doi: 10.12011/1000-6788(2018)07-1853-09 中图法分类号: TP18 文献标志码: A

基于多视角特征融合的移动信息服务模式挖掘

钟学燕 [1,3], 陈国青 [2], 孙磊磊 [2], 张明月 [2], 刘 澜 [1]

(1. 西南交通大学 交通运输与物流学院, 成都 610031; 2. 清华大学 经济管理学院, 北京 100084;
3. 西南石油大学 计算机科学学院, 成都 610500)

摘 要 大量移动手机应用 (Apps) 存在重叠性功能特征, 大量用户评论和多个标签, 给市场机会发现、开发应用集成和应用选择带来困扰. 本文提出基于文本挖掘和相似度网络融合的移动信息服务模式挖掘框架: 基于功能描述文本、用户评论、分类标签分别构建个体相似度网络; 将从不同信息视角得到的相似度网络进行非线性融合; 用聚类验证融合网络有效性, 将其用于发现不同移动信息服务模式. 最后实验抓取 2451 个 Apps, 多视角特征融合方法克服数据水平差异和噪音, 集成互补信息. 融合结果用于聚类, 在归一化互信息和准确率指标上都有显著提升, 准确发现地图导航、火车汽车票、打车专车、公交查询等主流移动信息模式. 研究成果为发掘市场机会和竞争者提供事实证据.

关键词 信息服务模式; 网络融合; 文本聚类; 多视角特征; 移动出行

Mining mobile information service patterns based on multi-view features fusion

ZHONG Xueyan[1,3], CHEN Guoqing[2], SUN Leilei[2], ZHANG Mingyue[2], LIU Lan[1]

(1. School of Transportation and Logistics, Southwest Jiaotong University, Chengdu 610031, China;
2. School of Economics and Management, Tsinghua University, Beijing 100084, China;
3. School of Computer Science, Southwest Petroleum University, Chengdu 610500, China)

Abstract Various mobile applications (Apps) have overlap functional features, a large number of user reviews, and multiple labels, which may cause a difficulty for market opportunity discovery, application integration and selection. This paper proposes a text mining and network fusion analysis framework for finding information service patterns. First, similarity networks are built from three views of function description, user reviews and labels. Then, different similarity networks are nonlinearly fused in an integral manner. Data of 2451 mobile travel applications were crawled, three networks got fusion to form a comprehensive view, overcoming different measurement and noise, taking advantage of complementary information. Fused networks were used for clustering. External evaluation result had significantly been improved including normalized mutual information and accuracy. Mainstream travel service patterns were found, such as map navigation, train and car tickets, taxi car, bus and so on. The research results provide factual evidence for exploration of market opportunities and competitors.

Keywords information service pattern; network fusion; text clustering; multi-view features; mobile travel

收稿日期: 2017-09-14

作者简介: 钟学燕 (1980–), 女, 汉, 四川犍为人, 讲师, 博士研究生, 研究方向: 智能交通与文本挖掘, E-mail: zhongxueyan@126.com; 通信作者: 陈国青, 男, 北京人, 教授, 博士生导师, 博士, 研究方向: 商务智能、模糊逻辑与数据模型, E-mail: chengq@sem.tsinghua.edu.cn.

基金项目: 西南石油大学科研启航计划项目 (2014QHS010); 西南石油大学人文社科专项基金 (2013RW007); 中国博士后科学基金 (2017M620054)

Foundation item: Scientific Research Starting Project of Southwest Petroleum University (2014QHS010); Humanities and Social Sciences Foundation of Southwest Petroleum University (2013RW007); China Postdoctoral Science Foundation (2017M620054)

中文引用格式: 钟学燕, 陈国青, 孙磊磊, 等. 基于多视角特征融合的移动信息服务模式挖掘 [J]. 系统工程理论与实践, 2018, 38(7): 1853–1861.

英文引用格式: Zhong X Y, Chen G Q, Sun L L, et al. Mining mobile information service patterns based on multi-view features fusion[J]. Systems Engineering — Theory & Practice, 2018, 38(7): 1853–1861.

1 引言

移动互联网和智能手机的爆发式增长正在改变用户的信息使用模式. 手机应用市场包含大量 App 下载使用和文本信息, 如功能描述、标签、评价等, 为发现市场结构 [1]、协同推荐 [2,3], 网络口碑营销 [4] 等提供大量不同来源和不同类型数据. 例如在发现主流信息服务模式并进行市场分析过程中, 这些数据从不同视角反映产品属性, 使得如何集成这些异质数据尤为重要. 虽然调查统计分析能发现现有市场主流移动信息服务模式 [5], 但聚类挖掘作为一种无监督学习算法 [6], 可将 App 对象自动划分为若干簇, 使得簇内对象尽可能相似, 簇间对象尽量相异 [7]. 聚类结果和下载量统计的关联可作为发现流行信息服务模式的现实证据.

特征选择和相似性度量是进行聚类、分类、检索等的基础, 多维特征的融合方法是保证数据集成质量的关键 [8]. 虽然多种不同类型数据使得从不同视角衡量对象相似度成为可能, 但由于不同类型数据的异质性和量纲的不一致, 在特征集成过程中存在挑战 [8]. 已有文献引入结构化数据, 文本数据 (功能描述 [9]、评论 [9]、标签 [10])、图像数据作为相似度度量 [11], 通过线性函数 [10]、sigmoid 函数 [10] 和深度学习 [11] 进行特征集成, 在此基础上选择不同方法进行聚类或推荐 [11,12]. 但冷启动使得评论、标签缺失, 导致利用线性函数或 sigmoid 函数进行特征集成加权在实际应用中有局限. 而深度机器学习需要大量的训练集使得方法应用受限. 本文基于网络融合方法结合网络结构特征, 捕捉来自不同数据源共享和互补信息, 消除各相似度矩阵分布不一致性, 克服数据缺失和训练集规模的问题 [8].

