

doi:10.3772/j.issn.1000-0135.2016.002.001

基于任务的图像检索相关性评价研究¹⁾

黄 崑 王珊珊 王凯飞

(北京师范大学政府管理学院, 北京 100875)

摘要 在回顾国内外图像检索相关性研究的基础上,探讨国内用户评价图像检索结果相关性的特点。选取十项相关性标准,设计一般、专指和抽象三类任务,招募30位被试,通过实验研究考察用户在完成三类任务时评价结果相关性的共性和差异。结果发现:主题、准确性、文本上下文、完全性以及吸引力是不同类型任务下重要性认同最高的五项相关性标准;任务类型会影响相关性标准的使用情况;其中,任务越专指,用户越倾向于根据主题、准确性进行结果评价;任务越抽象,用户则越倾向于根据图片的启发性、所激发的情感进行结果评价;一般性和专指性任务都更为重视结果图片包含检索需求涉及对象的完全性。并且,同一种任务类型中,用户对相关性标准重要性的评价还会随着图像检索目的、检索需求描述方式的不同而存在差异。

关键词 相关性 图像检索 任务 实验研究

Research on Relevance Evaluation in Task-based Image Retrieval

Huang Kun, Wang Shanshan and Wang Kaifei

(School of Government, Beijing Normal University, Beijing 100875)

Abstract Based on the review of the researches on relevance evaluation in image retrieval at home and abroad, it explored how domestic users evaluate relevance of searching results in image retrieval. It selected ten relevance criteria, designed generic, specific and abstract tasks, and recruited 30 participants for the experiment. By conducting an experimental study, it examined the commonness and differences of the use of the ten relevance criteria among the three different tasks. The findings indicated that topicality, accuracy, textual information, completeness and appeal were rated most important for all the three kinds of tasks. Besides, task would influence the use of relevance criteria. The more specific the task was, the more frequently the criteria of topicality, accuracy would be used, and the more abstract the task was, the more frequently the criteria of suggestiveness and emotion would be used. In addition, users would evaluate the relevance by completeness more frequently in both generic and specific tasks. Furthermore, within the same type of task, different searching purpose and different descriptions of searching requirements may lead to differences among the use of relevance criteria.

Keywords relevance, image retrieval, task, experimental research

收稿日期:2015年9月6日

作者简介:黄崑(通讯作者),女,1978年生,教授,主要研究方向:信息用户和信息检索,E-mail:huangkun@bnu.edu.cn。
王珊珊,女,1992年生,硕士研究生,主要研究方向:用户信息行为,E-mail:perfect_1992@126.com。王凯飞,女,1990年生,硕士研究生,主要研究方向:用户信息行为,E-mail:942040581@qq.com。

1) 本文系教育部人文社科青年基金项目“基于大众参与的图像情感特征标引机制与方法研究”(11YJC870010)成果之一。

1 引言

在信息检索系统中,信息资源集合与用户信息需求集合的匹配和选择通常建立在信息资源的特征化表示和用户信息需求的符号化表示基础上,检索相关性问题的产生。两类抽象表示的相关性判定研究主要是探讨如何在算法、系统相关性基础上,综合用户面临的现实问题情境需要,以及个人的知识背景和检索经验等,实现主题、认知、情境,乃至情感层面的相关^[1],以满足具有个体认知差异的不同用户群体,进一步提高检索结果对用户的有用性和适用性。自20世纪50年代开始,检索相关性研究逐步从以算法、系统为中心向以用户为中心转变,90年代出现过研究高潮。进入21世纪后,检索相关性依然是热点的研究问题之一。从检索相关性整个领域来看,主要以文本信息为主。不过,近年来,关注特定资源类型,如图像、音视频等非文本信息的相关性研究也逐步受到关注^[2]。图像语义的丰富性为用户筛选和评价检索结果提供了更多选择,而图像认知过程中用户理解的多样性使得不同用户对结果评价和筛选的一致性要低于文本信息。因此,虽然文本检索的相关性标准对于非文本信息资源具有一定的适用性,但是,图像用户在检索目的、检索需求表达方式、对图片的主观理解等方面都有其特殊性。所以,依然有必要针对图像信息检索中的相关问题进行专门研究。

图像检索相关性标准的研究关注人们找寻图片时筛选图片的标准,以及筛选过程的影响因素。因此,用户视角开展的图像检索相关性研究同时也是图像用户研究的内容之一。从图像用户研究来看,国内外研究者不仅从需求产生阶段考察过图像检索的目的和动机,从查寻阶段考察过用户的信息源使用情况、图像需求的表达方式、关键的检索行为及采取的提问调整策略,此外,也从结果筛选阶段考察过用户评价检索结果相关性的依据问题等^[3]。总体来看,以关注前两个阶段的研究为主,探讨结果筛选阶段的研究还较为有限。然而,在图像检索技术实现的多个重要环节,如检索模型的构建、相关反馈机制的设计,以及个性化检索结果的生成等,都与用户判断检索结果相关性的标准关系密切。因此,图像检索相关性评价研究对于促进图像检索技术的完善具有一定的参考和指导价值。本研究选取任务这一重要的情境因素^[4],针对图像检索问题,设计了一

般性、专指性和抽象性三类任务。通过实验研究,考察用户完成三类任务时,评价检索结果相关性的共性和差异,揭示国内用户评价图像检索结果相关性的特点及其与任务类型的关系。

2 国内外相关研究

相关性问题是信息检索领域活跃的研究问题之一,国内外研究者在不同时期都曾对检索相关性研究开展过较为系统的综述回顾工作。1997年,意大利乌迪内大学的 Mizzaro^[5]从相关性的含义、标准、特性及影响因素等方面分析和评价了156篇相关研究工作。美国罗格斯大学的 Saracevic 曾就相关问题陆续发表过多篇综述文章。2007年, Saracevic^[1]系统梳理了过去三十多年中开展的相关研究,从相关性的本质、理论、模型、行为和影响因素等5个方面进行了讨论,并进一步指出相关性具有动态性和情境依赖的特点,用户的认知和情感都会影响相关性评价标准的使用。在国内,西南大学的付玲玲^[6]、南京大学的成颖^[7]、河北大学的王雅坤和国家图书馆的成全^[8]、南京大学的李亚琴等^[9]、中国农业科学院的王健等^[10]都曾分别从相关性标准研究的方法论、研究阶段和研究主题等不同角度进行梳理。这些综述性工作较好地总结了不同时期的研究特点,勾勒出研究发展的脉络。可以看到,在经历了从技术视角向用户视角的转变后,检索相关性研究关注的信息资源类型从以学术信息为主,发展为包括网页、图像、音视频、口述资料、博客等更丰富的资源类型;在研究的用户群体上除了学生和教师,还对医护人员、新闻从业人员等特定专业、行业领域的用户进行研究。

