大数据下基于信息流的快速种子用户识别

谢杨晓洁 赵 凌

(四川师范大学数学与软件科学学院 成都 610066)

摘 要 针对大数据下的种子用户的精准识别,分析了影响用户成为种子用户的两大因素:时间优先和属性特征,以及种子信息传播的两大特征:传播时差和方向性。据此,提出了一种快速寻找种子用户的方法,即先将用户按属性特征分到不同的组中,通过分析所有组之间短信流通关系和传播时差找到信息流,即方向性,从而逐步缩小了搜索范围,再通过阈值筛选备选种子。最后验证备选种子,建立树状评价模型,设计种子用户的评价体系,由评价体系的最后得分寻找出种子用户。

关键词 大数据,种子用户,信息流,信息流浓度,树状评价模型

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.7.051

Precise Identification of Seed Users Based on Information Flow in Big Data

XIE Yang-xiao-jie ZHAO Ling

(College of Mathematics and Software Science, Sichuan Normal University, Chengdu 610066, China)

Abstract Aiming at the precise identification of data seeds under big data, we analyzed two major factors which impact users to become seeds users: time priority and attribute characteristics, and two characteristics of the dissemination of seed information: propagation time difference and directionality. Accordingly, we proposed a method to quickly find the seed users. First, users are put into different groups by the property features. Through analyzing the time difference and SMS circulation among all groups, we can find out the dissemination of information flow, that is to say, direction. Thus the search range is gradually narrowed, and alternative seed is filtered through threshold. We established evaluation model tree, designed seed users evaluation system, and used this evaluation system to calculate the final score to find out the seed users.

Keywords Big data, Seed user, Information flow, Information flow density, Tree network evaluation model

1 引言

随着科技的进步,信息社会已经进入了大数据时代。大数据的涌现不仅改变着人们的生活与工作方式、企业的运作模式,甚至还引起科学研究模式的根本性改变[1]。一般意义上,大数据是指无法在一定时间内用常规机器和软硬件工具对其进行感知、获取、管理、处理和服务的数据集合[2]。现代信息传递的方式信息也多种多样,短信是其中的重要方式之一,载体有手机、微博、QQ等通讯工具。识别短信种子用户,对于控制信息传播、发掘信息源头和分析传播路径具有重要的意义。人与人的交流不再局限于地域,特别是网络信息,常常以数据流的形式动态、快速地产生,具有很强的时效性,用户只有把握好对数据流的掌控才能充分利用这些数据。

现有的关于种子用户识别的文献没有充分考虑种子信息传播的特征:传播时差和强方向性,以及成为种子用户的因素(时间优先和属性特征)。例如 Chen 等[3],Ye 和 Wu[4],Tomar 等[5]建立的模型通过考察网页的指向和引用来识别重要网页的 pagerank 算法[6-8],这是最早使用的方法,比较有启发性,但是没有充分考虑种子信息传播的特征,损失了很大部分统计量。李永立等[9]通过对每一个用户一个月内发送的短信进行分析得到可能的种子信息,再用树形网络得到短信种子用户,他们首先考虑到了时间优先性,得到了较好的效果;但是由于数据杂乱无章,没有考虑到属性特征,逐一寻找起来费时费力,时间复杂度为 $O(kN\log N)$,当应用到大数据中时已无法处理。针对这一情况,本文提出了一种在大数据下快速寻找的方法。首先分析属性特征,用户之间的联系多为同行间的交流,一般受行业和地域限制,所以属性特征为行业属性和地域属性。按行业分组一方面能很好地将信息转发现象集中起来;另一方面,种子用户除了引起组内的大量转发也会往其他组传播,所以按行业分组是比较合理的方式。通过分析组间短信流通的情况,从而找到短信流动的方向。种子用户的特征是能引起大量的转发现象,形成信息流,如果能掌握信息流动

