doi:10.3772/j.issn.2095-915x.2021.04.003

基于异构数据特征向量的图文检索方法研究

骆有隆[1,2] 朱卉钰[1] 梁松宇[3] 张腾[3]

1. 武汉理工大学管理学院 武汉 430070；
2. 富媒体数字出版内容组织与知识服务重点实验室 北京 100038；
3. 武汉理工大学计算机学院 武汉 430070

开放科学（资源服务）标识码（OSID）

摘要：[目的 / 意义] 互联网中存在的信息多数是以文本、图像、视频等相结合的形式，即所谓的跨媒体数据。传统的检索方式包括以文检文、以图检图，它们已不能满足如今信息检索的需求，所以跨媒体数据检索应运而生，跨媒体数据检索的研究具有非常重要的研究意义及应用价值。[方法 / 过程] 本文利用 Doc2vec 模型和 VGG16 模型分别处理文本和图片数据得到各自的特征值，根据特征值的相对语义距离对文本和图片数据进行匹配，从而实现自动的给文字配图片或者为图片选择标签。[结果 / 结论] 本文提出的方法能够通过提取不同形式的数据特征对数据进行融合，有效提升了图片和文本数据自动匹配的命中率，实现了基于异构数据特征的图文检索。

关键词：跨媒体；图文匹配；检索；融合

中图分类号：G35

Research on Image and Text Retrieval Method Based on Feature Vector of Heterogeneous Data

LUO Youlong[1,2] ZHU Huiyu[1] LIANG Songyu[3] ZHANG Teng[3]

1. School of Management, Wuhan University of Technology, Wuhan 430063, China;
2. Key Laboratory of Content Organization and Knowledge Service for Rich Media Digital Publishing, Beijing 100038, China;
3. School of Computer Science, Wuhan University of Technology, Wuhan 430063, China

基金项目 富媒体数字出版内容组织与知识服务重点实验室开放基金“云计算环境下面向富媒体文本和图片数据的实体关系抽取与推理研究”(ZD2020-09/05)，教育部产学合作协同育人项目“智能制造仿真系统的研究与设计”（201901083006）。

作者简介 骆有隆（1974-），博士，副教授，研究方向为文本挖掘、社会网络分析、科技情报分析，E-mail：luoyoulong2006@aliyun.com；朱卉钰（1996-），硕士生，研究方向为财务管理、信息处理；梁松宇（1995-），硕士生，研究方向为文本挖掘、信息处理；张腾（1995-），硕士生，研究方向为信息系统、网络搜索。

引用格式 骆有隆 , 朱卉钰 , 梁松宇 , 等 . 基于异构数据特征向量的图文检索方法研究 [J]. 情报工程 , 2021, 7(4): 27-39.

doi:10.3772/j.issn.2095-915x.2021.04.003

Abstract: [Objective/ Significance] Most of the information existing on the Internet is in the form of a combination of text, images, videos, etc., which is the so-called cross-media data. Traditional retrieval methods include document retrieval and image retrieval. They can no longer meet the needs of today's information retrieval. Therefore, cross-media data retrieval has emerged. [Methods/Process] The research on cross-media data retrieval has very important research significance and application value. This paper uses the Doc2vec model and the VGG16 model to process the feature values of text and image data respectively, and uses the vector similarity distance metric to match text and image data. [Results /Conclusions] The experimental results show that the method proposed in this paper can fuse data in different modal forms, effectively improve the hit rate of automatic matching of graphic and text data, and realise a graphical text retrieval method based on feature vectors of heterogeneous data.
Keywords: Cross-media; image-text matching; retrieval; integration

引言

随着信息技术的高速发展，信息形式呈现多样化的趋势，如文本、图片、视频、音频等。人们对信息的获取、传播、分析和处理也逐渐从单一媒体形式转变为多种媒体形式。传统获取信息的方式是从海量数据中过滤出用户需要的信息，再将过滤结果返回用户，但无法获取和解析这些结果间存在的内在关系，缺乏从语义角度去挖掘隐藏在大数据深层次规律和知识的能力，用户只能从结果中自己去理解和筛选信息。跨媒体数据融合通过挖掘出深层次的不同媒体数据间的语义关联实现数据融合，以最大化满足用户的信息需求，从而提供更加优质的信息服务。因此，研究跨媒体数据之间的融合具有重要的实际意义和应用价值。

近年来，图文匹配在人工智能、机器学习等领域中逐渐兴起。为了给文本选取最适合的图像，在过去通常借助人工搜索的方式，通过判断文本内容来从海量图像中筛选出匹配的图像集合，但这会耗费人类大量的时间和精力。而图文匹配系统，则能大大减轻这种负担。

图文匹配必须同时关注文本和图像这两个不同模态下的数据，但因为不同模态数据存在于不同的特征空间且拥有不同的表征形式和分布特征，所以不能直接度量不同模态数据之间的相似性。现有的跨媒体融合着重解决不同模态数据间的语义关联问题，而较少关注到可以通过比较底层特征来实现跨媒体数据融合。本文通过分析跨媒体数据融合的一般过程，从借助深度学习技术获取文本特征和图像特征入手，解决跨媒体数据的一致性表达问题，从而实现图文相互检索的目的，为跨媒体数据融合之后的实体识别和关系抽取做技术储备。