根据文献调研情况来看,专门探讨图像检索相关性的研究主要分为两类:一类是探讨用户评价图像检索结果相关性的依据问题,通常以特定专业或者职业背景的用户为对象,考察他们评价结果相关性的标准。如2000年, Markkula 和 Sormunen^[11]招募了8名新闻记者,观察他们从照片检索需求产生到查找、选择的完整过程,并结合访谈来了解他们在日常工作中挑选照片的标准。结果发现,主题是他们最看重的评价标准;此外,照片拍摄有关的技术特征、书目信息、照片的感染性、表现力,以及个人的审美标准也常用于筛选照片;并且,新闻记者在实际选择照片时还会受到多种因素的影响,如需要考虑照片与报道文字、呈现网页的相称性,以及有关新闻报

道的政策和伦理道德约束等。2006年, Hung^[12]招募了30位具有新闻传媒行业图片编辑经验的被试,通过图像检索实验,发现用户使用到了37个相关性标准,并且图片的象征意义、视觉构成、主题、上下文等10个标准使用最多。2008年, Sedghi等^[13]考察了医护人员进行医学影像图片检索的相关性评价标准问题,该研究招募了26位医护人员,让他们对日常所需的医学影像图片进行检索和筛选,结合访谈和出声思维的方法收集被试在检索过程中运用到的评价标准。结果发现,医护人员在言语表达中提及了26个标准,可以归纳为视觉、医学、文本和其他四类标准;并且,视觉相关、图像质量和图片标题、注释等背景信息是医护人员评价图片相关性最常用的三个标准。2012年, Sedghi等^[14]在之前研究上继续推进,新发现了原创性、放大倍数、技术信息、方向信息等6个医护人员还会使用到的标准,它们在已有的文本和其他多媒体检索相关性标准中未被提及。

第二类研究探讨影响因素问题,分析图像检索相关性判断与用户、情境、检索过程等因素之间的关系。有的研究者从检索阶段与检索结果相关性评价的关系角度进行探讨,1999年, Hirsh^[15]曾研究小学生在检索前、后期评价图像检索结果相关性的特点,结果指出小学生用户在不同的检索阶段会采用不同的相关性标准,并且随着检索过程的深入,用户在后期使用到的相关性标准数量高于前期。该研究还从比较的角度考察了文本和图像相关性标准的差异,指出小学生们在评价文本信息时更注重主题相关,而评价图片时更关心是否符合自己的兴趣。2002年, Choi和 Rasmussen^[16]探讨了检索前与检索后用户评价相关性标准的变化,该研究招募了38位被试,选取了主题、准确性、启发性等9个相关性标准,被试在看到检索结果前、后分别对9个相关性标准的重要程度进行评分。结果发现,检索前、后用户对相关性标准的重要性评价存在显著差异:在检索前,被试普遍认为主题相关、准确性和完全性最重要;而在检索后,主题相关、时间跨度和可获取性被认为最为重要。

有研究者从任务因素与图像检索相关性评价的关系角度进行探讨。2005年, Hung等^[17]设计了一般、专指和主观三个任务,并招募了10位大学生参与实验,研究者在被试完成每一个检索任务后进行访谈,了解被试在检索过程中评价结果相关性的依据。结果总结了主题、情感、审美、文本等12个相关性标准,并指出主题、情感和审美是三类任务中共同

常用的标准;然而,用户在完成一般任务和主观任务时会更依赖个人情感和文本信息,完成专指任务时则更依赖图片中对象的外在特征(如图中对象的面部表情、姿势和外貌)。研究人员进一步将文本信息划分为标题、对象名称、地点和创建时间,结果发现地点在用户相关性判断中更加有用。同时,该研究还发现,仅有女性被试使用了审美这一标准,可能性别因素也会产生一定影响。类似地,2010年, Hamid和 Thom^[18]也设计了一般、专指和抽象三个检索任务,招募了12位被试参与实验,对预设的十项相关性标准的重要性进行打分,结果发现,主题、准确性、视觉冲击最为重要,而引申含义和文本信息用户在进行相关性判断时考虑最少。并且,主题在一般及专指性任务中更为重要,而准确性在专指性任务中更为重要,对于抽象性任务,图像视觉特征的重要性得分要远远低于在一般性和专指性检索任务中的得分,可能与抽象性任务更注重象征、引申意义,而非实际的图片中包含的视觉特征和对象。

综上所述,国内外研究者已经在图像检索相关性评价问题上开展了一系列研究,并且初步总结了用户评价图像检索结果相关性的各项依据,也揭示了相关性标准与任务、阶段、资源类型的关系,为进一步深入挖掘用户评价图像检索相关性的行为方式和特点奠定了基础。不过,这些研究大都以西方被试为主,国内在这一方面的研究显得更为有限,尽管国内在检索相关性问题上的讨论较为丰富,但是关注图像的研究数量并不多。因此,本研究从任务因素角度出发,考察国内用户在开展不同类型的检索任务时评价检索结果相关性的差异和共性,以丰富国内在检索相关性方面的研究,并且为图像检索系统的优化提供参考。

3 研究设计

3.1 研究问题

本研究主要探讨如下三个研究问题:

第一,用户在评价和筛选图像检索结果时倾向于使用哪些相关性标准?

第二,用户在完成不同类型的图像检索任务时,倾向使用的相关性标准有何共性和差异?

第三,用户检索图像的目的、图像检索任务的描述方式不同,用户倾向使用的相关性标准是否也不同?

