万维网信息可信性问题

孟小峰 艾 静 马如霞
中国人民大学

关键词：万维网数据管理 万维网信息 可信性 可信度传播机制

随着万维网（Web）应用的迅速发展，目前一个门户网站的信息量已经远远超过网站管理者所能处理的能力范围。同时，也改变了人们获取信息和进行消费的方式。人们更愿意通过网络获取各种资讯，寻找数码新品、电影和歌曲，网上购物、交易，甚至在网络上找工作、找配偶。万维网已成为人们获取信息的重要途径。

然而，互联网是一把双刃剑，它在给人们带来极大方便的同时，也带来了一系列的问题。从这些纷繁复杂的信息中甄别可信信息已经变得日益重要起来。在互联网发展的过程中，由于信息可信性（Credibility）问题所引发的麻烦和损失不胜枚举。例如：2003年3月29日，国内一家网站的编辑错误地把微软公司总裁比尔·盖茨遇刺身亡的虚假消息当成CNN（Cable News Network，美国有线电视新闻网）的新闻（如图1所示），此后该消息被新浪、搜狐等一些影响力巨大的门户网站在第一时间以醒目标题转发。在半个小时之后，微软公司出面澄清，这一骇人听闻的重大新闻只是个谣言。最新发布的2009中国未成年人互联网运用状况调查报告显示，被调查的中学生大多对网上信息的真实可信度持怀疑态度，认为网上大部分或绝大部分信息真实可信的只占26.6%。互联网是否可信已经是一个困扰人们很久的问题，在一定程度上也阻碍了互联网的发展。因此，针对信息可信性问题的研究是网络进一步健康发展的需要。

图1 微软公司总裁比尔·盖茨遇刺身亡的虚假新闻

万维网信息不可信的成因

造成万维网信息不可信的原因多种多样，主要包括5类：（1）信息具有时效性。信息的效用依赖于时间并且有一定的期限，其价值与提供信息的时间密切相关。例如，某人的单位信息和联系方式随着他的工作变迁在不断地发生变化，因此当我们寻找其联系信息时就会遇到一些过时的信息，这些过时信息需要我们进行甄别；（2）信息发布者由于自身专业知识不足而发布了一些错误的信息；（3）信息发布者为了自身的利益故意制造虚假信息。例如，电子商务网站附带的用户点评论坛，其中包含大量的用户评论信息以及使用感受等，对该网站的用户决定是否要购买某种商品有非常重要的影响作用（用户总是更信任同为消费者的其他用户）。因此某些别有用心的人（如某件商品的出售者）就有可能故意假扮用户来引导评价或打分；

（4）信息发布者发布一些具有导向性的信息。例如，关于“伊拉克战争是否正确”，不同的网站可能因为立场或政治观点的不同，而选择只发布有利于自己观点的信息。这些信息描述的事件确实是在伊拉克战争中真实发生过的，所以都不是虚假信息。然而，如果只判断网站上的信息描述的是否是现实世界中真实发生的事件，而不考虑网站的情感倾向性，也会造成用户的误判；（5）万维网为信息的传播提供了良好的土壤。它的开放性和信息共享性等特征使得信息的传播更加迅速和高效。当然，这也为低可信度信息的传播提供了便利条件。同时，这些不可信信息一旦在网络中传播开来，需要很长的时间来清除。

基于上述原因，万维网上的不可信信息可分为以下几个类别：过时信息、错误信息、虚假信息、超前信息、片面信息和带有特定感情色彩的导向性信息等。根据这些信息的危害程度将信息分为：高风险信息、中度风险信息和低风险信息。其中，高风险信息主要包括：有意发布的虚假信息和病毒等；中度风险信息包括：无意造成的错误信息、过时信息、超前信息和片面信息等；低风险信息主要包括带有特定感情色彩的导向性信息等。

万维网信息的可信度

伴随着万维网技术发展，万维网信息可信性的研究面临各种各样新的挑战。万维网技术的发展经历了三代历程：Web1.0、Web2.0和Web3.0，如图2所示。

图2[1]形象地反映了三代万维网之间的异同。在Web1.0中，信息发布者与用户的角色都是固定的，因此只需要从少量的信息发布者着手控制信息的质量，就可以在很大程度上保证信息的可信度。这一时期，信息的可信度研究已经悄然兴起，并且成为一项重要的研究课题。最初的研究主要是从网站的可信性开始，研究如何设计网站从而使其更加可信。在Web2.0中，用户不再是只能被动地做信息的接收者，也可以作为信息的创造者向网上发布自己的信息。网上论坛、博客和合作知识库等网络应用应运而生。这时候

