计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20131342
52(2):456-474, 2015

基于开放网络知识的信息检索与数据挖掘

王元卓[1] 贾岩涛[1] 刘大伟[2] 靳小龙[1] 程学旗[1]
[1](中国科学院网络数据科学与技术重点实验室(中国科学院计算技术研究所) 北京 100190)
[2](烟台中科网络技术研究所/中国科学院计算技术研究所烟台分所 山东烟台 264005)
(wangyuanzhuo@ict.ac.cn)

Open Web Knowledge Aided Information Search and Data Mining

Wang Yuanzhuo[1], Jia Yantao[1], Liu Dawei[2], Jin Xiaolong[1], and Cheng Xueqi[1]
[1](*CAS Key Laboratory of Network Data Science & Technology* (*Institute of Computing Technology*, *Chinese Academy of Sciences*), *Beijing* 100190)
[2](*Institute of Network Technology/ICT(YANTAI)*, *Chinese Academy of Sciences*, *Yantai*, *Shandong* 264005)

Abstract Network big data refers to the massive data generated via interaction and fusion of the ternary human-machine-thing universe in the cyberspace and available on the Internet. It has a few typical features, such as multi-sourced, heterogeneous, interactive, bursty, and noisy. It contains mainly unstructured data, and has strong real-timeness. Network big data implicitly contains tremendous highly-interconnected knowledge. Building up open Web oriented large-scale knowledge bases is an effective means for obtaining rich knowledge from network big data. This paper compares both the domestic and international mainstream open Web knowledge bases. We specifically analyze the core techniques and methods for constructing open Web knowledge bases, fusing multi-sourced knowledge, and updating the knowledge bases. Furthermore, we summarize the research status and main issues of open Web knowledge base based information search, data mining, and system applications from different aspects, including user intension understanding, query extension, semantic Q&A, clue mining, relationship referencing, and prediction of relationships and attributes. Finally, we look into the development trends and main challenges of open Web knowledge bases.

Key words network big data; open Web knowledge; ontology; information search; data mining

摘 要 网络大数据是指“人、机、物”三元世界在网络空间(cyberspace)中交互、融合所产生并在互联网上可获得的大数据.这些数据具有多源异构、交互性、时效性、社会性、突发性和高噪声等特点,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.建立面向开放网络的知识库是获取网络大数据中的丰富知识的有效手段.对当前国内外主要的开放网络库进行了比较,分析了相应的构建方法、多源知识的融合以及知识库的更新等关键技术.进一步从用户意图理解、查询扩展、语义问答、线索挖掘、关系推理以及关系和属性预测等方面出发,总结了基于开放网络知识库的信息检索、数据挖掘与系统应用的研究现状和主要问题.最后,对开放网络知识库的发展趋势和面临的主要挑战进行了展望.

收稿日期:2013-10-12;**修回日期**:2014-04-18
基金项目:国家“九七三”重点基础研究发展计划基金项目(2014CB340401,2013CB329601);国家自然科学基金项目(61173008, 61100175, 61232010, 60933005, 61402442);北京市科技新星计划项目(Z121101002512063);北京市自然科学基金青年基金项目(4154086)

关键词 网络大数据;开放网络知识;本体;信息检索;数据挖掘

中图法分类号 TP182

近年来互联网技术和应用模式快速发展,在改变人们生活方式的同时,也产生了巨大的数据资源.根据互联网数据中心的报告,2012 年全球的数据总量为 2.7 ZB(1 ZB 相当于 10 万亿亿字节),预计到 2020 年,全球的数据总量将达到 35 ZB,远远超过人类有史以来所有印刷材料的数据总量(200 PB).Google 每月处理的数据量超过 400 PB;百度每天大约要处理几十 PB 数据;Facebook 注册用户超过 10 亿,每天生成 300TB 以上的日志数据.网络空间(cyber space)中各类应用的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信息,根据其产生方式的不同可以分为 Web 内容数据、Web 结构数据、自媒体数据、日志数据.其中,Web 内容数据主要是通过互联网网页产生和发布的数据,它既可以是文字、文本、消息,也可以是图片音视频等,以及 HTML,Java scripts,Interstitial 间隙窗口、Microsoft Netshow,Flash 等所产生或解析的数据.如今,Web 内容数据量呈指数级增长,例如检索网页的总量达 500 亿,在线图书网页达 7 亿 5 000 万,其中英文维基百科数量达 427 万个页面,中文百科数据达 900 万个页面.Web 内容数据的特点既包括数据量巨大、内容信息丰富,还具有动态更新快、多源异构等.Web 结构数据是指 Web 页面间的结构数据,主要包括页面间的超链接关系和 Web 的组织结构.伴随着 Web 内容数据的增长,Web 页面间的链接关系也呈现出大规模增长的趋势.

自媒体数据主要是指通过以 Facebook,Twitter 等为代表的社交网络中产生的用户生成数据(user generated content,UGC),具有空前的规模性和群体性,数据总量巨大,数据变化非常快.1 min 内,Twitter 上新发的数据量超过 10 万条;Facebook 用户每天分享的内容条目超过 25 亿个,数据库中的数据每天增加超过 500 TB.此外,自媒体数据还具有十分复杂的内在关系,超过 10 亿的 Facebook 用户的好友关系和超过 5 亿的 Twitter 用户之间的关注关系构成了极为复杂的关系网络.

日志数据主要指各种网上服务提供商积累的系统和用户操作的日志记录,比如 Google、百度等搜索引擎提供商积累的用户搜索行为日志等.此类数据的特点是具有大量的历史性数据,同时数据增速极快、数据访问吞吐量巨大.以 Google 为例,目前有超过 200 个 Google 文件系统集群在运行,而每个集群大约有 1 000~5 000 台机器,每个谷歌文件系统(Google file system,GFS)都存储着高达 5 PB 的数据;成千上万的机器需要的数据都从 GFS 集群中检索,这些集群中数据读写的吞吐量可高达 40 GBps,每天都在产生着富含大量知识的数据.

这些网络大数据中,直接反映的往往是一个个孤立的数据和分散的链接,但这些反映相互关系的链接整合起来就是一个网络[2].该网络中的数据具有多源异构、交互性、时效性、社会性、突发性和高噪声等特点,不但非结构化数据多,而且数据的实时性强.数据自身的信息、数据间的关联信息以及网络的结构特征等都隐藏在这样的数据网络中,网络大数据往往以复杂关联的数据网络这样一种独特的形式存在.有效利用网络大数据价值的主要任务不是获取越来来越多的数据,而是数据的去冗分类、去粗取精,从数据中挖掘知识,对大数据网络后面的知识进行深入分析.社交网络中蕴含了丰富的用户的交友模式、用户的情感演化以及用户间消息的传播规律等知识;电子商务网站如 Amazon 等拥有的用户浏览、购买和评论商品的信息中揭示了用户的购买行为规律;搜索引擎服务商存储的大量的用户日志信息中隐含了用户的查询意图和查询请求,这些真实的查询记录能够帮助搜索引擎提供商深入理解用户真正关注哪些实体,想查询哪些实体以及对于每个实体,用户真正感兴趣的属性有哪些.

传统的信息检索技术致力于从海量数据中过滤对用户有用的信息,然后将这些过滤结果或称为知识返回给用户,但是这些结果间存在的内在关联并没有被很好地挖掘和解析出来,缺乏从语义角度去挖掘深层次的规律和知识的能力.用户只能从结果中自己去理解和筛选知识,更何况检索技术本身也存在缺陷,如查全率和精确度不高,仍然不能满足用户的需求,效果远不能使人满意.值得欣慰的是,目前搜索引擎正在从基于关键字或者基于文本内容检索这一浅层次的知识理解和挖掘工作,向表达、理解语义和关系这些深层次的知识挖掘方面发展.互联网的创始人 Lee[3] 在 XML2000 国际会议上正式提出语义 Web 的体系框架.而语义 Web 中“语义”

的核心就是知识共享,包括计算机与计算机、人与计算机之间的共享.在计算机与计算机、人与计算机之间以无偏差的方式传递信息[4].传统的 Web 资源中的语义信息或领域知识以机器难以处理的自由文本的方式存在,资源间的语义关系是以一种隐含的方式存在,这些语义信息由于缺乏明确的描述而丢失.

本体(ontology)是语义 Web 的基础[3],语义 Web 在本质上是基于本体的 Web,本体可以有效地进行知识表达、知识查询或不同领域知识的语义消解.本体还可以支持更丰富的服务发现、匹配和组合,提高自动化程度[5].本体[6]是领域概念及概念之间关系的规范化描述,这种描述是规范的、明确的、形式化的、可共享的."明确"意味着所采用概念的类型和它们应用的约束实行明确的定义;"形式化"指本体是计算机可读的(即能被计算机处理);"共享"反映知识本体应捕捉该领域中一致公认的知识,反映的是相关领域中公认的概念集,即知识本体针对的是团体而非个体的共识.本体定义为"给出构成相关领域词汇的基本术语和关系,以及利用这些术语和关系构成的规定这些词汇外延的规则的定义".Neches 认为:"本体定义了组成主题领域的词汇表的基本术语及其关系,以及结合这些术语和关系来定义词汇表外延的规则.

基于本体构建的开放网络知识库是一个面向开放网络的、动态的、可增量的知识库.开放网络是指知识的来源是多元化的,包括来自互联网的非结构化多语言文本数据(如时事新闻、电子邮件、微博、社交媒体网站的帖子、即时通信以及可以转换成文本的信息等)、半结构化的在线百科知识库、机器可读的结构化语言知识库(如各类词库、专名库、主题词表、标注语料库等).动态是指领域知识和语言知识的要素之间关系以及知识本身的属性是动态变化的,为了把握这种动态变化的规律与机理就需要采用概率化等计算方法判定其不确定性.可增量是指知识的规模可以满足用户对于知识获取的实时性要求而进行动态的扩展.

基于开放网络大数据构建知识库是国内外工业界开发和学术界研究的一个热点.目前,世界各国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[7-8], TextRunner[9], NELL[10], Probase[11], Satori[12], PROSPERA[13], SOFIE[14] 以及一些基于维基百科等在线百科知识构建的知识库 DBpedia[15], YAGO[16-18], Omega[19], WikiTaxonomy[20-21].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①;美国官方政府网站 Data.gov②;Wolfram 的知识计算平台 wolframalpha③;Google 的知识图谱(knowledge graph);Facebook 推出的类似的实体搜索服务 graph search 等.在国内,中文知识图谱的构建也有大量的研究和开发工作.代表性工作有:中国科学院计算技术研究所的基于 OpenKN(开放知识网络)的"人立方、事立方、知立方系统";中国科学院数学与系统科学研究院的陆汝钤提出的知件(knowware);上海交通大学最早构建的中文知识图谱平台 zhishi.me;百度推出的中文知识图谱搜索;搜狗推出的知立方平台;复旦大学 GDM 实验室推出的中文知识图谱展示平台等.

