中国科学: 信息科学 2015 年 第 45 卷 第 12 期: 1523–1535
www.scichina.com info.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS
社会媒体处理专题 · 论文

基于图排序的社会媒体用户的消费意图检测

刘挺*, 付博, 陈毅恒

哈尔滨工业大学社会计算与信息检索研究中心, 哈尔滨 150001
* 通信作者. E-mail: tliu@ir.hit.edu.cn

收稿日期: 2015–09–16; 接受日期: 2015–10–20; 网络出版日期: 2015–12–02
国家自然科学基金面上项目 (批准号: 61472107)、国家自然科学基金青年项目 (批准号: 61202277)、国家重点基础研究发展计划 (973 计划) (批准号: 2014CB340503) 和国家自然科学基金重点项目 (批准号: 61133012) 资助

摘要 消费意图是指用户为满足某种需要, 在一定购买动机的支配下, 通过文本内容表达出对产品或服务的购买意愿. 消费意图检测旨在推断用户在文本中是否表达出消费意图的含义. 我们定义微博中的消费意图必须包含两个重要元素, 分别是消费意图触发词和消费意图对象 (即需要购买的产品), 这两种元素直接引发用户的购买意愿, 是决定用户消费意图的重要特征. 本文提出了基于弱监督的图排序算法, 该方法适用于数据总量较大、已标注数据量相对较小的情形中, 并且可以使未标注数据和标注数据同时参与到图排序算法的学习过程中. 实验结果表明, 所采用的图排序方法对于消费意图检测是行之有效的.

关键词 消费意图 消费意图检测 社会媒体 图排序 弱监督

1 引言

随着网络媒体技术的发展和普及, Twitter、新浪微博等社会媒体成为了最普遍的信息发布、传播和共享的工具之一. 在这些用户生成的数据中蕴含着用户为了满足某种需求, 在一定购买动机的支配下, 通过文本内容表达了对某产品或服务的购买意愿, 我们将此意愿称之为消费意图. 本文以新浪微博 —— 一种国内最为流行的社会媒体网站为例, 来自动检测具有消费意图的短文本. 消费意图检测即判断用户发布的一条微博是否明确表达出对某产品产生了购买的意愿. 我们定义用户表达出的消费意图必须包含两个重要元素, 分别是消费意图触发词 (以下简称触发词) 和消费意图对象 (即需要购买的产品), 这两种元素直接引发用户的购买意愿, 是决定用户消费意图的重要特征. 表 1 中, 具有消费意图的文本明确指出了用户期望未来购买的目标, 如句子 2 明确表示他/她想买一个空气净化机. 其中, “想买” 即是具有触发消费意图的词语, “空气净化机” 表示消费意图对象. 有了这些信息作为用户消费意图的描述, 可以更好地理解用户消费行为进而为其提供精准的产品推荐.

解决消费意图检测问题最直接的方法是基于模板匹配的方法, 即按照消费意图定义利用 〈 触发词 〉 或 〈 触发词 + 名词短语 〉 的模板匹配方式挖掘具有消费意图的文本. 但该方法在应用中存在着几个明显的局限性. 首先是制定模板规则的限制, 通常规则的制定要基于一系列预处理与语言分析等过程, 导致人工编写规则费时费力, 并且所编写的规则可扩展性差. 其次是灵活性的限制, 一个模板所含有的词是固定的, 在使用时需要精确匹配模板, 但效果往往不甚理想, 例如: 将触发词作为模板抽取

引用格式: 刘挺, 付博, 陈毅恒. 基于图排序的社会媒体用户的消费意图检测. 中国科学: 信息科学, 2015, 45: 1523–1535, doi: 10.1360/N112015-00220

表 1　微博实例与其对应的标注类别

Table 1　Example of messages and their corresponding label from the sina weibo

序号	类别	实例
1	消费意图句	现在想买一款全触屏的手机, 大家有没有什么好的推荐.
2	消费意图句	我刚起来, 看镜子吓我一跳. 全脸是掉死皮. 北京天气实在太干燥. 我想买一个空气净化机. 朋友们, 能推荐给我几个好的机器? 欢迎转发这个信息.
3	非消费意图句	中国游客简直疯了, 成了日本旅游的绝对主力, 看到什么都想买, 是要把日本买空?
4	非消费意图句	推荐大家一个活动 —— 美团网: 0 元参与即可第一时间赢取最新款苹果现货 1 台! 现货! 是现货! 美国直送! 大家都来试试运气吧! http://sinaurl.cn/htvflR

具有消费意图的文本时, 由于触发词在不同的语境下含义不同, 类似触发词 “买”、“推荐” 等, 用户在句子中并没有表达消费意图的含义. 通过统计触发词和对实际数据的观察, 利用包含 ⟨ 触发词 + 名词短语 ⟩ 的模式抽取微博后, 发现用户发布的微博中约有 13% 的内容真正具有消费意图, 这些例子可以见表 1 中句 1 和句 2, 而抽取出的其余大部分微博并不具有消费意图, 如表 1 中句 3 和句 4 所示.

在以往的研究中, 研究者们通常将有指导的分类方法用于消费意图检测的任务中 [1~3]. 消费意图检测任务可定义为: 给定一个句子 s , 类别体系 $c = \{c_i, i = 1, 2\}$, 通过学习到的分类器 f 判定句子 s 是否属于类别 c_i, 具体表示为 $f : \{s\} \to c$. 在训练分类器 f 时, 通常基于 n-gram 特征、模板特征等来完成句子的消费意图检测. 但是, 基于有指导的方法依赖于大量的精确标注的训练语料, 这些人工标注语料的方法在实际中是相当费时费力的. 因此, 如何在有标注数据的基础上使用未标注数据来提高系统的性能成为一个非常重要的问题.