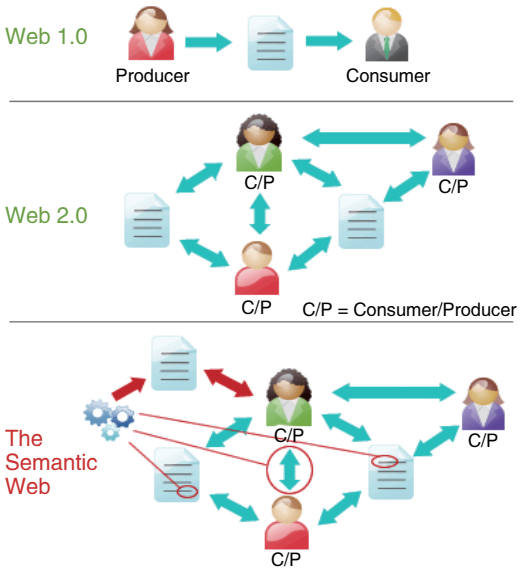

图2 Web发展的三代历程

万维网信息可信度的研究不但要考虑网站本身的特征对信息可信度的影响，还要考虑信息的生产者——普通用户的参与对万维网信息可信度的影响。用户的专业程度、历史信誉等因素都会对该用户所发布信息的可信度产生影响。Web2.0技术强调的是用户参与和交互。这一特点也使得用户之间在现实生活中的各种关系在网络中也逐步体现出来，如信任关系、合作关系等。研究这种人与人之间的相互信任关系，进而通过可信度的传播机制来计算信息的可信度也是可信度研究方面的一个重要议题。在Web3.0（语义网）中，参与者不仅仅是人，还有可能是计算机。语义网被看作是这一代网络的代表。语义网提供对数据语义的描述，使计算机能够“理解”万维网信息，实现计算机之间、计算机与人之间的智能交互，从而使万维网成为全球化信息共享的智能服务平台。计算机中的智能体（Agent）将成为网络的操作者和信息的传播者，所以需要研究智能体的可信度评估与计算等问题。与此同时，由于研究网络中的各个参与者之间的相互联系对信息可信度的评测具有重要的影响，因此不能孤立地进行研究。参与者之间的相互关系主要包括：网站之间的关系、网站与用户之间的关系、用户与智能体之间的关系等等。

下面从目前万维网信息产生过程中所涉及信息

本身、网站和用户三个方面分别论述信息可信度的研究现状。

面向信息本身的信息可信度

信息本身的可信度问题主要是从信息自身出发，根据其内容特征、相互之间的关联关系来研究信息的可信度。目前这方面的研究包括如下两个方面：

基于内容分析的可信度 本质是用信息本身的特征作为评价信息质量和可信度的标准。通常，基于内容分析的可信度计算方法主要包括两种：（1）基于拼写错误的评价方法。基本思想是使用文章中的拼写错误率作为评价这篇文章的数据质量和可信度的标准，认为网页中单词的拼写错误率与信息的质量是正相关的。这种评价方法的应用范围很广，因为它利用信息的文本特性来进行可信度及质量的评价，可以用于判断几乎任何网页的信息质量；（2）基于关键特性量化的评价方法。认为关键的信息特性对信息的质量、价值和可信度的评价有至关重要的作用，因此根据某一类具体的应用场景，通过对信息的几个基本特性的分析，将信息的可信度用量化的数值表示，进而计算出一篇文章（一个文档或一个网站）的信息可信度。从实际应用场景中抽象出数学模型是这一类方法的基本思路。

基于信息关联的可信度 信息之间并不是相互独立的，存在着各种关联，如超链接关联、引用关联、拷贝关联和语义关联等，这些关联在一定程度上影响着信息的可信度。目前这方面的研究大多基于超链接关联来分析页面的可信度，而对于其它关联关系则关注较少。基于超链接关联的信息可信度计算方法通常根据网页之间的超链接关系，如指向它的链接的重要度和可信度，来计算页面的可信度，这种思想类似于谷歌的网页排名算法PageRank。

尽管信息本身可信度的研究已经受到了一些关注，但仍存在如下问题：目前这方面的研究工作仍旧集中在文本信息的可信度上，对于图像、音频和视频信息的可信度研究相对比较少；在基于信息关联的可信度研究方面，主要集中在超链接关系的研究，对于信息间的隐含关联（如语义关联等）研究还很缺乏，解决这些隐含关联对于万维网信息可信度的影响是一个非常具有挑战性的工作。这些问题还有待进一步解决。

面向网站的信息可信度

网站可信性是人们最早关注的万维网信息可行性研究领域。其基本思想是网站的可信度与该网站中信息的可信度是正相关的，一般来说权威网站上面的信息比普通网站上的信息更为可信。网站的可信性可以从两个角度来判断：

