

国家、地区媒体形象的数据挖掘

—— 基于认知心理学与计算机自然语言处理技术的视角

葛 岩 赵 海 秦裕林等*

摘 要 从海量媒体报道中合理挖掘与国家、地区有关的信息,是大数据时代国家、地区形象研究需要解决的重要问题。研究解决该问题的理论思路、具体方法和实际挖掘实验的结果表明:(1)依据认知图式模型,人们对外部事物的知觉和判断通过认知图式完成,而图式中稳定存储的所涉事物的属性和及其默认值会影响注意力、记忆和判断,从而影响对外部事物的形象构建;(2)TF-IDF等文本分类技术能够挖掘媒体报道中稳定关注的事物和对这些事物的评价,挖掘结果可视为媒体认知图式中属性及其默认值的映射,能够帮助刻画、分析和追踪媒体报道中的国家、地区形象;(3)对《纽约时报》十年(2003—2012)中国报道总体(N=13,259)的挖掘实验显示,通过进一步完善,本文提供的理论模型和挖掘技术应具有操作的可行性、结果的客观性、数据的全面性和广泛的应用性。

关键词 国家形象 地区形象 媒体形象 数据挖掘 TF-IDF

作者葛岩,上海交通大学媒体与设计学院教授、凯原法学院双聘教授(上海 200240);赵海,上海交通大学计算机科学与工程系副教授(上海 200240);秦裕林,上海交通大学凯原法学院访问特聘教授,社会认知与行为科学研究院研究员(上海 200240)

中图分类号 G20

文献标识码 A

文章编号 0439-8041(2015)07-0163-08

背景与问题

国家、地区、城市、企业、产品的媒体形象是否正面,这是政府、企业和各种社会机构都会关注的问题。利益相关方无不希望拥有正面媒体形象,为自己带来显性或隐性的利益。^①欲研究媒体形象,需要知道形象是怎样的;制定形象传播战略或策略,需要以测量到的形象为依据;要知道改善形象所作的努力是否奏效,需要在传播计划形成之前,实施之后分别测量形象的变化。因此,如何从海量媒体报道中提取和分析有关信息是媒体形象研究的重要问题。

* 本文作者还有陈长阁、何俊涛、徐剑、卢嘉杰和李晓静。

① [美]西蒙·安浩:《铸造国家、城市和地区品牌:竞争优势识别系统》,葛岩、卢嘉杰、何俊涛译,上海:上海交通大学出版社,2010年,第9—11页。[以]埃里·阿夫拉汉姆、伊兰·科特:《地区危机传播:实用媒体策略——改善城市、国家、旅游地的形象》,葛岩、卢嘉杰、何俊涛、文炳森译,上海:上海交通大学出版社,2013年,第25—39页。

传播研究中,内容分析是最常见的媒体形象测量方法。比之于文本阅读,内容分析通过编码将非结构化的文本数量化,便于统计处理,结果也相对客观。不过,内容分析有两个弱点:第一,人工编码耗时费力,通常只能处理不多的样本和较具体的新闻事件,如《纽约时报》对汶川地震的报道,《泰晤士报》对昆明火车站恐怖袭击事件的报道。对于时间跨度大,报道量大的研究问题,比如《华尔街日报》十年间对中国环境问题的报道,《朝日新闻》五年来对上海报道,则难以处理。第二,内容分析中许多编码指标是客观的,如报道的字数、发表版面、是否配有图片和表格等,但还有一些指标相对主观,如报道内容是否客观,报道评价倾向是正面或负面。对后一种指标,编码员的主观偏差很难避免。在人力和时间相对节省,主观偏差得到有效控制的条件下,能否找到处理更多的样本乃至总体的方法?

针对上述问题,使用国家和地区的媒体形象为例,本文介绍我们所做的一些探索。

媒体形象与认知图式

媒体每天都会发表成百上千的报道,报道好事,也报道坏事。假定一个国家或城市的媒体形象存在于这些报道之中,怎样才能找到它们,将之整合到一个简单,直观,便于追踪和分析的“形象”之中?