3.2 相关性标准与任务设计

3.2.1 相关性标准的选取

在相关性标准的选取上,本研究沿用了 Hamid 和 Thom^[18]总结的十项标准,该研究整合了 Choi 和 Rasmussen^[16]、Hung^[12]研究中揭示的高频使用的相关性标准,去重后筛选出主题、准确性、启发性、吸引力、完全性、视觉特征、情感、文本上下文、引申含义和视觉冲击十项,如表1所示。

表1 图像相关性评价标准

相关性评价标准	说明
主题	我会选择与检索主题直接相关的图片。
准确性	我会选择准确表现了我的需要的图片。
启发性	我会选择具有启发性的图片。
吸引力	我会选择吸引人的图片。
完全性	我会选择图片中包括我所需要的重要细节的图片。
视觉特征	我会选择视觉性特征上符合需要的图片,如颜色、形状、纹理、分辨率等。
情感	我会选择能够激发情绪反应的图片。
文本上下文	我会选择图片的文本描述符合需要的图片。
引申含义	我会选择暗含、间接地与主题相关的图片。
视觉冲击	我会选择有视觉冲击力的图片。

本研究利用五级计分的李克特自陈量表,请用户在完成每一个检索任务后对表1中列出的十项陈述进行符合程度的评价。其中,“5”表示非常符合,意味着用户认可所评价的标准对于筛选结果的重要性;相反,“1”表示非常不符合,意味着用户不认可在评价结果相关性时所列的标准很重要。

3.2.2 任务类型的设计


在设计图像任务类型时采用了常见的三分法^[17,18],即根据 Shatford 提出的图像需求分类理论^[19],将图像检索任务分成一般性、专指性和抽象性三种。其中,一般性任务是指检索需求描述中使用事物名称的情况,如“山川”、“河流”;而专指性任务是指检索需求描述中更准确地使用事物确切的名称,如“喜马拉雅山”、“黄河”;抽象性任务则是指检索需求描述中包含感觉、情绪、情感等与人类主观体验有关的要求,如“愉悦”、“浪漫”,或者是包含“和平”、“法律”等抽象概念。

为了增强检索任务的真实性,本研究参考百度知道上与图像需求有关的问题及陈述方式^[20],模拟了七个任务,如表2所列。

如表2所列,在对任务进行三种类型划分的基础上,本研究还结合检索需求描述方式、图像用途理论,对各个类型下的具体任务进行了设计。

首先,在检索需求描述方式上,一方面在一般性和专指性任务中提供了基于文字找图和基于图片找图两种任务给出方式;其中,任务2给出了一张图片,在检索前问卷中先询问被试是否知道图中狗的

表2 图像检索任务列表

任务类型	任务编号	任务描述
一般性	1	我正在为一段广告设计配图,希望有蓝天,有白云,一片广袤的田野,田野上有人或者牛羊,图片大小最好在1M以上,高清的。
一般性/专指性	2	图  中的狗是什么品种?我需要它的图片。
专指性	3	我从淘宝上买了一双 New Balance 574 系列 ML574CPB 型号的鞋,怀疑是仿制品,想要正品图片验证一下。
	4	我想要哥伦比亚球员 J 罗在 2014 年世界杯比赛中的图片作为电脑桌面。
抽象性	5	为了参加一家外企用人面试,我想要换个发型,希望自己显得干练、有活力,请帮我推荐发型图片。
	6	我正在制作大学毕业纪念册,想要给合影照片找合适的背景图片,希望看起来有想当年的感觉。
	7	我需要素材图片,想要表现人间真爱,希望找点灵感。

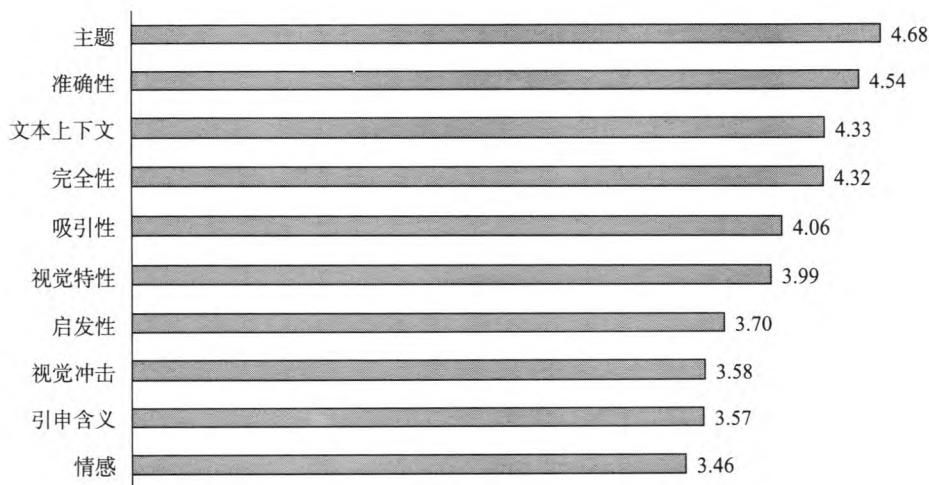


图1 10项相关性评价标准评价均值

品种,然后再进行图片查找。对于了解的被试,该任务为专指性任务,而对于不清楚的被试而言,则为一般性任务(即被试只能识别出图片中有一只狗)。因此,该任务用于分析和文字找图情况下的差异。另一方面,在抽象性任务描述中使用了图像情感语义的不同方面,包括对客观事物的风格描述(干练、有活力)、对客观事物引发的人类基本情绪描述(想当年),以及抽象概念(人间真爱)三方面,用于分析抽象性任务中任务描述涉及的图像特征与用户倾向的相关性标准的差异。

其次,在专指性任务的设计中,参考 Fidel^[21]提出的两极理论设计了两个任务,该理论将图像用途看作具有两极的连续线条,线条一端是将图像作为信息和数据来源使用,称为数据极,如用户查找到美国统计局发布的人口趋势变化图了解美国人口变化情况。线条另一端是将图像作为客体或物体的描述使用,称为对象极,如用户希望更换计算机桌面背景,或者是为演示文稿找插图。根据这一理论设计了两个专指性任务,一个是为了检验商品真伪而找图,是较为典型的将图片作为数据源使用,另一个是用于设置计算机桌面,是较为典型的将图像作为对象来源使用。这两个任务用于探讨专指性任务中图像用途的不同是否会引起用户倾向的相关性标准的差异。

