DOI:10.13530/j.cnki.jlis.190021

融合知识组织与认知计算的新一代开放知识服务架构探析*

孙 坦 刘 峥 崔运鹏 鲜国建 黄永文

摘 要 信息过载现象导致发现并获取有用信息变得越来越困难,用户急切需要精准的知识发现和问题解答服务,通过对当前实现精准知识发现的主要技术方法进行分析,本研究分别采用基于传统知识组织构建方法和基于深度学习的方法,面向湿地领域进行语义知识组织体系的构建和精准发现实验。实验分析证明,传统知识组织方法无法单独支持特定主题的精准知识发现,尽管基于词向量的深度学习方法可以有效弥补传统知识组织系统的局限,但会受到计算语料规模和质量的限制。因此,本研究最终提出融合知识组织与认知计算的基本思路和体系框架,分析了融合方案所涉及的关键技术,这对于构建新一代开放知识服务系统具有重要指导意义。图 3。参考文献 15。

关键词 知识组织 认知计算 知识应用 机器智能

分类号 G254

Analysis and Design of A New Generation of Open Knowledge Service System Integrating Knowledge Organization and Cognitive Computing

SUN Tan , LIU Zheng, CUI Yunpeng, XIAN Guojian & HUANG Yongwen

ABSTRACT

Information explosion has made it difficult to acquire useful knowledge, and users are more likely to expect precise results or answers directly. This paper provides a survey to gain an overall view of the current research status in precise knowledge discovering focusing on methodologies, technologies, and most distinctive characteristics, and proposes possible solutions with regards to some existing problems and future research in this area.

The authors selected the field of wetland to make tests on semantic knowledge base construction and precise knowledge discovery. Two main approaches are adopted and compared experimentally: one is traditional Knowledge Organization System (KOS), and the other is deep learning.

* 本文系国家科技图书文献中心专项“下一代开放知识服务平台总体设计及关键技术研发”和中国农业科学院科技创新工程项目“大规模 RDF 数据转换机理与存储研究”(编号:CAAS-ASTIP-2016-AII)的研究成果之一。(This article is an outcome of the project “Design and Research on A Next Generation of Open Knowledge Services System and Key Technologies” supported by Special Project of National Science and Technology Library and the project “Research on Large-scale RDF Data Conversion Mechanism and Storage” (No. CAAS-ASTIP-2016-AII) supported by Science and Technology Innovation Project of Chinese Academy of Agricultural Sciences.)

通信作者:孙坦,Email:suntan@caas.cn,ORCID:0000-0002-8257-5064 (Correspondence should be addressed to SUN Tan, Email: suntan@caas.cn, ORCID: 0000-0002-8257-5064)

The results indicate that traditional knowledge organization technique cannot support the precise knowledge of specific topics, while deep learning based on word vector can make some improvements on it, though the effect is limited by the scale and quality of computational corpus. Then the authors propose the framework and key technologies of integrated approach of KOS and cognitive computing.

The comparison and integration between new feature words obtained by deep learning and existing KOS is weak in this paper, and the verification of precise knowledge discovery should be made in practical application scenarios.

The results show that the integrated approach of KOS and cognitive computing can be expected to improve the precise knowledge discovery, and achieve high efficiency of knowledge calculation and high accuracy of intelligent question & answer.

As far as the authors know, this is the first study presenting the idea of a new generation of open knowledge service system. It shows the capabilities and limitations of current research in precise knowledge discovery and provides directions to researchers by pointing out that the integration of KOS and cognitive computing will effectively improve the low precision of machine algorithms led by lacking high-quality big data and semantic knowledge base. In addition, the proposed approach may break through the deep cross-boundary fusion and automatic knowledge acquisition of multi-source heterogeneous data, and then make great improvements on transformations from unstructured literature data to structured and semantic knowledge networks. 3 figs. 15 refs.

KEY WORDS

Knowledge organization. Cognitive computing. Knowledge application. Machine intelligence.

0 引言

数字化、网络化已经走过20余年发展历程,特别是数字出版和开放获取极大地催生了数字科技文献的爆炸式增长与快速传播[1-3]。但是,随着信息获取越来越方便,信息过载现象也有愈演愈烈之势,发现并获取有用信息变得更加困难。受限于数字知识表示与检索技术,海量信息资源与碎片化信息需求之间的矛盾已成为制约有效信息选择与知识获取的首要障碍。为克服这个障碍,实现知识互联的语义网技术被寄予厚望,从信息描述到信息抽取,从元数据到关联数据,从叙词表到本体、知识图谱,一系列语义表示、抽取、描述和组织的技术方法得以发展,但大规模语义知识组织体系的构建仍是亟待解决的难题。

近年来,机器智能时代乘着深度学习、认知计算之舟踏浪而来,推动知识组织、知识分类、探索式知识发现、基于语义分析的自动问答等智能知识应用迎来了颠覆式创新时代,人工智能技术将重新定义我们曾熟悉的知识组织和知识应用。作为优质可控的知识源,图书馆采集的各类文献无疑是智能知识应用最重要的数据源与知识供给。因此,如何利用深度学习、认知计算等技术将科技文献转换为可计算的知识,实现知识融合并支撑智能化的知识应用,已成为新一代开放知识服务系统必须解决的瓶颈问题。

