面向微博内容的信息抽取模型研究

郑 影 李大辉
(齐齐哈尔大学计算机与控制工程学院 齐齐哈尔 161006)

摘 要 社会媒体是人们用来分享意见、见解、观念和经验的平台或工具,目前已经发展成具有重大影响力的新媒体。而微博作为社会媒体的一个重要部分,对信息的传播起到了很大的作用。面向微博内容的信息抽取就是要从充满噪音的、零碎的、非结构化的微博内容的自由文本中提取有价值的结构化的信息,以利于从微博内容中有效地获取信息。提出了一种基于因子图的微博事件抽取方法来准确地抽取微博中所反映的事件。最后通过实验验证了该方法在性能和准确性上都比其他的方法要高。

关键词 社会媒体,微博,事件抽取,因子图
中图法分类号 TP393 **文献标识码** J

Research on Information Extration Model for Microblog Content

ZHENG Ying LI Da-hui
(Institute of Computer and Control Engineering, Qiqihar University, Qiqihar 161006, China)

Abstract Social media is the platform or tool that people use to share opinions, insights, ideas and experience. It has become the new media having great influence. Microblogging is an important part of social media, so it will play an important role in the information transfer. Microblogged content-oriented information extraction is to extract the valuable structred information from free text of full of noise, loose, unstructured microblogging content to facilitate effective access to information from Twitter content. This paper proposed a microblogging event extraction based on factor graph approach to accurately extract the events reflected in microblogging. At last we used some experiments to verify the effectiveness of the methods, and the results show that the performance and accuracy of this method is higher than other methods.

Keywords Social media, Microblog, Event extraction, Factor graph

社会媒体是人们用来分享彼此意见、见解、观念和经验的平台或工具。在不至混淆的情况下,社会媒体也泛指借助平台或工具性质的社会媒体所分享的内容。社会媒体内容可以是文本、图片、声音或视频。本文仅仅考察社会媒体中文本类型的媒体进行信息的抽取。最近5年来,社会媒体得到迅猛发展,特别是一批微博站点异军突起,例如美国的Twitter[2]和Facebook[1],以及中国的新浪微博和腾讯微博等,现已发展成为互联网上的巨擘。到2012年3月,Twitter的注册用户已突破了5个亿[3],而Facebook的注册用户则已超过8亿[4];通过Twitter,人们每天发布5千万条信息消息,而Facebook则约为6千万条;Facebook目前已经超过谷歌,成为北美流量最大的网站。

草根性是微博的重要特征之一,草根们的声音通过微博汇聚放大,开始凸显出不可忽视的影响力。例如,2007年洛杉矶大火刚刚爆发,该事件就通过Twitter消息在第一时间内进行了报道;2008年,奥巴马利用相关社会媒体在短短27天就募集到5500万美金的竞选经费,远远超过了竞争对手;2009年,伊朗总统选举期间,反对派支持者利用Twitter交流信息,并借助Twitter号召发起了一次10万人规模的大游行。

随着微博蓬勃发展而带来的一大隐患就是人们从微博内容中获得有用信息越来越困难。一方面是微博内容书写随意,噪音大,内容数据量达到海量级别,且处于不断更新之中。另一方面,缺少针对微博内容的有效的搜索工具。当前主流的搜索引擎,例如谷歌、微软和雅虎,对于微博内容的搜索还是处于一种有限的状态,且无法真正对其内容进行搜索。因此从微博中抽取社交事件面临着以下的挑战:1)微博一般很短,不能提供分类所需的足够信息;2)微博往往充满噪音,导致现有的自然处理工具(例如句块分析和依存语法分析工具)性能差而不能提供稳定的特征(这些特征在自动内容抽取任务的语料库上已经被证明是有效的)。本文提出利用因子图(一种概率图模型)从多条相似的微博中同时抽取社交事件,以应对这些挑战。

1 相关工作

这节将从信息抽取的相关研究以及在针对微博的社交事

到稿日期:2013-03-11 返修日期:2013-05-14 本文受齐齐哈尔大学青年教师科研启动支持计划项目(2011k-M03),黑龙江省自然科学基金项目(F201218)资助。

郑 影(1978—),女,博士生,讲师,主要研究方向为计算机图像处理、智能控制,E-mail:zhengying991@163.com;李大辉(1968—),男,博士,教授,主要研究方向为信号检测与识别。

件抽取上的相关研究进行叙述。

1.1 信息抽取相关研究

信息抽取是一种将非结构化的互联网结构化乃至知识化的手段。在某些特定的领域内，信息抽取获得了很大的成功。例如，在生物领域内，信息抽取技术被成功应用到基因和蛋白质名称抽取[5]以及与蛋白质间交互有关的事实的抽取[6]。本节把信息抽取技术分为两个发展阶段，分别介绍每个阶段的信息抽取的主要任务、主要研究方法以及典型系统。

