毕业论文（设计）任务书

题目　一种面向微博主题的自动画像系统

队　　别　智慧警务三队

专业（方向）　数据警务技术

姓　　名　魏志豪

中国人民警察大学

题目	一种面向微博主题的自动画像系统
论文（设计）任务	（一）标签化文本分类 互联网上的大量个人数据，可以轻松得知一个人的各式各样的独特标签，它可以是一个人的购物表象，通过类比推理，他也可以是一个人的犯罪表象。例如，通过聚类分析总结同一类犯罪人的特点，当他们的标签都趋向于同一方向时，利用大数据同步监测到新的一名用户加入到此画像群体中，此用户作为研究对象，判断其是否为新的犯罪人或者是以及在逃的犯罪人。具体内容是为文本预设标签，将待处理的文本准确分类到预设的标签中，情感分析、对话分类以及主题分类。利用朴素贝叶斯，K 近邻算法以及支持向量机。 （二）搭建深度学习模型 对文本进行数据清洗，对其进行特征表示，之后进行浅层学习模型和深度学习模型搭建，选取特定的评价标准准确度输出标签。对于每时每刻都在变化的主题画像，横向比对分析，不断定制事件标签，将标签的增长速度也能够衡量一个事件的属性，例如刻画已发生舆情事件的增长变化，去拟合新事件的发展变化，将画像的生长变化作为一个数据对比普通时间的生长变化。 （三）基于知识蒸馏的联合模型 传统主题模型建立微博画像存在主题一致性较低，缺乏解释性的缺点，通过使用知识蒸馏的方法将预训练模型学习到的知识迁移到学生模型中，学生模型具有较小的参数及较短的训练时间，以至于进行实时监控微博信息并及时训练生成实时画像。 （四）主题分类 构建用户标签画像，能够个性化推荐每个用户想要的商品，同样的道理，同一类犯罪嫌疑人往往会存在相同的共性，在此想法上，通过已有嫌疑人的特征画像去拟合全体画像，或许能够发现犯罪预备人或者是犯罪嫌疑人，在主题层次上，我们同样能够

	将已有的舆情事件数据分主题进行画像，将多个共性参数形成画像的多个标签，由于事件主题从无到有到产生舆情是一个过程，对此过程进行研究分析画像的演变过程，并与新事件主题进行拟合，以此来发现众多新事件中的可能会发生舆情的事件。 （五）公式超参数取值 通过大量的画像对比性分析，选出最优的参数，提高对于舆情事件的判断精准度，实现舆情事件的精准预测，捕捉其演变过程，对演变过程的参数变化进行分级取值，建立模型。传统主题模型建立微博画像存在主题一致性较低，缺乏解释性的缺点，通过使用知识蒸馏的方法将预训练模型学习到的知识迁移到学生模型中，学生模型具有较小的参数及较短的训练时间，以至于进行实时监控微博信息并及时训练生成实时画像。实时预警需要这样一款量身定制的新型模型，设计并采用这样一款模型让实时系统更为即时可用。 （六）实验环境模拟 通过爬取大量的主题进行标签化聚类分析、细分、建立模型、发现共同模式。分析研究引起舆论热潮的大量主题，将此通过机器学习、神经网络训练为可实战化利用的模型，通过对比即可及时预测事件的走势发展，防止舆情事件的发生，同时也可以建立一些虚拟的主题场景，验证我们的想法是否成立，成立即可逐步应用于警务实战化平台，预测监控以及防止舆情事件的发生扩大。 （七）实现价值 搭建全平台一体化分布式系统，管理员及各级用户都分有不同的权限，分配不同主题的画像查看与舆情预警。将所有舆情事件和安全事件的画像参数存储在数据库中，每十分钟更新数据库，并及时提取数据。利用深度学习和神经网络不断设计优化模型，不断适应新的事态发展趋势。将全部的舆情事件分主题的绘制画像，对参数优化，以及没有舆情的事件进行画像，实时对比同主题的新型事件进行超参预警。

