硕士学位论文

面向暗网的用户画像构建技术研究与应用

RESEARCH AND APPLICATION OF USER PROFILE TECHNOLOGY FOR DARK NET

车馨悦

哈尔滨工业大学

2020 年 6 月

国内图书分类号：TP393　　学校代码：10213
国际图书分类号：004.9　　密级：公开

工学硕士学位论文

面向暗网的用户画像构建技术研究与应用

硕 士 研 究 生：车馨悦
导　　师：李东 教授
申 请 学 位：工学硕士
学　　科：网络空间安全
所 在 单 位：计算机科学与技术学院
答 辩 日 期：2020 年 6 月
授予学位单位：哈尔滨工业大学

Classified Index: TP393
U.D.C: 004.9

Dissertation for the Master Degree in Science

RESEARCH AND APPLICATION OF USER PROFILE TECHNOLOGY FOR DARK NET

Candidate： Che Xinyue
Supervisor： Prof. Li Dong
Academic Degree Applied for： Master of Engineering
Speciality： Cyberspace Security
Affiliation： School of Computer Science and Technology
Date of Defence： June, 2020
Degree-Conferring-Institution： Harbin Institute of Technology

摘　要

近年来，为防止匿名用户在网络中进行散布不实信息，损害他人名誉或煽动恐慌情绪等网络犯罪行为，保护和引导健康的网络环境，网络安全法新增了网络用户实名制的规定。然而暗网目前仍然处于国家网络安全监管的盲区，大量不法分子利用暗网穿上“隐身衣”，进行各类非法行为。因此在“看不见摸不着”的暗网中挖掘用户信息，构建出暗网用户画像对暗网监管具有重要意义。

目前针对用户画像的研究大都只面向表层网络的用户，但在匿名、隐蔽、结构性弱的暗网中，用户画像的相关研究极少，构建出较为丰富用户画像更具有一定困难。本文将面向暗网，针对如何获取用户数据，通过稀疏的信息构建出暗网用户画像这一问题进行研究并加以应用。

本文搭建了基于 Tor 的暗网采集系统，获取用户在暗网中留下的用户信息、交易记录、社交言论等数据，通过部分暗网用户标识对用户数据进行关联扩充。将数据清洗后，在中英两种语言的暗网用户数据集上提取用户特征。对结构化数据进行解析，针对非结构化数据建立暗网关键词词库，结合命名实体识别、实体关系抽取等自然语言处理算法，得到由基本信息，社交行为，市场交易三个维度的用户基础属性构成用户的基础画像。为进一步挖掘用户属性，改善暗网用户属性过于稀疏的缺陷，通过建立适用于暗网数据的情感分析、立场分析、影响力计算、活跃度计算、商户销量预测等算法模型对用户的各类特征标签化，将用户基础画像扩充为深度画像。应用文中得到的暗网用户画像，基于用户画像的相似度，优化 Louvain 聚类算法实现了暗网用户虚拟群体的发现。

本文构建出包含 21 个用户属性的暗网用户画像，包括 5 个基本属性，8 个社交属性以及 8 个交易属性。对用户属性的准确率进行分析，均好于近期相关研究应用于暗网数据集的效果。

关键词：暗网；自然语言处理；文本挖掘；用户画像；虚拟群体发现

Abstract

In recent years, in order to prevent anonymous users from spreading false information in the network, damaging the reputation of others or inciting panic emotions, and other cyber crimes, the Cyber Security Law has added a real-name management method based on the user's real name.However, at present, the dark net is still in the blind zone of the national network security supervision. A large number of criminals use the dark net to put on the "invisibility cloak" to carry out various illegal acts. Therefore, it is of great significance to mine user information in the "invisible and intangible" dark network and construct the user portrait of the dark network for the supervision of the dark net.

At present, most of the researches on user portraits are only for the users of surface network, but in the anonymous, hidden and weakly structured dark net, there are few researches on user portraits, and it is more difficult to construct rich user portraits. In this paper, we will study and apply how to obtain users' data and build user profiles through sparse information for dark net.

This paper builds a Tor-based dark web collection system to obtain user information, transaction records, and social speech data left by users in the dark web, and correlates and expands user data through some dark web user IDs. After the data is cleaned, user features are extracted from the dark web user data sets in both Chinese and English. Analyze structured data, build a dark web keyword vocabulary for unstructured data, and combine natural language processing algorithms such as named entity recognition and entity relationship extraction to obtain a three-dimensional user base consisting of basic information, social behavior, and market transactions The attributes constitute the user's basic profile.

In order to further mine user attributes and solve the problem of sparse user attributes on the dark net, we tag all kinds of user characteristics through algorithms such as sentiment analysis, position analysis, influence calculation, activity calculation, and merchant sales forecast. Expand the user's basic portrait into a deep portrait.Using the dark net user portraits constructed in this paper, by calculating the similarity of the user portraits, a clustering algorithm is designed to realize the discovery of dark web user virtual groups.

In this paper, we build a dark network user profile with 21 user attributes, including 5 basic attributes, 8 social attributes and 8 transaction attributes. The analysis of the accuracy of user attributes is better than that of the recent relevant research applied to the dark net data set.

Keywords: dark net, natural language processing, text mining, user profile, virtual group discovery

目　录

第 1 章 绪　论

1.1 课题背景与研究意义

1.1.1 课题来源

本课题来源于实验室的某事业单位委托项目。目标是：对暗网网站进行大规模采集，针对暗网中的用户，利用多种文本挖掘的方法，从基本信息，社交，交易三个维度分析用户属性，以此构建出暗网用户画像，并根据用户画像相似度实现暗网用户虚拟群体的发现。

1.1.2 课题的研究背景与意义

近年来，互联网与大众的联系越来越密切，网络已经渗透到人类日常生活的政治文化、经济金融、时事热点等不同领域，线上操作变成了人们社交、消费、搜索信息等活动的主要方式，与此同时，用户的线上操作也在互联网中留下了信息丰富且数量庞大的用户数据。随着互联网不断地迭代升级，尤其是在今天，web2.0 技术已经得到极大发展，web3.0 也逐渐崭露头角，互联网中的信息量将继续飞速增长。然而，普通用户所接触的表层网络仅仅是露出水面的冰山一角，在表层网络之下更涌动着深不可测的大量暗网。

暗网的主要特征包括：高度隐匿性，交易便捷性，生态混乱，不稳定性等。隐匿性指用户在登录暗网时，需要经过严格的注册流程，并使用 Tor 等特殊浏览器，以此在洋葱路由的最深处，隐藏用户个人的身份、IP 地址、访问记录等信息。交易便捷性也是暗网的一个重要特性。在全球范围对金融交易的监管愈发严格的背景下，支持比特币、以太币等虚拟货币支付的暗网受到了罪犯们的青睐。在暗网中，各种虚拟货币支持人民币、美元等几大主流货币的兑换[1]，十分便捷。交易双方的信息在交易过程中均可以被加密，这都为犯罪分子的隐身提供了可乘之机。此外，暗网生态较为混乱。暗网用户大多具有极强的自由主义、甚至恐怖主义思想，他们发布的内容包括许多宣传极端政治、宗教思想、煽动社会动乱、颠覆国家政权等咨询，或者利用暗网传递情报、发布国家涉密信息等危害国家和社会的政治安全。

虽然暗网具有技术门槛高、自身稳定性不强等问题，但尽管如此，犯罪分子仍然对具有高隐蔽性的暗网有着十分迫切的需求，不法分子将暗网视为他们的犯罪天堂。他们在“看不见摸不着”的暗网中利用相应的技术手段躲避公安

部门的监管肆意违法犯罪。例如 2019 年 12 月，一个被公开在了暗网上的服务器被名为 Bob 和 Vinny 的两位研究人员发现，这个 ES 服务器中包含了人员数据实验室的近 30 亿条用户数据，包括用户在领英、Twitter、Facebook 等多个平台的个人资料，从这个数据源中可以提取出 4 亿多的唯一电话号码，6.5 亿个人邮箱地址，确定约 12 亿独特的人员。用户不需要任何身份验证就可以随意访问甚至下载服务器中存储的数据，约 4T 隐私数据的泄露成为了近年来最为严重的单源组织数据泄露事件之一。此外，极端组织也将暗网作为筹划恐怖活动的重要工具，他们在暗网中大肆传播极端主义思想，组织新成员招募，传授制枪制爆技术，筹集活动资金。除上述外，还有许多非法用户通过暗网穿上“隐身衣”，隐藏用户的身份，传递政治敏感情报、私售非法禁售物品，进行人口贩卖等大量非法交易。随着暗网技术的自身完善，犯罪分子们在暗网上的活动频率仍将逐渐提高，规模也将日益扩大。IntSights 在 2018 年 7 月的报告中指出[3]，利用暗网中的威胁情报可以更全面感知网络安全威胁发展态势，对重大事故提前设计应对措施，从而避免或减轻事故造成的声誉、财产甚至生命的损失。

在上述背景下，如何加大对暗网的监管力度，从暗网冗杂的数据中获取高质量的用户信息，分析威胁数据并对可疑用户进行刻画是亟需解决的问题。因此，本文依托相关项目以暗网用户数据为数据集，根据网络用户在互联网中留下的消费记录、社交内容、社会属性等数据，将用户的各类特征标签化。重点对暗网用户的基本属性、社交属性、交易属性三个维度进行提取，形成暗网的用户画像，并利用暗网用户画像实现暗网用户虚拟群体的发现。本文构建的暗网用户画像还可以应用于人物分析、搜索或追踪等领域，在安全层面，更有助于暗网舆情的监管和引导，情报获取与研究，反恐工作的有效侦查。对帮助安全人员预测或及时发现网络威胁具有重要意义。

1.2 国内外相关研究现状

1.2.1 暗网威胁数据获取研究现状

目前，越来越多的学者对暗网数据的采集获取展开研究，暗网资源的获取日益受到各领域专家与相关部门的关注，某组织基于暗网中数据信息的内容，按照不同领域构建了包括：化学、网络、金融等十几个类别的目录，从而对暗网资源进行分类。面对如此庞大的暗网资源，企业界主要通过两类方法实现暗网数据的采集。一类是为不同暗网网站定制更有针对性的爬虫，这一方法在搜索引擎发展历程里始终贯，每一次的升级都与这个问题相关，从内部通过改进算法的方式解决这个难题。另一方法如谷歌推出的 OneBox 服务，百度推出的

阿拉丁计划等，主要的搜索巨头都制定了相应的计划，由搜索引擎厂商提供相应的接口解决获取暗网数据信息的困难。

威胁数据的时效性决定了威胁数据的收集必须及时、迅速[5]，如何尽早发现网络威胁情报一直是近期研究的热点问题。现有的情报数据相关研究大多关注威胁在表网中出现后的分析和溯源[6-8]，未能在其出现、甚至肆虐前预知威胁信息，难以及时控制影响和降低损失。2016 年，Nunes 等人[9]对暗网的威胁情报获取展开研究，他们首先采用机器学习的方法根据数据内容是否与攻击行为有关对根据暗网数据进行二分类，然后进行人工筛查，从恶意攻击相关的内容中发现黑客技术、恶意工具软件的介绍等威胁数据的信息。但通过人工筛查的方式发现威胁数据耗时长、更新慢、成本较高，而且对与暗网中出现的未知威胁的应对性也不够理想[10]。此外，暗网中的威胁数据种类多样，二分类结果在研究人员对多类型威胁的快速发现和获取过程存在一定局限性。2018 年，Sapienza 等人[11]的研究通过统计热点词频，以发现威胁信息。该研究中，通过监控词频获得暗网中的高频热点词，对网络安全态势作出预测。但由于信息传播需要时间，当新型威胁成为热点时，数据的时新性已经被削弱。因此为使后续用户画像的分析预测得到更好的效果，如何及时获取将暗网中敏感用户的威胁数据，并在冗杂的暗网数据中筛选高质量的数据，仍需要进一步的研究与实验。

1.2.2 人物信息抽取研究现状

关于人物信息抽取的研究中，当前已有的研究工作主要聚焦在表层网络中的人物传记和人物搜索这两个方向上。Schiffman 等人的团队[12]于 2001 年第一次提出了人物传记这一概念，之后许多学者就在人物传记这一方面进行了大量的研究，提出了针对人物传记的抽取方法，例如基于本体和面向人物追踪方法，基于多文档摘要技术的方法等等。

国外的相关研究中，Schiffman[12]采用了多文档摘要的方法来提取人物传记，该方法结合了语言学和统计学的相关理论，从多个文档抽取关于某个对象的基本信息，例如姓名，性别，学历等得基本信息，生成传记性文本。Zhou 的团队[13]于 2004 年实现多文档人物传记摘要系统，该系统主要采用了分类算法的思想，对多个句子进行划分，将每个句子划分到其对应的类簇当中，在这个过程中，首先规定了与人物传记相关的句子类别，例如所在地区，工作经历，教育背景，社会关系等，接着通过分类算法得到最能表征人物特性的句子，最后多个句子进行结合生成人物传记。Filatova 团队[14]于 2005 年提出一个名为“元事件”的概念，“元事件”是由人物、时间、地点三类命名实体所组成的行为，

专门用于人物信息的抽取。Han 等人的团队[15]在 2007 年提出了对人物进行事件本体的构建，通过本体描述语言进行实现，事件的基本要素涵盖了事件的地点，时间，事件的参与者以及事件的具体内容等，进而完成人物所涉及的事件的抽取。

除了上述的工作，人物的信息抽取的工作还包括人物搜索引擎，这方面的工作开始得较晚，直到 2008 年人物搜索引擎才逐渐形成了雏形。目前市面上较为成熟的人物搜索引擎有雅虎旗下的人物搜索引擎，微软的“人立方”等，这两个系统的核心是人物间关系的抽取，并且聚焦于体育界、文娱界，政治圈和等多个圈层内的知名人士。雅虎人物搜索引擎主要依靠的是信息挖掘，从检索到的各个网页中进行人物简介，人际关系，机构关系和作品关系等多方面的个人信息的抽取。“人立方”则是基于海量的网页，从中识别出各类命名实体，并且进行实体间关系的抽取。在“人立方”中使用人名的关键词进行搜索时，可以得到搜索人物的个人主页，社会关系和新闻资讯等多个方面的信息，但由于其来源于海量的网页，这些网页中的信息数据通常都是杂乱的，且伴随着大量虚假信息，而系统也未建立良好的过滤机制，例如对重名没有进行筛选和区分，这样也就导致了抽取出的人物关系可信度较低，例如当我们在“人立方”中搜索与 A 具有母女关系的 B 时，出现了 A 与 B 是朋友关系的结果，而会出现这一结果的主要原因在于在某个社区论坛中出现了“A 和 B 看起来年级相仿，关系真的很亲密”这样的错误信息。除此之外，Tang 等人[17]针对学术领域的专家开发了 ArnetMiner 系统[18]，从这些专家的个人主页，社交网络和发表的论文等数据中挖掘人物的相关信息。Gundecha 等[19]针对 Facebook，LinkedIn 等社交网络提出了个人信息的挖掘工具。

国内对于相关内容也进行了一定研究，2012 年，乔磊等人[20]针对人物的出生日期、籍贯、政治面貌等属性总结规则，提出了基于规则的半结构化人物信息抽取算法并应用这一研究结果实现了人物信息抽取系统。2013 年，李红亮[21]基于触发词的人物信息抽取方法，提取出了百度百科网页中的人物属性信息。他先后通过分析文本制定触发词表，利用统计学相关原理，依据人名周围的词场自动发现候选规则，最后为进一步优化规则，对候选规则集进行支持度计算。周婷[22]研究了异构信息源的领域人物信息抽取，首先同样对爬取下来的网页利用 SVM 选出包含人物信息的网页，在制定人物属性抽取的规则库后，利用规则完成对高校教师领域的信息提取。此外，于锟[23]对简历中的求职者信息提出基于双层级联文本分类的提取方法。刘金红等人[24]结合语义分析、自然语言处理、隐马尔可夫模型等理论，提出基于语义上下文分析的人物信息提取算法。郝冬生[25]对个人主页和简历中的人物信息通过基于触发词和规则的方法进行了抽取。

以上针对人物信息抽取方面的研究在表层网络中取得了比较好的效果，采用的方法大多以人工总结规则或半自动获取规则为主，基于触发词或基于规则匹配，从简历、个人主页等属性丰富的文本数据中提取出名字、出生年月、性别、地址等基本信息。但是这类由于方法的规则并不能达到全部覆盖，因此不能有效识别出隐含在文本中诸多人物信息，召回率较低。此外，暗网中用户为隐藏自己的身份，不会主动留下个人的上述真实信息，因此需要另寻方法实现相关信息的抽取。在人物之间社会关系的抽取方面，上文提到的部分搜索引擎虽然实现了这一功能，但由于受到数据真实性无法保障、数据来源杂乱等因素的影响，抽取效果并不理想。且相关研究同样几乎没有应用于暗网中，针对暗网的用户信息抽取还有待研究。

1.3 本文主要研究内容

课题的主要研究方向是：以大量暗网网站为基础，针对暗网中的用户，从基本信息，社交，交易三个维度分析用户属性，以此构建暗网用户画像，并根据用户画像相似度实现暗网用户虚拟群体的发现。

本课题的主要内容主要包括：

（1）暗网数据大规模采集及数据集构建，由于暗网中的海量半结构化数据复杂且不规范，所以需要对暗网网页数据进行大规模采集，并对其中的数据进行预处理。同时由于暗网的隐蔽性导致采集到的数据属性稀疏，因此还需要通过与明网中数据关联来扩充属性，作为后续研究的数据集；

（2）暗网用户多维度基础画像构建，用户基础属性构成用户的基础画像，基础属性指对数据浅层次的解析后可提取出的属性；通过命名实体识别从用户相关文本中提取出各类实体，再通过实体关系抽取，得到用户相关的部分基础属性；提取结构化数据中有意义的字段，其中的基础属性作为用户标签，用于从基本信息、社交行为、市场交易三个维度构建出暗网用户基础画像，其它用户特征用于后续用户深度属性的分析；

（3）基于文本挖掘的暗网用户深度画像构建，用户深度画像是在基础画像中补充需要挖掘的深度属性，构成更加丰富的用户画像；利用(2)中抽取出的基础特征指数数据，对 PageRank 算法的 PR 分配加以优化，评估暗网用户的影响力；基于 AHP 思想设计计算模型分析活跃度；利用历史交易数据分析交易规模并通过 FTRL+XGBoost 算法预测商户未来销量；进一步挖掘数据集中非结构化的文本数据，分析用户的情感、政治倾向等属性。构建出暗网用户深度画像。

