文章编号:1000-5862(2021)03-0305-09

一种标签嵌入子空间的跨模态离散哈希学习

滕少华[1],郭兰君[1],张　巍[1],滕璐瑶[2]
(1. 广东工业大学计算机学院,广东 广州　510006;2. 维多利亚大学应用信息研究中心,维多利亚 墨尔本　3011)

摘要: 针对有监督的跨模态哈希检索存在计算成本高及准确度不高的问题,提出了一种标签嵌入子空间的跨模态离散哈希学习方法,将数据信息和标签信息同时嵌入到公共子空间中,通过以带标签信息的语义特征逼近公共子空间、并生成低松弛的离散哈希码,降低了计算成本,快速生成了具有丰富语义的公共子空间. 经 3 个标准数据集对比实验,结果表明其准确率均优于被比较的方法.

关键词: 标签嵌入; 子空间; 离散哈希
中图分类号: TP 391　　**文献标志码:** A　　**DOI:** 10.16357/j.cnki.issn1000-5862.2021.03.13

0　引言

跨模态检索意味着先从模态 A 中(如图像模态)查询,再从模态 B 中查询(如文本模态)得到最相关的结果,由于不同媒体类型数据之间存在“异构鸿沟”[1],所以导致图像、文本等不同媒体数据的特征表示不一致,无法直接度量它们的相似性. 虽然不同媒体数据表征异构,但是在语义上却是互相关联的[2],这也使得跨模态检索成为可能. 由于哈希方法生成的哈希码简短,占据存储空间小,能有效加速检索速度,并提高检索的准确性,因此哈希方法在跨模态检索领域中也变得越来越具有吸引力.

跨模态检索的关键问题是发现成对多模态数据之间的相关性,并利用机器学习方法进行学习. 如何在大规模数据集上实现快速、准确的哈希检索,如何发现并学习到异构数据之间的潜在语义,这已经成为了在跨模态检索领域中一个非常值得探讨的研究方向,近年来已有不少研究进展[3-4].

目前的哈希方法可分为有监督哈希和无监督哈希. 有监督方法保留了训练数据集的语义标签,在训练阶段时从标签信息中获得语义之间的相关性. 一般来说有监督方法比无监督哈希检索方法准确性更高,近年来有监督方法也得到了越来越多的关注. M. Bronstein等[5]提出了一个基于有监督的相似性框架,将模态数据映射到汉明空间中,并将其表示为正例和负例的二进制分类问题,增强了学习效果,但未考虑模态内的相关性,准确度不高. Lin Zijia等[10]通过最小化 KL 散度去近似学习具有成对哈希码的相似矩阵. Zhang Dongqing等[11]利用语义相关最大化,将标签语义嵌入到哈希学习过程中,通过语义标签之间的余弦相似度构造成对数据的语义相似性,用待学习的哈希码学习相似性矩阵. 以上 2 种方法的共性是保持哈希码的成对相似性,但当出现只有标签信息、没有成对数据集时,计算的时间复杂度会大大增加. Kan Meina等[12]用一种多视角判别分析方法,通过获取所有视角共享的公共空间,从而匹配不同视角的样本. Wang Kaiyue等[13]对不同模态的数据集 X_a 和 X_b 学习相应的投影矩阵 $\boldsymbol{U}_a$ 和 $\boldsymbol{U}_b$,利用 $\boldsymbol{U}_a$ 和 $\boldsymbol{U}_b$ 把 X_a 和 X_b 映射到类标签定义好的公共空间中,从而实现不同模态的检索. 在利用 CCA 方法后,因为在模态中的文本、图片特征维数高,需要进一步的降维操作. Gong Yunchao[15]等利用 CCA 方法对文本、图像及第 3 种视角标签共同投影到子空间中,最后在子空间中再利用相似度函数进行距离比较; V. Ranjan等[16]提出了对 CCA 方法进行进一步扩展,在学习子空间同时合并了数据集的多标签信息. 以上 2 种方法共同点是先利用 CCA 方法,再将标签信息和语义信息嵌入到子空间中,但它们的计算成本和时间较高. 为了解决上述问题,本文提出了标签增强的跨模态离散哈希检索方法,将标签信息和原始数据信息一起嵌入到子空间中,使

收稿日期:2020-10-17
基金项目:广东省重点领域研发计划(2020B010166006),国家自然科学基金(61972102)和广州市科技计划(201903010107,201802030011,201802010026,201802010042,201604046017)资助项目.
作者简介:滕少华(1962—),男,江西南昌人,教授,博士,主要从事大数据、数据挖掘、数字音频分析与处理、网络安全方面的研究. E-mail: shteng@gdut.edu.cn

得子空间同时保留了原始数据和标签信息的语义,更好地实现了它们在子空间中的语义相关. 另外,将标签信息 2 次嵌入到要学习的公共潜在子空间中. 本文方法直接嵌入标签信息到潜在子空间中,不同于之前直接将 2 个模态的数据投影到共同的潜在空间的方法,这样不能较好利用标签信息. 不同模态投影到的公共子空间里应包含共同的语义特征,将语义特征融合到子空间中,进一步学习哈希码. 这样获得的哈希码既包含模态的标签信息,也能将标签信息和语义信息联合起来,进一步获得更多的语义信息. 对潜在子空间采取正交约束,利用子空间生成离散哈希码 B,并且生成的哈希码包含了更多的语义信息. 在优化过程中未使用大规模矩阵,在保证检索准确率的前提下又能在较大程度上减少计算成本.

1 相关工作

众所周知,同一个主题相关的多模态数据通常是以不同的形式存在,如商品详情介绍通常是由文字描述和图像组成,因为来自不同形式的数据具有语义相关性,将原始数据转换为潜在的语义空间可以最大限度地利用它们. 矩阵分解将原始数据学习为潜在的低维空间,是学习数据潜在信息的最有用工具之一,Zhou Jile 等[17] 提出的 CMFH、Fei Wu 等[18] 提出的 LSSH 以及 Wang Di 等[19] 通过矩阵分解寻找潜在的低维空间,以适当地重构多模态数据并量化重构系数以获得二进制代码,LSSH 使用矩阵分解的强大功能来发现图像中隐藏的语义. 这些方法的良好性能说明了矩阵分解在多模式哈希学习应用中的有效性.

但是,传统的矩阵分解方法基本上是无监督的,无法利用标签信息. 因此,它不适用于监督学习问题. 为此,本文为跨模态哈希学习任务提出了一种标签增强的跨模态离散哈希学习方法,直接嵌入标签信息到潜在子空间中,不同模态投影到的公共子空间里应包含共同的语义特征,将语义特征融合到子空间中,可以进一步学习哈希码,这样获得的哈希码既包含模态的标签信息又包含着原始模态的数据信息. 以此呈现出来的语义表示更丰富,能挖掘到更深层次的语义信息.

2 标签嵌入子空间的跨模态哈希检索方法

为了跨模态检索,学习不同模态映射出的公共语义潜在空间,并基于公共潜在子空间中的表示生成哈希码,进行数据库内检索. 为了检索样本外数据,学习哈希函数,将它们映射到汉明空间中生成哈希码.

