信息诊断系统设计思路

曾姿颖; 黄煜; 张引; 宋韵雅; 周琳

Published in:
全球传媒学刊

DOI:
10.16602/j.gmj.20210003

Published: 15/03/2021

Document Version:
Publisher's PDF, also known as Version of record

第 8 卷第 1 期　　全球传媒学刊　　Vol. 8, No. 1: 35-62
2021 年 3 月　　Global Journal of Media Studies　　March 2021

信息诊断系统设计思路：人工核查、广众参与和人工智能的三合一运用

曾姿颖[1]，黄　煜[2]，张　引[3]，宋韵雅[4]，周　琳[5]

摘　要　当前对虚假新闻治理的研究较少出于技术逻辑的思路。真假新闻的界限在实践中通常微妙以及难以辨别，加上人们对真假新闻往往有着不同的见解和解读。本文基于来自香港的实证数据分析以及本地虚假新闻核查平台的初步实践经验，提出“人工核查—广众参与—人工智能”的协同核查模式，以最大限度优化社交平台虚假新闻的治理效果。

关键词　虚假新闻；信息诊断；人工智能；事实核查

DOI　10.16602/j.gmj.20210003

一、引言

社交媒体的出现和迅速发展为虚假信息传播提供了温床，这种现象引起了很多学者的关注。过往文献已有关注虚假信息的生产、传播及接收的研究，但暂时还没有一个完整的理论框架去系统研究读者如何接收与处理虚假信息，更没有一个与理论挂钩的诊断系统去应对虚假信息的传播。对此，本文从实践角度出发，通过对国内外学术文献的回顾与思考，尝试提出一个以理论为基础，并具有实际探索经验的信息诊断系统供学术同仁和事实核查人员参考。

Hansson et al. (2020)最近就针对虚假信息在健康灾难中的传播提出，虚假信息会使人们在六个方面变得脆弱。基于 Hansson 等人的主张，我们总结出虚假信息在一般情况下，可通过以下六种方式对人们造成伤害，导致社会脆弱性(social vulnerability)：(1)让人们形成错误的观念；(2)导致人们错误评估事件对于自身以及社会的风险；(3)鼓励人们使用错误甚至有害的措施应对风险；

1. 曾姿颖：香港浸会大学传播系助理教授。
2. 黄煜：通讯作者；香港浸会大学传理学院教授。
3. 张引：香港浸会大学新闻系助理教授。
4. 宋韵雅：香港浸会大学新闻系副教授。
5. 周琳：香港浸会大学传理学院高级研究助理。

(4)妨碍人们采取适当的防范措施应对风险;(5)诱使人们泄露有关自己的私密信息;以及(6)对信息相关的当事人带来骚扰甚至仇恨言论。

为应对虚假信息带来的伤害,本文将透过回答三个研究问题来说明一个信息诊断系统的设计(图 1):第一,在信息接收的过程中,何种因素会让读者觉得信息可疑?换句话说,什么原因会导致一个人对信息的真实性和作者意图产生疑问?第二,当我们需要诊断一个信息,去验证信息内容的时候,作为一个读者会怎样进行核查工作?第三,作为一个核查机构,又需要怎么去进行核查工作以应对以上两个研究问题带来的挑战?

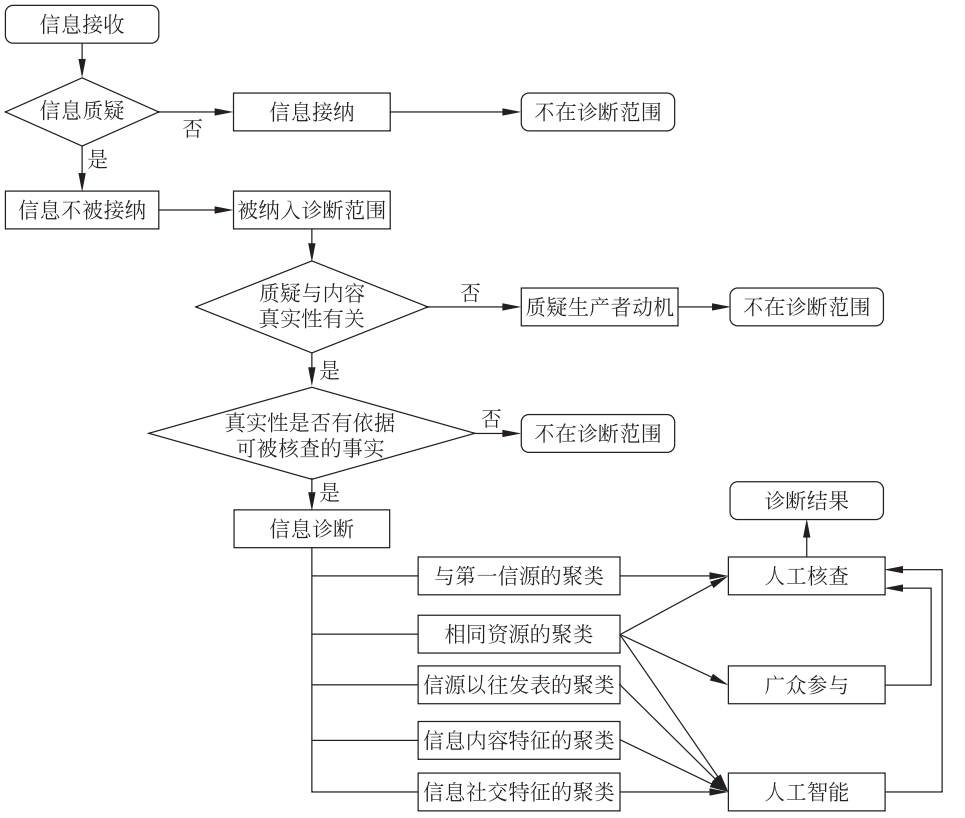

图 1 信息诊断系统设计思路总览

的确,在高度“全球化”和“媒介化”的今天,人们不但需要学会使用敏锐、快捷的方法来识别虚假信息,核查机构更需要开拓新的方法以应对传统核查方法的局限(Shu et al., 2017)。近年来,随着数据挖掘、机器学习、深度学习等技术的快速发展、日益成熟,各国的研究人员也都在试图寻找一系列自动化的方法以辅助人们检测和识别虚假新闻。因此,本文章提出的信息诊断系统也将以结合人工核查、广众参与和人工智能三种应用。

二、信息诊断的需求：读者对虚假信息的吸纳

鉴于现有文献对受众接收虚假信息的研究来看，人们对虚假信息，尤其是虚假新闻，带有错综复杂的看法。美国的 Nielsen & Graves（2017）就发现，人们对虚假新闻有着不一样的见解，从肤浅的、不准确的或能引起轰动的新闻到超党派内容、政客的谎言以及带有政治目的、意图改变舆论的信息，再到带有误导性的广告和赞助内容，都有可能被看作是虚假新闻。可见，人们对虚假信息的定义是多样化的，有着极大的差异，对真假新闻没有绝对的划分。虽然很多学者尝试去定义虚假信息，但普罗大众尚未对虚假信息有一个统一的定义。因此，虚假信息的研究除了研究信息的传播，更需要先了解人们对信息的理解，更重要的是影响他们对虚假信息接纳及不接纳的因素。

（一）读者对虚假信息的吸纳程度

一般而言，读者接触一个信息，有可能产生三种结果：(1)完全接纳信息内容；(2)对信息内容半信半疑；(3)完全不接纳信息内容。完全接纳信息内容的读者，大多不会对信息进行下一步核查，因为他们相信信息内容并接纳信息而忽略了核查的需要。只有对信息内容产生不同程度质疑的读者才会有意识地通过更多渠道去挖掘"真相"，以达到了解内容真实性的目的。这个吸纳过程带来一个重要的研究问题：到底什么因素会导致读者对内容产生质疑。对此，我们提出四个因素，并将其归类为对于内容真实性以及动机性的两大类因素。其中，真实性包括(a)内容客观与事实不符和(b)人们主观认为内容与事实不符，而动机性则包括(c)对内容生产者的动机存疑和(d)人们自身对内容生产者的不信任。

客观虚假对比主观虚假

一般来说，人们对信息内容产生质疑，主要是因为内容与读者对于该事件或事物的认知有所不一致，以至于对信息内容的真实性产生疑问。客观来说，内容有可能是真的与事实不符，也就是说读者的质疑是对的。同时，内容也有可能是准确的，可是内容跟读者自身的认知有差异，导致读者怀疑信息内容的真实性。

前者可以称为是合理的质疑，而后者则带出一个问题——什么因素会导致人们对信息内容产生错误的怀疑，把准确的内容看成是错误或存疑的。判断内容是否虚假的困难在于新闻往往提供给受众最新的消息，而新闻受众基本上很少会事先知道所提供的信息是否真实。换句话说，在接收新的消息时，人们必

须获得更多的知识才能准确判断内容的真实性。矛盾的是,受众很少会掌握有关新消息的知识点去判断真假,所以他们只能依赖自身有的知识基础去判断虚假信息。基于这个逻辑,受众对于内容真实性的判断就很大机会会被自身的观点左右。人们主观认为信息内容与事实不符的原因,很可能是由于信息内容与他们对事情/事物的观点不一样,因此,以下讨论以主观虚假为中心。也只有对新闻内容产生质疑的受众,才会进而去进行后续的信息核查。

