突发事件检测的 MapReduce 并行化实现*

卓可秋　虞　为　苏新宁
(南京大学信息管理学院　南京 210023)

摘要: 【**目的**】在大数据环境下，从文本流中准确且快速地检测出特定领域的突发事件。【**方法**】利用 Kleinberg 突发检测方法和 LDA 主题模型方法，将其扩展到 MapReduce 并行框架中，实现并行语料预处理、并行突发词检测、并行突发文档过滤和并行主题提取。【**结果**】对新闻文本流进行模拟仿真实验，结果表明，该并行方法在特定领域突发事件检测中准确率 P、召回率 R 和调和平均值 F 分别最高可达 87.50%、77.78%和 82.35%。【**局限**】基于 MapReduce 的并行方法难以实现大规模动态文本流在线(Online)实时(Real-time)突发事件检测。【**结论**】与传统串行突发事件检测方法相比，所构建的分布式并行化方法在保证检测结果正确性的同时，具有良好的可扩展性，性能得到较大提升。

关键词: 突发事件检测　MapReduce　分布式处理　LDA 主题模型
分类号: TP311.1

1　引　言

随着互联网的普及，各种媒体包括传统的和新兴的，都第一时间将收集到的信息传播到网络上。这些媒体每天都会产生上亿条的新闻信息流。这些新闻信息流，特别是文本流，蕴含着丰富的研究价值。从文本流中自动检测突发事件是很有价值的研究内容。这些文本流中的突发事件，在一定时间窗口内，随着事件的突然发生，表现为引起媒体的大量报道与人们的持续关注[1]。建立自动突发事件检测系统，能为有关用户(如应急部门或个人)获取网络上所呈现的突发事件提供极大便利。

本文提出一种基于 MapReduce[2]的检测方法，从文本流中挖掘出用户感兴趣的特定领域的突发事件，并在开源软件 Hadoop[3]上加以执行。实现了突发词计算的并行化，为大规模并行化的突发事件检测提供了实验框架。该方法主要分为 4 个步骤：并行语料预处理、并行突发词检测、并行突发文档过滤和并行主题提取。实验结果表明，该方法能够有效地挖掘特定领域的突发事件。在大数据环境下，这种基于分布式的并行计算框架在挖掘突发事件的可延拓性方面具有一定意义。

2　相关工作

在大数据环境下，基于 MapReduce 的突发事件检测涉及到主题检测、突发词检测和 MapReduce 并行计算等方面的研究。

主题检测最初是来自主题检测与追踪(Topic Detection and Tracking,TDT)[4]，其研究要早于突发检测。在主题模型的研究方面，先后出现了 PLSA[5]和 LDA[6]两个经典的主题检测模型。PLSA 和 LDA 两者都是概率生成模型，与 PLSA 相比，LDA 将主题混合权重视为 K 维参数的潜在随机变量，而非与训练数据直接联系的个体参数集合，在推理上采用 Laplace 近似、变分近似和 MCMC 等方法获取待估参数值，所以

通讯作者：虞为，ORCID: 0000-0003-1933-5380，E-mail: luckjp@163.com。

*本文系国家社会科学基金项目“基于关联数据的图书馆语义云服务研究”(项目编号:12CTQ009)、国家社会科学基金重大项目“面向突发事件应急决策的快速响应情报体系研究”(项目编号: 13&ZD174)、国家自然科学基金面上项目“面向知识服务的知识组织模式与应用研究”(项目编号: 71273126)和江苏省社会科学基金青年项目“基于语义云服务的数字阅读推广研究”(项目编号:14TQC003)的研究成果之一。

在处理新文档方面更有优势[7]。近年来，针对新的应用场景，特别是社交媒体的文本流，出现了一批基于这两个经典模型的改进方法[8-11]。此外，一些学者考虑到 LDA 主题模型在主题数难以控制的情况下，存在对主题生成结果影响较大的缺点，因此提出采用凝聚式聚类算法对文档进行主题表示[12-15]。凝聚式层次聚类算法通过抽取文本特征，以向量形式表达文本，通过计算文本之间相似度进行聚类，然而这种文本向量间的距离在高维空间中难以确定。同时，聚类结果也只是起到对文本进行类别划分的作用，并没有提取描述该类别的主题词。

在突发词检测方面，Kleinberg[16]提出一种基于隐马尔科夫模型(Hidden Markov Models, HMM)的状态机，建立一个文本流中消息到达时间的模型。这种消息到达的时间间隔符合指数分布。Ihler 等[17]基于时间变化的泊松过程模型构建一个时间序列计数数据(Count Data)无监督学习的框架，并发现了一些异常事件。随后的许多有关“突发词”的研究成果，都是基于对 Kleinberg[16]和 Ihler 等[17]的方法加以改进[11,18-20]。He 等[21]和 Xie 等[1]则利用物理上的速率、冲量和加速度作为一个指示器，从而检测突发事件的发生。邱云飞等[15]通过计算当前时间窗口内信息集与已知话题的相似度，过滤出潜在突发文档。王勇等[14]根据微博文本的特点提出以“热点性”、“突发性”和“重要性”三项指标抽取突发词。以上方法各有其适用性和局限性，通过易用性和准确性综合考量，本文选取 Kleinberg[16]的思想检测突发词。

