



(12) 发明专利

(10) 授权公告号 CN 101026802 B

(45) 授权公告日 2012. 10. 17

(21) 申请号 200710087413. 8

审查员 欧阳琦

(22) 申请日 2007. 03. 16

(73) 专利权人 华为技术有限公司

地址 518129 广东省深圳市龙岗区坂田华为
总部办公楼

专利权人 电子科技大学

(72) 发明人 尚明生 傅彦 邵刚

(51) Int. Cl.

H04W 88/02 (2006. 01)

(56) 对比文件

CN 1845550 A, 2006. 10. 11,

CN 1870601 A, 2006. 11. 29,

CN 1787655 A, 2006. 06. 14,

CN 1529966 A, 2004. 09. 15,

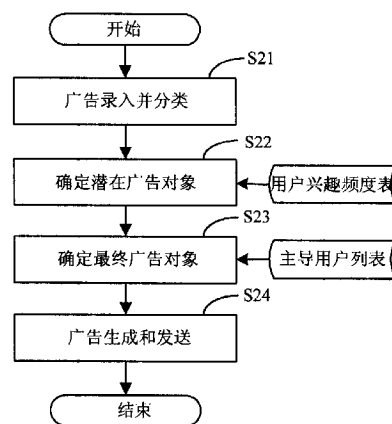
权利要求书 3 页 说明书 15 页 附图 5 页

(54) 发明名称

一种信息推送方法与装置

(57) 摘要

本发明公开了一种信息推送方法,包括:将第一信息按照第一类别集进行分类,建立所述第一信息与第一类别集中所属类别的第一映射关系;以及将信息源发送的第二信息按照第二类别集进行分类,建立发送所述第二信息的信息源与第二类别集中所属类别的第二映射关系;根据第一类别集和第二类别集中所属类别之间的对应关系,由与所述第一信息建立第一映射关系的第一类别集中各类别,匹配出第二类别集中的对应类别,并根据第二映射关系确定出对应信息源;将所述第一信息推送到确定出的对应信息源。采用本发明能够实现根据用户发送的短信(第二信息)确定出推送第一信息(广告)的相应通信终端,克服了现有技术中信息推送的盲目性。



1. 一种信息推送方法,包括:

信息推送装置将第一信息按照第一类别集进行分类,建立所述第一信息与第一类别集中所属类别的第一映射关系;

所述信息推送装置将信息源发送的第二信息按照第二类别集进行分类,建立发送所述第二信息的信息源与第二类别集中所属类别的第二映射关系;

所述信息推送装置根据第一类别集和第二类别集中所属类别之间的对应关系,由与所述第一信息建立第一映射关系的第一类别集中各类别,匹配出第二类别集中的对应类别,并根据第二映射关系确定出对应信息源;

所述信息推送装置将所述第一信息推送到确定出的对应信息源;

其中,所述信息源为通信终端,所述第二信息为通信终端发送的多条短信;在将信息源发送的第二信息按照第二类别集进行分类之前还包括获取信息源发送的第二信息的步骤,具体获取方法包括:接收短消息中心实时转发的通信终端发送的短信;或,从通信终端的原始话单文件中获取短信;或,监听并获取通信终端发送给短消息中心的短信,其特征在于,

该方法还包括:所述信息推送装置建立以通信终端标识为网络结点,以通信终端之间的短信收发为有向弧线,以交互的短信数量为弧线权值的有向网络;

所述信息推送装置根据所述有向网络,计算出每一个结点对应的通信终端对其余结点所对应的通信终端的主导域值;

所述信息推送装置按照所述主导域值排列所述通信终端标识,生成主导用户列表;

当多个相同的通信终端标识映射到第二类别集中的同一个类别时,统计出该相同通信终端标识发送的短信数量;对所述第二类别集中的每一个类别分别生成用户兴趣频度表;

根据所述第一信息映射的类别,确定出与第一信息关联的用户兴趣频度表;

删除包含在所述主导用户列表中、但不属于与第一信息关联的用户兴趣频度表中的通信终端标识;

按照所述主导域值重新排列所述主导用户列表中的通信终端标识,生成新的主导用户列表;

所述信息推送装置将所述第一信息推送到确定出的对应信息源的步骤包括:

所述信息推送装置根据所述第一信息的推送规模,在所述新的主导用户列表中按主导域值从大到小的顺序选择出若干个通信终端,将所述第一信息推送给选择出的通信终端。

2. 如权利要求1所述的方法,其特征在于,所述将第二信息按照第二类别集进行分类,具体包括:

周期获取通信终端发送的短信并存入短信数据库中,通过计算各短信文本的相似度,将多条相似的短信文本整合成一条短信文本;

对整合后的短信文本采用单分类算法进行分类,将每一条整合后的短信文本分入所述第二类别集中唯一类别。

3. 如权利要求2所述的方法,其特征在于,所述单分类算法包括但不限于支持向量机SVM算法。

4. 如权利要求2所述的方法,其特征在于,所述将多条相似的短信文本整合成一条短信文本,具体包括:

对短信数据库中存储的短信按照发送方为主关键字、接收方为次关键字进行排序；

设置进行文本整合的滑动窗口尺寸 w ，逐条读取所述短信数据库中排序后的短信，与滑动窗口中的 w 条短信文本进行相似度计算，对相似度大于或等于相似度阈值的短信文本进行整合；若当前短信与滑动窗口中的 w 条短信文本的相似度均小于所述相似度阈值，则作为滑动窗口中的新短信文本，且滑动窗向后滑动一格。

5. 如权利要求 4 所述的方法，其特征在于，所述相似度计算方法包括但不限于基于特征词向量的夹角余弦方法；所述短信文本整合包括：对同一通信终端发送的相似度大于或等于相似度阈值的短信文本直接按照特征词的频度对应相加并规范化。

6. 如权利要求 2 所述的方法，其特征在于，将每一条整合后的短信文本分入所述第二类别集中唯一类别后，建立起发送该短信的终端移动标识与第二类别集中该类别的第二映射关系。

7. 如权利要求 1 所述的方法，其特征在于，所述计算出每一个结点对应的通信终端对其余结点所对应的通信终端的主导域值，具体包括：

计算通信终端 i 对通信终端 j 的主导程度：

$$p_{i,j} = \lambda_1 a_{i,j} + \lambda_2 a_{j,i};$$

其中 $a_{i,j}$ 表示通信终端 i 发送给通信终端 j 的短信条数， $a_{j,i}$ 表示通信终端 i 从通信终端 j 接收到的短信条数；

$\lambda_i (i = 1, 2)$ 为常数，且 $\lambda_1 + \lambda_2 = 1$ ，表示发送和接收的不同权重；

计算通信终端 i 对有向网络中全部通信终端的主导程度之和，即

$$p_i = \sum p_{i,k}$$

定义通信终端 i 的主导范围 r 为：

$$r_i = \eta_1 d_{i,out} + \eta_2 d_{i,in}$$

上式中， r_i 表示通信终端 i 的主导范围， $d_{i,out}$ 表示通信终端 i 所发出短信的总条数， $d_{i,in}$ 表示通信终端 i 所接收短信的总条数， $\eta_i (i = 1, 2)$ 为常数且 $\eta_1 + \eta_2 = 1$ ，表示短信发送和接收的不同权重；

通信终端 i 的主导域 L_i 的计算方法为：

$$L_i = \gamma_1 \frac{p_i}{\text{avg}(p)} + \gamma_2 \frac{r_i}{\text{avg}(r)}$$

上式中， $\text{avg}(p)$ 为全部通信终端的平均主导程度， $\text{avg}(r)$ 为全部通信终端的平均主导范围， $\gamma_i (i = 1, 2)$ 为常数且 $\gamma_1 + \gamma_2 = 1$ 。

8. 一种信息推送装置，包括：

第一信息处理模块，用于将第一信息按照第一类别集进行分类，建立所述第一信息与第一类别集中所属类别的第一映射关系；

第二信息处理模块，用于获取信息源发送的第二信息，将所述第二信息按照第二类别集进行分类，根据分类结果，建立发送所述第二信息的信息源与第二类别集中所属类别的第二映射关系；

信息匹配模块，用于根据第一类别集和第二类别集中所属类别之间的对应关系，由与所述第一信息建立第一映射关系的第一类别集中各类别，匹配出第二类别集中的对应类别，并根据第二映射关系确定出对应信息源；

信息推送模块,用于将所述第一信息推送到确定出的对应信息源;
其特征在于,

所述第二信息处理模块还根据本地短信数据库中存储的短信,建立以通信终端标识为网络结点,以通信终端之间的短信收发为有向弧线,以交互的短信数量为弧线权值的有向网络;根据所述有向网络,计算出每一个结点对应的通信终端对其余结点所对应的通信终端的主导域值;按照所述主导域值排列所述通信终端标识,生成主导用户列表;

当多个相同的通信终端标识映射到第二类别集中的同一个类别时,统计出该相同通信终端标识发送的短信数量;对所述第二类别集中的每一个类别分别生成用户兴趣频度表;

根据所述第一信息映射的类别,确定出与第一信息关联的用户兴趣频度表;

删除包含在所述主导用户列表中、但不属于与第一信息关联的用户兴趣频度表中的通信终端标识;

按照所述主导域值重新排列所述主导用户列表中的通信终端标识,生成新的主导用户列表;

