一种约束上下文区间的关键词组合策略

杜刚，朱艳云，张晨，杜雪涛

（中国移动通信集团设计院有限公司，北京 100080）

摘　要　随着5G的商用，短信功能将逐渐被5G消息所替代。5G消息单条文本消息长度将可达上万字。现有关键字组合策略将由于文本长度的增加产生误匹配。本文设计了一种约束上下文区间的关键词组合策略，并给出了详细的实现方法。使用该方法可以高效实现在特定长度的上下文区间内进行关键词组合策略匹配，同时该方法能够截取命中策略的关键文本片段，帮助人工审核人员进行快速审核。

关键词　内容安全；关键词匹配；内容识别

中图分类号　TN918　　**文献标识码**　A　　**文章编号**　1008-5599（2021）09-0069-05

DOI:10.13992/j.cnki.tetas.2021.09.012

关键词组合策略是一种有效识别垃圾文本信息的方法，目前是运营商进行垃圾短／彩信治理的主要技术手段。一条关键词组合策略通常是由若干关键词和“与”、“或”逻辑运算符组成，定义了一种垃圾信息的词语特征。当文本信息满足策略定义词语特征时，则可判定文本信息为垃圾信息。举例说明，假设定义策略为(A|B)&(C|D)，则所有包含关键词A或B且包含关键词C或D的文本信息可认为是疑似垃圾信息。

随着5G的商用，传统垃圾短／彩信将被5G消息服务逐步取代，消息通信能力将全面增强。针对文本类消息，其支持的文本长度不再受到限制，长度可以仅包含一个字，也可以达到上万字。当消息较短时，若消息命中若干关键词，则关键词之间的最大距离不大于消息的第一个词到消息最后一个词的距离。因此，命中关键词之间的距离相对较近。词语之间距离越近，则关联度越大，更可能在表达同一个意思。但当消息较长时，策略中的关键词可能会分散在消息的各个部分，词语之间的距离可能很大，不同关键词可能是在表达不同的意思。这种情况下很容易产生误识别。

从垃圾信息散播者的角度看，将若干敏感词零散分布在大量文字中的各个角落完全达不到传播垃圾信息的目的。换言之，如果一段长文本中出现了违规内容，违规内容更可能相对集中在较短的上下文之中，如一些小说章节中夹杂着不良内容的广告信息，或存在色情露骨的描述。所以，在匹配长文本时，当策略匹配的关键词分散在文章的各个角落时，基本可以断定属于策略的误匹配。

例如策略“推出 & 积分 & 优惠”是一条广告类策略。当消息字数小于140个字符时，文本若命中该策略，意味着“推出”、“积分”、“优惠”3个关键词同时出现在一个相对集中的上下文当中，文本很可能是广告类信息。但当消息字数达到上万字时，即便文本命中该策略，策略中的3个关键词之间的平均距离有可能会超过300个字符，3个词语可能完全不在同样的语境下出现。例如一篇正常的长文可能先后包含了“推出新政策”、“积分落户”、“税收优惠”3个不同的上下文，结果恰好误匹配该条策略的3个关键词，最终被误判定为垃圾信息。

收稿日期：2020-08-31

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

值得注意的是，在短文本消息识别中也有可能出现策略中关键词所在上下文不同导致的误匹配问题，但由于消息本身字数受限，此问题不够凸显。随着消息的文字数量增加，策略误匹配问题发生的概率将随之增加。综上所述，无论是短文本还是长文本，都存在由于关键词所处上下文不同导致的误匹配问题。通过改进现有策略匹配机制，在考虑关键词是否符合逻辑出现的同时，考虑关键词匹配位置是否在特定长度的上下文区间，能够同时提高策略在各种长度文本中的垃圾信息识别效果。

1　约束上下文区间的策略匹配

文本的上下文是一个比较抽象的概念，其可能对应一个或多个句子。现有技术手段很难精确对一段文本的上下文进行拆分。本文通过限定策略关键词在文本中出现的上下文区间大小来约束关键词所在的上下文。上下文区间大小与关键词的位置有关，关键词的位置指从文本的开头到该关键词的第一个字符之前所包含的文字数量。如图 1 所示，“娱乐城”右上角的“2”代表位置信息。由于“娱乐城”距离文本的开头有两个字符，故位置为 2。可以将两个关键词的位置相减得到关键词之间的距离，可将此距离定义为这两个关键词的上下文区间大小。如图 1 中“娱乐城”与“现金”之间的位置之差为 7 个字符，则“娱乐城”与“现金”的上下文区间长度为 7。

