新闻事件分析系统 IEventMiner 的设计*)

雷 震 吴玲达 王 辰 贺 玲
(国防科技大学信息系统与管理学院 长沙 410073)

摘 要 本文介绍了基于事件的新闻报道信息组织系统 IEventMiner 的设计思想和实现方法,该系统不但可以对新闻报道中的事件进行探测和追踪,还可以进行事件相关故事单元(RSU)的检索与分类。对 IEventMiner 系统的设计思路以及各功能模块进行了详细的说明,并给出了原型系统。与相关研究工作相比,该系统不但具有更好的可扩充性,而且更加稳定。

关键词 新闻报道,事件探测,事件追踪,检索与分类

The Design of News Event Analysing System IEventMiner

LEI Zhen WU Ling-Da WANG Chen HE Ling
(College of Information System and Management, National University of Defense Technology, Changsha 410073)

Abstract The design of an event-based news report information organization system IEventMiner is introduced. The system can not only detect and track the events in news report, but also retrieve and classify event relevant story units (RSU). The design idea and each functional module of the system IEventMiner are described in detail, and the implementation of prototype system is also described. Compared with related work, the system has a better expandability and performs better in stability.

Keywords News report, Event detection, Event tracking, Retrieval and classification

1 引言

在当前这个"信息爆炸"的时代,越来越多的人从网上获得数据和信息。根据 CNNIC 的统计,用户从网上获得的信息中约有 84.38%的信息是新闻,包括 Web 文档新闻、Web 视频新闻和 Web 广播新闻等多种媒体形式。随着互联网的飞速发展,Web 上的数据量和信息量越来越大,信息超载的问题越来越突出,对新闻报道信息组织能力的要求也越来越高。目前多数网站对这些信息仍然采用人工的手段进行整理,不但耗费了大量的人力物力,而且效率也不尽人意。为了提高新闻报道信息组织的效率,改善繁冗的人工组织新闻报道信息的过程,我们研发了基于事件的新闻报道信息组织系统 IEventMiner,其目的是实现新闻报道中基于事件的多源信息组织任务。

IEventMiner 系统主要在事件探测与追踪[1]的基础上扩展考虑了与完整的新闻信息组织任务密不可分的事件 RSU 的检索与分类[2],使得所开发软件系统的功能更为完善,可处理的源媒体数据类型更加多样化,更符合当前对新闻报道信息组织的实际应用情况。该系统核心模块的设计主要来自对新闻报道中多种信息组织任务的内在规律的探究和分析,以及转换为具体的实现方法后的评估和完善,其设计过程不但考虑了系统的实用性和高效性,而且兼顾了系统的灵活性和可扩展性。

IEventMiner 以 Visual C++ 6.0 和 MATLAB 6.5 作为系统的主要开发工具,软件系统的运行环境为 Windows 2000 Professional。系统中充分利用了我们当前在基于事件的新闻报道信息组织领域的最新研究成果,并集成了中科院研制的汉语词法分析系统 ICTCLAS[3](Institute of Computing Technology, Chinese Lexical Analysis System)、Xercesc 的 XML Parser 和我室参与研制的辅助情报分析的新闻视频挖掘系统[4]中的新闻故事单元探测模块。

2 IEventMiner 系统的设计

2.1 IEventMiner 系统的设计思路

IEventMiner 系统的设计既要保证系统性能的可靠性,同时要兼顾用户的实际需求。IEventMiner 系统主要分为事件探测、事件追踪和事件 RSU 检索与分类 3 大部分。其中事件探测和事件追踪主要处理文本数据,包括网络新闻文档以及视频新闻数据转录过来的文档,而事件 RSU 检索与分类在进行事件报道切分时主要处理的是视频数据,进行 RSU 检索与分类时主要处理文本数据,另外在训练时主要以事件相关的网络新闻文档为训练语料。

