软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*,2016,27(11):2843−2854 [doi: 10.13328/j.cnki.jos.004870]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

# 社会媒体中用户的隐式消费意图识别*

付 博, 刘 挺

(哈尔滨工业大学 计算机科学与技术学院 社会计算与信息检索研究中心,黑龙江 哈尔滨 150001)

通讯作者: 刘挺, E-mail: tliu@ir.hit.edu.cn

**摘 要**: 不同于已有的显式消费意图识别的研究,提出了社会媒体中用户的隐式消费意图自动识别方法.该方法将隐式消费意图识别视作多标记分类问题,并综合使用了基于用户关注行为、意图关注行为、意图转发行为以及个人信息的多种特征.由于隐式消费意图识别难以评价,自动抽取了大量跨社会媒体的用户链指信息,利用该方法,共抽取出12万余对的用户链指.在此自动评价集上的实验结果表明,所采用的多标记分类方法对于识别用户的隐式消费意图是行之有效的,其中使用的各种特征对于提高隐式消费意图识别的效果皆有帮助.

**关键词**: 隐式消费意图;多标记分类;用户行为分析;用户链指;数据挖掘

**中图法分类号**: TP311

中文引用格式: 付博,刘挺.社会媒体中用户的隐式消费意图识别.软件学报,2016,27(11):2843−2854. http://www.jos.org.cn/1000-9825/4870.htm

英文引用格式: Fu B, Liu T. Implicit user consumption intent recognition in social media. Ruan Jian Xue Bao/Journal of Software, 2016,27(11):2843−2854 (in Chinese). http://www.jos.org.cn/1000-9825/4870.htm

## Implicit User Consumption Intent Recognition in Social Media

FU Bo, LIU Ting

(Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

**Abstract**: Unlike previous works such as explicit consumption intent recognition research, this paper presents a method that uses user behavior analysis to automatically recognize the implicit consumption intent. Specifically, the proposed method recasts implicit consumption intent recognition as a multi-label classification problem, which combines multiple features based on follower's behavior, intent behavior, retweets behavior, and user profiles. The paper proposes a method for the automatic extraction of a large user linkage across social media. With the proposed method, more 120000 user linkage pairs are extracted. Experimental results show that the multi-label classification-based method is effective for implicit intent recognition. Especially, the exploited features are all helpful for improving the recognition performance.

**Key words**: implicit consumption intent; multi-label classification; user behavior analysis; user linkage; data mining

近些年,随着电子商务平台的蓬勃发展,有越来越多的用户参与到网络购物的过程中.由于网络购物的特殊性,用户在消费前积极地在网络中收集信息、选择评价,随后对某类产品做出消费决策.为了更好地对用户的消费意图进行挖掘,并快速找出富有价值的用户,用户的消费意图分析(consumption intent analysis)应运而生.消费意图分析是指用户通过文本内容或行为方式表达出对某一产品或服务产生的购买意愿,是一项针对社会媒体用户消费行为进行识别和挖掘的研究任务,包含多项具有挑战性的研究任务,如显式消费意图识别、显式消费

* 基金项目: 国家自然科学基金(61133012, 61202277); 国家重点基础研究发展计划(973)(2014CB340503)

Foundation item: National Natural Science Foundation of China (61133012, 61202277); National Program on Key Basic Research Project of China (973) (2014CB340503)

收稿时间: 2014-11-20; 修改时间: 2015-03-11; 采用时间: 2015-07-01

意图中的消费对象识别、隐式消费意图(implicit consumption intent)识别等任务,具体情况见表 1.本文着重研究隐式消费意图识别.隐式消费意图是指用户并未明确发布信息表示购买,但通过自身行为表示出对某类产品产生了潜在的购买需求.隐式消费意图识别在众多应用领域中都有重要的意义,例如,在产品推荐研究中,隐式消费意图可用于解决推荐系统中的冷启动问题,为首次通过社会媒体连接到某电子商务网站的用户提供高水平的推荐;在产品推荐引擎研究中,隐式消费意图可用于指导推荐系统发现用户感兴趣的产品领域,以改善现有的产品推荐引擎;在媒体营销中,隐式消费意图可用于电子商务公司针对社会媒体富有价值的用户提供广告宣传.

**Table 1** Explicit consumption intention compared with implicit consumption intention

**表 1** 显式消费意图与隐式消费意图对比

| | 显式消费意图 | 隐式消费意图 |
|---|---|---|
| 问题定义 | 以文本内容形式,明确地指出其需要购买的产品或服务 | 用户并未明确发布信息表示购买,但通过自身行为表示出对某类产品具有潜在的购买需求 |
| 使用数据 | 文本内容 | 用户(人) |
| 实例 | 如:想买一部手机 | 如:用户对汽车用品具有潜在的购买意愿 |
| 判断依据 | 触发词(如:想买)+消费对象(如:手机) | 用户行为(如:关注/购买等) |
| 方法 | 基于模板或基于有指导的分类方法 | 融合多特征的有指导方法 |
| 结果输出 | 二元分类[1−3](即判断一条文本是否具有购买意愿) | 排序问题[4](即判断一个用户对某些类别/产品有购买意愿) |

在以往的工作中,研究者们通常将显式消费意图的识别作为消费意图分析的一项首要任务.显式消费意图是指用户以文本内容形式,明确地指出其需要购买的产品或服务.Goldberg 率先提出 buy wish 的概念[1](即,本文定义的显式消费意图,如:想买一部手机),Chen 等人[2]也提出过相似的概念“Intention Posts”.早期的一部分研究者将这项任务分为两个步骤:首先获取意图模板和词袋等特征,继而基于特征分类器来完成显式消费意图的识别[1].这种方法极大地提高了识别的准确率,但由于模板具有局限性及语料不平衡的限制,召回率不高.近期的一部分研究者侧重于对不平衡语料的处理,基于弱监督的方法或迁移学习的方法来识别显式消费意图句[2,3].此类方法假设在不同的领域下意图表达的方式具有相似性,这种方法可以获取大规模语料或意图词来提高系统识别的性能.此外,有研究[5]利用深度学习方法识别社会媒体上的显式消费意图文本内容,进而将其应用到电影票房预测任务上.然而,实际上只考虑显式消费意图句在意图分析中的应用是远远不够的,大部分的研究工作忽视了以下几个问题:

- 问题 1:出现在显式消费意图中的触发词和消费对象限制用户必须明确提出自己的消费需求.这种针对特定类别来判定用户是否具有消费意图的问题,并不一定能够全方位地覆盖用户的真实需求.同时,显式消费意图识别在应用中有一定的局限性,如受时效性、外界环境和心理因素等影响,会随时发生变化.
- 问题 2:具有显式消费意图的微博内容是以文本为处理对象,无法评估微博内容真实的商业价值.只有当文本信息/行为中的消费意图真正转化到了产品购买时,才是更有价值的消费意图文本/行为.
- 问题 3:显式消费意图往往由单条微博文本决定,而用户的消费意图由个人兴趣及生活需要等多方面原因共同决定.例如,对“手机(数码类别)”感兴趣的用户同时需要买“尿布湿(母婴类别)”.当追踪用户一系列潜在的消费意图才具有更为准确、细致的定位,能够从简单需求的满足延伸到心理预期需要的满足.

然而截至目前,国内外对社会媒体用户隐式消费意图的研究却很少.Zhang 等人[4]率先提出预测用户购买行为的研究(即本文定义的隐式消费意图识别),并借助于社会媒体(Facebook)用户信息与电商网站(eBay)用户购买行为之间的关联性预测用户购买.该方法的基本假设是:仅利用用户的社会媒体提供的信息,可用于预测用户将来购买的产品类别.例如,社会媒体用户喜欢时尚品牌的用户比喜欢汽车配件更可能购买时尚产品.基于这一假设,该方法首先从社会媒体同意分享到电子商务网站的用户信息,包括人口统计数据和个人兴趣,来预测用户的购买行为,然后把这一问题看作排序问题加以解决.实验结果说明,仅利用社会媒体信息可以很好地预测用户的购买行为.但此方法并没有将不同媒体或社区的用户连通而达到用户数据共享,没有考虑个人兴趣中的转发与回复等行为.

鉴于已有方法存在的缺陷,隐式消费意图识别具有重要的研究意义.本文提出了社会媒体中用户的隐式消费意图识别.事实证明:通过社会媒体中的用户行为信息,可以很好地识别电子商务网站上此用户会购买哪些产品类别(例如母婴用品).本文将隐式消费意图识别看作多标记分类(multi-label)问题,即一个社会媒体用户的隐式消费意图可以属于多个类别之中.这里主要的问题包括:

(1) 如何描述社会媒体用户的行为和此用户隐式消费意图类别之间的关系;

(2) 如何获取大量的自动标记的训练数据为多标记分类器所用.

鉴于此,本文将用户关注/转发等行为特征与社会媒体用户的隐性消费意图关联起来,使用了包括基于用户关注行为特征、用户意图转发行为特征、用户意图关注行为以及个人信息这 4 类特征学习一个多标记分类器.进一步地,本文还提出了一种自动采集大量社会媒体与电子商务网站用户链指(user linkage)[6]的方法(即,将一个自然人在不同社会媒体中的用户身份链接起来),对用户的隐式消费意图进行评价.利用本方法,本文共抽出用户链指 12 万余对.实验结果表明,本文提出的基于用户行为的方法对于用户的隐式消费意图识别是有效的.即用户行为能够较好地挖掘用户意图和购买行为之间的关系.

本文的贡献包含以下 3 点:

(1) 隐式消费意图识别作为一项有价值的研究,前人的工作却很少涉及.本文针对该问题提出了一种有效的方法并取得了较为满意的实验结果.与以往的方法不同,本文并不是限制用户必须明确提出自己的需求,因此,识别的意图更为多样化.

(2) 本文自动抽取了大量跨社区的用户链指信息,而用同名的方法仅可以获取到 15%的用户链指信息.利用链指信息可以自动地对隐式意图识别进行评价,解决了隐式消费意图训练语料不足及难以评价的问题.

(3) 本文提出将隐式消费意图识别作为多标记分类问题加以解决.实验结果表明,在自动标注的训练集上学习到的分类器是有效的.

## 1 用户隐式消费行为数据的采集

本文以新浪微博用户为例,研究社会媒体用户中的隐式消费意图识别.在真实的新浪微博用户行为数据中,用户以关注、发布、评论以及转发为主要目的,即使其中涉及用户的消费意图,也难以评判是否真正转化到了产品购买中.为预测用户的真实消费行为,以下基于社会媒体网站(新浪微博 Weibo.com)和电子商务平台(京东商城 JD.com)同意分享信息的用户的真实行为,以获取用户的隐式消费行为数据.

- 获取用户链指数据

用户链指[6]是指将一个自然人在不同社会媒体中的用户身份链接起来,本文将新浪微博与京东商城的用户身份链接起来.

- ✓ 首先,通过普通网页爬虫的方式获取用户分享到新浪微博中的京东商城产品评论信息(如图 1 所示),然后,利用新浪微博 API 可以获得用户的个人信息、关注用户及其历史微博;
- ✓ 其次,通过对所分享内容中涉及的购买产品的短链接、用户的评价时间以及用户产品内容评价进行匹配,在京东商城,对此产品评价页面上挖掘京东用户的个人信息(包括 id、用户名、会员级别)及购买历史;
- ✓ 最后,通过用户所标注的地点信息,判断所链指到的用户是否为同一用户.