其中,就规模而言,拥有概念最多的知识库是 Probase,目前核心概念约 270 万,概念总量达到千万级.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 Google 的知识图谱,目前规模是 5 亿个实体对象和 350 亿条实体间的关系信息,而且规模在随着信息的增长不断地增加;Probase 也是近几年比较热门的知识库.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.除此之外,比较有特色的还有国内搜狗的知立方系统,侧重与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.

本文首先分析了网络大数据环境下的 Web 内容数据、自媒体数据和日志数据中富含大量的语义知识,并介绍了国内外基于开放网络知识建立的知识库的整体现状.然后分析了开放网络知识的构建、多源知识的融合以及知识的更新等方法的主要特点,论述了开放网络知识库在信息检索和知识挖掘方面的主要作用,并介绍了当前知名系统和应用的情况.最后对开放网络知识库的研究与应用面临的挑战和未来研究的重点进行了展望.

1 开放网络知识库的构建

开放网络知识库的构建包括 3 个部分,即知识

① http://www.evi.com

② http://www.data.gov/semantic/index

③ http://www.wolframalpha.co

库的构建、多源知识的融合以及知识库的更新.由于基于本体的知识描述要素主要包括概念、实例、属性和关系,在本节,我们针对这4个要素展开论述,即如何从开放网页等数据源中识别、抽取这4个要素.这里,概念也称为类,是指对象的类型或事物的种类.实例也称为个体,是指基础的或底层的对象.属性是指概念或实例具有的属性、特征、特性、特点和参数.关系是指概念和实例之间的彼此关联的可能具有的方式.

1.1 知识库的构建

开放网络知识库的构建,就是要构建几个基本的构成要素,包括抽取概念、实例、属性和关系.从构建方式上可以分为手工构建和自动构建.手工构建是依靠专家知识编写一定的规则,从不同的来源收集相关的知识信息,构建知识的体系结构[22].比较典型的例子是知网(Hownet)[23]、同义词词林[24]、概念层次网络(HNC)[25]和中文概念词典(CCD)[26]、OpenCyc[27]等.自动构建是基于知识工程、机器学习,人工智能等理论自动从互联网上采集并抽取概念、实例、属性和关系[28-29].比较著名的例子是Probase,YAGO等.手工构建知识库,需要构建者对知识的领域有一定的了解,才能编写出合适的规则,开发过程中也需要投入大量的人力物力.相反,自动构建的方法依靠系统自动地学习经过标注的语料来获取规则的,如属性抽取规则、关系抽取规则等,一定程度上可以减少人工构建的工作量.随着网络大数据时代的到来,面对大规模网页信息中蕴含的知识,自动构建知识库的方法越来越受到广大学者和工程开发人员的重视和青睐.本节主要介绍通过自动的方式从大规模的开放网络数据构建的图形化知识库.如上文所述,自动构建的方法需要系统学习标注语料获取规则,根据系统是否具备在学习过程中扩展规则的能力,自动构建知识库的方法主要分为有监督的构建方法和半监督的构建方法2种.

有监督的构建方法是指系统通过学习训练数据,即标注网页,获取抽取规则,然后根据这些规则,提取同一类型的网页中的概念、实例、属性和关系[30].比较著名的方法有kernel methods[31-32].这类方法的缺点是规则缺乏普适性,由于规则是针对特定网页的,当训练网页发生变化,需要重新进行训练来获取规则.

半监督的构建方法是系统预先定义一些规则作为种子(seed),然后通过机器学习算法,从标注语料中抽取相应的概念、实例、属性和关系.进一步,系统根据抽取的结果,发现新的规则,再用来指导抽取相应的概念、实例、属性和关系,从而使抽取过程能够迭代地进行.这种方法又称bootstrapping(步步为营)的方法,也是目前主流的方法.该方法的最早版本是由Brin[33]提出的,但该版本在种子规则选择、种子规则质量评估、大规模标注网页预处理等方面具有一定的缺陷,后续出现很多对它的改进,比较著名的有华盛顿大学Etzioni领导的研究组开发的KnowItAll方法[7-8]和TextRunner方法[9]等.图1显示了KnowItAll的主要组件和工作流程.bootstrapping组件用于构造抽取规则(extraction rules)和鉴别性的短语(discriminator phrases),Extractor组件构造一个由规则中关键词触发的查询传给搜索引擎,应用抽取规则抽取特定信息.Assessor组件利用搜索引擎计算抽取结果的正确性,将正确的结果插入知识库当中.注意KnowItAll使用的bootstrapping技术是完全自动化的,它不需要事先标注的训练语料,从而避免了对大规模语料的预处理中遇到的标注错误和预处理不当等问题,使得抽取规则更为准确.但是KnowItAll在执行起来效率较低,这是因为需要

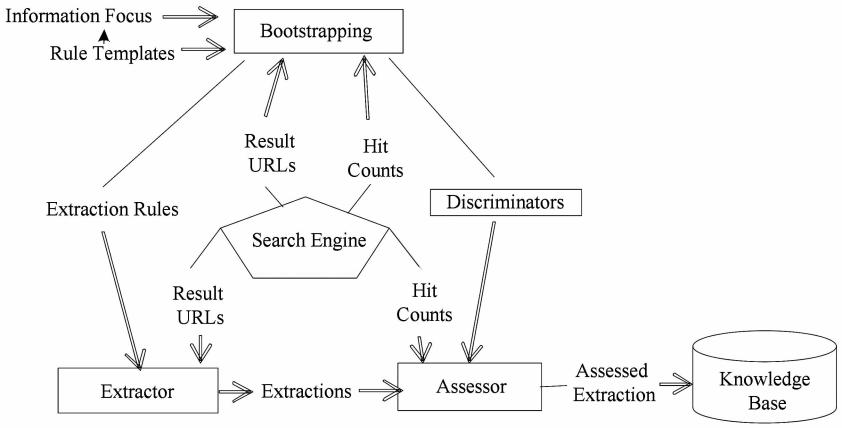


Fig. 1 Flowchart of the main components in KnowItAll.

图1 KnowItAll主要组件和工作流程图

大量的搜索查询,而且由于 KnowItAll 的输入是规则,系统 1 次只能处理 1 条规则,运行多条规则需要串行. 为了解决 KnowItAll 的效率问题,TextRunner[9]引入了 3 个组件,输入为语料样本的自监督学习器(self-supervised learner),可并行抽取的单通道抽取组件(single-pass extractor)和通过模型评估结果的冗余评估器(redundancy-based assessor).

如前文所述,bootstrapping 的方法具有识别新的规则或上下文的优势,这些新的规则和上下文又可以转而用于识别新的概念、实例、属性和关系等结果,从而极大地提高了召回率. 但是,如果这些抽取结果具有歧义就很容易产生错误,从而导致语义漂移(semantic drift)现象[34-35]. 为了克服这个缺陷,很多学者从不同角度对 bootstrapping 算法进行改进:如卡内基梅隆大学的研究人员开发出一个新的知识抽取系统 NELL(never-ending language learner)[15];微软亚洲研究院王海勋博士团队开发的概率化知识库 Probase 等[20]. 我们首先介绍 NELL,其系统架构图如图 2 所示[15]:

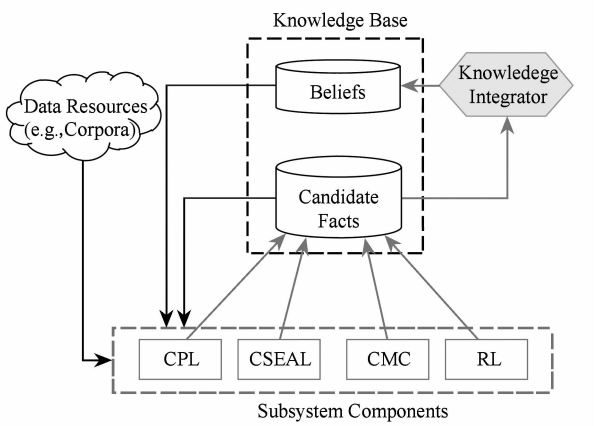


Fig. 2 Architecture of NELL.

图 2 NELL 系统框架

NELL 系统包括知识库、知识整合器(knowledge integrator)和子系统组件 3 个部分. 系统首先将外部数据资源输入到子系统组件中. 经过子系统组件中 4 个组件:CPL(coupled pattern learner),CSEAL(coupled SEAL),CMC(coupled morphological classifier)和 RL(rule learner)的处理,输出候选事实(candidate facts). 知识整合器对候选事实进行过滤,生成置信度较高的结果(beliefs). 最后将候选事实和结果作为输入,输入到子系统组件中,实现系统的迭代运转. 其中,CSEAL 组件提供了一种半监督的学习算法用于信息抽取,称为耦合的半监督的信息抽取学习方法(coupled semi-supervised learning for information extraction)[36],提供了用以过滤低质量的规则鉴别器,可以有效解决 bootstrapping 算法产生的语义漂移现象. 事实上,对于语义漂移现象,单纯依靠不断构建高质量的规则还是不够的,这只是停留在句法角度(syntactic),这样高质量的规则可能是很少的[11]. 所以需要从语义角度定义迭代抽取的方法,即直接建立在知识基础上的 bootstrapping 算法,应用知识去理解文本并获取更多的知识. 相比于 KnowItAll,TextRunner 和 NELL 这些基于句法的 bootstrapping 方法,基于语义的 bootstrapping 能够量化地反映出抽取对象间的语义关联,更好地从大规模网络数据中构建语义级的知识库,因而成为近年研究的热点问题.

微软推出的 Probase 是一种基于语义抽取构建的知识库[11,37],它通过迭代方式进行信息抽取的特点是:每轮迭代都会聚集一些置信度高的正确的知识,用以指导下一轮迭代的信息抽取. Probase 获取知识并不依赖于很多的句法规则,而是在每轮迭代中固定一些规则(具体为 Hearst patterns). 以抽取概念间的 isA 关系为例,Probase 主要通过 Syntactic-Extraction,SuperConcept Detection 和 Sub-Concept Detection 3 个模块迭代的对候选句子中的概念进行 isA 关系的抽取. 其中,SyntacticExtraction 用以从一句话中发现候选 super-concept 和候选 sub-concept. 如果候选 super-concept 多于 1 个,则运行模块 SuperConcept Detection 用以减少 super-concept 的个数. 然后,用 Sub-Concept Detection 模块过滤掉一些不太可能的 sub-concept. 这里,Super-Concept Detection 模块和 Sub-Concept Detection 模块运用了概率化的手段用以排除不正确的概念. 以 Super-Concept Detection 为例,记 $X=\{x_1,x_2,\cdots,x_m\}$ 为候选 super-concepts 集合,$Y=\{y_1,y_2,\cdots,y_n\}$ 为候选 sub-concepts 集合. 如果 $|X|>1$,为了过滤集合 X 中的元素 x_k,我们计算似然函数 $P(x_k|Y)$. 不失一般性,假定 super-concept 中 x_1 和 x_2 具有最大的似然函数,且满足 $P(x_1|Y)\geqslant P(x_2|Y)$. 我们计算这两个似然函数的比值 $r(x_1,x_2)$ 如下:

$$r(x_1,x_2)=\frac{P(x_1\mid Y)}{P(x_2\mid Y)}=\frac{P(Y\mid x_1)P(x_1)}{P(Y\mid x_2)P(x_2)}=\frac{P(x_1)\prod_{i=1}^{n}P(y_i\mid x_1)}{P(x_2)\prod_{i=1}^{n}P(y_i\mid x_2)},$$

其中第 3 个等号成立时,假设给定 super-concept,且各个 sub-concept 是相互独立的. 当 $r(x_1,x_2)$ 大于某一事先给定的阈值时,我们选择 super-concept 为 x_1. 否则,选择 super-concept 为 x_2. 这样就可以

实现集合 X 中元素的过滤功能. Sub-Concept Detection 的过程和上述过程类似. Probase 从多达 16 亿的网页数据中抽取出核心概念 270 万、概念间的关系 2 000 多万,基本实现了从大规模开放网络中自动构建知识库的需求.