人类认知的信息加工理论假设,面临认知与判断任务,外部信息进入人脑,记忆里储存的相关信息会被激活,通过“自下而上”(外部信息刺激)和“自上而下”(人脑中已经存储了的知识、观念、态度、价值等)的双向加工形成知觉与判断。^①据此,媒体报道中有关国家或地区的信息会在记忆中储存,日积月累。当面临需要判断的情境时,如选择购买意大利还是巴西生产的香水,判断美国代表在联合国大会上的发言是否旨在遏止中国崛起,相关国家的信息会被提取并加入判断过程。信息加工理论还说明,外部输入的信息不可能全部储存在记忆之中,大量信息会消失,曾经熟记的信息也可能因为不常提取和使用而遗忘。只有那些进入长时记忆并具有高易得性(accessibility)——即容易想到的信息——在相关情境中才可能被迅速提取,并对知觉和判断产生影响。^②

什么样的信息会进入长时记忆,稳定地存储在记忆中并容易被激活?记忆研究显示,激活程度决定了长时记忆中的信息被提取的可能性和速度。具体言之,信息是否被激活,激活程度如何,由当下的环境与被提取的信息的关联程度,以及该信息本身的重复出现率和最近出现的时间所决定。^③特定环境与特定信息的关联程度通常是具体化的,很难得出一般性结论。退而求其次,可使用简单但较为有效的指标来判断记忆信息激活强度:重复出现率。一般而言,经过不断重复,信息的记忆时间长度、易得性和被信赖程度都会有所提高。^④

依此去看媒体形象,在报道特定事物(如国家、地区、性别、种族、文化、群体、职业、企业、产品等)时,媒体会有较稳定的关注点、解释框架、评价乃至字汇。例如,论及中国政治,西方媒体会反复提及一些负面政治事件(如与西藏,台湾、20世纪80年代末政治风波有关的事件);对于美国在华连锁快餐店,我国主流媒体会稳定关注其食品安全问题,解释框架也稳定地设定在管理缺失、对中国市场歧视性对待等方面。媒体的这类行为被称作对报道对象的“特征建构”(trait construct),其功能是在为所涉对象和一组特征间建立稳定的联想关系,以致所涉对象出现时,受众记忆中的联想关系便自动激活,隐性或不知不觉地影响随之而来的判断和评价。^⑤使用广告所进行的品牌传播或是最典型的特征建构活动,如为高尔夫赋予“高大上”

①③ [美] 约翰·安德森:《认知心理学及其启示》(第七版),秦裕林、程瑶、周海燕、徐玥译,北京:人民邮电出版社,2013年,第58—59、169—200页。

② R. Fazio, M. Powell & C. Williams, "The Role of Attitude Accessibility in the Attitude-to-Behavior Process", *Journal of Consumer Research*, 16(1989), pp.280—288. J. Bassili, "Response Latency and Accessibility of Voting Intention: What Contributes to Accessibility and How It Affects Vote Choice", *Personality and Social Psychology Bulletin*, 21(1995), pp.686—695.

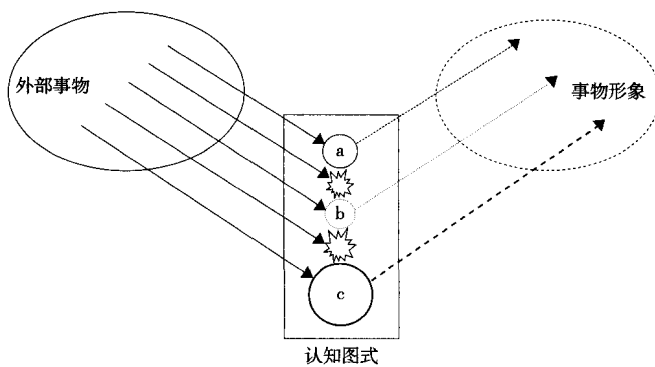
④ R. Holland, B. Verplanken & A. van Knippenberg, "From Repetition to Conviction: Attitude Accessibility as a Determinant of Attitude Certainty", *Journal of Experimental Social Psychology*, 39(2003), pp.594—601. 葛岩:《重复广告效果的分析》,《现代传播》2005年第2期。

⑤ S. Fiske, "Stereotyping, Prejudice, and Discrimination at the Seam Between the Centuries: Evolution, Culture, Mind, and Brain", *European Journal of Social Psychology*, 29(2000), pp.188—211. A. Cuddy & S. Fiske, "The BIAS Map: Behaviors from Intergroup Affect and Stereotypes", *Journal of Personality and Social Psychology*, 92(2007), pp.631—648.

