软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(8):1785−1804 [doi: 10.3724/SP.J.1001.2011.03857]

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

面向网络论坛的高质量主题发现*

陈 友[1,2], 程学旗[1+], 杨 森[1,2]

[1](中国科学院 计算技术研究所,北京 100190)

[2](中国科学院 研究生院,北京 100049)

Finding High Quality Threads in Web Forums

CHEN You[1,2], CHENG Xue-Qi[1+], YANG Sen[1,2]

[1](Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100190, China)

[2](Graduate University, The Chinese Academy of Sciences, Beijing 100049, China)

+ Corresponding author: E-mail: cxq@ict.ac.cn

Chen Y, Cheng XQ, Yang S. Finding high quality threads in Web forums. *Journal of Software*, 2011,22(8): 1785−1804. http://www.jos.org.cn/1000-9825/3857.htm

Abstract: This paper presents a general detection framework, and develops a variety of content and structure features to find high quality threads. The feature selection algorithm, which is a combination of genetic algorithm, Tabu search and a machine learning algorithm, is designed to attain a better assessment of key features. In this paper, an experiment is done that focuses on the Tencent Message Boards. The experimental results, obtained from a large scale evaluation of over thousands of real web forum threads and user ratings, demonstrate the feasibility of modeling and detecting high quality threads. The proposed feature extraction methods, feature selection algorithms, and detection framework can be useful for a variety of domains such as Blogs and social network platforms.
Key words: Web forum; high quality; feature selection; feature extraction; classification

摘 要: 提出了一种通用的高质量主题发现框架.在该框架下,利用特征抽取技术提取内容特征,利用结构特征去发现高质量主题.提出了一种基于遗传算法、禁忌搜索与机器学习的特征选择算法,用来评价被抽取特征的重要性.在腾讯论坛数据集上进行了大量的实验.实验结果表明,该框架能够很好地发现高质量主题.提出的特征抽取算法、特征选择算法以及高质量主题发现框架能够在很多 Web2.0 领域得到应用,例如,博客、社会网络平台等.
关键词: 网络论坛;高质量;特征选择;特征抽取;分类
中图法分类号: TP311 **文献标识码**: A

2000 年以来,用户产生的内容(user generated content,简称 UGC)已经开始在网络上流行起来,并且数量庞大.现在,越来越多的用户遇到问题时首先选择在网络上提出问题,然后寻求他人的帮助与解答.这种相互之间进行询问以及解答的方式是 Web2.0 的重要体现.UGC 已经成为人们日常生活中不可缺少的信息来源,它是 Web2.0 的一个重要部分.它具有两个典型的特点.首先,它能够被所有的网络用户创建,这些用户不分背景、身份,

* 基金项目: 国家自然科学基金(60933005, 60903139); 国家高技术研究发展计划(863)(2007AA01Z438)
收稿时间: 2009-08-27; 修改时间: 2010-03-04; 定稿时间: 2010-04-14
CNKI 网络优先出版: 2010-11-17 17:23, http://www.cnki.net/kcms/detail/11.2560.tp.20101117.1723.003.html

都可以创建.例如,医生、学生、律师等各行各业的人都可以参与进来,创建自己的 UGC.其次,UGC 的质量参差不齐,描述语言丰富多彩,有书面的、口语的以及时下流行的网络语言.假设一个病人在网络上寻求医治自身疾病的方法,给他提供帮助的人可能有医生、护士、IT 人士、律师等.在这些人中,可能医生的回答质量比较高,用语专业、规范,而 IT 人士可能回答得比较差,用语不规范,并且内容方面可能还有错误.从上述两个特点可以看出, UGC 缺乏严格的发表者身份控制,这就使得 UGC 中包含大量的用户,而这些用户产生的大量信息中可能含有高质量的内容,也可能含有低质量的垃圾信息.用户要想遍历庞大的 UGC 并且从中获取自己需要的信息,不但需要花费大量的时间,而且还不一定找到需要的有用信息.如何从庞大的 UGC 中发现高质量的内容,已经成为当前迫切需要解决的一项任务.

UGC 的平台很多,诸如博客、维基论坛以及产品评论等.网络论坛作为 UGC 的典型平台,近年来被越来越多的人关注.网络论坛中含有大量的板块,诸如“医药卫生”、“交通信息”、“历史频道”等板块.用户可以选择在自己感兴趣的板块下浏览帖子(post),并且可以发表帖子.用户发表的帖子类型有主贴与回帖.板块的组成元素是主题(thread),即板块的全部内容是以大量的主题形式来表现的.每一个主题含有标题(title)、主贴、回帖等.这种开放形式的交流平台使得网络论坛上存在大量的 UGC.网络论坛上的 UGC 与传统 Web 上内容的一个显著区别是:论坛上的 UGC 质量分布不均匀[1],UGC 的质量分布可以从质量很高的主题到质量很低的主题,甚至有些主题还是垃圾信息;而传统 Web 上的内容都是专业的写作人员进行撰写的,用语规范,因此它们的质量都很高.同时,网络论坛上的信息量大,这就使得用户很难快速地从这些海量信息中找到自己所需要的高质量信息.因此,网络论坛上的高质量主题发现成为当前 Web2.0 方面研究的热点.

本文提出一种通用的框架,以发现网络论坛上的高质量主题(thread).该框架包括特征抽取、特征选择以及分类器 3 部分.特征抽取的目的是抽取高质量主题与低质量主题之间的区别性特征.这些特征可以用来发现高质量主题.特征选择用来从抽取的特征中选择主要特征.利用主要特征进行高质量主题发现不但可以提高效率,而且可以提高发现的精度.分类器利用选择后的特征作为输入,以检测高质量的主题.本文的贡献有 3 点:

(1) 提出了一种适用于 Web2.0 的通用检测框架.该框架可以快速、自动地发现网络论坛上的高质量主题.
(2) 在特征抽取上,包含小波特征与突发特征的时序特征被挖掘出来.挖掘出的特征能够很好地发现高质量主题,并且这种时序特征的挖掘方法可以在很多领域得到应用.
(3) 实验分析以及实验结果表明了哪些特征以及哪些分类器可以很好地用于高质量主题发现.

1 相关工作

UGC 含有大量有用的信息,对 UGC 进行分析与处理可以扩大 UGC 的影响范围,促使更多的人参与到 Web2.0 中.从 UGC 中提取高质量的信息不仅可以帮助人们更好地理解自己需要的信息,而且依靠这些高质量的信息可以帮助他们做出正确的决定.UGC 中含有的大量噪音信息,使得针对 UGC 的自动化分析与处理技术面临更大的挑战.

我们从 UGC 质量评估的任务、质量评估需要的特征、质量评估使用到的技术以及在那些领域需要质量评估 4 个方面对 Web2.0 环境下的 UGC 质量评估做一个概要的介绍.表 1 列出了这 4 个方面的基本要素.

UGC 质量评估的任务含有 3 个要素:质量的等级、质量评估的层次以及质量评估的输入、输出.当前,针对质量等级的划分最典型的有两类:高质量与低质量[1],即 UGC 的数据样本被标注为高质量或者低质量.网络论坛上,UGC 质量评估的层次有两种:基于 post 的质量评估与基于 thread 的质量评估.评估系统的输入是已经给出质量等级的数据样本,质量评估的输出是对未给出质量等级的目标样本贴上相应的等级标签.质量评估过程中需要用到一些特征来鉴别样本的质量等级,当前使用较多的几类特征是统计特征、链接特征和关系特征.统计特征是对 post 或者 thread 中的特殊符号和特殊字词,如“?”、“re”、“的”等的统计;链接特征是利用论坛中 thread 内部 post 之间的链接关系或者是用户之间的链接关系而挖掘出的特征;关系特征是指 thread 中标题(title)与帖子(post)之间的重叠关系.在质量评估上使用到的技术是一些典型的分类器,如 SVM,Naïve Bayes,C4.5 等.还有一些是链接特征分析技术,如 HITS,Page Rank 等.UGC 的质量评估在 Web2.0 平台上得到了广泛的应用,如网络

论坛、博客以及问答系统.在这些平台上可以针对用户的书评、产品评论、影评进行质量评估.浏览者根据这些高质量高的评论增进了对产品的理解.

Table 1 A taxonomy of quality assessment for UGC

表 1 UGC 质量评估

Category		Description
Tasks	Classes	High quality/low quality
	Level	Post-Level or Thread-level classification
	Source/Target	Whether source/target of quality is known or extracted
Features	Surface	The number of token in a post, the percentage of sentences ending with "?", or the percentage of words in CAPITAL
	Lexical	Spelling error frequency or swear word frequency
	Link based	User authority scores calculated from user-user graph
	Relationship	Overlap between post title and replies, and number of replies corresponding to the title
	Forum specific	Whether or not a post contains HTML
Techniques	Machine learning	Techniques such as SVM, Naïve Bayes, C4.5, etc.
	Link analysis	HITS or Page Rank algorithms
Domains	Reviews	Product, movie and music reviews
	Web discourse	Web forums, Blogs and question-answer community

当前,针对 post 进行质量分析的技术比较多,其中,Weimer 等人[2,3]提出了一个系统,对网络论坛上的 post 进行质量评估.他们首先在 Nabble[4]论坛的 software 板块上搜集数据集,该板块下的所有 post 均被用户打分,打分等级从一颗五角星到五颗五角星.在整理训练集时他们认为,当 post 的平均级别多于 3 颗五角星时,该 post 可被认为是高质量 post.他们提出 post 质量评估系统的目的是,在不需要用户对 post 进行打分的前提下能够自动识别 post 的质量等级.他们首先在用户打分的数据集上挖掘出区别高质量与低质量 post 的统计性特征,如停用词的统计值等,然后在这些挖掘的特征基础上应用 SVM 分类器来分类出高质量的 post.在众多分类器之中,他们认为 SVM 分类器的性能最好,能够实现 89.1%的准确率.Nabble 论坛类似于问答系统(question and answer),用户提出一个问题可以得到很多类型的答案.从各种不同类型的答案中寻求最佳答案,是质量评估系统的任务.通过调研发现,问答系统中问题质量的好坏往往影响到答案的质量水平,因此对问题进行质量评估也是必须的.质量高的问题往往能够得到众多高质量的答案,而质量差的问题却很难得到高质量的答案.我们需要把论坛上的标题(title)与其相对应的 post 综合起来考察质量,以发现高质量的 thread.因为在论坛上,独立的 post 不易理解,只有将其放到 thread 中,意义才完整;并且,如果 post 离开其所在的 thread,则对其质量进行考察会因 title 与 post 之间关系的丢失而失去客观性.网络论坛上每天都有成千上万新的 thread 加入其中,如果用户把全部时间都花在新的 thread 的浏览与判断上,则会消耗他们大量的时间.因此,对 thread 进行质量评估,丢弃那些低质量的 thread,不仅可以帮助用户节约时间,而且可以让用户在极短的时间内找到自己需要的高质量 thread.在高质量 thread 中,用户可以获取对一个问题或是一个观点的详细解释.而这些,从单个的高质量 post 中是获取不到的.