2.1 符号介召

假使有一个 n 个样本 m 个不同模态的训练数据集 $\boldsymbol{X}^{(t)} = \{x_i^{(t)}\}_{i=1}^{n} \in \mathbf{R}^{d_t \times n}$ 是第 t 个模态的 $d_t(t=1,2,\cdots,m)$ 维度矩阵,$\boldsymbol{L}$ 是真实数据的标签矩阵,$L_{ij}=1$ 代表 j 样本属于第 i 类,反之,则不属于. 哈希码 $\boldsymbol{B} = (b_1, b_2, \cdots, b_n) \in \{-1,1\}^{k \times n}$ 全部为二进制编码,r 是哈希码的长度. $\boldsymbol{P}_t(t=1,2,\cdots,m)$ 和 $\boldsymbol{Q}_t(t=1,2,\cdots,m)$ 均属于映射矩阵.

2.2 子空间学习

给定一个实例(比如商品详情介绍),它拥有不同模态,既有商品图片,又有文字介绍. 图片和文字都用来描述商品,即不同模态的描述内容是相同的,换句话说,不同模态之间存在一致信息和相关性. 此外,不同模态之间还共享相同的标签信息,因此可以把标签视为不同模态之间的桥梁.

与 $\boldsymbol{P}_t\boldsymbol{X}_t - \boldsymbol{V}$ 直接将 2 个模态的数据投影到公共潜在子空间未使用标签信息不同,因为不同模态投影的子空间里包含着共同语义特征 $\boldsymbol{QL}$,首先学习标签信息和原始数据之间的关系,再将获到的信息嵌入到子空间内,这样不仅利用了标签信息而且将其和原始数据之间的语义相关性挖掘出来:

$$\min_{\boldsymbol{V}} \sum_{t=0}^{m} \| \boldsymbol{P}_t\boldsymbol{X}_t - \boldsymbol{QL} \|_F^2 + \| \boldsymbol{QL} - \boldsymbol{V} \|_F^2. \quad (1)$$

2.3 量化

为了减少哈希码 $\boldsymbol{B}$ 和公共嵌入空间 $\boldsymbol{V}$ 之间的量化误差,进一步最小化了 $\boldsymbol{B}$ 和 $\boldsymbol{V}$ 之间的误差,但在对 $\boldsymbol{B}$ 施加约束后,$\boldsymbol{B}$ 的生成会随之变得棘手和困难. 因此,增加对 $\boldsymbol{V}$ 的约束,从而放松对 $\boldsymbol{B}$ 的约束,完成后的公式为

$$\min_{\boldsymbol{B},\boldsymbol{V}} \| \boldsymbol{B} - \boldsymbol{V} \|_F^2,$$
$$\text{s.t. } \boldsymbol{B} \in \{-1,1\}^{r \times n}, \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = n\boldsymbol{I}, \boldsymbol{V}\mathbf{1}_n = 0_r.$$

2.4 目标函数

为了实现将标签信息和数据映射到同一个子空间中,增强子空间中的标签信息效果,并得到了降低约束的离散哈希码,使用以下目标函数来进行约束

$$\min \sum_{t=1}^{m} \beta \| \boldsymbol{P}_t\boldsymbol{X}_t - \boldsymbol{QL} \|_F^2 + \gamma \| \boldsymbol{QL} - \boldsymbol{V} \|_F^2 + \alpha \| \boldsymbol{B} - \boldsymbol{V} \|_F^2,$$
$$\text{s.t. } \boldsymbol{B} \in \{-1,1\}^{r \times n}, \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = n\boldsymbol{I}, \boldsymbol{V}\mathbf{1}_n = 0_r. \quad (2)$$

其中 $\boldsymbol{V}$ 是子空间,$\boldsymbol{P}$ 为映射矩阵,$\boldsymbol{X}$ 为模态,$\boldsymbol{Q}$ 为映射矩阵,$\boldsymbol{L}$ 是标签矩阵,$\boldsymbol{B}$ 为生成的哈希码矩阵. 给定

以上数据,目的是把不同模态数据的样本训练到不同的子空间,并整合到哈希码 B 中,进行进一步检索.

2.5　优化

由于假设问题情形有多种模态,所以为了描述方便,简化后用 X_1 和 X_2 直接代表 P_1 和 P_2,分别代表 2 种模态在子空间中生成的映射矩阵,式(2) 可整理为

$$\min \sum_{t=1}^{m} \beta(\boldsymbol{P}_t\boldsymbol{X}_t\boldsymbol{T}_t^{\mathrm{T}}\boldsymbol{P}_t^{\mathrm{T}} - 2\boldsymbol{P}_t\boldsymbol{X}_t\boldsymbol{L}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}} + \boldsymbol{Q}\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}) + \gamma(\boldsymbol{Q}\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}} - 2 - \boldsymbol{Q}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}) + \alpha(\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} - 2\boldsymbol{V}\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}),$$

$$\text{s.t. } \boldsymbol{B} \in \{-1,1\}^{r\times n}, \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = n\boldsymbol{I}, \boldsymbol{V}_{\boldsymbol{1}_n} = 0_r. \quad (3)$$

这一部分,使用 ADMM(Alternating Direction Method of Multipliers) 算法来解决优化问题,优化过程可分为 4 个步骤:

(i) 固定 $\boldsymbol{P}$、$\boldsymbol{Q}$、$\boldsymbol{V}$,求 $\boldsymbol{V}$. 通过固定其他变量并将式(3) 导数设置为 0,可得

$$\min \sum_{t=1}^{m} \gamma \| \boldsymbol{Q}\boldsymbol{L} - \boldsymbol{V} \|_F^2 + \| \boldsymbol{B} - \boldsymbol{V} \|_F^2,$$

$$\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = n\boldsymbol{I}, \boldsymbol{V}_{\boldsymbol{1}_r} = 0_r.$$

为了解决这个问题,首先将目标函数转换为具有 $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = n\boldsymbol{I}_r, \boldsymbol{V}_{\boldsymbol{1}_n} = 0_r$ 约束的矩阵迹的形式,可简化为

$$\max_{\boldsymbol{V}} \mathrm{tr}\left(\left(\alpha\boldsymbol{B} + \sum_{t=1}^{m} \gamma\boldsymbol{Q}^{(t)}\boldsymbol{l}^{(t)}\right)\boldsymbol{V}^{\mathrm{T}}\right),$$

$$\text{s.t. } \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = 0, \boldsymbol{V}_{\boldsymbol{1}_n} = 0_r.$$

为了简化表示,本文定义 $\boldsymbol{J} = \boldsymbol{I}_n - \boldsymbol{1}_n\boldsymbol{1}_n^{\mathrm{T}}/n$,$\boldsymbol{Z} = \alpha\boldsymbol{B} + \sum_{t=1}^{m} \gamma\boldsymbol{Q}^{(t)}\boldsymbol{l}^{(t)}$,为了找到 $\boldsymbol{V}$ 的解,首先执行 $\boldsymbol{Z}\boldsymbol{J}\boldsymbol{Z}^{\mathrm{T}}$ 的解

$$\boldsymbol{Z}\boldsymbol{J}\boldsymbol{Z}^{\mathrm{T}} = (\boldsymbol{W} \quad \ddot{\boldsymbol{W}}) \begin{pmatrix} \boldsymbol{\Omega} & 0 \\ 0 & 0 \end{pmatrix} (\boldsymbol{W} \quad \ddot{\boldsymbol{W}}),$$