MapReduce 是目前最为成功的、面向大规模数据的和基于集群而非单机的并行计算抽象方法。近年来，对 MapReduce 的理论和应用的研究取得很多成果。在应用的研究方面，MapReduce 已涉及自然语言处理、机器学习和大规模图处理等领域[22]。例如 Das 等[23]利用 MapReduce 框架实现了 MinHash 聚类、PLSA 和访问计算的并行化算法，用于 Google 新闻用户个性化推荐的协同过滤。Choi 等[24]设计出用于大规模 XML 数据标号的并行标号树算法，并在 MapReduce 平台上加以执行，结果显示该算法比传统算法在标号速度上提高了 17 倍。刘滔等[25]提出一种基于 MapReduce 框架的条件随机场模型训练并行化方法进行词性标注，大大降低了训练时间。Mahout 项目作为可以与 Hadoop 无缝衔接的可扩展机器学习库，已开源了 K-means、LDA、SVD、LR 分类器和朴素贝叶斯分类器等分类聚类的并行算法，使得用户能够更加方便快捷地创建分布式的应用程序[26]。

目前针对突发事件的检测方法大部分仅限于串行计算，个别并行算法也只是采用多线程单机模式，这两者都有一定的局限性。因为当文本流大到无法一次性载入内存的时候，这两者都很难在较短的时间内完成计算。而 MapReduce 分布式处理框架则能很好地解决这种大数据问题。为了能够从文本流中快速检测出突发事件，本文构建了基于 MapReduce 并行计算平台的突发事件自动检测模型，并以主流媒体的新闻文本流进行仿真实验，从中挖掘出与特定领域相关的突发事件。

3　突发事件检测并行算法

3.1　总体框架

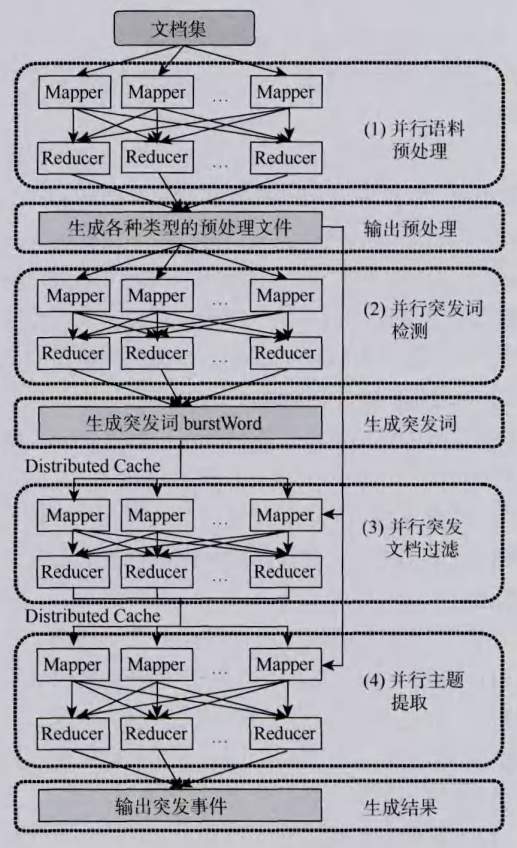

图 1　突发事件检测并行计算总体框架

本文提出的基于MapReduce框架并行突发事件检测方法，主要通过 4 个步骤完成突发事件的检测，整体流程框架如图 1 所示。控制程序 Driver 操控整个流程，将可并行操作的步骤以任务的形式提交给 MapReduce 平台，具体步骤如下：

(1) 并行语料预处理。并行扫描一次文档集，过滤标题和内容相同、只是时间上相差不到半天的重复新闻信息，输出去重后的文档集。再并行扫描一次去重后的文档集，将每个词进行词形还原(此处主要针对英文文档)，去除停用词。最终输出：每个词在时间窗口上的相关文档数和每个时间窗口上的总文档数；每个词和与其对应的索引号的映射表；每篇文档用词的索引号表示的新文档。并将这些文件以序列文件(SequenceFile)的形式存储在分布式文件系统中，用于并行突发词检测、并行文档过滤和并行主题提取。

(2) 并行突发词检测。基于 Kleinberg[16]的突发检测方法，针对每个潜在的突发词，根据在时间窗口内与其相关的文档出现的频次，判断该词是否为突发词，如果是则输出。需要说明的是，此处实现的是词与词之间的并行，但针对一个词的判断则是采用串行模式。