本文第 2 节描述之前的相关工作。第 3 节正式介绍了基于异构数据融合的图文检索模型。第 4 节通过实验检验图文匹配效果。第 5 节对文章内容进行总结。

1 研究现状

现有的图文匹配方法主要有两大类：第一类方法是将不同模态的数据映射到同一语义空间中，然后在该空间中对二者进行语义匹配。较早对该方法进行研究的是 Rasiwasia 等 [1]，他们将典型相关性分析方法和语义匹配方法融合，建立数据关联结构实现了跨媒体数据在子空间上的映射。国内较早关注这方面的李向阳等 [2]，

则提出了基于内容相关性的跨媒体检索方法，取得了不错的效果。Socher 等 [3] 提出通过语义依赖树递归神经网络（SDT-RNN）将语句数据映射到图像的语义空间中，然后通过图像的语义空间上的距离来度量图片与语句之间的关联。Wang 等 [4] 使用卷积神经网络（Convolutional neural network，CNN）提取图片特征，使用 WCNN 提取语句特征，然后将二者特征映射到同一公共空间中，形成相关或不相关的图像 - 文本对，并使用一对多学习策略对其进行学习。Karpathy 等 [5] 做了精度更进一步的工作，他们将语句的类型依赖关系树与图片的对象映射到同一公共空间，然后再对二者进行度量以确定其关联度。

另一类方法主要是利用像典型相关分析（Canonical correlation analysis，CAA），深度学习等方法来挖掘语句与图片的语义关联。王晓宇 [6] 运用基于词袋模型的逻辑回归分类器进行图文结合，然后进行了基于二类型文本相似度计算交叉融合的配图推荐和基于反向传播神经网络的配图推荐，也取得了不错的效果。Hodosh 等 [7] 提出了核典型相关分析方法（Kernel canonical correlation analysis，KCCA），利用该方法来寻找语句和图片共享的特征空间。Vendrov 等 [8] 提出了 Gated recurrent unit（GRU）方法来提取句子的特征，他们将句子和图片的关系看作是一种偏序关系，并在此偏序关系的基础上进行图文关联性的度量。Ma 等 [9] 将图文分别在词、片段和语句三种级别使用多通道卷积神经网络（Multimodal convolutional neural networks, m-CNNs）进行匹配，实现了图文在局部与全局的混合匹配。

跨媒体数据融合的一般过程 [10] 为：首先分别提取不同模态数据的特征信息，此时的特征信息是异构的，无法直接计算它们之间的距离；然后通过某种映射机制将处于异构空间中的跨媒体数据特征映射到同构空间中，就可以使用公共距离计算公式去度量异构数据间相似性。过程中每一环节的效果对最终结果都有重要影响。

1.1 文本特征提取

文本特征提取技术属于自然语言处理问题，是一种从文本中提取出关键信息，然后用提取出的关键信息表示文本的方法。现今主要包括三种研究方法：基于统计的方法、基于词向量的方法和基于主题模型的方法 [10]，如图 1 所示。

（1）基于统计的方法 [11-13] 中最具代表性的方法是词频 – 逆向文档频率（Term Frequency-Inverse Document Frequency，TF-IDF），它的思路相对简单，通过统计文档内和所有文档中词频和词间关系来表征文本，再根据特征项之间所含信息量的多少来衡量该特征项的重要性。

（2）基于主题模型的方法 [14,15] 可解释性强，通过主题建模技术产生抽象主题统计模型方法，但其受文档长度的限制，代表方法为潜在狄利克雷分布（Latent Dirichlet Allocation，LDA），它将每一篇文档视为一个词频向量，通过对每篇文档抽取主题得到的关于所有文档的概率分布。

（3）基于词向量的方法 [16] 应用广泛，通常采用分布式表示方法将词表示为一个定长且连续的稠密向量。词向量的出现使得词与词之

doi:10.3772/j.issn.2095-915x.2021.04.003

间有了距离的概念，让语义相近的词在距离上更近，相对而言是效果较好的方法。Word2vec 作为其中应用较为广泛的代表，它主要依据是词上下文之间的关系训练生成的词向量包含语义信息且能将其作为中间结果参与其他计算。然而 Word2vec 模型通过对词向量进行平均处理，忽略了词之间的排列顺序对情感分析的影响，即 Word2vec 只是基于词的维度进行“语义分析”的，而不具有上下文的“语义分析”能力。因此 Quoc Le 和 Tomas Mikolov 于 2014 年提出了一种处理可变长度文本的总结性方法，即 Doc2Vec 模型。该模型在 Word2Vec 模型基础上增加一个段落向量，涵盖了段落级的上下文语义信息，更加符合本文研究的需求。

研究方法		优点	缺点
基于统计的方法	TF、DF、TF-IDF	用词出现频率来衡量其重要性，易于理解，突出重要单词，实现难度小	仅考虑词频不能有效反映词汇的重要程度以及分布情况
	信息增益（IG）	选择最大的信息增益属性进行划分，能很好地衡量出文档分类效果	其结果会偏向取值较多的特征
	互信息（MI）	依据每个特征项与各个类别的平均关联程度选择出一定量的特征项	容易选出生僻词甚至是噪音词，这些词的 MI 大但是携带较少的类别信息
基于词向量的方法	Word2vec	包含语义信息，一定程度上解决了语义鸿沟问题，高效	需要大量训练语料，具有不可解释性
基于主题模型的方法	LDA	找到没有共同词的两篇文档之间的潜在联系、挖掘出文档中的潜在词	受文档长度限制，短文本无法挖掘出有用的词
	LDA2Vec	吸收了 Word2Vec 局部预测和 LDA 全局预测的优点，在单词和文档上构建表示	训练时间长，同样受文档长度影响

