基于动态流通语料库（DCC）的流行语发现技术研究*

王莹莹[1] 刘华秋[2] 杨尔弘[1]✉ 江宇轩[1]

[1] 北京语言大学

[2] 北京交通大学

ying_y_wang@126.com

摘　要：流行语的监测研究是中国语言监测工作中十分重要的一项任务。目前每年发布的年度中国媒体十大和分类流行语都是基于流行语发现技术和专家挑选的人机结合方式进行提取，但现有的流行语发现技术存在着数据统计粒度较大查询缓慢、以及候选流行语表数据量大和“非词”字串占比较高造成的人工工作量大的问题。针对现有技术的不足，本文提出了以“日”为周期使用时间序列数据库 InfluxDB 的数据存储方案、逻辑斯蒂曲线拟合和异常上升段识别两种流行语发现模型以及借鉴短语挖掘技术对候选流行进行质量评分和过滤的方案。对比实验结果表明：本文提出的方案可取得更好的效果，显著降低后续的人工工作量，提供更丰富的候选字串的统计信息以及更高效的数据查询方式，具有更高的实际应用价值。

关键词：流行语发现；时间序列数据库；曲线拟合;异常上升段识别；质量短语

Research and Implementation of Buzzword Detection Technology Based on Dynamic Circulation Corpus (DCC)

Yingying Wang[1], Huaqiu Liu[2], Erhong Yang[1]✉, Yuxuan Jiang[1]

[1] Beijing Language and Cultural University

[2] Beijing Jiaotong University

ying_y_wang@126.com

Abstract: Buzzword monitoring research is a very important task in Chinese language monitoring. At present, the top ten and classified buzzwords in Chinese Newspaper released every year are extracted based on man-computer cooperation, which consists of buzzword detection technology and experts selection. But the existing buzzword detection technology has the following shortcomings: large statistical granularity, slow query, large amount of candidate buzzwords and high proportion of "non-word" n grams. In view of the shortcomings of the existing technology, this paper proposes a data storage solution with "day" as the cycle and using time series database InfluxDB, two buzzword detection models comprising logistic curve fitting and abnormal rising segment recognition, and utilizing the phrase mining technology for quality scoring and filtering of buzzword candidates. The comparison experiments show that the technology proposed in this paper can achieve better results, significantly reduce the subsequent manual workload, provide more statistical information of buzzword candidates and more efficient data query method.

Keywords: buzzword; time sequence database; curve fitting; identification of abnormal rising segment; quality phrases.

1　前言

流行语是一种语言现象，指在某一时期、某一范围内广为传播、盛行一时的词语[1]。流行语的监测研究是中国语言监测工作中十分重要的一项[2]，自 2008 年初首次发布“年度中国媒体十大流行语”以来，已连续发布了十四年。自 2011 年开始作为“汉语盘点”活动的一部分，由国家语言资源监测与研究中心、商务印书馆等多家机构在现场联合向社会发布。每年的流行语及其解读也都会收录到《中国语言生活状况报告》[3]]中。流行语是所监测的语言形式与其所反映的社会热点的结合，以流行语盘点年度媒体的大事小情，记录风云变幻中的中国与世界[4]。

流行语发现是一项具有重要意义的基础性研究工作，其基于大量的语料，通过分析语料中字串的使用频率随时间的而变化发现语料中的流行语，运用计量手段，描述语言变化，反映社会状

*本文受国家语委“十三五”规划 2019 年度重点（科研中心）项目 ZDI135-105、国家语委“十三五”科研规划 2020 年度重点（科研中心）项目 ZDI135-131、北京语言大学语言资源高精尖创新中心项目 TYZ19005、北京语言大学 2020 年度研究生创新基金 20YCX143 的资助。

