ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.4, April 2008, pp.912−924
DOI: 10.3724/SP.J.1001.2008.00912

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

Blog研究*

杨宇航+, 赵铁军, 于 浩, 郑德权

(哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

Research on Blog

YANG Yu-Hang+, ZHAO Tie-Jun, YU Hao, ZHENG De-Quan

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: Phn: +86-451-86416225, E-mail: yhyang@mtlab.hit.edu.cn, http://mitlab.hit.edu.cn

Yang YH, Zhao TJ, Yu H, Zheng DQ. Research on blog. *Journal of Software*, 2008,19(4):912−924. http://www.jos.org.cn/1000-9825/19/912.htm

Abstract: Popularity of bloggers and the amount of information in the blogosphere increase fast. Blogs have constituted a dynamic and tightly social network by using frequent links and information interaction, and become an important source of information for the real world. Most researches on blog mainly concentrate on blog definition and identification, content mining, community discovery, importance analysis, blog search and spam blog identification. Methods and technologies of link analysis and natural language processing are used in most works, and some blog-specific methods are proposed. This paper analyzes and compares these researches on blogosphere. Problems of current topics are discussed, and finally future directions are proposed in this paper.

Key words: blog; content mining; community discovery; importance analysis; blog search; spam blog identification

摘 要: Blog信息源和信息量迅速增长,并已通过频繁的链接和信息交互在互联网上构建了一个动态且紧密的社会网络,成为现实世界一个重要的信息来源.目前,Blog 领域的研究主要集中在 Blog 的定义与识别、内容挖掘、社区发现、重要性分析、Blog 搜索和作弊 Blog 识别等几个方面.大部分研究采用或借鉴了链接分析、自然语言处理等方面的技术和方法,也提出了一些针对 Blog 领域的特定方法.分析和比较了 Blog 领域的相关研究,并且讨论了研究中存在的问题,展望了未来的研究方向.

关键词: Blog;内容挖掘;社区发现;重要性分析;Blog 搜索;作弊 Blog 识别

中图法分类号: TP393 **文献标识码**: A

Blog是Web Log的简称,在国内普遍被译为博客.Blog是一种作者与读者通过互联网以日志风格进行交互的中介,是一种崭新的信息传播和交互方式.与门户网站等传统Web信息相比,Blog有着更多的信息源,可以提供更丰富的信息,且信息源间的交互更加频繁,联系更加紧密.Blog的发展异常迅速,根据中文搜索引擎百度(http://www.baidu.com)的统计,截止到 2006 年 11 月,在中文互联网领域,Blog站点达到 5 230 万,Blog作者达到 1

* Supported by the Nantional Natural Science of China under Grant Nos.60435020, 60302021 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2006AA01Z150 (国家高技术研究发展计划(863))

Received 2006-10-09; Accepted 2007-08-24

987 万.根据Blog搜索网站Technorati(http://www.technorati.com)的统计,从 2004 年 10 月开始,Blog网站开始加速增长,截至到 2006 年 8 月 10 日,其监测的Blog网站已达到 5 000 万个,是 3 年前的 100 倍.Blog的读者也日益增多,由comScore[1]发起的用户在线行为研究表明,2005 年第一季度,有 5 000 万美国人访问过Blog网站,约占美国互联网用户的 30%,占美国人口的 1/6.可见,无论在国内还是在世界范围内,Blog都变得日益流行和普及.作为一种新兴的媒体,其影响力也日益扩大,在报道的即时性与多样性方面甚至已超过传统媒体,对现实世界正产生着巨大的影响.

不同领域的研究者也开始更多地关注Blog,并从不同角度开展了有关Blog的研究.媒体理论家着眼于Blog如何挑战传统媒体,社会学家则关注不同Blog社区的动态变化和Blog作者的行为动机.在计算机领域,研究者运用统计方法,借助基于内容和链接分析的技术挖掘Blog内容,抽取公众观点,发掘Blog社区特点并对Blog重要性进行分析[2].国际上著名的文本检索会议TREC以及WWW会议等也开始关注Blog领域的研究,提供相关的数据资源,并提出了针对Blog的评测任务.尽管针对Blog领域开展的研究越来越多,然而相关研究还主要处在探索阶段,研究方案和技术手段都比较分散,缺乏统一的实验平台和资源,还没有形成十分明确的研究趋势.本文将分析和比较这个新兴研究领域的最新研究进展,讨论相关研究中存在的问题,并展望未来的研究方向.

Blog 相关研究可划分为 Blog 定义与识别、内容挖掘、社区发现、重要性分析、Blog 搜索和作弊 Blog 识别这 6 个主要方面,本文的组织结构也围绕这几个方面展开.第 1 节介绍 Blog 定义和识别的研究.第 2 节介绍 Blog 内容挖掘的技术.第 3 节介绍 Blog 社区发现的现有方法.第 4 节介绍 Blog 重要性分析的主流技术.第 5 节介绍 Blog 搜索的研究现状.第 6 节介绍作弊 Blog 识别的相关研究.第 7 节简要概括目前存在的问题和未来研究方向.

1 Blog 定义和识别

Blog研究的前提条件是对研究对象即Blog本身的定义,然而到目前为止,关于Blog并没有一个公认而明确的定义,一些研究者和媒体工作者给出了相关的描述性定义.根据“Glossary of Internet Terms”的定义,一个Blog可以看作互联网上可获取的杂志,更新一个Blog的活动就是Blogging,维持一个Blog的人就是Blogger,即Blog作者.eGlossary把Blog定义为一个容易更新的个人网站,通常,每天都会有更新来表达自己的观点.Glance等人[3]把Blog定义为由包含日期信息且倒序排列的条目构成的网页,这个网页由Blog作者通过Blog发布工具进行维护和更新.Blog是一个任何人都容易更新和使用的网站,通过它,用户可以把自己的观点公诸于众,Blog也可以看作是反映大众观点的信息仓库[4].图 1 显示了典型的Blog站点结构及其链接关系.

由图 1 所示,Blog 站点可形式化地定义为:

(1) Blog 站点=(站点 URL,RSS,Blog 作者,站点名,Blog 条目(或文章))

(2) Blog 条目=(永久链接,Blog 作者,时间,标题,描述,评论)

(3) 评论=(Blog 作者,时间,评论内容)

Blog 信息是网络信息的一种,但它又有其自身的特点,因此,要研究 Blog,首先必须通过自动识别将其从其他网络信息中区分出来.对 Blog 的识别通常是根据 Blog 的结构特点和独有的内容信息进行判别.Tomoyuk 等人[5]认为,Blog可被理解为同一作者发布的网页信息,由一系列有日期信息并按照日期排序的文章构成.基于此,他们提出了基于对日期表示和HTML文档分析的Blog识别方法,具有一定的代表意义.根据对Blog的理解和描述,方法将包含符合一定特征的文章条目的页面作为候选Blog,而这些条目所符合的特征包括:每个条目需要在头部包含一个日期表示.这些日期表示应该有着一致的格式,并按照升序或降序排列;所有条目的日期序列是唯一的.除了Blog信息,BBS网页等也可能具有类似的特征.为了排除干扰信息,该方法将满足如下特征之一的网页判别为非Blog网页:页面中包含的典型关键词,例如‘bbs’、‘聊天’、‘回复’、‘re’等(含有这样关键词的很可能是聊天或BBS信息);含有未来日期表示的条目;两个相邻的条目间时间间隔过长(Blog和日志应该更新得更加频繁);条目中不含有动词、形容词等谓词(Blog条目应该包括关于一个事件的描述).

