意图知识图谱的构建与应用

陈成[1,2]，陈跃国[1]，刘宸[3]，吕晓彤[1,2]，杜小勇[1,2]
1. 中国人民大学数据工程与知识工程教育部重点实验室，北京 100872；
2. 中国人民大学信息学院，北京 100872； 3. 中国人民大学统计学院，北京 100872

摘要
政府治理的效果评估是一个难题。没有很好的评估方法和评估体系，政府治理的效果就不能得到很好的保障。提出从自然语言问答的角度理解网民在政府治理话题中的意图，并通过构建意图知识图谱，关联语义等价的问题和意图。不同意图又通过实体的相互关联，支持意图的关联和对比。给出了意图知识图谱的定义、构建框架和政府治理场景的使用范例，展示了意图知识图谱是解决政府治理的效果评估问题的一种有效方法。在政府治理的场景中，利用意图知识图谱可以分析对比同一治理话题下不同治理主体之间的意图场，从而深入剖析特定治理主体在特定治理话题下的效果，并发现治理中存在的问题。

关键词
意图理解；知识图谱；自然语言问答；实体识别

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2020014

Constructing and analyzing intention knowledge graphs

CHEN Cheng[1,2], CHEN Yueguo[1], LIU Chen[3], LYU Xiaotong[1,2], DU Xiaoyong[1,2]
1. Key Laboratory of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing 100872, China
2. School of Information, Renmin University of China, Beijing 100872, China
3. School of Statistics, Renmin University of China, Beijing 100872, China

Abstract
It is very difficult to evaluate the effects of government governance. Without a good evaluation method and evaluation system, the effects of government governance cannot be guaranteed. Understanding the intention of web users in the topic of government governance from the perspective of natural language question-and-answering was proposed. By constructing a knowledge graph of intentions, equivalent questions and intentions were associated. The definition, construction framework and usage examples in government governance were illustrated, showing that knowledge graph of intentions is an effective way to evaluate the effects of government governance. In the context of government governance, by using the knowledge graphs of intentions, the intention fields between different governance subjects under the same governance topic were analyzed and compared, the effects of specific governance subjects on specific governance topics were analyzed, and the issues remained in government governance were found.

Key words
intention understanding, knowledge graph, natural language question answering, entity recognition

1 引言

人工智能和大数据等技术在政府治理中发挥的作用日益突显，交通、医疗、城市规划、一站式政务服务、灾情疫情应急处理等越来越多的市政服务应用开始尝试利用大数据与人工智能技术。我国政府也在强调利用大数据等技术手段推进政府决策科学化、社会治理精准化、公共服务高效化。

随着政府治理大数据应用的开展，评估政府治理的效果成为迫在眉睫的事情。没有有效的评估机制和手段，就无法深入利用和改进大数据技术，无法有效地将政府治理与大数据技术深度结合，导致难以真正做到政府决策的科学化、社会治理的精准化以及公共服务的高效化。

一种可能的评估手段是从政府内部构建治理的评估机制，相对公众用户而言，政府具有更好的数据优势，能够通过政府治理的多种应用场景收集数据，通过自评估的方式评价治理的效果。然而，由政府自身主导的治理效果评价也可能会存在一些问题。首先，尽管政府具有较好的公信力，但也不能排除在数据采样或发布手段等方面出现问题，使得评估效果偏离实际，不能真正发挥导向作用；其次，很多政府治理场景需要跨部门的协作与评估，不同部门之间数据共享困难，难以协作评价治理的效果。因此，政府单方面的治理效果评估机制会存在一定的问题。

另一种政府治理效果的评估手段是从公众的角度、从用户的角度评估政府治理的效果。一般需要使用问卷调查的方式收集大量的用户反馈，才能有效地评估政府治理的效果。然而，使用问卷收集公众反馈进行有效评估是非常困难的事情，最突出的问题就是只能获得书面的社会信息，而不能了解生动、具体的社会情况。其次，问卷回复率和有效率低，对无回答者的研究比较困难，反馈人群的偏差和反馈数据的偏差会造成很大的评估偏差，从而无法实现有效评估政府治理效果的目的。

本文通过收集互联网上多种来源（如网页论坛、微博、百度知道和知乎等问答社区）的用户自然语言问题，并对问题进行分析处理，将其抽象成问题背后的意图。如果2个意图语义相同，或者包含相同的实体、词条，那么它们之间就可以建立关联，进而构建公众意图知识图谱。进一步结合政府治理的一些应用场景，分析公众意图在不同城市的不同政府治理话题下的分布差异性，分析公众意图的演化规律，进而评估一些政府治理的应用场景的治理效果。采用自然语言问题构建意图知识图谱，并用来评估政府治理效果有以下原因：第一，相比搜索引擎的用户搜索日志，自然语言问题更详细（搜索日志多为关键词），意图表达更精确；第二，用户搜索日志多为搜索引擎公司的私有数据，难以公开研究和应用，也缺少第三方数据的公立性；第三，相比网页和社交媒体的文本数据，用户的自然语言问题意图更明显、表达更精炼，更有利于精准识别用户的意图；第四，社交媒体上与政府治理相关的数据通常带有强烈的主观色彩，煽动性强，真假混杂，而自然语言问题通常是用户意图的自然表达，不是情绪的宣泄，用自然语言问题分析潜在的政府治理效果受到的干扰会更小；第五，社区的自然语言问题，除了问题，通常还有答案以及用户对问题及答案的反馈数据，这更有利于识别问题的意图以及问题的质量和重要性；第六，互联网上尤其是问答社区（如百度知道、搜狗问问等）拥有