开放网络知识库的构建仍然面临很多问题. 上述几个知识库的构建过程仍然无法解决大规模自动化抽取所面临的知识多样化精准抽取问题、知识来源的多样化选取问题、知识的深度语义关联理解问题等问题. 首先,在知识的多样化精准抽取方面,以概念间关系为例,上述知识库自动化抽取的概念间关系类型多以 isA 关系为主,缺乏对概念间其他关系的处理. 其次,在知识来源的多样化选取方面,上述知识库的构建尚未将网络大数据组成之一的自媒体数据作为一个重要的知识来源. 自媒体数据本身的变化快、富含短文本信息等特点,给知识库的构建带来了诸多难题. 最后,在知识的深度语义关联的理解方面,在构建知识库的过程中,如何将网络大数据背后蕴含的丰富语义关联更好地体现在知识库构建过程中,尽可能保留这些语义信息并体现在概念之间的关系以及实例间的关系中,目前的工作还没有深入的涉及. 上述问题的解决思路之一是充分利用自然语言处理领域的积累,如句法分析方法、篇章分析方法等,以及语言的可计算性理论等基础工作,理解网络大数据背后蕴含的语言规律,更好地指导开放网络知识库的构建.

1.2 多源知识的融合

多源知识的融合是为了解决知识的复用问题. 如上文所述,构建一个知识库的代价是非常大的,为了避免从头开始,需要考虑知识的复用和共享,这就需要对多个来源的知识进行融合,即需要对概念、实例、属性和关系的冲突、重复冗余、不一致进行数据的清理工作,包括对概念、实例进行映射、消歧,对关系进行合并等. 在本文中主要讨论概念间关系或分类体系的融合方法. 按融合方式可以分为手动融合和自动融合. 对于规模较小的知识库手动融合是可行的,但是一件非常费时、容易出错的融合方式. 相比于手动融合方式,建立在机器学习、人工智能和本体工程等方法的融合方式具有更好的可扩展性,相关工作包括 YAGO[16-18],Probase[38] 等.

YAGO 知识库将维基百科、WordNet 和 GeoNames 等数据源的知识整合在知识库中. 其中,将维基百科的分类体系和 WordNet 的分类体系进行融合是 YAGO 的重要工作之一. 维基百科的分类是一个有向无环图生成的层次结构[16],这种结构由于仅能反映主题信息,所以容易出错. 例如将足球运动员界定为足球. 所以 YAGO 仅保留了维基百科分类层次中的叶子分类节点,然后利用 WordNet 提供的基于同义词的分类体系构建其他分类节点. 为了解决这其中出现的词语歧义问题,YAGO 使用了基于共现和 WordNet 词频统计等手段构建了完整的分类体系. 由于 YAGO 是基于维基百科、WordNet 等结构化的、质量较好的知识源构建的,其概念空间在几十万的规模[18]. 从大规模半结构化和无结构化的网页中提取的海量的分类体系的融合面临更多的噪音、冗余和信息缺失现象,微软的 Probase 给出了一种基于概率化的实体消解(entity resolution)的知识整合技术[38],将现有结构化数据如 Freebase,IMDB,Amazon 等整合到 Probase 当中. Probase 将数据融合问题视为对已知分类体系进行匹配和映射的问题,这首先需要解决实体消解问题,即判定 2 个实体是否为同一个实体. 具体来讲,它将外部数据源的数据分为 positive evidence 和 negative evidence,然后将实体消解问题转化为寻找图的最优 multi-way cut 问题. 该实体消解方法可以解决实体属性信息不足的问题. 除此之外,为了解决大规模的实体消解的时间代价高的问题,Lee 等人[38] 引入了 2 种原则:CnD 原则和 BoF 原则,来保证该实体消解技术具有很好的可扩展性.

对多源知识的融合,除了分类体系的融合外,还包括对实体和概念的消解问题、实体和概念的消歧问题等. 面对海量知识库时,建立若干个针对不同领域、不同需求的有效的知识融合算法,快速进行多元知识的融合,是亟须进一步解决的问题之一.

1.3 知识库的更新

大数据时代数据的不断发展与变化带给知识库构建的一个巨大的挑战是知识库的更新问题. 知识库的更新分为 2 个层面:一是新知识的加入;二是已有知识的更改. 目前专门针对开放网络知识库的更新的研究工作较少,很多研究都是在数据库的更新角度展开的,如对数据库数据的增加、删除和修改工作的描述. 虽然对开放网络知识库的更新和数据库的更新有很多相似之处,但是其本身对更新的实时性要求较高. 目前这方面的工作从更新方式来讲分为 2 类:一是基于知识库构建人员的更新;二是基于知识库存储的时空信息的更新. 前者准确性较高,但是对人力的消耗较大. 后者多由知识库自身更新,需要人工干预的较少,但是存在准确率不高的问题. 我们分别介绍这 2 方面的工作,即知识库 YAGO2[16,39] 和 NELL[40-41] 的更新工作.

NELL系统实现自我更新是通过自我更正(self-correct)来实现的.截止2010年10月,NELL系统已经从互联网上抽取到44万个事实,准确率为87%[40].NELL系统并不是百分之百正确,有时它也会导出一些错误的结论,比如错误的关系等.对于这些错误,NELL系统的做法是定期(几周)由团队维护人员对系统的错误进行修正[40].

YAGO2是YAGO[18]的扩展.相比于YAGO,它存储的实体和事实都系统地构建了时间戳.对于实体,它记录了实体的存在时间.例如,通过人的出生年月和逝世年月计算出人的存在时间,通过组织机构等的创建时间和解体时间计算出组织机构的存在时间等.对于事实,记录发生的时间点或时间范围.例如记录现任美国总统奥巴马的上任时间和任期范围等.YAGO2还结合数据源GeoNames对每个实体和事实赋予空间维度.例如人的出生地、组织所在地等.同时,YAGO2还构建了一种时空知识表示模型SPOTL,方便对知识库中的时空信息进行查询.有了这些时间戳和查询语言,知识库的更新可以通过系统定期对实体和事实的时间戳进行更新来完成.

总体上讲,对知识库的更新仍然没有很有效的方法.尤其是在面对用户对知识的实时性更新需求方面,远远达不到用户的要求.在更新数据的自动化感知方面,缺乏有效的办法能够自动识别知识的变化,和能够动态响应这些变化的更新机制.

2 基于开放网络知识库的信息检索

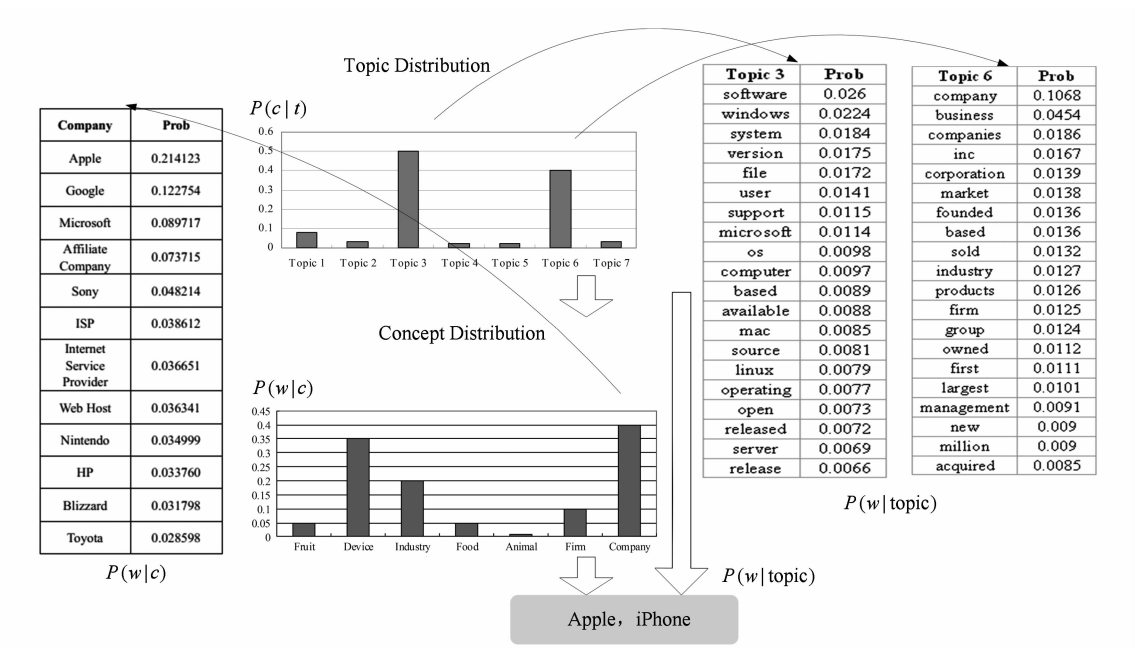
面向网络大数据海量、自由、免费、内容开放的特点,基于本体的开放网络知识库成为一种有效的知识结构化表示工具,进而为解决自然语言理解和Web信息检索中的问题提供了新的资源和方法.现有的大多数信息检索技术都是以单个词为处理对象[42].从用户角度来讲,仅仅使用几个关键词就把检索要求描述清楚是一件很困难的事情.检索需求的模糊不清必然会导致检索结果的混乱;从关键词检索技术本身来讲,简单匹配方法常常会检索出很多非相关文档.在文档表示技术方面,目前的方法主要是基于向量空间模型(vector space model)理论[43-44],以单个独立词为处理对象,并且假定词与词之间是相互独立的,然而在实际文本中词与词之间是相互关连的,这也是Web搜索技术发展的瓶颈所在.基于本体的知识库系统的引入作为一种网络开放文本的表示方法,为信息检索优化的各个研究领域提供了结构化数据基础.