时间安排	2022年10月15日—2022年10月17日：开题。 2022年10月22日—2023年3月15日：软件系统的编写。 2023年3月16日—2021年3月29日：中期检查。 2023年3月30日—2021年4月19日：指导教师、评阅人评定成绩。 2023年4月20日—2023年4月30日：学术不端行为检测。 2023年5月8日—2023年5月22日：毕业设计答辩，成绩评定。
任务下达日期 年　　月　　日 指导教师（签名）： 学　　生（签名）：	
教研室主任意见	签名： 年　　月　　日

中国人民警察大学

毕业论文（设计）开题报告

题目 一种面向微博主题的自动画像系统

学　　号	2019160053
姓　　名	魏志豪
队　　别	智慧警务三队
专业（方向）	数据警务技术
指 导 教 师	安晓伟 讲师

二〇二二　年 十 月

一、课题的来源

智能手机等终端设备带给人们便捷性的同时，也无时无刻将人们的想法遍及整个互联网，会造成信息的自我公开，随着互联网用户普及率即将到达 100%，人们的想法也随之拥有潜在的公开透明，当你的想法表达到互联网上，可能你喜欢的产品不一定出现在你的礼物中，但会出现在你的个性化广告中，你会拥有一台最懂你的智能手机。这些都是画像存在的结果。

微博主题画像是有关部门在监测网络舆情过程中根据已获得微博信息数据对某些特别关注的主题进行描述和刻画，类比于公安机关的侦察画像可了解到，侦察画像是公安机关寻找犯罪嫌疑人的有力法宝，以能够在现实空间中轻而易举地抓获准确的犯罪嫌疑人；而在互联网空间中，网络不是法外之地，一切违反法律的言论甚至谣言都应受到应有的处罚，于是我们很有必要采取措施去监控网络动态，及时捕捉微博敏感数据，及时处理数据并描绘出可疑分子，并及时的分析研判，最后达到发现可疑事件的微博主题发出预警。在中国知网，以“微博画像”为主题进行搜索可获得相关文献 255 篇，发表时间为 2011 年到 2022 年，梳理这些文献可以发现微博画像技术的研究主题内容及发展阶段。通过检索可以发现，大部分研究方向都偏向于微博用户画像的研究，而对于微博主题画像的研究则少之又少，所以特别需要对微博主题进行研究，通过整体全面地从大局上描绘出整个主题的画像，作为一种面向主题的模型，给主题加上标签，通过可视化形成画像，从中发现可疑点。“神笔警探”林宇辉，专注模拟画像，侦破无数大案，甚至在没有任何目击者的情况下，根据已有信息描绘极为相似的嫌疑人画像。互联网上可利用的数据更多，需要我们去挖掘利用。本课题以此为背景，本人结合专业所学知识，对微博主题画像进行研究，通过软件系统的形式实现研究目的，对我国当前互联网舆情治理具有重要的理论意义和现实意义。

二、本课题国内外研究动态

通过 CiteSpace 的分析国内外文献，我们可以清楚的得知，在微博画像领域中，哪些关键词占据着主流地位？不同的研究方向之间是如何相互关联的?在微博画像研究领域的发展历程中，哪些文献起着关键作用?微博画像中的知识基础和研究前沿是什么?研究前沿是如何演变的？并对关键词进行共词和聚类分析以发现研究热点。归纳了用户画像领域的主要研究方向及研究状况，为下一步的研究工作提出了建议。