所述信息匹配模块根据所述第一信息的推送规模,还在所述新的主导用户列表中按主导域值从大到小的顺序选择出若干个通信终端;

所述信息推送模块还将所述第一信息推送给从所述新的主导用户列表中选择出的通信终端。

9. 如权利要求8所述的信息推送装置,其特征在于,所述第二信息处理模块周期获取通信终端发送的短信并存入本地短信数据库中,通过计算各短信文本的相似度,将同一通信终端发送的多条相似的短信文本整合成一条短信文本;

对整合后的短信文本采用单分类算法进行分类,将每一条整合后的短信文本分入所述第二类别集中唯一类别;并建立起发送该短信的终端移动标识与第二类别集中该类别的第二映射关系。

10. 如权利要求9所述的信息推送装置,其特征在于,所述信息匹配模块还根据与第一信息关联的用户兴趣频度表中各通信终端标识对应的短信条数,以及第一信息与其映射的类别的相似度,确定出各通信终端对所述第一信息的兴趣度,生成针对所述第一信息的用户兴趣度列表;并根据第一信息的推送规模,在确定出的用户兴趣度列表中按照兴趣度从大到小的顺序选择出若干个通信终端;

所述信息推送模块结合从主导用户列表中选择出的通信终端和按兴趣频度从大到小的顺序选择出若干个通信终端,将所述第一信息推送给选择出的通信终端。

一种信息推送方法与装置

技术领域

[0001] 本发明涉及通信领域,尤其涉及一种向通信终端推送信息的方法与装置。

背景技术

[0002] 随着通信技术的发展,新业务不断出现。例如:现有技术中,通信终端(移动终端)的短信业务得到了迅猛发展。面对庞大的移动用户群体,一种信息推送业务的新形式即基于短信的广告已经在实际中使用。

[0003] 然而,现有技术中的短信广告多数是一种不分受众的短信群发。即:不对用户根据其兴趣爱好等加以区分,将短信广告群发给所有的用户。这种群发短信广告的方式具有如下明显的缺点:(1) 不能满足用户的特定需求,很难达到预期的广告效果;(2) 造成了很多垃圾短信,不仅浪费了公共通信资源,更引发了用户的普遍反感;(3) 真正有需求的用户可能收不到相应的短信广告。

[0004] 因此,要真正获得短信广告的最好效果,必须深入挖掘用户兴趣和爱好,得到用户即时的和潜在的需求,为用户提供个性化的短信广告服务。

发明内容

[0005] 本发明实施例提供一种信息推送方法与装置,通过分析用户发送的信息和待推送信息之间的关联,选择出有对应需求的用户进行信息推送。

[0006] 一种信息推送方法,包括:

[0007] 信息推送装置将第一信息按照第一类别集进行分类,建立所述第一信息与第一类别集中所属类别的第一映射关系;以及将信息源发送的第二信息按照第二类别集进行分类,建立发送所述第二信息的信息源与第二类别集中所属类别的第二映射关系;

[0008] 所述信息推送装置根据第一类别集和第二类别集中所属类别之间的对应关系,由与所述第一信息建立第一映射关系的第一类别集中各类别,匹配出第二类别集中的对应类别,并根据第二映射关系确定出对应信息源;

[0009] 所述信息推送装置将所述第一信息推送到确定出的对应信息源;

[0010] 其中,该信息源为通信终端,该第二信息为通信终端发送的多条短信;在将信息源发送的第二信息按照第二类别集进行分类之前还包括获取信息源发送的第二信息的步骤,具体获取方法包括:接收短消息中心实时转发的通信终端发送的短信;或,从通信终端的原始话单文件中获取短信;或,监听并获取通信终端发送给短消息中心的短信,该方法还包括:

[0011] 该信息推送装置建立以通信终端标识为网络结点,以通信终端之间的短信收发为有向弧线,以交互的短信数量为弧线权值的有向网络;

[0012] 该信息推送装置根据该有向网络,计算出每一个结点对应的通信终端对其余结点所对应的通信终端的主导域值;

[0013] 该信息推送装置按照该主导域值排列该通信终端标识,生成主导用户列表;当多

个相同的通信终端标识映射到第二类别集中的同一个类别时,统计出该相同通信终端标识发送的短信数量;对所述第二类别集中的每一个类别分别生成用户兴趣频度表;根据所述第一信息映射的类别,确定出与第一信息关联的用户兴趣频度表;删除包含在所述主导用户列表中、但不属于与第一信息关联的用户兴趣频度表中的通信终端标识;按照所述主导域值重新排列所述主导用户列表中的通信终端标识,生成新的主导用户列表;

[0014] 该信息推送装置将该第一信息推送到确定出的对应信息源的步骤包括:

[0015] 该信息推送装置根据该第一信息的推送规模,在该主导用户列表中按主导域值从大到小的顺序选择出若干个通信终端,将该第一信息推送给选择出的通信终端。

[0016] 本发明的另一方面还提供了一种可以执行上述方法的一种信息推送装置。采用本发明实施例,能根据用户发送的信息,分析其需求,再与待推送信息进行关联匹配,确定出特定的用户群体进行信息推送。满足了用户的特定需求,克服了信息推送的盲目性,避免了公共通信资源的浪费。

附图说明

[0017] 图 1A、图 1B 为本发明实施例提供的根据用户移动终端发送的短信推送广告给特定移动终端的实现流程图;

[0018] 图 2 为本发明实施例提供的短信预处理与整合流程图;

[0019] 图 3 为本发明实施例提供的短信分类示意图;

[0020] 图 4 为本发明实施例提供的用户兴趣频度表;

[0021] 图 5 为本发明实施例提供的根据短信数据库得到用户社区网络示意图;

[0022] 图 6 为本发明实施例提供的广告录入并对该广告进行分类的流程图;

[0023] 图 7 为本发明实施例提供的一种信息推送装置结构示意图。

具体实施方式

[0024] 本发明实施例提供一种信息推送方法,包括:

[0025] 将第一信息按照第一类别集进行分类,建立第一信息与第一类别集中所属类别的第一映射关系;

[0026] 获取信息源发送的第二信息,将第二信息按照第二类别集进行分类,根据分类结果,建立发送第二信息的信息源与第二类别集中所属类别的第二映射关系;

[0027] 根据第一类别集和第二类别集中所属类别之间的对应关系,由与第一信息建立第一映射关系的第一类别集中各类别,匹配出第二类别集中的对应类别,并根据第二映射关系确定出对应信息源;

[0028] 将第一信息推送到确定出的对应信息源。

[0029] 下面以向移动终端推送短信广告为例对上述方法加以详细说明。即:假定第一信息为广告信息(包括但不限于商品广告、业务广播或服务广告等);信息源为能发送信息的通信装置,如移动终端;第二信息为移动终端通过短消息中心发送的短信。在这种假定的使用场景下,需要对移动终端发送的短信进行短信分类,并对不同广告内容进行广告分类,通过这两种分类之间的对应关系,来确定出不同广告内容可以推送的移动终端。

[0030] 为简单起见,在本发明实施例中,假设短信与广告具有相同的分类,即每一个短信

类别与唯一的一个广告类别相对应。

[0031] 参见图 1A、图 1B,为本发明实施例提供的根据用户移动终端发送的短信,推送广告给特定移动终端的整体流程图。图 1A 包括如下具体步骤:

[0032] 步骤 S11、采集用户移动终端发送的短信,并存放到数据库中;

[0033] 步骤 S12、对用户的短信数据进行预处理与整合;

[0034] 步骤 S13、对整合后的短信文本分类;

[0035] 步骤 S14、建立发送短信的移动终端标识与各短信类别之间的映射关系,对每一个短信类别建立用户兴趣频度表;

[0036] 步骤 S15、根据短信数据库中存储的短信建立表征移动终端之间进行短信交互的社区网络;

[0037] 步骤 S16、根据建立的社区网络确定出主导用户列表。

[0038] 上述步骤 S12/13/14 与步骤 S15/16 没有逻辑上的顺序关系,可以并行处理。

[0039] 图 1B 包括如下具体步骤:

[0040] 步骤 S21、进行广告录入,并对广告进行分类;

[0041] 步骤 S22、根据建立的用户兴趣频度表,结合当前广告所属广告类别确定出针对该广告的用户兴趣度列表,即确定出潜在广告对象;

[0042] 步骤 S23、根据确定出的用户兴趣度列表(潜在广告对象)并结合主导用户列表,确定出最终的广告对象,即确定出最终推送该广告的移动终端;

[0043] 步骤 S24、根据移动终端的不同类型,生成不同表现形式的广告并发送给对应移动终端。

[0044] 下面对上述每一个步骤进行详细描述。

[0045] 对于步骤 S11,描述如下:

[0046] 首先建立一个空的短信数据库,该数据库可以由 Oracle 等现有的数据库管理系统创建。该短信数据库的表结构至少包括:短信发送方移动终端标识(ID)、接收短信方移动终端标识、短信发送时间和短信内容等信息,如下表 1 所示:

[0047] 表 1

[0048]

