



基于内容分析的短信种子客户挖掘模型与算法

黄志超¹, 陶俊才¹, 高胜保²

(1. 南昌大学信息工程学院计算中心, 江西 南昌 330029;

2. 中国电信股份有限公司江西分公司, 江西 南昌 330029)

摘要: 为从海量的短信记录中挖掘短信种子客户, 控制种子短信的传播路径, 提高其传播效率, 提出了一种基于内容分析的短信种子客户挖掘模型与算法。首先通过分析客户转发短信的兴趣性、随机性、单向性特征, 构建客户转发短信的树型模型; 其次, 通过定义和应用综合评价函数生成优化的种子客户挖掘模型, 并基于亲密群概念实现短信种子客户的挖掘; 最后, 使用电信运营商的实际数据进行实证分析, 验证了上述模型与算法的有效性。

关键词: 短信种子客户; 内容分析; 挖掘模型; 挖掘算法; 亲密群

中图分类号: TP311

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2016057

SMS seed customers mining model and algorithm based on content analysis

HUANG Zhichao¹, TAO Juncai¹, GAO Shengbao²

1. Computing Center, Information Engineering College, Nanchang University, Nanchang 330029, China

2. Jiangxi Branch of China Telecom Co., Ltd., Nanchang 330029, China

Abstract: In order to mining SMS (short message service) seed customers from massive text messages, control the spread of the seed messages path and improve the efficiency of its spread, a SMS seed customers mining model and algorithm was proposed, which was based on content analysis. First of all, by analyzing the interest, randomness and one-way characteristics of customer forwarding messages, the tree model of customer forwarding messages were constructed. Secondly, the optimal seed customers mining model was generated by definition and application of comprehensive evaluation function, and SMS seed customers mining was realized based on the concept of close group. Finally, by analyzing the actual data from telecom operators, the effectivity of the model and algorithm was verified.

Key words: SMS seed customer, content analysis, mining model, mining algorithm, close group

1 引言

随着互联网的发展和智能终端的普及, 短信、微博、微信等现代信息传播手段被广泛应用, 它们都具有使用便捷、传播快速等优点。相对而言, 短信对受众的年龄与知识

水平、移动终端功能的要求较低而具有特定的优势, 更适合信息广播、知识普及、亲情交流等场景。种子短信指承载某些特定信息的定制短信, 如: 政府发布的灾害预警、信息发布, 企业的客服信息, 增值服务商的亲情交互、幽默信息等。种子客户则是种子短信的首轮接收/转发者, 由于他们

收稿日期: 2015-06-13; 修回日期: 2016-01-07

基金项目: 国家自然科学基金资助项目 (No.61262049)

Foundation Item: The National Natural Science Foundation of China (No.61262049)

应当对特定种子短信具有尽可能高的兴趣与短信转发量,因此对于信息的传播效率与成本具有决定性的影响。然而,目前确定种子客户的方法基本凭借人工经验寻找和筛选,其效率和准确率极不理想。因此,从海量的短信记录中挖掘出优质的短信种子客户,对于控制种子短信的传播路径、提高传播效率、降低传播成本,具有重要意义。

短信传播具有极强的兴趣性、随机性、单向性特征。兴趣性指不同的短信客户对于不同类型的短信(新闻、体育、幽默、益智、养生等)往往具有较强的兴趣偏好,因而种子短信的生成/首发者(以下称为短信中心)一般需将种子短信分成多个类别,以提高客户转发同一类型短信的概率,这就决定了需要通过短信内容的关联分析来识别同一客户感兴趣的短信类别,并有必要针对不同兴趣的客户群分类建立不同的挖掘模型。随机性是指短信客户转发短信的对象可能是随意的,也可能具有相对固定的客户群和一定的规律,为了提高种子短信的转发效率与传播速度,就需要从所有转发种子短信的客户中挖掘出规律性较强和拥有相对固定客户群的种子客户。单向性特征是指种子短信在传播过程中不会发生直接回传的现象,因为接收者没有必要把内容相同的短信再回发给短信的发出者,而对于可能出现的循环回传的种子短信不可能再被转发,故其回传没有增加信息的传播,对于识别种子客户没有任何价值,必须予以剔除。

