知识组织与知识管理

面向社会文本流数据探测爆发主题方法浅析*

乐小虬[1]　洪　娜[2]

[1](中国科学院国家科学图书馆　北京 100190)

[2](中国医学科学院医学信息研究所　北京 100020)

【摘要】社会文本流数据富含上下文环境信息、语言不规范且参与用户数量庞大。针对这类数据开展爆发主题探测需要寻找新的思路。本文对社会文本流数据的概念、特点以及爆发主题表达形式进行系统性梳理,从文本内容、时间、社会三个维度阐述探测爆发主题的主要研究思路和基本流程,分析利用社会特征(如用户参与、上下文环境、社团结构)进行爆发主题探测的主要技术方法。

【关键词】社会文本流　爆发主题探测　社会网络

【分类号】TP393

A Survey of Burst Topic Detection Towards Social Text Stream Data

Le Xiaoqiu[1]　Hong Na[2]

[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

[2](Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China)

【Abstract】Social text streams have rich contextual information and huge participants who communicate with informal steams. It needs to find suitable solutions to detect burst topics from this kind of data. In this paper, the authors comb through the concepts, the characteristics of social text stream data and the presentation forms of burst topics. It also summarizes the main research ideas and the basic procedures of burst topic detection towards social text stream data in three dimensions: textual content, social, and temporal. The principal approaches to make use of social features, such as user participation, social context and community structure evolution, for burst topic detection are generally discussed.

【Keywords】Social text stream　Burst topic detection　Social network

1　引　言

随着社交网络应用的发展,用户成为网络内容的主要创建者,社会媒介活动以及用户间的交互变得异常活跃。在线参与到社交网络或论坛中订阅微博服务或维护博客数量巨大,由此形成的社会网络不仅反映了网络社区中人与人之间的社会关系,也蕴含丰富的社会活动信息[1]。分析和监测一段时期内社会文本流中主题的变化情况,探测和鉴别突发增加的主题是爆发主题探测的重要研究内容,它涉及话题探测与跟踪(Topic Detection and

收稿日期:2012-09-25
收修改稿日期:2012-10-12
*本文系国家社会科学基金项目"网络科技信息中爆发主题的监测与分析方法研究"(项目编号:09BTQ035)的研究成果之一。

Tracking, TDT)、舆情监测、趋势探测等多种应用。利用 Twitter 进行爆发词探测并实现趋势探测的研究框架如图 1 所示:

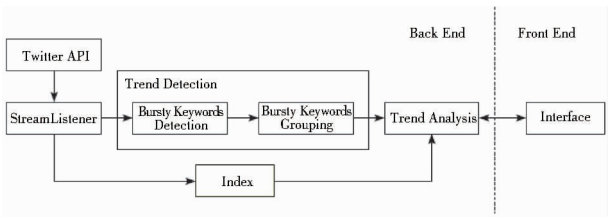

图 1 Twitter 趋势探测系统框架[2]

在文档流的爆发探测研究中,研究对象主要集中在文本内容和时间两个维度上。对于社会文本流数据,其应用形式和社会上下文环境均发生了很大变化,如果在分析时嵌入时间、信息流模式(交流模式)以及社会网络关系,不仅能在较好的粒度上更为准确地探测爆发主题,还能沿着内容、时间和社会维度进一步探究主题间的关系[3]。

本文将对已有研究中有关社会文本流数据的概念、特点以及主题的表达形式做进一步阐述,并对社会文本流数据中爆发主题探测的主要研究思路、基本处理流程以及如何有效利用社会特征的技术方法进行分析。

2 社会文本流数据

2.1 社会文本流数据概念

社会文本流数据(Social Text Stream Data)是指随时间而聚集的文本交流数据,文本流中每个文本片段同某些社会属性(如作者、评论者、发送者、接受者)相关联[3]。典型数据有博客、BBS、邮件列表、评论站点、网络论坛以及社交网站等。对于邮件数据集,交流即是邮件消息本身,而在博客数据集中,交流则是博客之间的评论。同文档流相比,社会文本流的输入输出单元为文本段,强调用户间的交互,而文档流是指按时间顺序到达的一系列文档,输入输出单元为文档,如新闻网页、科技文献,因而粒度上要比社会文本流大。在研究中,二者的界限较模糊,如果不考察社会信息,社会文本流有时也被看作文档流,如 Kleinberg[4] 在研究中就将邮件作为文档流,Fujiki 等[5] 也将 BBS 实体作为文档流。