的联想,为奢侈化妆品赋予美丽、青春、健康的联想。就其片面性、易得性和影响方式而言,被建构起来的媒体形象与社会心理学所谓“刻板印象”相似,故大量媒体研究文献常使用“媒体刻板印象”的概念来称呼媒体对于特定事物、族群的稳定的,带有偏见的报道方式。据此,有理由把媒体形象看作一种刻板印象,它是由媒体从业者、媒体制度对所涉对象的一种稳定报道模式。

从认知心理学的角度看,刻板印象是所谓认知图式(cognitive schemata)中的一类。认知图式被描述为表征事物属性知识的插槽结构(slot structure)。该结构有若干插槽,每个插槽表示事物的一个重要属性,每个属性有对应的默认值(default value)。^①例如,“鸟”的认知图式中可能“能否飞翔”“是否有羽毛”“筑巢能力”“生蛋与否”“攻击性”等属性插槽,与插槽对应,还有“会飞”“有羽毛”“会筑巢的”“能生蛋”和“一般无攻击性”等默认值。当被认为是“鸟”的事物出现,认知图式随之激活,人借助属性插槽和其默认值来知觉、记忆和判断。

图1 认知图式与形象



依据该模型,外部事物的信息进入知觉时,便会与图式中的插槽发生相互作用(图1)。与插槽(a, b, c)对应的信息更可能获得注意,更可能通过认知图式的加工投射到的对事物的认知或事物形象之中。其他信息则更可能被忽略,或被认知图式完全屏蔽。在投射过程中,插槽既有的属性默认值会影响该属性投射到形象时的权重。例如,外部信息通过默认值高的c插槽后,属性权重会被放大;通过默认值低的b插槽后,属性权重会被缩小。该模型显示,经过认知图式的加工,事物的原本信息状态与事物的形象之间

难以避免地发生偏离乃至歪曲。例如,一个未见到过鸵鸟的人看到“鸵鸟”这一词汇,“鸟”的认知图式中的属性插槽“能否飞翔”及其默认值“会飞的”可能会从记忆中提取,使她或他以为鸵鸟也能够飞翔。

利用认知图式可以极大提高认知效率。在注意力有限或仅能获得有限信息的条件下,认知图式能够帮助将注意力限定在外部事物的一些常见且重要属性上,能够依据有限的信息提供的类别知识推想出更多的有关知识。然而,类别化、固化的认知图式也有其弊端,它不但可能引发如“鸵鸟会飞”那样的判断,在社会认知领域,它可能仅通过对如性别、口音、阶层、肤色、国别、地域差别等属性的知觉,便联想到一系列其他属性,无论这些属性在所欲判断的具体事物那里是否重要,进而诱发基于刻板印象的负面或正面的歧视。^②

上述观点或可称为基于认知图式的形象理论模型(Schemata-based Image Theory, 或, SIT),它对形象的形成机制和构成要素有不错的解释力。对于媒体形象研究,对于从海量新闻报道中找到与国家、地区或产品、企业形象密切相关的信息,该模型能否提供有价值的思路?回答是肯定的。

假定对于某地区有足够数量的报道,比如800篇,且这些报道覆盖了足够长的一段时间,比如十年,若某些字汇在报道中反复出现,则说明这些字汇表征的东西受到媒体稳定、持续的重视。表征四川时,这些字汇可能包括“麻辣”“美食”“烹饪”等,果如此,便可认为川菜是四川形象构成中的重要属性。若与这些字汇一起出现频度很高的字汇是“喜爱”“流行”等,后者便可视作川菜的默认值。我们据此推测,从媒体报道中可以提取与认知图式类似的字汇模式,它透露出媒体选择了哪些属性和评价来表征一个国家或一个地区。

① W. Brewer & J. Treyns, “Role of Schemata in Memory for Places”, *Cognitive Psychology*, 13(1981), pp.207—230.

② S. Fiske, “Stereotyping, Prejudice, and Discrimination at the Seam Between the Centuries: Evolution, Culture, Mind, and Brain”, *European Journal of Social Psychology*, 29(2000), pp.188—211. A. Cuddy & S. Fiske, The BLAS Map: Behaviors from Intergroup Affect and Stereotypes”, *Journal of Personality and Social Psychology*, 92(2007), pp.631—648.