目前,信息检索服务已经逐渐从基于链接的信息检索引擎发展为基于知识的对象搜索引擎.对象(object)定义为实体(entity),对象之间的关联不仅仅是关键词或网页链接的关联,在基于本体的知识库系统中,概念、实例、属性等实体表征包含了更广泛的关系,既包含上下位等结构关系,又包括语义相似度等内容关系.利用知识库进行信息检索搜索引擎优化主要解决如下2个问题:一是利用若干关键词描述的检索要求不明确,进而造成了检索结果的混乱;二是如何利用上下文语境信息对查询词和文档进行更好的语义表示.本节结合以Probase和knowledge graph为代表的开放网络知识库系统,介绍信息检索优化的研究内容,包括意图感知,查询扩展和语义问答等.

2.1 意图感知

用户搜索意图可以理解为用户通过搜索希望获取到的信息,可以量化为用户希望得到的检索结果集.通过关键词搜索的方式,用户输入的每个词背后都隐藏着更深层次的查询意图,而这些查询意图往往需要深入挖掘才能够获得,需要对用户检索意图进行建模,在搜索日志分析的基础上进行包括用户意图分类模型构建和用户意图分类特征研究等.Andrei[45]将搜索者的意图分为:为了查找具体的某个网站地址的导航型(navigational)、为了获取某种信息,学到新知识的信息型(informational)和为了完成一个目标明确的任务的事务型(transactional)3个类别.现有研究表明,用户的意图具有隐性的特点,在其建模过程中通常涉及逻辑推理问题,引入本体思想,通过对各领域知识的深层次理解和规则推断,可以为意图感知提供新的方法保障.如今大规模知识库系统的出现,尤其是在知识库构建过程中融入了用户查询日志的相关语义内容,使得基于本体的用户意图感知大规模分析与应用成为可能.Cambria[46]提出让机器真正理解用户的意图,需要提供的不仅是关键词组合及其共现频率等,更需要提供知识层面的信息.融合知识库系统Probase和基于自然语言的语义网ConceptNet,使用基于奇异值分解(SVD)的多维度约简技术来挖掘和分析用户意图.Wen等人[47]基于Probase知识库推断查询的概念化含义,对应检索任务自动给出话题级别的查询标识.如图3所示,查询中的一个词项(term)可以同时被归类为一个属性(attribute)或一个实体(entity),通过基于贝叶斯的方法计算概率$P(instance|concept)$,即在给定的概念(concept)中某个实例(instance)的典型和流行程度,并以此作为话

题级别的量化判定依据.在计算上述概率的过程中融合了多种特征表示,包括概念特征(conceptual features),即概念化聚类生成的概念相似度;词汇特征(lexical features),即 N-char Jaccard 相似度;样本特征(template features),即从实际用户的查询记录中根据编辑距离生成查询关键词之间的相似度;时间特征(temporal features),即根据时间间隔提取用户在同一个查询任务内进行连贯查询修正对应的相似度.最终给出量化的概念分布和话题分布,并以此作为用户意图的判定依据.进一步,Guo 等人[48]提出了基于意图感知的特征表达与度量机制,用以建模查询意图多样性和不确定性,实现精准的意图理解,形成了查询意图感知的查询相似度量(intent-aware query similarity)的概念:首先结合搜索关键词和点击记录,利用话题模型自动学习潜在查询意图;其次,根据学习训练的结果,对应每个意图提取查询关键词表示;最后,对应不同的查询意图使用不同的度量方法,包括两两度量(pair-wise)方法如余弦相似度(cosine similarity)、基于图(graph-based)的度量如谱嵌入(spectral embedding)等.



Company	Prob
Apple	0.214123
Google	0.122754
Microsoft	0.089717
Affiliate Company	0.073715
Sony	0.048214
ISP	0.038612
Internet Service Provider	0.036651
Web Host	0.036341
Nintendo	0.034999
HP	0.033760
Blizzard	0.031798
Toyota	0.028598

Topic 3	Prob
software	0.026
windows	0.0224
system	0.0184
version	0.0175
file	0.0172
user	0.0141
support	0.0115
microsoft	0.0114
os	0.0098
computer	0.0097
based	0.0089
available	0.0088
mac	0.0085
source	0.0081
linux	0.0079
operating	0.0077
open	0.0073
released	0.0072
server	0.0069
release	0.0066

Topic 6	Prob
company	0.1068
business	0.0454
companies	0.0186
inc	0.0167
corporation	0.0139
market	0.0138
founded	0.0136
based	0.0136
sold	0.0132
industry	0.0127
products	0.0126
firm	0.0125
group	0.0124
owned	0.0112
first	0.0111
largest	0.0101
management	0.0091
new	0.009
million	0.009
acquired	0.0085

Fig. 3 Query intent awareness based on knowledge base.

图 3 基于知识库的查询意图感知

2.2 查询扩展

Web 搜索优化的另一个重要方向是查询扩展,即采用有效措施,在用户输入查询关键词基础上添加相关词,为判断检索文档相关性提供更多的信息.在基于本体的知识库系统中,语义扩展查询将原始查询,尤其是短文本查询概念化(conceptualization),提取更高精准度的查询语义,通过全局和局部的关联规则和用户查询日志扩展等方法在得到语义层面的扩展概念和实例.进而通过迭代修正查询关键词得到更好的检索效果.Song 等人[37]提出了基于概率化知识库 Probase 的短文本查询概念化问题,在概念规模达到人类知识级别的基础上,采用贝叶斯推断机制对关键词或短文本进行概念化,如图 4 所示,首先从查询中提取实例(instance),再结合 Probase 知识库,为每个查询实例生成对应的概念向量,进一步,在概念空间进行相似度检索,找到匹配的相关概念向量,即与原始查询语义相近的所有概念.与传统的潜在语义话题模型和基于 WordNet,Wikipedia,Freebase 单纯统计的方法相比,Song 的算法在查询理解方面有更好的效果.Kim[49]从上下文关联角度分析了查询概念化扩展的问题,结合概率化话题模型 LDA 提出了在词句级别的概念化算法,在单词级别(word-level)根据相似度对语义相关性进行判断,进而推测出相关的隐性的实例,在句子级别(sentence-level)使用多种线性加权组合来匹配查询关键词集合,并使用了大量现实查询点击记录来辅助参数学习.结果表明,在 Probase 知识库的语料规模上结合话题模型可以得到优化的查询扩展效果,并且仍然有进一步提升的空间.可见,在语义概念层面解决查询扩展问题,很大程度上取决于作为扩展依据的知识库体量规模和关系结构.基于本体的知识库系统为各种文本挖掘技术,尤其是话题模型、文本聚类等提供了新的计算空间和优化方法,并推动了一系列新算法的研究.

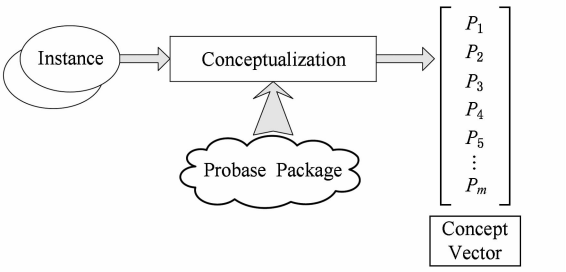


Fig. 4 Conceptual query extension based on knowledge base.

图 4 基于知识库的概念化查询扩展

2.3 语义问答

信息检索的一个目标是会帮助用户做事,而不是简单地给出建议;回答用户的问题,而不是简单地给出链接.其对应的研究领域为语义问答(question answering).语义问答是近 10 年来热门的研究方向,主要采用信息检索和自然语言处理相关技术,构建针对人类语言描述问题的自动应答系统.问答系统需要一个结构化知识的数据库作为信息支撑,通常采用知识库的方式[50-51].文献[52]给出了语义问答系统的技术路线图,提出了问题语义分类、问题处理、上下文关联、数据源选择、答案提取和形式化等一系列相关解决方法.其中对问答系统的语义分类体系(如表 1 所示)从自然语言处理的角度体现了类别与层次的详细划分,例如:"昨天是否下过雨?"属于确认型;"世界上人口最多的国家是?"属于概念型;"如何烤面包?"属于条件型;"我不知道到机场的路"属于声明型等.

Table 1 Taxonomy of Questions

表 1 问题的语义分类体系

Question	Abstract Specification
Verification	Is a fact true? Did an event occur?
Comparison	How is X similar to Y? How is X different from Y?
Disjunctive	Is X or Y the case? Is X, Y, or Z the case?
Concept completion	Who? What? When? Where? What is the referent of a noun argument slot?
Definition	What does X mean? What is the superordinate category and some properties of X?
Example	What is an example of X? What is a particular instance of the category?
Interpretation	How is a particular event interpreted of summarized?
Feature specification	What qualitative attributes does entity X have? What is the value of a qualitative variable?
Quantification	What is the value of a quantitative variable? How much? How many?
Causal antecedent	What caused some event to occur? What state or event causally led to an event or state?
Causal consequence	What are the consequences of an event or state? What causally unfolds from an event or state?
Goal orientation	What are the motives behind an agent's actions? What goals inspired an agent to perform an action?
Enablement	What object or resource anables an agent to perform an action?
Instrumental/Procedural	How does an agent accomplish a goal? What instrument or body part is used when an agent performs an action? What plan of action accomplishes an agent's goal?
Expectational	Why did some expected event not occur?
Judgmental	The questioner wants the answerer to judge an idea or to give advice on what to do.
Assertion	The speaker expresses that he or she is missing some information.
Request/Directive	The speaker directly requests that the listener supply some information.

在问题语义分类的基础上,进一步结合知识库系统进行信息融合提升问答的准确率,通过查询空间的降维提升在线处理速度.现阶段研究成果主要集中于特定形式的问题、特定领域内的问题等,如回答时间地点等问题、回答术语定义等问题、依据人物传记回答问题以及类型多样化的问题,如同一个问题多种语言表述,关于音频、图像、视频内容的问题等[53].另外,从单一的知识源或少数算法出发,很难让语义问答系统达到接近人类的水平,IBM 的沃森系统(Watson)[54]正是通过搜索与融合 DBPedia, WordNet, Wikipedia 和 Yago 等多种知识库,从多角度运用大量小算法,对各种可能的答案进行综合判断和学习,才具有性能的鲁棒性.因此,下阶段的研究将关注于融合多种不同的知识体系,针对共识的事实知识和特定领域的专业知识,建立统一的知识推理机制.

3 基于开放网络知识库的数据挖掘

基于开放网络知识库的数据挖掘包括 3 方面内容,即线索挖掘、关系推理和关系预测.

3.1 线索挖掘

利用开放网络知识库进行线索挖掘,是指利用知识库对 2 个原本无关联的实体或概念作关联,给出 2 个实体间的关联路径或关联的模式,或称为 storytelling[55-56].例如,如果要回答奥巴马和比尔盖茨之间的关系路径是什么,还有哪些实体和他们有关,这些相关性的证据是什么,都要通过实体关联来解决.相关的方法都是将知识库当作图来处理,图中每个点代表一个实体或概念,边代表实体间关系、概念间关系以及实体和概念间的关系.然后将关联任务转化为基于图的各种操作,例如寻找导出子图、特定的子结构、连通分支等.这里,我们主要介绍 2 个有代表性的方法,由 Hossain 等人[56]提出的基于团(Clique)的关联方法.