（4）暗网用户画像应用，对(3)中构建的用户画像的相似度进行计算，依据画像相似度实现暗网用户虚拟群体的发现。

将研究问题形式化描述如下：本课题输入为七元组（S，T，e_r，r_e，s_a，t_a，p_c），其中 $S=\{s_1,s_2,...s_a\}$为暗网社交论坛数据集合；$T=\{t_1,t_2,...t_b\}$为暗网交易市场数据集合。输出为人物画像集合 $Profiles=\{p_1,p_2,...,p_k\}$，$p_i=\{p\_basic_i$，$p\_social_i$，$p_transaction_i\}$为暗网用户 i 的用画像，其中 p_basic_i，p_social_i，$p\_transaction_i$ 分别表示用户的基本信息画像、社交画像以及交易画像。

（1）实体识别函数 e_r：将 $S=\{s_1,s_2,...s_a\}$，$T=\{t_1,t_2,...t_b\}$中的实体识别出来，得到实体集合 $Entities=\{e_1,e_2,...e_c\}$；

（2）实体关系抽取函数 r_e：抽取出 $Entities=\{e_1,e_2,...e_c\}$在 $S=\{s_1,s_2,...s_a\}$，$T=\{t_1,t_2,...t_b\}$的文本中存在的实体关系，得到实体关系集合 $Relations=\{r_1,r_2,...r_d\}$；

（3）社交属性分析函数 s_a：将 s_i，r_i 作为输入，其中 s_i∈社交数据集合 $S=\{s_1,s_2,...s_a\}$，r_i∈实体关系集合 $Relations=\{r_1,r_2,...r_d\}$，分别通过影响力计算函数 s_a_i，活跃度计算函数 s_a_a，情感特征分析函数 s_a_s，政治倾向分析函数 s_a_p 等函数，计算得出社交画像 $p\_social_i=\{influence_i, activity_i, sentiment_i, polity_i$ ……$\}$，其中 $influence_i, activity_i, sentiment_i, polity_i$ 分别为用户的影响力，活跃度，情感特征，政治倾向属性。

（4）交易属性分析函数 t_a：将 t_i，r_i 作为输入，其中 t_i∈交易数据集合 $T=\{t_1,t_2,...t_b\}$，r_i∈实体关系集合 $relations=\{r_1,r_2,...r_d\}$，计算得出社交画像 $p\_transaction_i=\{identity_i, orders_i, type_i, scale_i$ ……$\}$，其中 $identity_i, orders_i, type_i, scale_i$ 分别为用户的交易身份，订单数量，产品理性，交易规模属性。

（5）用户画像聚类函数 p_c：将 $Profiles=\{p_1,p_2,...,p_k\}$作为输入，计算画像的相似度，聚类得到虚拟用户群体集合 $groups=\{g_1,g_2,...g_e\}$，其中 $gi=\{p_1,p_2,...,p_f\}$。

基于对课题内容的分析，本课题具体的研究工作流程下图 1-1 所示。

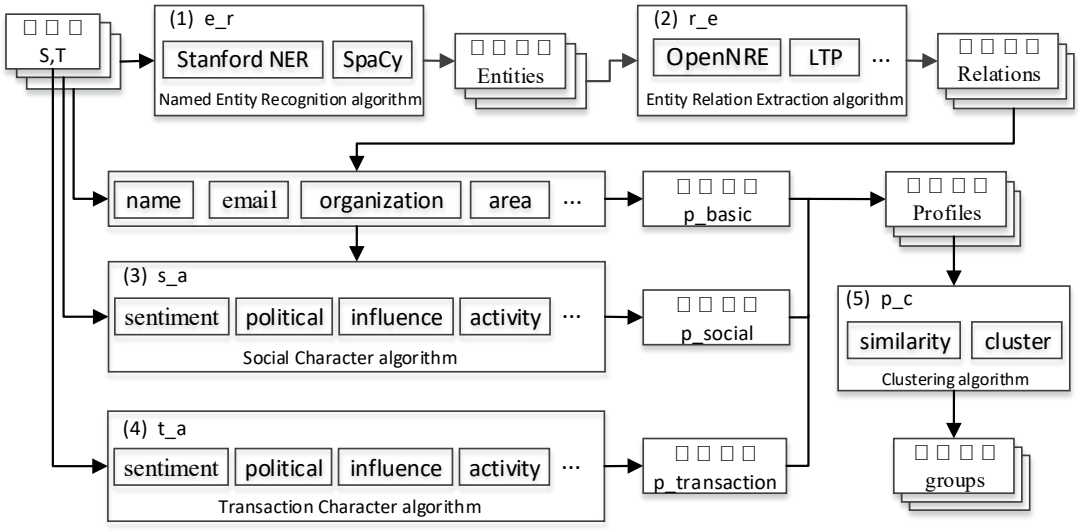

图 1-1 研究工作流图

基于主要研究内容制定的研究路线主要分为四个阶段，如下图 1-2 所示。

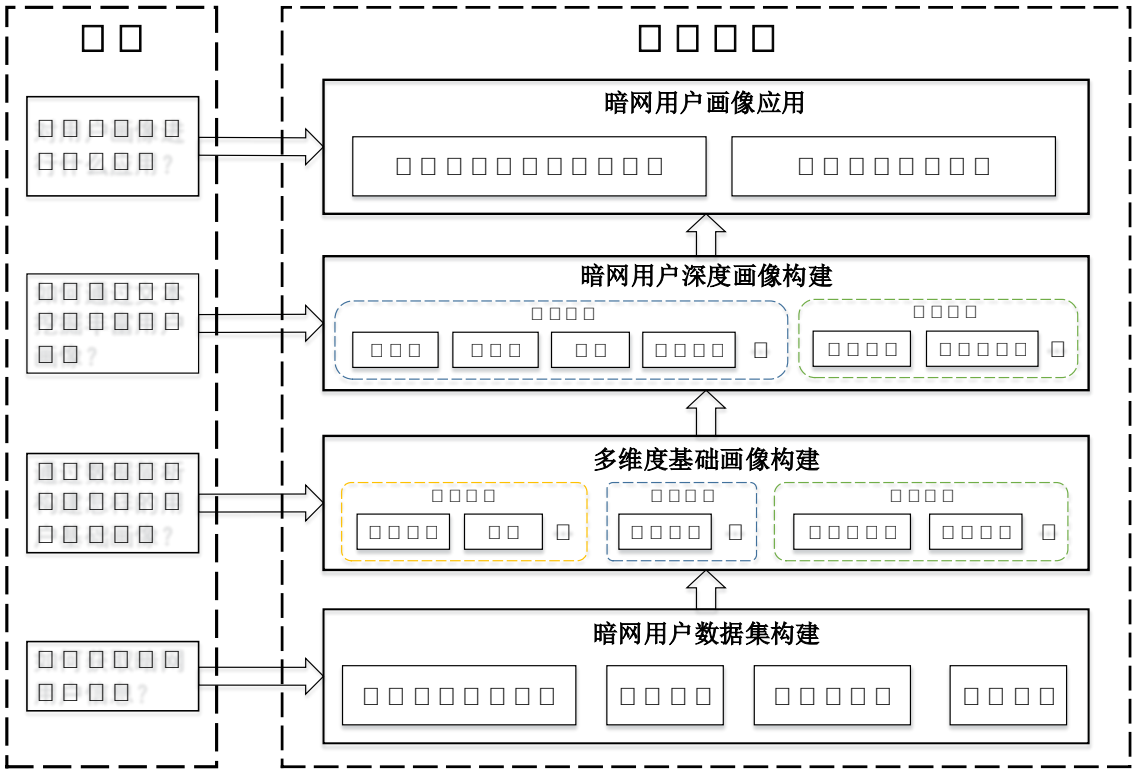

图 1-2 研究路线图

1.4 本文组织结构安排

论文具体组织结构如下：

第一章为绪论。本章介绍了课题的来源以及暗网用户画像的研究背景与重要意义，并对暗网用户画像相关的国内外研究现状综述分析。通过对相关的研究进行分析，总结了当前用户画像研究在暗网中的局限性，最后对论文内容进行了介绍。

第二章为暗网用户数据采集与数据集构建。介绍了课题中暗网数据的来源，设计了暗网爬虫系统实现对暗网网站进行大规模采集，对数据进行数据清洗、匿名化等处理后构建出本课题的数据集，最后对数据内容进行基本的统计分析。

第三章为暗网用户多维度基础画像构建。本章首先从结构化数据中提取出部分用户特征。其后对非结构化数据通过建立出暗网中的高频黑话词库，结合根据暗网数据特征添加的自定义规则对文本进行实体、识别与实体关系抽取，进一步提取用户特征，从基本信息、社交、交易三个维度构建出用户的基础画像。

第四章为基于文本挖掘的暗网深度画像构建。本章根据用户的社交言论，

通过情感分析、立场分析、影响力计算、活跃度计算等算法分析用户的深度属性，根据用户的交易记录分析用户的交易规模并通过 FTRL+XGBoost 算法对未来的交易量进行预测，将用户基础画像丰富为深度画像。

第五章为暗网用户虚拟群体发现。本章首先计算了暗网用户深度画像的相似度，其后设计了基于 Louvain 的优化聚类算法，实现暗网用户虚拟群体的发现。

第 2 章 暗网用户数据采集与数据集构建

在互联网这座网络冰山中，用户在日常中浏览使用的表层网络仅是海面之上的冰山一角，更加复杂、深不可测的暗网则隐藏于大众视野下。如何获取暗网中的信息，并将其处理为可用于课题研究的数据集是本文的第一项工作。本部分将介绍暗网数据来源、网页数据的采集及数据预处理的方法并展示实际结果。

2.1 暗网数据来源

从数据数量上，暗网中包含了表层网络所蕴含信息量的 400 倍以上的可访问的数据，仅排在前 50 位的大型暗网站点中具有的资源数量就达到了表层层网数据量的 50 倍。从数据质量上，暗网站点的月访问量也远高于表层网络站点，月数据访问量的平均值为表层网络的 1.5 倍。

虽然就整体而言，暗网的隐蔽性使得暗网的数据在数量和质量上都比表层网络更高，但也正因为暗网的隐蔽性，使得暗网中用户个人相关的信息十分稀疏。本文选取暗网中用户较为活跃的社交网站与交易市场作为数据来源[27]，从由于暗网不同于表层网络，各网站之间不具有较为明显的关联结构，因此目前大多数暗网研究中都采用人工收集相关网址的方法选择网站源[28]。本课题选取的网站主要包括 Zerotalk、Zerome、Freenet_fms_bbs、Dreamforum、HiddenAnswer 等社交网站，与 AgrthaMaket、ApollonMarket、EmpireMarket 等交易市场，数据来源分布如下图 2-1，图 2-2 所示。

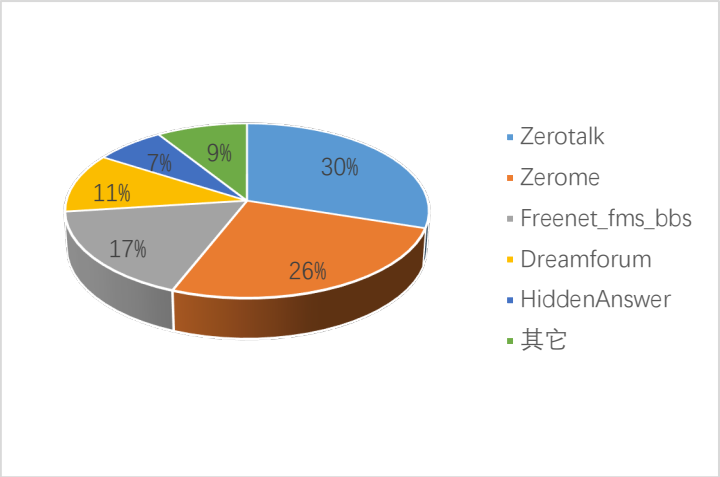

图 2-1 社交数据来源分布图

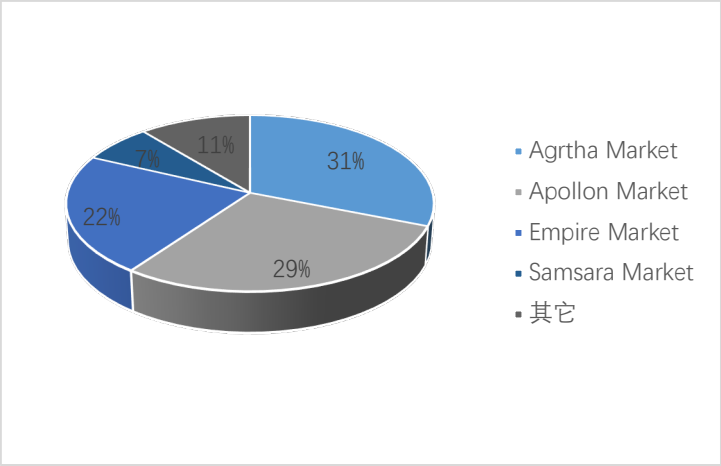

图 2-2 交易市场数据来源分布图

2.2 暗网数据大规模采集

暗网具有极强的隐蔽性，通过常规的网络方法并不能对暗网进行搜索和浏览。本课题构建了如图 2-3 的暗网采集系统，系统的组成部分主要为:Tor 网络层、任务采集层以及数据层。通过在系统中部署 Tor，架设 VPN 代理进入 Tor 网络，利用 Python 的 Requests 模块构造请求，通过 Selenium 框架模拟用户对浏览器的操作编写爬虫获取到网页数据，并使用 BeautifulSoup 模块与 select css 选择器对主要内容进行解析，获取的用户基本信息、社交言论、上线时间、比特币地址、交易描述等数据。

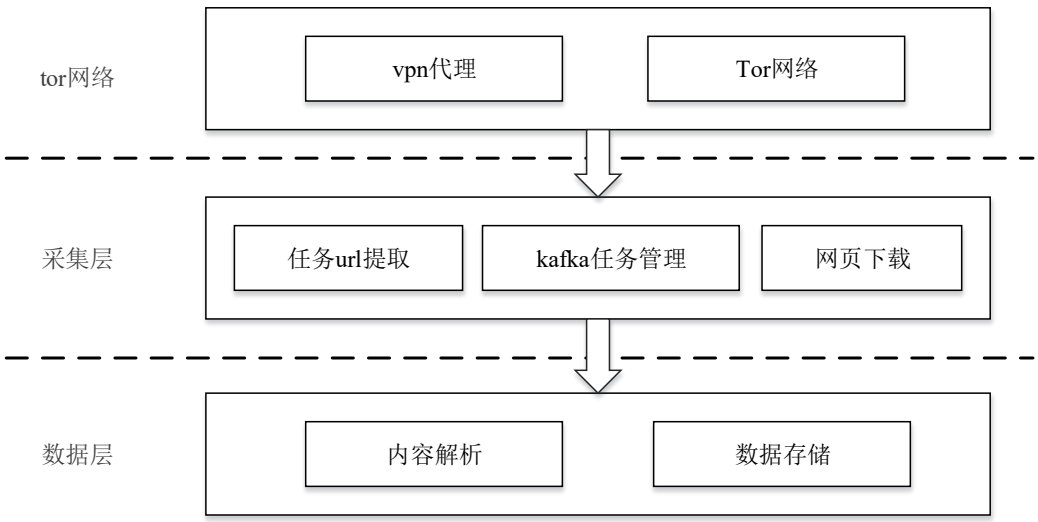

图 2-3 暗网采集系统

2.2.1 Tor 网络层

Tor 网络层主要由 Tor(The Second Generation Onion Rooter)实现。Tor 实现了客户端的匿名通信服务，当用户进行即时通信等活动时，Tor 可以隐藏用户的物理地址[29-33]。由于国内网络的部分限制，需要架设 VPN 代理进入 Tor 网络。通过前置代理，将流量先代理到国外的服务器上。根据要求在系统中部署 Tor，Tor 使用的协议是 Socks5，由于需要 Python 的 Requests 和 Selenium 两个模块进行 GET 和 POST 请求，而 Requests 模块对 Socks5 协议代理支持不够良好，所以需要使用 Cow 把 Tor 的 Socks5 协议转为 HTTP 协议，下图 2-4 即为 Tor 网络代理路线图。

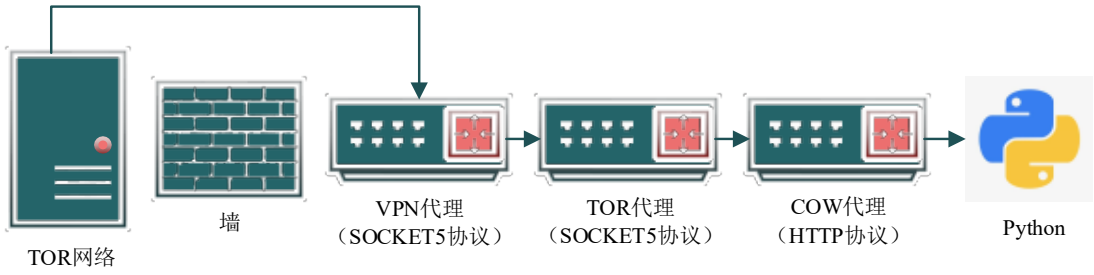

图 2-4 Tor 网络代理路线

2.2.2 任务采集层

任务采集层包括任务管理模块和数据采集模块。任务管理模块是主程序的命令控制端，也是爬虫模块与 Tor 网络模块的中间点；数据采集模块接收管理模块的采集任务并根据配置完成数据爬取。下图 2-5 为任务采集层的结构设计。