3.3 数据收集

本研究通过网络随机招募了30位被试,均为高校在读学生,其中,男性9人,女性21人,平均每日上网时间在6个小时以上的有25人(83.3%),表示经常和有时查找图片的有18人(60%)。实验过

程中,被试先填写个人基本信息;然后,根据要求顺序完成指定的七项任务,每项任务完成后随即对所列出的十项相关性标准的陈述进行符合程度的评价;七项任务全部完成后,对被试进行了访谈,询问被试对相关评价标准的看法。实验过程采取拉丁方实验设计避免检索任务顺序对研究结果的影响。

经过数据收集,30位被试均完成了指定的七项任务,共计收集到210次关于十项相关性标准的评价数据。

4 数据分析与讨论

4.1 总体情况

根据所列出的十项相关性标准,30位被试进行了逐一评价和反馈,各项标准的用户评价均值如图1所示。

在十项标准中认可度最高的前五项为主题(4.68)、准确性(4.54)、图片相关的文本上下文(4.33)、图片包含所需细节的完全程度(4.32),以及图片的吸引力(4.06),评价均值都高于4分。相比而言,用户对情感标准的评分均值(3.46)最低,不过,亦大于表示中立态度的3分,说明被试在评价图片相关性时,对情感标准的重要性还是有一定的认可度。可见,用户在筛选图像检索结果时,对问卷中列出的十项标准都有所运用,由此也揭示出图像检索过程是一个较为复杂的筛选过程,用户会综合运用多项标准进行决策。

在被试完成七个任务后进行访谈,询问被试除了十项列出的标准,是否还会使用其他标准评价结

果相关性。结果发现,有15人(50%)表示,当面对很多检索结果时,会进一步从尺寸、大小、清晰度筛选图片,并且希望图片中不要有水印,希望找到的图片可以直接进行编辑或者使用。有13人(43.33%)提到个人审美标准,他们认为图片应该看着舒服、漂亮,符合个人的审美喜好比较重要,有1人具体指出喜欢给人正能量、积极向上的图片。另有4人(13.33%)提到希望图片是吸引人的、特别的,2人(6.67%)提到新颖性很重要。不过七成以上的被试都指出,究竟使用什么标准,还和具体任务需求和图片用途有关,他们在实验中会从尽可能符合题意的角度去判断,比如有的任务对图片大小有要求,他们就会更注重运用这一标准筛选图片。可见,任务与相关性评价标准的侧重有一定的关系。

4.2 相关性标准重要性评价在三类任务类型间的差异分析

表3列出了通过方差分析对三类任务类型间相关性标准重要性差异的分析结果。根据评分均值将4分以上界定为被试对该标准重要性表示认可,低于4分则界定为不认可。因此,从三类任务涉及的重要标准数量来看,一般性任务和抽象性任务涉及的标准数量(6个)仅略多于专指性任务(5个)。综合来看,各类任务完成过程中,十项标准中有半数左右对用户进行结果评价具有重要作用。

表3 任务类型与10项相关性标准重要性评价的方差分析

相关性指标	F值	P	一般性 N=40	专指性 N=80	抽象性 N=90
			均值	均值	均值
主题	16.928	0.031*	4.68	4.81	4.53
准确性	13.990	0.030*	4.40	4.75	4.47
启发性	16.125	0.041*	3.48	3.55	4.00
吸引力	7.871	0.446	4.18	3.88	4.08
完全性	15.777	0.046*	4.40	4.38	4.20
视觉特征	13.188	0.106	4.08	4.13	3.72
情感	21.853	0.005*	3.28	3.30	3.89
文本上下文	12.713	0.122	4.35	4.50	4.20
引申含义	7.858	0.447	3.55	3.45	3.77
视觉冲击	9.964	0.268	3.60	3.53	3.60

*:0.05 水平显著

从三类任务的共性来看,主题、准确性、完全性和文本上下文是三类任务下评分均值都高于4的四项标准,对于评价结果相关性较为重要。从三类任务的差异来看,三类任务在主题、准确性、启发性、完全性、情感五个标准上,差异显著。对于专指性任务而言,主题和准确性两方面的评价均值都显著高于另外两类任务,这与专指性任务通常比较明确有关。而对于抽象性任务,启发性和情感性两方面的评分均值都显著高于一般性任务和专指性任务,这与抽象性问题的结果评价更为主观有关。对于一般性任务,在完全性标准上的评分均值上与专指性任务相近,但是要显著高于抽象性任务,这与前两类任务更易于明确图片中的组成对象有关。

4.3 相关性标准重要性评价在同类任务组内的差异分析

为了揭示用户在完成同类任务时对相关性标准重要性评价的差异,分别对一般性任务、专指性任务以及抽象性任务进行差异显著性分析,如表4、表5、表6所列。

表4 一般性任务在相关性标准重要性评价上的内部差异分析

相关性指标	t值	P	一般性任务1 N=10	一般性任务2 N=10
			均值	均值
主题	0.000	1.000	4.60	4.60
准确性	1.177	0.269	4.20	4.60
启发性	-1.500	0.168	3.50	3.10
吸引力	-0.709	0.496	4.10	3.80
完全性	0.361	0.726	4.40	4.50
视觉特征	1.000	0.343	3.70	4.10
情感	-0.688	0.509	3.60	3.40
文本上下文	0.688	0.509	4.40	4.60
引申含义	1.000	0.343	3.70	3.90
视觉冲击	-1.500	0.168	3.90	3.30

在一般性任务中包括两项任务,一项为广告设计寻找配图,称为“一般性任务1”,另一项为以示例图找相关狗的图,称为“一般性任务2”。根据被试在第二个问题中的反馈,20人能够说出狗的品种,10人表示不清楚。对于不清楚的10位被试而言,

以图找图的任务为一般性任务,考虑到配对样本 T 检验的要求,因此,仅以这 10 位被试完成的两项一般性任务的评价数据为基础进行分析。如表 4 所示,未发现两个一般性任务在相关性标准的重要性评价上存在显著差异,初步显示出图像需求以文字方式给出还是图片方式给出,并未影响相关性评价标准的使用。