通过此方法,本文获得了 2012 年 1 月~2012 年 12 月共 124 470 个新浪微博和京东相对应的用户.我们将京东购买历史中购买次数少于 10 次的产品和用户过滤,同时将新浪微博用户中关注数少于 10 个的用户过滤,以此得到的数据集为:用户数 91 346,产品数 110 771,产品记录 6 210 379.我们以京东上 10 个一级类别作为用户的隐式消费意图类别.

对于每个用户,数据集里存储了以下信息:

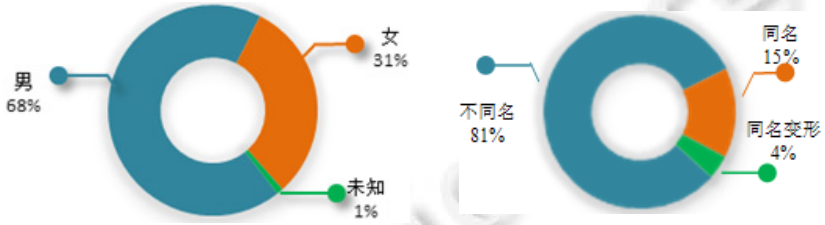
- 用户个人信息:包括用户名、性别以及地理位置.

- 用户关注信息:用户的关注及转发信息以及所关注用户的标签.
- 用户购买历史:用户在京东商城的个人购买历史记录.


Fig.1 Extracting user information based on user reviews

图 1 基于用户评价分享的用户信息抽取

图 2(a)表示所获取用户的性别分布情况,其中,女性分享用户比例略低于男性分享用户,占总用户数量的31%.尽管女性用户是互联网购买的主力,但新浪微博中男性用户多于女性用户,且男性用户更乐于分享自己的购买经历.图 2(b)表示链指用户的用户名对比情况,统计显示,仅有 15%的用户在跨社会媒体的网站上使用相同的用户名.这意味着信息不共享时,仅用同名的方法难以获得大量不同社区的用户.可以发现,由于用户名的冲突,为了达到在不同社会媒体中使用同一个用户名的目标,有 4%的用户会把自己的用户名编辑为统计意义上稀有的字符串,例如 hanjh 和 hanjh2012.




(a) 用户性别分布 (b) 用户名分布情况

Fig.2 Distribution of user's gender and usernames

图 2 用户性别和用户名分布

此外,针对不同性别用户在京东 10 个一级类别中的购买数量进行了统计,如图 3 所示.同样可以发现,女性用户购买较多的商品类别为个护化妆和母婴用品,而男性用户购买较多的商品类别为手机数码类产品[7].

表 2 显示了用户链指实例(其中,"–"代表无此项,圆括号中代表频次),数据集的基本信息见表 3.

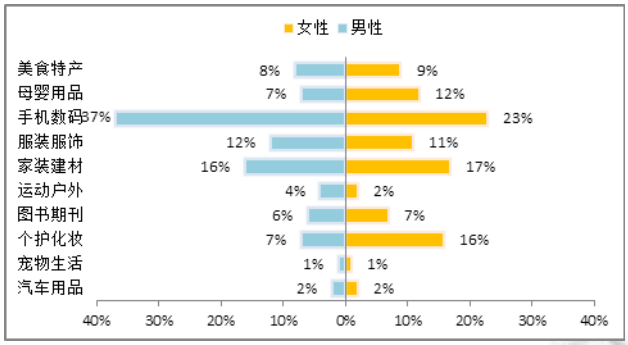



Fig.3 Percentage of purchase for different gender in each domain

图 3 不同性别用户在各个领域购买的比例

**Table 2** Example of user information

**表 2** 用户信息实例

| 社交媒体 | 新浪微博 | 京东商城 |
|---|---|---|
| 用户名(ID) | 致胜乾坤 | Suduqd |
| 性别 | 男 | – |
| 地理位置 | 北京朝阳区 | 北京朝阳区 |
| 新浪微博意图关注标签 | 旅行(75),吃(58),数码(43),育儿(16),汽车(8) | – |
| 京东购买历史类别<br>(隐式消费意图) | – | 手机数码(9),美食特产(17),运动户外(14),汽车用品(6),母婴用品(4) |

**Table 3** Basic statistical information of users in social media

**表 3** 社会媒体的用户基本统计信息

| 用户链指数 | 91 369 |
|---|---|
| 新浪微博关注标签数 | 2 019 682 |
| 新浪微博关注用户数 | 2 0995 272 |
| 新浪微博意图关注标签数 | 45 635 |
| 新浪微博类别用户数 | 869 889 |
| 京东类别数 | 10 |
| 京东购买次数 | 6 210 379 |

表 4 显示了本文所使用的数据集的基本统计信息,其中,$|D|$表示数据集中样本实例总数;$|L|$表示标记的个数;Label Cardinality 表示每个样本实例的平均标记个数, $Label\ Card=\sum_{i=1}^{|D|}|S_i|/|D|$; Label Density 表示在数据集中的总标签数, $Label\ Density=\sum_{i=1}^{|D|}|S_i|/|D||L|$.

