

一种基于结构与内容的网络水军检测模型

杨锡慧

(中国电信南京分公司网络运维部,江苏 南京 210008)

摘 要: 随着网络水军团体行为破坏性越来越强,如何有效检测网络水军成为一个迫切问题. 针对这一问题,提出了一种基于结构与内容的网络谣言水军检测模型 MSD,该模型认为重叠结构中多次传播谣言的节点为网络谣言水军. MSD 首先从单位时间内发送信息的节点中,挖掘传播能力强的节点(重叠结构);然后根据已知的谣言与重叠结构中的节点传播内容进行相似度检测,检测出嫌疑网络水军;最后根据疑似网络水军传播谣言的频率确定网络水军. 实验表明,MSD 模型较相关其他模型,在检测网络谣言水军方面具有更高的准确率.

关键词: 社交网络;网络水军;谣言信息;重叠结构

中图分类号: TP393.0

文献标识码: A

文章编号: 1000-7180(2017)12-0099-08

A Webspammer Detection Model Based on Structure and Content

YANG Xi-hui

(Network Operation and Maintenance Department, Nanjing Branch, China Telecom, Nanjing 21008, china)

Abstract: With the destruction of the Webspammer is more and more strong, how to efficiently and accurately detect the webspammer has become an urgent problem. To solve this problem, this paper proposes a webspammer Detection Model based on Structure and Content (MSD), the model considers the node in the overlaps structure repeatedly spread Internet rumors is the webspammer. First, In the unit time, from the node that has sent the information, MSD mining strong spread ability of the node (overlapping structure); Then, MSD can detect the the suspected webspammer according to the similarity between the rumor and the node spread content in the overlapping structure; Finally, determine the webspammer according to a suspected webspammer spread rumors frequency. Experimental results show that the MSD model is related to other model has higher accuracy in the detection of webspammer.

Key words: social networking; webspammer; rumor; overlapping structur

1 引言

以微博^[1]为代表的社交网络在新闻传播、网民互助、知识传播等方面表现出了积极作用. 然而,由于有组织的网络水军的出现,社交网络中出现谣言盛行、欺诈活动猖獗等现象,给社会政治经济文化活动带来损失. 因此,如何有效识别网络水军成为一个亟待解决的问题.

目前,网络水军的检测主要依据是微博内容特征、用户行为特征和网络特征三个单一特征,进一步

研究会发现,网络水军一般同时具有多个特征. 因而,仅从单个特征入手(偶尔传播过一次谣言),则无法准确检测出真正的网络水军.

针对从用户单个特征出发,无法准确检测出真正的网络水军问题,本文提出一种基于结构和内容的网络水军检测模型 MSD(A Model of Web spammer Detection based on Structure and Content). 它的基本思想是:用户要成为网络水军必须具备两个要素:其一要有成为水军的条件;其二要有成为水军的意愿.

本文认为,在微博回复转发网络中,是否处在重叠结构(有重大影响力的节点)中是用户成为网络水军的基本条件;有传播过谣言的历史记录,可以说明用户有成为网络水军的意愿。

因此,MSD 模型首先构建微博回复转发网络,从中挖掘出重叠结构;将重叠结构中转发过谣言用户看成疑似网络水军。为了避免计算风暴,只考虑谣言出现后单位时间内用户间的回复转发行为。

MSD 模型有如下优点:

(1)客观性. 本文认为重叠结构中的节点(即具有较大影响力的谣言发布者)是疑似网络水军的假设,比传统单一特征方法更客观。

(2)准确性. 重叠结构中节点如果多次传播谣言才确认为网络水军,比 1 次传播谣言就确认网络水军的方法,得到的结论更加准确。

(3)可行性. 构建回复转发网络时,只考虑单位时间内用户间的回复转发行为,只牵涉到部分用户的部分信息,可以显著减少计算量。而重叠结构是回复转发网络子图,用户间回复转发信息规模更小,计算量会进一步减少。

2 相关研究

针对如何检测网络水军的研究,相关学者取得了一系列成果。这些成果主要从内容特征、用户特征、网络特征和基于影响力分析四个方面进行网络谣言水军的检测和识别。

早期网络环境中,网络水军产生的内容具有显著的可识别特征。Ott M 等人^[2]利用评论自身的文本特征可以识别发布相似虚假评论的网络水军。文献^[3]提出基于内容特征的方法,发现论坛水军制造的垃圾内容,但该方法只在发现论坛垃圾页面时较为有效。由于该方法的核心是关注内容,而不是水军本身,所以检查网络水军最好结合用户行为特征进行分析。

