

多媒体信息检索中的查询与反馈技术

查正军<sup>1</sup>      郑晓菊<sup>1,2</sup>

<sup>1</sup>(中国科学技术大学信息科学技术学院 合肥 230027)

<sup>2</sup>(中国科学院合肥物质科学研究院 合肥 230031)

(zhazj@ustc.edu.cn)

Query and Feedback Technologies in Multimedia Information Retrieval

Zha Zhengjun<sup>1</sup> and Zheng Xiaoju<sup>1,2</sup>

<sup>1</sup>(School of Information Science and Technology, University of Science and Technology of China, Hefei 230027)

<sup>2</sup>(Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031)

**Abstract** In spite of the remarkable progress made in the past decades, multimedia information retrieval still suffers from the “intention gap” and “semantic gap”. To address this issue, researchers have proposed a wealth of query technologies to help user express search intent clearly as well as feedback technologies to help retrieval system understand user intent and multimedia data accurately, leading to significant improvements of retrieval performance. This paper presents a survey of the query and feedback technologies in multimedia information retrieval. We summarize the evolution of query styles and the development of feedback approaches. We elaborate the query approaches for retrieval on PC, mobile intelligent devices and touch-screen devices etc. We introduce the feedback approaches proposed in different periods and discuss the interaction issue in exploratory multimedia retrieval. Finally, we discuss future research directions in this field.

**Key words** multimedia information retrieval; retrieval intent; content understanding; query; feedback

**摘 要** 历经几十年的发展,多媒体检索取得了长足的进步,然而检索性能的提升依然受到“意图鸿沟”与“语义鸿沟”的制约.针对此问题,学术界提出了一系列查询技术帮助用户清楚地表达检索意图以及反馈技术帮助系统准确地理解用户意图与媒体数据,有效提升了检索性能.对多媒体检索中的查询与反馈技术进行了分析与讨论.分析了查询方式的演变与反馈技术的发展,综述了面向 PC 机、移动智能终端、触屏设备的查询技术,介绍了不同时期的反馈技术,探讨了探索式搜索中的交互问题,最后分析了该领域的未来研究趋势.

**关键词** 多媒体信息检索;检索意图;内容理解;查询;反馈

**中图法分类号** TP391

随着多媒体采集设备的日益普及和数据存储、互联网等技术的飞速发展,图像、视频等多媒体数据已经成为人们获取与传播信息的主要媒介,正在全

方位地渗透到人们的工作和生活中.多媒体数据规模庞大且呈爆炸式增长,其潜在价值巨大,是国家大数据战略资源的重要组成部分.在“国家中长期科学

收稿日期:2017-01-05;修回日期:2017-04-24

基金项目:国家自然科学基金面上项目(61472392);国家自然科学基金优秀青年科学基金项目(61622211);国家自然科学基金重点国际合作项目(61620106009)

This work was supported by the General Program of the National Natural Science Foundation of China (61472392), the National Natural Science Foundation of China for Excellent Young Scientists (61622211), and the Key Project of International Cooperation of the National Natural Science Foundation of China (61620106009).

和技术发展规划纲要(2006—2020)”中,数字媒体内容被列为“信息产业及现代服务业”领域的优先主题之一.面对海量多媒体数据,如何实现快速准确的信息检索,一直是多媒体研究领域的热点问题.多媒体检索是满足人们信息需求的主要途径,是众多智能媒体应用的基础.

最早的多媒体检索研究可以追溯到20世纪70年代末期,其主要依赖人工标注生成媒体数据的文本标签,利用文本匹配完成检索.人工标注费时费力,且存在标注缺失与错误等问题.自本世纪初始,随着计算机视觉、模式识别、机器学习等技术的进步,逐渐发展出多媒体内容自动标注方法,通过建立语义概念模型自动分析媒体数据内容,生成其语义标签<sup>[1]</sup>,用于大规模数据的管理与检索.与此同时,针对网络媒体数据,基于网页文本挖掘的多媒体检索技术发展迅速,成为当前主流商业图像视频搜索引擎的主要技术基础.此类技术通过自动分析网络图片视频的标题、环绕文字、URL等元数据(metadata),抽取出反映图片视频内容的关键字,建立数据索引,支持基于文本的多媒体检索<sup>[2-3]</sup>.由于网页元数据含有噪声,抽取的关键字往往与图片视频内容不相符.针对于此,自2008年起,媒体内容分析技术被逐步引入到基于网页文本挖掘的多媒体检索中,以提升分析与检索的精度<sup>[4]</sup>.

除上述基于文本的检索之外,基于内容的多媒体检索也广受关注,其通常以图像或视频示例作为查询,通过视觉特征匹配完成检索.该技术起源于20世纪90年代初期,迄今经历了兴起—没落—再兴起的发展轨迹.20世纪90年代,基于内容的图像视频检索(CBIR,CBVR)是多媒体领域的研究热点,研究人员相继研发出了QBIC<sup>[5]</sup>,VisualSeek<sup>[6]</sup>,MARS<sup>[7]</sup>等早期的图像视频检索系统,支持几千至几万幅图片视频的检索.受限于视觉特征表达能力,检索的质量难以保证,且缺少可扩展索引方案,难以支持大规模检索.因而,基于内容的检索在本世纪初陷入低谷.随着高判别力视觉特征的提出<sup>[8]</sup>、高维特征索引<sup>[9]</sup>和视觉词倒排<sup>[10]</sup>等技术的出现、计算能力的大幅度提升以及新型应用需求的不断涌现,基于内容的图像视频检索技术在近十年进入飞速发展期,产生了一系列新颖的检索技术,也促生了“以图搜图”技术的商业化.

历经了几十年的发展,多媒体检索在研究的深度与广度以及技术应用的渗透度和覆盖面等方面均取得了长足的进展.相关的基础理论和关键技术不

断发展,应用服务渗透至包括电子商务、市场营销、社会安防等在内的众多领域.然而,在多媒体检索中,用户时常难以清楚地表达检索意图且检索系统难以准确地理解用户意图,导致用户与检索系统之间存在“意图鸿沟”.另一方面,多媒体数据模式复杂、视觉形态各异、内容繁杂多样,导致计算机感知的底层特征与人们认知的高层语义之间存在“语义鸿沟”.如图1所示,“意图鸿沟”和“语义鸿沟”成为制约多媒体检索发展的瓶颈,限制了检索性能的提升,阻碍了检索技术的应用.克服此2类鸿沟是多媒体研究的重要任务,是实现多媒体检索跨越发展的必由之路.因此,研究人员开展了大量针对性研究,提出了一系列检索模型与技术.其中,有效途径之一是在检索环路中引入用户交互,采用“人在环路”(human in the loop)的方式进行交互式检索,收集并利用用户的交互反馈帮助检索系统理解媒体数据内容以及用户信息需求.研究人员主要在检索流程的查询端(入口)和结果端(出口)进行技术创新,提出查询技术帮助用户表达检索意图以及反馈技术帮助系统理解用户意图与媒体内容.

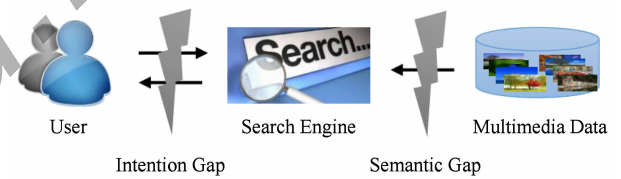


Fig. 1 The “Intention Gap” and “Semantic Gap” in multimedia retrieval

图1 多媒体检索中的“意图鸿沟”和“语义鸿沟”

本文将介绍多媒体检索中的查询与反馈技术的研究现状与进展.首先,介绍多媒体信息检索的整体框架与技术环节;进而,依次综述查询与反馈技术;最后,对未来的研究趋势进行展望与讨论.

