，他们之间相互作用、制约，由此形成一定的“关联关系”或简称“关联”。实体可以是名人、城市、球队、电影、地标性建筑、艺术品、概念、事件等，关联则可以是人与人、概念与地点、人与物品以及地点与物品等之间存在的关系。利用实体以及他们之间的关联，不仅可以提高搜索精度和优化搜索结果，还可以支撑语义分析、关联分析、知识搜索和智能推荐等高层的服务。

简单地说，实体对象与关联关系建模就是要从网络空间中抽取实体及关联信息，形成知识库。这是个工业界及学术界共同关心的问题，谷歌和百度的知识图谱、搜狗的知立方都是这类知识库的实例。表 1 显示部分公开的知识图谱及它们的规模。广义上讲，这个建模问题本质上是解决如何使用计算机进行大规模多源知识的获取、组织和使用的问题。它的必要性表现在以下方面。

1) 实体对象及关联关系建模是跨越语义鸿沟的关键，背景知识缺乏是语义鸿沟难以跨越的一个重要原因。现有机器可读的知识库在质量上和完整性方面仍然难以达到人类语义理解的基本水平，但近年来研究开发的基于知识图谱的知识库，相对于传统知识表示方法，在兼顾精准性的同时，在完整性方面取得了长足的进步，它为用户意图理解、语义消歧、信息整合等提供了必要的背景知识，使征服语义鸿沟又前进了一步。谷歌等搜索引擎已将基于知识图谱的知识库成功用于提高搜索结果准确性。

2) 实体对象及关联关系建模是知识有效运用的基础。网络空间所涉及的实体数巨大，已有的知识库中实体数已达千万量，关联数则以亿计，它们所形成的是典型的异构信息网络。实体对象及关联关系建模呈多模形态，常常需要用某种测度来表达实体及关系的出现频率、强度等信息；需要用边的方向表达关系的非对称性；需要用概率体现数据源的不确定性等。上述特征对于实体对象及关联关系模型提出更高的要求，设计良好的模型是其上进行高效查询、更新和推理的基础。

3）实体对象及关联关系建模是搜索智慧化的前提。实体对象及关系模型相对于领域本体和传统语义网络而言，其实体覆盖率更高，语义关系也更加全面而复杂。利用实体对象及关联关系可以对搜索结果行系统的语义分析，将用户查询映射到知识库的概念上，从而用于优化搜索结果。还可利用已

表 1 **公开的知识图谱**

名称	内容	规模	网址
DBpedia	维基百科	1 900 万实体、1 亿关系	dbpedia.org
Freebase	多种网络数据源	6 800 万实体、10 亿关系	www.freebase.com
Wiki-links	维基百科	4 000 万已消岐的关系	code.google.com/p/wiki-links
Data.gov	美国政府官方数据	64 亿关系	www.data.gov/semantic/index
WolframAlpha	计算知识	10 万亿实体	www.wolframalpha.com

知的实体对象及关系进行推理，产生新知识，这种能力是问题解答、自动服务生成、智慧方案形成等的技术前提。

以下介绍与讨论智慧搜索中实体对象及关联关系建模相关的关键技术与方法，其技术之间的关系如图 1 所示。

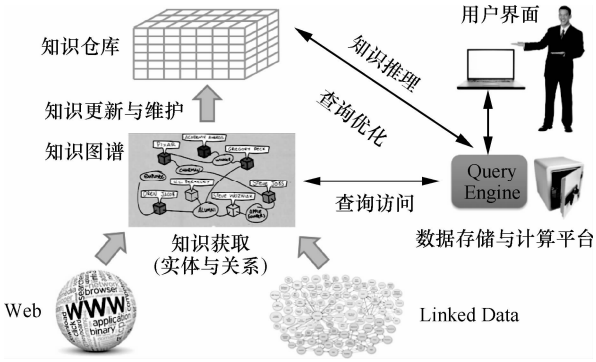

图 1　智慧搜索中实体对象及关联关系建模关键技术关联

2　知识图谱

实体或概念是世界中客观存在的事物，他们之间相互作用、制约，由此形成一定的关系。实体与关系建模本质上是解决如何使用计算机进行大规模多源知识的表示、获取和使用的问题。目前，实体对象及其关系建模工作较多地围绕知识图谱展开。

知识图谱是采用语义检索技术从多种信息源收集与某一主题相关的实体或概念，以及他们之间的关联所形成的网络图。图中的节点对应实体或概念，图中的弧对应实体或概念之间的关联关系。

大搜索借助知识图谱，通过深化现实世界中每个实体以及他们之间相互关系的理解，提高搜索精度和优化搜索结果。语言的歧义性会给搜索带来了困难，例如当用户输入查询词“苹果”，传统搜索引擎无法理解用户想要查询的是水果还是公司。基于知识图谱的智能搜索将所有这些可能性归纳分组，用户仅需点击其中一组即可看到针对特定含义的所有搜索结果。有了知识图谱，搜索引擎可以更好地理解用户的查询，从而提供与该查询更相关的内容，即根据不同的实体，展示最相关的事实。如图 2 所示，当用户搜索“Marie Curie”（居里夫人）时，不仅可以看到与居里夫人相关的网页，还可以看到有关居里夫人教育经历、科学贡献和社会关系等信息。利用知识图谱还可以提供语义分析、关联分析、知识搜索和智能推荐等知识服务。

