基于网上特定话题的多侧面跟踪技术及应用研究*

王兰成，娄国哲，严骏，张思龙
（南京政治学院上海校区军事信息管理系，上海 200433）

摘要：扩展对新媒体信息资源的获取和研究是数字图书馆系统进一步提升服务方式和质量的重要课题，而话题跟踪是其中一项关键技术。针对专门话题跟踪过程中的话题漂移及跟踪结果的杂乱无序问题，提出网络文本的客观中心和主观中心概念，给出话题主观中心的表示和更新方法，建立多中心模型按话题内容进行组织分析；提出先分类后聚类，依据客观中心跟踪话题和主观中心组织话题的方法对策，对传统KNN算法进行改进，把基于时间、地点、机构和关键词要素的KNN话题跟踪算法和基于内容要素的文本聚类过程进行合并，给出军事类话题多侧面跟踪组织的技术解决方案和实例。

关键词：网络文本；话题多侧面；话题漂移；跟踪技术
中图分类号：G203
DOI：10.3772/j.issn.1673-2286.2016.12.011

1 引言

当前，数字图书馆系统扩展对新媒体信息资源的采集整合，可进一步改进信息服务方式，提升信息服务质量。如当数字图书馆系统的用户浏览网络小说时，希望查找到相关论坛中网友的评论；又如，当数字图书馆系统的用户在检索经典文献时，希望读到最新跟踪报道。但是目前数字图书馆系统建设中较少提供这些功能，基于网络特定话题的跟踪是扩展数字图书馆系统上述功能的一项重要技术。

对于专门网络文本而言，敏感内容常引起热门话题，而与该话题相关的报道通常集中于该话题出现的最近一段时间。随着时间推移，直接的相关报道频率逐渐衰减至消失，同时网络媒体对该话题进行报道的侧重点会发生变化而产生话题漂移现象。例如，关于2010年4月中国海军公海训练内容的话题变化：第一阶段，主要描述训练的客观情况，参加训练的舰艇构成、训练时间的跨度、训练地点、训练内容；第二阶段，敏感国家/机构对中国海军训练进行评论，主要围绕训练目的、训练意图、军力变化等展开；第三阶段，部分国家/机构采取何种相应的政策和措施应对；第四阶段，话题消失。这种从内容出现，到对内容评论，再到各国政府的应对，话题发生漂移。本文将每个阶段的讨论内容称为话题侧面，一个完整的话题展示是由话题的多个侧面构成的[1]。在该军事话题的动态变化中，将网络中的动态变化划分为四个阶段，分别为萌芽、形成扩散与聚焦、爆发、消亡。显然，话题的漂移与网络内容的动态变化有着一一映射的关系，因此可通过对话题漂移的跟踪来实现对网络内容动态的全面把握。

话题跟踪是数字图书馆系统进一步对新媒体资源获取和分析的一项重要技术，其功能是识别出给定网络内容的后续报道。面对话题跟踪中的话题漂移问题及跟踪到的报道杂乱无序问题，一种解决方案是引入话题客观中心和动态话题主观中心的结构来描述

* 本研究得到国家社会科学基金2015年课题“信息化条件下档案社会化媒体信息资源的整合路径与机制研究”（编号：15BTQ078）和南京政治学院2015年军队专项科研课题资助。

话题。漂移过程中客观中心相对恒定，而主观中心不断变化，重点对传统K最近邻（K-Nearest Neighbor，KNN）分类算法进行改进[2]，提出根据话题客观中心对话题进行跟踪，根据主观中心对追踪到的话题进行聚类，形成话题的多个侧面，整体给出话题中心漂移问题及话题组织问题的解决方案，不仅能实现对话题的准确跟踪，还能实现对话题的细粒度把握，即对话题不同的侧面实现跟踪。从系统功能的角度而言，话题跟踪是跟踪已知的经典文献资料热点的技术手段。