(3) 并行突发文档过滤。扫描步骤(1)中所输出的新文档集，根据突发词以及突发词和文档的关联阈值(RMin∈1,2,3,…)，判断该文档是否属于潜在的突发文档。如果是潜在突发文档，则将其文档号 DocID 输出。关联阈值 RMin 是指文档中的词出现在突发词集中的最小个数，这个 RMin 的大小与所检测的语料有很大关系，如果是新闻文档则RMin就会高些，而如果是微博等短文本语料则 RMin 相对较小。也就是说，RMin 的大小与语料中单篇文档的长度直接相关，且呈正比关系。经突发词过滤后的潜在突发文档，在后续主题提取中降低文档规模和提高结果准确性方面都起到很大作用。

(4) 并行主题提取。根据潜在突发文档的文档号 DocID，过滤不相关文档，再进行主题模型训练和主题推断。根据 LDA 变分法所计算出的词–主题矩阵和主题–文档矩阵，得出使用主题词描述的突发事件以及与该主题相关的文档。

以下将详细介绍本文并行突发事件检测方法涉及的两个主要算法：并行突发词检测，并行主题提取。

3.2 并行突发词检测算法

本文的突发词抽取是基于 Kleinberg[16]的突发检测方法实现的。这种检测方法通过计算消息平均到达间隔的概率分布变化来完成。消息以概率的方式进行发送，第 i 个消息与第 i+1 个消息之间的时间间隔为 x，它符合无记忆性指数分布：$f(x;\phi)=\phi e^{-\phi x}(\phi>0)$。设 ϕ_0 为单位时间内正常状态平均消息数，ϕ_1 为突发状态平均消息数，则有 $\phi_1=S\cdot\phi_0, S>1.0$，其中 S 表示突发状态平均消息数与正常状态平均消息数之间的线性关系系数。如果 S 值越高，则只有更加明显的突发才能被检测到。

对于一个非负的时间间隔序列 $\mathbf{x}=(x_1, x_2,\cdots, x_n)$，目标就是找到一个最优的状态序列 $\mathbf{q}=(q_{i_1}, q_{i_2},\cdots, q_{i_n})$ 使得公式(1)取得最小值：

$$c(\mathbf{q}\mid\mathbf{x})=b(\mathbf{q})\ln(1-p)/p)+\left(\sum_{t=0}^{n}-\ln(f_{i_t}(x_t))\right) \qquad (1)$$

其中，b(**q**)表示状态转移的个数，p 表示状态转移的概率，$f_{i_t}(x_t)=\phi_{i_t}e^{-\phi_{i_t}x_t}$ 表示在时间间隔为 x_t、到达速率为 ϕ_{i_t} 的指数密度函数。

计算最优的状态序列，采用标准的 HMM 的前向动态规划算法完成。为了进行并行化突发词抽取，本文实现了关于突发词的 MapReduce 并行框架。该框架主要分为 Map 和 Reduce 两个阶段。

(1) Map 阶段

在Map阶段，每个Mapper节点都尽可能地为本节点上的每个潜在突发词计算出最优的状态序列。在计算过程中，过滤掉高频词和低频词，根据HMM的前向动态规划算法计算潜在突发词的最优状态序列。一般情况下，状态数q取值为 2，表示在任何时刻只存在两种状态，即正常状态和突发状态。在时间窗口t上，通过计算最小成本函数，即可得知当前最有可能处于哪种状态。这种最优的状态序列，如果存在突发状态且符合一定的突发时间跨度和突发度阈值，则将该潜在的突发词输出为实际突发词(burstWord)。具体Map过程如算法 1 所示：

算法 1 突发词检测 Mapper 类
输入：
键(Key)：潜在的突发词(word)
值(Value)：(1)每个时间窗口上与潜在突发词相关的文档数 N^t

(2) 每个时间窗口上文档集的总文档数 M^t

输出:

键(Key): 突发词(burstWord)

值(Value): null

```
Map
T ← number of time window   //时间窗口数
q ← number of states //状态数
φ1 ← k / n,     φ0 ← S·φ1  (S<1.0)
    //非正常状态和正常状态消息到达的速率
γ, β ← state change control parameter   //状态转移控制参数
π_{lowerstate→higherstate}(π_{H→L}) ← r·ln(T+1) − ln(β)
    //从低状态转移到高状态的成本花费
π_{higherstate→lowerstate}(π_{H→L}) ← 0
    //从高状态转移到低状态的成本花费
Cost_0(0) ← 0,   Cost_1(0) ← ∞
    //初始化第 0 个时间窗口的成本花费
for t←1 to T begin              //开始计算每个时间窗口上的状态
  for j←0 to q-1 begin
    Cost_j(t) ← −(Σ_{i=1}^{N^t} (−ln i) + Σ_{i=N^t+1}^{M^t} ln i + N^t·ln(1/φ_j) + (M^t − N^t)·ln(1 − 1/φ_j))
  min TotalCost_j(t) ← Cost_j(t) + (min TotalCost (t−1) + π_{L→H})
  end
  min TotalCost(t) ← min TotalCost_0(t)
    //默认正常状态是成本花费最小的
  for j←0 to q-1 begin
    if  min TotalCost(t) > min TotalCost_j(t)  then
      min TotalCost(t) ← min TotalCost_j(t)
      setPath(t) ← j
        //记录时间窗口为 t 时，花费成本最小的状态
    end
  end
end
if word.isValidBurstCandidate then
  //如果满足突发词条件，则发射出去
   Emit <word, null>
end
```