Stringhini 等学者从用户行为特征出发^[4-7],通过检测与正常用户行为特征的不同,来发现网络水军。Stringhini 等人^[4]在 3 个主流社交网络平台:Facebook、MySpace 以及 Twitter 中使用“诱捕器”收集实验数据,分析不同平台的网络谣言水军行为,提出相应的识别方法。但其识别的方法过于简单,无法挖掘出趋向于正常用户的网络谣言水军。Yang 等人^[7]深入分析了 Twitter 中网络水军的隐藏策略,并提出基于邻居节点特征发现 Twitter 中网络水军的方法。但从网络谣言水军的邻居节点入手,可能将

问题复杂化,因为网络谣言水军通常具有的庞大邻居节点。

随着网络谣言水军的行为逐渐复杂化,并趋向于正常用户,仅从单一特征进行识别的传统方法无法发现隐蔽的水军^[8]。因此可以从网络水军对周围环境产生的影响入手分析。

在网络环境层级,Las-Casas 等人^[9]根据网络水军活动时的网络特征进行识别。由于网络水军为尽可能达到效益最大化,会集中在一段时间内大量传播谣言,因此这段时间内的网络负载会突然加大,流量也会集中在某些链路中。该方法准确率较高,但该方法需要应用 ISP 数据,无法普遍推广。Song 等人^[10]基于关系特征的方法,认为网络水军和接收谣言的普通用户间的距离为 3~4 个节点,同时出度越大,流通性越好,越有可能是网络谣言水军。但此方法要在全网中进行搜索,计算量太大;另外只对距离为 3~4 个节点的用户对进行分析,准确率不高。文献^[11]介绍根据神经网络进行社团分类,其实是个子图分类问题,在全网范围内,子图的挖掘本身就是个问题。

另外,还可以从节点影响力这个角度进行网络谣言水军检测^[12-15],影响力越大,网络水军越容易完成它的目的。文献^[13]基于 PageRank 的思想,对网络中的用户进行评级。认为用户的评级越高,则传播虚假信息过程中起到的作用越大。这个方法可以准确检测出有影响力的节点,但不能说明这些节点,就是网络水军。文献^[15]把对特定成员影响力构成,起到重要作用的网络群体称为支撑结构。通过寻找支撑结构,来间接检测网络谣言水军。但在寻找支撑结构的过程中,对节点影响力的度量计算复杂,不利于实际运用。而且,即使找到了某些节点的支撑结构,若此支撑结构与其支撑的节点与其他社区孤立,它也无法将信息传播得更远。文献^[16]给出了一种重叠社团的计算方法,文献^[17]给出了用户影响力的计算方法,但是没进一步研究网络水军与他们间的关系。

尽管相关学者已经在网络水军检测工作上取得了大量成果,但是他们的工作仅从单一特征入手进行检测,准确率不高。本文提出的方法避免了从单一特征入手检测网络谣言水军的片面性,减少误判概率;只对重叠结构中节点行为进行分析,提高了计算效率和降低了计算规模,为净化网络环境提供了技术支持。

3 基于结构与内容的网络水军检测模型 (MSD)

3.1 相关概念

定义 1 转发回复网络. 对于 n 个社交网络节点, m 个节点间转发回复关系, 可分别表示为节点集 $V_t = \{v_1, v_2, \dots, v_m\}$ 与边集 $E_t = \{e_{i,j} \mid i, j \in V_t\}$, t 对应网络谣言出现的时间. 每一个节点 v_i 中, 有相应的内容 D_i . 若两个社交网络节点 v_i, v_j 之间发生过转发回复行为, 则存在一条边 $e_{i,j}$. 最终形成一个转发回复网络 $G_t(V_t, E_t)$.

定义 2 重叠结构. 基于转发回复网络 $G_t(V_t, E_t)$ 中的重叠结构 OS(Overlapping Structure), 由若干节点 v_i' 所组成. 每个 v_i' 被 2 条隶属于不同社区的边 e, e' 连接, 即:

$$OS(v_i') \Leftrightarrow \exists u_1, u_2, C_1, C_2 (u_1 \neq u_2 \wedge C_1 \neq C_2):$$

$$e = (v_i', u_1) \in C_1 \wedge e' = (v_i', u_2) \in C_2 \quad (1)$$

定义 3 嫌疑网络谣言水军. 对于 OS 中的节点 v_i , 在谣言出现时 1 小时内发布的信息内容集合 D_i , 定义其与谣言内容的相似度为 $S_{\theta_{D_i}, \theta_{D_i}}$. 若大于等于相似阈值, 则认为其为此谣言的嫌疑发布者, 第 k 条谣言检测出的嫌疑网络谣言水军团体可定义为

$$CO_k = \{v_i \mid v_i \in V_t, S_{\theta_{D_i}, \theta_{D_i}} \geq \sigma\} \quad (2)$$

定义 4 网络谣言水军. 对由 n 个谣言检测出的嫌疑网络谣言水军团体 CO , 进行求交集操作, 得到最终的嫌疑网络谣言水军团体 I_{CO} .

$$I_{CO} = CO_1 \cap CO_2 \cdots \cap CO_n \quad (3)$$

3.2 模型核心思想

MSD 模型认为重叠结构中节点传播能力强, 符合网络水军的身份特征. 通过文献[18]对节点传播影响力的分析可知, 某节点的传播能力与和其直接相连的社区的数目有关, 越多则传播速度越快, 且传播范围越广. 进一步分析, 若某节点隶属于多个社区, 则它的传播能力应该更强, 因为它不仅连接了多个社区, 还直接与这些社区里的成员有直接连接, 重叠社区结构(图 1 中的 5 号节点)就是具有这样特性的结构.