IEventMiner 系统的设计思路如图 1 所示,主要包括事件探测、事件追踪和事件 RSU 的检索与分类 3 部分。事件探测旨在将输入的新闻报道聚成不同的事件簇,并在适当的时候将首次报道的新闻事件识别出来。事件追踪则是根据由事件探测识别出来的几篇关于某个事件的新闻报道监控后续新闻报道,以发现与该事件相关的报道。事件 RSU 的检索与分类则是利用文本信息从高层语义分析的角度实现视频新闻 RSU 的检索与分类,训练中可以用到的正例样本数量远远多

*)本文得到国家自然科学基金(60473117)和国家"八六三"高技术研究发展计划基金(2001AA115123)资助。雷 震 博士生,研究方向:机器学习、数据挖掘;吴玲达 教授,博导,研究方向:多媒体信息系统;王 辰 讲师,博士,研究方向:机器学习、基于内容的检索;贺 玲 博士生,研究方向:数据挖掘、基于内容的检索。

于事件追踪中已知的训练正例样本数量，并且其训练语料选自与事件相关的网络新闻文档。通过事件探测，用户可以将属于同一事件的新闻报道聚合成簇，并可以发现以前没有报道过的新事件。借助事件追踪，用户能够实现对自己感兴趣事件的新闻报道追踪，从而了解该事件的进展情况。而事件RSU的检索与分类是以含有比镜头更多语义信息的事件RSU为检索和分类单位，通过提取事件相关媒体中的文本信息并利用机器学习方法自动建立事件类的模型，从而提供概念化的RSU检索与分类方式，是在一定程度上解决基于内容的视频检索中低级特征与高级概念之间语义鸿沟的一种有效途径。具体实现时，应根据实际的应用设计出相应的方法，来满足用户实现新闻报道信息组织的不同需求。