研究者发现,产品评论的质量好坏直接影响到产品的销售业绩,高质量的、积极的评论可以提高产品的销售业绩,并且大量的高质量产品评论可以开拓产品潜在的消费市场[5].众多的产品评论质量参差不齐,顾客要想浏览全部的产品评论需要花费大量的时间.这就需要提出一种自动化技术来鉴别高质量的产品评论,使得这些高质量产品评论可以帮助顾客买到自己需要的实用的产品.Kim 等人[6]利用 SVM 构建质量鉴别系统,该系统可以自动鉴别产品评论的有用性.在构建系统的过程中,研究者发现,评论的长度、用户对评论的打分是产品评论质量评估的重要特征.Kim 等人利用用户对评论的打分情况进行评论的质量鉴别,但在很多的网络论坛上,用户并没有对评论的质量进行打分,这就使得该系统的实用性比较差.

有相当多的学者通过挖掘 post 的内容特征来发现高质量的 post.最早的论文自动打分系统就是其典型的应用[7−11].这些系统使用的内容特征有词性、拼写语法错误、常用词等.这些特征应用在新闻语料上可以获得很好的性能,但是在 Web2.0 环境下效果不好.与其新闻语料相比,网络论坛上的文本内容存在大量噪音.

在基于 post 质量评估系统的特征抽取方面,除了抽取 post 的文本内容特征以及与文本内容相关的一些统计特征妈外,有很多研究工作从论坛中用户与用户之间的关系出发来挖掘权威性用户.找到这些权威性用户就

可以跟踪到他们发表的 post,而这些 post 往往被认为是高质量的 post.很多研究者[12,13]利用 HITS 算法[14]在网络论坛的 user-to-user 网络上寻找权威的用户,他们发现,post 的质量高低与发表该 post 的用户权威度息息相关.用户的权威度可以帮助质量评估系统选出高质量的 post,它可以作为 post 质量评估的一个重要特征.但是,基于该特征建立的 post 质量评估系统存在很高的漏报率与误报率.用户的权威性特征在社会网络领域[15]有广泛的应用,但在 post 质量评测上应用很少,并且效果不好.

一些研究者通过 thread 中 title 与 post 的关系来发现高质量的 post.研究者[16,17]利用 post 与 title 的重叠程度来帮助用户在问答系统中找到问题的最佳答案.他们用到的特征有:答案与问题的重叠度、答案的长度以及问题有多少备选答案.

Liu 等人[16]第一次把问题和答案联系起来评估答案的质量,在这之前的工作都集中在对单独 post 的考察分析上,没有利用问题与答案之间的关系特征.本文旨在发现网络论坛中的高质量 thread.首先,我们分析的对象是 thread 而不是独立的 post,这样不仅可以挖掘出 title 与 post 之间的关系特征,而且也方便用户理解 thread 中的 post.因为把 post 放在 thread 中相当于为其加入了上下文环境,这样更有利于用户正确深刻的理解;其次,我们不仅从内容特征来识别高质量的 thread,而且我们利用 thread 内部 title 与 post,以及 post 与 post 这种树形结构特征来识别高质量的 thread.

2 Thread 结构

图 1 展示了论坛的组成结构.它可以很好地帮助我们了解下面定义的概念.

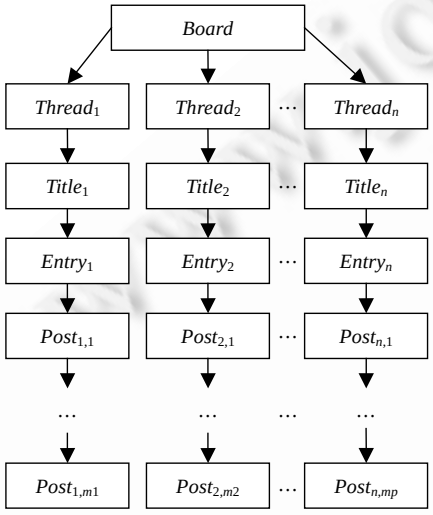

Fig.1 Structure of threads in Web forums

图 1 论坛中 thread 结构图

- 帖子(post):帖子是指作者就某个论题进行论述的文章.它分为两类:主贴与回帖.一个帖子具有 4 个特性:发表时间、作者、论题以及论述内容.
- 论题(title):thread 中的标题,是 thread 中第 1 个作者提出来的,并且在该 thread 中是唯一的,它代表着这个 thread 即将论述的主要内容,以后的 post 都是围绕这个 title 展开论述.一个 thread 内部的所有帖子共享同一个论题.
- 入口(entry):它是 thread 中 title 的详细论述,是由创建 title 的作者对 title 的一个详细解说,每一个 thread 中 entry 也是唯一的.实际上,它是 thread 中的第 1 个 post.
- 主题(thread):thread 由 title,entry,post 组成,并且 title,entry 是唯一的.它包含一系列的 posts,每一个 post 唯一属于一个 thread.

论坛是由很多板块构成的,如在水木社区中有“水木特区”、“个人 show”、“交通信息”等板块.在每一个板块中,文本的组织形式是 thread.每一个板块含有很多 threads,而这些 threads 是由 title,entry 以及 post 等构成的.

3 高质量主题识别框架

面向网络论坛的高质量主题识别框架包括特征抽取、特征选择、分类器.特征抽取是重点,也是难点.特征抽取的目的是发现并找到能够区别高质量主题与低质量主题的特征.网络论坛上的 UGC 含有大量的噪音,如果单纯地依据内容特征来识别高质量主题,效果就会不理想.本文不仅从内容特征入手,而且依据论坛上 thread 的树形结构来挖掘结构性特征,帮助框架识别高质量的主题.在结构特征挖掘上,通过考察 thread 在时间轴上的变化来区别高质量主题与低质量主题.首先,把 thread 的生命周期划分成等时间窗口的时间片,时间片可以是 1 天

或者 1 小时;然后,在每一个时间片上计算该时间段内新增用户数量、新增回帖(reply)数量以及新增回帖的文本长度;最后,把这些时间片段连接起来就形成了 3 个时间序列:user-series(新增用户数量时间序列),reply-series(新增回帖时间序列),size-series(新增回帖文本长度时间序列).每一个 thread 都形成 3 个这样的时间序列,然后在这 3 个时间序列上挖掘结构特征.通过粗略的观察我们发现,高质量主题与低质量主题在时间序列上的幅值以及突发性有很大的不同.大部分高质量主题时间序列上的幅值比较大,并且往往伴随着突发性;而低质量主题往往时间序列上的幅值越小,突发性也不明显.首先,利用小波变换技术来提取时间序列上的能量特征;然后,利用离散点检测技术来提取时间序列上的突发性特征.依据内容特征、能量特征以及突发性特征,框架可以很好地识别高质量主题.图 2 展示了高质量主题识别框架图.

识别框架由源数据预处理、特征抽取、特征选择、分类 4 部分组成.源数据预处理首先从论坛数据中抽取结构性数据:帖子的标题、作者、发表时间、内容以及该帖子所属 thread 的 id 号.这些数据抽取结束之后,对帖子的内容进行分词处理,分词结束之后,删除那些诸如“的”、“顶”、“re”等停用词;对 thread 内部的帖子流进行计算,形成 user-series,reply-series,size-series 这 3 个时间序列.

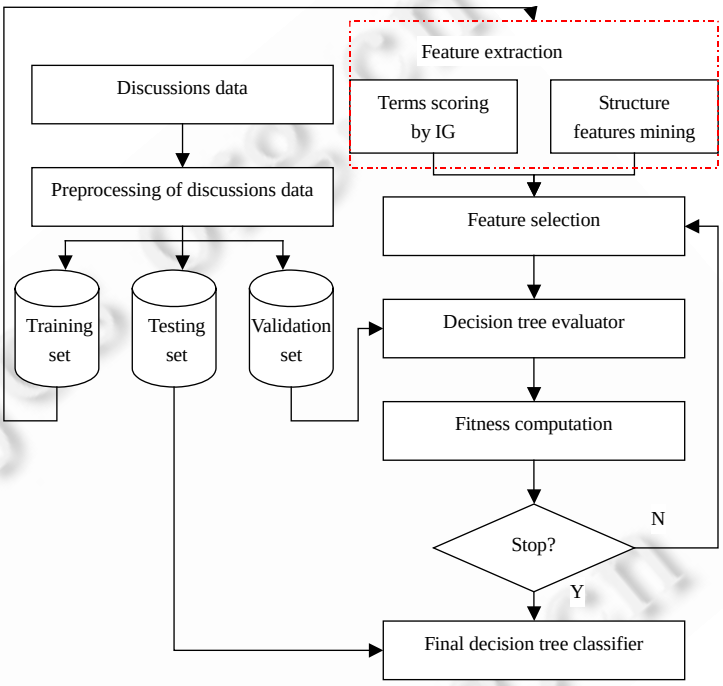

Fig.2 A framework of quality classification in Web forums

图 2 面向网络论坛的质量分类框架

经过预处理之后的源数据集被分成 3 个部分:训练数据集、测试数据集、验证数据集.首先在训练数据集上利用信息增益技术(information gain,简称 IG)抽取内容特征,利用小波变换与离群点检测抽取结构特征;其次,基于抽取后的特征形成分类器,利用分类器在验证数据集上的错误分类率来选择最优的特征集合;最后,在选择后的特征集合上训练分类器,在测试数据集上测试分类器的性能.在特征选择上,当分类器的分类错误率下降到一定的标准之后,就可以停止选择过程,得到的特征子集就是选择的最优特征子集.

3.1 内容特征

对预处理过的文本内容进行分词并且删除停用词之后,在 thread 内部对每个词进行词频统计.然后,利用信息增益理论计算整个词向量空间上每个词的信息增益值,把 thread 内部每个词的信息增益值乘以词频就得到该词在该 thread 中的信息增益值.Thread 内部所有词的信息增益值之和为该 thread 的信息增益值.Thread 的信息增益值可以作为区别 thread 是高质量还是低质量的内容特征.

