doi:10.3969/j.issn.1000-1247.2017.06.001

基于自然语言处理和图计算的情报分析研究

杨明川 胡 婕 杨哲超
中国电信股份有限公司北京研究院

摘要 通过自然语言处理技术，可以将海量情报信息中的实体进行结构化提取，并通过图计算的方式进行关联分析，从而为情报部门快速侦破案情提供线索帮助。文中论证研究了通过使用知识表示、基于长短时记忆神经网络的命名实体识别、图数据库等技术，针对情报数据进行信息提取、信息过滤、情报知识库建立，进行关联挖掘和分析。

关键词 情报分析 自然语言 处理图 计算 搜索引擎

1 引言

2016年是人工智能技术发展的重要一年，被产业界称为人工智能元年。随着像自然语言处理类的人工智能技术逐渐成熟，人工智能技术在情报分析研究领域的作用和价值日益凸显。面对情报部门每天产生的大量情报类的非结构化数据以及互联网上的海量文本信息，能否快速有效地提取出相关信息成为办案人员能否快速破案的决定性因素。对于传统的情报分析方法，主要存在三点问题。

(1)单一情报信息中可能蕴含着大量潜在线索，需要比对判断大量相关信息才能够得到。

(2)情报信息往往涉密，对于办案人员要求较高，很难通过劳动力外包的形式缓解案件激增的情况。

(3)情报分析工作对时效性要求往往较强。在情报分析领域，机器优势和人类专家优势的对比分析如图1所示。

2 基于自然语言处理技术的知识表示

情报关联分析任务作为一个复杂、融合的系统，其数据覆盖的深度与广度决定了其发现信息的价值。为了提高情报分析的可靠性，提升情报分析精准度，需要收集和处理大量翔实的情报资源。

情报分析的数据资源往往以文本数据的形式存在，如文献资料、网页信息、专利资源、政府公文，以及图片、视频文件等这样的非结构化数据。为了提高情报分析的真实性及准确性，往往要求情报分析人员收集全面的数据和信息，并从信息中整理、提取、加工得到有价值的情报。

事实上，传统的情报资源收集依赖于情报工作人员的智力加工。严重依赖于人工的方案在应对数据量庞大、变化迅速、类型多样且价值非常稀疏的数据时，往往无法有效收集、提取有价值的线索与信息。对于体量巨大的情报资源数据，避免依赖人工的情报分析方案，是提升情报分析工作效率的关键所在。

与此同时，传统情报分析的做法是将半结构化数据或非结构化数据转化为结构化数据再进行处理，或是通过人工分析非结构化数据。这一过程可能导致丢失非结构化数据中隐含的关系，进而导致分析结果的不确定性。非结构化数据中可能具有的价值点的发现，以及价值点的有效关联关系的建立，是提升情报分析工作价值的关键所在。

自然语言处理（Natural Language Processing）技术属于人工智能与语言学领域的交叉范畴，被用于对自然语言语料进行处理，目的是让计算机“理解”自然语言的内容，即把自然语言的数据内容使用计算机内部的机制表示出来，并进一步进行计算与处理。常见的自然语言处理任务包括句法分析、语音识别、文本分类、信息抽取、问答系统、机器翻译等。在情报分析研究任务中，通常涉及到自然语言构成的非结构化文本，包括语音记录、人员档案、事件描述、人员关系图谱等。使用自然语言处理技术对这些非结构化的数据进行处理，是发现价值信息的重要手段。

价值信息中，一个非常重要的内容为情报资料中关键人物或者机构的发现与提取。命名实体是一种标识某一个概念或者实体的特殊短语，在情报分析领域中，包括且不限于专有名词、人名、地名、公司名、组织机构名、过失罪责等。命名实体通常比词语和一般短语表达的信息更加精确。

命名实体识别（Named Entity Recognition，NER）任务是为了让计算机能够辨识出文本当中的实体。NER任务是很多应用的基础组件，比如信息抽取、问答系统等。

对于电脑而言，自然语言是输入的一串符号序列，而自然语言的形式，包括声音、文本等，都没有具体的含义。计算机的任务只是接收输入的数据，计算与处理，然后输出。如图2所示，对应自然语言形式输入的文本序列，命名实体的标注结果也是一组对应位置的序列。所以命名实体识别的任务可以被看作是序列数据标注问题。在该领域的算法模型主要有以条件随机场模型为例的统计自然语言模型，以及以长短时记忆神经网络为代表的深度学习模型两种类型。

命名实体识别用在情报分析系统领域，可以看作是一种将非结构化数据转换为结构化数据的方法。在情报分析任务中，命名实体识别任务的流程可以理解为：首先，输入一系列序列数据，计算机将其解读为一些内部的标识符号，例如，把每一个汉字标注1、2、3、4的表示形式；然后通过计算，找到某些上下文的字符形成命名实体（如地名、人名）的规律。当新的数据输入时，计算机能够根据之前在训练文本中获得的经验，预测新的数据中哪些字符的组合构成命名实体。