既有的关于挖掘短信种子客户的可参考文献很少,参考文献[1]中关于构建树型网络的思想比较具有参考价值。然而,由于以下主要原因使其不适用于本文提出的短信种子客户挖掘任务:首先,其挖掘目标(短信种子用户)是指短信的创建/首发者,而不是短信首次接收/转发者;第二,其基于短信的时域特征而不是内容进行关联分析,准确性不够;最后,其挖掘路径是在树型网络中自底向上进行,与本文的要求正好相反。本文的主要工作在于:定义并构建新的基于内容分析的短信种子客户挖掘模型,提出新的种子客户挖掘算法,并基于完整的实际短信数据进行实证分析,验证本文研究成果的有效性。

2 原始模型构建

2.1 建立树型模型描述种子短信的转发关系

短信中心创建并向客户发送种子短信,客户对感兴趣的种子短信进行转发,这种转发过程与关系可以用图1所示的原始模型来描述,其中的节点表示客户,边为客户之

间的转发关系。节点的属性包括:发送号码、接收号码、短信内容、发送时间。如前所述,在该模型中,不可能出现直接回传的短信,而对于可能发生的循环回传(图1中的回路),则可用最小生成树算法予以剪除,从而得到如图2所示的树型模型。其中,第0层节点为短信中心,其发出的短信称为种子短信。第1层的各节点是种子短信的首轮接收/转发者,即前文所定义的种子客户,以下的节点则为其其他客户。初始的种子客户只能根据经验通过人工筛选得到,之后则可应用本文研究成果自动挖掘优质的短信转发客户,作为种子客户群。

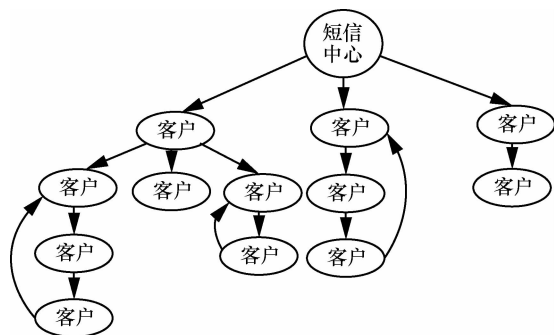


图1 种子短信转发的基本模型

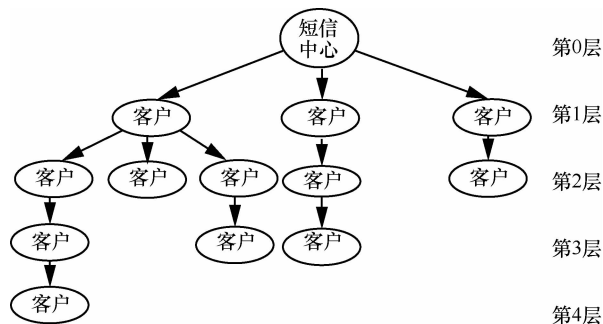


图2 种子短信转发的树型模型

2.2 模型的分类

上述模型是针对某一类种子短信而建立的,如前所述,针对客户转发短信的兴趣性特征,短信中心一般需将种子短信分成多个类别,从而针对不同兴趣的客户群分类建立不同的挖掘模型。有多种算法可用于实现种子短信的分类^[2-6],本文采用简单的KNN算法^[7]来实现。但为了提高挖掘种子客户的效率与准确率,需要进一步剔除冗余短信,为此需要对KNN算法进行下述调整:在分类结果中,一般存在与训练数据距离为0的测试数据,KNN算法的处理方法是直接将它们全部归入一类,实际上这些数据中包含着不少冗余数据,因此本文进一步采用内容匹配对其予以发现并剔除。



3 种子客户挖掘算法

从图2的树型模型中挖掘种子客户,提出短信种子客户挖掘(seed customer mining, SCM)算法,该算法分为3个阶段。

3.1 发现和剔除一次转发量过小的节点

图2中有不少节点的一次转发量很小,这些节点对于计算结果的影响很小,为了降低计算复杂度,应当发现和剔除这些节点,为此作如下定义。

- 一次转发量:某节点向其下邻层节点转发的种子短信数。
- 一次转发量阈值 α :需剔除节点的一次转发量的上限。

显然,当阈值 α 取值大于图2中最大的节点一次转发量时,优化模型将只剩下根节点;当其取值为0时,则优化模型将与图2相同,其实际取值需要结合使用者的需求、经验,通过实验来调整和确定。本阶段将计算所有节点的一次转发量,将其中一次转发量小于 α 的节点剔除,以便在以下处理过程中被忽略,从而生成一次优化模型。

