Chapter 7 文本分类与聚类（研究进展、现状&趋势）

文本挖掘（Text Mining）

- 从非结构或半结构化的文本数据中获取高质量的结构化信息的过程
- 目的是从未经处理的文本数据中获取有用知识或信息
- 任务
  - 核心任务
    - 文本分类（Text Classification）— 根据给定文档的内容或主题，自动分配预先定义的类别标签
    - 文本聚类 — 根据文档之间的内容或主题相似度，将文档 集合划分成若干个子集，每个子集内部的文档相似度较高，而子集之间的相似度较低
  - 概念/实体抽取
  - 情感分析
  - 文档摘要

文本分类、聚类

1. 文本表示（构建文本特征向量）

- 将无结构化的文本内容转化成结构化的特征向量形式，作为分类或聚类模型的输入
- 建立特征空间
  - 文本词袋（Bag of Words）模型
    - 每个文档被表示为一个特征向量，其特征向量每一维代表一个词项。 所有词项构成的向量长度一般可以达到几万甚至几百万的量级。
    - 不足：忽略了词与词之间的序列信息以及句子结构信息
    - 解决办法
      - 进行特征选择（Feature Selection）与特征提取(Feature Extraction)，选取最具有区分性和表达能力的特征建立特征空间， 实现特征空间降维
      - 进行特征转换 (Feature Transformation)，将高维特征向量映射到低维向量空间
      - 话题分析(Topic Analysis)
  - 向量空间模型（Vector Space Model）
    - 向量空间的每一维代表一个词项(词语或 N-Gram)，然后通过 TF-IDF 等方式就可以计算得到 文本在向量空间中的表示
    - 思想：如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力
    - 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。
    - TF-IDF
      - TF是词频(Term Frequency)，IDF是逆文本频率指数(Inverse Document Frequency)
        - 词频（TF）表示词条（关键字）在文本中出现的频率。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。
        - 公式：$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ 即：$TF_w = \frac{\text{在某一类中词条w出现的次数}}{\text{该类中所有的词条数目}}$
        - 其中 $n_{i,j}$ 是该词在文件 d_j 中出现的次数，分母则是文件 d_j 中所有词汇出现的次数总和；
        - 逆向文件频率 (IDF)：某一特定词语的IDF，可以由总文件数目除以包含该词语的文件的数目，再将得到的商取对数得到。如果包含词条的文档越少, IDF越大，则说明词条具有很好的类别区分能力。
        - 公式：$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}| + 1}$
        - 其中，$|D|$ 表语料库中的文件总数，这种idf_i 表示包含词条 t_i 的文件数目，如果该词语不在语料库中，就导致分母为零，因此一般情况下使用 $1+|\{j:t_i \in d_j\}|$
        - 即：$IDF = \log(\frac{\text{语料库的文档总数}}{\text{包含该词w的文档数}+1})$，分母之所以要加1，是为了避免分母为0
        - TF-IDF实际上是：TF * IDF — 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。
    - Ref: https://blog.csdn.net/asialee_bird/article/details/81486700

2. 特征降维

- 特征选择
  - 构造面向特征的评分函数，对候选特征进行评估，然后保留评分值最高的特征
  - 特征评分函数
    - 文档频率(Document Frequency, DF) — 在整个文本集合中，出现某个特征的文档的频率，DF 值低于某个阈值的低频特征通常为噪音特征或者信息量较小不具有代表性
    - 计算不同特征的类别区分度
      - 前提：给定事先标注了类别标签的文本集合
      - 信息增益(Information Gain) — 计算新增某个特征后信息熵的变化情况，用以衡量特征的信息量
      - 互信息(Mutual information) — 根据特征与类别的共现情况来计算特征与类别的相关度 =
        - 具体来说，词项 t 与类别 c 之间的互信息定义如下：$I(t,c) = \log \frac{P(t,c)}{P(t)*P(c)} = \log \frac{P(t \wedge c)}{P(t)*P(c)}$
        - 特征的互信息值越高，说明该特征与某个 类别的关联程度更紧密，用来进行分类的话区分效果就更好
      - 卡方统计（x^2 Statistics） — 计算特征与类别关联关系的方法，定义了一系列词项 t 与类别 c 之间共现或不同现的统计量（A、B、C、D）
        - 公式 — $\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C)(B+D)(A+B)(C+D)}$
      - 与 DF 相比，基于标注数据集合选取的特征更具区分性。对文本分类效 果提升显著，其中以卡方统计的表现最佳
- 特征转换（特征映射）
  - 主成分分析(Principal Component Analysis, PCA)
    - 计算特征变量之间的协方差矩阵，然后选择协方差矩阵特征值最大的若干个特征向量作为主成分
    - 利用这些特征向量，通过线性映射就可以将高维特征映射到低维空间中
  - 线性判别分析(Linear Discriminant Analysis, LDA)
    - 将高维特征向量映射到具有最佳区分度的低维空间，来达到压缩特征维度的效果
    - 保证转换后的表示具有最大的类间间距和最小的类内间距，意味着新的低维特征空间具有最佳的判别性