该关联方法在两节点间构造一条路径,路径上相邻两点以及一些其他点构成事先设定的大小的团结构.如图 5 所示,在关联实体 J. Escalante 和 A. Sufaat 时,构造的路径用粗线条标示出来,这条路径上的相邻两个点连同其他外部节点(如 Feb 22,2003, Arze, Cuban Intell, Escalante)构成大小为 6 的团.生成这样的团补充了很多外部信息,可以更好地解释 2 个实体间的关联,增加更多的线索.基于团的实体关联主要通过 3 个步骤来实现,首先利用 CHARM-L 算法[57]生成概念格(concept lattice),然后利用 A* 算法产生路径上的候选节点并评估这些候选节点以确定最终的路径.

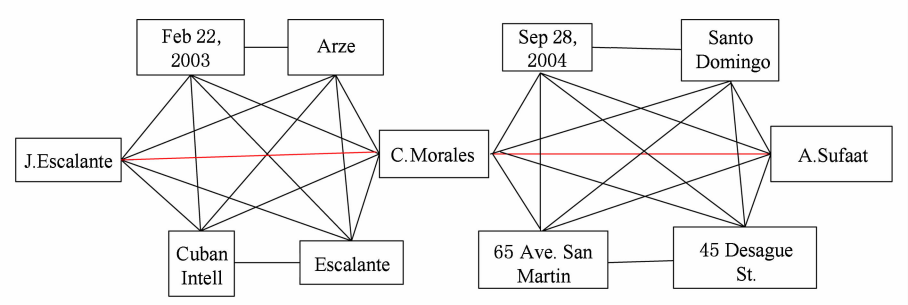


Fig. 5 Clique-based entity linking.

图 5 基于团的实体关联示意

相比于基于团的实体关联方法得到的只是一条路径的方法,Fang 等人[58]提出了基于关系解释(relation explanations, REX)的知识库中实体关联的方法,REX 可以返回关联 2 个实体的比树和路径更为复杂的连通结构,对关联给出更详细的解释.该系统主要包括解释枚举(explanation enumeration)和解释排序(explanation ranking)2 部分.在解释枚举部分,系统从知识库中海量的实体中通过定义覆盖路径模式集(covering path pattern set)这一典型的结构来快速识别出能够关联两个实体的候选实体,在解释排序环节,系统定义了基于结构,基于分布和聚集性等多重度量指标用以衡量解释枚举环节得到的候选实体的相关性(interestingness),并按这些指标对候选实体进行排序.

对大规模知识库中的实体作线索关联面临很多的挑战.由于知识库中存储的知识规模越来越大,对关联的效率、实时性和准确性的要求越来越高.由于知识库的构建并不完美,知识库中的数据存在噪声、错误、不一致等情况,如何开发设计出扩展性强、抗噪声能力高的线索挖掘方法是下一步研究的方向.

3.2 关系推理

基于开放知识库的关系推理,是指利用知识库

现有的实体间关系推断或推理实体间潜在的、隐含的关系. 例如,推理人物间的爷孙关系,前提是 A 是 B 的父亲,B 是 C 的父亲,则通过关系推理功能可知 A 是 C 的爷爷. 利用知识库进行实体间关系推理,目前流行的方法,主要有基于规则的方法即利用机器学习领域中的归纳逻辑编程(inductive logic programming)技术的方法,例如基于一阶 Horn 子句或一阶归纳逻辑(FOIL)[59-61]以及基于概率图的方法,如随机游走[62]、Markov 逻辑网络[63-64]等. 前者典型的工作有 NELL[60],后者包括 Schoenmackers 等人[63-64]提出的 Sherlock-Holmes 方法.

NELL 系统采用一阶 Horn 子句的方法推理实体关系. 例如,它会使用如下的 Horn 子句:

$$AthletePlaysForTeam(x,y) \wedge TeamPlayInLeague(y,z) \Rightarrow AthletePlaysInLeague(x,z)$$

来推断关系 *AthetePlaysInLeague*(HinesWard, NFL). 如果子句里 x, y 和 z 分别取值为 HinesWard, PittsburghSteelers 和 NFL, NELL 系统学习 Horn 子句规则的方法是 FOIL 算法的变形即 N-FOIL 算法. N-FOIL 将大量正例和反例作为训练数据输入到系统中,例如 *AthletePlaysInLeague*(HinesWard, NFL)是正例,而 *AthletePlaysInLeague*(HinesWard, NBA)是反例,然后利用分而治之(divide-and-conquer)策略学习满足训练数据的 Horn 子句. 每条 Horn 子句都是通过从对一般规则的特殊化来实现的,使得规则始终覆盖大部分正例和极少的反例. 当一个子句通过学习得到,则被该子句覆盖的例子从训练数据中移除,然后迭代进行下一个子句的学习,直到训练数据中没有正例. 之所以采取 N-FOIL 算法,是因为用 FOIL 算法学习一阶 Horn 子句的计算代价很高,不仅是搜索空间很大,而且有些 Horn 子句的得出代价很高. N-FOIL 算法通过预处理等技巧来缩小搜索空间,增强 FOIL 算法的扩展性. 尽管如此,NELL 系统仍然需要人工编写和指定初始的规则,而且对学习到的规则缺乏有效的评估机制以应对知识库中数据的噪音问题. 所以需要一个对规则的评分函数(scoring function)来应对噪音问题,由 Schoenmackers 等人[63]提出的 Sherlock-Holmes 方法提出了这样一个评分函数.

Sherlock-Holmes 是基于开放领域文本的关系推理方法,该方法分为线上部分 Sherlock 和线下部分 Holmes. Sherlock 学习推理规则,将规则输入到 Holmes 推理引擎中,利用这些规则对推理查询作出回答. 如图 6 所示. 具体地讲,Sherlock 输入已有的实体关系或事实,利用 Hearst pattern(如 such as 模式)识别出具有代表性的实体的类别和这些类别的实例,然后发现类别间关系,通过穷举搜索如 MCMC(Markov chain monte carlo)抽样方法推出推理规则,并通过统计相关性检验的方法计算每条推理规则的置信度,输出通过评分函数计算的带权重的 Horn 子句逻辑规则. Holmes 将 Horn 子句逻辑规则、知识库中的实体间关系以及使用规则表示的查询作为输入,对满足对查询条件的实体间关系构建证据树,把证据树转化为 Markov 逻辑网络,使用 Markov 逻辑网络推理相应节点的概率,输出满足条件的隐含关系及其置信度. 这里的 Markov 逻辑网络推理是指在 Markov 网络中加入边从而产生一系列的团结构(Cliques). 根据 Markov 网络的理论,每个这样生成的团 k 都具有一个势函数 φ_k,返回基于团中变量状态的非负值. 记状态 x 的概率为 $P(x)$,则 $P(x)$满足:

$$P(x) = \frac{1}{Z}\prod \varphi_k(x_{\{k\}}),$$

其中,Z 是正规化因子,$x_{\{k\}}$ 是团 k 中所有变量的状态. 对于原先证据树中的叶子节点,其概率可以通过知识库计算出来,对于需要推断的节点概率事先赋予一个指数先验. 最后利用 loopy belief propagation 方法进行近似推理.

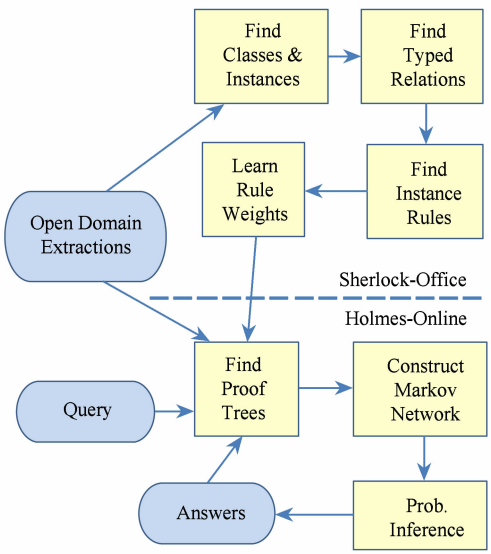


Fig. 6 Illustration of Sherlock-Holmes.

图 6 Sherlock-Holmes 方法示意

Holmes 推理新关系采用的 Markov 逻辑网络方法需要事先由人来编写少量推理规则. 而 Lao 等人[65-66]提出了基于路径排序算法(path ranking

algorithm,PRA)的知识库关系推理方法[61],可以从训练数据中自动发现推理规则,而且采用了有效的近似推理方法,从而比 Markov 逻辑网络的方法具有更高的效率. PRA 方法相对于每个查询点 x,给关系图中的其他节点 y 作排序. PRA 方法先枚举生成所有的关系路径,将每条这样的路径当作训练的专家(ranking experts),产生一条随机游走. 当达到稳定状态时,给每个点 y 赋予一个概率值,根据这些概率值对节点 y 进行排序. 最后 PRA 通过 logistic 回归对这些专家进行加权统计. 具体而言,对于知识库中的每个关系 R 给定概念 x,我们试图寻找其他所有的概念 y,使得 x 和 y 满足关系 $R(x,y)$. 定义关系路径 P 为关系 $R_1,R_2,\cdots,R_l$ 的序列. 为了强调每一步关系的类型,P 还可以写成 $T_0 \xrightarrow{R_1} \cdots \xrightarrow{R_l} T_l$,其中 $T_i=range(R_i)=domain(R_{i+1})$. 记 $domain(P)=T_0$,$range(R)=T_l$. 我们对每条关系路径 $P=R_1,\cdots,R_l$ 和种子点 x、路径型(path constrained)随机游走给每个点 y 都定义了一个分布 $h_{x,P}(y)$. 当 P 是一条空路径时,如果 $x=y$,则 $h_{x,p}(y)=1$,否则 $h_{x,p}(y)=0$. 当 P 不为空时,假定 $P'=R_1,\cdots,R_{l-1}$,则 $h_{x,P(y)}$ 迭代定义如下:

$$h_{x,P}(y)=\sum_{y'\in range(P')} h_{x,P'}(y')\cdot P(y\mid y',R_l),$$

其中,$P(y|y',R_l)=\dfrac{R_l(y',y)}{|R_l(y',y)|}$ 是从节点 y' 到节点 y 的所有长度为 1 的随机路径的类型是 R_l 的边的概率,标示了是否存在一条类型为 R 的从 y' 到 y 的边. 更一般地,给定关系路径集合 $P_1,\cdots,P_n$,可将每个 $h_{x,P}(y)$ 进行如下线性加权:

$$\theta_1 h_{x,P_1}(y)+\theta_2 h_{x,P_2}(y)+\cdots+\theta_n h_{x,P_n}(y),$$

得到节点 y 的相对于节点 x 的打分 $score(y,x)$,其中 θ_i 为路径 P_i 的权重,$score(y,x)$的定义如下:

$$score(y,x)=\sum_{p\in P} h_{x,P}(y)\theta_p,$$

其中 P_l 是长度为 l 的关系路径集合. Lao 等人[61]提出的基于知识库的关系推理方法对 PRA 方法作了 2 方面的改进,以应对知识库中大量的关系和提高推理的准确性.