2 特定话题的模型及其漂移分析

扩展检索新媒体资源需要全面跟踪有关网络信息，并关注其后续发展及相关内容。显然在不同阶段文本内容的侧重点各有不同，但通过研究发现，每篇报道都有对主题内容本质特征最基本的叙述，一般本质特征通常是对专题内容的客观描述，由时间、地点、机构等主体构成；而内容的评论、政策措施等，均因内容而引起相应主体的情感、行为反应，形成主观态度。客观描述决定网络讨论话题，而主观态度决定网络讨论的内容。本文给出如下话题的客观特征和主观特征概念。

（1）话题的客观特征。话题演变过程中，在话题各侧面均出现，用来标识该话题基本语义信息的特征称为客观特征，由{D，P，E，K}要素构成。其中，D为时间要素，P为地点要素，E为机构要素，K为关键词要素。

（2）话题的主观特征。在一个话题中，除去共有的客体特征，用来标识各侧面相对独特、新颖的特征称为主观特征，由{C}要素构成。其中，C为内容要素。

通过对话题漂移现象分析，可看到虽然对特定内容的报道讨论是动态变化的，但这些报道讨论不能脱离具体内容而存在，没有内容就谈不上话题，话题和内容不可分割。报道讨论中均包含对特定内容本质特征最基本的描述，这些构成话题的客观中心，在话题漂移过程中，客观中心基本保持恒定，因此可依据客观中心进行话题跟踪。话题的客观中心由时间要素、地点要素、机构要素和关键词要素复合构成。话题的主观中心是对话题某个侧面讨论内容的概括和总结，是最能体现该侧面语义特征项的集合。通过动态话题主观中心描述话题多个侧面，通过话题内容的多中心结构展示话题的各侧面，可展现话题漂移的过程。话题的主观中心由内容要素构成。

话题建立时，即形成话题的客观和主观中心，有些话题在初始阶段就逐渐消亡，只形成一个初始中心，没有演化出新中心。而大多数话题都会发生演变，话题主观中心随着内容的发展而不断变化，由初始中心开始，不断生成新中心，每个中心代表话题讨论不同的侧重点。以中俄两国举行“海上联合”军事演习为例，初始阶段主要围绕演习的背景情况；随着两国军事机构对演习内容的深入分析，讨论的核心集中于演习对区域形势的影响与中俄两国军事合作能力分析；接下来，报道侧重于中俄两国的政府和军队对该热点话题的应对（如政策措施、军力调整等）。在上述案例中，话题重点讨论了三个中心，即训练概况、形式分析和政策应对。每个中心代表本次内容的一个话题侧面，它们都是组成这次内容的重要方面。通过采用一个客观中心和三个主观中心的话题模型可全面表示此次训练内容。

赵华等提出一种面向动态演化的双质心话题模型[3]，话题由初始质心和当前质心表示，以分界点为界，当新的分界点出现时，已有的初始质心和当前质心合并成新的初始质心，促使分界点建立的文档形成新的当前质心。本文通过客观中心和动态话题主观多中心结构建立话题模型，主观中心的个数没有限定，完全根据话题演变的情况动态调整，话题演变出几个内容侧面，模型就建立几个中心，在结构上同话题漂移的结果保持一致，如图1所示。

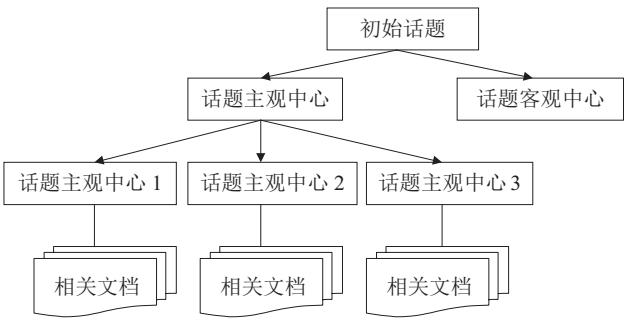

图 1 话题多中心模型

该话题模型最大的特点是用话题主客观中心展示话题变化过程中的变与不变，一方面体现客观中心的恒定，另一方面还原话题多侧面的逻辑结构。重要的是，它能并行维护和更新话题的每个主观中心。“并行”指不管后续文档到来的时序，本模型能按照文本内容对它们进行分析，确定其属于哪个主观中心，或是用该文档建立新的主观中心，并非所有模型都具有“并行”特性。建立多中心模型是解决话题漂移问题的有效方法，