接下来的问题是,有没有一种技术能够合理地抽取上述属性和默认值?

媒体形象的基本算法

依据认知图式-刻板印象模型,通过文本阅读方法也可以提取属性和默认值一类要素。不过,这意味着工作量巨大,且阅读者的主观偏差难以避免。在大数据时代,这种方法更显得力不从心,甚至匪夷所思。值得庆幸的是,文本挖掘技术提供了解决这一问题的可能性。^① 本文所分析的是与我们所做的国家、地区媒体形象研究直接相关基本算法,词频-逆文本词频(term frequency-inverse document frequency, TF-IDF),一种用于文本检索与分类的技术。^② 其表述如下:

$$TF-IDF_{(w)} = TF_{(w)} * \log D/TF_{(w)}$$

TF-IDF 的基本逻辑简洁明了:若某词 w 在一类文献中出现频率很高,但在其他数量为 D 的不同类型文献中出现频率很低,则判定该词具有类别区别力。^③ 举例来说,需要了解某报纸对基督教在中国状况的报道,关键字汇有“基督教”,“在”和“中国”。假定一篇围绕该主题的典型报道有 1000 个字汇,三关键词分别出现 2, 35, 5 次,词频分别为 0.002, 0.035, 0.005。三词频加和为 0.42。直观地看,若某报道中三者的 TF 较高,该报道与主题“基督教在中国”的相关性可能也会较高。但这种看法有明显偏颇,例如,任何报道都可能出现大量“在”,但该词与主题类别并无必然关系。这类字汇可排除不计,即设“在”的 TF 等于 0.000,“基督教”和“中国”的 TF 和于是为:0.002+0.005=0.007。

进一步分析,若研究总体是对关涉中国的报道,“中国”在每一篇或绝大多数报道中可能都会出现,“基督教”出现频率却很低,但后者与主题“基督教在中国”的关联性应该更高。一个词反复出现,固然可能表明它的重要性,但如果该词在不同的类别的文件中也频频出现,它便失去了类别特征,对区分主题来说变得不再重要。这类字汇使用 IDF 辨别。按照 IDF 的逻辑,如果某类文献数量为 D ,一个重要字汇 w 在 D_w 份报道中出现,则 D_w 越大, w 此类报道中的权重越小; D_w 越小, w 的权重越大。IDF 用 D/D_w 以 10 为底的对数来计算,即 $\log(D/D_w)$ 。例如有 10000 篇涉及中国的报道,“中国”出现了 5000 次,“基督教”出现 30 次。“中国”的权重为 $\log(10000/5000)=\log(2)=0.301$,“基督教”的权重为 $\log(10000/30)=\log(333.33)=2.523$ 。就“基督教在中国”这一主题而言,虽然“中国”出现的频率更高,但“基督教”的重要性是“中国”的 8.38 倍。如此,TF 和 IDF 对字特定主题类别的关联性给出了量化评估: $TF_{\text{基督教}} * IDF_{\text{基督教}} = 0.002 * 2.523 = 0.005$, $TF_{\text{中国}} * IDF_{\text{中国}} = 0.005 * 0.301 = 0.0015$, 当设定 $TF_{\text{在}} = 0.000$, 短语“基督教在中国”的权重和则为 $0.005 + 0.000 + 0.0015 = 0.0065$ 。

逻辑上,不同主题对应的字汇权重模式理应不同,而权重高的字汇应反映出媒体报道特定主题时最关注的该主题的属性。据此我们推测,TF-IDF 应能挖掘出国家、地区形象基本属性的字汇表征模式。

媒体形象的挖掘实验

为检验上述推测,我们采集了《纽约时报》(2003—2012)中国报道为研究总体($N=13,259$),使用 TF-IDF 等算法,挖掘了中国国家和地区字汇表征模式,或媒体形象(见图 2)。

