# 基于微博的用户画像分析

魏嘉毅 2620160022
李安琪 2620160031
刘艳 2620160026

## 一、引言

大数据时代的到来，对数据进行挖掘分析出每个用户的兴趣爱好、行为习惯，构建出用户的信息全貌，对于个性化服务有很大的帮助。用户画像作为大数据的根基，它完美地抽象出一个用户的信息全貌，为进一步精准、快速地分析用户行为习惯、消费习惯等重要信息，提供了足够的数据基础，奠定了大数据时代的基石。在个性化的信息检索、推荐系统等应用中用户画像也发挥着很大的作用，现在也是学术界和工业界的关注热点。

随着社交网络的发展，如国外的 Twitter、Facebook，抑或是国内的新浪微博、腾讯微博等，越来越多的用户参与其中。针对社交媒体的特性，设计构建一套完整的用户画像模型是很有必要的。

用户画像，即用户信息标签化，就是企业通过收集与分析消费者社会属性、生活习惯、消费行为等主要信息的数据之后，完美地抽象出一个用户的商业全貌，这是企业应用大数据技术的基本方式。用户画像为企业提供了足够的信息基础，能够帮助企业快速找到精准用户群体以及用户需求等更为广泛的反馈信息。用户画像对企业能提供很大的帮助，那么提高用户画像的精确度和全面性是很有必要的。结合当前社交网络的发展，越来越多的用户以各种形式在社交平台上表现自己，研究社交媒体环境下的用户画像是很有意义的。

并且当前的基于社交媒体的用户画像还存在很多有待改进的地方，如用户属性描述不深入、不全面，没有做到及时更新等，更需要我们进行深入研究，解决发现的问题，构建深入全面的用户画像，为个性化的推荐系统、信息检索等服务提供较全面详细的信息。

## 二、理论基础

### 1、用户画像

用户画像是根据用户社会属性、生活习惯和消费行为等信息而抽象出的一个标签化的用户模型。构建用户画像的核心工作即给用户贴“标签”，而标签是通过对用户信息分析而来的高度精炼的特征标识。除去“标签化”，用户画像还具有的特点是“低交叉率”，当两组画像除了权重较小的标签外其余标签几乎一致，那就可以将二者合并来弱化权重较小的标签的差异。

打标签的重要目的之一是为了让人能够理解并且方便计算机处理，如可以做分类统计：喜欢日料的用户有多少？喜欢日料的人群中，男、女的比例是多少？也可以做数据挖掘工作：利用关联规则计算，喜欢日料的人通常还会喜欢什么样的穿衣风格？利用聚类算法分析，喜欢日料的人年龄段分布情况？大数据处理离不开计算机的运算，标签提供了一种便捷的方式，使得计算机能够程序化处理与人相关的信息，甚至通过算法、模型能够“理解”人。当计算机具备这样的能力后，无论是搜索引擎、推荐引擎、广告投放等各种应用领域，都将能进一步提升精准度，提高信息获取的效率。

用户画像的作用大体可以概括为以下几类：1）精准营销，分析产品潜在用户，针对特定群体利用短信、邮件等方式进行营销；2）用户统计，比如统计某个主题的用户关注度；3）数据挖掘，构建智能推荐系统，利用关联规则计算，喜欢日料的人通常喜欢什么样的穿衣风格，然后进行推荐；4）进行效果评估，完善产品运营，提升服务质量，其实这也就相当于市场调研、用户调研，迅速定位服务群体，提供高水平的服务；5）对服务或产品进行私人

定制，即个性化的服务某类群体甚至每一位用户；6）业务经营分析以及竞争分析，影响企业发展战略。

2、行为建模

行为建模是对上一步收集到的数据进行处理，进行行为建模，以抽象出用户的标签。建模的方法主要有遗传算法、聚类、贝叶斯和神经网络方法等。遗传算法通过模拟自然进化过程搜索最优解，采用遗传结合、遗传交叉变异以及自然选择等操作实现建模。依据每个个体的适应度函数值，种群经过选择、 交叉和变异等操作一代代向更优良、更适应环境的方向进化，从而逼近最优解。当用户的兴趣发生变化时，通过遗传进化满足用户新的需求。