假设 S 是训练集中的样本,这些样本都被标注成高质量或者低质量.S_i 是训练集中类为 C_i 的样本.s 是训练集中样本总数目.m 是训练集中类的数目,并且训练集中类 C_i 的样本数为 s_i.于是,区分一个样本的期望信息是

$$I(S_1,S_2,...,S_m)=-\sum_{i=1}^{m}\frac{S_i}{S}\log_2\left(\frac{S_i}{S}\right) \tag{1}$$

词 t 贡献的平均信息值为

$$I(S_1,S_2,...,S_m \mid t) = -\frac{s_{it}}{s}\sum_{i=1}^{m}\frac{s_{it}}{s}\log_2\left(\frac{s_{it}}{s}\right) \tag{2}$$

其中,s_{it}是含有词 t 的类别为 C_i 的样本数目.词 t 的信息增益值为

$$G(t)=I(S_1,\ldots,S_m)-I(S_1,\ldots,S_m|t) \tag{3}$$

Yang 等人[18]把词 t 的信息增益值扩展为

$$G(t) = I(S_1,...,S_m) - I(S_1,...,S_m \mid t) - I(S_1,...,S_m \mid \bar{t}) \tag{4}$$

公式(4)计算词 t 的信息增益值不仅考虑了词 t 出现在样本中,同时,词 t 未出现在样本中也被考虑进去,能够更加全面地表达词 t 在分类中的贡献能力.

本文只涉及两个类:高质量与低质量.S_1 为高质量样本,S_2 为低质量样本.本文词 t 的信息增益值计算如下:

$$Gain(t) = \begin{cases} G(t), & \text{if } \frac{s_{1t}}{s} \succ \frac{s_{2t}}{s} \\ -G(t), & \text{otherwise} \end{cases} \tag{5}$$

词向量空间中的词可以按照公式(5)计算得出的值进行排序.如果 thread 中含有大量排序靠前的词,则说明该 thread 被分成高质量 thread 的可能性比较大.在词的信息增益值基础上,可以计算得到 Thread 的信息增益值,其计算如下:

$$Score(thread) = \sum_{j=1}^{N} Gain(t_j) \times f(thread, t_j) \tag{6}$$

其中,$f(thread,t_j)$是词 t_j 在 thread 中出现的频率,N 是 thread 中不同的词的数目.从公式(6)可以看出,thread 的信息增益值越大,说明该 thread 被分成高质量 thread 的概率越大.

3.2 结构特征

通过观察发现,高质量 thread 和低质量 thread 在 thread 内部的 post 流上存在很大的差异性.在观察发现的基础上,我们依据 thread 内部的 post 流创建 3 个时间序列来描述 thread 在时间轴上的变化.Thread 内部 post 流在不同时间片上的新增创建者数目、新增回贴数目、新增回帖的文本长度各不相同.我们通过 3 个时间序列 S_{user},S_{reply},S_{size} 来记录 thread 在所有时间片上的信息.S_{user} 对应 user-series,S_{reply} 对应 reply-series,S_{size} 对应 size-series.通过对时间序列的挖掘,我们可以抽取区别高质量 thread 与低质量 thread 的结构性特征.

图 3 展示了高质量主题与低质量主题在 3 个时间序列上的差异性.

图 3 左侧为高质量主题的 3 个时间序列,横轴单位为小时,即时间片的大小为 1 小时.从图中可以看出,高质量主题的持续时间为 1 807 小时,低质量主题的持续时间为 286 小时.幅值上的差异性在高质量与低质量主题中表现得也很明显.这只是我们给出的一个普通实例,它不能代表所有的高质量主题与低质量主题之间的差别.有些低质量主题的持续时间也很长,并且幅值也很大.因此,持续时间与幅值只能作为区别高质量主题与低质量主题的一个因素,而不是全部因素.在实验中,我们把序列幅值之和与持续时间的商定义为时间序列的流量,如果以流量为标准来区别高质量主题与低质量主题,则当 3 个时间序列的阈值为 8.3KB,36.7KB,12.4KB 时,高质量主题具有最高召回率 58%.即,当 3 个时间序列以流量的阈值作为分割点,大于阈值的为高质量主题,小于阈值的为低质量主题,则最多有 58%的高质量主题可以召回.这说明还有很多高质量主题在 3 个时间序列上的流量值并不大,而这些高质量主题可以利用前面介绍的内容特征或者后面即将介绍的突发性特征来鉴别.流量是一个直观的特征,易于理解,并且它依赖于当前创建的时间序列.为了提高框架的识别能力,我们在时间序列上应用小波包变换技术来抽取小波特征,它能够在时域与频域上对时间序列进行切割,更易于发现高质量主题与低质量主题在细节处的差异性,不同于流量特征的粗略性与概要性.

Thread 中的 post 流在很长时间内流量很小,而在极短的时间内流量很大的这种现象可以称为突发现象.thread 中 post 流的这种突发现象可以作为鉴别高质量主题与低质量主题的一个显著特征.图 4 是一个高质量主题与低质量主题在突发方面的比较图.

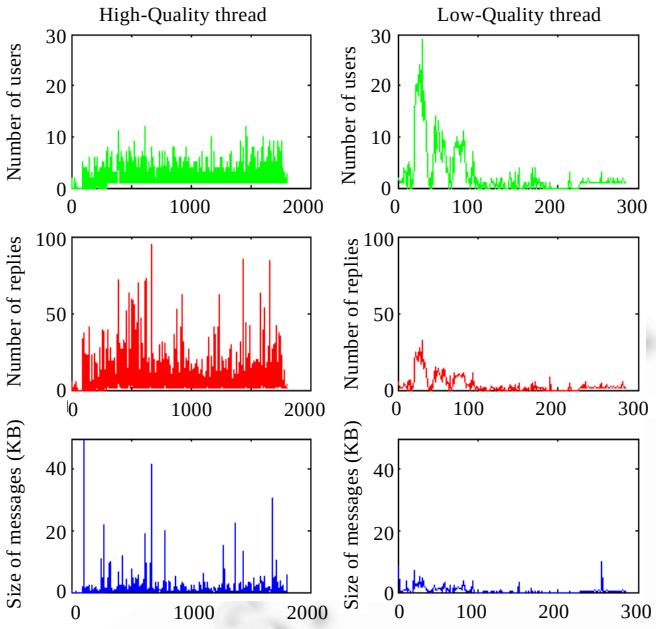

Fig.3 Comparison of three series on amplitude and duration between a high quality thread and a low quality thread

图 3 高质量主题与低质量主题的 3 个时间序列在幅值与持续时间上的比较

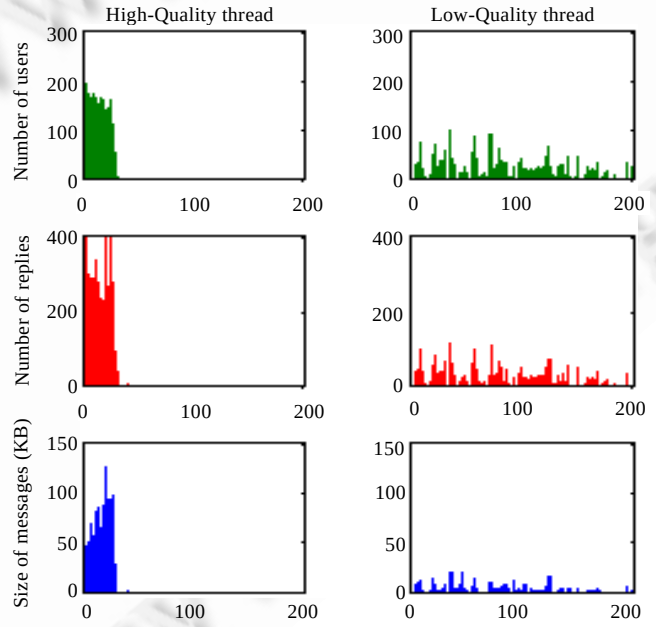

Fig.4 Comparison on bursts between a high quality thread and a low quality one in terms of three time series

图 4 高质量主题与低质量主题在 3 个时间序列上的突发性比较

图 4 中,一个时间片为 1 天.图 4 的高质量 thread 和低质量 thread 也只是我们列出的一个普通例子.是为了说明高质量主题与低质量主题在突发性上的不同:高质量主题突发性强,时间序列曲线跳跃性大;低质量主题突发性弱,曲线相对平滑.突发性只能鉴别部分高质量主题,其中有很多高质量主题不具有突发性,也有很多低质量主题具有突发性.因此,突发性特征只能作为识别框架的一个基本特征,而不是全部特征.要想全面地鉴别高质量主题与低质量主题,还需要综合考虑内容特征与结构特征.在实验中,我们利用突发性检测算法可以计算每个时间序列的突发值.在实验训练数据集上,我们利用阈值法可以召回 62%的高质量主题.其中,在 3 个时间序列上使用的阈值分别为 43.83,41.363 3,50.408.这表明,当主题的 3 个时间序列的突发值大于相应的阈值时,其可以划分为高质量主题;否则为低质量主题.在召回的高质量主题中存在很多低质量主题,这说明有些低质量主题也存在着突发性.

3.2.1 小波变换

小波变换[19,20]是一种信号的时间和尺度分析方法,它不仅继承和发展了短时傅里叶变换的局部化思想,而且克服了时间窗口大小不随频率变化、缺乏离散正交基的缺点,是一种理想的信号处理方法.小波变换较好地解决了时间分辨率和频率分辨率的矛盾.小波变换的窗是可调的时频窗,在高频时使用短窗口,在低频时使用长窗口.即以不同的尺度观察数据,以不同的分辨率分析数据,这充分体现了自适应分辨分析的思想.在低频部分具有较高的频率分辨率,在高频部分具有较高的时间分辨率和较低的频率分辨率,很适合用于捕获信号间的细节差异,被誉为分析信号的显微镜.小波变换的时间复杂度为线性时间 $O(N)$.

利用小波对时间序列进行分解与合成时,有两种方法可以采用.一种方法是用二进制小波变换及其逆变换,这种方法分解得到的两个序列均与原时间序列长度相同,但不存在分解尺度的选择.另一种方法是基于 Mallat 算法的多尺度小波分解与合成[21,22],这种方法每分解一次得到的两个序列均是分解前时间序列长度的一半,通过单支重构可使各分解序列恢复到原序列的长度.这种方法可以得到不同分解尺度下的细节系数与逼近系数,相当于将信号分解成各特定频率范围内的时间信号.本文采用 Mallat 算法.