到稿日期:2016-02-05 返修日期:2016-04-10 本文受国家自然科学基金重大研究计划:可信网络交易软件系统试验环境与示范应用(91218301),四川省教育厅重点项目:基于三阶段 DEA 方法对我国地区 R&D 投入绩效的评估及四川省 R&D 投入绩效分析(13sa0137)资助。

谢杨晓洁(1989-),女,硕士,助教,主要研究方向为数理统计大数据处理及算法优化,E-mail:835198517@qq.com;赵 凌(1964-),女,副教授,主要研究方向为大数据处理及算法优化。

的大致方向和时间，就可以放掉大部分无用的信息，在剩下的数据中寻找起来针对性强，效率也比较高，时间复杂度为 $O(N)$。

2 数据的预处理

2.1 数据分块

因为短信用户间联系范围受属性特征影响较大，所以依据属性的不同将信息发送数据分到不同的组 A_{ij} 中，$i,j\in\lambda$，$\lambda=\{1,2,\cdots,n\}$。其中，n 为属性特征的总类别个数，i 表示信息发送者的属性特征，j 表示接收者的属性特征。组 A_{ij} 中发送短信的数量记为 Λ^{ij}。现将具有 A_{ij} 和 A_{ji} 的组的短信数量进行两两相减得到信息流通差 a_{ij}，即比较两组间的信息流通量。公式表示为：

$$a_{ij}=\begin{cases}\Lambda^{ij}-\Lambda^{ji}, & i\neq j\\ \dfrac{\Lambda^{ij}}{2}, & i=j\end{cases}$$

其中，组 A_{ij} 中发送短信的数量记为 Λ^{ij}。由此得到一张信息流通情况的表格，这样就可以很直观地看到组内以及组间收发短信的情况，如表 1 所列。

表 1 信息流通情况表

行业	1	…	n
1	a_{11}	…	a_{1n}
…	…	…	…
n	a_{n1}	…	a_{nn}

2.2 数据清洗

由信息流通差可以得到组间平均信息流通差，记作 $\bar{a}$，公式为：

$$\bar{a}=\frac{\sum a_{ij}}{n^2}$$

通过分块后得到的组 A_{ij} 需要进行一次清洗，即删掉一部分组。清洗准则如下：

1)若 $a_{ij}\geqslant\bar{a}$，则信息是由 i 往 j 传播的，保留。

2)若 $a_{ij}<\bar{a}$，则信息没有传播，删除组 A_{ij} 和 a_{ij}。

2.3 数据整理

现对保留下来的组进行整理，方法是将 a_{ij} 下标数字首尾相同的放入同一集合 $\Omega_i=\{i,j,\cdots,r\}$，$i,j,r\in\lambda$。由此，可以得到信息传播的大致走向，如图 1 所示。

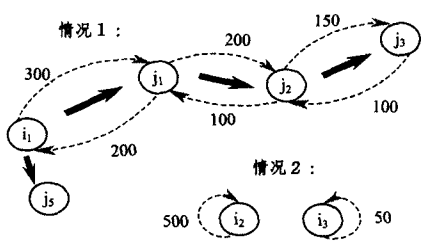

图 1 信息传播的大致走向图

情况 1：从图 1 中可以看到属性特征 i 为种子信息的位置，一直传到属性特征 j_3，那么 j_2 组和 j_3 组就为中转信息所在的位置。这时可得到一个信息流集 $\Omega_{i_1}=\{i_1,j_1,j_2,j_3,j_4,j_5\}$。

情况 2：分析属性特征 i_2 和 i_3，i_2 和 i_3 跟其他组没有交流，但 i_2 组自身有大量的信息在收发，所以可以考虑该组内有种子用户的可能性，得到集合 $\Omega_{i_2}=\{i_2\}$。而 i_3 的内部短信量较少，对其不予考虑，直接排除。

对于属性特征，若数据量在某个小区域内范围较小，就按行业属性进行上述处理即可。若为大数据范围覆盖全国大多数区域时，先按区域属性进行处理，找到信息发送的源头区域，在该区域内用行业属性再进行一次上述处理。

