文章编号：2095-2163(2019)03-0049-05 中图分类号：TP309 文献标志码：A

一种用户连续查询中隐私风险评估的方法

马永东[1]，王文涛[1]，王银款[2]
（1 东华大学 计算机科学与技术学院，上海 201620；2 上海航天控制技术研究所，上海 201109）

摘 要：用户在线查询服务中的隐私风险评估与隐私保护有着等同的重要性。关于用户查询隐私保护的研究引起了广泛关注，而对于用户查询隐私风险评估的研究较少。其中，连续查询作为查询一种重要表现，合理地对用户连续查询进行隐私风险评估，能够有效抵抗用户查询中隐私泄露。因此，本文基于隐马尔可夫模型（Hidden Markov Model，HMM）首次提出了一种能够动态评估用户连续查询隐私风险的方法，通过分析用户连续查询时存在的重要特征，以概率的方式评估用户每次查询时的隐私风险大小。最后，为了验证该方法的有效性，采用美国在线（AOL）真实的用户查询日志数据进行分析和证明，实验结果表明该方法具有较高的风险评估准确率，同时评估时间符合实际的用户查询需求。

关键词：用户连续查询；查询隐私风险；HMM 模型；动态评估

A method of privacy risk assessment in user continuous query

MA Yongdong[1], WANG Wentao[1], WANG Yinkuan[2]
(1 School of Computer Science and Technology, Donghua University, Shanghai 201620, China;
2 Shanghai Aerospace Control Technology Research, Shanghai 201109, China)

【Abstract】 The assessment of user privacy risk is of equal importance to user privacy protection. However, research on user query privacy protection has attracted widespread attention, and there has been very little research on user continuous query risk assessment. Among them, continuous query is an important performance of the query, and the user's continuous query privacy risk assessment can make the user's query privacy effectively protected. Therefore, this paper proposes a scheme that can dynamically evaluate the privacy risk of users' continuous query based on Hidden Markov Model (HMM). By analyzing the important features of users' continuous query, the privacy risk of users' each query is evaluated in a probabilistic manner. Finally, in order to verify the effectiveness of the method, AOL's real user query log data is used for analysis and verification. The experimental results show that the method has high risk assessment accuracy and the evaluation time meets the actual user query requirements.

【Key words】 user continuous query; query privacy risk; Hidden Markov Model; dynamic risk assessment

0 引 言

近年来，在线查询极大提高工作及学习效率，但也随之带来了各种隐私泄露问题[1-4]。面对这一状况，国内外众多学者提出了不同的解决方法，如匿名查询[3]、覆盖查询[4]、安全通信[5]、数据混淆[6]等。以上方法在一定程度上实现了隐私保护，但对隐私保护成本以及隐私保护合理性却未能给予适当关注。其主要原因是用户的每次查询中隐私风险是不同的。特别是在用户连续查询的场景下，如在查询一个复杂的医学问题时，需要用户多次查询或进行更长时间的查询。若采用现有的查询隐私保护方法，很容易造成隐私保护强度过高，如文献[7]中假设用户的每次查询都是高风险查询，对用户的每次查询采用相同的隐私保护方法进行保护，但这却会影响用户实际的查询精确度和查询时间效率。而查询风险评估与隐私保护是同等重要的[8]，目前仅有部分工作考虑了对用户查询时隐私泄露的风险评估[9-10]。

针对上述问题，本文结合实际查询场景，通过对用户在连续查询时存在的两大重要特性，即：递进关系（Progressive Relationship，PR）和共现关系（Co-occurrence Relationship，CR）进行分析。其中，PR 是指用户对当前的查询结果不满意时，将另外补充新词来缩小查询范围，进行再次查询。CR 是指当用户的 2 个或 2 个以上的查询频繁在同一个 Session 中共现时，则可以认为此时的查询之间存在共现关系，且该查询间有着紧密的相关关系。采用 HMM 建立用户查询动态风险评估方案，对用户的每次查询进行风险大小评估，就可以根据用户查询风险进行合理的隐私保护，实现隐私保护的同时，在很大程度上也节省了隐私保护成本。

作者简介：马永东（1994-），男，硕士研究生，主要研究方向：数据隐私保护、个性化信息检索；王文涛（1992-），男，硕士研究生，主要研究方向：云计算安全；王银款（1986-），男，学士，工程师，主要研究方向：数字信号处理。

