Statistics and Application 统计学与应用, 2022, 11(1), 111-119
Published Online February 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.111013

Hans汉斯

基于美颜相机APP用户画像精准营销策略研究

韩　璐

燕山大学理学院，河北　秦皇岛

收稿日期：2022年1月14日；录用日期：2022年1月24日；发布日期：2022年2月14日

摘　要

随着大数据时代的到来，互联网技术飞速发展。为在竞品当中脱颖而出，获得竞争优势，过度竞争趋势在互联网企业中逐渐显现，由此引发公司营销成本的上升以及营销绩效的下降等许多问题。针对上述问题，本文依靠美颜相机用户数据，利用LDA模型建立用户画像，通过提取主题词得到相应的主题分布；将选取的对应词扩充到特征空间中，完善用户特征，再利用SVM分类算法区分用户基本属性，进而构建用户画像。根据建立用户画像结果，给出精准营销策略建议。

关键词

用户画像，LDA，SVM，精准营销

Research on Precision Marketing Strategy of User Portrait Based on Beauty Camera App

Lu Han

College of Science, Yanshan University, Qinhuangdao Hebei

Received: Jan. 14th, 2022; accepted: Jan. 24th, 2022; published: Feb. 14th, 2022

Abstract

With the advent of the era of big data, Internet technology has developed rapidly. In order to stand out among competing products and gain competitive advantage, the trend of excessive competition gradually appears in Internet enterprises, which leads to many problems, such as the rise of marketing costs and the decline of marketing performance. To solve the above problems, this paper relies on the user data of beauty camera, uses LDA model to establish user portrait, and ob-

文章引用: 韩璐. 基于美颜相机 APP 用户画像精准营销策略研究[J]. 统计学与应用, 2022, 11(1): 111-119.
DOI: 10.12677/sa.2022.111013

tains the corresponding subject distribution by extracting subject words; Expand the selected corresponding words into the feature space, improve the user features, and then use the SVM classification algorithm to distinguish the basic attributes of users, so as to construct the user portrait. According to the results of establishing user portrait, give suggestions on precision marketing strategy.

Keywords

User Portrait, LDA, SVM, Precision Marketing

1. 引言

Alan Cooper 首次提出用户画像的概念，最初可理解为用户信息的标签化。用户画像的建立，需结合实际数据，制定符合实际数据情况的标签数据集，再针对数据进行分析。互联网的快速发展带来大量数据，这些数据中蕴含着无限商业价值。通过收集用户的行为数据，将行为数据可视化，关联用户的使用习惯以及兴趣偏好等，可以帮助美颜相机企业开发更舒适的 APP 使用体验，进而改进 APP 运营机制，向用户提供更优质的服务。陈志明等构建多属性用户画像模型，充分利用知乎网站后台用户行为数据，包含用户的兴趣爱好、知识能力等[1]。在此基础上，王乐等提出利用多类型特征算法排除网络行为数据复杂性因素，解决了特征选择不确定问题[2]；李军政利用 VSM 模型，根据检索数据提取特征词，关联用户基本属性，由此建立用户画像模型[3]。以上模型建立虽能得出相应的用户画像，但是因为短文本识别精准度不高以及检索词关联度高等问题[4]。VSM 模型在建立用户画像模型上存在着一定的缺陷，无法有效表示用户特征[5]。因此，本文提出利用 LDA 模型，利用美颜相机 APP 历史搜索记录数据，通过提取短文本主题，并扩充语义特征，建立精度更高的用户画像。基于精准的用户画像，帮助互联网企业建立精准的营销对策。

2. 基础模型

2.1. LDA 模型

本文先利用 LDA 模型进行短文本分类，将已经清洗完成的美颜相机用户搜索记录文本进行语义分类。LDA 模型一般在识别不明显语义文本信息中广泛使用，是无监督模型的一种。它的作用体现在将文档分为多个主题分布[6]。LDA 模型中文本主题分布为 $\Theta = \text{Dirichlet}(\alpha)$，其中，$\alpha$ 为超参数，通常 $\alpha = 50/T$。主题词也服从 Dirichlet 分布，可表示为 $\varphi = \text{Dirichlet}(\beta)$，其中 β 为超参数，通常 $\beta = 0.01$ [7]。