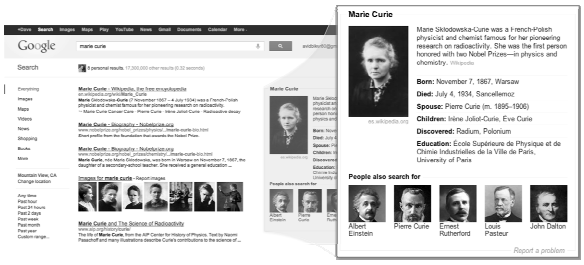

图 2　知识图谱优化搜索结果的例子（摘自 Google 搜索结果）

3　知识获取

知识图谱需要各种自动化知识获取方法来补充相关的知识（即实体及其关系），其中存储的知识越丰富，则解决问题的能力也越强。关联信息发掘是一种面向任务的信息获取方式，是指以一定的策略和方法去采集、获取、发掘用户需要的数据与信息的过程。关联信息发掘的工作过程如下。首先，以已有的知识图谱为引导，把所有可能的数据源都搜集起来，包括互联网上的网站、社交网络、网络服务等，以及物联网、视音频监控的数据等，并且针对每一种数据源设计相应的数据获取方式，如网络爬虫方式、API 数据获取方式等；之后，对所有数据源进行分类，类别层次是一个多层次多维度的分类过程，根据用户需求的变化，数据源类别层次应能做相应调整；当接受到用户的定向获取任务时，根据用户需求确定数据源的类别，并在相应类别的数据源中进行基于任务的数据获取；最后，对所有数据源获取的数据进行结果的综合，包括去重、清洗、结果融合等，并把最终结果返回给用户，并且对其中共性的内容用于更新已有知识图谱。

例如：通过搜索意图理解确定用户关心“达芬奇”相关的信息，则在互联网上获取维基百科、FreeBase 以及普通网页上关于达芬奇的介绍、照片，与达芬奇相关的音视频等信息，另外，通过深入分析，还可以把达芬奇的作品如“蒙娜丽莎”的相关信息、图片，以及同时期的艺术家“米开朗基罗”的相关信息等一起获取过来，之后再对获取的信息进行去重、清洗等预处理操作，最后把处理后的数据返回给用户。

关联信息发掘的关键技术除了传统数据集成任务所需的数据清洗、数据去重、数据融合、冲突消解和数据转换等数据预处理技术外，还包括直接和间接信息发掘技术。

3.1 直接获取

直接信息获取来源包括：互联网、物联网、视频监控、社交网络、专业领域数据等。

1) 互联网数据获取是指对互联网中的大数据进行高度并行的自动采集，并迅速收集到系统中的数据获取过程。互联网数据获取包括网页类获取和服务类数据获取2种方式，其中，网页类服务获取主要采用网络爬虫自动获取网页上的内容，网络爬虫可以按照一定的策略自动在互联网上蔓延以获取更多相关信息；服务类数据获取主要采用服务接口调用的方式获得网络服务数据。

2) 物联网数据获取是指通过RFID数据采集技术或者无线传感器网技术等方式获取物联网数据。RFID 数据采集技术是通过标签阅读器和标签接收器，定时或实时地收集人、物体、设备、环境、状态等基本信息。无线传感网技术是由许多在空间中分布的传感节点组成的一种无线通信计算机网络，这些传感节点协作地监控不同位置的物理或环境状况（如温度、声音、振动、压力、运动或污染物），其应用涉及军事、城市公共安全、公共卫生、安全生产、智能交通、智能家居、环境监控等领域。

3) 视频监控数据获取是对于视频监控系统和互联网上的视频数据进行收集并集成到系统中的过程。视频监控系统一般拥有大量的视频监控设备，视频监控设备产生的视频数据通过专用网络实时传输至视频监控系统的数据存储设备上，对于已存储的视频数据可以通过其调用接口进行获取。互联网上的视频一般具有特定的数据格式和相应的文本说明，可以通过网络爬虫利用合理的爬取策略来获取视频数据。

4) 社交网络数据获取是指对于各类社交网站中的相关数据进行自动收集并迅速集成到系统的过程。社交网络数据有表层和深层网络数据2类，如科研合作网络DBLP属于表层网络，而新浪微博属于深层网络。对于表层网络中网页信息的获取，可以直接使用爬虫程序对这些存储信息的网页进行解析，从标签属性值中抽取需要的信息。与表层网络相反，深层网络将页面信息存储在后台数据库中，只有通过查询接口查询才能由服务器动态生成并返回或者获取权限后才能查看，并没有超链接指向这些网页，不能被传统的搜索引擎索引到。因此，获取这些数据主要包含2种方式：一是通过查询接口查询由服务器动态生成并返回查询结果；二是仅对注册用户开放的信息，只有登录后才可查看专有网络信息。

5) 专业领域数据获取是根据需要，收集与某专业领域相关信息的过程。以医疗健康数据获取为例，它是对于医疗健康相关的信息系统和互联网上有关医疗健康的大数据进行高度并行的自动采集，迅速收集到系统中的数据获取过程。医疗健康信息系统包括医院信息系统、放射信息系统、实验室信息系统、医学影像存档与通信系统、临床信息系统、公关卫生信息系统、电子病历信息系统等，而互联网上有关医疗健康的数据有医学新闻博文、专业期刊杂志等。

3.2 间接获取

基于用户的搜索需求，间接信息发掘通过与智慧搜索知识推演系统的交互，基于知识推演给出深层次的搜索任务，从而获得更多面向任务的数据，并对获取的数据进行融合，最终满足用户的搜索需求。

间接信息发掘主要包含以下步骤。

1) 以用户的搜索需求和直接数据获取技术得到的数据作为输入，将其提交给智慧搜索知识推演系统。

2) 知识推演系统根据用户的搜索需求和已经获得的数据进行推演，如果该搜索需求仍不存在知识推演系统中，则将其返回给间接信息发掘系统。

3) 间接信息发掘系统根据当前收集相关数据和查询需求，发出新的查询请求，并将收集到的数据返回给智慧搜索知识推演系统。

4) 知识推演系统对用户的搜索需求和获得的信息进行推演，判断其是否满足用户的搜索需求。如果满足，则直接返回，推演结束；如果不满足，则重复步骤2)到步骤4)，直到获取的数据满足用户的搜索需求。

