学科漫谈

微博数据可视分析

■袁晓如

袁晓如

北京大学信息学院研究员、信息科学中心副主任

微博是基于用户关系的信息分享、传播、获取的平台,它内容简短,以不到 140 字公开的短消息,用户能够通过它交换一些小规模的信息,诸如短句、个人照片、视频链接等。它允许用户及时更新自己的个人信息并与他人交流,维护自己的人际圈。微博提供通过手机和电脑随时随地的发布途径,对社会的活动和个人的生活方式产生了重大的影响。从世界的各个角落发布的每一条微博,如同无数的社会化的传感器,记录着全球每时每刻发生的点点滴滴。微博使世界上的每一个人都成为信息源,并使之在全球传播,这使得微博所承载的信息量大大增加。从这聚集成的信息洪流中,提供了另一个隐约窥见世界全貌的途径。

微博可视化的现实需要

研究微博上的信息具有十分重要的意义。首先,微博集合了海量的新闻、事件和信息,并且每天都在更新,每天都在流传,并对现实的社会产生巨大的影响。尤其是在突发事件的信息传播上,微博更是超越了传统媒体,成为了信息快速传播的渠道。最早爆料出本·拉登死讯的并不是各大媒体,而是 Twitter。

其次,微博上的信息不仅发布及时,而且也是现实社会生活的缩影。挖掘微博上的信息有利于分析现实世界的情况。东南路易斯安娜大学的助理教授 Aron Culotta 曾经通过追踪一些与流感有关的关键词,如“flu”“headache”等,进行流感爆发趋势的预测。他利用发布于 2009 年 9 月到 2010 年 5 月间的近 5 亿条信息建立起了一个预测模型。通过该模型的预测结果与美国疾病预防控制中心的统计数据惊人地相符。

虽然微博信息不一定精确,但它的时效性强,不需要花费大量的人力物力去收集信息,这大大方便研究人员进行快速分析。当然,通过微博搜集到的海量数据也是传统数据收集方法所不可比拟的。

另外,每个用户在微博上还维护这一个人际交往圈,现实生活中的好友、网络好友、新朋友、朋友的朋友……这形成了一个错综复杂的人际网络,并逐渐对其自身造成潜移默化的影响。因此,微博上的人际关系也是一个十分有趣的分析内容。

微博上的信息海量、复杂且多样,传统的数据分析方法已经很难适应这一特点。而利用可视化的工具,对微博数据进行可视化、可视分析并加以人机交互,是一个十分有力且具有广大前景的研究方向。

标签云与 Wordle

标签云是一种使用广泛的可视化方法,它根据标签的热门程度来确定其字体大小,在许多网站、博客上都能见到它的身影。Wordle 是一种比较流行的、将文本中关键词可视化的方法。它极具视觉美感,可以在短时间内在感官上给人冲击,吸引读者,并能让使用者轻易地抓住文本中最主要的关键词。Wordle 同样是使用字体大小来表示词语的权重,通过把关键词按照一定外轮廓紧密地排列达到美观的效果。

微博转发

微博的一个重要传播特性是用户可以转发感兴趣的微博,从而形成链式的传播。一条微博可以在短时间内被成千上万的用户转发。北京大学可视化与可视分析实验室开发的 WeiboEvent 工具(http://vis.pku.edu.cn/weibova/weiboevents)(图一)就是可以便捷地可视化一条微博如何被其他人转发传播。通过几种不同的可视化方法,可以分析挖掘转发随着时间变化的状况以及参与转发的重要用户。

网络和地图

网络是社交网络可视化中经常使用的一种表现形式。通常情况下,它用“点”表示人,用“线”表示人与人之间的关系。将一个复杂的社交网络用可视化的形式表现出来,可以比较直观地展示网络中的人际关系情况。再加以人机交互的手段,可以挖掘出一些深藏在数据背后的信息。

TweetWheel 是 Twitter 上另外一个的好友关系可视化应用。它将好友排列在圆周上,互相认识的好友间都连一条曲线,便形成了这样一个美妙的圆盘,方便用户对好友间关系进行探究。如图一的好友关系好像“一盘意大利面条”,把“面条”从“盘子”里提溜出来,就是一组好友关系。

地图是一种简便、直观,也是目前非常流行的展现地理信息的可视化方式。它主要根据地理位置的不同,将不同地区的数据展示在地图上。随着移动互联网的爆炸性发展,我们可以越来越方便地获取到更加精确的地理位置信息。在这大量地理数据的背后,还有很多有意思的东西正等待着人们去挖掘发现。其中,微博上基于地理信息的可视化就有很广阔的前景。