模型中 Θ 和 φ 常使用近似估计技术进行估计，如 EM 算法、Gibbs 采样算法。

以 Gibbs 采样算法为例，估计流程为：

1) 随机分配文本主题。
2) 计算主题概率。
3) 重复采样，直到结果收敛。
4) 得到文本主题分布 Θ，主题词分布 φ。

采样公式为：

$$p\left(t_i = s \mid t_{-i}, w_i\right) = \frac{n_{m,-i}^{(s)} + \alpha}{\sum_{t=1}^{T} n_{m,-i}^{(t)} + T\alpha \sum_{j=1}^{N} n_{s,-i}^{(j)} + N\beta}$$

其中，$n_{m,-i}^{(s)}$ 表示文本 d_m 分到 s 的频次，$n_{m,-i}^{(s)}$ 表示词 w_k 分到 s 的频次。

采样后，文本主题分布为 Θ，主题词分布 φ 可进行估算。

$$\theta_{m,s} = \frac{n_m^{(s)} + \alpha}{\sum_{t=1}^{T} n_m^{(t)} + T\alpha}$$

$$\theta_{s,k} = \frac{n_s^{(k)} + \beta}{\sum_{j=1}^{N} n_s^{(j)} + N\beta}$$

2.2. SVM 分类模型

利用 LDA 模型进行文本语义分类之后，为提高语义分类的精确度，本文利用 SVM 模型进行特征扩展。SVM 模型是根据样本数据建立超平面，并区分样本[8]。计算样本与超平面之间的距离，距离最近的样本会影响分类的广泛度，扩大广泛度可提高模型的精准度，因此要尽可能让样本到超平面的距离变大。超平面公式为：$\omega^{\mathrm{T}} x + b = 0$，可以推出参数 ω 和 b 影响超平面，距离公式为：

$$L = \frac{\left|\omega^{\mathrm{T}} x + b\right|}{\|\omega\|}$$

所有样本当中对广泛度影响力最大的是支持向量，不同支持向量的距离公式可以表示为 $\gamma = \frac{2}{\|\omega\|}$，求得最优向量坐标 (ω, b) 即可使得 γ 最大，即广泛度最大[9]。

3. 用户画像构建

3.1. 获取数据

数据主要分为静态信息数据、动态信息数据两类[10]。静态信息数据是指在一段时间内保持稳定不变的用户信息内容，它也可作为用户基础属性数据，如性别、年龄等基本属性。而动态信息数据则是指正在变化的信息内容，包括当前用户的访问情况、浏览行为、应用偏好等，在一定程度上也体现了使用者的软件应用习惯、兴趣等属性。

3.1.1. 用户属性数据获取

美颜相机用户属性数据属于静态信息数据，包含基础属性、生活状况、社会标签、心理情况等。在相当一段时间中一般不会产生变化，比如性别、社会阶层、薪资条件等。通过 APP 后台发放调查问卷即可获得详细数据。

3.1.2. 用户行为数据获取

美颜相机用户行为数据属于动态信息数据，是建立用户画像的重要数据。需要通过 APP 后台搜集用户在使用软件时的流程数据。包括用户访问轨迹，历史搜索，ICON 点击量等，并对相关数据进行分析。

3.2. 模型构建

为还原用户信息，根据用户行为信息，标准化标签构建用户画像，主要分为如下三个阶段：一、对

用户基础信息、检索数据，APP 内行为数据等基本数据进行预处理。二、建立合适的用户标签。三、构建用户画像并绘制可视化标签云。

以用户搜索记录数据为例，先依靠建立的 LDA 模型进行文本主题的提取，获得主题词之后，对照用户的特征进行拓展，接着在 SVM 模型基础上区分用户的基本属性，进行用户画像的构建，模型框架如图 1 所示。