5) 将满足用户搜索需求的结果返回给用户。

例如，用户搜索“2014年全球总体失业率是多少”。使用直接数据获取技术会得到一些零散的与失业相关的数据，无法满足用户搜索需求。此时，间接信息发掘系统将用户的搜索需求以及已经获得零散数据提交给智慧搜索知识推演系统。知识推演系统推演得出全球的总体失业率可以通过综合不同国家和地区的失业率数据得到，因此，将各国的失业率作为查询需求返回给间接信息发掘系统。间接信息发掘系统进行查询并将得到的数据返回给知识推演系统。系统推演发现，

除了美国，其他各国 2014 年的失业率数据都可以得到。知识推演系统进一步推演得出通过查询美国每个季度的失业率来综合得到的美国年平均失业率。因此，将这一查询请求提交给间接信息发掘系统。间接信息发掘系统进行查询并将查询得到的数据返回给知识推演系统。知识推演系统推演发现将所有数据融合即可得到满足用户搜索需求的数据。因此，知识推演系统将最终融合后的数据返回给间接信息发掘系统，间接信息发掘系统将结果返回用户。

4　知识仓库

知识图谱包罗万象，可以看成是比较初级和粗糙的知识。为了能够支持高层的智能搜索、分析和推理服务，需要对知识图谱中所包含的数据进一步深度加工。在知识图谱中，一个实体可能存在着数量众多的关联关系，并且具备相同特征的实体又散布在图谱的各处，而基于知识图谱的具体处理和分析任务往往仅涉及部分子图和某些实体的部分关系。如何在语义层面对知识图谱中存储的知识进一步的组织和建模成为最大程度地发挥知识图谱作用的关键。这个层次的建模需要支持对知识图谱中符合某一语义定义的实体进行快速聚合，并且能够从多个维度对相关的实体集合进行分析，从而有利于发现各种规律或现象。此外，预先对知识图谱中的数据从不同维度进行组织和聚合，从而形成知识仓库，能够加快完成各种查询和分析的任务。

知识仓库是在整个知识图谱上，或者在满足预先定义或动态生成模式所形成的目标对象和关联对象所形成的子图上，通过系统地加工、汇总和整理所得到的结构化数据环境。知识仓库采用基于图的索引和分布式处理等技术能够对图中的对象从不同维度（或属性）和层次进行聚合（aggregate）、钻取（roll up / drill down）和旋转（pivot）等操作，以利于其上进行联机分析处理、数据挖掘，进而快速有效地从大量数据中分析出有价值的资讯。如图 3 所示，根据定义模式从知识图谱中定位和收集目标人群及其关联人群之后，可以通过地域、性别、年龄 3 个不同维度对目标人群和关联人群所组成的网络，结合其他相关信息进行焦点对象发现、多维度统计分析、对象行为预测、网络结构相似度分析等。

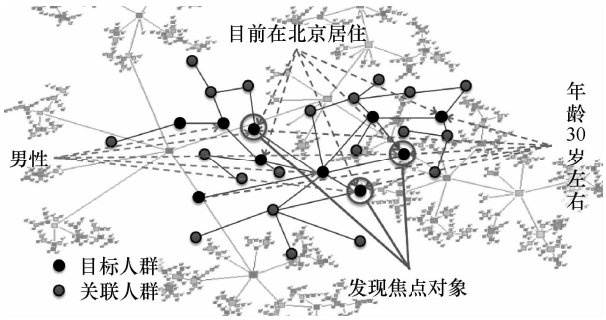

图 3　基于知识仓库多维分析的例子

5　查询与推理

知识图谱上的查询处理是管理和使用知识图谱的前提，也是获取蕴含于知识图谱中语义信息的基本操作。例如获取概念间的语义距离，获取一个或者一组实体的概念描述，获取句子的主题，对多义词进行消歧等任务，都可以转化为在知识图谱上的查询操作。知识图谱上的推理是从已知的知识产生新知识的过程。例如：从“配偶 ＋ 男性”推理出“丈夫”概念、从“应天是南京明朝时的名称 ＋ 建康是南京古称”推理出“应天和建康是同一城市在不同时期的称谓”。推理可以用于补充知识图谱的知识，也可以根据需要即时执行。

大数据是智慧搜索的处理对象，将搜索响应时间控制在合理的范围之内是系统成功的关键因素之一。知识图谱作为大数据的数据源之一，往往包含千万量级的实体和关系。为了提高知识图谱的查询性能，需要将知识图谱划分成若干子图，并且存储在不同的设备，然后通过分布式处理、并行计算、查询优化、索引技术来缩短查询的完成时间。知识图谱的推理一般采用基于规则的方法，规则既可以是基于数理逻辑学的逻辑规则，也可以是基于认知心理学的产生式规则。规则既可以人工定义，也可以通过学习获得。由于知识来源于动态、开放的网络，具有不可靠性，因而规则推理系统一般需要具备处理不确定推理的能力。

如果充分利用网页链接关系蕴含的信息是 Web 搜索引擎超越传统信息检索系统的基础，那么如何高效利用网络空间巨规模实体关联信息，将是智慧搜索取得成功的基础。智慧搜索能带来巨大的价值，不仅仅是因为利用了更多种类的数据或某一类型更大量的数据量，更主要在于其将充分发掘不同实体对象的跨域关联信息。

实体关联可以采用表和图 2 种方式来表达。相

比之下，图更适合表达稀疏、高维、海量的关联数据，表则会面临极高的连接、查询和存储的开销。因此，图是智慧搜索系统面向网络数据的一种最合理的表达抽象。智慧搜索支撑平台主要需要提供巨规模实体关联数据的存储和处理能力。

6 知识更新与演化

知识的演化与更新是指知识在时间轴上不断发展的一种动态变化，代表了知识的流动和变迁，即通过往知识仓库中添加新节点，并与网络中已有的节点进行连接，从而实现对知识的演化和更新。

