# DLDE社会新闻事件发现系统

陈杰，

北京理工大学

# 摘要xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

**关键词：**事件发现

# 1 引言

随着信息传播技术的进步，网络媒体正在逐渐代替传统的信息媒介成为人们获取信息的主要途径，而网络新闻平台作为网络信息传播的主要载体在近几年得到了极大地发展，各大网站都推出了自己的新闻平台，由于网络的实时性和快速传播等特性，这些平台对人们日常生活正在产生越来越强烈的影响，但是这些网络新闻平台也存在一些问题：目前网络中的信息规模急剧膨胀,而这些平台提供的信息大多是凌乱无序的,这使得对有价值的信息的发现和管理变得越来越难,特别是当用户需要了解事件的发展过程以及相关事件时,他们必须花费额外的时间和精力去进行搜索,用户体验较差。在这种情况下，一种能够自动将网络中相关新闻组织起来，形成相应事件的工具成为人们的迫切要求，因此，事件发现系统应运而生。

事件是指一段时间内人们关注的具有相同主题的新闻的集合，它一般是由事件发展过程中产生的新闻或具有主题相关性的新闻报道构成，因此，在一个事件集合中的新闻至少具有以下特点之一：第一，所有新闻具有相同的主题，即都是对同一件事情的报道和描述；第二，集合中的新闻描述的是某一事件的发展过程及其动态，比如有关雅安地震的新闻以及后续的一系列关于地震中的伤亡情况和救援情况也可属于同一事件；第三，集合中的新闻是对主题相关的事情的报道和描述。

本文在研究了已有的相关工作基础上，针对目前相关系统中存在的问题，设计了一种基于J2EE的事件发现系统。该系统面向互联网中各大主流新闻网站的新闻报道，能够自动地将主题相关的新闻聚合成为对应的事件，方便人们更有针对性的选择自己想要了解的信息，具有较高的实用价值，并且系统还能将新发布的新闻及时聚合到相应的事件中，具有一定的实时性。

本文接下来的结构如下：第二部分介绍相关的研究工作；第三部分详细描述我们设计的系统的框架和功能结构，并简单介绍用到的相关技术；第四部分，实验过程和结果分析；第五部分是我们对未来工作的展望。

# 2 相关工作

事件发现是话题检测与跟踪（TDT）技术在实际领域中的应用。目前大部分事件发现系统都是基于文本聚类技术和向量空间模型（Vector Space Model，VSM）来实现的，在这种情况下，研究者们提出了一系列的非常实用的话题跟踪和检测技术，比如，层次聚类技术、Single-Pass 聚类技术、增量K中心点聚类技术等。

[1]提出一种可伸缩的层次主题检测模型。通过使用密度函数来初始化类中心，论文[2]中提出一种改进的增量K中心点聚类方法，并使用该方法用于新闻事件的检测。与论文[3]中不同，A. P. Porrata等人结合分类和层次聚类的方法，提出一种改进的增量层次聚类方法来用于主题发现。黄胜[4]等人通过在聚类中使用命名实体提出了一种新的新闻事件检测方法——基于层次聚类和命名实体的新闻主题检测技术。[5]通过分析微博、互联网论坛等社交网络的特点，提出一种能够从存在噪声数据的文本数据集中发现事件的技术。在进行事件发现时，除了考虑新闻文本本身以外，还可以通过分析其他辅助信息提高事件发现的可能性，[6]通过对用户查询、新闻标题以及博客内容进行分析，提出一种基于用户查询指导的新闻事件发现方法。[12]认为文本中不同域的词项对于计算相似度时的贡献度不同，比如位于标题中的词项要比位于正文中的词项在相似度计算中的共现更大，基于这一结论，提出一种改进的层次聚类方法用于话题检测以及一种改进的single-pass聚类方法用于话题追踪。

对于新事件的检测,即当遇到一个新的新闻时,将它归于哪个事件也是事件发现的一个功能模块。Papka等人提出了一种Single-Pass聚类的方法[8]，对于一篇刚获取的新闻报道，首先进行预处理并表示成权值向量，然后计算该新闻与所有新闻的相似度，如果与所有新闻的相似度值均小于阈值，则将该新闻划分为一个新的事件。文献[9]中将刚获取的新闻与已经发现的事件进行比较，如果该新闻与事件的相似度大于阈值，则将该新闻归于相似度最高的事件，否则，将该新闻划分为一个新的事件。文献[10]结合Single-Pass聚类和新闻要素提出一种基于动态进化模型的事件检测算法，张阔[11]等人在考虑词元的基础上，对于不同类型的词元动态更新权重，进而计算新闻相似度的方法。