**Tabel 4** Statistics of data in our dataset

**表 4** 数据集中的数据统计信息

| Dataset | $\|D\|$ | $\|L\|$ | Label cardinality | Label density |
|---|---|---|---|---|
| | 91 369 | 10 | 6.169 | 0.617 |

## 2 基于多标记分类的隐式消费意图识别

### 2.1 问题定义

本文将隐式消费意图识别问题看作多标记分类问题[8−14],其形式化定义为:对任一用户 $u$,设 $L$ 为隐式消费意图类别,$\{(u_i,Y_i),1\leqslant i\leqslant|D|\}$为给定的训练数据集合,其中,$Y_i\subseteq L$ 是用户 $u_i$ 的类别标记集合.目标是学习一个分类器 $h$,判断未知用户的隐式消费意图类别集合,即对任意的 $u_i$,分类器预测隶属于该实例的类别标记集合 $h(u_i)\subseteq Y_i$.此时,分类器的输出对应于某个实值函数 $f:U\times L\to\mathbb{R}$,对于给定的样本 $u_i$ 及其对应的类别标记集合 $Y_i$,分

类器将在隶属于 $Y_i$ 的类别标记上输出较大的值,而在不属于 $Y_i$ 的类别上输出较小的值,即当 $y_1 \in Y_i$ 以及 $y_2 \notin Y_i$ 时,有 $f(x_i,y_1)>f(x_i,y_2)(y_1 \in Y_i, y_2 \notin Y_i)$ 成立.

### 2.2 特征选择

本文在实现隐式消费意图识别时共使用了 4 类特征,下面对这些特征以及特征值的计算方法作详细描述.

#### 2.2.1 特征 1(用户关注行为特征(followers behavior,简称 $F_{FB}$))

用户关注行为是指用户因在新浪微博中对某用户感兴趣而产生的关注,这种单向关注行为通过关注其他用户来获得其发布的文本信息.由于微博用户发布的微博内容属于自然语言文本,对微博内容进行处理会带来较大的噪声.而用户产生关注的原因常是受职业、兴趣爱好等因素的影响,这些因素可以通过用户标签(user tag)表现出来[15,16].因此,本文将一个待分类用户的关注列表中的所有标签(以下简称关注标签)定义为用户关注行为特征.通过观察用户关注标签我们发现,如果一个标签在用户标签集合中出现的频率 $tf(t,u)$ 较高,同时在其他用户关注此标签($U_t$)中较少出现,则认为这个标签对于当前用户具有较高的重要度.而对于类似电影/音乐等大众类标签,即使描述此类标签出现在当前用户里面的频率很高,但这样的标签并不具有区分度.因此,对于每个标签赋予权重 $w_t$,这里的权重基于以下公式计算:

$$w_t = tf(t,u) \times \log\left(\frac{|U|}{|U_t|}\right) \tag{1}$$

其中,$tf$ 表示用户 $u$ 在关注用户中标签 $t$ 的出现次数;$U_t$ 表示标签 $t$ 在所有用户的关注标签列表中出现的个数;$U$ 为常数,为语料中所有用户的个数.可见,该公式与通常使用的 tf.idf 原理相似,即,如果 $tf$ 越大而 $U_t$ 越小,其权重 $w_t$ 就越大.

#### 2.2.2 特征 2(用户意图关注行为特征(intent followers behavior,简称 $F_{IFB}$))

特征 1 是通过计算一个用户关注列表中所有用户标签来度量用户的隐式消费意图.由于全部标签的通用性,使用特征 1 可能并不一定与用户的隐式消费意图具有直接的关联性.作为对特征 1 的补充,本文引入了特征 2,将一个待分类用户的关注列表中的与消费意图相关的标签抽取出来,作为用户意图关注行为特征.这里,与消费意图相关是指一个微博帐号具有明显的类别倾向性,如图 3 所示.此特征有助于识别隐式消费意图.其根本原因在于,具有隐式消费意图的用户可能会先对所要消费的商品/类别产生关注后,等待时机进行购买.

基于对用户标签和隐式消费意图类别的观察,提出如下两种情况:

情况 1:一个标签可能只属于一个消费类别下的人群使用,如,“育儿”、“汽车”.

情况 2:一个标签可以属于多个消费类别的人群使用,如,“时尚”、“星座”.

基于以上的两种情况,我们需要挖掘出类似于情况 1 中的标签(以下简记为意图关注标签),即过滤情况 2 中的标签.本文采用 Bootstrapping 的方法,利用微博标签搜索引擎抽取出意图关注标签.其目的在于,在标签短小导致的语境不足和意图未知的情况下,借助于同类别下的标签组合,迭代挖掘出与消费类别相关的用户,通过这些类别相关用户收集大量的意图关注标签.该方法主要通过 3 个步骤,可以得到最终的与种子标签类别一致的意图关注标签,方法框架如图 4 所示.

步骤 1:标签处理模块.

标签处理的主要目的包括两个方面:一是选取种子标签,二是将种子标签组合为查询关键词.由于单一标签检索回来的大部分用户标签并不一定属于相同类别下的意图标签(如,用户给自己打标签为程序员、育儿),可以简单地利用查询扩展的方式对同类别标签组合进行标签检索.具体来说,输入 $n$ 个类别(本文 $n=\{C_j|j=1,2,\ldots,10\}$,与京东 10 个类别一致)的种子消费意图标签集合 $T_S, T_S=\{t_1,t_2,\ldots,t_n\}$,其中包括每个类别 $C_j$ 的种子消费意图标签集合 $t_{ij}$.我们将种子标签组合成查询 $Query=\sum_{j=1}^{k} t_{ij}(t_{ij} \in C_j)$ 后,将其放入微博标签搜索引擎中进行检索.本文对每个类别种子标签集合人工定义了 7 个标签作为种子标签,然后对每个类别中的种子标签随机抽取 $k$ 个标签(本文中 $k$=3)作为查询,得到包含随机 $k$ 个标签组合的所有用户.这里假定利用标签组合挖掘到的用户类别与种子标签所属类别一致.

步骤 2:类别用户发现模块.