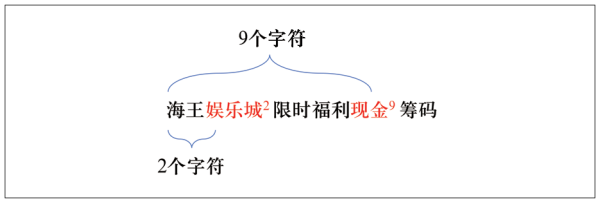

图1　关键词位置的定义示意图

对于仅包含“与”逻辑的策略，策略中的关键词都必须出现在文本中才算成功匹配。可以通过限定所有策略匹配的关键词之间的最大距离来约束关键词所在上下文。如策略“推出 & 积分 & 优惠”匹配了一段文本，其中“推出”、“积分”、“优惠”3 个词在文本中出现的位置分别为 8、35、108，则匹配关键词的最大距离为“优惠”与“推出”之间的位置之差为 100 个字符。本文将这个最大距离定义为策略的上下文区间大小，可以限定一个最大上下文阈值，仅当策略的上下文区间小于最大上下文阈值时，才认为策略成功匹配文本。

当策略存在“或”逻辑时，获取匹配关键词的最大距离问题将变得复杂。如策略“(推出 | 上架)& 买一赠一”。其中“推出”、“上架”、“买一赠一”3 个关键词在文本中出现的位置为 4、104、114。由于“推出”和“上架”是或的关系，二者命中一个即算匹配。则实际存在两种匹配方案，一种是“推出 & 买一赠一”，一种是“上架 & 买一赠一”。前者的上下文区间大小为“买一赠一”与“推出”之间的位置之差为 110 个字符，后者上下文区间大小为“买一赠一”与“上架”之间的位置之差为 10 个字符。若最大上下文阈值为 50，则“推出 & 买一赠一”不满足上下文要求，被抛弃。“上架 & 买一赠一”满足上下文要求，被保留。综上所述，当存在“或”逻辑时，需要考虑各种匹配方案，并将所有小于最大上下文阈值的匹配方案保留。

现实当中的策略通常会存在“与”、“或”逻辑的相互嵌套，可以通过将复杂策略拆分成仅包含“与”逻辑的策略后保留所有小于最大上下文阈值的匹配方案。但一条比较复杂的策略通常可以拆分出数百条“与”逻辑策略，当存在上万条策略时，消息与策略的匹配速度会严重下降。为了提高上下文判定效率，本文将判定上下文过程融入到关键词匹配的逻辑计算过程中，从而避开穷举“与”逻辑策略带来的巨大计算量。

2　匹配方法

2.1　关键词查找与定位

在实践中，一条文本通常需要与包含上万条策略的策略集合进行匹配。在进行消息与策略集合中的策略匹配前，需要先对策略集合中的所有关键词建立索引，以加快匹配

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

速度。本文推荐使用 AC 自动机对关键词建立索引。使用 AC 自动机建立索引后，仅需遍历一遍文本内容，即可对所有关键词进行快速的匹配与定位，匹配速度与策略的数量无关。

策略关键词索引如图 2 所示，首先将策略集合中的关键词形成关键词集合，其次将关键词集合中的关键词构建 AC 自动机索引。该过程属于文本与策略匹配的预处理阶段，在策略关键词没有改动的情况下，通过关键词集合构建的 AC 自动机可以被反复使用，无需与每一条文本信息匹配时重新构建。