3 模型的建立

3.1 时间的划分

本文对短信数据以 T 小时的时间周期为单位进行分批处理，这是符合客观事实的，因为种子信息的传播具有时效性，所以分批分时段处理更为合理。这里假定种子信息在发给接受者后，t_0 分钟内为转发的有效时间。

将每个时间周期 T 中的 A_{ij} 每隔 t_0 分钟划为一段，统计各时段 t_k^{ij} 的短信发送数量 Λ_k^{ij}，其中 $k=1,2,\cdots,\dfrac{T\times 60}{t_0}$。从中找到起始时段 Γ_i^0 和峰值时段 Γ_i^1，公式为：

$$\Gamma_i^0=\{T_{ij}^0\mid T_{ij}^0=t_k^{ij},\Lambda_k^{ij}>\overline{\Lambda^{ij}},\Lambda_{k+1}^{ij}>\Lambda_k^{ij}\}$$

$$\Gamma_i^1=\{T_{ij}^1\mid T_{ij}^1=t_k^{ij},\Lambda_k^{ij}>\Lambda_{k-1}^{ij},\Lambda_k^{ij}>\Lambda_{k+1}^{ij}\}$$

其中，$\overline{\Lambda^{ij}}$ 为 A_{ij} 中所有 Λ_k^{ij} 的均值。现在，需要找到种子信息的路径来定位起始组，即种子信息所在组。首先将之前得到的集合 Ω_i 中的 Γ_{ij}^0 和 Γ_{ij}^1 放在一起排序，挑出满足以下条件：

$$\begin{cases}T_{ij_1}^0<T_{j_1j_2}^0<T_{j_2j_3}^0\cdots\\ T_{ij_1}^0<T_{j_1j_2}^1,T_{j_1j_2}^0<T_{j_2j_3}^1,\cdots\end{cases}$$

的路径即为 $i\to j_1\to j_2\to j_3\to\cdots$，由此一个集合 Ω_i 里可能有多条路径，或长或短，都要保留。路径的起始组就为 A_{ij_1}，可以根据 A_{ij_1} 里两个时段的位置来推断种子信息发送的时间范围 T^*，因为种子信息发出到大量转发有一个时间差，也即在出现峰值的时间前在起始段附近，即 $T^*=(\min(T_{ij}^0),\max(T_{ij}^1))$。此时就需要具体到组头 X 中 T^* 时段每一条信息的发送情况。种子用户的一大特征是广度，即群发数量比较大，所以将该时段中发出短信量大于该时段平均短信量的用户挑出来作为备选种子用户 $V_i(i=1,2,\cdots)$。

3.2 判断种子用户

种子用户的另一大特征为深度，即信息发出后能多次转发，所以对备选的种子用户还要进行一次检验。方法是将每一个 V_i 里的用户信息在满足转发时间的发送情况下全部找出，得到它的短信传播的树状结构，如图 2 所示。

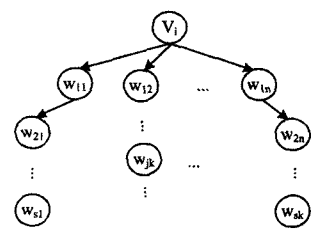

图 2 短信传播的树状结构

在该树状结构中，短信的首个发出者为备选种子 V_i，将所有短信的转发者记为 w_{jk}，$j=1,2,\cdots,s,k=1,2,\cdots$，其中 j 为所在层数，k 为 j 层上的第 k 个转发者。V_i 的群发数量(广度)对应第 2 层的节点数目，记为 n；w_{jk} 的群发数量记为 g_i；该短信的转发度量对应树的层数(深度)，记为 s；短信在一条链上传播的节点数(即层数)记为 l，其中 $l_{\max}=s$。