1 构建语义知识组织与精准知识发现的主要技术方法

Jakus等[4]认为,随着互联网和大数据的不断发展,海量数据的指数增长导致人们利用计算机查找专门数据和有用信息的能力迅速降低,形成开放信息环境中海量异构信息与用户

碎片化、个性化需求之间的矛盾:如何从海量数据源中快速有效提取所需数据?如何找到那些代表着问题答案或有助于解决问题的、必要和潜在可用却又未知的信息片段?如何集成那些潜在的信息片段?其主要瓶颈有:①搜索引擎的算法主要依赖关键词匹配,而不考虑关键词的各种不同含义及其上下文语境;②多数知识呈现出非结构化或异构化形态,致使其不能被进一步处理。笔者将上述矛盾转换为3个科学问题:精准知识发现、探索式信息分析和基于协同推理的知识计算。结合传统自然语言处理和人工构建的知识组织是前机器智能时代解决该问题的主要技术方法,基于神经网络算法的深度学习则是机器智能时代的主流技术方法。然而,在当前阶段,上述两种技术方法均存在不足与挑战。

本文所指知识组织,包括了同义词环、分类表、叙词表、本体、知识图谱等。在不同时期,受技术发展局限,知识组织的形成方式和所具有的功能不同。在1.1中知识组织指的是分类表、叙词表、本体,一般由人工构建,本文称为传统知识组织方法;在1.2中知识组织主要指的是知识图谱,一般由机器自动或半自动构建。

1.1 结合传统知识组织和自然语言处理技术的实验分析

缺乏统一数字知识表示和语义缺失是实现精准知识发现、知识关联分析与计算的首要障碍。传统知识组织方法和自然语言处理技术可以利用词表、本体等构建语义知识模型,进行统一知识表示与语义关联,通过基于实体识别与信息抽取的句法分析、位置分析、特征分析等自然语言处理技术辅助增强语义。该技术路线的一个典型商业应用案例是Springer Nature发布的SciGraph,其依托于nature.com语义出版平台提供的语义模型(核心本体和领域本体)以及RDF数据集(期刊、论文、作者等),支持用户进行语义搜索和分面呈现与关联[5]。这无疑是在文献元数据层面上一次成功的创新与进步,但是对于解决精准知识发现、探索式分析和知识计算仍然是不够的。笔者选择湿地领域设计了下列实验,旨在测试和发现面向综合及交叉学科领域,传统知识组织体系在精准知识发现方面的作用、不足与挑战。

1.1.1 现有知识组织体系的复用

(1)实验设计

收集现有湿地领域的词表、辞典、百科、分类等知识组织系统,形成一个湿地领域相关术语数据集;从Web of Science发表的文献中筛选出湿地数据集合;探索现有湿地知识组织系统的术语对湿地文献的覆盖率。

湿地数据集:湿地的定义和分类没有统一的标准,表现形式各异,笔者根据《湿地公约》和《GB/T 24708-2009 湿地分类》以及中国实际情况,梳理出表示湿地含义的词汇,通过检索发现,其中9个代表湿地类型("wetland、peatland、bog、swamp、mangrove、marsh、mire、fen、everglades"),约占湿地文献总数的80%,故以此为基础来形成湿地科学数据库。时间段选取1900—2014年,从WoS中共检索到9种湿地类型文献82 684篇,经人工判读确定66 832篇为湿地文献。

实验所采用的基础知识组织体系为STKOS超级科技词表,STKOS[6]是一个融合术语表、叙词表等各种知识组织素材,以科技术语为基本单元,以概念为核心,以来源词表的原有关系为依托,通过概念与来源词表术语进行语义关联的词网络。它由基础词库、规范概念集和范畴体系3个层次构成,基础词库涵盖了1 438万个来源科技术语和609万个基础术语,规范概念集选取于201部核心表,形成了覆盖理工农医学科领域的61万个概念,范畴体系具有38个一级大类的等级结构分类体系,能从学科类目的角度对概念进行汇聚和分类。

(2)实验过程

①选择湿地的一种类型,以关键词"mire"或"mires"在WoS上检索1900—2014年间相关湿地论文(1990年之前文章没有作者关键词和文摘)。经过机器和人工判读,获得论文2 451篇,

对作者关键词进行去重处理,提取 4 709 个作为目标对象集合。②分析 STKOS 超级科技词表的范畴类目中与湿地相关的类目,从水文学和水生态学等 51 个类目下获得术语 2 642 个,进一步扩展获取环境保护、环境监测等 152 个类目,获得术语 6 910 个。③将步骤 1 形成的目标对象集合与步骤 2 形成的知识组织体系进行术语比对,匹配 419 个,仅占目标对象集合作者关键词数量的 8.9%,进一步扩大比对范围,用目标对象集合与超级词表 STKOS 里的所有术语(包括概念优选术语 61.5 万和非优选术语 131 万)进行比对,匹配术语 2 151 个,占目标对象集合作者关键词总数的 45.68%。

