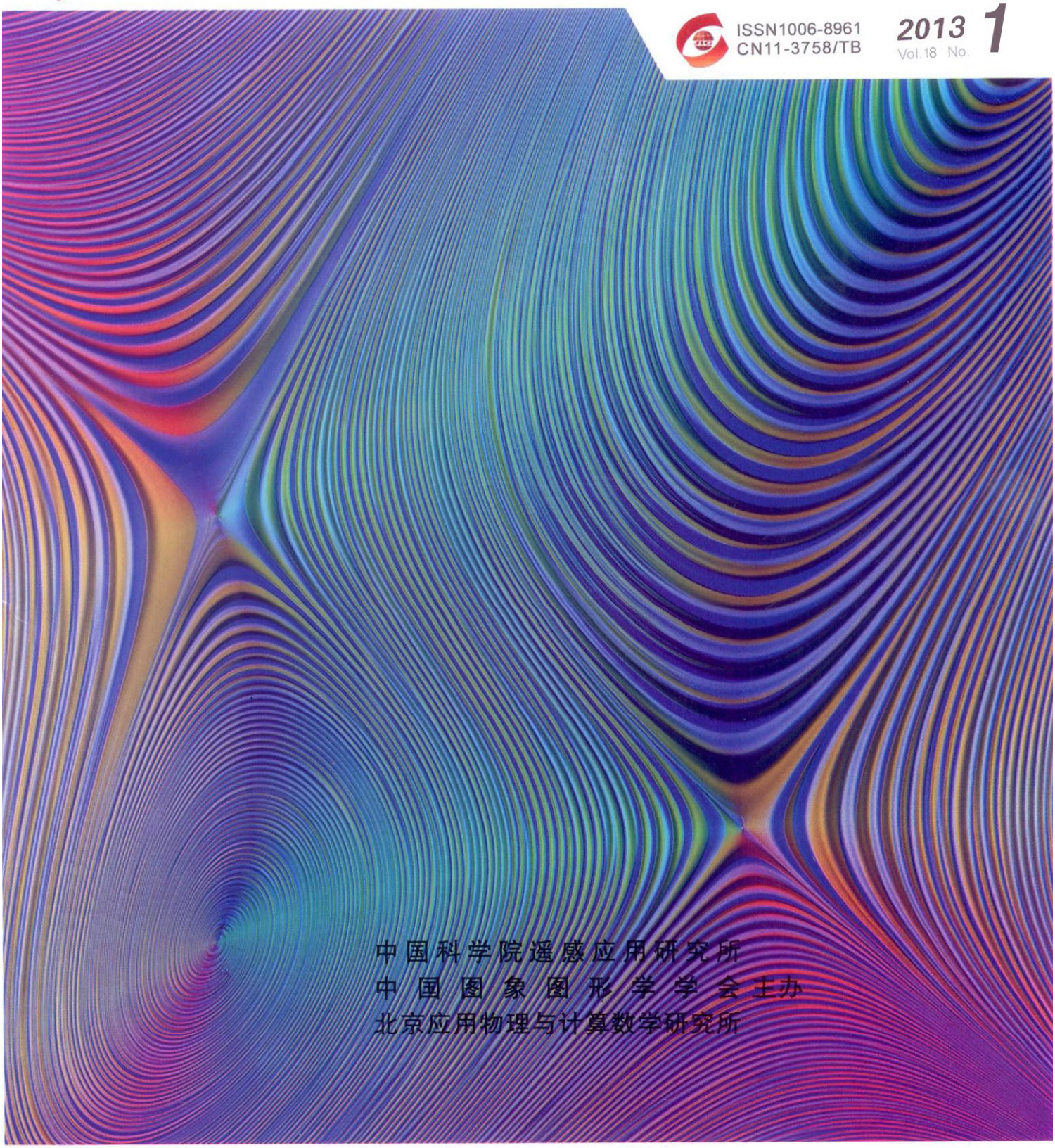
Journal of Image and Graphics

中国图象图形学报

ISSN1006-8961
CN11-3758/TB

2013 1
Vol.18 No.



中国科学院遥感应用研究所
中国图象图形学学会主办
北京应用物理与计算数学研究所

中国图象图形学报

Zhongguo Tuxiang Tuxing Xuebao

2013 年 1 月　第 18 卷　第 1 期(总第 201 期)

目　　次

Journal of Image and Graphics

(**Monthly, Started in 1996**)

Vol. 18 No. 1 January 2013

Contents

第 18 卷 第 1 期
2013 年 1 月

中国图象图形学报
Journal of Image and Graphics

Vol. 18, No. 1
Jan., 2013

中图法分类号: P208 **文献标识码**: A **文章编号**: 1006-8961(2013)01-0123-07

论文引用格式:张恒才,陆锋,陈洁. 微博客蕴含交通信息的提取[J]. 中国图象图形学报,2013,18(1):123-129.