动机性推理(motivated reasoning)和敌对媒体效应(hostile media perception)的研究正好能解释这种主观虚假的情况。根据动机性推理的文献,人们倾向于接受与自己观点一致的信息,在处理信息时,会更愿意相信支持自己所持立场的论点。他们会下意识不断寻找支持自己观点的佐证,重复巩固现有的观点,同时也会选择性忽视与自身立场不同的理据,避免既定成见受到挑战。这种系统性、有偏差的推理倾向在心理学为"确认偏误"(confirmation bias)。

另一方面,在传播学,以动机性推理为基础的敌对媒体效应指出,带有不同立场的人,尤其是对立的双方,会对媒体内容有截然不同的认知(Vallone et al., 1985)。换句话说,人们由于自身的立场不同,会对新闻造成不同的解读,以至于在同一篇新闻上看到不同程度甚至不同方向的偏见。因此,事件的支持方会认为媒体信息反对自己的立场。与此同时,事件的反对方也会认为媒体信息反对自己的立场。由于双方都认为信息反对自己的立场,这意味着受众对于同一篇信息的偏见感知带有分歧。

在一项经典的研究中,研究者邀请自我认同的亲阿拉伯参与者和自我认同的亲以色列参与者观看有关1982年贝鲁特大屠杀的新闻播报(Vallone et al., 1985)。虽然新闻播报的内容平衡报道大屠杀的两方,但亲阿拉伯和亲以色列的参与者同时认为该新闻播报不利于他们的观点。Vallone et al.(1985)提出,双方的极端立场导致了对平衡信息的偏见以及双方拒绝接纳,个中原因是报道内容不符合他们强大的既有信念体系。可见,个人所持的立场能影响信息偏见程度的判断。

一直以来,敌对媒体效应的研究大都以感知到的媒体偏见以及新闻可信度为因变量,学者们一直在研究带有不同观点的受众如何得出不同新闻偏见的结论(Gunther & Schmitt, 2004)。最近的研究开始把这种已被娴熟运用的理论推广到有不同立场的受众如何应对新闻虚假程度(Tsang, 2020b)。研究指出,对事件持相反观点的受众,的确会对于同一篇新闻,感知出有明显差异的虚假程度。因此,虚假信息的研究必须厘清一点,人们对于信息的质疑是由于客观信息含错误,还是出自受众自身的观点引起的主观错误。

总括来说,由于每个人都有可能根据自己的观点去感知不同程度的新闻偏

见/虚假程度并进行不同的信息处理，因此学术研究应向前迈进，为媒体受众提供某些客观标准作为指标。学术上需要更多客观性的调查，客观评价媒体的表现，这种调查将有望帮助受众做出更好的信息可信度判断，尤其是对于每个信息源的可信度。可以预料，客观性作为媒体指针，还可以提供媒体生产者一个质量的指标，推动产业通过达到指标要求来提高媒体质量。

（二）对信息生产者的动机质疑

以上的讨论只对信息真实性产生质疑进行了阐述，除了信息的客观错误、信息与自身观点不符，还有的就是人们对内容生产者的不信任。这种不信任也可以分为由自身观点产生的不信任以及普遍对于内容生产的不信任。就主观的不信任来说，Tsang(2020a)的研究表明，读者除了会对内容的真实性产生质疑，阅读信息的时候也会有意识或无意识地判断信息背后的动机。其实，作为一名新闻受众，很难仅通过阅读一则新闻内容去判断信息来源的意图，辨别所提供的信息是出于营利目的还是政治目的。就是说，读者一般难以单纯从一个新闻作品去提取新闻发布者的意图信息。因此，读者一般只能靠阅读内容本身去猜测作者的意图，而这个猜测一般会被读者本身的观点影响，与之前动机性推理的讨论一样。也就是说，受众对于信息生产者的动机极大程度会受他们对于内容的同意程度所牵制。的确，近期的研究发现，让带有不同观点的读者去评价同一则新闻时，读者对于新闻的虚假程度也取决于读者本身的立场(Tsang，2020a)，比如说新闻的动机是保护当权者还是试图影响公众舆论。受众对负责该新闻报道的记者的动机评估不同，可以导致双方对新闻虚假程度的判断亦有所不同。

另外，除了受众自身的观点，人们普遍对于内容生产者或者对于个别内容生产者的信任度也是一个关键因素。比如说，民意调查表明公众对媒体的信任度一般偏低。根据 Gallup 的民意调查，显示只有不到一半的美国人信任大众媒体 (Jones，2018b)。该民意调查还发现，62％的美国成年人认为他们在传统新闻媒体中遇到的新闻是有偏见的，而 80％的美国成年人认为这种情况亦出现在社交媒体中(Jones，2018a)。对主流媒体的这种不信任被称为媒体怀疑论(media skepticism)。这种现象可部分归因于虚假新闻的增多(Boczkowski，2017)。因此，人们很大机会会以消极负面的态度看待媒体发布的信息，但 Boczkowski (2017) 认为，对媒体的普遍质疑带来的并不只有负面影响，也可带来消除虚假信息的动力。换句话说，由于对媒体的不信任，人们会更加积极质疑他们所看到的信息，从而不会尽信虚假信息的一面之词。

相比内容真实性，动机的猜测并没有实际的方法去核查，因为只有内容生产者本人才能掌握自己发布信息的动机。因此，以下有关核查的讨论，不管是

由受众进行的核查，还是机构进行的核查，都只关注内容的真实性。

客观信息诊断弥补主观信息诊断的不足

虽然研究者大都明白不能过于依赖受众对于信息真实性的判断，但现在大多数新闻可信度研究仅集中于对新闻可信度的主观测量（Tsfati & Ariely，2014），而没有太多客观对于媒体信息的评判标准与报告。其实，由PEW和Gallup等民意研究组织进行的大量民意调查也仅关注可信度问题，询问受访者的信任程度以及在哪个方向信任和/或看到新闻报道有偏见，而我们主张需要更多的客观信息诊断去弥补主观信息诊断的不足。换句话说，单凭民意调查数据（研究受众对媒体和媒体信息的可信度）是有缺陷的，因为结果都是受众主观的判断，极大可能受自身观点的左右；同时，客观的判断在市面上非常的欠缺。实际上，事实核查的崛起可以看作是朝这个方向发展的一种措施，也符合我们提出的对信息进行客观的诊断，向新闻受众展示事实是非。

总括来说，客观信息诊断的建立可以成为两方面的指导方针，一来可指导受众用客观标准衡量新闻的质量，二来可以监测以及规范媒体的表现。由于人们大都受自身观点的影响，一个客观的诊断结果可以为他们带来一个标准去评判真假信息。再来，媒体现在的表现都基于人们主观的意见，比如说从问卷得到大众对媒体表现的满意程度等，这些也是依赖人们主观的判断。就算大多数人认为满意的个别媒体，并不代表质量一定就比其他媒体好。也就是说，热度高的或者知名媒体机构，不一定为大众带来更真实更完整的信息。市场上需要有一套判定真假信息的手册，弥补现在的虚假新闻大肆传播的漏洞。

三、信息诊断：信息聚类及途径

对于受众来说，因为往往不能直接与第一信息来源甚至第二信息来源有接触，他们更偏向用信息聚类的方式去进行核查。换句话说，如果人们对信息产生质疑，他们会从网上、身边的人或者其他媒介获取更多的相关信息，尝试对比信息以核对较先前看到的信息内容的真实性。我们将这种寻找相关信息做对比的过程定义为内容聚类，即把属性一致的内容整合起来进行对比分析，查看可疑内容是否存在错误。此外，我们提出一共五方面的属性聚类：(1)相关议题现有资源的聚类；(2)信息与第一信息来源的聚类（第一信息来源指的是与信息内容相关的当事人）；(3)信息来源以往发表的聚类（信息来源的信任度）；(4)信息内容特征聚类；及(5)信息社交特征的聚类。

（一）相关议题现有资源的聚类

在实际情况中，靠近事实来源并非轻而易举之事，因而独立的可靠信息来

源信息之间的聚类核证十分重要。基于人们对于外界的认识都是从其他人身上或者媒体所获得，我们假设对信息有不同程度质疑的受众会去寻找跟信息议题有关的现有资源，比如说可以透过搜索引擎进行搜索、向身边的人进行询问，而这些方法都是透过聚类更多相关的资源，然后与存疑的信息进行对比。以搜索引擎为例，人们可以找到一系列网站提供的相关信息，再而从中找出可解答自己疑问的信息，透过更多的阅读来核对自己刚刚接收的信息内容是否正确。类聚到越多确认信息真实性的资源，该信息的相关议题现有资源聚类结果重合度越高，代表信息越可信。相反，如果网上信息不同意被质疑信息的内容，该信息的相关现有资源的聚类重合度会减低，而越多网上信息提出的与被质疑的信息的内容不同，代表信息越不可信。