1.1.2 统计和语言学方法相结合

(1)实验设计

词表构建的基本方法是基于统计模型,利用文献作为语料,通过统计领域术语在整个文献中的词频来识别领域相关性高的领域概念,领域相关度计算的方法有互信息、KF-IDF、C-Value 等。通过语言学的方法进行语料识别获取候选概念,利用词性特征,对候选词进行词干还原、变形等操作,与已有的词库做相似度对比,过滤置信度低的候选概念。以湿地研究的分支领域湿地遥感为对象,来构建湿地遥感知识组织体系,以验证其对知识精准发现的支撑能力。

(2)实验过程

湿地遥感数据集:以 Web of Science 数据库为数据采集源,以主题=wetland or mire or swamp or marsh or bog or peatland or everglades or mangrove or fen 进行检索,时间 1900—2014 年,采用数据库分类体系“遥感”类别筛选出文献 1 571篇。

湿地遥感词表构建:将 1 571 篇论文的标题和作者关键词进行分词处理,获得 5 834 个词块,基于词频统计进行抽取,经过科研人员审核确认了 1 684 个与遥感相关词块,与 STKOS 词表中术语进行比对,原型化处理后进行精准和包含匹配,并由专家进行遴选补充;依据来源词表的专业权重,继承和裁剪了概念关系,并参照《遥感大辞典》设计所属分面,构建 217 个概念、409 个术语的叙词表。

根据用户的真实需求检索湿地遥感的相关文献。用构建的湿地遥感叙词表,检索之前构建的湿地数据集,获得文献 5 493 篇,经过人工判断,最终确定“湿地遥感”研究主题的文献为 3 697篇,其他 1 796 篇文献研究主题为“非湿地遥感。

1.1.3 实验结论

上述两个实验结果表明,在学科交叉领域,基于现有知识组织系统进行专门知识组织体系定制时,与领域文献作者关键词的匹配程度很低,领域作者关键词有超过半数未被现有的知识组织系统覆盖。因此,抽取现有的知识组织体系中的术语无法有效表征和描述学科交叉领域,更无法依靠这些术语来进行精准搜索和发现。

通过标题和作者关键词的抽取,能够获得表征和描述领域的术语;但在搜索和发现应用中,仅依靠术语本身的语义,会造成语义的偏移,利用这些术语检索的论文集合,含有这样一些论文:采用遥感数据作为基础数据,利用一些空间分析方法或者统计分析、模型等来研究湿地,这部分论文是遥感应用类论文,而不是研究湿地遥感。基于术语(词概念)的标识和共现的方法是无法有效区分应用类论文和研究类论文的。

在湿地遥感知识体系的构建过程中,笔者尝试通过描述和构建领域模型来构建湿地遥感研究本体/知识图谱,但湿地遥感属于交叉学科领域,涉及很宽的内容范围。既包括了遥感基础,如物理基础、数学基础、地学基础内容,也包括遥感技术本身,如遥感器、遥感平台、图像处理等,以及将遥感应用到湿地研究部分。关于领域内容的 T-Box 抽象,无论是人工构建,还是通过现有数据识别和转换进行构建,都存在很大难度,因而最终形成的只是一个词概念网络。

1.2 基于深度学习的技术方法及实验分析

2015 年,谷歌公司与牛津大学的一项合作

研究表明,无监督学习和认知计算能够有效教会机器阅读和理解[7]。Elsevier 公司则在融合知识组织与词向量计算方面开展了成功的探索,其最新产品 Knowledge Graphs for Life Sciences 可以将来自 emtree、Uniprot、InChi、Reaxys Tree Structures 等的知识组织体系,与其全文数据库文献抽取的关系融合成统一的知识图谱,并通过构建词向量矩阵计算和预测新的语义关系,进一步补充更新和优化知识图谱[8]。这项研究为我们优化实验中遇到的传统知识组织系统与文献关键词融合,以及增强语义关系控制提供了可借鉴的解决办法,但如何融合以及如何支持在线动态服务是应用于精准知识发现的关键问题。

基于这一目标,笔者开展了一系列方法实验,将基于词向量、段落向量和文档向量的深度学习方法应用于文档特征词、领域术语、关键词或事实的抽取,弥补传统知识组织和自然语言处理的局限,有利于构建高质量的语义知识库,以及精确认知用户的检索意图。

1.2.1 分组文档高辨别力特征词抽取

实验设计:将前述实验中获得的湿地遥感文献集合划分为相关集合 A 与不相关数据集合 B,分别在数据集 A 和 B 中运用深度学习进行相似度计算,得出相应的特征词,并比较两组特征词的关键差异。由此提取出所有文档有别于一般语料的特征词集合,同时也分别提取出两组文档各自具有辨识力的特征词集合。

实验过程:利用 Scaled F-score 方法[9]获取各个文档集合(A+B,A,B)特征词。考虑到相关术语具有相对较高的类别特定精度和类别特定术语频率的基本特征(即类别中术语的特定百分比是某些术语),选择精度和频率的调和平均值作为特征词提取指标。