### 1 多媒体信息检索

多媒体信息检索的基本流程如图2所示.在离线阶段,需要完成多媒体数据预处理、媒体内容分析、数据索引等,为高效准确地在线检索奠定基础.具体而言,1)预处理步骤需要完成数据选择等任务.受存储与计算资源的限制,无法对全量数据进行分析与索引.尤其在处理互联网媒体数据时,全网媒体数据规模极其庞大,需要从中选择能够满足大部分用户需求的数据进行分析与索引.2)分析步骤的主要任务是生成图像/视频等多媒体数据的特征表示

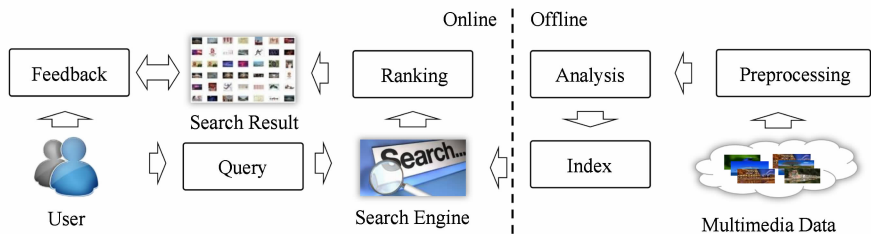


Fig. 2 The pipeline of multimedia retrieval system

图 2 多媒体检索系统流程图

(支持示例搜索)以及分析多媒体数据表达的语义信息,如概念、事件等(支持语义搜索)。近年来,图像/视频内容分析技术发展迅速,特征表示从人工设计特征演变为基于深度学习的特征表示、识别的语义元素从早期的少量语义概念发展到成百上千的概念集合、从简单的基本事件发展到复杂的综合事件。分析方法从模型驱动的基于分类器的方法<sup>[11-13]</sup>、数据驱动的基于搜索的方法<sup>[14-15]</sup>,发展到当前基于深度学习的方法<sup>[16-18]</sup>。3)在图像/视频的语义描述的基础上,索引步骤可以采用文本检索中的倒排技术生成图像/视频数据的语义索引,支持基于关键字的语义搜索。针对基于示例的内容搜索,需要解决图像/视频高维特征的索引问题。常用的索引技术包括 Hash 索引<sup>[19]</sup>、基于视觉词袋(bag of visual words)的倒排索引<sup>[10]</sup>、近邻图索引<sup>[20]</sup>等。

在线检索阶段,搜索引擎根据用户递交的查询,从索引中查找出与查询相关的媒体数据,利用排序技术生成结果序列。其中,查询的形式从经典的“关键字+查询框”衍生出多样式、多模态查询。排序技术由依据视觉相似度排序<sup>[21]</sup>发展为基于机器学习的排序方法<sup>[22]</sup>。交互式检索系统支持用户针对检索结果提供相关性反馈,利用用户反馈改进检索结果。常用的反馈技术包括相关样本反馈<sup>[23]</sup>、部分相关样本反馈<sup>[24]</sup>、新兴的属性反馈<sup>[25]</sup>等。

2 查询技术

2.1 查询推荐

图像/视频检索的经典查询方式为关键字查询,检索系统根据用户输入的关键字查找索引,将查找结果按照相关性排序返回给用户。然而,用户输入的查询往往不能精确表达其搜索意图。究其原因:1)用户输入的查询通常仅为 1~3 个词,表达的信息有限;2)查询词存在歧义、模糊等问题;3)用户对检索目标缺乏认知,无法构建准确的查询词。相关研

究<sup>[26]</sup>表明,多达 75% 的查询词不能清晰地表达用户意图。这就导致检索系统难以准确地理解用户意图,进而难以提供满足用户信息需求的搜索结果。

为帮助用户构造合适的查询以准确地描述其信息需求,检索系统普遍采用查询推荐技术,根据用户输入的查询,向用户提供一系列与原查询语义相关的候选查询<sup>[27]</sup>。传统的图像/视频检索系统借鉴文本检索中的查询推荐技术,利用文档、查询日志、点击链接等数据,针对不同性质的数据设计相应的分析模型,如查询流图模型<sup>[28]</sup>、词项转移图模型<sup>[29]</sup>、排序学习模型<sup>[30]</sup>等,从数据中挖掘出关键词之间的语义联系,生成若干候选查询词。例如,基于文档的查询推荐方法利用统计模型从包含查询词的文档数据或人工编辑语料(如 Wikipedia, Wordnet 等)中挖掘出与用户查询词相关的词或短语,利用其构建推荐查询。基于查询日志的方法通过分析搜索引擎的查询日志,挖掘查询之间的关联关系,发现过往搜索中出现过的关联查询,利用其构建推荐查询。查询日志是众多用户在使用搜索引擎进行查询操作时的日志记录,记录了用户的搜索行为,例如使用的查询、点击搜索结果等。大量的查询日志蕴含着查询间的丰富关联,现有方法通过分析不同查询之间的各类型关联,如查询在搜索过程(session)中的共现频率、查询共有的相同或相似点击 URL 的数量、查询出现频率随时间分布的相关性等,计算查询间的关联强度,指导查询推荐的生成。

在图像/视频检索中,用户的检索目标为图像或视频片段,其语义内容远比若干查询词复杂。因此,仅推荐查询词往往不能帮助用户构建合适的查询以清楚地表达信息需求。针对多媒体检索中查询词存在模糊、歧义等问题,Zha 等人<sup>[31-32]</sup>提出了联合图片和文字的视觉查询推荐技术,针对用户的查询词,自动推荐若干语义相关的新查询词以及描述新查询词的图片,形成了“词-图”相结合的多模态查询推荐,如图 3 所示。该技术利用图片查询的视觉呈现帮助

用户明确信息需求,综合考虑了图片的典型性与多样性,从多侧面对新查询进行视觉呈现,便于用户构建查询,进而结合用户选择的“词-图”新查询,融合视觉与文本特征改进检索.相比于查询词,多模态的查询更加有助于用户信息需求的表达以及媒体内容的查找,因而更加适用于多媒体信息检索.主流商业搜索引擎均提供多模态查询功能.例如谷歌、百度等搜索引擎支持基于关键词与图片示例的混合查询,支持用户在基于关键词检索返回的图片集中选择感兴趣图片作为查询示例,进一步查找相似图片.多模态查询被广泛应用于各类多媒体应用,如商品图片搜索<sup>[33]</sup>、多媒体问答<sup>[34]</sup>等.



Fig. 3 The interface of visual query suggestion system<sup>[31]</sup>  
图3 视觉查询推荐系统交互界面<sup>[31]</sup>

2.2 查询交互

提供便捷有效的查询交互是帮助用户清楚表达检索需求的另一途径.研究人员设计了若干新颖的查询方式,突破了多媒体检索中“查询词/示例+查询框”的传统模式. Zavesky 和 Chang<sup>[35]</sup> 研发的 CuZero 视频检索原型系统提供“Semantic Panel”帮助用户构建关键词组合查询,如图 4(a)所示.系统自动推荐相关的语义概念,呈现于“Semantic Panel”,其中每个概念所占的矩形框尺寸对应于该概念在查询中的权重.用户根据其信息需求,调整矩形框尺寸更新各概念在检索中的权重,便捷地构建合适的概念组合查询. Xu 等人<sup>[36]</sup> 设计了基于“Concept Map”的检索系统,支持用户在“Concept Map”的不同位置输入若干关键词,形成包含关键词相对位置关系的精细查询.针对每个关键词提供若干幅图片示例供用户选择,帮助用户进一步描述检索需求,如图 4(b)所示. Wang 和 Hua<sup>[37]</sup> 研发了基于“Color Map”的检索系统,支持用户在“Color Map”的不同位置涂鸦颜色,形成对目标图像的主体颜色及其空间分布的描述,用以检索具有相似色彩分布的图像,如图 4(c)所示.



Fig. 4 The interfaces of CuZero, Concept Map and Color Map systems<sup>[35-37]</sup>  
图4 CuZero, Concept Map and Color Map 系统交互界面<sup>[35-37]</sup>

近年来,随着手机、平板电脑等移动智能终端的普及以及移动互联网的发展,图像视频检索逐渐从PC端延伸到移动端.移动智能终端与图像视频检索技术的有机融合,改变了信息检索、获取及利用的方式,用户可以更加便捷地获取多样化的媒体信息.利用移动设备的摄录功能,用户可以快速、方便地采集兴趣目标的图像/视频作为查询输入,利用移动视觉搜索技术查找关联信息<sup>[38]</sup>.移动视觉搜索拥有巨大的应用前景,互联网巨头,如谷歌、百度、阿里巴巴等相继推出了移动视觉搜索服务.移动视觉搜索涉及多方面的研究内容,如系统架构、紧凑视觉描述子、

视觉匹配、检索模型、结果评价、视觉对象知识库等.围绕这些内容,研究人员开展了大量的研究工作,推动了移动视觉搜索的进步.本文仅介绍查询交互方面的相关工作.目前移动视觉搜索应用中的查询需求大多与用户日常生活相关,如搜索相似/相同的商品、图书、人物、食品、景点等.由于移动设备拍摄的查询图像往往包含复杂的背景和丰富的前景,待检索目标在查询图像中主体不突出,导致搜索系统难以展开具有针对性的信息查找,同时也造成移动端计算资源、网络通信资源的浪费.针对于此,研究人员利用智能设备的交互便捷性,开发了多种面向移动

视觉搜索的查询方法,支持用户在查询示例上进行交互,明确检索目标.例如 Sang 等人<sup>[39]</sup>设计了一种查询交互方式,支持用户在拍摄的图像上通过裁剪、画线、套索等操作圈出兴趣目标,利于目标查找.实验结果表明:套索操作是一种较为自然而有效的交互方式. Kawano 与 Yanai<sup>[40]</sup>开发了基于手机拍照的食物识别/检索系统. Kiapour 等人<sup>[41]</sup>研究了基于手机街拍的服装检索任务,其中的查询交互部分支持用户采用包围框指明待检索对象. You 等人<sup>[42]</sup>开发的移动搜索系统支持用户在查询图像上通过涂鸦线条区分背景和待检索的前景. Ngo 等人<sup>[43]</sup>针对用户的查询图片自动推荐图片中的若干区域作为待检索对象. Yu 等人<sup>[44]</sup>研究了基于手机拍照的地点搜索技术,提出了一种自动的查询推荐方法,指导用户拍摄最佳的查询示例. Zhao 等人<sup>[45]</sup>挖掘“用户-地点-查询”三元关系,提出一种张量函数学习算法,用于向用户推荐查询. 用户与系统间的查询交互能够帮助用户明确检索的主体目标,提高检索的成功率,改善用户体验.