例如，维基百科可以将用户发布的知识作为一个新的节点，通过将这一节点加入已有的知识网络，从而实现对知识库中原有知识网络结构的动态更新。又如，随着计算机技术的发展，以前所未闻的可穿戴式计算机应运而生，从而赋予了移动式计算机新的技术内涵。可穿戴式计算机为可穿戴于身上外出进行活动的微型电子设备，对于这种以前未在知识网络中出现的新知识，如何将其添加到知识库中，从而实现知识的演化和更新呢？其实，可以根据可穿戴式计算机的定义，利用知识网络中实体之间的关系，采用数据挖掘中的相关技术，如聚类技术，将其划入相应的知识社区中，从而实现知识网络的动态更新，最终更新知识库系统。

知识更新演化过程既反映知识网络的时序结构变迁，又体现知识和概念的内在涵义流变，演化模型是知识网络内在作用模式及作用过程的抽象表达。对演化过程的探讨既是分析知识网络结构的基础，也是探讨知识热点形成及创新趋势形成的基础。

7 数据存储与计算平台

数据存储与计算支撑平台用于存储、管理泛在网络空间的数据，支持智慧搜索的查询、统计和分析处理，包括高效知识提取和秒级搜索匹配等。

支撑平台的挑战主要包括 2 个方面。1）数据普适化，包括文本数据、音视频、地理数据、社交媒体关系数据、物联网数据等，这些大规模的非结构化数据需要通用的存储和计算模型来进行有效管理。2）查询、挖掘和分析多样复杂（如关键字查询、大图查询、时空查询、聚合查询、聚类分类、时序挖掘等），且具有严格的反馈时间要求。对普适化网络空间数据的存储、组织和管理是保证实体关系、知识抽取、搜索匹配能力的核心问题。

关系数据库管理系统和数据仓库系统已经在过去几十年中发展成为一项较为成熟的技术，主要用于管理结构化数据，无法有效存储组织形式松散的网络文本、多媒体等非结构化数据，并且由于大量的加锁操作和日志登记限制了数据更新性能。随着网络检索和大数据技术的快速发展，近些年在非结构化数据管理方面进一步形成了基于 GFS 和 HDFS 等分布文件系统的 NoSQL 家族，典型产品包括 HBase、Cassandra、MongoDB、Redis、Neo4J 等，以及著名的 Map-Reduce 分布式计算框架。这些数据库普遍采用列存的方式来达到更好的数据压缩，数据库集群具有较好的可伸缩性，并且提供了传统搜索引擎所需的简单索引。但是这些 NoSQL 数据库无法有效支持新一代“智能搜索”，其主要原因如下。

1) 智能搜索是一种情景敏感、基于语义内容的智能检索，根据不同搜索需要以多层次多维度的方式快速定位数据，因此现有数据库的数据索引和查询优化需要以可兼容的方式扩充，为大搜索处理提供底层的定制支持。

2) 现有数据库主要基于单一的数据模式，例如图模式、键值对模式以及关系模式等，分别对应着单一模式的数据。但是智能搜索集成了泛在网络空间数据，因此需要在模式层进行整合，以更加高效的方式管理普适化的巨规模网络数据。

3) 智能搜索需要对网络数据深加工，构建知识图谱并在此基础上发掘领域知识，需要执行大量的复杂挖掘和机器学习算法，大量的迭代处理无法在常规的 NoSQL 框架之下有效运行。随着内存存储能力的快速提升，可以在系统架构中引入内存计算框架来解决该类需求。

因此在智能搜索系统构建中，需要研发面向智能搜索的通用数据存储与计算平台，以分布式框架作为底层支撑，充分利用新型硬件效能（如内存计算、固态硬盘等，显著降低数据扫描的 I/O 代价），更加合理地组织管理泛在网络空间的异构数据，保证大搜索中各类复杂查询、统计分析、数据挖掘、知识抽取的快速处理。

8 相关工作

互联网上的搜索引擎已有 20 多年的历史，从最初的人工归类，到自动关键字搜索，一直到最近

的知识性搜索服务。下面围绕实体对象及关联关系在网络搜索中的应用，分析国内外研究现状。

1) 知识库在网络搜索中的使用

到目前为止，实体对象及其关系建模工作较多地围绕知识图谱（knowledge graph）展开。知识图谱简单地说就是一个“主谓宾”三元组的集合，其中“主”和“宾”是实体对象，“谓”是关联关系。2012 年 5 月 Google 发布了其基于知识图谱智能化搜索功能，通过对搜索进行系统的语义分析，使用户的每个查询关键词都能映射到知识库的概念上，从而用于优化搜索结果。知识图谱相对于本体和传统语义网络而言，实体对象覆盖率更高、语义关系也更加全面而复杂。目前学术界与工业界均呈现出一股构建和使用知识图谱的热潮。除 Google 之外，微软、百度、搜狗等公司都推出了各自的知识图谱，典型代表包括 KnowItAll[2]、TextRunner[3]、Probase[4]、YAGO[5]、DBpedia[6]、Freebase[7]等。

当前知识图谱的研究工作主要从构建与应用 2 个方面展开。知识图谱构建从其数据源来看可分为 2 类：一类是万维网的页面，另一类是相对结构化的在线百科。以前者为来源的典型知识图谱包括 KnowItAll[2]、TextRunner[3]和 Probase[4]。KnowItAll 基于规则模板抽取实体或概念之间的关系；TextRunner提出了自监督学习方法改善了KnowItAll需要人工定义规则的缺点；为了进一步提高关系抽取的准确性，Probase 采用基于语义的迭代方法抽取出更多更准确的 ISA 关系。而以在线百科为数据来源的知识图谱包括 YAGO 和 DBpedia 等。各类知识图谱已经在各类应用中发挥威力。Google 利用 Freebase 为用户提供更加智能化的搜索结果[8]。微软利用 Probase 理解 Web 表格[9]和查找话题[10]。苹果公司利用知识图谱进行智能问答[11]；利用 YAGO 增强地图的实时性[12]；利用 DBpedia 推荐音乐[13]、标签识别[14]以及信息抽取[15, 16]等。

国内也有研究团队从事这方面的研究，比如中科院计算所在知识抽取方面做了大量的工作，有基于图和图上推断的 CIIGA 方法[17]，在非结构化的文本中抽取实体并连接到知识库中，可以对现有的知识库做大量的补充。OpenKN[18, 19]可用于取大量新的实体和概念，进而不断对知识库进行更新。