对知识库中的实体关系进行推断是近几年研究的热点,也是难点问题,如何应对在超大规模知识库中对关系进行精确的推理,以及对推理结果的可信度进行有效判定,是下一步工作需要解决的问题之一.

3.3 关系预测

对大规模知识库的实体关系作预测,试图对实体间关系的时序变化作出定量和定性的预测,包括预测是否会产生关系、关系的类型变化、关系的置信度变化、关系发生的频次等. 和关系推断不同,对关系的预测是对未来实体间可能发生的关系进行判定. 这方面有代表性的工作都是基于机器学习等方法实现的,包括有监督的学习方法和无监督的学习方法. 例如依托社交网站构建的实体关系网络,存在预测 2 个用户在未来是否会有链接关系存在即 Link Prediction 问题. 针对该应用场景下的关系预测问题,有监督的学习算法是将预测问题视为分类问题. 比较有代表性的工作包括[67-68,76]. 尽管有监督的学习算法是目前比较流行的算法,但是它在分类特征选择方面经常会面临不平衡性(imbalance),必须依赖于实验数据集的先验分布之知识,因而预测结果准确率不高. 为了解决这一问题,有学者提出了基于非监督的学习算法的关系预测方法. 这类方法试图在 2 个用户之间定义一个合理的相似度指标,越相似的用户越有可能产生关系. 非监督的方法不用考虑关系预测的数据集的先验分布知识,因而准确度较高. 经典的相似度指标包括 Adamic-Ada 指标[69]、优先连接或择优连接(preferential attachment)指标[70]、Katz 指标[71]、结构相似度指标[72]等指标,以及行为相似度的指标[73]等. 基于上述几个指标的非监督关系预测方法在真实数据集上的效果比较已有大量的研究工作,例如 Jia 等人[73]使用在线微博数据发现基于行为相似度和结构相似度的关系预测方法在 F1-值等指标上相比上述其他几个非监督关系预测方法取得了更好的效果. 其他相关比较工作可参照文献[72,74]等.

对知识库中实体间关系进行预测可以在一定程度上推动对知识库的更新工作. 反过来知识库的更新又可以验证关系预测的准确性. 所以将知识库的更新和关系预测进行更好的结合是未来研究的方向.

4 系统应用

开放网络知识库的发展为下一代智能搜索和深入信息挖掘与分析提供了重要的基础. 目前已经从数据积累阶段逐步向产品产出阶段转化. 知识图谱和实体搜索已经成为 Web 搜索的标志性技术,本节从智能搜索引擎和商业情报分析系统 2 个方面来介绍现有系统应用,并对未来可能产生的新型应用模式进行了分析和展望.

4.1 智能搜索引擎

近年来,国内外的大型搜索引擎相继推出了基于知识库的搜索新产品,主要支持人物与知识的关系搜索.

4.1.1 人物关系搜索

在通用搜索引擎中,用户输入查询关键字后,返回的是包含该关键字的所有网页结果.而人物关系搜索除了提供网页结果之外,还能够将这些网页之中包含的人名、地址、机构等信息提取出来,并将所有与关键字相关的这些信息按照网络流行度或关系亲密度进行排序.这种信息过滤及聚合方式为信息浏览提供了很大的便利.目前包括微软人立方、雅虎人物搜索以及 Facebook 的社交图谱都提供人物关系搜索功能.

1) 微软人立方

人立方[①]关系搜索是微软亚洲研究院开发设计的一款新型社会化搜索引擎,它能从超过 10 亿的中文网页中自动地抽取出人名、地名、机构名以及中文短语,并通过算法自动地计算出它们之间存在关系的可能性.此外,人立方关系搜索还自动地找出人名之间最可能的关系描述词、与人名最可能相关的称呼、作品等词条等.人立方搜索的创建理念来自于"6 度空间",只要随便输入一个人物,人立方搜索将给出该人物的关系、网页、资讯、简介等众多内容.目前,人立方已经推出了关系百科、读心游戏、微博关系图以及 6 度搜索等特色人物关系搜索功能.

2) 雅虎人物搜索[②]

中国雅虎在对全能搜索引擎的升级中推出了人物搜索.在人物搜索功能中,用户不但可以得到被搜人的全部资料,连该人的人际关系都清晰地显示在一张图片中.通过"人际网络"搜索能够通过搜索分析描述出人与人之间的关系.比如搜索李开复,能看到其家庭、工作等各方面的关系,李开复女儿在其中也有显现.该系统能够提供人物检索功能和以人物为中心的社会关系网络展示,并且人物之间社会关系的可信度较高,然而仍然未对网页信息进行过滤和归纳,而且呈现的人物之间社会关系的信息比较少.

3) Facebook 社交图谱搜索[③]

社交网站如 Facebook 也通过构建知识库系统:社交图谱结合自身海量社交数据的特色和优势,推出个性化人物社交关系发现等站内搜索服务.社交图谱搜索是帮助用户查找相关人物、照片、场所和兴趣的新方式.通过一系列过滤器,例如"场所类型"、"被某人赞过"或"好友曾去过",用户可以使用结构化搜索工具,在自己的网络中进行查找,并发现新的社交关系.社交图谱搜索利用全球 10 亿用户社区的力量,回答用户关于人物、照片、场所和兴趣的问题,是 Facebook 吸引用户互动的一个关键,是社交网络未来"智能、原创、基本的拼图".

4.1.2 知识关系搜索

知识关系搜索引擎是知识管理的一种实现理念与工具,承担了"知识汇聚、知识发现、知识分类、知识聚类、知识门户的构建",最终通过搜索引擎技术完成知识管理的使命.依托知识库的构建,目前包括 Google、百度、搜狗、中国科学院等搜索引擎公司或研究机构相继推出了知识关系搜索功能.

1) Google 知识图谱

为了让用户能够更快更简单地发现新的信息和知识,Google 搜索发布了知识图谱[④],可以将搜索结果进行知识系统化,任何一个关键词都能获得完整的知识体系.知识图谱会从 Freebase、维基百科或全球概览中获得专业的信息,同时还通过大规模的信息搜索分析来提高搜索结果的深度和广度.和之前的搜素结果相比,"知识图谱"将在 3 个方面大幅度提高 Google 搜索的最终效果:①找到正确的结果.由于一个关键词可能代表多重含义,所以知识图谱会将最全面的信息展现出来,让用户找到自己最想要的那种含义.②最好的总结.有了知识图谱,Google 可以更好地的理解用户搜索的信息,并总结处相关的内容和主题.当你搜"玛丽·居里"时,你不仅可以获得这个关键词的所有相关内容,还能获得居里夫人的详细生平介绍.③更深、更广.由于"知识图谱"会给出搜索结果的完整知识体系,所以用户往往会发现很多不知道的东西(知识).当你搜索一个即将去旅行的地方时,可能你会发现一个以此命名的餐馆,甚至还可能发现还有一本小说就叫这个名字,并且已经改编成了同名电影.

① http://renlifang.msra.cn

② http://www.yahoo.cn/

③ http://www.facebook.com/

④ http://www.google.com.hk

2) 百度实体搜索

百度实体搜索[①]的技术方向和 Google 不同:百度主要侧重“深度”,而 Google 则强调“广度”.像“不掉毛”“濒临灭绝”这些细致的属性,传统的实体信息提取技术是无法涵盖的.但这种深入的属性数据挖掘一方面得益于大量的网页数据,另一方面也受限于互联网数据里大量的噪音影响,是一个技术难度和收益都比较大的方向.从目前的结果来看,这些深入挖掘出的属性数据在质量方面的表现是不错的,比如“不掉毛的狗”的结果中,除了雪纳瑞等较为常见的不易掉毛的狗以外,甚至可以覆盖到“冠毛犬”.假如在保证数据的质量的前提下覆盖领域可以不断扩大,对于知识类的问题,机器可以像 IBM Watson 一样直接给出超出人类平均水平的解答.

3) 搜狗知立方

搜狗致力于建设中文的大规模知识库[②]:通过知识库构建、信息抽取、本体映射、查询语法分析、实体消歧等技术的研究及运用,切实推进大规模领域丰富的通用知识库的建立,并成功运用在搜索引擎领域.成功研发并发布了知立方产品,基于对互联网数据的整理及知识化,使搜索引擎拥有推理计算的能力,帮助用户更快地获取信息.如搜索“冯小刚电影”,结果最上方就会直接展示所有冯小刚导演的影片,接下来是冯小刚的最新新闻及电影在线观看地址等搜索结果,页面右侧则展现冯小刚个人简介,所拍电视剧及著作等,以知识图谱的形式来理解用户的搜索意图.

4) 开放知识网络

中国科学院计算技术研究所提出了一种面向网络大数据的开放的、自适应的、可演化的知识计算引擎——OpenKN[76].通过知识库构建、知识验证与计算、知识存储、知识服务与应用,实现了一个全生命周期的知识处理过程.通过自适应的知识演化过程和自适应的知识获取策略,OpenKN 具备了捕获新数据的能力,而不同的获取策略又保证了知识的实时更新.OpenKN 的可演化性是基于一个称之为可演化知识网络的知识表示模型来实现的.该模型描述了一个异质网络,其点和边都带有时空信息,并被赋予了一系列函数或算子,这使得 OpenKN 能够对知识演化进行识别、定位和评估.OpenKN 的这些特点使用户在搜索知识间关系的同时,能够进一步查看关系产生、成立甚至消亡的时间和空间信息.例如查看人物与学校间关系时可以获得人物的毕业或入学时间以及人物的入学地点等,查看人物间配偶关系时可以获得人物间婚姻关系的持续时间等.

4.2 商业情报分析系统

近年来,通过构建以知识图谱为代表的大规模领域知识库,结合传统网络数据统计方法,国内外的搜索引擎也陆续推出了商业情报分析相关的产品、支持流行趋势、排名关键词等分析,以及在线自动语义问答等功能.

4.2.1 流行趋势分析系统

Google Trends[③]、百度指数[④]等,反映关键词在过去 30 d 内的网络曝光率及用户关注度,统计每天的变化趋势,发现、共享和挖掘互联网上最有价值的信息和资讯,直接、客观地反映社会热点、网民的兴趣和需求.企业通过 Google Trends、百度指数等可以及时准确地了解其所处行业中的流行趋势分析,发现新的商机,挖掘企业新的增长点.

4.2.2 排名关键词分析系统

针对行业业务竞争情报需求,Google 和百度推出了排名关键词工具,微软在知识库 Probase 的基础上进行关键词的概念化和智能排序,为用户提供主要核心关键词和概念的竞争对手排名状况,了解竞争程度,辅助商业决策.如百度搜索风云榜[⑤]提供了更多行业榜单,并从地域、人群等多个视角对实时热搜数据进行多维度展示.可见,在情报收集、系统扩展与集成、发现隐含情报信息等方面,知识库的引入拓展了结构化数据空间,提升了数据的使用价值.