况[1]。目前每年所发布的“年度中国媒体十大流行语”的发现技术采用人机结合的方法，对全切分处理得到的字串序列进行统计、建模、排序、筛选，得到流行语候选表，再多阶段地利用专家知识从中提取得到年度十大（综合类）和各类别流行语[3]。首先，从信息存储上来说，采用这种方式提取时，统计和存储的字串信息的粒度越细，信息数据量越大，查询性能越低下。然而，利用专家知识提取流行语的环节要以统计信息为基准，但目前以“月”为单位的字串信息的存储和查询方式难以高效地为专家所用。其次，得到的候选流行语表的数据量的大小直接决定了专家挑选时人工工作量的多少，排序靠后的字串仍需大量的人工挑选。最后，因在数据处理过程使用了全切分方案，因此在候选流行语表中有着较高占比的“非词”字串，也需人工甄别。目前所使用的流行语发现技术得到的候选流行语表有数十万的字串需专家进行多阶段地人工挑选，效率和性能都略显低下。

针对以上三个问题，本文提出了以下解决方案：数据存储字串序列的以“日”为周期，并使用时间序列数据库 InfluxDB 存储字串的统计信息；对流行语的字串频率历时变化曲线进行建模，提出了曲线拟合、异常上升段识别两种流行语发现模型；后处理阶段对流行语候选表进行质量评分，过滤质量评分过低的非词字串。采用以“日”为周期的字串信息存储方式与时间序列数据库 InfluxDB 的存储媒介，既能保留数据更细粒度的信息，也能解决了细粒度信息数据量过大引发的查询性能低下的问题。我们在中国主流平面媒体动态流通语料库上进行了对比实验，实验结果表明：与目前所使用的平面媒体流行语发现技术（基线模型）相比，本文提出的数据存储方案能提供更丰富的候选字串的统计信息以及更高效的数据查询方式，两种流行语发现模型在$TopN$召回率和$NDCG_p$两个指标上得到的结果更好，流行语质量评分模型也能在$TopN$召回率稍有降低的情况下显著降低流行语候选表的数据量，减轻后续的人工工作量。

2 实验方案

本文延续使用何伟等人对汉语流行语的定义和特征解释，即流行语指的是“在某一时期、某一范围广为传播、盛行一时的语汇”，以及“‘广为传播、盛行一时’是流行语与其他语汇的区别属性, 它的特点是以零或低频次为起点, 很快上升到高频次, 被广泛使用并持续一定时间；‘某一时期’、‘某一范围’则为不同类别的流行语提供了时空观照点”[5]。基于流行语“广为传播、盛行一时”的区别属性，为捕获字串的使用突增周期，本文设计了一套从数据的统计、存储到流行语发现模型，再到流行语质量过滤模型的完整方案，自动从语料中提取出流行语候选表，供后续的专家人工挑选。

2.1 数据统计和存储方案

在何伟对汉语流行语的定义中提到：“‘语汇’是流行语的类别属性, 它应该是一种语言符号, 可以是一个已有的词, 也可以是一个新词, 或者是由若干词组成的短语”[5]。为保留所得的流行语候选有“语汇”的类别属性，我们并未使用某一特定的汉语分词方案，而是采用全切分方案[5-7]对语料进行切分。语料经全切分得到 2-7 字的字串序列，例如“我爱北京天安门”，就能切分得到“我爱”“我爱北”“我爱北京”“我爱北京天”“我爱北京天安”“我爱北京天安门”以及“爱北”“爱北京”等字串。

切分完成后，对字串的频次和文档频等基本信息进行统计。我们抛弃以“月”为周期计算字串的字串频和文档频的数据统计方案和使用 SQLServer 数据库的存储方案，采用以“日”为周期的统计方案，使用“日期序号”计算累计字串频以及进行标准化处理。我们首先定义了“日期序号”：将一段给定的时间内的第一天编号为日期序号 1，第二天编号为日期序号 2，以此类推可得每一天的日期序号。其次，区别于已有研究中直接使用字串频数据来进行计算和筛选，我们计算了累计字串频，即从日期序号的第一天到这一天该字串的字串频的加和。最后我们对累计字串频进行标准化处理，即把所有日期序号的累计字串频都缩放到 0-100 之间，第一天的累计字串频（纵轴）为 0，最后一天为 100。图 2-1 是字串“霍金”和“中国”的日期-标准化累计字串频折线图，当字串“霍金”在某个日期使用突增时，图像在这一点也变得陡峭，也可以看到在某段时期内广为传播时图像近似是一个 S 型曲线。而全年使用都十分高频无明显突增的字串“中国”的图像则近似是一条直线。字串的日期-标准化累计字串频图，既提供给流行语挑选专家更形式化的参考去判断该字串是否会被收录为流行语，也能作为流行语发现模型的输入去更精确地发现流行语候选。