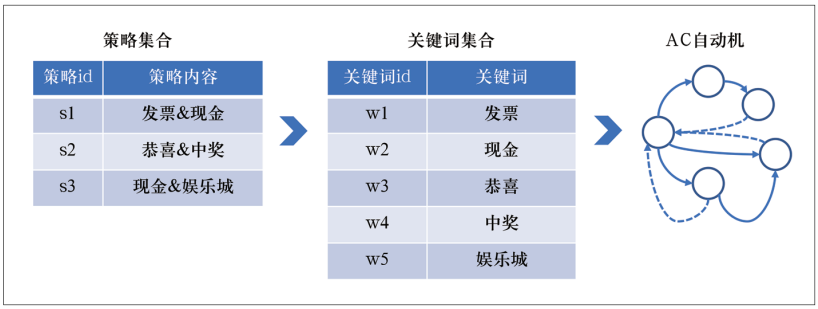

图2　对策略中的关键词建立索引示意图

使用 AC 自动机对文本消息中的策略关键词进行匹配与定位如图 3 所示。图中的 AC 自动机由策略中的关键词构建而成，其中已经包含了所有策略关键词的知识。利用自动机的快速索引功能可以快速定位文本消息中出现的策略关键词。如图中“娱乐城”和“现金”分别出现在了短信的第 2 和第 9 的位置。

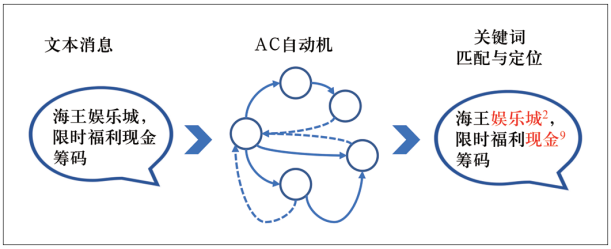

图3　对文本消息中策略的关键词进行匹配与定位示意图

2.2 策略二叉树表示

为了能够解析复杂的策略嵌套逻辑，需要将策略表达式转换为逻辑嵌套的二叉树结构。如图 4 所示，任何一个嵌套复杂的策略可以转换为相对应的二叉树结构。其中二叉树的分支节点为逻辑运算符，叶子节点为关键词，分支节点的深度代表逻辑嵌套的深度。通过对二叉树进行后序遍历并结合堆栈可以实现组合逻辑的运算。限于篇幅这里不再赘述。

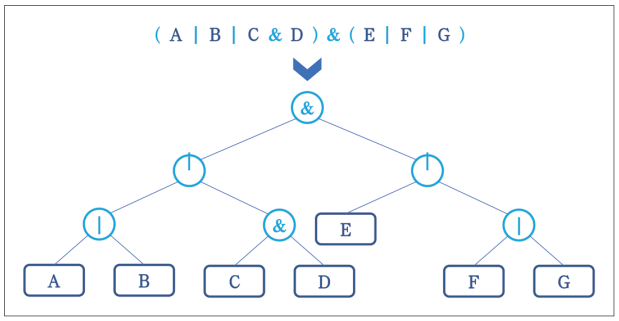

图4　表示关键词组合逻辑的二叉树

2.3 策略粗匹配

在获取文本中匹配的策略关键词及其位置后，可以用匹配的关键词在策略集合中快速筛选出可能匹配的策略，这一过程称为粗匹配。策略粗匹配可通过如下两个步骤实现。第一，当策略集合中的策略不包含匹配的关键词时，则无需与短信进行逻辑匹配；第二，当策略中包含匹配的关键词数量小于策略最小关键词匹配量时，则无需与短信进行逻辑匹配。

这里定义的策略最小关键词匹配量是保证策略命中的最小关键词数。如策略“(A|B)&(C|D)”一共有 4 个关键词。其中“与”逻辑左侧 A、B 任选一词，右侧 C、D 任选一词即可命中策略，故策略最小关键词匹配量为 2。对于任意复杂嵌套逻辑的策略，需要按如下步骤完成最小关键词数的计算。

（1）将策略使用二叉树结构表示，统一用 1 替换叶子节点的关键词。

（2）后序遍历该二叉树，按如下规则从叶子节点开始向上计算分支节点的取值：当分支节点是“或”逻辑时，该分支节点的取值为其孩子节点取值的最小值；当分支节点是“与”逻辑是，该分支节点的取值为其孩子节点