(2) Reduce 阶段

经过 Map 阶段，每个符合条件的突发词会通过 Reduce 阶段直接发射出去，如算法 2 所示。突发检测结束之后，其结果按照突发词的形式存储在分布式文件系统中，用于下一步过滤相关文档。

算法 2 突发词检测 Reducer 类

输入:

键(Key): 突发词(burstWord)

值(Value): null

输出:

键(Key): 突发词(burstWord)

值(Value): null

```
Reduce
  Emit(burstWord, null);
```

3.3 并行主题提取

需要对突发词和经突发词过滤后得到的相关突发文档进一步进行主题提取。通过主题模型提取出的主题词，能够自动地概括文档所描述的主要事件，以减少工作量。同时，由主题模型所推断出与主题相关的文档，使得用户在需要查看该主题所对应的具体消息时，变得十分方便。本文采用的并行主题模型是基于标准的 LDA 模型实现的。LDA(Latent Dirichlet Allocation)称为隐含狄利克雷分布，是 Blei 等[6]提出的一种用于离散数据集合的生成概率模型。其基本思想认为文档是由潜在的一些主题随机组合而成的，而每个主题又由词组成，即一个文档中可能包含多个主题，一个词也可能同时属于多个主题。本文使用 Nallapati 等[27]和 Zhai 等[28]的方法对 LDA 进行并行化，并实现相关的 MapReduce 算法。LDA 并行计算的运行框架如图 2 所示:

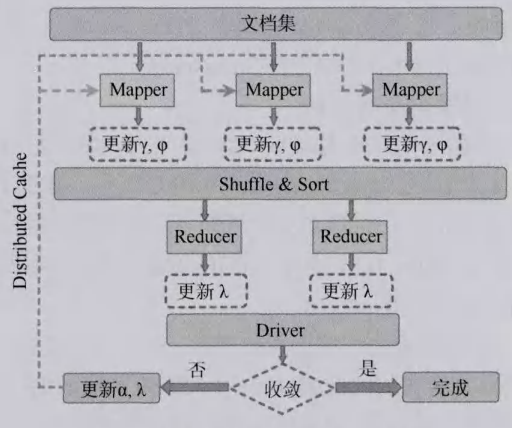

图 2　并行 LDA 计算的运行框架

4 实验与结果分析

4.1 数据来源与评价标准

在事件报道方面，主流媒体较社会化媒体能够获得更加可靠的信息。而且主流媒体在面向如综合新闻、财经、体育等特定领域的事件传播中也较社会化媒体更有针对性。因此，本文选择主流媒体的新闻信息作为突发事件检测的实验数据来源。通过 Yahoo News 实时推送接口 RSS①，采集从 2014 年 4 月 5 日至 7 月 7

① Yahoo News 的 RSS Feeds 获取接口为: http://news.yahoo.com/rss/.

日共三个多月的新闻信息，最终获得 12 023 条新闻数据。在所采集的 Yahoo News 数据中，通过人工识别，总共包含 9 个较大的突发事件。需要说明的是，所选取的突发事件主要是社会影响性较大(特别是社会危害性较大)的领域事件，文档集中突发事件数的确定或多或少存在着主观性，但已尽可能客观准确地筛选出文档集中所包含的特定领域的突发事件。设置一天作为时间窗口，对数据中的突发事件进行检测。

在人工筛选突发事件的过程中，将消息时间跨度阈值设为 10 个时间窗口，以表示该事件的持续影响性较大。如土耳其矿难，持续时间只有 6 个时间窗口，相对于测试集中的 94 个时间窗口来说，这种事件的影响不是很大，所以未将这类事件归为本实验的较大突发事件。还有如某事件，由于在本语料集中所呈现的信息是分别为 4 月 30 日、5 月 22 日和 6 月 21 日发生的独立事件，单从语料的字面含义很难看作同一事件，所以也未将这种语义上可以看作同一事件的事件，纳入较大突发事件集合中。评价标准采用信息检索领域常用的三大评价指标：准确率 P(Precision)、召回率 R(Recall)和调和平均值 F(F-measure)。

4.2 检测结果与分析

实验环境采用 Hadoop 云计算平台，共有 11 个节点：1 个主控节点(NameNode)和 10 个工作、数据节点(DataNode)。每个节点的配置如下：单核处理器 Intel Core i5-3470，1GB 内存，20GB 磁盘空间，百兆网卡，操作系统为 Ubuntu 10.10。

