

Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心



用户画像建模

硕士研究生 祁佳俊

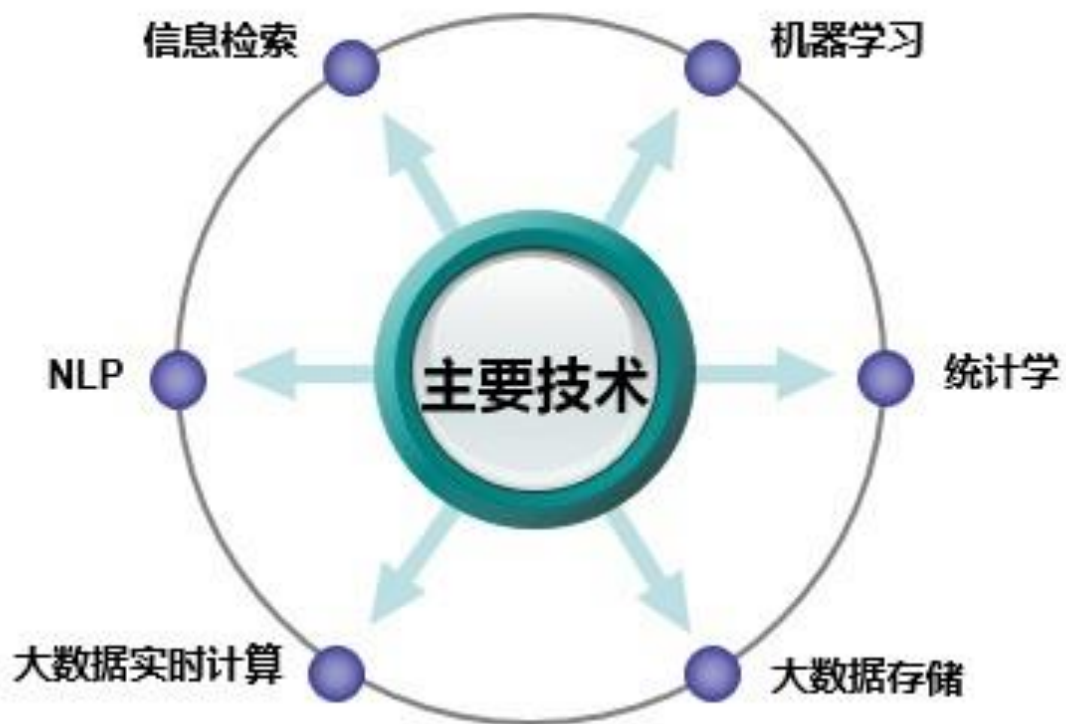
2021年03月07日

- 背景简介
- 基本概念
- 算法原理
- 优劣分析
- 应用总结
- 参考文献

- 预期收获
 - 1. 熟悉用户画像建模相关知识概念
 - 2. 了解LDA-RCC模型原理
 - 3. 了解用户画像的应用

- 用户画像（User portrait）：
 - 由交互设计之父Alan Cooper提出，是建立在一系列属性数据之上的目标用户模型。
 - 从用户群体中抽象出来的典型用户，本质是一个用以描述用户需求的工具。

- 主要技术：



用户画像建模



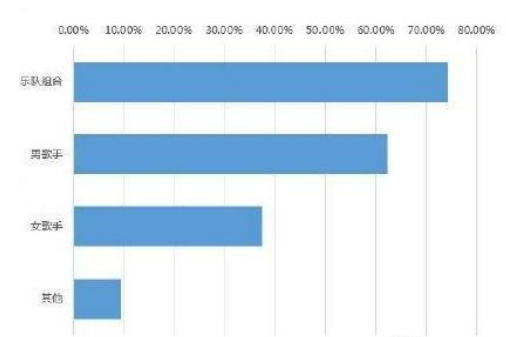
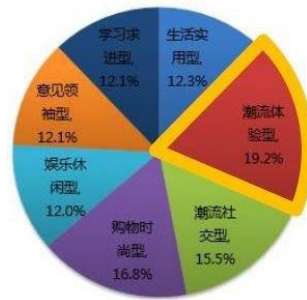
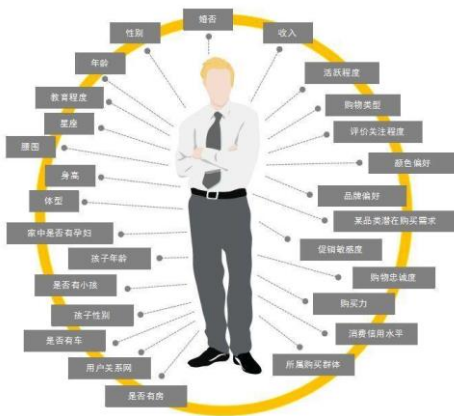
基本概念

