

Figure : image-20211014173327646

# 用户画像知识图谱

用户画像是特定用户的个人信息呈现，通常由一系列精确的属性及数值表示。以下主要介绍如何以图谱方式构建用户画像。

## 用户画像知识表示

Chatbot领域的用户画像图谱相对适合**自顶向下**的构建方法。

将每一个用户都看作一个实体，用户的基本属性视作实体属性，与该用户关联的人，事，物看做相连的实体。

## 知识抽取和挖掘

### 数据来源

数据来源方式多种多样，想从中获取所需信息，有**显式**和**隐式**两种方法。

* 显式获取：由用户数据直接获取的信息。例如年龄，性别，所在城市等。
* 隐式获取：是指不能直接获取用户数据，需要结合分析归纳或利用机器学习模型隐式获得的信息。如注册账户为男性，但是浏览商品都是女性，因此需要进行分析推断。

为了更好地补全用户画像，需要进行跨平台的用户信息融合。

以智能音箱产品为例：用户在使用的过程当中，需要下载相应的APP来绑定音箱，在APP中控制音箱，也可以与它直接进行语音交互，可以利用的数据就有： \* 人机交互语音数据： 对语音数据进行特征抽取，辨别出用户的性别等… \* 人机交互文本数据：例如日志log数据，将其从语音转换为文本，利用信息抽取获得信息 \* 触发的音箱功能的数据：通过用户的行为轨迹来挖掘用户的喜好特征 \* 用户的位置信息：根据地理位置进行推荐

### 知识抽取方法

大多是从非结构化的文本中抽取知识的过程，本质上是一个信息抽取的任务。

信息抽取包含：命名实体识别(NER)，关系抽取（RE）,事件抽取（EE)等子任务；还需要一些自然语言处理的技术：中文分词，词性标注，情感分析，句子主干提取等。 常用的工具包：斯坦福的CoreNLP，哈工大的LTP

定制领域特殊的NLP模型需要的自然语言处理技术有：传统的基于规则的方法，分类模型方法，序列标注模型方法，半监督学习方法，关键词提取方法。

#### 基于规则的方法

基于规则的方案： 一方面是因为规则方法在特定任务下准确度高，方便定制，另一方面是因为很多任务在初期没有训练数据，很难解决冷启动的问题。

规则模板通常需要由领域专家撰写，便于场景化的快速定制，且无须积累大量数据去训练模型。

#### 分类模型方法

基本思路如下： \* 确定分类的标签：尽可能让所有标签包含一个全集的情况，并且标签和标签之间的边界尽可能明确。 \* 特征选择：从词汇，短语，句子的角度出发， \* 词汇层面的常用特征有词性、词频等； \* 短语层面的常用特征有是否特定固定搭配、短语语法结构； \* 句子层面的常用特征有句法分析、依存分词等 \* 训练分类模型

#### 序列标注模型：

应对关系抽取和事件抽取等任务时会很有用，核心是**为输入的文本序列预测一个隐藏状态序列，而这个隐藏状态序列会标注出句子中哪些是待抽取的内容，哪些是无关的文字**。

常用的状态标签体系：BIO（Beginnig-Inside-Outside），BMES(Begining-Middle-End-Singleton)，BIOES（Begining-Inside-Outside-End-Singleton）

不同的任务会选择不同的标签体系，如分词、词性标注任务的每一个词均有对应的状态标签，因此通常选择BMES标签体系，而抽取类任务大多采用BIO或者BIOES标签体系。

经典模型：隐马尔可夫模型（HMM）、最大熵马尔可夫模型（MEMM）和条件随机场（CRF）模型， 基于统计的机器学习方法更优的序列标注模型，包括卷积神经网络（Convolutional Neural Network）、循环神经网络（Recurrent Neural Network）、注意力机制（Attention Mechanism）、Transformer以及它们的变种。 一些高质量预训练模型进行微调开展下游序列标注任务：BERT、XLNet、GPT，业界效果最好：LSTM+CRF