同时也认为, 网络水军出现的时间有一定规律性, 且不认为无意中传播一条谣言就认定为嫌疑网络谣言水军. 当有推广任务时, 网络水军会“全体”出动, 共同对所推广的话题进行传播, 使其成为热点. 这也体现了, 网络谣言水军团体在传播内容方面的特性——特定时间内(有推广任务时), 他们所发

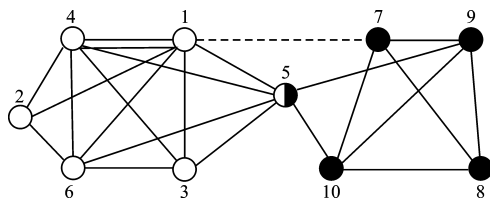


图 1 重叠结构

送的内容有高度语义相似性.

综上所述, MSD 模型核心思想是: 在重叠结构中, 将多次推广过谣言的节点视为网络水军. 因此, 先从结构特征入手, 找到传播信息能力强的结构(重叠结构). 再结合内容特征, 筛选出疑似网络谣言水军. 最终找出多次传播过谣言的疑似网络谣言水军, 确定网络谣言水军.

3.3 模型的定义和表示

3.3.1 模型总体框架

MSD 模型旨在挖掘社交网络中的重叠结构, 并通过内容特征筛选出重叠结构中疑似网络谣言水军节点, 最后根据疑似网络谣言水军在多次谣言传播中出现的频率, 检测出网络谣言水军. 主要步骤包括数据预处理、重叠结构挖掘、内容特征检测与基于结构与内容的检测. 如图 2 所示.

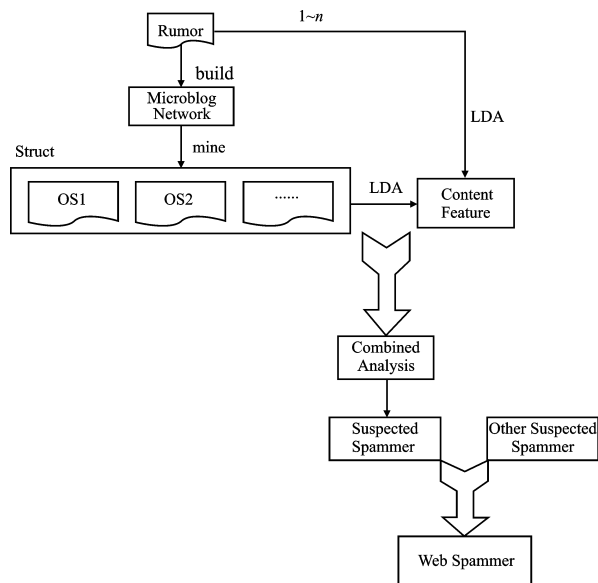


图 2 MSD 处理流程

(1) 数据预处理. 根据谣言出现的时间, 搜集相应时间内新发布消息的节点信息, 根据这些节点的转发回复历史记录, 构建转发回复网络.

(2) 重叠结构挖掘. 根据本文给出的算法, 挖掘微博转发回复网络中的重叠结构, 得到多个待检测的重叠节点.

(3)内容特征挖掘. 采用经典的语义主题生成模型 LDA,对谣言信息和重叠节点所传播的内容进行内容特征提取.

(4)基于结构与内容的检测. 如果重叠结构中节点传播内容与谣言语义相似,那么将其归入嫌疑网络谣言水军中. 如果嫌疑网络谣言水军在多个谣言中以较高频率出现,它是网络谣言水军的概率就较高.

3.3.2 重叠结构

从重叠结构节点关联多个社区,因此通过他们传播的信息,可以到达多个社区. 正因为重叠结构节点传播能力强,从而构成了网络水军的主体.

(1)数据预处理

由于网络谣言水军要起到推广特定内容的目的,在平时就需要与其他用户保持一定的“联系”,所以社交网络在一定时间内是结构稳定的,其中的重叠结构基本保持不变. 由文献[19]中的数据分析可知,转发行为一般发生在 1 h 以内. 因此本文的转发回复网络,利用谣言出现时间点 t 后 1 h 内,社交网络中的转发回复记录来构建. 既减少了数据的规模,避免在全网络进行检测工作;又排除了很多与此谣言无关的数据,提高检测的准确率.