聚类是将一个数据集划分为若干组或类的过程，并使得同一个组内的数据对象尽可能地相似，而不同组中的对象则尽可能地不相似。

贝叶斯方法分为贝叶斯分类和贝叶斯网络两种建模方法。贝叶斯分类方法通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，选择具有最大后验概率的类作为该对象所属的类。贝叶斯分类是基于各类别相互独立这一假设来进行类别计算的，但是实际上变量间的相互依赖情况是很常见的。贝叶斯网络就是用于描述随机变量之间依赖关系的图形模式，是一个具有概率分布的有向弧段，由代表事件或变量的节点和代表节点之间的因果关系或概率关系且不构成回路的有向弧段组成。

神经网络方法中比较典型的有 BP 神经网络和自组织神经网络。BP 神经网络模型由多层神经元层组成，其训练过程是输入学习样本，使用反向传播算法对网络的权值和偏差进行反复的调整训练，使输出的向量与期望向量尽可能地接近，当网络输出层的误差平方和小于指定的误差时训练完成。BP 神经网络对未知数据具有较好的预测分类能力，但学习时间较长，只适用于时间允许的应用场合。自组织神经网络不必提供学习样本，它以实际的神经网络中的侧抑制现象和生物体在认知过程中的自学习的“无师自通”现象(即无监督学习)为根据。随着深度学习的出现，现在较多采用卷积神经网络来训练数据集，特别是将多个数据源进行结合训练得到的结果也比较好 [11] 。

在进行建模的过程中，建模时间也是一个关键的因素。按照时间尺度，可以分为长期建模和短期建模，前者描述用户较长时间的、比较稳定的兴趣爱好，后者描述用户近期的、临时的兴趣爱好。但是一般都是建立一个混合模型，依靠用户的历时数据来挖掘用户稳定的、波动范围小的长期兴趣，依靠最近的数据来挖掘用户个性化的、波动范围大的短期兴趣。

按照更新方式，可以分为静态建模和动态建模。静态建模构建的用户模型不随时间的演变而发生变化，保持稳定；而动态建模则考虑了随时间的演进，用户原有的兴趣的衰减、变化和新兴趣的生成等。很多研究者对用户模型的遗忘和更新进行研究，提出基于时间窗口和遗忘机制的模型来解决兴趣偏移的问题。但是还是会存在一定的问题例如长期存在某个兴趣，但是只是有一段时间对其关注比较低，但是如果模型建立不合理会认为不存在这个兴趣了。

3、文本预处理

文本预处理即对文本进行初步加工，去掉一些没有意义的信息，留下对文本有标志意义的信息。一般的文本预处理包括分词、去停用词、去无意义的高频词等。对于某些特殊的短文本有可能还需要进一步的处理，例如去掉链接和特殊用语等。

分词就是根据特定的语言习惯或统计理论等规则将句子进行分割，使得分割后的每个小段都是有意义的一个词。但是对于中文来说，它不同于英文，因为在英文中含有空格，空格就是英文分词的分隔符，而中文是以字为基本的书写单位，词语之间没有明显的区分标志，所以中文分词相对于英文来说要困难。然而分词的好坏对中文自然语言处理实验的结果会有很大的影响，所以处理好分词是进行实验的关键。目前分词算法比较多，主要可以归纳为基于词库匹配的分词方法、基于知识理解的分词方法和基于词频统计的分词方法。

一般我们将那些经常出现但是包含的信息量少，自身也没有明确的意义，而且对于文本也起不到作用的词列为停用词，比如“了”、“吧”、“的”、“在”、“是”等。但是停用词的存在会占用存储空间、降低关键词的密度，在主题提取的过程中，有可能在各个主题中都出现这些停用词，而且权重还特别大，这样会影响用户属性的确定，因此去除停用词是很有必要的。实现去停用词的这个过程，需要先建立一张停用词表，然后在文本进行预处理的时候就可以参照这个停用词表来剔除一些词。