对数据集进行并行预处理后，就进入突发词检测阶段。在突发词的检测过程中，突发度阈值μ和突发词时间跨度阈值大小的不同设置，对各突发类型的突发词检测具有较大影响。突发度阈值μ是指消息突发状态从低层向高层转变所需花费的最小成本。本文认为大部分事件一般处于正常状态，只有少部分重大事件在不确定的时间发生，从而产生突发状态。μ值设得越高，则只有更加明显的突发词才能被检测到；反之，则那些突发强度不是很高的突发词也会被检测到，从而增大相关文档的数量，同时主题数也会相应增加，影响实验结果。

将状态转移控制参数γ，β都设为 1.0，状态数 q 取 2，时间窗口跨度阈值取 10，高频词设为频次大于总文档量的 2.50%，低频词设为频次小于总文档量的 0.20%，将突发度阈值μ设为 0.00000000，得到各突发词。取突发度最大的前 30 个突发词，其结果如表 1 所示。再将主题数 K 取 9，突发度阈值μ取 0.00000000 至 0.99999000 得到突发事件检测的不同准确率、召回率和调和平均值，实验结果如表 2 所示。

表 1 Top30 各突发词及其突发度

突发词	突发度	突发词	突发度
show	0.99999663	fire	0.99999409
miss	0.99999640	dead	0.99999356
government	0.99999629	10	0.99999348
city	0.99999602	attack	0.99999334
kill	0.99999592	southern	0.99999329
leave	0.99999548	student	0.99999315
search	0.99999540	building	0.99999309
people	0.99999535	make	0.99999297
official	0.99999516	house	0.99999293
police	0.99999499	state	0.99999291
home	0.99999478	center	0.99999289
president	0.99999477	murder	0.99999282
washington	0.99999472	water	0.99999276
file	0.99999460	australia	0.99999266
ukraine	0.99999415	authority	0.99999260

表 2 不同突发度阈值μ对突发事件检测的影响

突发度阈值μ	P(%)	R(%)	F(%)
0.00000000-0.99998484	77.78	77.78	77.78
0.99998484-0.99998884	55.56	55.56	55.56
0.99998884-0.99999000	33.33	33.33	33.33

由表 2 可见，当突发度阈值取 0.99998484 以下时，突发事件检测的调和平均值 F 都处于恒定的 77.78%；而将突发度阈值增大时，调和平均值 F 逐渐下降。前者说明即使增大相关文档数，结果的召回率也很难提高，这是因为将主题数 K 设为 9，在主题检测时至多只能检测到 7 个相关的突发事件，该问题的原因将详细阐述。而后者将突发度阈值增大，那么相关的潜在突发文档就会被过滤更多，从而相关的突发事件也会变少，因此在准确率和召回率方面都有所下降。另外，阈值μ取在一个区间内，得到的准确率、召回率和调和

平均值都处于一个恒值，这是因为检测时将主题数设为固定值，而检测出的突发事件结果数在μ的不同区间内也一致，只是表达突发事件的主题词有些许变化，但变化的主题词仍不影响人们对突发事件的识别。当然，突发度阈值取越小，用于表达突发事件的主题词就越容易被越多的噪声词所干扰，造成一部分不相干的主题词在表达突发事件的时候比例较高。也就是说，虽然准确率、召回率和调和平均值对不同突发阈值μ不是很敏感，但实质上用于表达突发事件的主题词已发生局部变化。因此，突发度阈值μ的有效选取，对突发事件检测系统效果的最佳发挥具有重要作用。

获得突发词之后，需要根据突发词在文档中出现的次数过滤一些不相关的文档，以利于无监督的 LDA 主题检测。这个过滤的过程是根据文档中出现突发词的频次进行比对，如果文档出现突发词的频次超过一定的阈值 RMin，则将此文档认为是相关的突发文档。

在 LDA 主题检测时，两个超参数α和η的设定在时间方面对主题模型的训练和推断具有一定的影响。它们初始值的设定与具体的语料有关，两者值越小，说明想要表达的主题也越少，即一篇文档和一个词赋予一个主题的概率就越大，因此选择合适的α和η可以提高 EM 计算收敛的速度。本实验的语料是新闻数据，一般一篇文档主要表明一个事件，特别是突发事件的主题词一般比较少，同时可能属于多个突发事件，为了简化参数估计的难度，经前期多次测试，将α的各分量统一设为 25/K，η的各分量也统一设为 13/K。

