**项目编号：**

1111

**大学生创新训练项目申请书**

**申报级别：国家级 省级□ 校级□**

**项目名称：基于深度学习的用户密码猜测及强度评估系统**

**项目负责人： 梁肖**

**手机：15619026193 Email： qingyuanxingsi@163.com**

**项目成员：**

**指导教师：**

**申请经费： 10000元**

**起止时间： 2017年4月-2018年4月**

**申报日期： 2017年3月20日**

**填 写 说 明**

1、本申请书所列各项内容均须实事求是，认真填写，表达明确严谨，简明扼要

2、本申请书为A4，左侧装订成册。可网上下载、自行复印或加页，但格式、内容、大小均须与原件一致。

3、主持人所在学院认真审核, 经初评和答辩，签署意见后，将申请书（一式两份）报送教务处实践教学管理中心。

* 1. 基本情况

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 项目  名称 | | **基于深度学习的用户密码猜测及强度评估系统** | | | | | | | |
| 所属  学科 | | 系统工程 | | | | | | | |
| 申请  金额 | | 10000 元 | | | | | | | |
| 项目实施时间 | | 起始时间： 2017年 4月 完成时间： 2018年 4月 | | | | | | | |
| 申  请  人  或  申  请  团  队 |  | 姓名 | 学号 | | 年级 | 所在院系/专业 | 联系电话 | | E-mail |
| 主持人 |  |  | |  |  |  | |  |
| 成员1 |  |  | |  |  |  | |  |
| 成员2 |  |  | |  |  |  | |  |
| 成员3 |  |  | |  |  |  | |  |
| 指导  教师 | 姓名 |  | 研究方向 | |  | | | | |
| 年龄 |  | 手机 | |  | | | 行政职务/专业技术职务 |  |
| 主持人曾经参与科研的情况 | | | |  | | | | | |
| 指导教师承担科研课题情况 | | | | 指导教师填写 | | | | | |
| 指导教师对本项目的支持情况 | | | | 指导教师填写 | | | | | |
| 项目简介  200字 | | | | 基于深度学习的用户密码猜测及强度评估系统使用深度学习方法，可根据用户个人信息猜测用户可能使用的密码，也能用于评估用户密码的强度，个人可使用此工具评估自己密码的强度，公司也可使用该工具作为在线注册的组件，提升注册用户账号的安全性。用户只需向系统提供自己的个人信息和密码，系统即可判断用户密码的强度，所提系统具有较高的准确性。 | | | | | |