图 **1** 文本特征提取技术

1.2 图像特征提取

图像特征提取是图像处理中最初级的运算，指使用计算机提取图像中属于特征性信息的方法及过程，主要分为两种方法：视觉底层特征提取和神经网络的方法 [10]，如图 2 所示。

（1）视觉底层特征涵盖了图片颜色、形状、纹理、空间关系等各个方面的特征 [17,18]，颜色特征是在符合人眼视觉感知的 HSV 颜色空间下提取的，包括饱和度、亮度等指标；形状特征一般有两种表示方法，一种是轮廓特征，另一种是区域特征。图像的轮廓特征主要针对物体的外边界，而图像的区域特征则关系到整个形状区域；纹理特征主要采用滤波器等方法来提取图像的局部特征；空间关系特征主要是指从图像中分割出来的多个目标之间的相互的空间位置或相对方向关系，实际应用中它不能有效地表达场景信息，通常需要与其他特征相结合。

（2）神经网络方法由于其高精度而受到人

们关注，其中最为广泛使用的当属卷积神经网络[19]（Convolutional Neural Networks, CNN），它在对图像语义处理上与其它视觉底层特征相比能力更强。卷积神经网络由一系列的卷积层、池化层、激活层和全连接层组成，理论上，随着网络层次的增加，网络模型可以提取更复杂的特征，从而能够取得更好的效果，但实际上会出现梯度消失、网络“退化”等问题。而在ICLR2015会议中提出的VGG16模型[20]在图像识别领域准确率能达到极高的水平。

研究方法			优点	缺点
视觉底层特征提取	颜色特征	RGB、HIS、HSV、CMYK	对图片大小、方向都不敏感，在一些情况下表现出相当强的鲁棒性	只用颜色特性很难完整而准确地描述一个具体物体，不适合人的视觉特点
	形状特征	SIFT特征提取算法	适用于不变特征的提取，提取的特征比较稳定，具有较好的区分性，信息量丰富，适用于海量特征数据处理	特征描述符的维数过大以及耗时过长，实时性不高，对边缘光滑的目标无法准确提取特征点
	纹理特征	灰度共生矩阵、随机场模型方法、小波变换	反映了图像中的同质现象，具有旋转不变性以及良好的抗噪性	容易受分辨率变化、光照变化等的影响
	空间关系特征	姿态估计	可加强对图像内容的描述区分能力	对图像目标的旋转、图像目标的反转以及尺度变化都较为敏感
神经网络方法		卷积神经网络CNN	无需手动选取特征，对高维数据的处理无压力，尤其对图像这种高度非结构化、分布复杂的数据具有很强的处理能力	需要大量的训练数据、对计算计性能要求高

图 2 图像特征提取技术

1.3 图文匹配

传统的匹配任务大多是在同一语义空间下进行，比如搜索引擎中网页匹配就是通过用户输入文本分词后得到的关键字来进行直接匹配获取的，甚至大多数的图片搜索引擎也是通过对用户输入图片进行关键词标注，或为图片生成描述文本，最后就将其转换为文本关键词匹配来得到近似图片[21]。

图文匹配在推荐系统、机器学习等领域中都起着不可忽视的作用。Yan等[22]提出使用深度网络来表示图像和文本，然后借助带有深度典型关联分析的联合隐藏空间学习来解决图文匹配。Ma等[23]通过在图文匹配任务中构建图像特征抽取网络并提出使用预训练的卷积神经网络来初始化。Wang等[24]则基于深度学习方法来构建了一个图文联合隐藏空间学习的一般框架，还提出了图像和文本都存在各自的结构保持约束以及图文匹配的双向排名约束。

2 基于异构数据融合的图文检索模型

由于异构数据通常以多种形态表示，例如文本、视频、音频等，但是不同形态的数据可能表达了同一个主题，也就是说他们的高层语义十分相似，而将这些不同模态之间的数据进行聚

doi:10.3772/j.issn.2095-915x.2021.04.003

类，建立起异构数据之间的联系是非常困难的。故本文提出的异构数据融合的图文检索内容重点在于从文本上下文联系与图像自身的底层语义特征出发，通过 Doc2vec 模型提取文本特征向量以及 VGG16 模型提取图片特征向量，提出一种基于异构数据特征向量的图文检索模型。

本文提出的基于异构数据特征向量的图文检索模型主要为了实现在给定单独图片的情况下，从图像数据库中挑选出与其最匹配的文本描述，即以图配文。

2.1 异构数据融合模型

Word2vec 模型能够捕获词汇上下文语义信息并得出两个词语间的相似程度，主要包括两种训练模型 [25]——连续词袋模型（Continuous Bag-of-Words，CBOW）和跳字模型（Skip-gram）。其中 CBOW 的目标是根据上下文来预测当前词语的概率。Skip-gram 刚好相反，它是根据当前词语来预测上下文的概率。

