智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

三元组深度哈希学习的司法案例相似匹配方法

李佳敏，刘兴波，聂秀山，郭杰，尹义龙

引用本文:
李佳敏, 刘兴波, 聂秀山, 等. 三元组深度哈希学习的司法案例相似匹配方法[J]. 智能系统学报, 2020, 15(6): 1147–1153.
LI Jiamin, LIU Xingbo, NIE Xiushan, et al. Triplet deep Hashing learning for judicial case similarity matching method[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(6): 1147–1153.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202006049

DOI: 10.11992/tis.202006049

三元组深度哈希学习的司法案例相似匹配方法

李佳敏[1], 刘兴波[1], 聂秀山[2], 郭杰[1], 尹义龙[1]
(1. 山东大学 软件学院, 山东 济南 250101; 2. 山东建筑大学 计算机科学与技术学院, 山东 济南 250101)

摘 要: 在数量庞大的司法案例文书中进行相似案例匹配可以有效地提升司法部门的工作效率。但司法案件文本不仅长, 而且文本自身还具有一定程度的结构复杂性, 因此司法案例文本匹配与传统自然语言处理任务相比, 具有较高的难度。为解决上述问题, 本文基于三元组深度哈希学习模型提出了一种司法案例相似匹配方法, 首先使用预训练的 BERT 中文模型分组提取文书的特征; 再利用文书三元组相似性关系, 训练深度神经网络模型, 用于生成文书的哈希码表示; 最后, 基于文书哈希码的海明距离来判断是否为相似案例。实验结果表明, 本文采用哈希学习方法极大地降低了文书特征表示的存储开销, 提高了相似案例匹配的速度。

关键词: 司法案例; 案例匹配; 相似检索; 哈希学习; 深度学习; 神经网络; BERT 模型; 三元组
中图分类号: TP391 **文献标志码:** A **文章编号:** 1673−4785(2020)06−1147−07

中文引用格式: 李佳敏, 刘兴波, 聂秀山, 等. 三元组深度哈希学习的司法案例相似匹配方法 [J]. 智能系统学报, 2020, 15(6): 1147–1153.
英文引用格式: LI Jiamin, LIU Xingbo, NIE Xiushan, et al. Triplet deep Hashing learning for judicial case similarity matching method[J]. CAAI transactions on intelligent systems, 2020, 15(6): 1147–1153.

Triplet deep Hashing learning for judicial case similarity matching method

LI Jiamin[1], LIU Xingbo[1], NIE Xiushan[2], GUO Jie[1], YIN Yilong[1]
(1. School of Software, Shandong University, Ji'nan 250101, China; 2. School of Computer Science and Technology, Shandong Jianzhu University, Ji'nan 250101, China)

Abstract: Matching similar cases in a large number of judicial case documents can effectively improve the efficiency of the judicial department. However, the text of judicial cases is not only lengthy, but also exhibits a certain degree of structural complexity. Therefore, the text matching of judicial cases is more difficult compared with the traditional natural language processing tasks. To solve the above problems and challenges, this paper proposes a judicial case similar matching method based on the triplet deep Hashing learning model. First, a pre-trained BERT model is used to extract the features of the documents in groups. The triplet similarity relationship of the documents is then employed to train the deep neural network model to generate the Hashing code representation of the documents. Finally, the Hamming distance based on the Hashing code of the documents is used to determine whether they are similar cases. Experimental results show that the Hashing learning method greatly reduces the storage cost of the documents' feature representations and improves the speed of similar case matching.

Keywords: judicial cases; case matching; similarity retrieval; Hashing learning; deep learning; neural network; BERT model; triples

随着社会的发展, 各种司法案件数量也在快速增加。相似案例匹配技术受到了广泛关注, 传统匹配工作方式还是司法工作人员筛选大量司法文书, 只能根据经验或者主动搜索去匹配对比相关司法案例文书, 这些方法往往不能做到面面俱到, 所以可能会产生不同的匹配结果, 同时成本较大、效率低。因此, 现有方法从庞大的司法案

收稿日期: 2020−06−29.
基金项目: 国家重点研发计划项目 (2018YFC0830100, 2018YFC0830102).
通信作者: 尹义龙. E-mail: ylyin@sdu.edu.cn.