Figure 1. User portrait model framework
图 1. 用户画像模型框架

3.2.1. 数据预处理

先将获得的数据中不具备完整属性的数据记录进行清洗。再进行分词处理，将短文本转化成主题词短语，获得关键文本内容，一般可采用 jieba 分词方法，过滤无意义词汇，降低特征词的词性维度，去除已停用的词汇，其结果如表 1 所示。

Table 1. Word segmentation results
表 1. 分词结果

原短文本	分词结果
纯欲高级感美颜滤镜	纯欲高级滤镜
可爱粉色俏皮感兔子贴纸	可爱贴纸
泰式混血夸张感美颜特效	夸张特效

3.2.2. 建立用户标签

先建立多维属性标签，再利用文本文义提取来抽象用户信息，对应建立的标签属性进行用户画像的构建。以用户行为属性为例，美颜相机用户行为属性框架如图 2。

3.2.3. 构建用户画像

用户通过检索 ICON 可直接获得满足自身需求的贴纸、滤镜，搜索词条与用户需求关联度高，用户需求则与用户兴趣爱好、基本属性关联度高。例如高收入人群对奢侈品图标贴纸使用频次更高；学生群体更偏向于使用可爱粉紫色贴纸滤镜；日活活跃的用户对新功能 ICON 点击率更高；男性相比女性来说，则更倾向于搜索运动风格、二次元贴纸，因此通过检索内容建立模型来描述用户属性标签能够构建相对精准的用户行为画像。在将长文本转化为短文本的方面，VSM 模型无法有效联系上下文，精准把握语义，并且无法针对短文本提取主题词，因此建立的用户画像的精确度会大大降低。本文在建立用户画像过程中融入 LDA 模型中。举例来说，对于“复古感调色”和“复古滤镜”，VSM 无法区分两个文本语义的相似性，然而实际上这两个短文本之间差异不大，LDA 模型则可以关联两个短文本，从而解决特征稀疏问题。根据 LDA 模型将文本语义精简为主题词分布，利用向量表示短文本，基于用户检索内容主题词差异相对较大，若检索词主题公共部分不重叠，即查询词主题之间相似度为 0，那么主题分类则不符合要求。由此延申，通过对主题文本特征扩展，关联潜在语义主题文本，全面表达文本特征。流程如下：