2.2 社会文本流数据特点

与一般的文本流数据相比,社会文本流数据具有以下几个特点:

(1)数据粒度小,表现为文本描述短、词汇重叠。例如评论、QnA 社团、Tweeter 通常都是短句、书签、标签甚至仅仅是关键词[6,7]。

(2)内容具有丰富的上下文环境特征。文本块在信息发送者 - 接收者或作者 - 评论者之间富含社会关系,每个文本块包含时间属性,文本块内容对上下文环境敏感,词义不仅依赖于文本块内容内部所处的环境,而且还依赖于社会参与者(如发送者、作者、接收者、评论者)和其他相关文本块(如先前文本块的时间和内容信息)[3]。

(3)语言不规范,非常随意和口语化,数据干扰多,帖子质量参差不齐,广告垃圾和欺骗信息常常混在其中[8]。

(4)参与用户数量庞大。

3 社会文本流数据中主题的表达

徐戈等[9]认为在自然语言处理中,主题(Topic)可以看成是词项的概率分布。通常主题是隐含变量,因此又称隐形主题(Latent Topic)或者隐形语义(Latent Semantic)等[9]。在标准的主题模型如 LDA(Latent Dirichlet Allocation)中,对于文档通常做这样的假设:一篇文档含有多个主题,文档可用主题分布进行表达,文中每个词有一个隐含的主题标签(Label)[10]。对于文档流而言,鉴别爆发主题的方法可以先发现单个爆发词的爆发模式,然后发现在相同文档中共现且有相同爆发期的词的分组,用分组中排列靠前的词系列表示主题,通过人工标记等手段形成显形主题。

表 1 TimeUserLDA 模型排列的前两个爆发主题[10]

爆发时期	靠前词(Top Words)	Tweets 实例	主题标记(Label)
Nov 29	vote, big, awards, bang, mama, win, 2ne1, award, won	(1) why didnt 2ne1 win this time! (2) 2ne1. you deserved that urgh! (3) watching mama. whoohoo	Mnet Asian Music Awards (MAMA)
Oct 5 – Oct 8	steve, jobs, apple, iphone, rip, world, changed, 4s, siri	(1) breaking: apple says steve jobs has passed away! (2) google founders: steve jobs was an inspiration! (3) apple 4 life thankyousteve	Steve Jobs death

然而,对于社会文本流数据,由于帖子内容短,一条帖子几乎就是一个单独主题。主题的表达通常与每个帖子的隐含变量相联系,一个短的词系列可分配单

独主题,例如,在 Diao 等[10]的实验结果中(见表 1),两个不同的爆发时期对应了两个排列靠前的词系列,其中的词是由 Tweet 帖子中的文本内容计算而来,其隐含主题由人工标记而来。在具体研究中,社会文本流数据中主题的选取与研究的目的和数据的特点有关,例如,在 Delicious 网站中,每个页面对应若干个标签(Tag),Ramage 等[11]将一个标签作为一个主题。Chen 等[6]在研究从网络论坛中抽取爆发主题时,将术语(Term)作为主题。

4　主要研究思路

利用社会文本流数据探测爆发主题的研究伴随互联网应用的发展而不断变化(见表 2)。在社会文本流数据出现前,爆发主题的研究主要针对两类数据:静态文档集和文档流。对于静态文档集,主要从文本内容的角度考察其中的词或术语,度量方法主要是计算词的频率或概率,典型模型有向量空间模型(Vector Space Model,VSM)、隐性语义索引(Latent Semantic Indexing, LSI)和 LDA,这类研究的目的主要用于信息检索或主题抽取。对于文档流数据,研究大多集中在文本内容和时间两个维度上,经典模型为 Kleinberg[4]提出的一种爆发词探测的两种状态自动机模型,通过在文档流中限定一个较小的时间间隔来观测包含某个词的文档的到达频率,识别超出平均到达频率的时间段及词的爆发强度。He 等[12]将爆发主题表示为一组爆发词的集合,提出了通过主题文档数和爆发时长计算主题爆发强度的方法。