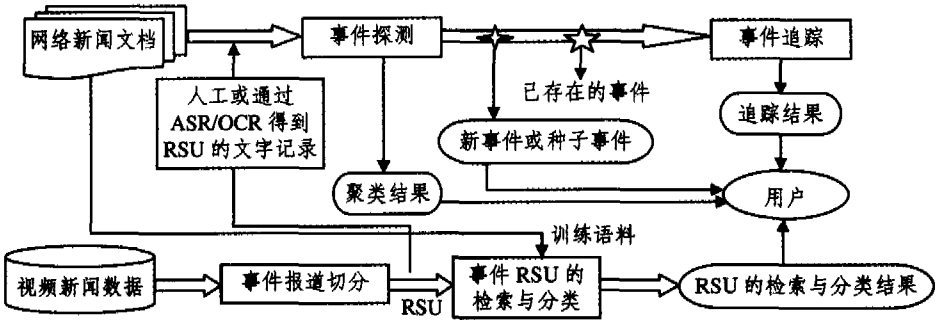

图1 IEventMiner系统的设计思路

2.2 IEventMiner系统的总体结构

基于事件的新闻报道信息组织系统IEventMiner的总体结构如图2所示。

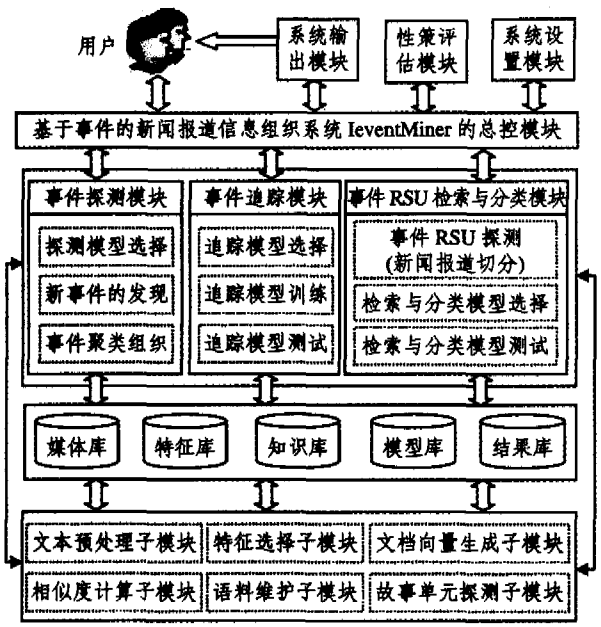

图2 IEventMiner系统的总体结构

该系统包含3个核心模块，即事件探测模块、事件追踪模块和事件RSU检索与分类模块。围绕处理对象和处理模型，所有的相关数据均集中于总体结构图中的5个数据库中。这5个库是媒体库、特征库、知识库、模型库、结果库。其中，媒体库存储原始新闻视频数据和以文本形式表示的新闻语料；特征库存储特征选择之前的原始特征和特征选择后的特征以及用于视频故事单元探测的部分视频低层特征；知识库存储领域相关的数据知识模型来支持基于事件的信息组织进程。最关键的是模型库，模型选择得合适与否直接影响信息组织的性能，它用来存储用于基于事件的不同类型信息组织任务的各种模型相关数据，例如模型的参数设置、选择等。结果库用来存储不同类型信息组织任务的结果以及模型性能评估结果。

3个核心模块代表了IEventMiner系统所能完成的3个不同类型的信息组织任务，其实现离不开图2中底部所示的6个子模块的支持，即文本预处理子模块、特征选择子模块、文档向量生成子模块、相似度计算子模块、语料维护子模块和故事单元探测子模块。下面分别对各功能模块进行介绍。

2.3 IEventMiner系统各功能模块定义

(1)文本预处理子模块

收集某个事件的相关网页之后，要首先对其进行预处理。该模块主要包括两个解析过程：首先是将包含标记符号的半结构化HTML文档解析成只包含有用信息的文本文档。从网上下载的半结构化HTML文档包含了一定的格式化标记，这些标记暗示了它们标示信息的重要性。然后需要对文本进行常规的分词操作，并去除停用词。

(2)特征选择子模块

由于学习算法的复杂度将会随着文档的规模呈指数级增长，因此在生成文档向量之前还需要进行特征选择，以降低特征空间的维数。实际操作时，需要根据不同的问题采用不同形式的特征选择方案，用户可以根据自己的需求以及不同的情况进行挑选。对于事件RSU的检索与分类，常采用的特征选择方法包括特征增强法(TS)、互信息法(MI)、信息增益法(IG)、CHI法等[5]。而对事件追踪以及事件探测，这些特征选择方法并不适用，可结合常用的特征权值计算方法，如"ntc方案"、"ltc方案"等[6]，并结合阈值剔除法来达到降低特征空间维数的目的。尤其是"ntc方案"，是特征表示方法中一个简单且费用较低的公式，其效果和信息增益、CHI相当，并优于其它方法(如互信息法、特征增强法等)。确定一种特征选择方法之后，在文本特征集库中进行分类有效特征的选取，生成特征子集，降低特征空间维数，并为后续任务做好准备。

(3)文档向量生成子模块

文档向量生成过程其实就是在特征子集中计算出每个特征的权值。每个词对该文档都有不同的支持度，这种支持度用权值来表示。向量在每一维上的分量对应该特征在这篇文档中的权值。通常来讲，词对文档的支持度可以通过词频(TF)和反比文档频率(IDF)等来计算。另外，对于半结构化HTML文档，要考虑到文本中不同位置、不同词性和不同长度的特征往往具有不同的重要性，利用标记增大重要部分的权值。

(4)相似度计算子模块

相似度计算是基于事件的新闻报道信息组织的核心问题

之一。在我们的NREMS系统中，尝试了多种相似度计算方法。例如在事件追踪中，我们采用余弦夹角公式来计算文档向量间或文档向量和事件模板向量之间的相似度，而在事件探测中则采用了欧氏距离计算向量间的相似度。