小波基函数的选择是小波分解的关键,目前已存在几种不同的基小波,如 Haar 小波、Daubechies(dbN)小波系、Meyer 小波和 Symlets 小波.由于不同的小波基在正交性、紧支撑性、平滑性甚至对称性上表现出不同的特性,对于同一信号采用不同的小波基函数,分析效果是不同的.

Daubechies 已证明,具有紧支撑和对称性的正交小波仅有 Haar 小波,但 Haar 小波是不连续的,频域局部性差,结构简单,常用于理论研究中,而且 Haar 小波是消失矩为 1 的 Daubechies 小波.根据 Daubechies 小波具有正交性、视频紧支撑、高正规性和具有 Mallat 快速算法的特点,对于本文的时间序列信号分解与重构具有很好的特性,因此本文采用 Daubechies 小波.

根据 user-series,reply-series 以及 size-series 的特点,我们采用四阶消失矩的 Daubechies 小波:db4.消失矩越大,其支撑长度就越大,通常是支撑长度不少于 $2\times N-1$.其中,N 是消失矩.消失矩越大,对应的滤波器越平坦,而且小波函数的振荡也越强,此时,光滑函数在利用小波展开后的零点也就越多.也就是说,小波消失矩的大小决定了小波逼近光滑信号的能力.

1992 年,Wickerhauser 和 Coifman 等人在小波变换的基础上提出了小波包变换(wavelet packet transform,简称 WPT)的概念,进一步地发展了小波变换的理论.与小波变换相比,小波包变换能够同时对信号的高频和低频部分进行分解处理,为信号提供一种更加精细的分解方法,能够对包含大量中高频信息的信号进行更好的时频局部化分析;并能够根据被分析信号的特征,自适应地选择相应频带,使之与信号频谱相匹配,从而提高了信号的时频分辨率.小波包分解的完全二叉树如图 5 所示,其中,A 代表低频逼近信号,D 代表高频细节信号,数字 1,2,3 表示小波包分解的层数(也即尺度数).原始信号 S 经过小波包分解之后可以表示为

$$AAA3+DAA3+ADA3+DDA3+AAD3+DAD3+ADD3+DDD3.$$

本文采用 db4 对 thread 的 3 个时间序列进行小波包分解与重构.实验中采用 3 层分解,层数的选择是根据经验确定的.根据图 5 可知,信号经小波包进行 3 层分解之后,在第 1 层产生 2 个小波系数,其中一个是低频信号的小波系数,另一个是高频信号的小波系数.在第 2 层产生 4 个小波系数,第 3 层产生 8 个小波系数.原始信号 S

经过3层小波包分解之后形成14个小波包系数:$C_{10},C_{11},C_{20},C_{21},C_{22},C_{23},C_{30},C_{31},C_{32},C_{33},C_{34},C_{35},C_{36},C_{37}$.分解之后,对分解的系数进行小波包重构,形成重构系数.重构后的系数为 $RC_{10},RC_{11},RC_{20},RC_{21},RC_{22},RC_{23},RC_{30},RC_{31},RC_{32},RC_{33},RC_{34},RC_{35},RC_{36},RC_{37}$.例如,$C_{30}$ 被重构成 RC_{30},C_{31} 被重构成 RC_{31}.经过小波包分解与重构之后,原始信号 S 可以表示为

$$\begin{aligned} S &= RC_{10} + RC_{11} \\ &= RC_{20} + RC_{21} + RC_{22} + RC_{23} \\ &= RC_{30} + RC_{31} + RC_{32} + RC_{33} + RC_{34} + RC_{35} + RC_{36} + RC_{37} \end{aligned} \tag{7}$$

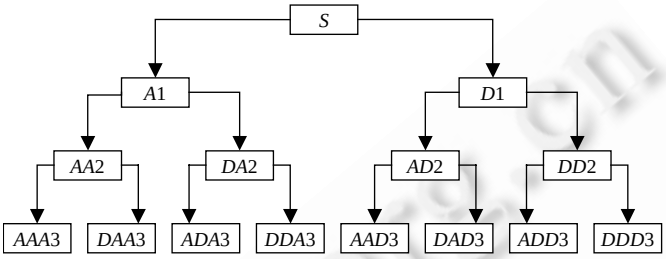

Fig.5 Wavelet packet decomposition fully binary tree

图5 小波包分解完全二叉树

对重构后的每一个小波包系数计算其熵值,如第3层的每个重构系数熵值计算如下:

$$en_{3,j}=wentropy(RC_{3,j},\text{'log}energy\text{'}),\ j=0,1,2,\ldots,7 \tag{8}$$

针对每一层的系数,我们把它们的熵值相加形成一个小波特征.这样,3层系数就可以形成3个小波特征.这3个特征可以用来鉴别高质量主题与低质量主题.针对每个时间序列信号提取的3个小波特征 $wf_i(i=1,2,3)$ 计算如下:

$$wf_i = \begin{cases} \sum_{j=0}^{1} en_{i,j}, & i=1 \\ \sum_{j=0}^{3} en_{i,j}, & i=2 \\ \sum_{j=0}^{7} en_{i,j}, & i=3 \end{cases} \tag{9}$$

下面给出图3中的高质量主题与低质量主题的两个时间序列,即两个 S_{user} 的重构系数图,如图6所示.图中第1行是高质量主题与低质量主题的原始信号,第2行~第9行是第3层的8个重构系数(3,0),(3,1),…,(3,7).

为了便于显示,我们只列出了第3层的8个系数(3,0),(3,1),…,(3,7).从重构系数图上可以看出,高质量主题与低质量主题在第3层的8个重构系数上存在很大的差异性.这些都是直观的、感性的.我们通过每个重构系数的熵值把这些系数量化,通过熵值的差异来突出高质量主题与低质量主题之间的差异性.针对3层小波包重构树,我们利用公式(9)计算每层的熵值和,每一层构成一个熵值特征.这样,一棵小波包重构树就产生3个小波特征 $wf_i(i=1,2,3)$.每一个thread有3个时间序列 $S_{user},S_{reply},S_{size}$,而每一个时间序列均产生一棵小波包重构树,即每一个时间序列对应一组小波特征.S_{user} 产生的3个小波特征记为 $wf(user)$;S_{reply} 产生的3个小波特征记为 $wf(user)$;S_{size} 产生的3个小波特征记为 $wf(size)$.图3中高质量主题与低质量主题在特征 $wf(user),wf(user),wf(size)$ 上的比较见图7($S_{user},S_{reply},S_{size}$ 各自对应一棵小波包树,每棵树有3层).从量化后的这些特征可以清楚地看出高质量主题与低质量主题的差异性.特征图显示的只是图3列出的一个典型例子.实际数据集中存在很多高质量主题与低质量主题通过小波特征是无法鉴别出来的,这就需要我们综合考虑多方面的特征,综合识别高质量主题.所以,除了小波特征以外,我们还设计了内容特征、突发特征来帮助识别高质量主题.我们把每一个主题在3个时间序列 $S_{user},S_{reply},S_{size}$ 上产生的9个小波特征记为 $F_{wavelet}$,则

$$F_{wavelet}=wf(user)\cup wf(reply)\cup wf(size) \quad (10)$$

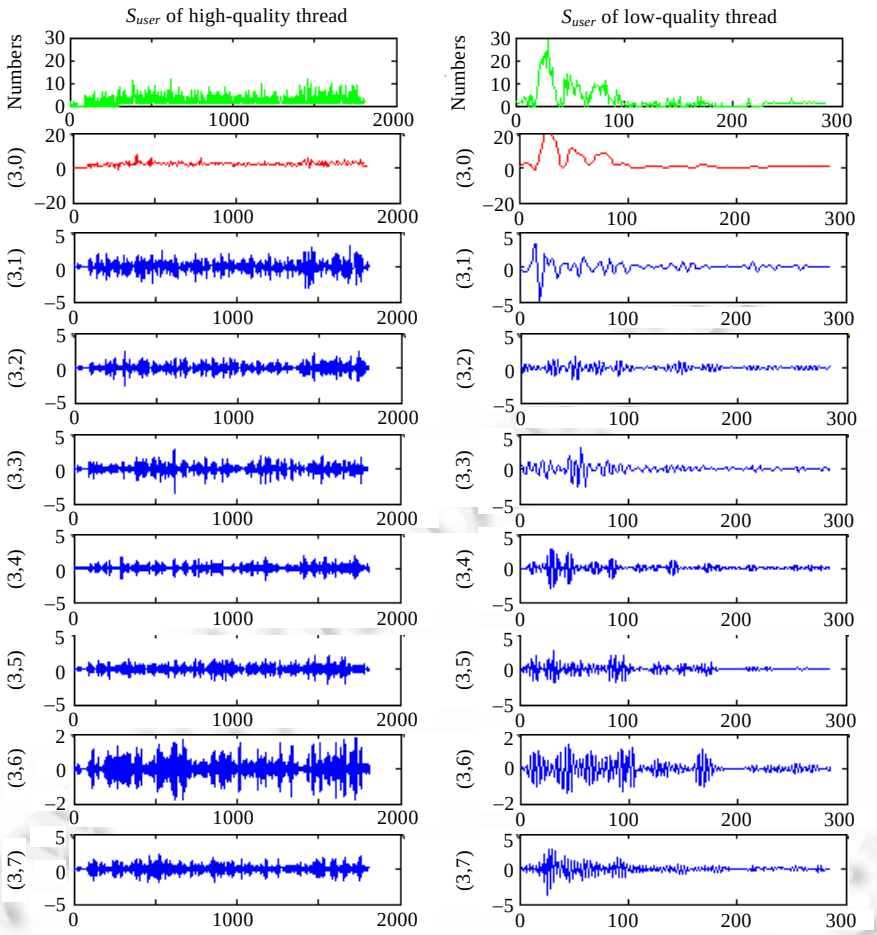

Fig.6 S_{user} and their corresponding reconstructed eight coefficients at level three of wavelet packet tree

