

Twitter 的 Follow 关系和 Retweet 关系对比

曾雪, 吴跃

(电子科技大学 计算机科学与工程学院, 成都 610000)

摘要: 研究 Twitter 在线社交网络中, Follow 关系和 Retweet 关系在传播用户影响力和表征用户同质性这两方面的差异。为研究两者在传播用户影响力上的差异, 定义了 V_f 变量和 V_r 变量分别度量 Follow 关系和 Retweet 关系的作用; 为研究两者在表征用户同质性上的差异, 分别基于 Follow 关系和 Retweet 关系构造出对应的社交网络图, 并采用 wvRN 算法分别对两个网络内的用户进行分类。通过对比用户的 V_f 变量值和 V_r 变量值发现, Retweet 关系在传播用户影响力方面的作用优于 Follow 关系; 通过对比分类结果发现, Follow 关系比 Retweet 关系更能表征用户的同质性, 基于 Follow 关系的分类精度比基于 Retweet 关系的分类精度高 20%, 分类结果同时揭示不同类别的用户体现出了不同的关注和信息互动特性。基于上述研究说明 Follow 关系和 Retweet 关系所携带的信息是不同的。

关键词: 在线社交网络; 网络数据分类; 同质性; 推特网

中图分类号: TP391 **文献标志码:** A **文章编号:** 1001-3695(2014)01-0192-04

doi: 10.3969/j.issn.1001-3695.2014.01.044

Comparison between Follow and Retweet on Twitter

ZENG Xue, WU Yue

(School of Computer Science & Engineering, University of Electronic Science & Technology of China, Chengdu 610000, China)

Abstract: This work aimed to compare the differences between Follow relationship and Retweet relationship in propagating the Twitter users' influence and measuring the Twitter users' homophily. In order to find their differences in the first aspect, this paper defined two verities: V_f was defined for the Follow relationships and V_r was for the Retweet relationships. In order to analyze the differences in measuring the homophily, it built two social network separately based on Follow relationships and Retweet relationships. Then it adopted the wvRN algorithm to classify users in these two networks. By comparing the V_f and V_r values for a group of user, it can be found that the Retweet relationship win in propagating the user's influence. By comparing the classification results, it can be found that the Follow relationship contains more homophily information, that the accuracy based on Retweet is 20% higher than that based on Retweet. But the results also illustrates that different kinds of users have different communication characteristics. This research shows that Follow relationships and Retweet relationships contain different information.

Key words: online social network; networked data classification; homophily; Twitter

0 引言

在社交网络高度发展的今天, 在线微博服务成为人们进行实时信息交流的工具。人们通过在微博上关注自己感兴趣的用来获取遍布世界各地的信息; 与此同时, 人们还将自己从本地获悉的信息以文字、图片、视频、动画等多种形式发布到微博上。在微博上传播的信息涵盖了人们的日常活动、政治言论、社会新闻和艺术文化等众多领域。随着注册微博服务的用户日益增多, 在线微博网站已经不单单是人们进行实时信息交流的工具, 更逐渐成为一种新型的公共传媒平台。Twitter 微博网站便是众多在线社交网络服务中的一个。作为当前在美国乃至全世界最为流行的在线微博服务网站, Twitter 截止 2012 年底已经拥有 10 亿活动用户。美国几乎所有的政治、商业机构、民间团体都拥有 Twitter 主页, 包括奥巴马在内的美国政要也注册了 Twitter 账户作为其官方信息发布平台。Twitter 之所

以被如此广泛地运用和关注, 得利于其实时的信息传播力。由于其活跃用户众多, 影响力大, 其用户所构成的社交网络在当前在线社交网络中具有代表意义, 因此已经有大量的研究针对 Twitter 网展开。

Michelson 等人在文献[1]中利用 Twitter 用户的 Mention 关系挖掘各个用户当前感兴趣的话题, 并据此为每个用户创立账户文档, 以完善用户信息; Masckassy 等人在文献[2]中研究了 Retweet 关系在信息传播过程中所起的作用。文献[3]从 Twitter 用户列表中提取用户感兴趣的话题, 发现相互关联的用户倾向于关注相同的话题。文献[4]分别运用基于文本的分类方法和基于网络的分类方法对 Twitter 上的热门话题进行归类, 发现基于网络的分类方法取得的分类效果略好。文献[5]研究了一种基于标签的用户搜索引擎, 该系统首先从 Twitter 用户列表中提取标签, 依据用户信息将这些标签赋予用户后, 系统将基于这些标签搜索匹配到的用户。文献[6]研究了