类别用户发现是基于意图关注标签挖掘其同类别下的用户.利用步骤 1 中所构建的查询,使用基于 cookie 模拟登陆的方式爬取了包含此查询的用户名单页面,获取了用户名单的 ID 号.为了过滤噪声用户,我们仅选择了加 V 用户和草根大号用户(以下简记为类别用户).主要原因在于,类别用户通常将个人标签设定与自己相关的领域,如图 5 所示.这里,加 V 用户是指新浪认证用户,是对个人用户真实身份的确认;草根大号是指未经过新浪认证加 V 用户,但拥有相当多的粉丝用户,本文定义草根大号粉丝数/关注数>1000 时为草根大号用户.这里需要说明的是,本文对某一用户的类别倾向性没有限制.即,并没有限定某个用户只能属于一个用户类别.但通过对实际的数据观察我们看到,在对步骤 1 中 $k$ 和步骤 2 用户限制后,大多数用户属于一个用户类别.

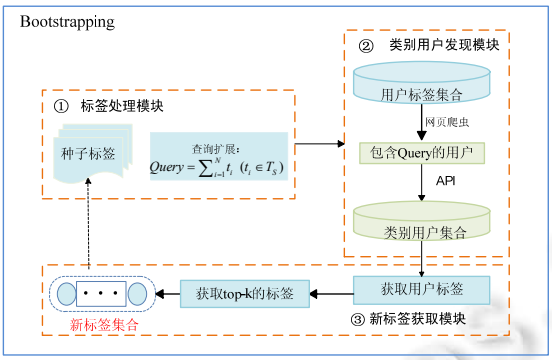



Fig.4 Framework of the follower's intent tags extraction

图 4 意图关注标签抽取框架图


Fig.5 Example of the user's tags in the baby field

图 5 母婴领域用户标签示例

步骤 3:新标签获取模块.

新标签获取的主要目的是获得更多、更准确的消费意图类别标签.从获得的类别用户的集合可以得到指定用户的标签列表.我们采用基于频率的方法抽取该类别下的消费意图标签.经过步骤 2 中的第 1 次筛选后,当候选标签经常出现在某一意图关注标签类别的集合中时,它们就可能是该类别的意图标签.具体来说,对于一个指定的意图关注标签类别,当一个候选标签出现在该类别中,且数量超过设定的阈值后(本文中取标签出现在该类别中的次数>10),我们就将这个候选标签作为该意图关注标签类别的标签.实验结果显示,这个简单的抽取方法能够达到很高的准确率.

通过此方法,我们获得了不同类别下的标签数量总计为 45 635 个,如图 6 所示.这类特征属于布尔值特征,即,用户关注标签中包含某个类别下的标签时,特征值为 1;否则,不包含为 0.

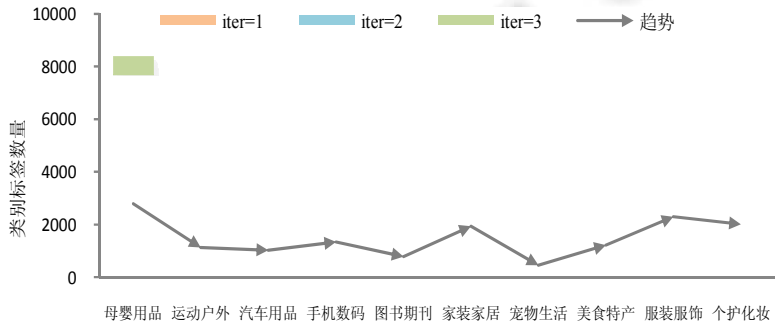



Fig.6 Number of tags from different categories

图 6 不同类别的标签数量

2.2.3 特征 3(用户意图转发行为特征(intent retweet behavior,简称 $F_{IRB}$))

用户意图转发行为是指用户在关注行为后产生的与消费相关联的转发行为,即对类别用户发布的微博进行转发的行为.我们观察到,一些用户喜欢参与转发类活动,特别是与类别用户的转发性行为.一是可以满足自己的消费需求(如,有奖转发);二是对商家的某促销活动感兴趣而进行的转发行为.有统计得出,用户在微博上浏览、转发或关注的产品,通常会同时到电子商务网站中进行购买.因此,本文通过步骤 2 获得的类别用户,利用新浪微博 API 判断一个用户是否与某类别用户具有转发行为.利用上述方法,共获得了 869 889 个类别用户,如图 7 所示.这类特征也是属于布尔值特征,即用户转发某类别下的用户微博时,特征值为 1;否则为 0.

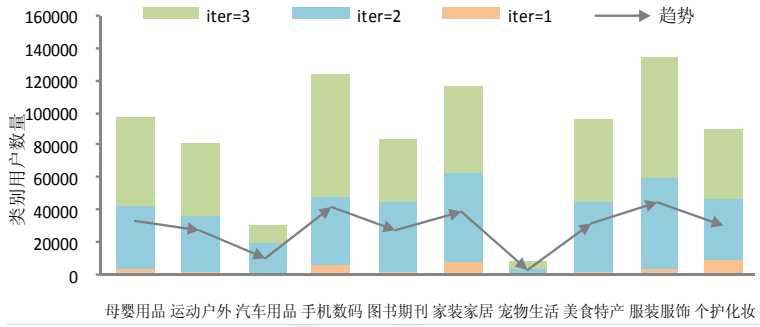



Fig.7 Number of users from different categories

图 7 不同类别的用户数量

2.2.4 特征 4(用户性别特征(user gender,简称 $F_{UG}$))

许多学者基于用户性别等人口统计学(demographic)特征研究其对购买类别的影响[7].基于此,本文抽取用户在微博中添加的性别信息作为特征 4.