在本文中，我们设计了一个基于J2EE的新闻事件发现系统，在使用爬虫抓取新闻网页、网页文本信息抽取、文本相似度计算等步骤以后，使用无向图的方式将具有相同主题的新闻网页聚合在一起生成对应的事件。

# 3 系统架构和功能设计

在本节中，首先对系统框架和流程做概要描述，然后对系统的关键功能模块和使用的关键技术进行重点讨论。

## 3.1 系统框架

本系统采用J2EE架构进行设计，系统主要分为四个模块：新闻爬取模块、网页文本解析模块、文本相似度计算模块和事件发现模块，其中前三个模块都是为最终的事件发现做准备。系统的具体流程为：首先，使用网络爬虫爬取指定新闻网站的所有新闻网页并以文本的形式保存在本地；第二，对新闻文本进行解析，过滤文本中的无关信息（html标签等），抽取出我们需要的结构信息，如标题、正文、作者等信息；第三，对已经过滤好的新闻文本按照结构建立倒排索引；第四，通过分析第三步中建立的索引信息计算文本的相似度；最后，根据文本的相似度信息将主题相同的新闻聚合成事件。系统各模块的关系以及系统的整个流程如图1所示。

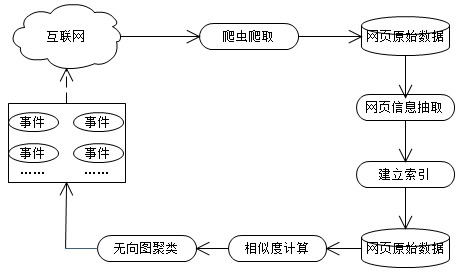


图1 事件发现系统流程图

## 3.2 功能模块设计

### 3.2.1 新闻采集模块

新闻采集模块主要是由网络爬虫(Crawler)来构成，它是整个系统的数据来源，决定着整个新闻系统的内容是否丰富以及新闻能否及时的更新。它的采集策略和采集性能将直接影响整个系统的效果。它的工作原理如下：根据指定的初始网址集合来获取对应的网页内容，从该网页内容中抽取出所包含的所有URL地址并分离出已经抓去过的URL地址，然后将新的URL地址放入待抓取的队列中，爬虫将重复上述过程直至待抓取队列为空或者达到其他预设的停止条件。

加入爬虫流程图

以社会新闻为主的新闻采集的主要采集策略是对抓取到的网页内容进行正文分类，将属于社会新闻类的网页内容保存，把不属于社会新闻类的新闻URL进行记录，从而实现对社会新闻类网页的准确采集。

训练数据分类训练数据通过对主流的新闻媒体网站的不同频道进行采集，然后将各自的频道名称作为标注数据。网页内容分类采用朴素贝叶斯的方法，首先将网页进行正文抽取（将在下一节进行具体描述），将正文进行分词并统计各个词对应的词频。然后根据各个词在不同类新闻中的数目和各个分类的总体数目进行处理。公式如下：

加入朴素贝叶斯的公式

### 3.2.2 网页正文抽取

由于爬虫抓取的网页文本是网页原始数据，其中包含大量与主题内容无关的信息，如html标签、javascript脚本等，因此在实际使用之前需要对数据做预处理。预处理工作包括两步：信息抽取和建立倒排索引。

网页信息抽取我们使用一种结合网页分块思想和基于密度方法的抽取算法[7]。首先构建原始网页数据的Dom树结构，然后使用转换算法将Dom树转化为更容易处理的BLE&IE 块，其中，BLE(Block Level Element)是指在显示时能够创建块区域或大片文本的元素，通常包含文本、IE元素或其它BLE元素，例如,<p>、<div>、<li>等；IE（Inline Element）是指在HTML中用来定义文本的元素，它们只能包含文本或其它IE标签。而BLE&IE 块是由原网页的DOM结构子树通过特殊操作转换而来，其根节点为BLE元素，根节点的直接子节点的排列符合正则表达式(IE|text)\*p?q\*|l\*|t\*，其中p、q、l和t分别代表特殊的BLE元素。