严谨来看，类聚结果重合度的大小并没有绝对真实的含义，因为二手信息可以有很多，而人们不一定会核查市面上所有的相关信息来得出这个对于相关议题现有资源的聚类的结果。有些人可能看完一个相关信息就诊断结束，而某些人可能看十个以上的相关信息才会得出对于信息真实性的结论。因此，呼应前面所提的主观虚假和客观虚假，这个相关议题现有资源的聚类结果为透过客观真实来判断信息的真实性。

（二）信息与第一信息来源的聚类

如果要严谨遵循信息来源的可靠、独立性，则必须获取第一手材料以进行核查。虽然大多数人难以获得第一信息来源或者正好是事件当事人，但也不排除的确有人或者机构有这个能力，例如新闻媒体、事实核查机构、商业机构、政府部门等。这些”中介”就未必单靠相关信息重合值来判断信息的真实性，他们还可以直接与当事人核对信息内容进行核查。这种聚类模拟的不只是量，比如说信息内容提到三位当事人，三位当事人都确定内容无误，信息与第一信息来源聚类的重和度可以相当高，也代表相对可信。同时这聚类的结果也会受到信息来源分量的影响，比如说那三个当事人，一个为最主要的，可以和最主要的当事人核对信息内容，信息与第一信息来源的聚类结果会相对只和没那么核心的当事人核对来得高。这就代表，信息内容提到的当事人对真实性判断的影响，也因此对于聚类结果有非常直接的干扰。例如比起一个负责管理案发现场的警员，一个目击整个案件经过的证人更能提高信息与第一信息来源的聚类结果。就是说，核查人员必须先识别与内容有关的关键人物，再去对比口供进行核查，如果信息内容与口供内容一致，代表两者重合，信息真实性也相应提高。

（三）信息来源以往发表的聚类

再者，人们也可以聚类信息来源的发布历史来帮助判断该信息的真实性。

如果某信息生产者以往发表的内容从未有过错误信息，那该信息的信息来源可信值就增加。相反，如果某信息生产者有发表错误信息的记录，该信息的信息来源可信值就会减少。其实，有越来越多的假新闻由机器人写作，其写作和传播速度往往高于人工发布的假新闻，因此及早检测到机器人账号显得尤为必要(Shu et al.，2020)，进行更自动化的虚假信息诊断，比如说利用类似的信息来源特征可以很好地识别异常账号。一般地，信息来源特征包括用户注册信息、用户行为信息、用户可信度等三个维度。

第一，用户注册信息包括用户名、地理位置信息、是否实名验证、有无简介、注册时间、链接等。有研究发现真实账号的注册时间是均匀分布的，而机器人账号会在一周内的某几天集中注册(Gurajala et al.，2016)。同时，通过分析社交账号中使用的装饰图片链接发现，尽管机器人账号采用了各异的链接，但是链接到的图片出现了大面积雷同(Gurajala et al.，2016)。第二，针对用户行为信息，有研究统计了用户的活跃时间，发现真实用户的活跃时间是不均匀的，在一天内有些时间几乎没有活跃者，而虚假账号的活跃程度常常与时间无关、呈现均匀分布(Gurajala et al.，2016)。也有研究将用户在某一时间窗内的活跃次数除以用户平均月活跃次数得到“异常指数”，据此判断是否有可疑行为(Zhao et al.，2014)。最后，对于用户的可信度，不少研究使用用户在传播网络中的属性来评价其可信度，例如好友人数、关注人数、好友/关注比例、好友是否为实名、好友相似度、自中心网络入度、自中心网络出度等等(Castillo et al.，2011；Zhao et al.，2014)。这些特征借助用户在网络中的地位来度量其可信度，应用效果较好。

这样看来，虚假信息发布者总是希望这些信息可以快而广地触达用户(Shu et al.，2020)，并吸引广泛的关注(Sunstein，2014)。鉴于此，我们除了可以观察信息来源，还可以透过研究传播路径以及用户反馈等特征以辅助虚假信息的检测。本节将从“新闻内容特征”与“新闻社交特征”两个维度展开，介绍目前应用较广、效果较好的几类特征及其提取方法，介绍数据采集、整理的基础及难点。

(四)信息内容特征的聚类

对于信息内容本身自带的特征聚类，目前可以大致分为四个类别：从语言和语义特征、写作风格特征、多媒体特征，到基于知识的特征。简单来说，就是先透过聚类真假信息，以学习虚假信息内容的特征，再用个别信息去对比这些以往聚类特征得到的成果以识别虚假信息。

1.语言和语义特征聚类

语言学特征是文本分析的基础，在字词、语句中均有较为成熟的特征。除

了基础的总字词数、单词中字母数量（适用于英文等）等（Shu et al.，2017）基本的统计特征外，一些特殊字符也常被使用（Castillo et al.，2011），比如感叹号、多重感叹号、问号、空集符号（∅）、用户提及（@）、标签（#）、表情字符、emoji、大写单词、粗体单词、引用，等等。同时，表示程度的词、否定词、疑问词、代词、关联词等等也可以作为特征。语文探索与字词计算（Linguistic Inquiry and Word Count，LIWC）作为一种文本与心理学分析的重要工具，在假新闻检测中也多有应用（Castelo et al.，2019；Zhou et al.，2020）。该工具建立了一套系统的词义词典，采用字词计算的方式进行文本特性的分析（黄金兰等，2012）。

目前，简体中文版的词典文心（Textmind）收录了共计 71 个类别、7444 个词汇，涵盖心理特性（如情感历程词、正向情绪词、焦虑词）、个人特性（如视觉词、身体词、家庭词、金钱词）、语言学特性（如后置词、语助词）等范畴。研究者可以将分词后的文本依据 LIWC 目录作分类，并计算得到文本在各类别中的数量与所占比例，从而得到文本在不同维度上的分布情况。类似地，有研究利用“武断用语”“社交分享用语”“呼吁提醒用语”等特定元素实现了较为有效的虚假新闻检测（Jhu-Jyun et al.，2020）。一些语句特征，如平均句子长度（Castillo et al.，2011）、句子的复杂程度（de Marneffe et al.，2006）等也会运用到虚假新闻的检测中。

另外，越来越多的研究者将深度学习运用到模型中，实现特征的自动抽取。这些方法通常将文本用向量表示，称为嵌入（embedding），输入下游模型中进行进一步训练。词袋模型（bag-of-word），n 元语法（n-gram）是表示文本的最基本的方法，但这些方法难以捕获到语义信息。随着 word2vec，GloVe 等预训练词向量的出现，配合循环神经网络（RNN），长短期记忆网络（LSTM）用于文本表示，丰富了嵌入的语义信息。近年来预训练语言模型也逐步发展，BERT、XLNet、RoBERTa 等预训练语言模型也被运用到文本的表示中，进一步提升了虚假新闻检测的效果。

2. 写作风格的特征聚类

相比语言学特征，写作风格特征可以更好地分析评判新闻的意图，即该新闻是否有意误导受众（Zhou & Zafarani，2020）。尽管虚假新闻发布者试图模仿正常新闻的写作方式以取信读者，但在以深度学习为代表的新技术加持下人们还是能够提取区分虚假新闻的特征。除了可以聚类内容的特定元素外，还可以根据信息的修辞结构与叙事特征来做判决。观察比对大量虚假新闻与不实信息可以发现，虚假新闻常常包含某些特定元素，例如在固定位置出现链接、卷标（#）或用户提及（@）等元素。除此以外，修辞结构理论（Rhetorical Structure Theory，RST）能够通过分析文本中不同意义单元的功能关系从而获得文本内容的一致性（Mann & Thompson，1988）。有研究利用

该理论并结合空间向量模型，将文本的修辞结构与叙事特征转化为空间向量（Rubin et al.，2015）。

3. 多媒体特征聚类

需要注意的是，越来越多的虚假新闻藏匿于图片、音频、视频而得以广泛传播，因此提取、处理多媒体特征亦成为关键的趋势。随着卷积神经网络（CNN）、循环神经网络（RNN）、对抗生成网络（GAN）等技术在图像、视频领域的广泛应用，许多研究者也致力于将这些神经网络中的思路或技术迁移至虚假新闻检测的任务中。因此，多媒体特征的聚类也将会是一个学术界以及业界的重点发展方向。

4. 基于知识的聚类

基于知识图谱的聚类通常依赖于外部知识库，知识被定义为一个SPO三元组（主语，谓语，宾语），称为知识图谱。例如，（北京，首都，中国）这个三元组代表了“北京是中国的首都”。给定一个陈述，这种方法通过计算它是否能从知识图谱中找到或从知识图推导出来判断它的真实性。有研究将事实核查视作一个在知识图谱中找到最短路径的问题。他们提出了一种分析路径长度的方法来度量一个陈述的真实性（Ciampaglia et al.，2015）。同样基于计算图的思想，有研究提出了一种新的无监督方法“知识流”（knowledge stream），通过从知识图谱中找到两个实体间的知识流来核查一个陈述（Ciampaglia et al.，2015）。不同于上面计算路径的方法，Pan et al.（2018）利用知识图谱嵌入（knowledge graph embedding）的方法来进行虚假新闻检测。

