基于动态认知的微博用户行为关系网络构建方法

赫熙煦[1,2]，陈雷霆[1,3]，张 民[4]，孙青云[4]

(1. 电子科技大学计算机科学与工程学院 成都 611731；2. 电子科技大学信息中心 成都 610054；
3. 东莞电子科技大学电子信息工程研究院 广东 东莞 523808；4. 电子科技大学图书馆 成都 610054)

【摘要】构建微博用户的社会关系网络是分析微博数据的重要基础手段之一。由于微博用户在信息的发布和传播过程中具有不确定的行为特性，导致常见方法无法有效地完成微博用户行为关系网络的建模。该文以不确定理论为基础，提出了基于Rough Set的动态认知技术，对微博的海量不完备信息进行处理，完成对用户行为的计算分析，构建了微博用户行为关系网络。并以此为基础，结合用户操作、主题与情感分析方法，对微博中的网络事件发展进行了分析。

关 键 词 动态认知； 主题检测； 用户行为关系网络； 微博
中图分类号 TP391.41 文献标志码 A doi:10.3969/j.issn.1001-0548.2018.02.016

A Method to Construct Weibo User Behavior Relationship Network Using Dynamic Cognitioin

HE Xi-xu[1,2], CHEN Lei-ting[1,3], ZHANG Min[4], and SUN Qing-yun[4]

(1. School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 611731;
2. Center of Information, University of Electronic Science and Technology of China Chengdu 610054;
3. Dongguan Institute of Information Engineering, University of Electronic Science and Technology of China Dongguan Guangdong 523808;
4. Library of University of Electronic Science and Technology of China Chengdu 610054)

Abstract To construct the social network of Micro-blog users has become one of the most important method to analyze micro-blog data. However, due to the uncertainty of the behavior of users in the process of information release and dissemination, it is hard to construct effectively the social network of Micro-blog users. Based on the uncertainty theory, the paper proposes a rough set based dynamic cognitive technology to handle the incomplete and massive information of micro-blog, complete the calculation and analysis of behavior of users, and construct the social network of Micro-blog users. On this basis, the development of network events in micro-blog is analyzed combined with the method with operation, theme and emotion analysis.

Key words dynamic cognition; theme detection; user behavior relationship network; weibo

博客、微博、微信等自媒体产生了海量数据，并加速了信息传播，特别是对突发事件及重大事件的传播产生了重要影响。关注突发事件及重大事件的网络传播规律及网民行为关系，有助于舆论的正面引导，维护社会稳定。因为此类数据大多数以海量不确定数据来呈现，所以进行深度的数据挖掘和分析难度较大。本文提出在海量微博数据上进行不确定性数据挖掘和分析，进而构建微博用户行为网络，来实现更深入的数据价值获取。

文献[1]通过复杂网络及网络动力学理论分析Twitter用户关系网络特性中的可行性，并完成了Twitter用户关系网络基本参数的计算。文献[2]通过对新浪微博用户关系网络的研究，发现了该网络是典型的复杂网络，具有小世界、无标度和高聚类的特性。文献[3]认为新浪微博网络结构满足幂律分布。

Rough Set理论是一种处理不确定性信息的基础理论。基于Rough Set的认知挖掘是当前的研究热点。文献[4]描述了社会网络与粒计算的关系。文献[5]提出一种动态维护近似W.R.T对象的方法，并添加属性到粗糙集决策理论的框架中。文献[6]使用粗糙集和粒计算等相关技术对社会网络进行了建模。文献[7]使用粗糙集解决社交网络中的分类和聚类问题。文献[8]使用模糊集，对海量社交网络数据进行情绪分析，并使用facebook进行了验证。文献[9]使用模糊综合评价方法对CPM算法进行改进，对微博主题进行发现。

收稿日期：2016 – 12 – 24；修回日期：2017 – 12 – 19
作者简介：赫熙煦(1982 –)，男，高级工程师，主要从事计算机网络、高性能计算和数据挖掘方面的研究.