具体方法为:给定一个词 $w_i \in W$ 且类别 $C_j \in C$,定义词汇 w_i 属于一个类别的精度为:

$$prec(i,j) = \frac{\#(w_i,c_j)}{\sum_{c \in C} \#(w_i,c)}$$

函数 $\#(w_i,c_j)$ 代表词汇 w_i 在标记为类别 c_j 的文档中出现的次数或标记为 c_j 且包含词汇 wi 的文档数量,同样方法定义一个词汇出现在一个类别的频率:

$$freq(i,j) = \frac{\#(w_i,c_j)}{\sum_{w \in W} \#(w,c_j)}$$

两个值的调和平均值定义为:

$$H_\beta(i,j) = (1+\beta^2)\frac{prec(i,j) \cdot freq(i,j)}{\beta^2 \cdot prec(i,j) + freq(i,j)}$$

$\beta \in R+$ 是一个比例因子,如果 $\beta<1$,则偏向词频,如 $\beta>1$,则偏向精度,如 $\beta=1$,则两者相等。F-score 等价于 $\beta=1$ 的调和平均值。用该方法可计算出一组文档中的特征词汇。如果将之与通用语料对比,又可计算出一组文档与通用语料对比突出的特征词汇包括哪些。利用该方法分别计算全部文档(A+B)与通用语料相比的特征词集合,以及 A、B 两组文档自身的特征词集合。实验结果表明,与通用语料相比,全部两组文档中抽取的特征词大多为与湿地有关的词汇,而经专家确认的文档集合中提取出来的特征词大多为与遥感相关的词汇。分别取 A、B 文档的 Top 50 特征词绘图(见图 1),从图中可发现两组文档中共同的特征词、不同组别文档特征词的分布特征,以及特征词与文档集合的相关程度。

实验分析:如图 1 所示,相关文献集合 A 与非相关文献集合 B 均是湿地领域的相关研究论文,但两者间确实存在显著区分特征词,相关文献数据集 A 的区分特征词更多与遥感相关。该实验表明,基于 F-Score 的方法能够有效发现某一主题或类别文献区别于其他主题类别的显著特征词,同时也能作为领域术语抽取的有效方法。

1.2.2 关键词抽取及事实抽取

现代自然语言处理技术可利用基于图的算法,实现文档关键词及事实的抽取,以及语义知识库的构建。如通过 PageRank[10]、Textrank[11]、sgrank[12]等算法抽取文档的关键词,以及文档中包含的事实及三元组信息。仍然采用之前的湿地遥感文档进行试验,利用 sgrank 算法对第一

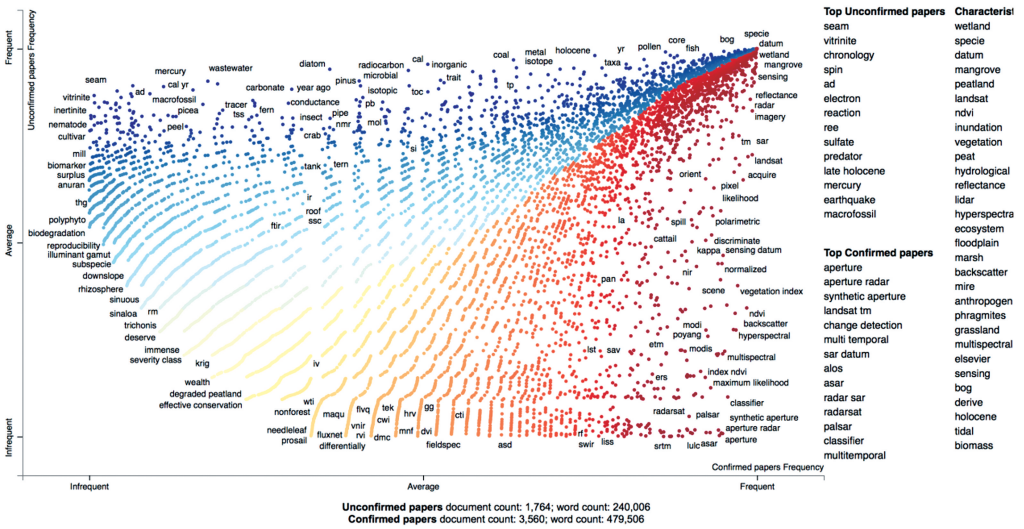

图 1　相关文档与非相关文档特征词分布

篇文档进行关键词抽取。示例如下:

原文:In this paper we first present the methodology and accuracy assessment of a 500m spatial resolution land cover map in insular Southeast Asia. This is followed by an analysis of the usability of the map for monitoring of deforestation/forest degradation The map was produced using Moderate Resolution Imaging Spectroradiometer (MODIS) images (acquired 1(st) Jan2(nd) Jul 2007), elevation information and peatland maps. The map covers the Malaysian Peninsula and the major islands of Sumatra, Java and Borneo, in addition to numerous smaller islands. The classification scheme of 12 classes reflects the special characteristics of land cover in insular Southeast Asia. We conclude that the map is suitable for monitoring the extent of forest cover in this region, but it misses the important aspect of forest degradation due to its inability to present varying forest degradation levels. Further development of the methodology is in progress to overcome this limitation.