因其不受话题产生时间影响，而完全按话题的内容进行组织分析。

3 多侧面话题的跟踪方法与信息组织

3.1 改进的KNN话题跟踪方法

话题多侧面跟踪算法的中心思想即采取先分类再聚类的方法，对新采集的网络文本进行分析[4]。首先，采用基于话题客观中心要素{D，P，E，K}的分类算法对文本进行分类，判断是否是已知话题的后续报道，如果满足相似度条件，则把文本归类于已知话题；然后，对跟踪到的网络文本进行聚类计算，主要采用话题主观中心要素{C}进行聚类，忽略其他要素影响，依据相同侧面的报道其内容要素相似度高，不同侧面的报道其内容要素相似度低的原则，形成话题的多个主观中心。

基于文本分类的话题跟踪技术主要包括：特征提取、向量表示、分类方法研究以及话题跟踪器的构建四部分。在传统的KNN算法中，因其是对整个文本的向量空间模型（Vector Space Mode，VSM）进行计算，当发生话题漂移时，跟踪话题的能力明显下降。同时，考虑训练正例稀疏的问题和采用要素的网络文本表示方法，传统的KNN算法不完全适用。以下给出改进的基于内容客观中心的KNN跟踪方法。

（1）跟踪策略的改进。首先，对话题跟踪加入时间窗策略。在跟踪过程中，只需考虑话题“运动周期”内所有网络报道，这样有效降低计算的复杂度，提高话题跟踪的效率，也避免将过去发生的类似报道判断为当前关注的话题；其次，仅考虑{P，E，K}要素的功能作用，分别对三个要素进行KNN分类，判断出与初始话题{P，E，K}要素的类别关系；最后，结合多要素复合比较策略判断出后续网络文本的类别关系。

（2）相似度计算方法的改进。采用要素的话题模型决定了要使用基于相关要素的相似性比较算法[5]。依次以{P，E，K}要素为计算对象，分别选出在训练集中与后续{P，E，K}要素最相似的k个文本。在这k个文本中分别抽取属于正例$r(s_1, k)$的所有文本，将这些文本与新文本对应要素的相似度值求和，作为新文本与正例$r(s_1, k)$对应要素的相似度值；同样，抽取属于反例$n(s_1, k)$的所有文本，将反例文本与新文本对应要素的相似度值求和，作为新文本与反例$n(s_1, k)$对应要素的相似度值；然后计算判定要素是否是属于追踪话题。

改进的KNN话题跟踪方法如下：

①采用命名实体识别技术对训练文集中的网络专门文本的要素进行提取，用要素表示网络文本；

②在新文本s_1到达后，判断该文本的时间是否在话题“运动周期”的时间窗内，如果在则执行下一步，否则将该文本判定为“不属于”，处理下一个新文本；

③根据命名实体、特征词对新文本进行分词处理，确定新文本的要素；

④分别计算新文本的{P，E，K}与训练集文本{P，E，K}的相似度值，对于每个不同的要素分别选出与要素对应的最相似k个文本；

⑤计算新文本{P，E，K}要素与正例、反例的相似度值：当 $p_{x\in(P,E,K)}(s_1, T)\geq 0$ ，判定新文本对应要素属于该话题；当 $p_{x\in(P,E,K)}(s_1, T)<0$ ，判定新文本对应要素不属于该话题；

⑥根据复合比较策略判断新文本是否属于追踪话题。

3.2 话题多侧面组织算法

对跟踪到的同一话题下的内容报道，按内容的不同侧面中心进行组织，这是话题跟踪的基础性工作，也属于无指导的文本聚类。与依据{D，P，E，K}要素聚类不同，对于该部分需要聚类的文本而言，是已知话题跟踪到的结果，主要根据内容要素C进行聚类，也即对本文定义的主观特征进行聚类组织。

对于动态增长的跟踪到的网络文本，采用Single-pass聚类算法进行聚类。依据话题多中心模型和内容要素相似度计算方法，首先，明确话题多中心话题模型结构，准确把握已有话题多侧面不同的主观中心；其次，通过要素向量的思想获取新闻文档的内容向量、判断文档所属主观中心；最后，结合Single-pass增量聚类算法给出基于话题主观中心的话题多侧面组织算法。以下给出算法的整体流程：

