广告中违法关键词的识别

**摘 要：**

**【目的】**针对目前违法广告泛滥的问题，提出计算机对广告中违法关键词的识别方法。

【方法】专家依据新修订的《广告法》提取广告违法种子词汇，再利用同义词林扩充种子词汇、语义相似度过滤扩充词汇，从而得到广告违法关键词库；基于构建好的广告违法关键词库，在字符串匹配的基础上，考虑上下文词性序列的影响，最终识别出违法关键词。

【结果】实验表明，该方法基于扩充后的违法关键词库，能够有效的识别出广告中的违法关键词，综合评价指标F值达到59.91%。

【局限】该方法针对违法广告中含禁用词的问题，未考虑广告中句子违法的情况。

**关键字：**违法广告识别 广告法 词性序列 词库

# 1引言

随着广告行业迅速的发展，违法广告泛滥的问题日益严重，并严重影响到了消费者的利益和社会的稳定。2015年9月1日起，新修订的《中华人民共和国广告法》（简称《广告法》，下同）正式施行。这是《广告法》实施20年来的首次修订。此次修订从管理的层面看将更加规范和严格，对虚假广告、未成年人代言等都做出了详细的规定，保证了广告行业的有效监管和规范[1]。因此，本文从广告主、广告发布者、广告经营者角度出发，根据新修订的《广告法》实现对广告中违法关键词的识别，有效的促进了广告行业的健康发展。

目前，对于文本广告的研究主要集中在判断一则广告是否违法。文献[2]基于贝叶斯分类原理，把广告分为正常广告和违法广告，通过对贝叶斯算法的改进，有效提升了违法广告的识别度。文献[3]提出针对工商监管领域的违法广告鉴别算法，采用基于关键词匹配和概率潜在语义模型相结合的方法对违法广告进行识别。以上方法虽然有效的识别出了违法广告，但未明确指出广告中违法的词语。从广告主、广告发布者和广告经营者角度出发，若能指出广告的具体违法内容，有助于其快速适应新修订的《广告法》。因此，本文提出对广告中违法关键词的识别方法。

广告中违法关键词的识别，可以归纳到信息检索研究领域的敏感词检测[4]。敏感词检测涉及到敏感词库、过滤机制这两个部分。对于敏感词库，百度、腾讯、360等公司都公布了较为全面的敏感词库。对于敏感词过滤机制的研究，主要分为两方面。一方面是对字符串的模式匹配研究。文献[5]介绍了几种常用的字符串模式匹配算法。它们分别是：朴素的模式匹配算法[6]、KMP匹配算法[7]、BM匹配算法[8]等。上述算法的核心都是字符串的精确匹配，其改进方向都是注重计算效率的提高。但广告违法关键词的识别对检测效率要求不高，主要关注违法关键词检测的准确率。另一方面是为了规避检测，敏感文本中夹杂了特殊符号，导致过滤中模糊匹配能力不足的问题。文献[9]针对此问题，提出基于本体[10]的敏感词过滤方法，有效的识别了敏感词的不同形式。而广告文本中人工干扰[11]这种问题较少发生，使用基于字符串匹配的方法就可以识别违法关键词，但此方法忽略了词语在不同上下文中拥有不同含义的情况。本文在简单的字符串匹配基础上，考虑文本中上下文词性序列的影响，对广告违法关键词进行识别。

综上所述，本文借鉴处理敏感词检测问题的核心思想。通过构建广告违法关键词库，利用上下文的词性序列，识别出广告违法关键词。目前尚未有专业的广告违法关键词库公布出来。因此，本文通过专家研究广告法提取出种子词汇，利用哈工大的同义词林[[1]](#footnote-1)扩充种子词汇、语义相似度过滤，得到广告违法关键词库。在对违法关键词识别过程中，本文在字符串匹配的基础上，利用上下文的词性序列识别违法关键词。

# 2方法设计

对于广告违法关键词的识别，分为两步：第一，构建违法关键词库；第二，利用构建好的词库，采用字符串匹配、词性序列过滤的方法识别出广告的违法关键词。

## 2.1构建词库

研究认为，违法关键词库是识别广告违法关键词的基础，其的好坏能够直接影响最终的识别效果。然而，目前尚未有专业的违法关键词库，所以构建一个合理的违法关键词库很有必要。构建违法关键词库分为三步：第一步，从广告法中提取种子词汇；第二步，通过利用《哈工大同义词林扩展板》扩充种子词汇，使用基于知网的语义相似度过滤词汇；第三步，种子词汇和扩充词汇构成违法关键词库。