### 2.3 分类器的选择

在传统的二分类问题和多分类问题中,每个样本只属于某一个类.然而在很多实际问题当中,每个样本可能同时属于多个类别.如,一篇文档既可以同时属于“体育”和“足球赛”类别,一幅图片可能同时标记为“城市”和“建筑”类别.这些问题不同于传统的二分类问题和多分类问题,称为多标记分类问题(multi-label)[9].可用于解决多标记分类问题的机器学习算法和工具很多,这些算法从总体上来看大致可分为两类[12,13]:一是根据某种策略将一个多标记分类问题转化为一系列单分类问题来解决,而对现有的分类算法本身不做改进;二是根据对多标记分类问题的特点对现有的分类算法进行改进,使得其可以应用到多标记分类问题中,其中包括基于决策树(decision tree)、支持向量机(support vector machine,简称 SVM)、最近邻(*K*-nearest neighbor,简称 KNN)等模型的分类器.在对现有算法的改进中,MLKNN(multi-label *k*-nearest neighbor)分类器[8]对多标记分类问题具有良好的分类效果.因此,本文选择 MLKNN 分类器来实现隐式消费意图的识别.限于篇幅,这里只能粗略地介绍 MLKNN 的原理.该算法的基本思想是,对基本的 KNN 算法进行改进,统计每个测试样本的 $k$ 个最近邻的类别标记信息,利用最大后验概率来决定测试样本的最终类别.本文在后续实验中对 MLKNN 分类器与 SVM 分类器进行了比较,结果发现,MLKNN 分类器在隐式消费意图这一具体问题上明显要优于 SVM 分类器.

## 3 实验与分析

由于隐式消费意图识别的方法需要训练语料,而目前国内外并没有公开发布的语料,本实验采用第 2.1 节自动获取的标注数据,将用户在京东商城的购买历史作为评价数据.在评价时,类似于文献[4,17],我们采用下面的黄金标准(gold standard),对于任意一个用户 $u$,按照其购买的产品数量的类别进行排序,如下所示:

$$gsRank(u,y_i)=\frac{purc(u,y_i)}{\sum_{y\in L}purc(u,L)} \tag{2}$$

其中,$purc(u,y_i)$是指用户 $u$ 在类别 $y_i$ 上购买的次数,$L$ 是指全部的 10 类隐性消费意图类别.

对此类别排序进行估计:

$$y_i>y_j \Leftrightarrow gsRank(u,y_i)>gsRank(u,y_j) \tag{3}$$

这里,如果用户在类别 $y_i$ 上购买的次数多于在类别 $y_j$ 上购买的次数,则 $y_i>y_j$. 对每个用户来说,理想的分类器输出结果是与黄金标准一致.

### 3.1 评价方法

为了评价多标记分类问题,我们使用文献[10,12,13]中的评价指标.主要包括两种类型的评价方法,分别是基于实例(example-based)和基于排序(ranking-based)的方法.其中,基于实例的评价指标是衡量分类器在单个测试样本上的分类效果,然后返回其在整个测试集上的均值作为最终的结果;基于排序的评价指标用以度量分类器标记类别的排序性能.基于第 2.1 节的符号表示,给定多标记分类器 $h(\cdot)$,以及多标记数据集合$\{(x_i,y_i)|1\leqslant i\leqslant |D|\}$,$Y_i\subseteq L$ 是样本 $x_i$ 的真实的类别标记集合,$f$ 为多标记分类器 $h$ 对应的实值函数,$S_i=h(x_i)$为预测集合.

#### 3.1.1 基于实例的评价方法

**定义 1(汉明损失(Hamming loss))**. 是多标记分类问题中较为常见的一个评价指标,该指标用于考察样本在单个类别上的误分类情况,即,衡量预测所得类别集合 $S_i$ 与样本实际类别集合 $Y_i$ 之间的不一致程度.具体定义为

$$HammingLoss=\frac{1}{|D||L|}\sum_{i=1}^{|D|}|S_i\Delta Y_i| \tag{4}$$

这里,$\Delta$表示两个集合的对称差(symmetric difference).

#### 3.1.2 基于排序的评价方法

**定义 2(错误率(one-error))**. 该指标描述了样本预测类别排序中,排在第 1 位的类别不是其实际类别的可能性,具体定义为

$$One\text{-}Error=\frac{1}{|D|}\sum_{i=1}^{|D|}\delta(\arg\min_{y\in L}rank_f(x_i,y)) \tag{5}$$

其中,$rank_f(x_i,y)$为与实值函数 $f(x_i,y)$对应的排序函数.该排序函数将所有的实值输出 $f(x_i,y)$映射到标记集合$\{1,2,\ldots,|D|\}$上,使得当 $f(x_i,y_1)>f(x_i,y_2)$成立时,$rank_f(x_i,y_1)<rank_f(x_i,y_2)$也成立.这里,$\arg\min_{\lambda\in L}rank_f(x_i,y)$ 表示样本 $x_i$ 中排位最靠前的类别,当 $\arg\min_{\lambda\in L}rank_f(x_i,y)\in Y_i$ 时,$\delta(y)=1$;否则,$\delta(y)=0$.

**定义 3(覆盖率(coverage))**. 该指标衡量了平均每个样本的预测类别排序中,平均需要在排序类别中跨越多少预测类别后,才能覆盖样本全部的真实类别.具体定义为

$$Coverage=\frac{1}{|D|}\sum_{i=1}^{|D|}\max_{\lambda\in Y_i}rank_f(x_i,y)-1 \tag{6}$$

**定义 4(排序损失(ranking loss))**. 该指标衡量样本预测类别排序中,不相关类别排在相关类别前的概率的平均值.具体定义为

$$RankingLoss=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{1}{|Y_i||\overline{Y}_i|}\left|\{(y_1,y_2)|f(x_i,y_1)\geqslant f(x_i,y_2),(y_1,y_2)\in Y_i\times\overline{Y}_i\}\right| \tag{7}$$

其中,$\overline{Y}_i$ 代表 $Y_i$ 在集合 $L$ 中的补集.