对于大量的文本资料，由于其数据格式不固定，同时作者的写作水平、文化背景、表述习惯和逻辑能力的区别，如果依赖于情报人员的阅读和判断，往往需要较多的人力投入，且效率低下。通过机器快速的计算和处理，从中识别出情报任务关注的关键信息，是一个较为理想的选择。

传统的中文命名实体识别任务需要大量的手工特征和指定领域的知识来达到较好的表现。传统的方法包括最大熵模型、隐马尔可夫模型、支持向量机模型和条件随机场模型，可以分成生成模型和判别模型两大类。

生成式模型的思想为观测无穷多的样本，生成概率密度模型，然后进行预测。生成模型以隐马尔可夫模型为代表，这个模型通过学习给成对的观测变量和标签序列赋一个联合概率。这一参数通过在训练集上最大化联合似然函数得到。隐马尔可夫模型最大的缺陷在于不支持长距离的依赖。

判别式模型的思想为观察有限样本，学习得到判别函数，然后基于判别函数来进行预测。判别模型以条件随机场模型为例，支持观测序列中任意的依赖关系，具有更优的表达能力，同时可以支持数据量较少的训练集。

作为序列数据，能够识别到其中长距离的依赖关系对于一个模型来说是至关重要的。神经网络模型特别是多层的神经网络，由于其强大的学习能力，比起传统的方案，已经能够在许多循环神经网络模型更适用于长距离的依赖关系的识别，非常适合于序列数据的处理操作。近年来，长短时神经网络在自然语言处理领域得到广泛应用，主要原因在于这个模型非常适用于处理序列数据，并支持长距离的依赖。长短时记忆神经网络模型的简单示意如图3所示。

具体操作为：首先输入数据通过一次softmax操作，先通过忘记门f_t的操作，决定是否考虑长距离的依赖关系，即哪些信息被丢弃，哪些信息被保留；然后，通过输入门i_t的操作，更新细胞状态；最后，输出门o_t的计算结合细胞状态和输入数据，计算输出至隐藏层h_t，并作为下一次更新状态的输入。

长短时记忆神经网络通过主动选择传入的信息，来避免由于储存过多历史信息带来的计算问题，以支持长距离的依赖关系建模。然而在实际应用中，通过对网络结构设计的优化以及对网络输出结果的再次加工，能够有效提升序列标注任务的准确度。如图4所示，对于一个长短时记忆神经网络，进行前向与后向两边扫描的改进，并对网络输出结果加入一个条件随机场再处理，求得标注的全局最优。