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

取值之和。

（3）按此规则不断计算，最终根节点的取值为策略的最小关键词匹配量。

策略“(A|B|C&D)&(E|F|G)”的最小关键词匹配量的计算过程如图 5 所示。策略的最小关键词匹配量的计算可以在文本匹配前预先计算完成作为一个策略的属性保存，无需在粗匹配过程中反复计算。

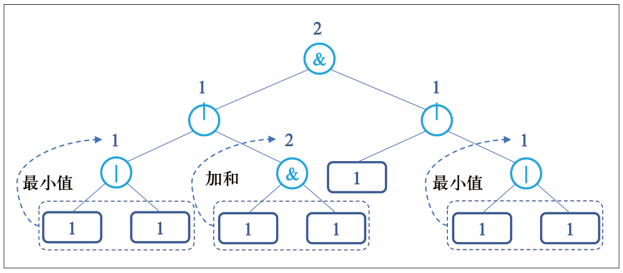

图5　策略最小关键词匹配量计算过程示例图

2.4 策略细匹配

通过策略的粗匹配可以快速筛选出文本可能会命中的策略。针对筛选出的每一条策略，可以使用如下步骤完成策略的细匹配过程。

（1）将策略使用二叉树结构表示，将二叉树叶子节点中加入布尔值。当叶子节点的关键词是文本匹配的关键词时，叶子节点加入 True，并且将关键词的位置信息也保存其中。否则，加入 False，没有位置信息。策略“(A&B&D|E&F)&(G|H|J)”的二叉树表示如图 6 所示。假设文本命中了策略关键词 A、B、D、E、F、H、J，未命中关键词 G，则 A、B、D、E、F、H、J 节点处为 True，G 节点处为 False。此外，在标记布尔值的基础上，还需将命中关键词对应的叶子节点在文本中的位置也保存在叶子节点中（图 6 中标红节点）。

（2）后序遍历二叉树，按如下规则从叶子节点开始向上计算分支节点的取值。

当分支节点是“或”逻辑时，该分支节点的布尔值取值为其两个孩子节点布尔值进行“或”逻辑运算的结果，同时可将孩子节点的上下文区间信息全部拷贝到分支节点。如图 6 中的情况 1 所示，分支节点有两个上下文区间分别是其两个孩子节点的上下文区间，当分支节点的孩子节点是叶子节点时，可以构造新的上下文区间，区间的开始和结束相等，具体示例如图 6 中的情况 2 所示。

当分支节点是“与”逻辑时，该分支节点的布尔值取值为其两个孩子节点的布尔值进行“与”逻辑运算的结果。若计算后为 False，则该分支节点的上下文区间为空。若计算为 True，则需要计算该分支的上下文区间。该分支的上下文区间为其孩子上下文区间的最小覆盖区间。如图 6 中的情况 3 所示，孩子节点为叶子节点，上下文区间分别为 [33,33] 和 [100,100]，能够覆盖住两个区间的最小覆盖区间为 [33,100]。当孩子节点存在多个上下文区间时，需要将“左孩子”与“右孩子”的上下文区间进行两两组合后再计算最小覆盖区间。如图 6 中的情况 4 所示，分支节点的“左孩子”和“右孩子”各有两个上下文区间，可构成 4 种组合方式，如图 7 所示。最终可以计算出 4 个最小覆盖区间作为分支节点的上下文区间。

