一种基于汉字笔顺特征的关键词变体匹配方法

王红雨，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　近年来，垃圾短消息呈现出包含大量拆分字和形近字的现象，这种短消息可以绕过监控系统的关键词审查。由于拆分字和形近字数量众多，变化灵活，将其全部加入关键词库将令关键词库变得冗余。对此，本文提出了一种基于汉字笔顺特征的关键词变体匹配方法。基于汉字笔顺特征，首先合并垃圾短消息中的拆分字；然后通过建立索引表，快速查找出短消息中包含的疑似关键词；最后提出了“金字塔匹配法”匹配关键词。本文提出的方法有效降低了关键词库的冗余度，提高了关键词匹配效率。

关键词　关键词变体匹配；合并拆分字；金字塔匹配法

中图分类号　TN918　　**文献标识码**　A　　**文章编号**　1008-5599（2020）12-0014-05

DOI:10.13992/j.cnki.tetas.2020.12.003

当前使用手机、微信等发送短消息已经成为不可或缺的沟通方式，然而一些诈骗分子、广告类机构却向用户发送大量诈骗短消息、推广类消息，严重影响了用户的使用体验[1]。相关垃圾短消息治理机构提出了使用黑名单和关键词过滤等方式，对垃圾短消息进行了大力治理，并取得了良好的成效。然而，不法分子使用拆分字和形近字生成关键词变体，通过向用户发送包含关键词变体的短消息来绕过监控系统的关键词检查。如关键词“炸金花”,可将其中的“炸”改写为“火乍”、“咋”、“柞”、“诈”等,还可以将其中的“花”改写为“垅”、“娓”、“椛”等。各种改写方法排列组合后可形成数量繁多的关键词变体，这使得关键词库维护和更新工作变得复杂。

1　相关工作

文献[2]提出了一种关键词变体的提取和匹配方法。该方法首先通过字符区位去除短消息中的噪音字符，然后使用拼音文件将分词后的文本和关键词转换为文本整数串，接着提取拼音替换和谐音替换的变异关键词，最后建立形近字词库，比较分词后的文本中字符与形近字词库中的关键词的每个字符是否为同一组形近字。该方法通过建立的形近字词库实现了关键词变体的正常提取。

文献[3]提出一种确定关键词变体的方法。该方法首先将待匹配文本拆分为多个文本字符串；然后利用汉字在多种编码形式下的字形相似关系，计算出每个文本字符串的异构图特征；接着使用得到的异构图特征，通过机器学习模型计算出文本字符串与待匹配关键词的相似度；最后，根据相似度判断文本字符串是否为待匹配关键词变体。这样便能判断出待匹配文本中是否包含待匹配关键字的变体。

现有的技术方法将这些关键词变体配置关键词库，一方面会令关键词库数据量增大且冗余度增加，另一方

收稿日期：2020-11-05

面增加关键词更新和维护的成本。而计算汉字编码距离或者使用机器学习模型判定形近字的方法难以确定合适的阈值，且并未提升匹配效率。对此，本文提出了一种基于汉字笔顺特征的关键词变体匹配方法。

2 一种基于汉字笔顺特征的关键词变体匹配方法

本文考虑到关键词变体通常由拆分字和形近字组成，从汉字笔顺特征角度出发，提出了一种基于汉字笔顺特征的关键词变体匹配方法。

如图 1 所示，本文提出的关键词变体匹配方法主要包括两个子过程：合并短消息中包含的拆分字和关键词变体匹配。

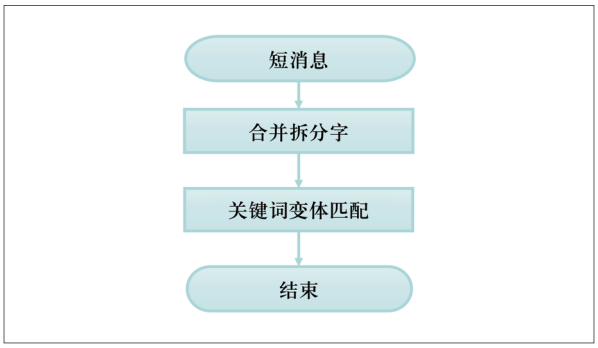

图1 原始短消息处理流程图

2.1 合并拆分字

汉字组合结构有很多，如上下结构、上中下结构、左右结构、左中右结构和半包围结构等。在垃圾短消息中，不法分子常对左右结构和左中右结构字进行拆分，因为这两种结构拆分后可以利用人们从左到右阅读习惯在脑中重新链接起来，可读性更高。故拆分字的合并重点为左中右组合和左右组合两种组合方式。

