浅析知识图谱于智能搜索领域中的应用

By Peng Zhang

2020-10-22

目录

- 引论
- 搜索的发展、现状及面临的问题
- 智能搜索的概念及发展
- 知识图谱的概念及发展
- 智能搜索与知识图谱

引论

- 在大数据时代，人们需要在最短的时间内，得到最有效的数据信息。
- 知识图谱是当下科技热点，并有众多应用分支；但针对智能搜索的应用方面，其技术目前尚在摸索中前行。
- 本文旨在对最新的智能搜索技术、知识图谱应用，以及两者之间如何有效地融合，进行了分析、探讨与预研。

搜索引擎的发展

搜索引擎的四个时代（张俊林《这就是搜索引擎》）

1. 史前时代：分类目录的一代

这个时代成为“导航时代”，Yahoo和国内hao123是这个时代的代表。

2. 第一代：文本检索的一代

文本检索的一代采用经典的信息检索模型，如布尔模型、向量空间模型或者概率模型，来计算用户查询关键词和网页文本内容的相关程度。

3. 第二代：链接分析的一代

这一代搜索引擎充分利用了网页之间的链接关系，并深入挖掘和利用了网页链接所代表的含义。

4. 第三代：用户中心的一代第三代即理解用户需求为核心的一代搜索引擎。不同用户即使输入同一个查询词，但其目的可能不一样。比如同样输入“苹果”作为搜索词，一个追捧iPhone的时尚青年和一个果农的目的会存在巨大的差异。

传统搜索引擎的核心技术

传统搜索引擎的核心技术常见包括：分词技术、网络蜘蛛、索引技术和词频指数。知识图谱或知识计算引擎被认为是下一代搜索引擎。

传统模式搜索的问题

- 以往传统的搜索引擎，是基于关键词或字符串的，并没有对查询的目标（通常为网页）和用户的查询输入进行理解。
- 传统搜索引擎在搜索准确度方面存在明显的缺陷，即由于HTML形式的网页缺乏语义，难以被计算机理解。
- 传统搜索引擎的搜索模式单一，即缺乏高级详细的搜索条件，大多也缺乏诸如语音、图像、视频识别等方面的搜索手段。
- 传统搜索引擎搜索的结果，需要搜索者在浩若烟海的结果页里逐页打开并分析鉴别最终提取有用的部分，效率低下。

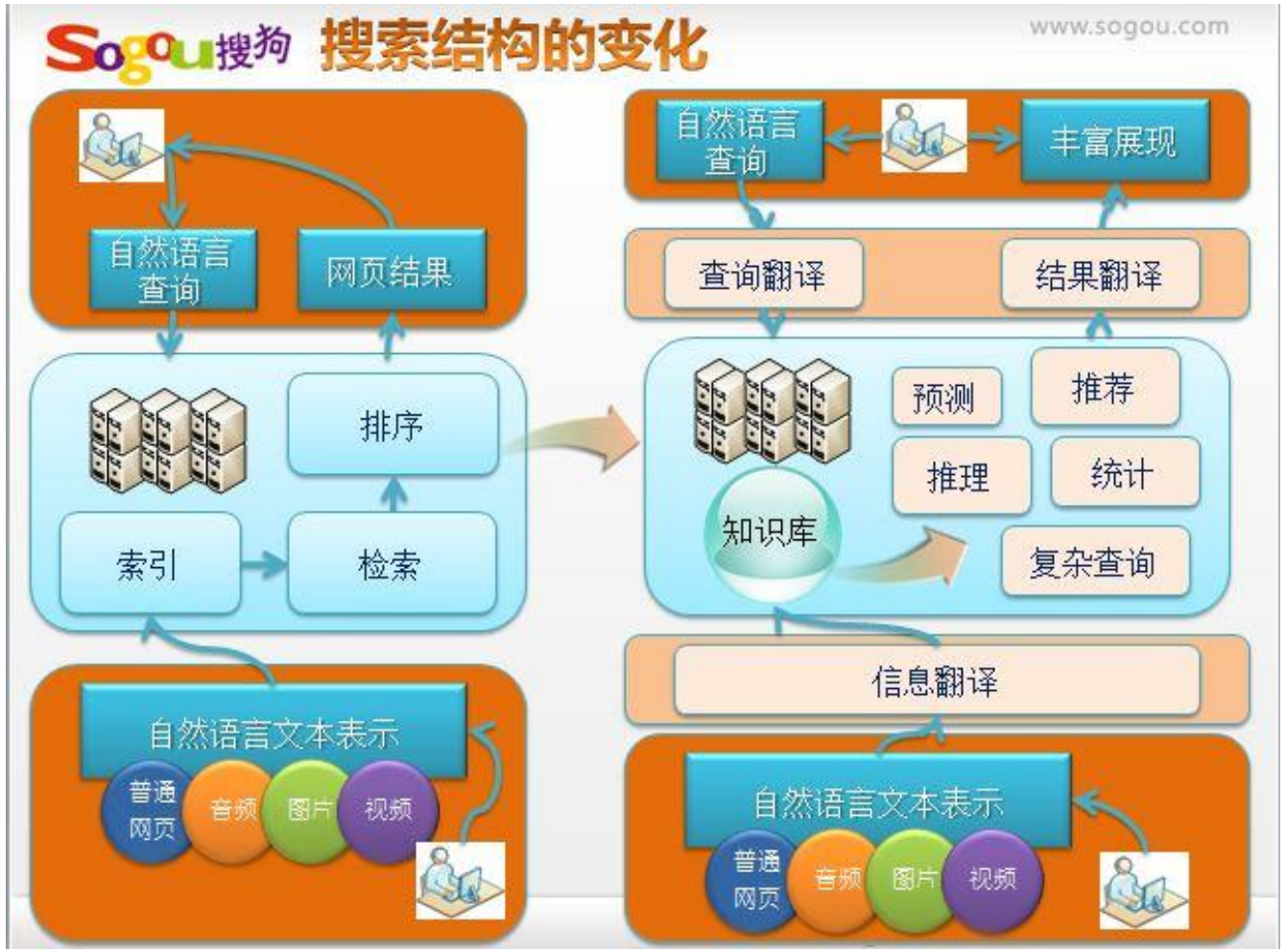
智能搜索的概念及发展--总论

表1　搜索引擎发展阶段

发展阶段	特点	定位
传统搜索	广泛采集信息，关键词匹配，简单结果呈现	网络世界的重要入口、导航，本身不提供知识
垂直搜索	纵向垂直深度搜索，信息聚合展示	领域垂直细分，面向主题，提供更专业更精细的服务
智能搜索	知识搜索、语义分析、语境分析	智能洞悉用户需求，解决用户问题而不只是信息查找