4、LDA 主题模型

LDA 是隐含狄利克雷分布的简称，是一种主题模型，它可以将文档集中每篇文档的主题按照概率分布的形式给出。同时它是一种无监督学习算法，在训练时不需要手工标注的训练集，需要的仅仅是文档集以及指定主题的数量即可。此外 LDA 的另一个优点则是，对于每一个主题均可找出一些词语来描述它。

LDA 是一种典型的词袋模型，即它认为一篇文档是由一组词构成的一个集合，词与词之间没有顺序以及先后的关系，这样将把问题简化，进而将文本的文字信息数字化为数字信息，方便之后的建模。一篇文档可以包含多个主题，文档中的每一个词都由其中的一个主题生成。即把每篇文档表示成主题的概率分布，每个主题表示成单词的概率分布。

5、BTM 主题模型

LDA 主题模型被广泛应用于挖掘语料潜在主题，在长文本领域取得了不错的成绩。但是对于短文本，会产生数据稀疏的问题。于是有人通过引用外部语料（比如搜索片段，背景信息等）来将短文本扩充成长文本后应用主题模型进行建模计算，但是符合扩充条件的语料有时并不容易获得，并且最终结果很大程度上依赖用来扩充的语料信息。针对上述方法存在的缺陷，提出了专用于短文本的 BTM 主题模型对短文本进行建模。

在 BTM 中，语料库中所有的词对共享一个主题概率分布，主题是互异词项的概率分布，BTM 直接对语料库中所有词对中词的生成过程进行建模，而没有对文档直接进行建模，所以 BTM 无法直接获得文档主题分布，但是这个概率分布可以用推理得到。

BTM(Biterm Topic Model)模型通过共现词对模式（biterm：一个短文本窗口中的无序共现词对）来加强主题模型的学习，利用整个语料库的丰富的信息抽样主题，推断整个语料库全局的主题分布，不仅保持了词之间的相关性，同时因为一篇文档中不同共现词对相互独立，也可以推断一篇文档对应的不同主题概率。BTM 直接对共现词对建模作为主题的语义传输单元，比单个词能更好地揭示主题。共现词对是指文本通过预处理之后共同出现在同一个文本片段中的任何并且无序的两个不同的词。

# 三、动态用户画像模型

由于用户在社交媒体环境下发布或者转发的内容多为短文本，特征稀疏，所以想要从中挖掘用户的属性就会显得很困难。我们只有对每个短文本中有限的词进行深度挖掘，才能更好地捕获文档之间的隐含关系。上一章中详细介绍的 BTM，正好是适用于短文本的，使用 BTM 可以生成短文本的主题，在此基础上进行动态的用户画像建模，最后得出用户的属性。

本模型的主要的步骤即文本预处理、主题提取和属性挖掘。动态用户画像模型的基本流程如下所示：

（1） 选择数据集。在社交媒体平台进行用户相关数据的爬取。

（2） 文本预处理。包括对文本进行分词、去停用词、去无意义的高频词等，获得较为规则的数据集。

（3） 主题提取。用 BTM 对分好词的文本进行建模，通过多次实验确定主题数目 K 的值，然后获得 K 个主题及其在文档中的分布情况，还可以获得每个主题下的主题词及其每个词各自的权重。

（4） 用户属性挖掘。通过统计用户文本中主题词的词频来确定用户的属性，并且加入时间窗口和衰减函数来获取动态的用户属性。经过以上步骤，用户的动态属性就基本得到了，即构建了用户画像。利用用户画像可以对用户进行分析并为其提供个性化的服务。

# 四、实验结果及分析

为了在现实平台上面评价我们的模型，我们在新浪微博平台随机选择了 2100 名用户，将他们在 2015 年一年内发布或者转发的微博以及用户的个人信息都爬取下来作为我们的数据集，一共有 640000 条微博。接下来我们对数据进行一定的筛选，把微博量少于 200 的用户信息给过滤了。在剩下的用户中选择 100 个发布或转发的微博量大于 5000 的用户作为训练集来进行主题提取，再选择 100 个用户作为测试集来进行用户属性挖掘和模型评价。