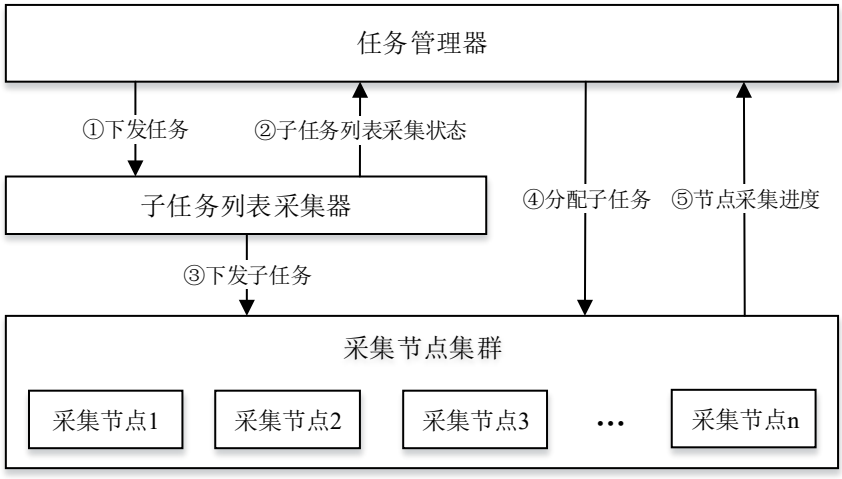

图 2-5 Tor 网络代理路线

在多个站点部署爬虫，任务控制模块接收采集任务，配置采集参数，解析任务，并将任务通过消息队列①分发到各采集节点，此处的采集节点为子任务列表采集器，用来爬取包含待采集内容的 url、html 等唯一标识。采集到子任务传入临时存储分节点，同时，通过消息队列②将子任务列表的采集状态返回给任务管理器，采集任务的完成状态包括采集成功，或由网络延时、账号封堵、程序异常等原因造成的爬取异常[34]。任务控制模块将未完成的任务再分配到其它采集节点进行二次采集，并设置采集次数阈值 3，超过阈值将任务在数据库中标记为不可完成。此外，对每个节点设置等待时限，超时严重时抛弃该节点，一定时间内不再为它分配任务。完成任务列表的采集后，将采集到的任务标识作为消息放入消息队列③，任务管理器根据子任务数量与采集节点的存活数量，将子任务通过消息队列④分发给相应采集子节点，具体采集过程与获取任务列表类似，采集节点每完成一个子任务的采集，通过消息队列⑤将任务的完成状态返回给任务管理器。

2.2.3 数据存储层

采集层爬取到的数据被存储在 ES(ElasticSearch)中，将社交数据，交易数据和用户信息分别存入三个文档中，文档中的数据存储结构设计分别下表 2-1、表 2-2、表 2-3 所示。

表 2-1 暗网社交数据存储结构设计

字段名	字段类型	描述
spider_name	keyword	爬虫名
domain	keyword	域名
url	text	url
page_index_id	keyword	全网爬虫记录
net_type	keyword	网络类型
topic_type	keyword	消息类型：post，comment
crawl_tags	keyword	该消息网站自己的标签
title	text	主贴标题
topic_id	keyword	话题原始的 id
user_id	keyword	发表这条信息的用户 id
user_name	keyword	发表这条信息的作者用户名
comment_user_id	keyword	评论者 id，消息类型为 comment 时有值
comment_id	keyword	评论的 id，消息类型为 comment 时有值
raw_content	text	原始爬取内容

表 2-1（续表）

字段名	字段类型	描述
content	text	清洗后的文本内容
clicks_times	long	点击次数
commented_times	long	回复次数
crawl_time	date	爬虫时间
publish_time	date	该条消息发布时间，页面上的时间戳
raw_publish_time	keyword	原始该条消息发布时间
thumbs_up	long	支持人数、被赞数
thumbs_down	long	反对人数、被踩数
emails	text	邮箱
bitcoin_addresses	keyword	比特币交易地址
eth_addresses	keyword	以太币交易地址
is_recognized	boolean	实体识别状态
recognize _time	date	实体识别时间
entity	list	实体列表
entity.property	keyword	实体类型
entity.value	keyword	实体内容
is_ extracted	boolean	实体关系抽取状态
extract_time	date	实体关系抽取时间
relation	list	实体关系列表
relation.property	keyword	实体关系类型
relation.value	keyword	实体关系内容
gmt_create	date	数据创建的时间点
gmt_modified	date	数据上一次修改的时间点

表 2-2 暗网交易数据存储结构设计

字段名	字段类型	描述
crawl_time	date	爬取时间
domain	keyword	域名
net_type	keyword	网络类型
spider_name	keyword	爬虫名
goods_name	keyword	产品名
goods_id	keyword	产品 id
url	text	产品链接
goods_info	text	产品信息
goods_img_url	keyword	产品图片
crawl_category	keyword	网站自己打的产品分类

表 2-2（续表）

字段名	字段类型	描述
sold_count	long	销量
price	keyword	价格
user_id	keyword	产品发布 userid
user_name	keyword	产品发布用户名
goods_area	keyword	供货地
publish_time	date	该条消息发布时间，页面上的时间戳
raw_publish_time	keyword	原始该条消息发布时间
sku	keyword	商品库存
bitcoin_addresses	keyword	比特币交易地址
eth_addresses	keyword	以太币交易地址
is_recognized	boolean	实体识别状态
recognize _time	date	实体识别时间
entity	keyword	实体列表
entity.property	keyword	实体类型
entity.value	date	实体内容
gmt_create	date	数据创建的时间点
gmt_modified	date	数据上一次修改的时间点

表 2-3 暗网用户数据存储结构设计

字段名	字段类型	描述
spider_name	keyword	爬虫名
domain	keyword	域名
net_type	keyword	网络类型
user_name	keyword	用户名
user_description	text	用户描述
user_id	keyword	用户 id
url	text	用户主页链接
register_time	date	注册时间（标准格式时间）
raw_register_time	keyword	原始注册时间
emails	text	用户个人邮箱
bitcoin_addresses	keyword	比特币交易地址
eth_addresses	keyword	以太币交易地址
raw_last_active_time	text	最近活跃时间原始数据
last_active_time	date	最近活跃时间（标准格式时间）
ratings	keyword	网站声望
User_level	keyword	用户等级数值表示

表 2-3（续表）

字段名	字段类型	描述
member_degree	keyword	用户等级原始数据
pgp	keyword	交易市场类的标识身份
crawl_time	date	爬取时间
topic_nums	long	用户发布帖子总数
goods_orders	long	若有交易记录，交易的订单量
crawl_topic_nums	long	用户发布帖子总数 通过聚类索引
identity_tags	keyword	交易身份（卖家、买家）
is_portraited	boolean	人物画像分析状态
portrait_time	date	人物画像分析时间
profile	text	人物画像相关
profile.basic	list	用户 id
profile.basic.property	keyword	基本画像
profile.basic.value	keyword	基本属性内容
profile.social	list	社交画像
profile.social.property	keyword	社交属性类型
profile.social.value	keyword	社交属性内容
profile.transaction	list	交易画像
profile.transaction.property	keyword	交易属性类型
profile.transaction.value	keyword	交易属性内容
gmt_create	date	数据创建的时间点
gmt_modified	date	数据上一次修改的时间点

2.3 暗网数据预处理

本文采集的数据可以分为结构化数据与非结构化的文本数据，结构化数据与非结构化数据在本研究的数据中所占比例如图 2-6 所示。

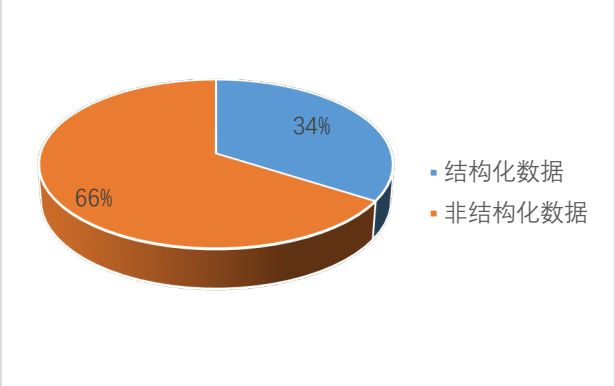

图 2-6 数据类型分布图

对于结构化数据，将不同源数据中表达方式不同的属性，进行统一规范化

处理。如时间类数据存在“年-月-日 时:分:秒”，“年-月-日”，“月-日-年”、“X 小时前”等多种格式；用户等级中存在数字等级、文字描述、积分数量等数据格式。本课题中将时间统一为“年-月-日 时:分:秒”的格式，将用户等级中未采用数字描述等级的数据追溯到原始网站，根据其积分制度或会员标识说明，将其等级量化至“0-5”，越接近 5 表示用户等级越高，0 表示初注册几乎无网络行为的用户。

同时，因为暗网特性导致采集到的数据中涉及的暗网用户属性十分稀疏，所以通过通讯方式、比特币地址、用户名等用户标识信息进行暗网中多源数据融合，并尝试与明网中的用户信息进行关联，收集的明网数据中数据的存储格式各不相同，部分数据如下图 2-7 所示。

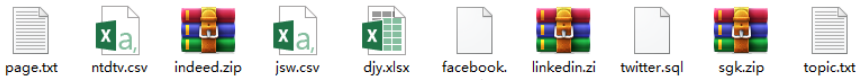

图 2-7 部分关联数据

以多源数据融合的方法实现用户属性的扩充，流程如图 2-8 所示。

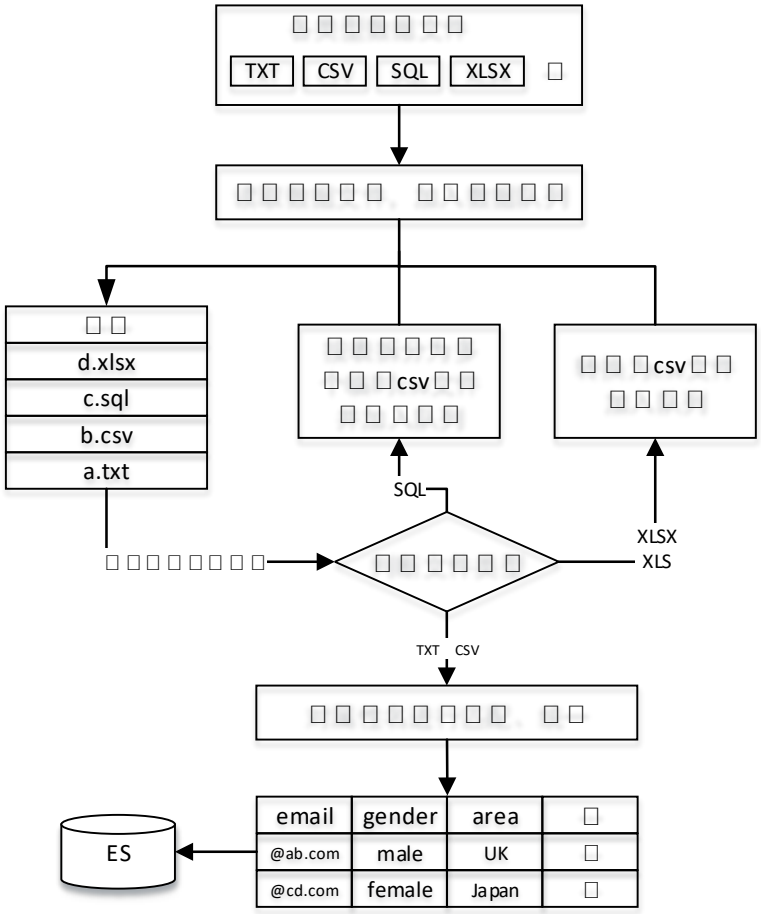

图 2-8 多源结构化数据融合流程图

与结构化数据相比，文本数据缺乏固定的结构，同时其表现形式采用人类的自然语言，计算机对其理解的程度有限，很难深入的理解并处理其蕴含的具体语义。对于非结构化的文本中存在的乱码、特殊符号等内容，通过正则表达式匹配等方式对文本数据进行清洗。清洗后的内容便可用于各特征的提取，特征指以计算机可识别、可计算、可理解的特征项，如数值、向量等，对文本的元数据进行表示，从数据中选取 3000 条优质数据，将 6 人分 2 组针对后续用户属性分析中需要的特征进行打标签，若同一数据的两组标签一致，则将标签结果作为该数据的最终标签，若不一致则加入第三方讨论出一致结果，从而达到对文本数据的结构化处理。打好标签的数据即可用于后续算法的分析，训练集与测试集按照 4:1 的比例进行划分。

2.4 数据匿名化处理

虽然本研究中所使用数据均来自于用户在暗网中公开展示的个人信息，但鉴于数据中可能存在用户个人私密信息，为保护用户隐私，本研究对实验中所使用的数据中的信息进行加密或匿名化处理。对易混淆数据保留上下文，参考完整性则可保留一致性。对于能够唯一标识一个人的属性，采用泛化的方法，比如某人的社交网络账号为一串字符串，本文将该字符串的若干位用星号替换。通过上述等方法对敏感人物进行假名处理、对敏感话题降低敏感性，在用户个人信息不被泄漏的基础上，最大程度地保留用户虚拟身份的属性特征，从而能够保证实验结果的准确性。

2.5 暗网数据统计分析

本研究共采集暗网数据 90104 条，其中用户信息、社交数据、交易数据的入库数量如图 2-9 所示。

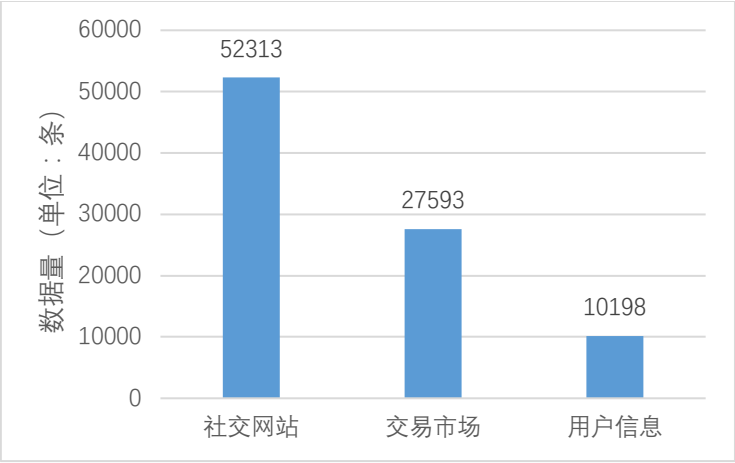

图 2-9 暗网数据量

暗网数据采集系统每月更新一次数据，采集数据量随时间的变化情况如下图 2-10 所示。

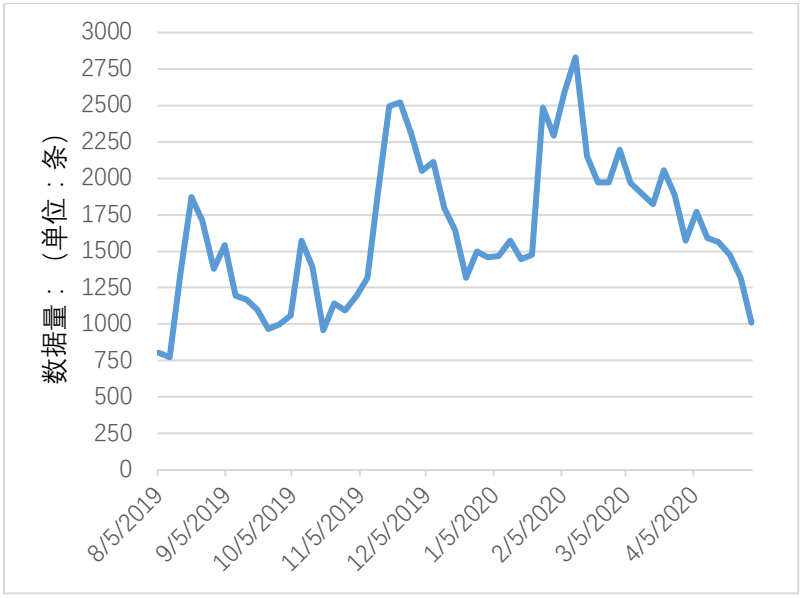

图 2-10 暗网数据量变化曲线

社交网站在某一阶段的数据量与该阶段内的舆情事件发生情况与规模具有明显的联系。当舆情事件发生后，网民们会通过社交网站对事件进行评论沟通，使舆情事件从传播扩散，到成为网络舆论热点，甚至最终会引发舆论风波。不同时间由于热点事件的影响力不同，采集数量曲线也有所差异，各热点事件相关数据数量占比随时间变化曲线如下图所示。

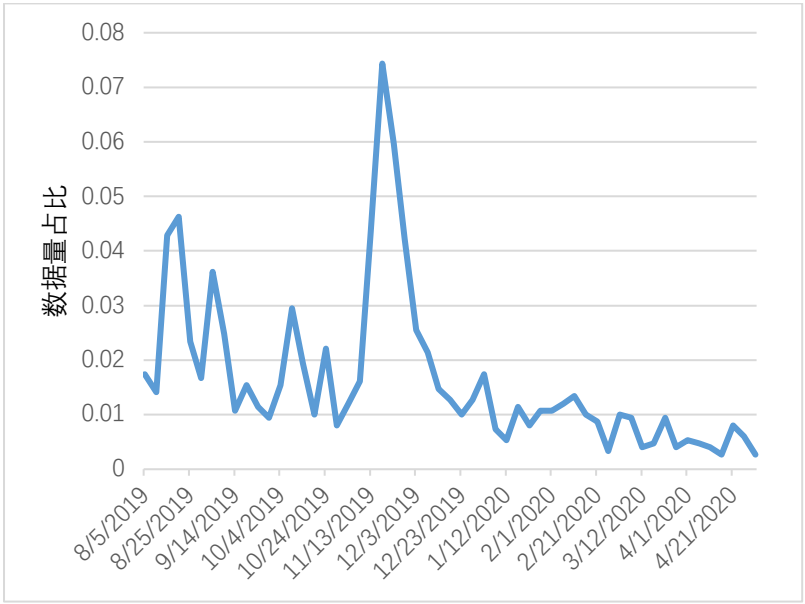

图 2-11 某暴乱事件数据数量变化曲线

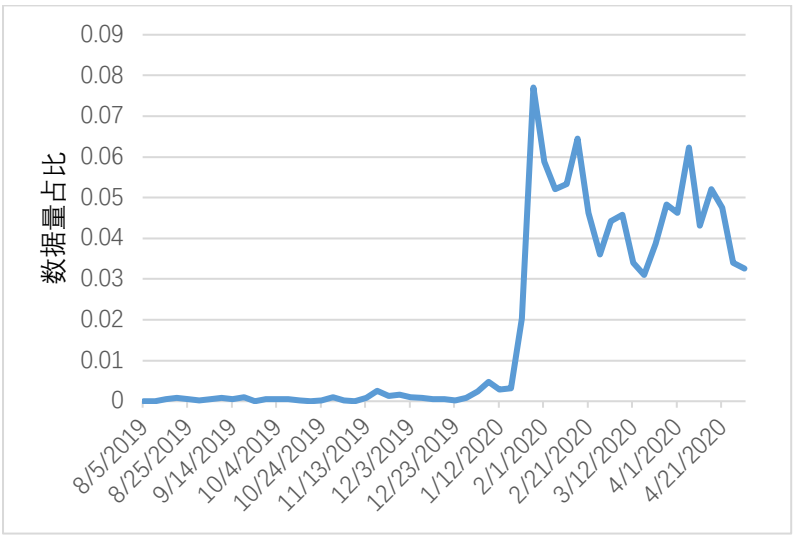

图 2-12 **肺炎数据数量变化曲线

暗网交易市场的交易量与重大事件的发生也有较大关联。在疫情快速蔓延的情况下，全世界的各个行业都受到不同程度的冲击，许多企业甚至面临着破产、倒闭等危险，损失十分惨重。但与此同时，暗网作为犯罪分子的集聚地吸引了许多非法商家，他们借由疫情的契机，贩卖口罩、防护服等十分紧缺的防疫物资，甚至提前上线了全球各国都仍在研发阶段的**疫苗。防疫物资相关的交易量变化曲线如图 2-13 所示。