有些研究者则尝试利用从可信来源，如百科全书，被证实的权威新闻，来获取证据辅助虚假新闻检测。Thorne et al.（2018）就定义了事实抽取和验证的任务：从维基百科抽取和合成证据用于事实核查。而Popat et al.（2018）提出DeClarE，一个从原始新闻中挑选出和陈述相关的显著词作为证据，来辅助事实核查。最后，Ma et al.（2019）通过共现模型和自然语言推断的方法来获取句子级别的证据用于事实核查。这种聚类依赖的是内容里的知识点作为聚类的方式。

（五）信息社交特征的聚类

1. 传播路径特征的聚类

信息社交特征的聚类可以分为传播路径和用户反馈特征两方面。传播路径按照传播节点与边又可以分为同构信息网络（homogeneous information network）和异构信息网络（heterogeneous information network）。同构信息网络中的节点和边属于相同类型。例如，可以将用户作为节点，将用户间的通信

关系作为边构成一个同构信息网络。传播树（propagation tree，也作传播图）就是一类典型的同构信息网络，能够刻画新闻的发布与转发关系（Zhou & Zafarani，2019）。常用特征包括级联深度（cascade depth）、级联尺寸（cascade size）、级联最大宽度（cascade max-breadth）、结构性病毒式传播（structural virality）等（Goel et al.，2016；Vosoughi et al.，2018）。研究发现虚假新闻的级联深度、最大宽度等指标的值较真实信息更大（Goel et al.，2016），同时虚假新闻传播形成的网络密度更高（Zhou & Zafarani，2019）；对比之下，真实信息触达同样多的用户所需时间比虚假新闻更长（Vosoughi et al.，2018）。

除了直接使用这些特征，还可以利用图核函数（graph kernels）方法学习得到捕捉任意两张图之间相似性的核函数，并利用支持向量机等传统核方法完成图分类（Draief et al.，2018）。近期，有研究基于核方法，采用数据驱动的方式由传播树自动生成相关特征（即子传播树）（Ma et al.，2017）。同时，也有研究从用户构成的邻接矩阵中利用非负分解的方法学习得到用户的潜在表示（Shu et al.，2019）。

而异构信息网络中节点和边可分属不同类型。例如，可以将新闻发布者、新闻文本、新闻用户、评论作为节点构成一个异构信息网络。有研究结合用户可信度、新闻文本标签与新闻发布者的倾向等特征得到异构网络的表示（Shu et al.，2019）。

2. 用户反馈特征的聚类

用户反馈通常指用户对一个新闻事件作出的反应，如点赞、评论、转发等，通常会出现于新闻的评论区和社交平台（如微博、推特）上。当人们在遇到谣言时，会更倾向于转发和评论它，在微博平台上收集与新闻事件相关的帖子，可以从这些帖子中学习时序信息和文本信息来用于虚假新闻检测（Ma et al.，2016）。同时，虚假新闻为了吸引眼球，往往会包含一些煽动性观点，促使用户产生激烈的回复，因此可以利用这些回复辅助虚假新闻检测（Ruchansky et al.，2017），但这种方法依赖于用户反馈，对于早期的虚假新闻检测效果不好。Qian et al.（2018）提出了使用条件变分自编码器（CVAE）的方法生成用户的评论，这种方法将用户评论作为监督信号来辅助虚假新闻检测模型的训练，使得虚假新闻在早期也能够得以检测。

（六）信息诊断的挑战与机遇

虽然我们可以从五个聚类值去推断信息的真实性，值得注意的是，信息来源以往发表的聚类、信息内容特征的聚类，以及信息社交特征的聚类都只能当作是参考值，因为它本身没有牵扯对于信息内容真实的核查。和对新闻从业人

员的专业要求一致，信息与第一信息来源的聚类比相关信息的聚类重要。核查需以跟当事人或者询问第一手数据为主，以相关信息重和值为辅。需要注意的是，相关信息的聚类是不能被忽视的，因为相关信息的聚类不但可以帮忙核对当事人所说的内容是否属实，而且在找不到或得不到当事人回复之时，相关信息的聚类可以大派用场，而不至于因找不到第一手数据而放弃整个核查工作。

对于信息来源以往发表的聚类、信息内容特征的聚类，以及信息社交特征的聚类，需要有质量的数据集来应对特征选取、模型的学习。而目前常用的开源数据集的概况如表1所示。可以看出，目前流行的数据集尚缺乏中文内容，直接迁移在这些数据集上得到的虚假新闻检测模型可能并不适用于中文语境，因此这是中文虚假信息诊断所普遍面临的痛点。然而近期，台湾资策会利用台湾事实查核中心积累的中文核查内容与用户提交的信息，联合开发了“不实信息快筛平台”，是中文虚假信息诊断中较为成功的案例(Jhu-Jyun et al.，2020；Shiang-Jiun et al.，2020)。总而言之，人工智能推动更自动化的诊断是大趋势，但针对第一信息来源的聚类以及相关信息的聚类不容忽视。

表1 目前常用的开源数据集梳理

数据集	特征	标签	数据源	内容特征	传播特征
Burfoot Satire News Dataset (Burfoot & Baldwin, 2009)	True, satire		English Gigaword Corpus, Google	✓	
Credbank Dataset (Mitra & Gilbert, 2015)	Topic, event, content, credibility annotation, tweet ID, tweet post time	Credibility	Twitter	✓	✓
Emergent (Ferreira & Vlachos, 2016)	Claim, source, headline, stance	True, false, unverified	事实查核网站，Twitter中的事实查核账号	✓	
BuzzFeed News (Home & Adalı, 2017)	Title, text	Fake, real	Facebook, BuzzSumo	✓	
Benjamin Political News Dataset (Home & Adalı, 2017)	Title, text	Real, fake, satire	多种报刊	✓	
Fake News Challenge Dataset	Headline, body text	Unrelated, discuss, agree, disagree	Fake News Challenge	✓	

续表

数据集	特征	标签	数据源	内容特征	传播特征
LIAR (Wang, 2017)	Statement, speaker, context, justification	True, mostly true, half true, barely true, false, pants on fire	PolitiFact.com	✓	
FakeNewsNet (Shu et al., 2019)	News content (2 categories, 3 features), social context (4 categories, 12 features), spatiotemporal information (2 categories, 4 features)	Fake, real	PolitiFact and GossipCop (for News Content); Twitter (for Social Context and Spatiotemporal Information)	✓	✓

四、香港民众日常新闻接触与主观信息诊断概况

了解人们对新闻的可信度、虚假新闻的特征等问题有何不同看法，对如何建立一个高效的信息诊断系统十分重要。为此，我们分别于 2019、2020 年间进行了两次全香港的问卷调查。调查通过电话访问进行，采用随机电话拨号的方式，邀请 18 岁以上能听说粤语的香港市民参与。两次电话调查的样本数分别为 1211 和 1223 个成功样本。为了提高民调结果的准确性，调查数据在分析前都先以“性别×年龄×教育程度”综合加权法进行加权；而加权的比例依据香港政府统计署相应年份的人口普查结果设定。以下是两年主要调查结果的报告。

调查数据发现，受访者评价一个新闻机构公信力标准的重要程度依次为：“准确报道事件真相”(87.3%受访者认同)、“不偏不倚报道”(84.0%受访者认同)、“关注国际局势与本港社会的联系”(81.7%受访者认同)、“关注公众利益”(78.8%受访者认同)、“信息全面及表达不同观点意见”(76.5%受访者认同)、“监督权贵”(69.4%受访者认同)。

比较不同信息来源的可信度，接近四成(39.5%)受访市民不认同社交媒体上的消息比传统媒体(包括电视和报纸)的信息更准确，约三成市民(34.6%)对此表示困惑。相对而言，人们更倾向怀疑其他用户在社交媒体上所分享的信息——接近一半的受访者(48.5%)表示，不能轻易断定在互联网上获得信息是否可信。更有约两成受访者认为“大多数网上消息都是不可信的”。而传统新闻媒体间比较，电子媒体与纸媒在信息准确度上的差异则并不明显。这一组数

据反映，面对网络信息泛滥，公众仍然倾向依赖传媒机构的专业把关。因为当媒体信息被证实为虚假时，传统媒体相对于社交媒体，更容易问责。然而，正如上文提到，市民大众很少会事先知道新闻事件所涉及的信息是否真实，他们需要获取更多的信息才能判断内容的准确性。从受众的角度来看，只有对新闻内容有质疑的受众，才会进而决定去进行后续的信息核查。其实，媒体怀疑论在香港社会一直存在；而且在新媒体环境下，媒体信息的可信度总体呈下降趋势。

