● 逯万辉[1,2]，马建霞[1]，赵迎光[1,2]
(1. 中国科学院　国家科学图书馆兰州分馆，甘肃　兰州　730000；2. 中国科学院　研究生院，北京　100049)

# 爆发词识别与主题探测技术研究综述*

摘　要：作为话题检测与追踪和舆情监测中的一项基础性工作，识别并处理爆发词对突发检测具有重要的作用，本文综述了该领域目前的研究现状和已有的研究成果并对其进行比较分析，总结了其中亟待解决的关键问题并进行了重点探讨，为后续研究指明了方向。

关键词：爆发词；热点话题识别；语义合并；综述

**Abstract:** As a basic work of topic detection and tracking and public opinion monitoring, identifying and processing burst word plays an important role in burst detection. This paper reviews the current research status in this field and makes a comparative analysis of the achieved results, summarizes and focuses on the key problems to be solved urgently, and points out the direction for future research.

**Keywords:** burst word; hot topic detection; semantic mergence; summary

近几年，网络舆情监测一直受到众多关注，真实、客观、完整和及时地获取信息是作出科学决策的基础。但是在信息爆炸的背景下，网络中每天都会出现大量的信息，涵盖众多领域，持续或长或短的时间，分析、处理并监测网络舆情对了解外部环境变化具有重要意义。爆发词(Burst Word)作为热点问题的直观表现，识别并处理爆发词对事件监控、机会发现和情态预测都具有重要意义。此外，在文献情报研究领域，针对学科热点和技术新兴点的知识发现也备受关注，正确、有效地捕捉潜在科技爆发词对科学研究趋势预测、研究热点和研发机会发现均有重要的研究意义和现实意义。

爆发词是指那种在一段时间大量出现的有意义的代表话题走向的词。因此，关于爆发词的识别，不单要考虑词频，还要结合上下文信息选取正确的有意义的词汇，同时也要计算该词汇出现的时间、消亡的时间和持续的时间段，进而实现话题探测和追踪的相关研究。本文将从主题探测与追踪的研究现状开始，对爆发词的特征识别与主题探测的关键问题进行分析归纳，提出后续的研究方向。

## 1　研究现状

突发主题的识别在许多文本挖掘领域都有重要应用，正确识别并处理爆发词是其基础性工作，也是其中的难点工作。关于爆发词的研究始于话题检测与追踪(Topic Detection and Tracking，TDT)领域，因此，本节内容先从话题检测与跟踪的研究现状入手，回顾话题检测与追踪的发展，进而引出爆发词识别的研究现状。

### 1.1　TDT研究进展

话题检测与追踪就是通过程序的方法自动处理新闻，将新闻切分成连续的事件报道的过程[1]。该技术可以用来监控各种语言资源，在新话题出现时发出警告，在信息安全、金融证券、行业调研等领域都有广阔的应用前景。

有关TDT的研究始于1996年，当时美国国防高级研究计划署(DARPA)根据自己的需求，提出要开发一种新技术，从1998年开始，在DARPA支持下，美国国家标准技术研究所(NIST)每年都要举办话题检测与跟踪国际会议，并进行相应的系统评测。由于话题检测与跟踪相对于信息检索、数据挖掘和信息抽取等自然语言处理技术具有很多共性，并且面向具备突发性和延续性规律的新闻语料，因此逐渐成为当前信息处理领域的研究热点。

与一般的信息检索或者信息过滤不同，TDT所关心的话题不是一个大的领域或者某一类事件，而是一个很具体的“事件”(Event)。在TDT研究的初级阶段，话题(Topic)与事件的含义基本相同。一个话题指由某些原因、条件引起，发生在特定时间、地点，并可能伴随某些必然结果的一个事件[2]，主要解决时间—事件的识别和抽取与映射的问题。随着研究的不断深入和外部环境的不断变化，现在有关话题的概念要相对宽泛一些，它包括一个