(2)重叠结构挖掘

目前,重叠社区挖掘研究引起了越来越多的关注,本文采用基于链接聚类的算法,首先对网络中的边集进行硬划分,再将边划分结果转化为相应的节点结构. 代表算法有 LINK^[20],但 LINK 方法找到的重叠社区,存在“过度重叠”问题(见图 1). 它将孤立边(1,7)看成是独立的链接社区,因此节点 1 和节点 7 都被认为是重叠节点. 但这样找到的重叠社区,并不满足本文的要求. 因为节点 1 和 7 连接的节点较节点 5 少,因此它的传播能力比真正的重叠社区(5 号节点)弱,是网络谣言水军的可能性小. 文献[21]引入核心边的概念,将与核心边直接密度可达的边,归入同一个链接社区,排除了孤立边. 但算法伪代码中只是形成了节点社区,并没有最终对重叠结构进行挖掘. 本节将对文献[21]中的算法伪代码进行改进,给出了节点社区进行重叠结构挖掘的算法,用于对社交网络中的重叠结构进行挖掘.

定义 5 边相似度. 若两条边存在一个公共节点,则这两条边之间存在相似度,例如,边 $e = (u, v)$ 和边 $l = (u, w)$ 之间的相似度,定义为

$$S(e, l) = \frac{|\Gamma(v) \cap \Gamma(w)|}{|\Gamma(v) \cup \Gamma(w)|} \quad (4)$$

其中,为节点自身和其所有的邻居节点构成的点集.

定义 6 邻居边. $e = (u, v)$ 的邻居边记为 $N(e)$,即:分别与节点 u 和 v 相连且不包括 e 的边集合.

定义 7 $N_\epsilon(e)$. e 的 ϵ 领域 $N_\epsilon(e)$ 是指在 e 的邻居中,与 e 的相似度大于或等于 ϵ 的边集合,即:

$$N_\epsilon(e) = \{l \in N(e) \mid s(l, e) \geq \epsilon\} \quad (5)$$

定义 8 核心边. 若一条边的 ϵ 领域数 $|N_\epsilon(e)| \geq \mu$,则称 e 为核心边.

定义 9 直接密度可达. 若 e 为一个核心边且 l 属于 e 的 ϵ 领域时,称 l 直接密度可达于 e ,即:

$$DirReach(e, l) \Leftrightarrow e.Core = true \wedge l \in N_\epsilon(e) \quad (6)$$

(3)重叠结构挖掘算法

算法 1 给出了挖掘重叠结构的方法. 包括 3 个阶段:第 1 阶段将边集进行聚类得到链接社区;第 2 阶段将得到的链接社区 LCS 转化为相应的节点社区集合 CS;第 3 阶段找到重叠结构,即点集 OS. 算法 2 是算法 1 中 FindLinkComs 函数的伪代码.

算法 1. Overlapping structure discovery algorithm.

input: $G_t(V_t, E_t)$, ϵ, μ ;

output: Overlapping structure OS.

/* Phase 1, the edge set clustering */

1. $LCS \leftarrow \emptyset, CS \leftarrow \emptyset$;

2. $\forall e \in E: eIsClassified \leftarrow FALSE$;

3. FOR EACH $e \in E$ DO

4. IF ! $eIsClassified$ THEN

5. IF $e.IsCore$ THEN

6. $eIsClassified \leftarrow TRUE$;

7. $LCS = LCS \cup FindLinkComm(e)$;

8. ENDIF

9. ENDFOR

/* Phase 2, convert link community into node community */

10. FOR EACH $LC \in LCS$ DO

11. $C \leftarrow \emptyset$;

12. FOR EACH $e = (u, v)$ IN LC_i DO

13. $C \leftarrow C \cup \{u, v\}$;

14. $CS = C \cup C_i$;

15. ENDFOR

16. ENDFOR

/* Phase 3, look for overlapping node and

```
get overlapping structure * /
17.  $CS \leftarrow \emptyset$ ;
18.  $\forall e \in LCS: e.IsClassified \leftarrow \text{FALSE}$ ;
19. FOR EACH
 $e \in LCS(e \in LC_x, e = (v, u) \in C_i)$  DO
20. IF !  $e.IsClassified$  THEN
21.  $e.IsClassified \leftarrow \text{TRUE}$ ;
22. IF  $(v \in C_j \vee u \in C_j) \wedge C_i \neq C_j$  THEN
23. IF  $v \in C_j$  THEN
24.  $OS \leftarrow OS \cup \{v\}$ ;
25. ELSE
26.  $OS \leftarrow OS \cup \{u\}$ ;
27. ENDIF;
28. ENDIF;
29. ENDIF;
30. ENDFOR;
31. RETURN OS.
```