（一）国内研究动态

2022 年 6 月，中国互联网络信息中心(CNNIC)发布了第 50 次《中国互联网络发展状况统计报告》，据报告中显示截至 2022 年 6 月，我国网民规模达 10.51 亿，较 2021 年 12 月增长 1919 万，互联网普及率达 74.4%，较 2021 年 12 月提升 1.4 个百分点。我国手机网民规模达 10.47 亿，较 2021 年 12 月增长 1785 万，网民使用手机上网的比例为 99.6%，与 2021 年 12 月基本持平。如此大规模的用户使用网络，并不断在网络上留下他们所生成的评论、留言、点赞，以及上网过程中留下的浏览记录、搜索痕迹等，造成了大量的数据残留，从而导致用户无法迅速便捷的在网络上查找到所需信息。因此，学者开始思考如何有效从大规模的数据中挖掘它所隐藏的价值，从而缓解甚至消除这些问题。用户画像在这时逐渐被学者们所关注，相关的研究文献也在不断增多。2011 年开始至 2022 年，研究文献数量持续增加，到现在有 255 篇有关微博画像的研究。如图一关键词出现的频次越高，则该节点越大;节点之间的连线越粗，则说明关键词之间的共现强度越大。从图中可以看出，在用户画像的研究领域中出现频次最多的关键词是“用户画像”，其次是“大数据”和“社交网络”，此外。还有“数据挖掘”“推荐系统”“个性化推荐”“协同过滤”“画像”“机器学习”“hadoop”“用户”等关键词提及频次也较高。这些关键词的内容显示了用户画像研究领域研究的主体内容。可见国内很大一部分研究仍是为了用户的个性化推荐，使得网站更受使用者的青睐。而很少有将其应用于网络舆情探测预警的研究。

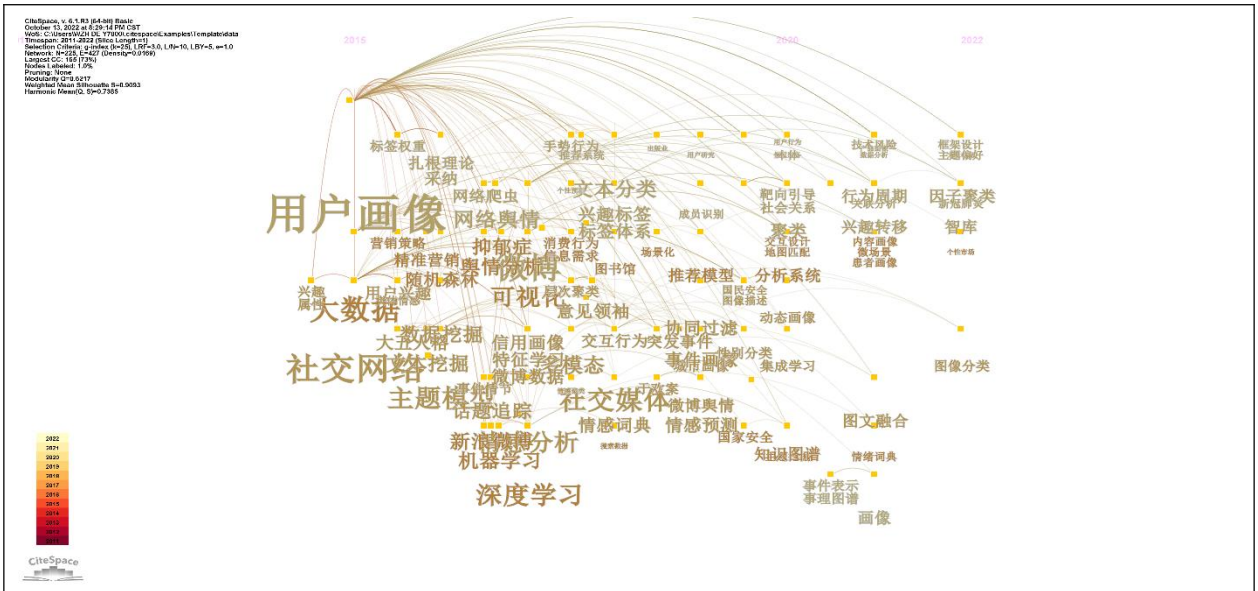

图 1

（二）国外研究动态

国外所用的 PERSONA 是 Allen Cooper 提出来的一种通过调研和问卷获得的典型用户模型，用于产品需求挖掘与交互设计的方法。Persona 是虚构出的一个用户用来代表一个用户群。一个 persona 可以比任何一个真实的个体都更有代表性。一个代表典型用户的 persona 的资料有性别、年纪、收入、地域、情感、所有浏览过的 URL、以及这些 URL 包含的内容、关键词等等。一个产品通常会设计 3～6 个用户模型代表所有的用户群体。而另一个单词，叫 Profile，是利用已经获得的数据，用来勾勒用户需求、用户偏好的数据分析方法。