[illegible]

- 什么是用户画像？
 - Alan Cooper定义用户画像“基于用户真实数据的虚拟代表”；
 - 一个从海量数据中获取的、由用户信息构成的形象集合。
- 用户画像**定义**：
 - 一个或一类真实用户的虚拟抽象，是基于一系列实际数据的虚拟用户模型；
 - **用户信息标签化**。



- 为什么构建用户画像？
- 便于计算机处理；
 - 分类统计：统计喜欢红酒用户的人数以及性别分布
 - 关联规则提取：喜欢红酒的用户还喜欢什么？
 - 聚类算法：喜欢红酒的用户年龄分布
- 实现数据可视化。



- 如何构建用户画像？
 - 用户数据分析：对用户相关数据进行分类，分类思想：封闭性的分类方式；
 - 静态信息数据：用户相对稳定的信息，eg：用户身份属性
 - 动态信息数据：用户变化的行为信息，eg：用户浏览网页信息
 - 用户画像建模：根据用户信息，构建用户画像模型，产出标签权重。
 - 标签：表征内容，用户偏好、需求的内容
 - 权重：表征指数，用户的需求度

- 用户画像分类
- 方法层面：
 - 定性画像：定性描述+类型细分（可深入挖掘使用场景，但缺少数据验证）；
 - 定量画像：数据收集+聚类分析（有数据佐证，可分析得到精确数据结果，但统计要求高，难以挖掘用户情感倾向）；
 - 定性+定量画像：类型细分+数据验证（工作量大，成本高）。
- 应用层面：
 - 健康画像；
 - 企业/个人信用画像；
 -

- 论坛或评论区的文本都是短文本；
- 问题：传统文本建模方法从短文本中提取主题时存在严重的稀疏性问题。



LDA-RCC模型

用户画像建模



算法原理

T	从用户发布的短文本中构建用户画像
I	论坛用户发布的短文本数据
P	1.文本预处理; 2.用户标签提取; 3.用户标签聚合; 4.用户画像生成
O	用户画像

P	解决传统文本建模方法从短文本中提取主题时存在的严重稀疏性问题
C	模型根据用户的数据自动提取主题个数，并分析用户偏好
D	提出动态文本建模LDA-RCC模型
L	ICANN 2020

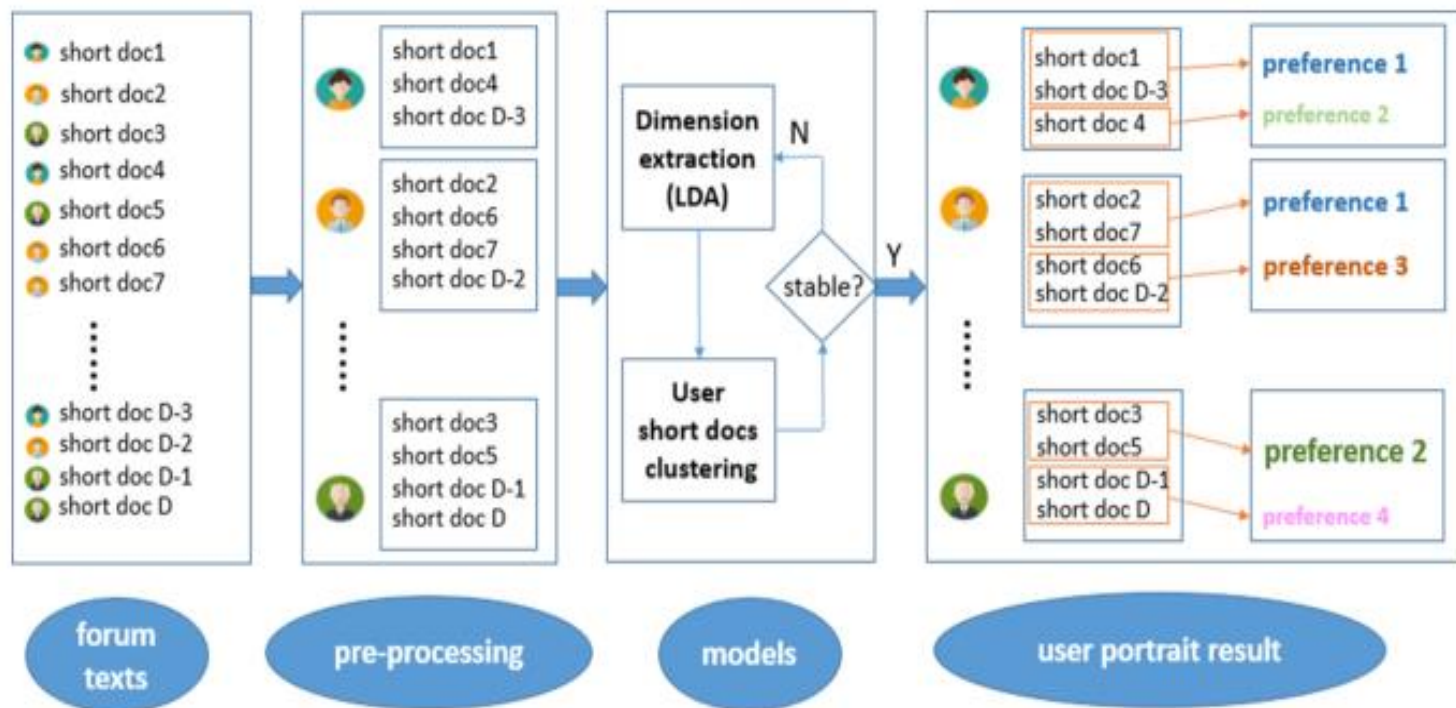
- 核心思想：
 - 提出基于实际文本的动态文本建模LDA-RCC模型，解决从短文本中提取主题时存在的稀疏性问题，用于分析用户偏好，构建用户画像；
 - LDA-RCC组合模型通过文本建模方法-潜在狄利克雷分配（LDA）和鲁棒连续聚类方法（RCC）的迭代组合，能有效的处理短文本，并自动提取主题个数，通过聚类分析，得到用户偏好。

