OpenKN——网络大数据时代的知识计算引擎

王元卓[1] 贾岩涛[1] 赵泽亚[2] 程学旗[1]
[1]中国科学院计算技术研究所
[2]信息工程大学

关键词：网络大数据 知识计算 知识网络

近年来，互联网技术和应用模式的快速发展在改变人们生活方式的同时也产生了巨大的数据资源。预计到2020年，全球的数据总量将达到35ZB(1ZB=2^{70}B)，其中75%来自个人（主要是图片、视频和音乐），远远超过人类有史以来所有印刷材料的数据总量(200PB[1])。随着互联网、物联网、云计算等技术的迅猛发展，网络空间(cyberspace)中各类应用层出不穷，引发了数据规模的爆炸式增长,形成了网络空间的大数据(简称网络大数据）[1]。

网络大数据中包含大量有价值的数据，根据其产生方式的不同可分为Web内容数据、Web结构数据、自媒体数据、日志数据等。如何从网络大数据中获得有价值的知识，并对其进行深入的计算和分析，已成为国内外工业界和学术界研究的热点[2]。目前，世界各个组织建立的知识库多达50余种，相关的应用系统更是达到了上百种。其中，有代表性的知识库或应用系统有KnowItAll[3], TextRunner[4], NELL[5], Probase[6], Satori [7], PROSPERA[8], SOFIE[9]以及一些基于维基百科等在线百科知识构建的知识库DBpedia[10], YAGO[11], Omega[12], WikiTaxonomy[13]。除此之外，一些著名的商业网站、公司和政府也发布了类似的知识搜索和计算平台，如Evi公司的TrueKnowledge知识搜索平台[2]、美国官方政府网站Data.gov, Wolfram的知识计算平台WolframAlpha、谷歌的知识图谱Knowledge Graph、脸书(Facebook)推出的实体搜索服务Graph Search等。

就规模而言，拥有概念最多的知识库是Probase，目前其核心概念约有270万个，概念总量达到千万级。它是基于概率化构建的知识库，支持针对短文本的语义理解。包含实体最多的是WolframAlpha,有10万亿个实体。近年来，影响力比较大的知识库或知识搜索服务有谷歌的知识图谱，包含5亿个实体对象和350亿条实体间的关系信息，而且规模也在随着信息的增长不断增大。除此之外，比较有特色的还有国内搜狗知立方系统，侧重于基于图的逻辑推理计算，包括利用语义网的三元组推理补充实体数据、对用户查询词进行语义理解以及句法分析等。

本文将提出一种面向网络大数据的、开放的、自适应的、可演化的、可计算的知识计算引擎——OpenKN。

OpenKN的整体架构

图1描述了OpenKN作为计

[1] 1PB=2^{50}B。
[2] http://www.evi.com。

算引擎的主要架构。OpenKN主要由知识库构建(knowledge base construction)、知识验证与计算(knowledge validation and verification, knowledge computation)、知识存储(knowledge repositories)、知识服务与应用(knowledge services and application)4个模块组成。这些模块实现了一个全生命周期的知识处理，从知识获取、知识融合、知识验证与计算、知识存储到知识服务与应用的知识处理工作流程。

