基于 LSA 模型的微博用户分组与基于 TF-IDF 的用户画像分析

2018 年 6 月 6 日

1 背景

1.1 LSA 模型

LSA 模型之前一节课已经用到，这里不再赘述。在本次实验中，考虑到 LSA 与 PCA 的本质都是 SVD 矩阵分解，我们可以利用 sklearn 库中的 PCA 来实现隐层语义分析，并达到降维的目的。

1.2 KMeans 分类算法

K-means 算法是最为经典的基于划分的聚类方法，K-means 算法的基本思想是：以空间中 k 个点为中心进行聚类，对最靠近他们的对象归类。通过迭代的方法，逐次更新各聚类中心的值，直至得到最好的聚类结果。

1.3 用户画像

用户画像，又称人群画像，是根据用户人口统计学信息、社交关系、偏好习惯和消费行为等信息而抽象出来的标签化画像。构建用户画像的核心工作即是给用户贴“标签”，而标签中部分是根据用户的行为数据直接得到。

1.4 TF-IDF

TF-IDF(Term Frequency-Inverse Document Frequency, 词频-逆文件频率)。是一种用于资讯检索与资讯探勘的常用加权技术。TF-IDF 是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的

重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。我们在对用户实现了分类之后，可以再利用 TF-DF 对每一类用户寻找最重要的标签作为代表。

词频 (term frequency, TF) 指的是某一个给定的词语在该文件中出现的次数。这个数字通常会被归一化 (一般是词频除以文章总词数), 以防止它偏向长的文件。（同一个词语在长文件里可能会比短文件有更高的词频，而不管该词语重要与否。）

$$TF_w = \frac{count(word, doc)}{size(doc)}$$

逆向文件频率 (inverse document frequency, IDF) IDF 的主要思想是：如果包含词条 t 的文档越少, IDF 越大，则说明词条具有很好的类别区分能力。某一特定词语的 IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到。

$$IDF_w = log(\frac{num(doc)}{num(doc, word) + 1})$$

分母之所以要加 1，是为了避免分母为 0. 某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。

在这次的实验中，我们把分好的 N 个类的用户的所有的标签组成一个“doc”，这样我们就可以一共获得 N 个 doc，从而可以计算每个 doc 里面的每一个标签词的 TF-IDF 值。

1.5 “Stop-Labels”

我们的实验中，提到了一个停用标签的概念。考虑到许多如同，“音乐”、“电影”这样的标签，并不具有十分显著的标签信息。我们在进行实验的时候可以将其剔除。

2 实验