大量的公众用户问题，具有良好的数据基础。基于以上考虑，笔者认为从互联网上爬取的海量自然语言问答数据非常适合构建意图知识图谱，适合在这样的知识图谱上分析政府治理的效果。现代汉语词典将意图定义为“希望达到某种目的打算”。意图知识图谱是对指定领域的问题意图的结构化整理，以知识图谱的形式呈现，便于计算机处理以及人们理解分析。基于此，本文创新性地提出构建意图知识图谱，并给出一个概要性的解决方案。

那么，意图知识图谱与一般意义的知识图谱有什么相同和不同之处？知识图谱需要具有节点和边的概念，传统知识图谱的节点表示信息实体或者实体的属性值，边表示2个被连接实体的关系或者一个实体的某个属性。在意图知识图谱中，节点与传统知识图谱相同，是一个具体的信息实体或词条。意图的形式化表达逻辑上是由多个实体或词条构成的集合，它以超边（hyper edge）的形式连接其包含的多个实体和词条，进而表达一个意图。边则可以有多种类型，如实体和词条之间的边，与传统的知识图谱相同，这些边用于实体之间的关联，表示实体的属性、实体和概念之间的上下位本体关系或者对某个实体属性施加的动作；意图超边和实体/词条节点之间的连接，表示某个意图超边包含了这些实体/词条，是对意图的描述；意图之间的边可以表示意图之间的等价关系（具有相同的意图）或者是在具体应用下构建的意图之间的相关关系（意图之间的某种关联）。

有了上述的意图知识图谱，就可以在给定的由实体和词条形成的意图下，分析它们涵盖的意图范围，提取与查询相关的子知识图谱，根据其等价关系进行聚类，对不同意图之间的关联性进行分析，形成意图场（子知识图谱）。因此，意图知识图谱是异构信息网络的变种，它有不同类型的信息实体、不同类型的边（包括超边）。下面将研究如何构建意图知识图谱以及如何将意图知识图谱应用到政府治理的场景中。

2 意图知识图谱的构建

从问题句子构建意图知识图谱的过程与传统知识图谱的构建过程不同，需要先做问句的抽取和提问意图主体的识别，以便下一步的意图聚类和分析。在很多政务治理场景下，意图分析强调意图的时效性，因此无法直接以现有的百科知识图谱为基础进行构建。**图1**给出了意图知识图谱的构建框架，它主要包括数据获取、信息抽取、知识融合、质量评估和知识更新4个模块。

2.1 数据获取

按数据来源，数据可分为3类：万维网；新浪微博、腾讯微博等短消息广播平台；百度知道、搜狗问问、知乎等问答社区。本节将从收集技术、清洗整理、数据抽取这3个方面评析万维网、微博、问答社区数据获取的质量和效率。

万维网数据包含了来自政府网站、民生论坛、少量的问答社区和少量的微博的消息，十分纷繁复杂，因此假设使用搜索引擎可以收集到万维网上所有与中文政务相关的问题，就可以较为全面地反映百姓对政务的各种意图。但是万维网上的问题句比较稀疏，与微博平台相同，大部分句子不是问句，需要做前期的问题句子过滤。关键词过滤法最大的弱点是难以过滤反问句，也有一系列研究反问句识别的技术。笔者以“北京”为主题爬取了万维网上约