2.3 草图查询

随着触屏技术的发展与触屏设备的普及,手绘草图成为用户表达信息的便捷方式. 依靠记忆与模仿勾勒草图进行信息表达是人类与生俱来的能力,利用草图进行多媒体信息检索是一种自然的人机交互方式,具有广阔的应用前景. 用户勾勒的草图具有高度的抽象性与不确定性. 例如用户描绘的对象轮廓是对检索对象高度抽象的描述,且存在不同程度的不规则形变. 针对同一对象,不同用户描绘的草图往往差异较大. 因此,以用户勾勒的充满创造力的草图作为查询,给检索提出了更大的挑战.

较之于基于关键字/查询示例的检索,基于草图的检索技术尚处于初步的研究阶段. 草图检索需要处理特征表示、索引结构等方面的难题<sup>[46]</sup>. 现有的草图特征表示方法根据其特征提取单元的不同可以总结为基于笔划描述的特征表示、基于组合图元的特征表示以及基于形状的特征表示<sup>[47]</sup>. Cao 等人<sup>[48]</sup>开发了 MindFinder 草图检索系统,如图 5 所示,针对由笔划形成的草图线条,构建一种融合像素坐标与方向角信息的边缘像素词典,生成包含边缘与方向信息的词袋模型,形成了简化的形状特征描述,同时保持了轮廓的空间信息. 针对用户描绘的由一个或多个图元构成的草图查询,可采用基于组合图元的特征表示与检索,首先识别草图与图像中的基本图元,通过不同层次的图形元素抽象,形成统一的特

征表示,进而利用图元之间的空间关系进行检索,如图形元素的相对位置关系、相对方位、相对旋转等<sup>[47]</sup>. 基于形状的特征主要提取草图轮廓的全局或局部描述,形成草图的外在形状特征. Eitz 等人<sup>[49]</sup>采用词袋模型,实验评估了形状内容描述子、星点描述子、改进的标准方向梯度直方图描述子等形状特征. 实验结果表明,改进的标准方向梯度直方图描述子具有相对较好的检索效果.

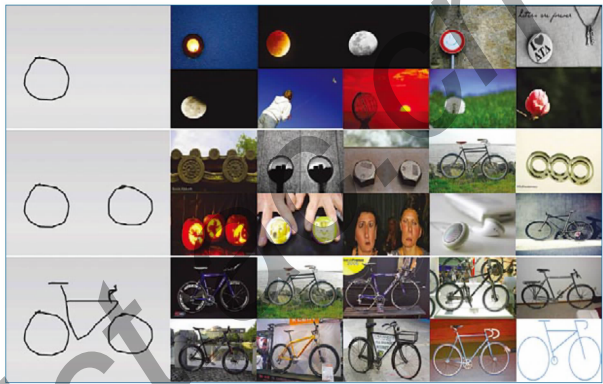


Fig. 5 Illustration of sketch queries and the corresponding top 10 search results of MindFinder system<sup>[48]</sup>

图5 MindFider 系统草图查询及搜索结果示例<sup>[48]</sup>

在草图索引方面,早期的草图检索系统采用线性的索引结构,只能处理小规模数据库. 近年来,为支持面向草图查询的大规模图片索引与快速查找,研究人员相继提出了一些新的草图索引技术. 例如, MindFinder 系统<sup>[48]</sup>采用一种类似文档倒排的 Edgel Index 索引结构,实现了在 200 万幅图片数据库上仅需几百毫秒的检索效率. Xiao 等人<sup>[50]</sup>开发了 IdeaPanel 交互式草图检索系统,支持用户根据检索返回图像修改草图重新检索,实现百万级图片实时交互式检索. Sun 等人<sup>[51]</sup>采用基于 K-中心聚类的局部敏感 Hash 算法(K-medoids locality sensitive hashing),支持大规模索引,同时采用多探寻(multi-probe)策略,有效减少了 Hash 表数量,大幅节省了存储开销,实现了 20 亿规模图像数据库的实时草图检索.

基于特征匹配的草图检索技术一般要求用户绘制的草图接近检索目标,对于几何形状与检索目标差异较大的草图,难以获得准确的检索结果. 针对此问题,研究人员提出基于草图语义的检索方法,对草图进行语义分类,以类别作为关键字检索图像,进而利用形状等视觉特征改进检索结果. Schneider 和 Tuytelaars<sup>[52]</sup>提出了基于 Fisher Vector 的草图识别方法. Sun 等人<sup>[53]</sup>针对用户草图类内差异大、类间

区分度小等问题,提出了一种基于查询自适应的形状主体模型用于草图识别与检索. Yanik 和 Sezgin 等人<sup>[54]</sup>将主动学习方法应用于草图识别,降低识别模型训练对标注样本的需求量. Yu 等人<sup>[55]</sup>设计了 Sketch-a-Net 深度神经网络,将深度学习技术应用于草图识别,取得了显著的识别效果. Sangkloy 等人<sup>[56]</sup>构建了一个包含大量物体图片和草图的数据库,含有 125 类物体的 12 500 幅图片和 75 000 余幅草图,以及图片与草图间的对应关系. 该数据集可以用于训练跨模态卷积神经网络,学习图片与草图的共享特征空间,有效支持草图检索与识别. 将草图语义融入草图检索能够降低对用户绘图的相似度要求,改善用户体验,提升检索的鲁棒性与准确性,为基于草图的多媒体检索开拓了新的发展方向.

## 2.4 跨媒体查询

多媒体和互联网的空前繁荣促使从不同渠道获取的文本、图像和视频等不同形态的媒体信息及与之相关的自然、社会属性信息紧密混合在一起,彼此间存在错综复杂的交叉关联,形成一种新的媒体表现形式,即跨媒体<sup>[57]</sup>. 在跨媒体信息环境下,用户提交一种媒体对象作为查询,检索系统不但可以返回相同种类的相似对象,而且还能返回其他种类的媒体对象,形成更为全面丰富的信息呈现,如利用图像查找语义相关的音频或视频片段<sup>[58]</sup>等. 面向跨媒体查询,检索系统需要克服不同媒体之间的“鸿沟”,最大限度地挖掘不同媒体之间相互表达、相互补充的语义关联性和协同效应,构建不同种类媒体数据的一致性表达与相似性度量,建立能够有效处理跨媒体查询和查找跨媒体信息的模型.