第一阶段的信息抽取是面向特定领域和针对特定关系类型，其主流方法是数据驱动的统计方法。本文把这种类型的信息抽取称为传统的信息抽取。

可以把传统的信息抽取分为如下 4 个子任务：实体边界确定、实体类型确定、关系抽取(将一组实体组合成一条记录)和聚类(判定实体或关系是否相同)。前两个子任务合起来被称为命名实体识别，而识别命名实体的方法可分为基于规则的和基于数据驱动的这两大类。实体抽取之后是关系抽取。关系抽取的任务是把一组实体组合为一条记录。关系抽取的方法分为两大类：基于规则的和基于数据驱动(或机器学习)的。关系抽取任务还需要对抽取到的关系的正确性进行评价。在评价关系的正确性时，一般考虑如下几个因素：关系的出现次数、抽取规则的可信度以及与已知事实的相似度。传统信息抽取的第四个阶段是信息融合。这一阶段要判定两条记录是否为同一个记录。该阶段典型的方法有基于马尔可夫逻辑网(Markov Logic Networks，MLNs)[7]的方法。KnowItAll[8]是传统信息抽取系统的典型代表。

目前，信息抽取发展到了开放式信息抽取的阶段。开放式信息抽取处理的对象是整个互联网，对领域、实体类型和关系类型都没有任何限制。开放式信息抽取的目标实体可以是任何名词短语，目标关系也是在学习过程中自动发现的。它的另一个突出特点是采用与 KnowItAll 类似的领域无关的知识库，通过所谓的自主监督学习的方式(Self-training)[9]，自己标注训练数据、训练模型。表 1 概括了开放式信息抽取和传统信息抽取的主要不同。TextRunner[10]是第一个公开报道的开放式信息抽取系统，也代表了目前开放式信息抽取的最好水平。

表 1 传统信息抽取和开放式信息抽取的对比

	传统信息抽取	开放式信息抽取
输入	待处理语料＋训练数据	待处理语料＋领域无关的知识
关系	预先指定	自动发现
开发时代价	O(R)，R 为关系数	O(1)
运行时代价	O(RD)，D 为文档数	O(D)
依赖工具	句法分析器＋命名实体识别器	名词短语识别器

总之，无论是传统的信息抽取还是近期提出的开放式信息抽取，都主要是为传统媒体而非社会媒体设计的。如前所述，社会媒体中的文本噪音大，书写随意，与传统媒体中的文本有着显著的不同。这导致现有的信息抽取技术(无论是基于规则的还是基于数据驱动的)直接应用到社会媒体内容时，其性能大幅度下降。本文的研究重点就是如何改造现有的信息抽取技术使之适用于社会媒体内容。

1.2 面向微博的社交事件抽取相关研究

Harabagiu 等[11]提出了一个针对句子层次的 ACE 事件提取系统，该方法主要是结合模式匹配和统计模型来进行事件抽取。Liao 和 Grishman[12]使用跨事件信息开发来对现有 ACE 事件提取进行研究。现在同样也有大量针对 TIMEML 事件的研究，例如，STEP[13]使用一个富文本集、形态依赖以及 WordNet 结合特征来建立的 SVMs 模型。最近，Llorens 等[14]分析语义角色对 TimeML 事件标识的贡献。除了 ACE 以外，TimeML 事件识别同样存在。例如，Yu 等[15]利用关联的语言模式再加上单个词作为特征对句子进行分类，将变成负面生活事件集成。

Sankaranarayanan 等[16]抽取 Twitter 中的一些突发事件来建立一个新的处理系统，称为 TwitterStand；Sakaki 等[17]制订的基于单一推文的特征来对 tweets 进行分类(例如，一个 tweet 的关键字以及词的数量等)，从而检测特定类型的事件(地震)。Benson 等[18]提出一个图形化的模型针对多个消息的信息进行聚类从而提取正规的娱乐事件。

Apoorv 和 Rambow[19]第一次从一个公司的 ACE 数据中研究社会事件提取，并提出使用内核 SVMs 来实现。我们的工作主要是受该工作的启发。然而，有两个显著的差异。首先，我们使用图模型而不是内核 SVMs 来推断所有候选事件的标签。模型的优点之一是它允许我们跨多个 tweet 信息，从而弥补了只在单个 tweeter 上进行信息聚合的缺点。第二，我们的模型采用一些简单的语言特征，而不是分块以及依赖于一些相关分析的特征。这是因为推文很短并且经常是随意书写的，所以当前自然语言处理工具对于 tweeter 来说并不是很好[20]，这意味着这种高级语言的功能对 tweets 来说是不可靠的。