在将原网页的DOM结构转化为BLE&IE Block后，我们使用基于密度的正文抽取算法处理这些块结构，从而将正文从噪声数据中分离出来。一个块元素的密度是指元素中文本长度（TextLength, Ttl）与标签长度（TagLength, Tgl）的比值。这一正文抽取算法基于这样一个事实：一般网页正文中总是包含大量的文本字符，但需要较少的字符描述它们的标签，而噪音数据正好相反。当计算出所有块结构的密度以后，只有那些密度高于密度阈值的节点才会被视为正文候选。计算节点密度并判断节点是否为候选正文节点的过程如表1所示：

|  |
| --- |
| **Algorithm 2 Pseudo code of get\_ content\_candidates(N)** |
| 1. **Input**: node *N* in BLE&IE block  2. **Output**: {*TtL*, *TgL*}  3. init {*TtL*, *TgL*} with data from N;  4. if *N* is a BLE:  5. for each child *n* in N:  6. {*TtL'*, *TgL'*}= **get\_ content\_candidates** (*n*);  8. increase {*TtL*, *TgL*} by {*TtL'*, *TgL'*};  7. get the density of *n* according to {*TtL'*, *TgL'*};  9. if density of *n* is less than *Threshold:*  10. remark *n* as a noise;  11. return {*TtL*, *TgL*}; |

表1 正文节点判断过程

在分离出网页正文以后，我们对网页正文建立倒排索引，以方便后续模块的分析和处理，因为构建索引不是本文的重点，因此本文中不做详细介绍。

### 3.2.3 新闻事件发现

新闻事件发现模块的功能是从大量的新闻中发现对同一个新闻事件进行报道的新闻，并对这些新闻进行重要程度上的分析，将事件定义为一组具有相同或者相关主题的新闻。通过计算出所有新闻之间的相关性，然后将相关性大于最小相似度阈值的新闻聚合成事件，因此事件发现分为两个步骤，一是文本的特征表示与数据预处理，二是对文本进行聚类。

### 文本特征表示与数据预处理

每一篇新闻正文首先要经过一系列的数据预处理，包括分词、去掉停词、词性标注以及实体识别。我们采用向量空间模型方法来表示每一个新闻，向量的元素中不包含形容词和副词。向量中每一项表示某个词项在对应文档中的权重。词的权重采用经典的TF-IDF来计算，权重计算公式如下：

Weight(t，d)= （1）

其中，dt表示在文本d中词项t出现的频率，|d|表示文本d包含的所有词项的个数。Nt表示整个语料集中包含词项t的文本数，N表示整个语料集中的文本数。

在对文本使用权值特征向量表示以后，我们使用下面的余弦公式计算两个文本之间的相似度：

Similarity(di，dj)= （2）

其中，di和dj表示文本集中的任意两个文本。

1. **新闻聚合**

根据公式（2）可以计算出文本集中任意两个新闻文本的相似度，并生成其对应的文本相似度矩阵，然后利用该相似度矩阵生成一个无向图，图的节点为新闻，边就是两个新闻的相似度。通过对无向图进行切分来发现具有相同主题的新闻得到事件，当有新的新闻节点加入的时候，通过判断该新闻节点与已有事件的质心进行相似度判断，从而决定该新节点是否属于这个事件。具体方法是将文本以及它们的相似度关系映射成为无向图,图中的每个顶点表示一个文本,如果两个顶点之间存在一条边,则表示这两个顶点之间存在关联,边上的权值表示两个顶点之间的相似度,对此我们定义一个相似度阈值参数min\_sim，只有当文本之间的相似度大于该阈值时，才会将两个文本顶点用带有权值的边连接起来，最终存在于同一个无向图中的节点所代表的新闻文本构成一个事件，效果如图2所示：