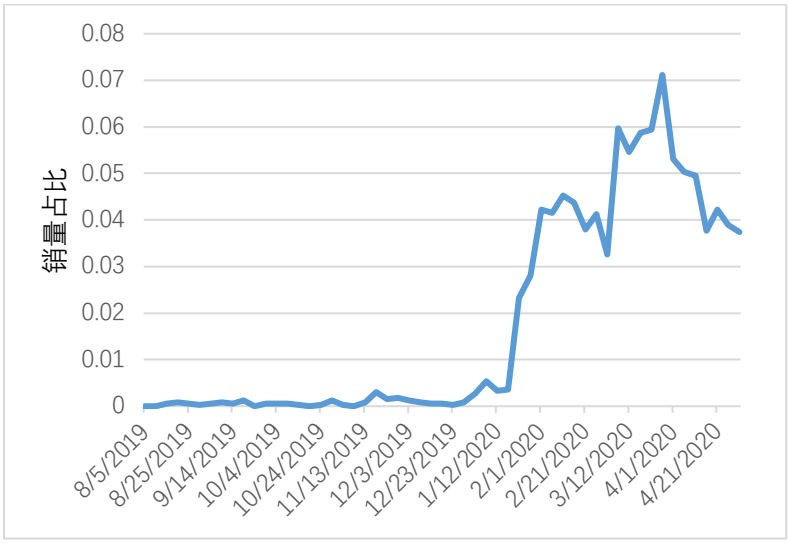

图 2-13 防疫物资销量变化曲线

此外，暗网隐私数据的交易量也与疫情期间层出不穷的数据泄露事件有明显关联，交易量随时间变化如下图 2-14 所示。例如某全球大型医疗机构的大量关键数据，在遭受黑客攻击后被放在暗网中售卖；大多办公人员在疫情期间的居家必备远程办公软件 Zoom 也遭遇了数据泄露，账号密码、邮箱、个人会议号与会议密码等 50 万以上的 Zoom 用户隐私信息，在暗网上以不足 1 美元的超

低单价甚至赠送的方式大肆出售。这使得疫情期间暗网中部分行业的交易量不降反升。

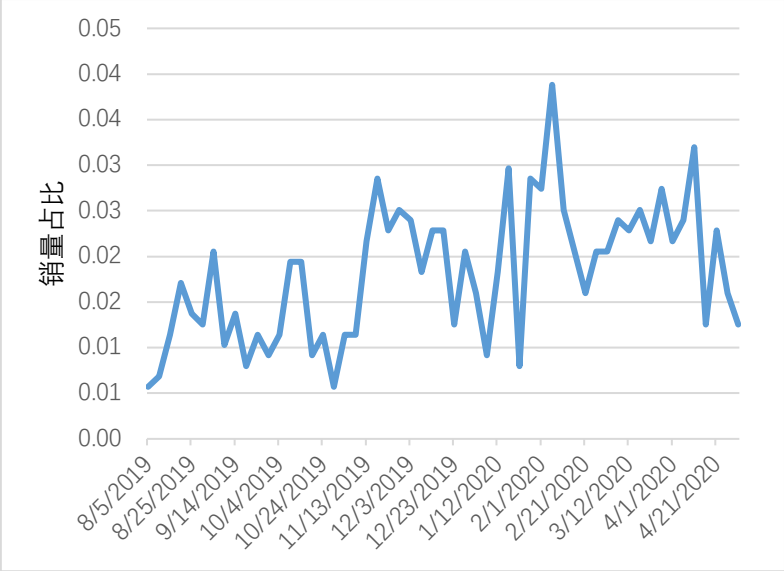

图 2-14 泄露数据销量变化曲线

2.6 本章小结

本章完成了暗网用户画像数据的采集与数据集的构建。通过合理选择数据来源，定制暗网数据采集系统实现暗网数据的大规模数据，对采集的暗网数据进行数据清洗、匿名化等处理，最终对数据进行统计分析并构建数据集。暗网用户数据作为后续用户属性分析与用户画像构建的基础，具有非常重要的作用。

第 3 章 暗网用户多维度基础画像构建

在采集到的不同来源的暗网数据中，由于暗网的隐蔽性导致数据中所包含的用户属性都十分稀疏，为使构建出的用户画像更加丰富，不仅需要提取数据中显性的用户属性与基础特征，更要对数据内容充分挖掘，获取到更多深度属性。其中显性的基础属性的获取除了将可直接提取的用户基本信息抽取出来外，还可以简单地统计、分析已有数据，提取出可以用于进一步分析用户特征的相关属性进行识别与计算。

在本部分将介绍暗网用户基础画像的构建，即用户基础属性的抽取。通过统计计算、命名实体识别、实体关系抽取等方法，得到暗网用户的基础属性和用于深度属性分析的基础特征。

3.1 基础属性与特征提取

基础属性指对数据浅层次的解析后可提取出的属性。在浅层网络中，通过简单的解析可以获取到用户的大量基础属性，如出生年月、居住地、教育背景、工作经历，甚至用户的宗教观点，政治倾向等可以方便的获取到。而在暗网中，为了更好的隐藏自己的身份，躲避网络的监管，用户尽可能避免留下与自己的真实身份相关的信息，因此本节获取到的基础属性与特征较为稀疏，居住地等基础属性须通过后续实体识别等算法加以分析。

本节根据源网站内容与采集的数据结果，通过对从结构化数据与半结构化数据进行清洗与统计，解析暗网用户的基础属性。结构化数据的信息相对完整、清晰，通过统一规范化处理等操作可以较容易地得到属性；半结构化数据包括用户在社交网站中按规定格式填写的自我介绍一类内容，与结构化数据相比而言，这类数据的内容较长，但相比于非结构化的文本数据内容格式更加规范，可以通过正则匹配等方式提取出用户属性。

从三个维度对暗网用户解析数据，可以中获取用户的提取出与课题相关的用户属性与特征共 15 个，其中包括 10 个用户基础属性，5 个用户基本特征。表 3-1 统计了名称与所属类别。其中可以从结构化数据中提取到的基本属性有：用户名、用户 ID、通讯方式（email）；社交属性有：用户等级、注册时间、最近活跃时间、发帖数；交易属性有：交易身份（买家或卖家）、订单数量、交易品类型、比特币交易地址、以太币交易地址、PGP 交易标识。此外，统计对采集到的数据可以计算出用户的回帖数与用户发帖得到的评论数，结合数据爬取

时间、账户注册时间以及最近活跃时间等数据，可以用于后续对用户影响力、活跃度等属性的计算。

表 3-1 用户基本属性特征表

序号	维度	名称	属性或特征
1	基本属性	用户名	属性
2		用户 id	属性
3		通讯方式	属性
4		用户等级	属性
5	社交	注册时间	特征
6		最近活跃时间	特征
7		发帖数	特征
8		评论数	特征
9		回帖数	特征
10	交易	交易身份	属性
11		订单数量	属性
12		交易品类型	属性
13		比特币交易地址	属性
14		以太币交易地址	属性
15		PGP 交易标识	属性

3.2 相关实体识别

对于暗网中难以获取到的用户位置等属性，可以采用先通过命名实体识别算法识别出文本中的地点、组织机构、时间等实体信息，再通过实体关系抽取判断实体抽取结果中实体间的关系，从而抽取出与用户相关的地址与组织机构。由于数据的语言类型主要为英文和中文，因此采用对支持中英文的斯坦福命名实体识别和对英文文本识别率较高的 SpaCy 进行实体识别，同时为了能够成功识别到暗网中用黑话表述的实体，通过建立具有针对性的暗网关键词库以及自定义规则提高识别的准确率。实体识别的输入为社交数据集 S 或交易数据集 T 中的文本数据，输出为识别结果集合 Entities={e_1,e_2,...e_c}，实体 e_i 为三元组{name，type，position}，其中 name 为实体名称，tpye 表示实体类别，position 记录实体单词在句子中的位置，用于后续的实体关系抽取。

3.2.1 中文命名实体识别

本课题中以斯坦福命名实体识别算法为基础，实现对中文文本的实体识别。斯坦福命名实体识别（下文中简称 Stanford NER）由斯坦福大学提出，它可以提供支持不同语言的模型，可以实现对本课题中的中英文数据的命名实体识别，Stanford NRE 支持识别的实体类型主要包括：名字、数字、时间三大类，识别结果中又细分为人名、组织机构，钱、百分比，日期、时间等类型。斯坦福命名实体在 ACE 和 MUC 评测会议的评测语料等不同语料上使用三个 CRF 组合标注序列训练。DATE 等数字实体在进行规范化后，通过基于规则的系统进行识别。本课题的研究中采用了 Python 封装的 Stanfordcorenlp[35]，首先对表层网络中语法标准的新闻数据进行时实体识别，识别结果如下图 3-1 所示。

```
美国与塔利班正在就塔利班被关押的5000名囚犯和阿富汗政府军被关押的1000名囚犯处理
事宜进行对话，但双方却在此问题上含糊不清，一方要求美军全面撤离阿富汗领土，而华
盛顿则仅答应部分撤军，以确保阿富汗不会被用作攻击美军的平台。根据其中一项协议的
草案内容，在签署协议后的135天之内，美国将把驻阿美军人数从现有的约1.3万人减少到
8600人，作为对应条件，塔利班则保证将对恐怖组织基地组织和伊斯兰国进行打击。
[('美国', 'COUNTRY'), ('塔利班', 'ORGANIZATION'), ('塔利班',
'ORGANIZATION'), ('5000', 'NUMBER'), ('阿富汗', 'COUNTRY'), ('1000',
'NUMBER'), ('一', 'NUMBER'), ('美军', 'GPE'), ('阿富汗', 'COUNTRY'),
('华盛顿', 'STATE_OR_PROVINCE'), ('部分', 'NUMBER'), ('阿富汗',
'COUNTRY'), ('美军', 'ORGANIZATION'), ('一', 'NUMBER'), ('签署',
'MISC'), ('协议', 'MISC'), ('后的', 'MISC'), ('135', 'NUMBER'), ('天',
'MISC'), ('美国', 'COUNTRY'), ('美军', 'DEMONYM'), ('1.3万',
'NUMBER'), ('8600', 'NUMBER'), ('塔利班', 'ORGANIZATION'), ('基地',
'ORGANIZATION'), ('组织', 'ORGANIZATION'), ('伊斯兰国', 'GPE')]
```

图 3-1 Standford NRE 识别结果

然而目前，命名实体识别只在有限领域的部分实体类型中得到了比较理想的效果[36]，如识别图 3-1 所示的明网中语法规范的英文新闻语料中所包括的人名、地名、组织机构名。但这些技术无法很好地迁移到暗网中，尤其是暗网中的特定领域中，如军事、医疗、毒品等。因为暗网数据与表层网络中的新闻等数据不同，其句子结构、标点符号、单词拼写等语法语义都非常杂乱无章，同时因为暗网内容的独特性，数据中还存在大量常规预料库中不包含的暗网黑话，导致斯坦福命名实体识别在暗网数据中的识别效果极差，下图 3-2 展示了对暗网数据的识别效果。

```
集体农庄庄员伊万在河里捉到一条大鱼，高兴的回到家里和老婆说："看，我们有炸鱼吃了！""
 没有油啊。""那就煮！""没锅。""烤鱼！""没柴。"伊万气死了，走到河边把鱼扔了回去。
 那鱼在水里划了一个半圆，上身出水，举起右鳍激动地高呼："斯大林万岁！"
[('伊万', 'MISC'), ('一', 'NUMBER'), ('伊万', 'MISC'), ('一', 'NUMBER'),
 ('斯大林', 'MISC'), ('万', 'NUMBER'), ('岁', 'MISC')]
国内的杀毒软件竟然成为传播病毒的主要工具，使用Windows系统的用户，别用任何国产的杀
 毒软件，如果实在没的用，就安装原版Windows10用windows10自带的
[('10', 'NUMBER'), ('windows', 'MISC'), ('10', 'NUMBER')]
Github计划在中国设立子公司，这真的是作死。github主站离着被屏蔽不远了。什么996ICU
 ，什么evil huawei，什么翻墙软件，都得给屏蔽掉。
[('Github', 'PERSON'), ('中国', 'COUNTRY'), ('996', 'NUMBER')]
```

图 3-3 Standford NRE 暗网数据识别效果示例

统计本课题中暗网数据集应用 Stanford NER 的识别结果，针对主要类型实体分别计算准确率、召回率与 F-测度 3 个指标，得到 Stanford NER 在暗网数据中的识别效果如下图 3-4 所示。

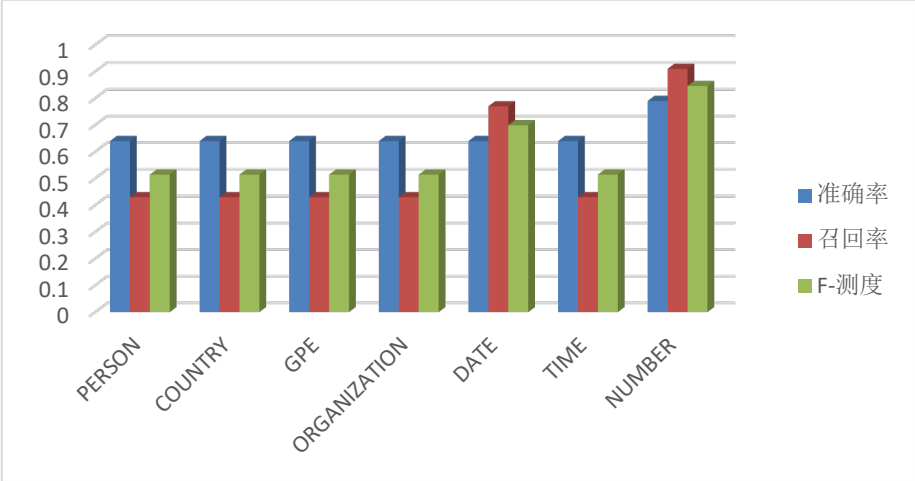

图 3-4 Standford NRE 暗网数据识别结果

3.2.2 英文实体识别

建对于课题中的英文文本，以目前世界上最快的工业级自然语言处理工具 SpaCy NER 为基础，实现命名实体识别。SpaCy NER 在 OntoNotes 5 公开语料库上进行了命名实体识别的训练，能够支持日期，时间，金额，人名，地名，组织机构名，事件，产品等类型实体的识别。由于 SpaCy 对中文数据的支持率较差，但对常规的英文数据的识别率较高，且处理速度非常快，具有很好的性能。因此本课题中仅对英文数据使用 SpaCy 进行命名实体识别。SpaCy 在规范数据中识别结果如下图 3-5 所示。

```
European authorities fined Microsoft a record $5.1 billion on Wednesday for
 abusing its power in the mobile phone market and ordered the company to alter
 its practices
[('European', 'NORP'), ('Microsoft', 'ORG'), ('$5.1 billion', 'MONEY'),
 ('Wednesday', 'DATE')]
Allen said that When Sebastian started working on self-driving cars at Google
 in 2007, few people outside of the company took him seriously. I can tell you
 very senior CEOs of major American car companies would shake my hand and turn
 away because I wasn't worth talking to, in an interview with Recode earlier
 this week.
[('Allen', 'PERSON'), ('Sebastian', 'PERSON'), ('Google', 'ORG'), ('2007',
 'DATE'), ('American', 'NORP'), ('Recode', 'PERSON'), ('earlier this week',
 'DATE')]
```

图 3-5 SpaCy 识别效果示例

但在图 3-6 所示的识别结果中可以看到，面对暗网数据的不规范，SpaCy 对暗网数据的识别率与 Stanford NER 同样低。

```
marquis is for heroin and cocaine. different test would be needed for mdma
[('marquis', 'PERSON')]
 Check out my review on the Avengers Disasocciative Forum. I sell on Empire mostly.
 dannyboys If you buy off me on market first with first order, we can talk. Need to
 build some trust. Empire: dannyboys I have very reputable feedback on Avengers -
 see disassociative vendor thread or check me out on Empire: dannyboys  100%
 feedback.
[('the Avengers Disasocciative Forum', 'ORG'), ('Empire', 'GPE'), ('first',
 'ORDINAL'), ('100%', 'PERCENT')]
```

图 3-6 SpaCy 暗网数据识别效果示例

统计本课题中暗网数据集应用 SpaCy 的识别结果，针对主要类型实体分别计算准确率、召回率与 F-测度 3 个指标，得到 SpaCy 在暗网数据中的识别效果如下图 3-7 所示。

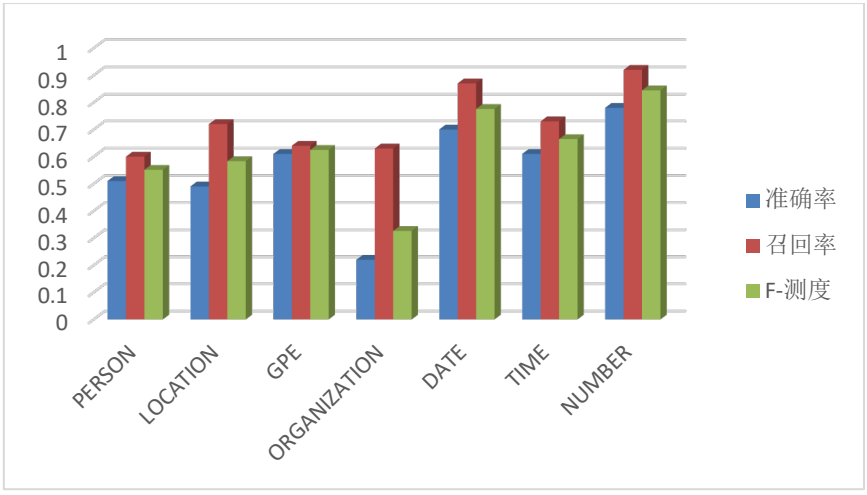

图 3-7 SpaCy 暗网数据识别结果示例

3.2.3 建立暗网语料库

暗网命名实体识别的几个难点包括：对特殊领域中的命名实体识别局限性、命名实体的开放性与复杂性以及命名实体在文本中表述歧义性与多样性[37]。造成这些难点的一个重要原因之一就是大部分知识库中实体标记不够完整，缺少实体的类型信息，因此对于每个没被标记的实体，要充分利用它的文本描述与属性等信息来预测缺失的类型[38]。