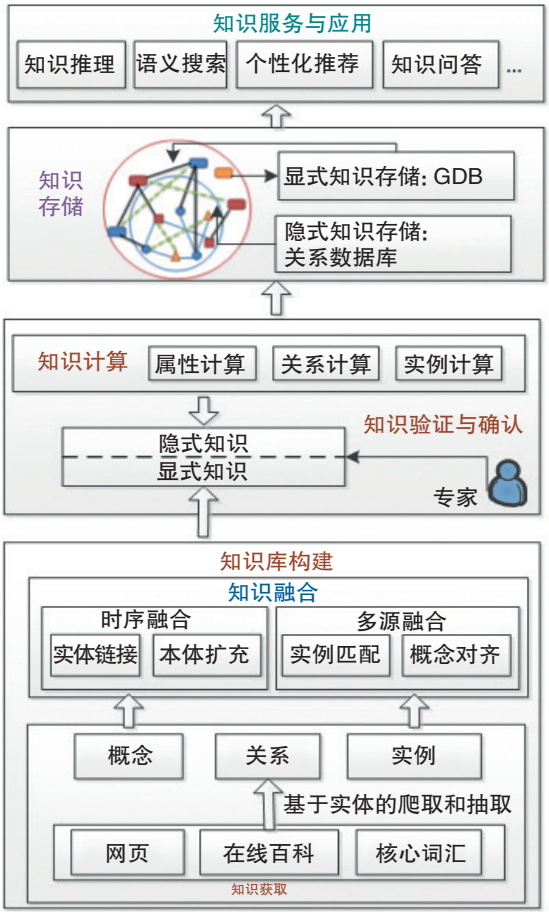

图1 OpenKN架构

知识库的构建 知识库的构建从逻辑角度讲，包括知识获取和知识融合两个方面。其中知识获取的主要目的是从开放网页、在线百科和核心词表等数据中抽取概念、实例、属性和关系。知识融合的主要目的是实现知识的时序融合和多数据源融合。图2进一步描述了知识库构建的思路。OpenKN构建的知识库包含两部分：(1)存储众所周知的常识性知识的通用基础库(general foundation base)，这些知识可从维基百科等在线百科中直接抽取获得。(2)特定领域的知识库，从左至右依次为领域1到领域n。基于每一个领域知识的特点不同，每一个特定领域知识库又可进一步划分为三部分：导出的通用基础库(induced GFB)、领域基础库(domain foundation base)和领域网络库(domain Web base)。具体地讲，导出的通用基础库是指从常识知识中选取的和领域相关的知识构成的知识库。领域基础库是用来描述领域相关的其他基本知识。领域基础库中的知识主要来自领域字典、核心词汇表等。为了获取当前最新最实时的领域知识，领域网络库用来从开放的互联网网页中抽取领域相关的最新知识。在图2中，橙色点和黑色点代表从网页中抽取获得的知识，点之间的边代表知识间的关联关系。随着网页数量的不断增加和内容的不断更新，领域知识库可实现自适应增长(self-grew)[14]。上述这些知识库的构建共同完成知识获取的全过程。此外，我们利用已有的公开知识库，如Freebase，YAGO等实现了知识融合。在完成OpenKN的知识库构建工作后，我们得到的知识称为显式的知识。

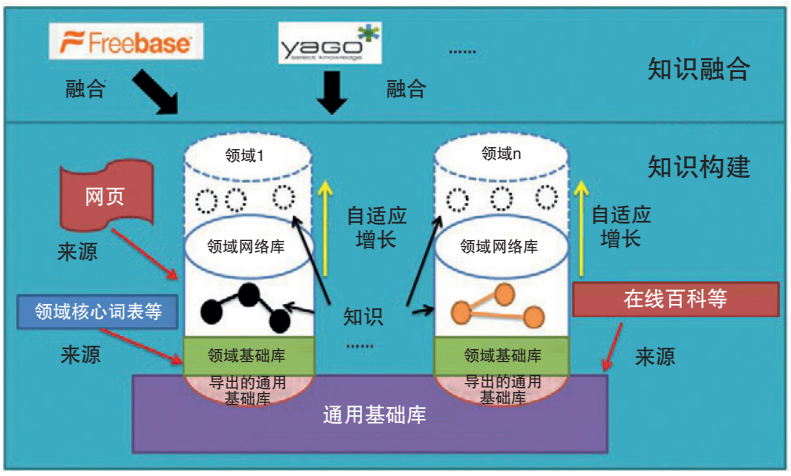

图2 知识库构建

知识计算 除了显式的知识，通过 OpenKN 的知识计算功能，包括属性计算、关系计算 [15,16]、实例计算等，我们还可以进一步获得隐式的或推断的知识。

知识验证与处理 为了检验显式知识和隐式知识的完备性、相关性与一致性，我们需要对知识进行校验，这称为知识验证过程。主要是专家或特定的知识计算方法检查冗余的、冲突的、矛盾的或者不完整的知识。

知识存储 经过验证的海量知识，在 OpenKN 里存储在一个基于图的数据库 (Graph DataBase, GDB) 以及关系数据库中。其中，GDB 中存储的是显式的知识，关系数据库中存储的是隐式的知识。GDB 作为大数据存储基础设施，支持大于 100 亿条知识的存储。与传统的数据库模型（如 Neo4j, Titan）相比，GDB 通过定义点和边的图数据模型来存储知识，这里的点和边都有各自唯一的 ID 并且支持一系列的多值属性。GDB 描述了一个与现有的图模型不同的异构网络，称为可演化知识网络。