发送方 ID	接收方 ID	发送日期、时间	短信内容
--------	--------	---------	------

[0049] 由于短信的发送方和接收方既可以是普通用户,也可以是连接在短信中心的实体(Short Message Entity,SME),但 SME 的短信并不代表用户个人的兴趣,因此本发明实施例中关心普通用户的点对点短信,所以本发明实施例中采集普通用户发送的短信。用户发送的短信可能具有多种不同的形式,如普通的文本信息和包含声音、图像及视频的彩信等,本发明实施例以采集普通用户的文本短信为例。

[0050] 短信采集具体实现方法可以是多种多样的,例如:

[0051] 采集方式一:接收短消息中心实时转发的通信终端发送的短信;

[0052] 采集方式二:从通信终端的原始话单文件中获取短信;即将计费服务器上的原始话单文件作为数据源,逐条从原始话单文件中读入每条短信;

[0053] 采集方式三:监听并获取移动终端发送给短消息中心的短信。

[0054] 以上短信的获取方式仅为举例,本发明对此不作限定。

[0055] 根据需要,可以设置短信采集时间周期,例如周期可以为一天,一周或者一个月。该时间周期结束时,将采集的短信数据供后续步骤分析和处理。

[0056] 对于步骤 S12,描述如下:

[0057] 由于短信字数的限制(一般 < 70 个汉字),以及输入不便等各种原因,通过步骤 S11 得到的短信数目极大,而且内容主题分散,这将可能造成后续的文本分类过程时间复杂度的急剧攀升,同时严重影响用户需求的准确性。为此,首先剔除群发短信号码。具体方法为:根据数据采集时间设定一阈值 k ,如果某个移动号码发送的短信条数超过该阈值,就判定该号码为短信群发号码,需要将该号码发送的所有短信数据从短信数据库中删除。对于号码发送短信条数的判断可利用数据库管理系统的统计功能实现。阈值的选取一般应取明显异常的值,例如,如果短信采集时间为一天,则可取阈值 $k = 300$;如果采集时间为一月,则可取阈值 $k = 2000$ 。

[0058] 其次,针对短信字数较少,往往需要连续几条短信才能表达明确的内容,而且与不同的接收对象所交流的主题并不一定相同的问题,利用短信文本的时间相关性和对象相关性,按照短信内容进行聚类。时间相关性和对象相关性可以通过对短信数据库排序得到,其中主关键字为短信发送方的移动终端号码,次关键字为短信接收方的移动终端号码。短信聚类的好处是使得短信文本的数量极大减少,同时使得文本主题相对集中,便于后续短信分类。

[0059] 为降低聚类算法的复杂度,本发明实施例提出一种基于滑动窗口的文本整合方法。具体为:预先确定一个合适的窗口尺寸 w ,新的短信文本仅需和最近的 w 个已整合短信文本进行相似度计算,对相似度高于阈值的最相似短信文本进行整合。通过适当调整 w 值,该算法在保证效果的同时使得时间复杂度可控。

[0060] 图 2 所示为本发明实施例的短信预处理与整合流程图。具体包括:

[0061] 步骤 S30、设置群发度阈值 k 、滑动窗口尺寸 w 和相似度阈值 d ;

[0062] 步骤 S31、以发送方号码为主关键字、接收方号码为次关键字对短信数据库进行排序;

[0063] 步骤 S32、删除数据库中短信发送数量超过阈值 k 的所有记录,即删除群发短信记录;

[0064] 步骤 S33、判断短信数据库中是否还有未处理短信,若有,继续下列步骤;否则,结束处理流程;

[0065] 步骤 S34、读入下一条短信;

[0066] 步骤 S35、提取短信的特征向量;

[0067] 步骤 S36、计算与前 w 条短信的相似度;

[0068] 步骤 S37、判断相似度是否大于相似度阈值 d ;若是,执行步骤 S38;否则,执行步骤 S39;

[0069] 步骤 S38、将该短信与最大相似度的文本整合,转至步骤 S33;

[0070] 步骤 S39、将该短信作为滑动窗口中的新文本,同时滑动窗口向后滑动一格;转至步骤 S33。

[0071] 上述流程需要事先指定群发度阈值、相似度阈值和滑动窗口尺寸,这些参数可根

据需要进行适当调整。

[0072] 文本相似度的计算方法如下：

[0073] 对两个文本 S1 和 S2, 令它们的所有特征词构成的向量空间为 $V = \{X_1, X_2, X_3, \dots, X_n\}$, 其中 X_i 为特征词。设文本 S1 的特征向量 $V_1 = (\omega_1, \omega_2, \dots, \omega_n)$, 其中 ω_i 为特征词 X_i 在文本 S1 中的频度；文本 S2 的特征向量 $V_2 = (\varphi_1, \varphi_2, \dots, \varphi_n)$, 其中 φ_i 为特征词 X_i 在文本 S2 中的频度。则两个文本的相似度按照下式计算：

[0074]

$$Sim(S_1, S_2) = \overline{V_1} \cdot \overline{V_2} = \frac{\sum_{i=1}^n \omega_i \cdot \varphi_i}{\sqrt{\sum_{i=1}^n \omega_i^2} * \sqrt{\sum_{i=1}^n \varphi_i^2}}$$

[0075] 文本的合并方法为直接按照特征词的频度对应相加并规范化。设文本 S1 和文本 S2 的特征向量表示如上, 将其各特征项所对应的特征向量相加, 然后对其规范化。新文本的发送时间为新近被合并文本的发送时间。

[0076] 合并后的文本向量规范化方法优选采用最小最大规范化方法。设直接按照特征词的频度对应相加后所得到的新文本特征向量为 $V = \{v_1, v_2, \dots, v_n\}$, 其中 v_i 为特征词 X_i 在新文本中的频度。设规范化以后所得到的特征向量为 $V = \{\phi_1, \phi_2, \dots, \phi_n\}$, 其中 ϕ_i 为特征词 X_i 在规范化后的新文本中的频度, 其计算方法如下：

$$[0077] \quad v = \frac{\phi_i - \text{Min}\phi_i}{\text{Max}\phi_i - \text{Min}\phi_i}$$

[0078] 在文本特征向量合并的同时, 记录该新文档对应的原始短信数目。具体实现时只需在每次合并时将新文本包含的短信数量加 1 即可。

[0079] 经过短信预处理及整合后的短信文本格式如下表 2 所示：

[0080] 表 2

[0081]

发送方 ID	发送日期、时间	短信文本向量	原始短信条数
--------	---------	--------	--------

[0082] 整合后的短信文本可以存入数据库中, 也可以存为文件或者其它形式。

[0083] 由于利用了短信特有的时间相关性、对象相关性和内容相关性, 经过整合的短信文本主题相对集中, 同时极大的减少了短信数量, 使得整合后的短信文本更易于后续分类。

[0084] 对于步骤 S13, 描述如下：

[0085] 短信文本分类用于将移动终端发送的短信文本分类到预先定义的短信类别中。中文文本分类技术主要包括多分类器集成学习的方法、支持向量机 (SVM)、KNN 方法、朴素贝叶斯方法、决策树、神经网络、最大熵模型等, 它们都可用于本发明实施例的分类过程。由于 SVM 的分隔面模式有效地克服了样本分布、冗余特征以及过拟合等因素的影响, 具有很好的泛化能力, 在效果和稳定性上相对其它方法具有优势, 因此本发明实施例优选 SVM 方法作为分类算法。然而本发明显然不局限于使用 SVM 算法。

[0086] 本发明实施例利用 LIBSVM 软件包来实现 SVM 分类操作。

[0087] 首先, 从整合后的短信数据库中选择出若干条短信文本作为训练文本集, 并对这

些训练文本进行人工分类。训练文本的选择需要使得各个类别的文本数量差别不大。具体实现时,可以预先指定每个类别的文本数目,例如 100,然后逐条从短信数据库中读入短信文本,对其进行人工分类。如果该类别的文本数量不足,则将该文本进行类别标记并放入训练集中;如果该类别文本数量已达到指定数目,则简单的丢弃该文本,重新从短信数据库中读入下一条文本。

[0088] 得到训练文本后,需要提取该训练集文本的特征,并使用向量空间模型 (Vector Space Model, VSM) 将训练文本集中的文本表示为其对应的特征向量。特征向量的提取有多种方法,例如采用 $tf \times idf$ 方法,该方法的具体实现可以参考文献 Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys, 2002, 34(1) : 1-47。

[0089] 经过上述处理后,训练数据集可表示如下:

[0090] $T = \{T_i | T_i = (W_i, c_i), c_i \in C\}$

[0091] 其中, W_i 为训练文本集中第 i 个训练文本的特征向量, C 为该特征向量的人工分类类别集 (即第二类集)。第 i 个文本的特征向量 W_i 表示如下:

[0092] $W_i = (w_{i1}, w_{i2}, \dots, w_{in})$

[0093] 其中 w_{ik} ($k = 1, 2, \dots, d$) 为特征项 k 对文本 i 的贡献程度, n 为特征向量的维数。人工分类类别集 C 表示如下:

[0094] $C = \{c_1, c_2, \dots, c_m\}$

[0095] 其中, m 为类别数。

[0096] 接下来利用 LIBSVM 工具进行文本模型的训练,其步骤如下:

[0097] 1) 设置系统参数。该参数可通过 LIBSVM 软件包提供的 `svm_parameter` 方法进行参数设定。本实施例中,选用 C_SVC 类型的支持向量机,其核函数 (Kernel Function) 使用径向基函数 (Radical Base Function, RBF):

