知识图谱与用户画像

- 1 介绍
- 用户画像
- 用户画像和知识图谱
- 2 数据收集以及数据建模
- 数据来源
- 数据建模
- 3 知识图谱和用户画像的结合
- 构建细粒度的用户知识图谱
- 通过用户浏览信息中找到关联关系
- 通过时序信息刻画用户的兴趣转移
- 构造通过知识图谱构建兴趣点之间的关系
- 热门视频对用户画像的影响

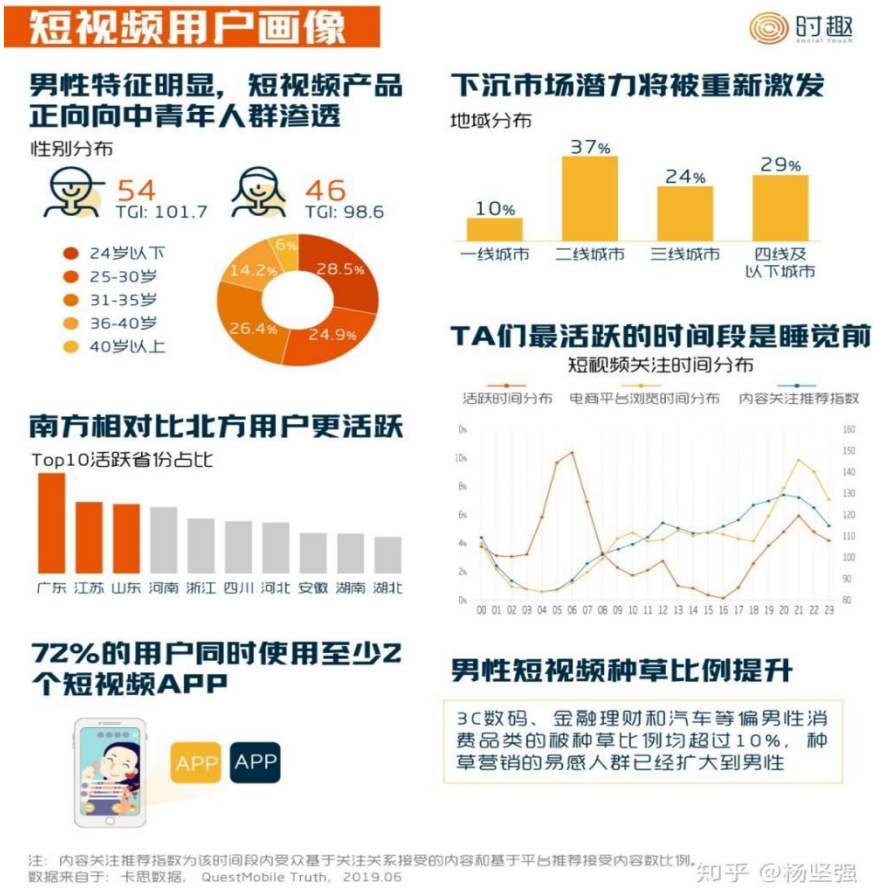
1 介绍

用户画像技术是把用户的特征更具象化表达；知识图谱可以把不同的特征的关联关系具象化表达成图。下一部分的内容是定义用户画像以及它的使用场景和知识图谱在用户画像上的应用。

1.1 用户画像

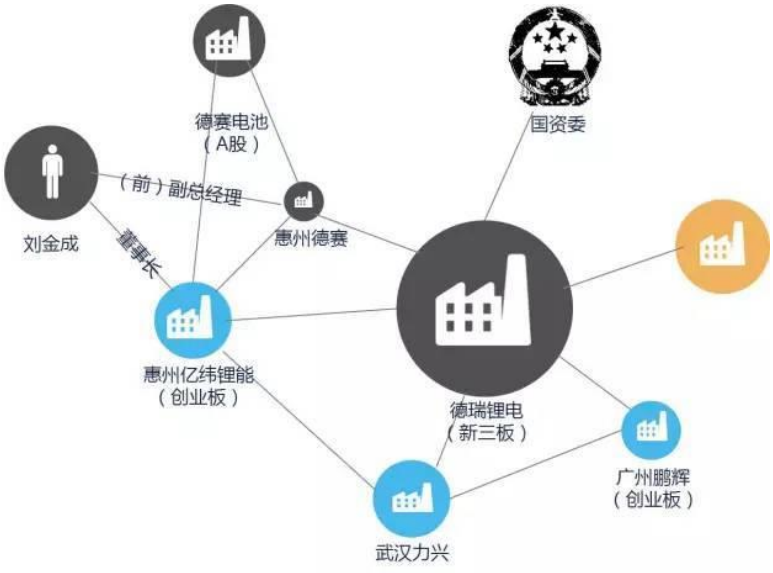
用户画像技术是真实用户的虚拟代表，是建立在一系列现实世界中真实用户数据之上的数模型：对用户的社会属性，个人爱好，生活习惯和消费行为等数据进行采集和积累，并且在明确的业务应用场景下，根据提前设定好的算法对符合业务需求的特定用户的消费目标,行为习惯和观点等方面进行画像和分析。将用户多种类型的数据抽象成一个标签化的用户模型、以挖掘深层次，能触及用户需求的信息。基于画像的标签方法就是为用户打一组标签，每个标签给一个权重，权重代表了用户在这个方面兴趣的强烈程度。比如豆瓣用户的标签云等。