这是比较典型的 Blog 识别的方法,即首先收集网页头部有日期表示、规范且顺序排列的网页,随后根据关

键字(如‘bbs’等)和其他一些特征对网页进行过滤,排除 BBS 和聊天等类似 Blog 页面的干扰,以得到 Blog 网页.然而,这种朴素的方法并不能保证较高的召回率.例如,有的 Blog 页面在讨论 BBS 相关的事件、有的 Blog 更新速度较慢等.一种可预见的改进方法是对 Blog 的特征信息根据其重要性赋以一定正的权值.类似地,对 Blog 的干扰信息赋以负的权值.再把发现的所有特征信息进行线性加权,运用学习方法通过训练得到各种特征信息的系数和最优的阈值,通过权值和阈值的比较识别 Blog 信息.此外,在充分挖掘特征信息的基础上,支持向量机等学习系统也可望在 Blog 识别任务中取得良好的效果.

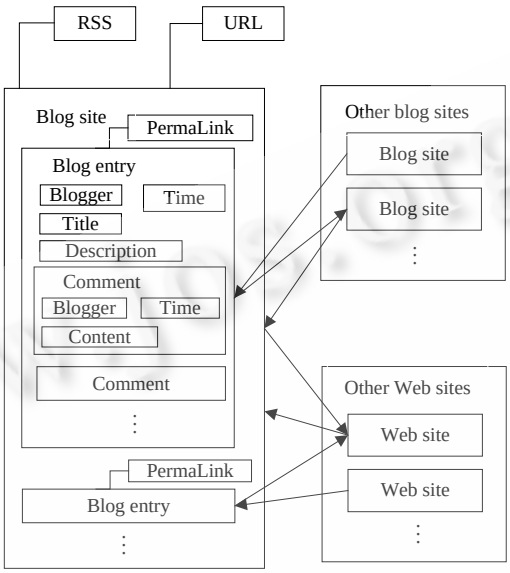


Fig.1 Typical blog site

图 1 典型 Blog 站点

2 Blog 内容挖掘

Blog 内容挖掘是指利用自然语言处理和数据挖掘等技术,从 Blog 社区自动发现和抽取信息的过程.采用的方法主要包括基于词频统计和基于相似度计算的内容分析方法,处理的对象主要包括 Blog 文章、评论和标签信息等.

Blog社区中经常讨论一些热门话题,对这些话题的发现有助于发掘Blog社区的特点,并从中获取相关信息.目前,这类研究主要是通过统计词频,提取热门关键词表以反映一段时间的流行主题.Tomoyuki等人[5]借鉴了用于鉴别文本流中词的爆发(bursty phenomena)的文本挖掘算法[6],运用简单的模式匹配算法以及手工构建的词表,通过词频统计发现爆发现象,进而获取反映热门话题的关键词.Mizuki等人[7]通过抽取高频关联词项,利用词频的动态变化构建阶段时间内的词频向量表示和描述话题.Glance等人[3]从Blog数据集中找到信息量大的短语,计算短语出现的频率,从而发现Blog的趋势,并根据短语在同一Blog条目中的同现概率,通过相似度计算,对信息量大且出现频率高的短语进行聚类以实现主题挖掘.

Tomohiro等人[8]也运用关键词匹配和词频统计等技术,根据阶段时间的统计信息,结合现实事件和Blog信息进行分析,将Blog社区被关注事件的变化模式分为周期模式、逐渐增长型模式、敏感模式、蔓延模式及其他几种模式.以上几种模式能够较好地反映Blog所关注事件的变化过程,但对模式的定义还只是描述性的模糊定义而不是量化的精确定义,这给判别带来了困难.

在一些研究[9]中,用Blog工具编写和发布的网页被定义为Blog,其他类似的页面被定义为网络日志.运用词频统计的内容分析方法以及RSS(really simple syndication,聚合内容)自动发现技术,通过对自动提取的热门关

键词的分类和比较分析可以得出Blog和网络日记的区别.通过区分构成不同的数据集,运用爆发鉴别算法自动提取不同数据集中的热门主题词,根据主题分类发现,Blog和网络日志关心的主题类别有比较明显的区别,即在内容方面有所不同.除了在Blog和网络日志之间存在区别以外,不同的Blog之间也存在区别.Mike[10]通过对链接和高频词的统计揭示了关注外部事件,对新闻敏感的Blog站点和关注自身事件,类似日记的Blog站点之间的区别.Adar等人[11]通过计算Blog网页对间的链接和文本的相似程度发现,相互间有链接的Blog对获得的相似度分值明显高于未链接的Blog对间的相似度,说明有链接关系的Blog更倾向于关注和讨论相同或相似的话题.

Blog与现实世界相互影响,密不可分.研究[3]表明,Blog文章中某产品的词频变化趋势与Amazon的销售额有着相关性.天气和假期对Blog用户的心情有显著的影响,一些固定的、季节性的、周期性的变化也能在Blog文章中有所体现[12].此外,还有研究通过抽取地名和邮编来挖掘Blog作者的地理信息,刻画了其地理分布情况[13],并用GIS(geographic information system,地理信息系统)的形式呈现出来[14].

除Blog文章以外,标签和评论中也包含了大量的信息.Folksonomy是人们随意选择关键词的人工分类方法,用户可以根据自己的理解给每篇Blog文章加上一个标签(tag),反映了很多人的观点.尽管它比其他人为的分类体系更加灵活,但仍然不能处理网络上的所有Blog文章.为了解决这样的问题,有研究[15]尝试将Folksonomy的过程自动化,即为Blog文章自动选择标签,为用户浏览相关信息带来了方便.

评论是Blog用户之间的潜在链接,评论还可作为Blog重要性和文章受关注程度的度量.Trevino[16]和Gumbrecht[17]研究了评论对“Blog体验”的重要性,得到了相似的结论,即大多数Blog用户认为Blog评论对Blog自然交互是至关重要的.抽取评论丰富了Blog作者和读者间的社会网络,对于研究Blog间的交互模式有所帮助[18,19].此外,Blog评论也是搜索引擎中作弊链接的主要来源之一[20].因此,对Blog评论信息的挖掘和分析也十分重要.然而到目前为止,相关的研究相对还较少.

3 Blog 社区发现

Blog 已经通过频繁的链接和信息交互在互联网上建立了一个迅速发展的社会网络,传统的排序方法不足以描述体现社会关系的 Blog 社区.因此,一些研究者开展了 Blog 社区发现方面的工作,并对 Blog 社区中的相关信息与传统媒体以及现实世界间的相互影响进行了分析.

3.1 传统的Web社区发现技术

传统的Web社区发现技术按不同的实现途径可分为基于HITS(hyperlink-induced topic search)算法的技术、基于有向二分图的技术和基于网络流量的技术[21].

HITS算法[22]用于识别hub和authority网页,一个hub网页链向很多authority网页,一个authority网页被很多hub网页指向.在单纯基于HITS算法的社区发现模型中,一个社区被看作是由hub和authority构成的双向图,对不同的主题产生根集合后计算网页的分值,最后用主特征向量和非主特征向量分别表示主要的和次要的社区[23].IBM Almaden实验室开发了用于资源半自动编辑的ARC(automatic resource compilation)系统[24]和用于互联网搜索的CLEVER系统[25].两个系统也都是以HITS算法为核心,并通过增加对网页内容信息的利用,在一定程度上克服了HITS算法的主题漂移问题.