上述知识图谱方面的工作，增加了搜索的智能性，在提高用户体验方面有着深远的影响。知识图谱的研究及开发也产生了大量的自然语言处理以及机器学习方面的理论和方法，极大地推进了领域的成长。文献[20]的工作主要点在于利用数据融合等方法，提高知识图谱的质量，在去除歧义、多名、错误等方面，有了长足的进步。但如 Sarma 等[21]和 Kuzey 等[22]指出，现行知识图谱技术偏重已知的实体，对不断涌现的新兴实体及其关联，尤其是事件性的关联，仍没有相应方法。

2) 知识库存储及查询相关研究

RDF 作为语义万维网技术的资源表示标准，许多知识图谱都选择 RDF 或者类似 RDF 的方式来表示知识。目前 RDF 查询研究重点在于查询语言的有效实现方法，但对查询模型的语义缺乏必要考虑。早期 RDF 查询多实现在关系数据库系统之上，利用关系表存储 RDF 数据，再将 RDF 查询转换为对应的 SQL 查询。其中典型的查询与存储系统包括：Sesam[23]、Jena2[24]、3store[25]、RDFSuite[26]。近期的焦点在于进一步提升 RDF 查询性能。如 Eugene[27]使用 RDF_MATCH 表函数，Abadi[28]利用垂直分片，Hexastore[29]通过常数倍的额外索引开来提升 RDF 查询性能。近 RDF 查询研究的核心是 SPARQL 查询语言，提高查询性能关键是减少 Join 操作的开销，MonetDB[30]和 Hexastore[29]都提出了 SPARQL 的 Join 优化算法。Medha[31]则利用流方式在压缩的 RDF 数据上生成最终结果而避免创建代价较高的中间连接表。Markus 等[32]研究了 SPARQL 查询的静态优化问题，定义和分析了基本图模式选择的启发式策略。Angela 等[33]和 Thomas[34~36]利用图挖掘技术计算并记录 RDF 图中的频繁最优路径来估计不同 Join 顺序的代价，用于查询优化。Huang 等[37]通过分割 RDF 数据和分解 SPARQL 查询来提高查询效率。Binna 等[38]设计了内存数据库 SpiderStore 来管理 RDF 数据和快速执行 SPARQL 查询。Weaver 等[39]提出了并行的 RDFS 闭包计算方法，而 Urbani 等[40]使用 MapReduce 实现类似的计算。Myung 等[41]和 Rohlo 等[42]研究使用 MapReduce 实现 SPARQL 查询。Manish Gupta 等研究 Top-k 子图的查询[43]。

在国内，北京大学和中国人民大学在 RDF 数据管理方面做了较多研究工作。比如，gStore[44]是一种由图作为存储方式的能够有效在动态 RDF 数据集上处理 SPARQL 查询的方法，Zou 等[45]提出了基于 RDF 数据的解决自然语言自动问答的方法，Yang 等[46]提出了自动分割 RDF 数据的方法来提升

查询效率并同时考虑了减少数据冗余，Du 等[47]研究了在集群环境下 RDF 数据分割和替换的策略，Bian 等[48]还提出了基于实体属性表单来补充知识库中 RDF 数据的方法。

RDF 本质上以“主谓宾”的方式表达实体之间的关联关系。理论上，这个形式有很强的表达能力，但对复杂实体（比如事件性实体时）一般采用隐含式表达。比如，在“事件本体模型”[49]中，事件作为实体，和事件有关的实体与此事件实体的关联（事件 S 涉及实体 A）即可用“主谓宾”模式建立，而事件的时间、地点，则也作为实体与事件实体简单关联。Trame 等[50]对怎样用 RDF 表示事件有所讨论，结论是简单的 RDF 很难自然地表达事件。即使是时间这个属性（也有把时间概念作为实体），基于 RDF 的表达也不够自然[50]。智慧搜索对各类显性及隐性实体必须用简单的方法，使之与人类一般认知规则相配，以便查询。

由于现行各类知识以简单 RDF 形式存储，故大量的图查询模型及技术可以应用知识库查询处理。目前大图查询研究工作主要围绕可达性查询、最短路径或距离查询、图匹配查询以及关键字查询开展。这些研究一般剥离图数据本身的领域背景，只在抽象的图查询模型上开展研究。图算法固然在知识查询方面有其作用，但当知识库在简单图上进行扩充，得以表达事件类实体时，需要考虑在知识库上的其他操作。

目前图查询算法大致有 4 类。1）可达性查询。这一问题主要研究特定约束条件下的可达查询，这些约束一方面使问题更为复杂，另一方面也为高效剪枝创造了条件。基本的约束是节点或边上的标签约束[51,52]和更为复杂的正则表达式约束[53]。2）最短距离或路径查询。当前主流方案都采用基于摘要（sketch）的框架。其基本思想是为每个节点创建固定大小的摘要，利用摘要估计节点之间的距离。目前有 2 类摘要方法：一是以到一组路标（landmark）节点的最短距离作为节点的摘要[54~58]；二是以节点在几何空间中的坐标作为摘要[59,60]。这些方案以线性空间索引实现常量时间的查询回答。第 1 类方法的研究侧重于提高距离估计准确性。第 2 类方法的研究集中于几何空间的选择。Zhao 等[59, 60]先后提出基于欧式空间和双曲空间最短距离查询方案，并证实基于双曲空间优于欧式空间。3）图匹配查询。这一问题的研究主要围绕 2 个核心问题开展：非精确匹配意义下的子图查询、大图上的子图查询。在非精确匹配方面，Fan 等[61]率先提出基于图模拟的图匹配，将子图匹配中边到边的严格映射放松为边到给定长度内的路径之间的映射。Zou 等[62]进一步改进图模拟高效算法。Ma 等[63]则提出了强模拟以进一步强化匹配约束。为了处理大图，Sun[64]、Ma[65]分别提出了相应的分布式子图查询方法和图模拟算法从而支持快速大图匹配。4）关键字查询。这类问题是寻找图中含有关键字的点和边，各研究的差异主要在于返回子图的结构约束不同，比如 r 半径斯坦纳（Steiner）图[66]，r-极大团[67]。针对 r 半径斯坦纳图，Li 等[66]给出了一种基于图划分的快速查询方法。Kargar[67]针对基 r-极大团的图上关键字查询提出了一个返回 top-k 的近似算法。