其中 $\boldsymbol{\Omega} \in \mathbf{R}^{r'\times r'}$,$\boldsymbol{W} \in \mathbf{R}^{r'\times r'}$ 分别是正特征值和对应特征向量的对角矩阵,$\ddot{\boldsymbol{W}}$ 是 $r - r'$ 剩余的向量,正好对应于零特征值. r' 是 $\boldsymbol{Z}\boldsymbol{J}\boldsymbol{Z}^{\mathrm{T}}$ 的秩,通过对 $\ddot{\boldsymbol{W}}$ 的施密特正交化,能够轻易得到正交矩阵 $\overline{\boldsymbol{W}} \in \mathbf{R}^{r\times(r-r')}$,进一步地,定义 $\boldsymbol{O} = \boldsymbol{J}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{\Omega}^{-1/2}$,随机正交矩阵 $\overline{\boldsymbol{O}} \in \mathbf{R}^{r\times(r-r')}$,若 $r' = r$,则 $\overline{\boldsymbol{O}}$、$\overline{\boldsymbol{O}}$、$\ddot{\boldsymbol{W}}$ 为空,根据文献 [20],得到方程的最优解为

$$\boldsymbol{V} = \sqrt{n}(\boldsymbol{W} \quad \overline{\boldsymbol{W}})(\boldsymbol{O} \quad \overline{\boldsymbol{O}}).$$

(ii) 固定 $\boldsymbol{P}$、$\boldsymbol{B}$、$\boldsymbol{V}$,求 $\boldsymbol{Q}$. 当固定 $\boldsymbol{P}$、$\boldsymbol{B}$、$\boldsymbol{V}$ 时,

$$\min \sum_{t=1}^{m} \beta \| \boldsymbol{P}_t\boldsymbol{X}_t - \boldsymbol{Q}\boldsymbol{L} \|_F^2 + \gamma \| \boldsymbol{Q}\boldsymbol{L} - \boldsymbol{V} \|_F^2,$$

$$\text{s.t. } \boldsymbol{B}^{(t)} \in \{-1,1\}^{r\times n}, \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = n\boldsymbol{I}, \boldsymbol{V}_{\boldsymbol{1}_n} = 0_r. \quad (4)$$

在求解 $\boldsymbol{P}$ 的过程中,$\boldsymbol{P}_1$ 和 $\boldsymbol{P}_2$ 都要参与计算,可进一步整理为

$$\min \sum_{t=1}^{m} \beta(-2\boldsymbol{P}_1\boldsymbol{X}_1\boldsymbol{L}^{\mathrm{T}} + \boldsymbol{Q}\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}} - 2\boldsymbol{P}_2\boldsymbol{X}_2\boldsymbol{L}^{\mathrm{T}}) + \gamma(-2\boldsymbol{V}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}} + \boldsymbol{Q}\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}),$$

$$\text{s.t. } \boldsymbol{B}^{(t)} \in \{-1,1\}^{r\times n}, \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = n\boldsymbol{I}, \boldsymbol{V}_{\boldsymbol{1}_n} = 0_r.$$

将目标公式 $\boldsymbol{Q}$ 的倒数趋近于 0 得

$$\boldsymbol{Q} = \beta(\boldsymbol{P}_1\boldsymbol{X}_1 + \boldsymbol{P}_2\boldsymbol{X})\boldsymbol{L}^{\mathrm{T}}/(2\beta + \gamma)\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}.$$

(iii) 固定 $\boldsymbol{Q}$、$\boldsymbol{B}$、$\boldsymbol{V}$,求 $\boldsymbol{P}$. 当固定 $\boldsymbol{Q}$、$\boldsymbol{B}$、$\boldsymbol{V}$ 时,$\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$ 作为 2 种模态的数据都参与计算,公式整理为

$$\min \sum_{t=1}^{m} \beta \| \boldsymbol{P}_t\boldsymbol{X}_t - \boldsymbol{Q}\boldsymbol{L} \|_F^2, \quad (5)$$

进一步可更新为

$$\min \sum_{t=1}^{m} \boldsymbol{P}_1\boldsymbol{X}_1\boldsymbol{X}_1^{\mathrm{T}} + \boldsymbol{P}_2\boldsymbol{X}_2\boldsymbol{X}_2^{\mathrm{T}} - \boldsymbol{Q}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Q} - \boldsymbol{Q}\boldsymbol{L}\boldsymbol{X}_2^{F},$$

将目标函数 $\boldsymbol{P}$ 的导数趋近于 0,可得

$$\boldsymbol{P}_1 = \boldsymbol{Q}\boldsymbol{L}\boldsymbol{X}_1^{\mathrm{T}}/\boldsymbol{X}_1\boldsymbol{X}_1^{\mathrm{T}}, \boldsymbol{P}_2 = \boldsymbol{Q}\boldsymbol{L}\boldsymbol{X}_2^{\mathrm{T}}/\boldsymbol{X}_2\boldsymbol{X}_2^{\mathrm{T}}.$$

(iv) 固定 $\boldsymbol{Q}$、$\boldsymbol{P}$、$\boldsymbol{V}$,求 $\boldsymbol{B}$. 为了求解 $\boldsymbol{B}$,固定其余参数,式(1) 可转变为

$$\min \sum_{t=1}^{m} \alpha(\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} - 2\boldsymbol{V}\boldsymbol{B}^{\mathrm{T}}),$$

$$\text{s.t. } \boldsymbol{B}^{(t)} \in \{-1,1\}^{r\times n}. \quad (6)$$

因为 $\mathrm{tr}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}) =$ 常数,所以公式简化为

$$\max \sum_{t=1}^{m} (\boldsymbol{V}\boldsymbol{B}^{\mathrm{T}}), \text{s.t. } \boldsymbol{B}^{(t)} \in \{-1,1\}^{r\times n}.$$

由此可得

$$\boldsymbol{B} = \mathrm{sgn}(\alpha\boldsymbol{V}).$$

2.6　哈希函数学习

对于样本外的数据,要进行哈希函数学习,为了捕捉到数据之间的非线性特征,使用核回归作为哈希函数,将一些线性不可分的特征映射到高维线性可分,不同模态的数据可以映射成高维线性可分特征. 模态的哈希函数可定义为

$$\min_{\boldsymbol{P}_t} \| \boldsymbol{B} - \boldsymbol{P}_t\varphi(\boldsymbol{X}^{(t)}) \|_F^2 + \mu \| \boldsymbol{P}_t \|_F^2, \quad (7)$$

其中 $\boldsymbol{P}_t$ 为第 t 模态的哈希函数,$\varphi(\cdot)$ 是 RBF 核函数,形式为 $\varphi(\boldsymbol{X}^t) = \exp(-\| x^t - a_i \|^2/(2\sigma^2))$,$\{a_j\}_{j=1}^{q}$ 是随机从训练样本 $\boldsymbol{X}^{(t)}$ 中选取的 q 个锚向量,σ 是一个参数.

$$\boldsymbol{P}_t = \boldsymbol{B}\boldsymbol{X}^{(t)\mathrm{T}}(\boldsymbol{X}^{(t)}\boldsymbol{X}^{(t)\mathrm{T}} + \sigma\boldsymbol{I}^{(t)})^{-1}.$$

在给出查询 t 模态的查询序列 $\boldsymbol{D}(l)$ 的情况下，可以使用下面的哈希函数获取到相应的哈希码为

$$\boldsymbol{H}_t(\boldsymbol{D}^{(t)}) = \operatorname{sgn}(\boldsymbol{P}^{(t)}\varphi(\boldsymbol{D}^{(t)})). \tag{8}$$

2.7 收敛算法

根据以上步骤，交替更新所有参数，直到收敛或者达到最大迭代次数为止.