近年来,大量的跨媒体表达与度量方法被相继提出,主要包括子空间学习方法、度量学习方法、主题模型方法以及新兴的基于深度学习的方法. 其中,子空间学习方法旨在构造一个能够表达不同种类媒体数据的共同子空间,使得不同种类的媒体数据在此空间中具有可比性,从而可以采用传统的度量计算查询对象与检索对象的相似度,进行跨媒体检索. 早期的子空间学习算法有典型相关分析(canonical correlation analysis, CCA)<sup>[59]</sup>、双线性模型(bilinear model)<sup>[60]</sup>、跨模态因子分析方法<sup>[61]</sup>等. Mahadevan 等人<sup>[62]</sup>将流形学习引入共同子空间学习,最大化不同媒体数据间的相关性,同时保持数据的局部近邻关系. Sharma 等人<sup>[63]</sup>提出一种广义多视图分析框架,通过引入类别信息以提升子空间的判别性. 度量学习方法旨在建立不同媒体数据之间合理的距离测

度,使得相似的数据度量距离小、不相似的数据度量距离大. 其可以利用数据之间的相似/不相似关系,也可以利用数据的排序信息进行距离测度的学习. Zhai 等人<sup>[64]</sup>提出一种多视图距离测度学习算法,保持数据分布的局部光滑性与全局一致性. Lu 等人<sup>[65]</sup>提出一种跨模态排序算法,利用隐结构 SVM 模型学习距离测度,支持多种排序准则的排序优化. Wu 等人<sup>[66]</sup>提出一种基于双向排序学习的跨媒体表示模型,有效利用图像到文本以及文本到图像的双向排序信息. 主题模型方法利用主题学习模型挖掘不同媒体数据之间的相关性与一致性. Zheng 等人<sup>[67]</sup>提出一种监督的文档神经自回归分布估计模型,在传统的文档神经自回归分布模型中引入语义类别监督信息,提升了隐主题特征的判别力,学习了视觉单词、文本单词和语义类别之间的共同特征表达. Liao 等人<sup>[68]</sup>提出一种非参数贝叶斯多模态主体模型,构建了一种复合非参数贝叶斯多模态先验用于刻画模态内部相似性与模态间相关性. Wang 等人<sup>[69]</sup>提出了一种多模态共同主题强化模型,建立跨模态联合概率图模型,建模不同模态数据隐含主题之间的相关性,在图像与文本的跨模态检索任务上进行了实验验证. 近年来,深度学习技术被应用于跨媒体数据特征学习,利用深度神经网络的特征抽象能力,学习不同种类媒体数据的统一特征表达. Wei 等人<sup>[70]</sup>采用卷积神经网络进行图像特征学习,将获得的深度特征用于跨媒体检索,实验对比了深度特征与传统视觉特征的检索性能. 实验结果表明,深度特征具有相对较高的检索精度. Ma 等人<sup>[71]</sup>提出一种多模态卷积神经网络,建模不同颗粒度的文本片段与图像之间的关联关系,进行文本和视觉的联合特征学习,提升了图像与文本双向检索的性能.

随着跨媒体数据规模的增长,如何构建高效的跨媒体索引以支持大规模快速查找成为跨媒体检索需要解决的另一重要问题. 跨媒体 Hash 是解决此问题的有效途径. 其通过设计 Hash 函数将不同种类的媒体数据映射到共同的 Hash 空间,尽可能地维持数据的近邻关系,进而通过比对数据的 Hash 值进行快速查找. Wu 等人<sup>[72]</sup>提出一种稀疏多模态 Hash 方法,采用超图表达模态内部相似性与模态之间相关性,通过超图正则化稀疏编码学习多模态数据的联合词典,进而生成多模态数据的 Hash 编码. Ding 等人<sup>[73]</sup>提出一种基于排序保持的跨模态 Hash 算法,利用数据的排序作为监督信息,设计了一种基于回归的排序保持损失函数,学习跨模态

Hash 函数与 Hash 码. Cao 等人<sup>[74]</sup>提出了一种端到端的跨模态深度 Hash 算法,采用混合深度网络结构,包含一个卷积网络和一个递归网络,分别用于学习视觉 Hash 函数和文本 Hash 函数,以及一个融合网络用于学习视觉与文本模态的共同汉明空间.得益于深度神经网络的建模能力,基于深度学习的跨媒体 Hash 方法具有良好的实验效果和应用潜力.

## 3 反馈技术

### 3.1 相关反馈

在检索流程中引入用户反馈是提升检索精度的有效途径.检索系统支持用户在输入查询后继续参与检索过程,对当前检索结果标记出与其检索意图相关/无关的样本,明确其信息需求,系统进而根据用户的反馈改进检索模型,调整检索策略,更新检索结果.通过用户与系统的交互,系统能够实时地、动态地了解用户的信息需求及其对数据的语义标记,提升系统对用户需求以及数据的理解能力,增强检索结果中相关样本的响应而抑制无关样本的出现,使得检索结果逐步贴近用户的期望,最终满足用户的检索需求.

20 世纪 90 年代末,Rui 和 Huang 等人<sup>[23]</sup>提出了基于相关反馈的交互式多媒体检索方法,为多媒体检索的发展开辟了新的道路.基于用户反馈的交互式检索广受关注,国际权威视频检索竞赛 TRECVID 一直将交互式视频检索作为每年的评测任务之一<sup>[75]</sup>.迄今,研究人员针对多媒体检索中的相关反馈技术开展了大量研究,以最大化反馈信息获取与利用且最小化用户交互量为目标,相继提出了一系列相关反馈技术,包括基于查询点移动策略、权值更新策略、机器学习、主动学习等反馈技术,推动了交互式多媒体检索的发展.早期的反馈技术主要采用查询点移动策略<sup>[76]</sup>和权值更新策略<sup>[77]</sup>.前者根据用户反馈修改检索空间中的查询点位置,使其尽可能地靠近相关样本且远离无关样本,依据新的查询点重新排列检索结果;后者根据反馈调整表征查询的各特征向量权重,形成新的查询特征表示,更新检索结果.此类方法主要关注如何根据用户反馈提高检索结果的排序质量,对检索模型没有本质的改进.

近年来,研究人员将机器学习理论与方法引入相关反馈中,将检索转化为不同类型的监督学习问题,设计相应的机器学习模型,基于用户标记样本训练模型,指导新的检索结果的生成<sup>[78]</sup>.例如考虑到

支持向量机(SVM)在有限样本条件下良好的推广能力,Zhang 等人<sup>[79]</sup>利用 SVM 从用户反馈的相关/无关样本中学习二类分类模型.Chen 等人<sup>[80]</sup>只考虑相关样本,采用单类 SVM 模型.Tong 等人<sup>[81]</sup>提出了基于 SVM 模型主动学习的相关反馈算法,选择临近 SVM 分类边界的样本供用户反馈标记,达到最大限度地减少模型解释空间尺寸的目的,实现在有限用户反馈条件下的尽可能大的信息收益.Zha 等人<sup>[82]</sup>提出了一种结合样本分布结构性的主动学习算法,刻画了样本分布的局部几何结构以及邻近样本的语义相似性,估计样本对模型的改进作用,同时考虑了样本的相关度、局域密度、不确定性、多样性等信息,综合多种信息遴选最佳待标注样本供用户标注,仅需较少的用户标注,有效提升了模型的性能.交互式图像视频检索的初始结果中往往仅有少数甚至没有相关样本,尤其在检索复杂查询时,导致相关反馈效果不佳或失效.针对于此,Yuan 等人<sup>[24]</sup>提出了“部分相关”反馈方法,支持用户在检索结果中标记相关/无关样本的同时,标记与其需求虽非整体相关,但却“部分相关”的样本,自动挖掘此类样本中与用户需求有关的信息,对用户需求进行建模,改善检索结果.在实际应用中,用户往往需要获取包含复杂内容的图像视频数据,面向复杂查询的相关反馈成为新的研究热点.

### 3.2 属性反馈

随着新反馈技术的不断提出,多媒体检索的性能获得了逐步提升.然而,计算机感知的底层特征与人们认知的高层语义之间存在“语义鸿沟”,依然影响着检索系统对用户意图的建模精度以及对多媒体数据的理解准确度,制约了多媒体检索的发展.为克服“语义鸿沟”,研究人员提出利用视觉属性作为图像视频内容的中层语义描述,连接底层特征与高层语义.视觉属性即对象固有的视觉特性,描述对象组成部分、形状、材质等,如鼻子、腿、方形、毛绒的等<sup>[83]</sup>.视觉属性比语义概念易于通过底层特征建模,比底层特征易于被人们理解.得益于其固有优势,视觉属性被广泛应用图像视频分析与检索中.研究人员提出了一系列属性建模方法<sup>[84-86]</sup>,基于属性模型的输出形成图像视频的中层特征表达,用于分析与检索.Douze 等人<sup>[87]</sup>融合属性特征与 Fisher 向量进行图像检索.Scheirer 等人<sup>[88]</sup>构建多属性空间,获得更好的属性特征,提高检索精度.Liu 等人<sup>[89]</sup>将属性应用于服装图像分析与检索.

