群聊业务中的跨消息关键词匹配技术研究

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，以群聊为媒介的不良信息和诈骗信息传播变得愈发猖獗。相比于点对点通信，群聊具有多人高交互性的特点，一些不良信息需要通过结合多条消息才能识别，需要研究跨多条消息的关键词匹配技术。本文重点对跨消息关键词匹配流程中的消息缓存和关键词匹配两个步骤进行了深入研究，并给出了高效的实现方案。该方案对群聊业务监控策略的设计和实现具有重要的参考价值。

关键词　关键词匹配；自动机；内容安全

中图分类号　TN918　　**文献标识码**　A　　**文章编号**　1008-5599（2022）02-0030-05

DOI:10.13992/j.cnki.tetas.2022.02.003

相比于普通消息，群聊消息具有更强的传播力，群聊消息中的不良信息破坏力更强，基于群聊的诈骗活动也会随之增加，因此需要对群聊中的文本内容进行有效监控，来实现各类不良信息的快速识别。在群聊消息引入前，运营商治理不良文本信息主要以单条消息作为识别对象。然而在群聊业务中，单条消息长短不一且彼此穿插，可能需要定位多条消息中的多个关键词才能够判定不良信息。若仅对群聊消息中的单条消息进行关键词匹配，会漏掉藏匿在多条群聊消息中的不良信息。因此有必要对多条群聊消息进行缓存，并进行跨多条消息的关键词组合逻辑匹配。

1　群聊跨消息关键词组合策略匹配流程

如图 1 所示，群聊跨消息关键词组合策略匹配流程可以分为以下 4 个步骤。

（1）消息队列缓存。该步骤将群聊中的多条消息合并为单条缓存消息，作为后续关键词匹配的基础。为控制缓存消息无限增长，可配置最大缓存消息数量，如缓存最近 100 条群聊消息。

图1　群聊跨消息关键词组合策略匹配流程

（2）关键词匹配。该步骤在缓存消息之上进行策略关键词的查找和定位。为后续策略逻辑的计算提供依据。由于策略关键词可能成千上万，此步骤对算法要求较高，直接影响了不良信息识别的性能。

（3）策略逻辑计算。该步骤检查消息中出现的策略关键词是否满足策略定义的布尔逻辑。该步骤需要解析策略中定义的布尔逻辑，并进行布尔计算。在处理群聊消息时，该步骤的计算方法可沿用现有垃圾短信监控系统的策略逻辑计算方法。本文不做重点讨论。

收稿日期：2020-11-10

（4）关键用户定位。当缓存的群聊消息命中策略逻辑后，需将命中关键词的相关消息提取出来，并定位消息的发送者，从而方便对用户进行处置。此步骤要求消息在缓存过程中不但要保存消息的内容，还需要保存消息发送者信息。

由上分析可知，群聊跨消息关键词组合策略匹配流程的关键在于消息队列缓存和关键词匹配环节。在消息队列缓存环节恰当地保存发送者信息，可以方便地进行后续关键用户定位。在关键词匹配环节快速定位关键词，则可以快速地判断策略逻辑是否满足。

2　群聊消息缓存

为了跨多条消息进行关键词组合逻辑匹配，首先需要将多条消息合并为一条消息。若直接将多条消息内容拼接为单条消息，则每条消息的发送者信息将丢失。为了保留该信息，可在合并消息时在每条消息头部加入头信息。头信息可记录消息的发送者和时间等消息元信息，可通过特殊字符串进行标记，方便提取和识别。同时通过头信息也方便将合并的缓存信息再分解为原始多条信息。

具体地，可使用“###”作为头信息的开始和结束标志。若群聊中有两条信息内容分别为“contentA”和“contentB”，发送者分别为C和D，则两条信息加入头信息后变为“###C###contentA”和“###D###contentB”。两条消息合并后，信息内容变为“###C###contentA###D###contentB”。通过正则表达式识别头信息可以快速地将合并消息还原为多条消息。

在进行消息合并时，可以设定最大缓存消息数量。通过统计合并消息中的头信息数量来计算当前合并消息中的实际消息数量，也可以通过额外的计数器进行统计。当消息达到最大缓存消息数量后，可将位于缓存消息头部的消息删除，并在尾部追加新消息，从而在较小的计算开销下实现类似于消息队列的功能。