算法 1 优化算法

Input: 训练矩阵 $\boldsymbol{X}^{(t)}$，标签矩阵 $\boldsymbol{L}$，编码长度 r，最大迭代次数 T，参数 α、β、γ，

初始化: $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{V}$ 都是随机生成的矩阵，嵌入 $\boldsymbol{X}^{(t)}$ 和标签值映射到非线性子空间 $\boldsymbol{P}^{t}$ 中.

(i) 使用式(4) 更新 $\boldsymbol{V}$;

(ii) 使用式(5) 更新 $\boldsymbol{Q}$;

(iii) 使用式(6) 更新 $\boldsymbol{P}$;

(iv) 使用式(7) 更新 $\boldsymbol{B}$;

结束.

返回值: 哈希编码矩阵 $\boldsymbol{B}$，

使用式(8) 构造映射矩阵 $\boldsymbol{P}^{(t)}$;

返回哈希函数 $\boldsymbol{H}(\cdot)$.

3 实验部分

为了验证方法性能，在 4 个基准数据集上进行了实验，分别是 Wiki-pedia[21]、Nus-wide[22]、MIRFlikr-25k[23]，它们已在主流方法中被广泛采用. 在本节中，将本方法与 7 种相关方法进行比较，并分析相关性能.

3.1 Wiki-pedia

跨模态检索系统的评估需要使用拥有成对文本和图像的文档语料库 Wiki-pedia. Wiki 数据集的数据来源来自 Wiki 网站上的文章，囊括了 29 个类别，在每篇文本后都附属了相关的一张或多张图像，图像与文本是配对的，有着相应的类别标签，这些图像来自 Wiki commons. 最后，使用了其中标签数量最多的 10 种类别，共包含了 2 866 个文档，并进一步细化成了测试集和训练集.

3.2 Nus-wide

Nus-wide 数据集是 NUS´S 实验室创基于 Flickr 的网络图像数据，共含 269 648 幅图像和 Flickr 用户标注的 5 018 个唯一标签. Nus-wide 是最大的真实世界网络图像集，整个数据集有 81 种类别. 在 269 648 幅图像中，其中 161 789 幅为训练图片，107 859 幅为测试图片.

3.3 MIRFlikr-25k

MIRFlikr-25k 是从社交摄影网站 Flickr 上提取出的 25 000 幅图像，在收集图像过程时，尽量多收集了有关图像的信息，包括创建者信息和图像的标题等，并随之收集到了图像标签. 图像标签共有 2 种形式: 一种是从用户获取的原始标签，另一种是清除了原始标签后再次进行注释. 在 25 000 幅图片中，15 000 幅为训练图集，10 000 幅为测试集，在每 5 个图像划分时，前 3 个指定为训练图像，后 2 个为测试图像.

3.4 评价指标

在跨模态检索任务中，用 I(Image) 表示图像，T(Text) 表示文本，I 表示图像检索文本，T 表示文本检索图像，用 M_{AP}(Mean Average Precision) 评估跨模态任务的准确度，给定一个任务，A_P 可定义成

$$A_P(r) = \frac{1}{n_r}\sum_{i=1}^{n_q} P(i)\,\delta(i), \tag{9}$$

其中 n_r 是检索真实世界中的相关实例，n_q 是检索出来结果的大小，$\delta(i) = 1$ 表示目标检索与结果一致，反之，$\delta(i) = 0$ 表示无关. $P(i)$ 是 $\operatorname{top}(i)$ 的实例检索的准确度，因此，M_{AP} 可以写成

$$M_{AP} = \frac{1}{N}\sum_{r=1}^{N} A_P(r), \tag{10}$$

其中 N 是查询序列集，M_{AP} 值越大表示方法性能越好.

3.5 对比方法及简介

将本文方法与 7 个相关子空间检索方法进行了比较.

CMFH 通过矩阵因子分解方法从不同形式中获得模态的潜在语义特征，获得多个实例的公共子空间，并生成具有仿射投影的统一哈希码，是一种保留了模态相似性的哈希学习方法.

SMFH[24] 通过图形正则化保留多模态原始特征之间的相似性，同时，将可用的语义标签合并到学习过程中，以此提高哈希码质量.

SCM 是语义相关最大化算法，用标签重构相似矩阵，将语义标签无缝集成到大规模数据建模的哈希学习过程中，以进行跨模态哈希函数学习.

DCH[25] 通过联合矩阵分解学习统一哈希空间，生成统一二进制代码，同时保持离散约束和类标签的判别性，提高了所生成哈希码的判别性.

SePH 首先通过使用标签信息构造相似性矩阵来学习哈希码，再通过逻辑回归学习哈希函数.

FSH[26] 对不同模态之间的融合相似性进行建模，并将其嵌入到公共汉明空间中，再生成相应的哈

希码.

SCRATCH [27] 结合矩阵分解和语义嵌入保持模式内和模式间的相似性,将交叉模式特征内核化,再利用矩阵分解将内核化的特征提取为语义,并离散地生成二进制码.

以上 7 种方法中,CMFH 和 FSH 属于非监督方法,其他 4 种属于有监督方法,按照论文作者提供的代码和论文建议参数来实现所有的基准方法,通过相同的评估方法和相同数据集计算了不同方法的 M_{AP}. 为了更好评估性能,通过 10 次平均实验结果获得 M_{AP} 值,实验均在 Intel(R) Core (TM) i7-4790 CPU 3.6-GHZ 8G 内存,64 位操作系统上进行,表 1 显示了实验的迭代收敛细节,表 2 展示了与不同方法的训练时间比较.

本文方法在 NUS-wide 和 MIRFlickr 上均取得了较好结果,而 Wiki 的数据集效果就没那么好. 这是无法改进的,原因之一是模态的文本特征维数较低.

NUS-wide 和 MIRFlickr 数据集中每一个方法的结果如表 2 所示,精度-召回率曲线如图 1 ~ 图 3 所示. 本文的方法在 T2I 和 I2T 的效果对比的 5 种方法中,M_{AP} 值更高.