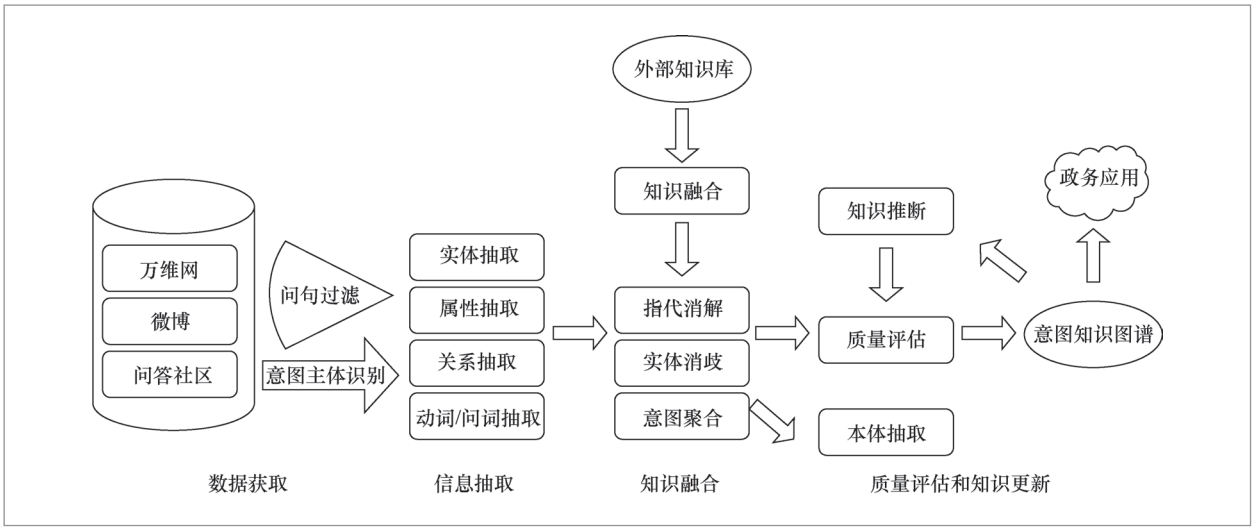

图1 意图知识图谱的构建框架

80万个网页，以“吗”“怎么”“什么”等疑问词为关键词过滤得到1 635 670个30字以内的短问句，随机抽样2 000条检测问句，识别准确率为96%。也可以手工标注问句/非问句二分类数据集，抽取句子的词法句法特征，用逻辑回归、SVM或LSTM等深度学习模型训练问句/非问句二分类模型。

万维网数据纷繁复杂，而研究者一般只关心特定领域的政府治理意图，因此需要做特定领域、限定主题的网络爬虫。主题爬虫技术在搜索引擎、领域语料库构建方面已经得到了广泛应用[1]。在构建指定主题的意图知识图谱过程中，主题爬虫从与事先定义的主题高度相关的统一资源定位符（uniform resource locator，URL）种子集合出发，根据一定的分析算法，对爬行网页进行主题相关分析，过滤与主题不相关的网页，并判断之后的爬虫走向，在不断抓取网页的过程中，将与主题相关的链接放进待爬行队列中，重复这个过程直到意图知识图谱没有较大更新、模型收敛为止。此时图谱构建初步完成，达到可用的程度。

微博的数据量较大，笔者于2020年1月27日晚以“新型肺炎”为关键词搜索微博，得到11 474 392条微博结果。知乎网站上“新型肺炎”话题下仅有500个问题，说明问答社区的时效性较差。百度知道和搜狗问问2个平台允许的搜索翻页数量有限，较难深入爬取，最多也只能得到600~700个问题。问答社区和微博都有格式规范的时间戳信息，方便提取。而网页文本比较混乱，从网页文本提取网页正文的发布时间几乎是不现实的。在爬取的80万个网页中，仅有15%的超文本传输协议响应（HTTP response）带有标记更新时间的字段。问题的发布时间、点赞和浏览数可以反映其意图的流行度。微博和问答社区有规范的点赞量、浏览数、转发量信息，而网页的浏览数一般较难获取。

不同数据源的特点对比见**表1**。

对于一个给定的应用目标和主题，可以融合多种来源的数据构建知识图谱，以期得到更全面的民意评估。作为参考，Wang等人[2]在完善微软概念图谱的过程中，利用已有概念图谱、网页文本、Bing搜索日志和外部知识库进行了多个数据源的加权融合。

表1 不同数据源特点对比

对比项	收集难度	数据量	时效性	流行度判断
万维网	高	大	低	难
微博	较高	大	高	易
问答社区	低	中	中	易

2.2 信息抽取

传统的信息抽取是指从非结构化的文本中识别实体，并发现实体的属性、实体之间的关系，在此基础上形成本体化的知识表达。笔者认为在意图知识图谱构建过程中，问句中关键动词或意图词组的识别和实体、属性一样重要，需要优先考虑。问句中的关键动词往往是提问意图的落点，如户口的“迁入”“迁出”、社保的“办理”“代缴”。对于部分无法用单个动词表示的意图，如“退款申请”“只退一半”这样包含动词的意图词组，需要进一步研究词组挖掘的方法进行识别。也有部分提问意图落在静态的实体或属性上，如猪肉“价格”。可以尝试结合词向量技术改进现有的词组挖掘方法，进行意图词组的抽取。

实体抽取也被称为命名实体识别。短文本的实体识别方法主要分为以下2类。一类是基于词典和规则的方法，是命名实体识别中最早使用的方法。人工构建准确度高，但昂贵费时，目前已有不少研究提出自动构建的方法。参考文献[3]提出一种自动归纳抽取规则的算法，参考文献[4]提出利用机器学习框架长期维护规则集。另一类是基于机器学习模型的命名实体识别，其预先对一部分文档进行实体标注，利用这些文档进行机器学习模型的训练，然后用训练好的模型对没有遇到过的文档进行命名实体识别和标注。2000年年初，隐马尔可夫模型（hidden Markov model，HMM）、条件随机场（conditional random field，CRF）等概率图模型得到广泛应用。近几年，随着深度学习的兴起，结合条件随机场的双向长短期记忆网络（Bi-LSTM-CRF）一度成为研究热点。2018年年底，谷歌公司提出自然语言预训练模型BERT[5]，在11项自然语言处理（natural language processing，NLP）任务中取得卓越表现。最近出现了很多结合BERT的实体抽取方法。

