Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

面向语义万维网“大规模分布式不完备推理平台LarKC国际专题会议”综述*

□ 李颖 焦淑娟 / 中国科学技术信息研究所 北京 100038

摘要：LarKC项目是开发面向语义万维网的大规模分布式不完备推理平台，LarKC国际专题会议又是全面掌握与免费应用LarKC推理平台的平台。基于第4届LarKC国际专题会议的第一手信息，文章概要介绍LarKC项目诞生的背景与任务、框架及开发进展、历届LarKC国际专题会议，最后作出总结。

关键词：LarKC，大规模分布式不完备推理平台，大规模知识加速器，大规模异构知识源，语义万维网

DOI：10.3772/j.issn.1673—2286.2011.02.014

1 LarKC项目的背景与任务[1-3]

语义万维网（Semantic Web）已发展到了推理层面，而目前的推理系统面临着知识处理规模上的瓶颈，为突破这一难关，诞生了LarKC项目。LarKC的使命是开发大规模知识加速器，即设计一个大规模分布式不完备推理平台。分布式体现在数据集分布在万维网（目前主要处理的是RDF格式的数据）、本地等不同资源；不完备的概念意味着“在有限时间内，基于海量数据的确定性推理几乎是不可能的”，只能“在不完全数据上进行令用户足够满意的推理”；平台的含义是：LarKC把基于语义Web的问题求解组件都以插件的形式组织在一起，通过管道（Pipe line）进行调用。这种体系结构设计的意义在于，当人工智能的研究者面对海量数据的问题求解时，不需要将任何事情都从头做起。LarKC项目虽然在2008年才设立，是“刚出世的婴儿”，却得到了强劲的发展与应用。它满足了大规模异构知识源处理领域的多种需求，如电信服务、生物医学研究、医药开发等。

LarKC发音为“lark”，全称The Large Knowledge Collider，这个名字的由来是受到了欧盟原子能研究组织开发的大规模强子对撞机（Large Hadron Collider）的启发，是欧盟第7框架计划（FP7，The Seventh Framework Programme）的项目。实施期间为2008年4月1日-2011年9月30日，经费约1000万欧元。其成员来自11个国家，合计13个著名研究机构及企业，中方的合作者为北京工业大学国际WIC研究院。

LarKC的主要任务如下：

◆ 扩充现有基于逻辑的语义万维网推理方法：通过运用信息检索、机器学习、信息论、数据库、概率推理等学科的理论研发新的推理方法；

◆ 利用受认知科学启发的方法与技术：如传播激活（spreading activation）、注意（attention）、强化（reinforcement）、习惯（habituation）、关联推理（relevance reasoning）、有限合理性（bounded rationality）；

◆ 构建分布式推理平台：计划在高性能计算集群及互联家庭计算机平台上实现。

2 LarKC架构及项目进展[3-5]

2.1 LarKC架构

创新与特性：

LarKC大规模集成计划是要突破现在语义计算中存储、查询、推理等方面技术上的局限性。这样一个

* 基金项目：中国科学技术信息研究所重点工作项目“汉语科技词系统建设与应用工程”（ZD2010-3-2）、中国科学技术信息研究所学科建设项目“知识工程”（XK2010-5）、“十一五”国家科技支撑计划项目“科技文献信息服务系统应用示范”（2006BAH03B06）基金支持。

基本设想的架构必须超越现在严格建立在逻辑上的典范。通过融合推理与搜索，以及重视有限理性这样的概念，达到在万维网上推理所要求的范式转换。因为只有部分的推理结果在一些应用领域中有用，所以，在推理进程的许多阶段可以采用不完全推理来达到使推理进程明显加速，包括从选择公理到基于这些公理进行不完全推理等不同阶段。

LarKC平台具有插件式的系统结构，使得该平台能够集成各种不同领域的技术及启发式策略，例如数据库、机器学习、认知科学、语义万维网等。LarKC平台将搜索与逻辑推理集成在一起，通过并行达到可扩展性：通过对高性能计算集群上并行进程的紧密整合，或通过更松散连接的大范围分布式计算，实现扩展性；其结果不是要创建一个适合于所有应用领域的推理引擎，而是要创建一个能将不同模块插入进来以达到对规模和效率的权衡取舍，来满足不同应用领域的不同要求的一个平台。由此产生的插件结构允许机构内部或外部的研究者和用户通过构建自己的插件来进行不同形式的平行或相近的推理；应用实例：这个平台和一系列预制的插件模块主要应用于以下三方面：实时分析、分析解释城市基础设施相关数据，以及为早期临床药品研制和致癌风险因素方面相关研究工作提供数据整合。