### 2.1.1提取种子词汇

广告法是判别广告中是否有违法关键词的主要依据。法律法规以规则的形式存在，而规则分解成一系列的条件。根据条件，专家从中提取出种子词汇。其具体关系如图1所示：



图1法律法规关系

通过专家解读广告法，把提取出的规则分两类：客观规则和主观规则。主观规则指从法规中无法提取出违法关键词，而此类违法行为只能通过语义感知、现实调查等办法判断。例如，酒类广告规定不得出现饮酒的动作，但是文本广告中无法准确描述哪些词表示“饮酒动作”的含义，只能通过人工语义感知的方法来判断。所以，主观规则不在本文的讨论范围内。而客观规则指专家通过研读广告法，可以从中提取出种子词汇的规则。

客观规则是对违法行为的描述，由若干个条件组成。其表现形式有两种：第一种由一个条件表示，如图1的条件1。条件1表示所有类别广告需遵守的规定。第二种是由广告类别和条件表示，如图1的广告类别和条件2。广告类别代表某类广告，条件2表示这类广告下需遵守的规定。例如，针对所有类别广告的法律：广告不得使用“国家级”、“最高级”、“最佳”等用语。其可以描述为，如果广告中含有“最佳”、“最好”、“顶级”等极限用语，就违反了广告法。而针对某类广告的法律：保健食品广告不得含有表示功效、安全性的断言或者保证的内容。其可以描述为，如果是保健食品广告，并且广告中含有“治愈”、“安全无副作用”等词，则违反了广告法。综上所述，本小结的主要工作：通过把法律法规转换为规则，分解成条件，根据条件提取出种子词汇。

1.所有类别广告下的种子词汇

专家研究广告法，把客观规则中所有类别广告需遵守的规定提取出6大类种子词汇。它们分别是：极限用语、时限用语、权威性词语、欺骗消费者的词语、敏感用语、专利用语。其解释及举例如表1所示：

表1 所有类别广告下的种子词汇

|  |  |  |
| --- | --- | --- |
| 名称 | 依据 | 种子词汇举例 |
| 极限用语 | 广告法第九条第三项不得使用“国家级”、“最高级”、“最佳”等用语 | 1. 与“最”有关：最赚、最优、最优秀、最好等； 2. 与“一”有关：第一、 中国第一、唯一等； 3. 与“级”有关 ：国家级、最高级、世界级等词语； 4. 与“首/家/国”有关： 首个、首选、独家、独家配方、首发、全网首发、全国首发等词语； 5. 与品牌有关：大牌、金牌、名牌、王牌、领袖品牌等； |
| 时限用语 | 广告法第八条规定各类促销活动中，时间须有具体活动日期 | 例如“随时结束”、“仅此一次”、“随时涨价”、“马上降价”、“最后一波”等这些无法确定时限的词语， |
| 权威性词语 | 广告法第九条第二项规定不得使用或者变相使用国家机关、国家机关工作人员的形象 | “国家某领导人或机关推荐”、“国家某机关专供、特供” “专家推荐”、“质量免检”、“老字号”、“中国驰名商标”等与权威有关的词语 |
| 欺骗消费者的词语 | 广告法第二十八条定义， 广告以虚假或者引人误解的内容欺骗、误导消费者的，构成虚假广告。 | 1. 与虚假有关：“史无前例”、 “祖传”、“100%”等； 2. 与欺诈有关 ：“点击领奖”、“恭喜获奖”、“全民免单”等； 3. 涉嫌诱导消费者 ：“秒杀”、“再不抢就没了”、“不会更便宜了”、“万人疯抢”、“卖/抢疯了”等； |
| 敏感用语 | 广告法第九条第八项规定：广告不得含有淫秽、色情、赌博、迷信、恐怖、暴力的内容。 | 这类内容为敏感内容，目前有专业的敏感词库。如迷信用语：“带来好运气”、“增强第六感”、“化解小人”、“招财进宝”、“健康富贵”、“提升运气”等； |

2.某类广告下的种子词汇

专家通过分析客观规则中某类广告需遵守的规定，把其分为9大类广告。其分别是：一般食品、化妆品、保健食品、酒类、医疗药品器械、投资理财、教育以及培训、房地产、农药。其各类的种子词汇如表格2所示：