表 1　基准方法的 M_{AP} 值

任务	方法	WIKI				MIRFlicker-25k				NUS-WIDE			
		16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
I2T	CMFH	0.261 4	0.235 7	0.243 1	0.253 2	0.585 2	0.584 9	0.584 8	0.585 3	0.391 7	0.390 5	0.394 7	0.393 2
	SMFH	0.245 9	0.256 1	0.256 1	0.267 4	0.628 0	0.634 5	0.638 5	0.649 0	0.540 8	0.556 4	0.567 5	0.5678
	SCM	0.234 1	0.241 0	0.245 6	0.257 5	0.628 0	0.634 5	0.638 5	0.649 0	0.540 8	0.556 1	0.560 8	0.561 2
	DCH	0.342 1	0.372 1	0.382 3	0.382 4	0.671 0	0.671 6	0.681 7	0.689 0	0.594 1	0.574 1	0.598 8	0.614 5
	SePH	0.277 1	0.301 1	0.309 8	0.320 1	0.670 8	0.676 6	0.681 5	0.683 2	0.584 0	0.600 2	0.603 1	0.609 1
	FSH	0.243 1	0.259 1	0.267 3	0.274 5	0.617 4	0.623 1	0.625 4	0.631 5	0.500 2	0.510 1	0.520 4	0.523 4
	SCRATCH	0.351 4	0.381 7	0.385 9	0.382 2	0.710 1	0.709 3	0.721 4	0.734 5	0.620 4	0.641 5	0.653 2	0.662 4
	Our	0.371 0	0.377 5	0.389 4	0.400 2	0.714 7	0.726 4	0.737 2	0.742 4	0.633 4	0.651 1	0.661 7	0.669 4
T2I	CMFH	0.490 1	0.520 1	0.531 2	0.541 4	0.587 2	0.589 9	0.601 2	0.604 1	0.401 2	0.412 1	0.402 1	0.412 3
	SMFH	0.471 4	0.505 6	0.518 9	0.489 7	0.574 1	0.552 1	0.561 4	0.558 9	0.432 1	0.424 8	0.462 4	0.492 4
	SCM	0.232 4	0.2514	0.248 2	0.262 2	0.624 1	0.631 2	0.630 6	0.625 9	0.524 2	0.523 4	0.525 4	0.526 8
	DCH	0.701 2	0.715 8	0.720 7	0.718 1	0.749 8	0.748 7	0.764 5	0.789 7	0.720 1	0.719 8	0.715 7	0.727 5
	SePH	0.629 0	0.649 7	0.670 1	0.672 3	0.723 1	0.730 1	0.734 2	0.738 5	0.670 1	0.680 3	0.691 2	0.700 2
	FSH	0.240 2	0.256 4	0.249 5	0.263 2	0.615 9	0.616 8	0.616 7	0.629 5	0.509 8	0.520 4	0.532 4	0.521 4
	SCRATCH	0.740 1	0.743 4	0.750 8	0.760 2	0.774 2	0.785 1	0.797 9	0.810 2	0.730 4	0.764 4	0.772 4	0.783 2
	Our	0.742 6	0.748 2	0.759 4	0.757 9	0.783 4	0.809 4	0.823 9	0.832 7	0.748 8	0.771 4	0.783 7	0.790 7

表 2　不同 nus 样本值的训练时间　　s

方法	大小					
	2 000	5 000	10 000	15 000	20 000	25 000
CMFH	32.80	91.20	206.50	345.20	960.50	654.10
SMFH	2.25	4.67	8.64	12.414	16.52	18.60
SCM	13.10	16.57	18.90	26.77	39.89	32.01
DCH	1.97	6.05	9.87	13.56	17.58	21.58
SePH	186.52	531.56	1 985.50	4 754.40	8 421.00	13 415.00
FSH	13.25	32.89	64.58	100.42	142.20	185.20
SCRACH	1.88	3.65	6.45	10.21	14.58	17.24
OUR	1.328 0	1.382 4	1.703 4	2.042 2	2.258 0	2.710 0

3.6　收敛性分析

通过在 MLRFlickr-25k 和 NUS-wide-25K 数据集上使用 32 bit 哈希码来评估其收敛性,本文方法的收敛特性如图 4 所示. 由图 4 可知,迭代次数在 16 次左右就开始收敛,因此将实验的迭代次数设置为 16,这足以获得最佳效果.

3.7　时间成本分析

表 2 显示了不同方法在 NUS-wide 上的训练时间,把 NUS-wide 的数据样本从 2 000 个样本增加到 25 000个样本,哈希码的长度设置为 32 位. 从表 2 能看出本文的方法是最快速的一个,因此可以说实现了更快速地检索.

3.8　参数敏感性分析

下面对 α、β、γ 等参数的敏感性分析,通过固定其他参数值来记录唯一参数的 M_{AP} 值,图 5 ~ 图 7 记录了 M_{AP} 值的变化.