[0098] $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$

[0099] 设定 RBF 核函数的参数 γ 的初始值为 0.5; `svm_type` 属性,具有 C_SVC, NU_SVC, ONE_CLASS, EPSILON_SVR, NU_SVR 共五个可选值,本实施例选用 C_SVC; `C` 属性,表示分类的类别数,设置为类别集的元素个数,即为 m ; `kernel_type` 属性,具有 LINEAR, POLY, RBF, SIGMOID, PRECOMPUTE 共五个可选值,本实施例选择使用 RBF; `shrinking` 属性,本实例将其值设置为 1。此外,从计算机运算的角度上出发,本实施例设置缓存大小为 40MB,运算的精度为 0.001,这些参数分别对应着 `svm_parameter` 的 `cache_size`, `eps`, `shrinking` 属性。综上所述,本实施例选择的参数为:

[0100] `svm_type = C_SVC;`

[0101] `C = m;`

[0102] `kernel_type = RBF;`

[0103] `cache_size = 40;`

[0104] `eps = 0.001;`

[0105] `shrinking = 1`

[0106] 2) 训练属性设置。

[0107] 设定 SVM 的参数后,将训练数据集做为 SVM 的输入,经过训练后产生 SVM 的分类器

的分类模型。在 LIBSVM 软件包中,使用 `svm_problem` 来描述当前的分类问题。设置 `svm_problem` 的 `l` 属性为训练数据集 `T` 的元素个数,`x` 和 `y` 属性分别设置为训练数据集 `T` 的训练文本特征向量集和对应的训练文本的类别集。

[0108] 在使用 LIBSVM 时,`svm_problem` 的 `x` 属性是一个二维的 `svm_node` 数组。将其第一维大小设置为训练数据集 `T` 的元素个数,第二维设置为训练数据集 `T` 中训练文本特征向量的维数。训练数据集 `T` 中的每一个元素对应着 `x` 中的一行。对于 `svm_problem` 的 `x` 属性中的第 `i` 行 `j` 列元素 `x[i][j]`,设置其 `index` 属性为 `j+1`,同时设置其 `value` 属性为训练数据集中第 `i` 个训练文本的特征向量的第 `j` 维数值。

[0109] `svm_problem` 的 `y` 属性是一个一维数组,其大小为训练数据集 `T` 中的元素个数。对于 `y` 的第 `i` 维,设置其值为训练数据集 `T` 中的第 `i` 个训练文本的类别 `ci`。

[0110] 3) 训练 SVM 分类器模型。

[0111] 在 LIBSVM 软件包中,调用 `svm` 的静态 `svm_train` 方法便可以完成 SVM 分类器的训练工作。该方法使用 `svm_problem` 和 `svm_parameter` 作为参数,这两个参数在前面的步骤中均已经设置完成。`svm_train` 方法的返回值为 `svm_model` 类型的对象,该对象即为 SVM 分类器模型。

[0112] 4) 短信分类。

[0113] 通过上述步骤就完成了 SVM 分类器的构造任务,接下来开始短信文本分类。在对未知文本分类之前,需要将文本 `d` 按照 VSM 模型表示为其特征向量:

[0114] $w_d = (w_{d1}, w_{d2}, \dots, w_{dn})$

[0115] LIBSVM 软件包提供了利用 SVM 分类器模型来预测未知文本类别的功能。对于未知文本 `d` 的特征向量 `wd`,按照训练数据集中的训练数据的录入方式录入,只是不设置其对应 `svm_node` 的 `value` 属性。

[0116] 录入待预测的文本的特征向量后,调用 `svm` 的静态 `svm_predict` 方法,便可以完成预测工作。该方法使用 `svm_model` 和 `svm_node` 数组作为参数。`svm_model` 为步骤 3) 中生成的 SVM 分类器模型,`svm_node` 数组则对应着待预测类别的文本的录入数据。`svm_predict` 方法将返回通过 `svm_model` 预测的文本的类别。

[0117] 上述步骤将短信文本分类后,每一条短信文本就归属于某个特定的类别。在实现时通过事先建立类别文件,如果判断出某条短信的类别,则将发送该短信文本中的移动终端号码记入分类类别中。即建立起发送该短信的移动终端标识和对应类别的映射关系。对数据库中全部短信按照上述分类方法进行分类后,得到如图 3 所示的短信分类示意图。

[0118] 图 3 中,短信类别共分为 `m` 种,在每种类别下包含若干个发送对应类别短信的移动终端标识,如用户 1 移动终端标识、用户 2 移动终端标识、用户 3 移动终端标识、用户 4 移动终端标识等。

[0119] 对于步骤 S14,描述如下:

[0120] 通过上述步骤得到的短信分类结果具有如下特征:

[0121] 1、某个移动终端可能被划分到多个类别中,如图 3 中的用户 1 移动终端被分类到类别 1,类别 2 和类别 `m` 中;

[0122] 2、某个类别中可能包含同一个移动终端标识多次,如类别 1 包含“用户 1 移动终端标识”两次;

[0123] 3、某个类别中包含的移动终端是无序的,如类别 1 下的“用户 1 移动终端标识”和“用户 2 移动终端标识”之间不存在任何顺序;

[0124] 4、分类结果中包含大量数据,对于每一个需要分类的整合后文本,在结果集中均会存在一个结果与之对应;即包含与整合后短信文本对应移动终端标识,以及一条整合后短信文本所包含的短信数量;如类别 1 中,包括两条用户 1 移动终端标识,其对应的短信数量分别为 8 和 12。

[0125] 由于分类结果包含了大量的数据,且每个类别中的移动终端标识均是杂乱无章地排列的。这样的数据不能直接表达不同对应移动终端用户对于某个类别的兴趣,从而对广告短信投放的正确性产生一定的影响。

[0126] 为了解决上述问题,需要对各类别中包含的移动终端标识在该类别中出现的情况进行统计,并计算该用户发送此类短信的条数,根据短信数量对移动终端用户降序排列。

[0127] 对于上述 SVM 分类器的分类结果,产生如图 4 所示的用户兴趣频度表。

[0128] 在图 4 所示的用户兴趣频度表中,同一个类别中不会出现相同的移动终端标识。在某个类别下,频度越大的移动终端,对该类感兴趣的程度就越大。对于实际应用来说,一个类别下包含同一个用户的次数通常是超过 1 的,因此,该结果较 SVM 分类后的数据结果集需要更少的存储空间。

[0129] 还可以通过对不同时间出现的分类结果赋以不同的权值来计算用户的感兴趣程度。由于短信文本是按照时间先后顺序排列,时间较早的短信较早出现在分类结果中,通过对早期的分类结果赋以较低的权值,而对后期的结果赋以较高的权值,这样计算出来的兴趣度能更好的反映用户最近的兴趣和需求。当短信时间较长时,优选加权的兴趣计算方法。

[0130] 对于步骤 S15,对根据短信数据库建立起社区网络的方法描述如下:

[0131] 本发明实施例从用户移动终端短信的收发行为来挖掘社区网络。由于在短信交流的虚拟世界里,交流频繁的用户一般关系较为密切,交流很少的用户则关系较为疏远,因此用户之间有无短信往来、短信往来的频繁程度决定了其在社区中的影响程度和影响范围。

[0132] 本发明实施例使用有向网络 $G = (V, \{E\}, W)$ 来表示用户社区,网络的结点 $v \in V$ 表示用户的移动终端,网络的边(即结点之间的有向弧线) $e \in E$ 表示用户之间的短信收发关系,边 e_i 上的权值 $w_i \in W$ 为用户之间短信数量。图 5 所示为从短信数据库中得到的用户社区的一个实例。图 5 中用 ID1、ID2、ID3、ID4、ID5 分别表示不同的移动终端标识(ID)。

[0133] 社区网络建立步骤如下:

[0134] 1) 初始化时网络为空,短信数据记录指针为 $i = 1$;

[0135] 2) 从短信数据库中读入第 i 条短信的发送方移动终端标识(如号码)和接收方移动终端标识(如号码);

[0136] 3) 判断发送方移动终端标识和接收方移动终端标识是否为网络结点标记,如果未标记,则建立结点并标记该结点为相应的移动终端标识,并从发送方到接收方建立一有向弧,其上标记权值 1;否则标记发送方结点到接收方结点权值为原权值加 1。

[0137] 4) 如果短信数据库还有数据,则转至 2),重复上述步骤,否则结束。

[0138] 上述方法得到的社区网络可能非常庞大,最坏情况下所有用户的移动终端都将存在直接或者间接的联系,这样就会得到全部移动终端用户属于一个社区网络的情况。此外,用户可能偶尔输入错误的号码,这样的误发短信并不能表示用户之间存在密切联系,从而

导致得到的网络并不能反映用户之间的联系情况。

[0139] 为防止上述两种情况,提出如下的解决办法:

[0140] 第一个方法:可以在该社区网络中寻找强连通分量。强连通分量定义为网络中所有结点之间互相可达,而可达定义为结点之间存在一条有向的简单路径。

[0141] 第二个方法:是仅考虑联系比较密切的用户之间的关系,而不考虑联系较稀少的用户之间的关系。在具体实现时,只需将权值小于某个阈值的边从网络中删除。该阈值可根据系统实际情况选取,通常取值为 2 ~ 5 即可。