收稿日期：2019-02-13

1 相关工作

本节主要对现有用户查询隐私保护技术和查询风险评估两个方面给出优缺点分析。对此可做研究阐述如下。

在查询隐私保护技术中,Roger[5]提出了一种在查询时隐藏用户身份与第三方监控的方法。文献[11-12]在此基础上提出了协作方案,以便让每个用户提交由其他用户生成的查询,使得用户隐藏在各自的用户组内。然而,该解决方案的响应时间比较慢,严重依赖后端系统的可用性,因此该方法无法得到广泛的应用。Balsa 等人[13]提出了一种混淆的私有 Web 搜索(OB-PWS)方案,通过将生成的虚拟查询和用户的真实查询一同发送到搜索引擎,以防止对搜索概要文件的准确推断,并提供查询可拒绝性。Howe 等人[14]提出了一种随机发布虚拟查询方法,在程序生成的诱饵流中混淆用户的查询来实现 Web 搜索中的隐私保护方法。Domingo 等人[12]提出并验证了一套保证用户查询隐私的方法和协议。但是以上方法在查询效率方面无法满足用户需求。此外,Arampatzis 等人[15]还提出一种使用模糊或混乱的查询来替换私有用户查询,从而近似于目标搜索结果。接下来,经过排序的查询结果用于覆盖私有用户感兴趣的内容。但只能落入预定义字典的查询中,极大地限制了用户灵活性。后续工作中进一步考虑了生成的查询与真实查询之间的语义相关性[16],注入与原始查询术语具有相似特异性的诱饵术语[17]来维护用户查询的匿名性和通用性。以上方法中均没有解决用户每次查询时的风险高低的问题。

查询风险评估与隐私保护起着同等重要的作用[17]。然而,只有有限的工作考虑了对用户查询时隐私泄露的风险评估。Peddinti 等人[9]基于机器学习分类器对 TMN(TrackMeNot)提供的隐私保障进行了评估。Gervais 等人[10]通过机器学习算法学习用户的原始查询和假查询之间的可链接性,对 TMN 和假查询生成等查询混淆技术进行了评价。Howe 等人[14]通过对现有 6 种混淆技术的隐私特性进行定性分析,却没有对这些技术做出定量的分析和比较。Chow 等人[18]提出了 2 个可以用来区分 TMN 虚拟查询和实际用户查询的方法。然而,现有研究很少对用户在线查询进行动态隐私泄露风险评估,但是这对隐私保护却至关重要。因此,本文针对用户连续查询提出一种动态风险评估的方法,为用户查询隐私保护提供一个较好的解决思路。

2 基于动态风险评估模型

2.1 预备知识

隐马尔可夫模型(Hidden Markov Model,HMM)是一种统计模型,用来描述一个含有隐含未知参数的马尔可夫过程,已广泛应用于模式识别、词性标注和信息提取方面[19]。

HMM 是根据可观察的参数确定该过程的隐含参数,继而利用这些参数展开后续分析。在 HMM 中,输出状态并不是直接可见的,每一个状态在可能输出的符号上都有一概率分布。因此输出符号的序列能够透露出状态序列的一些信息。为更清晰描述整个 HMM 过程,设用户查询序列为 $X=(x_1,x_2,\cdots,x_n)$,其中 x_i 代表每一个节点,且每个节点 x_i 都存在着一个转移概率 P,那么,在一个隐马尔可夫模型 M 上,一个查询序列 X 被观测的概率为在所有可能路径上的概率之和。这里可将其写作如下数学形式:

$$P(X \mid M)=\sum_{q_1,\cdots,q_n\in Q^l}\prod_{k=1}^{n+1}P(q_{k-1}\rightarrow q_k)\,P(x_k \mid q_k)\,, \tag{1}$$

其中,q_0 为初始状态,q_{n+1} 为终止状态,可得一个状态序列 $V(X \mid M)$,研究推得其所具有观察序列的最大概率为:

$$V(X \mid M)=arg\max_{q_1,\cdots,q_n\in Q^l}\prod_{k=1}^{n+1}P(q_{k-1}\rightarrow q_k)\,P(x_k \mid q_k)\,. \tag{2}$$