本文拟采用Rough Set理论，对微博的主题和用户情绪进行动态认知，进而构建微博用户行为关系网络，得出微博事件发展演化的路径。

1　微博用户行为关系网络的概念

行为关系网络是一种Web社会网络，它是描述用户行为关系的抽象网络。微博事件演化过程在行为关系网络中，以时间顺序进行表达。本文以新浪微博用户作为研究对象，选择发布、评论、转发和回复4种操作方式进行研究，构建了某一网络事件中微博用户之间形成的行为关系网络。

网络事件的演化是用户行为相互影响作用的结果。用户的行为特征在一定程度上反映了用户的活跃程度、理性程度和兴趣模型。通过主题跟踪，可以将同一话题相关的事件按照时间顺序关联起来，同时监控事件发展的空间(用户行为)变化。

因此，本文使用事件监测算法对互联网内特定的用户群数据进行分析处理，形成事件关联网络发展脉络；将已识别的事件进行训练后得到微博事态发展模型，收集后续相关事件进行时间和空间的关联分析，最终形成如图1所示的行为关系网络。

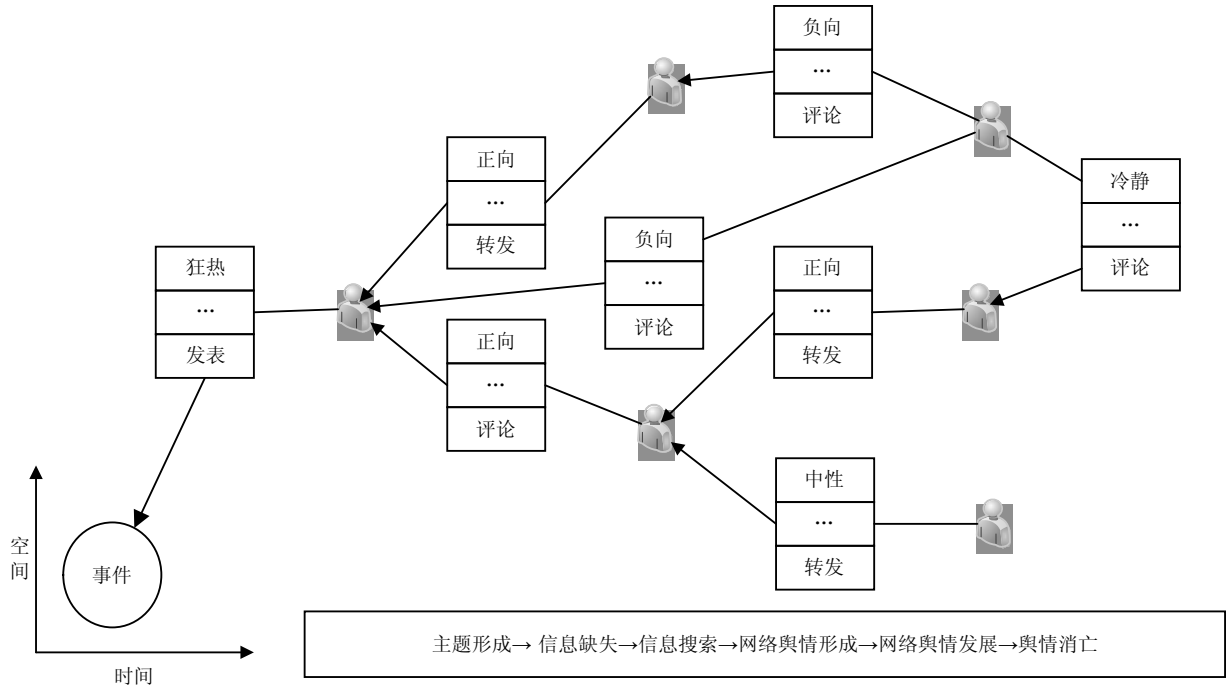

图1　用户行为关系网络示意图

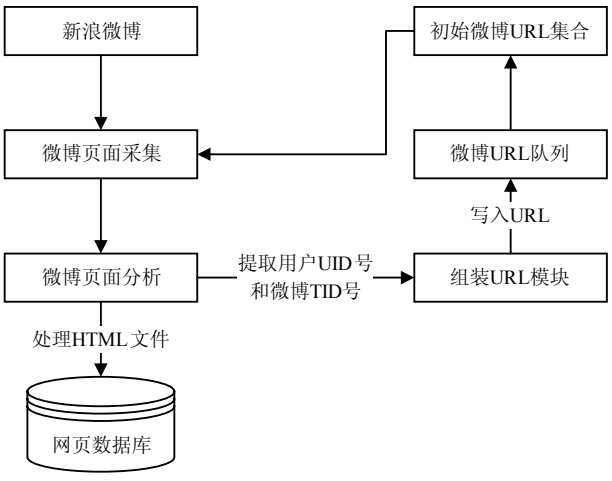

图2　微博数据采集过程

本文中微博数据采集采用了文献[10]中所提出的基于模拟登录的数据采集方案。数据采集过程如图2所示。

2　用户主题意向与情绪提取方法

认知用户行为需要对用户参与微博事件的过程中，所进行的操作类型、发表的内容及包含情绪等信息进行建模分析，形成用户行为的动态认知。依据该动态认知构建用户行为关系网络，完成对事态发展的监测和预判。