利用 sgrank 算法抽取的关键词如下:

('spatial resolution land cover map', 0.2233-806270276347),

('insular southeast asia', 0.1581417312219-6432),

('moderate resolution imaging spectroradiometer', 0.07616515112141577),

('forest degradation',0.05606877656651209-5),

('numerous small island', 0.0344258279951-76614),

('methodology', 0.029614372795914428),

('accuracy assessment', 0.02706393575167-1746),

('major island', 0.022961773788720698),

('peatland map', 0.022802937683504818),

('malaysian peninsula', 0.02218240919996-905)。

同样利用基于图的算法可以抽取文档中包含的事实陈述及三元组信息。以下示例为从文献摘要中抽取与词汇“wetlands”相关的事实陈述及三元组信息。

原文:Multi-Polarization ASAR Backscattering from Herbaceous Wetlands in Poyang Lake Region, China, Wetlands are one of the most important eco-

systems on Earth. There is an urgent need to quantify the biophysical parameters (e. g., plant height, above ground biomass) and map total remaining areas of wetlands in order to evaluate the ecological status of wetlands. In this study, Environmental Satellite/Advanced Synthetic Aperture Radar (ENVISAT/ASAR) dual polarization Cband data acquired in 2005 is tested to investigate radar backscattering mechanisms with the variation of hydrological conditions during the growing cycle of two types of herbaceous wetland species, which colonize lake borders with different elevation in Poyang Lake region, China. Phragmites communis (L.) Trin. is semiaquatic emergent vegetation with vertical stem and bladelike leaves, and the emergent Carex spp. has rhizome and long leaves. In this study, the potential of ASAR data in HH, HV, and VV polarization in mapping different wetland types is examined, by observing their dynamic variations throughout the whole flooding cycle. The sensitivity of ASAR backscattering coefficients to vegetation parameters of plant height, fresh and dry biomass, and vegetation water content is also analyzed for Phragmites communis (L.) Trin. and Carex spp. The research for Phragmites communis (L.) Trin. shows that HH polarization is more sensitive to plant height and dry biomass than HV polarization. ASAR backscattering coefficients are relatively less sensitive to fresh biomass, especially in HV polarization. However, both are highly dependent on canopy water content. In contrast, the dependence of HH and HV backscattering from Carex community on vegetation parameters is poor, and the radar backscattering mechanism is controlled by ground water level。

抽取出的事实：

(Wetlands, are, one of the most important ecosystems on Earth)

抽取出的三元组：

(wetlands, support, diversity),

(wetlands, support, avifauna),

(wetlands, play, role)。

从以上实验结果可以看出，基于机器学习算法从某种程度上能够实现文档关键词、文档中所包含的事实陈述以及三元组的自动抽取。虽然其抽取结果的精度仍有待验证和优化，但这也至少给我们提供了一种辅助的或半自动化的手段，并有望实现文档更深层次的自动语义知识标注。

1.2.3 基于人机交互的迭代式认知搜索

结合语义知识库，并利用主题抽取、词向量等技术，通过对文献摘要的分析获取用户知识检索意图，帮助人们用更科学的方法迭代构建检索式，提升知识发现的完整性和相关性。笔者开发了认知搜索与知识发现系统，以实现上述功能(见图 2)。

图 2 中对利用 Food Traceability(食品质量安全追溯)在 WoS 中进行初始检索，获得 1 261 篇文献的元数据，利用 LDA 算法提取检索结果集合中的主题，在进行多维尺度空间算法降维后构建主题间距图，对每个主题内最显著的 30 个术语(约占全部术语的 4.3%)进行优先呈现，并将利用词向量计算的上下文相关词和近似词显示在下方，供用户参考选择。用户通过选择优化检索词和检索式，迭代直至得到能够明确清晰地表达用户搜索需求的意图树，并重构检索式。本实验通过迭代形成二次检索式 TS = (traceability AND recall) OR TS = ("supply chain" AND traceability AND "data fusion")。

利用上述检索式再次检索 WoS 网站得到 480 篇文献，研究检索的结果文献可发现，某种程度上第二次检索在相关性和完整性上都更加贴近用户知识获取意图。从以上实验结果可以看出，基于人机交互的迭代式认知搜索可以有效帮助用户逐步明确认知意图，实现精准发现，并为进一步的知识计算提供可能。