2 相关概念描述

2.1 微博社交事件抽取任务描述

给定一组微博 $T=\{t_m\}$，其任务是输出 $E_m=\{e_m^i\}_i^{N_m}$，$m=1,\cdots,|T|$。其中：e_m^i 表示从第 m^{th} 个微博中抽取的第 i^{th} 个社交事件；N_m 表示从 t_m 微博中抽取到的社交事件总数。与 Apporv 和 Rambow 对社交事件的定义一致，本工作把一个社交事件定义为一个三元组 $e_m^i=(p_1,p_2,y)$，其中 p_1 和 p_2 分别表示参与该事件的两个人，y 表示该社交事件的具体类型，取值可为 Interaction、Observation 或 None(N)。

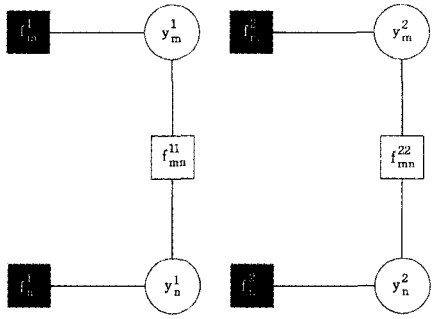
本工作假定 p_1 和 p_2 是可互换的，也就是说 (p_1,p_2,y) 和 (p_2,p_1,y) 代表相同的事件；进一步假定，微博中的人名已经被准确识别出来。

作为示例，假定输入为“… Love is in the air? This pic of the day looks like [Demi Lovato]$_p$ and [Wilmer Valderrama]$_p$ are getting pretty close …”和“…[Demi Lovato]$_p$ & [Wilmer Valderrama]$_p$ are caught dating …”，其中[…]$_p$ 表示一个人名。那么期望的输出是 $E_1=\{$(Demi Lovato，Wilmer Valderrama，I)$\}$和 $E_2=\{$(Demi Lovato，Wilmer Valderrama，I)$\}$

2.2 因子图相关描述

因子图是一种基于无向图的概率模型。每一个候选的社交事件都对应着一个随机变量。该随机变量的值标识该事件的类型。遵循 Apoorv 和 Rambow 对社交事件的定义，本文

定义3种类型:交互型、观察型和一个表示该候选不是一个真正的社交事件的特殊类型"无"。本文中用 x_m^i 和 y_m^i 分别代表来自微博 t_m 的第 i^{th} 个候选事件及关联的随机变量。接下来,对于任两个候选项涉及两个相同的实体且来自相似的微博,引入一个连接这对候选项的因子。用 f_{mn}^{ij} 表示连接 y_m^i 和 y_n^j 的因子。这里,t_m 和 t_n 表示两条相似的微博。两条微博是相似的,当且仅当:1)它们属于同一个时间范围;2)正文的相似性高于某个阈值(0.3)。图1给出了因子图的一个例子。



圆圈代表随机变量,其值为对应的社交事件候选的真正类型。正方形代表因子,其中黑色和白色分别表示一个和两个随机变量相关的因子

图1 用于识别社交事件的因子图

利用微博的冗余性克服单条微博的不足,在其他针对微博的抽取任务中也得到了应用。例如,本文就是用该策略对微博进行命名实体识别和语义角色标注。值得强调的是,此时不是通过聚类加两遍标记而是通过概率图模型来利用冗余性。概率图模型是概率论与图论相结合的产物,为各种统计推理和学习提供了一个统一的灵活框架。信息抽取中得到了广泛应用的条件随机场、马尔可夫逻辑网等都是概率图模型的实例。在人工标注过的数据集上,本文提出的方法的F1为77%,而基准系统(基于支持向量机的分类器)的F1为54%。

3 基于因子图的社交事件抽取

本文提出的方法是基于因子图来进行社交事件的抽取:它同时确定一组相关的社交事件选项的真实类型。本节先总体介绍该方法,然后详细介绍它的各个关键模块。

3.1 联合推断方法概述

本文工作提出的方法包含两个步骤。第一步,产生社交事件候选。对每条微博 $t_m \in T$,为 t_m 中的每个人名对创建一个社交事件候选(p_1, p_2, y)。注意,即使 p_1 和 p_2 对在一条微博中出现多次,也只会为相距最近的对产生一个社交事件候选。用 x_m^i 表示来自 t_m 的第 i^{th} 个社交事件候选,$x_m^i \cdot p_1$ 和 $x_m^i \cdot p_2$ 分别表示该候选事件的 p_1 和 p_2,y_m^i 则表示该候选社交事件的真实类型。例如,针对"$[\text{Michael Lohan}]_p^1$ accuses $[\text{Lindsay Lohan}]_p$ of smoking crack and $[\text{Donald Trump}]_p$ Accuses $[\text{Jon Stewart}]_p$ of "Racist", $[\text{Michael Lohan}]_p^2$ again.",会生成如下所示的候选事件:($[\text{Michael Lohan}]_p^1$, $[\text{Lindsay Lohan}]_p$),($[\text{Lindsay Lohan}]_p$, $[\text{Donald Trump}]_p$),($[\text{Lindsay Lohan}]_p$, $[\text{Jon Stewart}]_p$),($[\text{Donald Trump}]_p$, $[\text{Jon Stewart}]_p$),($[\text{Donald Trump}]_p$, $[\text{Michael Lohan}]_p^2$)以及($[\text{Jon Stewart}]_p$, $[\text{Michael Lohan}]_p^2$)。