3) 数据立方模型

数据立方（data cube）的概念于 1996 年由 Gray[68]引入数据分析领域。数据立方建立在关系数据库之上，为分析者提供简单易懂的概念模型和操作界面，把数据分析的操纵权从程序员手里夺走，交还给了分析用户，为数据分析研究和产业做出了革命性的贡献。对于这个成功，究其深层原因，是将数据以接近用户习惯的认知方式呈现给用户：将数据以多维度的形式，每个维度对应一类概念（如时间、空间），而每个概念又可以以不同粒度来观察数据。

研究人员已将数据立方相关的概念用于其他分析工作。如 Jiawei Han 所带领的研究团队开展了文本数据的多粒度特性方面的研究[69, 70]，支持文本数据多粒度分析，将大量的文本信息组织成层次结构，而后数据分析可以利用上卷、下钻等操作在不同粒度上进行访问。近期，该研究团队又在图数据上引入 OLAP 数据立方的概念，研究图立方（graph OLAP 和 graph cube）[71, 72]对图数据分析的用途。

9 结束语

智慧搜索将会为人们带来崭新的搜索方式——知识服务，它是指从各种知识来源（包括知识图谱和知识仓库）中按照用户的个性需求有针对性地提炼知识，并且用来解决用户问题的高级阶段信息服务过程。与传统信息服务强调信息资源获取（如文献检索）不同，知识服务侧重于提供个性化、面向解决方案的服务。它根据用户问题语义和上下文环境分析确定用户的需求，通过多源信息和知识的重

组与融合形成符合需要的知识产品。

实现大搜索的愿景，目前还面临许多的挑战，但同时也带来众多的研究机会。目前急需解决的难题包括：根据查询的需求，从包括海量实体以及关系的泛在网络空间中准确地获取数据；全面和深度地理解用户的真实搜索意图；融合多渠道、多模式和实时复杂的数据，挖掘和发现其中潜在、有价值的信息；确保大搜索使用过程安全可信；根据用户的搜索意图，基于知识仓库对关联知识进行推理和求解，形成若干可行的智慧综合解决方案。

大搜索是新一代具有“智慧”的搜索，能准确洞察和理解用户的搜索意图，在海量、多源、异构、多态、不确定的数据中，实现对与人物、物体和内容等相关信息的对象级搜索，为用户提供最贴切的搜索结果。这势必影响我国的社会、经济和生活等各个方面，具有广阔的应用前程。

参考文献：

[1] 方滨兴,等. 大搜索技术白皮书[M]. 北京: 电子工业出版社, 2015.
FANG B X, et al. Big Search Technology White Paper[M]. Beijing: Electronic Industry Press, 2015.

[2] ETZIONI O, CAFARELLA M, DOWNEY D, *et al*. Web-scale information extraction in knowitall:(preliminary results)[A]. Proceedings of the 13th International Conference on World Wide Web[C]. ACM, 2004. 100-110.

[3] YATES A, CAFARELLA M, BANKO M, et al. Textrunner: open information extraction on the web[A]. Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations Association for Computational Linguistics[C]. 2007. 25-26.

[4] WU W, LI H, WANG H, et al. Probase: a probabilistic taxonomy for text understanding[A]. ACM SIGMOD International Conference on Management of Data[C]. ACM, 2012. 481-492.

[5] SUCHANEK F M, KASNECI G, WEIKUM G. Yago: a core of semantic knowledge[A]. 16th International Conference on World Wide Web[C]. ACM, 2007. 697-706.

[6] AUER S, BIZER C, KOBILAROV G, et al. Dbpedia: a Nucleus for a Web of Open Data[M]. Springer Berlin Heidelberg, 2007.

[7] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[A]. ACM SIGMOD International Conference on Management of Data[C]. ACM, 2008. 1247-1250.

[8] SINGHAL A. Introducing the Knowledge Graph: Things, Not Strings Official Blog (of Google)[EB/OL]. http://googleblog.blogspot.com /2012/05/introducing-knowledge-graph-things-not.html.Retrieved.

[9] WANG J, WANG H, WANG Z, et al. Understanding Tables on the Web Conceptual Modeling[M]. Springer Berlin Heidelberg, 2012. 141-155.

[10] WANG Y, LI H, WANG H, et al. Toward Topic Search on the Web[R]. Technical report, Microsoft Research, 2010.

[11] Apple-Siri-frequently asked questions. Apple[EB/OL]. http:// www. siriuserguide.com/siri-faq/.

[12] HOFFART J, SUCHANEK F M, BERBERICH K, et al. YAGO2: exploring and querying world knowledge in time, space, context, and many languages[A]. 20th International Conference Companion on World Wide Web[C]. ACM, 2011. 229-232.

[13] PASSANT A. Dbrec—music recommendations using DBpedia[A]. The Semantic Web-ISWC 2010[C]. Springer Berlin Heidelberg, 2010. 209-224.

[14] GARCIA A, SZOMSZOR M, ALANI H, et al. Preliminary results in tag disambiguation using DBpedia[A]. Collective Knowledge Capturing and Representation[C]. California, 2009.

[15] Wu F, Weld D S. Automatically refining the wikipedia infobox ontology[A]. 17th International Conference on World Wide Web[C]. ACM, 2008. 635-644.

[16] KASNECI G, RAMANATH M, SUCHANEK F, et al. The YAGO-NAGA approach to knowledge discovery[J]. ACM SIGMOD Record, 2009, 37(4): 41-47.