以“日”为周期计算字串信息尽可能多地保留了字串的统计信息，但同时所产生的统计数据量也远超以月或周计算所产生的数据量。为此我们提出了使用时间序列数据库 InfluxDB[1]进行存储的方案，并基于字串信息的查询特点设计了数据库 Schema，将字串信息的日期存入数据库 point 的 time 部分，字串本身存入 point 的 tag 部分，字串的字串频和文档频存入 point 的 field 部分。这样字串和时间都将被索引，使用字串和时间进行的查询高效化。

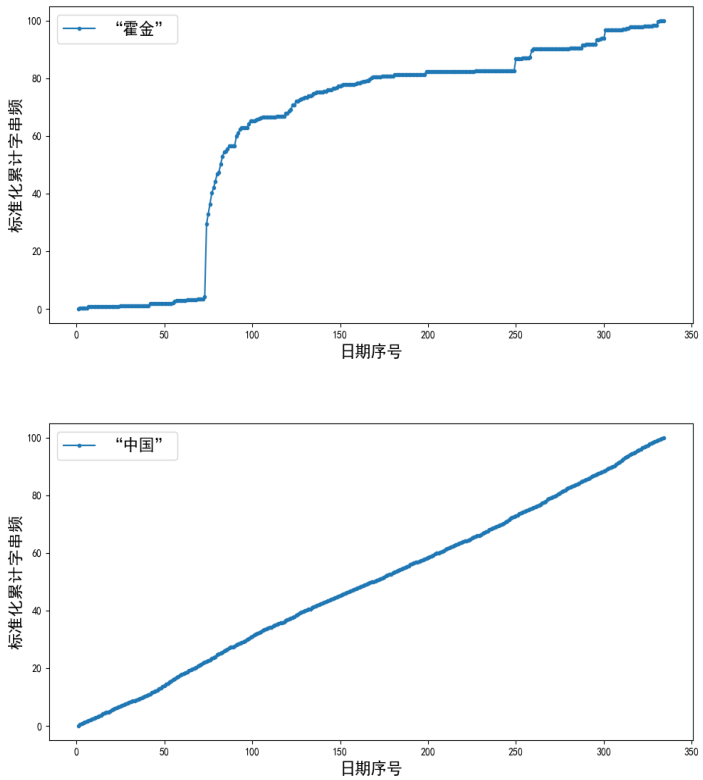

图 2-1 字串“霍金”和“中国”的日期序号-标准化累计字串频折线图

2.2 流行语发现模型

语料经全切分、数据统计和存储后，将根据流行语的特性进行建模，对全切分得到的候选字串表进行排序、筛选，得到流行语候选表。目前所采用的流行语发现模型采取的是对以“月”为周期的字串频次、频差及其他复合计量指标等统计信息综合建模，以字串频次总和排序得出流行语候选表，再供专家人工提取。这种方法仅是筛选了在某个月份使用有增加的字串，既不能确定是否是突然变化，也有可能过滤掉持续时间达数月的字串，不够准确。此外，这种方法得到的候选字串数据量较大，后续的专家挑选的工作量十分繁重。为尝试解决上述问题，本节内容根据流行语“广为传播、盛行一时”的区别属性，基于字串的日期-标准化累计字串频图提出了两种流行语发现模型：曲线拟合模型和异常段上升识别模型。

2.2.1 逻辑斯蒂曲线拟合模型

基于字串的日期-标准化累计字串频折线图，首先可以使用曲线拟合的方法判断字串是否符合流行语的“广为传播、盛行一时”属性，这样的流行语的图像特点是近似 S 型曲线，因此我们使用最小二乘法将日期-标准化累计字串频折线图拟合为逻辑斯蒂（Logistic）曲线，模型函数见式（2-1），其中b和a为待估参数，b与曲线的最大斜率相关，b越小，曲线的最大斜率越大，这个字串的使用突增就越大，就越可能被挑选为流行语。考虑到诸如“黄金周”等周期性流行语的存在，我们使用分段拟合的方式。选择一个n天的观察窗口，在观察窗口内进行曲线拟合（如图 2-2），每隔n天再拟合一次，拟合完成后观察所有拟合误差（这里使用的是使用均方误差）小于 100 的