基于二分有向图的Trawling算法与主题无关,该算法利用二分有向图的技术从一个大的数据集里发现社区[26].基于流量的技术则把Web上的社区定义为一组站点的集合,指向社区内的链接大大多于指向社区外的链接,模型中的图分割基于最大流量和最小切分[27].以上算法都是仅对单个社区进行识别,社区图表算法不仅可以用于社区发现,还提出了识别多个社区及社区间的多种关系的方法[28].

3.2 Blog社区发现技术

Blog社区有一些基本特点,即社区内的链接密度远大于社区之间的链接密度,两个社区通常有着不同的信息源,关注不同的话题.最早有关Blog社区的研究包括不同社区的比较[8]、社区演化[18]以及信息的传播模式[29]等.Blog社区发现与传统的Web社区发现技术基本相同,即主要将社区发现作为一个图的问题来研究,Blog和不

同Blog之间的关系分别由图中的节点和边来表示.然而,研究者对于Blog和Blog之间的关系有着不同的看法和认识,因而从不同的角度采用和提出了不同的社区发现方法,有代表性的方法包括基于Blog重要性、基于网络流量和基于Blog间相互感知的社区发现方法.

3.2.1 基于 Blog 重要性的社区发现

Belle等人[30]认为,探究Blog领域的关键性挑战在于理解Blog社区间的不同以及识别重要的代表,将Blog排序以及它们之间的社会关系结合起来,提出了一个有利于理解Blog社区的框架.Belle等人着眼于在社区中扮演重要角色并且得到大多数社区成员信任的Blog作者间的社会网络而不是整个Blog网络,其目的是基于同一主题,挖掘重要的Blog社区.研究基于这样的经验和直观印象,即能够从重要的社区中提取有用的和可信的讨论.例如,如果两个知名的Blog作者有着共同关注的主题,他们将更容易意识到对方并很可能开始交流从而产生有价值的讨论供社区的读者参考.系统基于用户给定的查询检索相关的Blog网页,随后排序,并从相互链接和相关的Blog中发现社区,生成一个关注用户查询主题的社区,通过山形视图将该社区各个层次的细节呈现给用户.

3.2.2 基于网络流量的社区发现

SPB(shortest-path betweens)算法[31]是一种基于流量的社区发现算法,通过找到任意两节点间的最短路径,去掉由最短路径累计流量最大的边来实现社区分割,在抽取现实世界的社区时取得了很好的效果.然而,Blog有意无意地趋向于包含很多无关的节点和边,由于SPB算法没有使用邻接矩阵中的信息,因而不能检测到无意义的节点和边,也就无法在Blog社区发现方面取得理想的效果.

在SPB算法的基础上,Ishida提出了WP(weakest pairs)算法[32].在WP算法模型中,Blog页面的集合F(设共有n_f个页面)和非Blog网页的集合T(设共有n_t个页面)作为节点,F中节点指向T中节点的链接作为边,从而构成了一个有向图.关系矩阵$R(n_f*n_t)$显示了一个Blog是否引用一个非Blog网页,如果第i个Blog引用了第j个非Blog网页,则r_{ij}为 1,否则为 0.$F=R*R^t$,$T=R^t*R$,F表示任意两个Blog引用相同的非Blog网页数,T表示任意两个非Blog网页被同一Blog同时引用数.然而,这种直接基于共同引用数量的相似度定义有很大的噪声,出度或入度本来就大的网页将会表现出和大部分网页都有较高的相似度分值.因此,对关系矩阵F和T都进行归一化得到F^{rel}和T^{rel},使矩阵每行元素之和都为 1,然后定义两个网页i和j之间的关系强度为$FS(i,j)=f_{ij}+f_{ji}$,$TS(i,j)=t_{ij}+t_{ji}$.这种定义使得出度或入度本来就大的网页与其他网页之间的关系弱化.根据归一化后的关系矩阵F^{rel},T^{rel}和关系强度定义,找到强度值为 0 之外的关系最弱的网页对.之后,找到关系最弱的网页对之间的最短路径,计算最短路径中边的频率,去除频率最高的边,得到的每个子图则被认为是一个社区.定义信息丢失和不完整性为衡量算法的两个指标,结果表明,对于节点较多和边密度较大的图,WP算法比SPB算法更加有效.

3.2.3 基于 Blog 间相互感知的社区发现

Lin等人[33]认为,社区的形成来源于Blog作者的行为,且行为必须是交互的.由此提出了基于Blog间相互感知的社区发现方法,特定的行为类型及其发生的频率和时间决定了不同的相互感知度,以此作为图中边的权值.方法利用个人Blog行为和语义链接结构,使用相互感知特性和基于排序的社区抽取算法发现社区.Zhou等人[34]使用PageRank算法选择种子,然后通过联合扩散确定社区成员.以图论的评价标准作为基线,并引入了新的评价标准,以便更好地体现动态和时间特性,将该方法分别应用于不同的数据集,并取得了较好的效果.

3.2.4 几种方法的比较

首先,各种方法对社区的理解和认识不同,从而导致了图中的节点和边所代表的意义不同.WP 算法中的节点为单个网页,边为网页间的链接;其余两种算法模型中的节点均为 Blog 站点,并且都认为节点并非同等重要.在基于 Blog 重要性的算法中,边的权值由 Blog 重要性决定;在基于相互感知的算法中,边的权值由相互感知度决定.从分割方法上看,WP 算法采用基于流量的方法,通过去除负载最大边抽取社区;另外两种算法则通过基于排序的聚类算法抽取 Blog 社区.此外,在基于 Blog 重要性的算法中,根据用户给定的查询检索相关的 Blog,以此为基础进行社区发现.在基于相互感知的算法中,使用全局链接信息而不是局部的,不同的链接类型代表不同的关系.由于各种方法对社区的理解、预期的目的、使用的数据集和评价方法均不相同,因此,很难对各种方法的实验结果进行客观的比较和评价.

4 Blog 重要性分析

4.1 传统的网页重要性分析

网页重要性评估的思想认为,每个网页被量化的价值通过一种递归的方式来定义,由所有链接指向它的网页的价值程度所决定.基于此,通常采用链接分析的方法对网页重要性进行分析.PageRank[35]算法和HITS[22]算法是两种最具代表性的链接分析算法,具有代表性的相关系统包括基于PageRank算法的Google搜索引擎以及IBM Almaden实验室开发的基于HITS算法的ARC(automatic resource compilation)系统[23]和CLEVER系统[24].

在HITS算法模型中,网页被分为权威性网页和中心性网页,其原因在于某些同主题权威性网页之间由于竞争关系等原因,缺乏相互之间的链接关系,而通过中心性网页能够更好地描述互联网的组织结构.