- 智能搜索引擎，主要通过自然语言处理和知识图谱等人工智能技术，来实现人工智能在搜索引擎产品的落地。它更注重与其他科学相融合、个性化搜索、智能化比较高。
- 智能搜索引擎比传统搜索引擎具有更多的优点。首先是更加的智能化，其在搜索内容时会自动筛选到所收集的信息和与此相关的信息，根据相关的信息系统进行自动整理，并且还可以把各种信息综合到一个数据集当中。其次系统会根据用户输入条件，推荐给用户感兴趣的信息。而且它具有语义分析、搜索补全、智能关联等功能。另外，智能搜索引擎得到的结果，不是孤立的文本信息，而是关联各个实体的多媒体化的有用信息。

智能搜索的概念及发展--搜索引擎范例



- Knowledge Graph从以下三方面提升Google搜索效果：

1、找到最想要的信息。

2、提供最全面的摘要。

3、让搜索更有深度和广度。

智能搜索的概念及发展--引擎框架

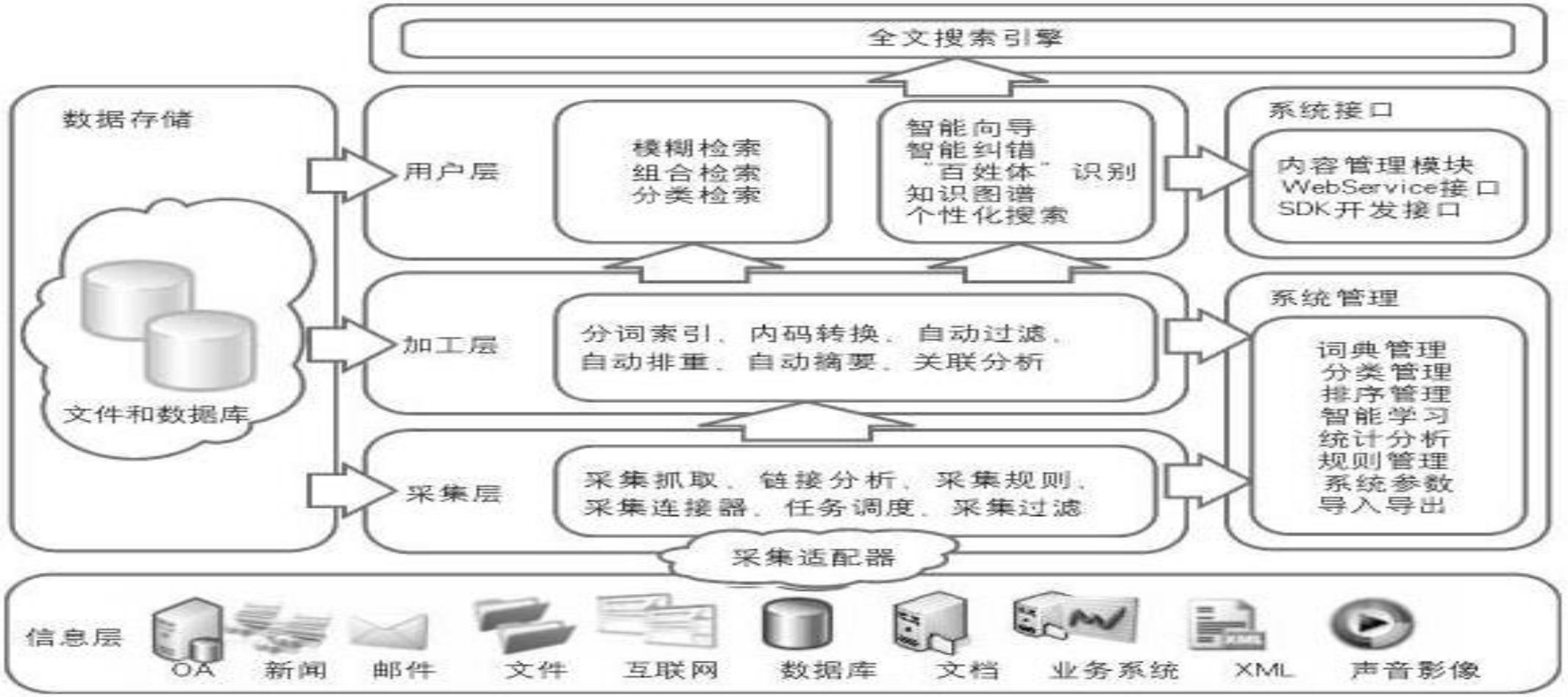
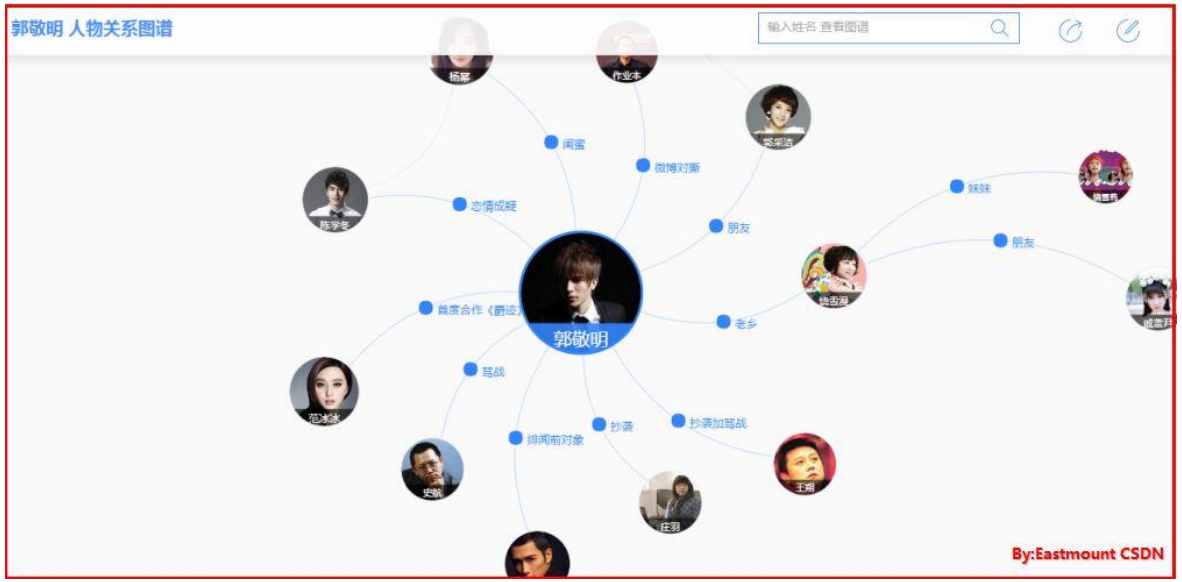
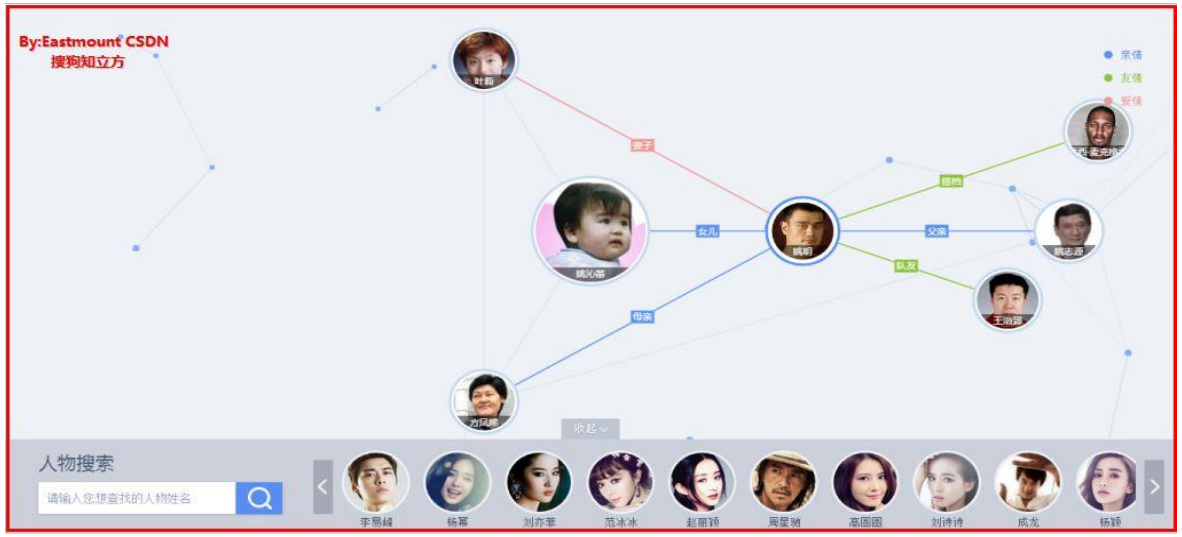