属性抽取方法有2种：一种是将从百科网站上抽取的结构化数据作为可用于属性抽取的训练集，然后将该模型应用于开放域中的实体属性抽取过程；另一种是根据实体属性与属性值之间的关系模式，直接从开放域数据集上抽取属性。

意图知识图谱关系抽取的目标是构建意图词组、实体以及属性之间的语义链接，关系的基本信息包括参数类型、满足此关系的元组模式等。关系抽取方法可分为开放式实体关系抽取和基于联合建模的实体关系抽取2类。目前，针对传统的实体、属性、关系抽取已有丰富的研究和较为成熟的解决方案，而开放领域的问题意图词组的准确识别暂未见相关研究，数据的有限性会为抽取精度带来很大的挑战。

2.3 知识融合

通过信息抽取，从原始问题句子文本数据中获取到了实体、关系、实体的属性信息以及表达问句意图的词组。有了这些信息碎片，下一步就是要将它们拼接成完整的意图知识图谱，给抽取出来的要素赋予结构。这一过程有3个问题需要解决。

- 实体链接：将从文本中抽取得到的实体对象链接到知识库中对应的正确实体对象。

● 意图融合：在从多个问句抽取出来的实体、属性、动作、关系中识别出相同或相似的意图，将相同或相似的意图抽象成超图的超边。

● 知识合并：通过现有的外部知识库（如百度百科）或领域关系数据库，对意图知识图谱进行补充和纠错，完善现有图谱的构建流程。

实体链接任务的基本思想是首先根据给定的实体指称项，从知识库中选出一组候选实体对象，然后通过相似度计算将指称项链接到正确的实体对象。在传统知识图谱构建中，针对实体消岐和共指消解已有较多研究。意图知识图谱在实体链接方面与传统知识图谱并无不同，参考实现即可。

意图融合的第一阶段是对单个问句包含的单个意图的处理。从单个问句中抽取出的实体、属性、动作、关系自然形成一条意图路径，类似于异构信息网络的路径代表的现实含义。超图是指每一个边可以包含2个以上的点的图，包含多个节点的边为超边，超边的作用是融合意图。对多个相似意图进行融合对比，可形成超边。例如：“之前申请信息是通过工作居住证申请的，现在想换成社保，信息如何更新”“持有北京居住登记卡算工作居住证吗”，这其中多个信息实体和意图动作融合成一个与工作居住证相关的意图超边；“每次延期审核都要检查近60个月的个税记录吗”“我名字有曾用名，税务审核个税时会有影响吗”，这是关于个税证明的意图超边，反映出对税务证明各方面的意图。

知识合并过程是指借助外部知识库完善现有意图知识图谱的过程。此过程分为2步：第一，数据层的融合，包括实体、属性、关系、意图超边的融合，要避免数据冲突和冗余；第二，在模式层将新的本体融入现有本体库。

与北京车牌摇号相关的意图知识图谱样例如**图2**所示，数据来自北京市小客车指标管理调控信息系统。不同意图通过相同的动作或相同的实体互相关联，一个意图以单条路径或者超边的形式表达，形成与北京车牌摇号相关的意图知识图谱。

对意图知识图谱的所有点、边要素的类型进行总结，具体如下。

● 节点：概念、实例、属性、意图词组。概念和实例在本文中未做严格区分，统称为实体。

● 边：本体层级关系、概念实例关系、实例属性关系、实例和意图词组之间的连边、属性和意图词组之间的连边。还有能够完整表达一个问句意图的意图超边，关联所有需要的节点，如**图2**中虚线圆圈表示的意图超边。

2.4 质量评估和知识更新

在意图知识图谱构建初期，数据准备工作难以完全由机器自动完成，需要大量的领域专家进行人工数据标注。近年来人在回路（human in the loop）的技术路线备受学术界和工业界的关注[6]。众包数据抽取方法首先通过自动的数据抽取方法生成候选结果，然后让众包工人进行验证，这种方法取得了明显高于自动数据抽取方法的效果。

质量评估也是知识库构建技术的重要组成部分。质量评估对知识的可信度进行量化，通过舍弃置信度较低的知识保障知识库的质量。从逻辑架构方面看，意图知识图谱的更新包括概念层的更新和数据层的更新。概念层的更新是指提取新的问题句子后获得了新的信息实体和意图，并将其对齐或添加到知识库的概念层中。数据层的更新时主要是指新增或更新实体、属性、关系、动作，对数据层进行更新时需要

考虑问题句子来源的可靠性、时效性、数据的一致性，选择各数据源中可信度高的数据加入知识库。

3 意图知识图谱的应用

3.1 市政服务场景的应用

与政务相关的自然语言问题一般包括主体与背景2个部分。主体包含地点、政策、行为等主要实体和动作，背景中则包含一些复杂的信息（如用户的个人信息等）。在百度知道上，有关北京的一个具体问题“北京市户口名下有外地牌照车还可以在北京申请摇号吗”中，问题的主体是“北京申请摇号”，背景则是“北京户口名下有外地牌照车”。问题包含多个实体，但要理解问题的意图，最重要的实体还是地点（北京）、政策（摇号）、行为（申请）。用户询问的意图是在“北京户口”且“名下有外地牌照车”的背景下“是否可以申请”。在意图知识图谱上，一个意图可以由多个实体和行为连接而成的路径构成，可以由路径还原问题的主干，而实体和行为之间的联系可以透露丰富的信息。