由此，针对第 V_i 个用户的树状结构饱满程度评分 P_{V_i}，由下式给出：

$$P_{V_i}=\frac{s\cdot n}{(l_{\max}-\bar{L})(g_{\max}-\bar{G})}$$

其中，$\bar{L}$和$\bar{G}$分别是该树状结构内所有用户 l 和 g 的平均值。式中 $s\cdot g_1$ 刻画了树状图框架的大小，值越大框架也就越大。而$(l_{\max}-\bar{L})(g_{\max}-\bar{G})$刻画的是树状图的饱满程度，值越小越饱满。所以整体上值越大时，该用户为种子用户的概率就越高。

3.3 时间复杂度

对该模型的时间复杂度分析如下：数据预处理的计算时间复杂度为 $O(N)$，3.1 节的计算时间复杂度为 $O(T\times 60/t_0)$，3.2 的计算时间复杂度为 $O(l_{\max})$，则总的计算时间复杂度仅为 $O(N)$。其中，N 表示数据样本短信发送的个数。

4 应用实例

4.1 数据来源

本文中用于分析的数据以中国移动在深圳于 2012 年 2 月 1 日全市手机短信用户发送的数据为例。有两张表，一张表为用户信息情况，里面有用户 id 和用户的行业（用数字标识，总共有 55 个行业类别）。另一张表为短信发送情况，里面有短信发出者 id、短信接收者 id 以及短信发送时间。

4.2 具体步骤

首先，将短信发送情况表按发送者和接收者的行业不同分到不同的块 varX-Y(X,Y=1,…,55)中，得到信息流通差表，其中$\bar{a}=11$，筛掉 $a_{ij}<11$ 的数据后，结果如表 2 所列。

表 2 信息流通差表

行业	1	2	15	28	47	57	58
1	10896	—	−13	—	10	167	11
2	—	78	—	—	—	−18	—
15	13	—	135	—	—	11	—
28	—	—	—	49	—	23	—
47	−10	—	—	—	—	−11	—
57	−167	18	−11	−23	11	106023	—
58	−11	—	—	—	—	—	50

筛掉无关的组，由此找出信息传播的大致走向，如图 3 所示。

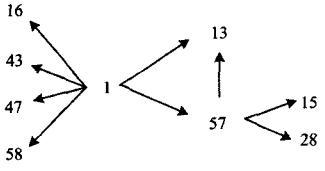

图 3 信息传播的大致走向

其次，对余下来的表 varX-Y 进行时间的划分，这里时间周期 T 定为 24h，选择 t_0 为 20min，即将每个表分为 54 段，得到各个段的起始时段 Γ_X^0 和峰值时段 Γ_X^1，用来判断图 3 中的路径是否满足本文 3.1 节的 2 个条件，由此得到几条较为合理的信息传播路径{1−57−15,1−57−28,1−13,1−58,57−1}，此时起始组为 1 和 57，其中 $\Gamma_1^0=\{7,11,15,19\}$，$\Gamma_1^1=\{12,14,16,21\}$，$\Gamma_{57}^0=\{9,11,14,24,34\}$，$\Gamma_{57}^1=\{10,12,15,25,35\}$，所以在表 var1-X 和 var57-X 中去找种子信息发送时间的范围是：

$T^*=\{(7,12),(11,14),(15,16),(19,21),(9,10),(14,12),(24,25),(34,35)\}$

在该范围中找到 9 个备用种子用户为：

$V_i=\{688612,751869,1255966,1065357,\cdots\}$

再将这些用户用树状评价法进行打分，其相关情况如表 3 所列。

表 3 9 个备用种子用户的 ID

用户 id	群发数	所在表	得分
688612	35	varl-1	135
751869	30	var 1-1	98.5
1299652	12	var 57-57	92
1298271	10	var 57-57	86
1255966	18	varl-1	36
1065357	14	varl-1	21
596644	10	varl-1	20
744816	10	var 57-57	15
122249	12	var 57-57	12

选出得分最高的种子用户的 ID 即前面 4 个，将他们作为种子用户。从表 3 中还可以看出，种子用户和非种子用户的得分具有很好的区分度（两类之间的最小差距），在同一时间段中的不同性质的用户的得分差距也很大。此外，在表 4 中，通过运行时间和备选种子个数可以明显看出该算法的优良性。