为了充分利用标注数据和未标注数据提升系统的性能, 本章使用了基于弱监督的图排序算法. 该方法可以适用于总数据量较大、已标注数据量相对较小的情形中, 并且可以使未标注数据和标注数据同时参与到图排序算法的学习过程中. 本文提出的图排序算法是在 RelevenceRank 方法上进行的改进, 类似的, 利用人工选定一些关键词作为种子节点, 在构建图时为种子节点赋予非零初值, 该方法认为当一个节点与越多的具有消费意图的节点相关联时, 该节点具有消费意图的可能性越大. 与 RelevenceRank 方法不同之处在于, 本文方法利用自动获取的种子集合 (具有消费意图的文本)、未标注集合, 以及文本之间消费意图表达的相似性构建图, 将未标注数据和标注数据的关系描述为一个无向图, 其中图上节点由具有消费意图的词组合形式构成, 每一个具有相似性关系的节点对连接成图上的一条边, 最终利用图的结构将节点权重值传递给其相邻节点, 以此来为每一个节点计算权重值. 这种图排序的方法可以理解为基于文本之间的消费意图表达相似性关系进行节点权重值的学习.

2　相关工作

2.1　基于弱监督学习的相关方法

目前, 有指导的机器学习方法被成功地应用于消费意图检测的研究中 [2]. 一般来说, 有指导的学习方法要求训练语料的规模足够大, 准确率和一致性足够好, 否则会大大影响学习到的分类器的效果. 然而, 人工标注语料是一个费时费力的过程, 这一做法也限制了语料的规模和覆盖范围. 近年来, 有许多学者对弱监督学习 (或称为 “半监督学习”) 展开了研究, 并尝试将其应用到自然语言处理、社会计算、情感分析等领域的研究中去.

弱监督学习方法是把未标注的数据作为学习对象, 通常采用自举法 (bootstrapping), 辅以少量种子 (seed) 或已标注数据为目标, 来进行自学习或自训练. 需要指出的是, 本文着重介绍弱监督样本标记的获取方法. 弱监督样本的标记也称为弱标记, 是指在获取标记的手段限制下, 样本的标记信息存在缺失或标记信息包含噪声, 从而导致样本的标记有可能不是真实标记, 有学者证明了弱监督样本是有价值的. 例如, Angluin 等 [4] 在近似可学习理论的基础上提出了噪音学习理论, 认为如果大多数样本的标记正确, 则引入的噪声所带来的弊处可以被使用大量训练样本带来的利处所抵消. 也就是说, 弱监督样本的引入可以在不增加人工标注工作量的条件下, 扩大样本的数量以提高模型的泛化能力.

由于弱标记样本普遍存在于社会媒体中, 因而, 信息检索、社会计算等许多研究领域都可以应用弱标记样本提高学习性能. Dai 等 [5] 将搜索引擎查询日志中的弱标记样本用于在线商业意图 (online commercial intention) 识别的研究中. 该方法的核心思想是利用点击具有商业意图网页的查询构建训练数据集, Davidov 等 [6] 使用含有哈希标签和表情符的弱标记数据获取 Twitter 中的情感数据. 该方法利用 50 个哈希标签和 15 个表情符抽取短文本, 以此获得的弱标记数据避免了人工标注.

2.2 基于图排序的文本抽取相关方法

PageRank 算法是 Google 用来标识网页重要度的一种方法 [7]. 其基本思想是一个网页的质量和重要性可以通过链向它的其他网页的质量和数量来决定, 也就是说当有越多或越重要的网页指向某网页, 那么该网页的重要度越高. 受到 PageRank 方法的启发, 该思想在各个研究领域均有重要应用. 例如, 在信息抽取领域, Mihalcea 和 Tara 提出一种基于图的文档摘要排序算法 (TextRank), 他们的方法中考虑了文档中词之间的共现关系, 即与重要词经常一起出现的词也很重要 [8]. 在信息检索领域, Yang 等提出了有指导的 RelevanceRank 图排序算法 [9], 用以解决查询和文档之间的相关性. 我们的图排序方法借鉴了 RelevanceRank [9] 的图排序方法, 使得标注数据和未标注数据可以同时参与到图排序的过程中. 为了更好地理解本文方法中的图模型, 我们在这里简要介绍这种图排序的方法.

RelevanceRank 算法 [9] 是一种基于弱监督的指导方法, 即利用人工选定一些关键词作为种子节点, 在构建图时为种子节点赋予非零初值, 基于词语之间的相似性关系从文档中挖掘出更多的与该词语相关的关键词. 该算法中对每个节点的权重值计算如下:

$$RR_{v_i}^t = \sum_{e=\langle v_i,v_j\rangle} \alpha_{v_j} \cdot RR_{v_j}^{t-1} \cdot T[v_j,v_i] + RR_{v_i}^0 \sum_{v_j\in \mathrm{V}} (1-\alpha_{v_j}) RR_{v_j}^{t-1}, \tag{1}$$

$$T[v_i,v_j] = \begin{cases} \frac{w_{\mathrm{p}_e}}{\sum_{e'=(v_i,w)} w_{\mathrm{p}_{e'}}} & \text{if } \exists e=\langle v_i,v_j\rangle \in E, \\ 0 & \text{otherwise}, \end{cases} \tag{2}$$

$$RR^0(v_i) = \begin{cases} \frac{s(c_{v_i})}{\sum_{v_i\in S} s(c_{v_i})} & \text{if } v_i \in S, \\ 0 & \text{otherwise}, \end{cases} \tag{3}$$