3.2 发现和剔除综合评分过低的节点

在一次优化模型的基础上,进一步发现和剔除其中综合评分过小的节点,得到二次优化模型,为此作如下定义。

- i 子树:以任意子节点 i 为根形成的种子短信转发子树。
- 评分阈值 β :需剔除节点的综合评分上限,其取值方法类似于 α 。

节点 i 的评分需要综合考虑 i 子树各层节点的转发短信数及其与节点 i 的距离,为此定义综合评价函数如下:

$$M_i = \sum_{j=1}^m \frac{L_j}{j}, 0 < i \leq n \quad (1)$$

其中, i 为第 i 个节点; M_i 为第 i 个节点的得分; j 为 i 子树的第 j 层; L_j 为第 $j-1$ 层向第 j 层转发的种子短信数; n 为节点的个数; m 为 i 子树的层数。在实际情况中,客户可能转发同一类型中的多条短信,从而一个客户可能存在多个节点评分,将其相加作为该客户的最终评分。根据式(1),显然可得出以下结论:

- i 子树的各层转发短信数越大,节点 i 的评分越高;
- 若 i 子树中某一节点与节点 i 的距离越远,则其转发短信数对于节点 i 的评分的贡献越小。

为便于估算和调整阈值 β ,使用下面给出的式子对节点综合评分进行归一化处理,使 β 的取值范围为[0,1]:

$$y_i = \frac{M_i - \text{Min}}{\text{Max} - \text{Min}}, 0 < i \leq n \quad (2)$$

其中, y_i 为客户 i 归一化的结果, M_i 为客户 i 的评分, Min 为客户评分的最低分数, Max 为客户评分的最高分数, n 为客户的个数。

通过计算一次优化模型中每个节点的综合评分,将其中评分小于 β 的节点剔除,即得到二次优化模型。显然,当 β 取值为0时,则二次优化模型将与一次优化模型相同;当其取值为1时,则二次优化模型只会保留评分最大的少数节点。

3.3 基于亲密群概念挖掘种子客户

(1) 基本思路

以二次优化模型为基础,进一步发现亲密群,并将其子节点予以剔除。定义如下。

- 亲密群:存在必然转发关系的父子节点。
- 亲密群阈值 γ :判定父子节点是否为亲密群的评分标准,小于 γ 者不是亲密群,否则为亲密群,其取值范围为[0,1],取值方法类似于 α 。

例如,节点 A 在接收到种子短信后一定转发给节点 B ,则 A 、 B 构成一个亲密群。显然,需将节点 B 剔除。

(2) 算法过程

本阶段的算法过程如图3所示,其中,图3(d)中集合 W 由式(3)得到:

$$W = L \times C = \{ \langle x, y \rangle | x \in L \wedge y \in C \wedge x \neq y \} \quad (3)$$

集合 C 和集合 L 分别为图3(b)和图3(c)所示; $\langle A, B \rangle$ 表示客户 A 转发短信给客户 B ;计数 Q 表示图3(a)中的父子节点出现的次数。

图3(f)中集合 S 由式(4)得到:

$$H_{ij} = \frac{P_{ij}}{R_i} \times 100\%, 0 < i \leq m, 0 < j \leq n \quad (4)$$

其中, H_{ij} 为父子节点出现的概率, P_{ij} 为客户 i 转发给客户 j 的种子短信条数, R_i 为客户 i 转发的种子短信条数, m 为集合 L 的元素个数, n 为集合 C 的元素个数。

(3) 算法实现

输入: $Users$ 表示二次优化模型中的节点, 树型结构为 T , 阈值为 γ , MinC 表示节点出现的最小次数。

输出: 亲密群结果集 Result , 种子客户群 EndUsers 。

步骤1 计算 $Users$ 中每个节点出现的次数即 $C = [C_1, C_2, \dots, C_n]$, 其中 n 表示 $Users$ 个数;