合并拆分字过程首先判断短消息中相邻的 2 个或 3 个字是否为 1 个左右结构或左中右结构的字拆分之后的字，如果是，则将其合并成 1 个字。如短消息“金月月鸟娱乐城”中相邻的 3 个字“月月鸟”为“鹏”拆分后的字，则将其合并为“鹏”。合并拆分字过程可进一步分解为 3 步操作。首先是对短消息进行必要的预处理，内容是去除掉短消息中的非中文字符；其次，是检测短消息中是否存在左中右结构的拆分字，并对拆分字进行合并；最后，是检测短消息中是否存在左右结构的拆分字并进行合并。下面详细介绍如何检测 / 合并左中右结构的拆分字，左右结构的拆分字检测 / 合并步骤与之相似。

如图 2 所示，以短消息“金月月鸟娱乐城”为例，展示了左中右结构拆分字的检测与合并过程。由于是左中右结构，故参与合并的汉字有 3 个。图中左侧使用长度为 3 的滑动窗口穷举了短消息中任意 3 个相邻汉字组合的可能，其中每一行是一种组合的可能性。针对每一行，都需要检测窗口中的 3 个汉字是否能够合并成新字。具体做法是将窗口中 3 个汉字的笔顺进行组合，看是否恰好是某个字的笔顺，如“月月鸟”恰好能够合并为“鹏”字。

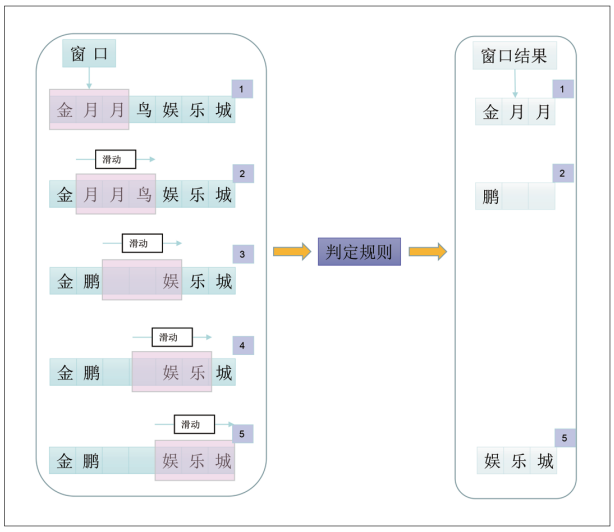

图2 左中右拆分字合并示意图

但同时要考虑到两种例外的情况。第一种情况是窗口中的汉字即便能够合并成一个字也不能将其进行合并。这种情况在处理左中右结构的拆分字时并不多见，更多见于处理左右结构的拆分字上。如“女子”可合并为“好”字，但若将“女子”合并为“好”，则很有可能改变了短消息的语义。第二种情况是窗口中的汉字笔

顺不能直接合并一个字，但仅与某个字的笔顺有微小的差别，此时应将窗口中的字合并成一个左中右结构的汉字。如图 3 所示，“王古月”的笔顺与“瑚”字的笔顺仅差一笔，原因在于“王”在变为部首后，最后一笔由“横”变形为“提”。