图1　智能搜索引擎系统整体框架

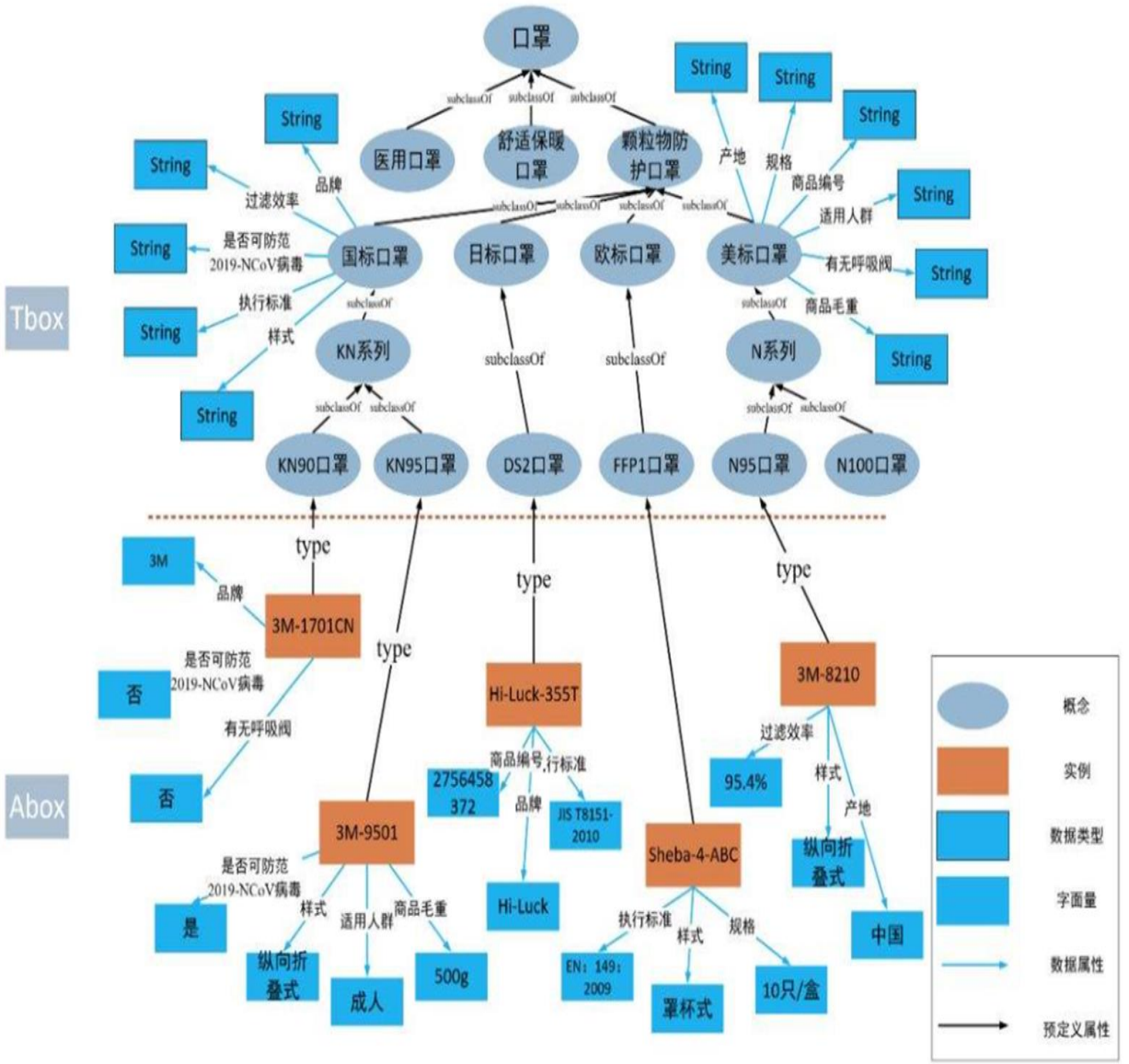
知识图谱的概念及发展—总论





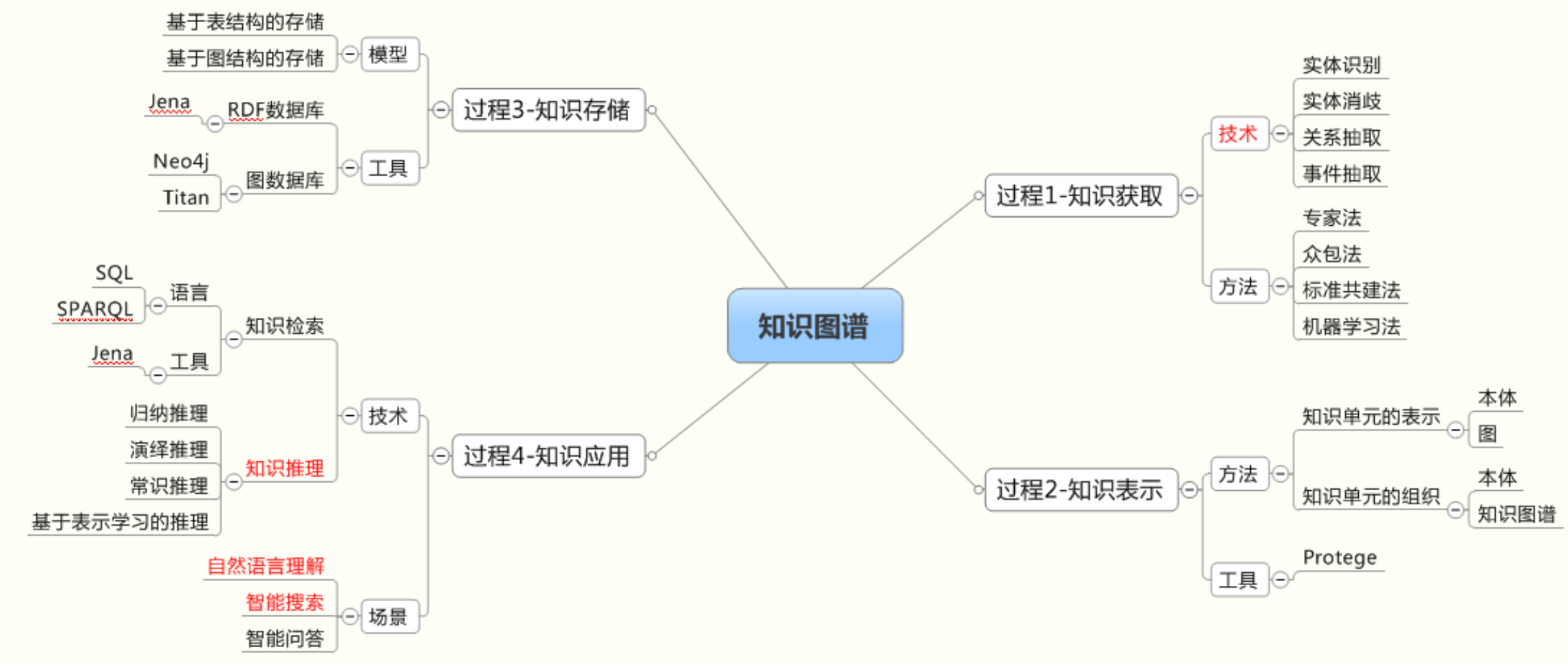
- 知识图谱于2012年5月正式被提出，其目标在于改善搜索结果，描述真实世界中存在的各种实体和概念，以及这些实体、概念之间的关联关系。从而改善搜索结果。紧随其后，国内搜狗提出了“知立方”、微软的Pro-base和百度的“知心”。
- 具体来说，知识图谱是通过将应用数学、图形学、信息可视化技术、信息科学等学科的理论与方法与计量学引文分析、共现分析等方法结合，并利用可视化的图谱形象地展示学科的核心结构、发展历史、前沿领域以及整体知识架构达到多学科融合目的的现代理论。

知识图谱的概念及发展—理解知识图谱1



知识图谱的概念及发展—理解知识图谱2

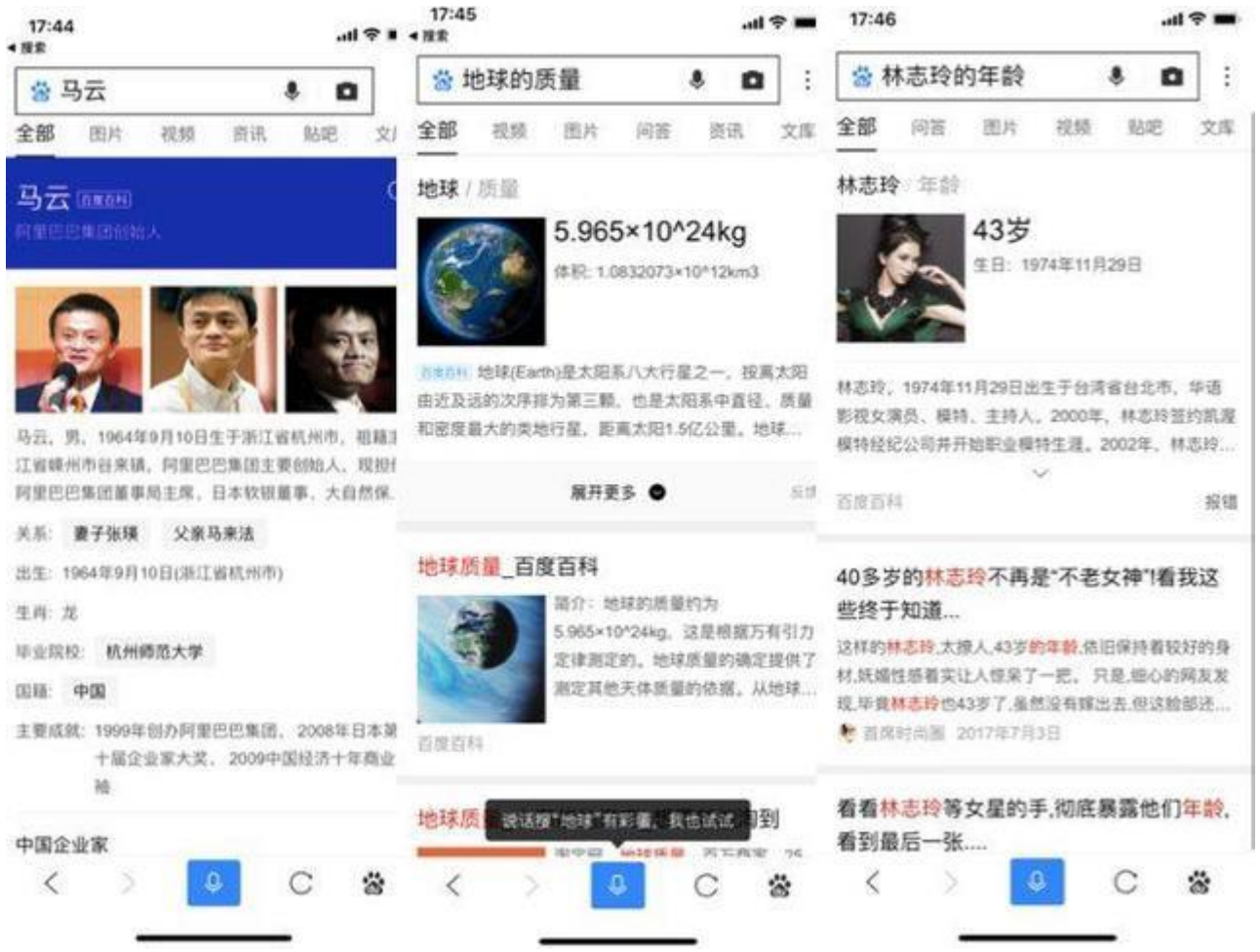
知识图谱的概念及发展—知识图谱体系



智能搜索与知识图谱—智能化搜索的发展1

- 几年前，谷歌发布了一个全新的书籍搜索产品：“Talk to Books”，用户可以通过对话的方式得到一本书籍的推荐，比如输入：“What is the best programming language？”（什么是最好的编程语言？），就会被推荐《C Programming for Arduino》。这个产品是典型的知识图谱技术的应用，它让搜索引擎可以理解用户的问题和每一本书的内容，进而进行精准匹配——就像有人在豆瓣给你荐书一样。

