DOI:10.3969/j.issn.1009-6868.2016.02.008
网络出版地址:http://www.cnki.net/kcms/detail/34.1228.TN.20160302.1631.004.html

基于概念网的媒体大数据分析和结构化描述方法

Topic Network-Based Media Big Data Analysis and Structural Description

张宝鹏/ZHANG Baopeng[1]
彭进业/PENG Jinye[2]
范建平/FAN Jianping[2]
(1. 北京交通大学 计算机与信息技术学院,北京 100044;
2. 西北大学 信息科学与技术学院,西安 710069)
(1.School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China;
2. School of Information and Technology, Northwest University, Xi'an 710069, China)

中图分类号:TP393 文献标志码:A 文章编号:1009-6868 (2016) 02-0033-005

摘要: 提出基于概念网的媒体大数据结构化描述和分析的技术框架,该框架可以针对不同的数据获取来源,通过层次式多角度概念描述模型融合数据的视觉特征、实例和概念关联的语义,并提出面向单一媒体和多媒体文档的跨媒体概念提取及基于结构的语义对齐方法,从而有效支持媒体大数据的语义关联分析及多领域的智能应用。

关键词: 概念网;媒体大数据分析;概念抽取;结构化描述;可视化

Abstract: In this paper, we propose that a topic network-based enabling technology framework for big media analysis and structural description. And it proposes a hierarchical concept description model with multiple perspectives for different sources data to integrating semantic of visual, instance and concept correlation. And cross-media concept extraction method for single media and multimedia document and their structure-based semantic alignment method are also proposed, which can efficiently support the big media analysis and smart application in many domain.

Key words: topic network; big media analysis; concept extraction; structural description; visualization

随着互联网的普及和迅速发展,各类在线社交网络(如Facebook、Twitter、新浪微博、腾讯网等)的飞速发展,网络数据资源越来越多样化,并呈爆炸式增长。这种大数据的势态引发了多行业、多领域的时代性变革。大数据思想的重要在于[1]:人们可以在很大程度上从对于因果关系的追求中解脱出来,转而将注意力放在相关关系的发现和使用上。目前,在互联网中,大量文本、图像、音频、视频等媒体大数据迅速增长,其中蕴含了很多人类社会活动的基本规律,公共卫生、商业乃至思维模式因此酝酿着重大的机会和挑战。基于大数据的研究逐渐成为各国政府重点发展的国家战略,及时、准确地获取并理解这些数据及其关系不仅仅可以为政府在社会生活、金融服务、医疗卫生等方面发现和处理民生问题,辅助政府决策,同时也为互联网经济的发展提供有效的客户和经济规律的知识辅助,提供商业智能决策支持。

收稿时间:2016-02-18
网络出版时间:2016-03-02

尽管媒体大数据成长迅速,应用广泛,但其数据量大、种类繁杂、价值密度低以及时时刻刻不断变化的特点,使得存储、统计、分类以及调用都非常困难[2],其价值远没有得到充分的利用和开发。而人工智能领域的一些理论和比较实用的方法,已经开始用于大数据分析方面,推动两个领域技术和应用融合的加速,但依然只是初期。目前谷歌、百度等通用的搜索引擎提供了基于文本描述的多媒体的检索机制,但对于大数据背景下的多种媒体数据来说,还缺乏准确文本描述,需要不同的算法分析、理解其内容的语义,实现相应的文本描述,从而为搜索引擎所用。另外,媒体数据间的异构性特点,使得当前单一媒体的搜索引擎无法有效支持大数据条件下异构媒体间的数据语义关联检索。因此,从媒体大数据智能应用的角度来看,其表示、理解及检索是重要的环节,而根据异构媒体间语义关系实现媒体大数据的智能的模式发现是解决这些问题的关键点。

1 媒体大数据分析和描述的关键问题

根据媒体大数据深度分析的目

标，以及其支撑媒体搜索引擎、媒体消费和关联分析的需求，尽管当前异构媒体的关联和分析技术有一些相关研究，但有些关键问题还没有得到解决，包括：

（1）媒体数据标注的不确定性及歧义性

除了大数据的4个V（Volumn、Variety、Velocity、Value）之外，为充分利用大数据蕴含的知识信息，一个重要的问题是解决媒体数据标注的不确定性、歧义性，这种不确定的标签数据包括：