2.2 参数确定

在建立 HMM 模型时,结合用户的 PR 和 CR 查询特征,从转移概率和观测概率角度进行分析。研究可得阐释分述如下。

(1)转移概率。是指用户给出先前查询数据序列后,再经过若干时间后得到用户的另一查询数据的条件概率。如相连的两节点 q_1,q_2 之间存在着一个转移概率 $P(q_1\rightarrow q_2)$,由于用户在连续查询场景下,用户查询数据可区分的风险取决于用户之前的查询数据,如果考虑同一主题中的先前数据,则该数据的信息增益会变高。设 X_t 为 HMM 中个人可识别敏感信息或者是敏感主题,则 X_t 节点之间的转移概率为 $p(X_t \mid X_{t-1})$,可以对节点之间已发生的转换次数进行加权计算,研究可推得其数学公式如下:

$$\alpha=\frac{1}{count(X_t \mid X_{t-1})}, \tag{3}$$

然后根据加权转移概率的方法计算用户查询中

的隐私风险，即 $\alpha * p(X_t \mid X_{t-1})$ 。

(2)观测概率。是指某个节点可能发生的查询行为，如用户 u_i 经过 q 查询了 e 的概率为 $P(e \mid q)$ 。该值基于用户的历史查询数据进行分析和计算获得，每个节点包含一组具有观测概率的观测值，将这些观测概率建模为不同用户在先前数据 $(p(u_i \mid X_t))$ 中找到的给定数据 X_t 的概率。用户查询特定主题的数据越多，则对用户兴趣数据的推断精确度越高，该查询风险也越高。同理，采用加权计数的方式确定查询风险，研究可推得其数学公式如下：

$$\beta = \frac{1}{count(u_i \mid X_t)}. \tag{4}$$

因为用户越均匀，查询隐私风险就越高，即 $\beta * p(u_i \mid X_t)$ 。

2.3　查询动态风险评估

在用户连续查询场景中，假设用户当前的查询为 X_T，只与前一查询 X_{T-1} 相关。例如用户在查询一个复杂问题(医学知识等)时，当前的查询结果不满足用户的查询需求时，用户会添加或删除前一次的查询内容，而修改后的查询与修改前的查询之间有密切的联系，所以是满足一阶马尔可夫性质。

结合 2.2 节参数设定，设用户 u_i 查询序列为 $X_1, X_2, \cdots, X_n$，则针对该用户查询序列输出的观测结果为 $Y_1, Y_2, \cdots, Y_n$。可得用户 u_i 查询序列和观测结果的联合分布为：

$$p(Y_1, Y_2, \cdots, Y_n \mid X_1, X_2, \cdots, X_n) =$$

$$p(X_1)\, P(Y_1 \mid X_1) \prod_{t=2}^{n} P(X_t \mid X_{t-1})\, p(Y_t \mid X_t), \tag{5}$$

可以计算出用户 u_i 的查询序列($X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$)所产生的整体隐私风险为：

$$p(X_1, \cdots, X_n \mid u_i) = \min(HMM \mid u_i) \times \alpha \times$$

$$p(X_1) \times (\beta \times p(u_i \mid X_1)) \prod_{x=2}^{n} \alpha \times p(X_x \mid X_{x-1}). \tag{6}$$

其中，$(HMM \mid u_i)$ 表示用户 u_i 的所有路径的隐私概率列表，这些列表包括用户观察概率大于 0 的节点，最后，可得用户查询序列为 $X_1, X_2, \cdots, X_n$ 时的查询风险为 $p(X_1, \cdots, X_n \mid u_i)$ 。

已知用户查询时的转移概率和观察概率，通过多属性效用理论对用户的查询进行风险等级划分，本文将用户的查询风险一共划分为 5 个等级。等级越高则查询风险越高，在隐私保护时需要增加隐私保护强度；查询风险等级越低，则用户查询内容中涉及的用户隐私较少，因此，在隐私保护时无需采用高强度的隐私保护方法。本文的查询风险等级划分符合“GB/T 33132-2016 信息安全技术、信息安全风险处理实施指南”和“GB/T 31722-2015 信息技术、安全技术、信息安全风险管理”的需求[20]。在本文研究中设定的查询风险等级见表 1。

表 1　用户查询的风险等级划分

Tab. 1　Risk level of user query

风险等级	标识	描述	查询风险概率
1	很低	I	$0.0 \leqslant P < 0.2$
2	低	II	$0.2 \leqslant P < 0.4$
3	中等	III	$0.4 \leqslant P < 0.6$
4	高	IV	$0.6 \leqslant P < 0.8$
5	很高	V	$0.8 \leqslant P \leqslant 1.0$