因此, 研究以移动出行信息服务模式挖掘为例, 基于手机市场获得的结构和非结构化数据, 主要关注两个关键点: 现有移动出行信息服务主要有哪些模式? 哪些模式更容易被用户接受? 先进的出行信息服务系统 (advanced traveler information system, ATIS) 能实时提供道路交通、公共交通、换乘、气象、停车场等信息, 使出行者在全过程中得到信息支持 [13,14]. 大量移动出行手机应用 (application, App), 如百度地图、滴滴、车来了等移动实时等特性, 使其成为出行主要信息工具, App 数量众多也为用户选择带来麻烦. 研究通过手机应用市场抓取的出行领域相关 App 的多维度文本数据, 提出基于相似度网络融合的移动信息服务模式文本挖掘框架. 有助于挖掘移动出行信息主流服务模式, 找出各服务模式中的 TOP K.

2 总体研究框架

总体研究框架包含四个关键过程, 如图 1. 第一步, 多视角特征获取和预处理是原始数据准备过程. 第二步, 多特征相似度计算是根据不同视角特征的特点选择合适的度量方法得到配对 App 的相似度矩阵. 第三步, 多特征相似度网络融合是根据三个相似网络最近邻信息, 运用文献 [8] 相似度网络融合算法计算最终总体相似度矩阵. 第四步是基于融合后相似度矩阵的聚类并发现模式, 相似度网络融合结果可以提高聚类精度和检索准确率.

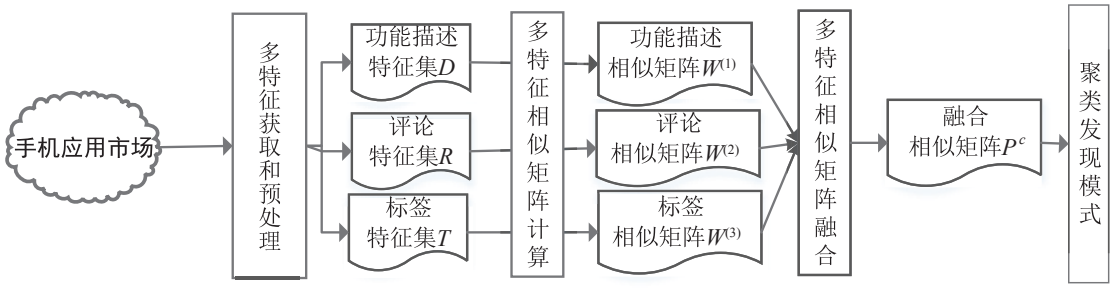


图 1 基于相似度网络融合的移动信息服务模式挖掘框架

3 多特征获取和预处理

3.1 多特征选择

通过编写爬虫程序抓取手机应用市场文本信息, 选择三个维度刻画 App 相似性: 功能描述维度、评论维度、标签维度.

功能描述维度利用功能描述文本中提取的功能特征刻画 App. 功能描述是应用开发商对产品的详细描

述, 包含大量功能特征词项, 可以作为产品功能提取的来源 [9]. 产品描述的准确性将直接影响用户对应用的采纳行为, 开发商会尽力完善并保证信息的准确性和完备性. 一般包括功能、非功能信息描述、版本更新内容. 定义功能描述特征集 $D=\{d_j\}, d_j$ 为第 j 个 App 描述文本, $j\in\{1,2,\cdots,n\}, n$ 为获取 App 总数.

评论维度是利用用户的评价信息提取功能特征刻画 App. 评论包含用户对不同产品部件或属性意见, 可从评论中提取产品特征 [15]. 如评论 "常常出现定位失败!!" 中可提取 "定位" 功能特征. 定义评论特征集 $R=\{r_j\}, r_j$ 为第 j 个 App 的获取评论的整合文本, $j\in\{1,2,\cdots,n\}$, n 为获取 App 总数. 数据处理中把每个 App 抓取所有评论整合为一个文本.

标签维度是根据开发商定义的产品标签刻画 App. 标签是开发商对产品特征的精炼表达, 如百度地图的标签 {导航、地图、定位}. 标签能传递内容信息和识别资源属性, 可用于聚类和推荐 [16]. 定义抓取的标签数据集 $T=\{t_j\}$, t_j 为第 j 个 App 的标签文本集合, $j\in\{1,2,\cdots,n\}$, n 为获取 App 总数.

最后, 一个 App 的特征刻画由一个三元组构成 $x_j=\{d_j,r_j,t_j\}$. 出行信息服务 App 数据集定义为 $X=\{x_j\}, j\in\{1,2,\cdots,n\}$, n 为获取 App 总数.

3.2 数据预处理

功能描述和评论信息是由多个段落构成的文档, 需要进行文本预处理, 分词, 词性标注, 特征抽取及停用词处理. 文本预处理是清除特殊符号, 包括 "@, (), [ ], 数字, http" 等. 分词是把中文语句划分成短语词汇, 采用结巴分词包, 增加用户自定义词典 (如凯立德、电子狗、高德、易烊千玺等). 词性标记是指为分词结果中的每个词标注一个正确词性的程序, 也即确定每个词是名词、动词、形容词或其他词性的过程, 研究采用斯坦福大学的词性标注工具. 特征抽取是抽取功能描述和评论信息中的名词 [17] 和动词 [9] 作为最后的特征数据集, 停用词处理是剔除特征数据中存在的大量功能无关词汇, 如 "软件、程序、互联网" 等.

标签维度中, 每个 App 标签集合由 m 个词项标签组成, $m\geq 0$, 不需要分词词性标注等处理.

4 多特征相似度计算

4.1 功能描述特征相似度计算

相似度计算的输入数据为预处理的功能描述特征文本. 采用向量空间模型进行文本表示, 选择 TF-IDF 模型计算特征权重, 相似度计算采用两个向量的夹角余弦值来评估相似度.

向量空间模型 [18] 中每个文档被表示成一个功能词项向量 $\boldsymbol{d}$. 采用 TF-IDF 来确定词项权重, 每个文档被表示成 $(tf_1\log(n/df_1), tf_2\log(n/df_2), \cdots, tf_m\log(n/df_m))$, 其中, tf_i 是指第 i 个词项在文档中的频率, df_i 是包含第 i 个词项的文档数. 考虑不同文档的长度, 每个文档向量需要规范化到单位长度 ($||d_{tf-idf}||=1$).

