文章编号：2095-2163(2020)09-0069-05 中图分类号：TP391.1 文献标志码：A

融合信息熵与多权 TF-IDF 的营销评论关键词提取算法

李 璐，何利力
（浙江理工大学 信息学院，杭州 310018）

摘 要：针对传统分词算法、传统提取关键词算法对现代营销活动中以客户为中心，分析客户评论，提取重要客户的需求具有局限性等问题，提出融合信息熵和多权 TF-IDF 关键词提取算法。该算法首先运用结合互信息和左右熵分词算法对标题、用户评论进行分词，产生新词；再运用 TF-IDF 算法抽取评论关键词、标题关键词，根据关键词的位置因子、词性因子、词长因子加以不同的特征权重，避免忽视标题和评论的不同重要性，提高结果精度；利用余弦相似度对两者的关键词进行相似度的比较，从而确定该评论的质量。实验结果表明：从互信息、左右熵、词语的位置，词性和词长几个方面考虑，可以提高提取关键词的效率，可以有效地筛选重要评论，为挑选重要客户提供了条件。

关键词：TF-IDF 算法；特征权重；互信息；左右熵；余弦相似度

Keyword extraction algorithm integrating information entropy and multi-weight TF-IDF

LI Lu, HE Lili
(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

【Abstract】 Aiming at the limitations of traditional word segmentation algorithms , traditional keyword extraction algorithms for customer-centric analysis of customer reviews and extraction of important customer needs in modern marketing activities, a fusion information entropy and multi-weight TF-IDF keyword extraction algorithm is proposed. The algorithm first uses the combination of mutual information and the left and right entropy word segmentation algorithm to segment the title and user comments to generate new words. Then the TF-IDF algorithm is used to extract the review keywords and title keywords based on the keyword's position factor and classification factor, which is added with different feature weights to avoid ignoring the different importance of the title and the comment and to improve the accuracy of the result. The cosine similarity is used to compare the similarity of the keywords of the two keywords to determine the quality of the comment. The experimental results show that considering the mutual information, left and right entropy, word position, part of speech and word length, the efficiency of extracting keywords can be improved, and important comments can be effectively screened, making it easier to select important customers.

【Key words】 TF-IDF algorithm; feature weight; mutual information; left and right entropy; cosine similarity

0 引 言

随着“互联网+”技术日趋成熟，基于“互联网+”营销企业需要根据不同属性对用户进行类别划分，为不同类别用户制定不同的营销策略，评论的质量可作为用户的一个属性。评论属于自然语言，人为对评论的质量评估，是可行的，但评论数量过大，人为评估速度慢，无法满足现营销企业的需求。自然语言处理针对结构复杂的文本信息进行处理，其中关键词的提取是基础与核心技术，在检索信息、文本分类、信息匹配、话题跟踪、自动摘要、人机对话等领域有广泛的应用[1-3]。

在自然语言处理领域中处理提取评论关键词的方法大致可以分为两类：监督学习，无监督学习[4]。监督学习是从特定的训练数据集训练出函数模型，根据函数模型判断该词语是否属于关键字类别，对训练集的要求较高，通常需要人工预处理。在无监督学习中，无法预知样本类型，需要根据样本数据间的内在结构对样本集进行聚类，使同一类别数据差距最小化，不同类别数据差距最大化[5]。常见的主流无监督关键字提取方法可以分为基于 TF-IDF 数值统计的关键词提取、基于 LDA 主题模型的关键字提取、基于词图模型的关键字提取 3 种类型[6-7]。上述方法都有各自的优点和局限性。

本文主要针对 TF-IDF 展开相关研究，综合考虑评论信息中词语的位置、词性、词长 3 种影响因子，对每种影响因子赋予一定的权重，最后加权得到最终的特征权重，获取权重最大前 5 的词语作为该短文本的关键词。通过余弦相似度来衡量评论与标题关键词

基金项目：国家重点研发计划(2018YFB1700702)。

作者简介：李 璐(1996-)，女，硕士研究生，主要研究方向：智能软件与数据处理；何利力(1966-)，男，博士，教授，博士生导师，主要研究方向：数据分析、企业智能。

通讯作者：何利力 Email：1923400237@qq.com

收稿日期：2020-07-22

的相似度,获取重要评论。该方法可识别垃圾评论、重要评论,可用于企业对用户某一属性的衡量。

1 相关技术

相关技术研究包括 TF-IDF、信息熵、Trie 树、词语的权重、余弦相似度这 5 个方面。设定一个文本集合 D, 集合中包含 N 个文本,每个文本都包含标题 title 和评论 comment 两部分[7]。comment 内容是由评论句子组成,评论句子是由多个词语组成。