考虑到视觉属性的优点,研究人员提出了基于

属性的反馈技术,利用用户对属性的反馈构成其检索意图的中层语义描述,利用属性作为连接用户检索意图与图像视频数据的中间桥梁。Zhang 等人<sup>[90]</sup>提出了一种属性反馈方法,自动挖掘有助于改进当前检索的属性,支持用户在属性上进行相关/无关反馈,改变了只支持用户在图像/视频样本上进行反馈的传统机制,如图 6 所示。用户在属性上的反馈构成了对其检索目标的属性描述,如“有腿、有翅膀、没有轮子”等,有利于系统理解用户需求。针对某一相关的属性,支持用户标记检索返回图像在该属性上与用户期望的图像是否相似,如某幅检索返回图像中的“鼻子”与用户期望的相似/不相似,以此得到对用户需求更为精细的属性描述。进而,检索系统基于属性模型,结合用户反馈,改善检索结果。Zhang 等人<sup>[91]</sup>进一步构建了一种集语义概念与属性于一体的语义树,根据概念之间的层级关系以及属性与概

念的从属关系,将概念与属性有机地组织起来。基于概念与属性模型,形成对图像内容的层次化语义描述,多粒度地刻画了图像内容。在此基础上,提出了样本与属性混合反馈机制,支持用户标记相关/无关样本及相关/无关属性,利用用户反馈有效提升了检索结果与用户意图的相关度。Kovashka 等人<sup>[92]</sup>和 Yu 等人<sup>[93]</sup>提出了基于相对属性的交互式图像检索系统。相对属性是指就某一属性而言,某一图像与其他图像相比在该属性上的相对比较关系,例如某一图像中的鞋子根部比其他图像中的鞋根“更高”。相对属性较二值属性能够表达更丰富的信息,符合用户对检索需求的表述,如图 7 所示,用户想搜索一双与之相似却更华丽的鞋子。检索系统针对每个属性训练一个排序模型,用于判断任意 2 幅图像在该属性上的比较关系,基于属性的排序模型,结合用户相对属性反馈,更新检索结果序列。

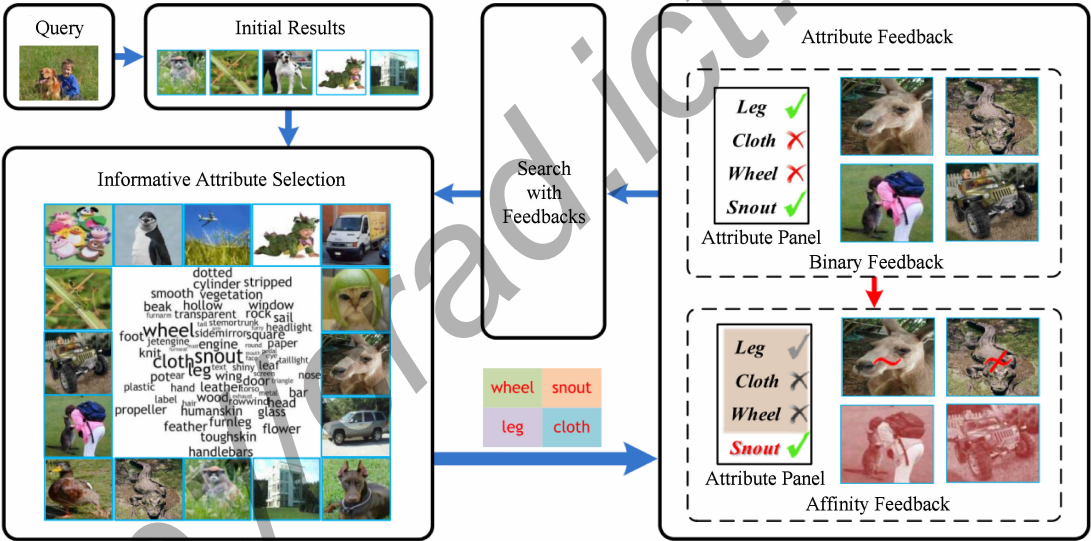


Fig. 6 The framework of attribute feedback system<sup>[90]</sup>

图 6 属性反馈系统框架图<sup>[90]</sup>

描述物体的属性拓宽为描述场景、事件等各类分析目标的属性,如以场景为分析目标时场景内的物体即为属性,以事件为目标时事件的组成部分即为属性,属性反馈的类型由二值反馈发展为相对比较反馈等更细粒度的信息反馈。

3.3 隐式反馈

充分利用用户隐式反馈数据是提升检索系统性能的另一有效途径。用户的检索历史与交互行为是隐式反馈信息的主要来源之一,通常包括输入的查询、点击的网页、停留的时间等。隐式反馈数据蕴含着用户的偏好,为理解用户检索意图提供了线索。尽管隐式反馈数据存在大量噪声,不如显示反馈精确,



Fig. 7 Illustration of relative attribute feedback<sup>[92]</sup>

图 7 相对属性反馈示例<sup>[92]</sup>

随着研究的推进,属性的获取由人为定义与手工标注发展为自动挖掘与识别,属性集合的规模由几十种属性增长至成百上千种属性,属性的范畴由

但在实际应用中,隐式反馈远比显示反馈丰富,大量存在于检索系统中,具有数据规模大、应用场景广等优势,同时也不要要求用户对检索结果进行反馈,减轻了用户操作负担。

近年来,基于隐式反馈的信息检索成为研究热点,大量的研究工作围绕着如何挖掘利用隐式反馈数据改进检索而展开。作为最常用的隐式反馈数据,用户的点击数据(click through)记录着用户在搜索过程中对文档、图像等对象的点击历史,从一定程度上反映着文档、图像等与用户查询及检索意图的关联强度。大量的研究表明,利用点击数据可以有效提升信息检索的精度。点击数据早先在文本检索中得到研究与应用。例如,Agichtein 等人<sup>[94]</sup>通过实验表明,在网页检索中利用用户点击数据可以使得检索结果的相关度提高约 30%。Joachims<sup>[95]</sup>利用点击数据训练排序 SVM 模型用于提升检索系统的排序质量。Jiang 等人<sup>[96]</sup>以微软 Bing 搜索引擎的日志数据为基础,提出一种回归模型预测多粒度的搜索结果满意度。在图像视频检索中利用点击数据的研究虽然起步相对较晚,但进展迅速。点击数据有助于克服图像视频检索中的“意图鸿沟”和“语义鸿沟”<sup>[97]</sup>,被广泛用于图像视频排序、重排序、分类等多个环节,展示出良好的实验效果。Hua 等人<sup>[97]</sup>以商业图像搜索引擎的点击数据为基础,构建了一个公开的大规模图像搜索点击数据集。Jain 和 Varma<sup>[98]</sup>利用点击数据训练高斯过程回归模型,预测检索返回图片的归一化点击率,对检索结果进行重排序。Yu 等人<sup>[99]</sup>结合点击数据与图像视觉特征进行图像搜索重排序,分别利用点击数据和多种视觉特征构建语义流形与视觉流形,提出一种基于多视图超图学习的重排序算法,在重排序学习中融合了语义流形与视觉流形。O'Hare 等人<sup>[100]</sup>综合使用点击数据与鼠标悬停记录,提出了多种隐式反馈特征,结合排序学习框架,实验验证了隐式反馈特征对网络图像搜索的改进作用。Wu 等人<sup>[101]</sup>提出使用点击图表达点击数据,充分挖掘数据对象之间的隐式关联。图中的顶点对应查询词或图片,联接边强度表示图片与查询词之间的点击频率。结合点击图,采用随机游走模型学习多模态特征,提升了检索精度。Jiang 等人<sup>[102]</sup>采用多层感知机和双通道递归神经网络从点击数据中“端到端”地学习用户查询与图像/视频的语义概念之间的映射关系,用于提高图像/视频检索的质量。在点击数据取得成功应用的同时,也应注意到点击数据具有明显的长尾现象,存在大量用户点击频度

低的长尾查询。如何提高长尾查询的检索质量是有待解决的问题。

不同于上述工作利用检索系统记录的隐式反馈数据,另有一些研究工作利用脑机接口、眼动仪等外部设备采集用户在检索过程中的各类行为数据,用于指导检索的进行。例如 Wang 等人<sup>[103]</sup>利用基于 EEG 的脑机接口系统采集用户浏览图片时的脑电信号,进而分析出用户感兴趣的目标。Papadopoulos 等人<sup>[104]</sup>利用眼动仪跟踪和记录用户浏览图片时的注视行为,提出了多种注视信号特征,用于提高图像检索结果与用户信息需求的相关度。Kauppi 等人<sup>[105]</sup>在检索过程中综合使用脑磁信号与眼动数据以提高检索的质量。