[0142] 经过上述处理,该有向网络将包含若干个连通分量。而从有向网络得到连通分量可通过多种方法,例如深度优先遍历算法求得。

[0143] 在具体实现时,该网络可采用邻接矩阵或者邻接表等作为存储结构,本实施例优选邻接表作为存储结构。在该存储结构中,表头结点存放在一个向量,该头结点至少包括存放用户移动终端号码和指向第一条邻边的指针的域;表结点表示一条边,至少包括下一个邻接点的指针和该边的权值两个数据域。

[0144] 对于步骤 S16,根据社区网络确定出主导用户,具体包括:

[0145] 在确定主导用户时,本实施例通过定义用户的主导域来保证短信的覆盖面足够广,同时主导用户的数量控制在合适的数量范围内。其中主导域的计算需要依据用户的主导程度和主导范围而定。

[0146] 用户 i 对 j 的主导程度 p 定义为用户 i 的移动终端 i 和用户 j 的移动终端 j 之间的短信交互频度,按下式计算:

$$[0147] \quad p_{i,j} = \lambda_1 a_{i,j} + \lambda_2 a_{j,i}$$

[0148] 其中 $a_{i,j}$ 表示弧 $\langle v_i, v_j \rangle$ 上的权值,即移动终端 i 发送给移动终端 j 的短信条数, $a_{j,i}$ 表示弧 $\langle v_j, v_i \rangle$ 上的权值,即移动终端 i 从移动终端 j 接收到的短信条数; λ_i ($i = 1, 2$) 为常数,且 $\lambda_1 + \lambda_2 = 1$,表示发送和接收的不同权重。由于发送方具有更大的主动性,更能体现其影响力,同时其接收的条数也能反映其影响效果,所以在此暂取 $\lambda_1 = 0.8$, $\lambda_2 = 0.2$,在充分实践后可再更改取值。

[0149] 用户的主导程度定义为该用户对各个用户主导程度之和,即

$$[0150] \quad p_i = \sum p_{i,k}$$

[0151] 用户的主导范围 r 定义为:

$$[0152] \quad r_i = \eta_1 d_{i,out} + \eta_2 d_{i,in}$$

[0153] 上式中, r_i 表示通信终端 i 的主导范围, $d_{i,out}$ 表示通信终端 i 所发出短信的总条数, $d_{i,in}$ 表示通信终端 i 所接收短信的总条数, η_i ($i = 1, 2$) 为常数且 $\eta_1 + \eta_2 = 1$,表示短信发送和接收的不同权重;同样暂取 $\eta_1 = 0.8$, $\eta_2 = 0.2$ 。

[0154] 移动终端 i 的主导域 L_i 的计算方法为:

$$[0155] \quad L_i = \gamma_1 \frac{p_i}{\text{avg}(p)} + \gamma_2 \frac{r_i}{\text{avg}(r)}$$

[0156] 上式中, p_i 为移动终端 i 的主导程度, $\text{avg}(p)$ 为全部移动终端的平均主导程度, r_i 为移动终端 i 的主导范围, $\text{avg}(r)$ 为全部移动终端的平均主导范围。 γ_i ($i = 1, 2$) 为常数且 $\gamma_1 + \gamma_2 = 1$,实际应用中可根据情况调整主导程度和主导范围之间的权重。

[0157] 一旦求出用户移动终端对应的主导域,将 L_i 降序排列即可得到网络中用户主导域

的大小顺序值。例如,对图 5 中所示社区网络,相应的计算结果如下表 3 所示(其中主导程度和主导范围的平衡系数取值为 $\gamma_1 = 0.4$, $\gamma_2 = 0.6$) :

[0158] 表 3

[0159]

用户移动终端标识	ID1	ID2	ID3	ID4	ID5	avg
主导程度 p	2.8	1.2	4.2	0.4	2.4	2.2
主导范围 r	2.0	1.0	2.6	0.4	2.0	1.6
主导域 L	1.259	0.593	1.739	0.223	1.186	

[0160] 由上表 3 可知,五位用户的最终主导域高低顺序为 :ID3, ID1, ID5, ID2, ID4。此顺序列表的一个典型的结果形式如下图 4 所示 :

[0161] 表 4

[0162]

用户移动终端标识	主导域
ID3	1.739
ID1	1.259
ID5	1.186
ID2	0.593
ID4	0.223

[0163] 通过上述步骤 S11-S16,已经根据用户移动终端之间的短信交互,建立起了针对某一短信类别(与广告类别相对应)的用户兴趣频度表;并根据建立的社区网络,确定出主导用户列表。

[0164] 下面具体描述广告分类方法,以及如何根据获得的用户兴趣频度表及主导用户列表来最终确定出广告推送的对象。

[0165] 针对步骤 S21,具体描述如下 :

[0166] 参见图 6,首先,为广告录入并对该广告进行分类的流程图。广告信息录入时可以只录入广告信息,此时该信息需要以文本信息表示,该广告信息需要进一步进行分类;在录入时也可根据需要录入该广告信息的类别信息,此时的类别信息和短信类别信息一致,都是预先定义好的。如果广告被指定了类别,则广告形式可以为文本,也可以为文本以外的其它形式,如可能包含视频或者图像或者音频等信息。

[0167] 广告录入时既可以手工逐条录入,也可以事先将广告信息存放到文件或者数据库文件中批量录入。

[0168] 如果录入广告未指定类别,则需要对其进行分类。广告文本的分类可采取多种方法实现。由于一个广告可能属于多个商品类别,因此不能采取 SVM 等单分类算法实现。本

实施例采用如图 6 所示的分类算法实现单个广告文本到多个类别的分类。具体的分类步骤如下：

[0169] 步骤 S40、读取一条广告信息；

[0170] 步骤 S41、确定是进行自动分类还是人工分类，如果是人工分类，执行步骤 S42；否则，执行步骤 S43；

[0171] 步骤 S42、根据预先定义好的广告类别，对当前广告录入其所属的广告类别中，结束当前广告的分类；

[0172] 步骤 S43、提取广告文本的特征并表示为 $W_{d'} = \{w_{d'1}, w_{d'2}, \dots, w_{d'n}\}$ ；

[0173] 步骤 S44、将训练数据的第 i ($i = 1, 2, \dots, m$) 类向第 j ($j = 1, 2, \dots, n$) 维投影，得到第 i 类第 j 维的重心 $Center_{ij}$ 以及投影范围 $Range_{ij} = (R_{ij}^-, R_{ij}^+)$ ， R_{ij}^- 和 R_{ij}^+ 分别表示训练数据集中第 i 类第 j 维文本距中心的负向和正向半径。具体方法如下：

[0174] 将第 i 类训练数据集 $T = \{T_i | T_i \in T \text{ 且 } T_i \text{ 类别为 } c_i\}$ 分别向第 j ($j = 1, 2, \dots, n$) 维投影，在第 j 维的投影中得到数据：

[0175] $T_{1j}, T_{2j}, \dots, T_{kj}$

[0176] 其中， T_{ij} 表示第 i 类训练数据集 T 中的第 i 个文本特征向量的第 j 维， k 为 T 中元素个数。计算第 j 维的重心 $Center_{ij}$ ：

$$[0177] \quad Center_{ij} = \frac{\sum_i T_{ij}}{|T|}$$

[0178] 同时计算投影范围 $Range_{ij} = (R_{ij}^-, R_{ij}^+)$ ，其中， $R_{ij}^- = \max_s (Center_{ij} - T_{sj})$ ，

[0179] $R_{ij}^+ = \max_s (T_{sj} - Center_{ij})$ 。

[0180] 步骤 S45、计算等效半径 R_{ij}^{Equal} ，具体方法为：

$$[0181] \quad R_{ij}^{Equal} = \alpha_{ij} R_{ij}^- + (1 - \alpha_{ij}) R_{ij}^+$$

[0182] 其中， $\alpha_{ij} = \frac{n_{ij}^-}{n_{ij}^- + n_{ij}^+}$ ， n_{ij}^- ， n_{ij}^+ 分别表示位于 $Center_{ij}$ 左右的文本数；

[0183] 步骤 S46、计算广告和各类别的距离 S_i ：

$$[0184] \quad S_i(w_{d'}) = \sqrt{\sum_{j=1}^k \frac{(w_{d'j} - Center_{ij})^2}{(R_{ij}^{Equal})^2} + \sum_{j=k+1}^m \frac{w_{d'j}^2}{\beta^2}}$$

[0185] 其中， $1/\beta^2$ 为距离系数，分类器函数对该变量并不敏感，本实施例中取 $\beta = 10$ 。计算 $S_i(w_{d'})$ 的值，得到广告特征向量对于类别 i 的距离值。该值越小，表明广告越靠近相应的类别；

[0186] 步骤 S47、最后确定该广告的所属类别；一种简单的实现方法是取距离最小的 k 个类别为该广告类别，例如取 $k = 3$ ；优选的实现方法是根据距离值从小到大排序，然后检查相邻的两个距离值的变化情况，若变化程度突然变大，则认为该广告属于该变化前面的若干个类别。