市民对媒体可信度的评分只反映他们对信息真实性的印象判断，并未体现其他因素而导致的信息筛选（例如政治立场、市场定位）所带来的影响。其实，市民在接收传媒信息时，能意识到媒介信息其实是一系列信息筛选过程的产物。有超过一半（55.7%）的受访者表示，在看新闻的时候，他/她总会关心报道隐藏了什么内容。而对新闻议题的喜好，则影响了受众主动寻找数据、核查信息的可能性。有七成半的受访者在看到自己关心的新闻时，会尝试利用其他途径搜索更多补充数据。有超过四成（42.3%）的受访者“有时会用好长时间思考一则新闻”。但值得注意的是，这种自觉意识并非体现在全面的新闻接触当中。在当下信息过剩、注意力缺乏的时代，市民大众对新闻时事的关注并不全面，接近一半（48.9%）的受访者通常只会关注自己感兴趣的新闻。有接近两成（19.2%）受访者表示，大部分的新闻他们看完以后就会不记得。

对普罗大众而言，虚假新闻往往没有清晰、确切的定义，只能依靠一定的语意特征以及个人经验进行判断。受访者大多认同“在未交代事实全部的情况下就做出结论”（62.5%）以及“标题与内文不符”（62.3%）两种报道手法属于虚假新闻。而“用词夸张”（44.2%）以及“包含过去或已知信息”（24.4%）两种报道手法是否属于虚假新闻则有待商榷。而市民普遍同意“为了误导公众而捏造新闻”（90.8%）、“未厘清事实之前就发布重大新闻”（89.4%）、“发布捏造视频或照片”（88.1%）、“倾向某一政治立场的事实”（79.4%）等新闻事例会让香港市民对社会现实的理解造成困扰。

个人新闻阅读行为及主观诊断偏向，往往受教育程度影响（表2）。卡方分析（ANOVA）结果显示，教育程度越高的受访者，越经常“通过多个信息来源核查新闻真实性”，“与他人讨论新闻”，“寻找新闻报道没有包括的事实”。而教育程度越低的受访者则越经常“只浏览新闻标题”。由此可见，不同人群对于潜在虚假信息的态度以及愿意为核查事实而付出的努力都存在显著差异。因此，如何有效发动群众力量、利用社会资源进行事实核查，须有针对性。教育程度的高低、专业知识的多寡在很大程度上制约了市民在日常生活中处理虚假信息的主观能动性、核查信息细节的能力，以及寻求他人协助的可能性。

表 2　不同教育程度受访者个人新闻阅读行为及主观诊断偏向之比较

		平均值	F 值(自由度)，p 值
通过多个信息来源核查新闻真实性	小学或以下	2.49	F＝67.79 (2)，p＜0.001
	中学程度	3.27	
	大学或以上	3.74	
与他人讨论新闻	小学或以下	2.46	F＝49.92 (2)，p＜0.001
	中学程度	3.20	
	大学或以上	3.50	
利用自己的常识判断新闻正确性	小学或以下	3.40	F＝35.90 (2)，p＜0.001
	中学程度	3.89	
	大学或以上	4.16	
寻找新闻报道没有包括的事实	小学或以下	2.04	F＝69.24 (2)，p＜0.001
	中学程度	2.77	
	大学或以上	3.24	
只浏览新闻标题	小学或以下	2.88	F＝6.32 (2)，p＜0.01
	中学程度	2.70	
	大学或以上	2.52	
看到假新闻的时候选择置之不理	小学或以下	3.64	F＝5.12 (2)，p＜0.01
	中学程度	3.61	
	大学或以上	3.35	

注：研究采用 5 点李克特量表测量上述行为的频繁程度，1 表示基本不会，5 表示经常会。表中数值为不同教育水平组别的平均数。

五、讨论：客观信息诊断的需求

仅依赖于信息可信度的主观理解的研究虽然有助于理解个人如何形成虚假信息的认知，但不足以解决虚假信息在数字信息千变万化的环境中的持续变化。困境是，尽管新闻媒体/信息素养包括教育受众准确评估信息真实性的知识和能力，但主观意识强的人将有可能继续全然接纳虚假信息。根据有关敌对媒体效应的文献(Vallone et al.，1985)，持相反立场的人们会感知一则完全相同的新闻内容有相反方向的媒体偏见(Gunther & Schmitt，2004)以及显著不同的虚假程度(Tsang，2020a，b)。总之，人们倾向于看到媒体对个人立场怀有敌意，从而容易依赖个人观点感知媒体信息真实性。因此，除了市面上林林总总的媒体信息可信度的民意调查，本文提出需要投放更多的资源来创建一个较客观的信息诊断系统，以供受众用作参考点评判信息的真实性。简而言之，一个以聚类为主的信息诊断设计，可以帮助受众区分真实和错误的信息。

其实，对于客观信息诊断的需求，有一个重大问题是，谁可以决定哪些新闻

是虚假的，哪些又是真实的？在事件和相关人物都被大众贴满标签的时代，谁可以被信任去执行一个如此重要的任务（Berezow，2017）？就算是第三方事实核查的人员，也有可能在特定问题上有特定偏见或议程。这些问题让我们思考，到底什么时候可以信任事实核查机构，以及什么因素可以提高大众对事实核查机构的可信度。对此我们提出两个条件：一、独立性；二、透明度。

这就说到为什么大部分事实核查机构都反复强调其独立性，因为这是事实核查机构取得公众信任的基础。核查机构必须得在公众面前保证客观性，没有与党派有瓜葛，更没有接受会影响其客观性的金钱来往。如果事实核查机构不能确保其独立性，即不受政党或其他利益关系影响，那机构将不能避免跟现有的传统媒体一样，被贴上各样的标签（史安斌，2020）。被贴上标签可能带来两个严重的后果。第一，这会引起“选择性接触”（selective exposure），只有认同其核查机构的标签的人才会订阅该核查账号，从而流失很多的受众；第二，因为大众对核查机构的信任度低，核查结果将不会被大家认真对待，由于没有了公信力，核查也将失去效力。因此，机构需要把关其独立性，一旦失手会直接影响受众对于核查内容的认同。

另外，核查机构亦需要把关其透明度，即需要把核查过程公开透明地呈现给大众，受众甚至可以跟着核查员的程序自己核查从而得出一样的核查结果。高程度的透明度不但可以让大众了解核查的程序，更可以自己靠阅读每个核查步骤去判断信息的真实性。这样一来，比直接告诉受众结果更有公信力。正因为如此，每个地区都需要有能提供独立且透明的核查服务。

六、客观信息诊断的实现：建立”三合一”事实核查模式

对于香港人来说，除了本地的核查机构，最具影响力的核查机构为国际事实核查网络（IFCN）。主要原因是Facebook是本地最多人用的社交媒体，而该机构是Facebook的合作伙伴。一旦事实核查机构能跟着IFCN提出的标准进行核查工作，而又能通过他们的考核，并签订承诺书会遵守IFCN所定的工作守则以后，就能对外宣称机构为IFCN认证的机构。IFCN除了对机构的营利方式，对核查过程和方法的透明度以及资金和组织的透明度都有要求。一旦被IFCN认证，就可以成为Facebook的事实核查伙伴，核查结果将可以直接影响Facebook平台的信息传播。比如说，Facebook将会减少被评为虚假的信息的曝光量，代表看到该信息的人会减少。又比如说，Facebook会拦截掉多次发布虚假信息的账号。较积极的合作方式还有，信息的下方会出现相关的核查内容的链接，Facebook用户可以透过链接看到核查机构发布的相关报告，查阅虚假信息的核查信息。

现在对于信息的事实核查存在几个痛点。第一是核查很费时，核查结果验证完成时，已经错过了新闻热度的黄金时期，很难及时对虚假新闻进行有效干预。第二是越来越多研究指出事实核查有可能会有反效果（back-firing effect），所以这也响应到刚刚所说机构需保持独立性和透明度的重要性。针对香港的本地市场，香港浸会大学传理学院最近就推出了事实核查服务，提供透明、独立的核查报告供大众参考。利用"人工核查—广众参与—人工智能"三合一的协同治理模式，提升社交平台虚假新闻的治理效果。

客观信息诊断的实现案例：香港浸会大学事实查核中心

香港面临着社会运动与新冠疫情的双重打击，虚假新闻的肆虐持续地冲击着媒体平台和社会稳定。自 2019 年社会运动席卷以来，虚假新闻的煽风点火助长了不同政见的两极分化，社会稳定受到极大冲击，彭博新闻于 2019 年 11 月指出，香港虚假新闻和政治宣传泛滥，两极分化的言论助长暴力和不信任感，令香港社会的分化已经到了难以和解的地步。而紧随其后的新冠疫情，更为虚假新闻提供了乘虚而入的机会，从健康领域的虚假新闻，例如抗疫消息、新冠疫苗科普，上升到政治层面的煽动与混淆是非，都令香港社会陷入前所未有的虚假信息漩涡中难以脱身。

关于虚假信息的传播研究与发展应对方法是香港浸会大学传理学院近一年的学术重点，学院自始至终坚持在新闻真假难辨的泥淖中高举"唯善为真"的明灯。然而停留在学术理论层面是远远不够的，为了真正让学术理论带来实践意义，学院着手创办香港浸会大学事实查核中心。该中心不仅是香港第一个由大学独立创办并运营的事实查核中心，也打开了业界在事实核查领域互相交流的窗口，除了产出高质量、专业的事实核查结果，也积极与同业者创办工作坊和论坛，供业界人士和新闻学生提供了更广阔的交流平台，促进事实核查的发展。