加入消息头的消息会隔断多条消息的内容，若两条消息内容衔接处恰好命中关键词则会被遗漏。如群聊中用户A先后发出两条消息分别为“cont”和“ent”，两条消息内容合并后可组成消息“content”，然而加入消息头后变为“###A###cont###A###ent”，直接在此合并消息上做关键词匹配将导致无法命中关键词“content”。为此，需要设计一种能够自动忽略消息头信息的关键词匹配算法，以解决这种跨消息关键词匹配问题。

AC自动机是目前业内使用比较广泛高效的关键词匹配算法，本文基于AC自动机原理设计并实现了一种可自动忽略消息头的关键词匹配算法。

3　AC自动机原理

给定一条消息和一个关键词集合，AC自动机能够从消息中快速匹配出关键词集合中的关键词。不同于逐个检查关键词集合中的每个关键词是否在消息中出现，AC自动机首先将关键词集合中的关键词构建自动机。具体过程如图2所示，左侧为关键词集合，右侧为对应的AC自动机结构。任何一个AC自动机都有一个初始状态0。自动机在任意状态下接收一个字后，会发生状态转移。当自动机处于初始状态时，输入“常”，则自

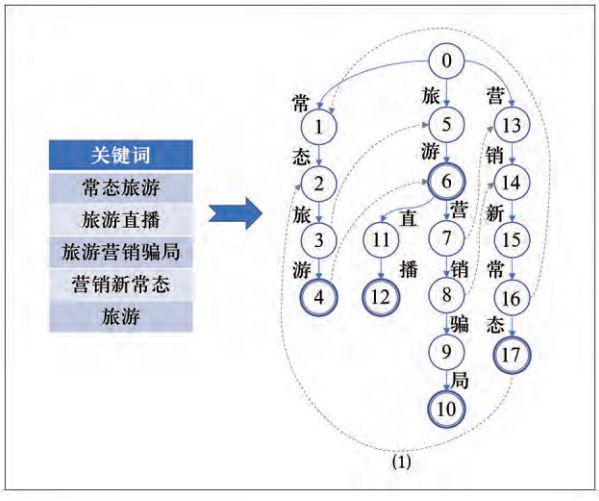

图2　AC自动机原理示意图

动机会转移到状态 1。又如当自动机在状态 6 时，输入“直”可以转到状态 11。

从状态 0 开始，可从左到右依次输入关键词的文字，从而构建一个状态转移序列。如“常态旅游”创建了状态转移序列 0，1，2，3，4。当两个关键词存在相同前缀时，则前缀共用相同的状态转移序列，如“旅游营销骗局”和“旅游直播”共享状态转移序列 0，5，6。当自动机转移到加粗的状态时（4、12、6、10、17），代表自动机已经成功匹配了关键词集中的某个关键词。如状态 4 代表自动机成功匹配了“常态旅游”这个关键词，此时状态机会输出“常态旅游”。

由图 2 可知，自动机的每个状态所能接收的字是不同的。如状态 6 可接收“直”和“营”，状态 3 只能接收“游”。当自动机输入的字不在当前状态可接收字的范围时，会产生错误匹配。在错误匹配发生时，需要将自动机转移到合适的状态以便继续匹配。自动机通过失败指针来实现错误匹配时的状态转移，失败指针的转移在图中用虚线箭头表示。失败指针转移指向的位置应该为当前成功匹配串的后缀与关键词集中关键词前缀的最大公共串所在状态。如失败指针（1）所示，当自动机处于状态 17 时，自动机已经成功匹配串“营销新常态”，当此时再输入字“旅”时，由于“旅”不是状态 17 可接收的字，故自动机将转向失败指针指向的状态。由于“营销新常态”的后缀“常态”与“常态旅游”的前缀“常态”相同，且“营销新常态”更长的后缀“新常态”不是任何关键词的前缀，故“常态”是最长的公共子串。因此状态 17 的失败指针应该指向状态 2，该状态是自动机由初始状态接收到“常态”两个字符后应该转移的状态。当自动机转到状态 2 后，再尝试输入“旅”，则可以成功转移到状态 3。

假设消息内容为“营销新常态旅游”，将字依次输入自动机后，自动机首先到达状态 17，而后通过失败指针转移到状态 2，并进一步接收“旅游”二字转移到状态 4，从而实现对“营销新常态”和“常态旅游”两个关键词的成功匹配。