微博客蕴含交通信息的提取

张恒才, 陆锋, 陈洁

中国科学院地理科学与资源研究所资源与环境信息系统国家重点实验室, 北京 100101

摘 要: 微博客消息中可能蕴含大量描述城市道路的交通信息,如交通状况、交通事件、交通管制等,提取这些交通信息能够为传统的固定式传感器和浮动车采集交通信息手段提供有效补充。然而,微博客消息描述的模糊性、差异性及非结构化特征,使得从海量微博客消息中快速准确地提取和甄别交通信息成为难题。提出一种从微博客消息中快速提取和融合交通信息的技术方法,首先对采集到的微博客消息进行分词解析和路网匹配,然后采用基于神经网络的模糊 C 聚类方法对描述路段交通状态的微博客消息定量化结果进行分析,获取各路段置信度最高的交通状态描述,最后得到各路段的交通畅通度水平。基于新浪微博客和北京路网的实验过程验证了本文技术方法的有效性。

关键词: 微博客;交通信息;分词;模糊聚类;畅通度;置信度

Extracting traffic information from massive micro-blog messages

Zhang Hengcai, Lu Feng, Chen Jie

State Key Lab of Resources and Environmental Information system, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, *China*

Abstract: Micro-blog messages usually contain a great deal of traffic information such as traffic conditions, traffic events and traffic controls, which can be useed as a complement to conventional traffic information collection technologies like fixed sensors and floating cars. However, due to ambiguous narrating, uncertainty, and the unstructured characteristics of micro-blog messages, extracting traffic information from micro-blog messages is rather difficult. In this paper, we propose an approach for extracting traffic information from a large amount of micro-blog messages. First, we build a traffic information table by semantically extracting traffic related words from micro-blog messages and matching each word onto the corresponding road segment of the road networks. Then, according to the traffic information table, we evaluate the highest confidence level of traffic condition for each road segment by using a neural network based Fuzzy-C-Means (FCM) clustering method, to obtain the most confident road conditions. Experiments on Beijing road networks with a large number of Sina micro-blog messages verify the effectiveness of the presented approach.

Key words: micro-blog; traffic information; word segmentation; fuzzy clustering; clear degree; degree of confidence

0 引 言

交通信息采集与处理是出行信息服务的关键技术,也是智能交通系统的重要组成部分。交通信息包括道路交通流、道路路况、交通限制、交通管制、交通事件、交通天气与路面环境信息等。快速、及时、准确地获取实时变化的交通信息,能够缓解交通拥

收稿日期:2012-02-06;**修回日期**:2012-07-30

基金项目:国家高技术研究发展计划(863)基金项目(2012AA12A211);国家自然科学基金项目(40871184, 41101149)。

第一作者简介:张恒才(1985—),男,中国科学院地理科学与资源研究所地图学与地理信息系统专业博士研究生,主要研究方向为互联网空间信息搜索、轨迹数据管理与数据挖掘。E-mail: zhanghc@lreis.ac.cn

通讯作者:陆锋,E-mail: luf@lreis.ac.cn

堵、提高交通运输效率,保障交通安全,改善环境质量,方便公众出行。目前交通信息采集和发布技术发展很快,交通信息服务形式多样,为提高城市出行效率及城市管理水平提供了很大帮助。

目前较成熟的交通信息采集技术包括固定传感器技术(感应线圈、视频监控和微波探测),安装 GPS 和无线通信设备的浮动车技术、移动通信终端信令分析技术等[1-4]。这些技术虽然得到了广泛应用,但难以有效捕捉突发性交通事件信息、临时交通管制信息、新增交通限制信息、针对某特定地点的实时交通信息及其他难以通过传感器探测的交通环境相关信息,仍存在很大的应用局限[5]。

近几年兴起的微博客(micro-blog)技术为交通信息采集提供了新的思路。微博客作为一种新型的信息分享、传播以及获取平台,具有更新速度快、参与人数多、用户分布广泛的特点。微博客平台(如 Twitter、新浪微博、腾讯微博等)的出现,正在改变人们传递信息的途径,已逐渐成为拥有大量用户的全新的人际交流方式及信息分享方式[6]。由于微博客消息大多与城市生活密切相关,是城市生活状态的实时反映,因此,微博客消息中蕴含着丰富的交通信息,涵盖各种交通信息类型,具有很高的时效性,而且随着微博客用户数的增长,来自普通大众的、志愿发布分享的实时交通信息将会同步增多,使得短时间内获取大量实时交通信息,尤其是其他技术手段难以采集的实时交通信息成为可能。可以预见,微博客将可能成为获取实时交通信息的新型技术手段,并与其他采集方式形成有效互补。

然而,微博客消息的描述具有很强的模糊性和非结构化特征,且多个不同微博客用户消息所描述内容之间可能存在差异甚至矛盾,从微博客消息中提取交通信息必须解决中文分词、消息甄别、时效性检验、消息融合等问题,才能有效获取城市道路路段的实时交通状态。针对上述问题,围绕微博客消息中文分词、消息蕴含交通信息的形式化和定量化描述、多消息的模糊聚类等关键环节展开,提出从微博客消息中提取和融合交通信息的技术方法。

1 微博客消息中文分词

自然语言分词是微博客消息提取信息的关键环节。目前主流的自然语言分词方法有基于词库的分词方法(最大匹配、逐词遍历方法、双向扫描方法等[7])与基于统计模型的分词方法(基于加权有限状态转换机、隐马尔科夫模型(HMM)等分词方法[8])。本文首先对中文微博客消息源进行过滤,只保留与交通信息相关的微博客消息用于交通信息的提取过程。由于微博客消息短小,最多不超过 140 个汉字,且本文从微博客中提取的交通信息专业性较强,存在较少的理解歧义,因此,在分析大量实时路况信息的语料库所构建的地址词库、方向词库、事件词库以及附属定位词库的基础之上,采用前期提出的交通信息跨阶分词算法[9-10]实现微博客消息的中文分词过程。该算法充分考虑了交通信息自然语言描述词库记录长度特点,采用依词库性质变化的多阶跨越中文分词算法,提高了分词效率。