表2 某类广告下的种子词汇

|  |  |  |
| --- | --- | --- |
| 广告种类 | 依据 | 种子词汇举例 |
| 一般食品（不含保健食品） | 广告法第十七条 | “清热解毒”、“清热袪湿”、“滋阴补阳”、“减肥”等医疗用语 |
| 化妆品 | 广告法第十七条 | “祛斑”、“去除皱纹”、“纯天然”等易与药品相混淆的用语 |
| 保健食品 | 广告法第十八条 | 1. 不得使用“具有\*功效或长期使用会达到\*效果”、“安全无副作用”等词语； 2. 不得使用“化解死细胞”、“瘦身”等效果的词语； 3. 不得使用“预防癌症”、治疗等人类尚未攻克的疾病的用语；不得使用商品是“健康必需品”的用语。 4. 食品广告中涉及特定功效的，不得利用“专家”、“学者”、“某机构”的名义或者形象做宣传。 5. 食品广告不得含有明示或者暗示可以“替代母乳”等词语。 |
| 酒类 | 广告法第二十三条 | 明示或者暗示饮酒有“消除紧张、焦虑”、“增加体力”等功效的词语。 |
| 医疗药品器械 | 广告法第十六条 | 1. 医疗器械通用名称除符合规则规定的相应要求外，不得含有“型号、规格”、“人名、企业名称、注册商标或者其他类似名称”的词语； 2. 不得含有“有效率”、“治愈率”等的用语；不得含有排他性词语，如“独占鳌头”、“某某专有”等。 |
| 投资理财 | 广告法第二十五条 | 1. 不得对含有“返利”、“无风险”、“保收益”、“投资回报率高达某某”等对未来收益做出保证性承诺的词语； 2. 不得利用“学术机构”、“\*\*协会”、“专业人士”等的名义或者形象作推荐、证明。 |
| 教育以及培训 | 广告法第二十四条 | 1. 不得含有“通过率”、“保过”、“过线率”、“名校直升”等对对培训的效果做出明示或者暗示的保证性承诺的词语； 2. 不得含有“有命题组成员”、“考试机构”等参与培训的暗示性词语； 3. 不得利用“\*\*机构”、“\*\*协会”等的名义或者形象作推荐、证明。 |
| 房地产 | 广告法第二十六条 | 1. 不得含有“紧邻\*\*”、“\*\*分钟直达\*\*”、“坐拥\*\*”、“学区房”、等利用规划或者建设中的交通、商业、文化教育设施以及其他市政条件作误导宣传的词语； 2. 不得含有“升值”、“年回报率”、“稳定回报”等做出升值或者投资回报的承诺的词语； 3. 不得含有“最低价”、“低至\*\*元起”等违反国家价格管理规定的词语； |
| 农药类 | 广告法第二十一、二十七 | 1. 不得含有不科学地表示功效的断言或者保证，如：“无害”、“无毒”、“无残留”、“保证高产”等； 2. 不得含有“无效退款”、“保险公司保险”等承诺； 3. 不得利用“\*\*专家、教授”、“机构推荐”等的名义或者形象作证明； 4. 不得含有“显著药效”、“发芽率”等说明有效率的词语。 |

### 2.1.2扩充词库

在2.1节中，通过专家提取的方法，已经得到了一个由种子词汇构成的违法关键词库（以下简称STOP）。为了实现违法关键词库的扩充，本文提出两层过滤算法完成这一过程，具体步骤如图2所示：



图2 STOP词库扩充

1. 在第一层中，利用《哈工大同义词林扩展板》中词语的同义词关系，将STOP中没有的同义词、相关词等词语加入词库中。
2. 在第二层中，通过基于知网计算语义相似度的计算方法，将第一层中不符合要求的词语进行过滤，最大限度的保持扩展结果的准确性。

#### **1.同义词扩充词库**

第一层的同义词扩充词库，利用《同义词词林扩展板》。哈工大的同义词林扩展板是由哈工大社会计算与信息检索研究，一共包含77343条词语。其采用的编码方式如下表3所示：

表3词语编码表

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 编码位 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 符号举例 | D | a | 1 | 5 | B | 0 | 2 | = \ # \ @ |
| 符号性质 | 大类 | 中类 | 小类 | | 词群 | 原子词群 | |  |
| 级别 | 第一级 | 第二级 | 第三级 | | 第四级 | 第五级 | |  |

表中的编码位是按照从左到右的顺序排列。第八位的标记有3 种，分别是“=”、“#”、“@”。 “=”代表“相等”、“同义”。末尾的“#”代表“不等”、“同类”，属于相关词语。末尾的“@”代表“自我封闭”、“独立”，它在词典中既没有同义词，也没有相关词。根据最后一位标记的特点，先筛选出第八位的标记是“=”、 “#”的，即词典中表示同义词或者相关词的一行，记为TYCL集。

