学号： 153549

东南大学

工程硕士学位论文开题报告

院（系、所） 软件学院

工 程 领 域 软件工程

研究生姓名 陶涛

校 内 导 师 陶军

校 外 导 师 彭艳兵

开题报告日期 2016.11

入学年月 2015.9

　制

一、学位论文开题报告

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 论文题目 | | 基于半监督学习的微博水军识别的研究 | | | | | | | | | | | |
| 研究方向 | | 软件工程 | | | | | | | | | | | |
| 题目来源 | | 国家 | 部委 | 省 | 市 | | 厂、矿 | | 自选 | 有无合同 | | 经费数 | 备注 |
|  |  |  |  | | √ | |  |  | |  |  |
| 题目类型 | | 工程项目规划管理 | 工程设计 | 工程技术 | 技术攻关改造推广 | | 应用基础性研究 | | 其它 | | |  | |
|  |  | √ |  | |  | |
| 1. **课题来源、选题依据、课题研究目的、工程应用价值**    1. **课题来源**   本课题是本人参与实习单位的项目舆情分析和预测，通过对舆情平台微博数据采集与分析，针对目前微博传播中水军泛滥的情况下提出对舆情平台中的微博用户进行水军识别的工作。   * 1. **选题依据**   当前，以Twitter、Facebook、微信、百度贴吧、人人网、新浪微博等为代表的在线社交网络在新闻传播、网民互助、品牌营销、知识信息传播等方面表现出积极作用。基于社交网络的信息传播也越来越深入和广泛。研究表明，社交网络中的信息传播与传统媒体中的信息传播相比，呈现出规模大、实时、快速等特点。其对公民日常生活、国家经济和公共安全的影响越来越深入。然而近年来，社交网络上出现谣言信息盛行、欺诈活动猖獗等负面因素，特别是有组织的网络水军的出现，更加放大了这些负面因素，从而动摇了社交网络的安全基础，最终会影响社交网络的发展前景。微博凭借传播速度快、用户多、便捷等特点风靡世界，已成为最具影响力的社会媒体之一．随着微博平台用户量、数据量剧增，传播效应不断扩大，微博已成为“网络水军”的主战场。2013年4月“雅安地震”，微博一方面成为最有力的信息传播媒体，各种“大V”、政务微博、平民账号等充分利用微博的信息扩散能力，帮助救灾。但另一方面，也有不法分子利用微博传播谣言，欺骗公众，造成社会的不稳定和民众恐慌，带来极坏的后果。如何快速有效的识别出社交网络中的网络水军，遏制垃圾信息的传播，对于维护公民生活的正常秩序和国家公共安全具有重要意义。 | | | | | | | | | | | | | |
| * 1. **课题研究目的**   本课题是通过挖掘社交网络中的用户行为、关系特征和文本内容，进行综合分析，发现潜藏在复杂网络环境中的水军行为特征，由于社交网数据非常庞大而标记数据非常少，故采用半监督学习方法，从而实现网络水军识别。   * 1. **工程应用价值**   本课题选自实习公司的实际项目。目前在公司的舆情分析和预测项目中发现存在大量水军，本课题项目的目的就是研究并实现对海量微博数据里面的水军识别，用于后期的舆情监测、引导和预警等功能。  传统的水军识别主要采用统计机器学习的算法，对账户进行分类，但无法有效地解决目前公司的数据非常庞大而标记数据非常少的问题。为了有效解决不平衡数据分类的问题，本报告提出基于半监督学习的方法应用于微博水军识别中，更加准确的识别出社交网络中的水军用户，有助于解决网络水军扰乱互联网秩、危害社会安全的问题。 | | | | | | | | | | | | | |
| 1. **中外文献与资料阅读清单**（要求30篇以上） | | | | | | | | | | | | | |
| 篇数 | 34 | | | | | 其中外文文献与资料篇数 | | | | | 19 | | |
| 序号 | 题目（及作者） | | | | | | | 出处（书籍及出版社、期刊名及期刊卷期号等） | | | | | |
| 1 | 网络水军识别研究（莫倩，杨珂） | | | | | | | 软件学报.2014,25(7):1505~1526. | | | | | |
| 2 | 面向微博的概率图水军识别模型（韩忠明，许峰敏，段大高） | | | | | | | 计算机研究与发展,2013,50(suppl):180-186. | | | | | |
| 3 | 基于结构和内容的社交网络水军团体识别（周国强，金礼仁） | | | | | | | 南京邮电大学硕士学位论文,2016. | | | | | |
| 4 | 面向网络舆情数据的异常行为识别（郝亚洲，郑庆华） | | | | | | | 计算机研究与发展，2016,53(3):611-620. | | | | | |
| 5 | 多元表征军事信息可信度研究（张天宇，林鸿飞） | | | | | | | 计算机工程与科学，2011,33(9) | | | | | |
| 6 | 基于关系图特征的微博水军发现方法（程晓涛，刘彩霞等） | | | | | | | 自动化学报，2015,41(9) | | | | | |
| 7 | 基于综合指数和熵值法的微博水军自动识别（袁旭萍，王仁武等） | | | | | | | 情报杂志，2014,33(7) | | | | | |