表 2　不同类型数据源爆发主题探测研究思路比较

数据源类别	典型数据	考察维度	特征来源	度量	典型模型
静态文档集	数据库文献	文本内容	词/术语	频率/概率	VSM、LSI、LDA
文档流	网络新闻	文本内容 时间	词/术语 + 时间	变化率	Kleinberg、TOT
社会文本流	微博、论坛、标签	文本内容 时间 社会信息	词/术语 + 时间 + 社会信息(用户活动\上下文环境\社团结构等)	变化率	Kleinberg 改进模型、混合模型(如 TimeUserLDA)

对于社会文本流数据,研究者开始更多地关注其中的社会信息,如用户参与信息(Participations)、社会上下文环境信息(Social Context)和社团结构信息等。爆发主题探测倾向于综合利用文本内容、时间、社会三个维度的信息进行整体考察。在特征选取上,更多的是考虑多个维度间相互作用表现出的特征。在爆发度量方面与文档流差别不大,都是着眼于计算变化率。建模方法常采用 Kleinberg 改进模型或混合模型(如 TimeUserLDA)。

5　基本处理流程

综合已有文献,社会文本流爆发主题探测的处理流程大致分为数据获取、信息提取、特征选取与计算、爆发主题建模和结果评估 5 个阶段,如图 2 所示:

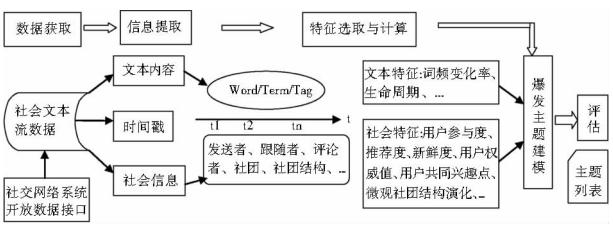

图 2　社会文本流中爆发主题探测处理流程

目前,大多数社交网络系统(如新浪微博、Tweet、Facebook、Delicious 等)都对外提供开放数据接口,既有历史数据集也有在线数据。数据经预处理后,从中抽取文本内容、时间戳和社会信息,经分词、过滤处理后得到包含时间信息的词(或术语/标签)和不同类型参与者(或社团)的信息,然后以这些信息为基础选取并计算爆发特征,如词频变化率、用户参与度、推荐度、新鲜度、用户权威值、社团结构演化等,然后利用这些特征构建爆发主题模型,并对模型运行结果(爆发主题列表)进行评估,调整模型参数(如阈值、权重等)以达到应用要求。

6　主要研究方法

社会文本流数据中爆发主题探测有许多研究方法,最为直接的就是探测社会网络文本块中关键词的爆发行为,如果一个关键词在文本流中出现不寻常的高比率即被视为爆发。例如,Mathioudakis 等[2]把爆发词作为趋势探测的入口点,只要关键词表现出爆发行为,就将它视为已出现新主题的迹象,并进一步对其追踪探测。但这种方法没有有效利用社会信息,社会文本流数据的爆发主题探测多从文本内容、时间、社会三个维度进行,由于目前在文本内容和时间两个维度上已有许多研究和论述,本文不再赘述,仅分析利用社会信息探测爆发主题的主要方法。

6.1 引入用户参与/活动信息

与文本内容相比,用户参与信息仅涉及用户 ID、发帖时间和发帖频率,噪音相对较小,且易于处理,因而成为爆发主题探测中的重要考察特征。对于用户参与信息的利用方法,Zhu 等[13]在对 BBS 做话题探测和跟踪研究中提出了 UF – ITUF (User Frequency – Inverse Thread User Frequency) 模型,该模型同计算内容相似性的 TF – IDF 模型相似,主要用来度量同时参与两个主题(Threads)的用户组的相似性。此外,Wu 等[14]在研究论坛中的话题检测问题时提出用户参与模型。