基于同义词扩充的方法：构建词集A，词集A是STOP与TYCL的交集，得到STOP集中词语的同义词或相关词。

#### **2.语义相似度过滤**

为了提高扩展词语的准确率，在第二层中设计了过滤词集A中的违法关键词汇的方法。通过计算词语之间的语义相似度，将符合条件的词语保留在扩展词库中，将不符合条件的词语进行过滤。

对于种子词汇和其同义词或相关词之间的语义相似度sim值，本文使用文献[12]提出的基于知网、面向语义、可扩展的相似度计算新方法。根据文献[12]可知，语义相似度范围为[0，1]。一个词语与本身的语义相似度为1，在任何上下文中不可相互替代的两个词语相似度为0。根据实验结果，本文设置阈值simmin=0.95。它代表的是语义相似度的最小值，若sim小于simmin则在词集A将其过滤。反之sim大于simmin，则在词集A中将其扩充到STOP中。扩展后的STOP称作STOPWORD。

## 2.2识别违法关键词

违法关键词多为一些形容词、名词、动词等修饰性词语，很少会用干扰符，或者改写来规避检测。因此基于匹配的方式就可以识别到此类违法情况。但是基于字符串匹配的方法，忽略了相同的词在不同的上下文中代表着不同的语义的情况。基于此，为了提高违法关键词识别的效果，本文提出一种基于上下文词性序列的匹配方法。

违法关键词在不同的上下文词性中拥有不同的含义。例如，“为了/p您/rr的/ude1第一/m次/qv使用/vn”和“也/d是/vshi每年/r夏天/t欧洲/ns销量/n第一/m的/ude1防/v晒/v产品/n”这两句广告语。简单的字符串匹配认为两句中的“第一”都是违法关键词。但是，只有第二句中的“第一”是违法关键词。从词性标注得出，“第一”的上下文词性是不同的。

因此，本文提出基于词性序列的违法关键词识别算法：给定两种构造好的词性序列，分别判断某类广告文本是否包含这两种词性序列，如果有则判定该词是违法关键词，否则该广告不包含该违法关键词。在算法中，两种词性序列分别代表：违法关键词左边词语词性和它本身词性；违法关键词右边词语词性和它本身词性。具体流程如图3所示：



图3基于词性序列的违法关键词识别算法

具体步骤如下所述：

1. 一则广告文本，首先对其预处理即分词、词性标注、去除标点符号等；
2. 对预处理的广告文本和违法关键词库进行字符串匹配，得到违法关键词的位置、它本身和左右词语的词性构成的两种词性序列。
3. 利用上一步得到的两种词性序列、违法关键词的位置。从该位置左边匹配是否包含左词性序列，右边匹配是否包含右词性序列。若左右词性序列都包含，该词是违法关键词。否则该广告没有违法关键词。
4. 最后，输出这则广告的违法关键词。

# 3实验

## 3.1实验数据与评价

在集成开发环境myeclipse下，本文使用java语言实现了构建词库、违法关键词识别方法。其中分词和词性标注使用中国科学院计算技术研究所提供jar包[[2]](#footnote-2)。通过研读广告法，专家提取种子词汇882个，经过实验扩充后违法关键词库达到1781个。

为了验证本文方法的有效性，利用爬虫软件抓取某电商网站化妆品类目下的广告语，选取其中的240篇违法广告作为本文实验数据[[3]](#footnote-3)。本文以多组人工标注、结果交叉验证的方式，标注出240篇违法广告的违法关键词语。

目前对于识别广告中违法关键词语的研究没有统一的评价标准，本文使用较为广泛的准确率P、召回率R以及F值来衡量识别效果。其中，用KA 表示人工标注的违法关键词集合，KB表示算法识别出的违法关键词集合。具体公式计算如下：

（1）

（2）

（3）

## 3.2 结果和分析

此外基于相同的实验环境以及实验数据，本文基于STOP、STOPWORD词库实现了基于字符串匹配、上下文词性序列的违法关键词识别方法，并对这四种实验结果进行了分析比较。

由表4、表5可以看出，基于扩充词库即STOPWORD词库的违法关键词识别效果优于基于种子词库即STOP词库的；基于上下文词性序列的违法关键词识别效果优于基于字符串匹配的；4种方法的召回率都高于准确率。由图四可以看出基于STOPWORD词库的上下文词性序列识别方法的F值最高；基于STOP词库的字符串匹配识别方法的F值最低。

表4 STOP词库的违法关键词识别效果

|  |  |  |
| --- | --- | --- |
| 评价指标 | 基于字符串匹配 | 基于上下文词性序列匹配 |
| 准确率 | 38.92% | 43.38% |
| 召回率 | 58.96% | 54.73% |
| F值 | 46.88% | 48.39% |