注意，AC 自动机的每个状态都有失败指针，其默认都指向自动机的初始状态 0，即代表当前匹配串与关键词最长公共子串是 0。鉴于图示清晰性，在图 2 中没有画出。

4　自动忽略消息头的 AC 自动机模型

标准的 AC 自动机通过给关键词建立索引而加速消息与关键词集合的匹配效率，但其不能直接用在群聊缓存消息的匹配过程中。因为其在匹配时无法自动忽略缓存消息中的消息头，实现跨消息的无缝关键词匹配。本文对 AC 自动机结构进行了修改，从而使其可以自动忽略合并消息中的消息头信息。为了达到目的，文本先创建一个能够识别消息头的 AC 自动机。此处定义消息头格式形如“＃＃＃A＃＃＃”，其中“A”为任意长度的字符串，且 A 中不包含“＃＃＃”。

识别消息头的 AC 自动机如图 3 所示。由于消息头以连续 3 个“＃”作为开始标识，故当自动机未出现连续 3 个“＃”时，其会通过失败指针（1）和（2）跳转回状态 0。仅当出现了 3 个连续的“＃”后，状态机进入状态 −3。状态 −3 代表自动机已经进入了消息头的内容区域，其可以是任意长度的字符串，但不能包含“＃＃＃”。此时出现任意非“＃”字符自动机都会通过

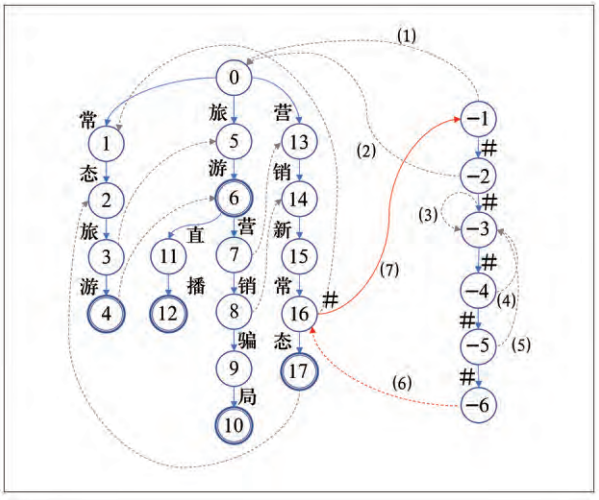

图3　识别消息头的自动机

失败指针（3）保持状态 −3。当出现“#”但并不是连续 3 个“#”时，也会通过失败指针（4）和（5）回到状态 −3。仅当 3 个连续的“#”出现后，自动机达到状态 −6，代表消息头结束。

在创建完成识别消息头的 AC 自动机后，可以将其嵌入到关键词集构成的 AC 自动机中，从而实现消息头的自动忽略。如图 4 所示，左侧为关键词集构成的 AC 自动机，右侧为识别消息头的 AC 自动机。图 3 中自动机的状态 0 转变为图 4 中状态 16，图 3 中失败指针（1）和（2）改为指向图 4 中失败指针（1）和（2）。

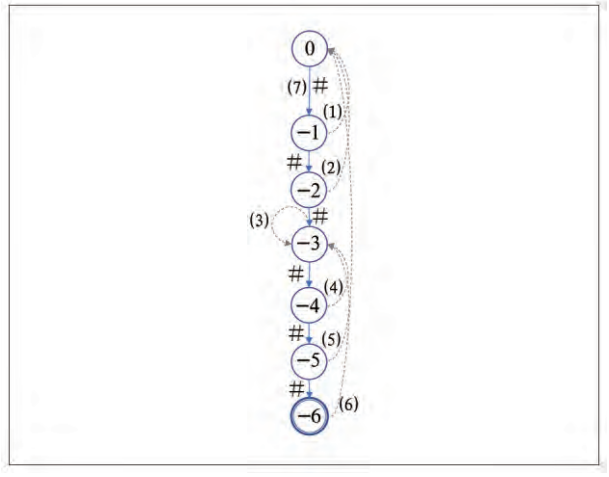

图4　改进的AC自动机模型

当消息为“营销新常 ###A### 态”时，“营销新常”可将自动机状态迁移到状态 16，状态 16 中接收“#”会直接转入识别消息头的 AC 自动机。当消息头识别完成进入状态 −6。状态 −6 接收任何字符会自动转回状态 16，此时再接收“态”，则可以将自动机转入状态 17，进而成功匹配了“营销新常态”这个关键词。但当消息为“营销新常 # 态”时，由于消息中的“#”并非消息头标记，自动机会进入状态 −1 后通过失败指针（1）返回状态 0，因此不会匹配“营销新常态”这个关键词。注意此处假设关键词集中的关键词不会以 # 开头，若存在以 # 开头的关键词，则需要将指针（1）和（2）指向自动机接收“#”对应的状态。建议在实际操作中使用不会出现在消息开头的字符作为消息头的标记，从而简化模型设计。