智能搜索与知识图谱—智能化搜索的发展2



- 相对于五年前十年前，搜索引擎更能理解你的意图——不论是自然语言、关键词、语音还是图片，都可以揣摩到你想要找什么内容的意图，同时更加智能地整合更适合的结果到一个页面上。
- 信息的聚合似乎还不能让用户感知到搜索引擎的“智能”，顶多是“丰富”。如果你搜索“太阳的质量”、“2的五次方等于多少”、“形容大海的成语”、“成龙的老婆是谁”，就会发现百度可以精准地理解你的问题，再给你个性化的结果，它不只是可以理解一段文字。

智能搜索与知识图谱—研究应用及发展方向

- 知识图谱是结构化储存人类知识的数据库，在信息检索、问答系统和智能对话等人工智能相关任务中广为运用。知识图谱可以看成是一张巨大的图，图中的节点表示实体或概念，而图中的边则构成关系。
- 运用了知识图谱的智能搜索引擎，则可以返回更加精准的结果。就知识图谱而言，研究语义分析方面至关重要。知识库的构建和知识搜索都需要语义分析。
- 知识图谱的研究和应用，可分为基于广度的通用知识图谱和基于深度的领域知识图谱。
- 搜索+知识图谱，是未来搜索引擎的发展方向，未来的搜索引擎，会以用户为核心，越来越智能化。
- 例如：日本近两年最新的医用机器人技术

智能搜索与知识图谱—公安领域论文分析

- **摘要：**为解决当前公共安全领域数据分布不均匀、结构复杂,大量数据之间存在信息壁垒,知识共享无法形成,公安知识不能被有效利用等问题,提出了一种基于公安领域知识图谱的构建方案。该方案采用知识建模、知识抽取、知识融合等核心技术对海量多源异构公安数据进行图谱构建,并在已构建的公安知识图谱基础上实现了家族族谱、知识智能搜索,并且具备亿级实体关系事件存储管理能力。因此,该方案对公安领域知识图谱的构建具有指导性的意义。