下一步,构建因子图 $G=(Y,F,E)$,其中:$Y=\{y_m^i\}_{m,i}$ 表示候选事件的真实类型的集合;F 是因子集合,包括两组因子,即$\{f_m^i(y_m^i)\}$和$\{f_{mn}^{ij}(y_m^i, y_n^j)\}$,$\forall x_m^i = x_n^j$;$E$ 表示所有边的集合,包括两类边,即连接 y_m^i 和 f_m^i 的边,连接 y_m^i 和 f_{mn}^{ij} 的边以及 y_m^i 和 f_{mn}^{ij} 的边。

$G=(Y,F,E)$ 按式(1)定义了条件概率 $P(Y/G,T)$:

$$\ln P(\frac{Y}{G},T) = -\ln Z(G,T) + \sum_{m,i}\ln f_m^i(y_m^i) + \sum_{m,n,i,j}\delta_{mn}^{ij} \cdot \ln f_{mn}^{ij}(y_m^i, y_n^j) \quad (1)$$

其中,$\delta_{mn}^{ij}=1$ 当且仅当 $y_m^i = y_m^j$,否则为0;$Z(G,T)$ 是规范化因子,其定义如式(2)所示:

$$Z(G,T) = \sum_{Y}\prod_{m,i} f_m^i(y_m^i) \cdot \prod_{m,n,i,j} f_{mn}^{ij}(y_m^i, y_n^j)^{\delta_{mn}^{ij}} \quad (2)$$

每个因子的对数可表示为式(3)定义的一组特征函数的加权和。

$$\begin{aligned} \ln f_m^i(y_m^i) &= \sum_k \lambda_k^{(1)} \phi_k^{(1)}(y_m^i) \\ \ln f_{mn}^{ij}(y_m^i, y_n^j) &= \sum_k \lambda_k^{(2)} \phi_k^{(2)}(y_m^i, y_m^j) \end{aligned} \quad (3)$$

其中,$\{\phi_k^{(1)}\}_{k=1}^{K_1}$ 和 $\{\phi_k^{(2)}\}_{k=1}^{K_2}$ 是两组特征。每个特征有一个实数值为其权重;$\Theta = \{\lambda_k^{(1)}\}_{k=1}^{K_1} \cup \{\lambda_k^{(2)}\}_{k=1}^{K_2}$ 表示特征权重的集合,也就是图模型 G 的参数。

本文提出的方法同时推断 Y。一旦 y_m^i 的值已知,可进一步输出 E_m:

$$E_m = \{(x_m^i, y_m^i) \mid \forall i, y_m^i \neq N\} \quad (4)$$

3.2 因子图训练

假定输入为 T,其中每个可能的社交事件都已经人工标出,训练的目的是确定模型参数 Θ,使得观察到的这些社交事件的对数似然最大。也就是求解式(5)定义的无约束线性化问题。

$$\Theta^* = \arg\max_{\Theta} \ln P(\frac{Y}{\Theta}, T) \quad (5)$$

为了求解该优化问题,先计算目标函数相对于模型参数的梯度:

$$\frac{\partial \ln P(\frac{Y}{T};\Theta)}{\partial \lambda_k^1} = \sum_{m,i}\phi_k^{(1)}(y_m^i) - \sum_{m,i}\sum_{y_m^i} p(y_m^i \mid T;\Theta) \cdot \phi_k^{(1)}(y_m^i) \quad (6)$$

$$\frac{\partial \ln P(\frac{Y}{T};\Theta)}{\partial \lambda_k^2} = \sum_{m,n,i,j}\delta_{mn}^{ij} \cdot \phi_k^{(2)}(y_m^i, y_n^j) - \sum_{m,n,i,j}\delta_{mn}^{ij}\sum_{y_m^i} p(y_m^i, y_m^j \mid T;\Theta) \cdot \phi_k^{(2)}(y_m^i, y_n^j) \quad (7)$$