图 6 高质量主题与低质量主题在 S_{user} 上进行小波包分解与重构之后形成的第 3 层 8 个系数

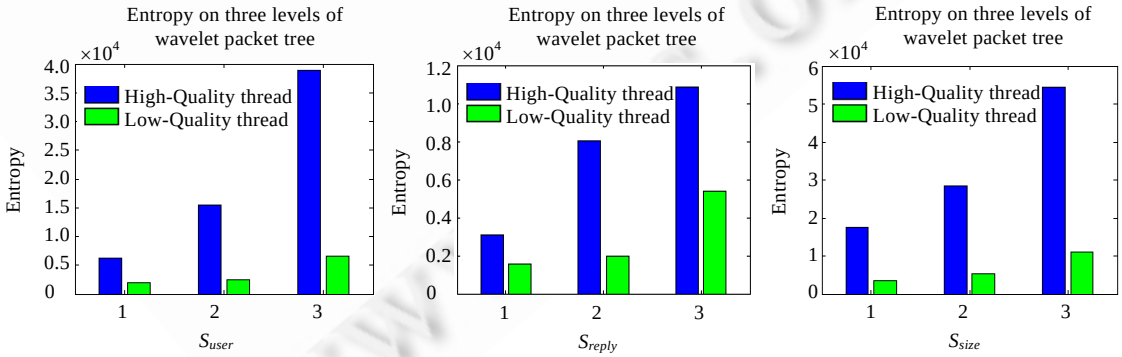

Fig.7 Wavelet features *wf*(*user*), *wf*(*reply*), *wf*(*size*)

图 7 小波特征 *wf*(*user*),*wf*(*reply*),*wf*(*size*)

3.2.2 离群点检测

从图 4 中可以看出,突发性可以作为区别高质量主题与低质量主题的一个重要特征.我们利用突发性特征可以提高高质量主题的识别性能.突发性特征的挖掘有两个关键问题:突发性检测和突发程度的测量.高质量主题与低质量主题的时间序列中均存在突发,如何检测到这些突发是首要问题;每个时间序列中突发点(burst)的值不同,只有计算了突发点的值,我们才能更准确地从量上去区别高质量主题与低质量主题.

突发性检测的应用比较多,突发检测应用最早并且最常用的是网络流检测[23,24].今天,突发性检测被广泛应用于时间序列挖掘与数据挖掘领域[25−27].以上这些突发性检测的应用实例都只是检测突发(burst),并没有给 burst 赋值.为了精确地给 burst 赋值,本文把对突发性检测的问题转移到离群点检测上来:时间序列上的 burst 相当于离群点.

离群点检测是数据挖掘领域的一个重要问题,并且已经被研究很多年了[28−30].离群点检测中的很多方法只是从时间序列中检测离群点,它们并不能比较每个时间序列中离群点的大小[31,32].最典型的离群点检测算法 LOCI[28]能够对时间序列中每一个离群点赋值,但是这种方法是基于局部密度来赋值的,并不是我们所期望的.图 8 显示了 LOCI 方法与我们所期望的离群点检测方法的差异性.我们期望对序列中每一个离群点都精确地计算出一个值,这样才能更准确地去鉴别高质量主题与低质量主题.

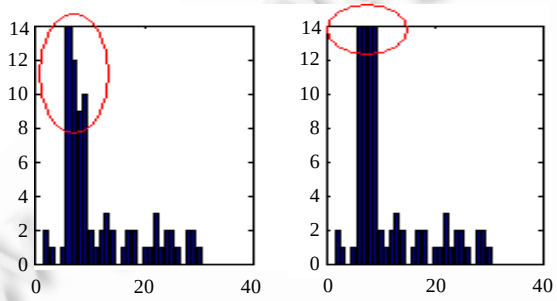

Fig.8 Comparisons on computation of sample grade between LOCI and our wish

图 8 LOCI 方法与我们期望的离群点检测方法在计算样本离群值上的差异性

LOCI 方法认为,图 8 中的左、右两个圈内的每个点的离群值是一样的,而我们期待是不一样的.因为 LOCI 依据局部密度来计算点的离群值,它认为同一密度的点的离群值是一样的.

我们首先把时间序列中的 burst 转换成离群点.转换分两步完成:首先,把时间序列的纵坐标变成横坐标.比如,S_{user} 的横坐标单位是天,纵坐标单位是用户个数,则经过第 1 步转换之后,用户个数就成为横坐标了.也就是说,横坐标上每一个点表示用户(user)的个数,用户的个数可以从 0 到一个很大的值.其次,把时间序列的横坐标转换成纵坐标.比如,S_{user} 的横坐标是天,经过第 2 步转换之后,天就成为纵坐标了.也就是说,纵坐标上每一个点表示天数,天数可以从 0 到一个很大的值.经过两步转换之后,转换后的序列上的每一个点(X,Y)表示:用户数为 X 的天数为 Y.图 4 的转换结果如图 9 所示.

图 9 中的圈为高质量主题的离群点集合,这些离群点对应原序列中的 burst.为了便于显示,我们把横坐标向右移动了 100 个单位.低质量主题中每天的新增用户数比较小,在 100 以内;而高质量主题中新增用户数较大,超过 100,并且新增用户数超过 100 的天数很少.也就是说,在极短的时间内用户数量激增,而在大多数天内,新增用户数量稳定,这在原序列中表现为突发.S_{reply},S_{size} 经转变后,情况也类似.

基于转换后的新时间序列 S'_{user}, S'_{reply},S'_{size} ,我们设计一种离群点检测算法来计算每个时间序列的离群值.每个新时间序列的离群值 $od(user)$,$od(reply)$,$od(size)$作为高质量主题识别的 3 个特征.离群点检测算法设计如图 10 所示.

设 R 是时间序列上点的集合,N 是时间序列上点的总个数,K 是序列上最大的离群点个数.$Ne(r_i,K,N-K)$,$K\prec N/2$ 是点 r_i 的第 K 近邻到第 $N-K$ 近邻点的集合.定义每个点的自距离为 $Self\_Dist[r_i]$,$r_i\in R$.其中,$Self_Dist[r_i]$是 r_i

和 $Ne(r_i,K,N–K)$中所有点的距离之和.$\mu[Ne(r_i,K,N–K)]$和$\sigma[Ne(r_i,K,N–K)]$是 $Ne(r_i,K,N–K)$中所有点的自距离的均值和标准差.根据切比雪夫不等式,序列上每个点的自距离都坚持了统一的标准,这样,不同序列上点的离群值可以进行比较.尽管 K 是指定的,但是 K 在算法中不是敏感参数.在我们的实验中,K 被设置成 $N/4$.这是因为当 K 在 $N/4$ 时,训练集上高质量主题检测可以达到最高的 $F1$ 值 0.62,此时,3 个新序列的离群值阈值分别为 43.83, 41.3633,50.408;而当 K 被设置成其他值时,3 个新序列的离群值阈值无论如何调整都达不到 0.62 的 $F1$ 值.如图 10 所示的算法表明,新序列中的每一个点的离群值都可以被计算出来,这满足了先前我们的期望.$Self\_Dist[r_i]$, $Outlier_val[r_i]$是算法中两个重要的参数,其中,$Self_Dist[r_i]$是绝对距离,$Outlier_val[r_i]$是相对距离.绝对距离突出了同一序列中离群点的值,相对距离突出了不同序列中离群点的值,这样更有利于在不同序列之间进行离群点值的比较.

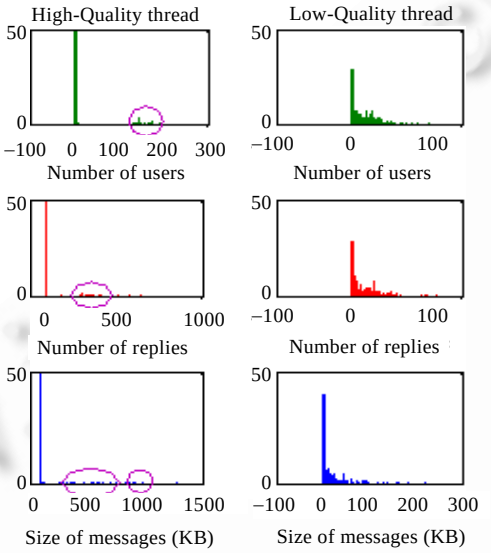

Fig.9 Results of time-to-frequency transformation for S_{user}, S_{reply}, S_{size} in Fig.4

图 9 图 4 中时间序列 S_{user},S_{reply},S_{size} 经过时间频率坐标转换之后形成的新序列

Input: $R=(r_1,r_2,\ldots,r_N)$; //data set
K. //the maximum outlier numbers
Output: $Outlier\_val[r_i]$, $r_i \in R$.
Method:
For each $r_i \in R$ do
Compute $Self\_Dist[r_i]$ // $Self\_Dist[r_i]=\sum_{r_j \in Ne(r_i,K,N-K)}(r_i-r_j)$
End for
For each $r_i \in R$ do
Compute $\sigma[Ne(r_i,K,N–K)]$, $\mu[Ne(r_i,K,N–K)]$

$$Outlier\_val[r_i]=\frac{Self\_Dist[r_i]-\mu[Ne(r_i,K,N-K)]}{\sigma[N(r_i),K,N-K]}$$

End for
End method

Fig.10 Outlier detection algorithm

图 10 离群点检测算法

根据序列中每个点的离群值,我们可以计算序列的离群值.根据切比雪夫不等式,当一个点的离群值是θ时,表明其是离群点的概率不低于$1-\frac{1}{\theta^2}$.在我们的实验中,常数 4 被设置成所有序列的离群点阈值,即点的离群值

大于 4 表明其为离群点,否则不是离群点.我们利用离群块来计算序列的离群值.离群块是指序列中的某个块,其块内每个点都是离群点.离群块的离群值是块内每个点的离群值之和.我们选择有最大离群值的块作为序列的离群值,这样,每一个 thread 的 3 个序列 S'_{user}, S'_{reply}, S'_{size} 的离群值 $od(user)$,$od(reply)$,$od(size)$就计算出来了.

我们的算法优点是计算简单,易于实现,并且能够对每一个离群点精确赋值;缺点是,参数 K 是一个经验参数.虽然参数 K 是经验性的,但它不是一个敏感参数.我们设计的离群点检测算法能够满足我们的期望:能够对每一个离群点精确赋值;而 LOCI 根据密度来计算离群点的值,得出的利群值是粗略的.在实验中,对 LOCI 提取的特征和我们的离群点检测算法提取的特征进行了对比,实验结果见表 2(训练集规模为 100~5 000).其中,LOCI 提取的突发特征为 $F'_{outlier}$,我们的算法提取的突发特征为 $F_{outlier}$;实验数据集来源于第 4.1 节的实验数据.