表 1 中不同的风险等级所代表的隐私风险高低是不同的。当用户查询中没有包含新的查询时，HMM 模型会根据用户历史查询对用户进行风险评估。一旦出现新的查询时，首先会进行模型训练，然后对用户查询进行评估。且本文中的风险评估是动态的，会随着时间的推移和用户的查询数据的积累，HMM 模型对用户查询隐私风险评估也在动态地变动与调整。

3　实验分析与评估

本文的实验分析环境为 Intel(R) Core(TM) i5-3337U CPU @ 1.80 GHz，RAM 为 8 G，系统类型为 Windows10，所有的算法和仿真实验均通过 python 而得到设计完成。

3.1　数据预处理

研究采用的实验数据为 2006 年 AOL 公布了 3 个月的用户真实查询日志，提供了超过 65 万用户的 2 000万个用户搜索查询。每一条 AOL 查询日志数据由 5 部分组成，分别是：用户 ID、查询字符串、查询时间、查询内容的排名以及查询内容的 URL。实验数据集明细详见表 2。

表 2　实验数据集明细

Tab. 2　Experimental data set details

数据详情	数据统计
查询中包含的实体数量	36 389 567
新查询的实例数	21 011 340
用户“下一页”的请求数	7 887 022
用户点击事件的数量	19 442 629
用户点击的查询数量	16 946 938
唯一查询的数量	10 154 742
唯一用户 ID 的数量	657 426

在实验评估前，首先对 AOL 数据集进行预处理，对无效查询、空查询等施以过滤处理。然后将清

洗后的数据集按 80%,20%比例进行划分,其中 80%用于模型训练,20%用于结果测试。此外,为了提高模型训练效率,采用了 k 均值聚类方法将训练数据分割成 k 个聚类,使用并行处理技术同时训练 k 个数据集,这样极大地提高了数据训练效率。

在高风险查询评估中,本文设定癌症、怀孕和酒精三个主题作为高风险查询主题。为了提取包含以上 3 个主题的查询,使用了自然语言处理包 NLTK[21] 和 Gensim[22] 进行主题建模和提取相关查询主题,对每个高风险主题识别一些同义单词。本文使用了 WordStream 提供的免费关键字工具,该工具利用最新的 Google 关键字 API 寻获了数百个相关的关键字结果[23]。然后对这些关键词进行主题建模,得到与主题相关的关键词。如对“癌症”敏感主题、应用主题建模后可得相关度高的单词为肿瘤、乳腺癌、甲状腺、白血病、胰腺等。

3.2 实验评估

3.2.1 查询风险分析

在数据预处理部分,本文从 AOL 数据中提取了“癌症、怀孕和酒精”作为高风险查询内容进行研究。如图 1 所示。图 1 中,preg 表示用户查询中包含“怀孕”的信息,随着查询次数的增多,其隐私风险也在增加。cancer 表示用户查询中包含“癌症”的信息,而且随着用户查询次数的增多,其隐私风险也在增加。同理,alcohol 也是如此。

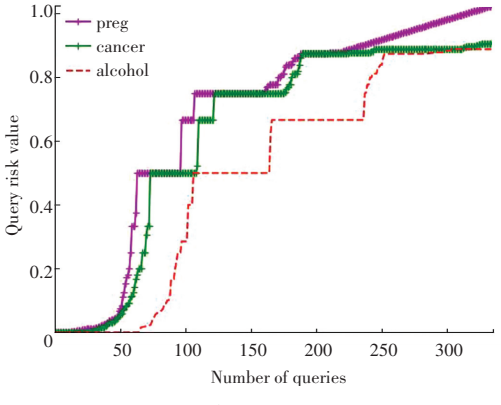

图 1 用户查询与风险关系

Fig. 1 User queries and risk relationships

从图 1 可得,3 个敏感主题均是随着查询次数的增多,查询隐私风险在逐步上升。结合表 1 中等级划分,当用户查询超过 50 次左右时,用户的查询风险等级为 2 级,表明实验结果符合风险等级的设定。