在动态用户画像模型的主题提取模块中，BTM 主题模型的参数 $\alpha$ =50/ K ,参数 $\beta$ =0.01 ，通过多次实验发现主题数目 K 设置为 10 比较合理，即我们通过主题模型可以得到 10 个主题。每个主题下的主题词的数目 m 设置为 20，即该 20 个主题词在该主题下是出现频率比较高的。主题词选得多，后面的词所占比例都不高，意义不大，而且容易和其他主题下的主题词重复；主题词选得少，代表性不够强，有些主题对应的主题词较分散。因此将主题词的数目设为 20 较为合理。

1、用户属性分布

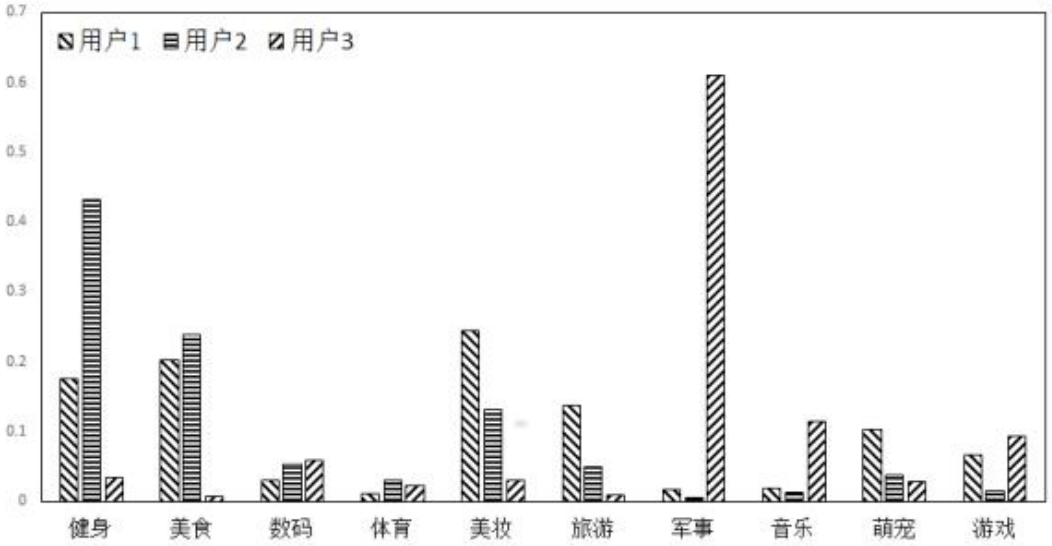
在用户属性挖掘模块中，阈值经过推理最终定为 T=0.19，取这个阈值可以使得挖掘出来的结果比较好，并且进行计算 F1 值也是最高的。衰减系数的公式中设置 $\mu$ =0.56， $\nu$ =0.06 ，参数设置这两个值使得曲线的变化趋势与用户属性的转移趋势较相似。时间窗口的长度设置为 3 个月，即以每个月作为一个时间段，每个时间段的结果通过其前 3 个月的数据来进行建模得到。

将作为训练集的 100 个用户的微博数据进行预处理后，采用 BTM 主题模型来进行主题提取。我们通过多次实验最终在主题数目选择 10 的时候效果较好，因此获得了 10 个主题以及每个主题下的 20 个主题词以及它们各自的权重，如下图所示。