例数据集中寻找相似的匹配结果，会耗费巨大的人力物力并且随着司法案件的进一步增多，工作量也会进一步增大。而当前机器学习等人工智能技术的发展为司法案例的匹配提供了可行、高效的技术和手段。把人工智能、大数据等信息处理技术引入司法案例匹配任务不仅可以提升相关司法部门监督执法能力，而且提高了办案效率，更为迈入新型司法发展道路打下基础。

对于司法案例文书的相似性匹配来说，传统的基于关键词匹配技术只能发掘出浅层的语义关系，并不能挖掘出具体司法案例之间的复杂语义匹配信息[1]。此外，传统的方法为了追求精度，通常将案件文书转化为实值表示，然后通过计算实值表示之间的距离来衡量相似度，并判定匹配程度。然而，当待匹配的司法案例文书数据规模较大时，遍历文书库的计算开销也会很大，因此这种方法不适用于大规模的相似案例匹配场景。随着机器学习的发展，机器学习模型越来越多地被用于文本之间的复杂信息进行匹配。近年来，最近邻搜索的代表方法——哈希方法因其低存储、计算效率高等优点引起广泛关注。

哈希方法[2]可以将文档、图像、视频等多媒体转换成一个紧凑的二进制编码，并保留原始数据间的相似性关系。该二进制编码(也称哈希码)间的距离度量使用的是海明距离，其可以通过硬件的异或运算进行快速求解。因此，哈希方法在存储和效率上具有极大的优势。但是，司法相似案例文书在特征空间中总是聚在一起，其映射而成的哈希码碰撞几率很大，难以做到精准识别。因此，针对现有问题，本文提出的方法通过引入三元组损失函数[3]来实现减小类内距离增大类间距离的作用，使得最终学习得到的哈希码更具区分性，提高相似司法文书匹配的准确度。同时，哈希方法因其二值的表示形式，可以大大提升检索的速度，对大规模司法案例数据匹配效率的提升具有较大的价值。

1　相关工作

哈希学习模型可以分成两类：无监督方法和监督方法。无监督方法仅使用无标签信息的训练数据来学习哈希码。经典的无监督方法包括：局部敏感哈希 (locality-sensitive Hashing, LSH)[4]，LSH 通过随机映射的方法产生哈希码；谱哈希 (spectral Hashing, SH)[5]是把哈希编码的过程看作图分割的问题，利用图割算法生成哈希；基于图的哈希方法 (Hashing with graphs, AGH)[6]利用锚点图建立一个易于处理的低秩邻接矩阵来进行哈希学习；迭代量化哈希 (iterative quantization, ITQ)[7]首先使用主成分分析进行降维处理，然后学习得到最优旋转矩阵来减少量化损失。

有监督哈希学习方法利用监督信息学习样本数据之间的关系，其性能一般优于无监督的方法。例如，二元重建嵌入方法 (binary reconstructive embedding, BRE)[8]通过最小化输入特征距离与相应哈希码距离的重构误差来学习哈希函数；最小损失哈希学习方法 (minimal loss Hashing, MLH)[9]通过最小化铰链类损失函数来实现哈希学习；核监督哈希 (supervised Hashing with kernels, KSH)[10]是一种基于核函数的监督哈希方法。

随着深度网络的发展，深度学习技术被广泛应用于计算机领域，如图像分类[11-13]、目标检测[14]等。深度学习的思想来源于对人工神经网络的研究，其结构就是包含多个隐含层的多层感知机。最早提出的基于深度学习的哈希方法是语义哈希，该方法首先训练受限玻尔兹曼机，然后通过训练好的模型对数据进行哈希编码。近几年，深度哈希学习算法在图像检索方面取得了很大进展[15-17]。与传统的哈希学习方法不同，深度哈希通过使用深度神经网络来代替线性投影的方法，进而学习二进制编码。例如，深度哈希 (deep Hashing, DH)[18]使用深度神经网络获取图像数据的分层非线性变换来学习二进制编码。Yang 等[19]提出监督语义保留深度哈希 (supervised semantics-preserving deep Hashing, SSDH)，该方法通过最小化分类误差的目标函数来训练模型，并且在深度网络隐含层的输出中得到哈希函数。深度学习方法可以很好地提取原始样本深层次特征，哈希学习方法在大规模的数据任务中具备独特的优势，因此深度哈希学习方法成为当前的研究热点。