1.1 TF-IDF 算法

TF-IDF 是常见的加权算法,通常用于资源检索与数据挖掘等方向,衡量文本集中一个特征词对包含该特征词的文本的重要程度,优于其它算法[6]。TF-IDF 是 TF 与 IDF 的乘积,TF-IDF 的词条提取函数如式(1):

$$W_{tf-idf} = TF(i) \times IDF(i), \tag{1}$$

其中, W_{tf-idf} 表示第 i 个词语的 TF - IDF 值, $TF(i)$ 表示该词的词频。主要思想是:如果该特征词 i 在该文本中出现的次数较多, $TF(i)$ 越大,则表明该词可能会较好地描述了该文本的主要信息,计算如式(2):

$$TF(i) = \frac{n_i}{n}, \tag{2}$$

其中, n_i 为该词 i 出现的次数, n 为所有关键词的总数。

$IDF(i)$ 表示逆文档频率,若包含该词 i 文档数越少, $IDF(i)$ 越大,说明该词 i 具有良好的类型区分作用,计算如式(3):

$$IDF(i) = \log\left(\frac{N}{df(i) + 1}\right), \tag{3}$$

其中, N 为文档总数, $df(i)$ 是为文档出现词语 i 的文档数。

TF-IDF 算法表明:在文本 comment 中出现频率足够高,而在整个文本集合 D 的其他文档中出现频率足够低的特征词是区别该文本 comment 最关键的词语[7-8]。TF 词频代表同类文本特征,不同类别文本的特征由 IDF 来表示。IDF 主要用于调整 TF,抑制噪声加权,但 TF-IDF 的结构过于简单,无法有效地反映单词的重要性和特征单词的位置分布,并且调整权限功能不是有效的,因此 TF-IDF 方法的准确性不高,且 TF-IDF 算法没有体现特征词的位置信息、词性、词长的重要性。对于一篇文档而言,不同结构的内容体现的信息是不同的,即权重也应按照不同的结构特征来分配,避免忽视文本结构问题[9]。特征词在不同的位置、词性、词长对文本内容的反映程度不同,其权重计算方式也应有所不同。因此,应该给文档中不同位置、词性、词长的特征词赋予不同的系数,并乘以特征词的 TF-IDF 值,以增强文本表达的效果。

1.2 互信息与信息熵

互信息反映两个词语的凝聚力,互信息的计算如式(4):

$$PMI(x,y) = \log\frac{p(x,y)}{p(x)p(y)}. \tag{4}$$

其中, $p(x,y)$ 为词语 x,y 的联合分布概率; $p(x)$, $p(y)$ 为词语 x,y 边缘分布概率; $PMI(x,y)$ 的单位为 bit。

根据互信息挑选的预选词,利用信息熵确定该预选词为新词。信息熵是一个具体事件发生所带来的信息[10],描述信息源的不确定度,熵是该预选词的所有可能取值,即所有可能发生预选词组合所带来的信息量的期望[11],来表示预选词的自由度。对于一个预选词所有可能的组合 X, 其信息熵为公式(5)

$$H(X) = -\sum p(x)\log_2 p(x). \tag{5}$$

其中, $p(x)$ 是 x 在系统事件中出现的概率。熵越大,则该预选词大概率为一个新词。

1.3 Trie 树

Trie 索引树是一种数据结构,是由非线性结构形式表示的键树,由首字散列表和字典索引树结点两部分组成,通常用于文本词频统计[12]。Trie 树可保存键值对映射关系,但 key 必须是字符串,除根节点,其它节点都只包含一个字符,每个节点的孩子节点包含的字符都不相同。其核心思想是通过最长公共前缀迅速查询到结果,空间换时间,降低时间复杂度。通过 Trie 树来存储和计算词语的信息熵,用于筛选出新词。

1.4 词语的权重

针对 TF-IDF 算法的局限性,引入词语权重。词语权重分为词语位置权重,词性权重,词长权重。活动的标题的 title 一般能概括活动的主要内容,则出现在标题中的词语成为关键词的概率更大;在评论中出现词语可能会反映该活动的隐藏关键词或活动相关关键词,则评论的词语也应该适当重视[13-14]。特征词位置的权重设置见表 1。

表 1 位置权重设置

Tab. 1 Location weight setting

位置	权重名	权重设置
标题	W_{ltitle}	7
评论	$W_{lcomment}$	2
其他	W_{lother}	1

中文中的词性可分为实词和虚词两类。实词一般包含：名词、动词、形容词、代词、数词、量词等；虚词一般包含：介词、连词、叹词、助词等[13]。关键词的词性通常是以名词或名词性短语为主，其次是动词、副词和其他修饰词。特征词的词性权重设置见表 2。