2 微博客消息融合

2.1 模糊 C 聚类算法

微博客消息多源于公众,受制于各种条件,微博客消息描述存在很大的模糊性,并且不同用户发送微博客消息可能存在差异。因此,需要对微博客消息进行模糊聚类,从不同用户发送的微博客消息中提取出可信度最高的交通信息。

聚类的本质是使同类之间的差别最小,类别之间的差别最大。模糊 C 聚类(FCM)由 Dunn 于 1973 年提出,利用隶属度来确定元素属于某个类别程度,特别适合处理模糊集合聚类问题[11]。本文采用 FCM 聚类算法实现微博客消息的模糊聚类过程。

假定微博客消息集为 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$,$n$ 为微博客消息数量,FCM 算法将微博客消息分为 p 类,每个类别的聚类中心为 $\boldsymbol{v}_j$,u_{ij} 为微博客消息 i 属于第 j 类的模糊隶属度,显然 u_{ij} 满足条件

$$\sum_{j=1}^{p} u_{ij}=1,\text{且 } u_{ij}\in[0,1]$$

$\boldsymbol{v}_j$ 与 u_{ij} 为

$$u_{ij}=\begin{cases}\left[\sum\limits_{k=1}^{p}\dfrac{\|\boldsymbol{x}_i-\boldsymbol{v}_j\|^{2/(m-1)}}{\|\boldsymbol{x}_i-\boldsymbol{v}_k\|^{2/(m-1)}}\right]^{-1} & \|\boldsymbol{x}_i-\boldsymbol{v}_k\|\neq 0\\ 1 & \|\boldsymbol{x}_i-\boldsymbol{v}_k\|=0\text{ 且 }k=j\\ 0 & \|\boldsymbol{x}_i-\boldsymbol{v}_k\|=0\text{ 且 }k\neq j\end{cases} \tag{1}$$

$$\boldsymbol{v}_j = \frac{\sum_{i=1}^{n} u_{ij}^m \boldsymbol{x}_i}{\sum_{i=1}^{n} u_{ij}^m} \tag{2}$$

FCM 算法的迭代过程是使模糊目标函数为最小化的过程,模糊目标函数计算公式为

$$J = \sum_{i=1}^{n}\sum_{j=1}^{p} (u_{ij})^m \| \boldsymbol{x}_i - \boldsymbol{v}_j \|^2 \tag{3}$$

式中,m 为模糊权重指数。

模糊聚类迭代过程如下:

1)给定类别数量 p,模糊权重指数 m,初始化聚类中心 $\boldsymbol{v}$,迭代次数 t;

2)根据式(1)计算模糊隶属度矩阵 $\boldsymbol{u}$,根据式(2)计算聚类中心 v;

3)根据式(3)确定模糊目标函数 J,计算约束条件

$$\| J^{(it+1)} - J^{(it)} \| \leqslant \varepsilon \tag{4}$$

式中,ε 为预先给定的允许误差,若满足则算法终止,否则转向步骤 2)。

2.2　RBF 神经网络优化

传统的 FCM 算法对噪声数据比较敏感,易陷入局部最优。由于不同交通信息类别差别较小、噪声数据较多,大大限制了 FCM 算法在微博客交通信息提取中的应用。鉴于此,提出基于径向基函数(RBF)神经网络的 FCM 聚类算法,利用 RBF 神经网络优化模糊聚类结果,提高聚类精度,实现对微博客消息中蕴含实时交通信息的融合。

RBF 神经网络是一种由输入层、隐含层及输出层组成的前向反馈网络。其特点是训练简洁,学习收敛速度快,对非线性连续函数具有较好的逼近性能[12-14]。RBF 神经网络的输入层作用仅限于数据传递,隐含层通过径向基函数对输入数据进行映射变换,实现低维输入数据到高维空间的转化。一般选择高斯函数作为隐含层的径向基函数,即

$$R_{i(x)} = \exp\left(\frac{\| \boldsymbol{x} - \boldsymbol{z}_i \|^2}{2\sigma^2}\right)$$

式中,$\boldsymbol{x}$ 为输入向量,$\boldsymbol{z}_i$ 为隐含层 i 个节点的高斯函数的中心,σ 为高斯函数的方差,从隐含层到输出层为线性调整,则神经网络的输出为

$$y_j = \sum_{i=1}^{k} w_{ij} R_{i(x)}$$

式中,y_j 为输出层的第 j 个输出节点,w_{ij} 为隐含层到输出层的连接权值,k 为隐含层的节点数。

2.3　交通信息提取

首先定义所有微博客消息记录集为 $\boldsymbol{M}_{\text{blog}}$,道路路段集合为 $\boldsymbol{R}$,交通信息对应时间点为 T_{traffic},道路路段的畅通度集合 $\boldsymbol{D}_{\text{clear}}$,对应置信度集合为 $\boldsymbol{Q}_{\text{support}}$,实时交通信息更新表为 $\boldsymbol{I}_{\text{infotable}}$。