知识图谱在公安领域的构建与应用研究　周鹏;胡运涛;苗世茂　第八届中国指挥控制大会论文集 中国会议 2020

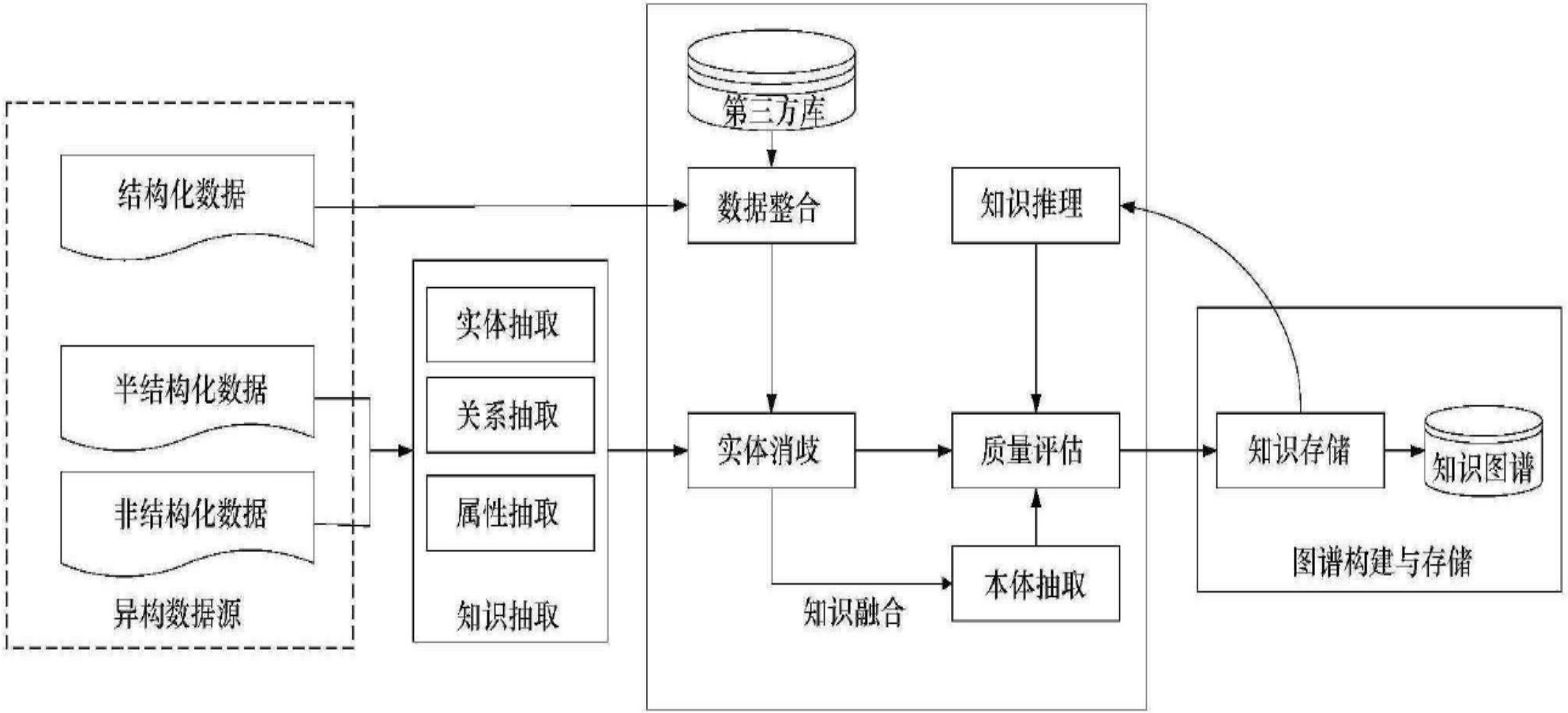


图 2　知识图谱构建技术流程

知识图谱在公安领域的构建与应用研究　周鹏;胡运涛;苗世茂　第八届中国指挥控制大会论文集 中国会议 2020

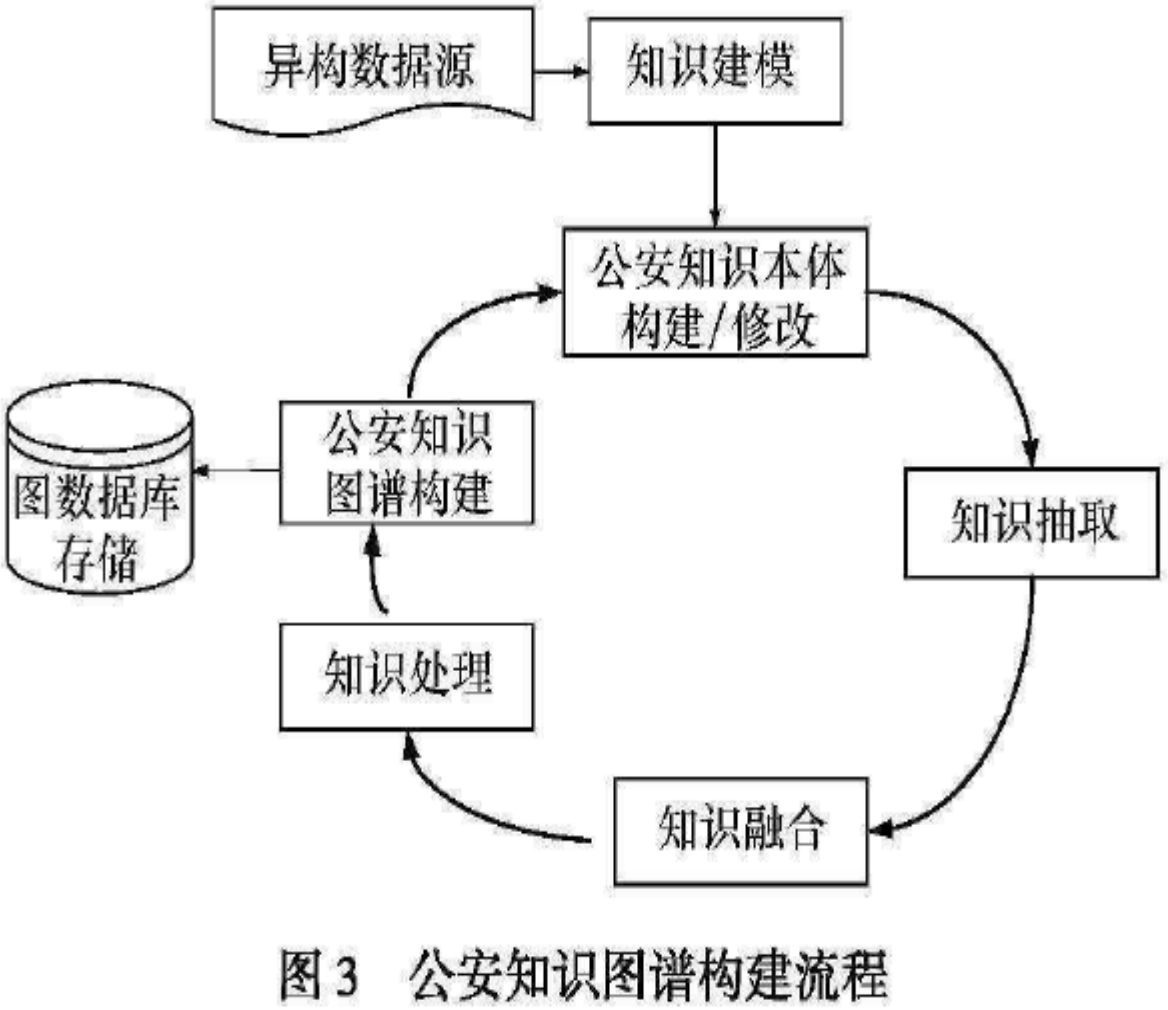


图 3　公安知识图谱构建流程

- 基于公安领域知识图谱的构建
- 信息的来源分析：公安内网数据，互联网信息。
- 该流程首先通过库属于识、公识过、谱修系。建构备的取，等求入本知及一模必件抽息理要存建成以了建定事库信处域并构形建成识确及体素识领谱从合构形知来性本要知安图程融的正公实义安通程三数库库后成熟

知识图谱在公安领域的构建与应用研究　周鹏;胡运涛;苗世茂　第八届中国指挥控制大会论文集 中国会议 2020

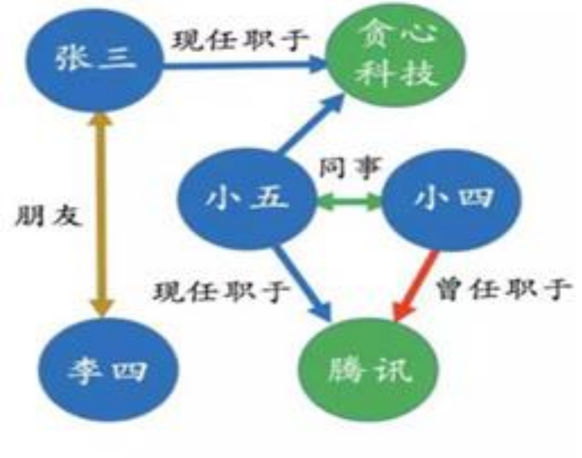
知识的抽取和融合是构建公安知识图谱的两个关键阶段。

1）知识抽取：从海量多源异构数据中进行实体、属性及关系提取的过程。比如，某新闻报道中“昨日晚10点，25岁的李女士在……路与47岁的张先生发生了一起追尾事故……所幸未造成人员伤亡”。该非结构化文本提供的直接信息中，可提取出两个实体人“李女士”和“张先生”以及“汽车追尾”的相互关系。若进一步关联，还可抽取出两者的实体车及“车与实体人车主所属”的关联关系，当然也可抽取出“25岁”和“47岁”描述实体人特征的属性信息。知识抽取依赖于实体识别技术。利用基于上下文的实体识别等NLP算法，并结合机器学习、规则引擎，可以准确识别人名、地名、组织机构名、身份证号、手机号等。其中基于CRF模型的人名和机构名识别，通过对训练语料进行数据加强，提供人名和机构名的识别召回率。基于实体词典和上下文概率的地名和外文名识别提高了实体抽取的精确率。