(5)语料维护子模块

语料维护是为基于事件的新闻报道信息组织提供已标记类别的训练文档集，可以采用手工标记的方法为每一个文档设定类别信息、时间信息、来源信息、标题信息等，也可以采用信息抽取技术自动获得每一个文档的相关信息。对于事件探测，类似于无监督的文档聚类，不包含训练集；对于事件追踪，需要包括反例事件和正例事件(追踪事件)两方面的训练文档集合，反例训练集的数量较大；对于事件RSU检索和分类，需要维护的正反例训练样本均比较大，不但包括从新闻网站上下载的语料，还包括从新闻视频提取或转换的文本语料。

(6)故事单元探测子模块

IEventMiner系统集成了我室参与研制的“辅助情报分析的新闻视频挖掘系统”中的“新闻故事单元探测模块”。该模块采用了一种融合多特征的新闻故事单元探测方法，在对新闻故事单元的结构进行分析的基础上，通过静音探测、标题字幕事件探测、镜头探测、口播帧探测来达到融合多种特征探测新闻故事单元的目的。该方法通过对镜头、播音员镜头、标题字幕事件以及静音的综合分析，能够准确地探测出新闻故事的边界，具有较好的适应性。

(7)系统设置模块

负责调整探测模型、追踪模型以及检索和分类模型在进行基于事件的新闻报道信息组织执行过程中的参数、提供实验数据路径、阈值设置、聚类类别、特征空间维数设定、终止条件、特征权值计算方案选择、特征提取策略、事件类别增添等。

(8)性能评估模块

利用查全率、查准率、漏报率、失报率、F1度量和归一化系统代价等评估方法并根据具体情况进行基于事件的新闻报道信息组织结果的评价，以便更好地评估各种算法之间的性能差异。

(9)系统输出模块

将基于事件的新闻报道信息组织结果以及多种性能评估指标输出到用户界面上，提交给用户。某些情况下还要绘出对应的图表。

(10)事件探测模块

事件探测是将输入的新闻报道归入不同的事件簇，并在新闻报道信息流中识别出以前没有报道过的新事件，或者说对一个新事件的首次报道。事件探测是一个非监督的学习过程，又可分为两种形式：回溯探测和在线探测。前者侧重于在一个已经按照时间顺序收集好的新闻报道集中发现以前未曾报道过的事件，后者则关注从新闻报道的信息流中实时地发现新事件的出现。由于我们的目的是使事件探测与识别系统能够自动发现过去发生的新事件，因此本文讨论的事件探测如不特别说明均指回溯探测。在我们的系统中，分别使用了Single-Pass法[6]、普通增量K均值法[6]和改进的增量K均值法进行事件的探测。系统的实验结果表明，我们所提出的改进的增量K均值法的探测性能在三者中是最好的。由于篇幅所限，关于改进的增量K均值事件探测算法将另文介绍。

(11)事件追踪模块

事件追踪旨在监控新闻媒体流，以发现与某一已知事件相关的后续新闻报道。在我们系统中提出了一种基于NEP-SVM的事件追踪算法，该算法首先借鉴主题提取的思想对传统的文档表示方式进行了改进，即通过简单的串匹配技术给能够更好地反映新闻主题的特征项分配更大的权值，然后修剪反例样本，根据距离和类标决定某个反例样本的取舍，然后使用SVM对修剪之后的样本集进行训练，最后通过参数训练将SVM的输出结果映射成概率[7]，从而确定某报道与事件相关与否，同时给出某报道与事件相关的置信度。另外，本系统还提出了一种用于事件追踪的基于K近邻特征线(KNFL)的分类方法，这种基于最近邻特征线(NFL)[8]的方法本质上可以看作是对K近邻(KNN)法[6]的推广，将改进后的KNN融入到NFL中形成KNNFL是为了更适合新闻事件的分析。研究结果表明，本文所提出的方法与传统的方法相比较，可以获得更好的效果。关于KNNFL算法的原理及实验结果，作者将另文阐述。