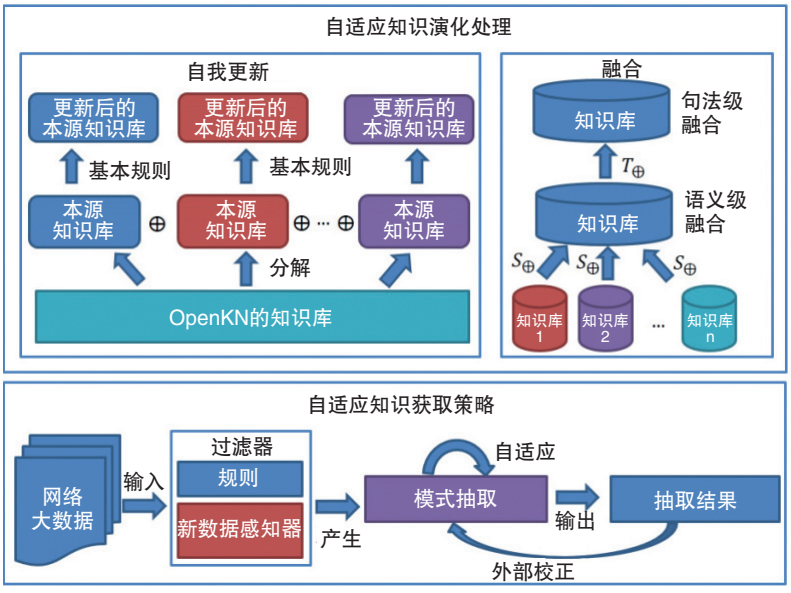

图3 OpenKN的自适应性

OpenKN 的两个主要特征——自适应性和可演化性，加在一起诠释了 OpenKN 的“Open”的含义。

OpenKN的自适应性

OpenKN 的自适应性主要体现在自适应知识演化处理和自适应知识获取策略两个方面。

如图 3 所示，自适应知识演化处理用来描述知识演化的规律，它分为知识库的自我更新和与其他知识库的句法——语义级融合两个阶段。在自我更新阶段，知识演化通过作用在知识库上的两个基本运算和一系列的规则完成，即逻辑加⊕和逻辑乘⊙运算，以及一系列基本的规则。这些规则被用于本源知识库 (primitive KB) 上的演化。这里的本源知识库定义为不可以被其他知识库通过逻辑加和逻辑乘来表示的知识库。如果我们把所有知识库组成的集合定义为一个向量空间，那么根据线性代数的基本知识，这些本源知识库实际上构成了该线性空间的一组基。对于向量空间的若干术语，可参考文献 [17]。另一方面，对于两个不同知识库的融合可分为两个操作，语义级的融合 S ⊕和句法级的融合 T ⊕。

自适应知识获取策略的主要目的是获取随时空演化的动态知识。如图 3 所示，自适应知识获取策略使用一个称为过滤器的组件来产生句法——语义级的抽取模板，例如 Such-As, Is-A，来对网络数据进行知识抽取。过滤器由规则和新数据感知器组成，其中规则保证不同类型的知识库中抽取得到知识的一致性，新数据感知器主要用于检测是否有新的数据产生以动态调整我们的抽取策略。抽取模板的调整是通过其自适应的调整和与抽取结果的反馈来迭代实现的。在自适应调整阶段，例如当 Such-As 模板遇到例外情况时，如句子“animals other than dogs such as cats”，它不仅可以从概率的角度发现这个特例，还可以通过模糊本体技术来识别这种情况，相关术语可参考文献 [18]。在抽取结果的反馈阶段，抽取模板通过抽取结果的正确性进行打分，

实现所谓的外部校正。

OpenKN的自适应性可以有效地满足网络大数据的快速变化带来的挑战。一方面，它可以使知识库具有捕获新数据的能力，另一方面，不同的规则如基本规则，保证了知识的实时更新。