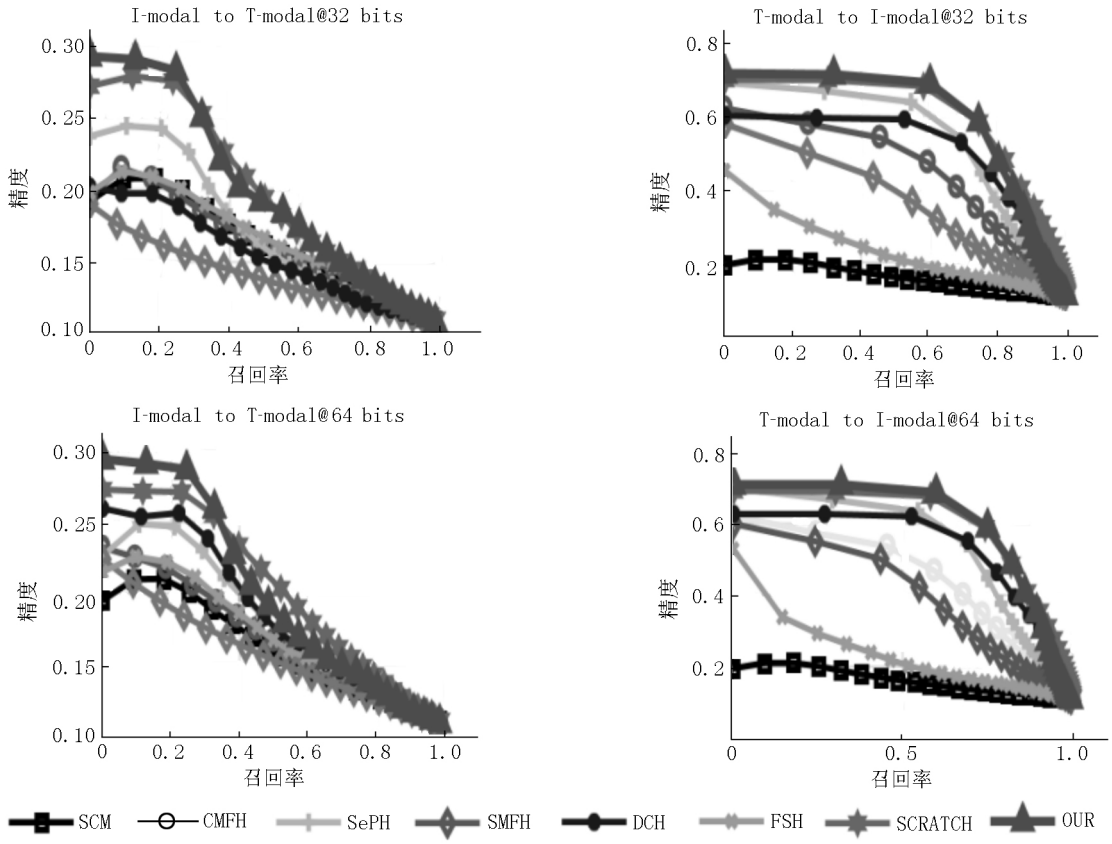

图 1　Wiki 长度变化的精度-召回率值

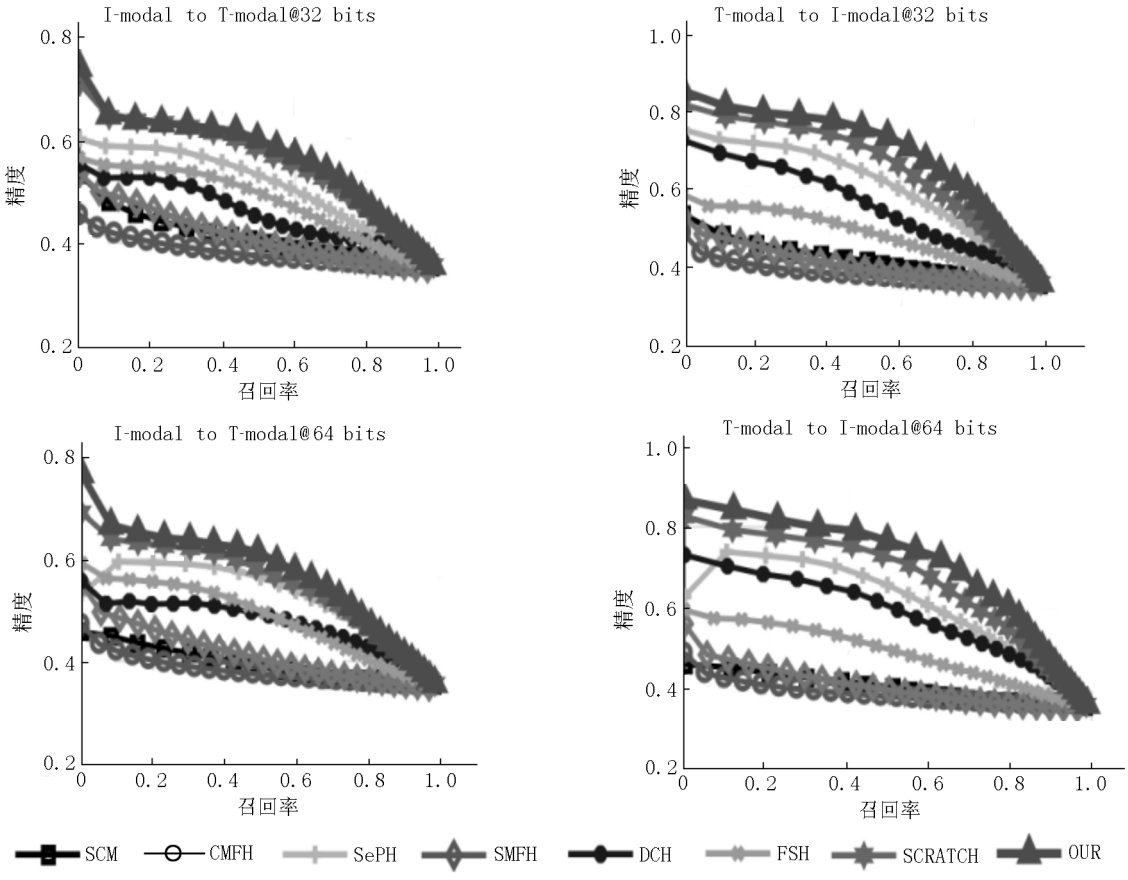

图 2　Nus-wide 长度变化的精度-召回率值

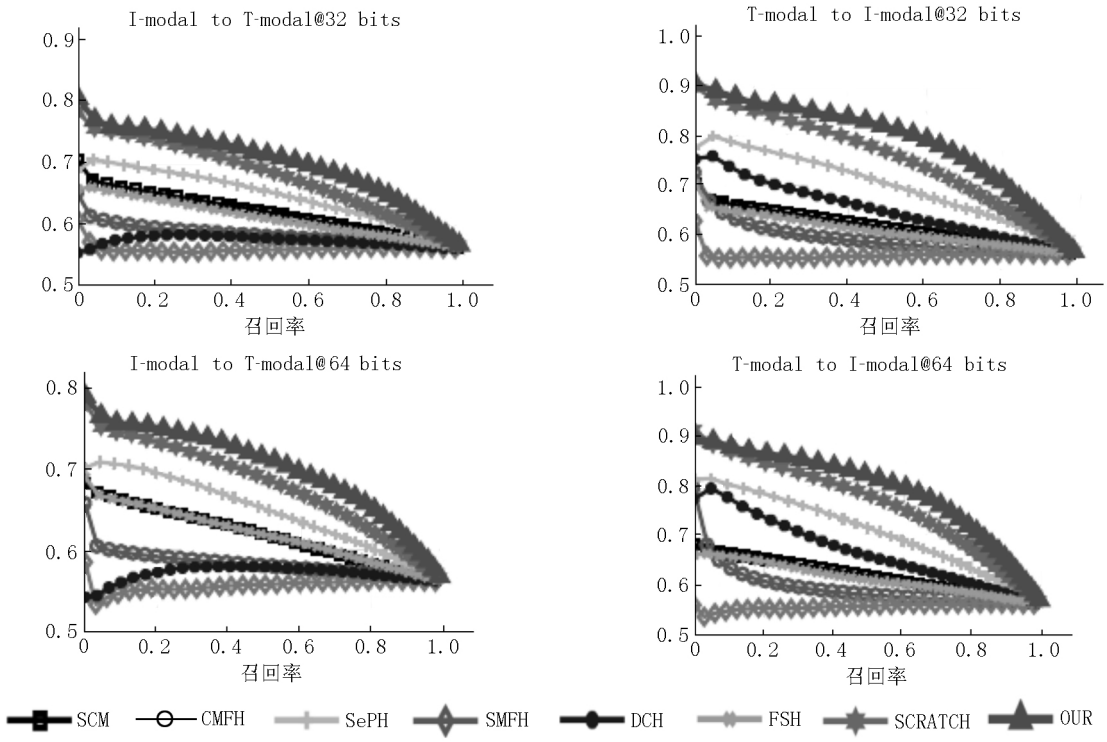

图 3　MIRFlickr 长度变化的精度-召回率值

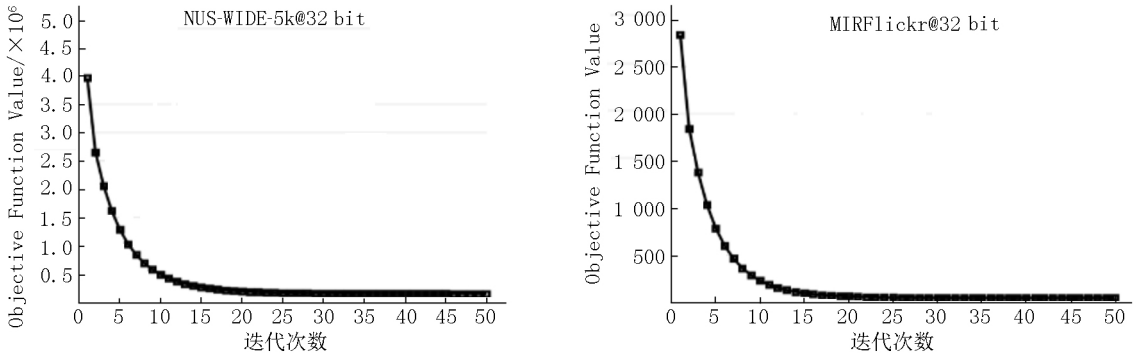

图 4　收敛性分析

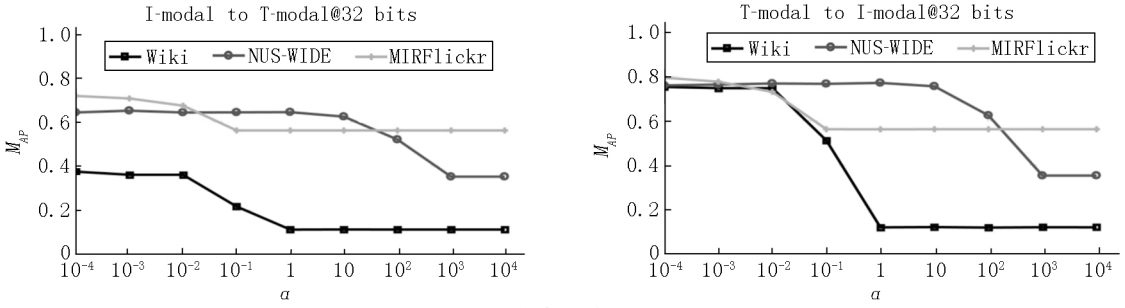

图 5　参数 α 与 M_{AP} 值

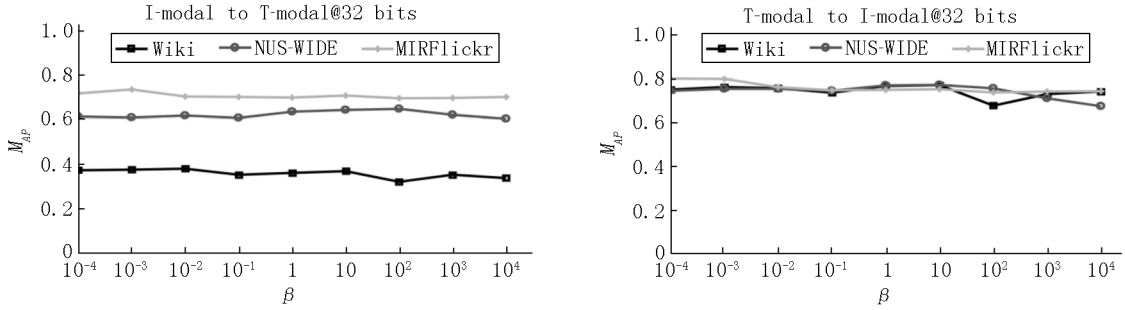

图 6　参数 β 与 M_{AP} 值

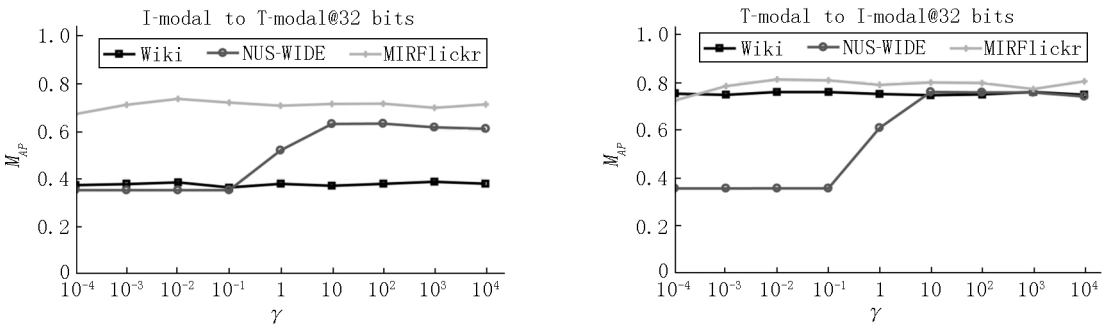

图7 参数 γ 与 M_{AP} 值

α 控制着哈希码和映射矩阵的生成,从图 5 明显可以看出,当 α 变大时,M_{AP} 值变得稳定并开始逐渐下降,特别是在 α 增加至 10 后,这是因为本文的方法在很大程度上并不受限于其他影响. 在实验中,将 Wiki 数据集的 α 设置为 1×10^{-5},NUS-wide 设为 1,MIRFlickr 设为 1×10^{-4}以获得最好效果.