步骤2 初始化结果集 $\text{Result} = []$, $\text{flag} = 0$, $\text{close} = []$, $\text{EndUsers} = []$, $k = 1$;

步骤3 For $i = 1 : n$

If ($C_i \geq \text{MinC}$)

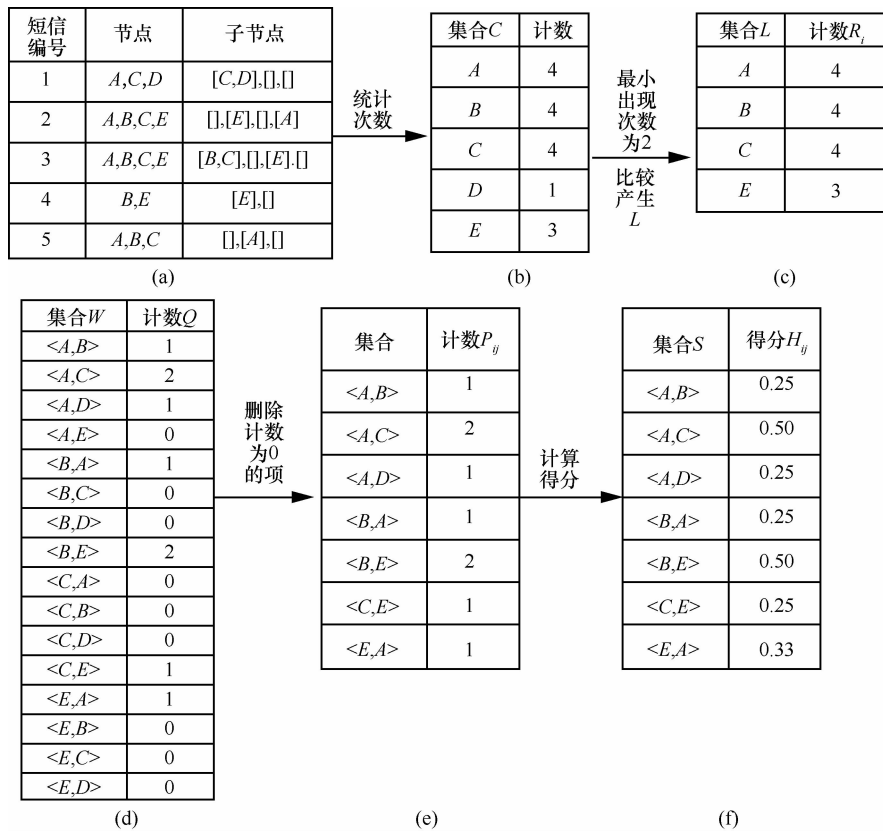


图3 本阶段算法过程

Children=Users[i]的孩子节点且该节点 $\in$ Users
 For $j=1:\text{length}(\text{Children})$
 $ct=T$ 中出现Users[i] $\rightarrow$ Children[j]的次数
 If ($ct \neq 0$)
 $ct = \frac{ct}{C_i} \times 100\%$
 If ($ct \geq \gamma$)//亲密群客户
 $\text{Result} = [\text{Result} \text{ Children}[j]];$
 End If
 End If
 End For
 End For
 步骤4 For $i=1:n$
 If ($i \notin \text{Result}$)
 $\text{EndUsers} = [\text{EndUsers} \ i];$
 End If
 End For
 步骤5 算法结束,输出结果。
 上述步骤中,步骤1和步骤2的时间复杂度为 $O(n)$,其中,

n 为二次优化模型中的节点数;步骤3的时间复杂度为 $O(n^2m^2)$,其中, m 表示二次优化模型中父子节点的子节点数,由于一般 m 远远小于 n ,所以该阶段总的时间复杂度为 $O(n^2)$ 。

4 实证分析

实证数据来源于某省电信分公司,从2015年2-3月份的短信中随机抽取共1万条手机短信和100条种子短信,格式见表1。

- 种子短信集合用作训练集,1万条手机短信用作测试文本集,使用KNN算法析取出3类种子短信:情人节祝福短信、除夕拜年短信、元宵祝福短信。
- 对3类短信分别构建树型转发模型。
- 发现和剔除一次转发量过小的节点(α 取值3),得到一次优化模型。
- 在一次优化模型的基础上,应用第3.2节中的评价式(1)、式(2)对于不同短信类型计算每个客户的评分,见表2。其中的类型1、2、3依次对应情人节种子短信、春节短信、元宵节短信,用序号表示不同的客户。将表2中的评分结果与阈值 β 比较,剔除小于 β 的节点(β 取值0.7),得到二次优化模型。