本文采用主题模型(topic model)对用户发表的内容和包含的情绪进行提取。它是源于隐性语义索引(latent semantic indexing, LSI)[11-12]，被广泛应用于主题挖掘、文本检索、文本分类、引文分析和社交网络分析等领域。

本文使用提取关键字等方法，对主题进行识别。广告性质的短语和一些微博没有评论内容会从待分

析数据中被剔除。对微博进行主题分析之前，还需要对其进行分词处理。此外，还在分词库中添加了一些常用的网络用语，以提高分词的准确性。在分词基础上，增加了停用词去除的代码，将对主题无影响的停用词从词库中去除，以提高主题分析的效率和准确性。其流程如图3所示。

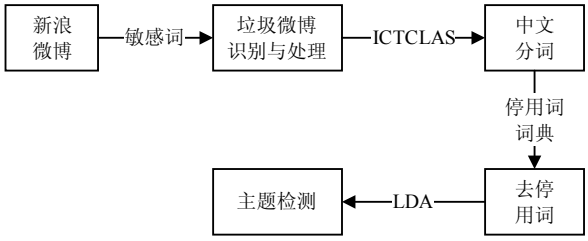

图3 微博主题检测流程图

在某个热点事件发生之后，互联网用户能够通过微博迅速获取事件信息，并进行反馈和传播。文献[13]通过对微博情感表达的研究，提出一种方法描述微博中正、负和矛盾的情感。文献[14]通过复杂系统理论处理在线个人情感，并探讨了微博背后的情感表达机制。

本文提出的方法能够自动分析微博数据中用户帖子所包含的情感倾向，监测用户群整体的情感变化趋势。首先，抽取事件中所包含的不同方面的关注点；然后，检测不同关注点相关的帖子中所包含的用户情感信息；接着，统计用户群对各个关注点的情感变化趋势。

以极性词典为基础，对情感极性进行判断。本文实验使用知网提供的正面、负面情感词及评价词词典。在微博的评论信息中，增加了一些流行的网络用语。情感词典中包含4 495个正极性词汇和4 376个负极性词汇。

3 基于Rough Set的用户行为动态认知算法

本文使用Rough Set理论来分析用户在微博事件中的主题意向、操作和情感等因素来获取对用户行为关系网络的认知，从而更好地处理信息模糊化难题。

对于一个属性分类集合 $K=(U,R)$，其中任意的属性子集 $X\subseteq U$ 和分类等价关系 $R\in \mathrm{ind}(K)$ 可以获得两个Rough Set基础子集：

$$R_*(X)=\{x\in U:[x]_R\subseteq X\} \tag{1}$$

$$R^*(X)=\{x\in U:[x]_R\cap X\neq\varnothing\} \tag{2}$$

上面形成的 $\mathrm{pos}_R(X)=R_*(X)$ 可以认为是由属性 X 获得核心域，而 $R^*(X)=\mathrm{pos}_R(X)\bigcup \mathrm{bn}_R(X)$ 则是属性 X 获得的可支持域，因此可得 $\mathrm{bn}_R(X)=R(X)-R_*(X)$ 是 X 的边界域。其中核心域 $R_*(X)$ 表示可以从属性 X 中获得的关于 K 所有的精确认识，而 $R^*(X)$ 表示可以从属性 X 关于 K 所有信息，包括不确定性信息。

在上近似集和下近似集合之间的元素是由于通过等价关系 R 并不能完全地确定其在子集 X 之中。对于这些元素可以称为 X 的 R 边界集，记为：

$$\mathrm{BN}_R(X)=R^*(X)-R_*(X) \tag{3}$$

定义 1 对于论域 U，等价关系簇 P 中如果存在一个等价关系簇 Q，且满足：

$$Q\subseteq P \tag{4}$$

$$\mathrm{IND}(Q)=\mathrm{IND}(P) \tag{5}$$

定义 2 对于论域 U 不同的等价关系簇 P 和 Q，称下式所求解为等价关系簇 Q 相对 P 的正域：

$$\mathrm{POS}_P(Q)=\bigcup_{X\in U/R} P_-(X) \tag{6}$$

定义 3 对于论域 U 不同的等价关系簇 P 和 Q，如果等价关系簇 P 中存在等价关系 r 满足：

$$\mathrm{POS}_P(Q)=\mathrm{POS}_{P\setminus\{r\}}(Q) \tag{7}$$

则称等价关系 r 为等价关系簇 P 中相对于等价关系 Q 中可以约简的；反之则是不可约简的。

时间顺序是观察事件发展的重要维度，故本文提出动态特征分析方法来构建属性。该方法将在每个属性上一个时间窗口，统计该窗口的内属性的变化率进而进行分析。

设论域 $U=\{x_1,x_2,\cdots,x_n\}$，其中存在属性域 $C=\{c_1,c_2,\cdots,c_m\}$，即对于每一个粒子可采用 m 个属性值来进行描述。不同的属性组合形成知识 R_k，依据知识 R_k 可以对当前的粒层形成凝聚，知识 R_k 之间所存在的蕴含关系可以对应生成相应的粒结构。