这两个词，都可以翻译为用户画像，但第一种，用于产品用研与交互设计，第二种，用于运营与数据分析。与国内相同的是，国内外都将其用于社会媒体以及用户的个性化推荐上，可以发现，在这个互联网高速发达的时代，不管是国内还是国外，都需要这样一个画像系统来监控网络上的风险。

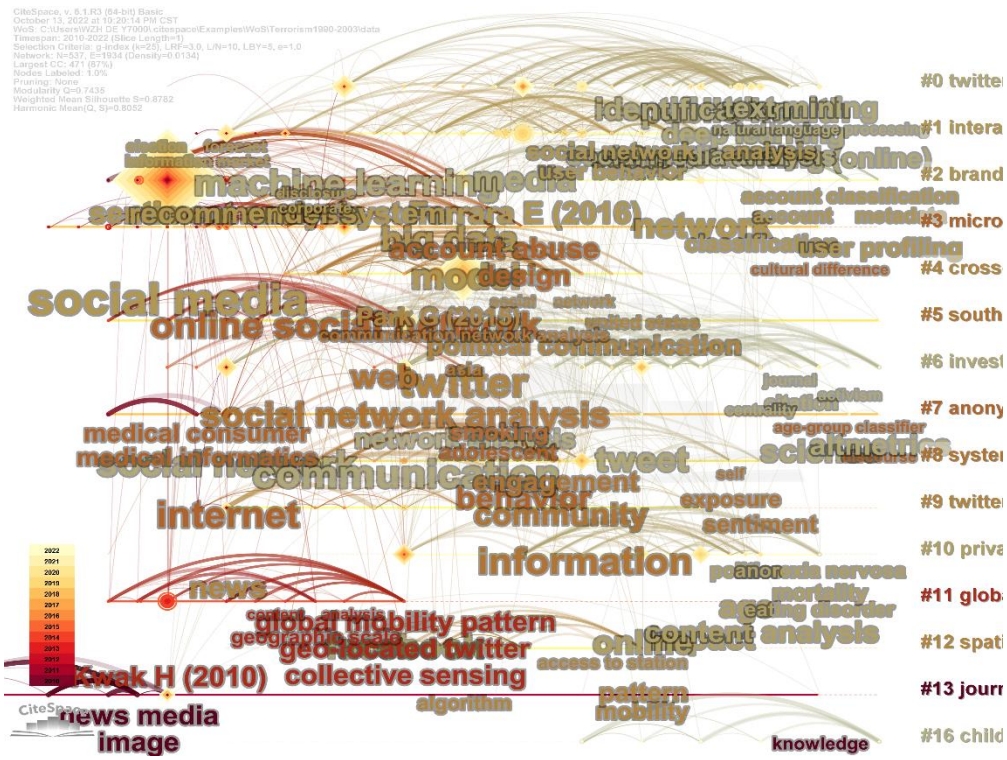

图 2

三、课题研究的目的及意义

大数据时代引发了社会媒体的变革，用户画像的引入不仅仅可以为实现精准服务提供思路和方向，也能够为公安舆情监控提供线索。老话常说，“水能载舟，亦能覆舟”，大数据即为海洋，覆舟之人，没于水下，我们通过精准描绘某个主题的画像，在众多主题中寻找敏感高危的话题，将符合预估值的画像拟合预设的画像，及时预警，将敏感主题及时清理；高危主题词中再进行下一步的拟合人物的高危预设画像（预描绘上千幅已有犯罪经历的人的人物画像，并通过机器学习，同互联网微博不断更新的数据进行对比拟合，发现同样犯罪的人以及将要犯同样罪的人），将其的犯罪预备动机及时捕捉，及时预警。基于此，本课题采用构建软件系统的方式，对我国当前互联网舆情治理、犯罪预测、公安侦查具有重要的理论意义和现实意义。