在市政服务的具体场景中，应用意图知识图谱可以更加高效地协助政府工作人员定位民众最关切、最亟待解决的问题，发现治理中存在的不足，从而及时满足民众的需求。**图3**展示了利用百度知道中与“深圳”“户口”相关的50条问题构建的意图知识图谱。可以清晰地看出，民众对于“深圳户口”的意图主要有“迁入”“迁出”“办理”“转移”“利弊”“市内迁移”6种，每个意图对应不同的实体，就“迁入”“迁出”来说，“深圳”与“北京”以及广东省的几个市之间存在相互关系。进一步，结合相关问题的个数，政策、地点等特定实体

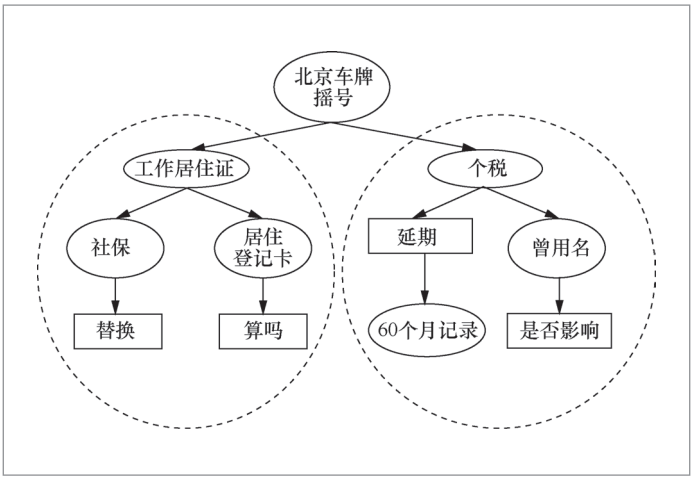

图 2 与北京车牌摇号相关的意图知识图谱样例

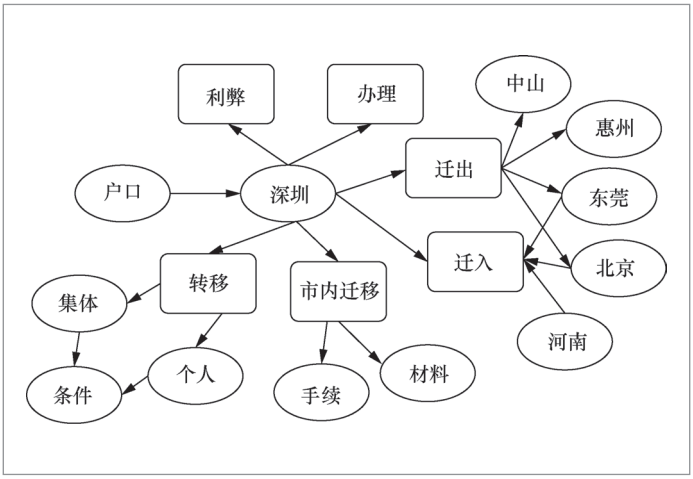

图 3 与“深圳”“户口”相关的问题的意图知识图谱

的出现次数、回答数、点赞量等，可以研究不同问题的热度，从而发现民众最关心和亟待解决的问题，发现政策中可能存在的漏洞，为政府高效性、针对性、智慧性地治理提供了便利。

意图知识图谱抽取的本体包含类、关系、函数、公理和实例5种元素。在政务领域下，由政务问题抽取的本体的类包含地点、政策、意图和人员背景4个部分，本体之间的关系有上下位、施加动作、相似问题

等，函数和公理则对类和关系进行约束，例如某些非区域性政策，政策与地区之间的关系不能被简单地刻画，因为不同地区可能会对政策有因地制宜的调整，从而产生完全不同的子图。

以本体组织指定领域的问句语料库，使自然语言变得可计算、可分析。抽取出的本体可以回溯原问句的主干部分。例如，问题“北京户口怎么办理”可以在意图知识图谱上转化为“北京”“户口”2个政策和“办理”这一意图，工作人员可以很容易地将该问题还原。在意图知识图谱的本体构建完成之后，可以处理复杂多样的关联分析，例如探索政策的地区差异、比较不同城市的治理能力等。政府工作人员可以结合实际需求进行交互式、探索式的数据分析。同时，可以利用交互式机器学习技术，学习推理、纠错、标注等交互动作，不断沉淀知识逻辑，提高系统智能性，方便政府对民情民意进行及时反馈。在可视化方面，图谱可以将相关信息更加清晰地展示给非专业人员，提高数据的可理解性。在数据分析方面，图谱可以帮助政府发现人民群众最关心和亟待解决的问题以及这些问题之间的关联性，提高政府的办事效率，更好地为人民群众服务。

3.2 智慧政府服务的比较

政策具有时效性和地域性，从而加大了比较不同地方政府之间工作效率的难度。而意图知识图谱的构建使得比较政府治理效果成为可能，这种比较可以通过问题的相似度、热度、时间等进行综合考虑。**图4**展示了深圳市宝安区政府参加“民心桥”问政活动时居民提出的部分代表性问题的意图知识图谱。数据来源于深圳市政府在线网站，从4年231个各类问题中，