## 4 探索式搜索

在信息检索中,往往存在用户不熟悉检索目标领域、不确定检索目标的路径,甚至不确定检索目标等情况。在此类情况下,用户需要进行信息探索。用户搜索行为分析相关研究表明,用户搜索行为中有相当比例属于探索式搜索<sup>[106]</sup>。探索式搜索的定义是:可用来描述一种开放的、持续的、多方面的信息搜寻的问题情景和具有机会性、反复性、多策略的信息搜寻过程。探索式搜索是交织着查找、学习、调查的反复的、启发式的交互过程。与传统搜索相比,探索式搜索的特征是最初的信息需求是模糊的,缺乏检索对象的相关知识,经多次交互,目标发生变化,检索终止的条件不清晰。探索式搜索更加依赖于用户与检索系统的交互。近年来,研究人员针对面向多媒体数据的探索式搜索技术开展了初步的研究<sup>[107-109]</sup>。如何增强用户与检索系统之间的交互能力,支持用户进行信息探索与发现,协助用户在繁杂的信息空间中完成高效的信息发现,是探索式搜索研究的关键。

## 5 总结与展望

随着图像视频等多媒体数据规模的爆炸式增长以及各类媒体智能应用需求的日益迫切,多媒体信息检索成为学术界的研究热点和工业界的关注焦点。发展多媒体检索中的查询与反馈技术是克服“意图鸿沟”和“语义鸿沟”的有效途径。本文在现有文献的基础上,介绍了多媒体检索查询与反馈技术发展演变脉络,综述了不同时期的技术革新。其中,

查询的输入方式由经典的“查询框”输入衍生出交互式查询构建以及草图勾勒查询. 随着新颖查询方式的应用, 查询的模态也不断被刷新, 由单一的文本或视觉模态发展为多模态查询、跨媒体查询. 多种查询模态的有效融合和不同种类媒体数据的跨越检索, 显著提升了多媒体检索的准确度与覆盖率. 另一方面, 反馈的形式由样本相关反馈衍生出语义属性相关反馈, 反馈的类型由二值反馈发展为多值反馈、比较反馈等更细粒度的信息反馈. 与此同时, 隐式反馈因具有数据规模大、应用场景广等固有优势, 吸引着的越来越多的关注. 在算法方面, 深度学习技术在查询与反馈的分析与建模中展示出良好的效果, 逐渐成为主流的查询与反馈算法.

过往的研究促进了技术的进步, 然而现阶段的技术远非完善, 尚存诸多问题亟待研究, 例如: 1) 用户检索行为分析与建模问题, 包括用户在检索中的客观行为分析、认知要素分析以及行为建模方法与演化机制等; 2) 多媒体检索与移动智能设备融合问题, 包括在检索中对移动智能设备多通道信息采集功能与环境感知能力的充分利用以及对多样化查询与多模态数据的综合处理等; 3) 用户、模型与系统协同问题, 包括探索更加合理的多媒体信息人机协同处理机制、更加有效的交互式分析与检索技术等.

## 参 考 文 献

- [1] Snoek C G M, Worring M. Concept-based video retrieval [J]. *Foundations and Trends in Information Retrieval*, 2008, 2(4): 215-322
- [2] Datta R, Joshi D, Li Jia, et al. Image retrieval: Ideas, influences, and trends of the new age [J]. *ACM Computing Surveys*, 2008, 40(2): 5
- [3] Zhang Lei, Rui Yong. Image search-from thousands to billions in 20 years [J]. *ACM Trans on Multimedia Computing Communications & Applications*, 2013, 9(1s): 36
- [4] Wang Xinjing, Xu Zheng, Zhang Lei, et al. Towards indexing representative images on the Web [C] //Proc of the 20th ACM Int Conf on Multimedia. New York: ACM, 2012: 1229-1238
- [5] Flickner M, Sawhney H, Niblack W, et al. Query by image and video content: The QBIC system [J]. *Computer*, 1995, 28(9): 23-32
- [6] Smith J R. VisualSeek: A fully automated content-based image query system [C] //Proc of the 4th ACM Int Conf on Multimedia. New York: ACM, 1970: 87-98
- [7] Huang T, Mehrotra S, Ramchandran K. Multimedia analysis and retrieval system (MARS) project [C] //Proc of 33rd Clinic on Library Application of Data Proc-Digital Image Access and Retrieval. Urbana, Illinois: Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 1997: 100-117
- [8] Lowe D G. Object recognition from local scale-invariant features [C] //Proc of the Int Conf on Computer Vision. Piscataway, NJ: IEEE, 1999: 1150-1157
- [9] Andoni A, Indyk P. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions [J]. *Annual Symp on Foundations of Computer Science*, 2006, 51(1): 459-468
- [10] Jiang Yugang, Ngo C W, Yang Jun. Towards optimal bag-of-features for object categorization and semantic video retrieval [C] //Proc of the 6th ACM Int Conf on Image and Video Retrieval. New York: ACM, 2007: 494-501
- [11] Snoek C, Sande K, Rooij O D, et al. The MediaMill TRECVID 2009 semantic video search engine [C] //Proc of TRECVID Workshop. New York: ACM, 2009: 1-14
- [12] Yanagawa S F, Chang L, Kennedy W, et al. Columbia university's baseline detectors for 374 lscm semantic visual concepts [R]. New York: Columbia University, 2007
- [13] Mei Tao, Zha Zhengjun, Liu Yuan, et al. MSRA at TRECVID 2008 high-level feature extraction and automatic search [C] //Proc of TRECVID Working Notes. New York: ACM, 2008: 1-11
- [14] Wang Xinjing, Zhang Lei, Jing Fei, et al. Annosearch: Image auto-annotation by search [C] //Proc of the 19th IEEE Conf on Computer Vision and Pattern Recognition (CVPR 2006). Piscataway, NJ: IEEE, 2006: 1483-1490
- [15] Zhao Wanlei, Wu Xiao, Ngo C W. On the annotation of Web videos by efficient near-duplicate search [J]. *IEEE Trans on Multimedia*, 2010, 12(5): 448-461
- [16] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [J]. *Advances in Neural Information Processing Systems*, 2012, 25(2): 1-9
- [17] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks [C] //Proc of the 27th IEEE Conf on Computer Vision and Pattern Recognition (CVPR 2014). Piscataway, NJ: IEEE, 2014: 1725-1732
- [18] Donahue J, Anne H L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description [C] //Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition (CVPR 2015). Piscataway, NJ: IEEE, 2015: 2625-2634
- [19] Kulis B, Grauman K. Kernelized locality-sensitive hashing for scalable image search [C] //Proc of the 12th Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2009: 2130-2137