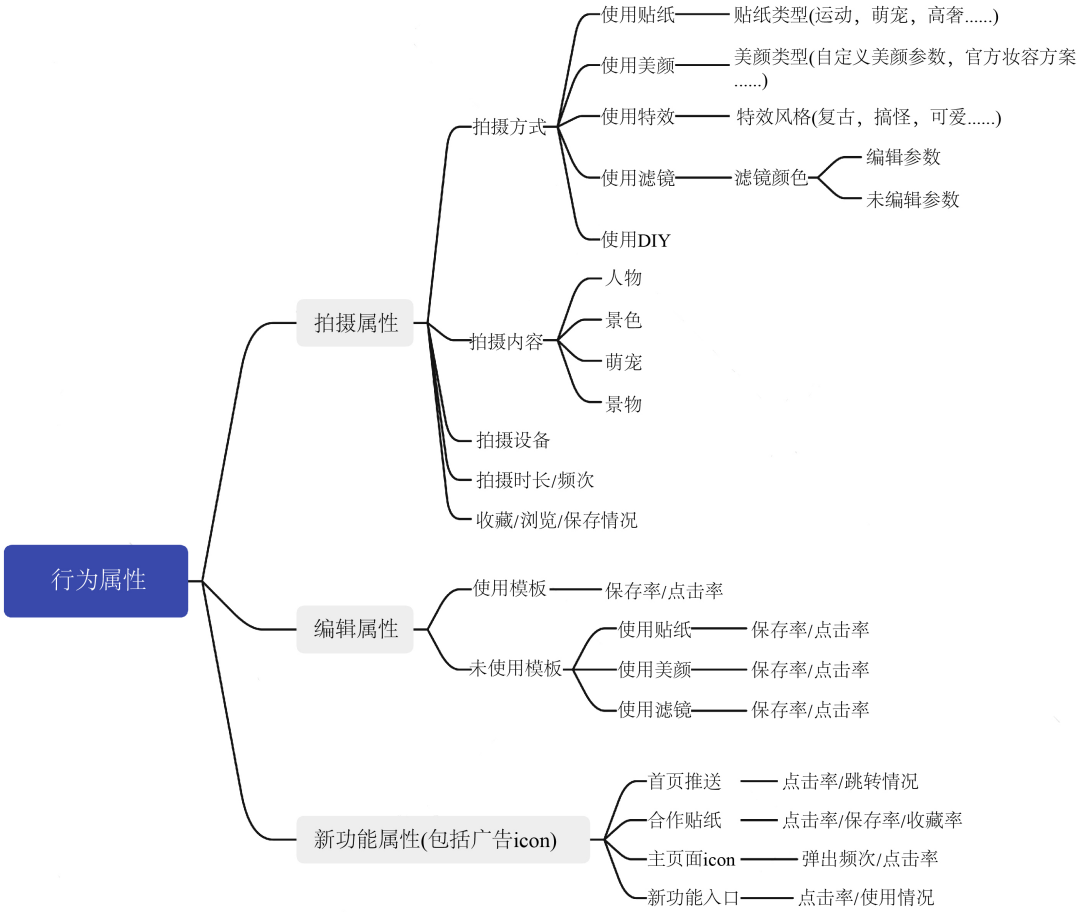

Figure 2. Behavior attribute framework
图 2. 行为属性框架

1) 以向量形式表达检索文本，特征值归一。

2) 特征选择提取的特征向量，把有意义的特征词看作原始特征词。

3) 针对用户检索数据集，得出检索词的主题分布 Θ，及主题词分布 φ。

4) 计算主题概率，概率最大主题为 s，s 的主题词即为用户的扩展特征，词 w_k 属于这个主题的概率就是 s 的特征值。此外，若 s 在最初的特征词集合中，则不重复添加。

5) 将扩展后的特征词的属性标签通过建立 SVM 模型进行分类。

判断拓展分类词分类结果，本文采取的评价标准为查准率 P、查全率 R 和 F_1 值[11]，计算性别、年龄、学历属性的分类精确率、召回率和 F_1 值，计算公式如下：

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$F_1 = \frac{2 \times P \times R}{P + R}$$

混淆矩阵是用来总结一个分类器结果的矩阵[12]。对于 k 元分类，即绘制 $k \times k$ 的表格，用来记录分

类器的预测结果。对于最常见的二元分类来说，其混淆矩阵是2×2的，如表2所示：

Table 2. Confusion matrix
表2. 混淆矩阵

	预测正类	预测反类
真实正类	TP (真实类)	FN (假反类)
真实反类	FP (假正类)	TN (真反类)

将基于LDA模型提取关键词方法与在LDA模型基础上进行特征扩展的方法提取用户基础信息，如性别、年龄、学历属性评价结果进行比对，如表3所示：

Table 3. Evaluation results
表3. 评价结果

LDA 模型			LAD 特征扩展		
精确率	召回率	F_1	精确率	召回率	F_1
0.753	0.752	0.751	0.812	0.814	0.813
0.534	0.601	0.543	0.578	0.612	0.593
0.556	0.598	0.571	0.596	0.632	0.612

实验结果可知，LDA特征扩展对文本关键词提取准确度更高，相比LDA模型准确率、召回率、F1值均提高2%左右，由于美颜相机用户检索短语的主题不密集，用户特征显现不明显，在此基础上，LDA模型经过特征扩展之后，可以相对全面表达短语特征，缓解美颜相机类软件中短文本特征稀疏问题，较好解决LDA模型在提取关键词过程中的一部分缺陷。