提取词频超过给定阈值的特征, 基于余弦相似度计算配对 App 向量之间相似度. 相比距离度量, 余弦相似度更多的是从方向上区分差异, 而非距离或长度上. 普遍用于高维空间, 特别是信息检索和文本挖掘. 给定文档向量 $\boldsymbol{d}_i$ 和 $\boldsymbol{d}_j$, 定义余弦相似度 $\cos(\theta)$ 为式 (1):

$$\cos(\theta)=\frac{\boldsymbol{d}_i.\boldsymbol{d}_j}{||\boldsymbol{d}_i||||\boldsymbol{d}_j||}. \tag{1}$$

当文档词项权重相同时值为 1, 没有任何相同词项权重时值为 0. 计算输出为功能相似度矩阵 $\boldsymbol{W}^{(1)}$.

4.2 评论特征的相似度矩阵计算

评论维度相似性计算与功能描述的预处理一致. 文本表示模型选择向量空间模型. 不同于功能描述文本的精炼和长度基本一致, 评论文本的相似性计算需要考虑冷启动及评论中大量无关信息影响 (例如情感评价词: 好用、垃圾、推荐等). 研究把抓取的一个 App 所有评论整合为一个文档, 通过分词和词性标注获取评论的名词和动词作为特征集合. 通过 TF-IDF 计算余弦相似度获得评论相似度矩阵. 计算评论相似度的算法与计算功能描述的算法一致, 不同的是输入数据为评论特征集. 输出为评论相似度矩阵 $\boldsymbol{W}^{(2)}$.

4.3 标签特征相似性计算

标签相似度计算是实现词项集合相似性的度量. 一种方法是基于中国知网 Hownet 网络计算词义的相似度, 但存在词义网概念不完全包含所有标签的问题. 本文基于 Jaccard 相似度 [19] 计算两个标签集合相似度. 给定两个 App x_i 和 x_j 的标签集合 t_i 和 t_j, 其交集大小与并集大小的比值可以作为相似度度量的很好

选择. Jaccard 相似度定义为式 (2)

$$\text{Jaccard}(t_i t_j) = \frac{|t_i \cap t_j|}{|t_i \cup t_j|}. \tag{2}$$

式 (2) 中, $t_i = \{tag_{i1}, tag_{i2}, \cdots, tag_{im}\}, t_j = \{tag_{j1}, tag_{j2}, \cdots, tag_{jn}\}, m \geq 0, n \geq 0$, 当集合 t_i, t_j 都为空相似度为 0. 例如高德地图标签集合 $t_1 =$ {周边信息, 男性, 路况, 导航, 地图, 地图导航}, 百度地图标签集合为 $t_2 =$ {定位, 地图导航, 出行, 周边信息, 导航, 地图, 街景}. Jaccard(t_1, t_2) 等于 4/9. 输出为标签相似度矩阵 $\boldsymbol{W}^{(3)}$.

5 多视角特征相似度融合

功能、评论、标签三个特征维度融合需克服至少两个计算挑战: 第一是不同特征维度数据水平差异, 数据收集偏差和噪音, 第二是综合不同类型数据提供的互补信息. 线性权重加权方法的权重选择依赖于专家经验和实验, 不能克服不同特征量纲及分布不一致. 使用非线性融合方法集成不同相似度网络在解决多维特征集成上具有很强优势 [8]. 采用文献 [8] 方法, 定义相似度网络为图 $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E}), \boldsymbol{V} = \boldsymbol{A}\{x_j\}, j \in \{1, 2, \cdots, n\}$, n 为 App 数目, 边 E 代表 App 之间相似度. 边权重表示成 $n * n$ 的相似度矩阵 $\boldsymbol{W}$, $\boldsymbol{W}(i, j)$ 可根据数据特点选择不同相似度度量方式. 本文三个相似度矩阵定义为: 功能描述相似矩阵 $\boldsymbol{W}^{(1)}$, 评论相似矩阵 $\boldsymbol{W}^{(2)}$, 标签相似矩阵 $\boldsymbol{W}^{(3)}$:

相似度网络融合 (similarity network fusion, SNF) 过程通过不断迭代融合各子图的最近邻信息. 功能描述相似矩阵 $\boldsymbol{W}^{(2)}$ 是一个紧密矩阵, 得益于开发商为吸引用户对 App 详细描述. 评论矩阵 $\boldsymbol{W}^{(2)}$ 和标签矩阵 $\boldsymbol{W}^{(3)}$ 由于部分数据缺失是一个稀疏矩阵. 为融合三个相似度矩阵, SNF 方法在相似网络节点集合上定义了一个完全核和一个稀疏核. 完全核是一个正态权重矩阵 $\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{W}$, 矩阵 $\boldsymbol{D}$ 是对角矩阵 $\boldsymbol{D}(i, i) = \sum_j \boldsymbol{W}(i, j)$, $\sum_j \boldsymbol{P}(i, j) = 1$. 然而, 这个规范化可能由于 $\boldsymbol{W}$ 的自相似问题而受限于数字不稳定性. 一种方式是进行一种更好的规范化, 如下所式 (3). 这个规范化能消除对角对象自相似水平影响, 依旧保证 $\sum_j \boldsymbol{P}(i, j) = 1$.

$$\boldsymbol{P}(i, j) = \begin{cases} \dfrac{\boldsymbol{W}(i, j)}{2\sum_{k \neq i} \boldsymbol{W}(i, k)}, & j \neq i, \\ \dfrac{1}{2}, & j = i. \end{cases} \tag{3}$$

让 $\boldsymbol{N}_i$ 代表 x_i 在图 $\boldsymbol{G}$ 的邻居数, 给定一个图 $\boldsymbol{G}$, 用 K 近邻度量局部吸引力, 如式 (4). 这个操作使得非近邻的配对点相似度为零.

$$\boldsymbol{S}(i, j) = \begin{cases} \dfrac{\boldsymbol{W}(i, j)}{2\sum_{k \in N^i} \boldsymbol{W}(i, k)}, & j \in N^i, \\ 0, & \text{otherwise}. \end{cases} \tag{4}$$

注意到 $\boldsymbol{P}$ 涵盖给定维度每个 App 之间所有的相似信息, 而 $\boldsymbol{S}$ 只是编码了 K 个最相邻 App 信息. 融合算法从 $\boldsymbol{P}$ 作为初始状态, 使用 $\boldsymbol{S}$ 作为核矩阵, 整个融合计算过程考虑图局部结构和计算性能.

