012 | 精读2017年KDD最佳研究论文

2017-10-30 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 10:34 大小 4.85M

前面我们介绍过 KDD 大会的时间检验奖，每年大会的另外一个亮点奖项就是最佳论文奖，有两类，一类是最佳研究论文，一类是最佳应用数据科学论文。今天我就先来说说前者。

大会每年都会在众多的学术研究论文中，选择最有新意和价值的研究论文，评选出最佳研究论文的第一名和第二名。从过去十多年的经验来看，KDD 历年的最佳研究论文，都会对之后很多领域的研究有开创性的影响。因此，不论是从阅读经典文献的角度，还是从学习最新研究成果的角度来说，认真分析和探讨每年的最佳研究论文都是一个不错的选择。

今天，我就带你认真剖析一下 KDD 2017 年的最佳研究论文《通过挖掘类比关系加速创新》（Accelerating Innovation Through Analogy Mining）。

作者群信息介绍

第一作者汤姆·霍普（Tom Hope）来自耶路撒冷的希伯来大学（The Hebrew University of Jerusalem），计算机博士，在读第三年。同时，他还是英特尔以色列的资深数据科学员，对深度学习的很多方面都有研究。目前他正在写一本基于 TensorFlow 的深度学习简明技术书籍。

第四作者达夫娜·沙哈夫（Dafna Shahaf）是霍普的博士导师，目前在希伯来大学计算机系任助理教授。达夫娜于 2012 年从卡内基梅隆大学博士毕业。她曾经在微软研究院以及富士通公司实习，并在斯坦福大学攻读博士后。达夫娜的论文曾获得 2010 年的 KDD 最佳研究论文，可以说她一直站在机器学习研究的前沿。

第二作者乔尔（Joel Chan）是来自卡内基梅隆大学人机交互学院的科学家。乔尔于 2014 年从匹兹堡大学毕业，获得认知心理学博士学位。他一直在人机交互领域进行研究。

第三作者安尼凯特·科图（Aniket Kittur）是来自卡内基梅隆大学人机交互学院的副教授。他于 2009 年从加州大学洛杉矶分校毕业，获得认知心理学博士学位，之后就一直在卡内基梅隆大学任教。

从整个作者群的情况来看，这篇文章是一个比较典型的机器学习技术与人机交互领域的交叉成果。

论文的主要贡献

我们先来看一下这篇文章的主要贡献。当然，要想深入理解这篇文章的贡献，我们还要先弄明白，这篇文章主要解决的是一个什么场景下的问题。

这篇文章主要阐述了帮助创新的一个重要步骤，那就是如何找到合适并且有效的类比案例。什么叫作类比案例？在人类发展的历史上，特别是科学技术的革新历程中，有很多重要的发明发现，都是因为当时的科学家借鉴了一些类似场景中的解决方案，或者是从这些场景中获取了灵感，进一步做出了创新。在这个步骤中，从类似场景中借鉴往往被称作类比。

比如，莱特兄弟从自行车中得到灵感，制造出了可以滑行的飞行器。再比如，诺贝尔生理学或医学奖得主萨尔瓦多·卢里亚从对赌博机运行的观察中，进一步发现了细菌基因突变的规律。同时，类比在很多国家的法律，不管是成文中或者是运行中都比比皆是。因此，我们可以看到，**如何找到合适的类比，并能从中获取灵感，可能就是创新的一个关键因素。**

时至互联网时代的今天，我们已经有很多数据库、文献库可以作为获取类比灵感的重要数据源泉。比如，谷歌学术搜索（Google Scholar）有上百万的论文和专利信息；OpenIDEO 有上百份关于社会问题的分析；Quirky 有超过两百万份的产品构想；InnoCentive 有超过四万份的社会、政治以及科技方面的解决方案；美国专利局有超过九百万份的专利信息，类似的例子还有很多。

与此同时，海量数据也为寻找灵感带来了很大的挑战。如何才能在这些百万级别的信息库里快速定位可能的类比场景，就成了阻碍创新速度的一大瓶颈。

找到类比场景有一个很大的挑战，那就是如何定义“类似”或者“相似”。如果仅从一些表面特征来看，很多场景之间的相似程度是很低的。因此，好的类比场景一定是能够找到，刨除表象相似以外的深层次相似的案例。另外一个挑战就是，寻找类比场景中，是否能有一个**非常丰富的数据结构**来支持推理，如果有，往往就能有比较简单的方法。

这篇论文的重点是如何从海量的无结构或者弱结构的文本数据中找到这样的类比。我们可以看出，这确实是一个巨大的挑战。

理解了这篇论文的目的，我这里就直接给你总结一下它的贡献，那就是提出了一种自动的在海量无结构的文本数据中挖掘类比场景的方法。这篇文章关注的是产品信息数据。作者们通过实际的数据，验证了提出的方法比之前的一些文本处理方法更加有效。

论文的核心方法

了解了这篇文章的目的和贡献后，接下来，我就来剖析一下作者们究竟提出了一个什么方法。

首先，作者们提出了一组叫“**目的**”（Purpose）和“**机制**”（Mechanism）的概念。什么叫“目的”呢？那就是当前的产品是要解决什么问题的。什么叫“机制”呢？那就是当前的产品是使用什么手段或者方法来解决这个问题的。对于一个产品，如果我们能够明确这个产品的目的和机制，找到类比就变得更加容易。

比如，我们可以针对某一个问题，相同的目的，采用不同的机制或者对不同的问题采用相同的机制。作者们认为，**这种对产品信息的分类符合很多工程设计的过程，是创新过程中的一个必要环节。**