收稿日期: 2013-03-20; 修回日期: 2013-04-22

作者简介: 曾雪 (1982-), 女, 四川成都人, 博士研究生, 主要研究方向为在线社交网络分析、关系数据分类 (xuezheng68@gmail.com); 吴跃 (1953-), 男, 教授, 博导, 主要研究方向为数据库、数据挖掘。

Twitter的信息传播实时性,并将Twitter上最具影响力的用户划分成三类,该研究指出,Twitter在信息传播方面比传统媒体更高效快捷。包含上述研究在内的很多现有资料都集中在挖掘Twitter上具有强影响力的核心用户或分析Twitter用户间的同质性上。然而,Twitter用户间存在多种关联关系,哪些关系对上述问题更具信息量?哪些关系能更好地帮助研究者找到有影响力的用户和对用户的类别进行划分?据笔者了解,目前对这些问题尚缺乏足够的研究。

鉴于此,本文将重点讨论Twitter在线社交网络中,Follow关系和Retweet关系在传播用户影响力和表征用户同质性这两个方面的差异。Twitter用户之间存在着多种联系,最常见的是Follow关系,除此之外由于用户之间的相互转发和提及等行为,还可以构成其他关系,如Retweet关系和Mention关系等。由于这些关系形成的因素不同,其中蕴涵的信息也不尽相同。文献[7~9]针对多个社交网络数据集中的不同关联关系展开研究,研究发现由于基于的关联关系不同,在经过完全相同的分类处理后,同一数据集上获得的分类效果间可能出现显著差异。这说明对同一个点集,选用不同类型的关联关系所构成的网络结构可能相差很大,因而其携带的信息也就存在明显差异。为了比较Follow关系和Retweet关系在传播用户影响力上的差异,本文定义了 V_f 和 V_r 两个变量来分别度量这两种关系对传播用户影响力的作用;为比较这两种关系在表征用户同质性方面的差异,本文分别基于这两种关系,采用wvRN分类算法对同一群用户进行了分类实验。实验结果表明,Retweet关系在第一个方面优于Follow关系,而在第二个方面Follow关系更具信息量。实验结果同时显示,隶属于不同类别的用户,其关注朋友和信息交互的模式是不同的。

1 Follow和Retweet在传播用户影响力上的差异

Twitter用户可以于任意时间在自己的主页上发布短小的消息,除个别设置了权限的用户外,所有消息都公共可见。Twitter用户之间可以单向或相互订阅,这种关系又称为关注(Follow);此外,还存在其他的信息交互关系,如提及(Mention)、转发(Retweet)等。由于篇幅关系,本文仅讨论最为常见的Follow关系和Retweet关系。

1.1 Follow关系和Retweet关系的出入度分布

Follow关系是指当用户 b 关注用户 a 时,用户 a 所发布的所有信息都能实时地被用户 b 获取。这时,称用户 a 和用户 b 之间存在Follow关系。

Retweet关系是指当用户 a 发布的原始信息被用户 b 转发时,用户 b 将该消息重新发布在自己的页面上,并引用用户 a 的用户名(@ a)。这时,用户 a 和用户 b 之间就存在Retweet关系。

直观上讲,Follow关系表示关注者(用户 b)对被关注者(用户 a)的整体情况感兴趣;而Retweet关系则表示转发者(用户 b)对被转发者(用户 a)所发布的特定信息感兴趣。由此可推想Follow关系与Retweet关系所携带的信息应当存在着一些差异。文献[10]提取了100 000名Twitter用户,并分别绘制了Follow关系和Retweet关系的出入度分布图,如图1、2所示。

图1中,(a)展示了Follow关系的入度分布;(b)展示了Follow关系的出度分布。由图可见,Follow关系的入度分布完全符合PowerLaw;其出度分布也基本符合PowerLaw规则。但是在 $10^{1.4}$ 和 10^3 这两点,分别出现了两次激增,这两个激增点分别是由于Twitter的用户推荐策略和关注限制策略造成的。