四、课题研究的主要内容

（一）标签化文本分类

互联网上的大量个人数据，可以轻松得知一个人的各式各样的独特标签，它可以是一个人的购物表象，通过类比推理，他也可以是一个人的犯罪表象。例如，通过聚类分析总结同一类犯罪人的特点，当他们的标签都趋向于同一方向时，利用大数据同步监测到新的一名用户加入到此画像群体中，此用户作为研究对象，判断其是否为新的犯罪人或者是以及在逃的犯罪人。

具体内容是为文本预设标签，将待处理的文本准确分类到预设的标签中，情感分析、对话分类以及主题分类。利用朴素贝叶斯，K 近邻算法以及支持向量机。

（二）搭建深度学习模型

对文本进行数据清洗，对其进行特征表示，之后进行浅层学习模型和深度学习模型搭建，选取特定的评价标准准确度输出标签。

对于每时每刻都在变化的主题画像，横向比对分析，不断定制事件标签，将标签的增长速度也能够衡量一个事件的属性，例如刻画已发生舆情事件的增长变化，去拟合新事件的发展变化，将画像的生长变化作为一个数据对比普通时间的生长变化。

（三）基于知识蒸馏的联合模型

传统主题模型建立微博画像存在主题一致性较低，缺乏解释性的缺点，通过使用知识蒸馏的方法将预训练模型学习到的知识迁移到学生模型中，学生模型具有较小的参数及较短的训练时间，以至于进行实时监控微博信息并及时训练生成实时画像。

（四）主题分类

通过构建用户标签画像，能够个性化推荐每个用户想要的商品，同样的道理，同一类犯罪嫌疑人往往会存在相同的共性，在此想法上，通过已有嫌疑人的特征画像去拟合全体画像，或许能够发现犯罪预备人或者是犯罪嫌疑人，在主题层次上，我们同样能够将已有的舆情事件数据分主题进行画像，将多个共性参数形成画像的多个标签，由于事件主题从无到有到产生舆情是一个过程，对此过程进行研究分析画像的演变过程，并与新事件主题进行拟合，以此来发现众多新事件中的可能会发生舆情的事件

（五）公式超参数取值

通过大量的画像对比性分析，选出最优的参数，提高对于舆情事件的判断精

准度，实现舆情事件的精准预测，捕捉其演变过程，对演变过程的参数变化进行分级取值，建立模型。传统主题模型建立微博画像存在主题一致性较低，缺乏解释性的缺点，通过使用知识蒸馏的方法将预训练模型学习到的知识迁移到学生模型中，学生模型具有较小的参数及较短的训练时间，以至于进行实时监控微博信息并及时训练生成实时画像。

实时预警需要这样一款量身定制的新型模型，设计并采用这样一款模型让实时系统更为即时可用。

（六）实验环境模拟

通过爬取大量的主题进行标签化聚类分析、细分、建立模型、发现共同模式。分析研究引起舆论热潮的大量主题，将此通过机器学习、神经网络训练为可实战化利用的模型，通过对比即可及时预测事件的走势发展，防止舆情事件的发生，同时也可以建立一些虚拟的主题场景，验证我们的想法是否成立，成立即可逐步应用于警务实战化平台，预测监控以及防止舆情事件的发生扩大。

（七）实现价值

搭建全平台一体化分布式系统，管理员及各级用户都分有不同的权限，分配不同主题的画像查看与舆情预警。将所有舆情事件和安全事件的画像参数存储在数据库中，每十分钟更新数据库，并及时提取数据。利用深度学习和神经网络不断设计优化模型，不断适应新的事态发展趋势。将全部的舆情事件分主题的绘制画像，对参数优化，以及没有舆情的事件进行画像，实时对比同主题的新型事件进行超参预警。