图 4 中仅以状态 16 为例展示了消息头识别自动机如何嵌入到关键词集自动机中，为了满足自动机在每个状态时都可以忽略消息头，需要为每个状态分配单独的消息头识别自动机。经过修改的 AC 自动机结构比修改前复杂很多，且冗余度较高。为了降低自动机的冗余度，减少不必要的内存开销，本文引入了动态 AC 自动机的概念。

传统的 AC 自动机是一个静态模型，即当通过关键词集构建好自动机后，自动机的结构不会发生变化。本文将传统 AC 自动机动态化，即当自动机处于不同状态时，其结构也相应发生改变。具体地，在构建 AC 自动机时，仅需要构建一个消息头识别自动机，当自动机处于不同的状态时，只需要将图 4 中标记为红色的两条线移动到自动机当前的状态，再开始接收字符即可。

如图 4 所示，自动机处于状态 16 时，连线（7）从状态 16 指向状态 −1，连线（6）指向状态 16。当自动机处于状态 17 时，仅需将连线（7）从状态 17 指向状态 −1，连线（6）指向状态 17 即可。通过如上自动机的动态变化，整个模型不需要为每个状态都分配独立的消息头识别自动机。所有状态仅共用一个消息头识别自动机，大幅降低了自动机的复杂度和内存消耗。在状态转移时，仅需要修改两个连接的起讫点，引入的额外计算开销极小。

5　增量关键词匹配更新

假设用于群聊消息缓存的队列最多缓存最近 10 条群聊消息，则在第 11 条消息到来时，需要删除队列头部消息，将第 11 条消息加入队尾，具体过程如图 5 所示。此时图中标绿色区域在消息缓存中并没有发生任何变化，其所能命中的关键词也不会发生变化。故在进行后续的关键词匹配时，该部分没有必要进行重复关键词匹配。

为了避免对重复内容进行多次关键词匹配，可设立

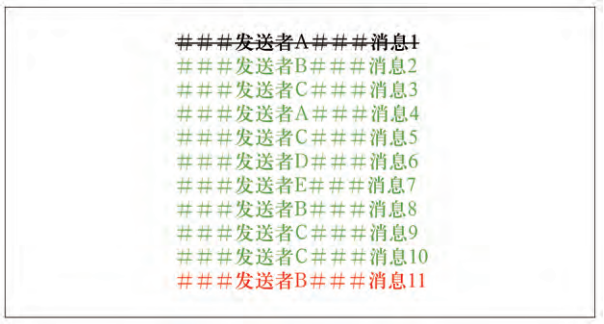

图5　缓存消息队列示意图

已匹配关键词缓存，用来缓存当前队列中已经匹配的关键词及其位置信息。当新消息到达时，为了计算出消息的增量变化带来的匹配关键词的增量变化，需要从如下3个方面考虑（以图5为例）。

（1）消息1的删除导致匹配关键词的减少。其中包含消息1中匹配的关键词和跨消息1和消息2匹配的关键词，需要将这些关键词从已匹配关键词缓存中删除。这部分关键词可以通过关键词缓存中匹配关键词的位置信息计算得出。

（2）消息1的删除导致现有匹配关键词位置的变化。该步骤仅需要将匹配关键词现有位置信息统一减去消息1的长度即可。

（3）消息11的加入导致匹配关键词的增加。其中包含消息11中匹配的关键词和跨消息10和消息11匹配的关键词，需要将这些关键词加入到已匹配关键词缓存中。这部分关键词可以通过缓存自动机状态得出。即将自动机匹配上一次缓存信息最终到达的状态进行保存，并将新消息的字符直接输入到该状态下的自动机中。则自动机会自动输出所需要的匹配关键词。

值得注意的是，当消息缓存队列填满时才需要删除消息1，故如上列举的前两个方面仅在消息缓存队列填满时才需要考虑。综上所述，可以避免对缓存消息中的相同内容进行重复匹配，自动机仅需要增量处理新消息的内容就可以实现匹配关键词的增量更新。