(一) 工作流程

1. 总体下载。首先,我们使用关键词,从英文报刊数据库中下载了《纽约时报》十年间全部与中国事务有关的报道(图 2,步骤 1)。

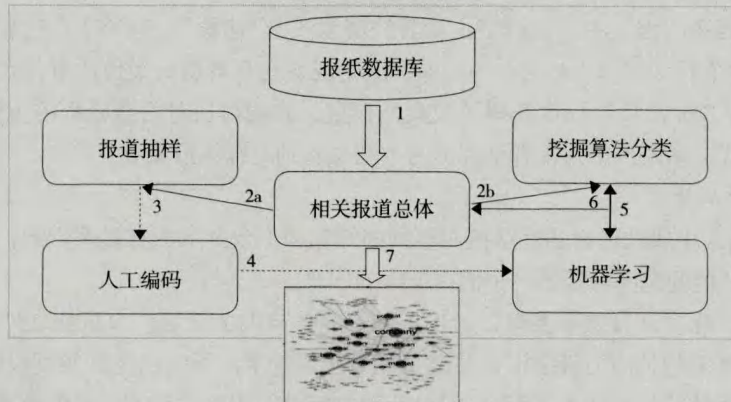
2. 样本抽取,算法分类。按照报道发表数量的年度分布,我们采取配额随机抽样的方法,从报道总体抽取供编码用的样本($N=1,576$)(步骤 2a)。同时,由软件系统按照 TF-IDF 等算法对总体报道做出字汇表征表示的分类(步骤 2b)。

① M. Berry & M. Castellanosc (eds), *Survey of Text Mining: Clustering, Classification, and Retrieval*, 2007, Second Edition, Springer, pp.v—vi.

② D. Blei, A. Ng, A. & M. Jordan, “Latent Dirichlet Allocation”, *Journal of Machine Learning Research*, 3(2003), pp.993—1022. H. Wu, W. Luk & K. Wong, “Interpreting TF-IDF Term Weights as Making Relevance Decisions”, *ACM Transactions on Information Systems*, 26 (2008), Article 13, pp.1—37.

③ 吴军:《数学之美》,北京:人民邮电出版社,2012 年,第 105—110 页。

图2 媒体形象挖掘流程



3. 人工编码分类。使用编码员对所有样本做量化处理（步骤3）。为避免编码员的主观偏差，编码表只要求标注报道的主题，如报道关涉中国整体形象不同方面（如政治、经济、国际关系、公共卫生、教育与科技、社会民生、文化、娱乐和体育等），关涉中国主要地区（省份、城市，如北京、上海、广东、西藏、山东、浙江、台湾、香港等），而且，每份报道由6个编码员背对背完成，并通过信度检验。

4. 机器学习。使用编码后的样本，令程序分析、学习人工分类的方法，总结出不同类别的字汇模式（步骤4）。

5. 分类模型优化。比照步骤2b和步骤4的分类结果，通过迭代计算减小二者的差别，获得达到一定精度的分类模型（步骤5）。

6. 总体分类。使用步骤5取得的模型，由程序阅读报道总体，做出不同主题报道的分类，获得与不同主题对应的字汇模式（步骤6）。

7. 形象生成。使用TF-IDF和其他量化指标，通过数据可视化处理，生成不同报道主题字汇表征模式的直观形象（步骤7）。

（二）形象属性测量

依据SIT，在认知图式中稳定出现的属性是形象的重要构成元素，最容易得到关注，对知觉与判断最具影响力。我们的工作显示，形象挖掘算法可较清晰地透露报道中最受关注的国家、地区属性。以上海和西藏两个地区为例，图3显示出有关上海报道中TF-IDF权重值最高的关键字汇，圆点的大小表示权重值的差别，连线表示关键字汇共同出现的连带关系。在上海形象中，“公司”“市场”“贸易”“经济”“增长”“商务”“投资”“货币”“出口”“全球”“国际”“美国的”等与经济发展和国际联系相关的字汇数量众多，表明报道