| 主题 | 主题词 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 健身 | 动作 | 0.0231 | 健身 | 0.0083 | 马甲 | 0.0075 | 教程 | 0.0072 | 瑜伽 | 0.0068 |
| | 腹肌 | 0.0059 | 减脂 | 0.0058 | 跑步 | 0.0057 | 拉伸 | 0.0053 | 腿部 | 0.0051 |
| | 训练 | 0.0047 | 锻炼 | 0.0042 | 练出 | 0.0041 | 脂肪 | 0.0040 | 小腿 | 0.0038 |
| | 视频 | 0.0036 | 瘦身 | 0.0035 | 大腿 | 0.0035 | 瘦腿 | 0.0034 | 肌肉 | 0.0032 |
| 美食 | 美食 | 0.0171 | 做法 | 0.0097 | 吃货 | 0.0079 | 美味 | 0.0065 | 好吃 | 0.0064 |
| | 爱看 | 0.0051 | 赶紧 | 0.0046 | 放入 | 0.0036 | 转需 | 0.0035 | 蛋糕 | 0.0028 |
| | 超级 | 0.0024 | 自制 | 0.0023 | 口感 | 0.0023 | 学习 | 0.0022 | 鸡蛋 | 0.0022 |
| | 生抽 | 0.0022 | 一道 | 0.0022 | 教程 | 0.0022 | 味道 | 0.0021 | 吃法 | 0.0021 |
| 数码 | 手机 | 0.0174 | 苹果 | 0.0083 | 小米 | 0.0076 | 摄像头 | 0.0068 | 处理器 | 0.0063 |
| | 魅族 | 0.0063 | 骁龙 | 0.0062 | 华为 | 0.0055 | 乐视 | 0.0052 | 像素 | 0.0046 |
| | 发布会 | 0.0044 | 识别 | 0.0043 | 屏幕 | 0.0042 | 三星 | 0.0040 | 直播 | 0.0039 |
| | 英寸 | 0.0036 | 抽奖 | 0.0033 | 发布 | 0.0032 | 售价 | 0.0030 | 新机 | 0.0030 |
| 体育 | 篮球 | 0.0101 | 曼联 | 0.0100 | 比赛 | 0.0074 | 科比 | 0.0068 | 范加尔 | 0.0068 |
| | 永不 | 0.0067 | 德约 | 0.0066 | 费德勒 | 0.0063 | 球员 | 0.0059 | 赛季 | 0.0050 |
| | 纳达尔 | 0.0048 | 澳网 | 0.0047 | 球迷 | 0.0045 | 网球 | 0.0043 | 阿森纳 | 0.0038 |
| | 曼城 | 0.0036 | 科维奇 | 0.0035 | 英超 | 0.0035 | 派送 | 0.0034 | 网坛 | 0.0033 |
| 美妆 | 化妆 | 0.0173 | 教程 | 0.0173 | 妆容 | 0.0104 | 美妆 | 0.0074 | 技巧 | 0.0061 |
| | 学习 | 0.0057 | 日常 | 0.0047 | 女生 | 0.0043 | 适合 | 0.0038 | 眼妆 | 0.0037 |
| | 好看 | 0.0037 | 护肤 | 0.0036 | 画法 | 0.0036 | 妹子 | 0.0036 | 实用 | 0.0035 |
| | 皮肤 | 0.0033 | 方法 | 0.0032 | 视频 | 0.0032 | 女神 | 0.0031 | 发型 | 0.0030 |
| 旅游 | 旅行 | 0.0224 | 攻略 | 0.0188 | 旅途 | 0.0102 | 见闻 | 0.0060 | 旅游 | 0.0057 |
| | 故事 | 0.0056 | 美食 | 0.0050 | 景点 | 0.0037 | 吃货 | 0.0031 | 最美 | 0.0029 |
| | 收藏 | 0.0028 | 实用 | 0.0028 | 转起 | 0.0027 | 丽江 | 0.0026 | 最全 | 0.0024 |
| | 童鞋 | 0.0024 | 拍照 | 0.0023 | 晚安 | 0.0022 | 美景 | 0.0022 | 韩国 | 0.0021 |
| 军事 | 中国 | 0.0096 | 航母 | 0.0060 | 美国 | 0.0057 | 海军 | 0.0048 | 南海 | 0.0045 |
| | 美军 | 0.0040 | 朝鲜 | 0.0034 | 导弹 | 0.0030 | 俄罗斯 | 0.0030 | 日本 | 0.0023 |
| | 解码 | 0.0023 | 战机 | 0.0022 | 台湾 | 0.0020 | 部署 | 0.0018 | 军情 | 0.0018 |
| | 头条 | 0.0017 | 印度 | 0.0017 | 潜艇 | 0.0016 | 演习 | 0.0015 | 北京 | 0.0015 |
| 音乐 | 一首歌 | 0.0140 | 音乐厅 | 0.0127 | 翻唱 | 0.0104 | 私人 | 0.0077 | 演唱会 | 0.0077 |
| | 演唱 | 0.0058 | 单曲 | 0.0056 | 现场版 | 0.0056 | 好听 | 0.0055 | 首歌 | 0.0052 |
| | 音乐 | 0.0047 | 歌手 | 0.0045 | 关注 | 0.0037 | 专辑 | 0.0034 | 主题曲 | 0.0033 |
| | 一首 | 0.0031 | 歌曲 | 0.0028 | 现场 | 0.0026 | 想不到 | 0.0026 | 首播 | 0.0025 |
| 萌宠 | 主人 | 0.0110 | 狗狗 | 0.0105 | 汪星 | 0.0080 | 主子 | 0.0070 | 猫咪 | 0.0062 |
| | 网友 | 0.0046 | 宝宝 | 0.0045 | 星人 | 0.0043 | 金毛 | 0.0034 | 宠物 | 0.0031 |
| | 视频 | 0.0029 | 可爱 | 0.0026 | 短腿 | 0.0025 | 二哈 | 0.0023 | 柯基 | 0.0022 |
| | 哈士奇 | 0.0022 | 主银 | 0.0022 | 真的 | 0.0020 | 技能 | 0.0019 | 表情 | 0.0019 |