表5 STOPWORD词库的违法关键词识别效果

|  |  |  |
| --- | --- | --- |
| 评价指标 | 基于字符串匹配 | 基于上下文词性序列匹配 |
| 准确率 | 46.78% | 52.81% |
| 召回率 | 72.41% | 69.23% |
| F值 | 56.84% | 59.91% |

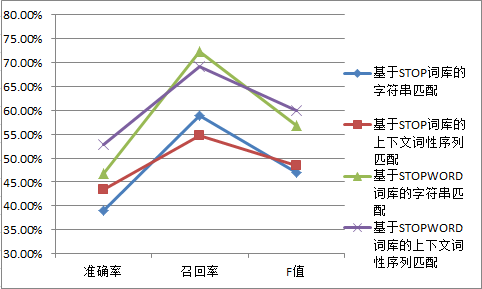


图 4 实验结果对比图

综上所述，经过扩充后的词库，有效的提高了违法关键词识别效果。基于上下文词性序列匹配的方法，考虑了相同的词在不同的上下文中代表着不同的语义的情况，有效的提高了综合评价指标F值。但通过表5可知 ，目前对于违法关键词的识别方法准确率仍然不是很高，存在一定误差，仍存在改进之处。经分析误差主要来源于以下几点：

1. 分词带来的误差。例如，“最好”在使用中科院的ICTCLAS分词时把它分成了“最”、“好”两个词，但其实“最好”在人工标注时表示为违法关键词。
2. 实验数据不易获得，导致数据量太少。
3. 违法关键词许多是由短语构成，比如“最重要”、“强力推荐”等。人工标注时按照短语形式标注，但计算机评价时把本算法与人工完全相同的违法关键词语计入评价指标中。

# 4结束语

针对目前违法广告泛滥的问题，本文从广告发布者、经营者等人角度出发，提出一种基于违法关键词库的违法关键词语识别方法。因为目前尚未公布较为权威的广告违法关键词库，本文通过专家解读广告法提取种子词汇，再利用哈工大同义词林扩充、语义相似度过滤，得到广告违法关键词库。基于构建好的违法关键词库，本文提出基于上下文词性序列的广告违法关键词识别方法。实验结果表明，该方法有效的识别出广告违法关键词，较好的解决了违法广告中含禁用词的情况。下一步工作将解决描述型的违法广告识别问题。

# 参考文献

1. 陈力丹，钱童，魏雨珂.新《广告法》实施后“广告新闻”现象分析[J]. 新闻界，2016（02）:9-13.
2. 王宁宁，杨瑜，杨姗姗. 基于贝叶斯算法的网络广告监督管理[J].电子商务，2016：14-17.
3. 张知临. 工商网上违法广告智能识别关键技术研究与实现[D].浙江，浙江大学计算机科学与技术应用学院，2013.
4. 邓一贵，伍玉英. 基于文本内容的敏感词决策树信息过滤算法[J]. 计算机工程，2014，40（09）:300-304.
5. AA Abdulrazzaq, NA Rashid, ANB Ali. Fast hybrid string matching algorithm [J]. International Journal of Digital Content Technology & its Applic, 2013, 7: 62-68.
6. Erhard Rahm, Philip A. Bernstein. A survey of approaches to automatic schema matching [J]. Rahm, E. & Bernstein, P. The VLDB Journal, 2001, 10: 334-350.
7. 张忠生，王俊人. 一种模式匹配促进实体统一的方法[J]. 山东师范大学学报(自然科学版). 2012，27（02）：35-39.
8. Robert S. Boyer, Stanford Research Institute, Menlo Park. A fast string searching algorithm [J].Communications of the ACM, 1977: 762-772.
9. 温都日娜. 一种基于本体的敏感词过滤方法研究[D].吉林，吉林大学计算机科学与技术学院，2014.
10. Y. Ding, S. Foo. Ontology research and development: part1-a review of ontology generation [J].Information Science, 2002, 28(02): 234-260.
11. 张坤，徐安凤. 网络环境下有害信息的识别与过滤技术[J].电脑知识与技术，2009，5（09）：2099-2100.
12. 夏天. 汉语词语语义相似度计算研究[J]. 计算机工程. 2007，33（06）：191-193.

1. 下载地址：http://www.ltp-cloud.com/download/ [↑](#footnote-ref-1)
2. 中科院分词软件包下载地址：http://ictclas.nlpir.org/downloads. [↑](#footnote-ref-2)
3. 数据公布地址： [↑](#footnote-ref-3)