在利用用户参与信息进行爆发主题探测方面,主要是从文本内容和用户参与信息相结合的角度上开展研究。Chen 等[6]在研究从网络论坛中抽取爆发主题时提出一种利用术语、时间和用户参与信息提取爆发主题的噪音过滤模型。其基本思路为采用频率分段来度量术语和用户特征。术语用频率分段(Frequency Segments)的概念进行刻画和分级,术语随时间顺序发生,如果术语在频率分段中没有爆发,则被视为噪音。然后用一个分段中术语或用户出现频率权重总和(Sumseg)和出现频率权重方差 (Devseg)两个参数刻画术语和用户,并基于 Rank 值过滤数据集中的噪音。用户的参与特征用轨迹刻画,并定义一种角色权重策略照顾发布标题的用户。用户轨迹定义为 yu = [yu (1), yu (2),…, yu (T)],其中 yu (t)是用户在时间 t 内的发布频率。在爆发特征提取时,利用频率分段构造 sum – dev 向量,采用 score(fi) = α × log(Sumseg) + Devseg 计算特征的分值。然后分别选定不同的阈值排列术语特征和选择核心用户。在度量术语和用户的特征相似性时,主要从特征相关性和提交交叠两方面进行计算。

通过分析大规模用户查询日志发现爆发模式是利用用户活动信息探测爆发主题的另一种途径。Parikh 等[15]研究了一种能够克服数据获取方式差异的爆发探测模型,并用此模型对每天的查询进行不间断的增量爆发探测。该模型将查询看作一个用于探测爆发和系统内流动的其他查询的查询混合体,主要考察同一查询的量变百分率,而非查询量的绝对变化。这种增量模型考察分批次到达的新查询,批次到达的时间间隔可以为每天、每小时或任意合理的时间。对于任意一个想要探测是否为爆发的查询 Q,只要构建基于该查询 Q 的碎片(Fraction)模型并将其与系统中的所有查询相比较即可。

此外,Liu 等[16]通过构建用户浏览网页的路径图以及计算用户在每一个网页上的浏览时间,引入用户反馈信息来对网页的重要性排序。李东方等[17]将 Web 2.0 中用户信息活动看作发生在互联网上的热量活动,用户的信息活动被表示为热量的传递,将互联网作为承载热量活动的系统,进行网络主题抽取。这些分析方法虽未直接用于爆发探测,但其思路为如何利用用户活动信息进行爆发主题探测提供了借鉴。

6.2 利用社会上下文环境信息

利用社会环境信息探测爆发主题主要是利用文本块的时间属性和社会属性,考察社会网络环境中数据对象的生命周期或新鲜度、用户权威值、用户共同兴趣点等特征,综合多种特征构建爆发主题探测模型。

(1)生命周期

将生命周期/新鲜度列入社会信息特征主要是基于以下常识:发帖的时间越新,越易引起人们的注意,引发的转发、跟贴、评论的用户回应就会更多,因而更容易形成爆发。

Yao 等[7]在研究大众分类(Folksonomies)爆发探测时考察了帖子的生命周期,方法是计算含时间戳的帖子覆盖(Time – aware Post Coverage)。对于一个标签来说,若具体的时间间隔为 xi,在时间 t 内总共有 n 条被其他用户做过标签的帖子,则用 ni(t)度量在 xi 间隔中时间 t 内的帖子覆盖,标签帖依新鲜度被赋予权重。Cataldi 等[18]用老化理论对术语的生命周期建模,如果术语经常在具体的时间间隔发生并且相对来说在过去很少出现,则被视为新兴术语。