如图 1 所示，本文提出一种基于三元组深度哈希学习的相似案例匹配方法，将司法案例文书转换成二进制编码形式，不仅解决了存储开销和匹配速度等问题，也很大程度上保留了原始文书之间的相似关系，该方法适用于大规模的相似案例匹配的场景。

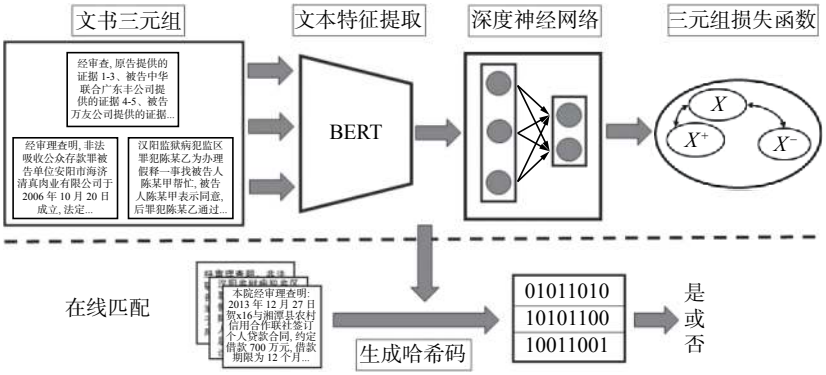

图 1　基于三元组深度哈希的架构

Fig. 1　Architectural overview based on the deep Hashing of triples

2　本文方法

基于深度哈希学习算法的优良性能, 本文提出了一种基于三元组深度哈希的司法相似案例匹配方法。该方法可以有效提升司法案例匹配效率。

假设 I 是文本空间, 哈希学习的目标是得到一个映射 $\boldsymbol{F}: \boldsymbol{I} \rightarrow \{0,1\}^q$, 这样输入的文书就可以编码成 q 位的哈希码, 同时保留文书的相似性。本文提出基于三元组的深度哈希的司法案例相似匹配方法如图 1 所示, 该方法共由三部分组成: 1) 使用中文自然语言处理模型 BERT 提取文书的特征表示; 2) 建立基于三元组文书相似度的损失函数项; 3) 利用深度神经网络学习一个映射, 用于按位生成哈希码。

2.1　特征提取

目前研究者通常采用向量空间模型来描述文本向量, 但是如果直接用分词算法和词频统计方法得到的特征项来表示文本向量中的各个维, 那么这个向量的维度将是非常的大。这种未经处理的文本矢量不仅给后续工作带来巨大的计算开销, 使得整个处理过程的效率非常低, 同时也会损害分类、聚类算法的精确性。因此, 为提升特征表示的精度和效率, 需要对文本向量做进一步净化处理, 在保证原文含义的基础上, 寻找对文本特征类别最具代表性的特征表示。解决这一问题最有效的办法就是通过特征来选择降维。本文使用中文自然语言处理模型 BERT[20] 进行司法案例文书的特征表示。在保留文书语义的前提下, 尽可能地降低文书的文字数量, 以便于降低文书特征表示的维度。

BERT 模型的目标是利用大规模无标注语料训练获得包含丰富语义信息的文本特征。BERT 模型的主要输入是文本中各个字/词的原始词向量, 该向量既可以随机初始化, 也可以利用 Word2-Vector[21] 等算法进行预训练作为初始值; 模型输出是文本中各个字/词融合了全文语义信息后的向量表示, BERT 模型如图 2 所示。

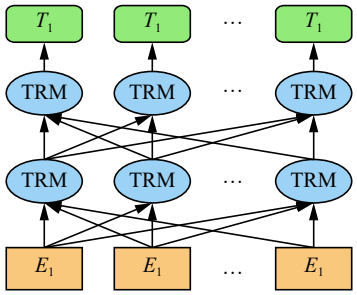

图 2　BERT 模型示意

Fig. 2　Schematic diagram of BERT

对于数据集中的司法案例文书, 本文首先利用文本预处理手段去除掉数字、标点符号、没有实际意义的虚词等文字; 然后, 将文书按 1 024 个汉字为一组, 输入到预训练好的 BERT 中文模型中, 得到 768 维的向量表示。重复此步骤, 直到得到文书各组的特征表示, 并将其拼接成完整的特征表示。特别地, 对于不够 1 024 个汉字的文书片段, 用数字 0 补齐。

