异常行为检测数据集快速构建方法

杜潘飞[1],王志辉[2],李雄伟[1],朱永旺[2]
(1. 陆军工程大学 石家庄校区,河北 石家庄 050003;
2. 河北建设投资集团有限责任公司,河北 石家庄 050001)

摘 要:文中提出一种快速构建异常行为检测数据集方法,该方法以一种半自动的方式完成数据集的构建,有助于减少构建过程中人工操作的工作量。首先以网络爬虫的方式自动地从互联网上搜索并下载包含指定动作的视频,之后以当前SOTA(state-of-the-art)的目标检测模型作为人物空间位置检测器,最后以人工标注和行为检测模型相结合的迭代方式完成人物行为的标注,其中需要手工完成的主要包括对下载的视频的挑选、人物边框核对以及一部分的行为标注,手工部分的工作量仅占整个任务的工作量的15%左右。实验表明,由该方法所构建的数据集可以作为异常行为检测模型的训练集使用,验证了该方法的有效性。通过该方法可以快速地构建一个大尺度、高质量的行为检测数据集,将有助于推动异常行为检测研究的深入开展。

关键词:数据集构建;行为识别;目标检测;半自动构建方法;异常行为
中图分类号:TP391.4 **文献标识码**:A **文章编号**:1673-629X(2021)09-0155-06
doi:10.3969/j.issn.1673-629X.2021.09.026

Fast Construction Strategy of Abnormal Action Detection Dataset

DU Pan-fei[1],WANG Zhi-hui[2],LI Xiong-wei[1],ZHU Yong-wang[2]
(1. Shijiazhuang Campus,The Army Engineering University of PLA,Shijiazhuang 050003,China;
2. Hebei Construction & Investment Group Co.,Ltd.,Shijiazhuang 050001,China)

Abstract:We propose a fast strategy to build abnormal action detection dataset. This strategy completes the construction of dataset in a semi-automatic way,which significantly reduces the workloads of manual operation in the construction pipeline. This method searches and downloads videos containing specified actions automatically from the Internet in the form of web crawler. The object detection model with the state-of-the-art performance is used as the person spatial detector. Finally,the human action annotations are completed iteratively by combining manual annotation and action detection model. What needs to be done manually is the selection of downloaded video,the check of person bounding boxes and a part of action annotations. In the experiment section,it demonstrates that the dataset constructed by this method is effective and can be used as the training set of abnormal action detection model. Furthermore,it shows that this method can quickly build a large-scale and high-quality action detection dataset,which will accelerate the development of research and application of abnormal action detection.

Key words:dataset construction;action recognition;object detection;semi-automatic construction method;abnormal action

0 引 言

自从AlexNet[1]在ImageNet图像分类比赛中取得标志性的进步以来,卷积神经网络(CNNs)在计算机视觉领域得到广泛的应用与发展。与此同时,从2D图像空间目标识别扩展到3D时空的视频行为识别也成为该领域的研究热点,视觉行为的识别不仅增加模型的复杂性,而且训练行为识别模型所需的数据量也显著地增加。异常行为检测作为行为识别中的一个特殊领域,在公共安全、自动驾驶、视频监控等应用中具有相当重要的意义,现代智慧城市的发展高度依赖于以人为中心分析的技术的进步。文献[2]中说明对于一个动作识别模型,越多的训练数据对模型的识别性能越有益,因此为了获得更好的检测精度,通常需要大量的训练数据。但构建一个大尺度、高质量的行为检测数据集是一个很大的挑战,其构建过程涉及视频收集、人物空间位置检测以及人物行为标注,因而如何快速地构建一个高质量的数据集对提升模型的检测性能具有重大的意义。早期的行为识别数据集HMDB-51[3]

收稿日期:2020-11-20 **修回日期**:2021-03-25
基金项目:国家青年科学基金(61602505)
作者简介:杜潘飞(1991-),男,硕士研究生,通讯作者,研究方向为计算机视觉;李雄伟,教授,硕导,博士,研究方向为信息安全。

和 UCF-101[4] 包含的数据量较少且都是对经过裁剪的视频片段做的标注,其每个视频片段中仅仅含有一个人物的一个动作,很明显这种方法构建的数据集不能满足实时运行的异常行为检测应用的需要。