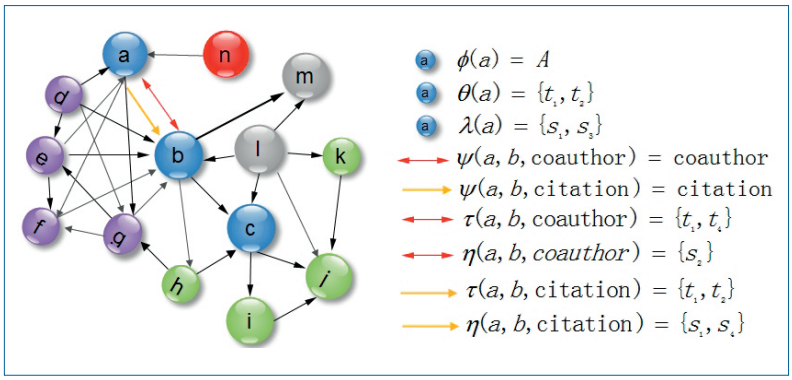

图4　可演化知识网络

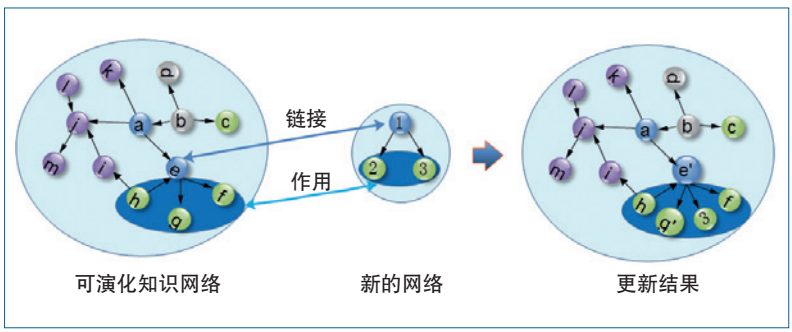

图5　可演化知识网络的更新

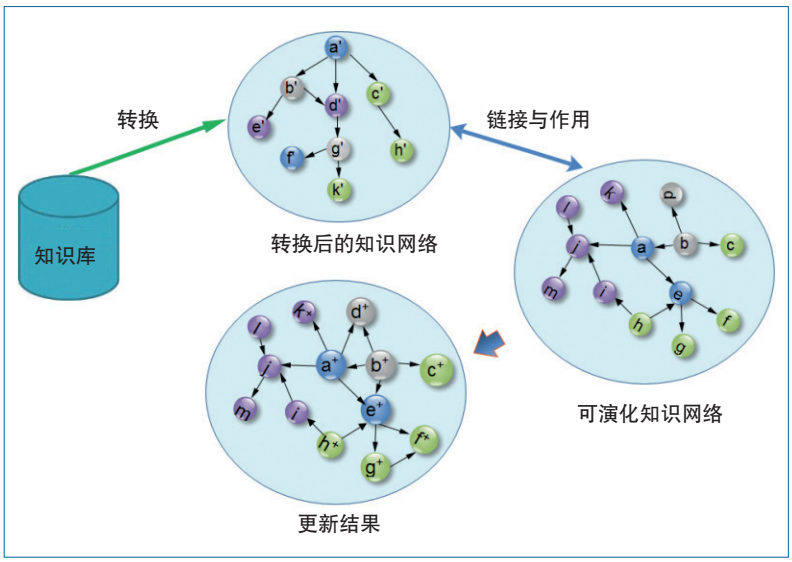

图6　可演化知识网络融合其他知识库

OpenKN的演化计算

为了证明OpenKN的可演化性特征，我们首先引入可演化知识网络这一知识表示模型。这一表示模型也是知识存储设施GDB搭建的基础。

可演化知识网络

这里的可演化知识网络是每一个点和边上的带有时空信息和一系列演化函数或算子的异构网络。更准确地讲，给定时间信息集合T和空间信息集合S，点的类型集合A，边的类型集合R，可演化知识网络$G_{T,S}$可定义为如下的8元组：

$$G_{T,S} = (V, E, \phi, \varphi, \theta, \tau, \lambda, \eta)$$

其中，V是点的集合；E是有向边的集合，即一系列关系对(u, v, r)，$u, v \in V$，$r \in R$，即每对点都被赋予了一个或多个关系；$\phi: V \to A$是一个定义在点集上的映射函数，表示在点集合中，每个顶点通过该计算函数可得到唯一的顶点类型$\phi(v) \in A$；$\varphi: E \to R$是一个关系映射函数，使得每对点之间的关系类型最多有$|R|$个。$\theta: V \to 2^T$是定义在点上的时间映射函数，用于计算特定点的时间戳进而描述点的生命周期。这里2^T是集合T的幂集；$\tau: E \to 2^T$是定义在边上的时间映射函数，返回某一指定边的时间戳用来描述该边存在的时间信息；$\lambda: V \to 2^S$是定义在点上的空间信息映射函数，返回某一指定点