[0187] 针对步骤 S22，确定用户兴趣度表，具体描述如下：

[0188] 在对广告文本进行分类后，需要确定对该广告感兴趣的用户列表，该列表的用户都是对给定广告感兴趣的移动终端用户，并按照用户的感兴趣程度从高到低排列。

[0189] 对于待投放的广告 A_i , 经过广告分类操作后, A_i 被分类到类别集 $R_i \subseteq C$ 中。对于 R_i 中包含的类别 $c_j \in R_i$, 根据步骤 S14 中确定出的用户兴趣频度表, 并结合广告分类的所属类别, 可以得到对该广告感兴趣的所有移动终端用户。具体方法如下:

[0190] 1)、利用广告分类方法将广告 A_i 分类, 得到其类别集合 $R_i = \{c_{i1}, c_{i2}, \dots, c_{ip}\} \subseteq C$ 以及广告 A_i 与 R_i 中每个元素相对应的相似度集合 $S_i = \{s_{i1}, s_{i2}, \dots, s_{ip}\}$;

[0191] 2)、对于 R_i 的类别, 在短信分类中, 其中与其建立起映射关系的移动终端标识即为对广告 A_i 感兴趣的移动终端。计算 S_i 与对 A_i 感兴趣的用户移动终端标识 U_j 在 R_i 中对应类别中出现的次数构成的向量 t_j 之间的内积 I_{ji} :

$$[0192] \quad I_{ji} = (S_i, t_j) = \sum_{r=1}^p (s_{ir} \times t_{jr})$$

[0193] 其中, $t_j = (t_{j1}, t_{j2}, \dots, t_{jp})$, t_{jk} 为 U_j 在 c_{ik} ($k = 1, 2, \dots, p$) 中出现的次数。 I_{ji} 即为移动终端标识 U_j 的对应用户对广告 A_i 的感兴趣程度;

[0194] 3)、按照 I_{ji} 大小降序排列, 得到对该广告感兴趣的用户列表, 即用户兴趣度列表 (以用户的移动终端标识代表用户)。

[0195] 针对步骤 S23, 确定最终广告对象, 具体描述如下:

[0196] 上述步骤 S22 中得到的用户兴趣度列表, 即为潜在的广告对象列表。为达到更好的广告效果并节省广告成本, 需要对广告对象进行精选:

[0197] 广告对象精选主要基于如下理由:

[0198] 1)、用户兴趣度列表中包含了大量结果, 包括对广告兴趣度很低的用户。如果向这些移动终端投放广告的话, 一方面不能引起用户的兴趣, 会被用户视为垃圾短信, 甚至将广告短信发送列入“黑名单”, 导致将来更多的广告短信都不能正常发送。另一方面, 发送大量的短信会占据很大的网络资源, 严重状况下甚至会导致网络拥塞, 影响正常短信的发送。

[0199] 2)、一般来说, 用户更加信任朋友或者亲人所推荐的商品而不是广告进行的宣传; 因此, 可以通过上述步骤 S16 中确定出的主导用户在该社区网络中转发广告, 这样不仅可以减少短信数量节省广告成本, 更由于社区网络成员的相互信任而达到更好的广告效果。

[0200] 3) 可以进一步根据各移动终端对应兴趣度和其主导域, 得到一个兴趣主导用户列表, 具体为:

[0201] 对于由兴趣度和主导域表示的用户的移动终端为: $U = \{(I_{i1}, L_1), (I_{i2}, L_2), \dots, (I_{in}, L_n)\}$

[0202] 将兴趣度和主导域作内积, 根据得出的兴趣主导程度生成新的兴趣主导用户列表。内积的形式为:

$$[0203] \quad I_{Li} = I_{ii} \times L_i$$

[0204] 其中, I_{Li} 为用户 i 的兴趣主导程度, I_{ii} 为通过上述方法确定出的移动终端 i 对应用户对 i 类广告的兴趣度, L_i 为通过上述方法确定出的用户移动终端 i 的主导域。此内积所决定的用户移动终端标识顺序, 即为给对应用户发送广告将能达到的理论上的广告效果的顺序。兴趣主导用户列表的一个典型的结果形式如下表 5 所示:

[0205] 表 5

[0206]

用户移动终端标识	兴趣主导程度
ID1	1.658
ID2	1.012
.....	
IDU	0.125

[0207] 广告运营商针对待发的广告,给定广告对象规模 N 。根据上述方法得到该广告的分类集 $R_i = \{c_{i1}, c_{i2}, \dots, c_{ip}\} \subseteq C$ 。最终的广告目标用户将从三个方面获取:用户兴趣度列表、主导用户列表以及兴趣主导用户列表。

[0208] 用户兴趣度列表中所包含的移动终端标识的对应用户,对特定类别的商品具有较大的兴趣,且有潜在购买可能。因此,实际应用中可以从用户兴趣度列表中选取兴趣度较大的 $N*40\%$ (可调整,本实施例中选取 40% 作为对商品兴趣度较大的用户数上限)个用户作为最终广告目标用户的第一部分;

[0209] 主导用户是一类具有社区代表性的用户,对短信的发送具有较大的兴趣。因此,实际应用中先从主导用户列表中去掉对当前广告所属类别不感兴趣的用户,再从剩下的主导用户列表中选取主导域较大的 $N*10\%$ (可调整,本实施例中选取 10% 作为主导用户数的上限)个用户作为最终广告目标用户的第二部分;

[0210] 最后,从兴趣主导用户列表中剔除掉已选择的第一和第二部分用户,再从中选取 $N*50\%$ (可调整)个用户作为最终广告目标用户的第三部分。

[0211] 以上三部分的总和即为根据最优原则选取出来的最终广告发送对象。

[0212] 针对最后一个步骤S24,生成广告并发送,具体描述如下:

[0213] 广告发送存在两种形式:

[0214] 其一,面向选择出的所有用户,广告内容和形式完全相同;

[0215] 其二,广告内容和形式存在个性化差异。

[0216] 由于短信群发技术已经非常成熟,本发明实施例可以直接利用现有的短信群发平台实行广告发布,因此可以将这两种形式的短信广告直接传递给现有的短信发送平台直接发布即可。

[0217] 在实际应用中,用户移动终端可能具有不同的特征和功能。例如,不同移动终端的屏幕可能具有不同的尺寸、支持不同数量的颜色;在功能上某些移动终端仅仅支持文本短信,而某些移动终端支持语音短信、图像短信甚至视频短信等。据此,一个可选实现方法是根据移动终端的不同特征功能,为其发布不同形式的短信广告,以尽可能的提升移动终端用户对广告的关注程度。

[0218] 在具体实施中,由于移动终端特征可能存在非常大的差别,对所有这些特征进行广告准备存在极大的开销,这些开销不仅包括准备各种不同短信形式的开销,而且在为不同移动终端选择广告形式时也存在极大的时间开销。因此可以只考虑两种基本的实现形式,也就是将短信广告形式限定为纯文本短信和MMS彩信两种形式。

[0219] 用户移动终端的特征可以通过多种方法得到,事实上移动无线访问协议

(Wireless Access Protocol, WAP) 应用中移动终端识别技术已经非常成熟,可直接利用该技术。

[0220] 以上用具体实施例描述了根据移动终端发送的短信,来确定不同类别广告的推送对象的完整流程。

[0221] 本发明上述实施例提供了根据用户移动终端发送的短信向其推送相应类型广告的方法,一种相应的信息推送装置 10 的具体结构实施例如图 7 所示,包括:

[0222] 第一信息处理模块 101,用于将第一信息按照第一类别集进行分类,建立第一信息与第一类别集中所属类别的第一映射关系;

[0223] 第二信息处理模块 102,用于获取信息源发送的第二信息,将第二信息按照第二类别集进行分类,根据分类结果,建立发送所述第二信息的信息源与第二类别集中所属类别的第二映射关系;

[0224] 信息匹配模块 103,用于根据第一类别集和第二类别集中所属类别之间的对应关系,由与所述第一信息建立第一映射关系的第一类别集中各类别,匹配出第二类别集中的对应类别,并根据第二映射关系确定出对应信息源;

[0225] 信息推送模块 104,用于将第一信息推送到确定出的对应信息源。

[0226] 上述第二信息处理模块 102 执行如下具体处理:

[0227] 周期获取通信终端发送的短信并存入本地短信数据库中,通过计算各短信文本的相似度,将同一通信终端发送的多条相似的短信文本整合成一条短信文本;

[0228] 对整合后的短信文本采用单分类算法进行分类,将每一条整合后的短信文本分入所述第二类别集中唯一类别;并建立起发送该短信的终端移动标识与第二类别集中该类别的第二映射关系;

[0229] 统计出映射到第二类别集中同一个类别的、同一通信终端发送的短信条数,并按照短信条数对通信终端进行排序,生成用户兴趣频度表;

[0230] 第二信息处理模块 102 还根据本地短信数据库中存储的短信,建立以通信终端标识为网络结点,以通信终端之间的短信收发为有向弧线,以交互的短信数量为弧线权值的有向网络;

[0231] 根据有向网络,计算出每一个结点对应的通信终端对其余结点所对应的通信终端的主导域值;

[0232] 按照主导域值排列通信终端标识,生成主导用户列表。

[0233] 上述信息匹配模块 103 获取第一信息处理模块中第一信息映射的类别,确定出与第一信息关联的上述用户兴趣频度表;并根据第一信息的推送规模,在确定出的用户兴趣频度表中按兴趣频度从大到小的顺序选择出若干个通信终端;由信息推送模块 104 将第一信息推送给选择出的通信终端;