表 1 客户短信数据格式

发送号码	接收号码	短信内容	发送时间
1807971****	1361707****	^_^情人节各种“情”来聚会,热“情”藏在鲜花里,深“情”粘在糖果里,...	2015 年 2 月 14 日 12:44:10
1807961****	1517966****	^_^羊年来到,祝您在新的一年里:所有的希望都能如愿,所有的梦想都能实现,...	2015 年 2 月 14 日 12:44:45
...	...	...	...

表 2 客户评分

类型	客户	评分	类型	客户	评分	类型	客户	评分
1	2	1	...	...	...	1	30	0.95
1	11	1	3	278	1	2	187	0.95
1	34	1	3	282	1	1	120	0.92
1	61	1	3	283	1	1	9	0.90
1	72	1	3	285	1	1	49	0.90
1	76	1	3	289	1	2	142	0.90
1	99	1	3	291	1	2	165	0.85
1	129	1	3	307	1	1	20	0.80
2	143	1	3	309	1	3	281	0.80
2	152	1	3	313	1	1	28	0.75
2	221	1	3	323	1	3	271	0.75
3	244	1	2	156	0.99	2	150	0.73
3	255	1	1	58	0.97	...	...	...
3	256	1	3	305	0.97			

· 进一步应用亲密群概念进行挖掘,得到的种子客户见表 3(γ 取值 0.65)。

由表 2 和表 3 可知,从 1 万条手机短信中挖掘出的种子客户数为 40 名。为进一步分析阈值 α 、 β 、 γ 对挖掘结果的影响,另外选取 2 万条短信与上述结果进行比较,相关情况如图 4、图 5 所示。图 4 为设定 $\beta=0.7$ 、 $\gamma=0.65$,种子客户数随 α 变化的结果,当 $\alpha \geq 2$ 时,曲线趋于平缓;图 5 为

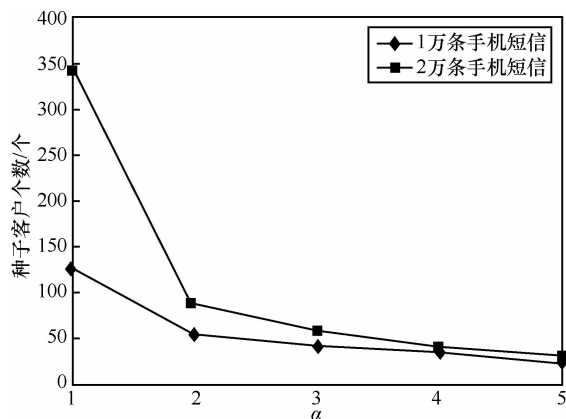
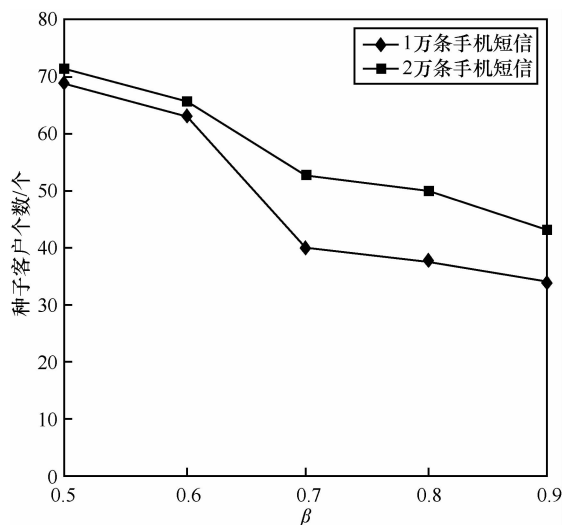
设定 $\alpha=3$ 、 $\gamma=0.65$,种子客户数随 β 变化的结果,当 $\beta \geq 0.7$ 时,曲线趋于平缓;这些变化趋势与实际情况完全吻合。因此,可以选择 $\alpha=2$ 、 $\beta=0.7$ 作为挖掘种子客户时的参考值,必要时可以根据需要调整。