设 $\boldsymbol{P}^{(1)}_{t=0} = \boldsymbol{P}^{(1)}$, $\boldsymbol{P}^{(2)}_{t=0} = \boldsymbol{P}^{(2)}$, $\boldsymbol{P}^{(3)}_{t=0} = \boldsymbol{P}^{(3)}$ 代表三个初始状态 ($t = 0$) 矩阵, SNF 方法关键步骤就是根据每个特征维度不断迭代更新相似度矩阵, 如下式 (5)~(7) 所示:

$$\boldsymbol{P}^{(1)}_{t+1} = \boldsymbol{S}^{(1)} \times \left(\frac{\boldsymbol{P}^{(2)}_t + \boldsymbol{P}^{(3)}_t}{2} \right) \times \left(\boldsymbol{S}^{(2)}\right)^{\mathrm{T}}, \tag{5}$$

$$\boldsymbol{P}^{(2)}_{t+1} = \boldsymbol{S}^{(2)} \times \left(\frac{\boldsymbol{P}^{(1)}_t + \boldsymbol{P}^{(3)}_t}{2} \right) \times \left(\boldsymbol{S}^{(2)}\right)^{\mathrm{T}}, \tag{6}$$

$$\boldsymbol{P}^{(3)}_{t+1} = \boldsymbol{S}^{(3)} \times \left(\frac{\boldsymbol{P}^{(1)}_t + \boldsymbol{P}^{(2)}_t}{2} \right) \times \left(\boldsymbol{S}^{(3)}\right)^{\mathrm{T}}. \tag{7}$$

式 (5)~(7) 的迭代过程体现了多视角相似度网络融合方法核心思想. 其理论假设为如果两个个体具有相同的邻居, 它们很可能属于同一个簇; 另外如果两个个体在一个特征视角相似性不高, 但是在另一个特征视角具有很高的相似性, 融合算法能保证这个相似性能通过融合过程进行传播. 由于 $\boldsymbol{S}$ 是 $\boldsymbol{P}$ 的一个 K 邻接图, 能减少对象个体之间的一些噪音. 在每次迭代后, 对 $\boldsymbol{P}^{(1)}_{t+1}$, $\boldsymbol{P}^{(2)}_{t+1}$, $\boldsymbol{P}^{(3)}_{t+1}$ 都用式 (3) 进行了规范化, 这样

保证每个个体在多次迭代后总是和自己最相似, 以及保证融合结果在分类和聚类问题中的排序和重要性. 在 t 次迭代后总的状态矩阵为式 (8):

$$\boldsymbol{P}^{(c)} = \frac{\boldsymbol{P}_t^{(1)} + \boldsymbol{P}_t^{(2)} + \boldsymbol{P}_t^{(3)}}{3}. \tag{8}$$

最后的 $\boldsymbol{P}^{(c)}$ 就是多维相似度融合矩阵, 可用于检索、聚类、分类, 本文聚焦聚类和检索. 如果存在更多特征, $m > 3$, 则式 (5)∼(7) 可拓展为式 (9), 最后 t 次迭代的融合矩阵 $\boldsymbol{P}^{(c)}$ 为式 (10):

$$\boldsymbol{P}^{(\nu)} = \boldsymbol{S}^{(\nu)} + \frac{\sum_{k\neq\nu}\boldsymbol{P}^{(k)}}{m-1} + \left(\boldsymbol{S}^{(\nu)}\right)^{\mathrm{T}}, \quad \nu = 1, 2, \cdots, m, \tag{9}$$

$$\boldsymbol{P}^{(c)} = \frac{\sum_{k=1}^{v} P_t^k}{m}, \quad \nu = 1, 2, \cdots, m. \tag{10}$$

相似度网络融合方法在样本空间而非特征空间集成数据. 算法的输入可以是配对相似度, 针对不同数据 (离散、连续), 可以采用不同相似度度量方法. 相似度网络迭代过程考虑不同特征的 K 最近邻信息, 对于量纲不一致的特征融合具有较强适用性. 算法敏感性测试结果显示相似度融合算法对于不同的超参数具有鲁棒性 [8]. 该算法劣势是进行大样本迭代需要大量计算资源.

6 基于多视角特征相似度融合矩阵的聚类

6.1 聚类方法选择

聚类方法是一种无监督学习方法, 旨在将数据对象集合分成几个子集, 使得同一个子集中数据对象尽量相似, 不同子集之间数据对象尽量相异. 在获得的融合相似度矩阵 $\boldsymbol{P}^{(c)}$ 基础上, 选择聚类方法获得聚类簇. 聚类方法层面, 本研究采用自上而下的重复二分 rbr, 直接划分 direct 方法, 自下而上的层次聚类 agglo, K-medoids[18], 密度峰值 (density peak clustering, DPC)[20] 进行聚类. rbr、direct, agglo 聚类采用 CLUTO 工具 (http://glaros.dtc.umn.edu/gkhome/views/cluto) 实现. K-medoids, DPC 采用 MATLAB 程序实现.

rbr 聚类是自顶向下聚类, 初始将所有对象作为一个类别, 通过执行一系列 $k-1$ 次重复二分得 k 个聚类. direct 与 rbr 相比, 同时直接计算出 k 个聚类, 但 k 大于 20, rbr 是更加适用. agglo 聚类是自底向上聚类, 初始将每个对象作为一个类别, 然后将这些类别按一定聚合条件聚合成较大类别, 直到得到 k 个聚类. rbr、direct, agglo 内部评价函数选择 $H_2 = \frac{I_2}{E_1}$ 进行全局优化 [7], I_2 是最大化簇内相似度, γ_1 是最小化簇间相似度.

K-medoids 聚类通过最大化对象与中心点的相似度来进行聚类 [21]. 密度峰值 DPC[20] 算法假设是类簇中心被具有较低局部密度的邻居点包围, 且与具有更高密度的任何点有相对较大距离.