(2)用户权威值或用户吸引力

这种特征主要用于度量用户的影响力或重要度。其基本思想与大规模超文本系统中网页权威性假设类似:在社会媒介中一个用户跟随另一个用户类似于 Web 中一个网页链接到另一个网页,二者都是一种推荐形式,跟随关系获得声誉而后又给出声誉。用户的追随者越多,他/她就越权威;他/她的追随者越权威,他/她就越权威。用户权威值的计算与 Google 的网页权威计算相类似,借助 Page Rank 算法分析其中的社会关系而获得[7]。有的研究则是选择 HITS 算法,基于社

团中的用户网络提取用户权威值[18]。

(3)用户共同兴趣点

Adamic 等[19]在同质性(Homophily)研究中表明,具有相似兴趣的人更有可能联系在一起。如果到访者有相似兴趣或者有相同兴趣的到访者的朋友紧紧连接在一起,那么到访者间的连接会对到访者的兴趣产生更强影响[20]。个体的行为通常在时间上是连续的,呈现出很强的个性化模式,用户对于某种事件的兴趣不会在短时间内发生戏剧性的改变,某一时刻出现的兴趣爆发模式在下一个时间点很可能依然保持很高的水平。当一个到访者对一事件高度感兴趣,他/她所写内容跟此事件更相关[21]。所以选取用户共同兴趣点作为社会特征也是爆发主题探测的重要途径。Lin 等[21]将发帖者的兴趣状态作为网络连接图中的吉布斯随机域(Gibbs Random Field)进行建模,通过设计一些潜在函数使个体的兴趣值更接近过去的状态和邻居的"赞同"意见,其权重框架由现实网络中的观察结果决定。

(4)利用多种特征探测爆发主题方法

利用多种社会环境特征进行爆发探测的方法依研究目标不同而存在差异。Yao 等[7]的做法是:首先选取单个特征,以 Kleinberg 的两种状态自动机模型为基础进行独立的单流探测,然后采用基于学习的混合模型集成多种爆发特征源,所用方法是 Rankboost,它能将许多仅适度精确的"弱"规则结合起来形成预想规则。对于一个单独的排列者(Ranker),用函数 fi 将实例 xi 映射到 R,这些给定的映射被称为排列特征(Ranking Features),一个具体标签的所有时间间隔就形成了一个实例空间 X,而 R 则被视为排列空间。实验中作者将标签频率、发布覆盖和用户吸引力这三种特征作为"弱"排列规则,并从这些初步的爆发探测结果中得到了更高质量的混合结果。

6.3　利用社团结构演化信息

社交网络中存在大量的社团(Community),社团成员间存在许多交互信息,当有趣的话题出现、凸显、消退时这些反应的顺序会在短暂的重大活动爆发期间发生。社交网络中社会网络空间易在连续时间下形成演化视图。分析微观社团结构及其连通性,研究社团演化过程中微观社团随时间变化形成过程以及活跃社团内正在进行的活动,以便发现社团内部链接产生的爆发周期密度或爆发社团,是利用社会网络信息探测爆发主题的重要途径。

这种思路将研究重点放在时间图构建和爆发社团追踪上。Kumar 等[22]在研究博客爆发演化时,首先定义了由博客形成的时间图,然后在此图上采用两步法完成爆发社团跟踪:社团提取,从博客图中提取密集子图,并将这些子图作为所有潜在社团(不管是否爆发),采用剪除扩充算法,剪除用于鉴定社团种子,扩充用于将种子培育成形成社团签名的稠密子图;爆发分析,以 Kleinberg 研究事件流中爆发所用方法为基础,通过无限状态自动机中的状态转换反映爆发的出现,对社团提取中获取的每个子图进行爆发分析以便从那些社团中鉴别和排列爆发。不同之处在于,Kleinberg 处理的基本事件相对简单,主要是定位特征(例如,邮件中是否包含给定的关键词),而 Kumar 的框架没有提供这种定位点,爆发社团并不是以单个博客或时间图边中术语进行刻画,而是从分析整个博客图结构后体现出来。

另一种方法是研究文本主题与网络结构间的相互作用。网络用户间的交流和文本内容均会随时间发生变化,从而导致网络结构和文本集同时演化。Lin 等[21]的实验证实在社会网络中探测或跟踪爆发主题及事件演化时,如果不考虑网络结构会显现两大缺陷:

(1)不能充分反应突然变化。如果在当前时间点没有某个特定用户的文本信息,用户的新状态被设为以前的状态,而不能借用用户邻居信息进行更为准确的新状态评估。

(2)反应局部噪音能力更加脆弱。一个拥有 10 个用户跟随的 Tweet 比没有跟随的 Tweet 更有影响力,如果同等对待这两种情形,则"孤立"Tweet 的影响力就被不恰当地放大。

所以研究者在做社会网络中流行事件跟踪研究时,综合考察了用户兴趣爆发、网络结构信息扩散以及文本话题演变,利用吉布斯随机域对图的历史状态影响力和依赖关系进行建模,合并所有的历史特征、文本特征、结构特征推理出联合分布。此外,研究者证明了经典 Kleinberg 爆发探测模型可以作为该模型的特例,因而对于综合利用网络结构信息探测爆发主题具有一定的参考价值。

7　结　语

社会文本流数据与文档流数据相比,其文本粒度

相对较小,输入输出单元为文本段,更强调用户间的交互。与一般的文本流数据相比,社会文本流数据富含上下文环境信息、语言不规范且参与用户数量庞大。社会文本流数据中主题的表示方式不尽相同,既有用词系列表示、隐含主题由人工标记的方式,也有直接将术语、标签作为主题的方式。当前,面向社会文本流数据探测爆发主题的思路倾向于综合利用文本内容、时间、社会三个维度的信息进行整体考察。在特征选取上,更多的是考虑多个维度间相互作用表现出的特征。从社会维度上看,研究者在利用用户参与信息、社会上下文环境信息以及社团结构信息方面开展了许多有益的探索。从研究者的实验结果来看,利用社会文本流数据中的社会特征、综合其他多种特征进行爆发探测的实验效果比选取单一特征的探测效果要好。

参考文献:

[1] Matsumura N, Goldberg D E, Llora X. Mining Directed Social Network from MessageBoard[OL]. [2012 - 09 - 20]. http://delivery.acm.org/10.1145/1070000/1062884/p1092 - matsumura.pdf? ip = 159.226.100.225&acc = ACTIVE%20SERVICE&CFID = 118941016&CFTOKEN = 80199382&__acm__ = 1348453917_c173fc996f56a0c611739ba100e23712.

[2] Mathioudakis M, Koudas N. TwitterMonitor: Trend Detection over the Twitter Stream [OL]. [2012 - 03 - 11]. http://delivery.acm.org/10.1145/1810000/1807306/p1155 - mathioudakis.pdf? ip = 159.226.100.225&CFID = 35057597&CFTOKEN = 95005305&__acm__ = 1310711607_fd4454ee954f38c1a4c767c8b5047820.

[3] Zhao Q, Mitra P, Chen B. Temporal and Information Flow Based Event Detection from Social Text Streams[C]. In: *Proceedings of the* 22*nd AAAI Conference on Artificial Intelligence*. 2007:1501 - 1506.

[4] Kleinberg J. Bursty and Hierarchical Structure in Streams[J]. *Data Mining and Knowledge Discovery*, 2003,7(4): 373 - 397.

[5] Fujiki T, Nanno T, Suzuki Y, et al. Identification of Bursts in a Document Stream[OL]. [2012 - 10 - 12]. http://citeseerx.ist.psu.edu/viewdoc/download? doi = 10.1.1.106.6773.

[6] Chen Y, Yang S, Cheng X Q. Bursty Topics Extraction for Web Forums[C/OL]. In: *Proceedings of the* 11*th International Workshop on Web Information and Data Management*. 2009:55 - 58. [2012 - 03 - 09]. http://portal.acm.org/citation.cfm? id = 1651587.1651600&coll = DL&dl = ACM&CFID = 19983188&CFTOKEN = 89593705.

[7] Yao J, Cui B, Huang Y, et al. Temporal and Social Context Based Burst Detection from Folksonomies[C]. In: *Proceedings of the* 24*th AAAI Conference on Artificial Intelligence*. 2010:1474 - 1479.