为解决这个问题,文中提出一种快速构建行为检测数据集的方法,其以一种半自动的方式来完成数据集的构建,具有很强的适用性和有效性,同时有助于节省人力成本,使用该方法生成的数据集不仅覆盖多个领域的视频,且都是无裁剪的,可应用于训练实时运行的行为检测系统。(文中使用的相关程序代码及生成的数据集可在 https://github.com/xiayule158/AbnormalAction/上获取。)

1 相关工作

随着计算机硬件算力的不断提升与存储设备价格的不断下降,计算机视觉在过去的十多年中得到快速的发展,其中行为检测更是在实际生活中具有很多潜在的应用,如自动驾驶、社会安防等。为实现让计算机自动识别视频中的异常行为,首先要解决的问题便是拥有大量已标注的异常行为检测视频数据集,而到目前为止这些数据集大多是以人工的手段标注完成,花费时间长。当前数据集可根据对视频中行为的标注方法不同而将其分为视频水平的标注和视频帧水平的标注,在这部分中介绍比较典型的公共数据集,简要地介绍其特点和构建的流程。

Charades 数据集[5],2016 年公开的包含 157 个日常室内动作的数据集,其通过 Amazon Mechanical Turk(AMT)这个众包平台发布任务来完成数据的收集工作;主要构建流程为:生成室内活动剧本、要求工作人员表演剧本中的活动并记录、确认记录的视频和剧本是否对应和标注出视频中动作发生起点以及终点作为时序标签。这个数据集是基于视频水平来完成标注的,每个视频中都包含一种动作,属于裁剪过的数据集,不适合在实时的异常行为检测应用中作为训练集使用。

Kinetics-400 数据集[6],于 2017 年 Google 发布,其视频来源于 YouTube 上的短视频,整个数据集包含 400 个类别的动作,每个类别包含至少 400 个识别片段,每个片段长度大约为 10 s。其构建流程主要为:确定动作列表、根据标题与动作列表是否匹配从 YouTube 中检索视频、通过 AMT 发布视频标注任务来完成标注和最后人工完成数据集的核对。该数据集同样属于视频水平的标注,且大部分的工作由人工来完成,同样需要较高的人工成本。

AVA[7],2017 年公开发布的基于视频帧水平标注的数据集,拥有 80 个原子动作类别。与基于视频水平标注不同的是,该数据集是对视频的中动作的关键帧做的标注,且对其中的每个人物标注多种行为,这种标注方法突破了对动作视频长度的限制,属于无裁剪的动作视频,适合作为实时检测场景应用的训练集。其主要的构建流程分为:动作词典生成、视频段选择、人物边框标注(由 Faster R-CNN[8] 人物检测器来完成)和人物动作标注(在众包平台发布任务完成)。

HiEve[9](Human in Events)数据集,2020 年发布的稠密场景监控视频数据集,其视频来源于 9 个不同的稠密场景,行为识别部分包含 14 类别,且其标注方法属于视频帧水平的标注,很适合作为实际场景中异常行为检测的训练集,但其标注过程仍属于手工标注的方法,工作量较大。数据集构建流程:选择几个有复杂事件的密集场所和直接从 YouTube 上通过异常行为关键字搜索来收集视频、手工核对消除冗余视频、手工标注视频中所有行人的边框和以 20 帧的间隔手工完成行人行为类别标签的标注工作。

本部分提及的数据集和文中收集的数据集比较如表 1 所示。从中可以看出:基于视频水平的标注方法虽然工作量较小,但不适合作为实时行为检测的训练集;相比之下,基于视频帧水平标注的数据集更适合作为实时检测的数据集,但目前基于帧水平的数据集标注过程很多步骤都是由人工来完成,工作量非常大。

表 1 动作识别数据集比较

数据集	类别	标注数	视频段数	时长/h	裁剪
HMDB-51	51	6 766	6 766	-	√
UCF-101	101	13 320	13 320	-	√
Charades	157	66 500	9 848	82	√
Kinetics-400	400	306 245	306 245	850.68	√
AVA	80	1 074 458	430	107	
HiEve	14	56 643	32	0.56	
Ours	14	73 652	2 100	210	