对于给定的论域 U 上，存在决策系统 S，等价关系簇 D 为决策属性集，等价关系簇 P 为条件属性集。那么，$\mathrm{SGF}(r_i,P,D)$ 是条件属性 r_i 在等价关系簇 P 的条件属性的重要度：

$$\mathrm{SGF}(r_i,P,D)=\frac{\mathrm{card}(\mathrm{POS}_{P\cup\{r_i\}}(D))-\mathrm{card}(\mathrm{POS}_P(D))}{\mathrm{card}(\mathrm{POS}(D))} \tag{8}$$

$I(r_i,D)$ 是条件属性 r_i 相对于决策等价关系簇 D 的互信息熵：

$$I(r_i,D)=H(D)-H(D\,|\,\{r_i\}) \tag{9}$$

从属性重要度和互信息熵的定义中可以看出，$\mathrm{SGF}(r_i,P,D)$ 越大，条件属性 r_i 所提供分类的信息量就越大，所获得粒度越大。

动态属性认知可以定义为：设有决策信息系统

$S=\langle U,C\cup D,V,f\rangle$，有属性等价关系簇 R 和属性等价关系簇 C，且 $R\subset C$。对于已获属性取值的样本 $R(x_p)$，R 相对于 C 的补集 R^c 中属性 a_j 的动态属性认知为：

$$\mathrm{sig}_I(a_j \mid R(x_p))=\frac{\mathrm{card}(\mathrm{POS}_{U/\{a_j\}|R(x_p)}(D))}{\mathrm{card}(\mathrm{POS}_{U/\{R(x_p)\}}(D))} \qquad a_j\in R^c \tag{10}$$

式中，对于 $a_j\in R^c$ 属性而言，条件属性 a_j 提供了动态属性认知。基于上述属性认知可以获得对应的粒层分析，形成相关的粒层用于智能分类等应用。

利用上述模型可以计算所提取每一个用户行为属性对整体事件的影响情况，进而获得每个用户在网络行为中的重要性，从而实现对网络行为关系的构建。

4　实验与分析

以“招商银行济南招聘”事件为例，进行微博用户操作、主题意向和情绪分析，从而构建用户行为关系网络。本文采用模拟登录方法，在2012年12月15日～2013年1月7日期间，获取到30 196条有效记录，包含14 569位微博用户。

通过分析，得到如图4所示的用户操作统计。

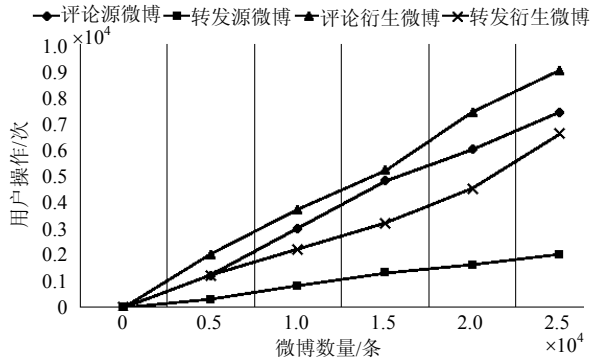

图4　招商银行济南招聘事件全过程用户操作统计

图4中，可以发现用户对源微博和衍生微博的评论次数大于转发数量，说明大多数用户在社会网络构建的过程中，有强烈的意愿来发表自己的意见，使得事件在很短的时间内形成较大的规模。

通过本文提出的动态认知方法，构建了如图5所示的行为关系网络。

用户行为关系网络图中结点按照出现的时间顺序进行编号。从中可以看出一些用户的操作特点，如对源微博多是进行评论，从这些用户结点的编号可以看出出现的时间顺序分布比较均匀。该现象说明随着评论的增加，用户会对和自己观点一致的衍生微博进行转发和评论，尤其是该微博的粉丝和出现时间接近的用户；对于回复的操作，该图中只出现了4个用户。一般来说回复操作是若干个用户对衍生微博中观点的讨论，甚至是争论。其中，正向情感1 512人次，负面情绪2 386人次，中性情绪11 003人次。