| 8 | 在线社交网络中异常帐号检测方法研究(张玉清,吕少卿,范丹趵) | | | | | | | 计算机学报，2015,38(10) | | | | | |
| 9 | 微博用户的时间行为研究（张杰斌，秦红） | | | | | | | 计算机科学，2016,43(6) | | | | | |
| 10 | 基于交互行为的在线社会网络水军检测方法（陈侃，陈亮，朱培栋，熊岳山） | | | | | | | 通信学报,2015,36(7). | | | | | |
| 11 | 马尔可夫逻辑网络研究（徐从富，郝春亮，苏保君，楼俊杰） | | | | | | | 软件学报，2011,22(8). | | | | | |
| 12 | 基于转发关系的微博话题演化算法（徐伟,赵斌,吉根林） | | | | | | | 计算机科学，2016,43(2). | | | | | |
| 13 | 微博消息传播中意见领袖影响力建模研究（王晨旭,管晓宏,秦涛,周亚东） | | | | | | | 软件学报，2015,26(6) | | | | | |
| 14 | 一种基于逻辑回归算法的水军识别方法（张良，朱湘，李爱平，王志华，鲁鹏） | | | | | | | 信息安全与技术，2015年4月. | | | | | |
| 15 | 一种面向语义重叠社区发现的 Link-Block 算法（辛宇,杨静,谢志强） | | | | | | | 软件学报，2016,27(2). | | | | | |
| 16 | Song J, Lee S, Kim J. Spam filtering in Twitter using sender-receiver relationship. | | | | | | | Heidelberg: Springer-Verlag, 2011. 301−317. | | | | | |
| 17 | Murmann AJ. Enhancing spammer detection in online social networks with trust-based metrics [MS. Thesis]. | | | | | | | San Jose: San Jose State University, 2009. | | | | | |
| 18 | Moh TS, Murmann AJ. Can you judge a man by his friends? Enhancing spammer detection on the Twitter microblogging platform using friends and followers. | | | | | | | Heidelberg: Springer-Verlag, 2010. 210−220. | | | | | |
| 19 | Krestel R, Chen L. Using co-occurrence of tags and resources to identify spammers. | | | | | | | Brookline: Microtome Publishing, 2008. 38−46. | | | | | |
| 20 | Sriram B, Fuhry D, Demir E, Ferhatosmanoglu H, Demirbas M. Short text classification in Twitter to improve information filtering. | | | | | | | New York: ACM Press, 2010. 841−842. | | | | | |
| 21 | Zhao YY, Qin B, Liu T. Sentiment analysis. | | | | | | | Journal of Software, 2010,21(8):1834−1848 | | | | | |
| 22 | Liu B. Sentiment analysis and subjectivity. | | | | | | | Boca Raton: CRC Press, 2010. 627−666. | | | | | |
| 23 | Benevenuto F, Rodrigues T, Almeida V, Almeida J, Zhang C, Ross K. Identifying video spammers in online social networks. | | | | | | | New York: ACM Press, 2008. 45−52. | | | | | |
| 24 | Parameswaran M, Rui H, Sayin S. A game theoretic model and empirical analysis of spammer strategies. | | | | | | | In: Proc. of the 7th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conf. (CEAS 2010), Vol.7. 2010. 1−7. | | | | | |