在专指性任务中包含 3 个具体任务,查验球鞋正品的任务为“专指任务 1”,查找球员赛事图片的任务为“专指任务 2”,以图找图中知道图中狗的品种的情况归入“专指任务 3”。同样考虑到配对分析的需要,因此,选取“专指任务 3”中的 20 位被试,根据他们对三个专指任务的评价数据进行分析,通过 K 个相关样本非参数检验分析得到表 5。

表 5 专指性任务在相关性标准重要性评价上的内部差异分析

相关性指标	卡方值	P	专指性 任务 1 $N = 20$	专指性 任务 2 $N = 20$	专指性 任务 3 $N = 20$
			均值	均值	均值
主题	0.875	0.646	4.80	4.80	4.85
准确性	2.24	0.326	4.55	4.90	4.85
启发性	0.039	0.981	3.45	3.45	3.55
吸引力	21.143	0.000 *	3.00	4.25	4.45
完全性	0.809	0.667	4.25	4.30	4.45
视觉特征	1.302	0.521	4.10	4.25	4.15
情感	11.804	0.003 *	2.80	3.65	3.45
文本上下文	3.059	0.217	4.75	4.40	4.45
引申含义	2.048	0.359	3.15	3.80	3.25
视觉冲击	10.719	0.005 *	2.95	4.10	3.50

*, 0.05 水平显著

表 5 反映出,3 个专指性任务在吸引力、情感和视觉冲击三项相关性标准的重要性评价上存在显著差异。其中,以图找图的“专指任务 3”和寻找球员图片用作桌面背景的“专指任务 2”两个任务中,吸引力标准的评分均值均超过 4 分,而进行正品验证的“专指任务 1”在该标准上的得分均仅为 3.00。并且,“专指任务 1”在情感、视觉冲击两项标准上的评价均分均不到 3,重要性认可较低。这与该任务属于将图片作为数据来源使用的目的有关,因此

用户会更注重从图中获得的数据的真实性,以辅助用户做出判断。相比而言,查找图片用作桌面背景的“专指任务 2”属于将图片作为对象来源使用的一种用途,并且具有潜在的审美、欣赏目的,因此会在图片是否吸引人、是否具有一定的视觉冲击力方面有更高的要求。这一差异也反映出,尽管三个任务都给出了确指的对象名称,同属于专指性任务,但是因为任务描述的检索目的不同,用户进行结果选择时所依据的相关性标准也会有所差异。同时,经过多重比较分析发现,尽管以图找图的专指任务 3 与基于文字找图的专指任务 1 在相关性标准使用上存在显著差异,但是与同是以文字找图的专指任务 2 在各项标准上的使用差异不显著。因此,是否基于文字或者图片给出检索需求描述,与用户在检索中评价检索结果的标准之间并没有显示出显著的关联性,与一般性任务下的发现是一致的。

抽象性任务同样包括 3 个具体任务,查找干练发型的任务为“抽象性任务 1”,查找具有想当年感觉背景的任务为“抽象性任务 2”,查找表征人间真爱的任务为“抽象性任务 3”,经过 K 个相关样本的非参数检验分析得到表 6。

表 6 抽象性任务在相关性标准重要性评价上的内部差异分析

相关性指标	卡方值	P	抽象性 任务 1 $N = 30$	抽象性 任务 2 $N = 30$	抽象性 任务 3 $N = 30$
			均值	均值	均值
主题	2.52	0.284	4.60	4.47	4.53
准确性	0.116	0.944	4.50	4.47	4.43
启发性	23.432	0.000 *	3.50	3.83	4.67
吸引力	4.622	0.099	4.00	4.40	3.83
完全性	0.473	0.789	4.13	4.17	4.30
视觉特征	5.890	0.053	3.87	4.00	3.30
情感	23.238	0.000 *	3.20	4.20	4.27
文本上下文	4.353	0.113	4.37	3.97	4.27
引申含义	8.747	0.013 *	3.30	3.97	4.03
视觉冲击	10.457	0.005 *	3.13	3.70	3.97

*, 0.05 水平显著

表 6 反映出,用户在完成三个抽象类任务时,对启发性、情感、引申含义和视觉冲击四项标准的重要

性认可上存在显著差异。在启发性标准上,查找反映人间真爱的“抽象性任务3”显著高于其他两类任务,并且,也高于本实验中其他六个任务在该标准上的评分均值,其他六个任务在该标准上的评分均值都不到4分。这一情况与该任务描述流露出图片查找目的是“希望找点灵感”有关,结合用户访谈数据发现,被试在完成过程中会尽量切合题意进行结果筛选,因此对启发性的评分较高(4.67)。可见,检索目的一定程度会影响用户筛选图片的标准选取。

在其他三个出现显著性差异的标准上,“抽象任务2”和“抽象任务3”对这几项标准的重要性评价均值都要显著高于“抽象任务1”,这与两方面原因有关:一方面,“抽象任务3”中的“人间真爱”属于一个抽象概念,与人类情绪、情感有密切关系,并且也需要有一定的引申和用户个人的主观解释。例如,访谈中有被试指出,当她看到检索任务中要求查找表现人间真爱的图片,头脑中就在想象什么能够体现人间真爱,然后就想到了亲情,因为她认为爱情不可靠,亲情更能体现人间真爱,所以直接用使用“亲情”来检索。结合被试实际提问数据也可以看到,30位被试中,有一半的用户仅使用“人间真爱”、“真爱”、“爱”等抽象提问进行检索,另一半用户则会结合使用“父爱”、“母子情深”、“地震救援”等对人间真爱的进一步诠释的关键词进行检索。无论是哪一种情况,被试在筛选结果时都遵循着个人对人间真爱的理解。类似地,“抽象任务2”中提出的“想当年”的感觉也较为主观,图片中应包含的对象并不明确,因此,这两个任务在情感、引申含义和视觉冲击上的评分均值都高于“抽象任务1”。

另一方面,尽管查找发型的“抽象性任务1”在任务描述中使用了“干练、有活力”这样的形容修饰性词汇,这类特征属于图像特征体系中的情感语义^[22],但是,可能用户会认为与人类情绪有关的才可称为情感,打动人的才是有情感的,而形容主观体验和感觉的未必都属于情感范畴。因此,“抽象任务1”在情感相关性标准上的评分均值并不高,仅为3.2。从这三个抽象性任务的差异分析可以看到,尽管三个任务中都不同程度包含了主观性的描述词汇,但是因为抽象程度,或者用户对情感标准理解的差异,使得在用户在三个任务中对相关性标准的重要性评价存在不同。