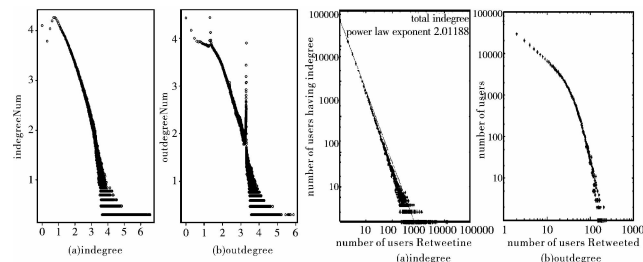


图1 关注(Follow)关系的
出入度分布

图2 转发(Retweet)关系的
出入度分布

图2中,(a)展示了Retweet关系的入度分布;(b)展示了Retweet关系的出度分布。由图可见,Retweet关系的入度分布完全符合PowerLaw,但是Retweet的出度关系却并不符合PowerLaw。对比图1、2可知,Follow关系和Retweet关系在出度分布上存在较大差异。Follow关系的出度分布显示,仅有极少数用户关注了大量的朋友。然而Retweet关系的出度分布显示,大多数Twitter用户都倾向于进行信息转发;通过进一步查看用户的转发对象还可以发现,很多用户转发的信息并不是其关注的朋友的信息,即转发者和被转发者之间仅存在单纯的Retweet关系而不存在Follow关系。这更进一步说明,Retweet关系可能隐含了不同于Follow关系的信息。

1.2 Follow和Retweet在传播用户影响力上的差异

Twitter用户通常会从与自己关注的朋友相关联的其余用户中挑选出新朋友加以关注。这种与其朋友间存在的关联关系可能是Follow关系,也可能是Retweet关系。为了度量这两种关系对用户添加新朋友所产生的作用,本文定义了 V_f 、 V_r 两个变量, V_f 用以估计Follow关系对传播用户影响力的作用; V_r 用以估计Retweet关系对传播用户影响力的作用。

定理1 假定用户 a 在时间点 t 所关注(Follow)的所有朋友共有 m 个,分别为 $a_1, a_2, a_3, \dots, a_m$,且这 m 个朋友所构成的集合为 $F_{a(t)}$,即 $F_{a(t)} = \{a_1, a_2, a_3, \dots, a_m\}$;又假设 a 的任意一个朋友 a_n 在 t 时刻关注的朋友构成的集合为 $F_{a_n(t)}$,则集合 $F'_{a(t)} = \{F_{a_1(t)} \cup F_{a_2(t)} \cup F_{a_3(t)} \cup \dots \cup F_{a_m(t)}\}$ 表示用户 a 因Follow关系而产生的所有潜在的朋友所构成的集合;又假设 a 在 $t+1$ 时刻的朋友集合为 $F_{a(t+1)}$,其在 $t+1$ 时刻新增的朋友即为 $F_{a(t+1)} - F_{a(t)}$ 。则

$$V_f = |(F_{a(t+1)} - F_{a(t)}) \cap F'_{a(t)}| / |F'_{a(t)}| \quad (1)$$

定理2 假定用户 a 在时间点 t 所关注(Follow)的所有朋友共有 m 个,分别为 $a_1, a_2, a_3, \dots, a_m$,且这 m 个朋友所构成的集合为 $F_{a(t)}$,即 $F_{a(t)} = \{a_1, a_2, a_3, \dots, a_m\}$;又假设 a 的任意一个朋友 a_n 在 t 时刻转发(Retweet)的朋友构成的集合为 $R_{a_n(t)}$,则集合 $R'_{a(t)} = \{R_{a_1(t)} \cup R_{a_2(t)} \cup R_{a_3(t)} \cup \dots \cup R_{a_m(t)}\}$ 表示用户 a 因Retweet关系而产生的所有潜在的朋友所构成的集合,又假设 a 在 $t+1$ 时刻的朋友集合为 $F_{a(t+1)}$,其在 $t+1$ 时刻新增的朋友即为 $F_{a(t+1)} - F_{a(t)}$,则

$$V_r = |(F_{a(t+1)} - F_{a(t)}) \cap R'_{a(t)}| / |R'_{a(t)}| \quad (2)$$

V_f 估算了当用户 a 的任意一个朋友关注了一个新用户

时, a 也随之关注该新用户的可能性; 而 V_r 估算了当用户 a 的任意一个朋友转发了一个新用户的消息时, a 也随之关注该新用户的可能性。 V_f 和 V_r 的值越大, 说明其对应的关系在传播用户影响力方面所起到的效果就越大。 本文将在实验部分针对同一群用户计算这两个变量并进行对比分析。