* 本文受中国科学院西部之光联合学者项目“基于计算情报方法的甘肃省战略新兴产业竞争发展研究”资助。

种子事件或活动以及所有与之直接相关的事件和活动[3]，主要研究事件之间的关联和相互影响程度的问题。与话题相对应的是主题（Subject），但是主题与话题不同，话题是与某个具体事件相关，而主题可以涵盖多个类似的具体事件或者根本不涉及任何具体事件。如“镍铬生产”是一个主题，而“关于镍铬生产过程中的排污治理”则是一个话题。

目前国内外在话题检测与追踪方面的主要技术有基于向量空间的方法[4-5]、基于统计语言模型的方法[6-8]、基于图论的方法[9-10]和基于突发检测的方法[11-13]。其中基于向量空间的方法主要是通过文本相似度计算和聚类技术实现话题的自动分类，进而发现热点话题；基于统计语言模型的方法是基于文本内容中词频及此项权重的计算为基础，从词项的出现概率出发研究热点问题；基于图论的方法来源于网络链接分析的有关方法，通过转移概率和转移矩阵来研究话题的演变，进而追踪话题发展；基于突发检测的方法应该说是以上几种方法的混合使用，通过对词项等文本内容的分析来发现爆发词和突发特征，并对这些特征进行组织处理，识别话题和主题，从而发现问题和追踪事态进展。

### 1.2 爆发词识别研究现状

针对爆发主题探测的研究，目前国内外关于主题探测的研究多集中于文本聚类、分类方法上，并探索了基于模糊聚类、层次化分类和周期分类、文献计量指标等方法的应用[14-17]，并开始逐步探索语义化的方法，运用语义模型来切分报道，并提出基于 SVM 的语义分类问题[18-19]。但不管是基于聚类的方法还是基于分类的方法，识别话题和主题都是其中核心和基础的工作，主题的识别以及语义化组织和基于语义化的话题识别与追踪将是下一阶段的主要研究方向。

关于爆发词的识别，多数的研究都基于词频的统计和词项权重的计算[20-21]，这些方法忽略了词义信息和上下文的语境，同时在特征提取和权重计算的研究上过于简单，在词串规整上也不够系统规范。爆发词识别作为主题探测的基础和新兴趋势监测的前兆，正确识别爆发词对话题识别和趋势预测具有重大影响。魏晓俊对基于词语的科技监测的方法进行了归纳，将它们分为词频分析方法、基于词网络关系的共词分析方法、基于词频变化率的突发监测方法、基于短语差异的分析方法[22]。其中突发监测算法是 Kleinberg 在 2002 年提出的[23]，它关注那些相对增长率突然增长的词，并认为话题的报道数量不是平滑增长，而是在不同水平之间跃迁，这种在一段时间内突然增长的词就是我们需要研究的爆发词。

针对以上关于爆发词识别与主题探测的研究总结和归纳，目前在爆发词识别中仍旧存在许多亟待解决的关键问题，包括爆发特征的识别、词的语义合并、爆发词出现时间（段）识别和关联事件的映射。接下来将对这些问题的研究进行探讨，以期为爆发词识别和主题探测技术的实现提供研究思路和技术线路基础。

## 2 关键问题探讨

### 2.1 爆发特征识别

爆发特征作为爆发词的候选基础特征项，正确识别爆发特征对最终爆发词的确定有重要影响。在对文本特征进行分析从而识别爆发特征的过程中，需要正确识别特征词项、未登录词和非结构性短语。

1）特征词项是具有信息价值的最小单位，英文文本中每一个词之间都有空格隔开，在词项处理中往往只需要进行词干还原（Stemming）和词形归并（Lemmatization），但中文中的最小单位是字，而具有意义的是词，因此就需要在识别爆发特征中首先进行词项切分，针对不同的应用和不同的领域，需要重点识别的词项内容也不尽相同，因此出现了命名实体识别、地名机构名识别、时间短语识别、术语识别等多种研究。这些研究的目的都是为了识别和抽取句子中最小的语义单元，并在一定程度上都取得了不错的研究成果。因此在进行爆发词识别时首先就需要先确定领域和明确目标，进行爆发特征的归纳和特征分析，为爆发词识别提供方向。