#### 半监督学习

在没有足够的训练数据来进行有监督学习模型的训练时，使用半监督学习来实现。

* 基于种子的启发式方法（Seed-based或Bootstrap）
* 远程监督方法（Distant Supervision）

##### 基于种子的启发式方法

准备一些高质量的实体-关系作为初始种子(Seed)，同时准备一个大规模语料库提供种子进行模板(pattern)学习。 学习的循环步骤：

* 以种子为基础，从大规模语料库中匹配
* 对这些句子的上下文进行分析，提取可靠的模板
* 通过模板去匹配语料，发现的更多种子
* 通过新抽取的实例挖掘更多模板 — 满足预设的收敛条件时，可以设置不再发现新的实体或者模板作为收敛条件。

该方法**构建成本低，适合进行大规模的数据构建**，并且可能发现新的隐含关系。 但是它**对初始种子的质量要求高，总体准确率较低**，在不断迭代过程中可能会发生**语义偏移**的情况，需要较多的条件约束来控制提取出来的数据的质量。

##### 远程监督方法（Distant Supervision）

由Mintz提出，结合了**监督学习和基于种子的启发式方法**的优点，用于做**关系抽取**任务。它能够处理关于监督学习时缺少标注数据的问题。

该方法使用在知识抽取的前提假设是：**如果两个实体之间存在某种关系，那么所有同时提到这两个实体的句子都能够描述这种关系**。其核心思想是，**利用实体在语料中抽取潜在关系，再用关系反向定位抽取实体**。

流程大致如下： \* 基于远程监督方法获取大量的标签数据，再使用机器学习、深度学习等有监督学习方法训练分类器，接着对程序自动标注的数据进行划分，将质量较高的自动标注数据加入训练数据，丢弃质量较低的部分给人工。

缺点： \* 前提假设并不总是成立的，会在学习时进入很多噪声，在后续的训练集中产生大量的wrong label。 \* NLP在前者产生的训练数据上工作，可能是的错误不断传递。

资料：[1] Distant supervision for relation extraction without labeled data 将远程监督使用到了关系抽取任务当中。 关于远程监督和标签的解释和Demo：https://mp.weixin.qq.com/s/BxzSbr4vh2Vv4mUC–l3Rg

##### 关键词提取

关键词提取主要应用于**信息抽取和文本挖掘**领域，**通过几个关键词将用户核心关注的信息提取出来**，从而更准确地表达文本信息。

主要分为两类：

* 监督学习方法：
  + 转换为分类问题
  + 转化为序列标注问题
* 无监督学习：
  + 基于统计特征：选择一些统计特征作为词语的关键性衡量指标，然后根据评分来对候选词进行排序。统计特征有 词语在文本中的权重和 词语之间的关联度(TF-IDF)等。
  + 基于词的图模型：构建文档的语言网络图，以词为节点，词与词之间的关系为边，权重由关联度来表示。
  + 基于主题模型：如果某一篇文章围绕一个中心思想，那么文章一定会有特定词语频繁出现。从数学角度出发，统计文章词语，判断包含几个主题和占比。常用：LDA,LSA,pLSA,Ida2vec

## 图谱建立

知识图谱分为两种构建方式： \* 自底向上：需要对所有的实体进行类别归纳，先归纳成最细致的小类，然后逐层往上，形成大类概念，该方式普遍适用于通用知识图谱的构建。 \* 自顶向下：需要为图谱定义数据模式，并从最顶层的概念开始定义，逐步往下进行细化，形成类似树状结构的图谱模式，最后将实体对应到概念中，此类构建方式通常适用于领域或者行业知识图谱的构建。

## 信息抽取案例

以音乐推荐作为案例，使用开源的CRF(条件随机场)作为信息抽取的模型。

领域命名实体识别的方法：基于词表匹配的方法，基于规则的方法，基于机器学习的方法，基于深度学习的方法。 \* 词表匹配：基于词典做最大匹配的方法快速匹配出文本中所包含的实体信息。 \* 基于规则：基于规则的方法通常需要编写特定规则模板来抽取所需要的实体，规则通常由领域的业务人员和熟悉数据的开发人员一同制定。 \* 基于机器学习：是NLP任务中典型的序列标注问题，可以通过机器学习中的序列标注模型来训练一个领域下的NER模块。 \* 基于深度学习：需要更多训练数据，但无需人工构造特征，而且网络自动学习语义特征的效果往往更好。