- 粗糙标注，例如图片中对象是在图片层次上给出的，而忽略了其区域性的语义；
- 抽象标注，指标签只从高层语义角度给出，缺乏具体语义关联；
- 无关标注，指标注和图像语义并无关联；
- 噪声标注，指错误的标注。

这些标签数据将误导数据驱动的机器学习方法，从而导致数据训练分类器在性能和准确率上的退化。目前很多项目开展了图像智能标注的工作，旨在提高标签的准确率，包括传统概率的方法[3-4]、场景限制下的综合方法[5]、深度学习方法[6]及面向大规模的方法[7]等，但面向媒体大数据的复杂结构，复杂的语义及智能化的需求使得当前技术还远远不能满足其需要。

（2）媒体大数据结构化描述及其机器学习的算法

媒体大数据包含大量的语义概念，而且语义概念之间有千丝万缕的关系；同时对于不同主域的应用环境，不同的语义关系需要不同的结构化描述。目前传统多媒体语义描述模型主要包括两种：词袋模型，其源于自然语言理解，适合于视觉的相似匹配，但与语义并没有直接的对应关系；基于特征-语义的分类模型，源于机器学习，其主要参考的是人类语义感知设计，提取难度较大，准确率不高。由于传统多媒体语义提取采用多类学习的方法，其中用两类分类器合成的方法，训练检测复杂度较高，训练难度大，而传统的多任务学习和结构化支持向量机（SVM）学习方法，无法真正发掘出概念间相似性结构的信息。两种方法必须要解决的问题就是面向媒体大数据的泛化能力。目前，基于深度学习的多媒体语义提取方法得到了空前的关注，如文本检索会议（TREC）的视频事件检测提出的基于卷积神经网的深度学习算法，微软的音、视频索引服务（MAVIS）的语音识别系统，Google的深度学习模型等，都获得了很好的效果。但它们主要对音频、视频或文本单一模态进行分析，没有充分利用多模态信息间的相互协同关系。

（3）媒体大数据的关联性分析

媒体大数据分析首先需要研究异构媒体的统一表示[8]，相似度计算及语义关联的分析方法。传统的异构媒体采用基于子空间的映射技术，包括典型关联分析（CCA）方法、概率潜语义分析（PLSA）方法等。在相似度计算方面，主要的度量方法是基于图模型的相似度度量方法和基于学习的相似度度量方法[9]，但目前两者主要都是依赖共生性假设，即如果两个多媒体文档包含同一个媒体对象，则它们具有相同语义，也可以说是基于概念和概念的相似性或简单的物理依赖。跨媒体数据中的内在语义关系和结构（概念相关性网络）并没有给予充分的考虑，并且概念间关系复杂，因此并不适用于媒体大数据的深度分析，而主流的机器学习方法可能无法直接解决其复杂、大规模学习问题。

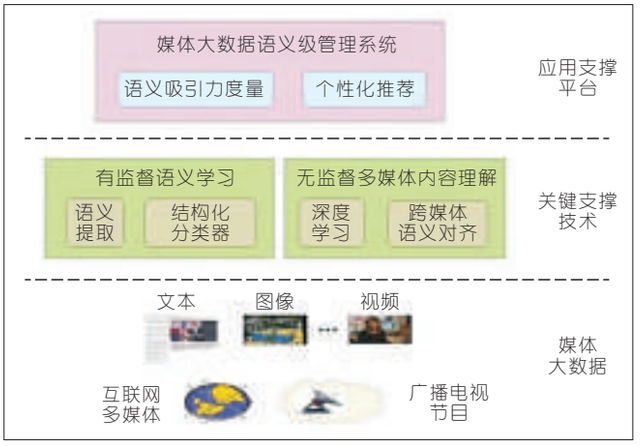
（4）媒体大数据的可视化与可视化分析

在媒体大数据的深度分析中，准确率和查全率是主要的分类器的评估标准，但由于学习分类器会过拟合，以及用于分类器训练和测试的样本是服从于同样的分布，因此评估标准会误导分类器的判定能力，也就是说不能显式地反映分类器和正确率和其辨识力。一种有效用于分类器评估的方法是可视化分类器的边界和类间的边缘，用户可以交互式地评估其正确率。因此，在机器学习过程中融合人的交互式操作，来改善分类器训练具有更高的应用价值。