选取 1000 条暗网数据采用 BIOES 法进行标注，B 标志当前词位于一个实体的开始位置，I 表示当前词在实体的内部，O 代表处于实体外部，E 代表当前词位于这个实体的结束位置，S 代表这个词本身就可以构成一个实体，标记示例如图 3-8 所示。使用暗网语料库重新训练实体识别算法模型，使更多暗网中特殊实体能够被识别出来，对提高命名实体识别算法的识别率十分有效。

中(B-ORGANIZATION)共(E-ORGORGANIZATION)代(O)表(O)中(B-COUNTRY)国(E-COUNTRY)
最(O)广(O)大(O)人(O)民(O)的(O)根(O)本(O)利(O)益(O)

图 3-8 自定义暗网实体标记示例

3.2.4 自定义规则

首先，为了提高识别的正确率，对英文数据采用 Stanford NER 与 SpaCy 双重校验的方法，以t_{spacy}表示 SpaCy 的识别结果，$t_{stanford}$表示 Stanford NER 的识别结果，校验规则如公式 3-1 所示，两种算法识别结果一致则作为最终结果，并将两种标签统一；否则结果不保留。

$$\text{type} = \begin{cases} t_{stanford}, & t_{stanford} = t_{spacy} \\ none, & otherwise \end{cases} \tag{3-1}$$

在提高较短的实体的识别率后，需要对较长的嵌套实体识别效果进一步优化。在实际应用的过程中，存在大量的嵌套实体，而现有的模型对嵌套实体的识别效果并不理想，当前大部分命名实体识别都会忽略嵌套实体，无法获取到深层次文本中粒度更细的语义关系。如下图 3-9 所示，在“中国驻俄罗斯使馆提醒确有需要乘坐国航航班回国人员，务必及时进行核酸检测，并持书面检测报告办理登机手续。”在这一句子中，“中国驻俄罗斯使馆”是一个典型的嵌套实体，其中“中国”和“俄罗斯”是两个独立的地名实体，而“中国驻俄罗斯使馆”却是一个完整的组织机构名。用现有的命名实体识别模型对这一例句进行识别，识别结果中只出现了 “中国”和“俄罗斯”两个地名实体，并没有识别出“中国驻俄罗斯使馆”这一完整实体。同样，“Dec 3 2019”也被识别为“3”、“2019”3 个独立的 NUMBER 实体。

图 3-9 嵌套实体识别示例

针对完整实体在结果中被拆分的情况，采用动态堆叠多个扁平命名实体识别层的方法，基于内部命名实体识别结果提取外部实体。在优化前，组织、时间等由多个单词构成的嵌套实体，在实体识别中被拆分成多个单独实体。在嵌套实体识别优化后，“Dec 3 2019”将作为一个完整的 DATE 实体被识别出来。合并后效果如下图 3-10 所示。可以看出，在完成嵌套实体优化的同时，对于识别出一段文本中给定的命名实体到底属于哪一个概念，找到文本中实体所对应的真实指代有突出效果，在一定程度上提高了命名实体的消歧效果。

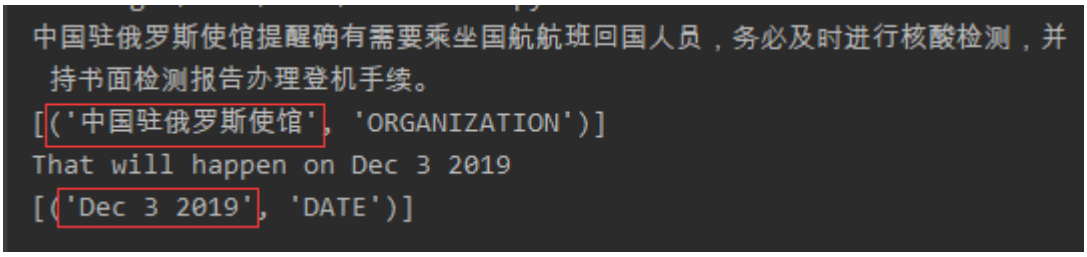

图 3-10 连续实体合并前示例

3.2.5 实验结果与分析

采用上述算法，上述实验与优化，本文对中文、英文数据所支持的识别的实体类型如表 3-2 所示。

表 3-2 实体识别类型表

类别	描述
person	people
org	organization:companies,agenciess,institutions
norp	nationalities or religious or political groups
loc	location:including mountion ranges, bodies of water
gpe	countries,cities,states
time	times smaller than a day
date	Absolute or relative dates or periods

通过建立暗网语料库，实现嵌套实体识别，识别结果双重校验等方法优化后，本文对暗网数据集中各类实体识别结果的准确率、召回率与 F-测度如图 3-

11 所示。

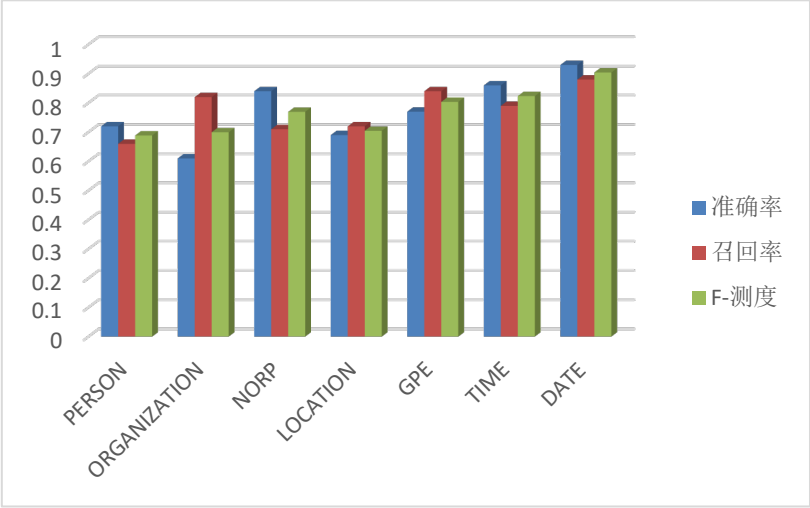

图 3-11 优化后识别结果

分别对 Stanford NER、SpaCy 以及本课题最终方法所有种类实体识别结果计算平均值，作为三种方法实体识别的效果，如下图 3-12 可看出本课题的优化方案效果明显好于前两者。

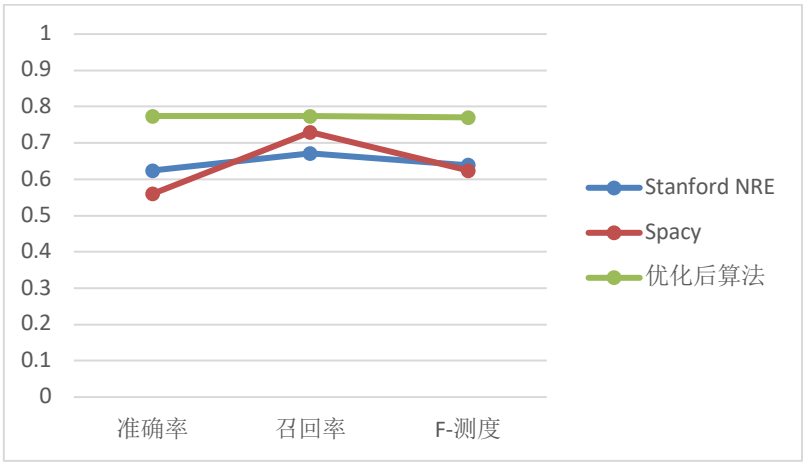

图 3-12 三种方法识别结果对比图

3.3 实体关系抽取

在实体识别阶段所识别出的实体集中只有一小部分是与用户主体相关，因此通过进一步对实体关系的判断，筛选出与用户相关的实体，并抽取出关系。实体关系抽取首先需要提前定义出实体关系集合，之后在实体识别的结果之上从文本数据中提取出实体关系。相比于从大量的文本段落数据中提取或组合出目标文本或段落的文本聚类、文本分类等技术，实体关系抽取则可以从文本句子这类更精细的文本数据中挖掘出用户所需要的语义关系，提供更细粒度的服

务。本文中实体关系抽取的输入为 $S=\{s_1,s_2,...s_a\}$，$T=\{t_1,t_2,...t_b\}$中文本与实体识别得到的实体集 $Entities=\{e_1,e_2,...e_c\}$与规则库 $Rules=\{rule_1,rule_2,...rule_d\}$，输出为实体关系集合 $Relations=\{r_1,r_2,...r_k\}$，其中关系 r_i 为三元组{head，tail，relation}，其中 head 为实体 A，tail 表示实体 B，relation 表示实体 A 与实体 B 的关系，方向由 A 指向 B。本课题采用 OpenNRE 与 LTP 结合自定义规则匹配等方法实现对实体关系的抽取。

3.3.1 中文实体关系抽取

对中文数据，由于目前支持中文实体关系抽取的模型很少且效果不佳，国内使用比较广泛的是哈尔滨工业大学的计算与信息检索研究中心研发的自然语言处理云服务系统[39]，它以语言技术平台（Language Technology Platform，简称 LTP）为基础，是一个较为全面中文自然语言处理工具，本可以基于 LTP 进行依存句法分析，基于语法分析，抽取出实体在句子中的依存关系；同时基于规则匹配的方法指根据现有的关系样例，人工制定人物关系抽取的规则库，然后将待处理的语句按照规则库中的模式进行匹配，如果匹配成功，则人物之间具有模式对应的关系。

本文通过对大量分词后的实体关系候选句进行观察及分析，结合汉语语法知识，总结了以下六条启发式规则:①关系词位于两个实体之间，中间无其他词语；②最直接的一种表示，以“实体名”是“实体名”的“关系名”为句型；③关系词处于两者之间且紧挨着第二个实体名的左边，中间无标点符号；④关系词处于两者之间且关系词必须与第一个实体名之间仅以“的”连接，中间无标点符号；⑤关系词紧接着第二个实体名右边，且两实体名之间为并列关系，即以“和”、“与”或中间无任何词连接；⑥人名指代的情况，关系词紧接着第二个人名的左边，且指代词位于关系词的左侧，指代词右侧的词语不能以标点符号间隔。最后一条规则主要是消除人名指代的问题，是六条规则中唯一允许有标点符号的情况，但标点符号只能出现在指代词之前。

根据以上六个启发式规则，结合暗网数据特点补充自定义规则库，通过建立正则表达式来构建规则库。用规则库中的每个正则表达式分别对句子文本进行匹配，在匹配时以规则出现频率从高到低的顺序进行，这将对算法准确率的提高起到一定作用，同时降低匹配到无关句子的数目。基于规则匹配的关系抽取算法如算法 3-1 所示。

算法 3-1 ：基于规则匹配的关系抽取算法

输入： 已完成实体抽取的句子文本集合 $S=\{s_1,s_2,...s_a\}$, $T=\{t_1,t_2,...t_b\}$;规则库 $Rules=\{rule_1,rule_2,...\ rule_d\}$ (按频率降序存储)

输出： 实体关系集合 $Relations=\{r_1,r_2,...r_k\}$，$r_i$ 为关系三元组{head，tail，relation}

```
BEGIN
获取句子文本集合 S={s1,s2,...sa}, T={t1,t2,...tb}
 WHILE（文本集合还有待抽取的句子）DO
   读取待抽取的句子 s，初始化 d=1
   WHILE（规则库里还存在未匹配的规则 rule d）do
     从规则库中取出 rule d 并匹配句子 s
     IF（匹配成功则）
       提取实体关系三元组 ri,跳出本循环，抽取下一句子
     ELSE
       d+1，进行下一次循环
   END WHILE
 END WHILE
END 基于规则匹配的关系抽取算法
```

通过上述的实体关系抽取算法，可以确定实体之间的关系描述词。通过对大规模语料抽取后，可以获得实体关系集合 $Relations=\{r_1,r_2,...r_k\}$，但其中的关系词可能存在同义词，为了确定实体间的具体关系，合并关系三元组中的关系词，同时对权重也进行相应叠加，最后选择具有最大权重的关系三元组表示两个实体的关系。

3.3.2 英文实体关系抽取

本课题中以 OpenNRE 为基础，实现对英文文本数据的实体关系抽取。OpenNRE 是一个基于 TensorFlow 的神经关系抽取的框架，支持“Cause-Effect”、“Country of Origin”等 80 中实体关系的抽取，将关系抽取的工作流程分为四个部分：向量化，编码器，选择器和分类器。

本课题中选取 CNN 和 bert 模型使用，并根据课题需要与抽取效果选取部分实体关系作为抽取目标。对于输出结果，tagging-based 方式有一些不足之处：其一，抽取出的不同三元组可能表示同一种关系，比如 A is followed by B 和 B follows A，需要将这两种关系合并。其二，部分实体关系在文本中可能不存在比较明显的三元组来代指其含义。因此，本课题中采用基于分类、聚类的方式

解决以上问题。通过分类或聚类，将表达了相同关系的句子放在一个簇中，然后给这个簇确定一个唯一的关系三元组。

由于本课题中存在算法输入为同一用户的多个文本数据的情况，此时会抽取出的用户关系候选集合。以 live in 关系为例，输入不同文本可能会导致出现包括 2 个及以上地区名称均与该用户为 live in 关系。如[{Tom，America，live in},{Tom，NewYork，live in},{Tom，California，live in}]。此时需要定义规则从候选地名集合中得到最终的用户居之地。本文采用频率与层级关系结合的方法确定最终目的地。通过统计频率选取频率最高的地名以缩小候选范围；此外，对地名进行规范，并根据地名库生成了关于洲、国、省、市、区的树，能够较好的处理地名之间的包含关系，在抽取居住地时，若 A 地被 B 地包含，并且 A 是发生地，抽取时仅将 A 抽取出来，而不把 B 地也作为居住地[40]。

3.3.3 实验结果与分析

统计各类型实体关系抽取结果，剔除“said to be the same as”等准确率较低，“owned by”、“occupation”、“occupant”等与课题相关性较弱的实体关系，最终只保留“work for”、“religion”、“live in”三种实体关系类型，分别用于提取暗网用户的“组织机构”、“宗教”、“地区”三个属性，三种的保留实体关系的准确率如图 3-13 所示。

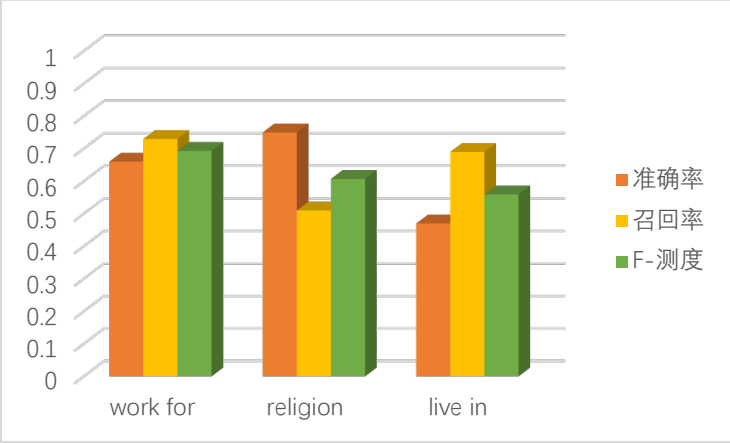

图 3-13 实体关系的准确率

3.4 暗网用户基础画像

经过基本特征抽取后，已取得 5 个用于后续分析的用户特征与 13 个用户画像中的属性。其中 5 个特征均为用户社交行为特征；13 个用户属性作为基本画像、社交画像、交易画像的分类如下表 3-3 所示。

表 3-3 用户画像属性表

序号	类别	用户属性
1	基本属性	用户名
2		用户 id
3		通讯方式
4		组织机构
5	社交	地区
6		用户等级
7		宗教
8	交易	交易身份
9		订单数量
10		交易品类型
11		比特币交易地址
12		以太币交易地址
13		PGP 交易标识

暗网用户基础画像如图 3-14 所示，分别用黄色、蓝色、绿色标注用户的基本信息画像、社交画像与交易画像。其中带有透视阴影的属性表示通过实体识别与实体关系抽取的得到分析结果。

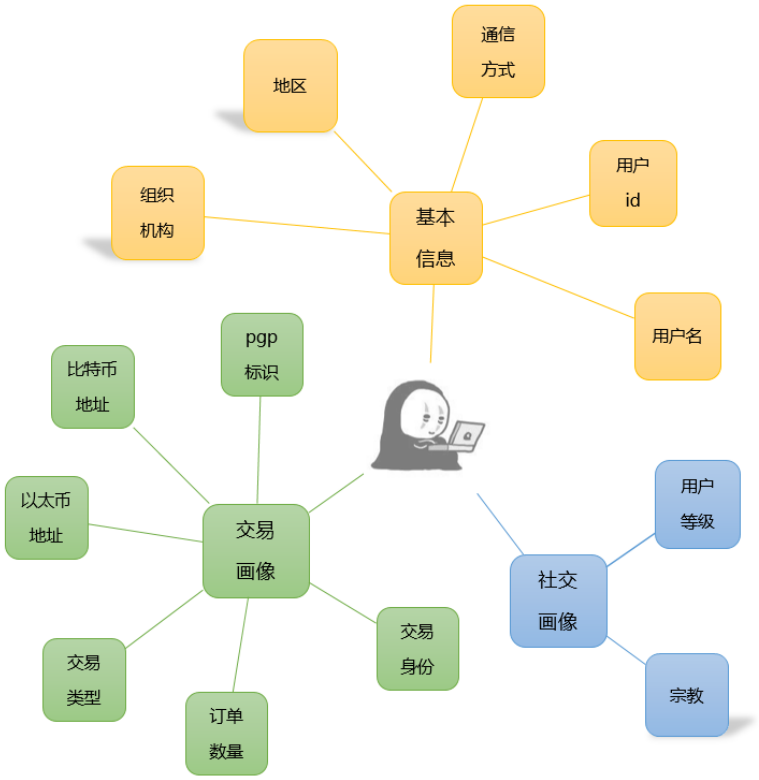

图 3-14 暗网用户基本画像

3.5 本章小结

本章利用暗网用户数据集构建出多维度暗网用户基础画像。对结构化数据通过解析与统计提取出部分用户属性及特征。对于非结构化文本数据通过建立出暗网中的高频黑话词库，结合根据暗网数据特征添加的自定义规则对文本进行实体、识别与实体关系抽取，进一步提取用户特征，从基本信息、社交、交易三个维度构建出用户的基础画像。

