

030 | 文档理解第一步：文档分类

2017-12-11 洪亮劫

AI技术内参

[进入课程 >](#)



讲述：初明明

时长 08:30 大小 3.90M



我们在前几周的专栏里讲解了最经典的信息检索（Information Retrieval）技术以及基于机器学习的排序学习算法（Learning to Rank），并且花了一定的时间分享了查询关键字理解（Query Understanding）这一关键搜索组件的核心技术要点。上周，我们还详细讨论了如何从线上和线下两个层面来评价一个搜索系统。

这周我们的分享将转移到搜索的另外一个重要部件：**文档理解**（Document Understanding）。也就是从文档中抽取各种特性，来帮助检索算法找到更加相关的文档。

文档理解最基本的一个步骤就是给**文档分类**（Classification），看这些文档表达什么类别的信息。今天我就来和你聊一聊文档分类的一些基本概念和技术，让你对这方面的开发与研究有一个基本认识。

文档分类的类型

如果我们把文档分类看做一个监督学习任务的话，那么在各式应用中就经常使用以下几种类型的文档分类。

第一个类别就是**二元分类**，或者称为二分文档分类，目的就是把文档分成两种不同的类别。比如，把文档分成“商业类”或者“非商业类”。

第二个类别自然就是**多类分类**，也就是判断文档是否属于好几种不同类别中的某一个。比如，把文档划归为“艺术”、“商业”、“计算机”或者“运动”类别中的某一类。

当然，在多类分类的下面，我们还可以分三个小类别。

第一个小类别，是“**多类 - 单标签 - 硬分类**”（Multiclass, Single-Label, Hard Classification）。什么意思呢？就是说每一个文档只能在多类分类问题中被赋予唯一的标签，并且所有互相的类别是不兼容的。

第二个小类别，就是“**多类 - 多标签 - 硬分类**”（Multiclass, Multilabel, Hard Classification），也就是说每一个文档可以被认为属于多个类别，然而每个这样的分类都是唯一确定的。

最后一个小类别则是“**多类 - 软分类**”（Multiclass, Soft Classification），也就是认定每个文档以概率的形态属于多个类别。

在这个分类基础上，还有一种分类的方法，那就是可以把所有的类别看做一个平面的结构（Flat）或者是有组织结构的。通常情况下，如果把文档分类到一个层次组织（Hierarchical Structure）里就叫“**层次分类**”（Hierarchical Classification）。在这样的情况下，一个文档同时属于这个层次结构上从根节点到叶子节点的所有类别。一般来说，上层节点相对于下层节点更加抽象。

文档分类经典特性

了解了文档分类的基本类型之后，我们接着来讨论文档分类所用到的经典特性。

我们最先会想到的当然是使用文档上原本的文字信息。最直接的文本特性可能就是每个英文单词，或者中文的词语。这种完全把文字顺序打乱的方式叫作“**词袋模型**”（Bag-of-

words Model) 。

从很多实践者的报告来看，“词袋模型”虽然不考虑文字的顺序，但是在实际使用中，依然不失为一种非常有效的特性表达方式。同时，在“词袋模型”中，每个词的权重其实可以用我们之前介绍过的 TF-IDF 或是语言模型（Language Model）对单词进行加权。关于 TF-IDF 以及语言模型，建议你回到我们前面讲过的内容去复习一下。

除了“词袋模型”以外，还有一些不同的尝试，是希望能够保留部分或者全部的词序。

比如，我们曾经讲过的“**N 元语法**”（N-gram）对文字的表达方法，就是一种非常有效的保留部分词序的方法。不过，N 元语法最大的问题就是极大地增大了特性空间，同时，每一个 N 元组被观测到的次数明显减少，这也就带来了数据的稀少（Sparsity）问题。

除了 N 元语法以外，近年来随着深度学习的推广，比较新的思路是用“**递归神经网络**”（RNN）来对序列，在这里也就是词句进行建模。有不少研究表明这样的效果要明显好于“词袋模型”。

除了文档上的原始文字以外，文档上的排版格式其实也是很重要的。有些字段有很明显的特征，比如一个文档的标题显然占据了举足轻重的地位。有一些文档有“章节”、“段落”等结构，其中这些小标题对文章的主要内容有很大的指导意义。于是，对文章的不同“字段”（有时候也叫做“域”）进行建模，对文档分类的效果可能会有比较大的影响。

另外，针对某些特殊文档，仅仅考虑文字的基本信息可能是不够的。例如，现代网页的原始 HTML 表达和最终在浏览器中呈现出来的效果很可能会有较大区别。因此，针对网页，我们可能还需要采用浏览器中最终呈现出来的视觉效果来提取特性。

对于孤立的文档来说，单个文档的信息可能是比较有限的。但是在互联网上，很多文档都不是孤立存在的。就拿普通网页来说，互联网的一个特点就是很多网页都通过各种链接连到一起。这些和当前网页相连的其他页面很可能就会为当前页面提供一些额外信息。