定义描述道路路段集合 $\boldsymbol{R}$ 中第 l 条道路路段 R_l 交通信息的微博客消息记录集为

$$\boldsymbol{V}^l = \{\boldsymbol{V}_1, \boldsymbol{V}_2, \cdots, \boldsymbol{V}_i, \cdots, \boldsymbol{V}_n \mid n \geqslant 1\}$$

式中,n 为包含路段 R_l 交通信息微博客消息记录个数,第 i 条微博客消息 $\boldsymbol{V}_i$ 采用 1 维向量表示,即

$$\boldsymbol{V}_i = (P_i, S_{p_i}, T_i^{\text{publish}}, L_i, Z_i, C_i^{\text{keep}}, C_i^{\text{forward}}, C_i^{\text{comment}}, B_i)$$

P_i 为发表该消息的用户标识;S_{P_i} 为用户 P_i 诚实度,且系统初始状态时 $S_{P_i} = 0$;T_i^{publish} 消息发表时间;L_i 消息描述的地点,可以是平面坐标或者自然语言描述;$Z_i = [0,1]$ 为消息发表客户端,移动客户端取值为 1,否则取值为 0;C_i^{keep} 为该消息收藏次数;C_i^{forward} 为该消息转发次数;C_i^{comment} 为该消息评论次数;B_i 为该消息内容;畅通度集合 $\boldsymbol{D}_{\text{clear}}$ 反映城市各路段交通畅通情况,本文采用模糊数学方法,以值域区间 $[0,1]$ 来反映各路段的交通通畅程度,畅通度越高代表路段越畅通;置信度集合 $\boldsymbol{Q}_{\text{support}}$ 中各元素反映了对应的路段畅通度结果的置信程度。

基于 RBF 的微博客消息 FCM 聚类与交通信息提取具体步骤如下:

1)根据所需产生交通信息的时间点 T_{traffic} 及时间偏移量 Δt 构建有效时间窗口 T_{interval}。

2)利用 T_{interval} 及道路路段集合 $\boldsymbol{R}$ 对微博客消息集合 $\boldsymbol{M}_{\text{blog}}$ 进行过滤分组,构建路段 R_l 所对应 $\boldsymbol{V}^l$。

3)利用交通信息描述词库解析 $\boldsymbol{V}^l$ 中微博客内容,提取对应的交通状态信息及方向描述,定量化表达微博交通信息 $\boldsymbol{V}_{\text{new}}$。

4)判断交通信息类型,文中包含两类交通信息,一类为交通限制、交通管制或其他交通相关信息,如突发性交通事件以及特定点交通状态(如交叉口或特定位置)描述信息等;另一类为路况型的交通信息,如道路畅通度或行驶速度信息。

5)设定初始交通状态类别数目 num,对分组定量化后的微博客消息 $\boldsymbol{V}_{\text{new}}$ 执行 FCM 聚类,得到每类交通状态模糊隶属度矩阵 $\boldsymbol{u}$ 及聚类中心,每条微博客消息提取的交通状态类别,判断是否达到设定聚类精度 ε,若“是”转步骤 9),若“否”转步骤 6)。

6)选取距每个类类别中心最近的 p 个微博客消息记录,构建新微博客消息样本数据记录集 $\boldsymbol{V}_{\text{sample}}$。

7)利用样本数据记录集 $\boldsymbol{V}_{\text{sample}}$ 训练 RBF 神经网络模型,得到训练好的神经网络 *net*。

8)将微博客消息 $\boldsymbol{V}_{\text{new}}$ 中的所有消息作为训练好的神经网络 *net* 的输入数据,根据神经网络输出序列将微博客消息重新分为 *num* 类。

9)求出每个交通状态类别中所有微博客消息提取交通信息样本的平均值 *mean* 及微博客消息数量 m,选取每类内 m 最大的交通状态类别的平均值为路段 R_1 在有效时间窗内的畅通度 D_{clear}。

10)采用所选定交通状态类别与微博客消息数量的比值作为该条交通状态估计的置信度 $\boldsymbol{Q}_{\text{support}}=m/n$。

11)将 D_{clear} 与 $\boldsymbol{Q}_{\text{support}}$ 插入 $\boldsymbol{I}_{\text{infotable}}$。