* 1. 立项依据（可加页）

|  |
| --- |
| 1. 项目实施的目的、意义   基于深度学习的用户密码猜测及强度评估系统使用深度学习方法进行密码猜测和密码强度评估。用户只需向系统提供自己的个人信息和密码，系统即可判断用户密码的强度，所提系统具有较高的准确性。使用该系统可帮助用户选择强度更高的密码，提高用户账号的安全性。使用深度学习方法对密码文本进行建模且在建模中考虑了用户的个人信息，可提高用户密码猜测的效率和用户密码强度评估的准确性。   1. 研究内容   随着互联网时代的到来，互联网网站和服务越来越多，而这些网站和服务广泛采用的用户认证方式就是文本密码，在可预见的未来，文本密码依然会占据不可替代的主导地位。尽管人们已经提出了多种不同的用户认证机制，然而没有一种能在不给用户引入额外负担的前提下保留文本密码的所有优势。此外，更为不幸的是，文本密码长期以来一直被认为是用户认证中较为薄弱的一环。为了便于记忆，用户密码往往并不是随机字符串，用户密码中或多或少会包含用户的个人信息，包括用户名、姓名、出生日期、身份证号、手机号、邮箱等等。换言之，用户密码往往只是从全部密码空间的一小部分中选取的，使其较易受到暴力破解或字典攻击。  为了增强用户密码安全，大部分网站和服务都引入了强制的密码策略；同时，很多网站和服务也提供密码强度评估工具来评估用户密码的强度，帮助用户选择强度较高的密码。然而，这些密码强度评估工具往往仅仅依赖于简单的启发式规则，在实际使用中被证明并不可靠，甚至在一定程度上会对用户造成误导（如将弱密码标记为强密码）。为了更为准确和可靠地评估用户密码强度，我们需要对用户如何生成密码有更深入的了解。  到目前为止，学术界已经从多个角度揭示了密码构成结构。传统字典攻击方法指出用户在生成他们密码时更倾向于使用较为简单的单词。语言这一维度也很重要，用户在构造密码时倾向于使用他们的母语。此外，尽管有可能不是单词，密码大多数也是朗朗上口，易于记忆的。进一步的研究也指出用户在密码中可能使用键盘和日期字符串。更进一步地，除了以上密码模式外，也有部分工作对密码构成中的语义模式进行了分析，然而这些工作均需要对密码文本进行显式的语义分割，对密码文本进行语义分割有可能很难而且并不是必需的。  通过以上分析不难看出，现有工作中考虑密码语义信息的较少，而且考虑语义信息的相关工作均依赖于对密码文本进行显式的语义分割或匹配。另一方面，近年来，深度学习领域取得了空前的发展，深度学习方法在自然语言处理、语音、图像等领域均取得了极好的结果。深度学习网络可用于对高维函数进行近似，近年来的研究也指出深度神经网络在序列生成任务上的有效性，此外深度学习网络也比较适合用于进行端到端学习，这使得深度神经网络成为解决上述问题的一个可行的选择。基于此，同时考虑到用户个人信息在密码文本中的广泛使用，本文所述系统主要从用户个人信息角度出发，对密码构成中的语义信息进行研究，提出一种基于深度学习的端到端的用户密码文本概率模型，以更为准确地对密码文本进行建模。同时，基于上述模型，本文提出了一种基于深度学习的用户密码猜测方法,该方法可用于更为高效地对密码文本进行猜测；此外本文也实现了一种基于深度学习的密码强度评估工具，该工具可更为准确地对密码文本强度进行评估。  下面是对主要工作的具体描述：  1）基于深度学习的密码文本建模  提出了基于深度学习的密码文本概率模型------bpe2char模型/char2char模型，所提模型着眼于用户个人信息，不依赖对密码文本进行显式的匹配或分割，可以进行端到端的学习。针对密码建模中存在的难点，本文对模型的编码器和解码器进行了重新设计，具体而言，解码器端，本文使用双尺度循环神经网络捕捉密码文本中可能存在的不同粒度的信息；编码器端，当编码基本单元为字符时，本文提出使用一系列的卷积、池化及高速公路网络层减少输入序列的长度，将其控制在合理的范围之内，接着将缩减后得到的序列作为双向循环神经网络的输入，以捕捉输入序列中的依赖关系。  2）基于深度学习的用户密码猜测和强度评估方法  基于上述模型，提出了基于深度学习的密码猜测方法和密码强度评估方法。基于深度学习的密码猜测方法使用改进的集束搜索算法生成候选密码，可极大地降低密码生成的时间复杂度；基于深度学习的密码强度评估方法使用基于蒙特卡洛的密码猜测次数估计方法对待测密码猜测次数进行估计，并将密码猜测次数作为密码强度的评估指标。此外，本文使用真实的用户密码文本样本对所提猜测及强度评估方法进行了测试，实验结果表明所提密码猜测方法猜测效率要远优于现有方法，此外，通过与若干主流学术界/工业界密码强度评估工具进行对比，所提密码强度评估方法的准确性要高于其它方法。  3）基于深度学习的用户密码猜测及强度评估系统的设计与实现  设计并实现了密码猜测及强度评估系统PGES，该系统实现了密码猜测及强度评估功能，系统也提供了密码强度评估接口，可供用户利用学术界及工业界的多种强度评估工具衡量用户密码的强度，帮助用户选择强度更高的密码。   1. 国、内外研究现状和发展动态 2. 用户密码猜测   当前主流的用户密码猜测方法主要包括基于规则的方法和基于概率模型的方法两类，基于规则的方法工业界使用较多，主要基于一些启发式规则对用户密码特征进行建模；而学术界主要基于统计概率模型对密码文本进行建模，并基于一定策略按概率从高到低的顺序生成候选密码，然后将（哈希后的）候选密码与用户真实密码进行比对，判断是否匹配即可。以下对相关方法一一进行介绍：  **基于规则的方法**；目前有很多软件工具可帮助用户生成密码猜测结果，这些工具通常基于词表（密码串或者单词），使用预定义的替换或转换规则生成最终候选密码串，这些工具通过预定义的替换或转换规则对用户常用密码构造模式进行建模。例如，一种替换规则可能是“在开始位置添加一个数字且将所有‘a’替换为‘@’”。在这类工具中，用户使用较多的两种是Hashcat和John the Ripper(JtR)。尽管这些方法并不直接基于统计学方法对密码构成进行建模，但由于它们能较快地产生较为准确的密码猜测结果，在实际应用中使用也较为广泛。  **马尔科夫模型（Markov Models）;** 马尔科夫模型于2005年被NARAYANAN等人首次提出，后续也有更多的学者对该模型进行了探讨，其核心思想是基于之前的字符或者上下文字符预测下一个字符的概率。继NARAYANAN提出马尔科夫模型之后，Castellucci对该模型进行了改进，提出可使用基于n-gram的马尔科夫模型使用之前的n-1个字符去预测下一个字符的概率。然而，以上工作均不能按概率从高到低的顺序枚举密码，为了解决这一问题，M. Durmuth提出了Ordered Markov Enumerator(OMEN)模型，该模型可按照概率从高到低的顺序生成候选密码。更进一步地，作者对OMEN模型进行扩展，引入OMEN+模型，OMEN+基于Markov模型在密码猜测过程中引入了个人信息，这也是首个在密码建模中考虑个人信息的工作，然而其实验数据规模很小，性能提升有效，而且其论文也没有正式发表。