2）未登录词（Out-of-Vocabulary，OOV）是指通过某种途径产生的、具有基本词汇所没有的新形式、新意义或新用法的词语，并且是在分词系统中未被分词词典收录的词语，包括各类专名（人名、地名、企业字号和商标号等）、某些术语、缩略语和新词等[24]，正确识别未登录词对分词系统性能提升具有重要意义，也是进一步识别爆发词的基础性工作。有研究表明，未登录词造成的分词精度失落至少比分词歧义大 5 倍[25]。

3）短语识别是中文信息处理技术的一个重要研究方向。中文信息处理技术可分为字处理技术、词处理技术和句子级处理技术，目前词处理技术已经相对比较成熟，词处理技术也在逐渐完善，而句子级处理技术仍处于探索和研究阶段。短语识别作为句子的组成部分和句子级处理技术的基础，识别短语结构具有重要的研究价值。在爆发词的识别中，由于中文词的多义性、随意性和不规则性，非结构性短语识别对爆发词语义判断也具有重要影响。

### 2.2 语义合并及关联关系挖掘

由于中文词语的多样性和表达的随意性，经常会出现一词多义或者多词一义的现象。特别是在网络舆情监控和主题追踪领域中，往往出于表达方式的考虑，作者很少会

频繁使用同一个词语。因此在识别爆发词时，就需要先对识别出来的特征词进行语义合并和归类处理，然后进一步统计词频才有意义。此外，在词语的语义处理过程中，往往同一词语在不同的语境中也有不同的含义。例如，在处理“apple”这个词语的时候，有可能谈论的是食物“苹果”，也有可能是电子产品，此外，在讨论食物苹果时，我们可能会谈到目前的物价水平、CPI 指数等；在谈电子产品时，极有可能涉及“乔布斯”、“Mac”、“iphone”等。因此，在处理“apple”这一概念时，首先，就需要根据语义环境进行推理，研究这一概念的具体含义；其次，还需要进行关联信息分析，发现概念背后的隐藏信息。因此，对爆发词的词汇关系和语义关系合并主要需解决以下问题。

1) 词语间的语义关系确定及词义消除。在这方面的研究中，国内外学者都进行了较多的探索，其中较成熟的成果有英文的 Wordnet 词汇数据库和以此为基础开发的关于中文的 Hownet 词典。Wordnet 中将词语关系概括为同义关系、上下位关系、部分整体关系（包含与被包含关系）、反义关系和近义关系。在该词典词义归纳的基础上，众多研究人员采用基于规则的方法来进行短文本语义相似度计算和合语义并算法研究，但是该方法只是在词汇间关系的识别上进行了处理，并没有解决词汇语义推理问题。目前，已有研究者探索了基于该词典的领域本体构建方法，并进一步研究了概念推理机制。

2) 词语关联关系发现。关联规则目前已经在诸多领域都有应用，特别是在数据挖掘中，已有比较经典的案例和较为成熟的算法。目前，较多研究者试图将其引入到其他领域中，并也取得了一定的效果。在网络信息处理中，已有基于关联规则的知识发现研究和人际情报网络构建研究，这些研究在词语共现的基础上加入了关联规则中置信度和支持度的分析，将关系的发现引入到了更加深入的层面，并针对网络信息的时序性、动态性特征，提出了基于时间序列的事件研究，识别并处理时间—事件关系就成为其中不可或缺的一部分。在爆发词间关系的探测上，也可以参考关联规则的有关算法，进行爆发词间关系发现的研究。