而 Doc2vec 模型除了增加一个段落向量以外，其几乎等同于 Word2Vec。和 Word2Vec 一样，该模型同样存在两种方法，包括句向量的分布记忆模型（Distributed Memory Model of Paragraph Vectors，PV-DM）和句向量的分布词袋（Distributed Bag of Words version of Paragraph Vector，PV-DBOW）。PV-DM 试图在给定上下文和段落向量的情况下预测词语的概率。而 PV-DBOW 则是在只给定段落向量的情况下预测段落中一组随机词语的概率。由于 Word2vec 模型只是基于词的维度进行比较的，而不具有基于上下文的语义分析能力，故综合考虑下选择使用 Doc2vec 模型进行文本特征提取。

VGG16 模型的网络结构如图 3 中配置 D 所示 [20]。

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

图 3 VGG 模型网络结构

D 配置所示模型总共有 16 层，包括 13 个卷积层和 3 个全连接层，首先经过含有 64 个卷积核的两次卷积后，进行一次 pooling，然后经过含有 128 个卷积核的两次卷积后，再进行 pooling，接着经过含有三个 512 个卷积核的两次卷积后，再进行 pooling，最后再经过三次全连接层。VGG16 模型通过增加深度能有效地提升性能，且从头到尾只有 3x3 卷积与 2x2 池化，简洁优美，卷积可代替全连接，可适应各种尺寸的图片。

2.2 异构数据匹配

本节实现了 Doc2vec 与 VGG16 模型融合并得到一种基于异构数据特征向量的图文检索模型。设 A 代表训练集中的文本集合，B 代表测试集中的图像集合，result_img 代表推荐结果的图像集合。模型的具体实现步骤为：

（1）给定一张在数据集中的图片 Y_{test}，利用 VGG16 模型提取出 Y_{test} 与数据集中图片的特征向量，接着通过与训练集中图片特征向量进行相似度比较，所用的相似度计算公式如式（1），*dis* 代表两个特征向量 *A*，*B* 的余弦距离，其中 ||*A*||，||*B*|| 分别代表这两个向量的模，*n* 表示向量的维数，A_i 与 B_i 代表向量 *A*，*B* 每一维上的值，最后得到与 Y_{test} 最相似的图像集 result_img，其大小为设定的范围 R，将其作为推荐候选图像集合，其中包括图像编号与图像相似度。

$$dis = \cos(\theta) = \frac{A \cdot B}{\|A\| \ \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{1}$$

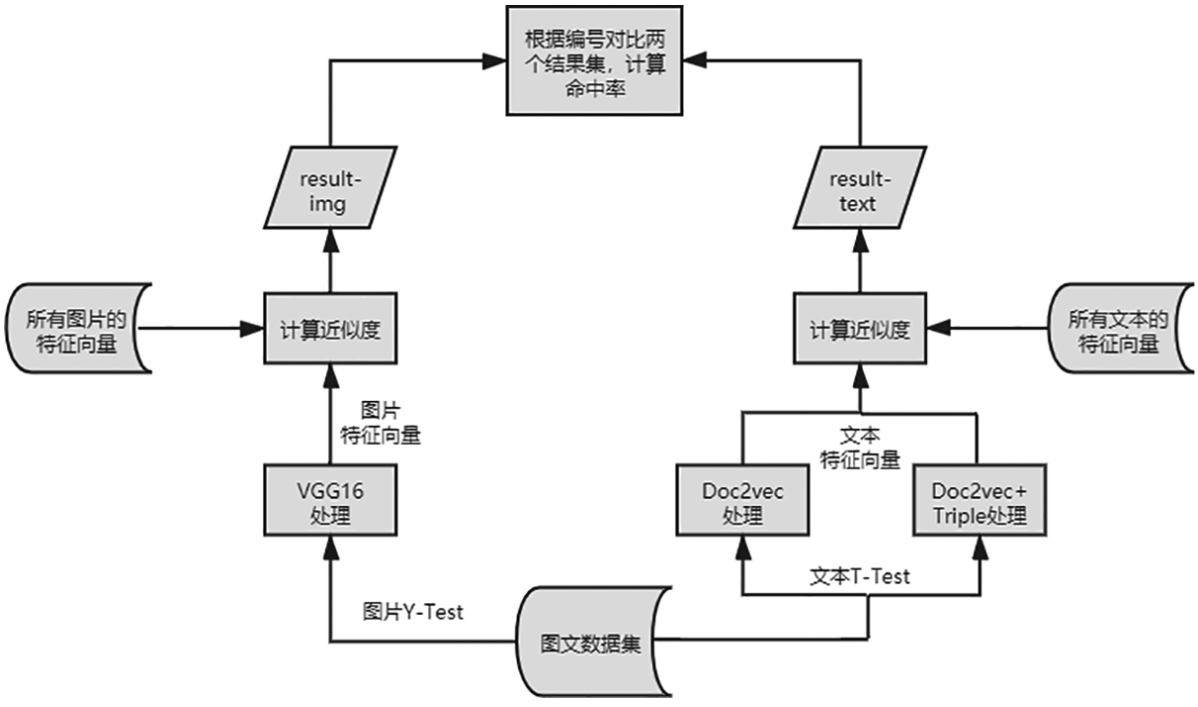

图 4 基于异构数据融合的图文检索模型

（2）根据 Y_{test} 对应的文本 T_{test}，通过预训练的 Doc2vec 模型或者 Doc2vec 与语义分析 Triple 模型得到与其最相似的文本集 result_text，其大小同样为 R。

doi:10.3772/j.issn.2095-915x.2021.04.003

（3）由于数据集中图像和文本编号相同，通过去除其本身编号后判断 result_img 中的图片编号是否同样存在于 result_text 中。若存在，则表示该图像配文命中成功，若不存在则表示命中失败，遍历所有测试图片，得到关于以图配文的评价命中率，具体公式见式（2）。