表 4 两种方法的比较

对比项	基于信息流	基于树形网络
运行时间	0.01s	4.3s
区分度	50	32
备选种子个数	9	2468
种子个数	4	5

结束语 本文基于树形网络分析的短信种子用户挖掘模型进行改进，将树形网络模型简化成树状评价模型，用于做最后的检验。充分利用成为种子用户的两大因素以及种子信息传播的两大特征来寻找备选种子。通过合理地划分，找到信息流的大致位置，进一步找到信息流的源头。逐步有目的地缩小寻找种子用户的范围，极大地减少了对数据处理的工作量，提高了效率。其次，分析种子用户的特征指标，得到树状模型。最后设计评论体系结构，考虑到 2 个指标之间的关联性，采用综合评价函数，将提炼出的几个指标有效地融合成一个能表示该用户是否成为种子用户的最终指标。

参 考 文 献

[1] Wang Yuan-zuo, Jin Xiao-long, Cheng xue-qi. Network big data: present and future[J]. Chinese Journal of Computers, 2013, 36(6): 1125-1138(in Chinese)
王元卓，靳小龙，程学旗. 网络大数据：现状与展望[J]. 计算机学报，2013，36(6)：1125-1138

[2] Li Guo-jie, Cheng Xue-qi. Research Status and Scientific Thinking of Big Data[J]. Bulletin of Chinese Academy of Sciences, 2012, 27(6): 647-657

[3] Chen J, Subramanian, Brewer E. SMS-based Web search for lowend mobile devices[C]//Proceedings of the 16th Annual International Conference on Mobile Computing and Networking. New York, 2010: 125-136

[4] Ye Shao-zhi, Wu S F. Measuring message propagation and social influence on Twilter. com [J]. Lecture Notes in Computing Sci-

ence,2010,6430:216-231

[5] Tomar V,Asnani H,Karandikar A,et al. Social network analysis of the Short Message Service[C]// 2010 Natinal Conference on Communications. 2010:1-5

[6] Ma Nan,Guan Jian-cheng,Zhao Yi. Bringing pagerank to the citation analysis [J]. Information Processing & Management, 2008,44(2):800-810

[7] Ding Ying,Yan E,Frazho A,et al. Pagerank for ranking authors in co-citation networks [J]. Journal of the American Society for Information Science and Technology,2009,60(11):2229-2243

[8] Franceschet M. Pagerank: standing on the shoulders of giants [J]. Communications of the ACM,2011,54(6):92-101

[9] Li Yong-li,Wu Chong,et al. A Tree-network Model for Mining Short Message Services Seed Users and Its Empirical Analysis [J]. Chinese Journal of Management Science,2012(20):49-54 (in Chinese)
李永立,吴冲,等. 基于树形网络分析的短信种子用户挖掘模型及其实证分析[J]. 中国管理科学,2012(20):49-54

(上接第 267 页)

象在所选择的 4 个分类模型中均有明显表现。这一结果暗示着网络节点的数量对分类模型的构建具有直接的影响。据分析,在大节点规模模板中,由于节点数量的增加,每个节点的体积(体素数量)减少,这将意味着所采集的功能影像数据更容易受到噪音信号的影响。而小节点规模模板能够更好地中和噪音,但节点体积的增大会造成脑区功能特化的模糊。所以,在节点规模的选择上,必须要考虑到二者的平衡,才能更好地表达脑网络基本拓扑结构。

结束语 本研究利用了不同节点规模定义的模板,对正常组及抑郁组进行了静息态功能脑网络拓扑属性的分析及对比,并进行了机器学习研究。结果发现,无论正常组还是抑郁组,不同节点规模下网络拓扑属性均表现出相似的变化规律。这一结论说明,节点规模对正常组和抑郁组脑网络拓扑属性具有相同的影响。同时,不同节点规模下的分类模型效果差异很大,当节点数目为 250 时,各分类器均表现出最好的效果。上述结论提示在进行脑网络指标分析特别是机器学习研究时,需要考虑到所采用的节点规模的影响。