4. 研究结果

4.1. 美颜相机用户整体画像

用户画像可以刻画美颜相机整体用户特征。美颜相机的用户多为14~20岁学生群体，女性占比88.08%，选择的滤镜风格以复古风居多，贴纸风格以可爱风居多，拍摄内容主要以人物居多，对于会员的购买情况，购买会员的用户占比不到20%，对软件的评分平均为2.873。由网络分析图3可知，用户对软件内贴纸滤镜的选择主要依靠以下五个角度：风格、效果、顺序、流行热词、清晰度。

4.2. 美颜相机用户特征画像

为精细化美颜相机用户画像，将用户所在城市划分为七个城市等级，分别为一线城市、准一线城市、二线城市、三线城市、四线城市、五线城市、其他城市。

由图4可知，美颜相机用户主要聚集在一线、准一线城市，城市越发达，用户量越大。以一线城市用户为例，可刻画具体用户画像如图5，一线城市用户主要以普通用户和VIP用户为主，女性，年龄在14~28岁之间，学历在高中以及本科人数居多，除学生群体外，企业就业人员收入水平在8000元以上，用户将美颜相机作为主要拍摄软件，使用设备为摄影类手机为主，其中手机品牌以苹果、华为、小米居多。在广告推送方面，用户接受度偏高，可接受每天2条推送左右，开屏广告对用户体验影响度不高。在滤镜、贴纸选择方面，主要选择类型为ins风、复古风、可爱风，在拍摄内容方面，主要以美食、人像

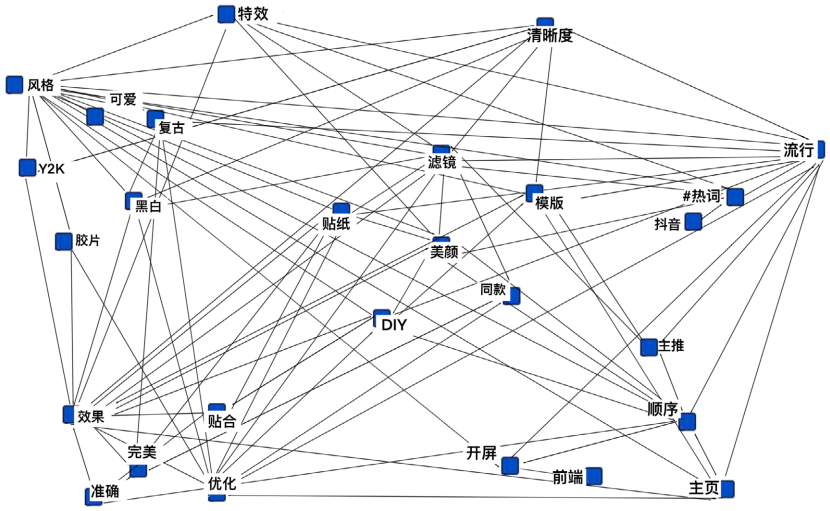

Figure 3. Network analysis of user portrait search words
图 3. 用户画像搜索词网络分析图

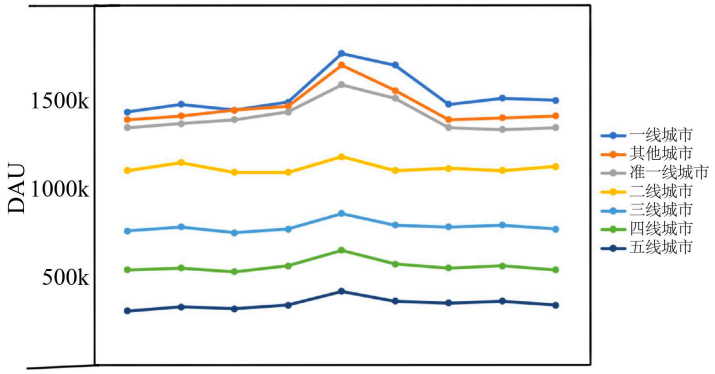

Figure 4. Distribution of beauty camera users in cities of each line
图 4. 各线城市美颜相机用户分布情况