在 PageRank 算法中,假定网页 $T1,\ldots,Tn$ 有链接指向网页 A,设 $PR(A),PR(T1),\ldots,PR(Tn)$分别表示网页 $A,T1,\ldots,Tn$ 的 PageRank 值,参数 d 为一个跳转系数,介于 0~1 之间,通常取 0.85,$C(A)$被定义为 A 的出度,则网页 A 的 PageRank 值 $PR(A)$可由式(1)计算得到.

$$PR(A)=(1-d)+d\left(\frac{PR(T1)}{C(T1)}+\ldots+\frac{PR(Tn)}{C(Tn)}\right) \tag{1}$$

4.2 Blog重要性分析

Blog领域与传统门户网站的区别在于,Blog作者关注的主题更加明确,并且通常代表一种相对单一的观点,对重要Blog作者的发现将方便用户的信息查询.因此,Blog重要性分析除了包括对Blog网页的重要性分析以外,还包括对Blog作者的重要性分析[4,36].

基于二分有向图或基于hub和authority的排序算法存在如下 3 个问题:(1) 托管网站之间的关联相互增强;(2) 自动生成链接;(3) 没有相关的节点,从而导致主题漂移[37].

Blog 重要性分析同样面临以上 3 个问题,且更加严重.Blog 作者可以更加容易地使自己发布的网页相互链接;Blog 系统可能自动地嵌入指向入口站点或来自商业网站的广告链接;由于 Blog 信息的多样性和随意性,Blog 容易链接指向更多不相关的网页或站点.这些无意义的链接将影响 Blog 重要性分析的结果并容易导致主题漂移.

部分研究采用了传统的链接分析方法,它们的区别主要在于,根据不同的理解和需要对有向图中的边进行不同定义,使用不同种类的链接,并赋以不同的权值.Belle等人[30]提出了一种基于查询和Blog条目对Blog排序的方法,在系统模型中,一个Blog由多个Blog条目构成,系统根据用户的查询通过关键词匹配找到相关的Blog条目,用类似PageRank的算法对Blog条目打分,将一个Blog所包含的Blog条目的分值加权平均,得到这个Blog的排序分值.iRank算法[11]通过推断Blog间所有可能的信息传播途径,得到潜在信息流图,并在此基础上应用PageRank算法计算Blog的重要性.这种算法赋于具有较高感染力的Blog以高的分值,体现了Blog作为信息传播者的有效性.

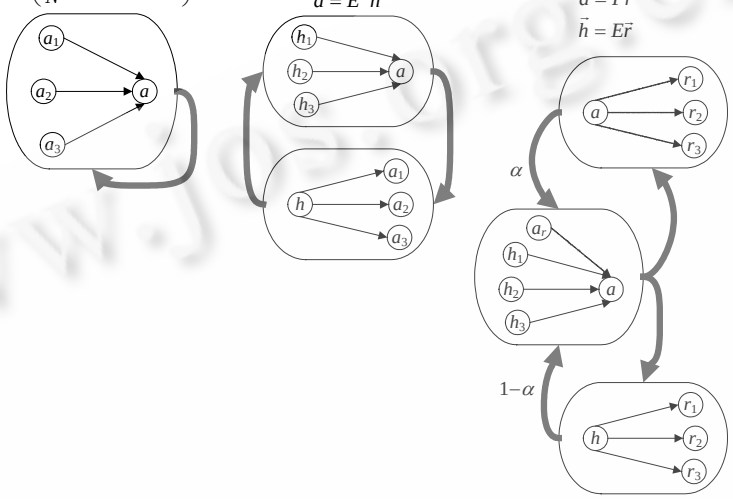
以上研究主要通过定义不同的链接关系形成不同的有向图,然后应用传统的链接分析方法进行重要性分析,还有的研究采用了不同于传统方法的链接分析算法,EigenRumor算法是其中具有代表性的方法.EigenRumor算法[38]通过对特征向量的计算衡量Blog作者的hub和authority值,一个Blog作者的authority分值越高,表示其有能力提供更好的Blog条目;hub分值越高,表示其能够更好地为社区贡献评论.算法综合发表该Blog文章的Blog作者的authority值以及对该Blog文章进行评论的其他Blog作者的hub值,衡量该Blog文章的重要性.基于对Blog作者的评价,该算法允许对一个好的Blog作者发表的但还没有其他Blog链接指向的文章评一个较高的分数.EigenRumor算法对传统链接分析方法的算法模型进行了改进,其与PageRank和HITS算法的比较见表1.EigenRumor算法的主要贡献有两个方面:(1) 直接分析Blog网页与Blog作者间的链接,在一定程度上解决了由于Blog文章间的链接较少而难以全面分析Blog网页重要性的问题,拓宽了被评分的Blog网页的覆盖率;(2) 一个由有着高authority分值的Blog作者提供的Blog网页一旦提交就能得到靠前的排序,而不像在PageRank和HITS算法模型中必须依靠链接入度来获得较高的分值,这符合Blog领域热衷于讨论新的话题和突发事件的特

点,有助于用户在第一时间找到有价值的信息.

Table 1 Comparison of EigenRumor, PageRank and HITS algorithms

表 1 EigenRumor 与 PageRank 及 HITS 算法的比较

	PageRank[35]	HITS[22]	EigenRumor[38]	
Entities	Web page	Web page	Agent/Object	
Link types	Evaluation (E)	Evaluation (E)	Evaluation (E) Provisioning (P)	
Scores	Authority ($\vec{a}$)	Authority ($\vec{a}$)	Authority ($\vec{a}$)	Agent
		Hub ($\vec{h}$)	Hub ($\vec{h}$)	Agent
			Reputation ($\vec{r}$)	Object
Algorithm	$\vec{a}=\left(\frac{d}{N}1_N+(1-d)E^T\right)\vec{a}$	$\vec{h}=E\vec{a}$ $\vec{a}=E^T\vec{h}$	$\vec{r}=\alpha P^T\vec{a}+(1-\alpha)E^T\vec{h}$ $\vec{a}=P\vec{r}$ $\vec{h}=E\vec{r}$	



除了常用的链接分析方法以外,Shinsuke等人[4]根据Blog作者发布信息的被引用次数以及对社区文章数量和内容的影响,采用基于链接、流行度及主题变化的指标判断其重要性.Shinsuke等人认为,重要的Blog作者可分为讨论的鼓动者(agitators)和总结者(summarizers),并提出了发现这两类Blog作者的方法.实验模型中,Blog线索(thread)是由起始的Blog页面和该页面上回复链接所指向的页面,以及源链接指向的且被线索中的条目所关联的页面构成.一个Blog作者获得线索中足够多的链接,在其发布文章后,线索中的文章数明显增多,且对一个Blog线索的主题有着重大的影响并使之发生改变时,被判定为鼓动者;而链接指向线索中足够多相关页面的Blog作者,则被判定为总结者.

几种典型的 Blog 重要性分析方法的比较见表 2,除了 iRank 算法对 Blog 站点进行评价以外,其中 3 种方法的评价对象都是 Blog 网页和 Blog 作者.对 Blog 的评分体现了不同方法的主要思想和对重要性的不同理解. Shinsuke 等人提出的方法根据 Blog 作者发布信息被引用次数以及对社区文章数量和内容的影响,采用基于链接、流行度及主题变化程度等指标判断其重要性.其余 3 种方法都采用了链接分析的方法:Belle 等人提出的方法继承了 PageRank 算法的思想,认为被更多其他 Blog 网页指向的 Blog 网页具有较高的重要性,利用 Blog 条目间的链接关系,通过 Blog 条目加权到 Blog 作者的重要性值;EigenRumor 算法对链接分析的算法模型进行了改进,根据发布 Blog 文章和对 Blog 文章进行评论的 Blog 作者的重要程度衡量该文章的重要性;iRank 算法与其他几种方法对重要性的理解不同,它通过计算所有潜在的链接对信息传播过程中起到重要作用的 Blog 作者赋以较高的权重.然而,iRank 算法和 EigenRumor 算法在 Blog 领域链接的稀疏性和链接动态结构的重要性方面有着相似性.