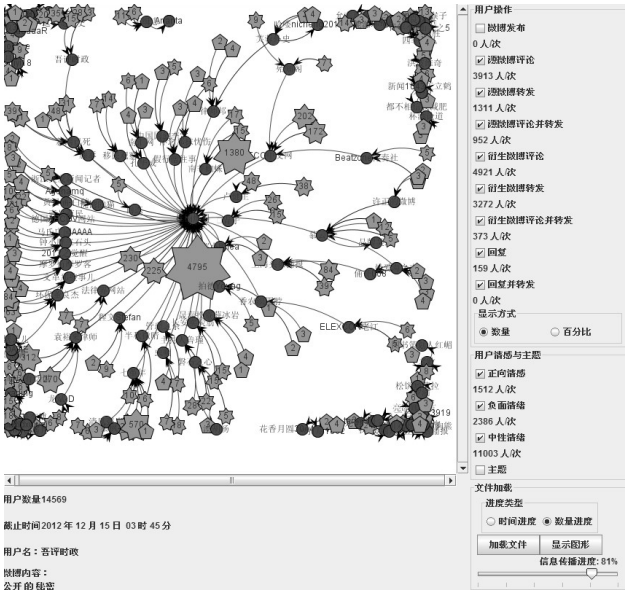

图5　招商银行济南招聘事件行为关系网络示意图

5　结 束 语

本文通过对微博网络事件的分析和研究，针对微博用户特性动态建模，形成动态认知，依据用户动态认知来形成关系网络，发掘网络事件发展和传播的潜在规律，实现用户行为关系网络建模的目标。本文所提出的方法对网络舆情的了解和网络事件发展的预判提供了一定的参考。

参　考　文　献

[1] TEUTLE A R M. Twitter: Network properties analysis[C]// Electronics, Communications and Computer. Cholula: IEEE, 2010: 180-186.

[2] KANG S, ZHANG C, LIN Z, et al. Complexity research of massively microblogging based on human behaviors[C]// Database Technology and Applications. Dalian, China: IEEE, 2010: 1-4.

[3] FAN P, LI P, JIANG Z, et al. Measurement and analysis of topology and information propagation on Sina-Microblog[C]//Intelligence and Security Informatics. Beijing, China: IEEE, 2011: 396-401.

[4] LIAU C J. Social networks and granular computing[J]. Encyclopedia of Complexity and Systems Science, 2009(1): 8333-8345.

[5] CHEN H, LI T, LUO C, et al. A decision-theoretic rough set approach for dynamic data mining[J]. IEEE Transactions on

Fuzzy Systems, 2015, 23(6): 1958-1970.
[6] YAGER R R. Intelligent social network modeling and analysis[C]//Intelligent System and Knowledge Engineering. Xiamen, China: IEEE, 2008, 1: 5-6.
[7] MITRA A, SATAPATHY S R, PAUL S. Clustering analysis in social network using covering based rough set[C]// Advance Computing Conference. [S.l.]: IEEE, 2013, 8628: 476-481.
[8] MUKKAMALA R R, HUSSAIN A, VATRAPU R. Fuzzy-set based sentiment analysis of big social data[C]// Enterprise Distributed Object Computing Conference. [S.l.]: IEEE, 2014, 1: 71-80.
[9] CHEN Xiao-lei, CHEN Xiang, CHENG Yi-jie. Community structure discovery and community topic analysis in microblog[C]//International Conference on Information Management, Innovation Management and Industrial Engineering. Xi'an, China: IEEE, 2013, 1: 590-595.
[10] 孙青云, 王俊峰, 赵宗渠, 等. 一种基于模拟登录的微博数据采集方案[J]. 计算机技术与发展, 2014, 24(3): 6-10.
SUN Qing-yun, WANG Jun-feng, ZHAO Zong-qu, et al. A microblog data collection method based on simulated login technology[J]. Computer Technology and Development, 2014, 24(3): 6-10
[11] DEERWESTER S, DUMAIS S T, FURNAS G W, et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391.
[12] XU Ge, WANG Hou-feng. The development of topic model in natural language processing[J]. Chinese Journal of Computers, 2011, 34(8): 1423-1436.
[13] HU Y, ZHAO J, WU J, et al. On exploring ambivalent expression in Weibo[C]//Service Systems and Service Management. Guangzhou, China: IEEE, 2015: 1-6.
[14] ZHOU J, ZHAO Y, ZHANG H, et al. Measuring emotion bifurcation points for individuals in social media[C]// System Sciences. Kauai, Hawaii: IEEE, 2016: 1949-1958.

编 辑 蒋 晓