算法 2. Link community find algorithm Find-LinkComs.

```
input:  $e, \epsilon, \mu$ ;
output: Link Community(LC).
1.  $LC \leftarrow \emptyset$ ;
2.  $Q.Append(e)$ ;
3.  $e.IsClassified \leftarrow \text{TRUE}$ ;
4. WHILE  $Q.IsEmpty() = \text{FALSE}$  DO
5.  $x \leftarrow Q.Serve()$ ;
6.  $R \leftarrow \{y | (E|DirReach(x, y, \epsilon, \mu))\}$ ;
7. FOR EACH  $l$  IN  $R$  DO
8. IF  $l.IsClassified \leftarrow \text{FALSE}$  THEN
9.  $l.IsClassified \leftarrow \text{TRUE}$ ;
10. IF  $l.IsCare$  THEN
11.  $LC \leftarrow LC \cup \{l\}$ ;
12.  $Q.Append(l)$ ;
13. ELSE
14.  $LC \leftarrow LC \cup \{L\}$ ;
15. ENDIF
```

3.3.3 基于结构与内容的网络水军检测

本文认为处在重叠结构中的节点,如果存在转发、复制、重写谣言等操作,视为疑似网络谣言水军.因此,根据上文得到网络重叠结构后,将节点转发、复制、重写的内容特征与当时的谣言特征进行相似性比较,确定疑似水军.最后,如果疑似水军对多个谣言进行了转发、复制、重写操作,超过了一定的阈值,那么就可以认为该疑似水军为

网络水军.

由于在进行谣言推广的过程中无论以何种形式(转发、复制、重写,特别是重写),谣言的主题都不会改变,并且社交网络节点发布的信息通常都涉及一个或多个主题,这些特征正好和 LDA 主题模型匹配.因此本文采用 LDA 主题模型来推断节点所发信息的主题分布.

(1) 嫌疑网络水军的挖掘

嫌疑网络水军是根据重叠结构中节点一次转发、复制、重写谣言而确定的.在得到社交网络转发回复网络的重叠结构 OS 之后,根据网络谣言出现的具体时间,我们需要挖掘出重叠结构中每个节点的内容特征.一般地,绝大部分的转发时间间隔在 1 小时内,所以 MSD 模型选取重叠结构中每个社交网络节点在网络谣言出现时 1 小时内发布的内容进行分析.

定义 10 内容 D 的主题分布:令主题集合 $C = \{C_1, C_2, \dots, C_\tau\}$, 则 $p(C_i | D)$ 表示内容 D 属于主题 C_i 的后验概率, 由这 T 个后验概率组成的向量 $(p(C_1 | D), p(C_2 | D), \dots, (C_\tau | D))$ 被称为内容 D 的主题分布.

定义 11 令集合 $D = \{D_1, D_2, \dots, D_d\}$ 表示重叠结构(OS) 内某个用户在网络谣言发布后 1 h 内, 用户发布信息的集合. D_i 表示已经确定的谣言信息.

定义 12 某用户发布内容的主题分布:用一个 T 维向量表示成 $(t_1, t_2, \dots, t_\tau)$, 其中 $t_i = \frac{1}{d} \sum_{j=1}^d p(C_i | D_j)$

推断过程中的变量如表 1 所示.

表 1 推断过程中的变量列表

变量	含义
T	主题个数
V	单词个数
α	<i>Dirichlet</i> 分布的参数, 用于生成一个主题 θ 向量
β	各个主题对应的单词概率分布
Z_{w_i}	单词 w_i 的主题
$Z_{D,-i}$	除了第 i 个单词的主题外, 在内容 D 中其他所有单词的主题集合
$n(j, D)$	内容 D 中属于第 j 个主题的单词数
$n(D)$	内容 D 中的单词数
$n(j, v)$	主题 j 得到某一单词 v 的频数
$n(j, \cdot)$	属于主题 j 的所有单词数

采用 Gibbs 抽样算法间接计算模型参数 $\theta_{D,j}$ (内容 D 属于第 j 个主题的概率). 假设 D 由 n 个单词组成, 记为 $\{w_1, w_2, \dots, w_n\}$, 对 D 中的单词 w_i , $Z_{w_i} = j$ (单词 w_i 属于第 j 个主题) 的概率计算如下:

$$\begin{aligned} & P(Z_{w_i} = j | Z_{D,-i}, D, \beta, \alpha) \\ &= \frac{P(Z_{w_i} = j, Z_{D,-i}, D | \beta, \alpha)}{P(Z_{D,-i}, D | \beta, \alpha)} \\ &= \frac{n(j, D) + \alpha - 1}{n(D) + T\alpha - 1} \times \frac{n(j, v) + \beta - 1}{n(j, \cdot) + V\beta - 1} \end{aligned} \quad (7)$$

对式(7)反复迭代, 并对所有主题进行抽样, 最终达到抽样结果稳定, 最终 $\theta_{D,j}$ 可以估计为

$$\hat{\theta}_{D,j} = \frac{n(j, D) + \alpha}{n(D) + T\alpha} \quad (8)$$

内容 D 的主题分布表示为 $\hat{\theta}_D = (\hat{\theta}_{D_1}, \hat{\theta}_{D_2}, \dots, \hat{\theta}_{D_T})$.