独立网站的可信度 这类研究只考虑网站自身的特征，根据该网站中数据的质量、网站所有者的权威性、网站的领域特征等方面信息来判断网站的可信度。关键特征分析的可信度评估方法是最早用来研究网站特征对信息可信度影响的方法。例如，美国斯坦福大学的研究者曾在2001年发布了一个关于网站特点与其可信度关联关系的大规模调查报告。通过对来自美国和欧洲各国的1400多名志愿者进行问卷调查，评估51个不同的网站特点对其可信度的影响。这一工作是一个大规模的用户调查，为进一步研究万维网的信息可信性打下了基础。调查结果对如何设计更可信的网站有指导意义。

基于网站间依赖关系的可信度 主要考虑网站之间的依赖关系（如：相似依赖、相异依赖等）对网站可信度的影响，目前这方面的研究还非常少。

面向用户的信息可信度

根据Web2.0的理念，网站通常是由建设者负责建立信息发布平台与架构，主要的信息资源是由网站的用户参与发布。这种形式的万维网应用有很多种，如：BBS（Bulletin Board System，电子公告板）和合作知识库等。这些万维网应用都是合作的信息源，它包含了许多不同发布者的观点和看法，是集体智慧的集合。它的信息比仅由服务商提供信息的网站更加客观，并且更能反映大多数普通用户的感受和看法。以上这些都得益于用户的参与，然而用户的参与也给信息的可信性带来了很大的影响。用户之间的信息交互将不同的用户联系起来，

形成了一个社会网络。人们在这个网络中进行信息交流，这种信息的交互行为也给信息的可信性研究带来了巨大的挑战。因此，从用户的角度来研究信息的可信度具有现实的意义，目前这方面的研究主要有两种评估方法：基于评分和投票的可信度评估机制和基于信任传播的可信度评估机制。

基于评分和投票的可信度 基本思想是基于发布者的可信度、帖子的信息内容特征值等信息，进行归类、评价和打分，从而对以帖子为基本单位的信息的可信度进行判断。常用的归类与计算可信度值的方法是机器学习方法。目前主要的方法分为三大类：（1）根据帖子本身包含的信息与信息特征值，利用机器学习算法对帖子的信息可信度进行归类计算；（2）考虑用户可信度的影响因素，根据用户的历史记录计算其可信度，并辨别论坛中的信息的偏向性；（3）以作者为中心，建立可信度管理模型，建立作者与版本、作者与作者、作者与不同的词条（文章）之间的关联关系。

基于信任传播的可信度 在可信度传播机制中，不再是关注信息本身的可信度，而关注的是信息发布者的可信度，然后再根据信息发布者的可信度来判断信息的可信度。可信度传播机制主要建立在信任网络的基础之上。在信任网络中，节点根据其应用场景的不同代表了不同的实体，边反映了网络节点之间的信任关系。然后，以信任网络中少部分节点的可信度值作为先验知识，对目前可信的节点所信任的节点，均认为是可信节点。利用网络结构和邻近知识进行可信度值的传播。而且，信任值的传播操作，通常是通过对信任值矩阵（Trust Value Matrix）的变换操作实现的，包括矩阵加法、乘法等。这种网状结构中的信任机制，基本上都是利用名誉的传播机制，将节点的局部知识通过一个中央机制综合起来，从而得到统观全局的每个节点的可信度情况。信任网采用这种机制，使原本只知道他周围邻居的可信度的每个用户，通过整个系统采用合适的传播机制和信任值计算机制，可以预测出系统中任意两个用户之间应该对对方赋予的信任值。在P2P（Peer to Peer）网络、社交网络和语义网的环境中，信息可信度的评估通常采用传播机制来实现。绝大部分方法都是在信任网的基础上，根据自身的特点做一些改进，然后用传播机制（可能包括信任值和不信任值的传播）在节点之间传播可信度经验值。根据不同网络各自特有的特点，增加不同的机制。此外，图挖掘算法也是常用的方法之一。节点之间的网状机构可以被看作是一个图，而挖掘算法常用于对未知节点的可信值做推断。因为挖掘算法可以找出事物之间的联系，从而推断出未知的情况。

挑战与展望

虽然目前已经有许多针对万维网信息可信性的工作，但是缺乏系统性，还存在一些需要解决的问题。

图像、音频和视频信息的可信性 依据信息的载体特点可以将万维网上信息分为文字信息、声音信息和图像信息三大类。其中关于文本信息的可信性问题的相关研究工作最多。然而，近年来图像信息和音频信息在万维网上所占的比例越来越大，它们承载的信息量也非常巨大，受到研究者们越来越多的重视。目前，关于图像信息和音频信息的相关研究工作越来越多，但关于这些信息的可信性问题的研究工作却较少。因此，如何准确、有效地计算和评估图像和音频所携带的信息的可信性及可信度值，必将是未来重要的研究课题。