Table 2 Average precision, recall, $F1$, and accuracy of $F'_{outlier}$-C4.5, $F_{outlier}$-C4.5 for varying amount of training data

表 2 不同训练集上训练的分类器 $F'_{outlier}$-C4.5,$F_{outlier}$-C4.5 的平均准确率、平均召回率、平均 $F1$ 以及平均正确率

Classifier	Precision	Recall	$F1$	Accuracy
$F'_{outlier}$-C4.5	0.37	0.46	0.41	0.34
$F_{outlier}$-C4.5	**0.58**	**0.62**	**0.60**	**0.56**

从表 2 的对比结果可以看出,我们的离群点检测算法提取的突发特征比 LOCI 提取的突发特征在高质量主题发现中具有更高的准确率、召回率、$F1$ 以及正确率.实验结果表明,针对每个离群点精确赋值比基于密度的离群点赋值更能满足高质量主题发现的需要.

结合上面的内容特征 $Score(thread)$、小波特征,$F_{wavelet}$ 就形成高质量主题识别的最终特征 F_{last}.

F_{last} 描述如下:

$$F_{last}=Score(thread)\cup F_{wavelet}\cup F_{outlier} \tag{11}$$

其中,$F_{outlier}=od(user)\cup od(reply)\cup od(size)$.

3.3 特征选择与分类器

特征选择的目的是对特征集合中每一个特征的重要性进行评估,剔除特征集合中那些会降低分类器性能的特征.特征选择的方法有两类:基于过滤器的特征选择方法和基于封装器的特征选择方法.其中,基于封装器的特征选择方法选出的特征比过滤器要好[33].本文采用一种遗传算法(genetic algorithm,简称 GA)和禁忌搜索(Tabu search,简称 TS)相结合的混合搜索策略 GATS[34].

很多传统分类器可用于我们的框架,如决策树、Naïve Bayes、SVMs、最大熵、k-NN 等.在这些分类器中,我们选择 3 个有代表性的分类器决策树、Naïve Bayes、SVMs 进行实验.利用 C4.5 算法构建的决策树可以使分类达到很好的效果,Naïve Bayes 可以快速地鉴定框架的好坏,SVMs 被公认为在很多任务上可以达到很好效果的分类器.

4 实验与评估

我们进行了几组相关的实验,验证上面抽取的内容特征、结构特征在高质量主题识别上的性能.为了说明建立在内容特征和结构特征上的分类器的分类性能,我们利用下面 4 个参数[35]:

(1) 正确率(accuracy): $\dfrac{\text{被正确分类成高质量主题与低质量主题的样本总数目}}{\text{数据集中总的样本数}}$;

(2) 准确率(precision,简称 P): $\dfrac{\text{被正确分类成高质量主题样本数目}}{\text{被分成高质量主题的样本数目}}$;

(3) 召回率(recall,简称 R): $\dfrac{\text{被正确分类成高质量主题样本数目}}{\text{数据集中高质量主题的样本数目}}$;

(4) $F1$: $\frac{2PR}{P+R}$.

本文所有的实验都是在 Windows 平台上运行的.机器配置为:Dual-Core AMD Opteron™ 处理器 2214HE, 2.21GHz,8.0GB RAM.

4.1 实验数据

我们的实验数据来源于腾讯论坛[36,37].腾讯论坛是一个公共的大型网络论坛,有将近 2.3 亿的用户群,内容覆盖领域广.为了实验的需要,我们从论坛中选取 6 000 个主题,其中 5 000 个作为训练集,剩下 1 000 作为测试集.6 000 个主题的统计信息见表 3.6 000 个主题包含 135 625 个回帖,有 7 642 个用户参与讨论.数据集中所有帖子总的大小为 30 382 574Byte.我们对 6 000 个主题进行高质量与低质量的标注,标注过程比较复杂,并且标注出来的结果并不能满足所有用户的期望,只能按照满足大多数用户的期望来给出标注结果.首先,我们给出高质量主题的一些具体特征;然后,请 5 位用户根据这些具体特征对 6 000 个主题进行标注.高质量主题的具体特征有:主题描述的是一个完整的事件,主题中包含的信息冗余度低,信息量大;主题中回帖内容能够很好地解决主题中主贴提出的问题;主题中帖子的语言描述规范,内容易于理解.针对这些特征,每一个主题均被 5 个用户标注,标注结果只有两种:高质量主题或者低质量主题,并且不可以既是高质量又是低质量.根据多数战胜少数理论,我们定义:如果一个主题有超过 60%的高质量标注结果,则其可被认为是高质量主题;否则为低质量主题.表 4 列出了 6 000 个主题的标注情况.

Table 3 Statistics of experimental data

表 3 实验数据统计信息

Name	Value
Number of threads	6 000
Number of reply posts	135 625
Number of authors	7 642
Size of posts (Byte)	30 382 574

Table 4 Distribution of labels on 6 000 threads

表 4 6 000 个主题的标注结果

Categories of high quality threads	Number of threads
High quality threads with less than 20% votes	3 126 (52.1%)
High quality threads with 40% votes	542 (9%)
High quality threads with 60% votes	234 (3.9%)
High quality threads with 80% votes	475 (7.9%)
High quality threads with all votes	1 623 (27%)

4.2 实验评估

我们提出了高质量主题识别框架.该框架包括特征抽取、特征选择、分类这 3 个主要部分.为了验证框架在正确率、准确率以及召回率上的性能,我们进行了几组相关的实验.首先把训练数据集划分成 8 个部分,每个部分的样本数量为 100,200,500,1 000,2 000,3 000,4 000,5 000.对于每个部分的样本,我们首先训练 3 个分类器:C4.5,SVM,Naïve Bayes,然后在测试集上测试 3 类分类器的性能.实验结果如图 11 所示.图中的 3 种分类器 F_{last}-C4.5,F_{last}-SVM,F_{last}-Naïve Bayes 都是在抽取的内容特征、小波特征以及突发特征上建立的分类器.其中,F_{last} 是分类器的输入特征,且 $F_{last}=Score(thread)\cup F_{wavelet}\cup F_{outlier}$.从图 11 可以看出,$F_{last}$-C4.5 具有最好的性能,特别是在训练集规模达到 500 时 $F1$ 就已经达到 0.94;并且随着训练集规模的扩大,$F1$ 的性能始终保持在 0.94 左右.这说明 500 规模的训练集对于 F_{last}-C4.5 分类器已经足够了,多余的训练样本对于其性能的提高没有帮助.表 5 列出了 3 种分类器在测试集上的平均分类正确率、准确率、召回率以及 $F1$.从表 5 中可以看出,F_{last}-C4.5 具有最高的 $F1$ 值 0.946.在 3 种类别的分类器中,决策树分类器具有最好的性能.Naïve Bayes 分类器虽然召回率较高,但是其准确率较低,即在召回高质量主题的过程中把很多低质量主题也当作高质量主题召回,这在实际应用中是不适用的.提取高质量主题的目的就是为了帮助用户节约浏览主题的时间,让他们不需要去浏览论坛中每一个主题,把低质量主题剔除浏览的范围,而只是去关注质量高的主题.但是,Naïve Bayes 分类器虽然保证了高质量主题被遗漏的概率很小,但是召回过多的低质量主题会影响用户浏览的速度与质量.SVM 分类器性能介于 C4.5 与 Naïve Bayes 分类器之间,但是在速度上,C4.5 与 SVM 相比有绝对的优势.表 6 列出了 3 种分类器的平均建模时间与平均测试时间(训练集规模为 100~5 000).从表中可以看出,SVM 分类器的训练时间与测试时间较长,而 C4.5 分类器与 Naïve Bayes 分类器的训练时间与测试时间相当.

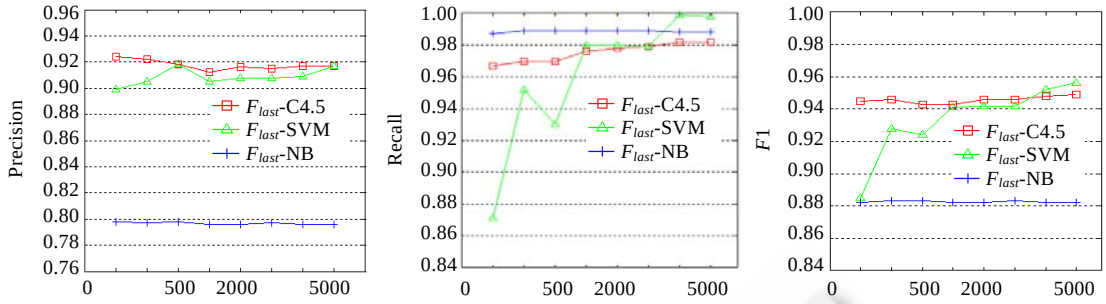

Fig.11 Precision, recall and $F1$ of F_{last}-C4.5, F_{last}-SVM F_{last}-NB for varying amount of training data

图 11 3 种分类器 F_{last}-C4.5,F_{last}-SVM,F_{last}-NB 在不同训练集上的分类准确率、召回率、$F1$

Table 5 Average precision, recall, $F1$, and accuracy of F_{last}-C4.5, F_{last}-SVM, F_{last}-NB for varying amount of training data

表 5 在不同训练集上训练的分类器 F_{last}-C4.5,F_{last}-SVM,F_{last}-NB 的平均准确率、平均召回率、平均 $F1$ 以及平均正确率

Classifier	Precision	Recall	$F1$	Accuracy
F_{last}-C4.5	**0.918**	0.976	**0.946**	**0.899**
F_{last}-SVM	0.908	0.961	0.934	0.877
F_{last}-NB	0.797	0.988	0.882	0.795

Table 6 Average building time and testing time for three different classifiers: F_{last}-C4.5, F_{last}-SVM, F_{last}-NB

表 6 3 种不同分类器 F_{last}-C4.5,F_{last}-SVM,F_{last}-NB 的平均建模时间与检测时间

Classifier	Building time (ms)	Testing time (ms)
F_{last}-C4.5	450	78
F_{last}-SVM	36 000	8 000
F_{last}-NB	590	96

4.3 特征选择

特征选择是为了验证抽取的特征之中,哪些特征对于主题的质量评估起到关键作用.GATS-C4.5 选取的前 6 个特征见表 7.