[1] InfluxDB 官方文档：https://docs.influxdata.com/influxdb/v1.7/

拟合曲线，如果其中任意一个曲线的斜率（b值）大于特定的阈值则将该字串记录下来，全部计算完毕后将所得字串按总频降序输出候选表。

$$f(x)=\frac{100}{1+e^{-b\cdot\left(\frac{x+a}{30}\right)}} \quad \text{(式 2-1)}$$

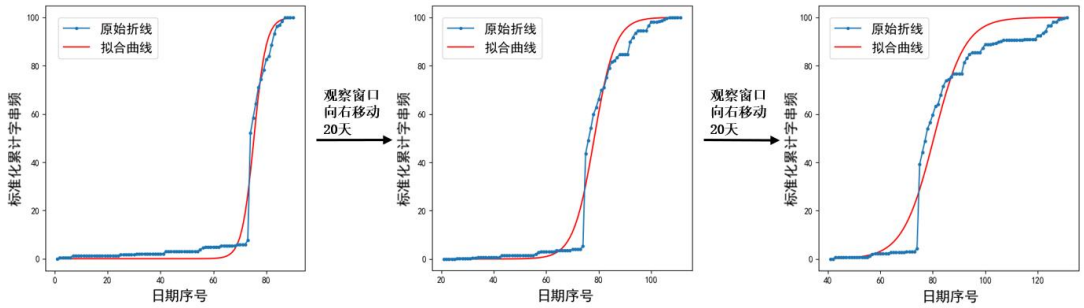

图 2-2 分段拟合示意图

2.2.2 异常上升段识别模型

曲线拟合的筛选模型较为耗时，我们尝试第二种方案：通过阈值识别出字串的日期-标准化累计字串频散点图的异常上升段来获取该字串的突增情况，计算最长异常上升段天数，将大于指定数量的字串筛选出来作为候选表。同样，我们选择一个n天的观察窗口，如果在窗口中标准化累计字串频上升超过了m，则将该窗口内的日期标记为异常上升。从左向右逐渐移动观察窗口，计算出每个字串的最长异常上升天数。根据流行语的定义，我们过滤掉异常上升天数少于 5 天（包括无异常上升）的字串，剩余字串按总频降序输出得到候选表。此外，模型还能给出每个字串的流行次数、流行天数、日最高频、日最高文档频、全年总频等详细的数据统计信息，供挑词专家参考。图 2-3 展示了字串“霍金”的日期-标准化累计字串频散点图中被标记为一次异常上升的部分。

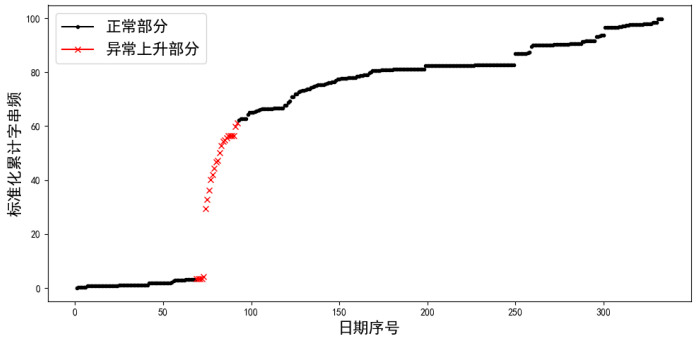

图 2-3 字串“霍金”的异常上升段识别的结果图

2.3 流行语质量评分模型

不管是现有的还是本文提出的流行语发现模型，都是以全切分的字串为基础数据，因此输出的流行语候选表中必定包含较大占比的非词字串。如“我爱北京天安门”经全切分得到的“爱北京天”“京天安门”等四字串都是此类的非词字串。本文借鉴短语挖掘技术来对流行语候选表进行过滤，所使用的是韩家炜团队的自动短语挖掘框架 AutoPhrase[8]，输出为一个按质量递减排列的短语列表。其中，对质量短语的定义是满足流行度、一致性、信息性、完整度的一个完整语义单词序列，质量评分则是这个单词序列成为一个完整语义单词序列的概率。这个定义与流行语的定义有重合之处，因此我们可以先获取到质量短语表，再以此来过滤流行语候选表中的低质量字串。