5 结束语

挖掘和利用短信种子客户,可以有效控制短信传

表 3 种子客户

类型	用户	类型	用户	类型	用户
1	2	2	150	3	278
1	9	2	152	3	281
1	11	2	165	3	282
1	20	2	187	3	285
1	28	2	221	3	289
1	30	3	244	3	291
1	34	3	255	3	305
1	58	3	256	3	307
1	61	3	262	3	313
1	72	3	264	3	323
1	76	3	268	3	341
1	99	3	271	3	343
1	120	3	272	...	...
2	142	3	275		

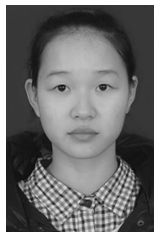
图4 种子客户数随 α 的变化关系图5 种子客户数随 β 的变化关系

播路径、提高传播效率、降低传播成本。为了从海量的短信记录中快速、准确地挖掘出种子客户,用作各类种子短信的首发客户,本文综合考虑客户转发短信的兴趣性、随机性、单向性特征,提出了一种基于内容分析的短信种子客户挖掘模型与挖掘算法。首先,针对种子短信按客户兴趣分类的特点,采用文本分类方法分别建立描述短信转发关系的原始树型模型;然后,基于本文提出的挖掘算法,对原始模型逐步优化,并提出亲密群的概念,最终挖掘出种子客户群。最后,使用实际数据进行实证分析,验证了本文研究成果的有效性。值得指出的是,本文研究成果也适用于微信等公共媒体的种子用户挖掘。

参考文献:

- [1] 李永立, 吴冲, 胡冬冬, 等. 基于树型网络分析的短信种子客户挖掘模型及其实证分析[J]. 中国管理科学, 2012(S1): 48-54.
LI Y L, WU C, HU D D, et al. The SMS seed customer mining model and empirical analysis based on tree network analysis[J]. Chinese Journal of Management Science, 2012(S1): 48-54.
- [2] HARRINGTON P. 机器学习实战[M]. 李锐, 李鹏, 曲亚东, 等译. 北京: 人民邮电出版社, 2013: 32-52.
HARRINGTON P. Machine Learning in Action [M]. Translated by LI R, LI P, QU Y D, et al. Beijing: Posts and Telecom Press, 2013: 32-52.
- [3] 李兵昌. 短信种子客户识别的研究[D]. 广州: 华南理工大学, 2013: 1-52.
LI B C. The research of SMS seed users' identification [D]. Guangzhou: South China University of Technology, 2013:1-52.
- [4] 陆旭. 文本挖掘中若干关键问题研究 [M]. 合肥: 中国科学技术大学出版社, 2008: 13-29.
LU X. Research on Some Key Issues in Text Mining[M]. Hefei: Press of University of Science and Technology of China, 2008: 13-29.
- [5] 黄娟娟. 基于KNN的文本分类特征选择与分类算法的研究与改进[D]. 厦门: 厦门大学, 2014: 1-16.
HUANG J J. Research and improvement on feature selection and classification algorithms for text classification based on KNN[D]. Xiamen: Xiamen University, 2014: 1-16.
- [6] 王博. 文本分类中特征选择技术的研究[D]. 长沙: 国防科学技术大学, 2009: 1-46.
WANG B. Related technologies research on feature selection for text categorization[D]. Changsha: National University of Defense Technology, 2009: 1-46.
- [7] HAN J W, KAMBER M, PEI J. 数据挖掘概念与技术第三版 [M]. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2012: 288-319.
HAN J W, KAMBER M, PEI J. Data Mining Concepts and Techniques, Third Edition[M]. Translated by FAN M, MENG X F. Beijing: China Machine Press, 2012: 288-319.

[作者简介]



黄志超(1990-),女,南昌大学硕士生,主要研究方向为数据挖掘、知识工程、软件工程等。

陶俊才(1956-),男,南昌大学教授,主要研究方向为软件工程、网络计算机与系统集成、模式识别、知识管理与决策支持。

高胜保(1966-),男,中国电信股份有限公司江西分公司网络运营支撑事业部副主任,主要研究方向为通信网络运营、网络信息安全、云及大数据分析等。