Table 7 Top six features selected by GATS-C4.5

表 7 特征选择算法 GATS-C4.5 选取的前 6 个特征

Feature
The third feature of *wf*(*user*)
od(*user*)
Score(*thread*)
od(*size*)
The second feature of *wf*(*reply*)
od(*reply*)

从表 7 可以看出,突发性特征 *od*(*user*),*od*(*size*),*od*(*reply*)全部被选中,这说明突发性是主题质量评估的一个关键因素.由此我们可以看出,高质量主题与突发总是存在着一定的关系,高质量主题时间序列中存在突发的概率大.前面介绍突发性理论部分我们就给出了一个事实:仅利用突发性特征作为高质量与低质量主题的判断标准时,62%的高质量主题可以被召回,并且在召回的高质量主题中仅有相当少量的低质量主题,准确率达到 93%.这从另一方面说明大部分低质量主题中含有很少的突发,特别是大突发.在前 6 个特征中,小波特征 *wf*(*user*),*wf*(*reply*)也被选中.我们知道,在 *wf*(*user*)中有 3 个特征,每个特征对应小波包树每一层的熵值和.*wf*(*user*)的第 3 个特征即为小波包树第 3 层上所有系数的熵值和.从选取的小波包特征可以看出,小波特征中大部分特

征被过滤掉,这说明小波特征中存在很大的冗余.*wf*(*size*)中的 3 个特征没有一个被选中,这就是典型的冗余.发现冗余的小波特征可以帮助我们在抽取特征时节约很大的计算开销.例如,我们在抽取小波特征时就可以不考虑时间序列 S_{size},这对于前面特征抽取的工作具有很强的指导意义.

为了比较基于选择后的特征建立的分类器与基于所有特征建立的分类器在性能上的差异,我们进行了几组深入的实验.实验中,我们利用上一节中性能最好的分类器 C4.5 建立基于选择后特征的分类器 F_{select}-C4.5 和基于所有特征的分类器 F_{last}-C4.5.实验依然在训练集的 8 个部分进行,实验详细结果如表 8(训练集规模为 100~5 000)、表 9、图 12 所示.同时,我们从相关工作中学习研究者提出的一些特征,以此作为基准特征与我们的抽取的特征进行了对比.基准特征包括:统计性特征[3],如主题中总的用户数、总回复数、帖子的平均长度;流量特征,如主题生命周期中的平均用户数、平均回复数;关系特征[18],如主贴与回帖之间的重复度.

Table 8 Average precision, recall, *F*1, and accuracy of F_{select}-C4.5, F_{last}-C4.5, BASELINE-C4.5 for varying amount of training data

表 8 在不同训练集上训练的分类器 F_{select}-C4.5,F_{last}-C4.5,BASELINE-C4.5 的平均准确率、平均召回率、平均 *F*1 以及平均正确率

Method	Precision	Recall	*F*1	Accuracy
F_{select}-C4.5	0.911	**0.99**	0.945	**0.903**
F_{last}-C4.5	0.918	0.976	0.946	0.899
BASELINE-C4.5	0.705	0.735	0.702	0.73

Table 9 Average building time and testing time for three different classifiers: F_{select}-C4.5, F_{last}-C4.5, BASELINE-C4.5

表 9 3 种不同分类器 F_{select}-C4.5,F_{last}-C4.5,BASELINE-C4.5 的平均建模时间与检测时间

Classifier	Building time (ms)	Testing time (ms)
F_{last}-C4.5	450	78
F_{select}-C4.5	380	77
BASELINE-C4.5	350	78

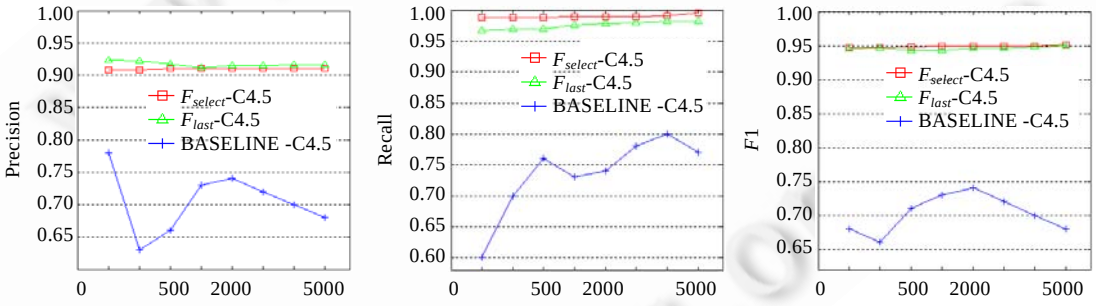

Fig.12 Precision, recall and *F*1 of F_{select}-C4.5, F_{last}-C4.5, BASELINE-C4.5 for varying amount of training data

图 12 3 种分类器 F_{select}-C4.5,F_{last}-C4.5,BASELINE-C4.5 在不同训练集上的准确率、召回率、*F*1

表 8 表明,基于选择后的特征建立的分类器,其性能与基于所有特征的分类器相当,甚至在某些性能上还有提高,如召回率从 97.6%提高到 99%,正确率(accuracy)从 89.9%提高到 90.3%.这表明,用较少的特征就可以达到与全部特征同样的效果,并且在某些性能上还有提高.无论是否经过特征选择,我们抽取的特征在性能上都比基准特征高出很多.表 9 表明,C4.5 分类器在不同特征子集上的建模时间与测试时间基本相同,这是因为特征子集空间 F_{last},F_{select},BASELINE 特征规模相当.表 8 和图 12 重点突出两点:首先,我们提出的高质量主题识别框架具有很好的性能,特别是基于抽取特征的分类器与基于基准特征的分类器相比,在 *F*1 上有很大的提高,从 70.2%~94.5%;其次,我们提出的框架可以在规模很小的训练集实现很高的正确率和 *F*1,而基于基准特征的分类器依赖于训练集的规模,并且性能不高.

4.4 Blog数据集上的评估

高质量主题识别框架是在网络论坛的平台下提出的,但是它适用于 Web2.0 环境下的各种平台,如 Blog 等.我们在 BLOG06[38]数据集上对高质量主题识别框架进行了评测.首先,利用特征抽取技术抽取包括内容特征、小波特征、突发特征等 13 个特征;然后,在这 13 个特征之中选取前 5 个重要特征;最后,针对 13 个特征和选择的 5 个特征分别建立分类器 F_{last}-C4.5,F_{select}-C4.5.选择的前 5 个特征见表 10.从表 10 中可以看出,小波特征 *wf*(*size*)排在最前面,这与网络论坛恰恰相反.网络论坛上的主题文本长度并不重要,基本上可以忽略.这说明,Blog 数据集中高质量主题与主题的文本长度紧密相关,那些经过详细描述的主题是高质量主题的概率较大.

Table 10 Top five features selected by GATS-C4.5

表 10 GATS-C4.5 特征选择算法选择的前 5 个特征

Feature
The first feature of *wf*(*size*)
od(*size*)
od(*user*)
Score(*thread*)
The third feature of *wf*(*reply*)

BLOG06 数据集是格拉斯哥(Glasgow)大学创建和发布的一个 TREC 数据集,我们从此数据集中提取 6 000 个主题,其中 5 000 作为训练集,余下的 1 000 作为测试集.在训练集中,把 5 000 个训练集划分成 8 个部分:100,200,500,1 000,2 000,3 000,4 000,5 000.对每一个部分的训练集都建立分类器 F_{last}-C4.5,F_{select}-C4.5,然后在测试集上测试这些分类器的性能.测试结果如表 11(训练集规模为 100~5 000)、表 12、图 13 所示.从图表中可以看出,两种分类器在 Blog 数据集上的性能都很好,F_{select}-C4.5 可以实现 79%的正确率和 88%的 $F1$ 值.我们提出的高质量主题识别框架同样适用于 Blog 数据集,特别是提出的特征抽取方法、特征选择方法具有通用性.

Table 11 Average precision, recall, $F1$, and accuracy of F_{select}-C4.5, F_{last}-C4.5 for varying amount of training data

表 11 在不同训练集上训练的分类器 F_{select}-C4.5,F_{last}-C4.5 的平均准确率、平均召回率、平均 $F1$ 以及平均正确率

Method	Precision	Recall	$F1$	Accuracy
F_{select}-C4.5	0.795	**0.99**	**0.883**	**0.795**
F_{last}-C4.5	0.80	0.971	0.88	0.793

Table 12 Average building time and testing time for two different classifiers

表 12 两种不同分类器的平均建模时间与检测时间

Classifier	Building time (ms)	Testing time (ms)
F_{last}-C4.5	450	78
F_{select}-C4.5	368	76

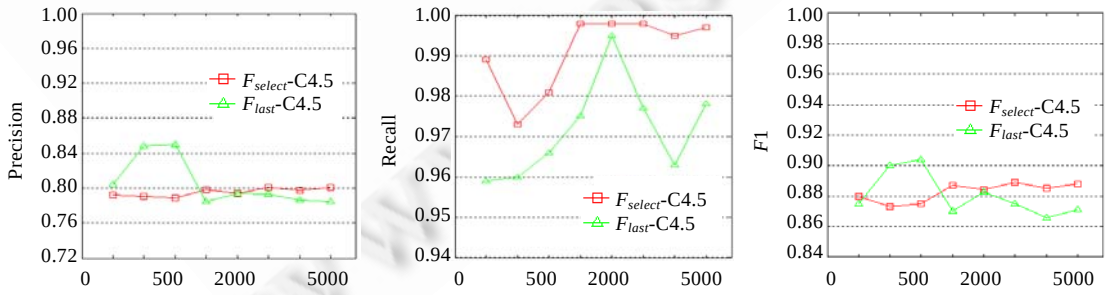

Fig.13 Precision, recall and $F1$ of F_{select}-C4.5, F_{last}-C4.5 for varying amount of training data

图 13 两种分类器 F_{select}-C4.5,F_{last}-C4.5 在不同训练集上的准确率、召回率、$F1$

5 结 论

本文提出一种通用的高质量主题识别框架,它包括特征抽取、特征选择、分类这 3 个主要部分.首先,在特征抽取上从 3 个方面对主题进行挖掘,抽取内容特征、小波特征以及突发特征;其次,针对抽取的特征模型采用特征选择算法选出重要特征;最后,在抽取的特征上建立分类器识别高质量主题.我们在腾讯论坛数据集以及 BLOG06 数据集上进行了大量的实验,实验结果表明,模型抽取的特征具有很高的性能,能够帮助分类器高效而准确地识别高质量主题.在识别高质量主题实验中,与基准特征 73%的正确率相比,基于抽取特征的分类器具有 90.3%的正确率,正确率提高了 17.3%.