3 实验结果分析

中国国内语言监测工作的数据资源基础是国家语言监测语料库，流行语的监测和发现基于其

中的通用语媒体语料库，分为平面媒体、有声媒体、网络媒体三个子库，每年以 10 亿字次的规模滚动建设，根据流通度来选择具有典型性和代表性的不同媒体中的语料。这些具有动态、流通特性的语料，记录了大众传媒的语言实态，反映了语言生活。

本文使用的数据资源是由国家语言资源监测与研究平面媒体中心提供的平面媒体语料子库——中国主流报纸动态流通语料库（Dynamic Circulation Corpus，简称 DCC）的 2018 年前 11 个月的语料，其中包含中国国内 15 份主流报纸媒体 2018 年 1 月 1 日至 11 月 30 日的全部新闻报道。2018 年发布的媒体年度十大流行语和其他七大类的分类流行语就是基于这份语料进行提取的，共 80 个，收录在《中国语言生活状况报告（2019）》[3]中。这 80 个流行语是在通过目前所采用的的数据处理和流行语发现模型得到的 2-7 字字串的候选流行语表的基础上，经专家人工挑选得到的。本文的实验即是以这些流行语为标准答案，对不同的流行语发现模型的结果进行评估。由于全切分方案得到的数据量较大，因此实验中仅选取包含流行语最多（35 个）的 4 字字串作为基准，共 683340 个字串。本文提出的两种流行语发现模型也仅提取 4 字字串与之对比。

本文所使用的评价指标有两种——$NDCG_p$（Normalize Discounted Cumulative Gain）和$TopN$召回率。$NDCG_p$衡量模型输出的候选表中字串的排序质量，计算公式如式（3-1）和（3-2），其中rel_i指候选表排序为i的字串的评分，最终发布的流行语分值为 1，其它字串分值为 0；p指候选表的前p项，在本文的实验中取值 15 万；$IDCG_p$指理想的DCG_p，即最终发布的流行语全部排在候选表首位时的DCG_p得分。$TopN$召回率的计算公式见式（3-3），指在候选表的前N条中，包含最终发布的流行语占候选表中所有流行语比率。在本文的实验中N的取值为 15 万。考虑到流行语候选表是后续专家筛选的基础，如果一个流行语在候选表中没有出现，则其在后续的筛选中也将很难被选出，因此在$NDCG_p$和$TopN$召回率两个指标中优先考虑$TopN$召回率。

$$NDCG_p = \frac{DCG_p}{IDCG_p} \tag{式 3-1}$$

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(1+i)} \tag{式 3-2}$$

$$R_n = \frac{TP}{TP + FN} \tag{式 3-3}$$

3.1 数据存储方案的性能分析

本节从数据处理和查询两个方面进行对比分析，比较现在的流行语监测方法中在用的以“月”为周期使用 SQLServer 数据库的存储方案和本文所提出的以“日”为周期使用 InfluxDB 数据库的方案。首先，我们对比了实验数据处理上的时间消耗、内存占用和插入数据库后的数据量，实验结果可以看出“日+InfluxDB”方案的总用时和所占内存都远大于已有的“月+SQLServer”方案。究其原因，一是按“日”进行统计，保留了更多且粒度更细的字串信息数据，更大量的数据要消耗更多的时间和内存。另外时间序列数据库 InfluxDB 会将 time 和 tag 部分在内存中建立索引以支持高速检索，tag 中存储了全切分得到的数十万字串的信息，导致占用更多内存。继续对查询性能进行对比实验，随机选择选取 6000 个字串使用 Python 进行数据库查询，实验结果表示，“日+InfluxDB”方案按月查询的平均单次耗时约为已有的“月+SQLServer”方案的百分之一，查询效率有着极大提升，且按日查询的速度下降幅度不大。综上，以“日”为周期使用 InfluxDB 时序数据库的数据存储方案虽在数据处理上的时间和内存消耗较多，但保留了更多的统计信息，且保证了数据查询的高效性。