2.3 文本分类模型

在 NLP 各类中文任务中，无论是稍早提出的 Cove、Elmo、GPT，还是 BERT 模型，其建模对象主要聚焦在原始语言信号上，较少利用语义知识单元建模，这个问题在中文方面尤为明显。

如果能够让模型学习到海量文本中蕴含的潜在知识，势必会进一步提升各类 NLP 任务效果。因此百度提出了基于知识增强的 ERNIE 模型。ERNIE 模型能够对海量数据中的实体概念等先验语义知识建模，学习真实世界中的语义关系，该模型通过对词、实体等语义单元的掩码，使得模型学习完整概念的语义表示。相较于 BERT 学习原始语言信号，ERNIE 直接对先验语义知识单元进行建模，增强了模型语义表示能力，因此采用 ERNIE 模型进行分类模型的训练及预测。

BERT 模型与 ERNIE 模型的对比如图 5 所示。

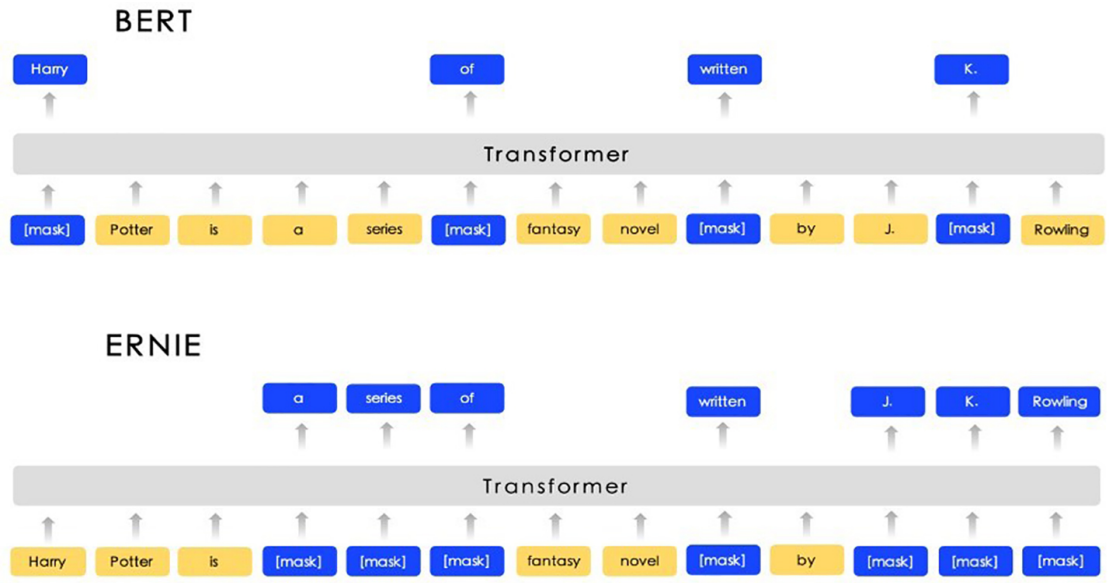

图 5 **BERT** 模型与 **ERNIE** 模型对比

3 实验设计与分析

3.1 实验数据及指标

为了验证本文设计的算法，实验数据选取了 Wiki 中文语料、THUCNews 中 84 万篇的新闻文档数据以及 2017 年由搜狐公司举办的图文匹配大赛中所用数据。后者所使用的数据包含了 100000 级别的训练集及 20000 级别的验证集。该数据集里的每一篇新闻文本描述都有与之对应的一幅配图。

本文设置平均命中率（Average Hit Rate，AHR）作为参考指标，AHR@N 表示从图像相似度匹配结果中的第一个开始遍历，看从第一个到第 N 个编号对应文本标题是否能够在文本相似度匹配结果中的前 N 个中找到，如果遇到了就将计数值 Count 加一，并结束小循环，最后直到大循环结束，根据公式（2）得到 AHR@N 的值，其值越大说明匹配结果越好。

$$AHR@N = \frac{\text{Count}}{N} \tag{2}$$

3.2 基于分类的图文匹配实验

3.2.1 实验流程

首先利用 keras 库中预训练好的 VGG16 模型计算出数据集中图像的特征向量，再通过内积运算得到所有图像间的向量距离，对于每幅图像的特征向量都能得到一个它与其他图像特征向量距离的序列，将其从大到小排序，得到与该图像最为相似的 R 幅图像。

由于数据集中的文本冗杂，故通过数据预处理将其规范化，得到关于每篇文档的标题，通过分析文档标题将其分为 10 个类别，包括财经、房产、股票、教育、科技、社会、时政、体育、游戏、娱乐，接着从 THUCNews 中获取了上述 10 个类别的 20 万条新闻标题数据（每类包含 20000 条数据），再通过 ERNIE 模型训练得到关于新闻标题文本的预训练模型，然后通过该预训练模型对数据集中的标题进行预测，得到数据集中的文本标题分类结果。

