商品属性词提取

NLP：短语挖掘、关键词提取

目录

1 概念定义

商品属性词一般属于评论的关键词，所以属性词提取方法与关键词提取基本相同。关键词是能够表达文档中心内容的词语，常用于计算机系统标引论文内容特征、信息检索、系统汇集以供读者检阅。**关键词提取**是文本挖掘领域的一个分支，是文本检索、文档比较、摘要生成、文档分类和聚类等文本挖掘研究的基础性工作。

2 问题分析

从算法的角度来看，关键词提取算法主要有两类：**无监督关键词提取方法**和**有监督关键词提取方法**。

3 无监督

3.1 基于统计特征的关键词提取（TF,TF−IDF）

思想：利用文档中词语的统计信息抽取文档的关键词；

方法：用于反映一个词对于某篇文档的重要性，过滤掉常见的词语，保留重要的词语；

优缺点：泛化能力强，无需人工标注。

3.2 基于词图模型的关键词提取(PageRank,TextRank)

思想：首先要构建文档的语言网络图，然后对语言进行网络图分析，在这个图上寻找具有重要作用的词或者短语，这些短语就是文档的关键词；

方法：把文本的词看做图中的节点，通过边相互连接，权重高的节点作为关键词；

优缺点：抽取的词多为高频词，具有局限性；

3.3 基于主题模型的关键词提取(LDA)

LDA 也称三层贝叶斯概率模型，包含词、主题和文档三层结构；利用文档中单词的共现关系来对单词按主题聚类，得到“文档−主题”和“主题−单词”2 个概率分布。

思想：利用主题模型中关于主题分布的性质进行关键词提取；

方法：分词−>去停用词−>构建词袋模型−>LDA 模型训练

优缺点：文档太短不利于训练 LDA；没有考虑词语间的关系

3.4 自动短语挖掘框架 AutoPhrase

方法：AutoPhrase 提取短语+人工审核结果

优缺点：基于 Wikipedia 场景；提取准确率较高，人力成本低

3.5 关键词匹配

方法：根据商品属性词库直接匹配对应文本；

3.6 Word2Vec 词聚类（DBSCAN/k−means）

思想：对于用词向量表示的词语，通过 DBSCAN/k−means 算法对文章中的词进行聚类，选择聚

类中心作为文本的一个主要关键词，计算其他词与聚类中心的距离即相似度，选择 top K 个距离聚类中心最近的词作为关键词，而这个词间相似度可用 Word2Vec 生成的向量计算得到。

方法：分词->Word2Vec->DBSCAN/k-means 聚类

4 有监督

4.1NER

思想：根据现有的商品属性表，训练 NER 模型

方法：用{评论，商品属性}来训练 NER

优缺点：需要大量人工标注数据来训练模型；

4.2 信息增益关键词提取

4.3 卡方检验关键词提取

卡方是数理统计中用于检验两个变量独立性的方法，是一种确定两个分类变量之间是否存在相关性的统计方法，经典的卡方检验是检验定性自变量对定性因变量的相关性。

思想：根据标注数据，利用卡方检验选择特征

4.4 转化成多标签问题（人工标注，训练模型）

4.5 CRF

条件随机场(Conditional Random Fields，简称 CRF)是给定一组输入序列条件下另一组输出序列的条件概率分布模型。例如，在词性标注任务中，输入序列为一串单词，输出序列就是相应的词性。

方法：通常选择三个特征：属性词的前后位置单词作为特征；利用句法分析抽取核心字句，将句法结构作为特征；将词性和属性词与情感词之间的词距作为特征。训练 CRF 模型

4.6 词性+规则

思想：根据既定规则，我们可以把句子分词、词性标注，然后提取出名词和形容词动名词之类。

方法：分词-->词性标注-->用规则提取

优缺点：简单、快速；缺点：精度不高，多用于评论标签（tag）提取

5 总结

文本的关键词提取在基于文本的搜索、推荐以及数据挖掘领域有着很广泛的应用。同时在实际应用中，因为应用环境的复杂性，对于不同类型的文本，例如长文本和短文本，用同一种文本关键词提取方法得到的效果并相同。因此，在实际应用中针对不同的条件环境所采用的算法会有所不同，没有某一类算法在所有的环境下都有很好的效果。

相对于上文中所提到的单一算法，一些**组合算法**在工程上被大量应用以弥补单算法的不足，例如将 TF-IDF 算法与 TextRank 算法相结合，或者综合 TF-IDF 与词性得到关键词等。同时，工程上对于文本的预处理以及文本分词的准确性也有很大的依赖。对于文本的错别字，变形词等信息，需要在预处理阶段予以解决，分词算法的选择，未登录词以及歧义词的识别在一定程度上对于关键词突提取会又很大的影响。

6 参考

[1] 任远远，王卫平．中文网络评论的产品特征提取及情感倾向判定[J].计算机系统应用,2014,(11):22-27.DOI:10.3969/j.issn.1003-3254.2014.11.005.

[2] Liu, Xiaoming. Domain-Specific Ontology Construction from Hierarchy Web Documents[R]. Beijing:IEEE, 270 2011. 160-163.

[3] Yanhui, Zhu. Research on Extraction for Feature Words from Chinese Product Comments[R]. Changsha:IEEE, 2010. 205-209.

[4]http://manu44.magtech.com.cn/Jwk_infotech_wk3/article/2011/1003-3513/1003-3513-27-5-49.html#outline_anchor_11

[5]https://blog.csdn.net/asialee_bird/article/details/96454544#%E5%8D%81%E3%80%81%E6%80%BB%E7%BB%93

[6] https://zhuanlan.zhihu.com/p/40631031

[7] https://blog.csdn.net/Sakura55/article/details/85122966