参 考 文 献

[1] ERDdS P, Wi A. On random graphs [J]. I. Publ. Math. Debrecen,1959,6:290-297

[2] Albert R, Barabási A L. Statistical mechanics of complex networks[J]. Reviews of Modern Physics,2002,26(1)

[3] Lynall M E,Bassett D S,Kerwin R,et al. Functional connectivity and brain networks in schizophrenia[J]. The Journal of Neuroscience,2010,30(28):9477-9487

[4] Stam C. Use of magnetoencephalography (MEG) to study functional brain networks in neurodegenerative disorders[J]. Journal of the Neurological Sciences,2010,289(1):128-134

[5] Horstmann M T,Bialonski S,Noenning N,et al. State dependent properties of epileptic brain networks:Comparative graph-theoretical analyses of simultaneously recorded EEG and MEG[J]. Clinical Neurophysiology,2010,121(2):172-185

[6] Liang Wang,Zhu Chao-zhe,He Yong,et al. Altered small-world brain functional networks in children with attention-deficit/hyperactivity disorder[J]. Human Brain Mapping, 2009, 30(2): 638-649

[7] De Vico Fallani F, Laura A, Febo C, et al. Evaluation of the brain network organization from EEG signals:a preliminary evidence in stroke patient[J]. The Anatomical Record, 2009, 292 (12):2023-2031

[8] Guo H,C C ,Cao Xiao-hua,et al. Resting-state functional connectivity abnormalities in first-onset unmedicated depression [J]. Neural Regen Res. ,2014,9(2):153-163

[9] Guo H,Cao X H,Liu Z F,et al. Machine learning classifier using abnormal brain network topological metrics in major depressive disorder[J]. Neuroreport,2012,23(17):1006-1011

[10] Tzourio-Mazoyer N,Landeaub B,Papathanassioua D,et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain[J]. Neuroimage,2002,15(1):273-289

[11] Collins D L,Holmes C J,Deters T M,et al. Automatic 3-D model-based neuroanatomical segmentation[J]. Humanbrain Mapping,1995,3(3):190-208

[12] Salvador R,Martinez A,Pomaral-Clotet E,et al. A simple view of the brain through a frequency-specific functional connectivity measure[J]. Neuroimage,2008,39(1):279-289

[13] Williams J B. A structured interview guide for the Hamilton Depression Rating Scale[J]. Archives of General Psychiatry,1988, 45(8):742-747

[14] Rubinov M,Sporns O. Complex network measures of brain connectivity:uses and interpretations[J]. Neuroimage,2010,52(3): 1059-1069

[15] Latora V, Marchiori M. Efficient behavior of small-world networks[J]. Physical Review Letters,2001,87(19):198701

[16] Watts D J,Strogatz S H. Collective dynamics of 'small-world' networks[J]. Nature,1998,393(6684):440-442

[17] Humphries M D,Gurney K,Prescott T J. The brainstem reticular formation is a small-world,not scale-free,network[J]. Proceedings of the Royal Society B:Biological Sciences, 2006, 273 (1585):503-511

[18] Guo Li-li,Ding Shi-fei. Rearch Progress on Deep Learning[J]. Computer Science,2015,42(5):28-33(in Chinese)
郭丽丽,丁世飞. 深度学习研究进展[J]. 计算机科学,2015,42 (5):28-33

[19] Pereira F,Mitchell T,Botvinick M. Machine learning classifiers and fMRI:a tutorial overview[J]. Neuroimage,2008,45(1 Suppl):S199-209

[20] Bullmore E,Sporns O. Complex brain networks:graph theoretical analysis of structural and functional systems[J]. Nature Reviews Neuroscience,2009,10(3):186-198

[21] Cox D D,Savoy R L. Functional magnetic resonance imaging (fMRI) "brain reading":detecting and classifying distributed patterns of fMRI activity in human visual cortex[J]. Neuroimage,2003,19(2):261-270