### 2.3 词语生命周期识别

爆发词的识别和处理，主要是研究某一时间段内词语频率的变化率，在该研究过程中，除了需要正确识别爆发特征之外，还需要对爆发特征的起始出现位置和结束位置进行正确标记，从而计算词语的出现时间（段）。在基于文本内容分析和处理的话题检测与追踪、舆情监测、时序文摘等的研究中，对时间信息的识别和抽取也是该研究中的基础性工作。有研究表明，时间信息在文本信息中所占的比重仅次于专有名词。作为文本语义理解、语块分析等信息抽取中一项关键的研究技术，解决时间信息的抽取对机器翻译和人工智能领域的推进也具有重大研究意义，是一项重要的、基础性的工作。

英文语料的时间信息表达方式分为：时相（Phase）、时制（Tense）和时态（Aspect），因此在研究英文时间关系时就需要将时态等这些内容进行还原，进一步确定时间序列。这一研究中 Stevenson 等进行了动词情态分类工作的研究[26]，从而将时间信息的抽取扩展到了隐性时间信息领域。在中文时间信息抽取研究中，徐永东等提出了基于规则的时间信息抽取、理解及时间语义的计算方法[27]，将承载时间信息的短语按照不同功能分解成了若干容易识别的语义单元，重点研究了时间表达式的语义计算，但并没有深入研究时间短语、事件短语间的映射关系。赵国荣对事件类时间短语的特点进行了重点研究，并通过规则匹配树的方法进行了实验验证[28]，但受训练语料库的限制，该方法的扩展性有限。

目前，关于时间信息抽取的主要方法有：基于规则的方法、基于词典和自动机的方法、基于错误驱动的方法和基于机器学习的方法。由于时间表达式具有种类多样、属性多样、规则多变等特点，基于规则的方法就很难处理大规模复杂语料，因此，基于统计的方法的优势就比较明显。

## 3 结束语

网络环境下的爆发词识别及主题探测技术在近几年已经出现了较多研究，本文通过比较这些研究所采用的方法、思路和研究结果，综述该领域的研究现状，并对若干亟待解决的关键问题进行了重点分析介绍，对爆发特征识别、语义合并和爆发词关联关系挖掘、爆发词生命周期识别等问题的研究进行了总结归纳，对其中的关键技术进行了探索，明确了今后的研究方向和研究目标。但是爆发词识别与主题探测是一个比较复杂的问题，牵涉网络信息处理与数据挖掘等技术，需要对领域知识有较深入的了解，交叉性强，处理难度大，还需要全面系统深入的研究。□

参考文献

[1] ALLAN J. Introduction to topic detection and tracking, topic detection and tracking: event-based information organization [M]. Kluwer Academic Publishers, 2002: 1-16.

[2] ALLAN J, CARBONELL J, DODDINGTON G, et al. Topic detection and tracking pilot study: final report [C] // Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, San Francisco, CA, Morgan Kaufmann Publishers, Inc, 1998: 194-218.

[3] Task definition and evaluation plan (TDT2002) [EB/OL]. [2011-09-10]. http://www.itl.nist.gov/iad/mig/tests/tdt/tasks/detect.html.

[4] KUMARAN G, ALLAN J. Text classification and named entities for new event detection [C] // Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, ACM, 2004: 297-304.

[5] NALLAPATI R, FENG Ao, PENG Fuchun, ALLAN J. Event threading within news topics [C] // Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, New York, NY, USA, ACM, 2004: 446-453.

[6] BLEI D M, LAFFERTY J D. Dynamic topic models [C] // Proceedings of the 23rd International Conference on Machine Learning, New York, NY, USA, ACM, 2006: 113-120.

[7] LI Zhiwei, WANG Bin, LI Mingjing, et al. A probabilistic model for retrospective news event detection [C] // Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, ACM, 2005: 106-113.