其中,两个边缘概率分布 $p(y_m^i \mid T;\Theta)$ 和 $p(y_m^i, y_m^j \mid T;\Theta)$ 均可通过熟知的多圈可信度传递算法(LBP, Loopy Belief Propagation)[21]进行估算。该算法可信度传递算法迭代应用多次,每次迭代时,每个节点(因子节点和变量节点)并行地往相邻节点发送消息,如图2所示。

其中:$v_n \to m$ 由式(8)定义,表示变量节点 n 在取值为 y_n 时传递给因子节点 m 的消息;$u_m \to n$ 由式(9)定义,表示因子节点 m 在变量节点 n 取值为 y_n 时传递给该变量节点的消息。在式(8)和式(9)中,$N(n)$表示与变量节点 n 相连的因子节点,例如图2中,变量节点 n 相连的因子节点为 $m\Delta m'$ 和

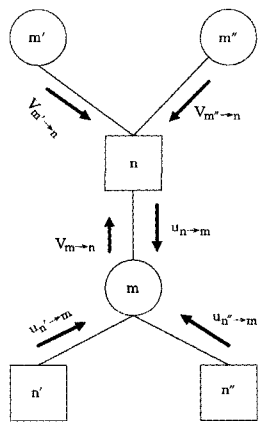
m''；$N(m)$表示与因子节点 m 相连的变量节点，例如图 2 中，因子节点 m 相连的变量节点为 $n\Delta n'$ 和 n''；Y_m 表示因子节点 m 连接的所有变量节点，例如图 2 中，Y_m 包括 $n\Delta n'$ 和 n''；f_m 表示因子节点 m 上定义的因子函数。

$$v_{n\to m}(y_n)=\prod_{i\in N(n),i\neq m}u_{i\to n}(y_n) \tag{8}$$

$$u_{n\to m}(y_n)=\sum_{\substack{Y_m\\ y_n}}f_m(Y_m)\cdot\prod_{i\in N(m),i\neq n}v_{i\to m}(y_n) \tag{9}$$

计算出消息后，根据式(10)就可以计算变量节点 n 的取值为 y_n 的边缘概率。

$$p(y_n)\propto\prod_{m\in N(n)}u_{m\to n}(y_n) \tag{10}$$



圆形表示变量节点，矩形表示因子节点。有两种消息：从变量到因子的消息以及从因子到变量的消息

图 2　因子图上的消息传递示意图

一旦算出梯度，就可以用基于梯度的标准的一维搜索技术例如最速下降法、共轭梯度法、拟牛顿法迭代计算出 Θ^*。拟牛顿法(Quasi-Newton methods)是求解无约束最优化方法最有效的一类算法[22]。它的一个突出优点是不需要计算二阶导数矩阵就具有超线性收敛速率，还具有 n 步收敛速率；其主要不足是所需存储量较大。本工作采用了拟牛顿法中著名的 BFGS 算法的受限内存版本(L-BFGS：the Limited-memory BFGS)[23]，它具有拟牛顿法的优点，又克服了其存储大的不足。

L-BFGS 算法按照式(1)计算下一个搜索方向，m 表示仅根据最近 m 词的迭代历史计算当前的下降方向，f 是待求最小值的函数，是当前海赛矩阵的逆矩阵的初始估计，$H_{(k)}^0$ 上标 T 表示矩阵或向量的转置运算。

3.3　因子图上下推理

给定一个测试集合 T，构建出因子图 G。假定该因子图的参数 Θ 已经解出，取值为 Θ^*，那么推理问题(也称解码问题)就是要找出 Y 的最有可能的赋值，也就是求解下面的优化问题：

$$Y^*=\arg\max_Y \ln P(Y|\Theta^*,T) \tag{11}$$

采用熟知的最大-乘积(MP，Max-Produce)算法来求解上述推理问题。该算法和 LBP 算法类似，只要把式(9)中的求和算子替换为最大化算子即可，如式(12)所示。

$$u_{m\to n}(y_n)=\max_{\substack{Y_m\\ y_n}}f_m(Y_m)\cdot\prod_{i\in N(m),i\neq n}v_{i\to m}(y_n) \tag{12}$$

3.4　社交事件分类特征

特征集合$\{\phi_k^{(1)}(y_m^i)\}_{k=1}^{K_1}$中的特征可分为两类：局部特征和全局特征。局部特征只与当前的微博 t_m 有关，包括：1)$x_m^i\cdot p_1$ 和 $x_m^i\cdot p_2$ 之间的单词数；2)$x_m^i\cdot p_1$ 和 $x_m^i\cdot p_2$ 是否在同一个句子；3)与 $x_m^i\cdot p_1$ 和 $x_m^i\cdot p_2$ 最近的动词，以及该动词相对于人名的位置，也即 L 或 R，其分别表示位于人名的左边和右边；4)t_m 是否含有哈希标签；5)$x_m^i\cdot p_1$ 的相邻词以及 $x_m^i\cdot p_2$ 的相邻词。