2　视频收集

在异常行为检测的研究过程中,快速且高质量地构建数据集是算法优化的前提,为此文中提出一种半自动化的构建方法,在保证数据集质量的同时,有助于降低时间成本。整个数据集的构建过程如图 1 所示(注:图中模块带阴影的流程为自动完成部分,否则为手动完成)。从图中可以看出,视频收集是完成数据集构建的基础,视频收集的过程主要包括行为列表确定、视频文件获取和视频文件清理三个步骤。

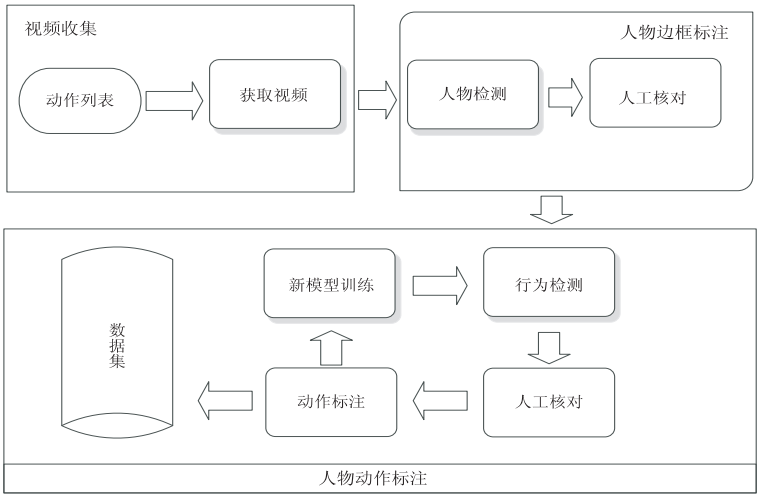

图 1　数据集构建流程

2.1　行为列表确定

文中主要考虑公共场所中的异常行为,选择对公共安全影响较大的动作作为候选对象。由于当前对公共安全中的异常行为并没有一个清晰明确的定义,因此如何构建一个描述异常行为的列表是非常困难的。为了确定候选的动作,参考现有的公开视频数据集(如:CASIA 行为分析数据库[10]、UCF-Crime[11])中的动作类别标签以及公开发表的该领域的学术论文,查阅现行法律、法规和条例等相关文件,最后确定了 14 种对公共安全妨害严重的行为作为候选列表对象,包括:行走、交通事故、抢劫、盗窃、追逐、纵火、砸车、醉酒、枪击、刀砍、打架、摔倒、跳跃、翻越护栏。

2.2　视频文件获取

在确定异常行为列表之后,以其的行为标签和同义词作为关键字从网络上搜索与之相关的视频。此外,为尽可能多地获取视频资源,对每个候选动作同时使用其他语言(如:英语、法语、日语等)作为关键词进行搜索。为提高视频的搜索和下载的效率,文中采取以 Python 语言实现的网络爬虫程序来自动地完成这项工作,减少人工参与的工作量;为提高视频下载的速度,文中采取多线程(设置为 8)的方式并行下载,对于每个类别搜索到的视频,最后将其以 mp4 的格式保存到以动作类别命名的文件夹下,实现视频的自动下载,这个过程花费 5 个小时共计下载 2 200 个视频文件,总计大小约为 18 GB。

2.3　视频文件清理

为构建高质量的行为检测数据集,视频下载完成之后首先需要对视频文件进行清理,移除其中下载出错而不能正常打开、不包含候选异常行为列表中动作类别、重复、模糊不清的视频文件。另外,为减少视频文件清理过程中工作人员的工作量,在这部分中采取行为检测器与人工核对相结合的方式来完成此项工作。

首先,对于其中下载出错而不能打开的视频文件,采用 OpenCV 这一计算机视觉工具来检验是否可以读取视频中的每一帧图像,若读取中发生错误,则说明其视频文件出现问题,直接将其移除。

其次,对于不包含候选列表中行为以及模糊不清的视频文件,采用行为检测算法来完成。文中使用 SlowFast[12] 和 AIA[13] 作为视频行为检测器来检测视频中是否包含行为列表中的行为。对于未能检测其中行为的视频,先将其移动到另外一个对应类别的文件夹中,以便后续的人工核对。