3.2.2 时间开销

时间开销主要是指当用户进行查询时,对用户的查询进行风险评估所花费的时间开销。

如图 2 所示,用户查询中 80%的查询时间开销低于 0.5 s,可得本文的评估时间在用户实际查询场景中是允许的。但在用户最初的查询中,查询时间偏高,其主要原因是刚开始需要模型训练,因此初始查询时间花费较多,随着用户查询次数的增多,用户的查询时间也会随之稳定、且时间逐渐变小。但是仍有一些查询时间高于 0.5 s,当用户查询中出现新的查询时,HMM 模型需要对数据重新训练,故而所需时间要长于一般查询时间,但整体查询时间是可以接受的。

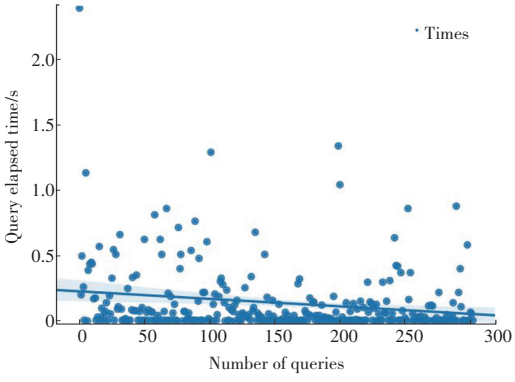

图 2 评估时间开销

Fig. 2 Estimate time cost

最后,研究得到不同用户查询风险变化趋势如图 3 所示。随着有效查询时间的增长,不同用户每次查询隐私风险不同。这是由于用户的查询内容不同,所包含的隐私信息是不同的,因此,用户查询风险也在动态变化。

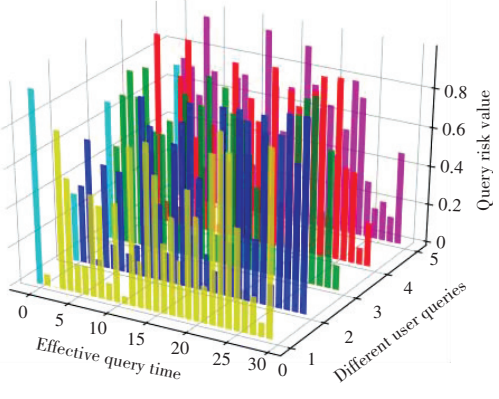

图 3 用户查询动态评估

Fig. 3 User query dynamic evaluation

4 结束语

用户查询风险评估对用户查询隐私保护起到非常重要的作用。本文提出了一种可以动态评估用户查询隐私风险的方法,对区分内容隐私风险高低具

有良好的借鉴价值。最后从时间开销、风险大小和动态评估效果三个方面进行了仿真实验验证，实验结果表明，本文方案是高效、且可行的。未来研究将对该评估方法与隐私保护算法相结合，提出一种基于动态评估用户查询风险的隐私保护方案，为用户查询隐私保护提供一种有益的解决思路和方法。

参考文献

[1] CHAIRUNNANDA P，PHAM N，HENGARTNER U. Privacy：Gone with the typing! Identifying Web users by their typing patterns[C]// 2011 IEEE Third International Conference on Privacy，Security，Risk and Trust（PASSAT）and 2011 IEEE Third International Confernece on Social Computing（SocialCom）. Boston，MA，USA：IEEE，2011：974-980.

[2] HANSELL S. AOL removes search data on vast group of Web users[N]. The New York Times，2006-08-08（C4）.

[3] SU J，SHUKLA A，GOEL S，et al. De-anonymizing Web browsing data with social networks[C]// Proceedings of the 26th International Conference on World Wide Web. Perth，Australia：ACM，2017：1261-1269.

[4] TOCH E，WANG Yang，CRANOR L F. Personalization and privacy：A survey of privacy risks and remedies in personalization-based systems[J]. User Modeling and User-Adapted Interaction，2012，22(1-2)：203-220.

[5] ROGER D. Tor and circumvention：Lessons learned[C]// Advances in Cryptology - CRYPTO 2011. Berlin/Heidelberg：Springer，2011：485-486.

[6] TEJA S，SAXENA N. On the effectiveness of anonymizing networks for Web search privacy[C]//Proceedings of the 6th ACM Symposium on Information，Computer and Communications Security，ASIACCS 2011. Hong Kong，China：ACM，2011：483-489.

[7] AHMAD W U，CHANG Kaiwei，WANG Hongning. Intent-aware query obfuscation for privacy protection in personalized Web search[C]//The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. Ann Arbor，MI，USA：ACM，2018：285-294.

[8] AHMAD W U，RAHMAN M M，WANG Hongning. Topic model based privacy protection in personalized Web search[C]// Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. Pisa，Italy：ACM，2016：1025-1028.