2）知识融合：简而言之就是将多个来源差异的知识整合成一个知识库，该过程就需依赖实体消歧、共指消解、实体对齐、实体属性值判定等手段来消除这种差异性。实体消歧指的是同样的词汇可能表示不同的实体，就需要基于语义从上下文中推测出哪个词代表哪个实体。比如，两个同名为张三的人可以根据其身份证或者年龄属性来判别；共指消解顾名思义就是同一个代词指代哪一个对象，也可以通过上下文语义来进行推测。实体对齐指的是如何将同一实体的不同描述指向同一实体，比如将“齐天大圣”和“美猴王”映射到“孙悟空”；实体的属性值判定指同一个实体从多个来源得到的属性描述差异就需要知识源的置信度来确定。实体相似性分析处理可利用快速相似度、规则计算以及分治的实体融合技术，基于覆盖、多值溯源策略有效地解决了结构化数据知识融合问题；基于多规的实体融合技术实现了身份ID、手机MAC地址、手机号、IMEI等多类实体和虚拟身份ID的关联融合。

表 1　知识写入图库性能指标分析

场　　景	数据量	线程	平均速度
单机多线程导入实体	2396019	64	43439/s
单机多线程更新实体	2396019	64	21078/s
单机多线程导入关系	200 万	64	20986/s
单机多线程更新关系	200 万	64	10984/s
单机多线程导入事件	480 万	64	28537/s



案例：社交网络

- 经过一系列的信息处理流程之后，碎片化的知识被整理成有序的知识网络，就可以存入图数据库了。公安知识图谱涉及节点数量众多且关系复杂，随着公安知识库地不断更新，知识图谱也需要相应地更新，为了保证新知识的入图不影响原有知识的准确性，同时也考虑到知识入图的效率问题。本文中使用的图数据库底层采用Janus Graph开发，整合但有别于传统的单机模式的图数据库，采用Hbase、Elastic search及Phoenix二级索引于一体的优化混合存储技术，且存储层和计算层都使用分布式架构设计，这样一来使得图库具有很好的横向扩展性，可轻松通过增加节点的方式提升系统存储和计算性能。目前使用的图库具有存储数十亿实体、数百亿关系的能力。
- 由此，形成公安知识图谱。从中可以通过智能搜索得到个人信息、同事朋友关系、所在单位，以及以前工作单位等有用的关联信息，有助于户籍管理和案件处理。