(12)事件RSU检索与分类模块

由于视频的领域较宽，视频的低级视觉特征和高级概念之间存在着较大的语义鸿沟，常导致检索效果不佳。有别于传统的视频检索方法，NREMS系统以含有比镜头更多语义信息的事件RSU为检索单位，通过提取事件相关媒体中的文本信息并利用机器学习方法自动建立事件类的模型，从而提供概念化的RSU查询方式。系统提出了组合特征选择方法和一种二阶段修剪KNN：TSP-KNN。组合特征选择方法相对于MI方法更适合事件相关故事单元的检索；二阶段修剪KNN先对训练集进行修剪，然后用KNN训练得到分类器。该方法解决了样本混叠以及多中心分布问题。我们的系统实验结果表明所提出的方法是有效的，明显地提高了事件RSU的检索性能。另外，系统提出了使用直推式支持向量机(TSVM)[9]进行事件RSU的分类，如果说RSU的检索是涉及某个具体事件的，RSU的分类则是一种相对宽泛的主题类别的训练与测试。

3 IEventMiner系统的特点

(1)集成化程度高

IEventMiner系统集成了中科院研制的汉语词法分析系统ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System)、Xercesc的XML Parser和我室参与研制的辅助情报分析的新闻视频挖掘系统中的新闻故事单元探测模块。其中ICTCLAS在“973”专家组机器翻译第2阶段的评测和2003年5月SIGHAN举办的第1届汉语分词大赛中，取得了可喜的成绩，是目前最好的汉语词法分析系统之一。ICTCLAS同时提供一套完整的动态链接库ICTCLAS.dll和相应的概率词典，开发者可以完全忽略汉语词法分析，直接在自己的系统中调用ICTCLAS。ICTCLAS可以根据需要输出多个高概率的结果，输出格式也可以定制，开发者在分词和词性标注的基础上继续上层开发。

Xerces-C++是Apache团队的开发成果，它不但严格遵循DOM与SAX规范，而且提供了良好的易用性和跨系统特性，并保证很高的执行效率，一直是国外很多项目的首选XML解析器。

而我室参与研制的辅助情报分析的新闻视频挖掘系统已经顺利通过了国家“863”专家组和“十五”专家组两次评审，均获得了较高的评价。

(2)特征权重的用户可配置性

研究过程中我们发现：不同事件类型的新闻文档采用不

同的特征权重计算方案会取得更好的效果。基于这种考虑，我们开发的 IEventMiner 系统可以让用户自己选择特征权重计算方案，以便更加灵活地配置特征加权体系。

(3)多种新闻媒体源数据处理能力

IEventMiner 系统不但可以处理网络新闻、视频新闻，还可以处理由人工、自动语音识别 ASR 或 OCR 技术从新闻广播、新闻专线和电视等媒体流的音频记录得到的文字记录。

(4)可扩展性强

在 IEventMiner 系统的开发过程中，我们采用了模块化的开发策略，并尽量使各个模块之间的耦合度最小，以便系统具有较强的可扩展性和灵活性。

4 IEventMiner 系统的实现

实验语料主要选自新浪、网易、搜狐、新华网和人民网等 5 家著名新闻网站的新闻报道，各网站已经对各热点事件进行了人工分类整理，因而在此基础上建立的实验语料库更具有说服力。我们建立的这个实验语料库共包含 1536 篇新闻网页，涉及“美军抓获萨达姆”、“驻伊美军枪击意大利女记者”、“美英士兵虐待伊拉克俘虏”等 16 个事件，时间跨度从 2003 年 12 月 25 日到 2005 年 3 月 16 日，实验结果是在 16 个事件涉及语料上实验的平均值。另外，还有部分实验数据是从中央电视台的新闻联播和新闻 30 分采集的视频数据，系统需要对视频数据进行相应的故事单元分割操作，并提取出对应于每个故事单元的文本信息以便用于事件 RSU 检索与分类。