图3 上海形象的字汇表征（关键词）
《纽约时报》（2003-2012）



图4 西藏形象的字汇表征（关键词）
《纽约时报》（2003-2012）

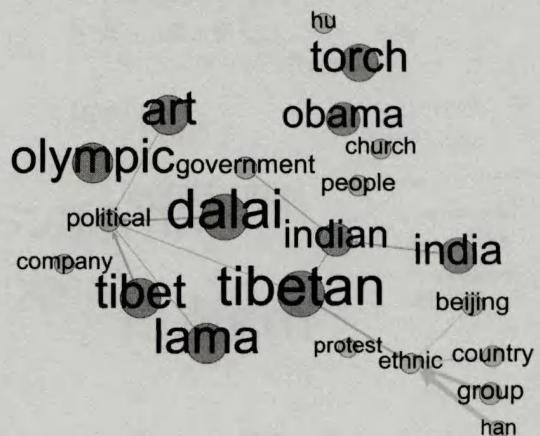


图3—7制图：上海交通大学社会认知与行为科学研究院 赵沁筠、印闯

十分关注上海这一城市的经济和国际化属性,或,经济和国际化是上海形象的基本构成要素。

对比上海,在对西藏的报道中,关键字汇均与经济无关,“达赖”“喇嘛”“艺术”“印度”“奥巴马”等获得很高的权重,“奥林匹克”“火炬”等关键字汇则说明北京奥运火炬传递事件在西藏报道中获得了异乎寻常的关注。此外,“抗议”一词也获得了较高的权重,表明报道对西藏骚乱的重视(图4)。比较上海和西藏的字汇表征方式,新闻报道为读者呈现出两类明显有别的媒体形象。

(三) 属性默认值测量

依据 SIT, 认知图式中属性已有的默认值对事物的知觉和形象的形成有显著影响, 因此, 通过挖掘获得的数据中能否发现默认值的对应物是另一个值得探索的问题。

我们的工作显示, 除了划分文本类型, 给出不同主题各自的关键字汇权重值之外, 稍加处理, TF-IDF 算法还可以计算关键字汇与其他关键字汇, 关键字汇与非关键字汇同时出现的频度, 即字汇的共现率 (co-concurrence)。对共现字的分析有两个层面: 一是分析两字汇共同出现的频度, 频度越高, 则字汇间联系的紧密程度越高; 二是估计共现字汇的价值或情感评价向度, 如关键词“法律”可能与非关键词“公正”之间存在较高的共现率, 而“公正”是正面向度的评价, 关键词“银行”可能与“腐败”之间存在较高共现率, 而“腐败”是负面向度的评价, 等等, 共现词对关键词常常有修饰作用。若将关键词视为属性的属性, 则其共现词的共现频度, 评价向度可视为形象属性的默认值。

以我们的挖掘实验为例, “市场”是上海报道中 TF-IDF 值最高的字汇之一 (图3), 它的共现字汇很多, 有其他关键词 (如“美国的”, “全球的”, “外国的”, “国际的”), 也有非关键词, 如“最快速成长的” (fastest-growing, 6), “黑色, 或地下的” (black, 4), “巨大的” (big, 4), “成熟的” (matured, 2) 等 (图5)。将“市场”看作上海认知图式中的一个重要属性插槽, 这些共现字汇透露出对于“市场”这一属性的评价向度, “快速增长的”“巨大的”“成熟的”趋于正面, “黑色的, 或地下的”趋于负面。共现频度高的评价性字汇, 如“最快速成长的” (共现频度=6), 意味着该字汇与“市场”的关系十分紧密。结合共现率与评价向度便可估算“市场”的默认值。

再如, 在构成西藏形象的关键词中, 除了“抗议”之外, 其他字汇没有明显的价值、道德或情感评价涵义。如果将这些关键词与它们的共现词一起分析, 则可发现“抗议”和“和平的”, “北京”和“强硬的”, “奥林匹克”和“反华”, “民众”和“抵抗汉人”, 和“武装起来的”, “政府”和“专制的”, “强大的”, “教堂”和“非法的”, “贫穷的”, “西藏”和“自由的”存在共现关系 (图6)。换言之, 关键词所表述的属性经作为属性默认值的共现词修饰之后, 《纽约时报》呈现出了一幅十分负面的西藏形象。

(四) 属性变化测量

依据 SIT, 属性插槽和默认值是稳定的, 但如同人们对事物的看法, 它并非一成不变。新闻媒体必须对世界上不断发生的事件做出反应, 有些事件的影响随时间很快消失, 有些影响稍大些, 会得到较长时间的

图5 “市场”在上海形象的共现字
《纽约时报》(2003-2012)



图6 西藏形象的字汇表征 (关键词和共现字)
《纽约时报》(2003-2012)