- [20] Wang Jingdong, Wang Jing, Zeng Gang, et al. Fast Neighborhood Graph Search Using Cartesian Concatenation [M]. Berlin: Springer, 2013: 2128-2135
- [21] Mei Tao, Rui Yong, Li Shipeng, et al. Multimedia search reranking: A literature survey [J]. ACM Computing Surveys, 2014, 46(3): 1-37
- [22] Li Hang. Learning to Rank for Information Retrieval and Natural Language Processing [M]. San Rafael, CA: Morgan & Claypool, 2011
- [23] Rui Yong, Huang T S, Ortega M, et al. Relevance feedback: A power tool for interactive content-based image retrieval [J]. IEEE Trans on Circuits & Systems for Video Technology, 2000, 33(2): 644-655
- [24] Yuan Jin, Zha Zhengjun, Zheng Yantao, et al. Utilizing related samples to enhance interactive concept-based video search [J]. IEEE Trans on Multimedia, 2011, 13(6): 1343-1355
- [25] Zhang Hanwang, Zha Zhengjun, Yan Shuicheng, et al. Attribute feedback [C] //Proc of the 20th ACM Int Conf on Multimedia. New York: ACM, 2012: 79-88
- [26] Strohmaier M, Kröll M, Körner C. Intentional query suggestion: Making user goals more explicit during search [C] //Proc of the 2009 Workshop on Web Search Click Data. New York: ACM, 2009: 68-74
- [27] Luo Cheng, Liu Yiqun, Zhang Min, et al. Query recommendation based on user intent recognition [J]. Journal of Chinese Information Processing, 2014, 28(1): 64-72 (in Chinese)  
(罗成, 刘奕群, 张敏, 等. 基于用户意图识别的查询推荐研究[J]. 中文信息学报, 2014, 28(1): 64-72)
- [28] Boldi P, Bonchi F, Castillo C, et al. The query-flow graph: Model and applications [C] //Proc of the 17th ACM Conf on Information and Knowledge Management. New York: ACM, 2008: 609-618
- [29] Song Yang, Zhou Dengyong, He Liwei. Query suggestion by constructing term-transition graphs [C] //Proc of the 5th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2012: 353-362
- [30] Zhu Xiaofei, Guo Jiafeng, Cheng Xueqi, et al. Query recommendation based on manifold ranking [J]. Journal of Chinese Information Processing, 2011, 25(2): 38-44 (in Chinese)  
(朱小飞, 郭嘉丰, 程学旗, 等. 基于流形排序的查询推荐方法[J]. 中文信息学报, 2011, 25(2): 38-44)
- [31] Zha Zhengjun, Yang Linjun, Mei Tao, et al. Visual query suggestion [C] //Proc of the 17th ACM Int Conf on Multimedia. New York: ACM, 2009: 15-24
- [32] Zha Zhengjun, Yang Linjun, Mei Tao, et al. Visual query suggestion: Towards capturing user intent in Internet image search [J]. ACM Trans on Multimedia Computing Communications & Applications, 2010, 6(3): 219-239
- [33] Lu Shiyang, Mei Tao, Wang Jingdong, et al. Exploratory product image search with circle-to-search interaction [J]. IEEE Trans on Circuits and Systems for Video Technology, 2015, 25(7): 1190-1202
- [34] Zhang Wei, Pang Lei, Ngo C W. Snap-and-ask: Answering multimodal question by naming visual instance [C] //Proc of the 20th ACM Int Conf on Multimedia. New York: ACM, 2012: 609-618
- [35] Zavesky E, Chang S F. CuZero: Embracing the frontier of interactive visual search for informed users [C] //Proc of the ACM Int Conf on Multimedia Information Retrieval. New York: ACM, 2008: 237-244
- [36] Xu Hao, Wang Jingdong, Hua Xiansheng, et al. Image search by concept map [C] //Proc of the 33rd Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2010: 275-282
- [37] Wang Jingdong, Hua Xiansheng. Interactive image search by color map [J]. ACM Trans on Intelligent Systems and Technology, 2011, 3(1): 12
- [38] Duan Lingyu, Huang Tiejun, Gao Wen. Technical research and standardization in mobile visual search [J]. Information and Communications Technologies, 2012, 6(2): 51-58 (in Chinese)  
(段凌云, 黄铁军, 高文. 移动视觉搜索技术与标准化进展[J]. 信息通信技术, 2012, 6(2): 51-58)
- [39] Sang Jitao, Mei Tao, Xu Yingqing, et al. Interaction design for mobile visual search [J]. IEEE Trans on Multimedia, 2013, 15(7): 1665-1676
- [40] Kawano Y, Yanai K. Foodcam: A real-time food recognition system on a smartphone [J]. Multimedia Tools and Applications, 2015, 74(14): 5263-5287
- [41] Kiapour M H, Han X, Lazebnik S, et al. Where to buy it: Matching street clothing photos in online shops [C] //Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition (CVPR 2015). Piscataway, NJ: IEEE, 2015: 3343-3351
- [42] You Quanzeng, Yuan Jianbo, Wang Jiaqi, et al. Snap n'shop: Visual search-based mobile shopping made a breeze by machine and crowd intelligence [C] //Proc of the 9th 2015 IEEE Int Conf on Semantic Computing. Piscataway, NJ: IEEE, 2015: 173-180
- [43] Ngo T D, Phan S, Le D D, et al. Recommend-me: Recommending query regions for image search [C] //Proc of the 29th Annual ACM Symp on Applied Computing. New York: ACM, 2014: 913-918
- [44] Yu F X, Ji R, Chang S F. Active query sensing for mobile location search [C] //Proc of the 19th ACM Int Conf on Multimedia. New York: ACM, 2011: 3-12
- [45] Zhao Zou, Song Ruihua, Xie Xing, et al. Mobile query recommendation via tensor function learning [C] //Proc of the 24th Int Conf on Artificial Intelligence. New York: ACM, 2015: 4084-4090

- [46] Wang Changhu, Zhang Lei. Charm and challenge sketches search [J]. China Computer Society Newsletter, 2012, 8 (12): 20-26 (in Chinese)  
(王长虎, 张磊. 草图搜索的魅力与挑战[J]. 中国计算机学会通讯, 2012, 8(12): 20-26)
- [47] Xin Yuxuan, Yan Zifei. Research progress of image retrieval based on hand-drawn sketches [J]. CAAI Trans on Intelligent Systems, 2015 (2): 167-177 (in Chinese)  
(辛雨璇, 闫子飞. 基于手绘草图的图像检索技术研究进展[J]. 智能系统学报, 2015 (2): 167-177)
- [48] Cao Yang, Wang Hai, Wang Changhu, et al. MindFinder: Interactive sketch-based image search on millions of images [C] //Proc of the 18th ACM Int Conf on Multimedia. New York: ACM, 2010: 1605-1608
- [49] Eitz M, Hays J, Alexa M. How do humans sketch objects? [J]. ACM Trans on Graphics, 2012, 31(4): 44:1-44:10
- [50] Xiao Changcheng, Wang Changhu, Zhang Liqing, et al. IdeaPanel: A large scale interactive sketch-based image search system [C] //Proc of the 23rd ACM Int Conf on Multimedia Retrieval. New York: ACM, 2015: 667-668
- [51] Sun Xinghai, Wang Changhu, Xu Chao, et al. Indexing billions of images for sketch-based retrieval [C] //Proc of the 21st ACM Int Conf on Multimedia. New York: ACM, 2013: 233-242
- [52] Schneider R G, Tuytelaars T. Sketch classification and classification-driven analysis using fisher vectors [J]. ACM Trans on Graphics, 2014, 33(6): 174
- [53] Sun Zhenbang, Wang Changhu, Zhang Liqing, et al. Query-adaptive shape topic mining for hand-drawn sketch recognition [C] //Proc of the 20th ACM Int Conf on Multimedia. New York: ACM, 2012: 519-528
- [54] Yanlk E, Sezgin T M. Active learning for sketch recognition [J]. Computers & Graphics, 2015, 52: 93-105
- [55] Yu Qian, Yang Yongxin, Liu Feng, et al. Sketch-a-Net: A deep neural network that beats humans [J]. International Journal of Computer Vision, 2016, 7(5): 337-341
- [56] Sangkloy P, Burnell N, Ham C, et al. The sketchy database: Learning to retrieve badly drawn bunnies [J]. ACM Trans on Graphics, 2016, 35(4): 1-12
- [57] Zhuang Yueting, Wu Fei, He Xiaofei. Cross-media retrieval and ranking [J]. Communications of the CCF, 2014, 10(7): 14-19 (in Chinese)  
(庄越挺, 吴飞, 何晓飞. 跨媒体检索与排序[J]. 中国计算机学会通讯, 2014, 10(7): 14-19)
- [58] Zhao Yao, Wei Shikui, Wang Shuhui, et al. Knowledge expression of cross-media-perception, association and consistency [J]. Communications of the CCF, 2014, 10(7): 8-13 (in Chinese)  
(赵耀, 韦世奎, 王树徽, 等. 跨媒体时代的知识表达—感知、关联及一致性表示[J]. 中国计算机学会通讯, 2014, 10(7): 8-13)
- [59] Hardoon D R, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: An overview with application to learning methods [J]. Neural Computation, 2004, 16 (12): 2639-2664
- [60] Tenenbaum J B, Freeman W T. Separating style and content with bilinear models [J]. Neural Computation, 2000, 12(6): 1247
- [61] Li Dongge, Dimitrova N, Li Mingkun, et al. Multimedia content processing through cross-modal association [C] //Proc of the 11th ACM Int Conf on Multimedia. New York: ACM, 2003: 604-611
- [62] Mahadevan V, Wong C W, Pereira J C, et al. Maximum covariance unfolding: Manifold learning for bimodal data [C] //Proc of the 25th Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2011: 918-926
- [63] Sharma A, Kumar A, Daume H, et al. Generalized multiview analysis: A discriminative latent space [C] //Proc of the 25th IEEE Conf on Computer Vision and Pattern Recognition (CVPR 2012). Piscataway, NJ: IEEE, 2012: 2160-2167
- [64] Zhai Deming, Chang Hong, Shan Shiguang, et al. Multiview metric learning with global consistency and local smoothness [J]. ACM Trans on Intelligent Systems and Technology, 2012, 3(3): 53
- [65] Lu Xinyan, Wu Fei, Tang Siliang, et al. A low rank structural large margin method for cross-modal ranking [C] //Proc of the 36th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2013: 433-442
- [66] Wu Fei, Lu Xinyan, Zhang Zhongfei, et al. Cross-media semantic representation via bi-directional learning to rank [C] //Proc of the 21st ACM Int Conf on Multimedia. New York: ACM, 2013: 877-886
- [67] Zheng Y, Zhang Y J, Larochelle H. Topic modeling of multimodal data: An autoregressive approach [C] //Proc of the 27th IEEE Conf on Computer Vision and Pattern Recognition (CVPR 2014). Piscataway, NJ: IEEE, 2014: 1370-1377
- [68] Liao Renjie, Zhu Jun, Qin Zenchang. Nonparametric Bayesian upstream supervised multi-modal topic models [C] //Proc of the 7th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2014: 493-502
- [69] Wang Yanfei, Wu Fei, Song Jun, et al. Multi-modal mutual topic reinforce modeling for cross-media retrieval [C] //Proc of the 22nd ACM Int Conf on Multimedia. New York: ACM, 2014: 307-316
- [70] Wei Yunchao, Zhao Yao, Lu Canyi, et al. Cross-modal retrieval with CNN visual features: A new baseline [J]. IEEE Trans on Cybernetics, 2017, 47(2): 449-460
- [71] Ma Lin, Lu Zhengdong, Shang Lifeng, et al. Multimodal convolutional neural networks for matching image and sentence [J]. Computer Science, 2015: 2623-2631