这里, $RR_{v_i}^t$ 表示节点 v_i 在第 t 次迭代的得分, 并且 $RR_{v_i}^t$ 的分值依赖于邻居节点第 $t-1$ 次权重值 $RR_{v_j}^{t-1}$ 的结果. 初始状态定义为: 对属于种子集合 S 中的节点 v_i, 其对应的初始状态见式 (3); 否则为 0. 节点 v_i 的分值以概率 $1-\alpha_{v_j}$ 的速度传递给种子节点, 以保证种子节点并不因迭代而损失分值. T 为 $m\times n$ 转移矩阵, 矩阵中的每个元素 p_{ij} 表示从当前节点 v_i 到其他节点 v_j 的概率. 当达到迭代收敛条件时, 根据每个节点的分值排序, 获得 top-k 个节点作为关键词集合. 在文献 [6] 中, 作者将此算法应用于查询检索中, 目的是使搜索引擎检索得到与当前查询更相关的检索结果.

3 基于图排序的消费意图检测问题描述

3.1 消费意图检测问题描述

消费意图检测的基本目标是根据文本特征和可获得的指导信息来确定用户发布文本的购买意愿, 即判断用户在文本是否表达了潜在的购买某种产品的意图. 我们定义触发词和消费意图对象是消费意图表达中最关键的描述性文字, 触发词 (Trigger) 是指能够清楚表达消费意图的动词, 如 “买”, “推荐” 等, 消费意图对象是用户期望购买的产品, 是消费意图得以满足的明确目标. 该问题可以形式化的定义如下: 输入具有消费意图的种子集合 V_S, $V_S = m_1, m_2, \ldots, m_i, \ldots, m_N$, 以及未标注的候选消费意图集合 $V_c = \{V - V_s\}$, 输出从集合 $\{V - V_s\}$ 中挖掘出具有消费意图的文本集合 V_{CI}.

3.2 基于图排序的消费意图检测系统框架

基于前面的讨论, 我们将每条微博作为图节点并赋予初始值, 以具有消费意图的文本之间的相似关系作为图上节点相连接成为边的条件, 目标是通过图排序的算法实现节点权重值从已标记的节点向未标记节点进行传播, 这样为每个未标记节点赋予最终的权重值, 从而完成基于图排序的消费意图检测任务.

本文方法的系统框架如图 1 所示, 这个框架包括 3 个部分: (1) 基于弱指导的方法获得有标记的集合作为种子集合 V_s; (2) 基于 〈 触发词 + 名词短语 〉 的模型匹配方式获得候选的具有消费意图的文本集合 V_c; (3) 通过以上两步获得的文本作为图上节点, 并根据节点之间的相似度为作为图上边连接条件, 通过不断迭代的过程计算未标注节点的权重值并按此分值排序.

具体的, 第 1 部分是利用少量的、具有消费意图含义的哈希标签 (如 # 求购 # 等), 在微博集合中抽取包含指定哈希标签的微博作为种子 (即正例). 在第 2 部分图模型构建及利用图排序方法获得高分值节点抽取消费意图句的步骤中, 将种子集合和候选微博集合作为图上全部的节点, 并把句中具有消费意图的表达定义为五元组形式, 利用五元组语义相似性建立节点之间的边, 通过图排序的算法将种子节点的分值按照节点间的消费意图相似关系扩散到未标注文本上, 最后根据节点分值的排序抽取出 top-k 个节点作为与种子集合具有相同标注的文本 (即正例).

4 基于图排序的消费意图检测

4.1 图模型构建

首先介绍图模型的构建方法. 形式化地, 所有的数据 $V = \{S_i, i = 1, 2, \ldots, N\}$, 由种子集合 V_s 和候选消费意图文本集合 $\{V - V_s\}$ 构成. 设 V 中所有的文本可以表示为图 $G(V, E)$, E 为图中边集合, E 中的每一条边连接着图中的顶点 (m_i, m_j), 边的权重值为 $A = \{a_{ij}, i, j = 1, 2, \ldots, N\}$ 反映了各个顶点之间的相似性. 然而全连通图需要考虑所有边的权重, 计算量较大, 为了简化计算, 我们采用了启发式的方法构建 ε 近邻图, 即当两个节点之间的距离有 $\|\mathrm{S}_i - \mathrm{S}_j\| < \varepsilon$ 时, 则连接节点 S_i 和 S_j 来构建图.

为了获得任意节点 S_i 最终的权重值 $CI(S_i)^t$, 首先定义节点权重 $CI(S_i)^t$ 初始状态 $CI(S_i)^0$ 为 N 维向量, 其中 t 为该向量的迭代次数. 当达到最终状态 $CI(S_i)^f$ 时, 获得图中各个节点的最终权重值得分. 对于 $CI(S_i)$ 的初始状态 $CI(S_i)^0$, 按照如下方式设定其中的元素 $CI(S_i)$: 对属于种子集合 V_s 的节点 (S_i), 其对应的初始状态 $CI(S_i)^0 = 1$; 而对属于候选消费意图文本集合 $\{V - V_s\}$, 其对应的初始状态 $CI(S_i)^0 = 0$.