4.2.3 统计分析系统

以 Google Analytics 和百度统计为代表的统计分析系统,将搜索引擎获取的数据进行深加工,呈现细节,通过微观信息揣摩用户行为和心理,进而为商业决策提供知识资源.统计分析系统主要功能包括事件目标分析、历史趋势分析、行业环境分析等.以百度统计为例,通过整合百度指数,搜索风云榜等不同功能产品的相关内容,在流量原因分析中给出更加优化的分析结果.相比之下,Google Analytics 在

① http://www.baidu.com

② http://www.sogou.com

③ http://www.google.com/trends/

④ http://index.baidu.com/

⑤ http://top.baidu.com

交叉数据分析、个性化过滤功能、转化路径分析等方面更加完善,功能更强.

网络大数据时代为知识库的系统应用提供了广阔的空间和可能.用户可以通过开放网络知识库及时获取海量网络数据中蕴含的丰富知识.同时,这种获取也要更大程度地满足不同人的不同的需求.因此,基于开放网络知识库的个性化的、可定制的、融入更多用户参与的系统应用,将会使开放网络知识库真正变成每个人身边不可获取的私人数据分析和处理专家,能够对数据进行深度分析、精准研判和及时的预测.

5 研究展望

基于开放网络大数据的知识库为人们深入利用网络大数据的价值提供有效的途径.目前,虽然在国内外已经有了一些以开放网络数据为基础的知识库,并兴起了一些新兴的应用,但无论知识库的构建、更新,还是应用都还不能完美地满足人们的应用需求,也就意味着每个方向都有极具挑战性的工作.展望未来,面对网络大数据的开放网络知识库的建设和应用的研究,还处于起步阶段,仍有大量问题需要研究和解决.

1) 开放网络知识库的创建和更新中融入群体智慧.由于开放网络知识库的来源数据具有冗余、噪音、不一致等特点,开放网络知识库的构建在数据预处理阶段面临很大的挑战.这就需要发挥群体智慧(collective intelligence),对数据进行预处理或者直接进行知识生产.例如维基百科等在线百科知识,就在很大程度上调动了每个个体的积极性,通过个体对在线页面的编辑,将无结构化数据直接转化为半结构化和结构化数据.对知识库的更新也是如此,对海量知识的更新可以通过每个个体将整个知识分割为若干部分,然后每个个体负责相关部分的更新工作,最后将每个部分的更新结果合成一个完整的更新结果,即采用众包的方式来完成.目前对知识知识库的构建和更新工作,在群体智慧的利用方面仍然有很大的空间.

2) 开放网络知识库的实时感知与自动更新.对开放网络知识库的实时感知是指开放网络知识库能够实时感知数据源的变化,包括数据的规模的增长,数据内容的变化等.在实时感知的前提下,对知识库内的知识作出自动的更新.这就需要知识库在构建时能够充分融入知识的时间信息和空间信息,实时跟踪这些时空信息的变化.不仅如此,结合知识自身的属性信息、关系信息,以及知识变化的一般规律,知识库还应具备对知识变化的判断能力,如建立一套知识自动更新的公理或命题逻辑,刻画知识变化的规律.这样知识库就能完全或部分摆脱手动更新的手段,真正实现智能化.

3) 通用知识库与领域知识库相结合,实现有效跨库映射.领域知识库是指知识库建模的是某个特定领域,或现实世界的一部分.领域知识表达的是适合于该领域的那些术语的特殊含义,可以用来构建针对特定任务的专业化知识库.通用知识库是指由若干个领域知识中普遍使用的共同对象构成的模型,收录核心词表,可以用来描述一系列领域中的对象.领域知识库可以通过对库中实体的概念化的过程而得到通用知识库,反过来,通用知识库可以通过对库中概念的实例化来得到领域知识库.构建通用知识库和领域知识库的一个好处在于,对具有相同通用知识库的多个领域知识库可以进行知识库的映射工作,如映射具有相同概念的实体,实现跨库间实体的关联和映射.

4) 实现知识库的跨语言融合.多语言知识库是目前研究的热点问题.由于单语种的数据源经常存在知识表达不完整的情形,融入多语种可以更好地弥补单语种的缺陷,同时可以充分利用多语种在表达知识方式上的互补性.这就需要在多个语种间实现知识的融合,即构建多语种知识间的映射关系.对多个语种中同时出现的知识进行匹配和关联,对仅在某些语种中出现的知识映射到其他语种.

5) 通过计算实现对潜在知识的推断和未来趋势的预测,对知识库中潜在的知识的推断和未来知识的预测是知识库智能化的体现,这既有利于知识库的自我更新,而且也有利于支持对知识库中实体的线索挖掘.通过对知识计算的内在规律和机理的研究,可以更好地理解知识间的关系、知识的形成和存在方式,以及知识变化和发展的内在规律,从而从根本上实现知识的推理和预测,支持建立在推断和预测基础上的应用.

6 总 结

网络大数据是指“人、机、物”三元世界在网络空间中交互、融合所产生并在互联网上可获得的大数据.这些数据具有多源异构、交互性、时效性、社会性、突发性和高噪声等特点,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、复杂关联的知识网络.有效利用网络大数据价值就是数据的去冗分类、去粗取精,从数据中挖掘知识,对大数据网络后面的知识进行深入分析.本文通过对当前国内外知名的开放网络知识库,以及其支持的应用进行了分析和论述.并从开放网络知识库的构建、以及基于开放网络知识库的对信息检索与数据挖掘方面的应用的方法和技术的现状进行了综述,并从人物关系分析系统、知识关系分析系统,以及商业情报分析系统等实际应用系统进行了介绍.最后,展望了开放网络库及其应用的未来主要研究方向.总之,网络大数据为知识库提供了更为丰富的知识来源,使得人们有机会通过开放网络知识,获得更为及时有效地应用.它与传统研究工作相比,在各个层面的差异都非常显著.尽管目前已经有一些探索性的研究工作,但是总体上来说,基于开放网络知识库的研究还不够深入,诸多问题尚待研究.

参 考 文 献

[1] Wang Yuanzhuo, Jin Xiaolong, Cheng Xueqi. Network big data: Present and future [J]. Chinese Journal of Computers, 2013, 36(6): 1125-1138 (in Chinese)
(王元卓,靳小龙,程学旗. 网络大数据:现状与挑战[J]. 计算机学报, 2013, 36(6): 1125-1138)

[2] Li Guojie, Cheng Xueqi. Research status and scientific thinking of big data [J]. China Academic Journal, 2012, 27(6): 647-657 (in Chinese)
(李国杰,程学旗. 大数据研究:未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J]. 中国科学院院刊, 2012, 27(6): 647-657)

[3] Lee T B. Semantic Web architecture [EB/OL]. 2000 [2013-07-25]. http://www.w3.org/2000/talks/1206-xml2k-tbl.2000-11-8

[4] Aditya P, Anand R, Hector G-M, Towards the Web of concepts: Extracting concepts from large datasets [C] //Proc of the 36th Int Conf on Very Large Data Bases VLDB'10. San Francisco, CA: Morgan Kaufmann, 2010: 566-577

[5] Wan Changlin, Shi Zhongzhi, Hu Hong, et al. Qos-aware semantic Web service modeling and discovery [J]. Journal of Computer Research and Development, 2011, 48(6): 1059-1066 (in Chinese)
(万长林,史忠植,胡宏,等. 基于本体的语义Web服务QoS描述和发现[J]. 计算机研究与发展, 2011, 48(6): 1059-1066)

[6] Gruber T R. A translations approach to portable ontology specifications [J]. Knowledge Acquisition, 1993, 5(2): 199-220

[7] Etzioni O, Cafarella M, Downey D, et al. Unsupervised named-entity extraction from the Web: An experimental study [EB/OL]. 2005 [2013-07-25]. https://homes.cs.washington.edu/~etzioni/papers/knowitall-aij.pdf

[8] 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

[9] 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

[10] Carlson A, Betteridge J, Kisiel B, et al. Toward an architecture for neverending language learning [C] //Proc of the 24th AAAI Conf on Artifical Intelligence. Menlo Park, CA: AAAI, 2010: 1306-1313

[11] 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

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

[13] 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

[14] 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

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

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

[17] Hffart J, Suchanek F, Berberich K, et al. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia [J]. Artificial Intelligence Journal, 2013, 194(4): 28-61

[18] Suchanek F, Kasneci G, Weikum G. YAGO—A core of semantic knowledge [C] //Proc of the 16th Int Conf on World Wide Web. New York: ACM, 2007: 697-706

[19] Philpot A, Hovy E H, Pantel P. Ontology and the Lexicon [M] //The Omega Ontology. Cambridge: Cambridge University Press, 2008: 35-78

[20] 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

[21] Ponzetto S, Strube M. Taxonomy induction based on a collaboratively built knowledge repository [J]. Artificial Intelligence, 2011, 175(9/10): 1737-1756

[22] Shi Zhongzhi. Knowledge Discovery [M]. Beijing: Tsinghua University Press, 2002 (in Chinese)
(史忠植. 知识发现[M]. 北京: 清华大学出版社, 2002)

[23] Dong Zhendong, Dong Qiang, Hao Changling. Theoretical findings of HowNet [J]. Journal of Chinese Information Processing, 2007, 21(4): 3-9 (in Chinese)
(董振东, 董强, 郝长伶. 知网的理论发现[J]. 中文信息学报, 2007, 21(4): 3-9)

[24] Mei Lijun, Zhou Qiang, Zang Lu, et al. Merge information in HowNet and tongyici cilin [J]. Journal of Chinese Information Processing, 2005, 19(1): 63-70 (in Chinese)
(梅立军, 周强, 臧路, 等. 知网与同义词词林的信息融合研究[J]. 中文信息学报, 2005, 19(1): 63-70)

[25] Huang Cengyang. Theoretical summary of HNC [J]. Journal of Chinese Information Processing, 1997, 11(4): 11-20 (in Chinese)
(黄曾阳. HNC 理论概要[J]. 中文信息学报, 1997, 11(4): 11-20)

[26] Yu Jiangsheng, Yu Shiwen. The structure of Chinese concept dictionary [J]. Journal of Chinese Information Processing, 2002, 16(4): 12-20 (in Chinese)
(于江生, 俞士汶. 中文概念词典的结构[J]. 中文信息学报, 2002, 16(4): 12-20)

[27] Lenat D, Guha R. Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project [M]. Boston: Addison-Wesley, 1989

[28] Xu Wenyan, Liu Sanyang. Logic for knowledgebase systems [J]. Chinese Journal of Computers, 2009, 32(11): 2123-2129 (in Chinese)
(许文艳, 刘三阳. 知识库系统的逻辑基础[J]. 计算机学报, 2009, 32(11): 2123-2129)