从表中我们可以看到一共提取的 10 个主题分别是：健身、美食、数码、体育、美妆、旅游、军事、音乐、萌宠和游戏，每个主题下有 20 个主题词，有些主题词会出现在不同主题中。例如“美食”这个词既出现在主题“美食”中，也出现在主题“旅游”中，因为在旅游的过程中必不可免地会涉及到美食。每个主题下的主题词还是较为合理的，例如“健身”这个主题下的前 5 个主题词为“动作”、“健身”、“马甲”、“教程”、“瑜伽”，都是与健身较为相关的。

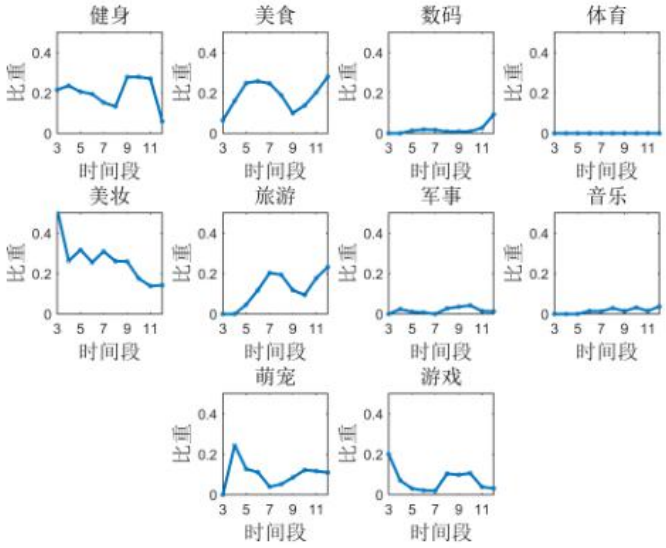
这 10 个主题是从 100 个用户的微博中提取出来的，但是每个用户的属性特点都不同，不同用户在 10 个主题的分布情况也是不一样的。同一个用户在 10 个主题的权重也不一样，可能对某些主题很感兴趣，但是对某些主题不存在很浓厚的兴趣。我们随机挑选了三个用户第一个月的数据，每个主题中不同用户的兴趣程度如图所示。




从图中可以看到不同用户的属性分布很不一样，用户 1 对于美妆方面的东西最感兴趣，其次是美食和健身以及旅游，对其他方面感兴趣程度相对较低，但是总体来说兴趣点还是较分散的，没有集中于某一方面；用户 2 最感兴趣是健身，其次是美食和美妆，其他都不怎么感兴趣。而且对于健身的喜爱程度相比其他几个高了好多；用户 3 对于军事最感兴趣，而且感兴趣的程度很高，可以认为在这 10 个属性中，对军事是独爱。这三个用户是随机挑选的，三者之间的兴趣相差很大，每个用户的属性分布相差很明显。