以上示例说明，机器学习和深度学习技术在图书情报领域的深入应用已渐成趋势，但方法

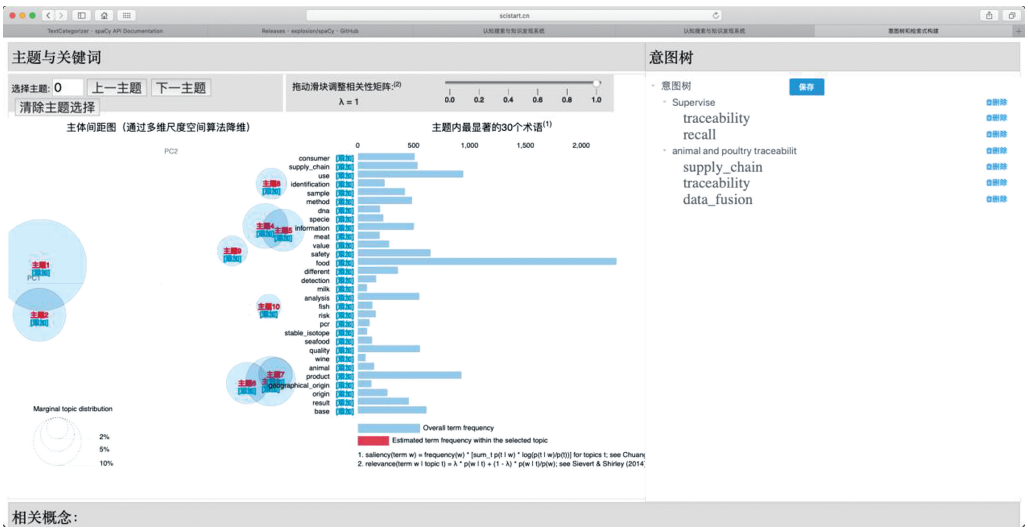

图 2 通过文献摘要分析获取用户知识检索意图并重构检索式

的精度、可控性及可解释性仍需专业人员深入调优和参与,方能得到满意的结果。

2 融合知识组织与认知计算的实现思路

2.1 基本思路

德勤咨询首席分析师 Roma 指出[13],认知计算能够自动从数据中提取概念和关系,理解其中含义,并独立地从数据模式和先前经验中进行学习——拓展人或机器可自行完成的工作。认知计算建立在神经网络和深度学习之上,它代表一种全新的计算模式,集成了信息分析、自然语言处理和机器学习领域的大量创新技术[14]。笔者通过前面两组实验分析证明,传统知识组织方法无法单独支持特定主题的精准知识发现,尽管基于词向量的深度学习方法可以有效弥补传统知识组织系统的局限,但又受到计算语料规模和质量的限制。因此,面向泛在精准知识发现和知识服务需求,笔者提出融合知识组织与认知计算的体系架构,如图 3 所示。

首先,利用文献计量学方法构建高质量小规模的初始语料集。利用文献计量学方法从作者、机构、项目、文献等多个指标预先构建面向领域或主题文献质量评价体系,以传统知识组织系统为基础采集并组织原始文献,基于文献质量评价体系优选抽取其中高质量文献,形成初始种子语料集。

其次,利用迭代式认知技术构建语义知识组织体系。运用词向量相似度计算、文档相似度计算等技术方法迭代生成认知意图树、显著特征词集和精准文献集。利用显著特征词对传统知识组织系统进行概念及概念(实体)关系扩展和优化,并利用基于图的深度学习方法抽取精准文献集中的知识单元及语义关系,构建语义知识库。

然后,在语义知识组织体系的基础上,利用深度学习、迁移学习等方法,突破语义智能检索、检索结果的多重因子排序、智能推荐计算、潜在关系挖掘、领域自动综述等关键技术,构建文本挖掘和认知计算引擎。

最后,集成上述关键技术,基于大数据与微服务架构构建新一代开放知识服务系统,嵌入含各种计量分析、演化分析、可视化分析、协同推理在内的认知搜索、知识发现、智能推荐及智能问答服务。