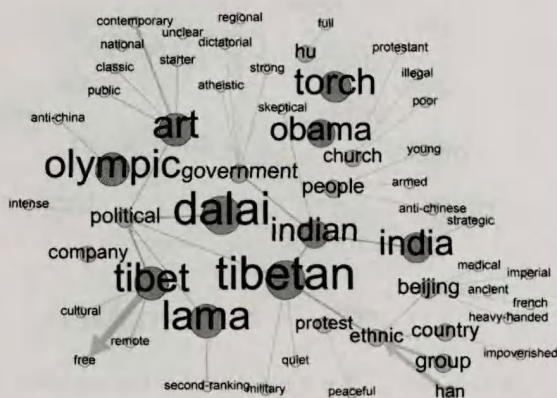


图 7 中国政治形象的字汇表征
《纽约时报》(2003-2012)

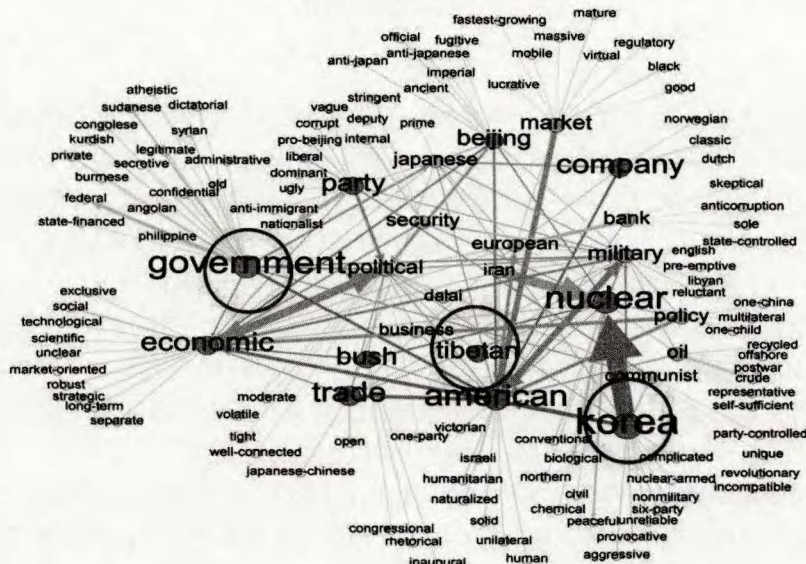
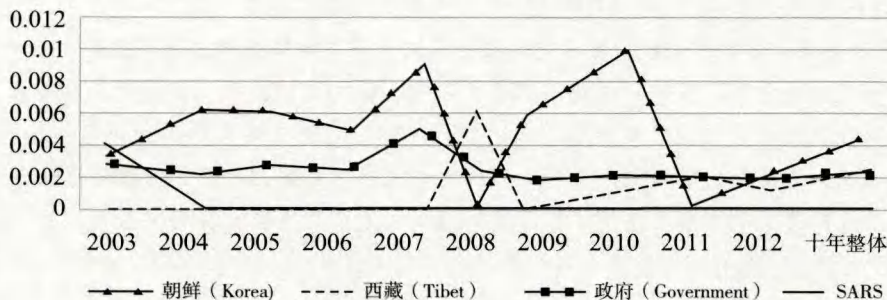


表 1 中国政治报道中“朝鲜”“西藏”和“政府”TF-IDF 值变化
《纽约时报》(2003-2012)



关注，还有一些影响很大，可能成为媒体的持续关注点，成为认知图式中稳定存在的属性插槽。我们的工作显示，挖掘方法能够区别媒体的短时反应与持续关注。具体言之，对于相同主题的报道，可以通过逐年测量结果与长时间跨度测量结果的比较，来确定那些成为媒体形象稳定构成元素的属性。

以中国政治形象为例。在《纽约时报》十年间中国政治主题的报道中，依 TF-IDF 权重排序，“朝鲜”，“政府”和“西藏（的）”三个字汇均位居前十（图 7），故可认为它们是中国政治认知图式中的属性插槽，是构成中国政治形象的重要部分。

然而，若逐年计算，十年间，“朝鲜”在七年的报道中的权重高于“西藏”，“西藏”在六年中并未出现。在 2008 年，“西藏”因奥运火炬传递事件受到特别重视，权重高于“朝鲜”。与这两个字汇比较，“政府”在历年中从未得到过最高权重，但它每年出现，从无例外。在十年整体的权重中，它与其他两字汇不相上下（表 1）。有理由认为，比之于“朝鲜”和“西藏”，“政府”在中国政治报道中是受到更多关注的属性，或，是更为强固稳定的属性插槽。