最后,由于目前发布的检测器算法模型中可检测的行为类别数并没有完全包含文中确定的行为类别,因此对于其检测到行为的视频需要人工核对其中是否包含。在这个过程中,对那些不包含候选行为、模糊的视频直接删除,而对于检测器误检的视频将其恢复到原始文件中。

至此已完成视频文件的清洗工作,共获得 2 100 个包含候选异常行为的视频文件,这个过程中人工参与的工作为核对行为检测器未能正确识别的视频文件共计 84 个,占总工作量的 4% 。

3　视频帧标注

构建的数据集是用于实时场景下监控视频中的异

常行为检测,因此采取基于帧水平的模式对视频中的动作进行标注,另外为了和当前通用的数据集在标注文件的格式上保持一致,文中采用与 HiEve 相同的标注形式。不同的是通过对数据的分析和观察认为其标注间隔过于密集,导致较大的工作量,因此,文中视频帧间隔采用 40 帧。本部分详细介绍视频帧水平的半自动标注方法,主要流程包括:视频帧人物边框标注、人物动作标注和标签人工核对。

3.1 人物边框标注

视频中动作检测首先要定位人物在视频帧中的空间位置,故而人物边框标注的问题便转化为目标检测的特殊情况,只需检测视频帧图像中的人物即可。为实现人物定位的自动化,文中采用目标检测模型作为人物检测器;另外考虑到公共场景的监控视频中可能会出现密集人群以及为了保证人物检测器的高召回率,目标检测器选择 CrowdDet[14] 和 Faster R-CNN。对于视频中的帧图像,执行以下的步骤完成人物边框标注:

首先,将帧图像 F 分别输入到人物检测器 Det_F、Det_C 中,得到各自预测的人物边框集合 B_F 和 B_C;为找到两个人物边框集合中的对应项,遍历集合 B_F 和 B_C 中所有元素组成边框对 $B_{\mathrm{FC}} = \{(b_F, b_c) \mid \forall b_F \in B_F, \forall b_C \in B_C\}$ 中的元素。

选择其中 IoU > 0.95 的项作为同一个人物边框的预测结果,取其两者的平均值作为人物边框值。在得到所有人物边框值之后,将其所在帧和边框值保存到本地文件中,作为后续核对的对象。

其次,为了保证人物边框标注的质量,在完成人物边框的自动提取之后,需要以人工的方式来核对提取的人物边框是否正确以及是否出现误判和遗漏项。在此将提取到的边框数值通过 OpenCV 展示在视频帧中,对其误判的人物边框删除,并对其未提取到的人物边框以手工的方式进行标注。

至此,便完成视频帧的人物边框标注工作,在这个过程中,仅有 5% 的人物边框需要手工添加和修改,其余都是通过人物检测器自动完成的,这在很大程度上节约了数据集标注的时间。

3.2 人物动作标注

在确定人物与视频帧中的空间位置之后,接下来要完成的便是人物动作识别及其时序定位,即人物动作标注。这是构建行为检测数据集的最关键部分,文中采取行为识别模型和手工标注相结合的方式来完成;另外,与 AVA 与 HiEve 数据集不同的是,文中确定的候选行为之间具有互斥性,因而对于其中的每个人物只标注一个动作,其具体步骤如下:

(1)将视频分别输入到目前公开可用的 SlowFast 与 AIA 行为检测模型做推理,得到部分人物行为标签的预测结果,其中包含人物动作的开始位置、结束位置、动作类别及其概率。

(2)人工核对 1 中预测的行为标签,对其中误判和未能成功识别的行为做 10% 的人工标注。

(3)使用 2 中人工核对与标注后的数据集作为训练集重新训练模型。

(4)重复 1 ~ 3 步,直到视频中约 95% 的人物的动作都被正确标注为止。

(5)对于剩下未能成功识别的人物,采用手工的方式来完成标注。

3.3 数据集概览

通过使用文中提出的方法所构建的异常行为数据集,共包含 2 100 个视频片段,总计时长约 210 个小时;14 个人物动作类别,总计标注数量为 73 652 个。其中手工标注的动作标签数量约为 10 300 个,占总标签量的 14%,相比于当前多个数据集的构建过程中手工标注量已经减少 30% ~ 50%。采用文中提出的构建方法构建的数据集的最终标注视频动作标签样本如图 2 所示,各类别动作所含标注数量分布如图 3 所示。