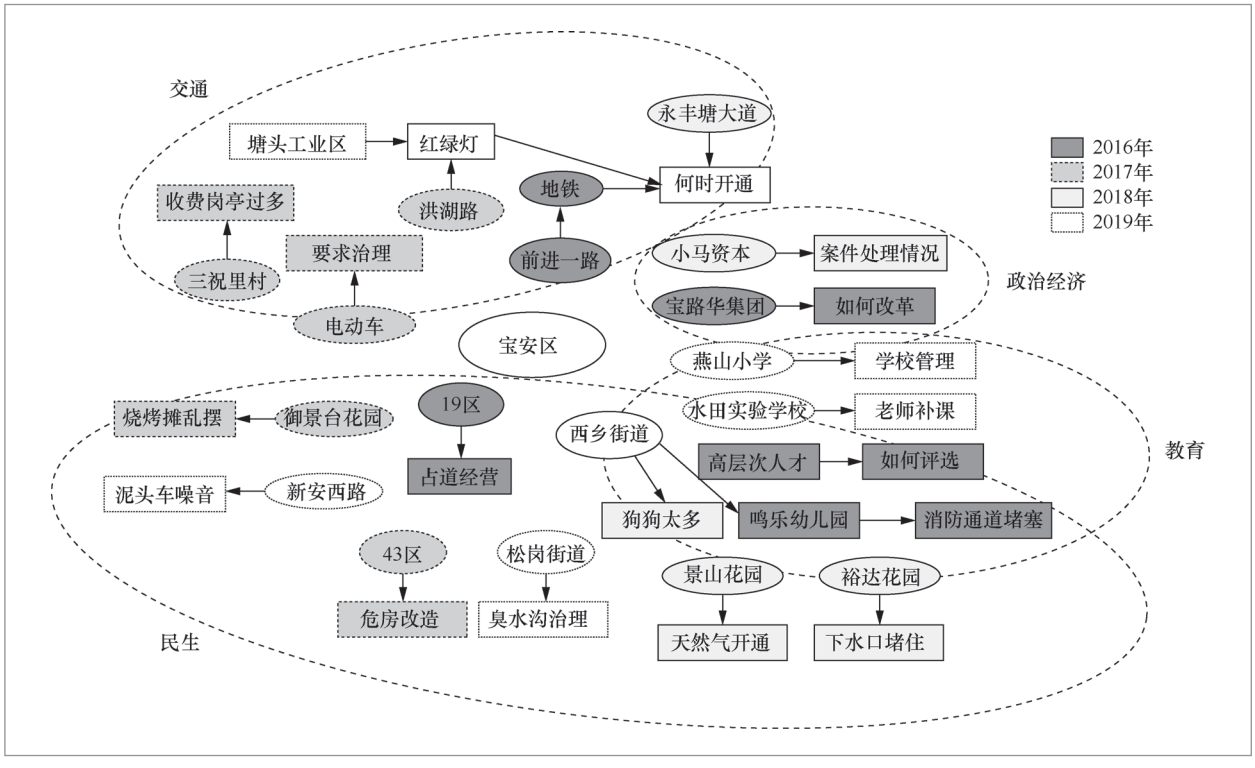

图 4 深圳市宝安区“民生桥”问政活动部分问题的意图知识图谱

使用简单随机抽样法每年选取5个问题，以此构建意图知识图谱（为了简洁起见，图中省略了宝安区和地理位置处于宝安区的实体之间的关系连接。如“宝安区→电动车”“宝安区→三祝里村”连边省略不画）。从**图4**可以看出，民众关心最多的主要是与生活息息相关的民生、交通、教育三大类问题，大部分问题的地点十分具体，明确到小区、道路和某些机构。同时，不同年份关注的问题侧重点略有不同，例如2016年的“高层次人才评选”、2018年的“小马资本案件”，显示了民众问政的时效性。最受关注的问题可以集中反映出政府工作的不足，使得政府机关能够及时做出反应，提高服务质量。

在智慧政务方面，意图知识图谱有广阔的应用空间。一是可以通过预计算和探索式、交互式数据分析系统的构建节约人力资源。可以在图谱构建的过程中进行预计算，从而避免重复进行简单的分析任务；探索式数据分析可以在不明确最终分析目标的思考过程中，辅助分析师发现、求证、推理，使数据分析工作变得简单；使用交互式机器学习可以积淀知识逻辑，进一步提高系统智能性。二是可以进行动态分析和实时舆情监测。意图知识图谱的结构随时间的变化能够反映出人民群众关心问题的改变和诸多社会经济因素，从而在智能交通、城市规划、灾情疫情应急处理等领域发挥作用。三是可以构建高效的政务问答系统。目前政府网站的政务问答大多基于人工操作，回复慢，解决问题能力差，而利用意图知识图谱在政务专业领域可以更好地识别用户问题的意图，从而提高政务问答效率。