表 3-1 两种数据存储方案的数据处理和查询性能对比实验结果

实验项	月+SQLServer 方案	日+InfluxDB 方案
总用时（分钟）	287	2762
所占内存（MB）	1606	40378
预处理后数据量（MB）	628	2400
按月查询平均单次耗时（秒）	0.684	0.00736
按日查询平均单次耗时（秒）	不支持	2.18

3.2 流行语发现模型的对比实验分析

本节在以“日”为周期+InfluxDB 时序数据库的数据存储方案的基础上，使用$NDCG_p$和$TopN$召回率两个评价指标评测两种流行语发现模型。值得注意的是，在语料预处理时我们已经基于停用词表过滤了一些垃圾字串。在统计了字串的频次信息后又进行了两次：过滤在语料中每一天的文档频均大于 8 时的常用字串；过滤在全部语料中的累计频次低于 50 的低频字串。

首先对逻辑斯蒂曲线拟合模型中的参数——观察窗口天数n和曲线斜率b值的阈值取不同值时的实验进行对比，结果如表 3-2。可以看出，当n取 60 并且b取 2.5 和 3 时$TopN$召回率最高，但在召回流行语的同时也召回了更多的垃圾字串，所以流行语的排名可能更靠后，即$NDCG_p$值低。基于上文提到的$TopN$召回率优先的原则，当n取 60 并且b取 3 时曲线拟合模型的实验结果最好。此外由于曲线拟合本身计算量较大，加之采用的分段拟合方式，一个字串在最坏情况下需要进行 14 次拟合，耗时较多，导出四字候选表用时约 8 天（单线程）。

接着对异常上升段识别模型中的参数——观察窗口天数n和最低字串频m取不同值时的实验结果进行了对比，结果如表 3-3。可以看出，n的增大会使得进入流行语候选表的字串增加，也会使$TopN$召回率增加；m的增大会使得进入流行语候选表的字串减少，$TopN$召回率降低。同样根据$TopN$召回率优先的原则，当n取 8 并且m取 9 时实验结果最好。异常上升段识别模型不似曲线拟合模型一样耗时过多，导出四字候选表用时约 1.5 天（单线程）。

表 3-2 n和b取不同值时曲线拟合模型的实验结果

n	b	$NDCG_p$	$TopN$召回率
60	2.5	0.26600	0.91176
	3.0	0.26647	0.91176
	3.5	0.26746	0.88235
70	2.5	0.26750	0.88235
	3.0	0.26778	0.88235
	3.5	0.26801	0.88235
80	2.5	0.26771	0.88235
	3.0	0.26800	0.88235
	3.5	0.26823	0.82353

表 3-3 n和m取不同值时异常上升段识别模型的实验结果

n	m	$NDCG_p$	$TopN$召回率
7	9	0.09304	0.41176
8	7	0.10257	0.47059
	8	0.22131	0.85294
	9	0.27572	0.88235
	10	0.26483	0.82353
	11	0.26953	0.82353
9	9	0.27273	0.82353
	10	0.27094	0.85294
	11	0.27503	0.85294
10	9	0.27063	0.85294
	10	0.26835	0.82353
	11	0.27260	0.85294

再将两种流行语发现模型的最好结果和目前所采用的基线模型进行对比，实验结果如表 3-4。可以看出，曲线拟合模型的$NDCG_p$和$TopN$召回率两项指标均高于基线模型，但是耗时约是基线模型的 16 倍；异常上升段识别模型的$TopN$召回率与基线模型持平，$NDCG_p$高于基线模型和曲线