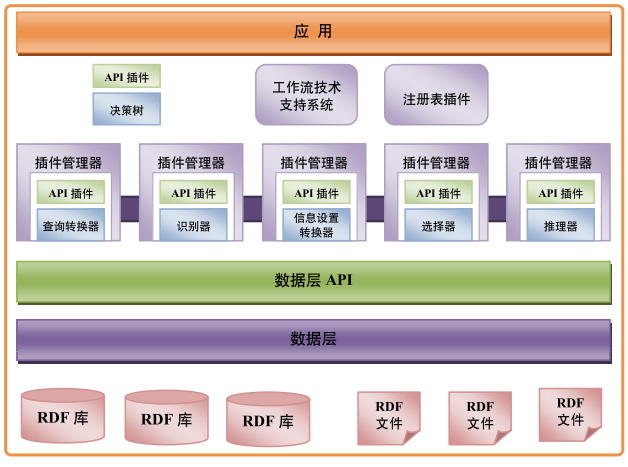

图1 LarKC架构

如图1所示，LarKC架构可概括为如下的要点：

◆ LarKC的结构设计和原型实现是协同进行的。最初的LarKC结构设计为简单的线性pipeline，现已演化为更加灵活的workflow方式。

◆ Pipeline / workflow采用灵活的插件结构，这种结构便于设计和测试各种新的推理技术。

◆ LarKC中的推理过程被分解为工作流中的若干步，每步对应一个插件，整体对应一个工作流，以解决大规模知识库上的推理问题。

◆ LarKC的目标是设计一个开放的、灵活的、可扩展的平台，又能够满足LarKC用例需求，是灵活性和性能的折衷。

LarKC平台的主要组成部分：

◆ 数据层及其API：用于存储和交换数据。

◆ 工作流支持系统：使决策树（Decider）插件能建立、执行和管理LarKC工作流。

◆ 插件API：用于和外部插件交互。

◆ 资源适配器：屏蔽处理异构的和分布式运行环境的复杂性，充分利用不同的计算资源来运行平台和插件。

◆ 插件注册组件：用于保存可用插件的信息。

总体结构：

◆ LarKC把推理过程分解为若干步，每步对应一类插件。

◆ 插件组合成工作流，处理大规模知识库上的推理任务。

◆ 平台已提供组建、实例化、监视和控制工作流所需要的支持。

原理：

◆ 插件式的体系结构：LarKC没有将其体系结构仅仅构建于逻辑之上，而是充分利用其他领域的方法：如认知科学（人类启发式）、经济学（有限合理性、开销/收益平衡）、信息检索（查准率/查全率之间的平衡），及数据库理论（大规模数据处理）。插件式的体系结构使得不同组件之间可以连贯地集成在一起。

◆ 分布式与并行推理：大规模知识加速器通过充分利用集群计算等并行硬件处理平台，由此可以达到处理大规模、分布式计算资源的目的，从而超越目前在语义计算领域以数据驱动的应用。

◆ 集成推理与搜索：

◆ 求解某个问题必要的公理与数据来自何方？（识别：IDENTIFY）

◆ 如何将这些数据进行抽象处理，以供随后的组件进行处理?（转换：TRANSFORMation）

◆ 哪一部分知识与数据是必要的？（选择：SELECTion）

◆ 何时结果已经“足够好”或“已经做到最好”?（决定：DECIDEr）

◆ 可从信息中通过演绎、非演绎推理方法自动获取什么结论?（推理：REASONer）

2.2 LarKC项目进展

基于意义（Meaning）的计算是未来计算技术发展的方向。目前语义计算最为紧迫的用例是实时处理近100亿RDF三元组。比如与语境密切相关的电信领域、个性化的移动服务、处理大规模医学文献等。LarKC为实现意义计算，致力于在万维网规模的知识源上进行海量、分布式、必要的不完备推理。海量推理是通过利用分布式计算平台实现的，充分考虑了分布式处理环境下的数据依赖等问题。LarKC不但能够实现大规模演绎推理，还通过插件式体系结构和复杂的工作流（workflows）集成更为丰富的数据处理能力。研究者们可以为这个平台设计具有不同功能的插件系统。

目前LarKC项目主要进展如下：

◆ 制药及生物领域的数据集成：数据集成问题仍然是制药及生物领域的挑战。“关联的生命数据（Linke Life Data）”集成了公开发行的生命科学数据集，这些数据集描述了基因、蛋白质、药物、疾病、病人之间的关系，目前包含了50亿RDF三元组。该数据集链接了20多个现有数据集，从而为以往研究中独立数据集之间的关联提供了“集成的蓝图”。

◆ 城市计算：目前城市环境相关的数据通过万维网平台分散地提供给用户，如：地图、时间、旅游胜地、交通信息等。不仅如此，某些国家、地区的政府将与城市相关的数据发布给公众，进行应用开发的趋势也越来越明显。LarKC项目目前专注于意大利米兰和韩国首尔的城市数据，基于这些海量数据正在进行交通路况预测、道路路标管理以及基于位置的服务等方面的研究。