高质量主题与低质量主题之间的差异性表现在很多方面,本文抽取的特征只是这些差异性的冰山一角.在未来的工作中,我们计划从其他方面,如主题内部构成的用户社区,来挖掘区别高质量主题与低质量主题的差异性特征.同时,在本文工作的基础上去挖掘高质量话题以及突发话题.

References:

[1] Agichtein E, Castillo C, Donato D, Gionis A, Mishne G. Finding high-quality content in social media. In: Proc. of the 1st ACM Int'l Conf. on Web Search and Data Mining. Palo Alto: ACM, 2008. 183−193. [doi: 10.1145/1341531.1341557]

[2] Weimer M, Gurevych I. Predicting the perceived quality of Web forums posts. In: Proc. of the Conf. on Recent Advances in Natural Language Processing. 2007.

[3] Weimer M, Gurevych I, Mühlhäuser M. Automatically assessing the post quality in online discussions on software. In: Proc. of the ACL 2007 Demo and Poster Sessions. 2007. 125−128.

[4] http://www.nable.com

[5] Feng DH, Shaw E, Kim J, Hovy E. Learning to detect conversation focus of threaded discussions. In: Proc. of the Human Language Technology Conf. of the North American Chapter of the ACL. New York: Association for Computational Linguistics, 2006. 208−215. [doi: 10.3115/1220835.1220862]

[6] Kim SM, Pantel P, Chklovski T, Pennacchiotti M. Automatically assessing review helpfulness. In: Proc. of the 2006 Conf. on Empirical Methods in Natural Language Processing. Sydney: Association for Computational Linguistics, 2006. 423−430.

[7] Page EB. Computer grading of student prose, using modern concepts and software. Journal of Experimental Education, 1994,62(2): 127−137. [doi: 10.1080/00220973.1994.9943835]

[8] Rudner LM, Liang T. Automated essay scoring using Bayes' theorem. Journal of Technology, Learning, and Assessment, 2002, 1(2):1−22.

[9] Burstein J, Wolska M. Toward evaluation of writing style: Finding overly repetitive word use in student essays. In: Proc. of the 10th Conf. on European Chapter of the Association for Computational Linguistics. Morristown, 2003. 35−42. [doi: 10.3115/1067807.1067814]

[10] Attali Y, Burstein J. Automated essay scoring with *e*-rater v.2. Journal of Technology, Learning, and Assessment, 2006,4(3): 21−41.

[11] Kim J, Shaw E, Feng DH, Beal C, Hovy E. Modeling and assessing student activities in on-line discussions. In: Proc. of the Workshop on Education Data Mining at AAAI. Boston, 2006. 12−18.

[12] Zhang JM, Ackerman MS, Adamic L. Expertise networks in online communities: Structure and algorithms. In: Proc. of the 16th Int'l Conf. on World Wide Web. Banff: ACM Press, 2007. 221−230. [doi: 10.1145/1242572.1242603]

[13] Campbell CS, Maglio PP, Cozzi A, Dom B. Expertise identification using email communications. In: Proc. of the CIKM. New Orleans: ACM Press, 2003. 528−531. [doi: 10.1145/956863.956965]

[14] Kleinberg JM. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999,46(5):604−632. [doi: 10.1145/324133.324140]

[15] Nakajima S, Tatemura J, Hino Y, Hara Y, Tanaka K. Discovering important Bloggers based on analyzing Blog threads. In: Proc. of the WWW Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Chiba: ACM Press, 2005. 22−28.

[16] Liu YD, Bian J, Agichtein E. Predicting information seeker satisfaction in community question answering. In: Proc. of the 31st Annual Int'l ACM SIGIR Conf. Singapore: ACM Press, 2008. 483−490. [doi: 10.1145/1390334.1390417]

[17] Lee CW, Day MY, Sung CL, Lee YH, Jiang TJ, Wu CW, Shih CW, Chen YR, Hsu WL. Boosting Chinese question answering with two lightweight methods: ABSPs and SCO-QAT. ACM Trans. on Asian Language Information Processing, 2008,7(4):1−39. [doi: 10.1145/1450295.1450297]

[18] Yang YM, Pedersen JO. A comparative study on feature selection in text categorization. In: Proc. of the 14th Int'l Conf. on Machine Learning. San Francisco, 1997. 412−420.

[19] Luo Y, Cheng LZ, Chen B, Wu Y. Study on digital elevation mode data watermark via integer wavelets. Journal of Software, 2005, 16(6):1096−1103 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/16/1096.htm [doi: 10.1360/jos161096]

[20] He YX, Cao Q, Liu T, Han Y, Xiong Q. A low-rate DoS detection method based on feature extraction using wavelet transform. Journal of Software, 2009,20(4):930−941 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3302.htm [doi: 10.3724/SP.J.1001.2009.03302]

[21] Mallat SG. A theory for multiresolution signal decomposition: The wavelet representation. IEEE Tran. on Pattern Analysis and Machine Intelligence, 1989,11(7):674−693. [doi: 10.1109/34.192463]

[22] Zhong S, Shi QY, Cheng MD. Multiscale stereo vislon based on wavelet transform. Journal of Software, 1995,6(11):281−291 (in Chinese with English abstract).

[23] Mallat S, Hwang WL. Singularity detection and processing with wavelets. IEEE Trans. on Information Theory, 1992,38:617−643. [doi: 10.1109/18.119727]

[24] Innanen, KA. Local signal regularity and lipschitz exponents as a means to estimate Q. Journal of Seismic Exploration, 2003,12: 53−74.

[25] Zhu YY, Shasha D. Efficient elastic burst detection in data streams. In: Proc. of the 9th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Washington, 2003. 24−27. [doi: 10.1145/956750.956789]

[26] Wang XH, Zhai CX, Hu X, Sproat R. Mining correlated bursty topic patterns from coordinated text streams. In: Proc. of the KDD. San Jose: ACM Press, 2007. 52−61. [doi: 10.1145/1281192.1281276]

[27] Kleinberg J. Bursty and hierarchical structure in streams. In: Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2002. 1−25. [doi: 10.1145/775047.775061]

[28] Papadimitriou S, Kitawaga H, Gibbons PB, Faloutsos C. LOCI: Fast outlier detection using the local correlation integral. In: Proc. of the 19th Int'l Conf. on Data Engineering. 2003. 315−326. [doi: 10.1109/ICDE.2003.1260802]

[29] Hodge V, Austin J. A survey of outlier detection methodologies. Artificial Intelligence Review, 2004,22(2):85−126. [doi: 10.1023/B:AIRE.0000045502.10941.a9]

[30] Yang J, Zhong N, Yao YY, Wang J. Local peculiarity factor and its application in outlier detection. In: Proc. of the KDD. Las Vegas: ACM Press, 2008. 776−784. [doi: 10.1145/1401890.1401983]

[31] Wei L, Gong XQ, Qian WN, Zhou AY. Finding outliers in high-dimensional space. Journal of Software, 2002,13(2):280−290 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/13/280.htm

[32] Zhou XY, Sun ZH, Zhang BL, Yang YD. A fast outlier detection algorithm for high dimensional categorical data streams. Journal of Software, 2007,18(4):933−942 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18/933.htm [doi: 10.1360/jos180933]

[33] Chen Y, Cheng XQ, Li Y, Dai L. Lightweight intrusion detection system based on feature selection. Journal of Software, 2007, 18(7):1639−1651 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18/1639.htm [doi: 10.1360/jos181639]

[34] Chen Y, Dai L, Cheng XQi. GATS-C4.5: An algorithm for optimizing features in flow classification. In: Proc. of the IEEE Consumer Communications and Networking Conf. Las Vegas: IEEE, 2008. 466−470.

[35] Williams N, Zander S, Armitage G. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. Proc. of ACM SIGCOMM, 2006,135(5):7−15. [doi: 10.1145/1163593.1163596]

[36] Tencent Web forums. http://bbs.qq.com/

[37] Chen Y, Cheng XQ, Huang YL. A wavelet-based model to recognize high-quality topics on Web forum. In: Proc. of the IEEE/WIC/ ACM Int'l Conf. on Web Intelligence and Intelligent Agent Technology. Sydney: IEEE, 2008. 343−351. [doi: 10.1109/WIIAT.2008.17]

[38] Macdonald C, Ounis I. The TREC Blogs06 collection: Creating and analysing a Blog test collection. Technical Report, TR-2006-224, Glasgow: University of Glasgow, 2006. 1–8.

附中文参考文献:

[19] 罗永,成礼智,陈波,吴翊.数字高程模型数据整数小波水印算法.软件学报,2005,16(6):1096−1103. http://www.jos.org.cn/1000-9825/16/1096.htm [doi: 10.1360/jos161096]

[20] 何炎祥,曹强,刘陶,韩奕,熊琦.一种基于小波特征提取的低速率 DoS 检测方法.软件学报,2009,20(4):930−941. http://www.jos.org.cn/1000-9825/3302.htm [doi: 10.3724/SP.J.1001.2009.03302]

[22] 钟声,石青云,程民德.基于小波变换的多尺度立体视觉方法.软件学报,1995,6(11):281−291.

[31] 魏藜,宫学庆,钱卫宁,周傲英.高维空间中的离群点发现.软件学报,2002,13(2):280−290. http://www.jos.org.cn/1000-9825/13/280.htm

[32] 周晓云,孙志挥,张柏礼,杨宜东.高维类别属性数据流离群点快速检测算法.软件学报,2007,18(4):933−942. http://www.jos.org.cn/1000-9825/18/933.htm [doi: 10.1360/jos180933]

[33] 陈友,程学旗,李洋,戴磊.基于特征选择的轻量级入侵检测系统.软件学报,2007,18(7):1639−1651. http://www.jos.org.cn/1000-9825/18/1639.htm [doi: 10.1360/jos181639]

陈友(1981－),男,安徽安庆人,博士,主要研究领域为数据挖掘,互联网挖掘,信息安全.

杨森(1983－),男,硕士,主要研究领域为网络数据挖掘.

程学旗(1971－),男,博士,研究员,博士生导师,CCF 高级会员,主要研究领域为网络科学与社会计算,互联网搜索与挖掘,网络信息安全,分布式系统与大型仿真平台.