**定义 5(平均精度(average precision))**. 该指标衡量样本预测类别中的平均精确度.具体定义为

$$AvePrecision=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{1}{|Y_i|}\sum_{\lambda\in Y_i}\frac{\left|\{y'|f(x_i,y')\geqslant f(x_i,y),y'\in Y_i\}\right|}{rank_f(x_i,y)} \tag{8}$$

在上述的这 5 个指标中,前 4 个指标的值越小,则说明系统性能越好;第 5 个指标(average precision)的值越

大,则说明系统性能越好.

### 3.2 基线实验

本文选取了以下的方法作为基线实验:

- MostPopularity(流行度).在很多产品推荐系统中,基于流行度的方法是经常使用的基线实验.流行度算法按照物品(类别)的流行度给用户推荐其最热门的几种物品/类别.
- SVM.为了证明MLKNN分类器在隐式消费意图识别中的有效性,我们将其与SVM分类器进行了对比.本实验使用的 SVM 分类器为 libsvm-2.82.我们利用本文提出的 4 类特征在自动标注的数据集上对SVM 分类器进行了实验(这里使用了 libsvm-2.82 默认的 RBF 核函数).

#### 3.2.1 对隐式消费意图识别结果的评价

由于国内外没有公开发布的隐式消费意图评价测试语料,本文利用自动标注的语料进行评价.具体来说,我们将全部链指的 91 369 个用户数据平均分为 2 份,其中一份用于训练,剩余一份用于测试.本文对用户的隐式消费意图类别进行了统计分析,发现 90%以上的隐式消费意图类别在 1 类以上,用户的平均隐式消费意图类别在 4 类.这说明识别出用户的隐式消费意图类别是分析用户隐式消费意图的一个重要的基本单元.因此,隐式消费意图类别的正确识别有助于消费意图相关任务的解决.

本节共对 3 种方法进行了实验对比,包括:(a) 基准系统;(b) SVM 分类方法;(c) MLKNN 分类方法.其中,基准系统是指根据购买类别的流行度进行排序,即对在训练数据集中的用户购买类别的情况进行排序.

表 5 给出了 3 种方法在测试数据集上的评测结果.其中,基准系统只给出了 51.2%的平均精度,这说明隐式消费意图识别并不是一个简单的任务.为了证明 MLKNN 分类器在隐式消费意图识别中的有效性,我们将其与 SVM 分类器进行了对比.从表 5 中可以看到,两类分类器的性能都远高于基准系统,其中,SVM 给出了 72.3%的平均精度,这表明利用提出的全部特征自动在训练数据集上学习得到的分类器是有效的.利用 MLKNN(取 $k$=10)的方法学习,使其得到了更好的性能.

**Table 5** Experimental results of different approaches for implicit consumption intent

**表 5** 各种隐式消费意图识别方法的对比实验结果

| 评价指标 | | Baseline | SVM | MLKNN |
|---|---|---|---|---|
| 基于实例 | Hanmming loss↓ | 0.383 | 0.264 | 0.147 |
| 基于排序 | One-Error↓ | 0.094 | 0.026 | 0.011 |
| | Coverage↓ | 8.982 | 7.625 | 6.656 |
| | Ranking loss↓ | 0.179 | 0.126 | 0.085 |
| | AvePrecision↑ | 0.512 | 0.723 | 0.835 |

#### 3.2.2 对分类特征的评价

首先,利用上述标注数据对本文提出的分类特征进行评价.为考察本文使用的 4 类特征是否对隐式消费意图识别都有作用,我们进行了 4 组实验,每组实验依次加入基于用户关注行为特征(特征 1)、用户意图关注行为特征(特征 2)、用户意图转发行为特征(特征 3)以及用户性别特征(特征 4).其实验结果见表 6.从表 6 中我们可以看到,随着每一类特征的加入,分类的平均精度都有明显的提高,其他 4 项评价指标都有明显的降低.尤其是当使用全部 4 类特征时,分类的平均精度达到最高,其他 4 项指标达到最低.这说明本文所采用的 4 类特征对于提高多元分类的性能都是有帮助的.也就是说,全部 4 类特征均有助于隐式消费意图的识别.

**Table 6** Contributions of the 4 kinds of features

**表 6** 4 类特征的贡献

| 特征贡献 | Hanmming loss | One-Error | Coverage | Ranking loss | AvePrecision |
|---|---|---|---|---|---|
| 特征 1 | 0.318 | 0.028 | 7.835 | 0.240 | 0.693 |
| 特征 1+2 | 0.194 | 0.017 | 6.963 | 0.172 | 0.817 |
| 特征 1+2+3 | 0.183 | 0.015 | 6.937 | 0.106 | 0.821 |
| 特征 1+2+3+4 | 0.147 | 0.011 | 6.656 | 0.085 | 0.835 |

## 4 结论与展望

本文提出了一种社会媒体用户中的隐式消费意图识别方法,并将隐式消费意图识别作为一个多标签分类问题加以解决.具体来说,本文利用 MLKNN 分类器解决隐式消费意图中的多分类问题,并综合使用了 4 类特征,即:(1) 用户关注行为特征;(2) 用户个人信息特征;(3) 用户消费意图转发行为特征;(4) 用户消费意图关注行为特征.在隐式消费意图识别评价方面,本文尝试将不同媒体用户中的用户身份链接起来,自动抽取了 12 万余对社交媒体网站和电子商务网站的用户数据.在此自动评价集上的实验结果表明,本文使用的分类器和 4 类特征对于隐式消费意图识别都是有效的.