本文提出的模型在犯罪分子知晓具体实现方法时，可能存在被不法分子规避的风险。犯罪分子可能模仿消息头的格式在消息内容中伪造消息头信息来避免敏感关键字被审查，如故意在消息内容中添加“### 敏感词 ###”等类似于消息头的内容。此时，自动机会误认为其为发送者信息，而不是消息内容，从而被忽略。解决如上风险的手段是在进行多条消息拼接前，提前检查消息内容中是否存在与消息头格式一致的内容。若存在，则将相关特殊字符进行替换或直接去掉。这样可以杜绝不法分子伪造消息头带来的风险。

6　结束语

本文针对群聊业务提出了一种高效的跨消息关键字组合策略匹配方案。相比于传统按单条消息进行关键词匹配的方案，本文引入了群聊消息缓存功能，并以消息头的方式保存每一条消息的发送者信息。在关键词匹配阶段，本文对现有AC自动机模型进行了增强，实现了自动忽略消息头进行关键词匹配的高效算法。同时，通过将原有的静态AC自动机动态化，可以进一步减少内存资源占用，使模型更加轻量化。最后，通过对缓存消息中匹配关键词的缓存和自动机最终状态的缓存，可以实现对关键词匹配的增量更新，避免对缓存消息中的重复内容进行反复匹配问题。该方法可以被工程化应用到现有的群聊业务监控系统中，以提高对群聊中各种不良信息的识别能力，具有较高的参考价值。

参考文献

[1] 范洪博，姚念民．高级AC自动机的快速构建方法[J]．计算机研究与发展，2013(12)．
[2] 侯整风，杨波，朱晓玲．一种适合中文的多模式匹配算法[J]．计算机科学，2013(11)．
[3] 侯整风，杨波，朱晓玲．一种节约内存的中文多模式匹配算法[J]．微型机与应用，2013(7)．
[4] 陈新驰，韩建民，贾洞．基于AC自动机的多模式匹配算法FACA[J]．计算机工程，2012(11)．
[5] 刘佳，杜雪涛．一种双向安全可信的网络架构[J]．电信工程技术与标准化，2011(10)．

（下转第40页）

实际中配置建议见表 8。

建议 40 个用户以内用第一种配置，大于 40 个用户的用第二种配置。高负荷场景需要增加调度用户数的使用，最后一种固定配置 3 个符号，在用户数不是太多且上传速率要求高的可以尝试固定配置 2 个符号。用户数小于 30 个，终端能力更强的可以配置固定 1 个符号。

参考文献

[1] 3GPP. Physical channels and modulation(Release9): TS 36.211[S]. 2009.
[2] 李小文，方前军，宋海贝．一种LTE系统中计算CFI值的方法[J]. 计算机应用研究，2011(10).

Study on influence of PCFICH configuration on UE throughput rate

YUAN Jun-jie, YIN Ye

(China Telecom Liuzhou Branch, Liuzhou 545000, China)

Abstract In LTE system, PCFICH channel configuration is mainly used to indicate the number of OFDM symbols occupied by PDCCH control channels in a sub-frame. The number of PDCCH control channels directly affects the number of dispatchable users and peak rate of cell. Under different scenarios in order to find out the optimal PCFICH channel configuration, and to achieve the highest user terminal throughput, in this paper, on the basis of theoretical calculation under different PCFICH configuration for the peak rate, testing in users, and connecting with the actual testing data for research, finally we conclude that users from low to high five scenarios PCFICH configuration recommendations.

Keywords PCFICH; PDCCH symbol number; peak rate; scheduling

（上接第 34 页）

Research on cross-message keyword matching technology in group chat business

DU Gang, ZHU Yan-yun, ZHANG Chen, DU Xue-tao

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

Abstract In recent years, the dissemination of bad information and fraudulent information through group chat has become more and more rampant. Compared with peer-to-peer communication, group chat has the characteristics of high interaction among many people. Some bad information needs to be identified by combining multiple messages, and it is necessary to study the keyword matching technology across multiple messages. This paper focuses on in-depth research on the two steps of message caching and keyword matching in the cross-message keyword matching process, and gives an efficient implementation scheme. The scheme has important reference value for the design and implementation of group chat business monitoring strategy.

Keywords keyword matching; automaton; content security