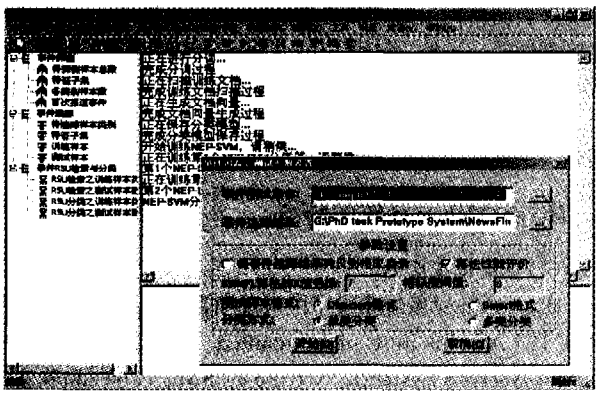

图 3 IEventMiner 系统的主界面

图 3 所示为 IEventMiner 系统的主界面。界面左面各条目显示的是三类基于事件的新闻报道信息组织任务。当前界面右上角的视图显示的是使用基于 NEP-SVM 的方法进行事件追踪的训练过程。由图可以看出，该过程主要包括分词、扫描训练文档、生成文档向量、保存分类模型、分别训练反例事件 NEP-SVM 分类器和训练追踪事件 NEP-SVM 分类器等几个过程。同时给出了进行事件追踪时的测试参数设置对话框，用户可以根据需要选择相应的事件测试集合目录和事件追踪结果目录，以及测试样本的格式和分类方式。

结论 随着互联网技术的迅速发展，自动高效地对新闻报道信息进行组织，尤其是从事件的角度解决此类问题已经成为一个极具潜力的研究方向，国内虽然刚处于起步阶段，但在国外已经开展得如火如荼。本文给出了一个基于事件的新闻报道信息组织系统 IEventMiner 的设计与实现，并从系统的结构设计、各功能模块定义、系统特点和系统实现等几个角度对 IEventMiner 做了介绍。该领域还有许多问题需要探讨，新闻报道信息组织系统的模型还有待进一步完善。在以后的工作中，我们还要添加更多的新闻报道信息组织模块，以丰富该系统的功能，同时将网页净化和 Web 数据挖掘技术有机紧密地应用到我们的新闻报道信息组织系统中，以提高系统的智能化程度，也是我们今后需要着重解决的问题。

参 考 文 献

1 Allan J, Carbonell J, Doddington G, et al. Topic Detection and Tracking Pilot Study: [Final Report]. In: Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, Morgan Kaufmann Publishers, Inc, 1998. 194～218

2 Lei Z, Wu L D, Lao S Y. A Method for Content-based News Story Classification in Data Mining. In: Proc. of the 11th ISPE International Conference on Concurrent Engineering, 2004. 265～270

3 刘群，张华平，俞鸿魁，等. 基于层叠隐马模型的汉语词法分析. 计算机研究与发展，2004，41(8)：1421～1428

4 谢毓香. 辅助情报分析的新闻视频挖掘技术研究：[博士学位论文]. 长沙：国防科学技术大学，2004

5 Yang Y M, Jan O P. A Comparative Study on Feature Selection in Text Categorization. In: Proc. of the 14th International Conference on Machine Learning, 1997. 412～420

6 Papka R. On-line New Event Detection, Clustering, and Tracking: [PhD Thesis]. University of Massachusetts at Amherst, 1999

7 Platt J. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: Advances in Large Margin Classifiers, MIT Press, 1999. 61～74

8 Li S Z. Content-based Classification and Retrieval of Audio Using the Nearest Feature Line Method. IEEE Trans on Speech and Audio Processing, 2000, 8(5): 619～625

9 Joachims T. Transductive Inference for Text Classification Using Support Vector Machines. In: Proc. of the 16th International Conference on Machine Learning, 1999. 200～209