香港浸会大学事实查核中心的事实核查工作围绕三部曲展开：人工核查、广众参与、人工智能，即以虚假信息传播的学术研究为基础，从提供专业高效的人工事实核查内容出发，利用广众参与模式的力量汇聚专业力量，并借助人工智能建立不实信息自动化检测机制，从而开辟香港事实核查的快车道，提高事实核查效率和公众的媒体素养，推动新闻业界与教育同行的协作。

（一）人工核查

人工核查是香港浸会大学事实查核中心的核心内容驱动，包括提取可疑信息和事实核查两大流程。提取可疑信息是人工核查的基础，香港浸会大学事实查核中心团队根据当下香港的热门话题及重要事件挑选关键词，定时借用脸书（Facebook）公开内容洞察工具 Crowdtangle 来收集信息。通过 Crowdtangle，

中心团队根据挑选出的关键词提取表现过热的帖文，随后团队事实核查专员逐一检查并筛选出可疑信息，再交由编辑和专业顾问分别进行二次筛选与确认。

中心筛选的过程中，对所谓的“可核查的可疑信息”有一套既定守则。首先，中心成员会先判定该内容是否可被核查，例如观点、评论等不在核查范畴之内，只有“事实”陈述可以被核查，与先前所说，只对事实真实性进行核查；另外，内容是否存在误导性或错误信息也十分重要；第三则要考虑该核查是否符合公共利益，如果内容是一些无关紧要的信息，则不必费心力进行核查；第四要看内容是否广泛流传，团队中心成员会根据帖文的热门程度、赞好/分享/评论次数来判断是否已经广泛流传；最后直接影响内容是否能够被推进到核查流程的，是考虑能否找到第一手数据，如果团队中心的资源可以找到第一手资料，例如访问到涉事人员、公司或取得特定领域的专业解答，那么该内容就具备“可被核查性”。有必要时，也会考虑采用第二手数据，不过须确保有多于一个第二手数据，而且数据源需要是有公信力的。

接着，具备可核查性的内容会分配给通过训练的事实核查专员，根据内容所涉及的领域，核查专员会对内容进行事实检验，包括但不仅包括翻阅官方数据、查询内容发布来源、访问专家、借助校友网络获得特定领域里专业人士的解答、以及到内容所涉及的地方进行现场调查，以保证核查内容的直接和准确。相等于，中心对信息与第一信息来源以及相关议题现有资源两方面进行类聚对比的动作。事实核查专员根据收集到的证据得出的类聚结果撰写核查报告，之后再交由编辑和专业顾问进行两层的审阅。审阅工作完成后，报告即可被发布。

报告发出后，用户可以通过以上渠道接收最新的事实核查内容。用户可透过发送邮件、在社交媒体账号留言等方式反馈他们对事实核查报告的见解与看法，如有异议亦可实时联络中心团队分享证据，中心团队的工作人员收到后会做进一步分析和核查，如对异议内容有所订正，便会发布订正消息到各相关平台，供用户做进一步跟进。

（二）广众参与

虚假信息既是一个传播议题，也是一个社会问题。依靠专业团队对不同媒体发布的信息进行事实核查准确可靠，但成本很高，需时较长，难以覆盖不同的社会议题。众人拾柴火焰高，如果社会上能建立一个有认受性的事实核查群体，并不断壮大，就可以逐步推广以“广众参与”为基础的事实核查机制，从而更好弥补仅仅依赖专业团队的缺点。为了逐步向这一个目标发展，香港浸会大学事实查核中心亦开发了手机应用程序 BU FactCheck，在 Andriod 和 IOS 商城供市民大众免费下载。应用程序的开发是为了建立一个可让公众参与的事实

核查互动平台，借助广众参与的模式汇聚更多社会资源与业界力量，提高事实核查的效率和议题覆盖范围，运用到事实核查的内容收集与核查过程中，有效地缩短了可疑信息的提交效率和证据收集周期。

移动终端是当下市民大众接触信息的最常用渠道。因此，手机应用程序的开发能够拓宽事实核查的接触面，为公众及专业用户提供一个与事实查核中心团队双向沟通的高效渠道。首先，公众用户可以透过该应用程序收取中心团队的最新判定推送，随时透过移动终端阅读详细的事实核查报告。此外，中心邀请具有新闻工作经验以及不同领域的专家（如公共卫生、金融）成为注册用户。除了公众用户享有的已核查信息浏览功能，注册用户更能够就一些待核查内容参与核查过程。注册用户可以凭借自身的从业经验和掌握的相关证据，对特定待核查内容进行评核。评核方式包括判定投票及证据上传两个步骤：为了与中心的判定分类保持一致，用户可选的判定选项包括“真实”“部分真实”“错误”或者“无法确定”；如注册用户掌握相关证据或希望提供专业的分析，亦可透过界面上传提交。这个过程进行了相关议题现有资源的聚类对比。香港浸会大学事实查核中心团队于后台接收到各方专家用户及时提交的相关证据后将会做进一步核实、验证，加快报告撰写的进程。如专家或知情人士对发布的已核查结果有异议，也可以及时报告中心团队做进一步讨论与核查。最后，用户如在工作生活中遇到可疑内容，欲委托中心团队进行核查的，也可通过流动应用程序及时上传可疑内容，中心团队接收后会进行可核查性判断，如具备可核查性，便会将内容推进到核查流程，以此拓宽可疑内容的收集覆盖面。

尽管广众参与不能完全取代程序更为严谨的人工核查，但它能够针对人工事实核查需时长、反应时间慢的痛点，让公众在更早的时间点开始关注可疑信息。而具有专业注册用户的参与，则拓宽了中心事实核查的广度和深度。尽管用户对可疑内容的评核投票目前只能被视为一个参考信息，但这一信息的准确性和可信度会随着更多人的加入和参与逐渐增加。因为当参与人数达到一个具有认受性的规模，广众参与的评核结果的可靠程度就会越来越高。

（三）人工智能

在虚假新闻泛滥的现状面前，单靠人工的力量进行内容筛选和核查工作略显单薄，香港浸会大学事实查核中心团队现正重点探索人工智能在事实核查领域的运用，建立一套自动化信息筛选和检测的系统，利用自动化、智能化加快事实核查的速度。

在目前的人工事实核查工作中，中心团队面临着以下几个方面的挑战：1）人工筛选可疑信息十分耗时：尽管通过训练的事实核查专员具备对热点话题可疑性的敏锐度，但相比数量万千的话题讯息来讲，难免在筛选可疑信息的工

作里力不从心；2)信息大量重复出现：当一个热点话题或事件成为民众探讨的焦点时，必定会有大量媒体进行报道，而报道的内容时常是重复的；3)核查过程严重依赖核查专员个人的经验：尽管在核查培训中，中心团队已经熟知虚假信息的常见形态，但核查专员各不相同的个人经验，会造成对同一信息的不同理解，主观因素相对难以控制。因此，中心需要一个客观、高效、能对信息进行预处理的人工智能方案来优化这个信息诊断流程。

短期计划来讲，中心团队正将研发一个可以根据多重标准对新闻数据进行排序的平台，这个平台通过信息可视化面板，对相同时间的新闻、帖子进行聚类。信息可视化面板是一个帮助用户监测、获取数据库中经过预处理的信息的接口。团队可以读取关键词和重要帖文等信息。这里“重要性”的衡量标准包括留言数量、点赞数量以及热度等。热度在这里是一项综合性指标，将信息的内容特征、传播渠道特征以及信息源特征等均考虑在内，可以灵活地选用不同标准筛选信息。同时，核查专员可以方便地查阅历史核查信息与相关关键词，这些信息可以提供多方位的线索，从而辅助专员决定进行怎样的核查。

面板包含三层，即主题层、关键词层、事件层。主题层将关键词归纳为固定的几个主题，如“社会”“医疗”“教育”。每个主题下的关键词是按照不同时间长度自动汇总生成的；而每个关键词又涵盖了若干相关的事件，这些事件将被展示在网络图部分。作为可视化面板的重要部分，网络图用于展示信息的流动过程。与同一个关键词相关的若干事件将以时间顺序按信息传播网络的形式呈现。网络中的节点表示不同的实体(如专业媒体机构、普通用户、自媒体人)。网络图与整个可视化面板能够帮助核查员实时监控与香港有关的海量媒体信息，洞悉其源头、内容与传播轨迹。

更长期的计划来看，中心团队旨在建立不实信息自动化监测机制，以利用计算方法来筛选优先核查内容，开发自动化的虚假信息监测器，提取虚假信息特征训练机器模型，真正实现以人工智能赋能核查为基础的自动化转型。这个转型主要针对信息来源以往发表的聚类、信息内容特征的聚类，以及信息社交特征的聚类。中心正在构建香港事实核查数据库，收集中文语境下(以广东话为主)社交媒体中的虚假信息来应对特征选取、模型的学习。

(四)信息诊断系统的实现——目前成果与实践

自中心运作以来，团队的报告产出维持在平均每周至少一篇，话题涵盖政治、经济、健康、时事等多个与民众息息相关的领域(图2)。截止到2021年2月12日，中心发表了一共30篇事实核查报告，其中判定为“错误”的有23篇，占报告比例的77%；判定为“无法确定”的有1篇；“真实”与“部分真实”各有3篇。这个发表比例说明中心团队在人工筛选可疑信息方面的机制相对高效，对错误