拟合模型，耗时较基线模型增长两倍。从评价指标角度来看，曲线拟合模型的性能更优；但从实际应用角度来看，曲线拟合模型耗时较长，在多线程条件下导出 2-7 字字串的 6 个候选表花费的时间约为两周，效率低下而无法应用于实际的流行语监测工作中。相较而言，异常上升段识别模型在提高$NDCG_p$的情况下仅少量增加了时间消耗，还能提供给后续的专家挑选步骤每个候选字串的流行次数、流行天数、日最高频、日最高文档频、全年总频等详细的数据统计信息。因此，异常上升段识别模型具有更高的实际应用价值。

表 3-4 不同流行语发现模型的实验结果对比

模型	$NDCG_p$	$TopN$召回率	导出四字候选表耗时
基线模型	0.25797	0.88235	约 0.5 天
曲线拟合模型	0.26647	0.91176	约 8 天
异常上升段识别模型	0.27572	0.88235	约 1.5 天

3.3 流行语质量评分模型的实验分析

本节应用自动短语挖掘框架 AutoPhrase，得到质量短语表，共 1,249,748 个质量短语及其质量评分。然后在异常上升段识别模型最好结果的流行语候选表（共 146,771 个四字字串）上，设置质量评分阈值，过滤掉候选表中低质量的字串。当质量评分阈值取不同值时的实验结果表 3-5。可以看出，当质量评分阈值为 0.2 时，候选流行语的数量减少了约 14%，但$TopN$召回率却未降低，当质量评分阈值继续升高为 0.4 时，候选流行语的字串数量减少了将近 50%，$TopN$召回率仅有小幅降低。这样的实验结果说明，在流行语发现模型之上应用质量评分模型进行过滤，能做到牺牲少量正确候选的同时大幅减少后续人工挑选的工作量，在实际的流行语监测工作中十分有意义。

表 3-5 质量评分模型的实验结果

质量评分阈值	候选表的字串数量	$TopN$召回率
0.2	126504	0. 88235
0.3	108700	0. 82353
0.4	76784	0. 82353
0.5	60583	0.76471
0.6	41382	0.67647

4 相关工作

流行语作为一种广受关注的语言形式，在社会各群体间广泛流传使用，其内部结构、语义、未来的发展均具有很高的研究价值。在过去的几十年间，科研人员对流行语的监测与发现进行了大量的研究，目前流行语监测与发现的相关技术主要采用基于统计的方法。

中国国内的研究中，张普[9]总结了流行语历史变化曲线的特点，为流行语动态跟踪与辅助发现打下了坚实的基础。之后的研究中，研究人员大都先将语料进行全切分处理得到字串序列并按月或周计算字串频和文档频，然后对流行语字串进行建模，将筛选结果组成流行语候选表。例如，何伟等人[5]提出了一种流行语时空监测模型，该模型定义了流行语分布空间、有效持续时间和分布空间的增长幅度等三个特征，并认为流行语的以上三个特征均处于特定的阈值范围内。杨尔弘[6]提出根据流行语字串频图像特点，依次采用穷尽式搜索、图形拟合和专家筛选的技术路线进行提取。吴保珍等人[7]则提出使用基于字串频次和文档频的向量空间模型、语言模型和垃圾串过滤规则对全切分结果进行逐层过滤，然后采用基于频次、频率和使用率的流行语评分模型对字串排序得到流行语候选表。

国外的研究中，Nakajima 等人[10]基于大规模博客数据进行了流行语早期监测的研究，通过关注相关博客主题人群的年龄段、各社区的分布情况等额外信息来提高博客流行话题发现的准确性。Mochizuki 等人[11]基于电视字幕语料库中戏剧和小报两个子语料库，对热门戏剧流行话题进行了研究。通过将两个语料库中的数据进行对比分析去除了戏剧语料库中的部分垃圾字串。网站 The Global Language Monitor 通过对全球纸质媒体、电子媒体和互联网数据进行分析，定期发布全球

英语流行语、最佳单词等，但并未公开具体技术细节。

上述国内外流行语发现技术的研究在一些方面仍然存在一定的不足。第一，国内的研究其最终计算所得的流行候选表中，一些流行语的排序比较靠后，因此候选表中排序靠后的字串也需要人工逐一进行筛选，人工的工作量依然很大。第二，国外 Nakajima 等人[10]和 Mochizuki 等人[11]的研究通过使用大量的额外信息或多个语料库来提高流行话题发现的准确性，这些额外的信息是很难获得的，因此其研究的通用性较低。