本文通过衡量重叠结构内的用户所发信息与谣言内容的主题相似度, 来判别用户是否转发、复制或重写了谣言消息以帮助其传播, 并将相似度较低的用户, 在此次检测中排除其是网络谣言水军的可能. 通过推导, 可以直接得到谣言 D_i 的主题分布. 根据定义 12, 计算出 OS 内所有用户的在 1 h 内发布信息的主题分布 $(t_1, t_2, \dots, t_T)$.

由于社交网络用户所发信息(微博、论坛帖等)一般涉及的主题较少, 即使是 1 h 内的信息, 网络谣言水军也只会将重点放在要推送的谣言消息上, 所以主题分布很稀疏. 针对这种情况, MSD 模型采用余弦相似性的方法, 来度量每个 OS 内的用户所发信息与谣言信息的相似程度:

$$S_{\theta_{D_i}, \theta_{D_j}} = \frac{\theta_{D_i} \cdot \theta_{D_j}}{||\theta_{D_i}|| ||\theta_{D_j}||} \quad (9)$$

求得所有的相似度结果后, 设定阈值 σ , 若 $S_{\theta_{D_i}, \theta_{D_j}} \geq \sigma$, 则将第 i 个用户加入嫌疑网络水军的聚类中, 最后得到由这条谣言信息检测出的嫌疑网络谣言水军团体(CO).

(2) 网络水军的确定

本文对从不同谣言出发, 检测得到的多个嫌疑网络水军团体, 进行求交筛选操作:

$$\begin{aligned} I_{\omega} &= CO_1 \\ I_{\omega} &= I_{\omega} \cap \bigcap_{i=2}^n CO_i \end{aligned} \quad (10)$$

式中, I_{ω} 为所有嫌疑网络谣言水军团体 CO 的交集, 初始值为 CO_1 . 由于 I_{ω} 中节点是重叠结构中且多次扩散谣言的节点, 可以认为他们就是网络水军.

4 实验分析

4.1 实验设定

本节在新浪微博平台上, 验证本文中模型和算法的有效性, 其中的方法通过使用 Java 和 Python 实现. 使用 LDA 进行的内容分析时, 本文采用了 Phan 等人^[22]提供的参数估计的工具 JGibbLDA, 进行 LDA 参数估计. 实验环境由一台 Windows 7 旗舰版操作系统、双核 3.2 GHz 英特尔 i5 处理器、4 GB 内存的 PC 组建.

对于潜在主题数量 T 的取值, 本文采用困惑度的方法进行选取, 计算公式为

$$Perplexity(M) = \exp \left\{ \frac{\sum_{m=1}^M \ln p(r_m)}{\sum_{m=1}^M N_m} \right\},$$

式中, M 为测试集; N_m 表示第 m 篇微博 r_m 的长度; $p(r_m)$ 表示模型产生微博 r_m 的概率, 如下公式所得:

$$p(r_m) = \prod_{i=1}^n \sum_{j=1}^T p(w_i | Z_{w_i} = j) p(Z_{w_i} = j | r_m) \quad (11)$$

在选取中, 困惑度越低意味着模型产生文档的能力越高, 主题数 T 的取值越接近真实情况. 在实验中, 对 T 取值为 10, 20, 30, 40, 50, 60, 70 的情况进行的测试, 随着主题数目的增加, 困惑度在不断的变化. 当主题数为 40 时, 困惑度达到最小值, 此时模型的性能最好. 因此选取 $T=40$ 作为主题的数目.

4.2 实验流程

第一步. 回复转发网络的生成. 从 2013 年的 7 月 13 日、8 月 9 日和 8 月 17 号中的 3 个谣言出发, 采用 Python 语言通过新浪微博公共开放端口, 分布抓取 3 个谣言的相关数据. 数据共计包括 19 746 个博主和其中 58 132 次转发回复关系. 其中包括谣言发布后 10 h 内发送过微博的博主, 以及由这些博主作为种子节点, 用深度优先的方式, 抓取与当前用户有过转发回复的节点. 由于六度空间理论, 本实验中的深度遍历不超过 3 层. 并且, 在抓取数据的过程中, 已排除了用户过少的独立图结构(与其他图结构无联系). 此数据集中, 每个节点代表新浪微博的一个博主的 ID, 如果一个博主转发或者回复了另一个博主的微博, 那么他们之间就存在一条边, 每条边不重复建立. 我们考虑微博中的评论信息并不直接为其他博主可见, 不能体现平时普通微博博主和网络谣言水军的结构关系. 因此我们选用了转发回复网络作为实验验证.