表 2　词性权重设置

Tab. 2　Part of speech weight setting

名词性	W_{cnonce}	5
动词性	W_{cverb}	2
形容词、副词	W_{cadj}	2
其他	W_{cother}	1

关键词过短无法体现包含信息，关键词过长，包含信息越多，则表示该关键词可以再次切分。研究表明，关键词的词长一般在[2,7]之间，词长过长过短需要过滤[15-16]。词长权重计算公式(6)：

$$W_{len} = \frac{i_{len}}{avg(len)}. \tag{6}$$

其中，i_{len} 是第 i 个词语的词长，$avg(len)$ 是平均词长。

综合上述多特征权重，词语权重计算公式(7)：

$$W_{word} = \alpha W_l + \beta W_c + \gamma W_{len}. \tag{7}$$

其中，W_{word}、W_l、W_c、W_{len} 分别为词语权重、词语的位置权重、词性权重、词长权重，α、β、γ 为系数分别为 0.6、0.3、0.1。

1.5　余弦相似度

本文目标是评测用户评论质量，需要与一个相对标准指标进行对比。余弦相似度是通过测量两类关键词向量的夹角余弦值来度量它们之间的相似性。余弦相似度计算公式(8)：

$$\cos\theta = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}. \tag{8}$$

评论中关键词的权重作为 A 向量，标准关键词权重作为 B 向量，计算两者之间的 $\cos\theta$。

2　实验及结论

2.1　算法步骤

提取关键词算法步骤为：

(1)文本获取：利用爬虫爬取活动标题、用户评论，写入文本中保存。

(2)文本预处理：清除文本中的噪声，例如：文本中的空格，表情符号，特殊符号等。

(3)分词：将文本分为标题、评论两部分，同时对这两部分进行分词，分词结果分为标题分词结果集和评论分词结果集，本文采用结合字典树和信息熵对文本进行分词。

(4)停用词过滤：由于停用词的普遍性，通常自身没有特定的意思，对文本主题的表达能力低。例如“的”，“啊”，“然后”，“哈哈哈”等词语以及标点符号，过滤停用词，消除对关键词提取的干扰。

(5)词性/词长过滤：对词性为语气助词、介词、连词、拟声词等词语过滤，过滤词长小于 2 大于 7 的词语。过滤这些词语可提高工作效率，避免增加工作量。

(6)利用 TF-IDF 算法计算词语的 W_{tf-idf}。

(7)根据式(7)计算词语的权重 W_{word}。

(8)计算词语的最终权重 $W = W_{tf-idf} * W_{word}$。

(9)根据余弦相似度公式计算评论与标题的相关度。

2.2　实验及结果分析

本文实验数据来自某微信公众号的活动评论。评论共有 91 120 条，去除只含表情、评论过短的评论，剩余评论为 83 680 条。本文中该公众号名都用“XXXX”来表示。

2.2.1　分词效果对比

互联网营销活动的标题和评论包含新词，传统分词算法可能无法实现新词的提取。本实验利用互信息和左右熵，以 Tire 树为数据结构提取新词。互信息是一个词语中包含的关于另一个词语的信息量，即两个词共同出现的概率。左右熵衡量预选词的自由度。左右熵越大，说明该预选词越有可能是独立词语。通过传统分词算法和基于互信息和左右熵的分词算法的分词结果见表 3。

表 3　两种分词对比表

Tab. 3　Comparison of two word segmentation

原句	传统分词	互信息和左右熵分词
美丽西子湖畔！魅力杭州出 XXXX 欢迎来杭赏美景品 XXXX	美丽/ 西子湖畔/ ！/ 魅力/杭州/ 出/平和/味道/欢迎/来/杭赏/美景/ 品/ 平和/味道	美丽/ 西子湖畔/ ！/ 魅力/ 杭州/ 出/ XXXX/ 欢迎/ 来/ 杭赏/ 美景/ 品/ XXXX/
杭州美丽古都，创新之城，智慧之都，心灵之城，XXXX，寻味杭州	杭州/美丽/古都/，/创新/之城/，/智慧/之都/，/心灵/之城/，/平和/味道/，/寻味/杭州	杭州/美丽古都/，/创新之城/，/智慧之都/，/心灵之城/，/XXXX/，/寻味杭州

从表 3 可以看出,传统分词将“XXXX”分成“XX”和“XX”两个词,将“寻味杭州”分为“寻味”和“杭州”,基于互信息和左右熵分词算法将“XXXX”、“寻味杭州”作为独立词语,这两个词语是与文本源——某公众号的活动相关。由此可见传统的分词算法无法识别新词,会导致 W_{tf-idf} 不准确。