2.2　基于三元组文书相似度的损失函数

在现有的监督哈希方法中, 辅助信息多采用成对类标签的形式, 用来表示样本对的语义相似或者不相似。这些方法中的损失函数通过成对类标信息来建立, 但这种处理方式只有成对样本之间相似性的精确数值, 缺乏了更丰富的语义信息。为了解决上述问题, 基于三元组的类标信息被提出来[20]。与成对类标信息不同, 三元组类标信息描述 3 个样本之间的相互关系。如图 3 所示, 一个三元组包含一个锚点样本、一个正样本和一个负样本, 并且锚点样本与正样本的相似度大于锚点样本与负样本的相似度。基于三元组的

类标信息比成对类标信息更容易获得，并且对于给出的成对相似的辅助信息，很容易地生成一组三元约束。

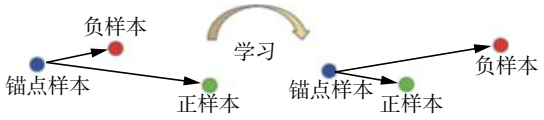

图 3　三元组关系示意

Fig. 3　Diagram of triple relationship

本文提出的方法使用三元排序损失[21]的变体来保持样本的相对相似性。具体地，给定 (I,I^+,I^-) 形式的文书样本三元组，其中文书 I 与文书 I^+ 的相似性大于文书 I 与文书 I^- 的相似性。本文方法的目标是学习到一个映射 $F(\bullet)$，使得二进制代码 $F(I)$ 更接近于 $F(I^+)$ 而不是 $F(I^-)$。基于三元组的损失函数为

$$\begin{aligned} &L_{\text{triplet}}(F(I),F(I^+),F(I^-)) = \\ &\max(0,\|F(I)-F(I^+)\|_2-\|F(I)-F(I^-)\|_2+g) \\ &\text{s.t.}\ \ F(I),F(I^+),F(I^-)\in[0,1]^q \end{aligned} \tag{1}$$

式中：g 是度量匹配文书对与不匹配文书对之间距离的阈值参数，本文使用多个阈值进行了实验，在这里取 g=0.5。对于给定的三元组 $\{I_i\}$，此损失函数可以最大化匹配文本与不匹配文本对之间的距离。

2.3　基于深度神经网络的哈希学习模型

在此模块中，用自然语言处理模型 BERT 提取的文书特征表示作为输入，利用深度神经网络学习一个从文书特征到海明空间哈希码的非线性映射，用于生成未知文书的哈希表示[22]。特别地，采用式 (2) 所示的损失函数构建哈希学习模型：

$$\begin{aligned} &L_{\text{triplet}}(F(I),F(I^+),F(I^-)) = \\ &\max(0,\|F(I)-F(I^+)\|_2-\|F(I)-F(I^-)\|_2+0.5) \\ &\text{s.t.}\quad F(I),F(I^+),F(I^-)\in(0,1)^k \end{aligned} \tag{2}$$

式中：F 为深度神经网络；I 为利用 BERT 模型提取到的特征表示；k 为哈希码长度。在本文的方法中，为了减小训练开销，降低模型复杂度，F 定义为含有两个隐含层的深度神经网络，如图 4 所示。

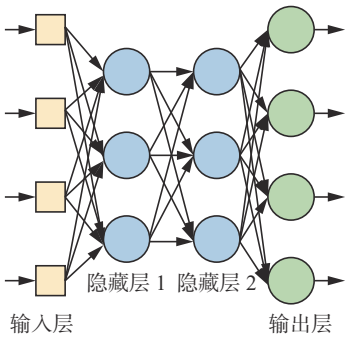

图 4　深度神经网络模型

Fig. 4　Depth neural network model

第一个隐含层采用 Relu 激活函数，与传统的 S 型激活函数相比，Relu 函数能够有效缓解梯度消失问题而且收敛速度快，其在深度神经网络中的使用加速了深度网络的直接监督式训练的突破。它的函数表达式为