[8] 陈友,程学旗,杨森. 面向网络论坛的突发话题发现[J]. *中文信息学报*,2010,24(3):29 - 36. (Chen You, Cheng Xueqi, Yang Sen. Outburst Topic Detection for Web Forums[J]. *Journal of Chinese Information Processing*, 2010, 24(3): 29 - 36.)

[9] 徐戈,王厚峰. 自然语言处理中主题模型的发展[J]. *计算机学报*, 2011,34(8): 1423 - 1436. (Xu Ge, Wang Houfeng. The Development of Topic Models in Natural Language Processing [J]. *Chinese Journal of Computer*, 2011, 34(8): 1423 - 1436.)

[10] Diao Q, Jiang J, Zhu F, et al. Finding Bursty Topics from Microblogs[OL]. [2012 - 10 - 02]. http://www.mysmu.edu/faculty/jingjiang/papers/ACL'12.pdf.

[11] Ramage D, Hall D, Nallapati R, et al. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi Labeled Corpora[C]. In: *Proceedings of the* 2009 *Conference on Empirical Methods in Natural Language Processing*, Singapore. 2009:248 - 256.

[12] He Q, Chang K, Lim E P. Analyzing Feature Trajectories for Event Detection[C]. In: *Proceedings of the* 30*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2007: 207 - 214.

[13] Zhu M, Hu W, Wu O. Topic Detection and Tracking for Threaded Discussion Communities[C]. In: *Proceedings of the* 2008 *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. 2008:77 - 83.

[14] Wu Z L, Li C H. Topic Detection in Online Discussion Using Non - Negative Matrix Factorization[C]. In: *Proceedings of the* 2007 *IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops*. 2007:272 - 275.

[15] Parikh N, Sundaresan N. Scalable and Near Real - Time Burst Detection from ECommerce Queries[OL]. [2012 - 04 - 16]. http://portal.acm.org/citation.cfm? id = 1401890.1402006&coll = DL&dl = ACM &CFID = 19983188&CFTOKEN = 89593705.

[16] Liu Y T, Gao B, Liu T Y. BrowseRank: Letting Web Users Vote for Page Importance[C]. In: *Proceedings of the* 31*st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Singapore. 2008:451 - 458.

[17] 李东方,俞能海,尹华罡. 一种 Web 2.0 环境下互联网热点挖掘算法[J]. *电子与信息学报*, 2010, 32(5):1142 - 1145. (Li Dongfang, Yu Nenghai, Yin Huagang. Mining Hot Topics on Internet Under Web 2.0[J]. *Journal of Electronics & Information Technology*, 2010,32(5):1142 - 1145.)

[18] Cataldi M, Caro L D, Schifanella C. Emerging Topic Detection on Twitter Based on Temporal and Social Terms Evaluation[OL]. [2012 - 05 - 15]. http://delivery.acm.org/10.1145/1820000/

1814249/a4 - cataldi. pdf? ip = 159. 226. 100. 225&acc = ACTIVE%20SERVICE&CFID = 118941016&CFTOKEN = 80199382&_acm_ = 1348454353_fc699f8fab306f6ff8ed2cd2b208f191.

[19] Adamic L A, Adar E. Friends and Neighbors on the Web[J]. *Social Networks*, 2003, 25(3):211 - 230.

[20] Backstrom L, Huttenlocher D, Kleinberg J, et al. Group Formation in Large Social Networks: Membership, Growth, and Evolution [C]. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2006: 44 - 54.

[21] Lin C X, Zhao B, Mei Q, et al. PET: A Statistical Model for Popular Event Tracking in Social Communities[C/OL]. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, DC, USA. 2010: 929 - 938. [2012 - 05 - 15]. http://delivery. acm. org/10. 1145/1840000/1835922/p929 - lin. pdf? ip = 159. 226. 100. 225&CFID = 35057597&CFTOKEN = 95005305&__acm__ = 1310719076_1a0cf689597ec51c79e4b05c7c614370.