2 媒体大数据关联分析的参考技术框架

针对目前在媒体大数据深度分析中所面临的问题，其未来发展的思路应该是基于内容语义的、全生命周期的支撑，因此我们提出了基于概念网的核心参考技术框架，如图1所示。针对媒体大数据处理的数据特点，我们需要考虑两种关键技术问题：有监督的媒体语义学习；无监督的多媒体内容理解。

目前媒体大数据的跨媒体概念的提取方法主要针对两种不同的媒体数据获取类型：一种是多媒体文档，主要是电视节目和网络媒体，包含图像、视频、音频和伴随文本描述等多种媒体形式。其关联关系隐含在多媒体文档中，重点解决的问题是多模态特征融合与跨模态关联分析的问题，而跨模态深度学习技术可以基于已有的图像、视频、音频及其对应的文本训练其语义概念检测模型，检测数据中的语义概念，并使用跨媒体语义对齐技术实现不同媒体语义概念的对齐。另外一种是单一视觉媒体，主要指监控录像和照片包含单一视频和图像，但没有伴随文本描述，需要进行多媒体数据中视觉语义概念的直接检测。其通过结合直接的标注数据和图像或视频的初级语义进行结构协同学习，提取语义概念并关联到对应的初级语义概念上，得到跨媒体语义。结构协同学习是基于概念相似性结构进行协同学习获得的分类模型的方法，其语义的统一于概念网络，有助于融合异构媒体的内容及关系特征，同时易于进行增量计算、测试修正及扩展。



◀图1
媒体大数据深度分析参考技术框架

该技术框架可以有效支持异构媒体大数据的可扩展应用，包括与当前搜索引擎的结合及面向不同应用领域的推荐系统等，如图2所示。

3 媒体大数据关联分析的关键技术

3.1 层次式多角度概念描述

多模态数据的语义提取并存储为语义库，需要一个能够描述所需语义信息，方便语义运算的语义模型作为数据语义存储和运算格式。由于相关的数据应用需要在高层语义、底层特征和实例样本等不同的层面处理海量数据及其语义，这要求语义描述模型要在统一的框架下存储所有这些信息。其难点在于：模型必须能够统一存储不同种类、不同层面差异巨大的媒体数据及其特征和语义。我们认为：应包括3层结构组成的描述模型，通过整合3个层次的关联（如图3所示），实现语义-实体-关系模型。其中位于语义层次的概念网应充分考虑大规模概念间的相关性，并提供能够对媒体大数据进行关联分析与结构化描述的新框架，从而用于指导训练大规模相关关联的分类器，并大幅度提高概念检测准确性。

3.2 基于多媒体文档的跨媒体概念提取

传统搜索引擎技术支持的图像-文本对应关系的获取具有很大的不确定性（如图4所示），而面向媒体大数据，语义对齐与关联分析可以利用视觉聚类、随机行走和概念语义网进行相关性重排以产生更准确的跨媒体语义对齐结果，并提取更准确的大规模跨媒体概念，同时利用视觉聚类可以进行跨媒体的语义消歧。这种跨媒体语义对齐方法可以为机器视觉研究提供大量的可靠标注的训练数据。

3.3 基于单一媒体的跨媒体概念提取

结构协同学习利用多个语义概念之间的相似性关系信息设计检测语义概念的分类器，通过充分利用这种相似性关系信息（该关系可以用结构表示），提升大类数媒体数据分类的性能和准确率。面向大媒体数据的大规模结构协同学习框架（如图5所示），首先将语义概念相似性网络进行分割以形成相似概念组，这一过程将最相似的语义概念分到同一组，而将差异较大的概念分到不同组，实现将语义概念中的相似结构表示为概念的分组情况。针对不同分类任务可选择不同分类算法和不同特征表示，有助于大量减少训练复杂度。同时，利用多层视觉树[10]来管理大量