$$f(x)=\max(0,x) \tag{3}$$

第二个隐含层采用 Sigmoid 激活函数，将输出映射到 0 与 1 之间。Sigmoid 函数公式为

$$f(x)=\frac{1}{1+\mathrm{e}^{-x}} \tag{4}$$

对于神经网络的优化训练，本文采用随机梯度下降法来实现。梯度下降法是为了找到最优的权重参数，最小化损失函数。通过对损失函数求导得到梯度方向，梯度下降的公式为

$$\theta=\theta-\eta\frac{\mathrm{d}J(\theta)}{\mathrm{d}\theta} \tag{5}$$

式中：η 为学习率即步长，η 决定了参数更新的快慢。学习率是一个固定的值，本实验中设 η 为 0.001，迭代次数为 120 轮，0.5 作为将实值转化为二值码 (哈希码) 的阈值。

2.4　新文书的哈希编码与匹配

基于预训练的 BERT 模型与三元组深度学习模型，生成文书的哈希表示，进而根据哈希码计算海明距离来进行相似性匹配，相关步骤如图 5 所示。

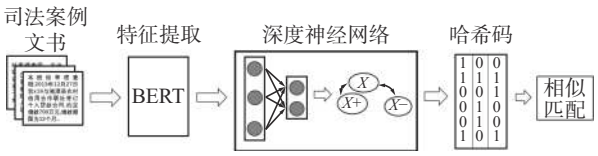

图 5　相似司法案例文本匹配框架

Fig. 5　Framework of similar judicial matching

具体地，给定新的文书，首先，进行特征提取，采用文本预处理方式，对文书内容进行缩减，然后按 1 024 汉字长度进行分组，并分别将每组汉字输入到预训练的 BERT 中文模型；然后将每组汉字的特征表示做融合拼接，得到文书的特征表示。其次，将文书的特征表示输入到预训练的哈希学习深度神经网络，得到文书的实数表示。进一步，采用 0.5 的阈值将实数表示二值化，即大于 0.5 的转化为 1，小于 0.5 的转化为 0。最后，一个汉字的文书转化为了长度为 K 的哈希码。在做文书匹配时，采用哈希距离的大小来衡量两个文书的相似度。算法的实现步骤：

1) 初始化算法的基本参数：深度神经网络参数 w，最大迭代次数 N=120；

2) 对于输入的案例文书数据集 Z 和 X, 通过 BERT 中文模型计算每个样本的特征向量;

3) 对于每一个案例文书三元组, 根据式 (2) 计算锚点样本与正负样本间距离, 最小化损失函数;

4) 使用 SGD 随机梯度下降算法更新参数 w;

5) 重复 2)、3), 直至循环次数达到最大迭代次数;

6) 在深度神经网络中输入测试案例数据集, 计算出测试集哈希码: $B = \mathrm{sign}(F(X;W))$;

7) 根据汉明距离计算两个文书的相似度, 输出匹配结果。

3　实验与结果

3.1　度量标准

该实验采用匹配准确度进行度量[23], 准确率公式定义为

$$A = \frac{p+t}{n} \tag{6}$$

式中: p 表示正确的匹配数目; t 表示正确的非匹配数目; n 表示总的样本数量。

3.2　数据集

本文将在“中国裁判文书网”公开的司法文书数据集上检验方法的性能。首先概要介绍一下实验中用到的数据集。

此数据集中每份数据由三篇司法案例文书组成, 对于每篇文书, 根据提供的事实描述。对于每份数据, 本文用 (d,d_1,d_2) 来代表该组数据, 其中 d、d_1、d_2 均对应某一篇司法文书。对于训练数据, 文书数据 d 与 d_1 的相似度大于 d 与 d_2 的相似度, 即 $\mathrm{sim}(d,d_1) > \mathrm{sim}(d,d_2)$。此数据集总共涉及 5000 组三元组对文书, 所有的文书三元组对都一定属于民间借贷。将 4 500 组文书三元组作为训练集, 500 组文书三元组作为测试集。

3.3　实验结果及其分析

本文在公开的司法文书数据集上进行了实验以检验算法的性能, 分别使用 48、64、96、128、256、512、768 bit 哈希码进行实验, 并将本文的方法与 SH[5]、PCA-ITQ[24]、PCA-RR[24]、MFH[25] 等哈希学习方法进行比较, 实验结果如表 1 所示。表 1 为数据集在不同编码位数下, 不同方法案例匹配的准确率。由表 1 可以看出, 本文的方法具有较高的准确度。