此外，Ma J也对马尔科夫模型的特性进行了研究，他们指出使用更长的上下文字符可提高猜测的准确率，但是过拟合的可能性也更高，而使用平滑和补偿方法可在一定程度上降低过拟合的可能性。他们也发现对于英文密码来说使用平滑的6阶模型在一般情况下可取得最优的结果。  **上下文无关文法（Probabilistic Context-Free Grammars,PCFG）**；上下文无关文法方法由Weir等人首先提出，其基本思想是密码遵循一定的密码模板结构，使用终结符将模板中的非终结符一一进行替换就能生成最终的密码。例如，“6个字母+2个数字”就可能是一种模板结构。特定密码的概率就是模板结构概率和终结符概率的乘积。  然而，Weir等人的工作中并没有考虑字符子串的概率，为了解决这个问题，Veras等人提出了一种改进的基于PCFG的密码猜测算法，该算法可捕捉密码串中的结构、语法及语义信息，其实验结果指出，使用自然语言词典去替换终结符也能提高猜测效率，对于长密码而言这一效果更为显著。此外，WierKELLEY等人研究发现训练模板结构和终结符时采用不同的训练数据源可提升猜测结果。KOMANDURI等人指出平滑技术可提升密码猜测的效果。   1. 密码强度评估   目前主流的密码强度评估方法主要包括两种，一种基于单纯的统计学方法，如使用香农熵或其它更为高级的统计学方法。然而，由于单纯的统计学方法往往需要相对较大的样本数，本节不再详细展开说明；第二种方法则是通过模拟密码猜测来评估用户密码强度，这些方法往往基于特定概率模型对密码文本进行建模，并提出特定的评价指标来评估密码的强度。具体而言，PCFG基于PCFG模型计算特定密码被破解的概率来评估其强度；BFM和Weir使用在特定猜测次数下被破解密码的比例来评估密码强度；PTG则通过考察超过特定密码概率阈值的密码的比例来评估密码的强度。此外，更一般地，DELL’AMICO等人提出使用蒙特卡洛方法对特定密码的猜测次数（guess number）进行估计并以猜测次数作为密码强度的衡量指标，该方法可较为准确和快速地在秒级别的时间内返回特定密码的估计猜测次数。   1. 神经网络   神经网络，是一类机器学习方法，可以用于对高阶函数进行近似。神经网络方法可模拟人脑神经元，一般用于分类或序列生成任务。借鉴之前使用神经网络进行序列生成的工作,本文工作也使用神经网络对密码文本进行建模。其基本思路是，给定上下文序列如***passwor***,神经网络可预测下一个出现的字符是***d***的可能性较大。  尽管密码生成和文本生成存在很强的相似性，然而很少有工作使用神经网络对密码文本进行建模。十年以前，CIARAMELLA 等人曾尝试使用神经网络将密码文本分成比较粗粒度的类别（强或弱），但是该工作并没有对密码生成的顺序进行建模，而且也没有考虑密码猜测的很多其他层面的因素。近年来，真正意义上考虑使用神经网络对密码文本进行建模的工作是William Melicher 等人的工作, 然而作者提出的模型中并没有考虑个人信息的因素。  另一方面，从理论上来说，神经网络与其他传统方法相比，具有较大的优势。与马尔科夫模型和概率上下文无关文法模型相比，神经网络可以生成之前并未出现过的密码序列，这也是本文认为神经网络适合用于对密码文本进行建模的重要原因。之前的基于概率的密码猜测方法（如马尔科夫模型）由于内存消耗较大，无法直接用于在客户端对密码文本强度进行评估。然而，与马尔科夫模型相关，神经网络可以使用更少的空间对密码文本进行建模。   1. 创新点与项目特色 2. 提出基于深度学习的密码文本建模概率模型； 3. 基于以上概率模型，提出基于深度学习的用户密码猜测方法，可提高用户密码猜测的效率； 4. 同样基于上述模型，提出基于深度学习的用户密码强度评估方法，可提高用户密码强度评估的准确性。 5. 技术路线   系统整体结构如下图所示，主要包括以下几个部分：数据获取子系统、密码建模子系统、用户密码猜测子系统、密码强度评估子系统以及可视化子系统。其中，数据获取子系统主要对多个网站的泄露密码数据集进行关联、分析、整理，得到用户个人信息/密码对，为密码建模子系统准备数据条件；密码建模子系统实现了包括bpe2char、char2char、PCFG、马尔科夫模型在内的四个概率模型对密码文本进行建模；基于密码建模子系统得到的密码文本概率模型，本文实现了若干密码猜测工具，其中，***Neural Network Enumerator***使用改进的集束搜索算法生成候选密码，将最后得到的密码串按概率从高到低的顺序进行排序即可得到最终的猜测结果。***PCFG Enumerator***和***Markov Enumerator***可直接依赖于一定的策略按概率从高到低的顺序生成密码；此外，同样基于以上概率模型，本文实现了若干密码强度评估工具，***Neural Network Meter***基于bpe2char/char2char模型对密码强度进行评估，***PCFG Meter***和***Markov Meter***则分别依赖于PCFG模型和马尔科夫模型，同时，本文也对工业界比较典型的密码评估工具进行了实现，包括***zxcvbn***及12306等。可视化子系统则是通过Web系统的方式提供若干密码强度评估工具供用户使用。     1. 拟解决的问题及预期成果 2. 拟解决问题   基于深度学习的密码文本建模方法  基于深度学习的用户密码猜测方法  基于深度学习的用户密码强度评估方法  基于深度学习的用户密码猜测及强度评估系统的设计与实现   1. 预期成果    1. 可以较快地根据用户个人信息生成候选密码    2. 用户可以通过web系统提交待测密码，系统可对用户密码强度进行评估，且准确性较高。 2. 项目研究进度安排   2017年4月，对数据进行整理，构建密码数据集；  2017年5月，设计并实现基于深度学习的用户密码文本建模方法 ；  2017年6月，设计并实现基于深度学习的用户密码猜测方法；  2017年7月，设计并实现基于深度学习的用户密码强度评估方法；  2017年8月，设计并实现基于深度学习的用户密码猜测及强度评估系统  2017年9月，对系统进行测试与完善。   1. 项目研究与实施的基础条件（1、2）    1. 与本项目有关的研究积累和已取得的成绩（如校基金支持等）   目前，初步设计完成了基于深度学习的用户密码文本建模方法。   * 1. 已具备的条件，尚缺少的条件及解决方法   已具备条件   1. 对项目有较为深入的研究，并对目前现有的密码文本建模方法进行了调研； 2. 基本完成了项目的整体设计； 3. 训练所需数据已清洗整理完成。   缺少的条件  （1）进一步提高系统的运行效率。  注：如前期受“西安交通大学本科生科研训练和实践创新基金”支持需说明本次申请项目与前期工作的区别和关系。 |