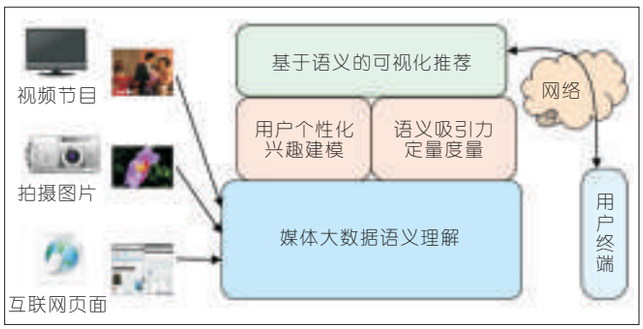
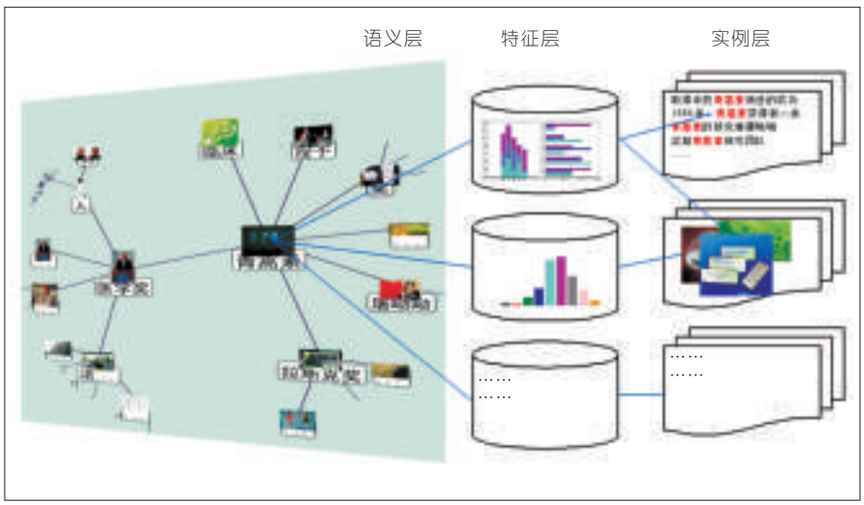
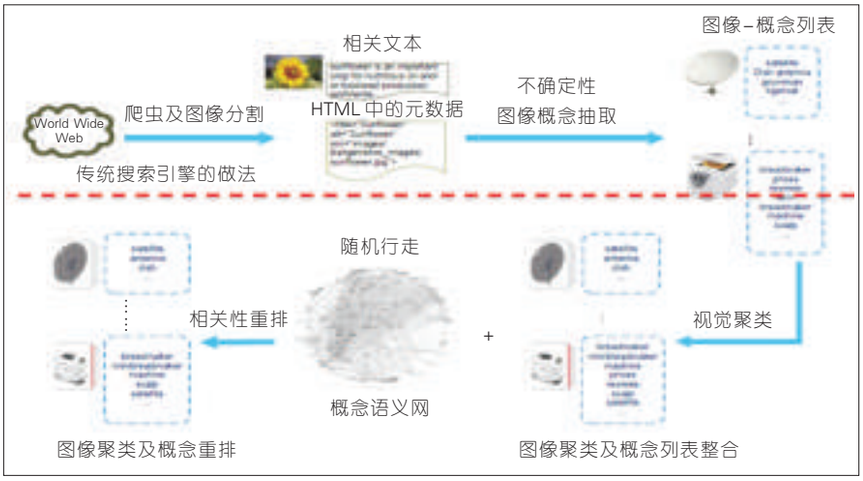