从图 6 可明显看出,随着 β 值的增加,除了 NUS-wide 外,其他数据集的 M_{AP} 基本保持平稳,β 控制着子空间的生成,而子空间的生成又受映射矩阵 Q、V 标签矩阵 L 的影响,因此仅随着 β 值的变化,M_{AP} 值并无明显递增递减变化. 为了获得最好实验效果,将 Wiki 数据集的 β 设置为 1×10^{-5},NUS-wide 设为 10, MIRFlickr 设为 1×10^{-2}.

由图 7 可以看出,随着 γ 值的增加,NUS-wide 的 $I\rightarrow T$ 和 $T\rightarrow I$ 的 M_{AP} 值有相当明显的提升,而 MIRFlickr、Wiki 则基本保持稳定不变. 这是因为 γ 值控制着映射矩阵 Q 和子空间 V 的学习,当 γ 增加时,即子空间 V 的生成质量依赖于 Q 和 V,NUS-wide 的 M_{AP} 随之增加也是合理现象. 为了获得最好的效果,设置 NUS-wide 的 γ 值为 100,Wiki 为 10,MIRFlickr 为 1×10^{-3}.

4 结论

本文提出了一种标签嵌入的跨模态离散哈希检索方法,本方法将数据信息和标签信息嵌入到公共空间中,既利用了标签信息又在最大程度上获得了语义和标签之间的语义相关性;为了提高计算效率,在优化过程中未使用大规模 $n\times n$ 矩阵,这不仅保证了检索准确率,而且又能够在较大程度上减少计算成本. 在 3 个数据集上进行的实验也证明了该方法的优越性,结果表明检索准确率优于目前 7 种方法. 下一步工作重点将放在子空间跨模态哈希检索上,以实现效率更高、准确率更高的检索效果.