[9] PEDDINTI S T，SAXENA N. On the privacy of Web search based on query obfuscation：A case study of trackmenot[M]// ATALLAH M J，HOPPER N J. Privacy Enhancing Technologies. PETS 2010. Lecture Notes in Computer Science. Berlin/Heidelberg：Springer，2010，6205：19-37.

[10] GERVAIS A，SHOKRI R，SINGLA A，et al. Quantifying Web-search privacy[C]// Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security（CCS ´14）. Scottsdale，Arizona，USA：ACM，2014：966-977.

[11] CASTELLÍ-ROCA J，VIEJO A，HERRERA-JOANCOMARTÍ J. Preserving user's privacy in Web search engines[J]. Computer Communications，2009，32（13-14）：1541-1551.

[12] DOMINGO-FERRER J，BRAS-AMORÓS M，WU Qianhong，et al. User-private information retrieval based on a peer-to-peer community[J]. Data & Knowledge Engineering，2009，68(11)：1237-1252.

[13] BALSA E，TRONCOSO C，DIAZ C. OB-PWS：Obfuscation-based private Web search[C]//2012 IEEE Symposium on Security and Privacy（SP）. San Francisco，CA，USA：IEEE，2012：491-505.

[14] HOWE D C，NISSENBAUM H. TrackMeNot：Resisting surveillance in Web search[C]//Lessons from the Identity Trail：Anonymity，Privacy，and Identity in a Networked Society. Oxford：Oxford University Press，2009：417-436.

[15] ARAMPATZIS A，DROSATOS G，EFRAIMIDIS P. Versatile query scrambling for private Web search[J]. Information Retrieval，2015，18(4)：331-358.

[16] SÁNCHEZ D，CASTELLÀ-ROCA J，VIEJ O A. Knowledge-based scheme to create privacy-preserving but semantically-related queries for Web search engines[J]. Information Sciences，2013，218：17-30.

[17] PANG H H，DING Xuhua，XIAO Xiaokui. Embellishing text search queries to protect user privacy[J]. Proceedings of the VLDB Endowment，2010，3(1)：598-607.

[18] CHOW R，GOLLE P. Faking contextual data for fun，profit，and privacy[C]// Proceedings of the 2009 ACM Workshop on Privacy in the Electronic Society，WPES 2009. Chicago，Illinois，USA：ACM，2009：105-108.

[19] RABINER L，JUANG B. An introduction to hidden Markov models[J]. IEEE ASSP Magazine，1986，3(1)：4-16.

[20] 国家市场监督管理总局国家标准技术审评中心. 全国标准信息公共服务平台[EB/OL]. http://www.std.gov.cn/gb/gbQuery.

[21] NLTK project. NLTK 3.4 documentation[EB/OL]. [2018-11-17]. http://www.nltk.org/.

[22] gensim[EB/OL]. [2019-01-31]. https://radimrehurek.com/gensim/.

[23] Bryant B D，Miikkulainen R. From word stream to gestalt：A direct semantic parse for complex sentences[R]. Austin，TX：University of Texas，2001.

（上接第 48 页）

[2] 杨少鹏. 基于光纤传感的生产井动液面实时监测技术研究[D]. 哈尔滨：哈尔滨工业大学，2013.

[3] FOURIER J B. Les temperatures du globe terrestre et des espaces planetaire[J]. Mémoires de l´Académie Royale des sciences de l´Institut de France，tomeVII，1827，7：569-604.

[4] BROWNRIGG D R K. The weighted median filter[J]. Communications of the ACM，1984，27(8)：807-818.

[5] ZALEVSKY Z，MENDLOVIC D. Fractional wiener filter[J]. AppliedOptics，1996，35(20)：3930-3936.

[6] 何希平，刘波. 深度学习理论与实践[M]. 北京：科学出版社，2017.

[7] 常亮，邓小明，周明全，等. 图像理解中的卷积神经网络[J]. 自动化学报，2016，42(9)：1300-1312.

[8] Bouvrie J. Notes on convolutional neural networks[EB/OL]. 2018.

[9] SIMONYAN K，ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv：1409.1556，2014.

[10] LI Yanghao，WANG Naiyan，SHI Jianping，et al. Revisiting batch normalization for practical domain adaptation[J]. arXiv preprint arXiv：1603.04779，2016.