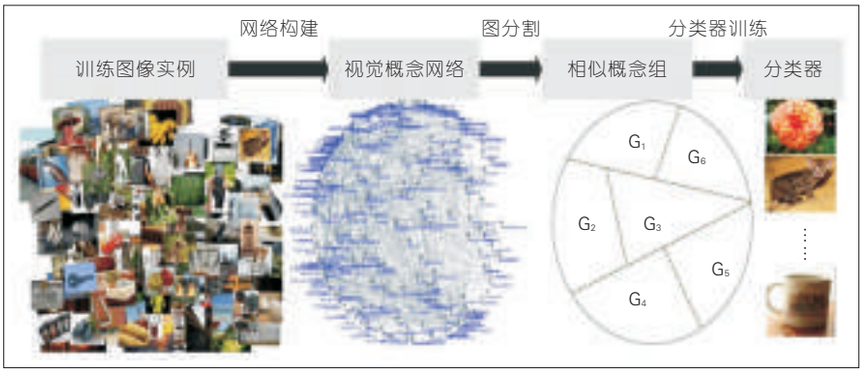
图2▶
基于语义分析的媒体大数据消费系统



▲图3 层次式多角度概念描述模型



▲图4 传统搜索引擎与跨媒体概念提取的技术思路对比



▲图5 大规模结构协同学习框架

分类器，实现快速提取大规模跨媒体概念。这其中一个重要的问题是：训练图像实例如何提取语义。目前，深度学习可以得到很好的特征提取及分类效果[11]，而更为有效的方法是将各种传统视觉特征作为先验知识模型加入到深度学习算法的训练当中。

3.4 跨媒体语义对齐

当前很多算法都是针对不同媒体的数据构建语义结构化模型。这些模型有的较好地关联到了高层语义，但因为缺乏相关的文本数据标注而无法关联到高层语义，只能通过深度学习算法获得大量抽象的语义概念及其关系。为了统一管理和挖掘媒体大数据，必须实现抽象的语义概念与具体的语义概念（语言）对齐。描述媒体的结构化语义信息的模型一般为图结构，我们需要研究语义对齐方法实现多个语义结构的对齐，提高语义信息的准确度。其难点在于：需要精确估计两个图的部分节点之间的相似度关系，但语义概念在不同媒体数据中的具体表现差异巨大，难以直接估算相似度。

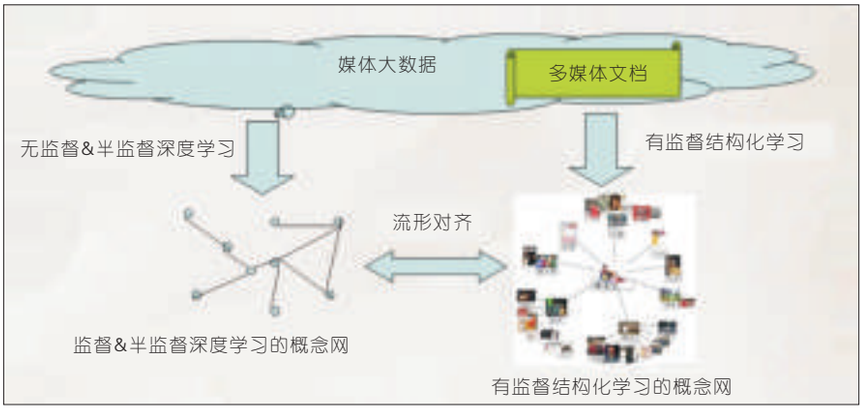
为了充分利用所有语义信息获得最高对齐精度，可以使用流形对齐方法，该方法对实现两个语义结构的对齐是个较好的选择。如图6所示：流形对齐算法综合计算两个语义空间的语义概念的相似度和语义概念的内在关联结构，从而实现两个语义空间的对齐，这比仅仅依据语义概念之间的各种相似性的方法具有更高性能。

为简化描述，下面我们把抽象的语义概念称为未标记实体，具体的语义概念称为语义实体。在使用流形对齐算法过程中，我们需要计算部分未标记实体和语义实体之间的相似度。我们提出了两种相似度计算方法：结构协同分类获得的语义概念包含对齐的图像视频数据，这些数据上也包括深度学习算法提取的未标记实体，通过统计未标记实体在某个语义实体对应的图像、视频数据中出现的概率，即可计算出未标记实体和语义实体的相似度；用结构协同学习获得的语义概念检测模型检测所有图像和视频关键帧，可以获得描述其语义的一个高维矢量，一对视觉实例间的语义相似度可以定义为其语义矢量之间的近似程度，未标记实体和语义实体的语义相似度则可基于两者对应的图像和视觉结构间的相似度进行计算。为了既可以体现跨媒体数据对齐的信息又利用结构协同学习的结果，有效的方法是将以上两种相似度加权组合获得未标记实体和语义实体之间的融合相似度，融合相似度可以用作流形对齐的节点间对应信息，从而实现大规模媒体数据的知识的融合和一致性处理。