算法 1　L-BFGS 下降方向计算

输入：k，表示进行第 k 次迭代

输出：下降方向 $d^{(k)}$

1. $q=g^{(k)}$
2. for i=k−1,k−2,…,k−m do
3. $\alpha_i=\rho_i(s^{(i)})^Tq$
4. $q=q-\alpha_i y^{(i)}$
5. end for
6. $r=H_{(k)}^0 q$
7. for i=k−m,k−m+1,…,k−1 do
8. $\beta=\rho_i(s^{(i)})^Tr$
9. $r=r+s^{(i)}(\alpha_i-\beta)$
10. end for
11. return $d^{(k)}=-r$

全局特征是在训练集和测试数据集上收集到的统计信息，包括：1)$x_m^i\cdot p_1$ 和 $x_m^i\cdot p_2$ 在同一个句子以及同一条微博中的共现次数；2)$x_m^i\cdot p_1$ 和 $x_m^i\cdot p_2$ 的相似度。该相似度由式(13)计算，其中 $P_1(P_2)$表示 $p_1(p_2)$和在微博中共现过的人名的集合体。

$$\mathrm{sim}(p_1,p_2)=\frac{|P_1\cap P_2|}{|P_1\cup P_2|} \tag{13}$$

特征集合$\{\phi_k^{(2)}(y_m^i,y_n^i)\}_{k=1}^{K_2}$包括：1)$t_m$ 和 t_n 的内容相似度，该相似度根据式(14)计算，在表示成词袋向量之前，微博中的停用词被去掉；2)t_m 和 t_n 的发布时间是否在同一个时间范围内，例如一天或者 12 小时；3)t_m 和 t_n 是否包含一个相同的哈希标签；4)t_m 和 t_n 是否包含一个相同动词；5)t_m 和 t_n 之间是否存在恢复和 re-tweet 关系；6)t_m 和 t_n 是否包含相同的链接；7)$x_m^i\cdot p_1(x_m^i\cdot p_2)$与 $x_n^i\cdot p_1(x_n^i\cdot p_2)$是否有相同的相邻词。

$$\mathrm{sim}(t_m,t_n)=\frac{\vec{t_m}\cdot\vec{t_n}}{|\vec{t_m}||\vec{t_n}|} \tag{14}$$

有如下 3 点值得说明。首先，所有的实数类型的特征值都被转化为 1 或 0。转化规则如下，假定原来的是数值为 r，那么转化后，该特征取值为 1，当且仅当 $P_{norm}(x>r)\leqslant 0.2$。这里假定实值特征 $P_{norm}(\cdot|\hat{\mu}\hat{\sigma}^2)$满足式(15)定义的正态分布；其次，特征抽取前做了预处理，包括停用词移除、推特元数据(如哈希标签、连接等)抽取；最后，把所有的单词都转换为小写的词根。例如，“Dating”和“dated”都被转化为它们的词根“date”，因而两者被认为是相同的。如此一来，在微博“…[Demi Lovato]$_p$ Is Dating [Wilmer Valderrama]$_p$…”中，与“[Demi Lovato]$_p$”和“[Wilmer Valderrama]$_p$”最近的动词是

"Dating"的原型"date"。

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}r_i,\hat{\sigma}^2=\frac{1}{n-1}\sum_{i=1}^{n}(r-\hat{\mu})^2 \tag{15}$$

4 实验

本节在一个小规模的手工标注的数据集上对本文提出的基于因子图的社交事件抽取技术进行了评价。实验结果显示该系统优于基准系统,从而表明使用因子图来从微博中抽取社交事件是有效的。

4.1 实验数据准备

首先从2011年12月1日和2011年12月5日间的英语微博中抽样了1923个至少包含一个人名的微博。经过规范化处理之后,请两人独立为选出的微博标出人名以及所有可能的社交事件。按Kappa系数计算的各个人物的标注一致性为:人名识别0.78,是否是社交事件0.69,社交事件是交互型还是观察型0.78。标注人员会讨论每个标注不一致的情况,直到达成共识。

最后,标识出5128个不同的人名,获得926条微博至少包含一对人名。这些微博共包括4631个社交事件候选项,其中分别有212个和241个交互型和观察型社交事件。126条微博被随机选出,作为开发集,其他微博用来做5重交叉验证。