2 基于 Follow 和 Retweet 的 Twitter 用户分类

用户的同质性分析在对用户进行产品推荐、划分用户群等多个应用领域都具有重要作用。 现有的大量研究表明, 社交网络中的用户具有同质性, 即具有相同特性的点更倾向于相互关联。 文献[8]研究了包括社交网络在内的多个关联数据集的同质性, 采用了一种称为 wvRN 的分类算法对标记后的关联数据集进行分类, 对每一个关联数据集都基于两种不同的关联关系分别进行分类预测。 通过对比基于不同关联信息而获得的分类结果后发现, 不同关联关系所携带的同质性信息是不同的, 某些关联关系更能表征关联数据的同质性, 从而基于这些关联关系所获得的分类效果就更好。

本章将介绍如何利用用户列表对用户预先赋予唯一的类别标志, 将讨论如何分别基于 Follow 关系和 Retweet 关系, 采用 wvRN 算法对 Twitter 的用户进行分类。 通过对比其分类结果, 可以获悉这两种关系所包含的用户同质性信息的差异。

2.1 利用用户列表对用户预赋类别标签

Twitter 用户通常拥有一个或多个用户列表(list), 用户通过对自己关注的朋友进行分类, 一个类别对应一个 list, 一个 list 中可以有 0 到多个朋友。 每个 list 都有一个 list 名, 对同一个用户而言, 其关注的朋友可能仅划分到一个 list 中, 也可能同时存在于该用户的多个 list 中。 不同的用户间, 由于划分的依据不同和对 list 的命名习惯不同, 其拥有的 list 往往不尽相同。 但是, 用户的 lists 之间往往存在一些共性类别, 如有关医疗、食物、运动、经济的类别等。 基于 Twitter 的用户列表, 本文采用了如下方法对用户进行预先分类:

a) 提取所有用户的列表信息, 将所有列表名称列入一个文档中;

b) 用文本分析工具对列表名进行划分, 将具有类似意义的列表名称划分到同一文档;

c) 将划分到同一文档的列表名命名为一个类别, 并将隶属于这些列表的所有用户都无区别地划归到该类别下;

d) 为了保证一个用户仅拥有一种类别, 删除所有隶属于两个及两个以上类别的用户。

通过上述处理, 可以得到一组列表, 且这些列表中的用户没有重合。

2.2 利用 wvRN 算法对用户进行分类

wvRN 算法是一种关联数据分类算法, 该算法假定相互关联的数据间具有同质性, 在分类过程中, 该算法可以仅仅根据网络结构信息进行分类。 所以该算法能够在最大程度上利用关联关系所携带的同质性信息进行分类。 分类效果好, 说明网络结构的同质性好, 其基于的关联关系所携带的同质性信息就多; 反之, 说明该网络结构的同质性较少, 其基于的关联关系所携带的同质性信息相对就少。

给定图 $G = (V, E, X)$, $v_i \in V$ 是该图中任意点, $e_{ij} \in E$ 是

连接点 v_i 和 v_j 的边, X_i 是点 v_i 的类别标签。 X_i 可取的分类值共 m 个, 分别为 $C_i = \{c_1, c_2, c_3, \dots, c_m\}$; v_i 属于且仅属于这 m 个类别中的一个。 假定 V_k 为已知类别的点集合, 则 $V_u = V - V_k$ 代表未知点标签的点集合。 wvRN 将基于 $G_k = (V, E, X_k)$ 来预测 V_u 中点的标签。 假设 N_i 代表点 v_i 的邻居集合, z 为平滑系数, 则

$$P(X_i = c | N_i) = \frac{1}{z} \sum_{v_j \in N_i} W_{ij} \times P(X_j = c | N_j) \quad (3)$$

其中: $P(X_i = c | N_i)$ 表示点 v_i 属于类别 c 的概率。 wvRN 将选择概率最大的类别作为点 v_i 的预测类别。

3 实验与结果分析

3.1 数据处理

本文随机提取了 30 000 个 Twitter 用户, 并提取了这些用户中的 Follow 关系和 Retweet 关系。 基于这两种关系分别构建了对应的用户关系网络, 并依照 2.1 节讲述的方法将这些用户进行了类别划分。 首先从这 30 000 个用户的用户列表中提取了所有用户创建的列表名称, 并将这些列表名称通过整理和合并后划分出九个大类别: apple、medical、design、photography、economy、politics、hockey、chefs、garden; 最后将所有的用户按其原来隶属的列表划分进这些大类别中, 并删除了同时隶属于多个类别的用户。 经过处理后的数据显示, 基于 Follow 关系的用户集 U_f 中共包含 25 830 个用户, 这些用户间存在 35 004 条边 (Follow 关系); 基于 Retweet 关系的用户集 U_r 中共包含 8 510 个用户, 这些用户间存在 11 214 条边 (Retweet 关系); 由于选择的关联关系不同, 所以 U_f 和 U_r 不完全重合, $U_f \cap U_r = 7 550$, 即两用户集拥有 7 550 个共同用户。 U_f 和 U_r 集合在九大类别中的分布情况如表 1 所示。