知识图谱在公安领域的构建与应用研究　周鹏;胡运涛;苗世茂　第八届中国指挥控制大会论文集 中国会议 2020

公安知识图谱的应用：家族图谱的智能搜索

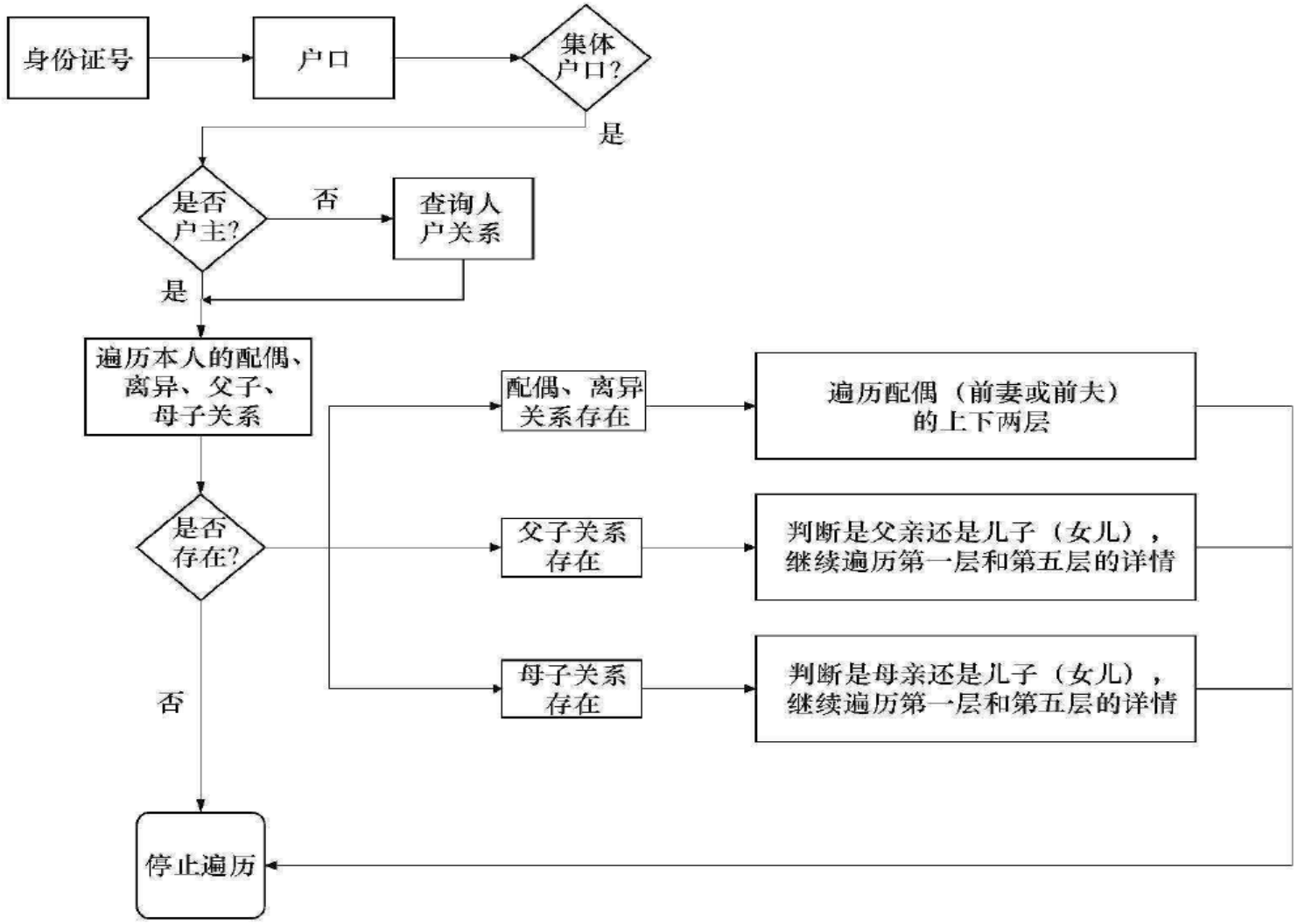


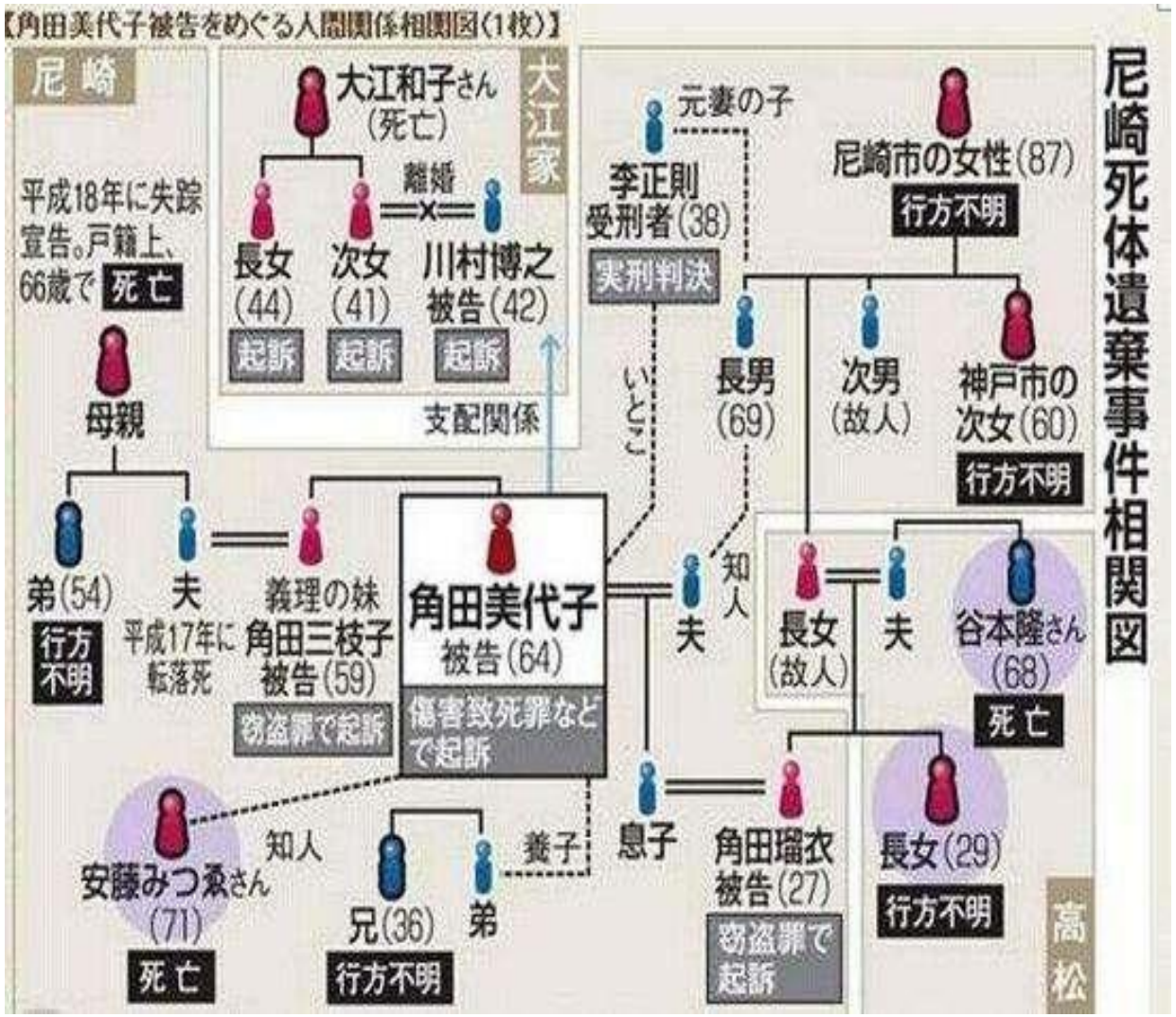
图 5　家族族谱检索关系扩展流程

如图5所示：根据输人的身份证号码，通过人户关系出查询户号（ 因迁移可能会出现 2 个或者以上的户号） 及与户主关系，针对户口的情形进行判断：本人是户主；直接查询本人的配偶、离异、父子、母子关系；否则可通过人户关系查询户主，继续做本人关系查询判断。在展示时显示上下两层关系，即以被查询人及配偶为核心，向上扩展两层，向下扩展两层，一共5 层，被查询人号码（本人）处于中间层（第三层）。第一层（ 上二层）为祖辈，第二层（ 上一层）为父辈，第三层（中间层） 为本人、配偶及兄弟等，第四层（下一层）为子辈，第五层（下二层）为孙辈。