第 4 章 基于文本挖掘的暗网用户深度画像构建

通过前面的工作已经得到了暗网用户的多维度基础画像，为解决暗网用户属性稀疏的缺陷，构建更丰富的用户画像，需要对用户数据进行进一步挖掘。文本挖掘可以在文本、语料数据中发现潜在的具有价值的信息及规律，本章采用情感倾向分析、观点立场挖掘以及影响力等其它属性计算等方法，挖掘出更深层次的用户属性，从而构建出暗网用户的深度画像。

4.1 活跃度计算

从严格意义上来说，活跃度并不是一个特定的学术词汇，而是在不同的情景下被赋予不同的意义。比如，在证券行业，股票活跃度一般被用以反映某只股票交易频率的高低。在互联网领域中，用户活跃度一般是被用来衡量用户活动的频繁程度。而在互联网中，社交网络用户的用户活跃度定义与门户网站的用户活跃度定义又不相同，门户网站的用户活跃通常以用户的登录频次和停留时间来记录。而社交网络上用户的活跃度被定义为该用户在社交网络上单位时间内各种行为的频次和。对于用户量大且活跃程度高的暗网而言，活跃度是暗网用户社交画像中的一个重要属性。

4.1.1 活跃度算法

许多文章[41-43]认为用户的活跃度与其在社交网络中进行的行为频率相关，在 Radicchi[44] 的研究中还得出用户的平均活跃度与用户在网络中发布的总信息数量呈线性关系的结论。通过实验证实，以上结论在暗网中同样成立。本课题中从用户的注册及登录账号，主动发布信息，与信息发布者交互等行为中提取暗网用户活跃度的影响因素，通过分析暗网用户的在社交平台的各类行为的频率建立用户的活跃度计算模型。此外，采用层次分析法来确定由这些因素构成的不同活跃度指标所对应的权重。AHP(The Analytic Hierarchy Process)是一种结合了定性分析和定量分析的多因素层次化决策方法，由美国匹兹堡大学的萨迪提出。通过分析多因素之前存在的关系利用较少的定量信息，为多因素、多层次、无结构特性的决策问题提供有效的决策方法。根据 AHP[45]方法流程依次构建层次结构模型，其后建立判断矩阵，最终经过层次排序及其一致性检验得

到各指标权重。

本文设计了以账号注册时间 d_r，最后一次活跃时间 d_a，数据采集时间 d_c，用户等级 l，用户发帖数 p，评论数 c 作为的输入，基于 AHP 的基本思想确定不同指标的权重的暗网用户活跃度计算模型，输出为用户活跃度ε。ε的值域为[0,1]，越接近于 1 表示用户的活跃度越高，活跃度计算模型公式为：

$$\varepsilon = \partial \mathrm{l} + \sum_{i=1}^{k} \alpha_i a_i + \delta \tag{4-1}$$

其中∂为用户等级权重，δ为调节因子，$\sum_{i=1}^{k} \alpha_i a_i$为社区活跃度计量指标，$\alpha_i$为第 i 个指标在活跃度中的权重，$a_i$ 表示第 i 个活跃度指标，其计算公式为：

$$a_i = \begin{cases} \sigma\left(\frac{n_i}{1+d_c-d_r}\right), & i \le 2 \\ \sigma\left(\frac{n_i}{100}\right), & 2 < i \le 5 \end{cases} \tag{4-2}$$

其中 n_i 可为用户发帖数 p，评论数 c，最后一次活跃时间 d_a，σ为 Sigmoid 函数。

4.1.2 暗网用户活跃度分析

应用上述算法于本课题数据集中的暗网用户进行活跃度分析，并对结果进行计算和统计，绘制 CDF 图如下图 4-1 所示。

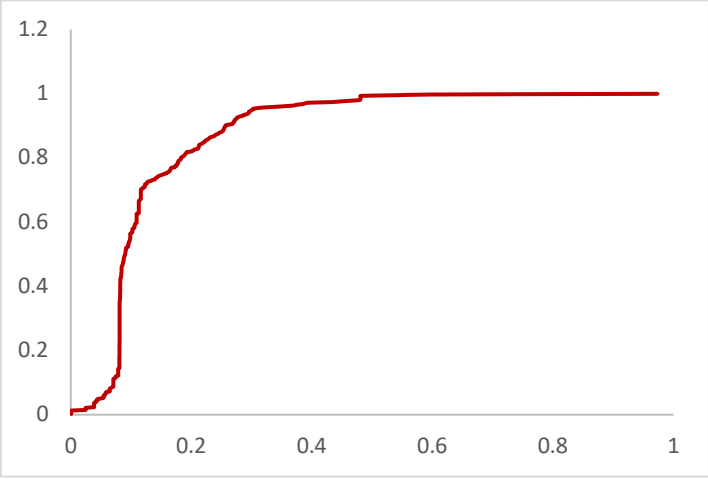

图 4-1 暗网用户活跃度 CDF 图

可以看出在暗网社交平台中，有很多沉默用户的存在，这些用户并不在暗网平台上发布信息或者与信息发布者发生交互行为，账号登录频率也较低，此类沉默用户占比率达 85%以上。而登录频率高，经常点赞、发表内容、分享观点的约活跃度较高的用户仅占约 10%。因此虽然暗网用户整体活跃度比表层网络更高，但在活跃用户分布率上，与 Cheng 等[46]对国外社交网络 Twitter 的用户

活跃度的研究结果基本一致。

4.2 影响力计算

在社交网站中用户的许多行为都会成为作用于影响力的因素，由于暗网中很少存在关注、粉丝等关系，因此需要更多的从用户行为中挖掘用户影响力的相关因素，同时不同的网站由于用户规模，用户评价等不同，其中用户的真实影响力也会随网站评分的高低变化。

4.2.1 影响力算法

目前主流的社交用户影响力计算方法主要基于 PageRank 算法实现，PageRank 算法运行在有网页与链接构成的有向图中，能够有效地判断网页的质量，是在网页排序的应用中非常经典的算法。PageRank 算法在两个假设下存在：当一个网站拥有越多的传入链接，则该网站越重要；当这些传入链接的起源网站拥有越高的重要性，则这些链接的含金量越高。这里的重要性便是该网站的 PageRank 值。可以对用户的影响力做出类似的推理，用户向越多人传递了思想，则其影响力越高；思想接受者拥有越高的影响力，则这次思想传播越有价值，越能体现思想源头的影响力。

然而，传统的 PageRank 算法关注的只是节点之间的关联关系，在 PR 值上采用平均分配的策略，考虑到的影响因素较少。但本课题将 PageRank 算法与用户行为因素结合，以用户间的评论、点赞关系构成有向图。由于在用户的实际互动中，不同用户之间的互动强度存在很大区别，因此本文考虑某用户和其它用户发帖的评论关系，增加可以调节的参数，使该用户对偏爱程度更高的用户在传递 PR 时值赋予更多比例的 PR 值，从而更合理地传递 PR。采用以下公式计算评论 PR 分配参数：

$$\text{comment}_{ab} = \frac{c_{ab} cu_a}{c_a} \tag{4-3}$$

comment_{ab}表示用户 a 传递给用户 b 评论 PR 参数，c_{ab}为用户 a 给用户 b 的评论总数，cu_a为所有被用户 a 评论过的用户总数，c_a表示用户 A 发出的评论总数。计算评论 PR 分配参数的实现流程如下。

算法 4-1 ：评论 PR 分配参数算法

输入： 社交数据评论类型列表

输出：评论 PR 分配参数comment_{ab}

BEGIN

通过社交数据表获取消息类型为 comment 的所有评论集合 comments

WHILE（comments 不为空）**DO**

Initialization $c_{ab}=0, cu_a=0,\ c_a=0,\ \text{comment}_{ab}=1$

IF 用户 a 评论了用户 b 的帖子

$\text{comment}_{ab}=\text{comment}_{ab}+1$

IF 评论用户的 id==用户 a

$c_a=c_a+1$

IF（评论用户的 id==用户 a）and （发帖用户的 id not in cus）

$cu_a=cu_a+1$

计算$\text{comment}_{ab}=c_{ab}cu_a/c_a$

END WHILE

END 评论 PR 分配参数算法

点赞与反对的 PR 分配参数也同样通过以上方法计算，再结合 Pagerank 算法计算出用户的评论影响力指标、点赞影响力指标与反对影响力指标。通过分析用户发表话题、评论，接收评论、赞踩等的情况，结合所在网站的评分，提出适用于评估暗网用户影响力的算法。算法输入还包括网站评分 r，用户等级 l 以及用户发帖数 p，评论数 c，输出值域为[0,1]的用户影响力τ，当τ接近于 1 表示用户具有较高的影响力，基于此建立出用户的影响力计算公式为：

$$\tau=\mathrm{r}\times(\mathrm{l}+\mu\frac{c+p}{1+c+p}+\vartheta\sum_{j=1}^{m}\alpha_j i_j) \tag{4-4}$$

其中μ、ϑ均为调节因子，$\frac{c+p}{1+c+p}$为言论量指标，$\sum_{j=1}^{m}\alpha_j i_j$为总影响能力，子影响力包括评论、赞、踩三个指标，$\alpha_j$为第 j 个子影响能力对应的权重，指标权重的计算方法同上一节活跃度模型中的计算方法相同。

4.2.2 暗网用户影响力分布

在本课题的暗网社交数据集中应用上述算法计算用户在暗网中的影响力，并对结果进行统计分析，绘制出暗网用户社交影响力的 CDF 图如下图 4-2 所示。从影响力分布图中可以看出，具有高影响力的意见领袖用户分布极少，暗网用户的影响力普遍较低，尤为特别的是，存在大量影响力为 0 的用户，在注册账号后几乎没有过发帖操作记录，或发表的帖子与动态未收到其他用户的评

论点赞等交互。

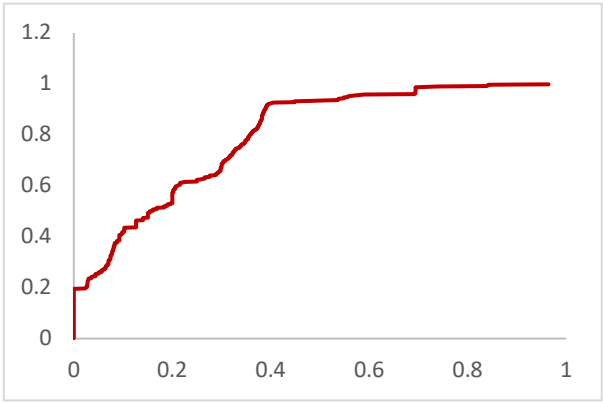

图 4-2 暗网用户影响力 CDF 图

4.3 言论情感分析

由于暗网的隐蔽性，许多用户借此躲避法律的约束，在暗网中肆意发表观点、宣泄情感，其中不乏大量极端言论，因此可以通过对用户挖掘言论内容，分析出用户个人的情感倾向。本节依然通过 langid 判断文本的语言类型，针对不同语言的文本采用不同的情感分析算法，中文数据使用 SnowNLP，对英文数据采用 TextBlob 和 Stanfordcorenlp 结合，同时结合自定义情感值库，对分析结果的准确率进一步提高。情感分析的输入为社交数据集 S 中的文本数据，输出为分析结果二元组 sentiment={polarity，subjectivity}，其中 polarity 为情感极性，subjectivity 表示情感主观性。

4.3.1 中文情感分析

对于数据中的中文文本数据本课题以 SnowNLP 为基础实现情感分析。SnowNLP 是一个 python 实现的类库，它可以方便快速的对中文文本内容进行情感分析，得到变化范围为[0,1]的正面情感的概率，0 表示 1 表示文本内容表达的情绪一定是负面的， 1 表示一定为正面情绪。在本课题中我们为了将中英文的识别结果统一规范化，使中文的情感分析结果也为变化范围为[-1,1]的情感极性，且情感极性与的情感质的变化规律也与英文中一致，我们将中文情感值定义为：

$$\text{sentiment} = (p - 0.5) \times 2 \tag{4-5}$$

其中 p 为 SnowNLP 输出的正面情感概率。

4.3.2 英文情感分析

本课题中以 TextBlob 为基础，结合 Stanfordcorenlp 进行双重情感极性验证，实现对英文文本的情感分析。TextBlob 是 Python 中常一个用于处理英文文本数据的库，其中的情感分析功能可以得到变化范围为[-1,1]的情感极性，-1 表示文本内容表达了完全负面的消极情绪，1 则表示文本内容表达了完全正面的积极情绪；以及变化范围为[0,1]的主观性， 0 表示完全客观，1 表示完全主观。

Stanfordcorenlp 直接将文本的情感分析为 Positive，Neutral，Negative 三个结果，通过 Stanfordcorenlp 的情感分类结果对原结果进行校准，将结果中的 Positive，Neutral，Negative 分别取值为 1，0 和-1，具体校准方法如下：

$$\text{sentiment} = \begin{cases} sp_t/2, & |sp_t| > 0.5 \ and \ sp_s sp_t > 0 \\ 0, & |sp_t| < 0.5 \ and \ sp_s sp_t < 0 \\ sp_t, & otherwise \end{cases} \tag{4-6}$$

其中sp_t为 TextBlob 输出的情感极性值，sp_s表示 Stanfordcorenlp 的情感极性分类结果。

4.3.3 自定义情感值库

在上述方法基础上上，人工收集暗网特征性较高的情感词构成情感种子词典，将正面情感词的情感值 s 设置为 1，负面情感词的情感值 s 设置为-1。通过同义词词林将种子词典进一步丰富[47]，待补充的同义情感词与种子情感词的相似度达到阈值φ，则将其归入情感词典，补充情感词的情感值定义为：

$$s(w_a) = \begin{cases} \max\limits_{w_s \in W} Sim(w_a, w_s) \times s(w_s), & \max\limits_{w_s \in W} Sim(w_a, w_s) \geq \varphi \\ 0, & \max\limits_{w_s \in W} Sim(w_a, w_s) < \varphi \end{cases} \tag{4-7}$$

其中 W={w_1,w_2,...w_n}为情感词集合，$Sim(w_a, w_s)$为词相似度计算函数。

人物的情感值由其参与讨论的性质来反应。一般通过一两个言论推断用户情感倾向是不准确的，但是从用户本人的大量的言论中进行统计分析得出的人物情感倾向是相对可靠和科学的，因此本文综合数据库中用户的全部言论文本中存在的情感词来推断的人物情感值，用户的情感值定义如下所示：

$$\text{sentiment} = \frac{\sum_{i=1}^{n} s(w_i)}{n} \tag{4-8}$$

4.3.4 暗网用户情感极性分布

将上述算法在应用于本研究的暗网数据集中，分析暗网用户的情感极性与主观性特征，并对情感极性的分析结果进行计算和统计，绘制出如下图 4-3 所示的 CDF 图。

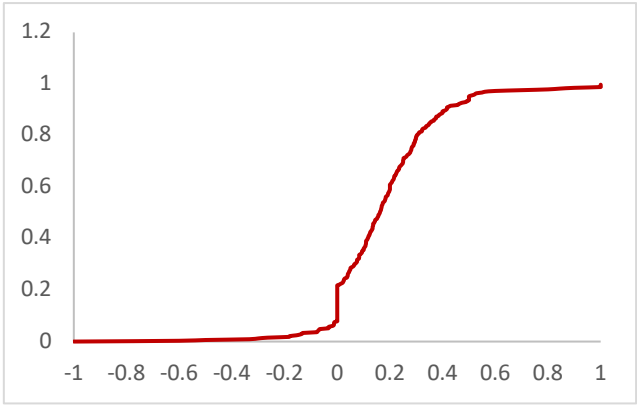

图 4-3 用户情感极性 CDF 图

4.4 观点立场倾向分析

相比于表层网络，许多用户更倾向于在能够隐藏身份的暗网中表达自己的政治观点等关于敏感话题的态度或传播恐怖主义等违法行为的言论，因此可以通过对目标主题进行立场分析，推断出暗网用户对特定观点的立场倾向。然而现有的立场分析方法大多基于辩论领域研究，其中的文本数据均带有明显且清晰的立场，同时辩手们通常会使用规律性较强的论证方式与句法结构；但在暗网的特殊环境中，立场表达更加隐晦，句法结构与表达方式也更加随性，因此相比于表层网络的立场分析，具有更大的难度。Sobhani[48]的研究表明，文本中表现出的情感与这段文本中作者想表达的立场有一定关联，因此本课题中加入文本的情感作为立场分析的特征，在一定程度上提高立场分析的效果。

4.4.1 特征提取

目前特征提取最为主流的方法是采用 TF-IDF 特征提取算法，该算法首先将文本无意义的词剔除，例如虚词、感叹词、停止词等，将去除后剩余的有效词作为文本的特征项，再根据统计学原理对特征项中的重要特征项进行提取。其中 TF 表示词频，意为词语在文档中出现的频率，IDF 意为逆文档频率，它与文档频率 DF 成反比，文档频率是指特征项在文档集合中出现的文档的数量。

TF-IDF 特征综合考量了一个词语在文本当中的代表性和区分度，如果一个有效词语在一个文档中出现的次数越多，那么这个词语就越能够代表这个文本的特征，越能表达一个文本的中心思想，即越具有代表性。而 IDF 越大，则表示这个词语出现在多篇文档当中过，表示越没有区分度。本课题中使用了 n-gram 模型得到单词向量，使用 TF-IDF 特征提取算法对 word 和 char 进行 TF-IDF 特征的抽取，即可得到 word 级和 char 级的 n-gram 特征矩阵。

此外，本研究中还加入情感极性也作为特征项，情感极性的分析主要通过上一节中介绍的方法实现。不同的是，在特征提取的过程中，取消对英文情感极性的双重验证，采用将两种方法的分析的原结果合并为二维情感特征的方式最大程度保留原始特征，用于后续立场分析。

4.4.2 分类判断

本课题结合项目需求，从暗网热点内容中选取 “分裂主义”、“恐怖主义”与“违禁药品”3 个主题的倾向进行立场分析，针对以上三个立场所标记的数据集中样例如表 4-1 所示。

表 4-1 政治倾向数据集中数据示例

类别	立场
How's the CCP membership treating you these days, wretch? Still upset that the sons of liberty will fight before they submit to tyranny?	分裂主义
谢谢国际上的好朋友，也一起把香*坚持民主的信念，传到全世界！	分裂主义
My team will wipe j***out of this earth with these	恐怖主义
Hi we are the Kill Stalk	恐怖主义
OG Kush was good, currently Bros Grimm Green Avenger is the daily diet	违禁药品
we will be getting new ketamine rocks/sugar, as well as new batch of mdma and pills	违禁药品

水平合并各种特征矩阵，合并的过程中先将稀疏矩阵对应的普通矩阵合并，再转化为稀疏矩阵。通过 SVM 分类器对本课题目标主题的立场进行二分类，代表对于目标立场的支持与否。