### 环境安装

下载CRF：[Google网盘地址](https://drive.google.com/drive/folders/0B4y35FiV1wh7fngteFhHQUN2Y1B5eUJBNHZUemJYQV9VWlBUb3JlX0xBdWVZTWtSbVBneU0?resourcekey=0-NW5cPRv1Xr2-Vfo_xlDTLQ)

下载其中0.58版本的tar.gz包，因为在CentOS下进行。

解压： tar -zxvf CRF++-0.58.tar.gz 编译安装：

cd /usr/local/CRF++-0.58  
./configure  
make  
sudo make install

配置ld.so.conf:

sudo vim /etc/ld.so.conf  
  
//添加一行：  
include /usr/local/lib  
  
//加载  
/sbin/ldconfig -v

不进行上面的步骤可能导致：ImportError: libcrfpp.so.0: cannot open shared object file: No such file or directory

查看CRF++版本：

crf\_learn -v

安装python的CRFPP包：

cd CRF++-0.58/python  
python3 setup.py build  
python3 setup.py install

测试是否安装成功：

//命令行输入：  
python  
  
>> import CRFPP

如果引入时报错ImportError: libcrfpp.so.0: cannot open shared object file: No such file or directory，则建立链接：ln -s /usr/local/lib/libcrfpp.so.\* /usr/lib64/，可能需要sudo。

### 训练

[参考使用教程](http://www.huaxiaozhuan.com/%E5%B7%A5%E5%85%B7/CRF/chapters/crfpp.html)

模型训练：

crf\_learn 模板文件 训练文件 模型名

模型测试：

crf\_test -m 模型 测试文件 > 测试结果文件

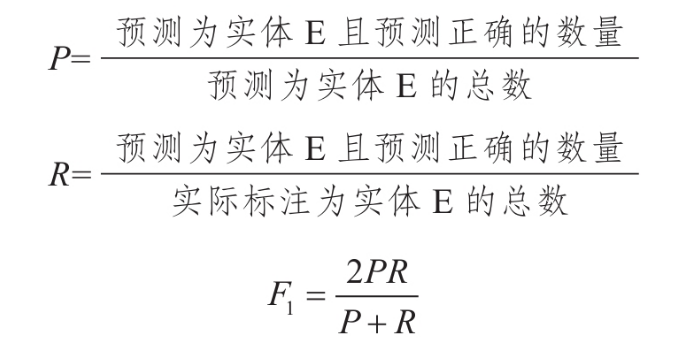
结果文件分三列：原句子的字符，字符的标准答案标签，模型预测的结果标签。

数据准备： 训练集：music\_ner\_train.txt，测试集：music\_ner\_test.txt，运行指令crf\_learn

按照三种实体：歌曲song—-SO，歌手artist—-AR，曲风genre—-GE进行标注。训练集：测试集 = 8：2

标签含义：

* 以‘B’开头说明是一个实体的开始，记录实体所在的句子行号line\_num，实体开始位置index，实体字符，实体标签；
* 以‘I’开头，表示处于一个实体中间，则只要添加当前字符到实体字符串中即可；
* 以‘E’开头，表示处于实体结束为止，添加字符，再将记录的信息添加到matched列表，并初始化one；
* 以‘S’开头，说明该字符单字成实体，该字符既是实体的开始，也是实体的结束，需要同时记录one的每一项信息后，添加到匹配列表，并重新初始化one。

对于模型的评估分三类：精确率(Precision,P)，召回率(Recall, R)，F1值(F1-score)。 

预测的返回结果以实体三元组列表的格式展示：，三元组的第一位是实体内容，第二位是实体在句中的开始位置，第三位是实体的类别标签。