①若当前文档是跟踪到网络文本中的第一个，则建立一个以该文档为初始中心的类，作为第一个类簇，然后继续处理下一篇文档；

②计算新文本内容要素与已有各类簇的相似度（采用余弦相似度计算），记录最大相似度及其对应话题侧面中心；

③若最大的相似度小于创新阈值，则建立一个新类簇，同时该文档为新类簇的初始中心，继续处理下一篇文档，转向①；

④若最大的相似度大于创新阈值，则文档归入最大相似度对应的类簇；

⑤更新中心向量，每当有新文档加入话题中心时更新相应的中心向量，继续处理下一篇文档，转向①。

4 应用及实例分析

某数字图书馆系统的用户在检索有关“反导系统”的经典文献资料时，希望查阅到网络最新的美国在韩国部署“萨德”反导系统跟踪报道。系统采用确定的测试数据与评测标准进行话题跟踪实验[6]，以验证本文提出跟踪方法的性能。在跟踪试验中，从数据集挑选“反导系统”话题进行跟踪测试，选取该话题最初始的450个内容报道构建初始话题，同时从测试集中除去这些内容报道，然后开始测试。先采用TRS文本检索系统对设定的话题进行跟踪，再应用本文提出的基于话题客观中心跟踪的方法对设定话题进行跟踪。其中，对于采用要素表示方法的网络文本，主要根据表示话题的客观中心要素进行跟踪，分要素进行KNN计算，判断各要素与初始要素的相似程度，当各要素的相似度计算完成后应用复合比较策略判断网络文本的所属类别。最后对两次试验跟踪的结果进行性能比较。

表1是采用两种跟踪方法的实验结果，图2是经过计算的两种方法的准确率、召回率、F1-measure的直观对比图。在跟踪试验中通过对图2的分析和比较，采用本文基于客观中心的改进KNN算法后，分类的准确率、召回率和F1值都有一定的提升，验证了本文跟踪算法的实效性。

表 1 两种跟踪方法的测试数据

反导系统	发现结果	返回结果	相关总文档数	准确率	召回率	F1-measure
TRS系统	2 310	3 206	3 922	0.721	0.589	0.648
本文方法	2 608	3 517	3 922	0.742	0.665	0.701