6.2 聚类方法的外部评价指标

聚类算法外部效果评价根据带有真实类别标签的数据对象. 评价指标选取归一化互信息 (normalized mutual information, NMI) 和准确率 *Accuracy*. 真实类标号来自 5 位交通领域专家, 在给定类标签基础上对下载量大于 1 百万且描述清晰 App 进行类别标注, 包含 1100 个 App. 根据投票数确定最后类别. 给定类标签包括 13 个类: 打车租车、交通信息、旅游景点、票务、公交地铁、航班机票、停车、地图导航、物流、电动车、驾驶辅助、自行车、其他.

归一化互信息是一种基于信息论的评价标准, 通过计算聚类结果 $\boldsymbol{c}$ 与真实类标号 $\hat{c}$ 之间的互信息 NMI 来评价聚类结果与真实类别标号的一致性 [22], NMI 值越高, 说明聚类结果越好. 定义如下式 (11).

$$NMI = \frac{I(\boldsymbol{c}, \hat{c})}{\sqrt{H(\boldsymbol{c})}\sqrt{H(\hat{c})}}. \tag{11}$$

$I(\boldsymbol{c}, \hat{c})$ 表示聚类结果 $\boldsymbol{c}$ 和真实类别标号 $\hat{c}$ 之间互信息. $H(\cdot)$ 表示单个类别向量的信息熵.

准确率 *Accuracy* 是一种更为直观的聚类性能的外部评价指标, *Accuracy* 越高, 聚类效果越好. 定义如下式 (12):

$$Accuracy = \frac{\sum_{i=1}^{\boldsymbol{N}} \boldsymbol{\delta}(\hat{c}_i, \boldsymbol{map}(\boldsymbol{c}_i))}{\boldsymbol{N}}. \tag{12}$$

其中 $\hat{c}_i$ 是数据对象 i 的真实类别标号, c_i 是聚类结果. 函数 $\delta(x, y)$ 用于比较两个变量的一致性, 如果 $x = y$, 则 $\delta(x, y) = 1$; 否则 $\delta(x, y) = 0$. $map(\cdot)$ 是将聚类结果的类别标号映射到真实类别标号. $map(\cdot)$. 首先找到

聚类结果中属于同一类的所有数据对象, 然后计算它们的真实类别标号频次, 并根据与真实类别标号的匹配赋以真实类别标号.

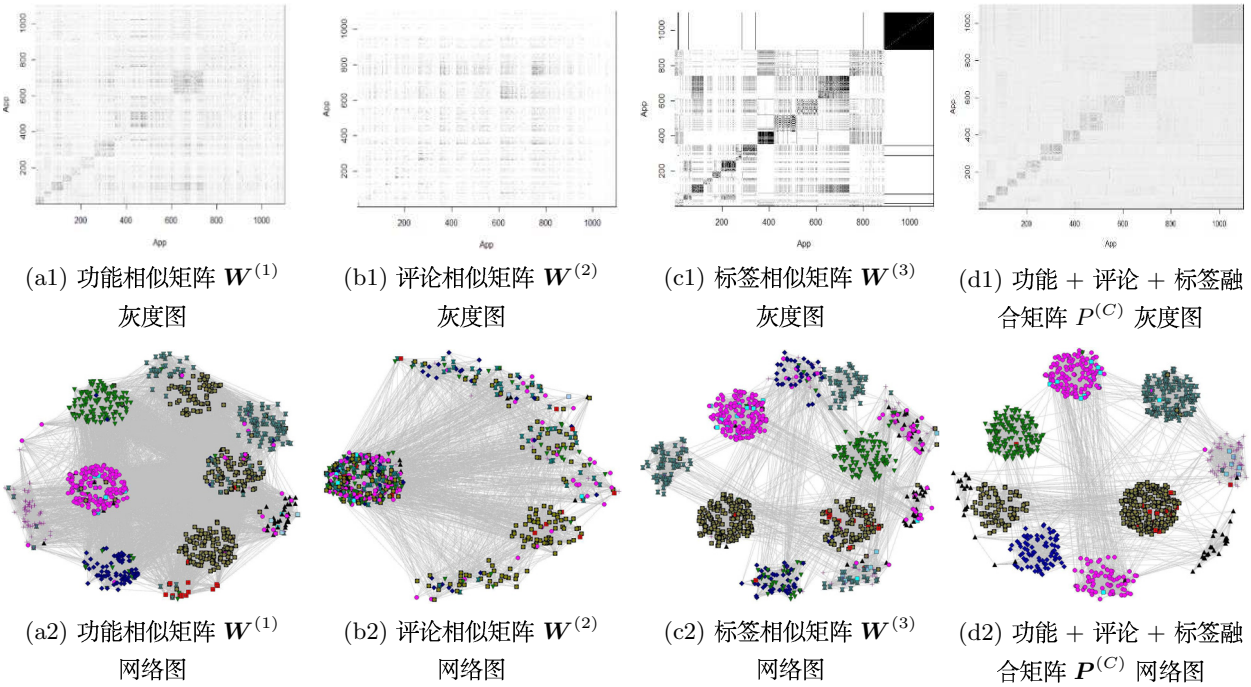
7 实验及结果

7.1 数据描述性统计

360 手机助手是国内最大手机应用分发平台, 本文选择该平台的 “地图旅游” 类 App (http://zhushou.360.cn/list/index/cid/102231/). 在 2017 年 5 月抓取该类所有 App 功能描述、评论、标签及下载量等信息. 共抓取 2451 个 App, 总评论条数为 116199, 单个 App 的最多评论条数为 4215, 最小数量 0, 平均数量 77 条. 评论数量 2000 以上有 15 个 App, 1000~2000 有 17 个, 500~1000 有 30, 100~500 有 87, 100 以下有 1358. 有标签 App 数量有 1761 个, 平均标签数为 3.