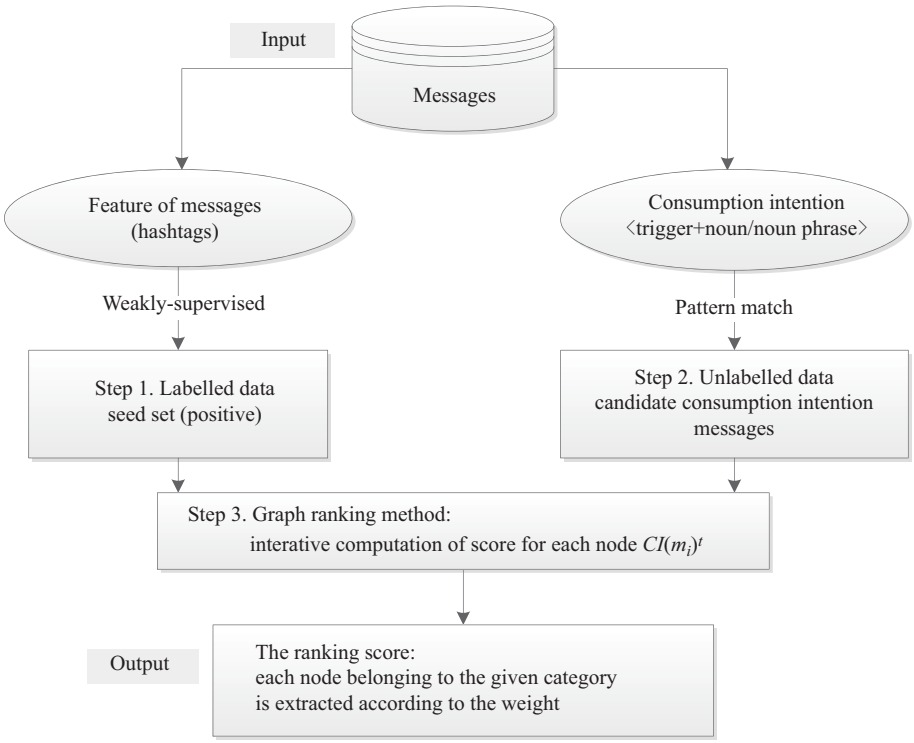

图 1 基于图排序的消费意图检测框架

Figure 1 Framework of consumption intention detection based on graph ranking

算法的目标是根据构建的图 G 和有标注的种子集合 V_s, 为无标注的候选文本集合 $\{V - V_s\}$ 中的节点计算最终权重值得分, 并输出 top-k 的节点作为抽取出的具有消费意图的集合 V_{CI}. 于是, 对图 G 中任意一个节点 S_i 的最终权重得分 $CI(S_i)^t$, 计算如下:

$$CI(S_i)^t = (1-\alpha)\cdot CI(S_i)^0 \sum_{v_j \in S} CI(S_j)^{t-1} + \alpha \sum_{\mathrm{e}=\langle S_i, S_j\rangle} A[S_i, S_j] CI(S_j)^{t-1}, \tag{4}$$

这里, 式 (4) 中 $CI(S_i)^t$ 代表第 t 次迭代时节点 S_i 的得分. 权重矩阵 A 中的元素 a_{ij} 根据节点之间的相似度计算得到, A 中每个元素的值为非负, 表示节点之间相关联的强度. α 是一个衰减因子 (damping factor), 取值范围是在 $[0,1]$ 之间. 对于每个节点, 衰减因子表示针对当前图上的节点 S_i 有 $1-\alpha$ 的概率随机游走到图中的其他节点上, 而有 α 的概率跳转到该节点的邻居节点上. 其中, 图在迭代的过程中, 需要尽可能保证初始种子节点的分值不因迭代发生变化, 即当随机游走到非种子节点时, $CI(S_i)^0$ 的分值赋予零, 而当随机游走到种子节点时, $CI(S_i)^0$ 的值赋予零值为非零的分值. 当达到迭代停止条件 $|CI(S_i)^t - CI(S_i)^{t-1}| < \theta$ 时, 表示经过迭代后所有图上节点所处的状态为最终状态.

基于以上的定义, 本章提出了基于图排序的消费意图检测算法, 如算法 1 所示.

4.2 图的节点选取

4.2.1 种子节点选取

图模型的计算一般都需要种子节点. 本文提出了一种基于弱指导的、由用户自然标注话题的字符串自动获取种子集合. 因此, 如何保证获取的种子节点的质量是值得研究的问题. 自然标注 [8], 即利用

表 2　哈希标签列表及对应获取到的训练集数量

Table 2　List of hashtags in training data set

Hashtags	Number	Accuracy (%)
# 求推荐 #	464	96.3
# 求购 #	543	98.6
# 粉丝问路 #	883	90.1
# 京东购物 #	101	92.7
Total	1991	

算法 1 基于图排序的消费意图检测算法

Require: 种子消费意图文本集合 V_s, 候选微博集合 $\{V - V_s\}$, k, 最大迭代次数 MaxIterations, 阈值 θ;

Ensure: 集合 $\{V - V_s\}$中的节点 S_i 迭代 t 次的得分 $CI(S_i)^t$, top-k 个具有较高得分的节点集合 V_{CI};

根据 V 和 V_s, 构建候选消费意图微博初始状态 $CI(S_i)^0$, 初始化 $CI(S_i)^t$, $t = 0$;

构建权重矩阵 A, 矩阵中每个元素根据节点之间的相似度计算得到;

根据式 (4) 迭代计算节点 $S \in V$ 的 得分 $CI(S_i)^t$, 直到达到收敛条件 $t =$ MaxIterations 或 $\left|CI(S_i)^t - CI(S_i)^{t-1}\right| < \theta$;

对每个节点 $S_i \in V$, 计算得分 $CI(S_i)$ 并排序;

输出集合 V_{CI}.

互联网上的各种资源, 如用户对网页、图像、博客、音乐等 “义务” 标注. 弱监督样本的引入可以在不增加人工标注工作量的条件下, 扩大学习样本集合, 提高模型的泛化能力. 结合微博自身的特点, 本章从哈希标签角度引入弱监督的标注数据, 解决自动获取高质量标注数据的问题.