表 1 两个用户集合的类别分布

	med	app	pol	gar	pho	des	hoc	eco	che
$F/\%$	6.0	4.4	21.1	1.7	57	5.5	2.3	1.1	0.8
F^*	1 549	1 140	5 440	430	14 768	1 428	588	282	205
$R/\%$	6.5	9.7	62.7	4.2	4.7	2.8	1.9	5.1	2.3
R^*	555	829	5 338	360	401	242	157	435	193

由表 1 所示, F 和 R 行分别列出了两个数据集在各类别中的占比; F^* 和 R^* 行分别列出了两个数据集在各类别中的绝对数量。 总的来看, 用户之间存在的 Retweet 关系比 Follow 关系少很多, 大多数用户关注了大量朋友, 但是仅仅会转发很少一部分朋友的信息。 从绝对数量上看, U_f 的样本数量在 medical、apple、photography、design 和 hockey 上都远远超过了 U_r ; 在 political、garden 和 chefs 上两者的样本数量接近; 在 economy 上 U_r 超过了 U_f , 大约是其绝对样本数量的两倍。 从样本占比上看, U_f 和 U_r 集合在九个类别上的分布都存在差异。 其中两者在 politics 和 photography 类别上的分布差异较大, 在其余类别上的分布差异较小。 这些差异反映了基于 Follow 关系和 Retweet 关系所构成的网络结构间的差异。 在下面的实验中, 将通过实验来对比基于这两种不同结构的网络所获取的分类结果, 并据此结果反过来观察不同用户群的关联特征。

3.2 实验结果及分析

从 15 000 个用户中随机提取 500 个用户, 分别计算了其

V_f 、 V_r 值并绘出其曲线图,如图3所示。由图可见,所有用户的 V_r 值都比 V_f 高出很多。这意味着,相较于关注(Follow)一个新用户而言,当一个用户转发(Retweet)了一个新用户的信息时,该用户的粉丝更倾向于去关注被转发信息(Retweet)的这个新用户。也就是说,一个用户如果想获取更多的关注,就应该想办法使自己的信息被更多地转发(Retweet)。所以在传播用户的影响力方面,Retweet关系比Follow关系所起的作用更大。在考量用户的影响力时,应该更多地依据Retweet关系来进行考察。

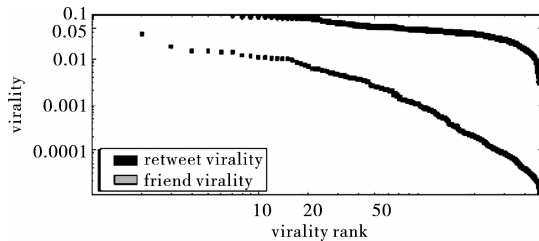


图3 用户的 V_f 和 V_r 值对比

为比较Retweet关系和Follow关系所携带的用户同质性信息的差异,本文采用wvRN分别对 U_f 和 U_r 中的用户进行分类,分类精度如表2所示。此外,表2列出了 U_f 和 U_r 的误差减少算子(error reduction)。为确保实验的正确性,所有实验结果都经过了100次交叉检验后得到。

表2 U_f 和 U_r 在九大类别上的分类精度

	med	app	pol	gar	pho	des	hoc	eco	che	ALL
F	0.79	0.72	0.93	0.21	0.98	0.92	0.82	0.63	0.53	0.92
R	0.34	0.41	0.92	0.16	0.25	0.44	0.36	0.62	0.47	0.72

如表2所示,基于Follow关系的分类精度在medical、apple、photography、design和hockey类别上都远远高于Retweet关系。这是由于 U_r 用户集在这几个类别上的样本数量都远远小于 U_f 用户集在这几个类别上的样本数量,由于样本过少,导致信息缺失较为严重,从而影响了分类结果。从另一方面也可以看出,隶属于这五个类别的用户之间的Retweet关联相对较少,说明其用户的信息互通相对少。这一现象在photography类别上体现最为明显,在 U_f 用户集中photography类别的分类精度远高于其他类别,而在 U_r 用户集中该类别的分类精度是最低的,这充分说明隶属于该类别的用户彼此之间虽然存在大量的关注,但缺乏信息的相互传递。