5 参考文献

[1] 彭宇新,綦金玮,黄鑫,等. Current research status and prospects on Multimedia content understanding [J]. 计算机研究与发展,2019,56(1):183-208.

[2] Fei Wu,Zhou Yu,Yang Yi,et al. Sparse multi-modal mashing [J]. IEEE Transactions on Multimedia, 2014, 16(2):427-439.

[3] Song Jingkuan, Yang Yang, Yang Yi, et al. Inter-media hashing for large-scale retrieval from heterogeneous data sources [EB/OL]. [2020-06-15]. https://doi.org/10.1145/2463676.2465274.

[4] Long Mingsheng,Cao Yue,Wang Jianmin,et al. Composite correlation quantization for efficient multimodal retrieval [EB/OL]. [2020-06-12]. https://doi.org/10.1145/2911451.2911493.

[5] Liu Hong,Ji Rongrong,Wu Yongjian,et al. Cross-modality binary code learning via fusion similarity hashing [EB/OL]. [2020-06-18]. 10.1109/CVPR.2017.672.

[6] Zhou Jile,Ding Guiguang,Guo Yuchen,et al. Latent semantic sparse hashing for cross-modal similarity search [EB/OL]. [2020-06-19]. https://dl.acm.org/doi/10.1145/2600428.2609610.

[7] Ding Guiguang,Guo Yuchen,Zhou Jile,et al. Collective matrix factorization hashing for multi-modal data [EB/OL]. [2020-06-19]. https://doi.org/10.1109/CVPR.2014.267.

[8] Fei Wu,Zhou Yu,Yang Yang,et al. Sparse multi-modal hashing [EB/OL]. [2020-06-19]. https://ieeexplore.ieee.org/document/6665155.

[9] Bronstein M,Bronstein A M,Michel F,et al. Data fusion through cross-modality metric learning using similarity-sensitive hashing [EB/OL]. [2020-05-19]. https://ieeexplore.ieee.org/document/5539928.

[10] Lin Zijia,Ding Guiguang,Hu Mingqing,et al. Semantics-preserving hash-ing for cross-view retrieval [EB/OL]. [2020-05-13]. https://ieeexplore.ieee.org/document/7299011?arnumber=7299011.

[11] Zhang Dongqing,Li Wujun. Large-scale supervised multi-modal hash-ing with semantic correlation maximization [EB/OL]. [2020-05-09]. https://cs. nju. edu. cn/lwj/paper/AAAI14_SCM. pdf.

[12] Kan Meina,Shan Shiguang,Zhang Haihong,et a1. Multi-view discriminant analysis [EB/OL]. [2020-06-19]. https://dl. acm. org/doi/10. 1007/978-3-642-33718-5_58.

[13] Wang Kaiyue,He Ran,Wang Wei,et al. Learning coupled feature spaces for cross-modal matching [EB/OL]. [2020-06-19]. https://dl. acm. org/doi/10. 1109/ICCV. 2013. 261.

[14] 费伦科,秦建阳,滕少华,等. 近似最近邻大数据检索哈希散列方法综述 [J]. 广东工业大学学报,2020,37(3):10-22.

[15] Gong Yunchao,Ke Qifa,ISARD M,et al. A multi-view embedding space for modeling internet images,tags,and their semantics [J]. International Journal of Computer Vision,2014,106(2):210-233.

[16] Ranjan V,Rasiwasia N,Jawahar C V. Multi-label cross-modal retrieval [EB/OL]. [2020-05-11]. https://ieeexplore. ieee. org/document/7410823.

[17] Zhou Jile,Ding Guiguang,Guo Yuchen. Latent semantic sparse hashing for cross-modal similarity search [EB/OL]. [2020-05-12]. https://dblp. org/rec/conf/sigir/ZhouDG14. html.

[18] Fei Wu,Zhou Yu,Yang Yang,et al. Sparse multi-modal hashing [J]. IEEE Trans Multimedia,2014,16(2):427-439.

[19] Wang Di,Gao Xinbo,Wang Xiumei,et al. Semantic topic multi-modal hashing for cross-media retrieval [EB/OL]. [2020-05-12]. https://dl. acm. org/doi/10. 5555/2832747. 2832791.

[20] Rasiwasia N,Pereira J C,Coviello E,et al. A new approach to cross-modal multi-media retrieval [EB/OL]. [2020-05-17]. https://dl. acm. org/doi/10. 1145/1873951. 1873987.

[21] Liu Wei,Mu Cun,Kumar S,et al. Discrete graph hashing [EB/OL]. [2020-07-19]. https://dl. acm. org/doi/10. 1. 1. 673. 2750.

[22] Chua T S,Tang Jinhui,Hong Richang,et al. NUS-WIDE: a real-world web image database from national university of Singapore [EB/OL]. [2020-05-10]. https://dl. acm. org/doi/10. 1145/1646396. 1646452.

[23] Huiskes M J,Lew M S. The MIRflickr retrieval evaluation [EB/OL]. [2020-05-14]. https://dl. acm. org/doi/10. 1145/1460096. 1460104.

[24] Liu Hong,Ji Rongrong,Wu Yongjian,et al. Supervised matrix factorization for cross modality hashing [EB/OL]. [2020-05-14]. https://dl. acm. org/doi/10. 5555/3060832. 3060868.

[25] Xu Xing,Shen Fumin,Yang Yang,et al. Learning discriminative binary codes for large-scale cross-modal retrieval [EB/OL]. [2020-05-16]. https://dl. acm. org/doi/10. 1109/TIP. 2017. 2676345.

[26] Liu Hong,Ji Rongrong,Wu Yongjian,et al. Cross-modality binary code learning via fusion similarity hashing [EB/OL]. [2020-05-10]. https://ieeexplore. ieee. org/document/8100155/citations#citations.

[27] Chen Zhenduo,Li Chuanxiang,Luo Xin,et al. SCRATCH: a scalable discrete matrix factorization hashing framework for cross-modal retrieval [EB/OL]. [2020-05-19]. https://dl. acm. org/doi/10. 1109/TCSVT. 2019. 2911359.

The Cross-Modal Discrete Hash Learning of Tag Embedding Subspace

TENG Shaohua[1],GUO Lanjun[1],ZHANG Wei[1],TENG Luyao[2]

(1. School of Computers,Guangdong University of Technology,Guangzhou Guamgdong 510006,China;
2. The Centre for Applied Informatics,Victoria University,Melbourne Victoria 3011,Australia)

Abstract: Because supervised cross-modal hash retrieval has still problems of high computational cost and low accuracy of retrieval,a cross-modal discrete hash learning method for tag embedding subspace is proposed,which embeds data information and tag information into the common subspace at the same time. The common subspace is approximated by semantic features with tag information,and discrete hash codes with low slack are generated,which greatly reduces the computational cost and quickly generates a common subspace with rich semantics. This method is compared with 3 standard data sets,and the results show that the retrieval accuracy is better than the compared method.

Key words: tag embedding; subspace; discrete hash

(责任编辑: 冉小晓)