技术流程如图 1 所示。

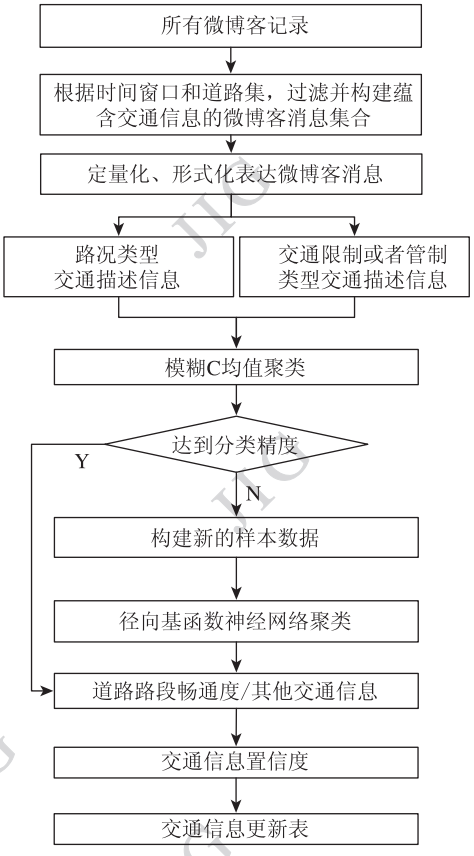


图 1 微博客蕴含交通信息提取流程

Fig. 1 Flow chart of extracting traffic information from micro-blogs

3 实验与讨论

3.1 实验分析

实验程序测试运行环境为 CentOS Linux 操作系统,CPU 为 4 核 Intel(R) Xeon(R) CPU E5520 2.27 GHz,内存 4 G,采用 JAVA 语言实现了群体微博客消息提取交通信息的方法。实验所采用的微博客消息来源于新浪微博客(http://weibo.com),主要内容包括:微博 ID、创建时间、信息内容、来源、是否已收藏、是否被截断、回复人 UID、微博 MID、图片地址、转发数、评论数、附加注释信息、地理信息字段、作者信息字段等[15]。提取 2010 年 9 月至 2011 年 10 月期间部分微博客用户发布的与北京城市交通有关的 39 506 条微博客消息,涉及 2 891 条路段。图 2 为获取的原始微博客消息记录样本。图 3 为某日不同时刻微博客消息解析形式化后的内容,图 4 为对应时刻产生的交通信息表。为验证本文所提出的技术流程的正确性,以 2011-10-28T19 为交通信息采集时间点,时间偏移量为 1h,进行微博客消息蕴含交通信息的提取与融合。图 5 为从新浪微博网站抓取的一条微博客消息附图,经过处理后该条微博客消息可形式化表达为:用户标识 P_i = 2 040 014 515;发表时间 T_i^{publish} = "2010-10-28T19:07:37";L_i = "长安街";客户端 Z_i = 1;收藏次数 C_i^{keep} = 0;转发次数 C_i^{forward} = 14;评论次数 C_i^{comment} = 3;B_i 该微博客消息内容。解析后可得到:道路路段 R_i 为长安街建国门内大街路段,交通信息为"非常堵",对应的方向为"由西向东"。图 6 为同一时刻 Google 地图(http://ditu.google.cn)与百度地图(http://map.baidu.com)所显示的交通路况(主要基于浮动车系统采集)。从该微博客消息和实拍照片中可以确定在 2011-10-28T19:07 左右,长安街东段建国门内大街路段发生道路拥堵,但是在 Google 地图与百度地图中却显示该时刻该路段为交通畅通状态。这显然是错误的。错误原因可能归因于该时刻途经该路段的浮动车较少,或者信息处理滞后所致。可以看出,就此条微博客消息而言,从中所提取的交通信息无疑具有更好的时效性,可以作为目前流行的交通信息采集手段的有效补充。

为了验证该算法的效率,在实验区域内随机选择了 10 条道路,进行微博客交通信息融合,设定交通信息提取时间点为 2011-09-07T18,共获取专业用户发送的微博客消息 113 条,经过滤后剩余 62 条。算法涉及微博客消息中文分词、交通信息形式化表达、模糊聚类、畅通度和置信度计算、记录插入数据库表等过程,经统计,本文算法处理一条蕴含交通信息的微博客消息平均耗时为 0.138 s,能够满足应用需求。

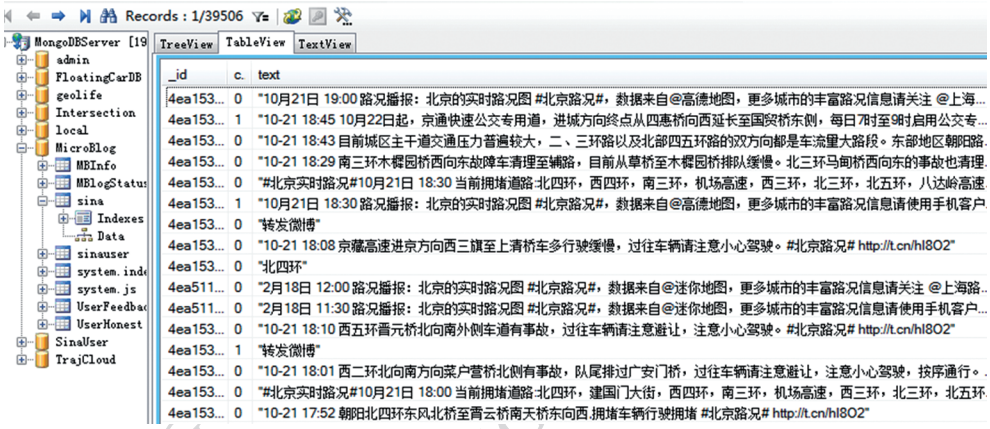
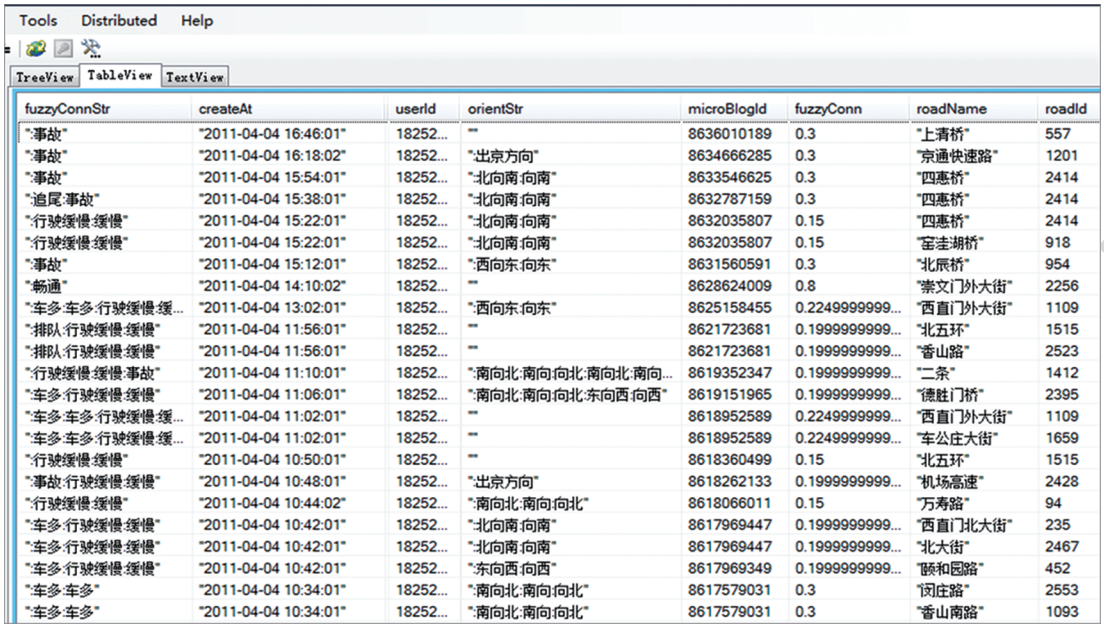