图 2 标注样本示例

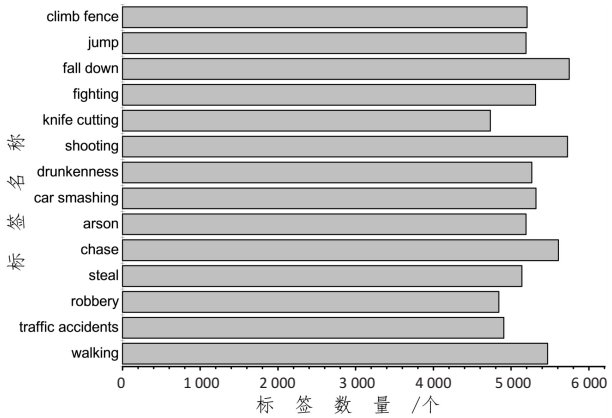

图 3　各个类别动作标注数量分布

从中可以看出各个类别所含的标注数量大致相同,满足了样本均衡的要求。

4　实　验

为验证文中提出的构建数据集方法的有效性,本部分以该数据集(为方便描述,将该数据集记为 OurData)和与之行为类别相近的 HiEve 分别作为训练集,CASIA 行为分析数据库(记为 CASIA)作为测试集,并采用当前性能处于行业领先的 SlowFast 与 AIA 作为行为检测模型来评估数据集的质量,OurData、HiEve 和 CASIA 数据集所包含的动作标签如表 2 所示。

表 2　OurData、HiEve 和 CASIA 所含动作标签比较

OurData	HiEve	CASIA
walking	walking-alone	walk
traffic accidents	walking-together	run
robbery	running-alone	bend
steal	running-together	jump
chase	riding	crouch
arson	sitting-talking	faint
car smashing	sitting-alone	wander
drunkenness	queuing	punch car
shooting	standing-alone	robbery
knife cutting	gathering	fight
fighting	fighting	follow
fall down	fall-over	gather
jump	down-stairs	meet
climb fence	crouching-bowing	overtake

为减少模型的训练时间,本实验中所有检测模型都使用与 Kinetics-400 数据集上预训练的权值作初始化,迭代次数都设置为 50,使用动量值为 0.9 的 SGD[15] 作为优化器,初始学习率为 0.01 并在 30 次迭代后衰减到 0.01,权值衰减设置为 0.000 1;实验中使用的深度学习框架为 PyTorch 且在 4 块 GeForce RTX2080Ti(11 GB)的 GPU 上完成,整个训练过程花费 25 天。实验结果如表 3 所示,从中可以看出:

表 3　各类别测试精度对比

Actions	HiEve		OurData	
	SlowFast/%	AIA/%	SlowFast/%	AIA/%
walk	60.1	61.2	61.1	63.4
run	25.2	25.3	14.3	15.2
bend	20.5	21.4	19.8	22.6
jump	3.1	3	11.2	13.1
crouch	15	15.5	11.1	9.8
faint	14.1	14.3	15.3	14.9
wander	20.3	19.4	19.5	20.3
punchcar	10.3	9.7	30.4	32.9
robbery	5.6	7.3	28.9	32.1
fight	24.5	26.7	25.6	28.2
follow	15.4	16.9	17.4	18.2
gather	40.3	45.2	30.8	30.2
meet	14.6	13.7	15.2	14.8
overtake	20.4	21.2	25.4	26.2
mAP[16]	20.67	21.48	23.28	24.42

(1)对于只在 HiEve 中含有的类别(run、gather、bend、crouch),在 HiEve 上训练的模型其检测精度都超过 OurData 训练的模型。

(2)对于只在 OurData 中含有的类别(jump、punch car、robbery、overtake),OurData 训练的模型的检测精度都超过 HiEve 训练的模型。

(3)对于 HiEve、OurData 中都含有的类别(walk、faint、fight),OurData 训练的模型的检测精度略微优于 HiEve 训练的模型,原因是 OurData 数据集中相应类别的标签数量多于 HiEve 中的标签数量。