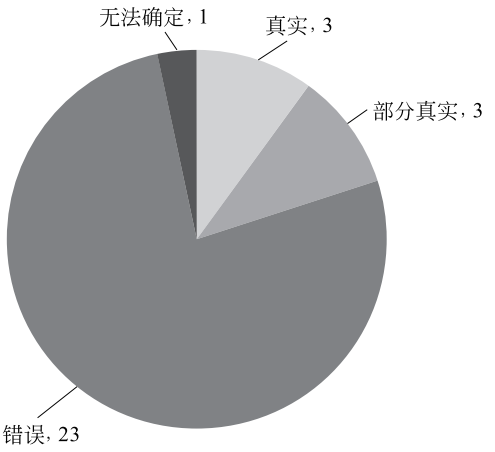

图 2　香港浸会大学事实查核中心现有报告诊断结果分布比例

信息的筛查水平较为专业，但仍有一定的提高空间；日后加上人工智能的辅助，对可疑信息的判断效率将会进一步提高。

中心发布的核查报告分别归类为时事、政治、经济与健康，从目前发布结果来看，时事占据的比例最高，一共有 14 篇（46.7%）；紧接着是健康类 9 篇；政治类 6 篇；经济类 1 篇（图 3）。对时事信息的较高核查率指明，虚假信息在时事信息中出现的频率有可能较高，此类虚假信息借由受众的较高关注度蒙混是非以达到某种传播目的，并且由于缺乏深度较容易成为过眼烟云而不被深究，因此，重点发展人工智能核查以提高核查的时效性，及时纠正虚假时事信息，对维护信息传播的真实性具有深远意义。

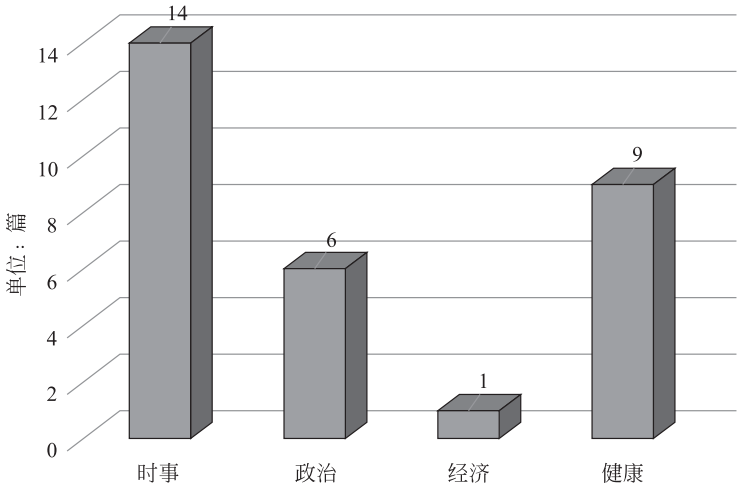

图 3　香港浸会大学事实查核中心现有报告分类

七、结论

随着虚假信息的崛起，大众必须有能力从海量的信息中识别虚假信息和可信来源，尤其是当虚假信息可以为社会带来社会脆弱性问题的时候（Hansson et al.，2020）。通过对国内外学术文献的回顾与思考，本文提出一个理论框架系统研究受众对于信息的接收与处理，点出依赖大众主观识别虚假信息的缺陷。虽然传播学学者早已认识到，对于一个同样的媒体信息，立场不同的读者会倾向有相反的解读（Vallone et al.，1985；Gunther & Schmitt，2004），从而得出显著不同的信息虚假程度，视信息与自己的观点持敌对态度（Tsang，2020b），但这种主观接收信息的因素并未在虚假新闻的情境下被广泛研究。因此本文作者认为，更多的学术研究应该关注（1）人们对虚假信息的定义与见解、（2）人们怎么处理他们自认为的虚假信息，包括动机性推理和敌对媒体效应在虚假信息接收的应用；及（3）客观信息诊断结果（核查报告）能否减少虚假信息对受众和社会带来的伤害，包括先前提到的社会脆弱性问题（例如，改正错误的观念，帮助人们正确评估事件对于自身以及社会的风险，减少对当事人带来的骚扰等）。

由于每个人都有倾向根据自己的立场去感知不同程度的信息虚假程度并进行不同的信息处理，学术研究应向前迈进，提供一套客观的信息诊断供大众参考。这种对于媒体和媒体信息的客观评价，不但可以帮助受众做出更好的信息可信度判断，其核查客观性将有助于提高记者和编辑对传统新闻标准的认识，连带自媒体对信息质量的要求。尤其是，他们得悉有核查机构会对他们发布的新闻报道进行核查工作时，他们将承受更大的压力以鞭策发布的信息质量，尤其是信息真实性。而且更重要的是，事实核查报告及媒体素养的推广可进一步使大众有能力判断信息的真实性，从而做出更明智而负责任的决定。以这个原则为基础，并结合先前提出的理论框架，本文作者尝试建构一个信息诊断系统以检测和识别虚假新闻，用香港浸会大学事实查核中心为案例，说明如何实现人工核查、广众参与、人工智能的三合一运用，辅助人们客观诊断信息。对于核查机构的要求，所谓的客观必须做到独立与透明，要不然核查结果将失去公信力和效力。

根据理论框架，不论是大众或者核查机构，都无法从信息内容直接针对虚假信息的动机去进行核查，因此核查主要围绕内容真实性。另外，利用信息聚类，我们提出五方面的聚类方法进行核查工作：（1）相关议题现有资源的聚类；（2）信息与第一信息来源的聚类；（3）信息来源以往发表的聚类；（4）信息内容特征聚类；及（5）信息社交特征的聚类。根据中心的第一手经验，人工核查通常利

用相关议题现有资源以及信息与第一信息来源的聚类；广众参与一般利用相关议题现有资源的聚类；而人工智能方面，除了与第一信息来源以外的聚类，都可以有所应用。学术界和业界人士应该继续发展人工智能在事实核查上的赋能，尤其是针对中文虚假信息的数据集，不单单是语言/语义特征、知识特征、信息传播特征、用户反馈特征等应用。

通过一套完整的信息诊断系统，大众和核查专员必须意识到可疑信息与第一信息来源的聚类的重要性。核查需以跟当事人或者询问第一手数据为主，以其他四方面的类聚为辅，因此广众参与和人工智能并不能取代人工核查，只能赋能核查工作。最后，至关重要的是，核查能够引起公众对虚假信息以及事实核查的讨论，即什么才算专业新闻，什么算虚假信息，谁有资格对信息进行评核。换句话说，对于虚假信息和核查相关的伦理问题，值得被更多的学者和利益相关者关注。这些讨论和研究都将有助于把媒体素养课程带进小区，真正做到服务社会大众的功效。从这个意义上来讲，新闻及信息素养不仅涉及具有批判性地接收信息的技能和了解自身对于信息认知处理存在的偏见，更包括认识与反思媒体信息、个人与社会之间的关系。

参考文献

黄金兰、林以正、谢亦泰、程威铨(2012)：中文版“语文探索与字词计算”词典之建立，《中华心理学刊》，第 54 卷第 2 期，185-201 页。

史安斌(2020 年 4 月 24 日)：史安斌：疫情后真相、后权威和后情感，《环球时报》，获取自 https://opinion.huanqiu.com/article/3xxbigEoaAS。

Berezow, A. (2017). *Should we ban fake health news?* American Council on Science and Health. Retrieved from https://www.acsh.org/news/2017/10/17/should-we-ban-fake-health-news-11975.

Boczkowski, P. (2017). *Fake news and the future of journalism*. NiemanLab. Retrieved from https://www.niemanlab.org/2016/12/fake-news-and-the-future-of-journalism/.

Burfoot, C. & Baldwin, T. (2009). Automaticsatire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* (pp. 161-164). Suntec, Singapore: Association for Computational Linguistics.

Castelo, S., Almeida, T., Elghafari, A., Santos, A., Pham, K., Nakamura, E. & Freire, J. (2019). A topic-agnostic approach for identifying fake news pages. In *Companion Proceedings of the 2019 World Wide Web Conference* (pp. 975-980). San Francisco USA: ACM. doi: 10.1145/3308560.3316739.

Castillo, C., Mendoza, M. & Poblete, B. (2011). Information credibility on

twitter. In *Proceedings of the 20th International Conference on World Wide Web* (pp. 675-684). Hyderabad, India: ACM. doi: 10.1145/1963405.1963500.

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F. & Flammini, A. (2015). Computational fact checking from knowledge networks. *PLoS One*, *10*(10), e0128193. doi: 10.1371/journal.pone.0128193.

de Marneffe, M. C., MacCartney, B. & Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa, Italy: European Language Resources Association.

Draief, M., Kutzkov, K., Scaman, K. & Vojnovic, M. (2018). KONG: Kernels for ordered-neighborhood graphs. arXiv preprint arXiv:1805.10014.