4.2 社交事件抽取评价指标

本工作把社交事件抽取的任务建模为一个分类问题,不是社交事件、交互型社交事件或观察型社交事件,因此采用评价分类系统性能时广泛采用的准确率、召回率和F1来评价在每个类上的性能。准确率衡量系统输出的标签正确的比率;召回率衡量标准数据集中的标签被正确标记的比率;而F1是准确率和召回率的调和平均,由式(16)所定义。平均准确率、平均召回率和平均F1被用来衡量系统的总体性能。每类事件所占比重和属于该类的实例数目成正比。

$$F1=\frac{2\cdot Precision\cdot Recall}{Precision+Recall} \tag{16}$$

4.3 微博社交事件抽取基准系统

和本工作提出的系统最相似的是Apoorv和Rambow提出的系统。但是该系统是在规范文本上训练的,并且使用了高级的语言学特征,例如和依存关系相关的特征。因此直接把系统应用到风格迥异的微博时,其性能较差。例如,在本工作人工标注的数据集上,它的F1是28.5%;作为对比,一个基于规则的系统在同样的数据集上的F1为35.0%。

遵循Apoorv和Rambow,本工作按如下方式构建了级联式风格的基准系统B_{ar}:首先构建一个分类器判定一个社交事件候选是否为社交事件;然后构建另一个分类器判定社交事件的类型(交互型或观察型)。两个分类器均基于支持向量机,都以$\{\phi_k^{(1)}(y_m^i)\}_{k=1}^{K_1}$为其特征。

另一个基准系统B_{nr}是本工作提出的系统的一个修改版本:它忽略了跨微博的特征$\{\phi_k^{(2)}(y_m^i,y_n^j)\}_{k=1}^{K_2}$。与$B_{ar}$不同,$B_{nr}$把社交事件检测和分类一体化实现。

为微博优化训练的词性标记器被用来识别动词。OpenNLP工具包被用来获得词的原型。

4.4 微博社交事件抽取基本结果

表2报告了本工作提出的系统和基准系统的平均准确率、召回率和F1。从该表可以看出,该系统整体性能超过了两个基准系统(统计显著测试$p<0.04$)。这意味着从多个微博中联合抽取社交事件是有效的。对实验结果的进一步分析表明,对相当一部分微博,基准系统未输出任何社交事件,而本工作提出的系统却能准确地输出它们中的社交事件。作为例子,考察下面两条内容相似且发布在同一个时间范围的微博:"…[Demi Lovato]$_p$ Is Dating [Wilmer Valderrama]$_p$ #1Dfacts …"和"… Love is in the air? looks like [Demi Lovato]$_p$ and [Wilmer Valderrama]$_p$ pretty close #1Dfacts …"。对第一条微博,所有的系统都成功地抽取到如下社交事件([Demi Lovato]$_p$,[Wilmer Valderrama]$_p$,I)。从第二条微博中抽取社交事件则比较困难,因为两个人名之间的动词"look"所提供的信息不足,并且这两个人名在人工标注数据集上仅出现两次,缺乏统计相关的证据。基准系统没有识别出任何社交事件,而本工作提出的系统能找出([Demi Lovato]$_p$,[Wilmer Valderrama]$_p$,I),因为该系统隐式地编码了下面的规则:相似的微博倾向于报道相似的社交事件。

表2 微博社交事件抽取总体结果

系统	准确率(%)	召回率(%)	F1(%)
FG	78	76	77
B_{ar}	69	48	56
B_{nr}	67	45	53

从表2还可以看出,B_{ar}稍优于B_{nr}(统计显著测试$p<0.05$)。这可能是由于社交事件候选的类型分布的不均导致的——90%以上的社交事件候选不是社交事件。因为数据分布的倾向性,导致B_{nr}比B_{ar}会更容易偏向"None"标签。

表3到表5报道了本工作提出的系统以及基准系统在每种类型上的分类准确率、召回率和F1。从该表可以看出,对任何一种类型的分类,基于因子图的系统都优于基准系统(统计显著测试$p<0.01$)。此外,对任何一种类型的分类,B_{ar}优于B_{nr}。这似乎表明B_{ar}比B_{nr}能更好地处理非平衡数据。

表3 针对非社交事件的分类结果

系统	准确率(%)	召回率(%)	F1(%)
FG	79	77	78
B_{ar}	68	49	57
B_{nr}	65	47	54

表4 针对交互型社交事件的分类结果

系统	准确率(%)	召回率(%)	F1(%)
FG	82	75	78
B_{ar}	67	49	56
B_{nr}	63	43	50

表5 针对观察型社交事件的分类结果

系统	准确率(%)	召回率(%)	F1(%)
FG	77	78	77
B_{ar}	71	46	56
B_{nr}	65	43	52

为了分别考察$\{\phi_k^{(1)}(y_m^i)\}_{k=1}^{K_1}$中的局部和全局特征的功能,修改本工作提出的系统,以仅使用局部或全局的特征方法来分析(也就是$\{\phi_k^{(2)}(y_m^i,y_n^j)\}_{k=1}^{K_2}$)。实验结果如表6所列。