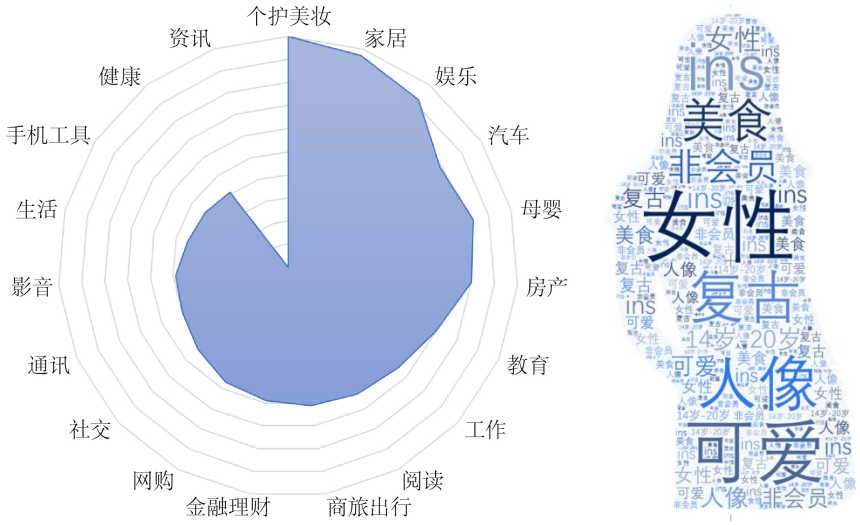

Figure 5. Specific user portraits in first tier cities
图 5. 一线城市具体用户画像

内容为主，在妆容选择方面，通常默认模板自带妆容，主动更换妆容频次不高，对于 DIY 等新功能入口点击率，用户点击率为 40%左右，调整入口位置，点击率变化不大，由此用户对于软件的使用更偏向惯性使用，对新功能探知欲不强。创作者专区模板情况数据情况为依据，使用模板具有特定风格、类型、时代的模板使用频次较高，用户自己做创作者的人数比例极低。新用户对于新功能入口、创作者专区等点击率较高，但保存率偏低，留存用户使用软件时间在 1~5 年之间，使用频次较稳定，点击率和保存率也保持稳定，一般拍摄时间为 10 点~13 点、19 点~23 点，经常出现反复性使用同一贴纸、滤镜的行为。软件内交流、分享、评论等行为参与度不高，模板收藏低于 100，创作者收藏低于 50。VIP 用户占比在 30%左右，其中其他平台关联 VIP 比例高达 70%以上，VIP 用户相较于普通用户，更愿意探索软件新功能，重视使用体验，针对软件的需求有明确的想法，以及对付费滤镜更新期待值也会更高，因此，提升 VIP 用户使用体验可以大幅提升付费型用户占比，对软件的盈利及营销都有至关重要的价值。

5. 基于美颜相机用户画像的精准营销对策

5.1. 建立完善的精准营销体系

精准营销策略的建立需要依靠真实数据，建立可靠用户画像，把握用户需求，为用户提供个性化服务，准确的内容推荐以及舒适的 APP 使用体验，在提高现有用户忠诚度基础上吸引新用户。在此基础上可以看出软件运营过程当中关键数据埋点的重要性，比如，新功能开放时 ICON 的点击率、保存率等等关键数据，在硬件设备发展逐渐扩大的现实条件下，软件的应用也应随硬件进行调试，比如智能手表，智能电视等设备的适配也在吸引新用户方面有很大的帮助。