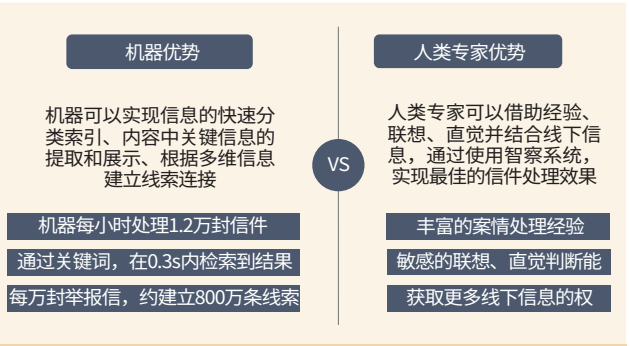

图1　机器优势和人类专家优势对比分析

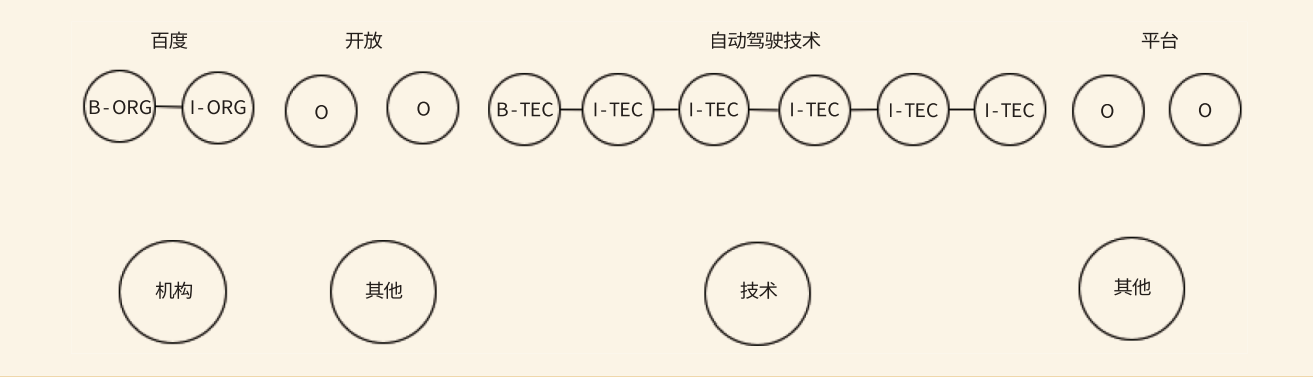

图2　命名实体识别示例

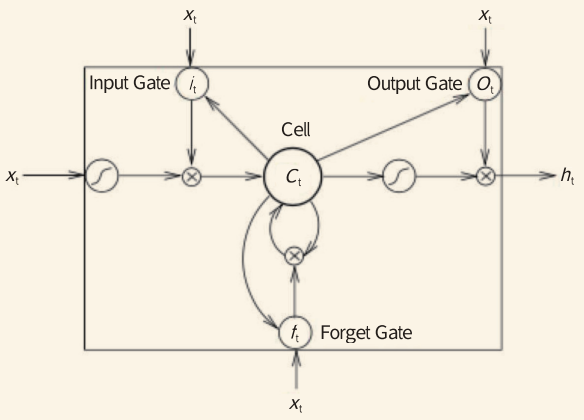

图3 LSTM网络模型细胞

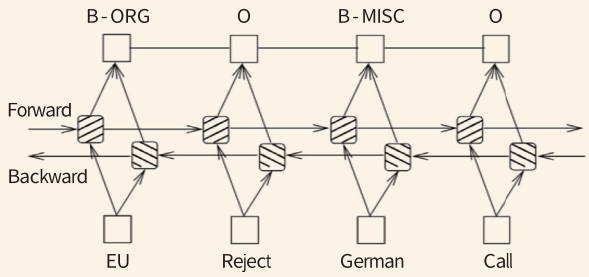

图4 长短时记忆神经网络

3 基于图计算的情报分析研究

随着关系数据库模型在非结构化表格数据上的不适应性逐渐凸显，一些NoSQL数据库不再将固定的表格模式作为存储的结构，并且避免表格的Join操作，以获得一些对特殊格式数据的存储查询更优的效果。

图数据是NoSQL的一种，主要支持图形结构的数据存储、查询与计算功能。其数据模型以节点和关系来体现，也可用于处理键值对。图数据库可以用来快速解决图结构数据上复杂关系的查询与计算问题。

情报分析是一个涉及到多个实体、多重联系的任务。在通过自然语言处理技术提取到文本中的关键信息之后，可以将这些信息作为一些关键信息点。实际的情报处理任务中，发现这些信息的关联是更加具有价值的工作。针对线索与人员、机构建立图结构关联关系，甚至对于企业等组织建立信息图谱，对于大量价值稀疏的数据处理任务都是有着巨大帮助的处理方案。

间接关系发现是图数据库关于线索发现问题的一大应用场景。图数据库可以很好地适应关联数据管理的场景。在关系数据库中，任何超出直接关联关系的浅遍历查询，都将因为涉及的索引数量而变得缓慢。而图数据库使用的是图遍历的方法，故所需的计算量远小于关系型数据库，可以达到非常快的查找速度。

除此以外，图数据库可以用于寻找两个节点之间的最短通路。

综上所述，虽然关系型数据库对于保存结构化数据来说依然是最佳的选择，但图数据库更适合于管理半结构化数据、非结构化数据以及图形数据。如果数据模型中包含大量的关联数据，并且希望使用一种直观、有趣并且快速的数据库进行开发，那么可以考虑尝试图数据库。

4 结束语

IDC统计结果显示，非结构化数据在现有存储系统中所占的比例已接近80%。与此同时，目前互联网上传播的数据中，超过90%为非结构化数据。

生活中人们每一天的日常活动，如阅读文章、观看视频、沟通交流，都在接收与产生非结构化数据。

这些数据包含价值潜力巨大的信息。事实上，人工标注的力量相对于这些数据的体量与价值而言，作用十分微小。在现今很多数据驱动的任务中，数据来源往往依赖人工提取，然后通过特征工程完成特征选择。

对于复杂、融合的系统而言，转化非结构化的数据为结构化数据，并发现其中价值的算法，是拿到大量数据真正价值的钥匙。

大量的非结构的、蕴含着大量价值的数据正在每时每刻地生成，而人工的力量并不能胜任标注这些数据集的重任。

如对本文内容有任何观点或评论，请发E-mail至ttm@bjxintong.com.cn。

作者简介

杨明川

现任中国电信股份有限公司北京研究院副总工程师，云计算与大数据研发事业部总监，参与多项国家重大专项研究，获得多项省部级科技奖，入选国家“863”计划专家库。

胡婕

北京邮电大学硕士，现就职于中国电信集团北京研究院人工智能团队，算法工程师，从事自然语言处理与图计算相关工作。

杨哲超

北京邮电大学工商管理硕士，现就职于中国电信股份有限公司北京研究院人工智能团队，具备多年技术领域项目管理经验，熟悉项目管理流程和风险控制，曾就职于微软（中国）有限公司，参与过Windows 7、8、8.1等全球性软件产品的研发，曾获得微软认证解决方案开发专家（MCSD），微软亚太研发集团Engineering Award，中国电信研究院2016年度青年才俊称号。