表 1　本文方法与其他算法准确度比较

Table 1　Accuracy comparison between the method and other algorithms

方法	哈希编码位数/bit						
	48	64	96	128	256	512	768
SH	0.4750	0.4820	0.4850	0.4760	0.5050	0.5120	0.5200
PCA-ITQ	0.5066	0.5096	0.5106	0.5196	0.5040	0.5132	0.5160
PCA-RR	0.5114	0.5126	0.5070	0.5186	0.5098	0.5048	0.5074
MFH	0.5244	0.5206	0.5230	0.5258	0.5240	0.5230	0.5322
Proposed	0.5790	0.5630	0.5590	0.5800	0.5870	0.5620	0.5690

本文方法采用预训练的 BERT 中文模型与哈希学习的三元组深度神经网络模型, 解决了传统相似司法案例文书匹配存储开销大、效率低等问题。该方法在不同哈希编码的位数下的准确率均高于其他方法, 证实了本文算法的有效性和优越性。

本文对算法的收敛性进行了实验, 实验结果如图 6 所示。由图 6 可以看出, 在不同的哈希编码位数下, 目标函数均具备良好的收敛性。

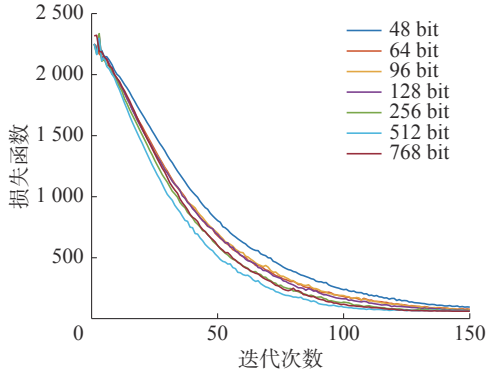

图 6　不同哈希编码位数的目标函数收敛情况

Fig. 6　Convergence of the objective function for different bits of the Hashing code

4　结束语

本文提出了一种基于三元组深度哈希的司法案例相似性匹配方法。该方法主要思路是通过三元组损失函数来训练深度神经网络，使得最终得到的哈希码保留文书样本原有的相似关系，并且具备相同语义数据之间的距离尽可能小，不同语义数据之间的距离尽可能大的特点。该方法采用哈希学习技术极大地降低了文书表示的存储开销，提高了相似案例匹配的速度，适用于大规模的相似案例匹配的场景。

本文的主要贡献如下：

1) 本文提出了一种基于三元组深度哈希学习的相似案例匹配方法，通过将司法案件文书转化为哈希码来进行快速相似度匹配。该方法在得到文书的特征表示的基础上建立基于三元组文书相似度的损失函数项，并利用深度神经网络生成未知文书的哈希表示，利用哈希码进行相似性匹配，提高了匹配效率。

2) 使用中文自然语言处理模型 BERT 提取文书的特征表示。在保留文书语义的前提下，尽可能地降低文书的文字数量，以便于降低压缩文书特征表示的维度。

3) 应用公开的法律文书数据集的实验结果表明本文提出的相似案例匹配算法提高了相似案例匹配的速度和准确度，适用于大规模的相似案例匹配的场景。

实验结果表明，本文提出的相似案例匹配方法在准确率和效率方面优于已有方法。

参考文献：

[1] 贾君枝, 毛海飞. 基于法律框架网络本体的语义匹配技术研究 [J]. 情报理论与实践, 2008, 31(1): 124–128.
JIA Junzhi, MAO Haifei. Research on the semantic matching technology based on the Chinese legal framenet ontology[J]. Information studies: theory & application, 2008, 31(1): 124–128.

[2] INDYK P, MOTWANI R. Approximate nearest neighbors: towards removing the curse of dimensionality[C]//Proceedings of the 30th Annual ACM Symposium on Theory of Computing. Dallas, USA, 1998: 604−613.

[3] LAI Hanjiang, PAN Yan, LIU Ye, et al. Simultaneous feature learning and hash coding with deep neural networks[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3270−3278.

[4] GIONIS A, INDYK P, MOTWANI R. Similarity search in high dimensions via hashing[C]//Proceedings of the 25th International Conference on Very Large Data Bases. Edinburgh, Scotland, 1999: 518−529.