其次通过中文维基百科语料集训练 Doc2vec 模型，再根据前面的分类结果将标题数据集划分，通过该预训练模型得到关于每个标题文本的特征向量，同样经过内积运算得到所有相同类别中标题之间的向量距离，对于每个标题的特征向量都能得到一个它与其他标题特征向量距离的序列，将其从大到小排序，得到与该文本标题最为相似的 R 份文档，通过设置不同的 R 值和判断图像最为相似的 R 幅图像对应文档标题是否出现在这 R 份文档标题中来检验图像匹配结果的有效性，最后我们能够通过评价不同类别的 AHR，得到该模型在不同新闻类别中的效果。

3.2.2 实验结果分析

表 1 显示了本实验在搜狐图文匹配大赛的验证集上获得的评估结果。我们可以从图 6 看出，R 的取值越大，平均命中率越高。在 R=50 时，在教育类新闻中图文匹配效果最好，平均命中率到达了 69.0%，在游戏类新闻中图文匹配效果最差，平均命中率只有 37.3%。

表 1 基于分类的图文匹配

新闻类型	平均命中率 (R=10)	平均命中率 (R=30)	平均命中率 (R=50)
财经	15.3%	38.2%	56.3%
房产	12.3%	25.4%	41.0%
股票	13.2%	31.5%	49.0%
教育	29.8%	58.5%	69.0%
科技	12.9%	26.7%	42.1%
社会	15.4%	29.0%	50.7%
时政	19.1%	32.6%	47.6%
体育	12.5%	23.5%	39.4%
游戏	11.6%	22.2%	37.3%
娱乐	14.8%	26.7%	40.2%

3.3 基于语义分析的图文匹配实验

因为图像与语句的高层语义信息是人类理

doi:10.3772/j.issn.2095-915x.2021.04.003

解的一种抽象信息，与底层的数据特征存在“语义鸿沟”，所以导致在房产与游戏类新闻中图文匹配效果较差。故我们进行了基于语义分析的图文匹配实验。

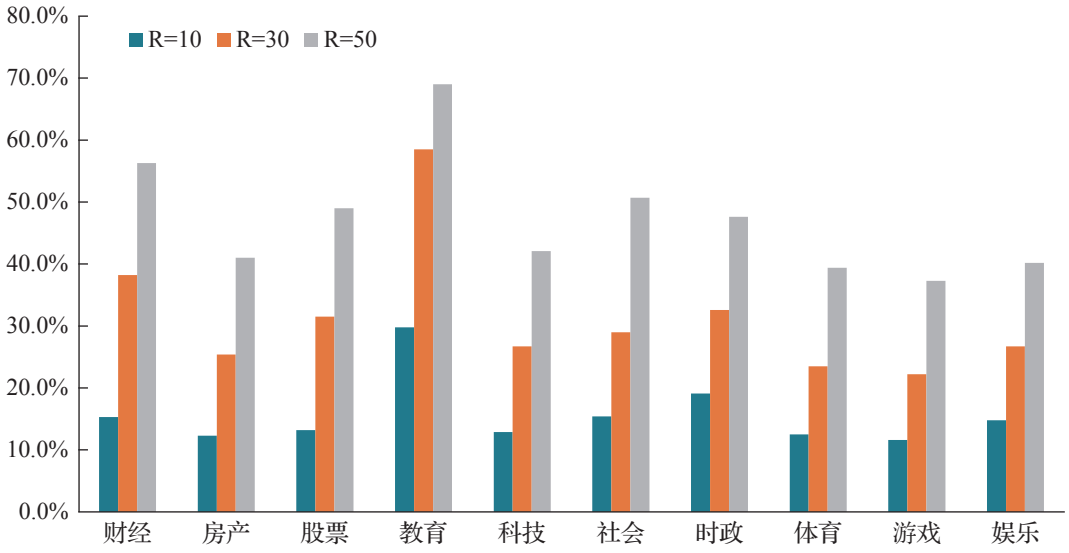

图 6 在基于分类的实验中，不同类别新闻在取不同 **R** 值时的平均命中率

3.3.1 实验流程

整体流程与基于分类的图文匹配实验步骤相似，在计算近似度时，要先对验证集利用基于分类的图文匹配实验中的预训练模型进行分类，然后使用预训练后的三元组提取模型 Triple 得到每个标题文本的三元组（实体 1，关系，实体 2）。利用三元组中实体或关系的特征向量距离计算近似度，近似度的计算主要分为以下两种情况：

（1）若两者的三元组都完整，没有缺失，则将实体 1 与实体 2 进行拼接，使用 Doc2vec 将其转化为特征向量，通过式（1）计算出两者近似度 S_1；将关系使用 Doc2vec 转化为特征向量，计算出两者近似度 S_2，通过式（3）计算出两者近似度 S。

$$S=\alpha S_1+\beta S_2 \quad (\alpha=0.6,\beta=0.4) \qquad (3)$$

（2）若有一方的三元组有缺失，则直接使用两者标题文本信息，使用 Doc2vec 模型将其转化为特征向量，通过式（1）计算出两者近似度。

3.3.2 实验结果分析

表 2 显示了本实验在搜狐图文匹配大赛的验证集上获得的评估结果。由图 7 我们可以看出，与上一个实验类似，R 的取值越大，我们的平均命中率也就越高。当 R 取 50 时，在教育类新闻中图文匹配效果最好，平均命中率到达了 60.4%，在体育类新闻中图文匹配效果最差，平均命中率只有 28.8%。通过图 8 我们可以看出，在 R 取 10 或 30 时，基于语义分析的图文匹配对于游戏类新闻的平均命中率稍有提升，但对于其它类别的命中率影响不明显。且在 R 取 50 时，基于语义分析的图文匹配的平均命中率明显低于基于分类的图文匹配，可能是因为三元组的提取丢失了部分语义信息，所以导致命中率的降低。