第二步. 重叠结构的挖掘. 在回复转发网络基础

上,根据本文提出重叠结构挖掘算法进行挖掘.

第三步. 随机选择 40 个热门主题,在此基础上,分别计算谣言信息的主题分布和重叠结构节点回复转发信息的主题分布,计算主题的相似度,根据相似度确定网络水军. 实验采用 LDA 分析重叠节点 1h 内的回复转发内容与谣言的主题相似度. 参照其他文献对超参数的选取,MSD 模型的超参数的选择为: $\alpha=50/T,\beta=0.1$.

4.3 MSD 的准确性

本文采用人工判断的方法,进行方法准确性的判断. 采用如下指标:

$$Accuracy=\frac{n}{N}$$

(12)

式中, N 为实验检测出的网络谣言水军数目, n 为两名工作人员在 N 中确认的网络水军数目. 工作人员确认水军的依据是用户所有的微博内容、微博转发数以及用户本身的粉丝数三个指标.

我们对比了以下 3 种方法:CAT、IM、我们提出的 MSD. 其中 CAT 是 Amleshwaram 等人^[23]综合用户的行为、内容、用户间关系等特征,实现对网络谣言水军的检测. IM 是 Wang 等人^[24]针对收集的评论数超过 7 000 的微博及其用户间联系,提出的影响力在线算法检测网络谣言水军.

我们先观察下 MSD 模型从单个谣言(嫌疑网络谣言水军)出发的检测效果. 对比结果如表 2 所示.

表 2 三种方法的准确率(Accuracy)

Method	7. 13	8. 9	8. 17
CAT	0. 78	0. 82	0. 75
IM	0. 83	0. 89	0. 72
MSD	0. 85	0. 92	0. 81

MSD 模型从三个谣言出发的检测效果,得到的对比结果如表 3 所示.

表 3 三种方法的准确率(Accuracy)

Method	CAT	IM	MSD
	0. 82	0. 89	0. 95

从实验结果可以看出,MSD 方法模型对本文提取的三个谣言,在还没有进行求交集的操作的情况下,其准确率已经高于其他两种水军检测方法. 进行求交筛选后,检测准确率进一步提高. 同时,以微博文本为测试集,计算潜在主题数量,提高了 MSD 方法模型的效果.

5 结束语

本文针对以往网络谣言水军研究工作中的不

足,提出了一种新的网络谣言水军检测模型 MSD,综合考虑了社交网络中的转发回复网络节点所在结构的特性和节点本身的内容特性. 模型不仅能提高网络谣言水军检测的准确率,还能找出谣言与其推广者的对应关系. 我们利用真实数据验证了基于结构与内容的网络谣言水军检测方法的准确性. 在今后的研究工作中,我们将继续优化 MSD 模型的效率、提高网络谣言水军检测的准确率.

参考文献:

[1] 朱明华,董科军,宋成. Blog 空间的特征初探[J]. 微电子学与计算机, 2005, 22(9):27-29.

[2] Ott M, Choi Y, Cardie C, et al. Finding deceptive o-pinion spam by any stretch of the imagination[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 309-319.

[3] Niu Y, Chen H, Hsu F, et al. A quantitative study of forum spamming using context-based analysis[C]//NDSS. San Diego, 2007.

[4] Stringhini G, Kruegel C, Vigna G. Detecting spammers on social networks[C]//Proceedings of the 26th Annual Computer Security Applications Conference. ACM, 2010: 1-9.

[5] Ghosh S, Korlam G, Ganguly N. Spammers' networks within online social networks: a case-study on Twitter [C]//Proceedings of the 20th international conference companion on World wide web. ACM, 2011: 41-42.

[6] Zhu Y, Wang X, Zhong E, et al. Discovering spammers in social networks[C]//AAAI. 2012.

[7] Yang C, Harkreader R, Gu G. Empirical evaluation and new design for fighting evolving Twitter spammers [J]. Information Forensics and Security, IEEE Transactions on, 2013, 8(8): 1280-1293.

[8] 莫倩,杨珂. 网络水军识别研究[J]. 软件学报, 2014, 25(7):1505-1526.

[9] Las-Casas P H B, Guedes D, Almeida J M, et al. SpaDeS: Detecting spammers at the source network [J]. Computer Networks, 2013, 57(2): 526-539.

[10] Song J, Lee S, Kim J. Spam filtering in twitter using sender-receiver relationship[C]//Recent Advances in Intrusion Detection. Springer Berlin Heidelberg, 2011: 301-317.