[29] Zhong Xiuqin, Liu Zhong, Ding Panping. Construction of knowledge base on hybrid reasoning and its application [J]. Chinese Journal of Computers, 2012, 35(4): 761-766 (in Chinese)
(钟秀琴, 刘忠, 丁盘苹. 基于混合推理的知识库的构建及其应用研究[J]. 计算机学报, 2012, 35(4): 761-766)

[30] Chen Liwei, Feng Yansong, Zhao Dongyan. Extracting relations from the Web via weakly supervised learning [J]. Journal of Computer Research and Development, 2013, 50(9): 1825-1835 (in Chinese)
(陈立玮, 冯岩松, 赵东岩. 基于弱监督学习的海量网络数据关系提取[J]. 计算机研究与发展, 2013, 50(9): 1825-1835)

[31] Kushmerick N, Weld D, Doorenbos R. Wrapper induction for information extraction [C] //Proc of the 15th Int Joint Conf on Artifical Intelligence (IJCAI'97). New York: ACM, 1997: 1-246

[32] Hsu C-N, Dung M-T. Generating finite-state transducers for semi-structured data extraction from the Web [J]. Information Systems, 1998, 23(8): 521-538

[33] Brin S. Extracting patterns and relations from the world wide Web [C] //Proc of the WebDB Workshop at 6th Int Conf on Extending Database Technology (EDBT'98). Berlin: Springer, 1998: 172-183

[34] Riloff E, Jones R. Learning dictionaries for information extraction by multi-level bootstrapping [C] //Proc of the 16th AAAI Conf on Artifical Intelligence. Menlo Park, CA: AAAI, 1999: 474-479

[35] Curran J, Murphy T, Scholz B. Minimising semantic drift with mutual exclusion bootstrapping [C] //Proc of the 10th Conf of the Pacific Association for Computational Linguistics. Melbourne: PACLING, 2007: 172-180

[36] Carlson A, Betteridge J, Wang R, et al. Coupled semi-supervised learning for information extraction [C] //Proc of the 3rd ACM Int Conf on Web Search and Data Mining, WSDM'10. New York: ACM, 2010: 101-110

[37] Song Y Q, Wang H X, Wang Z Y, et al. Short text conceptualization using a probabilistic knowledgebase [C] //Proc of the 22nd Int Joint Conf on Artificial Intelligence, IJCAI'11. New York: ACM, 2011: 2330-2336

[38] Lee T, Wang Z, Wang H, et al. Web scale taxonomy cleansing [J]. Proceedings of the VLDB Endowment, 2011, 4(12): 1295-1306

[39] Suchanek F, Hffart J, Kuzey E, et al. YAGO2: Modular high-quality information extraction with an application to flight planning [C] //Proc of the 15th GI Symp on Database Systems for Business, Technology and Web. Magdeburg: BTW, 2013: 515-518

[40] Lohr. S Aiming to learn as we do, a machine teaches itself [N/OL]. 2010-10-05 [2013-07-25]. http://www.nytimes.com/2010/10/05/science/05compute.html?ref=global-home

[41] Kyle V H. Right now a computer is reading online, teaching itself language [EB/OL]. 2010 [2013-07-25]. http://www.gizmodo.com.au/2010/10/right-now-a-computer-is-reading-online-teaching-itself-language/

[42] Liu S, Liu F, Yu C, et al. An effective approach to document retrieval via utilizing WordNet and recognizing phrases [C] //Proc of the 27th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval, SIGIR'04. New York: ACM, 2004: 266-272

[43] Zamiro O, Etzioni O. Web document clustering: A feasibility demonstration [C] //Proc of the 21st Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval (SIGIR'98). New York: ACM, 1998: 46-54

[44] Salton G, Yang C S, Yu C T. A theory of term importance in automatic text analysis [J]. Journal of American Society for Information Science, 1975, 26(1): 33-44

[45] Broder A. A taxonomy of Web search [C] //Proc of the 25th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval (SIGIR'02). New York: ACM, 2002: 3-10

[46] Cambria E. Isanette: A common and common sense knowledge base for opinion Mining [C] //Proc of the 6th Int Conf on Data Mining Workshop. Piscataway, NJ: IEEE, 2011: 315-322

[47] Wen Hua, Song Yangqiu, Wang Haixun, et al. Identifying users' topical tasks in Web search [C] //Proc of the 4th ACM Int Conf on Web Search and Data Mining, WSDM'13. New York: ACM, 2013: 93-102

[48] Guo Jiafeng. Intent-aware query similarity [C] //Proc of the 17th ACM Conf on Information and Knowledge Management (CIKM'11). New York: ACM, 2011: 259-268

[49] Kim D. Context-dependent conceptualization [C] //Proc of the 23rd Int Joint Conf on Artifical Intelligence (IJCAI'13). Menlo Park, CA: AAAI, 2013: 2654-2661

[50] Tang Yong, Lin Luxian, Luo Yemin, et al. Survey on information retrieval system based on question answering system [J]. Journal of Computer Applications, 2009, 28(11): 2745-2748 (in Chinese)
(汤庸, 林鹭贤, 罗烨敏, 等. 基于自动问答系统的信息检索技术研究进展[J]. 计算机应用, 2008, 28(11): 2745-2748)

[51] Wu Youzheng, Zhao Jun, Duan Xiangyu, et al. Research on question answering &evalution: A survey [J]. Journal of Chinese Information Processing, 2005, 19(3): 1-13 (in Chinese)
(吴友政, 赵军, 段湘煜, 等. 问答式检索技术及评测研究综述[J]. 中文信息学报, 2005, 19(3): 1-13)

[52] Burger J C, Chaudhri V, et al. Issues, tasks and program structures to roadmap research in question answering [EB/OL]. 2001[2013-07-25]. http://www.inf.ed.ac.uk/teaching/courses/tts/papers/qa_roadmap.pdf

[53] Maybury M T. New Directions in Question Answering [M]. Menlo Park, CA: AAAI /MIT Press, 2004

[54] IBM. Watson [EB/OL]. 2013 [2013-07-25] http://en.wikipedia.org/wiki/Watson_(computer)

[55] Kumar D, Ramakrishnan N, Helm R, et al. Algorithms for storytelling [J]. IEEE Trans on Knowledge and Data Engineering, 2008, 20(6): 736-751

[56] Hossain M, Butler P, Boedihardjo A, et al. Storytelling in entity networks to support intelligence analysts [C] //Proc of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD'12). New York: ACM, 2012: 1375-1383

[57] Zaki M J, Ramakrishnan N. Reasoning about sets using redescription mining [C] //Proc of the 11th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD'05). New York: ACM, 2005: 364-373

[58] Fang L, Sarma A D, Yu C, et al. REX: Explaining relationships between entity pairs [J]. VLDB Endowment, 2011, 5(3): 241-252

[59] Quinlan J R, Cameron-Jones R M. FOIL: A midterm report [C] //Proc of the 5th European Conf on Machine Learning (ECML'93). Berlin: Springer, 1993: 3-20

[60] Mitchell T M, Betteridge J, Carlson A, et al. Populating the semantic Web by macro-reading internet text [C] //Proc of the 8th Int Semantic Web Conf (ISWC). Berlin: Springer, 2009: 998-1002

[61] Cohen W W, Page D. Polynomial learnability and inductive logic programming: Methods and results [J]. New Generation Computing, 1995, 13(34): 369-409

[62] Lao N, Mitchell T M, Cohen W W. Random walk inference and learning in a large scale knowledge base [C] //Proc of the Conf on Empirical Methods in Natural Language Processing, EMNLP'11. Stroudsburg, PA: Association for Computational Linguistics, 2011: 529-539

[63] Schoenmackers S, Etzioni O, Weld D, et al. Learning first-order Horn clauses from Web text [C] //Proc of the 2010 Conf on Empirical Methods in Natural Language Processing (EMNLP'10). Stroudsburg, PA: Association for Computational Linguistics, 2010: 1088-1098

[64] Schoenmackers S, Etzioni O, Weld D. Scaling textual inference to the Web [C] //Proc of the Conf on Empirical Methods in Natural Language Processing (EMNLP'08). Stroudsburg, PA: Association for Computational Linguistics, 2008: 79-88

[65] Lao N, Cohen W W. Fast query execution for retrieval models based on path-constrained random walks [C] //Proc of the 16th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD'10). New York: ACM, 2010: 881-888

[66] Lao N, Cohen W. Relational retrieval using a combination of path-constrained random walks [J]. Machine Learning, 2010, 81(1): 53-67

[67] Backstrom L, Leskovec J. Supervised random walks: Predicting and recommending links in social networks [C] //Proc of the 4th ACM Int Conf on Web Search and Data Mining (WSDM'11). New York: ACM, 2011: 635-644

[68] Lichtenwalter R, Lussier J T, Chawla N V. New perspectives and methods in link prediction [C] //Proc of the 16th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD'10). New York: ACM, 2010: 243-252

[69] Adamic L A, Adar E. Friends and neighbors on the Web [J]. Social Networks, 2001, 25(3): 211-230

[70] Newman M. Clustering and preferential attachment in growing networks [J]. Physical Review E, 2001, 64(2): 025102

[71] Katz L. A new status index derived from sociometric analysis [J]. Psychometrika, 1953, 18(1): 39-43

[72] Yin Dawei, Hong Liangjie, Davison B. Strctural link analysis and prediction in microblogs [C] //Proc of the 20th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2011: 1163-1168

[73] Jia Yantao, Wang Yuanzhuo, Li Jingyuan, et al. Structural-interaction link prediction in microblogs [C] //Proc of the 22nd Int Conf on World Wide Web Companion. New York: ACM, 2013: 193-194

[74] Liben-Nowell D, Kleinberg J. The link prediction problem for social networks [C] //Proc of the 12th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2003: 556-559

[75] Zhao Zeya, Jia Yantao, Wang Yuanzhuo, et al. Content-structural relation inference in knowledge base [C] //Proc of the 28th AAAI Conf on Artificial Intelligence. Quebec: AAAI, 2014: 3154-3155

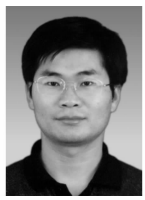
[76] Jia Yantao, Wang Yuanzhuo, Cheng Xueqi, et al. OpenKN: An open knowledge computational engine for network big data [C] //Proc of 2014 IEEE/ACM Int Conf on Advances in Social Networks Analysis and Mining. Los Alamitos, CA: IEEE, 2014: 657-664

Wang Yuanzhuo, born in 1978. PhD, associate professor. IEEE member and senior member of China Computer Federation. His current research interests include social computing, open knowledge network, network security analysis, stochastic game model, etc.

Jia Yantao, born in 1983. PhD, assistant professor. His main research interests include open knowledge network, social computing, combinatorial algorithms.

Liu Dawei, born in 1984. PhD, assistant professor. His main research interests include social network computing and media content analysis.

Jin Xiaolong, born in 1976. Associate professor, PhD supervisor. His research interests include social computing, multi-agent systems, performance modelling and evaluation, etc.



Cheng Xueqi, born in 1971. Professor, PhD supervisor. His research interests include network science, Web search & data mining.