[5] WEISS Y, TORRALBA A, FERGUS R. Spectral hashing[C]//Proceedings of the 21st International Conference on Neural Information Processing Systems. Vancouver, Canada, 2008: 1753−1760.

[6] LIU Wei, WANG Jun, KUMAR S, et al. Hashing with graphs[C]//Proceedings of the 28th International Conference on Machine Learning. Bellevue, USA, 2011: 1−8.

[7] GONG Yunchao, LAZEBNIK S, GORDO A, et al. Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(12): 2916–2929.

[8] KULIS B, DARRELL T. Learning to hash with binary reconstructive embeddings[C]//Proceedings of the 22nd International Conference on Neural Information Processing Systems. Vancouver, Canada, 2009: 1042−1050.

[9] NOROUZI M, FLEET D J. Minimal loss hashing for compact binary codes[C]//Proceedings of the 28th International Conference on International Conference on Machine Learning. Bellevue, USA, 2011: 353−360.

[10] LIU Wei, WANG Jun, JI Rongrong, et al. Supervised hashing with kernels[C]//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012: 2074−2081.

[11] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2012: 1097−1105.

[12] SZEGEDY C, LIU Wei, JIA Yangqing, et al. Going deeper with convolutions[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1−9.

[13] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 1026−1034.

[14] SZEGEDY C, TOSHEV A, ERHAN D. Deep neural networks for object detection[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA, 2013: 2553−2561.

[15] LIN K, YANG H F, HSIAO J H, et al. Deep learning of binary hash codes for fast image retrieval[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Boston, USA, 2015:

27−35.

[16] XIA Rongkai, PAN Yan, LAI Hanjiang, et al. Supervised hashing for image retrieval via image representation learning[C]//Proceedings of the 28th AAAI Conference on Artificial Intelligence. Québec City, Québec, Canada, 2014: 2156−2162.

[17] 李泗兰, 郭雅. 基于深度学习哈希算法的快速图像检索研究 [J]. 计算机与数字工程, 2019, 47(12): 3187–3192.
LI Silan, GUO Ya. Fast image retrieval based on hash algorithm in depth learning[J]. Computer and digital engineering, 2019, 47(12): 3187–3192.

[18] LIONG V E, LU Jiwen, WANG Gang, et al. Deep hashing for compact binary codes learning[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 2475−2483.

[19] YANG H F, LIN K, CHEN Chusong. Supervised learning of semantics-preserving hashing via deep neural networks for large-scale image search[J]. Computer Science, 2015, 10(12): 131−138.

[20] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA, 2019: 4171−4186.

[21] 汪静, 罗浪, 王德强. 基于 Word2Vec 的中文短文本分类问题研究 [J]. 计算机系统应用, 2018, 27(5): 209–215.
WANG Jing, LUO Lang, WANG Deqiang. Research on Chinese short text classification based on Word2Vec[J]. Computer systems & applications, 2018, 27(5): 209–215.

[22] LI Xi, LIN Guosheng, SHEN Chunhua, et al. Learning hash functions using column generation[C]//Proceeding of the 30th International Conference on Machine Learning, 2013: 142−150.

[23] NOROUZI M, FLEET D J, SALAKHUTDINOV R. Hamming distance metric learning[C]//Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2012: 1061−1069.

[24] GONG Yunchao, LAZEBNIK S. Iterative quantization: a procrustean approach to learning binary codes[C]//Proceedings of CVPR 2011. Providence, USA, 2011: 817−824.

[25] SONG Jingkun, YANG Yi, HUANG Zi, et al. Effective multiple feature hashing for large-scale near-duplicate video retrieval[J]. IEEE transactions on multimedia, 2013, 15(8): 1997–2008.

作者简介:

李佳敏, 硕士研究生, 主要研究方向为智能媒体处理。

刘兴波, 博士研究生, 主要研究方向为智能媒体处理、计算机视觉。

尹义龙, 教授, 博士生导师, 主要研究方向为人工智能理论与方法、机器学习、数据挖掘。主持国家自然科学基金重点项目 1 项、国家重点研发专项课题 1 项、面上项目 3 项、青年项目 1 项, 主持省部级科研项目 11 项。发表学术论文 300 余篇。