基于信息间隐含关联的信息可信性 目前基于信息间关联分析的可信度研究主要集中在网页之间的链接关联方面，而对信息之间隐含的其它一些关联的研究相对比较少。例如，信息之间的语义关联对可信度的影响。我们可以从语义的角度来区分信息内容的近似程度，从而利用这些相似度和信息的其它属性（如发布时间等）分析出信息之间的相互引用关系，进而根据这些引用关系来判断信息的可信度。

基于网站间依赖关联的信息可信性 网站间的依赖关系错综复杂，例如：超链接关系、拷贝关系和合作关系等。从网站的海量数据中发掘这些关系、分析这些关系之间的相互影响以及这些关系对信息可信度的影响程度和方式等都使得网站间依赖

关系的发掘具有非常大的挑战性。

构建多层次可信性关联 目前，在信息之间、网站之间、用户之间的相互关联对可信度的影响方面已经有了或多或少的研究。但是网站和信息、用户和网站、用户和信息之间也存在着很多关联，这些不同类型实体间的关联同样也影响着信息可信度。例如：网站中所有用户可信度影响了网站的可信度，进而影响着网站中信息的可信度，网站中信息的可信度反过来又可以影响网站的可信度，这些影响是相互渗透的。所以，通过构建多层次可信性关联的研究可以更加系统地研究信息的可信度。

用户发布信息的可信性验证 万维网中对已有信息的可信性评估只是一种亡羊补牢的做法，是对已经发布到万维网上的不可信信息进行甄别。然而，如图3所示，更好的办法应该是防患于未然，将不可信信息挡在万维网之外，因此我们需要在用户发布信息时就对用户和信息进行可信性验证。目前关于如何评估万维网上已有信息的可信性的研究工作有很多，然而用户发布的信息的可信性验证研究则关注较少。由于用户创造的消息种类和形式较多，如何验证这些用户的可信度及其发布信息的可信度，确保用户发布的信息是可信的，是一个非常具有挑战性的问题工作。

可信性评估算法基准测试 除了上述挑战性问题外，在万维网信息可信性研究体系中还缺少对可信性评估算法基准测试的研究。在关于万维网信息可信性的工作中，研究人员通常会提出一种或几种计算信息可信度值的算法，或者评估信息可信度值的评价机制，然后在数据集上通过实验证明这些算法和评价机制的准确性。然而，信息可信度算法和评价机制缺乏统一的数据源进行实验验证，而且没有基准测试（Benchmark）程序或规范，除了某些文章中有算法比较之外，很难知道不同的算法方法之间效果及效率的差异。因此，为万维网上的信息的可信度算法和评估机制提供统一的权威的实验数据集，以及基准测试是非常必要的。

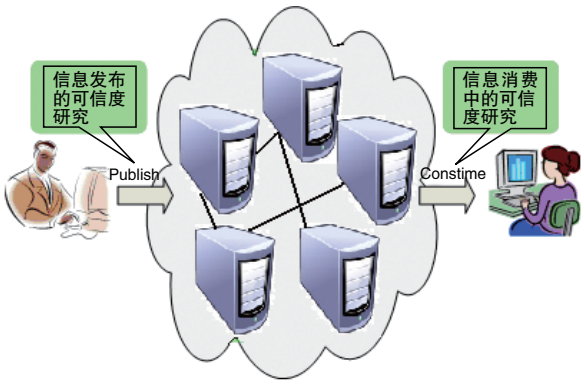

图3 万维网信息生产消费中的可信性问题

结语

万维网信息将在未来人们的生活、工作和研究中以及互联网的发展方面占有越来越重要的地位。本文分析了万维网信息可信性问题产生的原因、国内外信息可信性发展的现状，并在此基础上讨论了信息可信性问题所面临的挑战与未来的研究工作。万维网信息可信性研究对提高万维网中信息的质量和网络的进一步健康发展具有重要的意义。■

孟小峰

CCF理事。中国人民大学信息学院教授。主要研究方向为网络与移动数据管理，包括Web数据集成，XML数据库，云数据管理，闪存数据库和隐私保护等。xfmeng@public.bta.net.cn

艾 静

中国人民大学计算机应用技术专业硕士生。主要研究方向为Web数据管理。

马如霞

中国人民大学计算机软件与理论专业博士生。主要研究方向为社会网络和Web数据管理。

参考文献

[1] http://blogs.nesta.org.uk/innovation/2007/07/the-future-is-s.html