有了这种想法以后，很自然的下一个步骤就是如何从数据中学习到目的和机制，如何自动挖掘出海量产品信息的目的和机制。要想学习到这样的信息，作者们提出了一种依靠标签数据的监督学习（Supervised Leanring）机制。具体说来，作者们把文本信息中的每句话、短语交给亚马逊土耳其机器人（Amazon Mechanical Turk）上的在线工人，来标注每个文本信息是目的信息还是机制信息。也就是说，作者们依靠有标注的数据来训练提出的算法。

首先，我们有一组文本，每组文本都有这些文本的原始文字。**针对每个文档，我们都收集 K 个目的标注和 K 个机制标注。**这时，我们定义一组“目的标注”（Purpose Annotation）向量，其实也就是一组 0 或者 1 的向量。当文本原始文字中的某个字被标识为目的的时候，这个向量的相应元素置 1，反之置 0。类似的，我们也可以定义“机制标注”（Mechanism Annotation）向量。因为我们有 K 个标注，因此我们也有相应的 K 个“目的标注”向量和“机制标注”向量。这两组向量可以说是原始标签信息的一种向量的表达。

下一步就是从每一个有标签信息的文档里**产生唯一的目的向量和机制向量。**这篇文章采用的方法是，利用每个单词的**嵌入向量**（Embedding）来获得这个唯一的向量。

具体方法是这样的，首先，针对每一个标注（总共有 K 个），我们收集属于这个标注的单词的嵌入向量，并把这些嵌入向量都拼接起来。然后计算这组拼接好的向量所对应单词的**TF-IDF 值**（Term Frequency–Inverse Document Frequency，词频 - 逆向文件频率），并且取 TF-IDF 值最高的一些单词相对应的嵌入向量，加权平均以后，就得到了相应的唯一的目的向量或者是机制向量。作者们发现这种利用 TF-IDF 值加权的方法可以更加有效地表达文本的各种重要信息。注意，这个步骤是依赖于文档标签的，也就是说，我们只能对训练数据进行这样的构造。

到目前为止，我们描述了如何从文本到达文本对应的目的向量和机制向量的步骤。那么，如何基于这样的数据以及向量，来对未知的文档进行提取目的向量和机制向量呢？文章采用了深度模型**RNN**（Recurrent Neural Network，循环神经网络），具体说来是双向的 RNN 并且有一个**GRU**（Gated Recurrent Unit，门控循环单元）。

这里我就不复述细节了，总体的思路就是，根据文档的嵌入向量信息，我们希望得到一组文档的**隐含表达**（中间参数），然后可以从这个隐含表达来预测目的向量和机制向量。注意，因为需要预测两组目标，目的向量和机制向量，因此，这里至少需要分别有两组参数。

除了预测文档的目的向量和机制向量以外，作者们还提出了一个用**少数关键词**来解释这两组变量的机制。具体说来，就是设立一个新的学习目标函数，希望通过少数关键词所对应的嵌入向量来重构目的向量或者机制向量，让得到的误差最小。

方法的实验效果

作者们使用了 Quirky 数据集，通过亚马逊土耳其机器人标注了八千多组产品信息。首先，检测了利用学习到的目的向量和机制向量，是否能够比较容易地从海量数据中提取相应的类比信息。这里，作者们直接利用把目的向量和机制向量拼接的方式来表达问题。答案是，效果非常显著。**在前 1% 到 25% 的提取结果中，精度（Precision）和召回（Recall）都比之前的标准文本处理方法，比如 LDA、TF-IDF、全局的嵌入向量要好 10% 到 20%**，可以说这是非常有效果的。

作者们还测试了，通过提出的方法，是否能够为用户推荐比较好的类比场景。这里，文章又找了 38 个亚马逊土耳其机器人的虚拟工人来为 12 个产品思路打分。在不知道推荐方法的情况下，虚拟工人认为这篇文章提出的方法能够推荐更有新意、在深层次上更加类似的场景。这也部分解决了我们前面说到的文章希望解决的问题。

小结

今天我为你讲了 KDD 2017 年的年度最佳研究论文，这篇论文提出了一种自动的方法来挖掘类比信息，为快速创新铺平道路。一起来回顾下要点：第一，我简单地介绍了这篇文章的作者群信息，帮你了解到这篇文章是机器学习和人机交互研究的一个结合。第二，我详细地介绍了这篇文章想要解决的问题以及贡献。第三，我简要地介绍了文章所提出方法的核心内容。

最后，给你留一个思考题，这篇文章提出的是使用标注信息来获取目的向量和机制向量，我们有没有办法能够不使用标注信息，采用完全无监督的方式呢？

欢迎你给我留言，和我一起讨论。

拓展阅读：Accelerating Innovation Through Analogy Mining

上一篇　011 | 数据科学家高阶能力之如何系统提升产品性能

下一篇　013 | 精读2017年KDD最佳应用数据科学论文

精选留言 (3)

写留言

Luna
2017-10-31

2

你好，对于论文的核心方法部分有一点疑虑，双向rnn的输入以及输出分别是什么呢?

另外，提一个小小的建议，在讲解核心方法部分时，可否加一些简单的例子和图示?

谢谢!

展开

作者回复: 好的，尽量。

范深

2017-10-30

1

所以这篇文章的主要创新点是提出“目的”和“机制”的文本作用？似乎没有看到算法理论上的其他创新。那么它的推广价值在哪呢？

黄德平

2018-11-29

信息量好大，这一篇篇文章，想搞清楚所有细节很不容易，但是掌握核心思想也是很有帮助的