7.2 基于单一特征矩阵与网络融合矩阵的聚类算法效果对比

为展现基于单一特征及多特征融合构建相似度对聚类贡献, 基于功能描述相似矩阵 $\boldsymbol{W}^{(1)}$、评论相似矩阵 $\boldsymbol{W}^{(2)}$、标签相似矩阵 $\boldsymbol{W}^{(3)}$, 多视角融合矩阵 $\boldsymbol{P}^{(C)}$, 采用同样的 rbr 聚类算法进行聚类. 利用灰度图和网络图展示 1100 个评估 App 聚类效果, 如图 2. 灰度图中聚类用灰度水平函数进行编码 [8], 按照 rbr 聚类的类标签升序排列, 排序后 App 按类聚集, 颜色越深代表 App 之间相似度越高. 网络图用 Netdraw 绘制, 网络节点表示各个 App, 专家标注的同类 App 节点具有相同的形状和颜色. 边为基于相似度矩阵生成的最近邻矩阵, 每个个体最近邻数取 10, 其相似度值保留, 其余置 0. 这样可以通过网络图展示各特征相似度网络对类簇划分的贡献程度.



(a1) 功能相似矩阵 $\boldsymbol{W}^{(1)}$ 灰度图　(b1) 评论相似矩阵 $\boldsymbol{W}^{(2)}$ 灰度图　(c1) 标签相似矩阵 $\boldsymbol{W}^{(3)}$ 灰度图　(d1) 功能 + 评论 + 标签融合矩阵 $P^{(C)}$ 灰度图

(a2) 功能相似矩阵 $\boldsymbol{W}^{(1)}$ 网络图　(b2) 评论相似矩阵 $\boldsymbol{W}^{(2)}$ 网络图　(c2) 标签相似矩阵 $\boldsymbol{W}^{(3)}$ 网络图　(d2) 功能 + 评论 + 标签融合矩阵 $\boldsymbol{P}^{(C)}$ 网络图

图 2　基于单一特征矩阵与网络融合矩阵的划分聚类算法效果对比

基于功能描述 $\boldsymbol{W}^{(1)}$ 的聚类效果较好, 灰度图 (图 2-a1) 显示对角线聚类子簇相对清晰, 存在个别子簇不清晰, 分析原因可能是个别子簇共有特征不明晰. 网络图 (图 2-a2) 各子簇零星分布一些其他类别的 App, 结果验证功能描述能较好刻画产品属性.

基于评论 $\boldsymbol{W}^{(2)}$ 的聚类效果一般, 但也存在大类的聚集. 分析原因可能由于评论数量差异较大, 灰度图 (图 2-b1) 对角线子簇聚类不是特别明显. 但是网络图 (图 2-b2) 中部分子簇聚类较好, 说明 App 评论数量较多的情况下有助于发现相似 App.

基于标签 $\boldsymbol{W}^{(3)}$ 聚类, 标签为主观设定, 大类划分相对准确, 集成得到较好聚类. 灰度图 (图 2-c1) 可以看到对角线聚类子簇很清晰, 但由于部分 App 没有标签, 故存在空白. 网络图 (图 2-c2) 中各聚类簇边界明

晰, 只有少许子簇包含多类 App, 可能原因是用户标签定义和专家打标签的主观划分差异.

基于融合矩阵 $\boldsymbol{P}^{(C)}$ 的聚类融合功能描述、评论、标签三个视角特征, 灰度图 (图 2-d1) 显示对角线聚类子簇非常清晰, 网络图 (2-d2) 显示各子簇划分包含极少的其他类别 App. 四个相似度矩阵聚类效果的验证了特征融合可以提升聚类效果, 集成各视角互补信息.

7.3 基于网络融合矩阵的不同聚类算法效果对比

为对比网络融合相似度矩阵 $P^{(c)}$ 在不同聚类方法上表现, 实验选择三个对比相似矩阵: 基于 LDA 话题模型的功能描述相似矩阵 $\boldsymbol{W}^{LDA}$, 基于 TF-IDF 向量空间模型的功能描述相似矩阵 $\boldsymbol{W}^{(1)}$, 基于功能描述 $\boldsymbol{W}^{(1)}$ 和标签融合 $\boldsymbol{W}^{(3)}$ 的融合矩阵 $\boldsymbol{P}^{(1+3)}$. 其中 $\boldsymbol{W}^{LDA}$ 是采用 LDA 话题模型 [23] 对进行文本建模, 采用 JS (Jensen-Shannon) 距离计算配对 App 话题概率分布向量之间的相似性. 分别利用 rbr, direct, agglo, K-medoids, DPC 方法, 在聚类数目 k 为 13 和 20 下进行聚类, 采用 NMI 和 Accuracy 两个外部评价指标进行聚类效果评估, 聚类数目 k 为 13 评估结果如表 1, 聚类数目 k 为 20 评估结果如表 2.

从表 1 和表 2 的聚类评估指标来看, $\boldsymbol{W}^{LDA}$ 相似度矩阵在各种聚类算法上表现都处于劣势, 分析原因 LDA 模型根据离散数据集计算文档在主题上分布, 由于功能描述文本较短, 因此结果不够理想. 基于 TF-IDF 向量空间模型的功能描述矩阵 $\boldsymbol{W}^{(1)}$ 包含大量特征信息, 例如属性、子功能、子功能特征等, 在相似度网络融合中贡献较大. $\boldsymbol{P}^{(1+3)}$ 融合标签和功能两部分准确信息, 因此在部分聚类算法上表现较好, 如 agglo, K-medoids. 基于不同矩阵聚类效果的差异显著性检验, 基于融合矩阵 $P^{(c)}$ 与其他矩阵的聚类效果差异在 0.05 水平和 0.01 水平都显著, 验证融合矩阵能有效提升聚类效果, 可用于信息检索与聚类.