的空间信息用来描述该点活动的空间信息；$\eta: E \to 2^S$ 是定义在边上的空间映射函数，返回某一指定边的空间信息，来描述该边存在的空间信息。在可演化知识网络中，我们记录下了点和边的时空信息。需要指出的是，OpenKN 也可用于构建特殊领域的知识库，其构建方法与上述方法相同，此时定义中的点类型映射函数 ϕ 与边类型映射函数 φ 都是依赖于领域的。

以学术领域为例，我们构建了一个面向学术圈的可演化知识网络：在此网络中，点的类型可定义为作者 (A)、文章 (P)、会议 (C)、组织 (O) 和关键词 (K) 五类；边的类型包括作者间的合作关系和论文间的引用关系等。此外，每一个点都包含相应的时空信息，例如出生日期、出生地、归属地、毕业时间等。边也包含相应的时空信息，例如两个作者合作的年份、合作地点等。在图 4 中，我们以可演化知识网络中的一小部分作为例子进行阐述。此网络包含点集 $\boldsymbol{V}=\{a,b,c,d,e,f,g,h,i,j,k,l,m,n\}$，时间集合 $\boldsymbol{T}=\{t_1,t_2,t_3,t_4,t_5,t_6\}$，空间集合 $\boldsymbol{S}=\{s_1,s_2,s_3,s_4,s_5\}$。图 4 的右侧是一些函数以及相应的函数值，例如等式 $\psi\{a,b,coauthor\}=coauthor$ 表示两个点 a, b 之间存在合作关系；等式 $\phi(a)=A$ 表示点 a 的类型是 A；等式 $\theta(a)=\{t_1,t_2\}$ 和 $\lambda(a)=\{s_1,s_3\}$ 分别表示在点 a 上的时间和空间映射函数值；等式 $\tau\{a,b,coauthor\}=\{t_1,t_4\}$ 和 $\eta\{a,b,coauthor\}=\{s_2\}$ 分别表示在边 $(a,b,coauthor)$ 上的时间和空间映射函数值。

OpenKN 之所以称为可演化知识网络，主要是因为：一方面网络可以不断获取最新的知识，并进行自我更新；另一方面，知识网络可以将其他知识库中的知识转化为自己可以利用的标准形式，吸纳到自身的知识网络中进而形成新的知识网络。这两个过程如图 5 和图 6 所示。在图 5 中，最左侧的网络是一个确定的可演化知识网络，当从网页中获取了新的知识后，可以通过两个步骤将新知识融合到现有的网络中。首先，将新知识表示为一个知识网络，并将其与现有网络进行“链接”。其次，将新的知识网络中的点和边与已有网络中的相应的点和边进行“作用”，最终形成一个网络。在图 6 中，现有的知识库先转化为一个知识网络，然后如图 5 所示，和已有的知识网络进行融合。

可演化知识网络的演化特性构成一个完整的演化周期 [19]，包括演化识别与感知、演化定位、演化评估和管理等阶段。这种演化的非刚性同时保证了网络的时新性。

演化计算算子库

OpenKN 的演化计算可规范化为两类不同的算子或操作，即对点的操作和对边的操作。具体地讲，对点的操作单元可分为点的抽取、点的融合以及点的推理三个子操作，对边的操作也可分为相应的三个类似的子操作，即关系抽取、关系融合、关系推断。这里提到的所有操作均涉及到对点和边上的时间与空间信息的操作，并且这些操作与前文提到的自适应知识演化过程和自适应知识获取策略是一致的。即点和边的抽取实现了自适应知识获取，其他操作构成了自适应知识演化过程。OpenKN 的演化计算算子库首次将知识获取的整个流程中涉及的方法纳入到一个体系当中，便于深入理解每个方法之间的关系，为不同的方法及其之间的衔接与相互作用提供了一个全面的视角。

目前，OpenKN 这一知识计算引擎能够处理的点规模达到 3000 万，边的规模达到 10 亿级，同时处理规模仍在不断扩张中。

总结

网络大数据具有多源异构性、时效性、高噪声等特点，不但非结构化数据多，而且数据的实时性强。网络大数据背后蕴含着丰富的、复杂关联的知识。要有效利用网络大数据的价值就须进行数据的去冗分类、去粗取精，从数据中挖掘知识，对大数据网络背后的知识进行深入分析。本文提出了一种面向网络大数据的知识计算引擎——OpenKN。它的主要特点是自适应性和可演化性。这使得 OpenKN 可以更好地感知动态变化的网络知识，同时对潜在的和变化的时序知识进行推断和预测，更好地为网络大数据下的知识挖掘提供服务。■