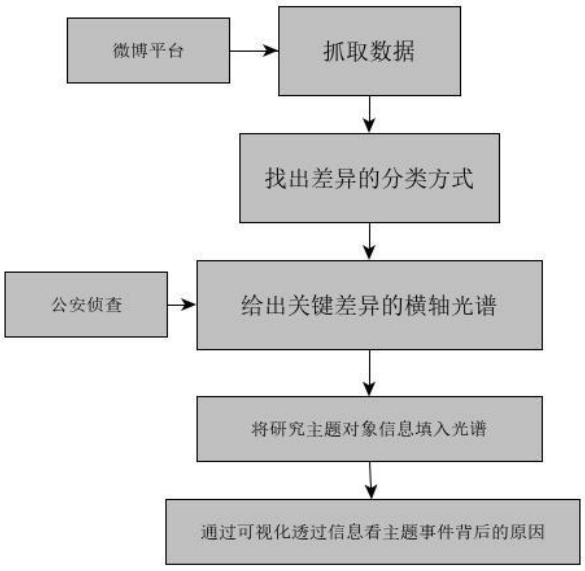

图 4

五、成果预测

构建面向微博主题的画像系统，集成实时爬取微博数据，数据清洗，利用有关文本特征表示的方法，例如 One-Hot 编码、词袋表示、TF-IDF 表示、Word2Vec 表示，对数据处理，对处理得到的分层数据进行知识蒸馏，对抗扰动训练，标签平滑性预测，搭建不同的主题模型，训练模型整体框架，最后分析得到微博主题兴趣画像的构建，先进行粗粒度主题画像建模，对粗粒度关键词提取，模糊聚类，及时计算对比某一新事件对于舆情事件的画像的拟合度，寻找最佳拟合的精确度，做到尽可能准确的预测舆情事件，当与舆情事件拟合度达到 x%时，归为舆情风险事件，并在主页实现大屏可视化数据，及时预警此类事件。

六、课题研究的条件及措施（宋体小四号，标准字符间距，行距 22 磅，段前 0 磅，段后 0 磅）

（一）课题研究的条件

1.通过在校期间的学习与熟悉了解本专业相关的理论知识，并通过毕业实习搜集资料；

2.学校内互联网和学校图书馆等信息资源提供的丰富资料以及调研；

3.学校丰富的专业教师队伍和实战部门教官提供的广泛的信息咨询和理论指导。

（二）课题研究的措施

1.通过互联网和学院图书馆查找相关书籍和广泛收集检索相关资料；

2. 与指导老师及专业老师交流探讨，修改、确定论文的提纲；

3. 利用毕业实习机会，亲身体验一线工作，获取直观认识；

4.返校后，按论文撰写计划和中期检查专家指导意见，修改完善论文。

七、课题进度计划

2022 年 10 月 15 日—2022 年 10 月 17 日：开题。

2022 年 10 月 22 日—2023 年 3 月 15 日：软件系统的编写。

2023 年 3 月 16 日—2021 年 3 月 29 日：中期检查。

2023 年 3 月 30 日—2021 年 4 月 19 日：指导教师、评阅人评定成绩。

2023 年 4 月 20 日—2023 年 4 月 30 日：学术不端行为检测。

2023 年 5 月 8 日—2023 年 5 月 22 日：毕业设计答辩，成绩评定。

八、参考文献

[1]彭媛媛,张海霞,连一峰,黄克振,刘倩. 一种基于多维度的网络空间人物画像方法[P]. 北京市：CN115114498A,2022-09-27.

[2]李继玲,李宝林,严宋如.疫情背景下快递物流服务的用户行为画像及主题挖掘研究[J].华东师范大学学报(自然科学版),2022(05):100-114.

[3]安璐,周雯静,李纲,余传明. 基于事件态势与用户认知的危机信息推送方法及系统[P]. 湖北省：CN114969560A,2022-08-30.

[4]刘浒,夏志杰,张晨芳.面向企业印象管理的社交媒体用户情绪演变及画像研究——以微博平台拼多多企业为例[J].情报工程,2022,8(03):112-125.