表 1　kway-13 不同算法聚类评估结果表

相似度矩阵	kway-13 NMI 对比结果					kway-13 ACC 对比结果				
	rbr	direct	agglo	K-medoids	DPC	rbr	direct	agglo	K-medoids	DPC
$\boldsymbol{P}^{C}$ (功能 + 评论 + 标签)	**0.86***	**0.81***	0.68	0.68	**0.50***	**0.96***	**0.94***	**0.86***	0.83	**0.64***
$\boldsymbol{P}^{(1+3)}$ (功能 + 标签)	0.68	0.75	**0.69***	**0.74***	0.48	0.84	0.91	0.85	**0.89***	0.61
$\boldsymbol{W}^{(2)}$ (功能)	0.69	0.65	0.55	0.44	0.25	0.88	0.85	0.78	0.65	0.39
$\boldsymbol{W}^{\mathbf{LDA}}$ (功能 LDA)	0.14	0.14	0.15	0.16	0.07	0.41	0.42	0.42	0.42	0.31

表 2　kway-20 不同算法聚类评估结果表

相似度矩阵	kway-20NMI 对比结果					kway-20 ACC 对比结果				
	rbr	direct	agglo	K-medoids	DPC	rbr	direct	agglo	K-medoids	DPC
$\boldsymbol{P}^{C}$ (功能 + 评论 + 标签)	**0.80***	**0.76***	**0.65***	**0.74***	0.46	**0.96***	**0.95***	**0.87***	**0.93***	**0.64***
$\boldsymbol{P}^{(1+3)}$ (功能 + 标签)	0.69	0.73	0.65	0.73	**0.49***	0.87	0.92	0.85	0.92	0.61
$\boldsymbol{W}^{(2)}$ (功能)	0.68	0.68	0.58	0.44	0.27	0.88	0.88	0.82	0.67	0.41
$\boldsymbol{W}^{\mathbf{LDA}}$ (功能 LDA)	0.17	0.17	0.17	0.17	0.09	0.43	0.44	0.43	0.45	0.36

7.4 基于融合矩阵聚类结果的模式发现和检索

360 手机助手的 “出行旅游类” 包含大量的 App, 研究实验结果是通过准确定义 App 之间相似性, 聚类发现主流信息服务模式. 从表 1 可看出, 聚类评估效果最好是 rbr 划分聚类, 因此选取该聚类获得的 13 个聚类簇命名和统计, 簇命名来自高频特征词, 每个簇 App 数量, 下载量汇总及簇内下载排行如表 3 所示.

从表 3 可以看出, 地图导航类服务模式无论 App 数量和下载量都是最多的, 高频特征词包括地图、导航、离线、语音、记录等, 公众在出行前和出行中普遍进行出行路径和行程规划, 是这类 App 下载较多的原因. 火车票汽车票类主要实现票务预定、购买、时刻表查询等, 极大方便中短途出行. 打车专车类下载量也比较大, 近两年共享经济发展迅猛, 滴滴出行、神州专车、首汽约车、一号专车抢占大量市场份额, 极大方便私人汽车出行. 公交查询类包括覆盖全国范围公交查询的 8684 公交、掌上公交、车来了等, 也有大量专门服务于特定城市的 App, 公交出行作为政府大力提倡的出行方式, 其服务提供方政府和商业企业并存. 航班机票类包括综合航班信息提供和各航空公司服务提供, 高频特征有机票、航班、酒店、航空、机场等. 汽车导航类服务更加精准, 主要针对小汽车出行的导航、加油、行车记录、车辆服务等. 停车类服务提供停车、车位、停

车场查询等功能. 拼车顺风车主要搭建共享出行平台. 地图街景类主要提供地图服务, 综合服务性比地图导航类差. 租车类主要提供汽车租赁业务. 定位 GPS 类主要依托 GPS 进行定位服务展开. 地铁类服务主要提供地铁线路查询. 还有一个簇类标签不明晰, 可能原因是网络融合算法最近邻数目选择影响.

表 3 基于 rbr 聚类结果的簇内下载 TOP5 及下载量统计

聚类簇标签	App 数量	下载量汇总	簇内下载排行 TOP5
地图导航	324	1, 395, 453, 693	高德地图、谷歌地图、高德导航、图吧导航、天翼导航
火车票汽车票	165	236, 388, 219	盛名时刻表、路路通、铁路 12306、智行火车票、高铁管家
打车专车	160	149, 270, 595	滴滴出行、神州专车、首汽约车、一号专车、嘟嘟巴士
公交查询	250	87, 483, 511	8684 公交、掌上公交、车来了、爱帮公交、酷米客
航班机票	86	64, 158, 074	航班管家、飞常准、春秋航空、航旅纵横、中国国航
汽车导航	163	59, 917, 755	悠悠导航、Google 地球、GPSTest、汽车 FM、GPSStatus
停车	79	40, 757, 278	贴条地图、宜停车、停车百事通、0 元停车、一点停
拼车顺风车	82	29, 715, 620	嘀嗒拼车、易到、天天用车、熊猫出行、51 用车
地图街景	164	23, 312, 095	Google 地图街景、奥维互动地图、搜搜地图、旅行离线地图、微话地图
租车	86	11, 047, 924	神州租车、宝驾出行、一嗨租车、快快优车、租租车
定位 GPS	166	6, 773, 070	GPS 工具箱、导航小蜜、车载导航、GPS 实景地图、导航地图
地铁	85	5, 715, 534	8684 地铁、地铁通广州地铁地铁终结者深圳地铁
类标签不明	641	2, 148, 426	标签不足, 评论较少, 类簇特征不明显
总计	2451	2, 112, 141, 794	

如果增加聚类数目, 可发现更多细分信息服务模式, 如共享单车, 充电汽车, 家人定位等. 通过聚类发现的这些移动信息服务模式, 有助于开发者准确识别同类竞争者, 平台管理者基于 App 相似性进行更准确推荐, 交通管理者识别各类交通信息资源需求状况, 为制定数据共享政策提供依据.