在今后的工作中,我们会尝试寻找一种方法来标定用户隐式消费意图中真正希望购买的具体产品,将更加方便对用户迅速推荐所需的产品信息.此外,我们也会考虑将本文提出的自动链指方法应用到其他不同社区中,将一个用户的不同形式、全面的信息都聚合起来,形成更丰富的个人信息,从而解决推荐系统和个性化系统中的冷启动问题.

**致谢** 在此,我们向对本研究工作提供帮助的老师和同学表示感谢.特别要感谢李一鸣、焦阳等同学在实验数据处理上的工作,还要感谢宋巍、伍大勇、张伟男等同学对本文初稿进行审阅并提出宝贵意见.

**References**:

[1] Goldberg AB, Fillmore N, Andrzejewski D, Xu Z, Gibson B, Zhu XJ. May all your wishes come true a study of wishes how to recognize them. In: Proc. of the 2009 Annual Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL). Morristown: ACL Press, 2009. 263−271.

[2] Chen ZY, Liu B, Hsu M, Castellanos M, Ghosh R. Identifying intention posts in discussion forums. In: Proc. of the 2013 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Morristown: ACL Press, 2013. 1041−1050.

[3] Fu B, LIU T. Weakly-Supervised consumption intent detection in microblogs. Journal of Computational Information Systems, 2013, 6(9):2423−2431.

[4] Zhang Y, Pennacchiotti M. Predicting purchase behaviors from social media. In: Proc. of the Int'l World Wide Web Conf. Steering Committee (WWW). New York: ACM Press, 2013. 1521−1532. [doi: 10.1145/2488388.2488521]

[5] Ding X, Liu T, Duan JW, Nie JY. Mining user consumption intention from social media using domain adaptive convolutional neural network. In: Proc. of the Association for the Advancement of Artificial Intelligence (AAAI). Menlo Park: AAAI Press, 2015. 2389−2395.

[6] Liu J, Zhang F, Song XY. What's in a name? An unsupervised approach to link users across communities. In: Proc. of the 6th ACM Int'l Conf. on Web Search and Data Mining (WSDM). New York: ACM Press, 2013. 495−504. [doi: 10.1145/2433396.24334 57]

[7] Chiu YB, Lin CP, Tang LL. Gender differs: Assessing a model of online purchase intentions in e-tail service. Int'l Journal of Service Industry Management, 2005,16(5):416−435. [doi: 10.1108/09564230510625741]

[8] Zhang ML, Zhou ZH. A *k*-nearest neighbor based algorithm for multi-label classification. In: Proc. of the 1st IEEE Int'l Conf. on Granular Computing (GrC). IEEE Press, 2005. 718−721. [doi: 10.1109/GRC.2005.1547385]

[9] Tsoumakas G, Katakis I. Multi-Label classification: An overview. Int'l Journal of Data Warehousing and Mining, 2007,3(3):1−13. [doi: 10.4018/jdwm.2007070101]

[10] Jiang Y, She QQ, Li M, Zhou ZH. A transductive multi label text categorization approach. Journal of Computer Research and Development, 2008,45(11):1817−1823 (in Chinese with English abstract).

[11] Schapire RE, Singer Y. Boostexter: A boosting-based system for text categorization. Machine Learning, 2000,39(2):135−168. [doi: 10.1023/A:1007649029923]

[12] Read J, Pfahringer B, Holmes G, Frank E. Classifier chains for multi-label classification. Machine Learning, 2011,85(3):333−359. [doi: 10.1007/s10994-011-5256-5]

[13] Tsoumakas G, Katakis I, Vlahavas I. Mining multi-label data. In: Data Mining and Knowledge Discovery Handbook. 667−685. [doi: 10.1007/978-0-387-09823-4_34]

[14] He ZF, Yang M, Liu HD. Joint learning of multi-label classification and label correlations. Ruan Jian Xue Bao/Journal of Software, 2014,25(9):1967−1981 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4634.htm [doi: 10.13328/j.cnki.jos.004634]

[15] Chen Y, Lin L, Sun CJ, Liu BQ. A tag recommendation method for microblog users. Intelligent Computer and Applications, 2011, 1(5):21−26 (in Chinese with English abstract). [doi: 10.3969/j.issn.2095-2163.2011.05.006]

[16] Jin YA, Li RX, Wen KM, Gu XW, Lu ZD, Duan DS. A survey on social annotation and its application in information retrieval. Journal of Chinese Information Processing, 2010,24(4):52−62 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2010.04.008]

[17] Zhang Y, Pennacchiotti M. Recommending branded products from social media. In: Proc. of the 7th ACM Conf. on Recommender Systems. New York: ACM Press, 2013. 77−84. [doi: 10.1145/2507157.2507170]

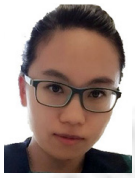
**附中文参考文献**:

[10] 姜远,佘俏俏,黎铭,周志华.一种直推式多标记文档分类方法.计算机研究与发展,2008,45(11):1817−1823.

[14] 何志芬,杨明,刘会东.多标记分类和标记相关性的联合学习.软件学报,2014,25(9):1967−1981. http://www.jos.org.cn/1000-9825/4634.htm [doi: 10.13328/j.cnki.jos.004634]

[15] 陈渊,林磊,孙承杰,刘秉权.一种面向微博用户的标签推荐方法.智能计算机与应用,2011,1(5):21−26. [doi: 10.3969/j.issn.2095-2163.2011.05.006]

[16] 靳延安,李瑞轩,文坤梅,辜希武,卢正鼎,段东圣.社会标注及其在信息检索中的应用研究综述.中文信息学报,2010,24(4):52−62. [doi: 10.3969/j.issn.1003-0077.2010.04.008]



**付博**(1983－),女,黑龙江海伦人,博士,主要研究领域为社会计算,信息检索.

**刘挺**(1972－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为社会计算,信息检索,自然语言处理.