[17] LIN H, JIA Y, WANG Y, et al. Populating knowledge base with collective entity mentions: a graph-based approach[A]. Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on[C]. IEEE, 2014. 604-611.

[18] JIA Y, WANG Y, CHENG X, et al. OpenKN: an open knowledge computational engine for network big data[A]. Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on[C]. IEEE, 2014. 657-664.

[19] 王元卓, 贾岩涛, 赵泽亚, 等. OpenKN——网络大数据时代的知识计算引擎[J]. CCF 通讯, 2014, 10(11): 30-35.
WANG Y Z, JIA Y T, ZHAO Z Y, et al. OpenKN—— knowledge computing engine in the big data era[J]. CCF Communication, 2014, 10(10): 30-35.

[20] LI Q, LI Y L, GAO J, et al. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation[A]. Proceedings of the 2014 SIGMOD[C]. 2014.

[21] SARMA D JAIN A A, YU C. Dynamic relationship and event discovery[A]. Fourth ACM International Conference on Web Search and Data Mining[C]. ACM, 2011. 207-216.

[22] KUZEY E, VREEKEN J, WEIKUM G. A fresh look on knowledge bases: Distilling named events from news[A]. 23rd ACM International Conference on Information and Knowledge Management[C]. ACM, 2014. 1689-1698.

[23] BROEKSTRA J, KAMPMAN A, VAN HARMELEN F. Sesame: an architecture for storing and querying rdf data and schema information[J]. Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential, 2003, 197.

[24] WILKINSON K, SAYERS C, KUNO H A, et al. Efficient RDF Storage and retrieval in Jena2[A]. The First International Workshop on Semantic Web and Databases[C]. 2003, 3: 131-150.

[25] HARRIS S, GIBBINS N. 3store: efficient bulk RDF storage[A]. Workshop on Practical and Scalable Semantic Systems[C]. 2003.

[26] ALEXAKI S, CHRISTOPHIDES V, KARVOUNARAKIS G, et al. The ICS-FORTH RDFSuite: managing voluminous RDF description bases[A]. SemWeb[C]. Hong Kong, China, 2001.

[27] CHONG E I, DAS S, EADON G, et al. An efficient SQL-based RDF querying scheme[A]. 31st International Conference on Very Large

Data Bases VLDB Endowment[C]. 2005. 1216-1227.

[28] ABADI D J, MARCUS A, MADDEN S R, et al. Scalable semantic web data management using vertical partitioning[A]. 33rd International Conference on Very Large Data Bases[C]. 2007. 411-422.

[29] WEISS C, KARRAS P, BERNSTEIN A. Hexastore: sextuple indexing for semantic Web data management[J]. Proceedings of the VLDB Endowment, 2008, 1(1): 1008-1019.

[30] SIDIROURGOS L, GONCALVES R, KERSTEN M, et al. Column-store support for RDF data management: not all swans are white[J]. Proceedings of the VLDB Endowment, 2008, 1(2): 1553-1563.

[31] ATRE M, CHAOJI V, ZAKI M J, et al. Matrix bit loaded: a scalable lightweight join query processor for RDF data[A]. 19th International Conference on World Wide Web[C]. ACM, 2010. 41-50.

[32] STOCKER M, SEABORNE A, BERNSTEIN A, et al. SPARQL basic graph pattern optimization using selectivity estimation[A]. 17th International Conference on World Wide Web[C]. ACM, 2008. 595-604.

[33] MADUKO A, ANYANWU K, SHETH A, et al. Estimating the cardinality of RDF graph patterns[A]. Proceedings of the 16th International Conference on World Wide Web[C]. ACM, 2007. 1233-1234.

[34] NEUMANN T, WEIKUM G. RDF-3X: a RISC-style engine for RDF[J]. Proceedings of the VLDB Endowment, 2008, 1(1): 647-659.

[35] NEUMANN T, WEIKUM G. The RDF-3X engine for scalable management of RDF data[J]. The VLDB Journal, 2010, 19(1): 91-113.

[36] NEUMANN T, WEIKUM G. Scalable join processing on very large RDF graphs[A]. Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data[C]. ACM, 2009. 627-640.

[37] HUANG J, ABADI D J, REN K. Scalable SPARQL querying of large RDF graphs[J]. Proceedings of the VLDB Endowment, 2011, 4(11): 1123-1134.

[38] BINNA R, GASSLER W, ZANGERLE E, et al. Spiderstore: exploiting main memory for efficient RDF graph representation and fast querying[A]. Proceedings of Workshop on Semantic Data Management (SemData@ VLDB) [C]. 2010.

[39] WEAVER J, HENDLER J A. Parallel Materialization of the Finite RDFs Closure for Hundreds of Millions of Triples[M]. Springer Berlin Heidelberg, 2009.

[40] URBANI J, KOTOULAS S, OREN E, et al. Scalable Distributed Reasoning Using MapReduce[M]. Springer Berlin Heidelberg, 2009.

[41] MYUNG J, YEON J, LEE S. SPARQL basic graph pattern processing with iterative MapReduce[A]. Proceedings of the 2010 Workshop on Massive Data Analytics on the Cloud[C]. ACM, 2010.

[42] ROHLOFF K, SCHANTZ R E. High-performance, massively scalable distributed systems using the MapReduce software framework: the SHARD triple-store[A]. Programming Support Innovations for Emerging Distributed Applications[C]. ACM, 2010.

[43] GUPTA M, GAO J, YAN X F, et al. Top-*K* interesting subgraph discovery in information networks[A]. 2014 International Conference on Data Engineering[C]. 2014.

[44] ZOU L, ÖZSU M T, CHEN L, et al. gStore: a graph-based SPARQL query engine[J]. The VLDB Journal—the International Journal on Very Large Data Bases, 2014, 23(4): 565-590.

[45] ZOU L, HUANG R, WANG H, et al. Natural language question answering over RDF: a graph data driven approach[A]. Proceedings of the 2014 ACM SIGMOD International Conference on Management of data[C]. ACM, 2014. 313-324.