8 结论和展望

研究采用多维特征定义产品相似度, 在样本空间而非特征空间集成多维特征数据, 进行了不同移动信息服务模式的挖掘. 首先选择功能、评论、标签三个维度刻画产品相似性, 弥补了单个维度信息缺失和信息量不足. 功能描述信息作为供应商产品简介, 极少存在信息缺失, 是相似度度量的主要维度. 评论信息由于冷启动, 存在数据稀疏, 但是对于拥有大量评论集的产品, 评论的引入提高了相似度度量的准确性. 标签信息作为用户自定义类别信息, 有助于进行大类的划分. 各维度特征的引入增强了对产品的综合刻画, 可根据各特征属性 (离散、连续) 选用不同的相似度度量方式, 使得方法的可扩展性强. 其次, 相似度网络融合方法能够通过非线性方式融合多维特征, 避免因为线性权重选择和不同特征数据量纲的不一致带来的总体相似度偏差. 相似度融合不断迭代融合各维度相似度矩阵的 K 近邻信息, 对相似度度量的噪音具有鲁棒性, 有效克服了数据异质性问题. 进一步研究需考虑数据分布情况选择合适 K 值构建稀疏核矩阵 S, 探讨各参数对算法的影响. 最后, 基于融合结果的不同聚类方法在归一化互信息和准确率指标上提升聚类效果, 验证了融合方法可提高聚类精度和簇内 topK 查找准确率, 未来还要考虑更多聚类算法适用性.

下一步研究将关注于三个方向: 1) 产品相似性度量的维度未来可考虑增加图片、视频等更多维度, 更加全面定义相似性; 2) 进一步探讨相似性网络分布、融合参数对融合结果的影响, 虽然融合矩阵对选定聚类算法都有提升, 但目前还没有一种聚类方法能够有效处理所有类型的类结构和属性; 3) 针对发现的信息服务模式自动提取服务特征, 结合用户标签、评论和已安装手机应用, 进行个性化推荐.

参考文献

[1] Oded N, Ronen F, Jacob G, et al. Mine your own business: Market-structure surveillance through text mining[J]. Marketing Science, 2012, 31(3): 521–543.

[2] 王伟, 王洪伟, 孟园. 协同过滤推荐算法研究: 考虑在线评论情感倾向 [J]. 系统工程理论与实践, 2014, 34(12): 3238–3249.
Wang W, Wang H W, Meng Y. The collaborative filtering analysis recommendation based on sentiment of online

reviews[J]. Systems Engineering — Theory & Practice, 2014, 34(12): 3238–3249.
[3] 李浩君, 张广, 王万良, 等. 基于多维特征差异的个性化学习资源推荐方法 [J]. 系统工程理论与实践, 2017, 37(11): 2995–3005.
Li H J, Zhang G, Wang W L, et al. The method of personalized learning materials recommendation based on multidimensional feature difference[J]. Systems Engineering — Theory & Practice, 2017, 37(11): 2995–3005.
[4] 王伟, 王洪伟. 特征观点对购买意愿的影响: 在线评论的情感分析方法 [J]. 系统工程理论与实践, 2016, 36(1): 63–76.
Wang W, Wang H W. The influence of aspect-based opinions on user's purchase intention using sentiment analysis of online reviews[J]. Systems Engineering — Theory & Practice, 2016, 36(1): 63–76.
[5] Jia Z Y, Li D, He F. Analysis and reviews on tourism and travel mobile apps of China[C]// 6th International Conference on Electronics, Mechanics, Culture and Medicine, 2016(45): 62–66.
[6] 周开乐, 杨善林, 丁帅, 等. 聚类有效性研究综述 [J]. 系统工程理论与实践, 2014, 34(9): 2417–2431.
Zhou K L, Yang S L, Ding S, et al. On cluster validation[J]. Systems Engineering — Theory & Practice, 2014, 34(9): 2417–2431.
[7] Zhao Y, George K, Usama F. Hierarchical clustering algorithms for document datasets[J]. Data Mining and Knowledge Discovery, 2005, 10(2): 141–168.
[8] Wang B, Mezlini A M, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale[J]. Nature Methods, 2014, 11(3): 333–337.
[9] Xu X, Dutta K, Datta A. Functionality-based mobile app recommendation by identifying aspects from user reviews[C]// Thirty Fifth International Conference on Information Systems, 2014.
[10] 顾晓雪, 章成志. 结合内容和标签的 Web 文本聚类研究 [J]. 现代图书情报技术, 2014(11): 45–52.
Gu X X, Zhang C Z. Using content and tags for web text clustering[J]. New Technology of Library and Information Service, 2014(11): 45–52.
[11] Zhang F, Yuan N J, Lian D, et al. Collaborative knowledge base embedding for recommender systems[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 353–362.
[12] Bauman K, Liu B, Tuzhilin A. Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews[C]// The ACM SIGKDD International Conference, 2017: 717–725.
[13] Natvig M K, Vennesland A. Flexible organisation of multimodal travel information services[J]. Intelligent Transport Systems, IET, 2010, 4(4): 401–412.
[14] 赵祥模, 惠飞, 史昕, 等. 泛在交通信息服务系统的概念、架构与关键技术 [J]. 交通运输工程学报, 2014(4): 109–119.
Zhao X M, Hui F, Shi X, et al. Concept, architecture and challenging technologies of ubiquitous traffic information service system[J]. Journal of Traffic and Transportation Engineering, 2014(4): 109–119.
[15] Liu B. Sentiment analysis and opinion mining[J]. Synthesis Lectures on Human Language Technologies, 2012, 5(1): 1–167.
[16] Xie H, Li X, Wang T, et al. Incorporating sentiment into tag-based user profiles and resource profiles for personalized search in folksonomy[J]. Information Processing & Management, 2016, 52(1): 61–72.
[17] 李光敏, 陈炽, 邢江, 等. 网络文本评论中产品特征抽取综述 [J]. 现代情报, 2016(8): 168–173.
Li G M, Chen C, Xing J, et al. Overview of extracting product feature from text reviews[J]. Journal of Modern Information, 2016(8): 168–173.
[18] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613–620.
[19] Jaccard J, Turrisi R, Wan C K. Interaction effects in multiple regression[J]. Newbury Park California Sage Publications, 1990, 40(4): 461.
[20] Rodriguez A, Laio A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492–1496.
[21] 孙磊磊. AP 聚类算法研究及其在电子病历挖掘中的应用 [D]. 大连: 大连理工大学, 2017.
Sun L L. Study on affinity propagation clustering algorithm and its application in mining electronic medical records[D]. Dalian: Dalian University of Technology, 2017.
[22] Strehl A, Ghosh J. Cluster ensembles-a knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research, 2003, 3: 583–617.
[23] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(1): 993–1022.