通过上述方法，对“分裂主义”、“恐怖主义”与“违禁药品”3 个主题的观点倾向进行立场分析。对三个话题的立场分析结果的平均准确率、召回率、F1 指标进行计算，结果如下表 4-2 所示。

表 4-2 话题立场分析结果

关系类型	准确率	召回率	F-测度
分裂主义	0.49	0.63	0.55
恐怖主义	0.62	0.51	0.55
违禁药品	0.53	0.47	0.49

4.4.3 自定义关键词库

可看出已有立场分析算法对本课题中的特定主体的立场分析效果并不理想，因此通过人工为每个主题创建关键词库，词库由暗网数据集中带有与主题相关的明显倾向性立场的词语，同时对出现频率较高的关键词通过同义词林进一步扩充构成相关词库，通过对关键词库与相关词库的匹配，提高本数据集中政治倾向的效果。优化后结果如图 4-3 所示，增加关键词库匹配后的各结果指标均好于原算法。

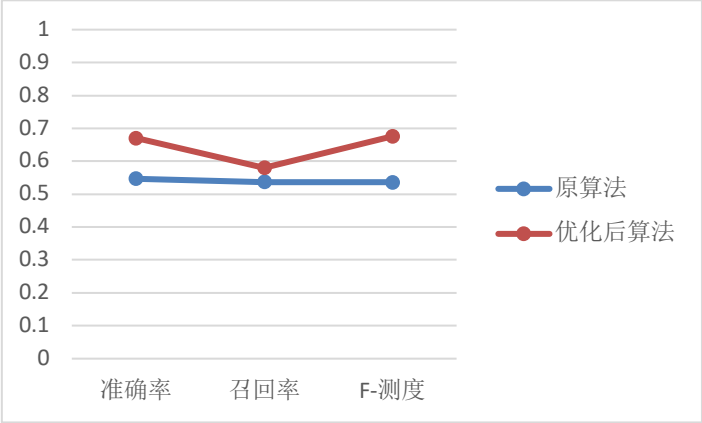

图 4-3 优化前后分析结果对比

4.5 交易规模分析与预测

在暗网中存在许多黑市交易，在基本特征抽取部分已经得到用户的交易身份，交易产品类型，交易量、以及各种交易标识等交易特征，本节将根据以上信息对用户交易规模进行分析，并采用 FTRL+XGBoost 算法模型，对未来的交易量进行估计预测。

4.5.1 交易规模分析

基于以上信息，根据交易身份，交易和交易量作为三个维度划分出用户的交易规模区间，从而可以将用户的交易规模标识为大型数据卖家、小型毒品买

家等。对卖家的规模统计如图 4-4 所示，其中小型卖家数量最多约占全部卖家的四分之三。

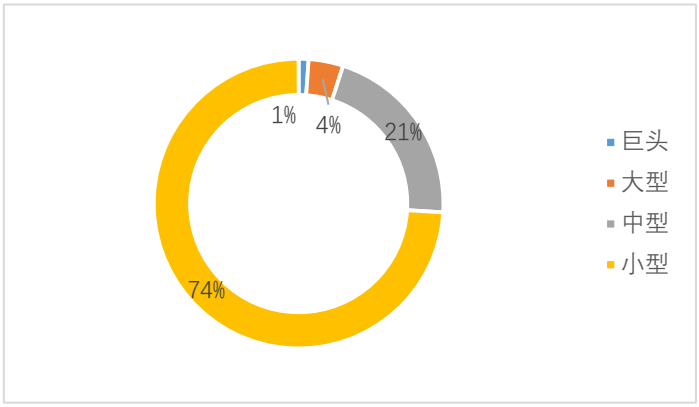

图 4-4 暗网卖家用户规模分布

在所占比重较高的小型卖家中，交易品类型的分布如图 4-5 所示，可以看出，不同于表层网络交易平台中的用户，暗网交易市场中的用户存在大量非法交易行为，涉及机密数据、色情视频、毒品、武器等方面，此外虚拟商品相比实物商品具有更大的交易量。

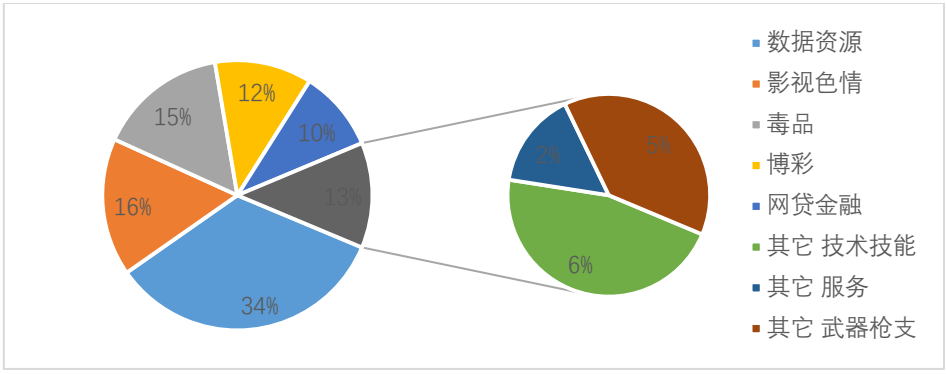

图 4-5 小型卖家交易品类型分布

4.5.1 交易量预测

暗网交易各个方面的数据进行整合，采用携程[49]提出的 FTRL +XGBoost 的机器学习算法模型，得出相关内容在各个时刻所处的态势，并根据相关的历史情况，对暗网商家的未来交易量趋势进行实现实时预测，满足对销量预测的准确性、可靠性和时效性的要求。

FTRL[50] (Follow the regularized leader)既能提高 OGD (online-gradient-descent)的精确度，又能获得很好的稀疏性，它是 2010 年谷歌提出的一个在线学习算法，FTRL 算法在精度、效率和泛化性上都提升很大，因此近年来被广泛

用于电商行业，取得了极好的效果。下面的公式即为 FTRL 的权重更新方法。

$$\omega_i^{t+1} = \begin{cases} 0, & if\left|z_i^{(t)}\right| < \lambda_1 \\ -(\lambda_2 + \sum_{s=1}^{t}\sigma(s))\left(z_i^{(t)} - \lambda_1 \sin\left(z_i^{(t)}\right)\right), & ootherwise \end{cases} \tag{4-9}$$

$$\sum_{s=1}^{t}\sigma(s) = \left(\beta + \sqrt{\sum_{s=1}^{t} g_{s,i}^2}\right)/\alpha \tag{4-10}$$

其中，ω_i是损失函数对第 i 维特征的梯度向量，$\sigma(s)$是学习率参数，α和β为超参数，λ_1和λ_2为 L1, L2 正则化系数。

XGBoost[51] (eXtreme Gradient Boosting)，是一种对 GBDT (Gradient Boosting Decision Tree)算法的改进算法。XGBoost 在 GBDT 的基础上，优化了并行训练，从而提高了效率，同时引入正则项防止过拟合。核心原理公式如下:

$$L^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in g_i} g_i\right)^2}{\sum_{i\in I_i} h_i + \lambda} + \gamma T \tag{4-11}$$

针对 XGBoost 训练和参数调优，需重点考虑下表 4-3 中所列参数，表中数据为最终调优结果。

表 4-3 最终调参结果

learning_rate	n_estimators	early_ stopping_ rounds	max_dept	min_child_weight
0.1	500	30	6	1.2

通过 FTRL+XGBoost 算法模型对未来的交易量进行预测后，结合交易类型规模对预测结果进行评价，以高、中、低作为交易量预测属性值。

4.6 暗网用户深度画像

通过上述算法对文本的进一步挖掘，可取得 21 个用户属性，包括 5 个基本属性，8 个社交属性以及 8 个交易属性，结合这些属性构建出暗网用户的人物画像。用户画像如下图 4-6 所示：分别用黄色、蓝色、绿色表示用户的基本画像、社交画像和交易画像，红色发光效果表示通过本节文本挖掘分析出的属性。利用暗网数据集构建出的用户画像，其中不同的暗网用户由于原始数据的来源与丰富程度不同，所以不同用户画像中的属性不完全相同，原始数据较稀疏的用户会存在部分属性的缺失。下图 4-7 所示的用户画像，根据宗教与对恐怖主义的立场属性推断，该用户极有可能为某教教徒且，支持恐怖主义，结合根据活跃度和影响力可以推断用户较长发帖参与讨论，但未引起较大影响。

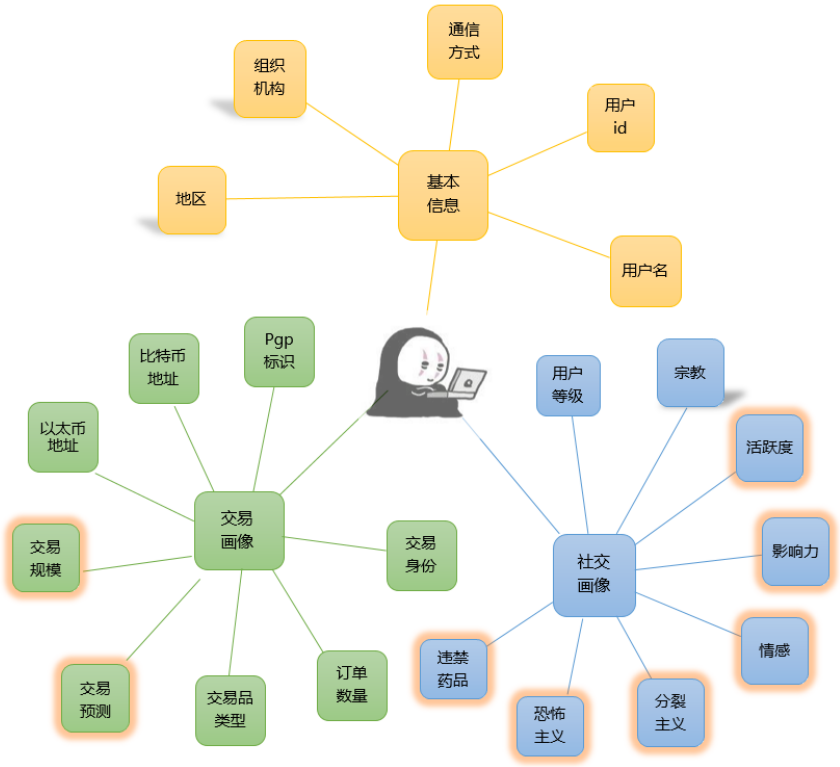

图 4-6 暗网用户画像结构

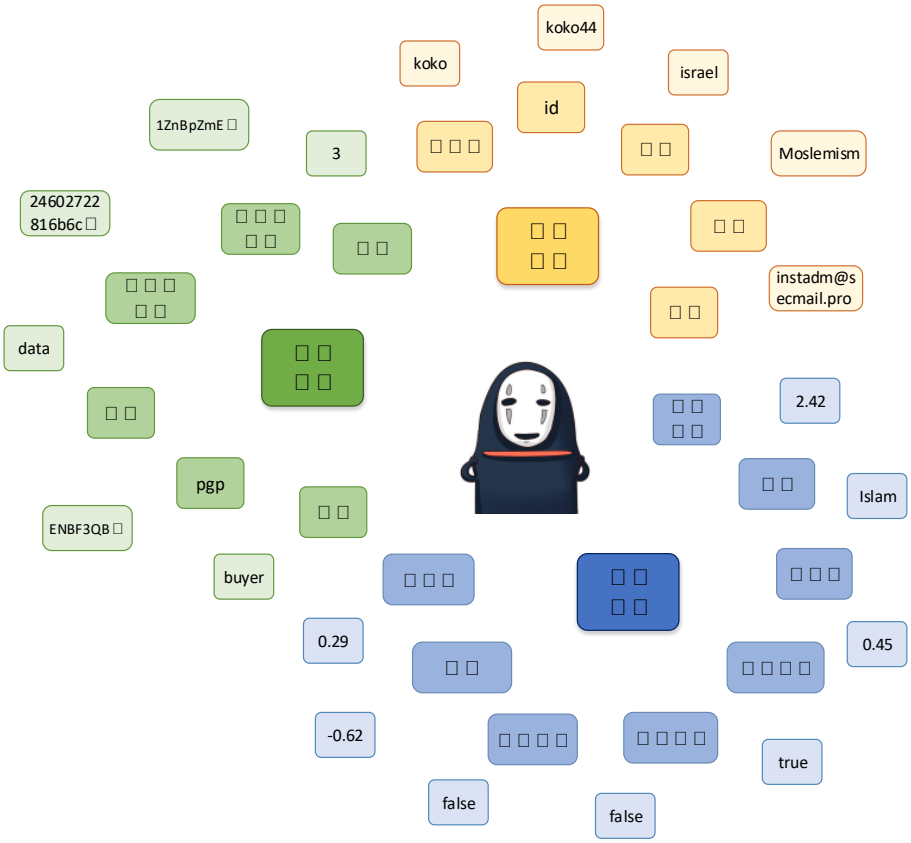

图 4-7 恐怖分子用户画像示例

下图 4-8 所示为另一暗网用户的画像示例，该用户有小规模售卖毒品的数据记录，但结合社交属性分析，用户在论坛中发布药品信息，没有参与其它话题的讨论，推测此账号极有可能为一个小毒品商户的贩毒专用账号。

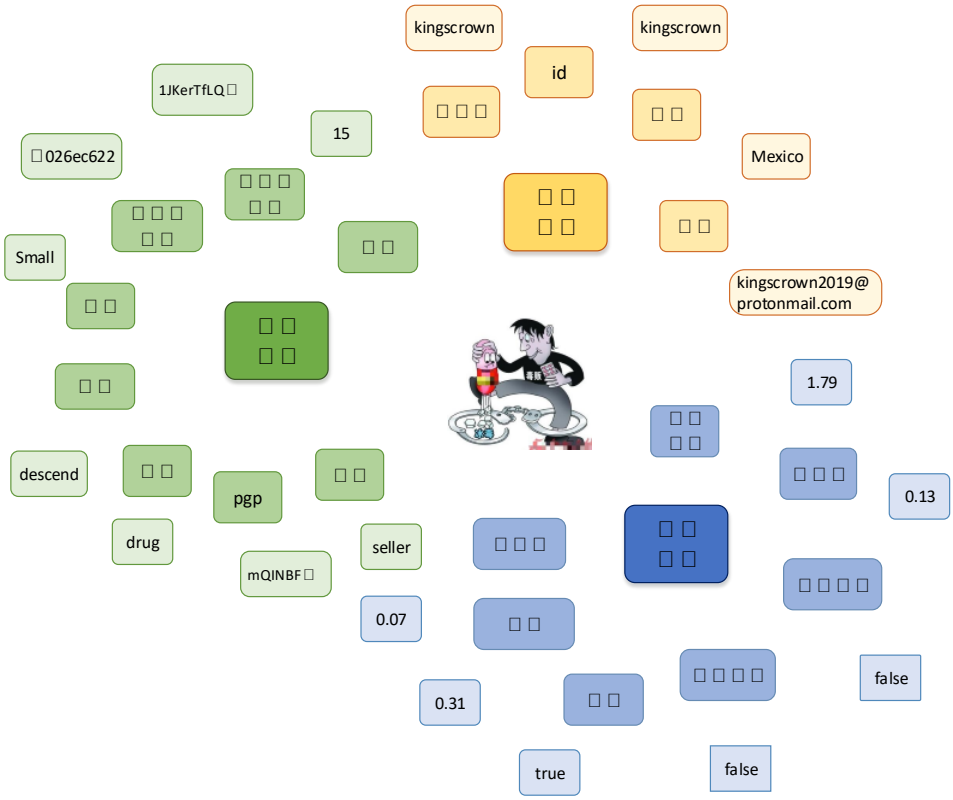

图 4-8 毒贩用户画像示例

4.7 本章小结

本章在前文工作中建立的暗网用户基础画像之上，对用户数据进行了更深层挖掘。在社交维度上，设计了适用于暗网的计算模型评估暗网用户的影响力、活跃度等属性；对非结构化的文本数据，分了析用户的情感以及对“分裂主义”、“恐怖主义”、“违禁药品”三个话题的立场倾向。在交易维度上分析商户的交易类别与规模，并预测了未来一段时间内的交易量。

第 5 章 基于暗网用户画像的虚拟群体发现应用

由于暗网中几乎不存在表层网络中具有的好友圈，关注与粉丝等网络结构，因此难以通过网络结构发现暗网中的用户虚拟群体。利用本课题中构建出的暗网用户画像，可以以计算暗网用户画像的相似度的方法，通过聚类实现暗网用户虚拟群体的发现。

5.1 用户画像相似度计算

为了计算画像相似度，使用 word2vec 方法把难以直接计算的用户画像信息转化为画像特征向量，word2vec 模型把每个词映射到一个向量，通过向量计算来表示不同词之间的关系。

通过 word2vec 的方法把暗网用户画像特征嵌入到连续的向量空间中，画像的相似度就可以通过向量的余弦相似度来度量，如果余弦相似度很高，则表示两个暗网用户的画像非常相似。由于不同属性对用户画像的重要性各不相同，因此在计算画像相似度时，为不同属性分配不同权重，如为用户 id、比特币交易地址等属性分配的权重较低，为政治倾向、情感极性等属性分配较高的权重。具体计算如公式下：

$$Sim_{profile}(A,B)=\frac{A_{profile}B_{profile}}{\|A_{profile}\|\times\|B_{profile}\|}=\frac{\sum_{i=1}^{n}\alpha_i A_i B_i}{\sqrt{\sum_{i=1}^{n}A_i^2}\sqrt{\sum_{i=1}^{n}B_i^2}} \tag{5-1}$$

其中$A_{profile}$和$B_{profile}$分别为用户 A 和用户 B 的画像，A_i表示用户 A 画像中的第 i 个属性，α_i表示为用户 A 画像中第 i 个属性分配的权重，n 为暗网用户画像中属性的数量。

5.2 用户虚拟群体聚类

为了发现具有较高相似模块度的群体，获得更好的聚类效果，本课题结合局部优化和多层次聚类技术设计了一种基于 Louvain 模块度[52]优化算法的聚类算法。Louvain 算法由 Blondel 和 Guillaume 等人提出，该算法通过迭代进行模块度局部优化和群体合并，直到模块度不再增加为止。本文定义群体相似系数Q_{group}作为群体聚类的度量指标，Q_{group}取值范围为[-1,1]，群体相似系数越大，代表群体内部相似性越高，Q_{group}的计算方法如下：

$$Q_{group}=\begin{cases}1 & n=1\\ \frac{1}{n}\sum_{i=1}^{n}Q_{node}(i) & n>1\end{cases} \tag{5-2}$$