王元卓

CCF高级会员。中科院计算所副研究员。主要研究方向为网络大数据知识计算、社会计算等。wangyuanzhuo@ict.ac.cn

贾岩涛

CCF会员。中科院计算所助理研究员。主要研究方向为知识计算、数据挖掘、组合优化。jiayantao@ict.ac.cn

赵泽亚

信息工程大学研究生。主要研究方向为知识工程和数据挖掘。zhaozeya@software.ict.ac.cn

程学旗

CCF杰出会员、杰出演讲者。中科院计算所研究员。主要研究方向为网络科学、网络与信息安全以及互联网搜索与服务。cxq@ict.ac.cn

参考文献

[1] 王元卓,靳小龙,程学旗. 网络大数据：现状与挑战.计算机学报, 2013;36(6): 1~15.

[2] 王元卓,贾岩涛,刘大伟等.基于开放网络知识的信息检索与数据挖掘.计算机研究与发展. 2014.

[3] Etzioni O, Cafarella M, Downey D, et al. Web-scale information extraction in knowitall: (preliminary results)[C]. *Proc of the 13th Int Conf on World Wide Web*. New York: ACM, 2004:100~110.

[4] Banko M, Cafarella M J, Soderland S, et al. Open information extraction from the web[C]. *Proc of the 20th Int Joint Conf on Artifical Intelligence, IJCAI'07*. New York: ACM, 2007:2670~2676.

[5] Carlson A, Betteridge J, Kisiel B, et al. Toward an architecture for never ending language learning[C]. *Proc of the 24th AAAI Conf on Artifical Intelligence*. Menlo Park, CA: AAAI Press, 2010:1306~1313.

[6] Wu W, Li H, Wang H, et al. Probase: A probabilistic taxonomy for text understanding[C]. Proc of the 2012 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2012:481~492.

[7] Gallagher S. How Google and Microsoft taught search to understand the Web. 2012[2013-07-25]. http://arstechnica.com/information-technology/2012/06/inside-the-architecture-of-googles-knowledge-graphand-microsofts-satori/.

[8] Nakashole N, Theobald M, Weikum G . Scalable knowledge harvesting with high precision and high recall[C]. *Proc of the 4th ACM Int Conf on Web Search and Data Mining*. New York: ACM, 2011:227~236.

[9] Suchanek F M, Sozio M,Weikum G . SOFIE: A self-organizing framework for information extraction[C] . *Proc of the 18th Int Conf on World Wide Web*. New York: ACM, 2009:631~640.

[10]Auer S, Bizer C, Kobilarov G, et al. DBpedia: A nucleus for a Web of open data[C]. *Proc of the 6th Int the Semantic Web and 2nd Asian Conf on Asian Semantic Web Conf, ISWC'07*. Piscataway, NJ: IEEE, 2007:722~735.

[11]Biega J, Kuzey E, Suchanek F M. Inside YAGO2s: A transparent information extraction architecture[C]. *Proc of the 22th Int Conf on World Wide Web*. New York: ACM, 2013:325~328.

[12]Philpot A, Hovy E H, Pantel P. Ontology and the lexicon [M]. The Omega Ontology. Cambridge: Cambridge University Press, 2008 35~78.

[13]Ponzetto S, Navigli R. Large-scale taxonomy mapping for restructuring and integrating wikipedia[C]. *Proc of the 21st Int Joint Conf on Artifical Intelligence, IJCAI'09*. San Francisco: Morgan Kaufmann,2009:2083~2088.

[14]Hailun Lin, Yantao Jia, Yuanzhuo Wang, and et al.. Populating Knowledge Base with Collective Entity Mentions: A Graph-based Approach. IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM) 2014.

[15]Yantao Jia, Yuanzhuo Wang, Xueqi Cheng,and et al.. OpenKN: An Open Knowledge Computational Engine for Network Big Data. IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM) 2014.

[16]Zeya Zhao, Yantao Jia and Yuanzhuo Wang. Content-Structural Relation Inference in Knowledge Base.AAAI2014.

[17]D. C. Lay. Linear algebra and its applications,1997.

[18]Y. Cai, C.-m. A. Yeung, and H.-f. Leung.Fuzzy computational ontologies in contexts. 2012.

[19]F. Zablith, G. Antoniou, M. d'Aquin, and et al..Ontology evolution: a process-centric survey. The Knowledge Engineering Review, 2013:1~31.