[22] Kumar R, Novak J, Raghavan P, et al. On the Bursty Evolution of Blogspace[C/OL]. In: *Proceedings of the 12th International Conference on World Wide Web*, Budapest, Hungary. 2003:568 - 576. [2012 - 05 - 15]. http://delivery. acm. org/10. 1145/780000/775233/p568 - kumar. pdf? ip = 159. 226. 100. 225&acc = ACTIVE%20SERVICE&CFID = 118941016&CFTOKEN = 80199382&__acm__ = 1348454514_530db4146e72606621c2c46ea5822704.

（作者 E - mail:lexq@ mail. las. ac. cn）

OCLC 将链接的数据添加至 WorldCat. org

OCLC 将 Schema. org 描述性标记数据附加到了 WorldCat. org 的页面上，这是 OCLC 向为 WorldCat 添加链接数据所迈出的第一步。目前，WorldCat. org 提供了网络上最大的链接书目数据集。将 Schema. org 标记数据添加到 WorldCat. org 中的所有书、期刊和其他书目资源中后，WorldCat 整个公共可用版本现在可以由在搜索索引和其他应用程序中使用该元数据的智能网络搜索器（如 Google 和 Bing）使用。

依靠基于 Web 服务的商业开发人员一直在探索利用链接数据潜能的方法。标记可以帮助搜索引擎和其他网络搜索器更加直接地利用支持许多在线服务的基础数据，2011 年由 Google、Bing 和 Yahoo！及后来加入的 Yandex 成立的 Schema. org 计划为之提供核心词汇。

OCLC 正与 Schema. org 社区合作以开发并将词汇扩展集添加至 WorldCat 数据。Schema. org 和图书馆特定扩展将为图书馆社区和消费者网络提供宝贵的双向桥梁。Schema. org 也在与许多其他产业合作，从而为其他特定使用案例提供相似的扩展集。

链接数据提供给全球图书馆社区的机会符合 OCLC 与图书馆协作创建 Webscale 的核心策略。将链接数据添加至 WorldCat 记录使这些记录更为有用，尤其是对超出图书馆社区范围的更广阔的网络上的搜索引擎、开发人员和服务。因此，搜索引擎将非图书馆组织链接到图书馆数据就更为容易。

OCLC 技术专员 Richard Wallis 说："Schema. org 引入了重要的新标准。"将图书馆信息与网络上广泛出版的丰富的数据资源兼容会将图书馆确立为链接数据领域主要的中心。这一功能展示了 OCLC 成员图书馆的丰富的书目和规范数据。

WorldCat 是由成千个成员图书馆在过去的 40 年创建的，是世界上最大的图书馆馆藏在线机构。OCLC 将继续代表成员图书馆与图书馆社区和更为庞大的开发人员社区合作，共同研究和探讨链接数据相关项目。

Zepheira 是一家专业服务公司，它将网络提升为平台以管理信息并在链接数据策略方面帮助 OCLC，该公司总裁 Eric Miller 指出："图书馆生成、维持并改进大量超出传统图书馆范畴的高质量数据。WorldCat 数据作为一种交换台与其他数据驱动资源之间互操作，可以更好地将学生、学者和商人连接到图书馆资源。"

OCLC 将 Schema. org 视作及时而且显著的改进，向为图书馆提供实惠的链接数据技术产品迈进了一步。OCLC 软件设计师 Jeff Young 认为："多年以来，OCLC 研究部门在将语义结构融入到网络中一直都是重要的参与者。Scheme. org 提供了一个便于使用的搜索引擎词汇表来说明复杂的数据环境。它使得各种社区很方便地加入网络上使用相同结构的权威资源，如 Dewey、VIAF 和 FAST 标题。"

为进一步展示作为提供链接图书馆数据的角色，OCLC 最近宣布 DDC 23 的全集（内有 23 000 多个可指定的数字和英文标号）现在作为链接数据进行使用。

（编译自：http://www. oclc. org/news/releases/2012/201238. htm）

（本刊讯）