从表2还可以看出, U_f 和 U_r 集合在political、garden、economy和chefs这四个类别上的分类精度相差不大,尤其在political和economy类别上的分类精度几乎完全一样。其中political的分类精度高达0.92;economy和chefs的分类精度相对较低,在0.5左右;而garden的分类精度很低,大约仅有0.2。由表1可知, U_f 和 U_r 集合在political类别的绝对样本数量接近且远远大于 U_f 和 U_r 集合中的其他类别,自然political类别样本中所包含的信息量也最大,所以其预测精度也最高。需要值得注意的是garden和chefs类别。这两个类别分别在 U_f 和 U_r 两个集合中的绝对样本数量都较小,chefs的样本数量甚至比garden的样本数量还少近一半,但是chefs的预测精度却比garden高一倍。这说明隶属于chefs的用户更倾向于关注和转发同类用户,而garden用户则对同类用户的关注少得多,更少传递同类用户的信息。

综合表1和2可知,整体来说,基于Follow关系的分类精度比基于Retweet关系的分类精度高20%,可见总的来说Follow关系比Retweet关系更能体现用户的同质性;但是,通过对比两者在不同类别用户上的分类结果可以看出,各类别的用户群在关注朋友和转发信息各具特征。这一结果说明,深入理解同一类别用户群在Follow关系和Retweet关系分布上的差异将有助于理解该用户群在社交网络上的信息交互特征。

4 结束语

本文重点分析了Twitter社交网络中的Follow关系和Retweet关系在传播用户影响力和表征用户同质性这两方面所携带的信息差异。文章定义了两个变量 V_r 和 V_f 用于度量这两种关系在传播用户影响力时的差异,相关实验表明,Retweet关系更利于传播用户的影响力,更倾向于从Retweet关系中发掘新的朋友。相较于那些拥有众多粉丝的用户,那些发布原始信息后被多次Retweet的用户更容易吸引新的粉丝。在表征用户同质性方面,分类结果显示Follow关系比Retweet关系携带了更多有用的信息,基于Follow关系的分类精确度比Retweet关系的分类精确度高出了20%。但与此同时,Retweet关系对深入理解各个用户群的信息交互特征亦具补充作用。

参考文献:

- [1] MICHELSON M, MASCKASSY S A. Discovering users' topics of interest on Twitter: a first look[C]//Proc of the 4th Workshop on Analytics for Noisy Unstructured Text Data. 2010: 73-79.
- [2] MASCKASSY S A, MICHELSON M. Why do people retweet? Anti-homophily wins the day[C]//Proc of the 5th International AAAI Conference on Weblogs and Social Media. California: The AAAI Press, 2011: 209-216.
- [3] KANG J H, LERMAN K. Using lists to measuring homophily on twitter[C]//Proc of AAAI Workshop Technical Report. California: The AAAI Press, 2012: 26-32.
- [4] LEE K, PALSETIA D, NARAYANAN R, et al. Twitter trending topic classification[C]//Proc of Workshop on Optimization Based Methods for Emerging Data Mining Problems. 2011: 251-258.
- [5] KOMAMIZU T, YAMAGUCHI Y, AMAGASA T, et al. FACTUS: faceted Twitter user search using Twitter lists[C]//Proc of the 12th International Conference on Web Information System Engineering. Heidelberg: Springer, 2011: 343-344.
- [6] HU Meng-die, LIU Shi-xia, WEI Fu-ru, et al. Breaking news on Twitter[C]//Proc of ACM Annual Conference on Human Factors in Computing Systems. 2012: 2751-2754.
- [7] ZENG Xue, WU Yue, ZHENG Li-hua, et al. Homophily modeling for a single class in relational classification[J]. *Advances in Information Sciences and Service Sciences*, 2012, 20(4): 667-674.
- [8] MASCKASSY S A, PROVOST F. Classification in networked data: a toolkit and a univariate case study[J]. *Journal of Machine Learning Research*, 2007, 8(5): 935-983.
- [9] MCPHERSON M, LOVIN L S, COOK J M. Birds of a feather: homophily in social networks[J]. *Annual Review of Sociology*, 2011, 27(1): 415-444.
- [10] WELCH Michael J, SCHONFELD Uri, HE Dan, CHO Junghoo. Topical semantics of Twitter links[C]//Proc of the 4th ACM International Conference on Web Search and Data Mining. 2011: 327-336.