3.3 意图检索及意图场的构建

意图知识图谱的使用用户主要分为2类，分别是政府内部工作人员和政府面向的社会公众群体。政府内部工作人员对意图知识图谱的使用集中体现在“了解民情、知晓民意”方面，通过由网络问答数据构建的网络知识图谱查找社会和政策漏洞，而对于政府面向的社会公众群体来说，最主要的是人机对话、智能问答等政府服务方面的应用。这2类用户虽然在使用意图知识图谱方面表现不同，但其根本意图都是检索。首先，对于政府内部工作人员，特别是专业技术人员来说，可以通过给定的种子实体和词条进行意图的检索。给定实体和词条后，在构建好的意图知识图谱中进行查找，得到相关的意图集合，并提取出与这些意图相关联的子意图知识图谱，把这些抽取得到的子知识图谱进行意图的关联分析对比，最终得到与用户输入的种子相关的意图场。其次，针对普通群众的意图检索，考虑到社会群体的多样性和图谱使用的友好性，可以将其检索输入形式进一步扩展为自然语言问题。运用相关自然语言处理技术[7]，对用户输入问题的实体和词条进行识别，再通过抽取得到的实体和词条进行意图的检索。意图检索流程如**图5**所示。

在意图检索的过程中，为了更好地满足和匹配用户查询的意图以及保障查询结果的完整性，要将与语义等价的意图从意图知识图谱中抽取出来。另外，类似于搜索推荐，可以在意图检索结果展示中加入用户检索意图的相关意图，这样既可以满足查询的高效性，也可以满足使用的友好

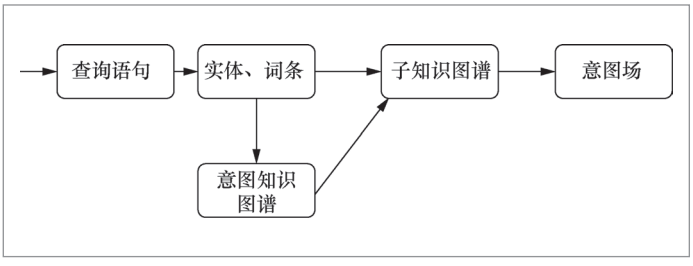

图 5 意图检索流程

性。根据用户输入的实体和词条的不同而产生不同的意图场，通过和其他意图场的对比分析，发现潜在问题。

3.4 意图空间的浏览

为了更好地引导用户浏览意图空间，高速有效地完成浏览目的，而不只局限于用户进行意图检索时输入的意图，需要在进行意图知识提取和局部意图知识图谱的构建过程中，通过语义分析进行语义等价判断[8]。在第3.2节中介绍的深圳市宝安区“民生桥”问政活动部分问题的意图知识图谱中，用户输入“塘头工业区交通”，通过与构建意图知识图谱时的问题进行语义等价判断，可以将包括红绿灯开通等一系列问题在内的答案反馈给用户。

由于自然语言的复杂性，用户的一次查询语句基本只能体现一个意图，为了实现政府工作的简便化和易操作性，通过对意图知识图谱中意图的相关性进行分析，在检索结果展示中加入相关性较强的非查询意图推荐。例如，在第3.2节介绍的意图知识图谱中，用户搜索意图为“西乡街道环境”，此处的环境大多是自然环境，但某些情况下也指居住环境、学校环境等，而居住环境和学校环境更多地出现在“设施”这一类中，因此在用户查询结果中，将加入“西乡街道设施”这一相关意图。即如果从用户查询语句中识别出的实体在某一意图中频繁出现（相关性极强），则可以将这一意图作为相关意图，并将其作为推荐的查询结果返回。

4 值得关注的研究点

（1）知识图谱的形式化表达方法

传统的知识图谱元素仅有实体、属性（值）、关系。来源于问题句子的意图知识图谱通常不关心属性值，只关心属性本身。与传统知识图谱不同，意图知识图谱必须体现的要素是意图。本文提出由一个实体或词条构成的集合（超边）表达一个意图，那么对于描述给定主题的意图知识图谱，如何对图中节点做聚合，形成意图超边，是一个新颖且有现实需求的研究点。由于图谱构建前期自然语言理解的不确定性，意图在图谱中或许也可表达为带有概率的节点和边集合。

（2）意图知识的更新维护

对于疫情、灾害类的时政消息，意图的时效性尤其重要。工作人员不仅要关注更新的准确性（如实体链接和知识合并），也要关注更新的有效性，将新的意图融合进现有的意图知识图谱，与相关旧意图进行对比，体现最具时效性的民情民意，给执政者新的启发。如何高效地完成准确更新，如何定义并解决有效性，还有待解决。

（3）意图知识图谱的搜索

意图知识图谱是基于大量问句构建的，初步解决了浏览原始问句意图信息过载的问题。但是面对具有万亿节点的图谱，搜索技术的研究也是必要的。搜索的界面和形式会直接影响用户理解搜索的信息。意图知识图谱的搜索需要帮助用户找到最想要的意图信息，提供更全面的摘要，使搜索更有深度和广度，既需要全面地总结出与搜索话题相关的意图，也要将搜索结果范围缩小到用户最想要的含义。