到目前为止,关于Blog信息源重要性评估本身还没有得到普遍认可的评价标准.理想的评价指标应该能够客观地反映其他信息源的推荐程度以及读者的关注程度,且能够方便而准确地获取以便于大规模的评价.基于

此,杨宇航等人[36]提出采用信息源的链接入度、评论数、Trackback数等作为Blog重要性分析的评价指标,具有一定的参考价值.

Table 2 Comparison of different blog importance analysis approaches

表 2 Blog 重要性分析方法比较

Algorithm	Evaluation object	Score type	Is link analysis-based	Edge definition	Main criteria
Shinsuke, *et al*.[4]	Blog entry /Blogger	Link-Based, popularity-based and topic-based discriminant	No		Number of entries and content change in the thread after the target entry published
EigenRumor[38]	Blog entry /Blogger	Authority Hub Reputation	Yes	Links between bloggers and their entries, comments and the target entries	Importance of the blogger and other bloggers who submitted comments to the entry
Belle, *et al*.[30]	Blog entry /Blogger	Authority	Yes	Links between blog entries	Importance of blog entries
iRank[11]	Blog site	Infection	Yes	Implicit links between blog sites	Initialization of the epidemics

5 Blog 搜索

Blog 搜索的首要问题是发现其区别于传统 Web 搜索的特点,分析研究专门用于 Blog 搜索的相关技术和系统的必要性,并探寻 Blog 搜索自身应该具备的功能和特点.

Broder[39]将Web搜索的查询分为 3 类:查找主题相关的Web页面的信息型查询(informational)、查找一个指定名称的站点或主页的导航型查询(navigational)以及查找一个服务入口以便进行下一步访问的事务型查询(transactional).然而,这样的分类体系并不能很好地适用于Blog领域,对大量Blog搜索引擎查询日志的分析[40,41]显示,Blog搜索的需求与传统Web搜索有着较为显著的区别:(1) Blog搜索大部分是关于名实体的查询,包括用户感兴趣的产品或某个领域的著名人物以及用户所处生活环境中的相关事物等(如所在的公司、同事等);(2) Blog 搜索的关注领域更多地集中于技术、娱乐和政治等领域;(3) Blog 搜索对即时事件有着特别的关注,这种现象也与 Blog 是即时事件的消息和评论源头的假设相吻合.而在用户行为方面,Blog 搜索与传统的 Web 搜索十分类似,用户通常也只关注排序最靠前的几个结果.

通过对 Blog 搜索日志的分析,发现了 Blog 搜索与传统 Blog 搜索的区别,验证了专门研究 Blog 搜索技术的必要性.有关 Blog 搜索的研究目前主要集中于 Blog 搜索系统的研究和开发、为搜索服务的内容挖掘以及相关反馈在 Blog 搜索中的应用等方面.

如今,很多Blog搜索引擎被开发并投入使用,其中有代表性的包括Technorati,Blogpulse,BlogWatcher,Bloglines等,Google等Web搜索引擎也提供了专门的Blog搜索功能,这些搜索引擎提供了多种不同类型的Blog搜索服务.Technorati提供了Blog文章、作者、图片、视频、音乐和事件等不同类型的Blog搜索功能.Blogpulse[3]较好地利用了Blog内容的特性,提供了支持话题和Blog作者搜索的功能,并能显示Blog作者之间的会话过程,Blogpulse还提供了引用、信息源以及相似的Blog站点等信息,便于用户找到具有相似兴趣或特征的Blog作者.BlogWatcher[5]在话题和口碑搜索方面具有独特性,能够抽取热门话题,并且通过图示表现有关用户查询的正、负面消息.BLOGRANGER[41]提供了多种Blog搜索服务,并通过词项共现和引用统计等对搜索结果进行分类,用户问卷调查显示了用户对BLOGRANGER系统在Blog作者搜索、话题搜索以及口碑搜索方面的满意度高于传统的Web搜索引擎.BlogHarvest[42]是一个用于抽取Blog作者兴趣、找到和推荐讨论相似话题的Blog,并提供以Blog作者为导向的搜索服务的Blog搜索引擎.BlogHarvest使用了基于话题相似度的聚类和基于词性标注的观点挖掘等自然语言处理技术,还自动生成包括作者兴趣、经常访问的站点和可视化的朋友网络在内的Blog档案,便于用户查找和提供个性化的服务.中文Blog搜索引擎也陆续出现,例如Souyo,Booso,Blogcn等,但发展相对缓慢,功能相对单一,且检索的Blog数量和覆盖率较小.

与传统的门户网站提供有限主题且观点类似的信息不同,Blog领域的话题更加分散且观点各不相同,因此,

Blog搜索不仅需要关注特定话题,还需要关注对话题的不同观点.观众的印象会根据镜头的顺序而有所改变,类似地,Blog读者也会因为Blog及其评论和Trackback组织顺序的不同而受到不同的影响.Daisuke等人[43]将电视和Blog两种媒体结合起来,通过视频场景切割和特征抽取,搜索话题相关且观点顺序与之相似的Blog信息.Osamu等人[44]也进行了与Blog作者观点有关的工作,抽取相关话题且具有倾向性的句子,不仅包括正面和负面的观点,还包括中性的请求、建议和想法等.由于这样的句子通常体现了作者的情感、要求和建议等,使得用户能够很快知道作者要表达的想法.

相关反馈技术在文档检索领域已有研究,用于根据用户需求生成合适的查询.相关反馈要求一个用户对相关的检索文档有着具体的标准,且生成的查询严格受限于现有的文档集.由于Blog包含了关于很多不同主题的极大数量的短文,这个问题在Blog搜索中更为严重.基于此,Yasufumi等人[45]假定用户搜索Blog空间时没有十分明确的目标,而是有着不同的兴趣,应用基于关键词地图的相关反馈和Blog搜索交互,提出的算法考虑了用户在关键词地图上感兴趣的多种话题,并基于这些话题进行子查询,结果更加符合用户需要,并增强了关键词地图的可读性.

6 作弊 Blog 识别

与传统互联网一样,Blog 领域同样充斥着大量的垃圾页面.一些网站设计或维护人员通过作弊手段误导搜索引擎,提高页面在检索结果中的排名,这些利用作弊手段的 Blog 被称为作弊 Blog(spam blog 或 splog).常用的作弊手段可分为基于内容和基于链接结构的两大类:基于内容的作弊手段通过自动生成或剽窃内容以及填充热门特征词等方法提高与查询词间的相关性;基于链接结构的作弊手段通过制造和发送无意义的链接和 Ping 扰乱链接关系,从而提高目标网页的重要性.传统的利用链接技术的作弊手段包括以克隆目录(directory cloning)为代表的制造出链(out-link)的手段以及蜜罐诱饵(honey pot)、渗入目录、交换链接、购买过期域名和制造链接工厂等制造入链(in-link)的手段.