主题数 K 的选取对检测结果的影响相当大，表 3 列出了不同主题数所对应的准确率、召回率和调和平均值，根据表 3 的前 25 项的数据可画出主题数 K 和调和平均值 F 的关系图，如图 3 所示，调和平均值 F 与主题数 K 不是呈正比或者反比线性关系,而是为向上凸的曲线关系。当主题数 K 取 8 时，准确率、召回率和调和平均值分别为 87.50%，77.78%和 82.35%，而 K 大于或小于 8 时，调和平均值都会变小。这说明本文方法至多能找到 80%左右的主题。设置 K 为 8，突发度阈值为 0.99998484，得到的突发事件检测结果如表 4 所示。

表 3　不同主题数 K 对突发事件检测的影响

主题数 K	P(%)	R(%)	F(%)
1	100.00	11.11	20.00
2	100.00	22.22	36.36
3	100.00	33.33	50.00
4	100.00	44.44	61.54
5	100.00	55.56	71.43
6	100.00	66.67	80.00
7	85.71	66.67	75.00
8	87.50	77.78	82.35
9	77.78	77.78	77.78
10	70.00	77.78	73.68
11	63.64	77.78	70.00
12	58.33	77.78	66.67
13	53.85	77.78	63.64
14	50.00	77.78	60.87
15	46.67	77.78	58.33
16	43.75	77.78	56.00
17	41.18	77.78	53.85
18	38.89	77.78	51.85
19	42.11	88.89	57.14
20	40.00	88.89	55.17
21	38.10	88.89	53.33
22	36.36	88.89	51.61
23	34.78	88.89	50.00
24	33.33	88.89	48.48
25	32.00	88.89	47.06
…	…	…	…
47	17.02	88.89	28.57
48	18.75	100.00	31.58
49	18.37	100.00	31.03
50	18.00	100.00	30.51

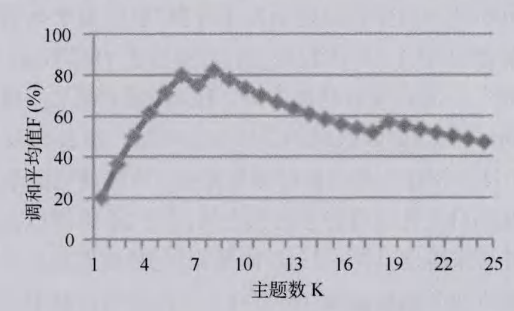

图 3　主题数 K 和调和平均值 F 的关系图

表 4　突发事件检测结果

突发时间段	突发主题	相应事件文档样例		事件描述
4.6-7.7	kiev, moscow, putin, pro-russia, protester, separatist, slovyansk, crisis, warn	(1) 2014-4-6 22:56 (2) 2014-4-7 15:55 (3) 2014-4-16 11:32	Pro-Russians storm Ukraine government buildings. Ukraine says Russia fomenting eastern unrest. Putin warns Ukraine on brink of civil war as Kiev sends army in.	乌克兰东部冲突
4.16-5.19	korea, coast, water, passenger, ferry, korean,ship,disaster, sink	(1) 2014-4-16 14:55 (2) 2014-4-17 19:51 (3) 2014-4-19 6:12	295 missing after S. Korea ferry sinks. Hundreds still missing in Korea ferry accident. Captain of sunken South Korean ferry arrested.	韩国沉船
4.16-6.3	girl, injure, nigeria, crash, kidnap, abducted, extremist, boko, haram	(1) 2014-4-16 19:16 (2) 2014-4-17 7:33 (3) 2014-5-10 23:15	Nigerian officials search for more than 100 abducted schoolgirls. Nigerian military: Most of more than 100 abducted schoolgirls freed. First lady's address on kidnapped Nigerian girls.	尼日利亚女孩被绑
6.10-6.23	iraq, baghdad, iraqi, control, fighter, Sunni, border, capture, levant	(1) 2014-5-12 4:37 (2) 2014-6-11 4:18 (3) 2014-6-19 22:40	Gunmen storm Iraqi military barracks, killing 20. Militants overrun most of major Iraqi city. Iraqi forces fight for control of oil refinery.	伊拉克内部军事冲突
5.27-7.6	cup, brazil, soccer, match, beat, watch, fan, score, rio	(1) 2014-5-27 0:12 (2) 2014-6-13 1:10 (3) 2014-6-18 16:17	AP PHOTOS: Brazil in countdown for World Cup start. Clashes in Brazil as World Cup begins. Spain playing for World Cup survival vs Chile.	巴西世界杯
4.18-7.2	judge, ban, strike, gay, ruling, supreme, marriage, rights, couple	(1) 2014-4-18 6:08 (2) 2014-5-20 1:32 (3) 2014-6-7 12:11	Judge in gay marriage case asks pointed questions. Ruling expected Monday on Oregon gay marriage ban. Gay couples begin getting married in Wisconsin.	同性恋结婚禁令正在部分地区取消
5.15-7.7	republican, capitol, hill, care, affairs, veterans, veteran, senate, committee	(1) 2014-5-15 1:24 (2) 2014-5-22 7:17 (3) 2014-6-6 4:24	Secret lists, deaths: Claims roil Veterans Affairs. Obama vows fix to veterans' health care troubles. Senators reach agreement on veterans' health care.	美国退伍老兵及其健康医疗