5 结论

流行语的发现与监测技术对于反映语言生活状况、描述语言使用实态、开发和利用语言资源具有重要意义。本文基于国家语言资源监测与研究平面媒体中心提供的中国主流报纸动态流通语料库，对现有流行语发现技术中存在的不足进行分析，从数据存储、流行语建模和质量评分三个方面提出了新的技术方案，并在 4 字字串的流行语候选表上进行对比实验。本文工作具体如下：

(1) 数据存储方案以“日”为周期，并使用时间序列数据库 InfluxDB 存储字串的统计信息。与现有的以“月”为周期使用 SQLServer 数据库存储的方案相比，虽然数据量、内存占用和时间消耗增加，但能存储更细粒度更丰富的统计信息以供流行语发现模型和专家挑选，同时也能保证高效的数据查询性能。

(2) 根据流行语定义中“广为传播、盛行一时”的特性，对流行语的字串使用进行建模。基于字串的日期-标准化累积字串频图像，提出使用逻辑斯蒂曲线拟合和异常上升段识别的流行语发现模型。通过与基线模型进行对比实验，结果表明：曲线拟合模型的$NDCG_p$和TopN召回率两项指标均优于基线模型，但时间消耗显著增加，约为基线模型的 16 倍；异常上升段识别模型的TopN召回率与基线模型持平，$NDCG_p$高于基线模型和曲线拟合模型，耗时较基线模型仅增长两倍。且异常上升段识别模型能够为专家挑选流行语提供每个候选字串丰富的频次和流行信息以作参考。因此，异常上升段识别模型具有更高的实际应用价值。

(3) 针对全切分带来的大量“非词”字串的问题，借鉴短语挖掘技术，使用自动短语挖掘框架 AutoPhrase 作为流行语质量评分模型过滤低质量评分的候选字串。实验结果表明，质量评分过滤能在仅牺牲少数正确流行语候选的同时，大幅减少专家挑选的工作量。

参 考 文 献

[1] 杨建国. 流行语的语言学研究及科学认定.语言教学与研究,2004(6):63~70.

[2] 侯敏,杨尔弘. 中国语言监测研究十年.语言文字应用,2015(3):12~21.

[3] 杨尔弘,肖丹,陈芳宇,等.2018,流行语里的世界与中国.中国语言生活状况报告（2019）.北京:商务印书馆.2019:228~236.

[4] 杨尔弘,李彦燕,王磊.2013,流行语里的世界与中国.中国语言生活状况报告（2014）.北京:商务印书馆.2014:260~267.

[5] 何伟,侯敏,文采菊.流行语时空监测模型研究.内容计算的研究与应用前沿——第九届全国计算语言学学术会议论文集.2007.

[6] 杨尔弘.支持语言监测的海量数据处理技术.术语标准化与信息技术,2010(2):38~43.

[7] 吴保珍,何婷婷,李立,等.基于全切分获取网络流行语方法研究.计算机应用研究,2009,26(4):1260~1262.

[8] Shang, J. , Liu, J. , Jiang, M. , Ren, X. , Voss, C. R. , & Han, J. . (2018). Automated phrase mining from massive text corpora. IEEE Transactions on Knowledge and Data Engineering, 30(10), 1825-1837.

[9] 张普. 基于 DCC 的流行语动态跟踪与辅助发现研究. 语言计算与基于内容的文本处理——全国第七届计算语言学联合学术会议论文集. 2003.

[10] Nakajima S., Zhang J., Inagaki Y., et al. Early detection of buzzwords based on large-scale time-series analysis of blog entries. Proceedings of the 23rd ACM conference on Hypertext and social media. ACM, 2012: 275-284.

[11] Mochizuki H, Shibano K. Re-mining Topics Popular in the Recent Past from a Large-Scale Closed Caption TV Corpus. International Journal of Future Computer and Communication, 2015, 4(2): 98.