在网购，娱乐乃至于金融场景下，用户画像都在数字化运营，兴趣发现以及推荐中起到了不可或缺的作用。
图例 1：用户画像在短视频平台上的应用。



1.2 用户画像和知识图谱：

基于知识图谱的用户画像技术利用采集到的大量真实用户数据，包括用户的手机 APP 行为数据，浏览器搜索词，娱乐消费数据等，构建面向用户画像的知识图谱。利用知识图谱提供的实体与实体之间的语义相似性和逻辑相似性，计算生成知识图谱的所有语料的词语与知识图谱中实体之间的相关性，得到语义有关的知识实体。同样计算得到相关实体与已知用户行为标签相似的标签表，并通过组合计算得到与标签对应用户的相关性的强弱，从而生成可以表示用户特性的用户行为标签关联组合。



图例 2：从金融数据中构建企业关系中的知识图谱

2. 数据收集以及数据建模：

类似于其他机器学习任务，用户画像任务是一项对数据依赖严重的任务。用户信息标签化，是企业通过收集与分析用户的社会属性，生活习惯，行为特征等，抽取出用户的商业全貌，实现精准营销。

2.1 数据来源：

1 人物属性数据：用户年龄，性别，所在省份城市，受教育情况，婚姻状况，生育状况，工作情况等。
2 行为数据：用户活跃度（比如用户更倾向于文章类别）
3 其他数据：用户的消费能力，用户的手机设备

2.2 数据建模：

1 关键词列表表示法（用户的兴趣有多个用户感兴趣的关键词所构成的关键词序列来表示）
2 基于布尔模型的表示方法（每个关键词在文档中只有两种可能，包含或者不包含。布尔模型的构造简单，易于理解，但是缺点明显）
3 基于向量空间的表达方法（基于向量空间模型的表示方法是指关键词和关键词权重的向量来表示用户兴趣模型）
4 基于本体的用户模型表示方法（给出构成相关词汇的基本术语和关系，利用这些术语和关系构成这些词汇的外延关系）基于知识图谱，可以发现用户兴趣关键词的上位词，从而进行推荐

3 知识图谱和用户画像结合

目标是通过知识图谱对于用户兴趣的发现，把用户的兴趣趋势挖掘出来。

3.1 构建用户兴趣细粒度的知识图谱。

一种构建基于用户兴趣细粒度的知识图谱方法。把现有开眼视频分类作为基础公司自行开发构建。例如：体育下可以被分为体操，健身，台球等。台球下可以分为中式九球和斯诺克等相关 tag。（优点：更加定制化；缺点：标注时间比较长）另一种方法是使用第三方知识图谱，现有知识图谱商用化案例有知立方等。（优点：比较节约时间；缺点：不够定制化）

3.2 通过用户浏览信息中找到关联关系

目的是通过用户浏览视频的历史信息，发现中等颗粒的标签，并且把带有中等颗粒标签的视频推荐给用户。假如体育大类可能包含太多种类的视频，比如滑雪视频，潜水视频。最好的情况是不要把滑雪的视频推荐给潜水的爱好者了；同时不要把潜水的视频推荐给滑雪的爱好者。

3.3 通过时序信息刻画用户的兴趣转移

现在传统用户画像中有些时候无法描述用户的兴趣变化，但是这是在现实情况下，用户很容易变化。设计用户标签的做法中最好加入时间衰减，最早点击的标签的时间衰减系数越高。

3.4 构造通过知识图谱构建兴趣点之间的关系（重点讲解）

通过兴趣标签之间的关联关系，使用 Trans-X 系统模型可以获得关系 embedding 和顶点 embedding 的两种 embedding 向量。可以通过计算兴趣标签的 embedding 之间的向量夹角，返回相类似兴趣标签推荐给用户。

3.5 热门视频对用户行为画像的影响

视频网站上，有很多视频是所有人都会点击观看的热门视频。这一定会对用户画像的结果产生影响。构建用户画像的时候，同时要降低热门视频对用户画像产生的影响。