- 总体流程：

- 1.数据预处理；
- 2.同一用户短文本聚合为一个伪文档；
- 3.利用LDA算法得到每个短文本的主题及概率；
- 4.采用RCC聚类方法将同

一用户主题相似的短文本聚集到一个集群中，并合成新的长文本伪文档；

- 5.计算前后两次聚类结果的归一化互信息（NMI），若大于0.6，停止迭代，输出主题作为用户画像标签；
- 6.转入步骤3，合成的新的长文本伪文档作为LDA算法输入。



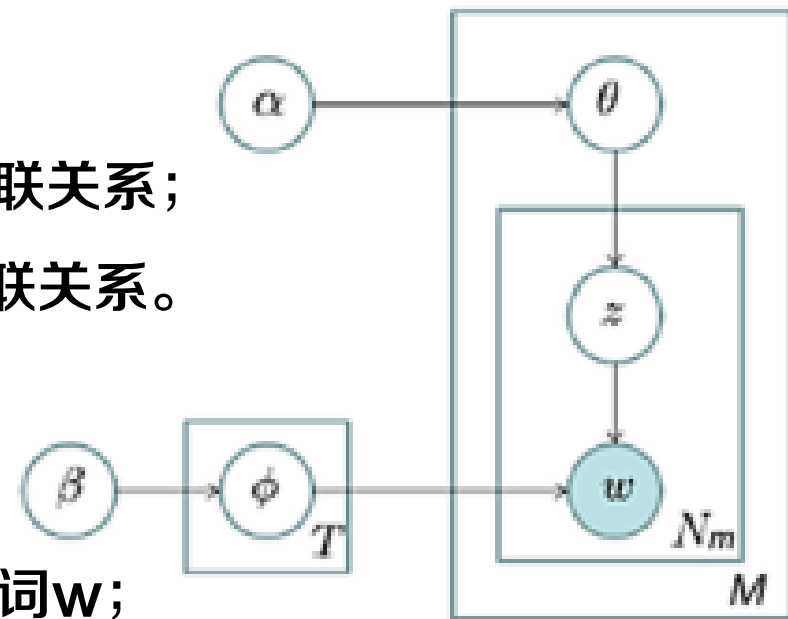
- 文本主题提取：

- LDA模型

- α 和 β 是服从狄利克雷分布的超参数；
 - 参数 θ 表示每个伪文档和 T 个主题之间的多项式分布关联关系；
 - 参数 ϕ 表示每个主题与 N_m 个单词之间的多项式分布关联关系。

- 生成过程：

- 1.对每篇文档，从主题分布 θ 中提取一个主题 z ；
 - 2.从上述抽取的主题所对应的单词分布 ϕ 中抽取一个单词 w ；
 - 3.重复上述过程直至遍历文档中的每一个单词。