5 讨论

由前述分析可以看到,在图像检索过程中,用户

筛选和评价检索结果图片相关性时,会综合地运用到主题、准确、启发、吸引等十项标准,并且体现了如下特点。

(1)用户在评价图像检索结果相关性时,重要性最高的前五项标准为主题、准确性、完全性、文本上下文和吸引力。尤其是主题和准确性标准,被试在访谈中也表示他们会首先关心检索结果是否符合题目要求,是否属于题目规定的主题。类似地,Hamid和Thom^[18]根据用户对相关性标准的等级评价结果发现,用户筛选图片时最关心的前两个标准是主题和准确性。又如,Hung^[17]根据用户完成图像检索任务后的访谈数据分析,发现“典型性”是用户最常用的相关性标准,该研究将“典型性”定义为图片能够典型代表用户需求主题的情况,与“主题”在含义上是一致的。Choi和Rasmussen^[16]在综合检索前后的相关性评价标准基础上也指出过,主题标准最为重要。相近的结果在网页检索研究中也有发现,如Savolainen等^[23]研究用户选择超链接和网页的相关性标准时指出,主题、专指度、熟悉程度以及全面性都是使用最多的评价依据。可见,主题和准确性在不同媒体类型检索中都被认同的重要标准^[24-26],这一点也体现了图像与文本在检索相关性评价的重要标准上具有相似性。尽管图像、音视频检索相对文本检索更为主观,但是,用户依然有一个内在的判断尺度。不过,准确性标准的重要性认同高,还可能与实验环境下被试希望尽可能表现出色有关。他们可能担心任务完成不佳而被研究人员认为能力低下^[27],因此,会尽力让自己找到准确符合要求的图片。在访谈过程中,有被试提到,对于一些抽象模糊的任务描述,曾经担心过找得对不对,希望自己能够找到更符合题意要求的图片。

此外,完全性、文本上下文和吸引力标准也是用户认为较为重要的标准,进一步支持了已有研究的发现,如Choi和Rasmussen曾指出结果图片对检索需求要求的细节包含的完全程度在检索前与检索后都被认为是重要的相关性评价标准^[16]。文本上下文标准的重要作用在医学图像检索研究中也曾得到过揭示^[13],只不过医学图片检索还会非常注重视觉特征,这一点与本研究的发现不同。在本研究中,视觉特征的重要性排名居中,完全性的重要性评分均值都要高于视觉特征,意味着用户更关注图片中是否包含目标对象,然后才是视觉特征。而在医学图像检索中恰好相反,这可能与医学图像用于病患诊疗、对图像中的病兆更为关注有关,所以视觉特征变

得更为重要。对于吸引力标准,Hamid 和 Thom 曾指出该标准对于不同类型的任务而言都属于重要的相关性标准^[18],尤其对于青少年用户,吸引力是他们选择图画的重要依据之一^[15]。由此也进一步说明,不同类型的用户群体,或者面向不同领域的图片进行检索时,图像检索相关性评价标准会有一些的变化。

相比而言,十项标准中重要性认可最低的是情感和视觉冲击标准。不过,虽然评分在十项标准中排名靠后,但是,评分均值依然高于 3 分,表明用户对这两项标准的重要性还是有一定的认可。十项标准都被用户应用于对结果的评价,这体现了评价选择过程的复杂性,而且用户在访谈中还提及了审美、新颖性、个性化等补充标准。

(2) 用户在执行专指、抽象程度不同的任务时,对相关性标准重要性的认可程度存在差异。一方面,从相关性标准的重要性认可来看,检索任务越专指,用户越倾向于根据主题相关、检索准确进行结果评价,这与专指任务对确切需求的吻合度要求更高有关。类似地,Hamid 和 Thom^[18]也曾指出,在专指检索任务中,准确性标准的重要性要高于一般任务,因为在专指任务中用户更清楚所要查找的对象。对于专指性和一般性任务,用户都会重视结果图片对需求对象包含的完全性,这与两类任务通常更容易明确结果图片中应包含的对象有关,加之在实验室环境下进行研究,被试为了追求查找结果准确性,也会注意结果图片对题目要求细节包含的完全程度。此外,任务越抽象,用户则越倾向于根据图片的启发性、所激发的情感进行结果评价。不过,在抽象性任务内部,当任务中出现的检索要求越接近人类基本情感,用户则越倾向于根据图片所激发的情感进行结果评价。所以,当检索需求中出现的抽象概念更能引申出情感方面的需求时,可能会比直接使用表达情感的描述词对用户产生的影响更大,这是由用户对情感标准的理解所决定的。这一发现也进一步支持了已有的发现,无论是采用出声思维法,还是采用量表评价的方法收集用户评价相关性的依据,不少研究都曾指出用户在评价结果相关性时会提及情感标准,对于图像、音视频等多媒体信息更是如此,如 Inskip 等^[28]曾指出用户经常根据情绪、心境对乐曲的相关性做出评价。所以,不少研究者已经从技术层面在探讨基于情感的图像、音视频特征提取和检索问题^[29,30]。

另一方面,从认可重要性的相关性标准数量来

看,本研究发现一般性任务和抽象性任务中重要性认可的标准数(6 个)略高于专指性任务(5 个),基本相当。这与 Hamid 和 Thom^[18]的发现有所不同,他们的研究指出专指性任务中用户会使用更多的相关性标准(7 个),高于一般性任务(4 个)和抽象性任务(5 个),认为与专指性任务更为明确有关。这一不同,可能与本研究针对三类任务又下设了 2~3 个具体任务有关,如本研究中寻找人间真爱的抽象任务 3,用户认可重要性的标准多达 7 个,而查验品牌鞋正品的专指性任务 1 被用户认可重要性的标准为 5 个。但是,查找球员赛场比赛的专指任务 2 被用户认可重要性的相关性标准数量也有 7 个,又要高于查找发型图片的抽象性任务 1。所以,在不同类型下,具体任务在认同的重要性标准数量上有一定的交叉,这可能与所设置任务描述的具体程度有关,本研究对检索目的进行了具体描述,可能引发用户由检索目的而引申出相关的标准,比如查找桌面背景图片的任务,用户通常会附加上在审美上的要求。因此,重要性认可的相关性标准数量并不完全取决于任务基本类型。