| 25 | Yang C, Harkreader RC, Gu G. Die free or live hard? Empirical evaluation and new design for fighting evolving Twitter spammers. | | | | | | | Heidelberg: Springer-Verlag, 2011. 318−337. | | | | | |
| 26 | Gargari SM, Oguducu SG. A novel framework for spammer detection in social bookmarking systems. | | | | | | | Washington: IEEE Computer Society, 2012.827−834. | | | | | |
| 27 | Zhu Y, Wang X, Zhong E, Liu NN, Li H, Yang Q. Discovering spammers in social networks. | | | | | | | Menlo Park: AAAI Press. 2012. 1−7. | | | | | |
| 28 | Hayati P, Chai K, Potdar V, Talevski A. HoneySpam 2.0: Profiling Web spambot behaviour. | | | | | | | Heidelberg: Springer-Verlag, 2009. 335−344. | | | | | |
| 29 | Hayati P, Chai K, Potdar V, Talevski A. Behaviour-Based Web spambot detection by utilising action time and action frequency. | | | | | | | Heidelberg: Springer-Verlag, 2010. 351−360. | | | | | |
| 30 | Hayati P, Potdar V, Chai K, Talevski A. Web spambot detection based on Web navigation behaviour. | | | | | | | Washington: IEEE Computer Society, 2010.797−803. | | | | | |
| 31 | Hayati P, Potdar V, Talevski A, Chai K. Characterisation of Web spambots using self organising maps. | | | | | | | In Journal of Computer Systems Science & Engineering, 2011,26(2):87−96. | | | | | |
| 32 | Hayati P, Potdar V, Talevski A, Smyth WF. Rule-Based on-the-fly Web spambot detection using action strings. | | | | | | | In: Proc. Of Collaboration Electronic Messaging Anti-Abuse and Spam Conf. (CEAS 2010). 2010. 13−14. | | | | | |
| 33 | Amleshwaram AA, Reddy N, Yadav S, Gu G, Yang C. CATS: Characterizing automation of Twitter spammers. | | | | | | | Washington: IEEE Computer Society, 2013. 1−10. | | | | | |
| 34 | Yang C, Harkreader R, Gu G. Empirical evaluation and new design for fighting evolving Twitter spammers. | | | | | | | IEEE Trans. on Information Forensics and Security,2013,(8):1280-1293. | | | | | |
| 1. **文献综述：国内外研究现状、发展动态**   网络水军检测和识别最初起源于邮件服务系统，当时的网络水军通过制造具有明显商业特点的垃圾邮件引起用户注意，引导用户点击商业广告站点。随着互联网的快速发展，社交网络中的网络水军识别已成为主流，陆续产生了大量的识别方法。按照网络水军识别方法采用特征的不同,将网络水军识别方法分为基于内容特征、基于用户特征、基于环境特征和基于综合特征的识别。  早期网络水军识别研究着重分析网络水军产生的内容,这是由于早期网络环境中,网络水军产生的内容具有显著的可识别特征,基于内容特征的网络水军识别研究涉及机器学习中的自然语言处理分支.该类含有观点的文本处理包括Sriram B等2010年提出的文本分类、Zhao YY等2010年提出的文本情感分析以及Liu B等2010年提出的文本倾向性分析等方面。  