* 1. 经费预算

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 开支科目 | 预算金费（元） | 主要用途 | 阶段下达经费计划（元） | |
| 前半阶段 | 后半阶段 |
| 预算经费总额 | 0 |  |  |  |
| 1.业务费 | 3000 |  |  |  |
| (1)计算、分析、测试费 | 0 |  |  |  |
| (2)资料印制费 | 500 | 打印复印等费用 |  |  |
| (3)会议、差旅费 | 1500 | 用于调研和交流 |  |  |
| (4)文献检索费 | 500 | 检索文献数据库等 |  |  |
| (5)论文出版费 | 500 | 成果发表费用 |  |  |
| 2.仪器设备费 | 4000 |  |  |  |
| 3. 试剂耗材费 | 2500 |  |  |  |
| 4. 图书购置费 | 500 | 购买图书资料 |  |  |
| 学校批准经费 |  |  |  |  |

**备注：因项目实施所需购置的器材（单价50元以上）归入仪器设备；因项目实施所需购置的器材（单价50元以下）归入试剂耗材，相关专业书籍（原则上不得超过总经费30%）。其他因项目实施所需的支出为业务费，包括调研差旅费、会议费（限1000元以内）、资料印制费、论文版面费等。**

* 1. 指导教师意见

|  |
| --- |
| 指导教师填写  **导师（签章）：**  **年 月 日** |

* 1. 学院意见

|  |
| --- |
| **教学副院长（签章）：**  **年 月 日** |

* 1. 学校专家组意见

|  |
| --- |
| **负责人（签章）：**  **年 月 日** |

* 1. 西安交通大学国家级大学生创新创业训练计划领导小组审批意见

|  |
| --- |
| **负责人（签章）：**  **年 月 日** |