[5]商瀑,李亚可,郭楼禄.基于用户画像技术的大数据侦查：一个框架的分析与设计[J].中国人民公安大学学报(社会科学版),2022,38(03):32-43.

[6]秦权. 微博环境下新冠肺炎疫情事件对网民情绪的影响分析研究[D].中北大学,2022.DOI:10.27470/d.cnki.ghbgc.2022.001223.

[7]李尹舒. 基于用户画像的大米电商精准营销策略研究[D].吉林大学,2022.DOI:10.27162/d.cnki.gjlin.2022.002038.

[8]孟芷薇. 基于微博平台的智库用户画像分析研究[D].新疆师范大学,2022.DOI:10.27432/d.cnki.gxsfu.2022.000335.

[9]胡华实. 建构，互动，遵从：城乡二元视角下青年对新型媒体的差异化使用及归因[D].华东师范大学,2022.DOI:10.27149/d.cnki.ghdsu.2022.003744.

[10]马焓宸. 大数据抓取在换装玩偶娃衣个性消费市场中的研究与应用[D].西南科技大学,2022.DOI:10.27415/d.cnki.gxngc.2022.000081.

[11]潘睿. 基于画像的公共图书馆抖音号运营策略研究[D].华东师范大学,2022.DOI:10.27149/d.cnki.ghdsu.2022.001205.

[12]黄传能. FM传媒用户节项目运营及营销策略研究[D].电子科技大学,2022.DOI:10.27005/d.cnki.gdzku.2022.001653.

[13]张虹.企业招聘，打好数据牌[J].人力资源,2022(04):47-49.

[14]王志刚,邱长波.基于主题的政务微博评论用户画像研究[J].情报杂志,2022,41(03):159-165.

[15]叶光辉,郭诚,徐彤,王灿灿.城市画像视角下的政务社交媒体资源保存研究[J].情报科学,2022,40(02):11-17+27.DOI:10.13833/j.issn.1007-7634.2022.02.002.

[16]HerdaGdelen, A., Zuo, W., Gard-Murray, A., & Bar-Yam, Y, "An exploration of social identity: The geography and politics of news-sharing communities in twitter", Complexity, Vol. 19, no. (2), pp. 10-20, 2013
[17]Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A., "Sentiment strength detection in short informal text". Journal of the American Society for Information Science and Technology, vol.61, no. 12, pp.2544-2558, 2010.
[18]H.U. Khan, S.M. Saqlain, M.Shoaib, M.Sher, "Ontology based semantic search in Holy Quran", International Journal of Future Computer and Communication, Vol. 2, no 2, pages 1-6, 2013.
[19]H.U.Khan, T.A.Malik, "Finding resources from middle of RDF graph and at Sub-query level in suffix array based RDF indexing using
RDQL queries", International Journal of Computer Theory and Engineering, Vol.4, no.3, pages 369-375, 2012.
[20]H.U. Khan and A. Daud, T.A.Malik, "MIIB: A metric to identify top influential bloggers in a community", Plos One, vol. 10, no. 9, 2015.
[21]U. Ishfaq, H.U. Khan, K. Iqbal, "Modeling to find the top bloggers using sentiment features", in Proc. International Conference on Computing, Electronic and Electrical Engineering, pp. 227-233, 2016.
[22]H.U.Khan, A.Daud, "Using machine learning techniques for subjectivity analysis based on lexical and non-lexical features".
International Arab Journal of Information Technology, Vol. 14, no. 4. 2017.

<table>
<tr><td colspan="2">指导教师意见

签名：
年　　月　　日</td></tr>
<tr><td>开题审核小组意见

组　长（签名）：
年　　月　　日</td><td>二级学院意见

二级学院（签章）：
年　　月　　日</td></tr>
</table>

注：1、可以根据内容适当调整表格大小；

2、开题报告一式四份，一份随学生毕业论文(设计)归档，学生所在二级学院、教研室、指导教师各留存一份。