[46] YANG T, CHEN J, WANG X, et al. Efficient S`PARQL query evaluation via automatic data partitioning[A]. Database Systems for Advanced Applications[C]. Wuhan, 2013.

[47] DU F, BIAN H, CHEN Y, et al. Efficient SPARQL query evaluation in a database cluster[A]. Big Data, 2013 IEEE International Congress on[C]. 2013. 165-172.

[48] BIAN H, CHEN Y, DU X, et al. MetKB: enriching RDF knowledge bases with web entity-attribute tables[A]. 22nd ACM International Conference on Conference on Information & Knowledge Management[C]. ACM, 2013. 2461-2464.

[49] RAIMOND Y, et al. The event ontology[EB/OL]. http://motools.sourceforge.net/event/event.html. 2007.

[50] TRAME J, KEßLER C, KUHN W. Linked Data And Time–Modeling Researcher Life Lines By Events[M]. Spatial Information Theory. Springer International Publishing, 2013.

[51] JIN R, HONG H, WANG H, et al. Computing label-constraint reachability in graph databases[A]. 2010 ACM SIGMOD International Conference on Management of data[C]. ACM, 2010. 123-134.

[52] XU K, ZOU L, YU J X, et al. Answering label-constraint reachability in large graphs[A]. Proceedings of the 20th ACM International Conference on Information and Knowledge Management[C]. ACM, 2011. 1595-1600.

[53] FAN W, LI J, MA S, et al. Adding regular expressions to graph reachability and pattern queries[A]. Data Engineering (ICDE), 2011 IEEE 27th International Conference on[C]. 2011. 39-50.

[54] GUBICHEV A, BEDATHUR S, SEUFERT S, et al. Fast and accurate estimation of shortest paths in large graphs[A]. Proceedings of the 19th ACM International Conference on Information and Knowledge Management[C]. ACM, 2010. 499-508.

[55] POTAMIAS M, BONCHI F, CASTILLO C, et al. Fast shortest path distance estimation in large networks[A]. 18th ACM Conference on Information and Knowledge Management[C]. ACM, 2009. 867-876.

[56] TRETYAKOV K, ARMAS-CERVANTES A, GARCÍA-BAÑUELOS L, et al. Fast fully dynamic landmark-based estimation of shortest path distances in very large graphs[A]. 20th ACM International Conference on Information and Knowledge Management[C]. ACM, 2011. 1785-1794.

[57] DAS SARMA A, GOLLAPUDI S, NAJORK M, et al. A sketch-based distance oracle for Web-scale graphs[A]. Proceedings of the Third ACM International Conference on Web Search and Data Mining[C]. ACM, 2010. 401-410.

[58] GOLDBERG A V, HARRELSON C. Computing the shortest path: a search meets graph theory[A]. Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms Society for Industrial and Applied Mathematics[C]. 2005. 156-165.

[59] ZHAO X, SALA A, WILSON C, et al. Orion: shortest path estimation for large social graphs[J]. Networks, 2010, 1: 5.

[60] ZHAO X, SALA A, ZHENG H, et al. Fast and scalable analysis of massive social graph [J]. arXiv preprint arXiv:1107.5114, 2011.

[61] FAN W, LI J, MA S, et al. Graph pattern matching: from intractable to polynomial time[J]. Proceedings of the VLDB Endowment, 2010, 3(1-2): 264-275.

[62] ZOU L, CHEN L, ÖZSU M T, et al. Answering pattern match queries in large graph databases via graph embedding[J]. International Journal on Very Large Data Bases, 2012, 21(1): 97-120.

[63] MA S, CAO Y, FAN W, et al. Capturing topology in graph pattern

matching[J]. Proceedings of the VLDB Endowment, 2011, 5(4): 310-321.

[64] SUN Z, WANG H, WANG H, et al. Efficient subgraph matching on billion node graphs [J]. Proceedings of the VLDB Endowment, 2012, 5(9): 788-799.

[65] MA S, CAO Y, HUAI J, et al. Distributed graph pattern matching[A]. 21st International Conference on World Wide Web[C]. 2012. 949-958.

[66] LI G, OOI B C, FENG J, et al. EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data[A]. ACM SIGMOD International Conference on Management of Data[C]. 2008. 903-914.

[67] KARGAR M, et al. A. Keyword search in graphs: finding r-cliques[J]. Proceedings of the VLDB Endowment, 2011, 4(10): 681-692.

[68] GRAY J, CHAUDHURI S, Bosworth A, et al. Data cube: a relational aggregation operator generalizing group-by, cross-tab, and sub-totals[J]. Data Mining and Knowledge Discovery, 1997, 1(1): 29-53.

[69] LIN C X, DING B, HAN J, et al. Text cube: computing ir measures for multidimensional text database analysis[A]. Data Mining, ICDM'08. Eighth IEEE International Conference on[C]. 2008. 905-910.

[70] ZHANG D, ZHAI C, HAN J. Topic cube: topic modeling for OLAP on multidimensional text databases[A]. SDM[C]. 2009, 9: 1124-1135.

[71] CHEN C, YAN X, ZHU F, et al. Graph OLAP: towards online analytical processing on graphs[A]. Eighth IEEE International Conference on Data Mining[C]. 2008.

[72] ZHAO P, LI X, XIN D, et al. Graph cube: on warehousing and OLAP multidimensional networks[A]. ACM SIGMOD International Conference on Management of data[C]. 2011. 853-864.

作者简介：

王晓阳（1960-），男，上海人，复旦大学教授，主要研究方向为时空移动数据分析、数据系统安全及私密、大数据并行式分析与计算。

郑骁庆（1979-），男，浙江杭州人，复旦大学副教授，主要研究方向为数据集成、自然语言理解、语义万维网。

肖仰华（1980-），男，江苏洪泽人，复旦大学副教授，主要研究方向为数据库、数据挖掘、海量数据处理、图数据库、图数据挖掘。