（4）意图知识图谱分析

意图知识图谱的核心要素是意图。意图分析的首要研究问题是怎样发现值得对比的相似、相关意图，形成有用的意图场。结合不同的分析场景需求，需要明确有用性的定义，在图谱中可进一步将意图场的识别抽象为子图发现或者聚类问题。如何将分析结果可视化地展现出来，让普通用

户也一目了然，值得研究者们进一步思考与实践。

（5）意图知识图谱探索

数据探索旨在当用户无法明确描述最终要寻找的目标时，帮助用户从数据中提取知识[9]。意图知识图谱构建了一个与提问意图相关的完整意图知识体系，因此用户在使用中往往会有意想不到的发现。在意图搜索分析过程中，用户可能会了解到某个新的意图或新的联系，促使其进行一系列的全新搜索查询；或者当用户不明确要搜索什么时，可以结合本体库或概念图谱的上下位结构，对意图图谱建立图分割，基于图分割为用户提供一个对指定话题意图全面概览的浏览探索入口，甚至是相关话题意图的逐步探索引导机制。

5 结束语

意图知识图谱是一个全新的研究领域，需要做的研究工作还很多。也可以将意图知识图谱理解为对意图进行管理的数据库，需要思考分析师与数据库中的意图元素如何交互、如何扩展新的查询语言或查询接口以及如何优化基于意图超边的数据存取算法。

参考文献：

[1] NOVAK B. A survey of focused web crawling algorithms[C]// SIKDD, October 15, 2004, Ljubljana, Slovenia. New York: ACM Press, 2004: 55-58.

[2] LEE T, WANG Z Y, WANG H X, et al. Attribute extraction and scoring: a probabilistic approach[C]// 2013 IEEE 29th International Conference on Data Engineering, April 8-11, 2013, Brisbane, Australia. Piscataway: IEEE Press, 2013: 194-205.

[3] TATAR S, CICEKLI I. Automatic rule learning exploiting morphological features for named entity recognition in Turkish[J]. Journal of Information Science, 2011, 37(2): 137-151.

[4] PETASIS G, VICHOT F, WOLINSKI F, et al. Using machine learning to maintain rule-based named-entity recognition and classification systems[C]// The 39th Annual Meeting on Association for Computational Linguistics, July 6-11, 2001, Toulouse, France. [S.l.]: ACL, 2001: 426-433.

[5] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[J]. Computer Science, 2018, arXiv: 1810. 04805.

[6] 范举，陈跃国，杜小勇. 人在回路的数据准备技术研究进展[J]. 大数据，2019，5(6)：3-18. FAN J, CHEN Y G, DU X Y. Progress on human-in-the-loop data preparation[J]. Big Data Resreach, 2019, 5(6): 3-18.

[7] ANTUNES F, FREIRE M, COSTA J P. Semantic web and decision support systems[J]. Journal of Decision Systems, 2016, 25(1): 79-93.

[8] VARELAS G, VOUTSAKIS E, RAFTOPOULOU P, et al. Semantic similarity methods in wordNet and their application to information retrieval on the web[C]// The 7th Annual ACM International Workshop on Web Information and Data Management, November 19-23, 2005, Bremen, Germany. New York: ACM Press, 2005: 10-16.

[9] IDREOS S, PAPAEMMANOUIL O, CHAUDHURI S. Overview of data exploration techniques[C]// The 2015 ACM SIGMOD International Conference on Management of Data, May 31-June 4, 2015, Melbourne, Australia. New York: ACM Press, 2015: 277-281.

2020014-11

作者简介

陈成（1995- ），男，中国人民大学信息学院博士生，中国计算机学会会员，主要研究方向为大数据分析、知识图谱，参与国家自然科学基金重点项目：政府治理大数据——行为知识图谱关键技术研究。

陈跃国（1978- ），男，博士，中国人民大学数据工程与知识工程教育部重点实验室教授、博士生导师，中国计算机学会高级会员，数据库专业委员会秘书长，主要研究方向为大数据分析系统和语义搜索，主持国家自然科学基金重点项目1项，广东省科技应用重大专项1项，近年来在国际重要期刊和会议上发表论文20余篇。

刘宸（1998- ），男，中国人民大学统计学院本科生，主要研究方向为数据挖掘。

吕晓彤（1997- ），女，中国人民大学信息学院博士生，中国计算机学会会员，主要研究方向为大数据分析和信息提取、信息检索等。

杜小勇（1963- ），男，博士，中国人民大学信息学院教授、学术委员会主任、博士生导师，中国人民大学校长助理，数据工程与知识工程教育部重点实验室主任。兼任教育部科学技术委员会信息学部委员，国家重点研发计划“云计算与大数据”专家组成员，中国计算机学会理事、教育工作委员会主任、大数据专家委员会主任，《大数据》副主编、全国信息技术标准化技术委员会大数据标准工作组副组长等。先后获得国家科技进步奖二等奖、北京市科技进步奖一等奖、教育部科技进步奖一等奖、中国计算机学会科学技术奖一等奖等奖项。

收稿日期： 2020-01-31
通信作者： 陈跃国，chenyueguo@ruc.edu.cn
基金项目： 国家自然科学基金资助项目（No.U1711261）
Foundation Item: The National Natural Science Foundation of China（No.U1711261）