表 2　基于语义分析的图文匹配

新闻类型	平均命中率(R=10)	平均命中率(R=30)	平均命中率(R=50)
财经	14.4%	35.4%	44.3%
房产	11.6%	23.7%	33.1%
股票	13.2%	29.8%	36.9%
教育	27.8%	56.4%	60.4%
科技	12.6%	23.2%	30.8%
社会	14.7%	28.1%	36.9%
时政	19.2%	32.2%	40.0%
体育	11.9%	21.4%	28.8%
游戏	13.6%	25.4%	31.5%
娱乐	15.2%	28.0%	33.1%

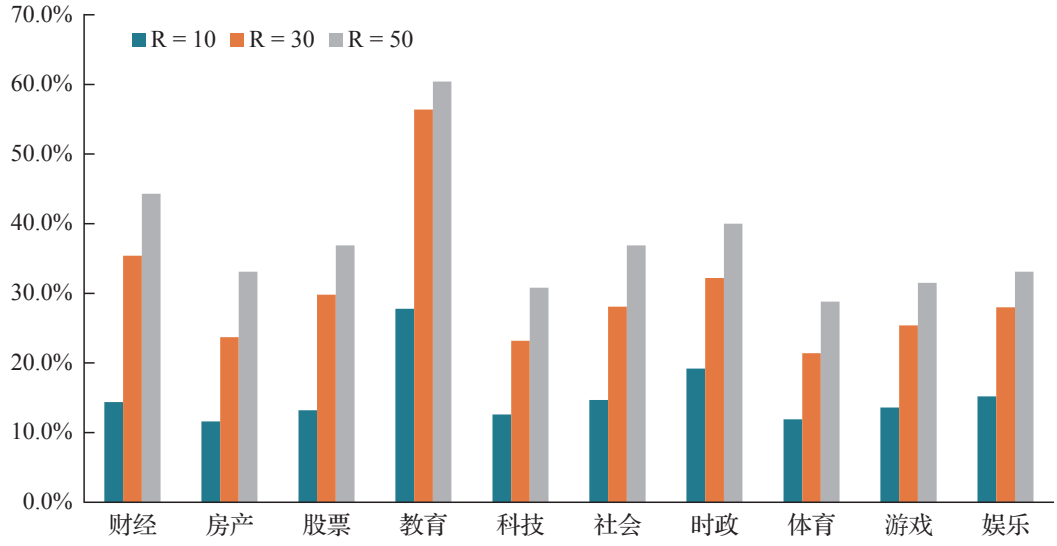

图 7　在基于语义分析的实验中，不同类别新闻在取不同 **R** 值时的平均命中率

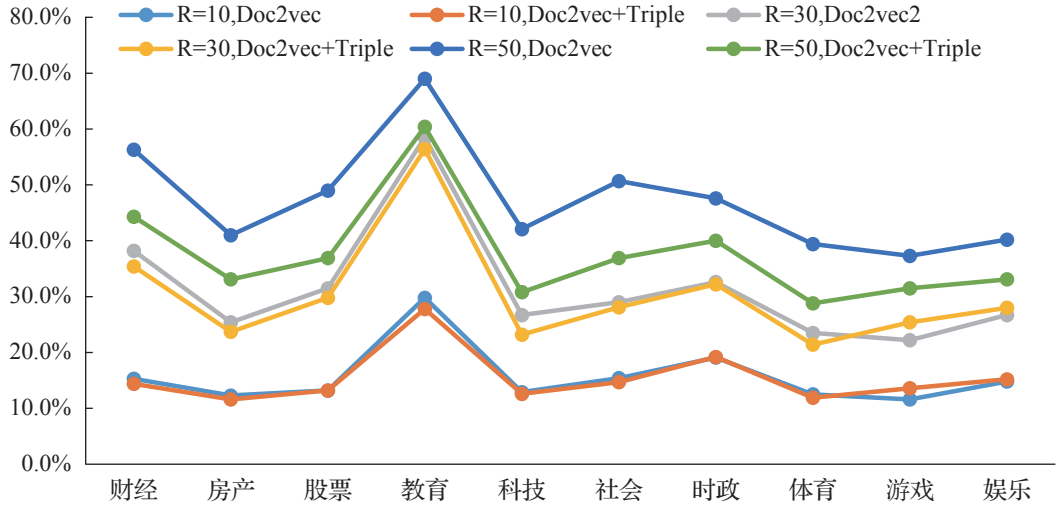

图 8　基于分类与基于语义分析的实验对比

doi:10.3772/j.issn.2095-915x.2021.04.003

4 结语

由于涉及不同模态的数据，基于异构数据的图文检索方法在实体抽取、知识图谱等领域意义深远，前景十分广阔。除了用户对图文匹配有需求之外，人工智能等领域也需要通过处理不同模态形式的数据来进行研究，这对之后的实体识别、知识推理和丰富知识图谱等方面有着不可替代的作用。