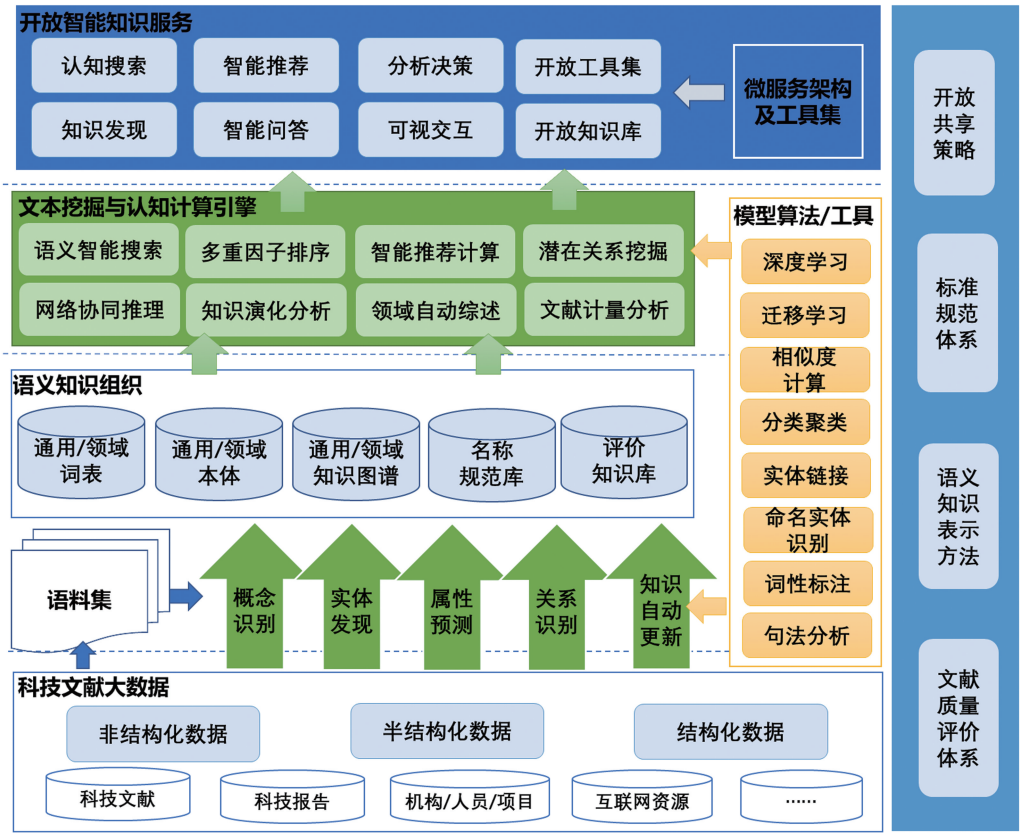

图 3　融合知识组织与认知计算的体系架构

2.2　关键技术分析

融合知识组织与认知计算的新一代开放知识服务系统，需要突破和集成优化的关键技术主要包括：语义知识表示技术、大规模语义知识库构建技术、认知计算技术、知识计算技术和可视化交互技术，具体如下。

（1）语义知识表示技术。以科技文献为来源和核心对象，构建知识单元语义描述模型和知识属性体系，知识单元类型包括：科学问题、科学假设、科学原理、技术与工艺、材料与方法、实验与数据、分析与结论等。采用各类知识单元语义关联的知识组织方法（如基于 hrLDA 主题模型和 word2vec 神经网络模型无监督本体学习方法），建设受控词表系统、领域本体、知识图谱等。

（2）大规模语义知识库构建技术。主要指基于科技文献、科学数据的大规模语义知识库构建技术，包括多源异构大数据融汇治理、基于特征测度的领域分析文献集构建、基于词向量扩展的数据集构建、语义知识库自动学习与进化更新、基于科学计量学的嵌入式知识源质量评价、可动态扩展的知识单元及其语义关系存储等，建立以语义知识组织系统和用户问题融合为标注语料的知识单元概念识别、实体发现、属性预测、实体关系抽取、关联等任务所需的深度学习技术策略。

（3）认知计算技术。认知计算更加强调机器如何能够主动学习、推理、感知世界，它会根据环境的变化做出动态的反应，所以认知更加强调它的动态性、自适应性、鲁棒性、交互性[15]。

认知计算技术可以有效支持智能问答,实现人机交互模式下用户认知意图精准刻画和机器自动问答,覆盖从信息采集、意图精炼、问题完善、语料获取、深度洞察到探索式分析的认知计算完整迭代过程。核心技术包括:结合上下文识别问句中的实体和术语、指代消歧、实体链接,融合深度学习和知识图谱实现对领域知识库的答案检索,融合问句和相关事实生成自然语言句子形式的答案等。

(4)知识计算技术。主要包括核心观点抽取、核心知识抽取、给定关键关系及其实体抽取、协同推理、基于共现和共引的合作关系分析、多重因子自适应排序技术、用户增强画像、知识演化分析、基于科学计量学的质量推荐等关键技术,支持知识图谱驱动的检索意图智能理解、认知搜索、领域知识画像和研究侧写、领域发展态势分析预测和智能知识问答等。

(5)可视化交互技术。主要指知识网络可视化呈现技术,以及递进式、探索式知识发现的可视化交互技术,包括:基于主题与范畴分类、领域本体和知识图谱进行关联的知识网络动态交互呈现,交互式宏观与微观知识分析,基于语义知识库和领域本体的人机交互可视化协同推理,可以实现知识演化进程、前沿发展态势、技术成熟度曲线、科技竞争力、主题分布与热点迁移、关键事件及关系展示、研究者集群及机构集群分析、交叉学科分析、知识点与文献互动分析、学科关键关系挖掘与展示等知识可视化应用。

3 结论与展望

综上所述,笔者认为通过融合知识组织与认知计算构建新一代开放知识服务系统,实现一系列关键技术的研究突破与优化集成,将具有几方面的创新性:①通过知识组织与认知计算的融合,可以有效应用知识组织体系优化知识抽取、概念识别、实体发现、属性预测的质量和效率,解决优质大数据和语义知识库缺乏而导致机器算法不够精准智能的难题,提升知识组织系统支撑下无监督学习的质量,提高知识计算效率、智能问答准确率和知识化服务智慧程度,实现受控知识组织体系与人工智能技术的双向驱动与优化促进;②融合系统能够突破多源异构数据的深度跨界融合、知识自动获取与可持续增长技术,实现基于海量科技文献、科学数据、科技报告等多模态专业科技信息向领域本体及知识图谱的大规模自动转化与可持续进化更新,实现无监督或半监督的非结构化文献数据向结构化、语义化知识网络的转变;③在认知搜索和问答服务进程中嵌入基于科学计量与文献计量学方法的知识源质量评价,能够支持精准和高质量知识源被优先推荐,实现集知识组织体系、文本大数据挖掘、用户增强画像和领域精品知识库四位一体的新型语义智能搜索与问答引擎,提供具备语境感知能力的语义发现、知识关联和个性化服务,显著改善知识获取效率和用户体验。