- 文本主题提取：
 - 生成结果显示在三个矩阵中：
 - MT矩阵，描述每个伪文档所属的主题及概率；
 - VT矩阵，描述每个单词所属的主题及概率；
 - DT矩阵，描述每个短文本所属的主题及概率。
- 用户标签聚合：
 - 鲁棒连续聚类模型（RCC）
 - DT矩阵的每一行是描述短文本的向量；
 - 聚类模型针对每个用户的短文本进行聚类，得到包含一个或多个短文本的集合，集合包含短文本越多，集合对应的主题标签权重越高，用户越在意这个标签。

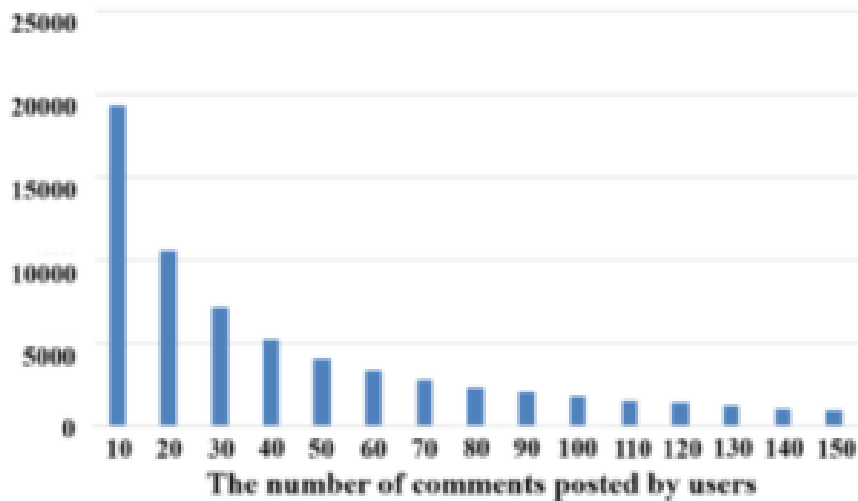
- 归一化互信息:

$$\text{NMI}(J_u^l, J_u^{l-1}) = \frac{MI(J_u^l, J_u^{l-1})}{\sqrt{H(J_u^l)H(J_u^{l-1})}}$$

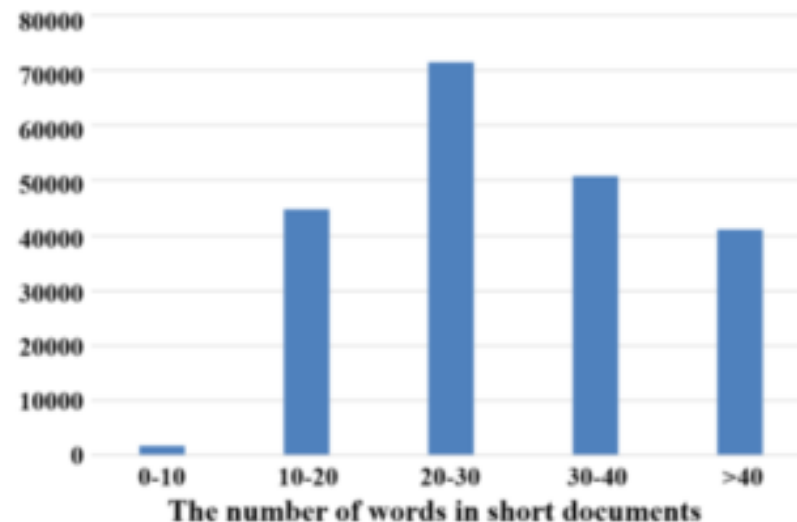
- J_u^l 是迭代第 l 次中用户 u 所有短文本主题标签的集合;
 - J_u^{l-1} 是迭代第 $(l-1)$ 次中用户 u 所有短文本主题标签的集合;
 - $MI(J_u^l, J_u^{l-1})$ 是用户 u 所有短文本主题标签迭代前后两次的互信息;
 - $H(J_u^l)$ 和 $H(J_u^{l-1})$ 是用户 u 两次迭代的聚类结果的熵。
- 当每个用户 NMI 值大于 0.6 时, 用户标签已经稳定。

- 数据集：
 - 在汽车之家网站上爬取的用户评论数据；
 - 用户数量和用户发布短文本数直方图：每个用户发布多条文本；
 - 短文本数量和包含字数直方图：文本包含字数不多，属于短文本。