Ferreira, W. & Vlachos, A. (2016). Emergent: A novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1163-1168). San Diego, California: Association for Computational Linguistics. doi: 10.18653/v1/N16-1138.

Goel, S., Anderson, A., Hofman, J. & Watts, D. J. (2016). The structural virality of online diffusion. *Management Science*, *62*(1), 180-196. doi: 10.1287/mnsc.2015.2158.

Gunther, A. C. & Schmitt, K. (2004). Mapping boundaries of the hostile media effect. *Journal of Communication*, *54*(1), 55-70. doi: 10.1111/j.1460-2466.2004.tb02613.x.

Gurajala, S., White, J. S., Hudson, B., Voter, B. R. & Matthews, J. N. (2016). Profile characteristics of fake Twitter accounts. *Big Data & Society*, *3*(2), 1-13. doi: 10.1177/2053951716674236.

Hansson, S., Orru, K., Siibak, A., Bäck, A., Krüger, M., Gabel, F. & Morsut, C. (2020). Communication-related vulnerability to disasters: A heuristic framework. *International Journal of Disaster Risk Reduction*, *51*, 101931. doi: 10.1016/j.ijdrr.2020.101931.

Horne, B. D. & Adalı, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. arXiv preprint arXiv:1703.09398.

Jhu-Jyun, H., Yen-Heng, T., Zi-Ying, C. & You-Chuan, Y. (2020). Using RoBERTa and linguistic features to detect fake news. In *The 10th International Conference on Frontier Computing* (*FC2020*) (p. 444).

Jones, J. M. (2018a). *Americans: Much misinformation, bias, inaccuracy in news.*

Gallup. Retrieved from https://news.gallup.com/opinion/gallup/235796/americans-misinformation-bias-inaccuracy-news.aspx.

Jones, J. M. (2018b). *U. S. media trust continues to recover from 2016 low*. Gallup. Retrieved from https://news.gallup.com/poll/243665/media-trust-continues-recover-2016-low.aspx.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K. F. & Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (pp. 3818-3824). New York: AAAI Press.

Ma, J., Gao, W. & Wong, K. F. (2017). Detect rumors in Microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*) (pp. 708-717). Vancouver, Canada: Association for Computational Linguistics. doi: 10.18653/v1/P17-1066.

Ma, J., Gao, W., Joty, S. & Wong, K. F. (2019). Sentence-level evidence embedding for claim verification with hierarchical attention networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2561-2571). Florence, Italy: Association for Computational Linguistics.

Mann, W. C. & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, *8*(3), 243-281. doi: 10.1515/text.1.1988.8.3.243.

Mitra, T. & Gilbert, E. (2015). CREDBANK: Alarge-scale social media corpus with associated credibility annotations. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media* (pp. 258-267). Palo Alto, California: AAAI Press.

Nielsen, R. K. & Graves, L. (2017). "*News you don't believe*": *Audience perspectives on fake news*. Reuters Institute for the Study of Journalism Report. Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2017-10/Nielsen&Graves_factsheet_1710v3_FINAL_download.pdf.

Pan, J. Z., Pavlova, S., Li, C. X., Li, N. X., Li, Y. M. & Liu, J. S. (2018). Content based fake news detection using knowledge graphs. In *International Semantic Web Conference* (pp. 669-683). Monterey, CA, USA: Springer.

Popat, K., Mukherjee, S., Yates, A. & Weikum, G. (2018). DeClarE: Debunking fake news and false claims using evidence-aware deep learning. arXiv preprint arXiv:1809.06416.

Qian, F., Gong, C. Y., Sharma, K. & Liu, Y. (2018). Neural user response generator: Fake news detection with collective user intelligence. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* (pp. 3834-3840). Stockholm, Sweden.

Rubin, V. L., Conroy, N. J. & Chen, Y. M. (2015). Towards news verification: Deception detection methods for news discourse. In *Proceedings of the Hawaii International Conference on System Sciences*.

Ruchansky, N., Seo, S. & Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. In *Proceedings of the* 2017 *ACM on Conference on Information and Knowledge Management* (pp. 797-806). Singapore: ACM.

Shiang-Jiun, C., Yi-Wei, M., Cheng-Mou, Chin-Shen, F. & Wei-Liang, W. (2020). Fake news detection on social media based on the propagation behavior, Taiwan Academic Network Conference (Tanent 2020), P1156.

Shu, K., Sliva, A., Wang, S. H., Tang, J. L. & Liu, H. (2017). Fake news detection on social media: A data mining perspective. arXiv preprint arXiv: 1708.01967.

Shu, K., Mahudeswaran, D., Wang, S. H., Lee, D. & Liu, H. (2019a). FakeNewsNet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media. arXiv preprint ArXiv:1809.01286.

Shu, K., Wang, S. H. & Liu, H. (2019b). Beyond news contents: The role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (pp. 312-320). Melbourne VIC, Australia: ACM. doi: 10.1145/3289600.3290994.

Shu, K., Bhattacharjee, A., Alatawi, F., Nazer, T., Ding, K., Karami, M. & Liu, H. (2020). Combating disinformation in a social media age. arXiv preprint arXiv:2007.07388.

Sunstein, C. R. (2014). *On rumors: How falsehoods spread, why we believe them, and what can be done*. Princeton: Princeton University Press.

Thorne, J., Vlachos, A., Christodoulopoulos, C. & Mittal, A. (2018). FEVER: A large-scale dataset for Fact Extraction and VERification. arXiv preprint arXiv:1803.05355.

Tsang, S. J. (2020a). *Issue stance and perceived journalistic motives explain divergent audience perceptions of fake news*. Journalism.

Tsang, S. J. (2020b). Motivated fake news perception: The impact of news sources and policy support on audiences' assessment of news fakeness. *Journalism* &

Mass Communication Quarterly. doi：10.1177/1077699020952129.

Tsfati, Y. & Ariely, G. (2014). Individual and contextual correlates of trust in media across 44 countries. *Communication Research*, *41*(6), 760-782. doi：10.1177/0093650213485972.

Vallone, R. P., Ross, L. & Lepper, M. R. (1985). The hostile media phenomenon：Biased perception and perceptions of media bias in coverage of the Beirut massacre. *Journal of Personality and Social Psychology*, *49*(3), 577-585. doi：10.1037/0022-3514.49.3.577.

Vosoughi, S., Roy, D. & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151. doi：10.1126/science.aap9559.

Wang, W. Y. (2017). "Liar, Liar Pants on Fire"：A new benchmark dataset for fake news detection. arXiv preprint arXiv：1705.00648.

Zhao, J., Cao, N., Wen, Z., Song, Y. L., Lin, Y. R. & Collins, C. (2014). #FluxFlow：Visual analysis of anomalous information spreading on social media. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1773-1782. doi：10.1109/TVCG.2014.2346922.

Zhou, X. Y. & Zafarani, R. (2019). Network-based fake news detection：A pattern-driven approach. *ACM SIGKDD Explorations Newsletter*, *21*(2), 48-60. doi：10.1145/3373464.3373473.

Zhou, X. Y., Jain, A., Phoha, V. V. & Zafarani, R. (2020). Fake news early detection：A theory-driven model. *Digital Threats：Research and Practice*, *1*(2), Article No.：12. doi：10.1145/3377478.

Zhou, X. Y. & Zafarani, R. (2020). A survey of fake news：Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, *53*(5), Article No.：109. doi：10.1145/3395046.

Theorizing an Information Diagnosis System：The Hybrid Uses of Manual Fact-checking, Crowdsourcing and Artificial Intelligence

Stephanie Jean Tsang, Yu Huang, Yin Nick Zhang, Yunya Céline Song, Lin Zhou

(*School of communication, Hong Kong Baptist University*)

Abstract　The emergence of social media has bred the dissemination of misinformation, which has attracted much scholarly attention. While the literature has mainly focused on the production, dissemination and reception of (mis)

information, there has been a lack of a comprehensive theoretical framework that sheds lights on how readers process misinformation as well as the absence of a practical framework to diagnose misinformation. This article reviews the current literature on misinformation and proposes an information diagnosis system by answering the following three research questions. (1) What causes a person to question the authenticity of information and the author's intentions? (2) How would a person verify a piece of suspicious information? (3) How should a fact-checking agency conduct verification to meet the challenges posed by the above two research questions? Two rounds of survey data collected in Hong Kong in 2019 ($N=1211$) and 2020 ($N=1223$) revealed that audiences held a diversity of definitions when it comes to misinformation and that they would only verify information perceived as suspicious. In addition, using HKBU FactCheck Service, a local fact-checking center, as a case study, we propose a hybrid information diagnosis system to verify misinformation to overcome the limitations of traditional verification methods. Given the rapid development of technologies, such as data mining, machine learning, and deep learning, it is necessary to take advantage of automation to complement the work of fact-checkers in misinformation detection. Hybrid applications of manual fact-checking, crowdsourcing, and artificial intelligence suggest the need to pursue fact-checking in a collaborative, interdisciplinary manner.

Key Words Misinformation; Information Diagnosis; Artificial Intelligence; Fact-checking

(编辑:吴璟薇)