笔者利用传统知识组织构建方法、基于深度学习的方法进行语义知识库的构建和精准发现实验,提出改进的实现思路——融合知识组织与认知计算两种技术方法,并对其体系架构和关键技术进行分析,具有诸多创新。但相关研究也还存在一些不足之处,如领域特征不完善,为了实现领域知识的精准发现,必须引入更多维度来表征领域特征,一方面需要加入外部特征,如作者、机构、期刊、评价等,另一方面需要对内容特征进行术语扩展和实体类型梳理;相关实验还不够深入完整,后续需要对通过机器学习获得的新的特征词与知识组织体系进行比对和融合,再次开展发现实验进行检验。未来一段时间,笔者所在研究团队将深入开展相关关键技术的研究突破与验证应用,以期推动大数据、人工智能时代的新型知识发现系统进入新模式和新常态。

参考文献

[1] Mesbah S, Fragkeskos K, Lofi C, et al. Semantic annotation of data processing pipelines in scientific publications [C]//14th Extended Semantic Web Conference (ESWC). Berlin: Springer, 2017:321-336.

[2] 苏静, 曾建勋.国内外语义出版理论研究述评[J].中国科技期刊研究, 2017, 28(1):33-38.(Su Jing,Zeng Jianxun. A survey of theoretical research on semantic publishing[J] Chinese Journal of Scientific and Technical Periodicals, 2017, 28(1):33-38.)

[3] Plan S. Making open access a reality by 2020[EB/OL]. [2018-10-20].https:// www.sciencee urope.org /making-open-access-a-reality-by-2020.

[4] Jakus G , Milutinovic V′, Omerovic S, et al. Concepts, ontologies, and knowledge representation[M]. New York: Springer , 2013.

[5] Springer Nature. SN SciGraph-a linked open data platform for the scholarly domain [EB/OL]. [2018-10-20]. http://www.springernature.com/gp/researchers/ scigraph.

[6] 孙坦,刘峥.面向外文科技文献信息的知识组织体系建设思路[J].图书与情报,2013(1):2-7.(Sun Tan, Liu Zheng. Methodology framework of knowledge organization system for scientific & technological literature[J]. Library & Information,2013(1):2-7.)

[7] Hermann K M, Kocisky T, Grefenstette E, et al. Teaching machines to read and comprehend [EB/OL] . [2018-10-13] .https://arxiv.org/abs/1506.03340.

[8] Allen B P. The role of metadata in second machine age[EB/OL].[2018-10-13] .http://dcevents.dublincore.org/IntConf/dc-2017/paper/view/526/648.

[9] Kessler J S. Scattertext: a browser-based tool for visualizing how corpora differ[EB/OL]. [2018-10-13]. https://arxiv.org/pdf/1703.00565.pdf.

[10] Musib N. Page rank algorithm and implementation[EB/OL]. [2018-10-13]. https://www.geeksforgeeks.org/page-rank-algorithm-implementation/.

[11] Wen Y J, Yuan H,Zhang P Z. Research on keyword extraction based on Word2Vec weighted TextRank[C]//2nd IEEE International Conference on Computer and Communications. Piscataway:IEEE,2016:2109-2113.

[12] Danesh S, Sumner T, Martin J H. SGRank: combining statistical and graphical methods to improve the state of the art in unsupervised keyphrase extraction[EB/OL] [2018-10-13].http://aclweb.org/anthology/S15-1013.

[13] Violino B. What is cognitive computing? Here's what you should know[EB/OL]. [2018-10-13].https://www.infoworld.com/article/3198633/primer-make-sense-of-cognitive-computing.html.

[14] 认知计算[EB/OL]. [2018-10-13]. https://baike.baidu.com/item/%E8%AE%A4%E7% 9F%A5%E8%AE%A1%E7%AE%97/1803591?fr=aladdin.(Cognitive computing [EB/OL]. [2018-10-13]. https://baike.baidu.com/item/%E8%AE%A4%E7% 9F%A5%E8%AE%A1%E7%AE%97/1803591?fr=aladdin.)

[15] Jones M T. A beginner's guide to artificial intelligence, machine learning, and cognitive computing[EB/OL]. [2018-10-13]. https://developer.ibm.com/ articles/cc-beginner-guide-machine-learning-ai-cognitive/.

孙 坦 中国农业科学院农业信息研究所,农业部农业大数据重点实验室,研究员。北京 100081。

刘 峥 中国科学院文献情报中心副研究馆员。北京 100190。

崔运鹏 中国农业科学院农业信息研究所,农业部农业大数据重点实验室,研究员。北京 100081。

鲜国建 中国农业科学院农业信息研究所,农业部农业大数据重点实验室,副研究馆员。北京 100081。

黄永文 中国农业科学院农业信息研究所副研究馆员。北京 100081。

(收稿日期:2019-04-04)