对 SARS 一词的分析显示出突发事件对中国政治形象的影响方式。逐年报道分析表明，在 2003 年，SARS 明显获得高于“朝鲜”“西藏”和“政治”的权重。之后，无论在逐年分析或十年整体分析中，SARS 均未再现身。由于 SARS 疫情的突发性、威胁性，以及对中国政府处理危机事件能力的挑战性，SARS 受到媒体的异常关注，对已有中国政治形象构成冲击，短暂影响了中国政治形象的表征，然而，SARS 不足以建立起稳定的属性插槽，对中国政治形象并无长久的影响力量（表 1）。

概括言之,通过挖掘获得的数据与 SIT 模型中包含的重要变量有对应关系,可视为媒体形象在文本中映射模式。

结 语

作为理论和技术上的初步探索, SIT 模型和相关技术还存在许多需要改进的方面。例如,如何找到合理简便的算法,整合质化(评价倾向)与量化(共现频率)的形象属性默认值,获得统一的量化结果,如何聚类分析形象构成中的关键词,通过一些关键词和变化来预测其他关键词的变化,等等,都是下一步研究需要回答的问题。不过,我们对《纽约时报》中国主题报道的数据挖掘实验已经显示:

第一,可行性:以 SIT 模型为思路,使用 TF-IDF 等挖掘技术能够提供新闻报道中关涉国家、地区形象的一组重要指标,可以数量化分析地区、国家形象的主要构成元素重要性,对于这些元素的评价,以及变化趋势。SIT 和响应技术具备操作的可行性。

第二,客观性:与文本阅读或内容分析等传统方法相比,挖掘技术极大地控制了研究者、编码员的主观偏差,其刻画的国家、地区媒体形象具备以往方法难以相比的客观性。

第三,全面性:与文本阅读或内容分析等传统方法相比,报道主题分类模型一经建立,挖掘技术便能够对大量的数据,乃至研究总体做出分析。其刻画的国家、地区形象具有传统方法难以相比的数据全面性。

第四,应用性: SIT 模型和与之对应的挖掘技术并非专用于报纸中国家、地区形象的测量和分析。就其原理而论, SIT 试图解决使用挖掘技术从大数据中提取关涉事物形象指标这样一个具有普遍性的问题。通过理论演绎和技术优化,它也应能处理网络信息,提取多种关涉形象数据的能力,为机构公关、企业经营和品牌管理提供有用的工具,具有广泛的应用价值。

[作者感谢上海交通大学电子工程学院吕宝粮教授协调本研究中传播学与计算机科学研究者的合作,感谢上海交通大学研究生冯竹青、曾珍、周婕、邹煜、陆捷、徐文婷、田野、胡莉明、宋姗姗、王啸啸,上海外国语大学研究生尹谜眉、周俊、臧金鑫等人的编码工作;感谢艾瑞咨询集团数据挖掘部经理萧嘉敏,工程师顾轩、熊文文,上海交大研究生赵沁筠、印闯对本文所用数据的可视化处理;感谢香港城市大学梁海博士、四川大学陈侠博士对本研究早期工作所做的贡献。本文为国家自然科学基金项目(11BXW022)阶段性研究成果,并获得上海交通大学文理交叉研究重点项目(10JCZ01, 13JCRZ03, 14JCRZ04)支持]

(责任编辑:周奇)

Mining National and Regional Images from Newspaper Reports

—— A Cognitive and Text Mining Approach

Ge Yan, Zhao Hai, Qin Yulin

Abstract: Focusing on mining national, regional images from newspaper reports, this study explores the theoretical and technical possibilities based on the results of a data mining experiment. The authors show that (1)The cognitive schemata containing property slots and their default values of an object influences attention, memory and judgment of the object, and thus the image construction of the object as well; (2)TF-IDF and other text mining algorithms are able to find and quantify properties and evaluations of nations or regions reported in newspapers, and they are approximately equivalents of the properties and their default values in the cognitive schemata, and can therefore be used to portray, analyze the national and regional images, monitor their variations; (3)A data mining experiment using all China related reports(N=13,259)from New York Times (2003-2012) largely validates the feasibility and objectivity of the cognitive schemata-based image model and technology, and implies a wide range of their application in the field of media image studies.

Key word: national image, regional image, media image, data mining, TF-IDF