哈希标签 (Hashtags), 又称哈希话题标签, 使用 “#...#” 符号以及其中的内容标记当前微博的关键字或主题. 例如: “#NBA 总决赛 # 詹姆斯上半场, 15 分 8 篮板 3 助攻” 是关于体育赛事 NBA 的话题.

本文借鉴前人提出的方法 [10], 使用哈希标签自动抽取具有消费意图的微博作为种子集中的正例, 哈希标签和选取如表 2 所示. 为了验证弱监督方法抽取的数据质量, 我们对获取的数据进行人工标注, 得到了每个哈希标签抽取出具有消费意图微博的准确率. 从表 2 中可以看到, 基于哈希标签抽取正例的准确率都在 90% 以上, 因而数据本身的噪声对消费意图检测任务的影响不大.

4.2.2　非种子节点选取

对于非种子节点的选取, 我们利用 〈 触发词 + 名词短语 〉 模式匹配方式抽取出的微博作为非种子节点集合. 本文利用文献 [3] 中的触发词列表, 其中包含 52 个消费意图触发词和 766 个非消费意图触发词. 在抽取未标数据集时, 如果句中的触发词与名词短语具有动宾关系时, 则抽取出这样的句子. 我们使用微博 API 自动抓取出 2012 年 3 月共 11854002 条微博, 对其进行预处理后, 基于以上的模式匹配方式共获得了 52 万条微博作为非种子节点集合.

4.3　图的边权重设置

在确定图模型中的种子节点和非种子节点后, 消费意图检测的问题就可以转化为确定图上非种子节点的类别, 也就是未标注类别的微博. 利用已获得的种子节点的类别和消费意图特征来确定未标注文本的最终类别.

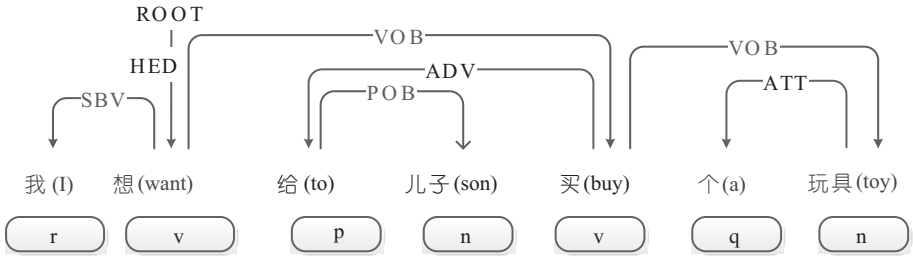

图 2 消费意图文本的依存句法结构示例

Figure 2 Syntactic structure samples of the consumption intention message

为了便于计算消费意图文本间的语义相似度, 我们将微博表示为具有消费意图含义的词组合形式. 具体的, 一条微博 $S = m_1, \ldots, m_k$, 每个 m_i 表示一条微博中的一句话 (即以句号为结尾), 利用 LTP 工具 [11] 对句 S 进行分词和依存句法分析后, 将 S 表示为与消费意图有关的五元组形式 Tuple$(m_i) = t(\text{neg}), t(\text{neg})_\text{vob}$, $t(\text{pos})\_\text{vob}$, SBV, POB), 其中 $t(\text{neg})$ 代表了句子中出现的非消费意图触发词, $t(\text{neg})\_\text{vob}$ 代表了非消费意图触发词和与其具有动宾关系的名词短语, $t(\text{pos})\_\text{vob}$ 代表了消费意图触发词和与其具有动宾关系的名词短语, SBV 代表了句中的主谓关系, 用以判定句中的施事者, POB 代表了句中的介宾关系, 用以判定句中的受事者. 图 2 显示了一条具有消费意图的微博经过依存句法处理后的结构示例. 实例化表示为 Tuple(m_i) =(NULL, NULL, 买玩具, 我想, 给儿子), 其中箭头的指向代表从父节点指向儿子节点, 如具有动宾关系的节点 “买玩具” 中, “买” 为父节点, “玩具” 为儿子节点.

显然, 这种精确到词的消费意图表示方法容易漏掉一些句法结构相似但不同表述的句子, 因此我们对此种表示进行泛化, 主要针对非消费意图触发词 ($t(\text{neg})$)、动宾搭配关系 ($t(\text{neg})\_\text{vob}$ 和 $t(\text{pos})\_\text{vob}$)、主谓关系 (SBV) 中的父节点和介宾关系 (POB) 中的儿子节点进行. 上述例子经过实例化后进一步泛化, Tuple(m_i) =(NULL, NULL, $t(\text{pos})$ 玩具, 我 SBV, 给 POB). 因而对于一条微博来说, 可将其表示为具有消费意图含义的词组合形式为 $CI(S) = \bigcup\limits_i \text{Tuple}(m_i)$.

对两条微博的消费意图相似度计算方法如下:

$$\text{sim}_{CI}(S_i, S_j) = \frac{CI(S_i) \cap CI(S_j)}{CI(S_i) \cup CI(S_j)}, \tag{5}$$

同上, $CI(S_i) \cap CI(S_j)$ 表示两条微博包含共同的消费意图表示的词组合个数, $CI(S_i) \cup CI(S_j)$ 表示两条微博中出现的所有消费意图表示的词组合个数. 基于以上的分析, 我们可以计算出图上所有节点之间的相似度, 如果相似度值大于已设定的阈值 θ, 则说明当前节点之间具有一定的相似度. 相应地, 我们在图模型中将这两个节点连成一条边, 并设置边权重为两者的相似度值.

5 实验结果与分析

本文的实验包括两部分: (1) 对图构建中设置不同边权重的方法与本文方法进行对比; (2) 对以往有监督和弱监督的方法进行对比.