注：F1-measure是根据准确率（P）和召回率（R）给出的一个综合评价指标，具体定义为 $F1=\frac{2RP}{(R+P)}$。

面对网上日益增多的特定内容报道，数字图书馆系统如何从这些网络文本中更有效并准确地获取信息，

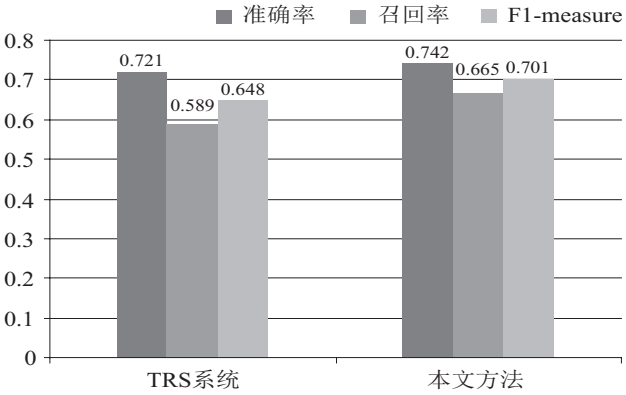

图 2 两种跟踪方法的评测对比

帮助用户关联信息将成为研究热点，而话题跟踪技术是其中一项关键技术。本文针对专门话题跟踪过程中的话题漂移及跟踪结果的杂乱无序提出网络文本的客观中心和主观中心概念，给出话题主观中心的表示和更新方法，建立多中心模型以按话题的内容进行组织分析；又提出先分类后聚类，依据客观中心跟踪话题和主观中心组织话题的方法对策，对传统的KNN算法进行改进，把基于时间、地点、机构和关键词要素的KNN话题跟踪算法和基于内容要素的文本聚类过程进行合并，给出军事类话题多侧面跟踪组织的技术解决方案和实例。针对专门话题内容的检测与跟踪工作，下一步还有很多工作，如要素特征的选择与抽取在很大程度上决定了网络文本分类和聚类的效果，本文的特征选择主要使用数学统计的方法，缺少对专门话题语义层面的处理，必然存在抽取质量和抽取精度的问题，影响检测和跟踪的准确率和召回率。通过对网络特定话题的多侧面跟踪技术及应用研究，是扩展对新媒体信息资源的获取和研究的技术保障，也是数字图书馆系统进一步提升服务方式和质量的重要课题。

参考文献

[1] 邱立坤,龙志祎,钟华,等.层次化话题发现与跟踪方法及系统实现[J].广西师范大学学报(自然科学版),2007(2):157-160.

[2] 张宁,贾自艳,史忠植.使用KNN算法的文本分类[J].计算机工程,2005,31(8):171-172.

[3] 赵华,赵铁军,于浩,等.面向动态演化的话题检测研究[J].高技术通信,2006,16(12):1230-1235.

[4] 郑希文.互联网话题演变与传播分析技术研究[D].哈尔滨:哈尔滨工程大学,2009.

[5] 王巍,杨武,齐海凤.基于多中心模型的网络热点话题发现算法[J].南京理工大学学报(自然科学版),2009(8):422-426.

[6] 洪宇,张宇,刘挺,等.话题检测与跟踪的评测及研究综述[J].中文信息学报,2007,21(6):71.

作者简介

王兰成，男，1962年生，南京政治学院上海校区军事信息管理系教授，博士生导师，研究方向：图书情报计算机管理、舆情传播与网络分析，E-mail：wanglancheng@163.com。
娄国哲，男，1982年生，博士研究生，讲师，研究方向：军事情报管理，E-mail：louguozhe@163.com。
严骏，男，1987年生，博士研究生，讲师，研究方向：军事情报管理，E-mail：chinash_yan@163.com。
张思龙，男，1987年生，博士研究生，讲师，研究方向：军事情报管理，E-mail：ahang35@126.com。

Research on Drift Analysis and Tracking Technique of Topic on Public Military Opinion

WANG LanCheng, LUO GuoZhe, YAN Jun, ZHANG SiLong
(Department of Military Information Management, Shanghai Branch of Nanjing Political College, Shanghai 200433, China)

Abstract: Extending the acquisition and research of new media information resources is an important subject of digital library system. It is to further improve the service mode and quality. The topic tracking technology is one of the key technologies. The objective and subjective text center network center concept are put forward to solve the problem of topic drift and special topic tracking results in the tracking process. We have given the topic representation and subjective center update method, and establish a multi center model according to the topic of organizational analysis. The first classification after clustering, countermeasures according to the method of topic tracking and objective center the subjective topic center organization are put forward. We improve the traditional algorithm of KNN with time, location, and topic keywords elements. The military side tracking technology solutions have been studied.

Keywords: Web Text; Topic Multi Side; Topic Drift; Tracking Technology

（收稿日期：2016-11-10）

《数字图书馆论坛》2017年征稿启事

《数字图书馆论坛》创刊于2005年，是由科学技术部主管、中国科学技术信息研究所主办的专业性学术月刊，是“中国科技核心期刊”统计源刊，是中国社会科学引文索引（CSSCI）扩展版来源期刊。

《数字图书馆论坛》是我国唯一一本以“数字图书馆”命名的刊物，一直关注国内外数字图书馆领域的相关研究和实践，报道主题涵盖信息检索、数字资源、知识组织、语义技术、数据挖掘、开放获取、用户服务等，既关注数字图书馆领域的理论成果和创新实践，又反映国内外业界的新趋势、新发展和新变革。

《数字图书馆论坛》既厚名家、更重新人，约稿和自由投稿相结合，对专家组稿以及刊用的博士生论文免收版面费。来稿将在1个月内给予录用答复，稿件一经发表即向作者寄送样刊并付稿酬。

投稿及开放网址：http://www.DLF.net.cn。