(注: 事件描述一栏为人工标注。)

除表 4 所列的突发事件外, 还有两个较大突发事件未能有效地被检测, 分别是“叙利亚政府与反对派在霍姆斯发生冲突”和“中国与越南紧张局势”。“叙利亚政府与反对派在霍姆斯发生冲突”事件由于该消息的集合中共同、频次较高且有自身特色的词不多, 仅为两个: “Syrian(叙利亚的)”, “Homs(霍姆斯)”。而且“Homs”出现的频次也不是很高, 再加上“Rebel(反对派)”这个词作为高频词已被剔除掉, 所以使得该事件没有具有事件自身特色的核心词, 因此给检测该突发事件带来难度。“中国与越南紧张局势”事件难以成功检测也有类似的原因: 该事件在不同时刻消息中主要共同且频次较高的词有“China(中国)”和“Vietnam(越南)”, “China”这个词在整个检测的消息文档集内从属于多种不同的主题, 不是该事件所特有的, 从而使得在主题检测时, 难以有效地将该事件的消息集聚成一类事件。

5　结　语

本文提出一种基于MapReduce实现的并行突发事件检测方法。该方法主要分为并行语料预处理、并行突发词检测、并行突发文档过滤和并行主题提取 4 个步骤。实验结果表明, 本文的突发检测方法在采集的 Yahoo News 国际主流媒体语料中, 准确率、召回率和调和平均值最高分别可达 87.50%、77.78%和 82.35%。

本文提出的基于MapReduce框架的并行突发检测方法是基于分布式环境构建的，所以具有很好的可扩展性，是大数据时代进行突发事件检测的一种重要的研究方法，有一定应用价值。

然而，该并行检测方法局限于对离线(Offline)静态新闻文本流进行模拟仿真式的批量处理检测，以获得突发事件。对于大规模动态文本流在线(Online)实时(Real-time)检测则是进一步采用流式大数据处理平台进行研究的内容。同时，由于未能获得中文主流媒体信息，针对中文环境下突发检测，本并行方案是否仍然有效则需进一步研究。

[1] Xie W, Zhu F, Jiang J, et al. TopicSketch: Real-Time Bursty Topic Detection from Twitter [C]. In: Proceedings of the 13th International Conference on Data Mining, Dallas, Texas, USA. IEEE, 2013: 837-846.

[2] Dean J, Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters [J]. Communications of the ACM, 2008, 51(1): 107-113.

[3] Hadoop [EB/OL]. [2014-07-15]. http://hadoop.apache.org/.

[4] Allan J, Carbonell J, Doddington G, et al. Topic Detection and Tracking Pilot Study Final Report [C]. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998: 194-218.

[5] Hofmann T. Probabilistic Latent Semantic Analysis [C]. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., 1999: 289-296.

[6] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. The Journal of Machine Learning Research, 2003, 3: 993-1022.

[7] 李文波，孙乐，张大鲵. 基于 Labeled-LDA 模型的文本分类新算法[J]. 计算机学报, 2008, 31(4): 620-627. (Li Wenbo, Sun Le, Zhang Dakun. Text Classification Based on Labeled-LDA Model [J]. Chinese Journal of Computers, 2008, 31(4): 620-627.)

[8] Wang X, Zhai C, Hu X, et al. Mining Correlated Bursty Topic Patterns from Coordinated Text Streams [C]. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2007: 784-793.

[9] Lin C X, Zhao B, Mei Q, et al. PET: A Statistical Model for Popular Events Tracking in Social Communities [C]. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2010: 929-938.

[10] Dubrawski A. Detection of Events in Multiple Streams of Surveillance Data [A].// Infectious Disease Informatics and Biosurveillance [M]. Springer US, 2011: 145-171.

[11] Diao Q, Jiang J, Zhu F, et al. Finding Bursty Topics from Microblogs [C]. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju Island, Korea. 2012: 536-544.

[12] 周刚，邹鸿程，熊小兵，等. MB-SinglePass: 基于组合相似度的微博话题检测[J]. 计算机科学, 2012, 39(10): 198-202. (Zhou Gang, Zou Hongcheng, Xiong Xiaobing, et al. MB-SinglePass: Microblog Topic Detection Based on Combined Similarity [J]. Computer Science, 2012, 39(10): 198-202.)

[13] 郭跬秀，吕学强，李卓. 基于突发词聚类的微博突发事件检测方法[J]. 计算机应用, 2014, 34(2): 486-490. (Guo Yixiu, Lv Xueqiang, Li Zhuo. Bursty Topics Detection Approach on Chinese Microblog Based on Burst Words Clustering [J]. Journal of Computer Applications, 2014, 34(2): 486-490.)