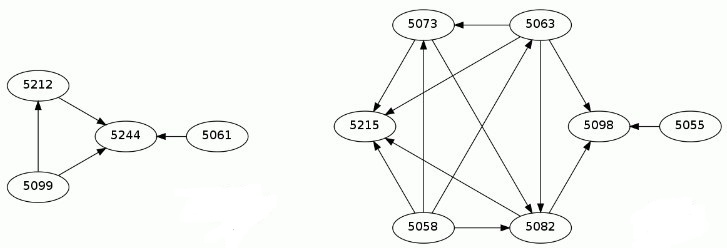


图2 使用无向图发现事件

来一段伪代码

### 3.2.4 事件摘要

事件摘要是指用较少且能表述出该事件重要信息的句子来概括该事件。本系统中我们采用基于质心的方法抽取文摘句，从句子与质心的距离和句子的位置两个方面计算句子的权重。

句子与质心的距离。

所谓质心就是一个事件中的核心词的集合。它代表了主题的中心思想，与质心越近，说明句子包含的重要信息越多，通过使用余弦相似度来表示句子与质心的距离，相似度越大，说明离质心越近。

首先，找出在所有文档中TF-IDF最高的k个词和短语构成文档集的质心Sc=(w1,w2,…,wk)。

然后，将文档集合中的所有句子与质心计算相似度，并按相似度得分进行排序，公式如下：



其中，是句子Si的向量表示，是质心Sc的向量表示。

一般情况下，位于文章开始位置的句子比后面的句子包含更多的信息，因此，应该为开始位置的句子赋予更高的权重。我们简单的使用句子在文档中位置的倒数来表示句子的位置权重，即，



于是，每个句子的权重可由以下公式计算：



**句子选取**

利用上述方法计算句子权重之后，得分最高的前N个句子可以作为文摘候选句。但这些句子之间仍存在冗余，为了选出最具代表性同时冗余信息又尽量较少的内容，还需去除这些冗余。

# 4 实验

1. 主题爬虫的对比试验
2. 事件发现的对比试验

# 5 总结

本文介绍了一种基于社会新闻的事件发现与跟踪系统的设计实现方案，利用主题爬虫的数据采集方法获得社会新闻，采用文本聚类的方法进行事件发现，并结合多文本摘要的方法进行事件的展示。后续的工作将重点对事件发现的算法和文本摘要的算法进行改进和优化，使得其在社会新闻事件发现方面的准确性和摘要的精准度进一步提高。

# 参考文献

[1] D. Trieschnigg, W. kraaij, Hierarchical topic detection in large digital news archives exploring a sample based approach, Journal of Digital Information Management, Vol. 3,2005.

[2] L. Zhen, W. L. da, L. Lei and H. Y. yan, Incremental K-means method

based on initialization of cluster centers and its application in news event detection, Journal of the China Society for Scientific Technical

Information, July 2006.

[3] A. P. Porrata, R. B. Llavori, J. R. Shulcloper, Topic discovery based on text mining techniques, Information Processing and Management, pp.

752-768,2007.

[4]S. Huang, X. P. Peng, Z. D. Niu and K. S. Wang, News Topic Detection based on Hierarchical Clustering and Named Entity, Natural Language Processing and Knowledge Engineering (NLP-KE), 27 Nov, 2011.

[5]S. Yang, X. Q. Cheng, J. Zhang and H. B. Xu, Detect Events on Noisy Textual Datasets, Web Conference(APWEB),2010 12th International Asia-Pacific, 6, April, 2010.

[6] A. X. Sun, M. S. Hu, Query-Guided Event Detection From News and Blog Streams, Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions, 9, 2011.

[7] S. Lin, J. Chen, Z. D. Niu, Combining segmentation-like approach and density-based, Tsinghua Science and Technology, June, 2012.

[8]Papka R, Allan J, On-Line new event detection using single pass clustering[Technical Report,UM-CS-1998-021] 1998

[9]W. Lam, H. M. L. Meng, K. L. Wong and J. C. H. Yen. Using contextual analysis for news event detection, International Journal of Intelligent Systems, April, 2011.

[10]Jia ZY, He Q, Zhang HJ, Li JY and Shi ZZ. A news event detection and tracking algorithm based on dynamic evolution model, Journal of Computer Research and Development, 2004(07)

[11]Zhang K，Li JZ, Wu G, Wang KH. A new event detection model based on term reweighting. Journal of Software, 2008,19(4):817-828.

[12] X. Y. Dai, Q. C. Chen, X. L. Wang and J. Xu, Online topic detection and tracking of financial news based on hierarchical clustering, Proceedings of the Ninth International Conference on Machine Learning and

Cybernetics, pp. 11-14, July 2010.