5.1 对基于图构建方法的评价

5.1.1 评价方法

采用 $p@N$ 指标来衡量消费意图句抽取的效果, $p@N$ 考虑的是抽取前 N 个微博中有多少是被标记为消费意图类别, 其值越大说明抽取出的结果越好, 计算如下:

$$P@N = \frac{\sum CI(t)}{N}, \tag{6}$$

其中, $\sum CI(t)$ 表示前 N 条微博中具有消费意图的微博数量, N 代表全部的微博数量.

5.1.2 对比实验系统

在图排序算法学习的过程中, 权重矩阵也是图构建的核心部分. 与本文提出的消费意图相似度构建边权重的方法进行了对比, 采用以下两种相似度方法作为对比实验系统, 来衡量未标记节点与种子节点的相似度, 分别是余弦相似度和 Jaccard 相似度. 这里, $S_i, (i = 1, 2)$ 代表任意两条已经过分词的微博消息 (Messages), 计算方法分别定义如下:

• 余弦相似度算法

也称为余弦距离, 该方法是用向量空间中两个 n 维向量夹角的余弦值 θ 作为衡量两条微博之间相似程度的度量. 其中, 已知两个微博向量 S_i 和 S_j 的余弦相似度计算为

$$\mathrm{sim}_{\cos} = \cos(S_i, S_j) = \frac{S_i \cdot S_j}{\| S_i \| \cdot \| S_j \|}. \tag{7}$$

• Jaccard 相似度算法

该方法是通过衡量两条微博中包含词数的交集在两条微博中包含词数的并集中所占的比例, 来评估它们之间的相似度. 已知两个微博向量 S_i 和 S_j 的基于 Jaccard 相似度计算为

$$\mathrm{sim}_J(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}, \tag{8}$$

其中, $|S_i \cap S_j|$ 表示两条微博中包含相同词的个数, $|S_i \cup S_j|$ 表示两条微博中包含的所有词的个数.

利用上述两种方法以及 4.3 小节提出的边权重计算方法构建图, 利用 4.2 小节获得的数据构建图中节点. 进而抽取具有消费意图的微博. 本文中最大迭代次数设为 1000 次, 阈值 θ 设为 0.00001, k 值取 2000. 再从以上两种对比方法的结果中随机抽样 2000 条微博用于评价准确率. 由此得到的 6000 条微博被混放在一起, 再提供给人工标注者. 这样做的目的是令标注者无从得知当前微博是由哪种方法抽取得到, 由此可避免标注者在数据标注时对某种方法的倾向性, 从而保证人工标注结果更加公正. 在标注者完成标注之后, 我们再将以上 3 种不同设置边权重的方法的抽取结果分开, 并按照上述 $P@N$ 的评价方法计算出 3 种方法抽取出的具有消费意图的微博的准确率.

5.1.3 实验结果评价与分析

本文基于图排序的消费意图检测方法需要考虑两种因素的影响, 分别需要考虑图的连通性和图上的边权重设置, 它们在图构建时起到了很重要的作用. 因此, 在实验中我们从图连通性和图上边权重的角度考察了图排序的方法, 以此证明该方法的切实有效性.

在本文实验中, 图排序算法中的衰减因子 α 是调整节点权重值对初始状态或者图结构的依赖程度. 在图 3 中显示不同范围的 α 值对应的抽取结果, 可以看到, 在 α 取值为 0.65 时, 实验结果显示抽

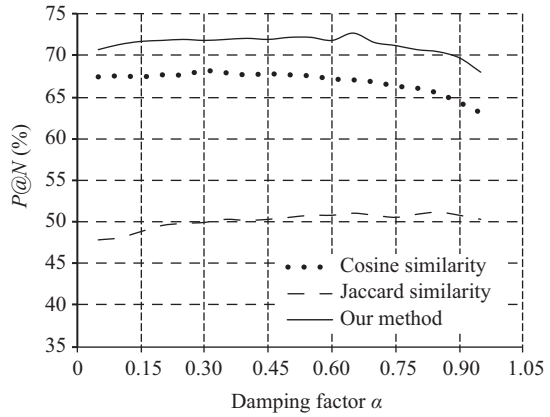

图 3 消费意图检测方法随衰减因子的变化

Figure 3 Change of consumption intention detection with the damping factor

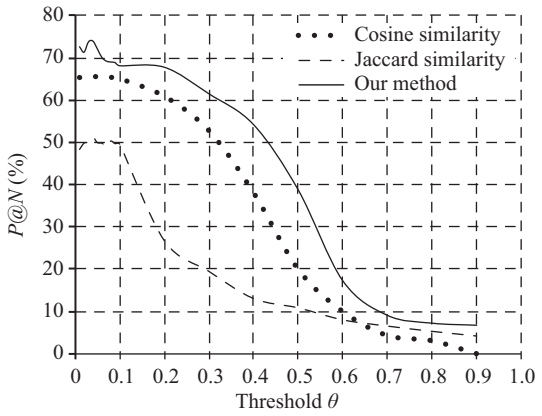

图 4 消费意图检测方法受阈值的影响

Figure 4 Change of consumption intention detection with the threshold

取的效果最佳, 因此本文中 α 取值为 0.65. 从图 3 中也可以看到, 在这个取值范围内实验抽取的效果基本相当, 并没有存在较大差异.