图 2　蕴含实时交通信息的新浪微博客消息

Fig. 2　Sina micro-blog messages containing traffic information



fuzzyConnStr	createAt	userId	orientStr	microBlogId	fuzzyConn	roadName	roadId
":事故"	"2011-04-04 16:46:01"	18252...	""	8636010189	0.3	"上清桥"	557
":事故"	"2011-04-04 16:18:02"	18252...	":出京方向"	8634666285	0.3	"京通快速路"	1201
":事故"	"2011-04-04 15:54:01"	18252...	":北向南:向南"	8633546625	0.3	"四惠桥"	2414
":追尾:事故"	"2011-04-04 15:38:01"	18252...	":北向南:向南"	8632787159	0.3	"四惠桥"	2414
":行驶缓慢:缓慢"	"2011-04-04 15:22:01"	18252...	":北向南:向南"	8632035807	0.15	"四惠桥"	2414
":行驶缓慢:缓慢"	"2011-04-04 15:22:01"	18252...	":北向南:向南"	8632035807	0.15	"玺洼湖桥"	918
":事故"	"2011-04-04 15:12:01"	18252...	":西向东:向东"	8631560591	0.3	"北辰桥"	954
":畅通"	"2011-04-04 14:10:02"	18252...	""	8628624009	0.8	"崇文门外大街"	2256
":车多:车多:行驶缓慢:缓...	"2011-04-04 13:02:01"	18252...	":西向东:向东"	8625158455	0.2249999999...	"西直门外大街"	1109
":排队:行驶缓慢:缓慢"	"2011-04-04 11:56:01"	18252...	""	8621723681	0.1999999999...	"北五环"	1515
":排队:行驶缓慢:缓慢"	"2011-04-04 11:56:01"	18252...	""	8621723681	0.1999999999...	"香山路"	2523
":行驶缓慢:缓慢:事故"	"2011-04-04 11:10:01"	18252...	":南向北:南向:向北:南向北:南向...	8619352347	0.1999999999...	"二条"	1412
":车多:行驶缓慢:缓慢"	"2011-04-04 11:06:01"	18252...	":南向北:南向:向北:东向西:向西"	8619151965	0.1999999999...	"德胜门桥"	2395
":车多:车多:行驶缓慢:缓...	"2011-04-04 11:02:01"	18252...	""	8618952589	0.2249999999...	"西直门外大街"	1109
":车多:车多:行驶缓慢:缓...	"2011-04-04 11:02:01"	18252...	""	8618952589	0.2249999999...	"车公庄大街"	1659
":行驶缓慢:缓慢"	"2011-04-04 10:50:01"	18252...	""	8618360499	0.15	"北五环"	1515
":事故:行驶缓慢:缓慢"	"2011-04-04 10:48:01"	18252...	":出京方向"	8618262133	0.1999999999...	"机场高速"	2428
":行驶缓慢:缓慢"	"2011-04-04 10:44:02"	18252...	":南向北:南向:向北"	8618066011	0.15	"万寿路"	94
":车多:行驶缓慢:缓慢"	"2011-04-04 10:42:01"	18252...	":北向南:向南"	8617969447	0.1999999999...	"西直门北大街"	235
":车多:行驶缓慢:缓慢"	"2011-04-04 10:42:01"	18252...	":北向南:向南"	8617969447	0.1999999999...	"北大街"	2467
":车多:行驶缓慢:缓慢"	"2011-04-04 10:42:01"	18252...	":东向西:向西"	8617969349	0.1999999999...	"颐和园路"	452
":车多:车多"	"2011-04-04 10:34:01"	18252...	":南向北:南向:向北"	8617579031	0.3	"闵庄路"	2553
":车多:车多"	"2011-04-04 10:34:01"	18252...	":南向北:南向:向北"	8617579031	0.3	"香山南路"	1093