随着网络环境逐渐复杂多样和用户辨别力的增强,使得制造传播具有显著特征内容的传统网络水军造成的影响不断降低。为了不断制造网络影响、妨害商业利益,网络水军逐渐衍生出多样的欺骗策略。其行为趋向于正常用户的行为,其发布内容也不再具有显著特征。通过分析变化的网络水军行为,基于用户特征的识别研究能够很好地发现潜藏的网络水军。因此,当前网络水军识别研究转向基于用户特征的识别,以实现从源头遏制网络水军和垃圾信息泛滥的目的。社交网络中,用户行为能够极大地反映出用户信息。例如,网络水军突发性行为模式与正常用户有明显区别。Benevenuto等人首先提出了在线视频分享站点中网络水军行为的统计数据,即,从著名互联网视频网站Youtube收集的网络水军行为统计数据。他们采用人工标记方法建立训练数据集,即,数据集中的用户由人工标明是否为网络水军。之后分析已识别网络水军的行为,定义其特征,并利用weka中3种特征选择算法评价各个网络水军行为特征的分辨力。采用传统监督分类方法,判断未知用户是否为网络水军。该方法是社交网络领域水军识别研究的代表性方法,即,基于用户行为特征识别网络水军。之后的网络水军识别研究都以该方法为基础,增加特征或优化识别方法来提高社交网络中水军识别准确率。Parameswaran 等人则提出了一种理论建模方法对水军行为及其偏离进行建模,他们发现:网络水军行为策略不断发生变化,并提出可以长期监控网络水军行为,以此建立黑名单来降低网络水军危害。Yang 等人首次深入分析了Twitter中网络水军的隐藏策略,并提出基于邻居节点特征发现Twitter中网络水军的方法。Gargari等人在识别特征中加入网络资源层级特征实现网络水军识别,利用网络水军使用资源模式相似这一特征提高网络水军识别准确率。Zhu等人指出国内的社交网站——人人网中的网络水军识别缺陷,并提出复杂社交网络中的网络水军识别方法。Hayati[28]等人首次提出了Web2.0水军机器人的概念,通过追踪其行为特征来发现Web 2.0平台中的水军机器人。该实验发现:水军软件机器人利用搜索引擎来寻找目标站点,并通过不断注册新用户账号,在短时间内发布大量垃圾信息。这些新注册账号的用户名利用机器生成,并且其个人主页只有较低的点击率,即,没有很好的交互行为。此外,Hayati 等人在文献[29]提出了两种具体的行为特征:水军软件机器人的行为时间及其频率,实现对网络水军行为的建模。Hayati 等人在文献[30]中提出利用水军机器人进行水军行为时的特殊特征进行识别,并在文献[31]中综合早期研究提出利用自组织神经网络发现网络水军机器人的特征,并根据其不同的行为将水军机器人分为4类。2010年,Hayati等人在文献[32]中提出利用动态水军机器人的行为动作设计基于规则的动态识别方法,限制高度疑似的水军行为继续发生,从而在网络水军制造垃圾内容时制止其行为,实现从源头遏制网络水军泛滥的目的。  Song等人[16]认为:基于用户行为特征的网络水军识别方法具有一定的滞后性和易伪造性,即,用户进行过类似水军的行为后,该类方法才能进行识别,并且该类用户特征很容易被网络水军修改掩饰.但是整个网络关系具有一定的稳定性,其特征不容易被用户行为所影响.因此,使用该类特征对网络水军识别具有更好的效果.文献度量了用户间距离和用户联系紧密度等特征,并使用几种不同分类器进行分类学习.该文通过实验得出如下结论:在Twitter中,大多数网络水军发布的垃圾意见都只有少数接受者.Murmann[17]在Twitter中利用具有直接  交互关系的邻居节点探测用户间信任关系,得到新的关系特征集.利用此特征集对用户进行可疑度排序,可疑度最高者即为网络水军.Moh 等人[18]同样利用Twitter中的用户社会关系,如其朋友及粉丝特征,通过不同特征矩阵得到该用户的可信任度,以此判断该用户是否为网络水军.Krestel等人[19]利用网络水军可疑度会在社交网络中传播的特点,利用图模型上的传播发现标签分享站点中的网络水军.该方法实现了标签分享站点中网络水军、标签以及网络资源间关系结构的建模.通过给定一些种子节点的可疑度,依据种子节点向外传播可疑度的特点,从而计算整个图模型中所有节点的可疑度,发现可疑的用户节点。  隐蔽的网络水军对用户展现出趋向于正常用户的特征,但其异常行为使其在网络环境层级表现出不同于正常用户的特点。Las-Casas 等人提出基于水军产生时的网络特征识别,并使用巴西宽带ISP的数据记录作为实验数据集。他们提出:从网络流量的角度却很容易发现网络水军特征.因为网络水军为尽可能减少工作负担并达到效益最大化,会集中在一段时间内大量制造垃圾意见,因此这段时间内的网络负载会突然加大,流量也会集中在某些链路中.与传统网络水军识别方法相比,该方法准确率较高.但该方法需要应用ISP数据,无法普遍推广.基于环境特征的网络水军识别研究依据网络水军进行危害行为时产生的环境特征,该环境特征是无法被网络水军修改掩饰,因此其识别准确率较高.但基于环境特征的网络水军识别研究大多需要相应的实验数据集,因此其可推广性较其他网络水军识别研究方法要低。  基于用户行为和关系特征的网络水军识别方法捕捉用户在社交网络中的行为和关系,并不分析其发布的内容.传统网络水军识别方法即依据用户产生内容判断其是否为水军,因此结合用户行为、关系及其发布内容的综合特征网络水军识别方法很常见。