3 LarKC专题国际会议[3]

LarKC专题国际会议尽管开始于1年半之前，但截止到2010年11月，在世界各地，以不同的形式共举办了4届，分别如下：

◆ 第1届：2009年6月1日，于希腊克里特岛，与2009年欧洲语义万维网大会（ESWC09）联合举行。

◆ 第2届：2009年10月25日，于美国华盛顿，与2009年国际语义万维网大会（ISWC09）联合举行。

◆ 第3届：2010年5月30日，于希腊克里特岛，与2010年欧洲语义万维网大会（ESWC10）联合举行。

◆ 第4届：2010年11月13-14日，于中国北京工大建国饭店，是LarKC在中国首次举办的专题培训班和2010年大规模知识加速器（LarKC）博士论坛。

第4届培训班（13日）目的是使语义万维网的研究者、实践者能够较早地接触并使用到LarKC平台的早期研究成果。培训班持续1天的时间，由LarKC项目的负责人、主要科研人员、开发人员担任讲授任务，并由若干开发人员指导参与者，现场进行了动手实验。参与本次培训之后，与会者获得了利用LarKC平台自己开发设计插件的基本技能，并可在LarKC平台上运行自己的程序。2010年大规模知识加速器（LarKC）博士论坛（14日）由7位来自不同国家（中国、意大利、德国、英国、荷兰）的研究者参与。LarKC项目的博士生及青年学者介绍了与海量语义数据处理相关的研究成果，气氛非常活跃。

4 LarKC总结[5]

LarKC是一个开源项目，在北京举行的第4届LarKC专题会议免费对参会者进行了系统化的培训，课程组织得很人性化，非常成功。它不仅让大家较为详细地了解LarKC项目，参与和推广应用了LarKC项目的成果；又为大家提供了一个很好的交流平台，让语义网和语义技术相关的研究和应用开发者们能够在一起很好地沟通交流，也为大家以后的进一步合作提供了机会。这让我们看到：基于语义技术的应用就在眼前！以荷兰阿姆斯特丹自由大学黄智生教授和北京工业大学国际WIC研究院的曾毅博士后为主设立的论坛，为中国的语义研究及其学术信息传播作出了建设性的贡献。

致谢：由衷感谢LarKC项目的中方技术负责人、北京工业大学国际WIC研究院的曾毅博士后，他在中国科学技术信息研究所知识组织与知识工程研究组的参会方面、在我们对项目理解与应用考量方面，给予了慷慨帮助，感谢他的超人的智慧与协作精神！感谢华人学者荷兰阿姆斯特丹自由大学黄智生教授，他的逻辑思维与推理智慧给我们带来了巨大的反思。

参考文献

[1] The Large Knowledge Collider [EB/OL]. [2010-11-15]. http://www.larkc.eu/.
[2] LarKC Wiki [EB/OL]. [2010-11-15]. http://wiki.larkc.eu/.
[3] LarK中文网站[EB/OL]. [2010-11-15]. http://www.wici-lab.org/wici/larkc/.
[4] 4th Early Adopters Tutorial - LarKC: the Large Knowledge Collider [EB/OL]. [2010-11-15]. http://www.larkc.eu/early-adopters/4th-early-adopters-tutorial/.
[5] 中国万维网联盟LarKC项目中文论坛[EB/OL]. [2010-12-01]. http://bbs.w3china.org/list.asp?boardid=80.

作者简介

李颖，博士，信息系统专业。近期研究课题：Topic Maps及RDF等语义/知识组织技术在金融领域的应用、基于数字对象唯一标识符信息融合等。E-mail: liying@istic.ac.cn
焦淑娟，硕士，软件工程。研究课题：蚁群算法、领域词系统等。E-mail: jiaosj@istic.ac.cn

Review for "International Workshop on LarKC – a Platform for Massive Distributed Incomplete Reasoning"

Li Ying, Jiao Shujuan / Institute of Scientific & Technical Information of China

Abstract: The Large-Scale Integrating Project LarKC is to develop the Large Knowledge Collider, which is a platform for massive distributed incomplete reasoning for the Semantic Web, and the International Workshop on LarKC is a platform for understanding and free adoption of LarKC. Based on first-hand information from the 4th International Workshop on LarKC, this article introduces the background and mission, framework and progress of the LarKC project, the previous International Workshop on LarKC, and finally, gives the summary.

Keywords: LarKC, Platform for massive distributed incomplete reasoning, Large knowledge collider, Massive heterogeneous information sources, Semantic Web

（收稿日期：2011-01-03）