随着Blog的流行,以作弊评论为代表的新的作弊手段被越来越多的作弊者使用,并已产生日益严重的影响.作弊评论是通过在支持用户动态编辑的网页上添加评论和回复,并在其中加入作弊链接的作弊行为[20],这种现象在Blog领域尤为突出.Blog等网络应用在为用户发布信息提供方便的同时,也难以避免地为作弊者提供了方便.传统的作弊手段通常需要作弊者具有一定的技术水平并掌握一定的资源,而作弊评论没有任何类似的约束或限制,作弊者只需将包含指向目的页面链接的信息作为评论,复制到任意Blog页面上即可.由于作弊过程大为简化,作弊评论已成为Blog领域最主要的作弊手段之一,商业搜索引擎正为这个问题寻找新的解决方法[46].

随着 Blog 领域作弊手段的日益增多并已造成严重影响,作弊 Blog 的识别与过滤已成为一个亟待解决的问题并受到广泛的关注,相关研究可主要分为人工识别方法和基于内容的识别方法.

6.1 人工识别方法

大部分作弊 Blog 的过滤方法是通过在 Web 信息的发布机制上采取措施,例如要求评论人注册、过滤评论中的 HTML 文本以及阻止对以前的文章进行评论等.这些方法对控制作弊评论起到了一定的作用,但同时也过滤掉了很多对用户来说很重要的正常评论,并对用户的使用造成了不便.此外,这些方法只能处理新的评论,对已经存在的评论无能为力.通过 IP 地址进行识别也是防止作弊行为的一类常用方法,然而这类方法需要持续的人工维护,且作弊者可以通过代理或假 IP 地址骗过过滤机制.2005 年,包括 Yahoo、MSN 搜索、Google 在内的搜索引擎宣布与 Blog 托管网站联合起来,通过在链接上添加特殊标记来防止作弊链接.然而,这种方法的使用同样会产生很多问题,它扰乱了 Blog 内部正常的链接关系,而且可能被网站管理人员滥用.

Han等人[47]认为,基于内容分析的方法其结果并不十分理想,总是有一些不确定的作弊信息需要人为地判别,因而提出了一种新颖的协作式垃圾信息过滤方法,依赖于对作弊链接的人为识别以及通过可信网络对作弊信息的共享.理想状态下,如果参与协作的用户数量很多,每个用户的参与可降低到可忽略的程度.然而,无法保证足够多的用户愿意参与其中,即便如此,任何用户有意或无意的错误识别将造成全局影响.

6.2 基于内容的识别方法

基于内容的识别和过滤方法主要来源于垃圾邮件的检测,以发现垃圾邮件和非垃圾邮件内容间的不同.这类方法也已应用到作弊 Blog 的识别和过滤中,通过对 Blog 内容甚至 Blog 中的链接指向页面的内容进行分析进而识别和过滤作弊 Blog.目前,这类方法主要基于特征词集合或正则表达式.由于特征词或规则集合需要手工完成,因此,这类方法与其他人工方法一样存在维护费时、覆盖率有限以及规则间相互冲突等不可避免的问题.

Kolari等人提出了采用支持向量机训练的作弊Blog检测模型[48,49],并分析了Blog和Splog在内容和链接关系方面的区别[50].然而,这种方法需要大量人工标注的训练语料和全局的链接关系,需要耗费大量人力且难以应用于在线识别.

Narisawa等人[51]利用作弊者通常会复制同样的内容达到作弊目的的特点,借鉴子串扩大(substring amplification)的方法[52],根据内容的重复频度以及正常文章和评论中的子串数服从Zipf分布的特性以检测作弊评论.然而这种方法仅使用了频度信息,很难分辨作弊Blog和包含同样高频内容的正常Blog信息.

Mishne等人[20]通过计算Blog文章和对应评论语言模型间的相似性识别作弊链接.这种方法的优势在于不需要训练和Web链接知识.然而,现实世界中评论通常很短,不可避免地导致了语言模型的数据稀疏,而且很多作弊者可能从内容上模仿正常评论,因此,实验结果尤其是对较短评论的识别效果并不理想.

杨宇航等人[53]在比较全面地分析作弊行为的基础上,提出了一种整合多种特征的作弊评论识别方法.与其他方法相比,该方法不需要任何先验知识和训练过程,既可以用于识别已经存在的作弊评论,也可用于Blog系统中进行在线识别,中文Blog领域的实验初步验证了方法的有效性.

总之,人工过滤方法需要大量的手工劳动,而目前基于内容的方法也需要大量手工标注的语料进行或人工维护特征词和规则集合,且由于特征选取的单一导致效果并不理想.此外,目前存在的方法只能单一地处理新产生或已经存在的 Blog 信息.为了满足实际应用需要,减少人工参与,能够同时对新旧 Blog 信息进行识别且性能稳定的方法是具有重要意义的研究方向.

7 存在的问题和未来研究展望

本文概述了目前 Blog 领域的 6 个主要研究方向,分别阐述了这 6 个研究方向的基本内容、常用方法和研究进展.总的来说,虽然相关研究才刚刚起步,但 Blog 领域已成为一个新兴的研究领域.许多相关领域的研究成果都可以供其借鉴,其中主要包括自然语言处理、链接分析、Web 挖掘和机器学习等方法和技术.然而,由于相关研究工作开展的时间很短,以及Blog领域自身的特殊性和复杂性,该领域总的来看还处在探索阶段,存在许多有待解决的问题.主要包括:

(1) 缺乏有效的方法.大多数关于 Blog 的研究都借鉴了相关领域的研究成果,直接沿用了相关方法或在其基础上进行修改,很多这样的移植并不十分成功,原因大多在于没有充分利用Blog领域的特点,而且由于缺乏开放的实验平台和统一的评价标准,很难得到有效的验证.

(2) 缺乏实验工具和平台.目前还没有公开发布并具有广泛影响的实验工具和平台,用于实验的公共数据集也很少,2006年的TREC评测(http://trec.nist.gov)中将Blog搜索作为一个新的任务,并提供了相应的数据集.此外,同年 WWW 会议(http://www2006.org)也提供了有关 Blog 的数据集.但这两个数据集所针对的任务和对语料加工的程度都十分有限.有必要关注相关的资源和实验平台的建立,这将在很大程度上促进Blog相关研究的发展.

(3) 缺乏统一的评价标准.关于Blog的评价标准还存在广泛的争议,在很多方面都没有取得一致.以Blog重要性分析的评价为例,关于哪种链接更能代表重要性,以及哪种Blog作者更具有影响力存在争论.缺乏统一的评价标准不利于Blog相关研究的进一步发展.所以,如何对相关研究进行定量的评价是一个重要的研究方向,也是一个迫切需要解决的问题.

总之,国际上 Blog 领域的相关研究刚刚起步,并开始活跃起来,在国内还鲜有介绍相关领域研究的文献[36,53,54],针对Blog领域的研究现状,以下几个方面是未来研究中值得关注的方向:

(1) 资源建设和评价体系建立.语料或数据集等资源的建设是相关工作顺利开展的基础,统一的评价体系

的建立,是对方法进行客观评价的前提,两者是不可忽视的基础性工作、是支持相关研究快速发展的必要保证.