(4)对于 HiEve、OurData 中都不含有的类别(wander、follow、meet),OurData 训练的模型的检测精度和 HiEve 训练的模型基本相同。

(5)总的来说,使用 OurData 训练的模型相对于使用 HiEve 训练的模型的整体行为检测性能要好(23.28>20.67,24.42>21.48)。

实验结果表明,OurData 可以作为异常行为检测模型的训练集来使用,证明文中提出的数据集构建方法是有效的。

5　结束语

详细描述了一种半自动构建实时异常行为检测数据集方法,通过该方法可以快速地构建一个大尺度、高质量的行为检测数据集,整个构建流程采用自动程序和手工核对相结合的方式执行,显著地减少了数据集

构建过程中人工参与的工作量。实验部分说明采用该方法构建的数据集是高质量的,可以作为行为检测及其相关模型的训练集来使用,希望该数据集构建方法可以成为研究视频中行为检测任务的一种有效的辅助工具。在接下来的工作中,将使用这一方法来完善异常行为检测数据集的构建,增加其中检测精度较低类别(jump、faint 等)的样本数量,并使用该数据集来进行异常行为检测算法的研究,以期进一步提升当前公共场景中异常行为检测模型的性能。

参考文献:

[1] ALEX K, ILYA S, GEOFFREY H. ImageNet classification with deep convolutional neural networks[C]//Proceedings of the 25th international conference on neural information processing systems. Lake Tahoe, USA: Curran Associates Inc, 2012: 1097-1105.

[2] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, USA: IEEE, 2017.

[3] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: a large video database for human motion recognition[C]//Proceedings of the 13th international conference on computer vision. Barcelona, Spain: IEEE, 2011.

[4] SOOMRO K, ROSHAN A, SHAH M. UCF101: a dataset of 101 human actions classes from videos in the wild[EB/OL]. [2020-11-16]. https://arxiv.org/abs/1212.0402.pdf.

[5] SIGURDSSON G A, VAROL G, WANG Xiaolong, et al. Hollywood in homes: crowdsourcing data collection for activity understanding[C]//Computer vision - ECCV 2016. Amsterdam, The Netherlands: Springer, 2016: 510-526.

[6] KAY W, CARREIRA J, SIMONYAN K, et al. The kinetics human action video dataset[EB/OL]. [2020-11-16]. https://arxiv.org/abs/1705.06950.pdf.

[7] GU Chunhui, SUN Chen, ROSS D A, et al. AVA: a video dataset of spatio-temporally localized atomic visual actions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake, USA: IEEE, 2018: 6047-6065.

[8] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015: 1137-1149.

[9] LIN Weiyao, LIU Huabin, LIU Shizhan, et al. Human in events: a large-scale benchmark for human-centric video analysis in complex dvents[EB/OL][2020-11-16]. https://arxiv.org/abs/2005.04490.pdf.

[10] The Center for Biometrics and Security Presearch. CASIA action database[EB/OL]. [2020-11-16]. http://www.cbsr.ia.ac.cn/.

[11] SULTANI W, CHEN Chen, SHAH M. Real-world anomaly detection in surveillance videos[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake, USA: IEEE, 2018: 6479-6488.

[12] FEICHTENHOFER C, FAN Haoqi, MALIK M, et al. SlowFast networks for video recogntion[C]//Proceedings of the IEEE international conference on computer vision. Seoul, Korea: IEEE, 2019: 6202-6211.

[13] TANG Jiajun, XIA Jin, MU Xinzhi, et al. Asynchronous interaction aggregation for action detection[C]//Proceedings of the 16th European conference on computer vision. Glasgow, UK: IEEE, 2020.

[14] CHU Xuangeng, ZHENG Anlin, ZHANG Xiangyu, et al. Detection in crowded scenes: one proposal, multiple predictions[C]//Computer vision - ECCV 2020. Glasgow, UK: Springer, 2020: 71-87.

[15] BOTTOU L. Stochastic gradient learning in neural networks[C]//Proceedings of Neuro-Nimes 91. Nimes, France: [s.n.], 1991.

[16] LIU Ling, TAMER M. Mean average precision[EB/OL]. [2020-11-16]. https://doi.org/10.1007/978-0-387-39940-9_3032.