[0234] 上述信息匹配模块 103 还根据与第一信息关联的用户兴趣频度表中各通信终端标识对应的短信条数,以及第一信息与其映射的类别的相似度,确定出各通信终端对所述第一信息的兴趣度,生成针对所述第一信息的用户兴趣度列表;并根据第一信息的推送规模,在确定出的用户兴趣度列表中按照兴趣度从大到小的顺序选择出若干个通信终端;由信息推送模块将第一信息推送给从用户兴趣度列表中选择出的通信终端。

[0235] 上述信息匹配模块 103 还根据第一信息的推送规模,在第二信息处理模块 102 生

成的主导用户列表中按主导域值从大到小的顺序选择出若干个通信终端；由信息推送模块 104 将第一信息推送给从主导用户列表中选择出的通信终端。

[0236] 综上所述,采用本发明实施例,能根据用户发送的信息(第二信息,上述实施例中以用户移动终端发送的短信为例),分析用户需求,再与待推送信息(第一信息,上述实施例中以向用户推送广告为例)进行关联匹配,确定出特定的用户群体,将第一信息推送给确定出的用户群体,满足了用户的特定需求,克服了第一信息(如广告)推送的盲目性,避免了公共通信资源的浪费。

[0237] 显然,本领域的技术人员可以对本发明进行各种改动和变型而不脱离本发明的精神和范围。这样,倘若本发明的这些修改和变型属于本发明权利要求及其等同技术的范围之内,则本发明也意图包含这些改动和变型在内。

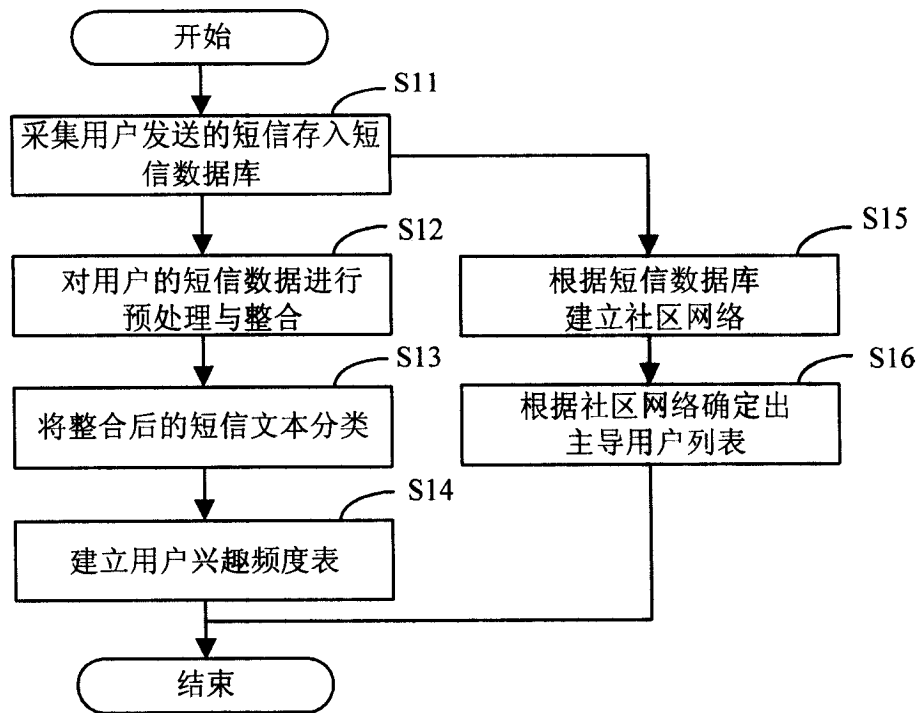


图 1A

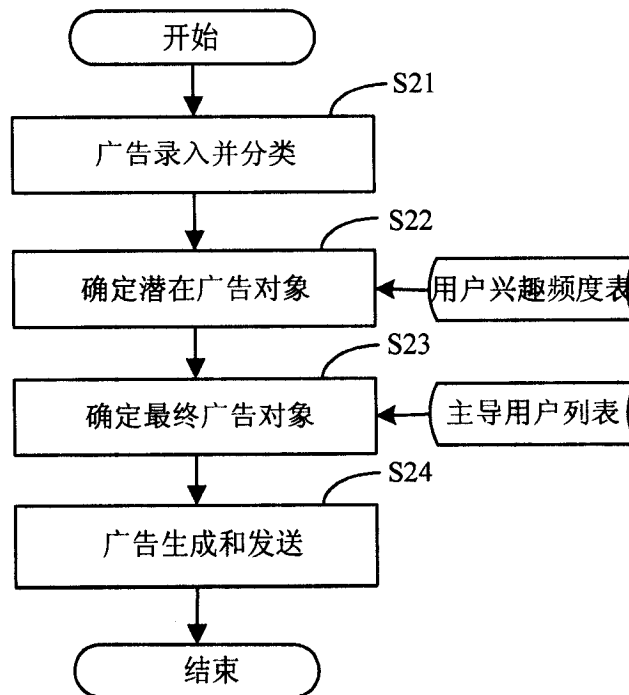


图 1B

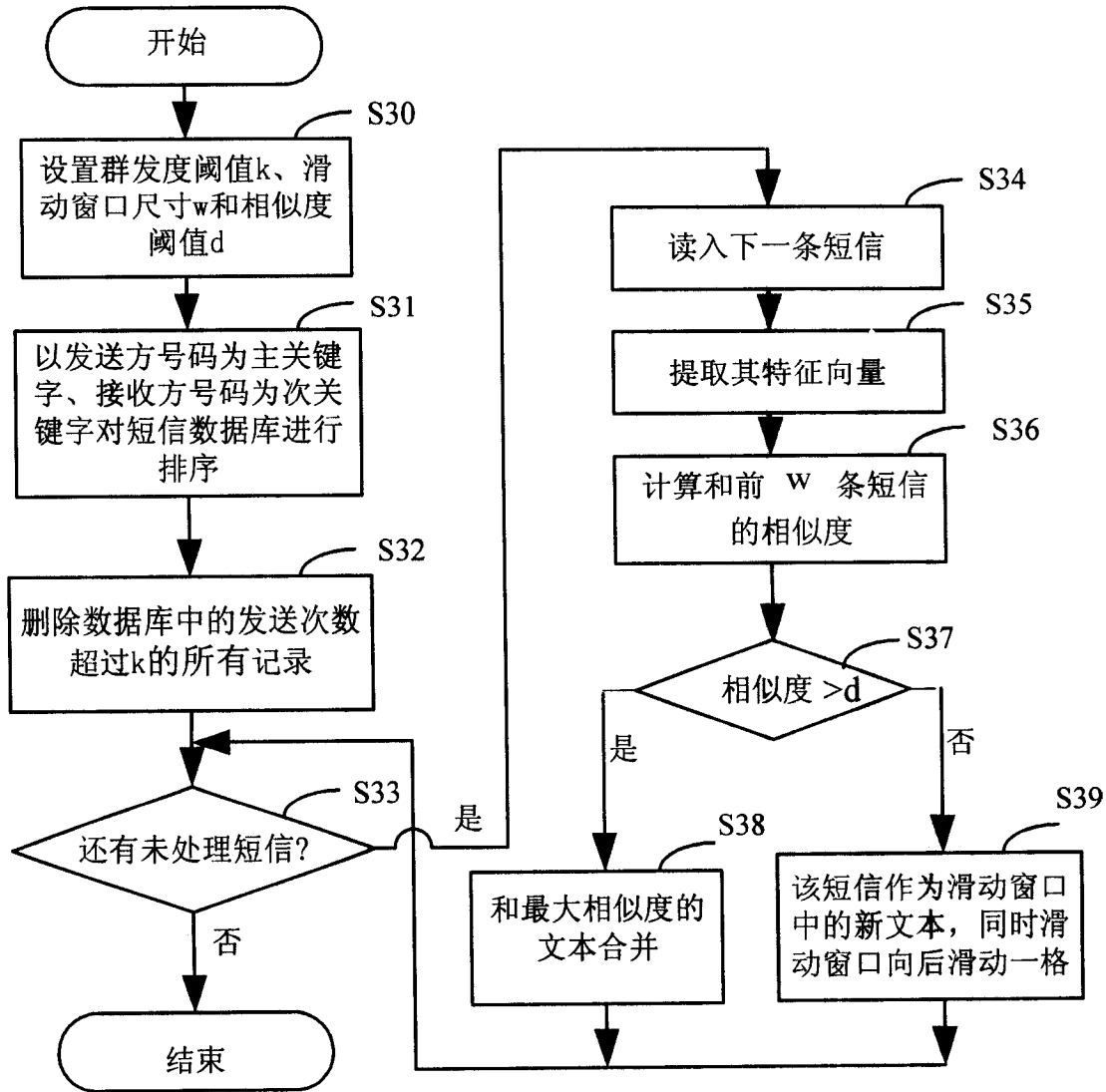


图 2

类别1	短信数	类别2	短信数	类别m	短信数
用户1移动终端标识	8	用户1移动终端标识	4	用户1移动终端标识	6
用户2移动终端标识	2	用户2移动终端标识	2	用户3移动终端标识	1
用户1移动终端标识	12	用户3移动终端标识	1	用户4移动终端标识	1
.....					