Amleshwaram 等人[98]综合社交网络中各方面用户的特征,如行为、内容、用户间关系等,实现了社交领域网络水军的快速识别,且其识别所需时间和资源都大为降低.他们对未知水军进行了聚类分析,发现了Twitter上盛行的一些网络水军团体.他们发现:Twitter 上大多数网络水军只有很少的Tweet,其主要目标是传播垃圾信息,制造网络影响。  目前社交网络数据量十分庞大，而标记数据相对很少，所以社交网络的水军识别是一种类别极度不平衡的数据分类工作。考虑到社交网络标注数据少带来的精度降低的问题，基于社交网络图的排名算法可以避免传统监督分类中由于特征选取不好带来的精度低的问题，本报告提出基于半监督学习分类模型和基于图的排名算法相结合的微博水军识别框架。 | | | | | | | | | | | | | |
| 1. **研究基础**   目前，对于微博水军的识别主要通过提取各种特征维度，然后运用机器学习的方法进行分类，以识别正常用户和水军，机器学习的技术已经很成熟了，特征维度的提取前面很多学者也做了很多相关方面的研究。解决分类数据不平衡问题的半监督学习方法也有很多改进的方法。实习所在公司的舆情分析和预测项目中发现有大量网络水军的存在，对网络水军的特征和行为模式比较熟悉。平时的工作和学习中对上述技术都有所涉猎和应用。  在文本处理方面的技术积累  爬出和数据抽取技术  公司的研究小组的技术支持 | | | | | | | | | | | | | |
| 1. **主要内容与关键技术**   本报告依据目前公司内部舆情分析和预测来编写，旨在提炼出微博水军的各种特征和行为模式，快速有效地识别出海量微博数据里面的水军用户。  **5.1主要内容**   1. 对微博舆情平台里的海量微博数据进行采集和预处理，提取每个用户的基本特征、内容特征、网络图特征以及邻居特征等特征维度，然后给少量特征明显的用户打上标签。 2. 通过标记数据的特征维度算出该用户的可信度，画出由标记节点与周边节点（关注与被关注）的可信度的传播的社交关系网络图，计算出图中节点的PageRank值。 3. 使用标记数据训练基于半监督学习的分类器，使用训练好的分类器对标记节点的周边节点进行分类，当分类器的预测与社交关系网络图的排名结果一致时认为该分类有效，给该节点打上标记。 4. 迭代进行步骤（2）、（3）直至分类器的分类误差满足设定条件时结束。 5. 使用优化后的分类器进行数据分类。   **5.2关键技术**  (1)基于图的排名算法  PageRank是一个基于图中链接的排名算法，算法初始为数据集中的每个数据点赋一个初始值，代表每个节点的重要度。通过链接的方向进行传播，最终计算出整个互联网每个网页的重要度值。PageRank的基本理论是将从网页指向另一个网页认为是一个链接，表示对另一个网页的支持。完整的PageRank公式是：    P1,p2,...pn是被研究的页面，L(pj)是链出页面的数量，而N是所有页面的数量,q是阻尼系数，一般取0.85。  如果给每个节点一个随机PageRank值（非0），那么经过不断的重复计算，这些节点的PR值会趋向于正常和稳定。使用PageRank算法计算出每个节点的可信度，可以有效的避免人工选取随机值和种子节点的干扰，结果可靠性高。  (2)半监督学习方法  由于本课题是利用少量的已标记数据进行大量的未标记数据的分类，选取半监督学习算法来实现分类。半监督学习方法主要包括半监督聚类以及半监督分类两种形式。其中，半监督分类首先基于少量的已标注数据构建出描述刻画数据的分类模型，再通过大量的未标注数据辅助分类模型的更新优化，在此过程中使用分类模型对数据进行类标号的标注从而达到分类的目的。半监督学习的基本设置是给定一个来自某未知分布的有标记示例集L={(x1,y1),(x2,y2),...,(x|L|,y|L|)}以及一个未标记示例集U={x1,x2,...,x|U|},期望学得函数f:X→Y可以准确地对示例x 预测其标记y。这里xi,x∈X均为d维向量,yi∈Y为示例xi的标记,|L|和|U|分别为L和U的大小,即它们所包含的示例数。  (3)代价敏感学习方法  代价敏感的学习方法是机器学习领域中的一种新方法，它主要考虑在分类中，当不同的分类错误会导致不同的惩罚力度时如何训练分类器。例如在医疗中，“将病人误诊为健康人的代价”与“将健康人误诊为病人的代价”不同；在金融信用卡盗用检测中，“将盗用误认为正常使用的代价”与将“正常使用误认为盗用的代价”也不同。通常，不同的代价被表示成为一个N×N的矩阵Cost中，其中N 是类别的个数。Cost[i, j]表示将一个i 类的对象错分到j 类中的代价。  按照对问题的解决方法的不同，对代价敏感学习的算法研究可以分成三类。  第一类代价敏感的学习方法关注于如何直接构造一个代价敏感的学习模型，对不同的分类器模型研究者们提出了不同的解决办法，它们包括基于决策树的代价敏感减值方法、代价敏感的Boosting 算法Ada-Cost以及基于Perceptron 分类算法的代价敏感的学习方法。  第二类代价敏感的学习方法基于对分类结果的后处理，即按照传统的学习方法学习一个分类模型，然后对其分类结果按照贝叶斯风险理论对结果进行调整，以达到最小的损失。  第三类代价敏感的学习方法基于传统的学习模型，通过改变原始训练数据的分布来训练得到代价敏感的模型。  本文采用第二类学习方法，对其分类结果按照贝叶斯风险理论对结果进行调整，以达到最小的损失。  (4)协同训练算法  标准协同训练算法(Co．training)是一种很有代表性的半监督学习方法，由A．Blum和T．Mitehell首次在文献中提出，最初其用于解决网页的半监督分类问题。Co．training算法首先根据己标注数据在两个不同视图的数据集上分别训练出分类器，在接下来的迭代过程中，两个分类器“互教互学”，分别为对方挑选出有着高标注置信度的无类别数据进行类别标注，并把添加了类别信息的数据添加到另一个分类器的已标注数据集中用以该分类器进行优化学习，直到满足终止条件退出算法。但是，很多实际应用很难满足Co．training算法给出的上述假定条件。因此，针对不需要满足Co．training算法假定条件的协同训练算法，国内外学者做了大量的研究。第一类为两个分类器的协同训练算法，研究的主要思路是不同基本分类器的采用，第二类为多分类器的协同训练算法，主要思路是将其与集成学习(Ensemble Learning)算法或多视图学习相结合进而利用该算法的优势提高Co．training算法的性能。代表性的算法有周志华等人提出的Tri-training算法以及其改进算法Co. Forest。  本文采用改进的Co.Forest算法，即引入代价敏感学习方法的Co.Forest算法，有效解决标记数据少的问题。 | | | | | | | | | | | | | |
| **6.研究方向、实施方案及可行性分析**  **6.1研究方向**  本课题的研究方向是分析目前舆情平台中的微博海量数据，并根据公司的研发项目给出一种快速有效的微博水军识别方法。  **6.2实施方案**  整个研究过程分为数据采集、数据存储、数据预处理、数据分析四部分，各部分的工作是相辅相成的，前面工作的质量直接影响后面工作的结果，后面工作的结果也能从一定程度上反映出前面工作的好坏。整体实施流程图如下：    6.2.1课题数据来源  (1)基于新浪API的数据获取方法  新浪微博开放平台为第三方提供了可用于获取微博信息的API接口以及方便微博API调用的支持多种计算机语言的软件开发包。利用该平台，用户只需要建立自己的应用，就可以调用微博的API，从而获得微博数据。其具体流程如下：首先通过创建账号和应用获取开发者身份，得到专属的AppKey和AppSecret；然后开发者向授权地址发送授权请求，通过OAuth2.0认证返回RequestToken授权；授权成功后，开发者再根据授权获得的Request Token得到Access Token；最后调用接口，获得JOSN数据流或XML文件，解析后完成微博数据的采集。采用这种方式只需要开发者按照说明文档就可以根据预先定义的接口，返回结构化标准数据，开发代价低、易于实现。但是采用这种方式会受限于官方规定的获取频率与数据获取方式，无法高效、全面地获取数据。  (2)基于微博爬虫的数据获取方法  微博爬虫是结合模拟登录、网页爬虫和网页内容解析3种技术的微博信息采集方法。首先根据微博的特点，分析出新浪微博的登录协议、数据请求过程和每个请求的URL与数据的对应关系，这可以通过网络抓包工具来实现；接着在上一步分析的基础上，通过获取cookie并创建session，实现程序对网页的模拟登录过程；实现模拟登陆后，以登录用户的微博信息为种子集合，通过HTTP协议使用GET方法来完成数据的采集并对获取得到的数据进行解析。这种方法通过程序模拟正常微博用户在用浏览器等客户端浏览微博的过程，不依赖于微博官方提供的授权与API，可以满足用户对数据多样性、灵活性的需求，同时不受限于官方规定的获取频率与数据获取方式，能够高效、全面地获取数据。本课题采取爬虫的方法获取数据。  6.2.2 数据存储  由于爬取的数据是新浪微博平台的海量用户数据，数据量大，因此存储于实习公司的oracle数据库里面。  6.2.3 数据预处理  对存储于数据库里面的数据进行预处理，去燥，提取出每个用户的ID，昵称，关注数、粉丝数、原创微博文本、转发微博文本、关注的账号，粉丝的账号等特征，在特征提取的时候对一些特征特别明显的用户可以打上标记，作为分析的标记用户，便于后面进行分析工作。  6.2.4 数据分析   1. 从标记数据出发，给每个标记数据赋予一个初始的可信度，基于标记数据的关注和被关注关系画出社交关系网络图，使用PageRank算法算出每个节点的可行度。 2. 使用标记数据训练基于改进的Co Forest算法半监督学习的分类器，用训练好的分类器对标记数据节点的周边节点进行分类预测，如果分类预测结果与图模型的预测一致，则认为预测结果可靠，作为新的标记数据。 3. 从新的标记数据中无放回采样，迭代进行上述步骤，直至分类器不能被进一步优化为止。 4. 使用最优化的分类器给剩余的没有标记的数据节点进行分类标记。   数据分析流程图：    **6.3可行性分析**  对于上述6.2中提出的方案，从数据采集到数据分析，每一步都有相应的比较成熟的技术作支撑，从技术层面来说是一套可行的方案。半监督学习方法也一直被认为是可以有效的解决类别数据不平衡的问题，本实施方案首次提出基于改进的Co Forest算法的半监督学习的微博水军识别，理论上是可行的。  可行性方面写的不够。  数据如何提取  分类器的了解  分类器要不要改一下，显示你的研究能力？ | | | | | | | | | | | | | |