在所有这些周围的页面中，有一类页面值得在这里提一下。那就是这些页面上会有链接指向当前我们需要分类的目标网页。这些链接往往有文字描述来叙述目标网页的一些特质，甚至有一些周围的文字描述也是有意义的。

比如，当前网页是微软公司的首页，上面也许因为有各种精美的图片而缺乏文字描述，而周围的页面上很可能就有“微软公司官方网站”等链接指向微软公司的首页。这样，我们就通过这些链接文字得出了“微软公司”的信息，然后如果我们又知道微软公司是软件公司，那么就比较容易对这个页面进行分类了。

根据这个思路，我们就可以尝试去使用周围文档中更多的信息。不过，值得指出的是，周围文档信息所带的“噪声”也是比较多的。已经有各类研究尝试去理解周围文档中更多有价值的信息，这里就不赘述了。

文档分类相关算法

根据我们刚刚讲过的不同文档的分类类型，就可以直接借用已知的、熟悉的监督学习各种算法和模型。

假如是简单的二分文档分类问题，那“对数几率回归”（Logistic Regression）、“支持向量机”（SVM）、“朴素的贝叶斯分类器”（Naïve Bayes Classifier）就都能够胜任工作。而针对多类分类问题，也是标准的监督学习设置，刚才说到的这几类算法和模型在一定的改动下也能够做到。

近些年，深度学习席卷很多领域。在文档分类领域，各类深度学习模型也都展示出了一定的优势。

需要注意的是，并不是所有的分类算法都“天生”（Natively）支持“概率的输出结果”。也就是说，如果我们需要对“多类 - 软分类”文档问题进行建模，那就会有一些问题。比如支持向量机就是这么一种情况。在默认的状态下，支持向量机并不输出每一个数据样例属于每一个类别的概率。

因此，这里就需要用到一些技巧。在实际应用中，我们经常采用的是一种叫“**普拉特调整**”（Platt Scaling）的办法。简单来说，其实就是把支持向量机的输出结果再当做新的特性，学习一个对数几率回归。

除了我们刚刚讲的利用基本的监督学习手段进行文档分类以外，另外一种方法就是我们前面说的利用周围有关系的文档，也就是所谓的“**关系学习**”（Relational Learning）。关系学习是说，希望利用文档与文档之间的关系来提高文档的分类效果。这一方面的很多方法都会利用这样的思想：**相似的页面很有可能是相同的类别。**

如果是在“层次分类”的情况下，相似的页面就很有可能在层次结构上距离比较近。这里，“相似”有可能被定义成文字信息相似，也有可能是在文档与文档之间所组成的“图”（Graph）上位置类似。

比如，某一个公司的很多子页面，虽然上面的文字本身有差异，但因为都是这个公司的页面，从大的文档页面网络上看，他们都代表这个公司的信息，因此在进行文档分类的时候，也很可能会把他们放到一起。

小结

今天我为你讲了现代搜索技术中又一个至关重要的环节，那就是文档理解中的文档分类问题。你可以看到文档分类所要了解的信息还是比较多的。

一起来回顾下要点：第一，简要介绍了文档分类的主要类型，包括二元分类、多类分类以及层次分类。第二，详细介绍了文档分类所可能用到的种种特性，比如文档上原本的文字信息、文档的排版格式以及周围有关系的文档。第三，介绍了如何利用监督学习以及其他的算法工具来完成文档分类的任务。

最后，给你留一个思考题，如果一个文档中既有图片也有文字，那我们该如何组织这些特性，然后放到我们的分类器中去学习呢？

欢迎你给我留言，和我一起讨论。


AI 技术内参

你的360度人工智能信息助理

洪亮劼

Etsy 数据科学主管
前雅虎研究院资深科学家



新版升级：点击「 请朋友读」，10位好友免费读，邀请订阅更有**现金**奖励。

© 版权归极客邦科技所有，未经许可不得传播售卖。页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 029 | 如何评测搜索系统的在线表现？

下一篇 031 | 文档理解的关键步骤：文档聚类

精选留言 (4)

 写留言



鬼猫猫

2017-12-13

 2

每篇都做一下思维导图当笔记

展开 

作者回复: 谢谢支持。



sky

2018-06-07

 1

利用深度学习把图片的特征学习出来，再把这些特征放到分类算法里面去训练，这样可以

吗



HaveTwoBru...

2019-05-04



先将文档的每段文字转化成一个词向量，然后按照顺序和图片转化的向量进行连接，一起作为分类算法的输入，这样做是不是可以保留语义信息和结构信息？

展开 ▾



georgesupe...

2017-12-12



要是能在每篇技术文档后附上代码Demo就完美了

展开 ▾

作者回复: 这些文档主要起抛砖引玉的作用，不过谢谢建议。