图 3　经解析和路网匹配后的微博客消息

Fig. 3　Analysis and road matching of micro-blog messages

_id	roadName	roadId	clearD...	support
4f0e7eb0445e0e708ae3e4a1	"崇文门外大街"	2256	0.8	0
4f0e7eb0445e0e708ae3dd61	"西五环"	400	0.35	0.5
4f0e7eb0445e0e708ae3dbea	"广渠路"	24	0.3	0
4f0e7eb0445e0e708ae3ded5	"学院路"	771	0.3	0
4f0e7eb0445e0e708ae3e03e	"清华东路"	1132	0.3	0
4f0e7eb0445e0e708ae3e1d2	"紫竹院路"	1536	0.3	0
4f0e7eb0445e0e708ae3e388	"京良路"	1974	0.3	0.5
4f0e7eb0445e0e708ae3e468	"成府路"	2198	0.3	0
4f0e7eb0445e0e708ae3e2b2	"国贸桥"	1759	0.2571...	0
4f0e7eb0445e0e708ae3e615	"建国路"	2626	0.2523...	0.6666...

图 4　动态交通信息更新表

Fig. 4　Dynamic updating the traffic information

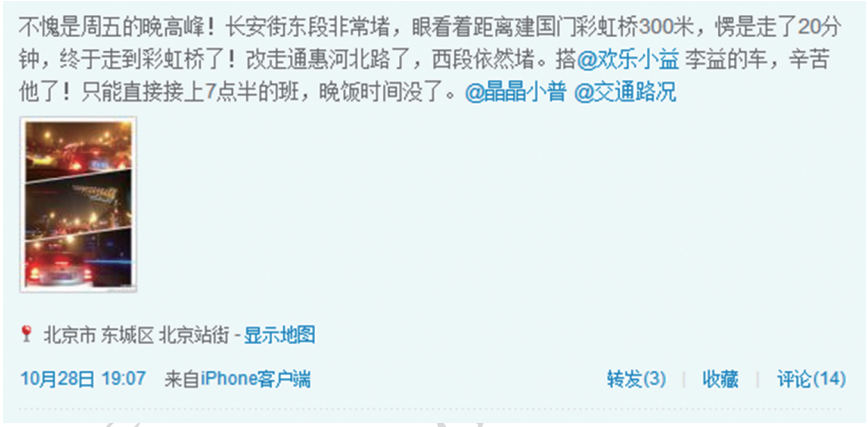
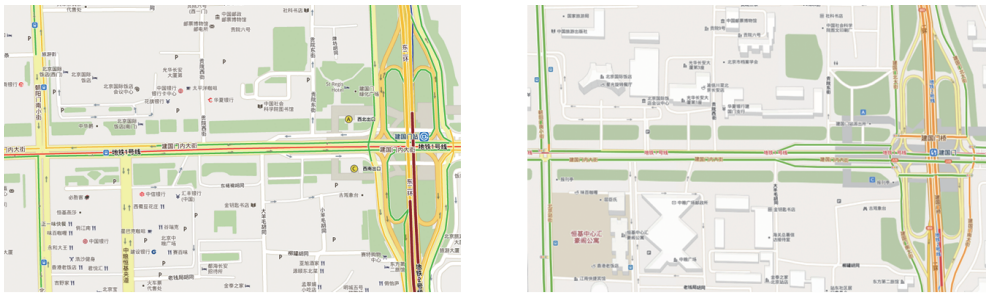


图 5 微博客用户发布的实时路况消息

Fig. 5 Real-time traffic released by micro-blog users



(a) 百度地图 (b) Google 地图

图 6 对应同一时刻的 Google 和百度地图路况

Fig. 6 Corresponding road traffic released by Google and Baidu maps

3.2 讨论

1)交通信息可以分为时效性较短的交通信息(如突发交通事故,临时交通限制等)与时效性较长的交通信息(如占道施工信息、交通管制信息等),但微博客发表时间往往具有滞后性,在微博客用户规模不大、交通信息微博客消息数量有限的城市环境中,会影响所提取的交通信息的时效性。因此,如果微博客用户和有效消息规模有限,本文提出的交通信息提取方法对时效性较长的交通信息提取,及其交通天气、路面状况(如道路遗撒、水淹结冰、路面破损等)等浮动车系统无法采集的交通信息提取更有效,随着微博客用户、特别是专业交通信息微博客的迅速普及,本文技术方法也将适用于动态性极强的实时交通信息提取。

2)通过分析实验收集的微博客消息记录,发现微博客消息发送具有突发性,发布高峰往往会集中在某个特定时段,早上的 8～9 点与 17～19 点是微博客消息发布的两个高峰时段,工作日与非工作日微博客消息的发送量也有很大差异。在本文交通信息提取算法的设计过程中,尚未考虑微博客消息发布的时序异质性特征及其对交通信息提取的影响,在后续研究中需要进一步考虑微博客消息发布时间的不均衡性对交通信息模糊聚类方法效率及正确率的影响。