二、论文工作进度与安排

|  |  |  |  |
| --- | --- | --- | --- |
| 起讫日期 | | 工作内容和要求 | 备注 |
| 2016.7.8-2016.8.31 | | 系统了解实习公司业务，为开题方向做准备 |  |
| 2016.9.1-2016.10.31 | | 初步确定开题方向，查找相关资料和文献，提出问题并寻找相应解决方案。 |  |
| 2016.11.1-2016.11.30 | | 编写开题报告 |  |
| 2016.12.1-2016.12.15 | | 提交开题报告，根据导师已经修改开题报告，完成开题 |  |
| 2016.12.16-2016.4.10 | | 完成论文初稿，对实现的功能进行测试验证。 |  |
|  | |  |  |
| 学校指导教师对开题报告的综合意见 | 指导教师（签字）  年月日 | | |
| 校外导  教师对开题报告的综合意见 | 指导教师（签字）  年月日 | | |
| 开  题  报  告  审  议  情  况  记  录 | １、审议小组意见（审议小组成员3-5人）  ２、投票表决结果  审议小组出席人  通过人  不通过人  ３、审议小组组长（签名）\_\_\_\_\_\_\_  审议小组成员（签名）\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  \_\_\_\_\_\_\_\_\_\_\_\_\_\_  年月日 | | |
| 院（系、所）意见  负责人（签字或印章）  年月日 | | | |