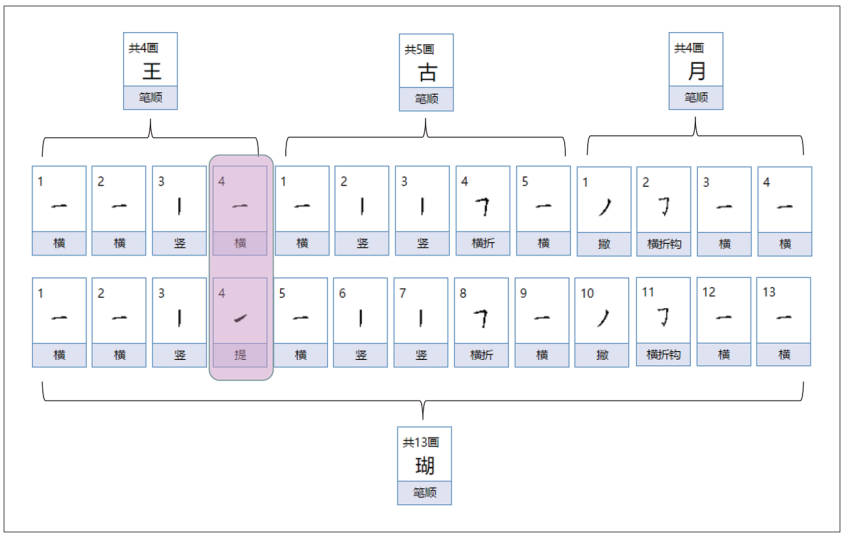

图3 “王古月”和“瑚”的笔顺对比示意图

针对第一种情况，可以通过设置常用词典的方法加以解决，即窗口中的汉字如果是常用词典中的词，则不能合并窗口中的字。针对第二种情况，可以计算窗口中汉字组合的笔顺与每个左中右结构字的笔顺的编辑距离，若编辑距离为 1，并且笔顺差异点出现在汉字之间的衔接笔顺上，则判定窗口中的汉字可以合并。如图 3 所示，“王古月”、“瑚”的笔顺序列的编辑距离为 1，且不同的一笔为“王”的最后一笔，故可以将“王古月”合并为“瑚”。

依据规则，将滑动窗口中能够合并的汉字组合进行合并，由于合并后 3 个字变成 1 个字，减少了短消息字数。为了便于继续穷举其它相邻汉字组合，将额外引入两个空格作为填充位。如图 2 中第 3 行所示，当窗口中出现填充位时，由于字数不足 3 个，不进行任何操作。

2.2 关键词变体匹配

将短消息中存在的拆分字合并后，下一步是进行短消息内容与关键词的匹配。如果拆分字合并后恰好是关键词库中的关键词，则直接精确匹配。然而，不法分子经常用形近字来替代关键词中的一个字或多个字，这样替换可以使精确匹配失效。如“诈金花”精确匹配“炸金花”会失败，人读后却可以联想到该形近词表达的意思。使用形近字的垃圾短消息给关键词精确匹配带来了难度。因此本文提出了更为适用的关键词变体匹配方法。

如图 4 所示，首先将汉字常用的 28 个笔画用 01 ~ 28 进行编号，建立汉字笔画编号表；然后以每个关键词包含的关键字作为索引，关键词 id 作为值建立一个索引表，通过此索引表能够快速得到每个关键词由哪些关键字组成，如查询关键字“炸”，则能够快速得到关键词库中包含“炸”的关键词为“炸金花”和“炸弹”；接着依据汉字笔画编号表对每个关键字进行笔顺分析，得到每个关键字的头 4 笔和末 4 笔，如通过查询汉字笔画编号表，得到“炸”的笔画 id 序列为 13−23−10−13−10−17−18−17−17，则其笔画序列的头 4 笔为 13−23−10−13，末 4 笔为 17−18−17−17（一般地，互为形近字的两字的头 4 笔或末 4 笔相同，如形近字“诈”和“炸”的末笔都为 17−18−17−17。）；最后，分别建立每个关键字的头 4 笔和末 4 笔对应关键字的索引表，通过查询此表可以得到头 4 笔或末 4 笔相同的疑似形近字。

经过上述准备工作，对于每条短消息都可使用本文提出的“金字塔匹配法”快速匹配关键词。如图 5 所示，关键词库中包含 3 个关键词“炸金花”、“诈胡”和“炸弹”，短消息为“来玩咋金花”，其中，“咋”为“炸”的形近字。