The number of users



The number of short documents



- **实验结果1:**
 - 三个用户的词云图;
 - 主题标签包含的短文本数越多，词云中的对应标签就更大更醒目。



- 实验结果2——聚类结果分析：
 - 一个用户的三个集合中六个短文本的主题分布；
 - 同一集合中的主题分布相似，不同集合中的主题分布不相似。

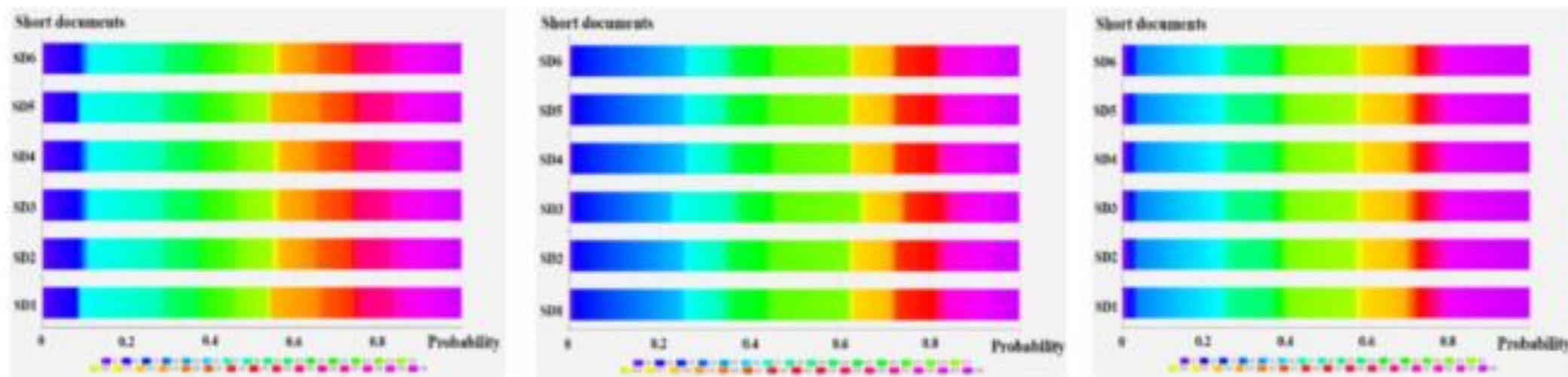


Fig. 6. Topic distribution of short documents in cluster A, B, C of user u .

- 实验结果3——话题质量比较：
 - 连贯性作为衡量话题质量的指标，越小质量越高
 - LDA直接处理短文本；
 - LDA处理同一用户短文本聚合的伪文档；
 - DMM算法处理短文本得到主题，并对短文本聚类。

Table 2. The coherence values of three algorithms.

Topic Number	LDA (with short texts)	LDA (with pseudo-documents)	DMM	LDA-RCC
20	−198.2602	−157.9289	−136.4822	−81.6170
30	−196.3452	−156.1905	−134.6653	−77.6984
40	−195.7384	−155.0746	−114.8646	−81.6295
50	−193.2462	−151.5630	−122.5287	−82.0068

- 实验结果4——用户画像对比：
 - 四个用户采用两种模型得到的用户画像，左：LDA-RCC，右：DMM；
 - LDA-RCC产生的词云密度更大，显示的信息更多
 - LDA-RCC对用户画像的处理更精确，DMM相对较粗糙。



Fig. 7. Comparison results of the two algorithms.

用户画像建模



优劣分析

- 优势：
 - 解决从短文本中提取主题时的稀疏性问题；
 - 提取的话题质量有较大提升；
 - 构建的用户画像更精确。
- 劣势：
 - 引入聚类算法增加模型运行时间。

- 用户画像应用：
 - 精准营销/个性服务：根据用户特征，分析用户潜在需求，通过短信、邮件营销，或提供个性化服务；
 - 数据挖掘：以用户画像为基础构建推荐系统、搜索引擎或广告投放系统；
 - 行业报告/业务决策：通过用户画像分析行业动态，例如人群消费习惯和地区消费差异，以指导关键决策；
 - ...

- [1] DING Z, YAN C, LIU C, et al. Short Text Processing for Analyzing User Portraits: A Dynamic Combination, Cham: Springer International Publishing, 2020.
- [2] WU T, YANG F, ZHANG D, et al. Research on Recommendation system based on user portrait: 2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIS)[C], 2020.March.
- [3] https://blog.csdn.net/trouble_maker_czt/article/details/82757005
- [4] https://www.sohu.com/a/231199100_295682

谢谢！

大成若缺，其用不弊。大盈
若冲，其用不穷。大直若屈。
大巧若拙。大辩若讷。静胜
躁，寒胜热。清静为天下正。