(3) 检索目的、任务描述方式会一定程度影响相关性标准的重要性认可情况。首先,检索目的对相关性标准的使用会产生显著影响。类似地,Sedghi 等^[31]也曾指出图像检索需求、目的不同会使得用户使用不同的相关性评价标准。一方面,本研究发现,当图片作为数据来源使用时,对准确性要求也会更高,而当作图片对象来源时,评价标准更为主观,也印证了 Fidel^[21]提出的两极理论中对数据极和对象极检索目的的差异的分析。另一方面,本研究设计的七个问题,都可以对应到 Conniss 等^[32]提出的七种图像用途,该研究曾在图像用户调研基础上总结了将图片用作插图、进行信息处理、信息传播、学习、启发灵感、审美和情感传递和感染七种常见用途。如在本研究中,查找插图启发灵感的抽象任务 3 就属于将图片作为启发性用途的任务检索,因此用户在评价相关性标准时对启发性的评分最高,也高于其他六个任务在该标准上的得分。同时,当图片用作情感传递(抽象性任务 2)、审美(专指性任务 2)时,情感标准的得分也较高。又如,Hung^[17]也曾指出因为新闻记者在找寻图片时,会考虑新闻图片表现的内容及其对社会公众的情感冲击,因此,情感标准是他们筛选图片的重要标准之一。可见,无论是被试被动地为了切合题意而使用相关性标准,还是用户在自然情境下使用相关性标准,检索目

的都发挥着重要的指导性作用。

其次,图像需求描述方式与相关性评价标准的使用有一定关系,主要体现在检索描述中涉及图像语义特征的类型,而检索需求的表现形式是文字还是图片并不会显著影响相关性标准的使用。一方面,本研究重点考察了图像检索需求描述中出现不同类型的情感语义的情况,结果发现抽象程度的差异,以及用户对情感标准的理解和认识都会引起评价结果相关性时倾向的标准不同,越是与人类基本情绪有关的检索要求,用户越是会考虑情感标准,而对人类基本感觉,抑或对客观事物的一般风格的检索要求,用户则不太会将其归为情感标准判断的范畴。因此,尽管同属于抽象性任务,仍旧存在相关性标准使用差异的情况。另一方面,本研究并未发现以图找图与以文字找图两种情况下用户对相关性标准的重要性评价存在显著差异。类似地,Fokumoto^[33]也曾指出,文字描述性任务和以图找图任务在相关性评价标准上不存在显著差异,尽管这两类任务在查看页面数量、检索策略使用数量、检索时间表现为前者显著高于后者。

6 结 语

本研究参考 Hamid 和 Thom 总结的十项图像相关性评价标准,设计了一般性、专指性和抽象性三种类型的七个具体检索任务进行实验研究,探讨不同任务类型下相关性评价标准使用的关系。本研究的发现揭示出,主题相关性、检索准确性、图片上下文文本的相符程度、完全性和吸引力是不同类型任务下重要性认可最高的五个相关性标准;不同专指程度的任务在相关性评价标准应用的程度上存在一定差异,任务越专指,用户越倾向于根据主题相关、检索准确进行结果评价;任务越抽象,用户则越倾向于根据图片的启发性、所激发的情感进行结果评价。一般性任务和专指性任务都更重视结果图片对需求对象包含的完全性。并且,相同专指程度的任务内部,可能因为图像用途、检索需求描述涉及图像特征的不同,而在相关性评价标准的应用程度上存在内部的差异,图像检索需求以文字方式给出还是图片方式给出并不影响相关性评价的依据。任务类型与图像检索相关性评价之间的关系对于优化和改进图像检索服务有一定的现实指导意义。例如,高频使用、重要性认可高的相关性标准不仅可以用于改进检索功能和界面设计,还可以用于改进检索结果输

出方式。对于用户认为重要而在检索系统中未能提供的相关性标准可以进行补充,如情感标准对于抽象性任务而言是重要的相关性标准之一,借助情感语义抽取技术,对图像在物理层、风格层、情绪层及审美多个层次的特征加以揭示,既丰富了检索入口的功能设置,又为检索结果的多样化输出提供了可能。又如,完全性亦是重要的相关性标准之一,然而基于内容检索技术准确抽取广泛图像中的对象和高级语义难以跨域“语义鸿沟”,因此,可以结合集成了大众智慧的图像标签,通过标签反映的图像内容特征,更便利地揭示图片中包含对象的种类和数量,从而为结果排序提供更多的选择。

本研究的不足主要有:尚未对用户的专业背景、对任务的熟悉度、任务难度等因素与相关性标准之间的关系进行讨论;并且,样本规模较小,预设的相关性标准也比较有限,所得到的研究发现在推广前还需要经过更大样本的检验,通过扩大被试和任务类型和数量,增加对更多自变量的考察,以提高研究发现的普适性。此外,本研究发现用户对相关性标准的重要性认可度总体上不错,从均值来看最低的标准也要高于3分,尽管这一结果体现了用户对多种标准综合运用的认可,但也可能与利用等级评估方法收集用户对相关性标准的评分时,通常用户打分会偏高有关^[5]。因此,还有待结合自然观察、出声思维、日记法等多种方法进行进一步研究,以得到更为自然、稳定的用户评价数据。

在未来研究中,还可以进一步挖掘国内用户评价图像检索结果相关性的标准,丰富和细化相关性标准的层次和类型,并从通用程度、阶段过程,以及用户类型、资源类型等角度进行更系统的理论归纳。此外,还应加强社交媒体环境下的相关性评价研究,如社会情感已经被发现在社交问答系统中影响着人们对最佳答案的选择^[34]。图像在社交传播中的使用也很广泛,因此,有必要在社交媒体环境,结合社交网络分析、协作式检索等理论和方法进一步探究对图像选择评价标准的影响。最后,加强用户视角的研究与技术实现的衔接,对主题相近而资源形态不同的用户选择标准进行研究,为垂直检索技术^[35]的改进提供更多理论支持,推动相关研究发现对实践的参考和指导作用。