图 3

类别1	次数	类别2	次数	类别m	次数
用户1移动终端标识	20	用户1移动终端标识	4	用户1移动终端标识	6
用户2移动终端标识	2	用户2移动终端标识	2	用户3移动终端标识	1
.....		用户3移动终端标识	1	用户4移动终端标识	1
					

图 4

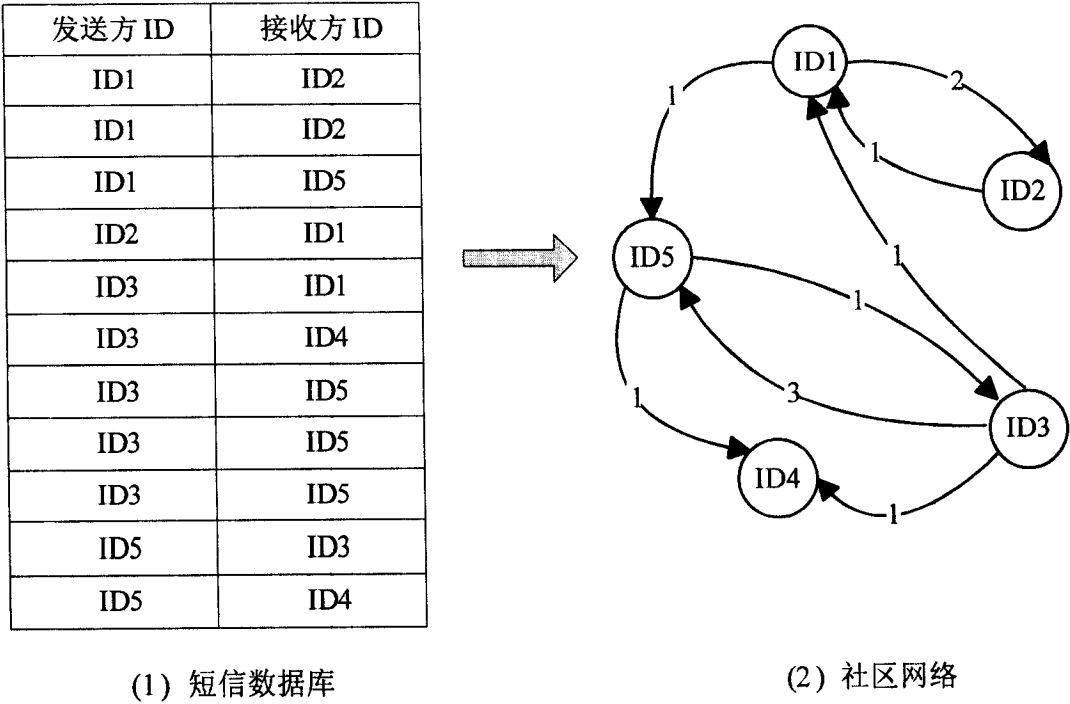


图 5

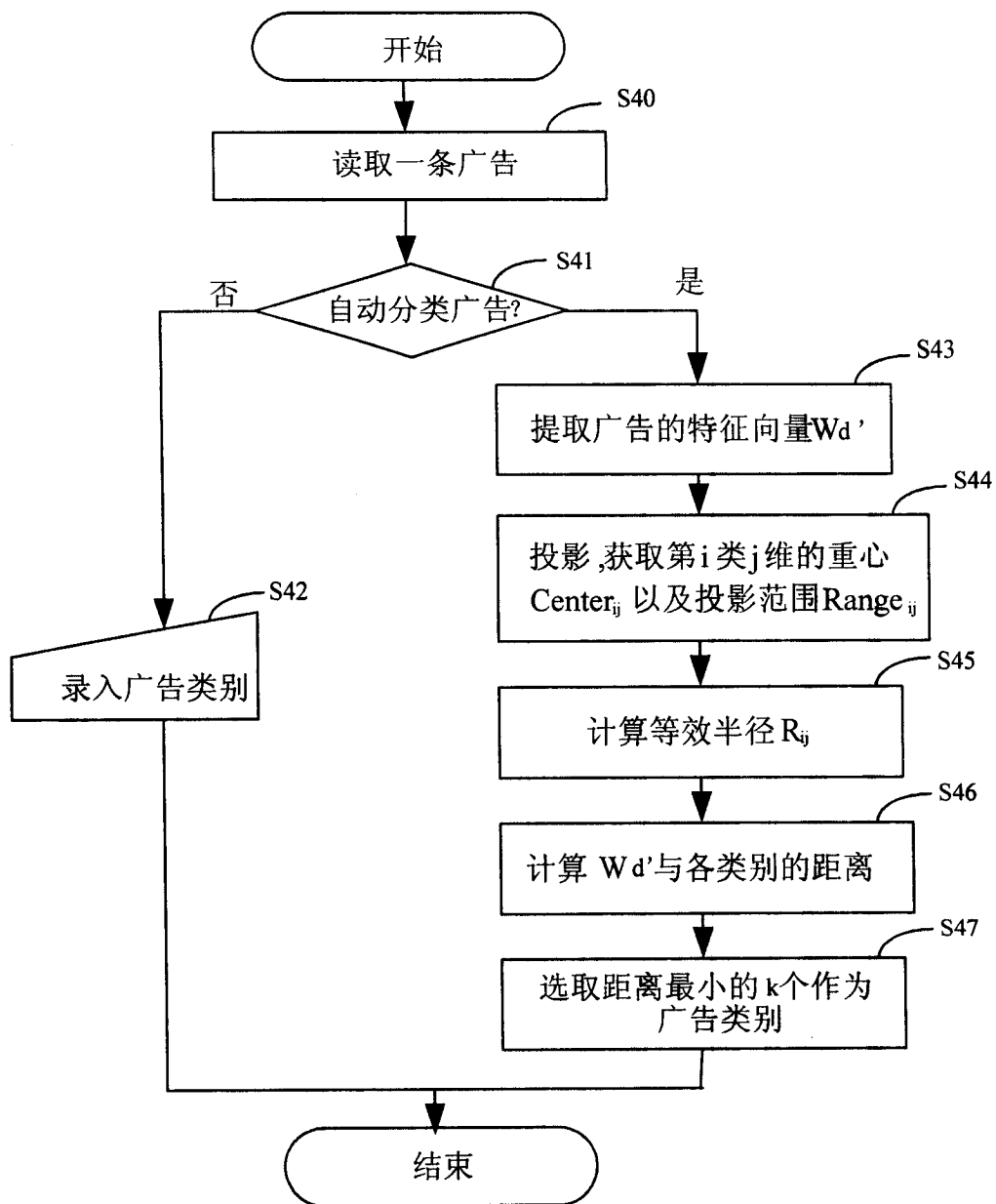


图 6

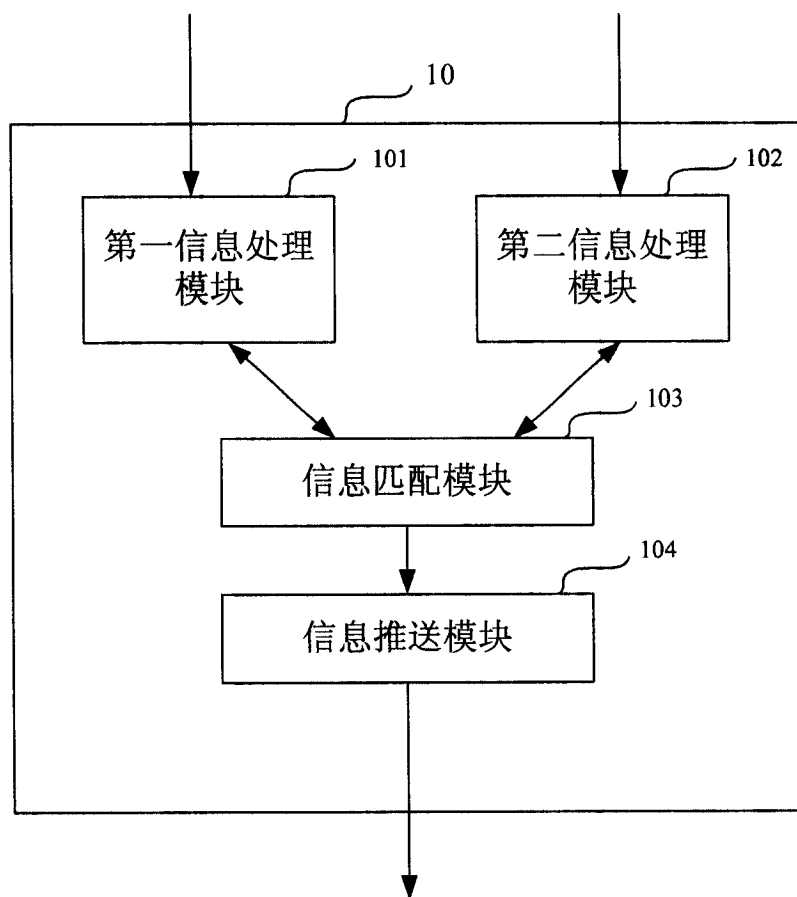


图 7