其中 n 为群体内节点个数，$Q_{node}(i)$表示节点 i 的相似系数的取值范围为[-1,1]，与群体相似系数相似，节点相似系数越大，代表节点与所在群体相似性越高，节点相似系数计算方法为：

$$Q_{node}(i) = \frac{p(i)-q(i)}{\max\{p(i),q(i)\}} \tag{5-3}$$

其中$p(i)$为用户节点 i 到同一群体内其它用户节点的相似度平均值，$q(i)$为用户节点 i 到距离最近的其它群体内用户节点的相似度平均值。

群体在完成一次局部优化后群体相似系数的变化定义为群体相似系数增益ΔQ_{group}，本文以ΔQ_{group}作为启发式信息，算法流程如图 5-1 所示：

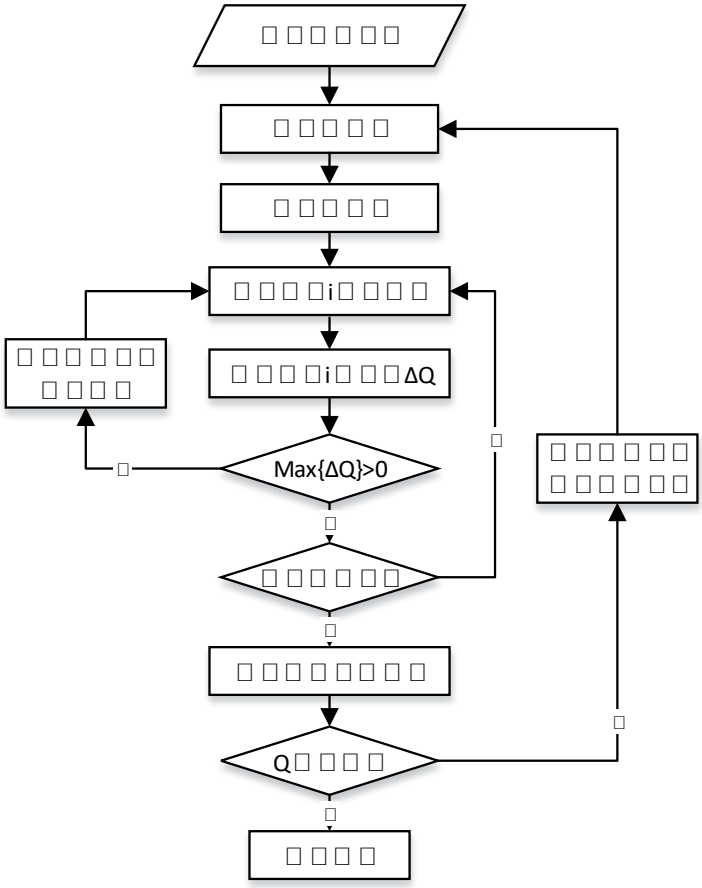

图 5-1 迭代流程图

具体步骤如下：

（1）相似度计算；每个暗网用户作为一个节点，根据公式(5-1)计算所有用户节点对的画像相似度。

（2）群体初始化；先将每个节点看做一个群体，将所有单一节点划分成不同的初始群体，并计算每个群体的Q_{group}。

（3）局部优化；使每个节点在其邻近区域内局部优化，逐一选择节点，计

算将它移动到其相邻群体中的增益ΔQ_{group}。如果最大增益大于 0，则将它移动到最大增益对应的相邻群体；否则，保持在原群体不动。

（4）重复执行步骤（3）；当所有节点的归属群体都不再变化时结束。

（5）构建新图；将得到的每个群体作为一个新的超级用户节点，更新这些节点间的相似度跳至步骤（2），直到获得最大的群体相似系数。

上述迭代步骤分为两个阶段 ：第一阶段：步骤（2）至（4），用于寻找各节点的归属群体，直到结果不再变化；第二阶段：步骤（5），这一阶段将构建新图，并重复执行第一阶段的操作，直到群体相似系数不再增加。

为了比较直接通过暗网用户原始数据与基于暗网用户画像的群体聚类效果，本文引入总相似比率 TSR(Total Similarity Ratio)，TSR 定义为平均群体内部相似度和平均群体间相似度的比值。下表列出了两种聚类方法当Q_{group}取最大值时 TSR 在不同数据下的最大值的对比情况。不难看出，基于暗网用户画像所发现的群体的总相似比率有所提高，有些数据可以提高到 15%以上。

表 5-1 总相似比率 TSR 对比

数据编号	基于原始数据	基于暗网画像	TSR 提高率%
1	6.360	6.741	5.991
2	5.481	6.337	15.618
3	7.057	9.479	34.321
4	6.370	6.749	5.950
5	4.308	4.781	10.980
6	6.919	7.896	14.121

5.3 本章小结

本章对课题中构建出的暗网用户画像应用到暗网用户虚拟群体的发现中。将用户画像信息转化为画像特征向量，采用计算余弦相似度的方法得到暗网用户画像的相似度，并设计基于 Louvain 模块度优化算法的聚类算法实现暗网用户虚拟群体的发现。

结　论

现阶段，对用户画像的研究大多还停留在表层网络的层面，主要是针对表层网络中的社交网络或其它单独平台进行研究，对多源数据融合的用户相对较少。而对暗网威胁数据的获取分析目前也主要用于情报获取，对属性稀疏的暗网用户缺乏个人属性的挖掘研究。

本文面向暗网，通过多种自然语言处理与文本挖掘算法分析暗网用户属性，从三个维度构建出暗网用户画像，并应用暗网用户画像实现暗网用户虚拟群体的发现。通过实验发现，本文设计的用户属性提取算法在暗网具有很高的适用性，暗网用户各属性分析结果的准确率相比现有相关研究均有提高，为暗网用户隐蔽，信息稀疏，难以构建用户画像的问题提供了初步解决方案。文中对各算法的优化设计，也同样可以应用到明网中。

本文主要工作内容可以分为以下几个方面：

(1)暗网数据大规模采集及预处理。对暗网网页数据进行大规模采集，并对暗网中复杂且不规范的数据进行预处理。针对暗网的隐蔽性导致采集到的数据属性稀疏的问题，通过与明网中数据关联来扩充属性加以改善。

(2)暗网用户多维度基础画像构建。从文本数据中通过命名实体识别从文本数据中识别出各类实体，再通过实体关系抽取，得到用户相关的部分基础属性；从结构化数据中解析出用户属性相关字段，从基本信息、社交行为、市场交易三个维度构建出暗网用户基础画像。

(3)基于文本挖掘的暗网用户深度画像构建。利用数据集中基础特征指数数据，建立计算模型评估暗网用户的影响力、活跃度、交易规模等属性，并预测未来交易量；对文本数据进一步分析用户的情感、话题立场等社交属性；在暗网用户基础画像的基础上构建出深度画像。

(4)暗网用户画像应用。通过计算暗网用户画像的相似度，并依据画像相似度实现暗网用户虚拟群体的发现，将暗网用户画像应用到暗网用户虚拟群体的发现中。

在以上的研究中，由于时间因素和条件限制，仍存在这一些不足，具体包括以下几点：

(1)本课题数据源几乎全部来自于暗网，由于暗网的复杂性，数据的真实程度难以保证。一些具有时效性的属性，需要不断更新，保证后续分析的准确性。

(2)暗网用户画像的属性相比于明网中略为稀疏，由于暗网的隐蔽性，在明

网中通过建立网站等平台容易获得的许多属性在暗网中只能通过分析推测获取。

(3)课题当前只以暗网数据中的少部分数据为研究基础，因此对于通过这些数据分析得到的结果能否代表全部暗网还有待进一步分析。

参考文献

[1] Kirkpatrick, Keith. Financing the dark web[J]. Communications of the ACM, 2017,60(3):21-22.

[2] Daniel Moore, Thomas Rid. Cryptopolitik and the darknet[J]. Survival Global Politics&Strategy,2016,58(1):7-38.

[3] IntSights. Financial services threat landscape report:The dark web perspective [EB/OL]:(2018-08-01) https: //intsights.com/resources/financial-services-threat-landscape-report-July-2018

[4] 刘伟,孟小峰,盟卫一.暗网数据集成问题研究[J].计算机学报.2007,30(9). 1475-1489.

[5] 王知津,范淑杰,王朋娜.竞争情报搜集与利用中的信息资产[J].图书馆学研究, 暗网数据集成问题研究,2011 (4): 2-6

[6] Konrad R, Philipp T, Carsten W, et al. Automatic analysis of malware behavior using machine learning[J].Computer Science, 2011,19(4):639-668

[7] Elike H, Xavier B, Andrew H， et al. Threat analysis of IoT networks using artificial neural network intrusion detection system[C].Proc of the 2016 Int Symp on Networks, Computers and Communications(ISNCC).Piscataway, NJ:IEEE, 2016

[8] Alan S, Richard E O, Tomasz R. Detection of known and unknown DDoS attacks using artificial neural networks [J].Neurocomputing, 2016, 172- 385-393.

[9] Nunes E, Diab A, Gunn A, et al. Darknet and deepnet mining for proactive cybersecurity threat intelligence[C]. Proceedings of IEEE Conference on Intelligence and Security Informatics(ISI). Piscataway, NJ:IEEE, 2016:7-12

[10] Dong Fangzhou, Yuan Shaoxian, Ou Haoran, et al. New cyber threat discovery from darknet marketplaces[C]. Proceedings of IEEE Conference on Big Data and Analytics(ICBDA).Piscataway, NJ:IEEE, 2018; 62-67.

[11] Sapienza A, Bessi A, Damodaran S, et al. Early warnings of cyber threats in online discussions[C]. Proceedings of IEEE International Conference on Data Mining Workshops (ICDMW). Piscataway,NJ:IEEE, 2017:667-674

[12] B.Schiffman, I.Mani.Concepcion.Producing biographical summaries: combining linguistic knowledge with corpus statistics[C]. Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'2001). New Brunswick, 2002, 450-457.

[13] L. Zhou, M. Ticrea, E. Hovy .Multi-document bio-graphy summarization[C].

Proceedings of EMNLP, 2005,434-441.

[14] E. Filatova, J. Prager. Tell me what you do and I'll tell you what you are: learning occupation related activities for biographies[C].Canada:HLT/EMNLP Vancouver, 2005:113-120.

[15] Y.J. Han, S.Y.Park, S. B. Park, et a1. Reconstruction of people information based on an event ontology[C]. Natural Language Processing and Knowledge Engineering, 2007, 446-451.

[16] 任宁. 大规模真实文本中的人物职衔信息提取研究[D].北京语言大学,2008.6-7.

[17] J. Tang, J. Zhang, L. Yao, et al. Arnet Miner: extraction and mining of academic social networks[C]. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, USA, 2008, 990-998.

[18] J. Tang, J. Zhang, D. Zhang, et al. Arnet Miner: an expertise oriented search system for web community[C].Semantic Web Challenge, Busan Korea, 2007.1-8.

[19] P. Gundecha, S. Ranganath, Z. Feng, et al. A tool for collecting provenance data in social media[J]. Acm Sigkdd International Conference on Knowledge Discovery&Data Mining, 2013:1462-1465.

[20] 乔磊,李存华,仲兆满,王俊,刘冬冬.基于规则的人物信息抽取算法的研究[J].南京师大学报(自然科学版),2012,35(04):134-139.

[21] 李红亮. 基于规则的百科人物属性抽取算法的研究[D].西南交通大学,2013.15-16.

[22] 周婷. 异构信息源的领域人物信息抽取研究[D].哈尔滨工业大学,2010.17-19.

[23] 于琨,管刚,周明,王煦法,蔡庆生.基于双层级联文本分类的简历信息抽取[J].中文信息学报,2006(01):59-66.

[24] 刘金红,陆余良,施凡,宋舜宏.基于语义上下文分析的因特网人物信息挖掘[J].安徽大学学报(自然科学版),2009,33(04):33-37.

[25] 郝冬生. 基于网页完整理解的人物信息抽取[D].吉林大学,2012.14-15.

[26] 李赵洁. 基于文本的人物画像挖掘技术的研究与应用[D].电子科技大学,2016.1-79.

[27] 张永超. 暗网资源挖掘的关键技术研究[D].西安电子科技大学,2013.1-5.

[28] 黄莉峥,刘嘉勇,郑荣锋,李孟铭.一种基于暗网的威胁情报主动获取框架[J].信息安全研究,2020,6(02):131-138.

[29] 王有文.第二代洋葱路由匿名系统 Tor 的性能改进研究[D].北京:北京邮电大学,2017.12-19.

[30] Peshave M, Dzhgosha K. How search engines work: And a Web crawler

application [D]. Springfield: University of Illinois.Springfield, 2005.
[31] Wood J. The darknet:A digital copyright revolution[J]. Richmond Journal of Law & Technology, 2018, 16 (4):15-17.
[32] 赵福祥.可靠洋葱路由方案的设计与实现[J].计算机学报, 2001, 24(5):96-102.
[33] 韩越.Tor 匿名通信系统路由技术研究[D].北京:北京邮电大学,2016.65-69.
[34] 王文龙. 面向突发事件案例库的事件抽取模型构建研究[D].南京大学,2015.13-21.
[35] Manning C D, Surdeanu M, Bauer J, et al. The Stanford coreNLP natural language processing toolkit[C] Proc of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2014:55-60.
[36] Yang B, Yih W T, He X, et al. Embedding Entities and Relations for Learning and Inference in Knowledge Bases[J]. International Conference on Learning Representations.2014.1-12.
[37] JU M, MIWA M, ANANIADOU S. A Neural Layered Model for Nested Named Entity Recognition [C]. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistic: Human Language Technologies, Volume 1 (Long Papers), 2018:1446-1459.
[38] IN H, HOU L, LI J, et al.Fine-Grained Entity Typing Via Hierarchical Multi Graph Convolutional Networks[C]. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019:4970-4979.
[39] Che W, Li Z, Liu T. LTP:a Chinese Language Technology Platform[C]. International Conference on Computational Linguistics:Demonstrations. Association for Computational Linguistics, 2010.1-4.
[40] H.E. Tingling, X.U. Chao, L.I. Jing, et a1.Named entity relation extraction method based on seed self-expansion[J]. Computer Engineering, 2006, 32(21):183-184.
[41] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion[C]. Twenty-ninth AAAI conference on artificial intelligence. 2015:2181-2187.
[42] Godoy D. Learning user interests for user profiling in personal information agents[J]. AI Communications,2006,19(4): 391-394.
[43] Domen Kosir, Igor Kononenlco, Zoran Bosnic. Web user profiles with time-decay and prototyping[J]. Applied Intelligence, 2014, 41(4):1081-1096.
[44] Brin S, Page L. The anatomy of a large-scale hyper textual Web search engine[J].Computer Networks and ISDN Systems, 1998, 30(1):107-117.

[45] Saaty T L. What is the analytic hierarchy process?[M]. Mathematical Models for Decision Support Springer-Verlag New York, 1988:109-121.

[46] Cheng A, Evans M, Sing H, Inside Twitter: An in-depth look inside the Twitter world [J]. Report of Sysomos, June, Toronto, Canada,2009.2-21.

[47] 梅家驹,竺一鸣,高蕴琦,等.同义词词林[M].上海:上海辞书出版社,1993,106-108.

[48] Maleszka B. A method for determining ontology-based user profile in document retrieval system[J]. Journal of Intelligent&Fuzzy Systems, 2016: 1-11.

[49] 钟小勇.基于 FTRL 和 XGBoost 组合算法的电商销量预测系统[J].信息记录材料,2020,21(01):1-3.

[50] McMahan H B. Follow the regularized leader and mirrordescent:equivalence theorems and L1 regularization[C]. Proceedings of the 14th International Conference on Artificial Intelligence and Statisitcs, 2011.1-9.

[51] Chen TQ,Guestrin C.XGBoost:A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and DataMining. San Francisco, CA, USA. 2016.785-794.

[52] Blondel V D, Guillaume J L, Lambiotte R, et al.Fast unfolding of communities in large networks[j].Jurnal of Statistical Mechanics Theory & Experiment,2008, 30(2):155-168.

哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《面向暗网的用户画像构建技术研究与应用》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：车馨悦　　日期：2020 年 6 月 28 日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：车馨悦　　日期：2020 年 6 月 28 日

导师签名：李东　　日期：2020 年 6 月 28 日

致　谢

至此，我的研究生生涯已进入了倒计时，在哈工大的 6 年时光也开始在我脑中不断回放。每一句“规格严格、功夫到家”，每一位尽职尽责的老师，每一个温暖贴心的朋友我都会铭记不忘。在此对给予过我帮助的人们表示感谢。

感谢我的导师李东教授。李东教授为人谦和、治学严谨。在学业上，时刻提醒我们要端正学术态度，专心搞好学术；在生活上，要求我们兼顾身心健康。我会记住您的教导，做到“温文尔雅、豪迈奔放”，追求卓越，成为更优秀的人。

我要感谢张宏莉教授。作为网安实验室的领导者，张宏莉教授为我们提供了先进的学习资源与浓厚的科研氛围。虽然事务繁忙，张宏莉教授仍每周定时和我们开会讨论，查看我们的研究进度，及时指出我们的问题与不足。感谢张教授！

感谢王星老师。非常幸运能够加入王星老师带领的项目组。组内的丰富的项目使我有机会接触到先进的技术，拥有了更多实践机会，也在每一次任务中都能感受到组内愈发强大的凝聚力。感谢星哥对团队的付出与对我的悉心指导！

感谢组内的伙伴们。像大哥哥一样的照顾我们主席王金麟；为我们提供很多生活妙招的沈卓学长；在我初来乍到时带我熟悉项目的孟超学长；最辛苦也是给我最多帮助的皓天哥；还有郜煜姐、董旭学长、庆丰学长、管志诚学长。也欢迎昌泽、伟杰、李川、田泽庶、项宇豪、张枫的加入。希望你们都能健康快乐。

感谢我的朋友们。感谢公主，六年来总是在我被问题难住时成为我的小老师为我答疑解惑；感谢语晨，在最辛苦的一次出差中与我作伴；感谢卢昆，用唱跳 rap 篮球为我们带来了轻松和愉快；感谢张济川作为项目组副组长，承担了许多任务，带我一起进步；感谢饶宇鑫，两年来大多数的快乐时光都有你的参与；感谢王志文，总是事事都为我着想；感谢我的闺蜜温慧，在每个压力最大的阶段给我鼓励，你也要变得自信！感谢你们的陪伴，希望你们以后遇到的都是可爱的人。

最要感谢的是我的父母。感谢父母给了我一个最温馨的家庭，成为我坚固的后盾，我会努力提升自己，不辜负你们的期望！

最后在哈工大百年校庆之际，感谢母校的对我栽培，祝母校宏图大展，再谱华章！