(2) Blog 特征的挖掘和利用.Blog 领域的很多问题和任务依赖于自然语言处理技术,充分挖掘和有效利用 Blog 的特征信息将有利于自然语言处理技术在相关任务中的有效应用.例如,利用标签信息改善 Blog 搜索和分类的性能、利用 RSS 信息提高 Blog 信息自动文摘的可靠性等.

(3) 系统研究的开展.目前,相关研究相对分散,然而 Blog 领域的相关问题之间具有紧密的联系,因此,系统性的研究非常重要.例如,Blog 社区的发现需要采用链接分析和内容挖掘等方法,同时结合作弊 Blog 识别和重要性分析去除垃圾信息的干扰,得到由关注相关领域且观点相似的 Blog 作者组成的社区.这样的社区既是相关内容挖掘和特定 Blog 搜索的重要信息来源,又是提供个性化的信息推送和分众服务的对象.

References:

[1] comScore Networks, Inc. Behaviors of the blogosphere: Understanding the scale, composition and activities of Weblog audiences. 2005. http://www.comscore.com/blogreport/comScoreBlogReport.pdf

[2] Thelwall M. Bloggers during the London attacks: Top information sources and topics. In: Proc. of the World Wide Web 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Edinburgh, 2006. http://www.blogpulse.com/www2006-workshop/papers/blogs-during-london-attacks.pdf

[3] Glance N, Hurst M, Tomokiyo T. BlogPulse: Automated trend discovery for weblogs. In: Proc. of the World Wide Web 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. New York, 2004. http://www.blogpulse.com/papers/www2004glance.pdf

[4] Nakajima S, Tatemura J, Hino Y. Discovering important bloggers based on analyzing blog threads. In: Glance N, Adar E, Hurst M, Adamic L, eds. Proc. of the World Wide Web 2005 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Chiba, 2005. http://www.blogpulse.com/papers/www2004glance.pdf

[5] Nanno T, Fujiki T, Suzuki Y, Okumura M. Automatically collecting, monitoring, and mining Japanesse weblogs. In: Glance N, Adar E, Hurst M, Adamic L, eds. Proc. of the World Wide Web 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. New York: ACM Press, 2004. 320–321.

[6] Kleinberg J. Bursty and hierarchical structure in streams. In: Hand D, Keim D, eds. Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Edmonton: ACM Press, 2002. 1–25.

[7] Oka M, Abe H, Kato K. Extracting topics from weblogs through frequency segments. In: Proc. of the World Wide Web 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Edinburgh, 2006. http://www.blogpulse.com/www2006-workshop/papers/wwe2006-oka.pdf

[8] Fukuhara T, Murayama T, Nishida T. Analyzing concerns of people using weblog articles and real world temporal data. In: Adar E, Glance N, Hurst M, eds. Proc. of the World Wide Web 2005 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Chiba, 2005. http://www.blogpulse.com/papers/2005/fukuhara.pdf

[9] Fujiki T, Nanno T, Okumura M. Differences between blogs and Web diaries. In: Adar E, Glance N, Hurst M, eds. Proc. of the World Wide Web 2005 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Chiba, 2005.

[10] Thelwall M. Bloggers during the London attacks: Top information sources and topics. In: Proc. of the World Wide Web 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Edinburgh, 2006. http://www.blogpulse.com/www2006-workshop/papers/blogs-during-london-attacks.pdf

[11] Adar E, Zhang L, Adamic L, Lukose RM. Implicit structure and the dynamics of blogspace. In: Glance N, Adar E, Hurst M, Adamic L, eds. Proc. of the World Wide Web 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. New York: ACM Press, 2004. http://www.hpl.hp.com/research/idl/papers/blogs/index.html

[12] Balog K, de Rijke M. Decomposing bloggers' moods: Towards a time series analysis of moods in the blogosphere. In: Proc. of the World Wide Web 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Edinburgh, 2006. http://citeseer.ist.psu.edu/balog06decomposing.html

[13] Lin J, Halavais A. Mapping the blogosphere in America. In: Glance N, Adar E, Hurst M, Adamic L, eds. Proc. of the World Wide Web 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. New York: ACM Press, 2004.

[14] Hurst M. GIS and the blogosphere. In: Adar E, Glance N, Hurst M, eds. Proc. of the World Wide Web 2005 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Chiba, 2005. http://citeseer.ist.psu.edu/733458.html

[15] Ohkura T, Kiyota Y, Nakagawa H. Browsing system for weblog articles based on automated folksonomy. In: Proc. of the World Wide Web 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Edinburgh, 2006. http://www.r.dl.itc.u-tokyo.ac.jp/~kiyota/paper/2006/WWW_2006_blog.pdf

[16] Trevino EM. Blogger motivations: Power, pull, and positive feedback. Internet Research 6.0. 2005. http://blog.erickamenchen.net/MenchenBlogMotivations.pdf

[17] Gumbrecht M. Blogs as "protected space". In: Proc. of the World Wide Web 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. New York, 2004. http://www.blogpulse.com/papers/www2004gumbrecht.pdf

[18] Kumar R, Novak J, Raghavan P, Tomkins A. On the bursty evolution of blogspace. In: Lawrence S, ed. Proc. of the 12th Int'l Conf. on World Wide Web. New York: ACM Press, 2003. 568−576.

[19] Wei C. Formation of norms in a blog community. In: Gurak L, Antonijevic S, Johnson L, Ratliff C, Reyman J, eds. Proc. of the Into the Blogosphere, Rhetoric, Community and Culture of Weblogs. Minneapolis: University of Minnesota, 2004.

[20] Mishne G, Carmel D, Lempel R. Blocking blog spam with language model disagreement. In: Adar E, Glance N, Hurst M, eds. Proc. of the World Wide Web 2005 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Chiba, 2005.

[21] Yang N, Gong ZD, Li X, Meng XF. Survey of Web communication Identification. Journal of Computer Research and Development, 2005,42(3):439−447 (in Chinese with English abstract).

[22] Kleinberg J. Authoritative sources in a hyperlinked environment. In: Tarjan RE, Baecker T, eds. Proc. of the 9th ACM-SIAM Symp. on Discrete Algorithms. New York: ACM Press, 1997. 668−677.

[23] Gibson D, Kleinberg J, Raghavan P. Inferring web communities from link topology. In: Proc. of the 9th Conf. on Hypertext and Hypermedia. New York: ACM Press, 1998. 225−234.

[24] Chakrabarti S, Dom B, Indyk P. Enhanced hypertext categorization using hyperlinks. In: Halevy AY, Ives ZG, Doan AH, eds. Proc. of the ACM SIGOMD Int'l Conf. on Management of Data. New York: ACM Press, 1998. 307−318.

[25] Chakrabarti S, Dom B, Gibson D, Kleinberg J. Automatic resource compilation by analyzing hyperlink structure and associated text. In: Cho J, Garcia-Molina H, Page L, eds. Proc. of the 7th Int'l Conf. on World Wide Web. New York: ACM Press, 1998. 65−74.

[26] Kumar R, Raghavan P, Rajagopalan S, Tomkins A. Trawling the Web for emerging cyber-communities. Computer Networks, 1999,31(11-16):1481−1493.

[27] Flake G, Lawrence S, Giles C. Efficient identification of Web communities. In: Simoff SJ, Za-ane OR, eds. Proc. of the 6th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2000. 150−160.