在前文我们提到, 为了简化计算, 采用了启发式的方法构建 ε 近邻图, 当两样本之间的距离有如下假设 $\|\mathrm{S}_i - \mathrm{S}_j\| < \varepsilon$ 成立时, 则连接 S_i 和 S_j 来构建图, 因而参数 ε 的设定对图构建时的连通性产生较大影响. 图 4 显示了上述 3 种相似度计算方法的图模型和阈值 ε 条件不同的方法的评价结果, 其中实验评价指标采用 $P@N$, 即消费意图句抽取结果排在前 N 个消费意图句的准确率. 可以看到, 我们的方法在候选消费意图微博抽取上的性能尤为显著, 是以上尝试的所有方法中最为有效的. 这也印证了以消费意图表述的相似性计算边权重的方法是按照消费意图文本的特性定义来抽取的, 此方法是非常有效的. 我们从图 4 中可以发现, 相比而言, 基于 Jaccard 相似度的方法要好于基于余弦相似度的方法. 其主要原因在于, 在不同领域人们表达消费意图的方式通常是非常相似的, 因而此方法要好于不区分消费意图表达方式的余弦相似度方法.

从图 4 也可以看到, 随着图上边权重的阈值 ε 的增加, 以上 3 种相似度计算边权重的方法的效果越来越差. 我们分析是由于阈值 ε 的增加导致图的连通性被减弱, 从而影响图上节点向相邻节点的传播, 导致损失了大部分有价值的边, 进而影响到了消费意图句抽取的准确率.

5.2 对基于图排序的消费意图检测评价

目前, 消费意图检测任务通常被看作为有指导的二元分类方法, 通过对文本进行特征选择, 然后通过学习到的分类器对文本作出二元判定. 为了与本章提出的图排序方法进行对比, 以下评价本文提出的基于图排序的消费意图检测方法, 并与传统的有指导和弱指导的方法进行了对比.

本文主要与文献 [3] 中基于二元分类的消费意图检测方法进行了对比, 该方法中主要使用了 3 类特征, 包括: 文本内容特征、微博特有特征和触发词特征, 具体见文献 [3] 中说明. 采用以上描述的特征选择方法将每条微博表示为词向量形式, 在词向量表示时这些特征都属于 "0-1" 布尔值特征. 本文对自动获取的训练语料和人工标注的测试语料进行特征抽取并表示为特征向量, 将人工标注的微博作为测试集部分, 根据训练集数据学习消费意图检测分类器 (采用 SVM 工具包 libsvm-2.82 [12]), 测试集用于评测该分类器的性能.

5.2.1 评价方法

由于测试数据不平衡, 这里仍采用文献 [3] 中的 ROC 曲线 [13] 作为评价指标. 主要原因在于当数据集中的正负样本分布变化时, ROC 曲线能够保持稳定. 通过用 ROC 曲线以及曲线所包围的面积 (AUC) 来评价分类器的可信度.

5.2.2 训练数据标注

目前, 国内外并没有公开发布的用于消费意图检测的测试语料, 本文利用人工标注的语料进行评价. 我们从全部爬取出的微博语料中随机抽取了 10000 条微博交由两名标注人员对数据进行标注, 然后通过计算标注结果的 Kappa 值来验证两名标注人员的标注一致性. 依照上述标注过程, 我们对 10000 条微博句预处理后, 标注得到 186 条具有消费意图的微博 (正例), 以及 7716 条不具有消费意图的微博 (负例), 其中 Kappa 值为 0.667, 说明两名标注者的标注结果比较一致.

5.2.3 对基于图排序的消费意图检测评价

为了与传统的方法进行比较, 本文分别将有指导和弱指导的消费意图检测方法作为对比实验系统.

• 对比系统 1, 基于有指导 (supervised) 的消费意图检测方法

该方法是目前广泛使用的消费意图检测方法. 按照上述人工标注的数据和选取全部特征构建分类器, 采用 5-fold 交叉验证的方法进行本实验.

• 对比系统 2, 基于弱指导 (weakly-supervised) 的消费意图检测方法

该方法是为了和本文提出的基于弱指导的图排序方法进行对比. 在构造训练集时, 我们利用在 4.2.1 小节基于哈希标签自动获得的正例语料共 1991 条微博. 尽管利用哈希标签自动抽取语料是有噪声的, 但统计发现此方法的噪声比较低 (小于 5%), 因而此方法可行. 为了保证训练语料的相对平衡性, 我们人工随机抽取了 4000 条微博, 过滤后剩下 2873 条作为不具有消费意图的微博. 利用全部特征在此训练集上构建分类器, 并在人工标注的结果上进行测试.

• 基于图排序 (graph ranking) 的消费意图检测方法

即本文提出的基于图排序的消费意图检测方法. 我们将利用哈希标签获得的种子节点和人工标注的 1000 条文本作为图中节点, 按照本文提出的图构建方式, 迭代计算人工标注节点的权重值, 并取 top-k 个微博作为具有消费意图的文本, 其余作为负例文本.

实验结果如图 5 所示. 从图 5 中可以看出, 本文提出的图排序方法实验效果最好, 其中 ROC 曲线面积达到了 0.96. 以往的有指导的统计学习方法, 需要人工标注一定规模的训练语料, 而人工标注语料的方法通常费时费力. 本文提出的方法仅需要少量的标注数据, 将无标注数据同时融入到图排序的学习中, 在这一过程中仅依赖于消费意图表示提供的信息, 并将这种信息通过图上邻居节点传递出去, 实验结果表明本文利用用户消费意图表达的特性构建图的方法是有效的, 可以很好地解决用户的消费意图检测问题.

从图 5 中可以看出, 弱指导的方法要好于有指导的方法, ROC 曲线面积值分别是 0.95 和 0.9 (t-test: $p < 0.0005$). 我们分析其中的原因, 主要是由于语料不平衡而且正例过少, 而使得基于弱指导的学习方法要优于基于有指导的方法, 因此有指导方法中的大部分正例会被系统误判为负例. 同时, 基于弱指导的方法有效地解决了训练语料获取的问题, 实验结果显示, 基于弱指导的学习方法是有效的.