[14] 王勇，肖诗斌，郭跬秀，等. 中文微博突发事件检测研究[J]. 现代图书情报技术, 2013(2): 57-62. (Wang Yong, Xiao Shibin, Guo Yixiu, et al. Research on Chinese Micro-blog Bursty Topics Detection [J]. New Technology of Library and Information Service, 2013(2): 57-62.)

[15] 邱云飞，程亮. 微博突发话题检测方法研究[J]. 计算机工程, 2012, 38(9): 288-290. (Qiu Yunfei, Cheng Liang. Research on Sudden Topic Detection Method for Microblog [J]. Computer Engineering, 2012, 38(9): 288-290.)

[16] Kleinberg J. Bursty and Hierarchical Structure in Streams [C]. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2002: 91-101.

[17] Ihler A, Hutchins J, Smyth P. Adaptive Event Detection with Time-Varying Poisson Processes [C]. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.New York: ACM, 2006: 207-216.

[18] Nakahara T, Hamuro Y. Detecting Topics from Twitter Posts During TV Program Viewing [C]. In: Proceedings of the 13th International Conference on Data Mining, Dallas, Texas, USA. IEEE, 2013: 714-719.

[19] Zhang L, Jia Y, Zhou B, et al. Detecting Real-Time Burst

Topics in Microblog Streams: How Sentiment Can Help [C]. In: Proceedings of the 22nd International Conference on World Wide Web Companion. 2013: 781-782.

[20] Koike D, Takahashi Y, Utsuro T, et al. Time Series Topic Modeling and Bursty Topic Detection of Correlated News and Twitter [C]. In: Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP), Nagoya, Japan. 2013: 917-921.

[21] He D, Parker D S. Topic Dynamics: An Alternative Model of Bursts in Streams of Topics [C]. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.New York, USA: ACM, 2010: 443-452.

[22] 李锐, 王斌. 文本处理中的 MapReduce 技术[J]. 中文信息学报, 2012, 26(4): 9-20. (Li Rui, Wang Bin. MapReduce in Text Processing [J]. Journal of Chinese Information Processing, 2012, 26(4): 9-20.)

[23] Das A S, Datar M, Garg A, et al. Google News Personalization: Scalable Online Collaborative Filtering[C]. In: Proceedings of the 16th International Conference on World Wide Web. New York: ACM, 2007: 271-280.

[24] Choi H, Lee K H, Lee Y J. Parallel Labeling of Massive XML Data with MapReduce [J]. Journal of Supercomputing, 2013, 67(2): 408-437.

[25] 刘滔, 雷霖, 陈荦, 等. 基于 MapReduce 的中文词性标注 CRF 模型并行化训练研究[J]. 北京大学学报: 自然科学版, 2013, 49(1): 147-152. (Liu Tao, Lei Lin, Chen Luo, et al. A Parallel Training Research of Chinese Part-of-Speech Tagging CRF Model Based on MapReduce [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2013, 49(1): 147-152.)

[26] What is Apache Mahout? [EB/OL]. [2014-09-27]. http://mahout.apache.org/.

[27] Nallapati R, Cohen W, Lafferty J. Parallelized Variational EM for Latent Dirichlet Allocation: An Experimental Evaluation of Speed and Scalability [C]. In: Proceedings of the 17th IEEE International Conference on Data Mining Workshops, Omaha, Nebraska, USA. IEEE, 2007: 349-354.

[28] Zhai K, Boyd-Graber J, Asadi N. Using Variational Inference and MapReduce to Scale Topic Modeling [OL]. Eprint arXiv, 2011. arXiv: 1107.3765.

卓可秋: 设计研究方案, 采集数据, 进行相关实验并分析结果, 论文起草与修订;
虞为: 提出研究思路, 最终版本修订;
苏新宁: 论文修订。

收稿日期: 2014-08-04
收修改稿日期: 2014-10-03

Parallel Implementing Bursty Events Detection Using MapReduce

Zhuo Keqiu Yu Wei Su Xinning
(School of Information Management, Nanjing University, Nanjing 210023, China)

Abstract: [**Objective**] In big data environment, this paper aims to accurately and quickly detect bursty events from the text stream. [**Methods**] Using Kleinberg bursty detection and LDA topic model, the method is extended to MapReduce framework to achieve parallel corpus predisposed, parallel detection of bursty word, parallel filtration of bursty document and parallel extraction of topic. [**Results**] The results of simulation experiments on the news text stream show that precision reaches 87.50%, recall reaches 77.78%, and F-measure reaches 82.35% with the parallel method to detect bursty events in specific areas. [**Limitations**] The MapReduce parallel method is difficult to achieve Online and Real-time detection of bursty events with large-scale dynamic text stream. [**Conclusions**] Compared with the traditional serial detecting method of bursty events, the distributed parallel method not only guarantees the accuracy of detecting results, but also has a good scalability.

Keywords: Bursty event detection MapReduce Distributed process LDA topic model