首先注意到，仅使用局部特征，基于因子图的系统就超过了两个基准系统。其次，局部特征的贡献似乎大于全局特征，可能的原因是实验所用的数据集较小而不能提供可靠的统计信息。最后，局部和全部特征结合起来获得最好的性能。

表6 使用不同特征集时微博社交事件的分类结果

系统	准确率(%)	召回率(%)	F1(%)
Local	67	58	62
Global	42	38	40

结束语 微博包含丰富的社交事件。对这类事件进行结构化抽取，对一系列的应用如社会关系网自动构建、名人动态自动报道等均有帮助。从微博中抽取社交事件所面临的主要困难在于单条微博不能提供足够的信息。这个困难源自微博短小杂乱的本性。本文提出用因子图(一种概率图模型)同时从多条相关的微博中抽取社交事件，以充分利用微博的冗余性克服单条微博信息量的不足。特别地，该方法首先产生社交事件候选，然后在因子图上同时确定每个候选项的类型标签。初步的实验结果表明，该方法是有效的。

下一步，打算从3个方面进一步改进现有的方法：1)进一步开发微博规范化模块；2)尝试语义角色标注相关的特征；3)加大标注数据，并对人名做规范化处理。

参 考 文 献

[1] Wikipedia. Facebook user statistics [OL]. http://en.wikipedia.org/wiki/Facebook,2013

[2] Wikipedia. Twitter user statistics[OL]. http://en.wikipedia.org/wiki/twitter,2013

[3] How many Twitter Users Are There 2012[OL]. http://www.howmanyarethere.org/how-many-twitter-users-are-there-2012/2/,2013

[4] How Many Facebook Users Are There [OL]. http://www.howmanyarethere.org/how-many-facebook-users-are-there-2012/,2013

[5] Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets[C]//Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications. Association for Computational Linguistics,2004:104-107

[6] Xiao J, Su J, Zhou G, et al. Protein-protein interaction extraction:a supervised learning approach[C]//Proc Symp on Semantic Mining in Biomedicine. 2005:51-59

[7] Richardson M, Domingos P. Markov logic networks [J]. Machine learning,2006,62(1/2):107-136

[8] Casella G, George E I. Explaining the Gibbs sampler [J]. The American Statistician,1992,46(3):167-174

[9] McClosky D,Charniak E,Johnson M. Effective self-training for parsing[C]//Proceedings of the main conference on human language technology conference of the North American Chapter of the Association of Computational Linguistics. Association for Computational Linguistics,2006:152-159

[10] Yates A,Cafarella M,Banko M,et al. TextRunner:open information extraction on the Web[C]//Proceedings of Human Language Technologies:The Annual Conference of the North American Chapter of the Association for Computational Linguistics:Demonstrations. Association for Computational Linguistics,2007:25-26

[11] Grishman R, Westbrook D, Meyers A. NYU's English ACE 2005 system description [C] // Proc. ACE 2005 Evaluation Workshop. 2005

[12] Liao S,Grishman R. Using document level cross-event inference to improve event extraction[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics,2010:789-797

[13] Bethard S,Martin J H. Identification of event mentions and their semantic class[C]//Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics,2006:146-154

[14] Llorens H, Saquete E, Navarro-Colorado B. TimeML events recognition and classification:learning CRF models with semantic roles[C]//Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics,2010:725-733

[15] Yu L C,Chan C L,Lin C C,et al. Mining association language patterns using a distributional semantic model for negative life event classification [J]. Journal of biomedical informatics,2011,44(4):509-518

[16] Sankaranarayanan J,Samet H,Teitler B E,et al. TwitterStand:news in tweets[C]//SIGSPATIAL,GIS'09. New York,NY,USA:ACM Press,2009:42-51

[17] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users:real-time event detection by social sensors[C]//WWW'10. 2010:851-860

[18] Benson E, Haghighi A, Barzilay R. Event discovery in social media feeds[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:Human Language Technologies-Volume 1. Association for Computational Linguistics,2011:389-398

[19] Agarwal A,Rambow O. Automatic detection and classification of social events[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics,2010:1024-1034

[20] Ritter A, Clark S, Etzioni O. Named entity recognition in tweets:an experimental study[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics,2011:1524-1534

[21] Murphy K P,Weiss Y,Jordan M I. Loopy belief propagation for approximate inference:An empirical study[C]//Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc.,1999:467-475

[22] Fletcher R. Practical methods of optimization [M]. 1987

[23] Nocedal J. Updating quasi-Newton matrices with limited storage [J]. Mathematics of computation,1980,35(151):773-782

[24] 杨武，宋静静，唐继强. 中文微博情感分析中主客观句分类方法[J]. 重庆理工大学学报：自然科学版，2013，27(1)：51-56