3)本文交通信息提取算法,运行时间耗费主要发生在 RBF 神经网络学习环节。主要原因在于 RBF 神经网络设计中,隐含层神经元的网络连接权值、高斯函数中心矢量等参数难以确定,对神经网络逼近造成影响。在后续研究中可以通过一些寻优算法来优化 RBF 神经网络中心参数,提高 RBF 神经网络寻优的速度与精度。

4)借鉴路段畅通度指标来反映微博客消息所蕴含的路段交通状态信息。其思想在于利用模糊集理论来处理动态连通关系。但是路段畅通度指标的影响因子众多,需要针对历史交通状态数据进行统计分析、获取专家经验等知识,才能获得较好的道路

路段畅通度和置信度计算结果。

4 结　论

从微博客消息提取交通信息的技术方法,通过对采集到的微博客消息进行过滤、自然语言分词解析和路网匹配,采用基于神经网络的模糊 C 聚类方法对描述路段交通状态的微博客消息进行统计分析,解决了微博客消息的描述模糊性处理和不同微博客用户发布消息的描述差异性处理问题,从微博客消息中获取了所描述路段的畅通度水平。微博客蕴含交通信息提取方法能够为现有的交通信息采集手段提供有效的补充。在后续工作中,将进一步研究微博客用户诚实度对交通信息融合的影响以及融合结果置信度函数的确定,使融合结果更加准确可信。

参考文献(References)

[1] Schneider W, Arsenal R. Mobile phones as a basis for traffic state information[J]. Intelligent Transportation Systems, 2005, 13(15):782-784.

[2] Zhang C B, Yang X G, Yan X P. Traffic data collection system based on floating cars [J]. Computer and Communications, 2006, 24(5):31-34. [张存保,杨晓光,严新平. 基于浮动车的交通信息采集系统研究[J],交通与计算机,2006,24(5):31-34.]

[3] Zhu T Y, Guo S M. A study on floating car based information processing technology [J]. Journal of Image and Graphics, 2009, 14(7):1230-1237. [诸彤宇,郭胜敏. 浮动车信息处理技术研究[J],中国图象图形学报,2009,14(7):1230-1237.]

[4] Guo D H, Cui W H. Trajectory mining for live traffic condition retrieving[J]. Journal of Wuhan University of Technology. Transportation Science & Engineering,2010,34(1):6-9. [郭丹怀,崔伟宏. 面向实时交通信息提取的车辆轨迹数据挖掘[J],武汉理工大学学报:交通科学与工程版,2010,34(1):6-9.]

[5] Lu F, Zheng N B, Duan Y Y, et al. Travel information services: state of the art and discussion on crucial technologies[J]. Journal of Image and Graphics, 2009,14(7):1219-1229. [陆锋,郑年波,段滢滢,等. 出行信息服务关键技术研究进展与问题探讨[J]. 中国图象图形学报,2009,14(7): 1219-1229.]

[6] Wu X, Wang J. How about micro-blogging service in China: analysis and mining on sina micro-blog [C]//The 1st International Symposium on From Digital Footprints to Social and Community Intelligence. New York, USA: ACM, 2011:37-42.

[7] Sun, M S, Zou J Y. A critical appraisal of the research on Chinese word segmentation [J]. Contemporary Linguistics, 2001, (1):22-32. [孙茂松, 邹嘉彦. 汉语自动分词研究评述[J]. 当代语言学. 2001, (1): 22-32.]

[8] Zong C Q. Statistical Natural Language Processing [M]. Beijing: Tsinghua University Press, 2008: 105-109. [宗成庆. 统计自然语言处理[M]. 北京:清华大学出版社, 2008:105-109.]

[9] Lu F, Liu H H, Chen C B. A cross-step word segmentation algorithm for understanding traffic information represented in natural Chinese language[J]. Geomatics and Information Science of Wuhan University,2009, 34(8):943-947. [陆锋,刘焕焕,陈传彬. 一种中文自然语言表达交通信息的跨阶分词算法[J]. 武汉大学学报:信息科学版,2009,34(8):943-947.]

[10] Chen C B, Lu F, Li H G, et al. Matching urban traffic information in chinese natural language with road network[J]. Journal of Image and Graphics, 2009, 14(8):1669-1676. [陈传彬,陆锋,励惠国,等. 自然语言表达实时路况信息的路网匹配融合技术[J]. 中国图象图形学报,2009,14(8):1669-1676.]

[11] Bezdek J C. Cluster validity with fuzzy sets[J]. Journal of Cybernetics, 1973, 3(3): 58-71.

[12] Jackson I R H. Radial basis functions: a survey and new results [C]//Proceedings of the 3rd IMA Conference on the Mathematics of Surfaces. Oxford UK: Clarendon Press, 1988:115-133.

[13] Powell M J D. The theory of radial basis function approximation in 1990[M]//Light W, ed. Advances in Numerical Analysis: Volume II: Wavelets, Subdivision Algorithms, and Radial Basis Functions. Oxford: Clarendon Press, 1992: 105-210.

[14] Schaback R. Comparison of radial basis function interplants [M]//Multivariate Approximation. From CAGD to Wavelets. Singgapore: World Scientific Publishing Co., 1995,293-305.

[15] Sina. micro-blog developer's guide[EB/OL]. (2011-10-25)[2011-12-10]. http://open.weibo.com/. [新浪. 微博开放平台文档[EB/OL]. (2011-10-25)[2011-12-10]. http://open.weibo.com/.]