3. 话题分析

- 假设文档与词语之间存在潜在的语义关系，将文档看成不同话题上的分布，将每个话题看成不同词语上的分布，即话题通过分析文档话题作为文档特征表示
- 目标：利用大规模文档集合，自动学习话题表示，构建"文档-话题"以及"话题-词 之间的关系
- 代表技术
  - 潜在语义分析(Latent Semantic Analysis, LSA)
    - 通过矩阵奇异值分解(Singular Value Decomposition, SVD)对文档-词语的同现矩阵进行分解，得到"文档-话题"矩阵以及"话题-词语"矩阵。
    - 缺点：LSA 并没有对两个目标矩阵中的取值范围设定限制，不具备概率分布的良好属性。
  - 基于概率的潜在语义分析 (Probabilistic Latent Semantic Analysis, PLSA)
    - 引入概率统计的思想，PLSA 学习得到的"文档-话 题以及"话题-词语"矩阵具有较好的概率分布属性
    - 改进：更直观地计算文档-话题以及话题-词语之间的语义关系，同时也避免了 LSA 中 SVD 的复杂计算过程
    - 缺点：PLSA 无法较好对新文档估计话题分布
  - 隐狄利克雷分布(Latent Dirichlet Allocation, LDA)
    - 层次化的贝叶斯模型， 通过为文档的话题分布、话题的词语分布分别设置基于 Dirichlet 的先验概率分布，从而使 模型具有较好的泛化推理能力，可以为新文档自动估计话题分布
    - 改进：与 PLSA 利用 EM 算法进行 参数估计不同，LDA 可以采用更高效的 Gibbs 抽样法和变分推断法来进行参数估计
    - 于 LDA 提出很多新的主题分析模型
      - 考虑文档之间关系的 RTM(Relational Topic Model)
      - 考虑主题之间 相关性的的 CTM(Correlated Topic Model)
      - 考虑话题随 时间演变的 DTM(Dynamic Topic Model)
      - 考虑文档作者信息的 Author-Topic Model
- 进行话题分析的结果，既可以作为文档特征进行文本分类或聚类，也可以用来分析大规模文档集合中的话题分布与演化情况 — 重要应用：话题检测与跟踪(Topic Detection and Tracking, TDT)，面向新闻媒体，进行新话题发现以及已知话题跟踪
- 以上主题模型均可用来进行有效的话题检测与抽取，而 DTM 等动态主题模型 也可以得到同一主题在不同时期的变化情况。

2.1. 文本分类

- 基于规则的分类模型
  - 首在建立一个规则集合来对数据类别进行判断
  - 规则可以从训 练样本里自动产生，也可以人工定义
  - 模型 — 决策树(Decision Tree)、随机森林(Random Forest)、RIPPER 算法等
- 基于机器学习的分类模型
  - 贝叶斯分类器(Naive Bayes)、线性分类器（逻辑回归） 、支持向量机(Support Vector Machine, SVM)、最大熵分类器
  - 以 Boosting、Bagging 为代表的集成学习分类模型组合方法能够有效地综合多个弱分类模型的分类能力 — 在给定训练数据集合上同时训练这些弱分类模型，然后通过投票等机 制综合多个分类器的预测结果，能够为测试样例预测更准确的类别标签
- 基于神经网络的方法
  - 多层感知机（Multilayer Perceptron, MLP）
    - 包括多层感知机在内的文本 分类模型均使用了词袋模型假设，忽略了文本中词序和结构化信息。
    - 对于多层感知机模型来说，高质量的初始特征表示是实现有效分类模型的必要条件。
  - DNN
    - 基于 CNN 和 RNN 的文本分类模型输入均为原始的词序列，输出为该文本在所有类别上的概率分布
    - 词序列中的每个词项均以词向量的形式作为输入
    - CNN — 面向文本的卷积操作是针对固定滑动窗口内的词项进行的
    - RNN
      - 与 CNN 相比，RNN 能够更自 然地考虑文本的词序信息
      - 改进 — LSTM、GRU、BiLSTM等
    - Attention
      - 引入选择注意力机制 (Selective Attention)，可以让模型根据具体任务需求对文本序列中的词语给予不同的关 注度。
      - Transformer
      - BERT

2.2 文本聚类

- 基于距离的聚类算法
  - 首先通过相似度函数计算文本间的语义关联度
    - 余弦相似度
    - 皮尔森系数
    - 闵氏距离（曼哈顿距离、欧氏距离、切比雪夫距离）
  - 然后根据文本间的语义相似度进行聚类
  - 层次法（层次距离）和划分法（K-means）
- 基于概率模型的聚类方法
  - 主题模型（Topic Model）
    - PLSA
    - LDA
  - 对文本集合学习概率生成模型
  - 假假设每篇文章是所有主题（聚集）上的概率分布，而不是仅属于一个聚集

总结&展望

- 面向互联网文本的分类聚类
  - 传统的文本分类与聚类任务更多聚焦在文本自身的 分析上，在封闭小数据集合上优化分类模型
  - 如何恰当地利用互联网本文等异构信息，构建高效的适用于互联网文本的分类与聚类模型，是文本挖掘领域面临的重 要挑战。
- 神经网络文本分类模型优化
  - 如何建立可解释的神经网络分类模型
  - 如何降低模型学习复杂度
  - 如何利用有限的标注样例取得更好分类效果
- 基于神经网络的文本聚类模型 — 如何充分利用深度 神经网络的强大语义表示能力，设计有效的目标函数，建立基于神经网络的文本聚 类模型，是文本聚类所面临的挑战。