3.5 基于概念网的媒体大数据关联性分析及其可视化

如果把语义概念之间的相似性用一个加权图表示，语义概念之间的相似性结构信息将形成一个语义概念相似性网络。这个网络的结构对应于语义概念之间的相似性结构，因此可以用于结构化学习指导分类器结构设计。构造语义概念相似性网络首先需要度量语义概念之间的视觉相似度，而语义概念之间的视觉相似度基于样本之间的相似度计算，样本之间的相似度要基于底层视觉特征计算。为了消除概念之间的相似性非常小却仍然有连接的现象，我们采用自底向上层次式聚类算法裁剪全连接的语义网络。

该方法可以有效表示主域的数据相关性。例如，用于描述新闻热点



▲图6 深度学习与结构化学习相结合的对齐方法

间的相关性的新闻概念网，如图7所示。这种概念网提供了一个对大规模媒体概念进行关联分析和结构化描述的新框架结构，同时也便于面向不同消费系统进行扩展应用。

4 结束语

当前多领域、跨领域的网络媒体数据呈大规模增长的态势，而异构媒体的智能关联、知识表示是合理利用数据并为行业提供智能化服务的核心研究问题，因此，突破媒体大数据的基于内容的结构化描述、关联与深度分析，形成媒体内容语义的全生命周期的技术框架，以支持个性化搜索与智能推荐、跨终端的多媒体内容呈现等关键技术的发展，对建立面向用户的智能服务平台，推进知识获取及推广，改善用户体验具有非常重要的意义。

参考文献
[1] 维克托.迈尔.舍恩伯格, 肯尼思.库克耶. 大数据时代: 生活、工作与思维的大变革[M]. 盛扬燕, 周涛, 译. 杭州: 浙江人民出版社, 2013
[2] ZHU W, CUI P, WANG Z. Multimedia Big Data Computing [J]. IEEE Multimedia, 2015, 22(3): 96–105. DOI: 10.1109/MMUL.2015.66
[3] FENG S L, MANMATHA R, LAVRENKO V. Multiple Bernoulli Relevance Models for Image and Video Annotation[C]//Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2004(2): II–1002–II–1009. DOI: 10.1109/CVPR.2004.1315274
[4] BARNARD K, DUYGULU P, FORSYTH D, et al. Matching Words and Pictures [J]. J Mach Learn Res, 2013(3): 1107–1135
[5] LI J L, SOCHER R, LI F F. Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework[C]//Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2009: 2036–2043
[6] FARABET C, COUPRIE C, NAJMAN L, et al. Learning Hierarchical Features for Scene Labeling[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence. USA: IEEE, 2012, 35(8): 1915–1929
[7] WESTON J, BENGIO S, USUNIER N. Large Scale Image Annotation: Learning to Rank with Joint Word–Image Embeddings [J]. Machine Learning, 2010, 81 (1):21–35
[8] ZHU S C. Statistical Modeling and Conceptualization of Visual Patterns [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(6): 691–712. DOI: 10.1109/TPAMI.2003.1201820
[9] 唐杰, 陈文光, 面向大社交数据的深度分析与挖掘[J]. 科学通报, 2015, 60(5): 509–519
[10] ZHOU N, FAN J. Jointly Learning Visually Correlated Dictionaries for Large–scale Visual Recognition Applications [J]. IEEE Transaction. on Pattern Analysis and MachineIntelligence, 2014, 36(4):715–730
[11] DEAN J, CORRADO G S, MONGA R, et al. Large Scale Distributed Deep Networks[C]// Proceedings of the 26th Annual Conference on Neural Information Processing Systems. Canada, 2012: 1223–1231

◀图7
基于概念网的媒体大数据的关联性表示

作者简介

张宝鹏，北京交通大学计算机与信息技术学院讲师；主要研究方向为多媒体理解及检索、大数据管理及挖掘等；已主持及参与完成国家级、省部级项目20余项；已发表论文20余篇。

彭进业，西北大学信息科学与技术学院教授、博士生导师，教育部新世纪优秀人才支持计划获得者，陕西省图象图形学会常务理事；主要研究方向为信号处理与信息安全；发表论文60余篇。

范建平，西北大学信息科学与技术学院特聘教授、博士生导师；主要研究方向为统计机器学习、大规模视觉识别、社交图像/视频分析、大规模图像/视频检索等；已发表论文160余篇。