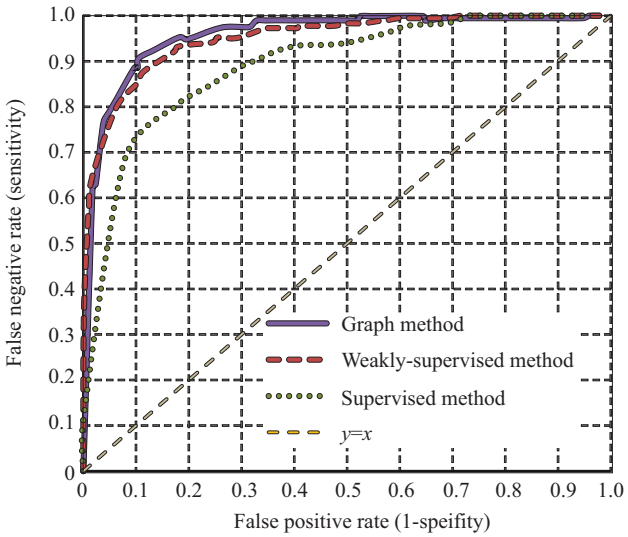

图 5 (网络版彩图) 各消费意图检测方法的对比

Figure 5 (Color online) The comparison among consumption intention detection methods

6 结论与展望

本文提出了一种改进的 RelevenceRank 方法应用于消费意图检测研究, 使得可以适用于没有明显链接关系的消费意图句的抽取中. 具体的, 在构建无向图时, 我们利用少量的、用户自然标注资源自动收集标注数据的方法, 解决了有指导方法需要大量人工标注训练数据的问题, 并且通过实验证明了自动获取到的训练数据质量较高. 然后基于定义的消费意图相似度构建图中边的权重, 因而可以获得大量的消费意图文本, 解决了利用大量未标注数据提升系统性能的问题, 并将有标注数据和未标数据间的相似性融合到图排序的过程中, 有效地实现了社会媒体用户的消费意图自动检测. 实验结果表明, 本文使用的基于图排序的方法对于微博用户的消费意图检测是有效的.

致谢 我们向对本研究工作提供帮助的老师和同学表示感谢.

参考文献

1 Goldberg A B, Fillmore N, Andrzejewski D, et al. May all your wishes come true: a study of wishes and how to recognize them. In: Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Boulder, 2009. 263–271

2 Chen Z, Liu B, Hsu M, et al. Identifying intention posts in discussion forums. In: Proceedings of of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, 2013. 1041–1050

3 Fu B, Liu T. Weakly-supervised consumption intent detection in microblogs. J Comput Inf Syst, 2013, 6: 2423–2431

4 Angluin D, Laird P. Learning from noisy examples. Mach Learn, 1988, 2: 343–370

5 Dai H K, Zhao L, Nie Z, et al. Detecting online commercial intention (OCI). In: Proceedings of the 15th International Conference on World Wide Web, Southampton, 2006. 829–837

6 Davidov D, Tsur O, Rappoport A. Enhanced sentiment learning using twitter hashtags and smileys. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Stroudsburg: Association for Computational Linguistics, 2010. 241–249

7 Page L, Brin S, Motwani R, et al. The PageRank citation ranking: bringing order to the web. Stanford InfoLab, 1999, 9: 1–14

8 Mihalcea R, Tarau P. TextRank: bringing order into texts. In: Proceedings of Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2004. 404–411
9 Yang Y, Bansal N, Dakka W, et al. Query by document. In: Proceedings of the 2nd ACM International Conference on Web Search and Data Mining, Barcelona, 2009. 34–43
10 Sun M. Natural language processing based on naturally annotated web resources. J Chinese Inf Process, 2011, 25: 26–32 [孙茂松. 基于互联网自然标注资源的自然语言处理. 中文信息学报, 2011, 25: 26–32]
11 Che W, Li Z, Liu T. LTP: a Chinese language technology platform. In: Proceedings of the 23rd International Conference Computational Linguistics. New York: ACM Press, 2010. 13–16
12 Chang C C, Lin C J. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Tech (TIST), 2011, 2: 27–37
13 Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning. New York: ACM Press, 2006. 233–240

Detecting consumption intention based on graph ranking in social media

LIU Ting*, FU Bo & CHEN YiHeng

Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin 150001*, China*
*E-mail: tliu@ir.hit.edu.cn

Abstract Consumption intention in a microblog indicates either an imminent or a future purchase, and this indicates a buying intention. Consumption-intention detection aims to determine the willingness of a user to make a purchase. The consumption intention in a microblog should contain two important elements, namely the trigger words and the consumption intention target (the product to be bought). These two elements directly indicate a user's purchase intention, which is an important feature of the user's consumption intention. In this paper, we propose to study the problem of detecting consumption intention in microblogs based on graph ranking. This method can be applied to cases where there is a large amount of data, the amount of labeled data is relatively small, and all data can be involved in the learning process of the graph-ranking algorithm. Experimental results show that the graph-ranking-based method is effective for detecting consumption intention.

Keywords consumption intention, consumption-intention detection, social media, graph ranking, weakly-supervised

LIU Ting was born in 1972. Currently, he is a Professor and Ph.D. supervisor at the School of Computer Science and Technology, Harbin Institute of Technology, and is a senior member of the CCF. His research interests include natural language processing, information retrieval, and social computing.

CHEN YiHeng was born in 1979. He is currently a Lecturer at the Harbin Institute of Technology. His current research interests include information retrieval, social computing, and text clustering.

FU Bo was born in 1983. She is a Ph.D. candidate at the Harbin Institute of Technology. Her research interests include information retrieval and social computing.