参 考 文 献

- [1] Saracevic T. Relevance: A review of the literature and a framework for thinking on the notion in information

- science. Part II: nature and manifestations of relevance [J]. Journal of the American Society for Information Science & Technology, 2007, 58(13):1915-1933.
- [2] 李亚琴, 孙建军, 杨月全, 等. 基于信息检索用户的相关性行为研究进展[J]. 情报科学, 2014(5): 157-160.
- [3] 黄崑, 白雅楠, 周晓分, 等. 图像信息需求研究综述[J]. 图书情报工作, 2014(6):135-141.
- [4] 李月琳, 胡玲玲. 基于环境与情境的信息搜寻与搜索[J]. 情报科学, 2012(1):110-114.
- [5] Mizzaro S. Relevance: the whole history[J]. Journal of the American Society for Information Science, 1997, 48(9):810-832.
- [6] 付玲玲. 信息检索相关性研究综述[J]. 情报探索, 2010(12): 77-79.
- [7] 成颖. 相关性判断研究综述(2000-2010)[J]. 情报杂志, 2011, 30(9): 79-84.
- [8] 王雅坤, 成全. 信息检索相关性研究综述及发展趋势[J]. 图书与情报, 2012(1): 88-94.
- [9] 李亚琴, 孙建军, 杨月全, 等. 基于信息检索用户的相关性行为研究进展[J]. 情报科学, 2014(5): 157-160.
- [10] 王健, 王志强, 刘茜, 等. 认知转向背景下用户相关性判断研究的方法论观察[J]. 图书情报工作, 2014, 58(18): 66-76.
- [11] Markkula M, Sormunen E. End-user searching challenges indexing practices in the digital newspaper photo archive [J]. Information retrieval, 2000, 1(4): 259-285.
- [12] Hung T Y. Search strategies for image retrieval in the field of journalism[D]. New Jersey: Rutgers University, 2006.
- [13] Sedghi S, Sanderson M, Clough P. A study on the relevance criteria for medical images [J]. Pattern Recognition Letters, 2008, 29(15):2046-2057.
- [14] Sedghi S, Sanderson M, Clough P. How do health care professionals select medical images they need? [J]. Aslib Proceedings, 2012, 64(4):437-456.
- [15] Hirsh S G. Children's relevance criteria and information seeking on electronic resources [J]. Journal of the american society for information science, 1999, 50(14): 1265-1283.
- [16] Choi Y, Rasmussen E M. User's relevance criteria in image retrieval in American history [J]. Information Processing & Management, 2002, 38(5):695-726.
- [17] Hung T Y, Zoeller C, Lyon S. Relevance judgments for image retrieval in the field of journalism: A pilot study [C]//Digital Libraries: Implementing Strategies and Sharing Experiences. Berlin:Springer Berlin Heidelberg, 2005: 72-80.
- [18] Hamid R A, Thom J A. Criteria that have an effect on users while making image relevance judgements[C]//Proceedings of the Fifteenth Australasian Document Computing Symposium. Melbourne, Australia: School of Computer Science and IT, RMIT University, 2010: 1-8.
- [19] Shatford S. Analyzing the subject of a picture: A theoretical approach [J]. Cataloging & Classification Quarterly, 1986:39-62.
- [20] Huang K, Niu X, Wang S S, et al. Chinese web users' daily image needs and seeking behavior in a Q&A community [J]. Chinese Journal of Library and Information Science, 2015, 8(1): 1-20.
- [21] Fidel R. The image retrieval task: implications for the design and evaluation of image databases[J]. The New Review of Hypermedia and Multimedia, 1997, 3(1): 181-199.
- [22] 黄崑, 王珊珊, 耿骞. 国外图像特征研究进展与启示[J]. 图书情报工作, 2015, 59(8): 138-146.
- [23] Savolainen R, Kari J. User-defined relevance criteria in web searching[J]. Journal of Documentation, 2006, 62(6):685-707.
- [24] Tang R, Solomon P. Use of relevance criteria across stages of document evaluation: On the complementarity of experimental and naturalistic studies[J]. Journal of the American Society for Information Science & Technology, 2001, 52(8):676-685.
- [25] Lawley K N, Soergel D, Huang X. Relevance criteria used by teachers in selecting oral history materials[J]. Proceedings of the American Society for Information Science & Technology, 2005, 42(1).
- [26] Crystal A, Greenberg J. Relevance criteria identified by health information users during Web searches [J]. Journal of the American Society for Information Science & Technology, 2006, 57(10):1368-1382.
- [27] Kelly D. Methods for evaluating interactive information retrieval systems with users [M]. Foundations and Trends in Information Retrieval ;2009:182-192.
- [28] Inskip C, Macfarlane A, Rafferty P. Creative professional users' musical relevance criteria [J]. Journal of Information Science, 2010, 36:517-529.
- [29] Lin K C, Huang T C, Hung J C, et al. Facial emotion recognition towards affective computing-based learning [J]. Library Hi Tech, 2013, 31(2):294-307.
- [30] Knautz, K, Stock, W G. Collective indexing of emotions in videos[J]. Journal of Documentation, 2011, 67(6): 975-994.

- [31] Sedghi S, Sanderson M, Clough P. A study on the relevance criteria for medical images [J]. Pattern Recognition Letters, 2008, 29(15):2046-2057.
- [32] Conniss L R, Ashford A J, Graham M E. Information seeking behavior in image retrieval: Visor 1 Final Report [R]. Library and Information Commission Research Report 95. Institute for Image Data Research, Newcastle upon Tyne, 2000.
- [33] Fukumoto T. An analysis of image retrieval behavior for metadata type image database [J]. Information Processing & Management, 2006, 42(3): 723-728.
- [34] Kim S, Oh J S, Oh S. Best-answer selection criteria in a social Q&A site from the user-oriented relevance perspective[J]. Proceedings of the American Society for Information Science & Technology, 2007, 44(1):1-15.
- [35] Arguello J, Diaz F, Callan J. Learning to aggregate vertical results into web search results[C]// Proceedings of the 20th ACM international conference on Information and knowledge management. New York: ACM, 2011:201-210.

(责任编辑 刘志辉)