使用提出的“金字塔匹配法”匹配“来玩咋金花”中包含的关键词变体“炸金花”的步骤如下。

（1）遍历短消息“来玩咋金花”中的每个字，对于每个字，结合图 4 提出的汉字笔顺编码表，得到每个字的笔画 id 序列。

（2）分别按每个字的头 4 笔和末 4 笔索引形近字，得到每个字在关键词索引表中的疑似形近字 id，如查询

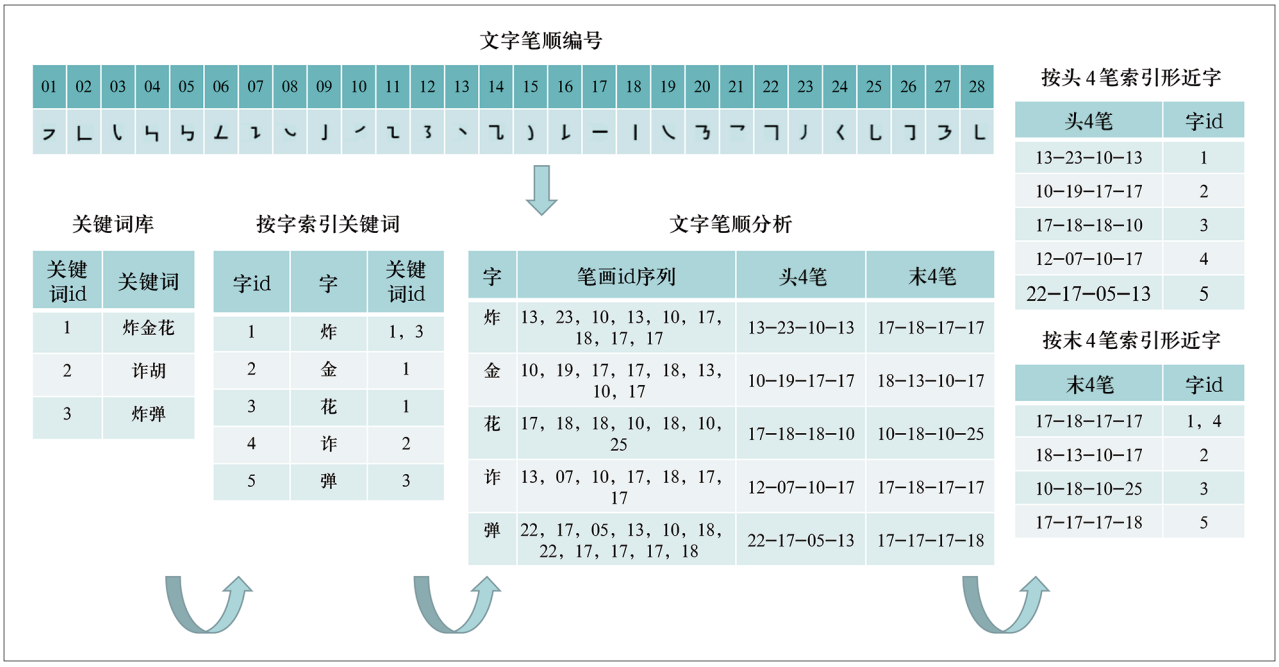

图4 关键词库、按字索引关键词、按头4笔或末4笔索引形近字示意图

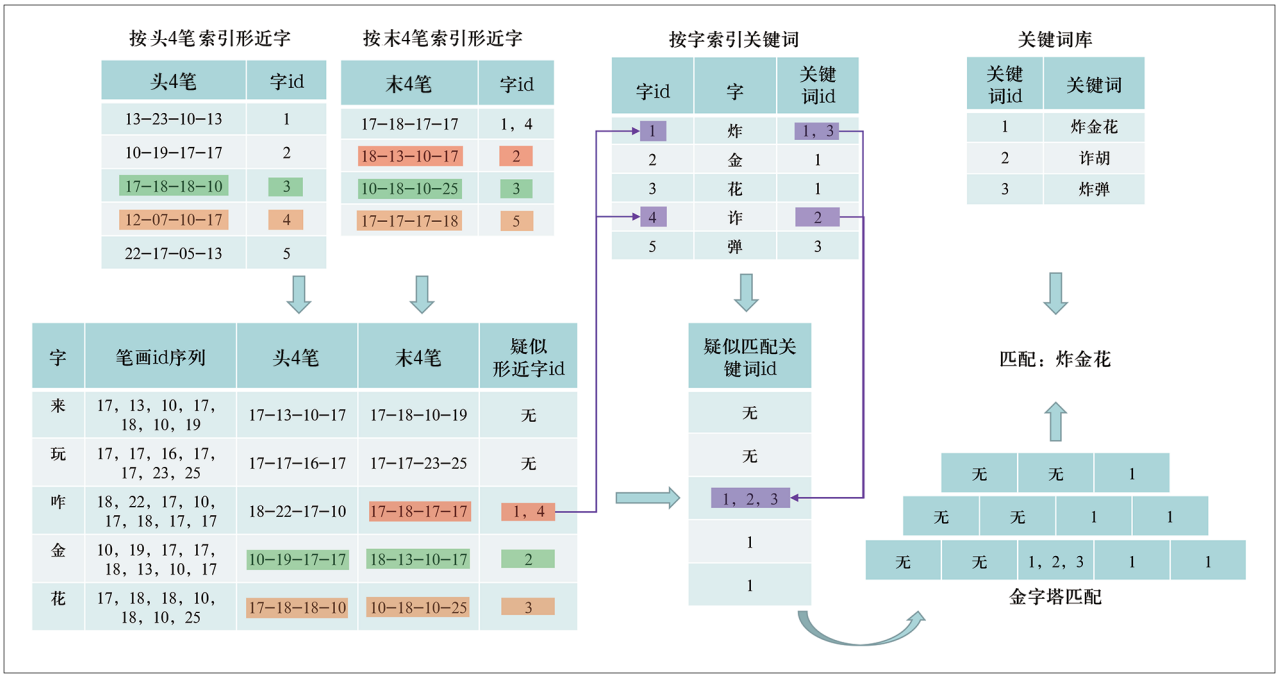

图5 形近字匹配及金字塔匹配法示意图

“咋”的末笔可以得到“咋”在关键词索引表中的疑似形近字 id 为 1（对应“炸”）和 4（对应“诈”），而“来”和“玩”未在关键词索引表中找到，则记为无。

（3）根据疑似形近字 id 得到包含这些疑似形近字的关键词 id。如根据 id 为 1 和 4 疑似形近字得到 id 为 1、3 和 2 的疑似匹配关键词。“来”和“玩”未查询到疑似

形近字，则其疑似形近字记为“无”，相应的，疑似匹配关键词记为“无”。

（4）经过前 3 个步骤之后可得到短消息“来玩咋金花”中每个字的疑似匹配关键词分别为“无”、“无”和 id 为“1、2、3”、“1”、“1”的关键词。将“无”和疑似匹配关键词 id 作为金字塔的底层，记作第 1 层；相邻的疑似关键词 id 计算交集，得到第 2 层；然后根据第 2 层的疑似关键词 id 查询关键词库中对应的关键词是否为两个字构成的关键词，如果是则匹配成功，代表短消息中包含此关键词，不是则匹配失败，表明第 2 层中无匹配关键词。同理第 2 层中的相邻疑似关键词 id 求交集得到第 3 层疑似关键词 id，接着计算第 3 层疑似关键词 id 在关键词库中对应的关键词是否由 3 个字组成，如果是则匹配成功，如本例中第 3 层的疑似关键词 id 为 1，查询关键词可知，对应的关键词“炸金花”由 3 个字构成，则匹配成功。直到最上面一层的疑似关键词 id 全部为空，匹配结束。

经过上述 4 步可成功匹配出短信息“来玩咋金花”中“咋金花”在关键词库中的形近词“炸金花”。“金字塔匹配法”避免了依次遍历关键词库，极大地提高了关键词变体的匹配效率。

3 结束语

本文提出了一种基于汉字笔顺特征的关键词变体匹配方法。该方法通过合并笔顺序列，能够有效处理短消息中将关键字拆分的情况。使用汉字的头 4 笔和末 4 笔索引形近字以及“金字塔匹配法”匹配关键词变体，避免了精确匹配方法中需要经常更新和扩展关键词库的问题。本文提出的方法简化关键词库的同时提高了垃圾短消息的查全率，且避免了依次遍历关键词库，提升了匹配效率。

未来工作将从两个方面展开研究。一是优化常用词词典，常用词典越全面，合并拆分字的效果越有效。二是深入研究关键词组合匹配的方法，本文的研究对象为单关键词匹配，未来将研究多个关键词的组合匹配，使用关键词组合匹配的方法将提升垃圾短消息过滤的准确率。

参考文献

[1] 黄文良. 垃圾短信过滤关键技术研究[D]. 浙江大学，2008.
[2] 傅彦，陈安龙，周俊临，等. 一种变异关键词的提取方法：CN101324883[P]. 2008-12-17.
[3] 高喆，康杨杨，陶秀莉，等. 关键词变体的确定方法和装置：CN110929477A[P]. 2020-03-27.

A variant keyword matching method based on the stroke order features of Chinese characters

WANG Hong-yu, DU Gang, ZHU Yan-yun, ZHANG Chen, DU Xue-tao

(China Mobile Group Design Institute Co., Ltd., Beijing 100080, China)

Abstract In recent years, spam short messages appear to contain a large number of split and similar characters, this kind of short message can bypass keyword filtering and be sent to users. Due to the large number and flexible changes of split words and similar words, adding them to the key database will make the database redundant. In this paper, a variant keyword matching method based on the stroke order features of Chinese characters is proposed. Firstly, the split words in spam short messages are merged based on the stroke order features of Chinese characters. Secondly, the suspected keywords contained in spam messages are indexed by an index table which is built using the characters of keywords. Finally, a pyramid matching method is proposed to match keywords. The method proposed in this paper can effectively reduce the redundancy of keywords database and improve the efficiency of keywords matching.

Keywords variant keywords matching; merging split characters; method of pyramid matching