5.2. 基于用户行为数据实施精准营销

用户画像是软件目标用户群体的整体体现，了解用户画像的根本目的就是根据用户群体的情况调整软件的运营决策，美颜相机在软件运营过程中，完善功能性 APP 使用流程，优化用户的使用体验，增大用户体量尤为重要。根据获得的用户画像信息，精准定位目标用户，不同用户群体采用不同的渠道进行内容、广告投放，在不改变用户软件使用舒适度的情况下，利益最大化，以用户喜爱的交互行为进行广告营销，以实现稳定用户群体的同时获得收益。例如，服饰类广告可以设计服饰品牌贴纸特效等，不同类型用户进入拍摄的 ICON 位置不同，从而使得广告投放点击率提升。

5.3. 即时更新数据，追踪用户画像变化情况

互联网数据时代催生用户使用软件情况变化迅速，流行风格以及软件受众都可能瞬息发生变化，此刻的需求并不是一成不变的，而是呈现变化的动态图谱，因此，要即使掌握用户的使用数据，实时有效分析。建立的标签具有一定的生命周期，不进行追踪就难以抓住用户的真实需求，因此标签必须实时更新。建立标签规则，如更新维度、更新权限等关键要素。增加新产生的标签，去除已经被淘汰的标签；调整触发机制以更新标签，设立更新条件。以此获得更为准确的用户画像信息。在完成拍摄、P 图等动作之后，还需定期发放问卷以调研用户的满意度，为软件提供下一步的产品、内容、服务等方面的策略调整。减少显性且频繁的广告营销，尤其是软件之间的跳转类型营销，频繁的软件切换会引起用户反感，影响使用体验。

6. 总结

本文主要介绍了如何根据用户检索内容建立精准用户画像，以帮助美颜相机互联网企业精细化定位人群，挖掘潜在用户。利用大数据建立模型的最终目的是服务于公司基于主体用户画像进行精细的营销

决策，以获得最大收益。用户画像连结互联网企业与用户，利用大数据建立模型，挖掘潜在用户并留住潜在用户，帮助互联网企业更加有针对性进行用户增长工作，是让互联网企业获得利益最大化的营销手段。

参考文献

[1] 陈志明, 胡震云. UGC 网站用户画像研究[J]. 计算机系统应用, 2017, 26(1): 24-30.

[2] 王乐, 倪维健, 林泽东, 等. 基于模型堆叠的上网行为日志用户画像方法[J]. 山东科技大学学报(自然科学版), 2018, 37(5): 70-78.

[3] 李军政, 黄海, 黄瑞阳, 等. 基于卡方检验和 SVM 的用户搜索画像技术研究[J]. 电子设计工程, 2017, 25(24): 6-10.

[4] 费鹏. 用户画像构建技术研究[D]: [硕士学位论文]. 大连: 大连理工大学, 2017.

[5] 张虹. 短文本分类技术研究[D]: [硕士学位论文]. 大连: 辽宁师范大学, 2015.

[6] Tan, C. (2013) Short Text Classification Based on LDA and SVM. *International Journal of Applied Mathematics and Statistics*, **51**, 205-214.

[7] 杨萌萌, 黄浩, 程露红, 等. 基于 LDA 主题模型的短文本分类[J]. 计算机工程与设计, 2016, 37(12): 3371-3377.

[8] 施瑞朗. 基于社交平台数据的文本分类算法研究[J]. 电子科技, 2018, 31(10): 69-70+75.

[9] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

[10] Sidorov, G. and Gelbukh, A. (2014) Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model. *Computer Sistemas*, **18**, 491-504. https://doi.org/10.13053/cys-18-3-2043

[11] Mikolov, T., Sutskeve, R.I., Chen, K., *et al*. (2013) Distributed Representations of Words and Phrases and Their Compositionality. *Proceedings of Advances in Neural Information Processing Systems*, [cs. CL] 16 October 2013, 3111-3119.

[12] Qiu, Z. and Wu, B. (2014) Collapsed Gibbs Sampling for Latent Dirichlet Allocation on Spark. *Journal of Machine Learning Research*, **16**, 17-28.