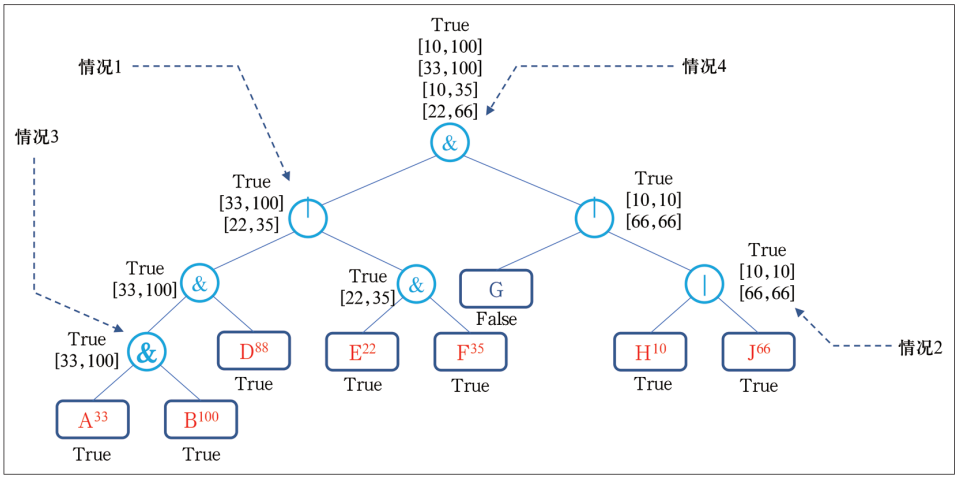

图6　策略细匹配过程示意图

将计算得到的上下文区间与最大上下文阈值进行

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

左孩子上下文区间	右孩子上下文区间	最小覆盖
[33,100]	[10,10]	[10,100]
[33,100]	[66,66]	[33,100]
[22,35]	[10,10]	[10,35]
[22,35]	[66,66]	[22,66]

图7 排列组合左右孩子的上下文区间求最小覆盖示意图

比较，仅保留区间大小小于最大上下文阈值的上下文区间。若没有区间满足要求，则分支节点的布尔值设置为False。由于采用后序遍历二叉树的模式，该步骤可以将不满足上下文长度限定的“与”逻辑进行剪枝，从而减少该分支上层分支节点的上下文区间数量和排列组合数量。

经过策略细匹配后，不但可以找到满足关键词组合逻辑和上下文区间长度限定的策略，同时还可以输出具体上下文区间对应的文本内容。通过上下文区间的起止点可以快速从文本中截取对应的上下文文本，作为匹配证据输出。在传统的关键词组合策略中，匹配的关键词可能离散分布在长文本的各个角落，因此得出的匹配证据文字量可能很大，审核人员需要阅读大量文本才能够定位所有匹配的关键词。使用约束上下文区间策略输出匹配证据长度永远小于最大上下文阈值（建议设置为100个字符）以下，匹配证据更易进行人工审核研判，可大幅提高审核效率。

3 结束语

5G消息支持长文本的发送。传统的垃圾短信关键词组合策略在长文本场景下会出现误匹配现象，其原因主要在于匹配的关键词可能并不处于相同的上下文。为了解决该问题，本文提出了一种约束上下文区间的关键词组合策略。通过设置最大上下文阈值，实现在关键词组合逻辑匹配的基础上，约束匹配的关键词必须集中分布在特定长度的文本片段中。使用该策略能够有效提高匹配关键词在同一上下文中的概率，进而提高长文本匹配的准确率。同时，本文提出的方法可以输出能够匹配策略的一个或多个文本片段，该功能一方面可以方便审核人员对长文本进行审核，另一方面可以方便策略优化人员评估策略的有效性，对策略进行优化更新。

参考文献

[1] 陈有伟，康磊．基于Trie树的关键词匹配算法在电子政务领域的应用[J]．智能计算机与应用，2019(5)．
[2] 金鹏飞，牛保宁，张兴忠．高效的多关键词匹配最优路径查询算法KSRG[J]．计算机应用，2017(2)．
[3] 程维刚，王宁，田勇．基于关键词匹配技术的相似试题检测方法研究[J]．北华航天工业学院学报，2015(3)．
[4] 史乙力．基于关键词匹配的网页文本过滤算法的研究和实现[D]．贵州大学，2009．

A keyword combination strategy constraining context interval

DU Gang, ZHU Yan-yun, ZHANG Chen, DU Xue-tao
(China Mobile Group Design Institute Co., Ltd., Beijing 100080, China)

Abstract With the commercialization of 5G, short messages will be replaced by 5G messages. The length of a single 5G message text message will reach tens of thousands of words. Existing keyword combination strategies will cause mismatches due to the increase in text length. This paper designs a keyword combination strategy that constrains the context interval. And gives a detailed implementation method. Using this method can efficiently achieve keyword combination strategy matching in a specific length context interval. At the same time, this method can intercept the key text fragments of the hit strategy and help manual reviewers to conduct quick review.

Keywords content security; keyword matching; content recognition

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net