[28] Toyoda M, Kitsuregawa M. Creating a Web community chart for navigating related communities. In: Proc. of the 12th ACM Conf. on Hypertext and Hypermedia. New York: ACM Press, 2001. 103−112. http://portal.acm.org/citation.cfm?id=504216.504244

[29] Gruhl D, Guha R, Liben-nowell D, Tominks A. Information diffusion through blogspace. In: Glance N, Adar E, Hurst M, Adamic L, eds. Proc. of the World Wide Web 2004. New York: ACM Press, 2004. 491−501.

[30] Tseng BL, Tatemura J, Wu Y. Tomographic clustering to visualize blog communities as mountain views. In: Adar E, Glance N, Hurst M, eds. Proc. of the World Wide Web 2005 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Chiba, 2005. http://www.blogpulse.com/papers/2005/tseng.pdf

[31] Girvan M, Newman MEJ. Community structure in social and biological networks. Proc. of National Academy Sciences, 2002,99: 7821−7826.

[32] Ishida K. Extracting latent weblog communities—A partitioning algorithm for bipartite graphs. In: Adar E, Glance N, Hurst M, eds. Proc. of the World Wide Web 2005 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Chiba, 2005.

[33] Lin YR, Sundaram H, Chi Y, Tatemura J, Tseng B. Discovery of blog communities based on mutual awareness. In: Proc. of the World Wide Web 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Edinburgh, 2006. http://ame2.asu.edu/faculty/hs/pubs/2006/www2006-discovery-blt-final2.pdf

[34] Zhou DY, Weston J, Gretton A, Bousquet O, Scholkopf B. Ranking on data manifolds. In: Thrun S, Saul L, Schoelkopf B, eds. Proc. of the Advances in Neural Information Processing System 16. The MIT Press, 2004. 169−176.

[35] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. In: Cho J, Garcia-Molina H, Page L, eds.Proc. of the 7th Int'l Conf. on World Wide Web. New York: ACM Press, 1998. 107−117.

[36] Yang YH, Zhao TJ, Zheng DQ, Yu H. Discovering important bloggers based on link analysis. Journal of China Information Processing, 2007,21(5):68−72 (in Chinese with English abstract).

[37] Bharat K, Henzinger MR. Improved algorithms for topic distillation in a hyperlinked environment. In: Croft WB, Moffat A, van Rijsbergen CJ, Wilkinson R, Zobel J, eds. Proc. of the 21st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 1998. 104−111.

[38] Fujimura K, Inoue T, Sugisaki M. The EigenRumor algorithm for ranking blogs. In: Adar E, Glance N, Hurst M, eds. Proc. of the World Wide Web 2005 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Chiba, 2005.

[39] Broder A. A taxonomy of Web search. SIGIR Forum, 2002,36(2):3−10.

[40] Mishne G, de Rijke M. A study of blog search. In: Lalmas M, *et al*., eds. Proc. of the 28th European Conf. on Information Retrieval. London: Springer-Verlag, 2006. 289−301.

[41] Fujimura K, Toda H, Inoue T, Hiroshima N. BLOGRANGER—A multi-faceted blog search engine. In: Proc. of the World Wide Web 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Edinburgh, 2006. http://www.blogpulse.com/www2006-workshop/papers/wwe2006-fujimura.pdf

[42] Joshi M, Belsare N. BlogHarvest: Blog mining and search framework. In: Proc. of the Int'l Conf. on Management of Data COMAD 2006. 2006. http://nikhilbelsare.googlepages.com/BlogHarvest_vldb_paper_0.6_final.pdf

[43] Kitayama D, Sumiya K. A blog search method using news video scene order. In: Proc. of the 12th Int'l Conf. on Multi Media Modeling. Beijing: IEEE Press, 2006. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1651369

[44] Furuse O, Hiroshima N, Yamada S, Kataoka R. Opinion sentence search engine on open-domain blog. In: Proc. of the 20th Int'l Joint Conf. of Artificial Intelligence. Hyderabad: IJCAI Press, 2007. http://www.ijcai.org/papers07/Papers/IJCAI07-443.pdf

[45] Takama Y, Kajinami T, Matsumura A. Application of keyword map-based relevance feedback to interactive blog search. In: Tarumi H, Li Y, Yoshida T, eds. Proc. of the 2005 Int'l Conf. on Active Media Technology (AMT 2005). New York: IEEE Press, 2005. 112−115.

[46] Henzinger MR, Motwani R, Silverstein C. Challenges in Web search engines. SIGIR Forum, 2002,36(2):11−22.

[47] Han SY, Ahn YY, Moon S, Jeong HW. Collaborative blog spam filtering using adaptive percolation search. In: Proc. of the World Wide Web 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Edinburgh, 2006. http://dspace.kaist.ac.kr/handle/10203/59

[48] Kolari P, Finin T, Joshi A. SVMs for the blogosphere: Blog identification and splog detection. In: Proc. of the AAAI Spring Symp. on Computational Approaches to Analyzing Weblogs. California: AAAI Press, 2006. 92−99.

[49] Kolari P, Java A, Finin T, Oates T, Joshi A. Detecting spam blogs: A machine learning approach. In: Proc. of the 21st National Conf. on Artificial Intelligence. California: AAAI Press, 2006.

[50] Kolari P, Java A, Finin T. Characterizing the splogosphere. In: Proc. of the World Wide Web 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Edinburgh, 2006.

[51] Narisawa K, Yamada Y, Ikeda D, Takeda M. Detecting blog spams using the vocabulary size of all substrings in their copies. In: Proc. of the World Wide Web 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Edinburgh, 2006.

[52] Ikeda D, Yamada Y. Gathering text files generated from templates. In: Proc. of the Workshop on Information Integration on the Web (IIWeb 2004). California: AAAI Press, 2004. 21−26.

[53] Yang YH, Zheng DQ, Yu H, Zhao TJ. Comment spam identification based on content analysis. In: Proc. of the NetSec2007. Qingdao, 2007. 288−294 (in Chinese with English abstract).

[54] Yang YH. Discovering important bloggers based on content and link analysis [MS. Thesis]. Harbin: Harbin Institute of Technology, 2006 (in Chinese with English abstract).

附中文参考文献:

[21] 杨楠,弓丹志,李忱,孟小峰.Web 社区发现技术综述.计算机研究与发展,2005,42(3):439−447.

[36] 杨宇航,赵铁军,郑德权,于浩.基于链接分析的重要 Blog 信息源发现.中文信息学报,2007,21(5):68−72.

[53] 杨宇航,郑德权,于浩,赵铁军.基于内容分析的作弊评论自动识别.见:第 4 届全国网络与信息安全技术研讨会(NetSec2007).青岛,2007.288−294.

[54] 杨宇航.基于内容与链接分析的重要 Blog 信息源发现[硕士学位论文].哈尔滨:哈尔滨工业大学,2006.

杨宇航(1983－),男,四川开江人,博士生,主要研究领域为自然语言处理,知识工程,基于内容的网络信息处理.

于浩(1971－),男,博士,副教授,主要研究领域为机器翻译与自然语言处理,机器学习与人工智能,网络智能计算.

赵铁军(1962－),男,博士,教授,博士生导师,主要研究领域为自然语言处理,机器翻译,基于内容的网络信息处理,人工智能应用.

郑德权(1968－),男,博士,副教授,CCF 会员,主要研究领域为跨语言信息检索,知识工程,自然语言处理.