2、用户属性动态变化

用户的属性是会随着自己内心的兴趣点的变化以及外部环境因素的影响而改变的。如果用户属性不改变，那么他每个月的属性都是一样的。我们随机选择了两位用户来观察其 2015 年一年来每个月属性分布的变化，前两个月的数据用来平滑数据，因此从第三个月来开始展示用户属性的变化。用户 1 的属性动态变化如图所示。

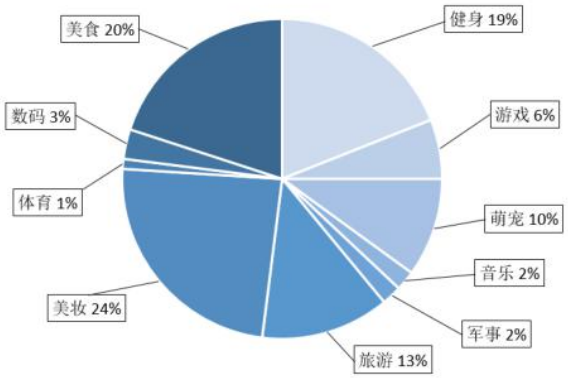



从图中可以看到该用户对于数码、体育、军事和音乐四个方面这一年来都没有感兴趣；对于萌宠和游戏方面感兴趣度低，但是兴趣变化平平，并不明显；但是在旅游和美食方面一开始关注度并不高，之后就处于上升状态，虽然中途也有降低，但是那个有可能是数据稀疏导致的，也就是这一年该用户开始关注美食和旅游；对于美妆和健身方面的关注度都处于下降趋势。总的来说，该用户这一年兴趣的变化还是挺明显的，对美食和旅游关注度有所上升，

对于健身和美妆的关注度有所下降。但是以第一个月的数据作为用户的属性挖掘，不及时更新就会无法准确地得到用户的用户画像。

3、个例分析

现在我们以单个用户为例来进行实验结果的展示。我们从众多用户中随机挑选一位用户，首先引入衰减函数机制来挖掘他最近的属性，其 10 个主题的权重分布如图所示。




从图上可知该用户主题的权重比较大的是美妆，占了 24%，其次是美食和健身，分别为 20%和 19%。其他几个主题的权重相对而言比较小。如对体育、音乐和军事感兴趣程度趋于 0。

接下来要进行主题匹配，设置的阈值为 0.19，则对上图的权重值进行筛选得到该用户的属性有“健身”、“美食”、“美妆”，即得到该用户动态的属性向量为 L1= [1,1,0,0,1,0,0,0,0,0]。该向量的意思是通过动态用户画像模型我们得知该用户对健身、美食和美妆比较感兴趣。

计算动态用户画像模型和静态用户画像模型的准确率、召回率和 F1 值如表所示。

| | 准确率 | 召回率 | *F1* 值 |
|---|---|---|---|
| 静态用户画像模型 | 0.6667 | 0.6667 | 0.6667 |
| 动态用户画像模型 | 1 | 1 | 1 |

从表格中的结果可以看到动态用户画像模型的结果比较好。其实因为是单个用户，通过微博我们就可以看到动态用户画像建模挖掘的属性比较准，而静态用户画像模型有一个属性“美食”没有挖掘出来，并且错挖掘了属性“游戏”。其实，静态用户画像挖掘的属性有可能是用户之前的属性，但是用户的兴趣点是在改变的，之前喜欢游戏，有可能现在不喜欢玩游戏了，转而关注美食了。那么，用静态用户画像模型来挖掘用户的属性就显得不合适，不能用之前的属性来描述用户现在的属性。

该用户的动态用户画像建模就基本完成了，我们成功地得到了她的最近一段时间的兴趣点是健身、美妆和美食。这个结果就可以用来给她推荐相关的微博，让她关注这些领域的专业人士等，也可以用于其他的个性化服务。