本文通过应用 Doc2vec 模型和 VGG16 模型分别处理得到文本特征和图片特征，以及使用向量相似度距离度量的方法计算得出文本和图片的匹配结果，实验结果表明 Doc2vec 模型能够较好地保存文本语义信息而 VGG16 模型提取的图片特征向量不尽人意。这极有可能是训练所用数据体量较小的缘故，在未来研究中，可以尝试扩大数据集的体量并采用其他神经网络模型，训练得到更能表征图片的特征向量。本实验的另一个局限在于，由于 Doc2vec 模型和 VGG16 模型训练得到的特征并未映射到同一相关子空间，无法直接对文本与图片进行比较。在后续研究中，可以尝试将本方法与子空间映射方法相结合，应能得到更优的结果。

参 考 文 献

[1] Rasiwasia N, Pereira J C, Coviello E, et al. A New Approach to Cross-Modal Multimedia Retrieval[C]. Proceedings of International Conference on Multimedia. Firenze: ACM, 2010:251-260.

[2] 李向阳 , 庄越挺 , 潘云鹤 . 基于内容的图像检索技术与系统 [J]. 计算机研究与发展 , 2001, 38(3):344-354.

[3] Socher R, Karpathy A, Le Q V,et al. Grounded Compositional Semantics for Finding and Describing Images with Sentences[C]. 52nd Annual Meeting of the Association for Computational Linguistics Conference PROGRAM. TACL, 2014:113-124.

[4] Wang J, He Y, Kang C, et al. Image-Text Cross-Modal Retrieval via Modality-Specific Feature Learning[C]. Proceedings of the 5th ACM on International Conference on Multimedia Retrieval June 2015. ACM, 2015:347–354.

[5] Karpathy A, Joulin A, Li F F. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping[J/OL]. arXiv Preprint, arXiv:1406.5679.

[6] 王晓宇 . 新闻自动推荐配图的方法研究 [D]. 呼和浩特 : 内蒙古大学 , 2019.

[7] Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: data, models and evaluation metrics[J]. Journal of Artificial Intelligence Research, 2015, 47(1):853-899.

[8] Vendrov I, Kiros R, Fidler S, et al. Order-Embeddings of Images and Language[C]. ICLR, 2016.

[9] Ma L, Lu Z, Shang L,et al. Multimodal Convolutional Neural Networks for Matching Image and Sentence[C]. IEEE International Conference on Computer Vision. IEEE, 2015.

[10] 钟庆虹 , 乔晓东 , 张运良 , 等 . 基于 LDA2Vec 和残差网络的跨媒体融合方法研究 [J]. 数据分析与知识发现 , 2019, 3(10):78-88.

[11] 鹿鹏 , 庄敏 , 龙刚 , 等 . 文本特征提取研究现状分析与展望 [J]. 科技创新与品牌 , 2017(4):70-74.

[12] 陈磊 , 李俊 . 基于词向量的文本特征选择方法研究 [J]. 小型微型计算机系统 , 2018, 39(5):129-132.

[13] 陈婧琳 . 基于特征学习和关联学习的在线商品跨媒体检索研究 [D]. 南昌 : 华东交通大学 , 2016.

[14] Moody C E. Mixing Dirichlet Topic Models and Word Embeddings to Make Lda2vec[J/OL]. arxiv Preprint, arxiv: 1605.02019.

[15] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003(3):993-1022.

[16] Mikolov T, Sutskever I, Kai C, et al. Distributed Representations of Words and Phrases and Their Compositionality[J]. Advances in Neural Information Processing Systems, 2013(26):3111-3119.

[17] 常芳 , 尚振宏 , 刘辉 , 等 . 一种基于颜色特征的自适应目标跟踪算法 [J]. 信息技术 , 2018(3):10-14.

[18] 叶雨晴 , 邱晓晖 . 基于 SIFT 与 K-means 的图像复制粘贴篡改检测 [J]. 计算机技术与发展 , 2018, 28(6):121-124.

[19] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks[C]. Proceedings of International Conference on Neural Information Processing Systems. Lake Tahoe: NIPS, 2012:84-90.

[20] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J/OL]. arXiv Preprint, arXiv:1409.1556v6.

[21] 郝志峰 , 李俊峰 , 蔡瑞初 , 等 . 面向图文匹配任务的多层次图像特征融合算法 [J]. 计算机应用研究 , 2020, 37(3):951-956.

[22] Yan F, Mikolajczyk K. Deep correlation for matching images and text[C]. Computer Vision & Pattern Recognition. IEEE, 2015:3441-3450.

[23] Ma L, Lu Z D, Shang L F, et al. Multimodal convolutional neural networks for matching image and sentence[C]. Process of IEEE International Conference on Computer Vision. Piscataway. NJ: IEEE Press, 2015:2623-2631.

[24] Wang L W, Li Y, Lazebnik S. Learning deep structure-preserving image-text embeddings[C]. Process of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway. NJ: IEEE Press, 2016:5005-5013.

[25] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv Preprint, arXiv:1301.3781.