2.2.2 关键词提取效果对比

采用准确率(*Precision*)、召回率(*Recall*)和 F_1 值来衡量关键词提取算法的优劣。准确率是指预测正确的样本数除以总样本数,召回率是实际为正确的被预测为正确样本的概率,则综合准确率和召回率这两个指标提出了 F_1 值,若 F_1 比较高,则说明该算法效果较好[5]。

准确率计算公式如式(9):

$$P = \frac{num_{correct}}{num_{total}}. \tag{9}$$

其中, $num_{correct}$ 表示符合主题的关键词数量, num_{total} 是关键词总量。

召回率计算公式如式(10):

$$R = \frac{num_{correct}}{num_{actual}}. \tag{10}$$

其中, num_{actual} 表示文本真实关键词数量。

F_1 值综合准确率和召回率两个指标,计算公式如式(11):

$$F_1 = \frac{2 \times P \times R}{P + R}. \tag{11}$$

通过准确率、召回率和 F_1 值对传统 TFIDF 算法和多权 TF-IDF 算法进行对比,结果见表 4。

表 4 两种算法指标对比

Tab. 4 Comparison of two algorithm indicators %

算法	准确率	召回率	F_1 值
经典 TF-IDF 算法	63.6	50.6	56.4
多权 TF-IDF 算法	81.8	69.2	75.0

本文通过计算基于传统 TF-IDF 算法和多权 TF-IDF算法提取的关键词权重与标题关键词权重的余弦相似度进行对比,提取余弦相似度排名前 5 的评论,表 5 为两种算法得到的不同重要评论排名结果。

这次文本标题是“网红带你寻味杭州! 抢! 千份好礼限时 8 h!”,通过人工标注关键词为:“寻味杭州”,“网红”,“好礼”,“8 h”,“千份”,活动标题的隐藏关键词“XXXX”,评论关键词与标题关键词进行比对,结果表明:多权 TF-IDF 算法优于传统 TF-IDF算法。

表 5 重要评论对比

Tab. 5 Comparison of important comments

排名	传统算法	多权 TF-IDF 算法
1	跟随网红寻味杭州,品鉴、品尝欢乐“XXXX”。	跟随网红寻味杭州,品鉴、品尝欢乐“XXXX”。
2	美丽的杭州,与 XXXX 发现不一样的美,寻味杭州 666	杭州美丽古都,创新之城,智慧之都,心灵之城,XXXX,寻味杭州
3	寻味杭州,XXXX,你值得拥有![得意]	最喜欢你,XXXX,寻味杭州! 不光是福利,还有视觉上的震撼。
4	心动不如行动,好想借着“XXXX”,去寻找网红小哥口中所描述的那个杭州[呲牙],期待中…[拳头]	杭州.中国八大古都之一,钱塘江畔,西子,最忆是杭州! XXXX,寻味杭州! 好想去看看!
5	XXXX,福利暖人心,好期待啊。	XXXX,网红,最美的杭州慢游。

3 结束语

本文针对现代营销活动中客户为中心,分析客户评论,提取重要客户的需求,提出利用互信息和信息熵结合的分词算法,互信息决定该词语是否成为预选词,左右熵决定该预选词是否为独立词语,在上述分词算法的基础上,融合词语的位置、词性、词长等多种因素对 TF-IDF 算法进行了改进,对每个影响因素分配相应的权重,加权处理,最后得到词语权值,取权值最大的 5 个词语作为评论的关键词,以人工标注的关键词为标准,对比两种算法,发现本文算法效果良好,可筛选出重要评论,获得重要客户,值得推广应用,也可应用到其他场景,如微博留言、帖子评论,可以有效的排除网络水军垃圾发言等。在研究的过程中也发现了一些不足和缺陷,在本文忽视了特征词的语义信息对关键词提取的影响,未来可对中文语义进行深入研究:中文语言中有许多词语存在相近语义或者多种语义。语义相同,词语不同的关键词,会被筛除,导致算法具有局限性,因此,研究不同特征词语义信息对提升关键词提取效果的影响是具有重要意义。

参考文献

[1] 杨凯艳. 基于改进的 TF-IDF 关键词自动提取算法研究[D]. 湘潭:湘潭大学,2015.

[2] Yi-Hui Chen, Eric Jui-Lin Lu, Meng Fang Tsai. Finding keywords in blogs: Efficient keyword extraction in blog mining via user behaviors[J]. Pergamon, 2014, 41(2): 663-670.

(下转第 76 页)