[11] 张兰华,孙岩,薛绍伟,等. 基于 BP 神经网络的社团分类研究[J]. 微电子学与计算机, 2011, 28(6): 197-200.

- [12] 周东浩, 韩文报. DiffRank: 一种新型社会网络信息传播检测算法[J]. 计算机学报, 2014, 37(4): 884-893.
- [13] 王永刚, 蔡飞志, Keong, 等. 一种社交网络虚假信息传播控制方法[J]. 计算机研究与发展, 2012, 55(增刊2): 131-137.
- [14] 张玥, 张宏莉, 张伟哲, 等. 识别网络论坛中有影响力用户[J]. 计算机研究与发展, 2013, 50(10): 2195-2205.
- [15] 韩毅, 许进, 方滨兴, 等. 社交网络的结构支撑理论[J]. 计算机学报, 2014, 37(4): 905-914.
- [16] 潘惠勇, 王鹏, 张慧乐. 基于 FCM 的复杂网络重叠社团结构发现算法[J]. 微电子学与计算机, 2011, 28(12): 111-114.
- [17] 肖宇, 许炜, 张晨, 等. 社交网络中用户区域影响力评估算法研究[J]. 微电子学与计算机, 2012, 29(7): 58-63.
- [18] 赵之滢, 于海, 朱志良, 等. 基于网络社团结构的节点传播影响力分析[J]. 计算机学报, 2014, 37(4): 753-766.
- [19] 丁朝云, 周滨家堰, 张璐敏. 微博中基于多关系网络的话题层次影响力分析[J]. 计算机研究与发展, 2013, 56(10): 2155-2175.
- [20] Ahn Y Y, Bagrow J P, Lehmann S. Link communities

reveal multiscale complexity in networks[J]. Nature, 2010, 466(7307): 761-764.

- [21] 朱牧, 孟帆蓉, 周勇. 基于链接密度聚类的重叠社区发现算法[J]. 计算机研究与发展, 2013, 56(12): 2520-2530.
- [22] Phan X H, Nguyen L M, Horiguchi S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections[C]//Proceedings of the 17th international conference on World Wide Web. ACM, 2008: 91-100.
- [23] Amleshwaram A A, Reddy N, Yadav S, et al. CATS: Characterizing automation of Twitter spammers[C]//Communication Systems and Networks (COMSNETS), 2013 Fifth International Conference on. IEEE, 2013: 1-10.
- [24] Wang K, Xiao Y, Xiao Z. Detection of Internet water army in social network[C]//2014 International Conference on Computer, Communications and Information Technology (CCIT 2014). Atlantis Press, 2014.

作者简介:

杨锡慧 女, (1974-), 硕士, 高级工程师. 研究方向为计算机网络、网络监控与维护. E-mail: 15335179448@189.cn.

(上接第 98 页)

级自动化仿真测试环境框架, 设计了自动化仿真测试环境的任务结构和物理结构划分, 并进行了工作过程的设计, 最终研制开发了 TEFAS 工具平台原型, 在工程应用中取得了较好的测试效果.

参考文献:

- [1] 龙祖利, 王子云. FPGA 测试技术及 ATE 实现[J]. 计算机工程与应用, 2011, 47(6): 65-67.
- [2] 刘子宜, 刘畅, 郑军. 基于软件测试技术的 FPGA 测试研究[J]. 电子技术应用, 2011, 37(5): 28-30.
- [3] Lemon R, Coulter B, Archibald S, et al. Interface test adapter development & maintenance using a Continuity/Insulation Automatic Test Station in large scale test systems[C]//AUTOTESTCON, 2009 IEEE. IEEE, 2009: 107-112.
- [4] Justin J, Lindstrom L, Jain A. Using a plug-in model to simplify and enhance ATE test software capabilities

[C]//AUTOTESTCON, 2012 IEEE. IEEE, 2012: 105-107.

- [5] 李艳, 陈陵都, 陈亮. 面向 FPGA 芯片开发的测试方法设计与实现[J]. 微电子学与计算机, 2014, 31(10): 22-27.
- [6] 周珊, 王金波, 王晓丹. 基于时序路径的 FPGA 时序分析技术研究[J]. 微电子学与计算机, 2016, 33(1): 76-80.
- [7] 刘畅, 刘斌, 阮镰. 航空电子软件仿真测试环境软件体系结构研究[J]. 航空学报, 2006, 27(5): 877-882.

作者简介:

高虎 男, (1986-), 硕士, 工程师. 研究方向为 FPGA 软件测试、软件测试环境. E-mail: gaohu_2009@163.com.

郑军 男, (1969-), 硕士, 研究员. 研究方向为软件工程、FPGA 与软件测试.

田曾昊 男, (1990-), 工程师. 研究方向为 FPGA 软件测试.