[8] MEI Qiaozhu, LIU Chao, SU Hang, et al. A probabilistic approach to spatiotemporal theme pattern mining on weblogs [C] // Proceedings of the 15th International Conference on World Wide Web, New York, NY, USA, ACM, 2006: 533-542.

[9] KUMAR R, MAHADEVAN U, SLVAKUMAR D. A graph-theoretic approach to extract storylines from search results [C] // Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, ACM, 2004: 216-225.

[10] ZHAO Qiankun, LIU Tieyan, BHOWMICK S S, et al. Event detection from evolution of click-through data [C] // Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, ACM, 2006: 484-493.

[11] FUNGG P, JEFFREY X Y, PHLLIP S Y. Parameter free bursty events detection in text streams [C] // Proceedings of the 31st International Conference on Very Large Data Bases, VLDB Endowment, 2005: 181-192.

[12] HE Qi, CHANG Kuiyu, et al. Analyzing feature trajectories for event detection [C] // Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, ACM, 2007: 207-214.

[13] LAPPAS T, ARAI B, PLATAKIS M, et al. On burstiness-aware search for document sequences [C] // Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, ACM, 2009: 477-486.

[14] 刘红霞，乔晓东，张运良. 新兴趋势监测指标体系探索 [J]. 情报杂志，2009，29 (S1)：93-97.

[15] 于满泉，骆卫华，许洪波，等. 话题识别与跟踪中的层次化话题识别技术研究 [J]. 计算机研究与发展，2006，43 (3)：489-495.

[16] 税仪冬，瞿有利，黄厚宽. 周期分类和 Single-Pass 聚类相结合的话题识别与跟踪方法 [J]. 北京交通大学学报，2009 (5)：85-89.

[17] 鲁明羽，姚晓娜，魏善岭. 基于模糊聚类的网络论坛热点话题挖掘 [J]. 大连海事大学学报，2008，34 (4)：52-54，58.

[18] MAKKONEN J. Semantic classes in topic detection and tracking [D]. Helsinki: Department of Computer Science University of Helsinki, 2009.

[19] KONTONSTATHIS A, GALISTSKY L, PORTTENGER W, et al. A survey o f emerging trend detection in textual data mining, 2003 [EB/OL]. http://wenku.baidu.com/view/f4a835e819e8b8f67c1cb937.html.

[20] 曾依灵，许洪波，白硕. 网络文本主题词的提取与组织研究 [J]. 中文信息学报，2008，22 (3)：64-70，80.

[21] ZHAO Wayne Xin, JIANG Jing, HE Jing, et al. Context modeling for ranking and tagging bursty features in text streams [C] // Proceedings of the 19th ACM Conference on Information and Knowdedge Management (CIKM' 10), 2010: 1967-1772.

[22] 魏晓俊. 基于科技文献中词语的科技发展监测方法研究 [J]. 情报杂志，2007 (3)：34-39.

[23] KLEINBERG J. Bursty and hierarchical structure in streams [J]. Data Mining and Knowledge Discovery, 2003, 7 (4): 373-397.

[24] 魏莎莎. 一种中文未登录词识别及词典设计新方法 [D]. 重庆：西南大学，2011.

[25] 黄昌宁，赵海. 中文分词十年回顾 [J]. 中文信息学报，2007，21 (3)：8-19.

[26] STEVENSON S, MERLO P. Automatic verb classification using distributions of grammatical features [C] // Proceedings of EACL99, 1999: 45-52.

[27] 徐永东，徐志明，王晓龙，等. 中文文本时间信息获取及语义计算 [J]. 哈尔滨工业大学学报，2007，39 (3)：438-442.

[28] 赵国荣. 中文新闻语料中的时间短语识别方法研究 [D]. 太原：山西大学，2006.

作者简介：逯万辉，男，1987 年生，硕士生。
马建霞，女，研究馆员，硕士生导师。
赵迎光，男，硕士生。

收稿日期：2011－11－23