- [72] Wu Fei, Yu Zhou, Yang Yi, et al. Sparse multi-modal hashing [J]. *IEEE Trans on Multimedia*, 2014, 16(2): 427-439
- [73] Ding Kun, Fan Bin, Huo Chunlei, et al. Cross-modal hashing via rank-order preserving [J]. *IEEE Trans on Multimedia*, 2017, 19(3): 571-585
- [74] Cao Yue, Long Mingsheng, Wang Jianmin, et al. Deep visual-semantic hashing for cross-modal retrieval [C] //Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1445-1454
- [75] TRECVID. TREC video retrieval evaluation [OL]. [2017-03-16]. <http://www-nlpir.nist.gov/projects/trecvid/>
- [76] Liu Danzhou, Hua K A, Vu K, et al. Fast query point movement techniques for large CBIR systems [J]. *IEEE Trans on Knowledge & Data Engineering*, 2008, 21(5): 729-743
- [77] Aksoy S, Haralick R M, Cheikh F A, et al. A weighted distance approach to relevance feedback [J]. *Electronic Test*, 2011, 4(4): 812-815
- [78] Huang T S, Dagli C K, Rajaram S, et al. Active learning for interactive multimedia retrieval [J]. *Proceedings of the IEEE*, 2008, 96(4): 648-667
- [79] Zhang Lei, Lin Fuzong, Zhang Bo. Support vector machine learning for image retrieval [C] //Proc of the 3rd Int Conf on Image Processing. Piscataway, NJ: IEEE, 2001: 721-724
- [80] Chen Yunqiang, Zhou X S, Huang T S. One-class SVM for learning in image retrieval [C] //Proc of 2001 IEEE Int Conf on Image Processing. Piscataway, NJ: IEEE, 2001: 34-37
- [81] Tong S, Chang E. Support vector machine active learning for image retrieval [C] //Proc of the 9th ACM Int Conf on Multimedia. New York: ACM, 2001: 107-118
- [82] Zha Zhengjun, Wang Meng, Zheng Yantao, et al. Interactive video indexing with statistical active learning [J]. *IEEE Trans on Multimedia*, 2012, 14(1): 17-27
- [83] Farhadi A, Endres I, Hoiem D, et al. Describing objects by their attributes [C] //Proc of the 22nd IEEE Conf on Computer Vision and Pattern Recognition (CVPR 2009). Piscataway, NJ: IEEE, 2009: 1778-1785
- [84] Vittayakorn S, Umeda T, Murasaki K, et al. Automatic Attribute Discovery with Neural Activations [M]. Berlin: Springer, 2016
- [85] Zheng Jingjing, Jiang Zhuolin, Chellappa R. Submodular attribute selection for visual recognition [J]. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 2016, DOI: 10.1109/TPAMI.2016.2636827
- [86] Li Yining, Huang Chen, Loy C C, et al. Human attribute recognition by deep hierarchical contexts [C] //Proc of the European Conf on Computer Vision. Berlin: Springer, 2016: 684-700
- [87] Douze M, Ramisa A, Schmid C. Combining attributes and fisher vectors for efficient image retrieval [C] //Proc of the 24th IEEE Conf on Computer Vision and Pattern Recognition (CVPR 2011). Piscataway, NJ: IEEE, 2011: 745-752
- [88] Scheirer W J, Kumar N, Belhumeur P N, et al. Multi-attribute spaces: Calibration for attribute fusion and similarity search [C] //Proc of the 25th IEEE Conf on Computer Vision and Pattern Recognition (CVPR 2012). Piscataway, NJ: IEEE, 2012: 2933-2940
- [89] Liu Ziwei, Luo Ping, Qiu Shi, et al. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations [C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition (CVPR 2016). Piscataway, NJ: IEEE, 2016: 1096-1104
- [90] Zhang Hanwang, Zha Zhengjun, Yan Shuicheng, et al. Attribute feedback [C] //Proc of the 20th ACM Int Conf on Multimedia. New York: ACM, 2012: 79-88
- [91] Zhang Hanwang, Zha Zhengjun, Yang Yang, et al. Attribute-augmented semantic hierarchy: Towards a unified framework for content-based image retrieval [J]. *ACM Trans on Multimedia Computing, Communications, and Applications*, 2014, 11(1s): 21
- [92] Kovashka A, Parikh D, Grauman K. Whittlesearch: Interactive image search with relative attribute feedback [J]. *Int Journal of Computer Vision*, 2015, 115(2): 185-210
- [93] Yu A, Grauman K. Just noticeable differences in visual attributes [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2015: 2416-2424
- [94] Agichtein E, Brill E, Dumais S. Improving Web search ranking by incorporating user behavior information [C] //Proc of the 29th ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2006: 19-26
- [95] Joachims T. Optimizing search engines using clickthrough data [C] //Proc of the 8th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2002: 133-142
- [96] Jiang J, Hassan A A, Shi X, et al. Understanding and predicting graded search satisfaction [C] //Proc of the 8th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2015: 57-66
- [97] Hua Xiansheng, Yang Linjun, Wang Jingdong, et al. Clickage: Towards bridging semantic and intent gaps via mining click logs of search engines [C] //Proc of the 21st ACM Int Conf on Multimedia. New York: ACM, 2013: 243-252
- [98] Jain V, Varma M. Learning to re-rank: Query-dependent image re-ranking using click data [C] //Proc of the 20th Int Conf on World Wide Web. New York: ACM, 2011: 277-286
- [99] Yu Jun, Rui Yong, Chen Bo. Exploiting click constraints and multi-view features for image re-ranking [J]. *IEEE Trans on Multimedia*, 2014, 16(1): 159-168

[100] O'Hare N, De Juan P, Schifanella R, et al. Leveraging user interaction signals for Web image search [C] //Proc of the 39th ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2016: 559-568

[101] Wu Fei, Lu Xinyan, Song Jun, et al. Learning of multimodal representations with random walks on the click graph [J]. IEEE Trans on Image Processing, 2016, 25(2): 630-642

[102] Jiang Lu, Cao Liangliang, Kalantidis Y, et al. Delving deep into personal photo and video search [C] //Proc of the 10th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2017: 801-810

[103] Wang J, Pohlmeier E, Hanna B, et al. Brain state decoding for rapid image retrieval [C] //Proc of the 17th ACM Int Conf on Multimedia. New York: ACM, 2009: 945-954

[104] Papadopoulos G T, Apostolakis K C, Dara P. Gaze-based relevance feedback for realizing region-based image retrieval [J]. IEEE Trans on Multimedia, 2013, 16(2): 440-454

[105] Kauppi J P, Kandemir M, Saarinen V M, et al. Towards brain-activity-controlled information retrieval: Decoding image relevance from MEG signals [J]. NeuroImage, 2015, 112(6): 288-298

[106] Marchionini G. Exploratory search: From finding to understanding [J]. Communications of the ACM, 2006, 49(4): 41-46

[107] Kai U B, Hezel N, Mackowiak R. ImageMap-Visually Browsing Millions of Images [M]. Berlin: Springer, 2015: 287-290

[108] Halvey M, Vallet D, Hannah D, et al. Supporting exploratory video retrieval tasks with grouping and recommendation [J]. Information Processing & Management, 2014, 50(6): 876-898

[109] Tsukuda K, Goto M. Exploratory video search: A music video search system based on coordinate terms and diversification [C] //Proc of the 2015 IEEE Int Symp on Multimedia (ISM). Piscataway, NJ: IEEE, 2015: 221-224



**Zha Zhengjun**, born in 1984. PhD. Professor, PhD supervisor in University of Science and Technology of China. His main research interests include multimedia analysis and retrieval, computer vision, and patten recognition.



**Zheng Xiaojun**, born in 1987. PhD candidate. Her main research interests include video analysis and retrieval.