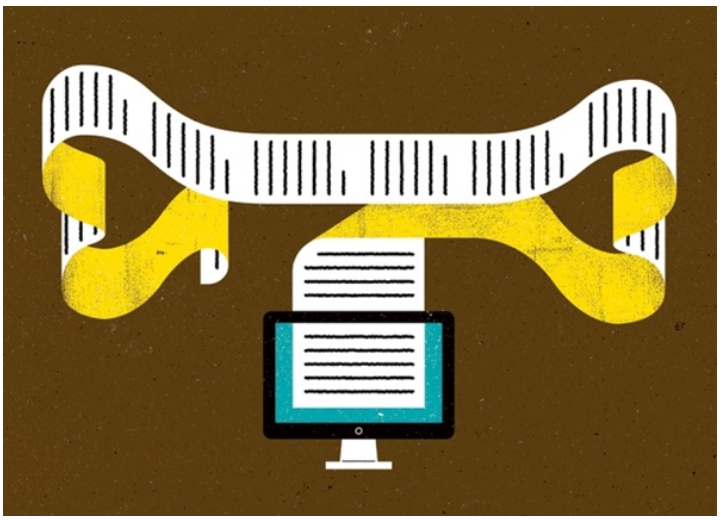
用计算机解放科学家双手

古生物学家拟用软件构建化石数据库

文章来源：中国科学报 张章 发布时间：2015-07-08 【字号： 小 中 大 】



图片来源：The Project Twins

对一个以记录历史为己任的领域而言，当提到组织其数据时，古生物学可谓十分有远见。维多利亚时代的自然历史博物馆就开始精心组织其手写的卡片，一直保存至今。而且，在过去15年间，研究人员已经将超过100万个化石的资料输入到数据库中，以便追踪生命历史的广阔发展趋势。目前，古生物学家正在探索能直接从研究论文中提取化石数据的新软件。

“我确信这将是未来趋势。”该项目合作者、美国威斯康星大学麦迪逊分校古生物学家Shanan Peters说，“构建数据库本身已经成为了过去。那些数据库将基于你感兴趣的问题而自动生成，计算机将担当重任。”

Peters是古生物学数据库（PBDB）的主要研究者，该数据库详细记录了约120万件化石的年代、位置和特征。自1998年启动以来，研究人员已经花费约8万小时录入从初期野外考察和约4万篇文章中提取的数据。PBDB已经产生了数百篇论文，并帮助古生物学家处理了若干除此之外无法回答的问题，例如新纪元灭绝率和某些恐龙的消失等。

PBDB是一个由专家创建的数据库：约380位科学家上传了有关32万个分类名称的约56万条已发表观点。但Peters十分好奇计算机能否自动编辑一个此类数据库。于是2013年，他开始与该校数据科学家Miron Livny和Chris Ré合作。那时，Ré已经开发出一个名为DeepDive的软件，该软件能挖掘书面文本，并提取出各种元素。在计算科学领域，文本挖掘目前已经司空见惯，并正慢慢开始用于基因组学和药物研发等领域。

DeepDive能以一种方式解析研究论文。“它能接受论文，并将其转化成文本。” Ré说，它能试着确定“什么是名词”“什么是动词”等问题的答案。下一步，该软件会尝试预测相关句子的概念（例如，对于古生物学而言，化石的名字和它们被发现的地点），并为每个判断分配一个可能性。但Ré表示，软件“通常是有缺陷的。这也是你需要相关领域科学家介入的地方”。

于是，Peters花费1年时间精练了软件，例如，它能知道在哪里寻找古生物学论文中新物种的名字和地理位置等信息。Ré将该过程称为与Peters的“不断往复的过程”，后者要求Ré团队想出自定义解决方案，以便能实现这些要求。“我很想说，答案是人们能按下按钮，并使用它，而且他们不再需要我们。”Ré说，但这一目标还未能实现。

作为原理论证，Peters和Ré使用他们称之为PaleoDeepDive的定制软件，创造了一个能文本挖掘的小规模PBDB，其中包含约1.2万篇论文。Peters表示，从某种程度上而言，计算机生成的数据库比PBDB更好。原因之一

热点新闻

发展中国家科学院第28届院士大...

14位大陆学者当选2019年发展中国家科学...
中科院举行离退休干部改革创新发展形势...
中科院与铁路总公司签署战略合作协议
中科院与内蒙古自治区签署新一轮全面科...
发展中国家科学院中国院士和学者代表座...

视频推荐

【新闻联播】“率先行动”计划 领跑科技体制改革

【共同关注】“首例基因编辑婴儿”事件：中科院发表声明——坚决反对

专题推荐

是，其中所有的信息都能连接到原始文本。“不过，当出现歧义，或文档合作者间存在偏差，计算机确实充满不确定性。”Peters说。不过，PaleoDeepDive还能设法从论文中的分类学名称中提取19.2万条观点。而PBDB的工作人员只找到了8万个。

另外，PaleoDeepDive在组织信息时也毫不费力。在发表于2014年12月的一篇论文中，Ré和Peters表示，在提取自电脑生成数据库中的100个语句随机样本中，92%的是正确的，这与PBDB的准确性相当。这两个数据库在第二个实验中也分数接近：科学家拿出5份文件，并要求它们从中获取准确元素。或许最让人印象深刻的是，PaleoDeepDive在评价过去5亿年物种多样性和灭绝率方面的工作。

“有点可怕，机器也能做到如此好。”PBDB执行委员会成员、乔治·梅森大学古生物学家Mark Uhen说。英国帝国理工学院古生物学家Jonathan Tennant则表示，“我认为这是古生物学领域长期以来最好的创新之一。”他每天都在使用PBDB，并认为文本挖掘将是收集大规模数据的有效方式，而之后再进行人工检查。“我不认为机器能取代人类。在分析学上保留人的视角非常关键。”Tennant说。

而PBDB联合创建者、澳大利亚麦考瑞大学古生物学家John Alroy似乎不太看好文本挖掘。他表示，DeepDive通常会过度评价物种存在的时期，这会导致对物种多样性的错误评估。

尽管存在诸多挑战，许多古生物学家仍将文本挖掘视为该领域的重大进步。“对于毕业生和博士后而言，手动将已发表的信息录入到数据库里十分浪费时间。”伦敦自然历史博物馆古生物学家Ross Mounce说。Peters希望PaleoDeepDive的努力能让他和同事有更多时间生成新数据，而非整理已有数据。“我认为这些机器阅读系统正在一点点将我们解放出来，让我们的工作重回野外和博物馆。”

（责任编辑：侯茜）