最终结果将以知识卡片形式展示，并且每个实体，都可以通过点击得到个人详细信息和关联信息。

知识图谱在公安领域的构建与应用研究　周鹏;胡运涛;苗世茂　第八届中国指挥控制大会论文集 中国会议 2020

• 公安知识图谱的应用：家族图谱的智能搜索帮助案件处理。



知识图谱在公安领域的构建与应用研究　周鹏;胡运涛;苗世茂　第八届中国指挥控制大会论文集 中国会议 2020

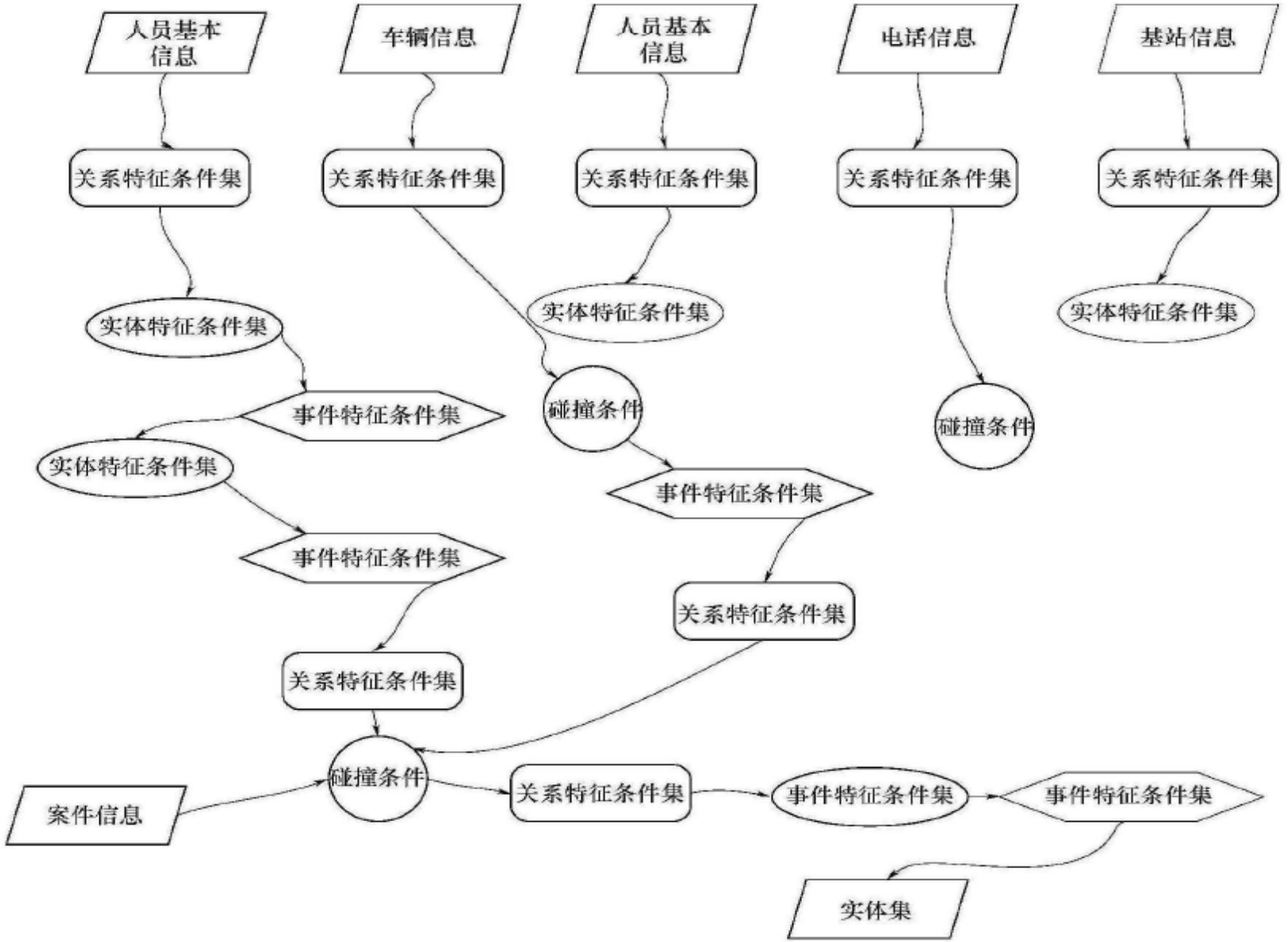


图 7　知识库检索和碰撞过程

公安智能知识检索：将一个查询分解成不同类型子查询进行分布式处理，更快速更智能地协助公安人员进行智能全息布控、重点人员管控、行为轨迹追踪（图7)。

知识图谱在公安领域的构建与应用研究　周鹏;胡运涛;苗世茂　第八届中国指挥控制大会论文集 中国会议 2020

万物互联的时代，更希望看到智能化的数据智能万联，可查即可得！

谢谢！