短文本关键词提取实践

数策 数策 2016-12-25

背景

在推荐系统中，最重要的工作是完成对用户兴趣的理解，并针对性的推荐内容、物品或者服务。推荐质量的好坏，通俗来讲，取决于推荐系统对于用户的理解和对物品的理解两方面。而对用户的理解在一定程度上也依赖于对物品的理解是否深刻。所以，加深对物品的理解，成了推荐系统优化中一项非常重要的任务。

推荐系统应用的一个重要领域是视频网站。在视频网站的推荐中，同样有对物品（内容）加深理解的需求。但不同于web page，每个视频的title都非常短，可以获取的信息有限，且获取有效信息的难度也更大（训练语料较少）。所以视频推荐中加深对视频内容的理解，是在更细分的短文本关键词提取领域的实践。

启动

本文是工业实践中以优化为目标的项目实践，不过分的追求算法背后的理论理解。项目启动借鉴敏捷开发的思想，组成小的项目组，统一思想，快速迭代和快速验证是整个项目推动中的核心指导思想。所以在项目实践中不追求绝对完美，有方便快捷的80分方案绝不采用需要更多资源和时间的90分方案。

方案

短文本关键词提取，简单分成两个大的部分：分词和赋权。

（一）分词

分词按照关键子任务分解，有切词、词性标注、命名实体识别、新词挖掘几个部分。每一个子任务，都可以细分出很多的方法和理论。比如切词的常见方法就分为字符串匹配、全切分和由字构词三种。而词性标注，最基础的也有基于规则的和随机标注算法。对于它们背后的细节不予深究。具体实践中，选用ansj_seg开源工具作为分词工具，ansj_seg是由ictclas优化而来的，全Java实现，效率和准确率据测试都比ictclas要更好一些，核心词库也是来自ictclas的核心词典，而且支持自定义词典和歧义纠正词典，便于日后添加规则。ansj_seg有词性分析，分词权重可以添加词性作为考虑因素。最让人惊喜的是，ansj_seg还有新词发现的功能，可以帮助我们补充词库。下面就实践中遇到的几个问题进行说明：

- 分词的效果严重受限于词典，所以前期产品辅助的重点在于词典的扩充。Ansj_seg支持自定义词典，这种贴心的可扩展性为我们的快速提升效果起到了重要作用。在补充词典中有两点体会：

a. 引入各种词库，各种方法引入词库：可以拿到的各种输入法的词库、各种百科的数据、所有垂直网站的数据、排行榜数据，以及内部可以提供的一切资源：明星库与全网作品库、超级手机浏览器中搜索和点击的关键词、编辑打的tag、图文版的tag等等等等。垂直领域的分词方案可以通过垂直网站信息解决。

b. 词库最好有一个分类体系。而不是所有的词库都混在一起。比如明星库、商业人物库、电影作品库、动漫作品库等。

- 准确度和新词发现之间的权衡

Ansj_seg提供几种分词调用方式，其中就有精准分词和Nlp分词。精准分词是在分词效率和精准性上实现很好权衡的分词方式。Nlp分词是一张充满惊喜性的分词方式，它具备新词发掘的功能，可以识别出词典中没有的词。具体使用哪种方式，根据不同业务的精准性要求有所不同。我们根据自己的精准性要求，忍痛放弃了Nlp的分词方法。

- 进一步切分

对于词典中有的“百变大咖秀第五季”这种长分词，按照最长匹配原则，只能分出“百变大咖秀第五季”，而不能分出“百变大咖秀”这个词，实践中为了解决这个问题，对长分词进行了进一步切分，切分出“百变大咖秀第五季”、“百变大咖秀”和“第五季”三个词。再通过权重等方式将“第五季”过滤。

实际分词中，就有全切分的方法，会切分出与词库匹配的所有可能的词，再运用统计语言模型决定最优的切分结果。全切分的方法主要是为了解决歧义的问题，但同时也解决了本问题。由于全切分出的词，需要语言模型选择出最优，较为复杂，本次未尝试。

- 添加歧义词典

歧义问题是分词面临的较大挑战。Ansj_seg支持添加歧义词典解决歧义问题。但由于歧义词典是强规则，添加不当会引入其他问题，所以需谨慎使用。

- 词库的识别顺序

自带词库和自定义词库，根据两个词库质量的不同，可以更改词库匹配的顺序。

（二）赋权

分词之后，需要对每一个词赋予权重。赋权方法我们选用的是LDA的主题模型。LDA的核心在于P(t|d)和P(w|t)。P(t|d)是指每个doc属于某个topic的概

率，P(w|t)是指每个word属于某个topic的概率。具体实现如下：

STEP1：随机初始化每个词的topic，并统计P(t|d)和P(w|t)；

STEP2：遍历训练语料，按照概率公式重新采样每个词对应的topic，并更新P(t|d)和P(w|t)；

STEP3：重复STEP2，直至模型收敛。

由于单纯使用LDA模型，调出一个靠谱的权重非常困难，基于项目快速迭代、快速验证的指导思想，制定了如下两个方案：

- 以LDA的权重为基础权重，利用TF-ITF进行加权或降权；
- 以TF-ITF的权重为基础权重，其中ITF中的T借助于LDA的topic，T也可以是具体的细分分类。

最终，选择了更优的第一种方案。

在上述实现中，为了快速获取更靠谱的权重，采取了多种方法进行修正：

1. 黑名单过滤：对于常用的无意义的词，快速通过技术手段导出top feature，加入黑名单，包括通过词性也可以导出一批黑名单；
2. 通过明星热度、大剧的播放热度、热门搜索词等，加权或降权；
3. 对《》,“”中的内容进行识别处理。
4. LDA的topic number是需要事先人工确定的，确定的数目对于效果有较大影响；
5. 对于某些分类，比如搞笑，实际上是不需要分词和赋权的，分词和权重结果需要有开关进行分类选择。

评估

分词和LDA的权重结果，均需要不断的进行评测，一套合理的评测标准有利于迅速发现问题和进行优化。在分词和LDA权重结果在30分到60分的过程中，评测人工一看，就可以轻易的看出问题，但是后期从60分优化到70分的过程中，一套合理的评测标准的必要性就愈发凸显。我们没有探索出有效的评测方法，仅仅列出一些想法，供大家参考，也欢迎一起讨论。

1. 由于评测的标准必须一致，所以最好是固定的一批人进行评测，有条件最好是建立评测团队，没有条件只能由产品兼任；
2. 评测团队按照分类挑选出足够的case，并基于人的理解，对各个case找出重要的分词和权重，这个case组成的库作为一个黑盒子，每次技术的结果出来之后就对这个库的视频进行跑分，根据分的大小决定是否有优化；
3. 2中的评测方法比较适合于分词的评测，权重的赋予由于有具体值的顺序和值的大小，不方便操作；
4. 2中评测库需要对技术封闭，不要造成按库优化；

5. 2中评测库对于时效性的内容，如果有失效，需要针对性解决；

6. 每次结果跑diff，看优化前后的影响面大小（同样最适用于分词的评测）；

7. 每次优化后，对badcase单独跑结果，评测是否有优化；

展望

短文本关键词提取实践还面临诸多难题，比如分词中的新词挖掘和歧义词处理，比如LDA赋权重中由于视频只能从较短的title中获取信息带来的困难。但从突破后对推荐系统带来的价值考量，此项工作值得更深入的投入。

更深入的讨论，待继续实践后进行更新和探讨。