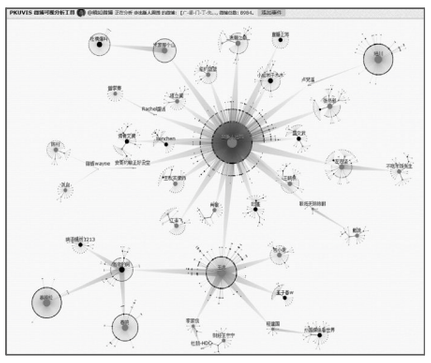
图二所示是在大约两周时间里,中国部分地区发布带有地理标记的微博的密度图。这幅微博人口活跃度“地图”和实际的城市发布大致吻合,我们可以清楚看到城市的发布,特别是沿着一些铁路大动脉的热点。

Bits Pics 是一个十分有趣的应用。作者 Eric Fischer 用它展示了用 Twitter 发消息和用 Flickr 发照片的用户的地理分布。地图上橙色的点表示使用 Flickr 发照片的用户,蓝色的点代表使用 Twitter 发消息的用户,而白点则表示两者均使用。

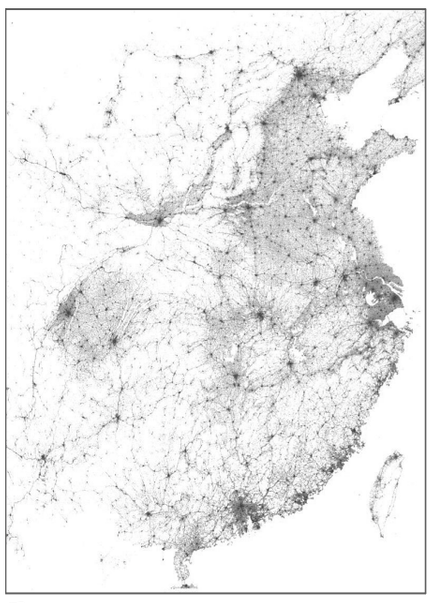
首先来看看全球的用户分布情况。我们可以看到,使用这两者的人多分布在美国和西欧地区,日本地区也有很多人在使用。我们可以很轻易地发现美国西部的人更偏爱发照片,而东部的人偏爱发微博消息,而那些比较明显的白色点大部分都是美国人口密集的大城市。在图上还隐约能看到几条横贯东西的白线,那些正是美国的高速公路。

从海量的微博数据中,提取与事件相关的地理空间信息,对社会和人们的日常生活都有着重要的意义。自然灾害、突发事件等事件的感知和应对,需要比专业测量更快速、更及时更新地地理空间数据;对日常生活中的话题、事件的地理分布的获取,又需要比专业测量更方便、低成本的方式。微博中的公众用户,就如同大量的社会化传感器,时刻发布着可能包含地理位置的各种事件的目击、描述和评论。通过提取微博中的地理空间信息,能够在一定程度上满足大众对各类事件了解的需求。北京大学发展的 ThemeMap 可视分析平台根据微博上大家对特定主题讨论的位置产生相应的主题地图,它结合了微博地理位置提取的自动化算法,和志愿地理信息系统的公众参与的思路,提供了对主题、事件的更好的地理位置提取和可视分析。它利用公众参与,能够更充分、准确地提取微博的地理位置,从而达到更好的可视分析效果;通过利用已有的微博数据和自动化算法,极大地降低了志愿者的参与难度和时间成本。

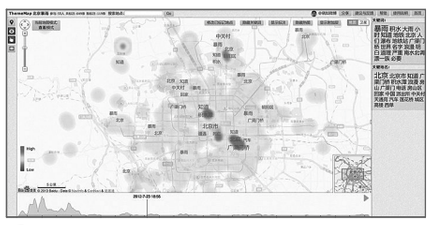
在 2012 年 7 月 21 日,北京经历了 61 年来最大的一场暴雨,超过 200 万人的生活受到了影响。在这场暴雨中,北京城中城郊的许多地点、道路产生了严重的积水,最深的淹水处有 5 米之深。暴雨发生之时、之后,新浪微博上爆发了许多谈论暴雨和积水的微博,其中许多谈论了积水发生的地点。图三是根据“北京 & 暴雨”为关键词,在暴雨发生阶段的微博产生的地图,基本反映了暴雨事件中主要严重积水地点。



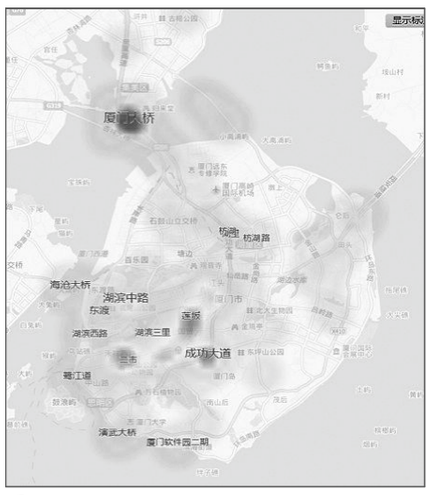
图一



图二



图三



图四

图四所示的为厦门堵车地图,是作者在微博上一名厦门交警的建议下以“厦门 & 堵车”为关键词创建的 ThemeMap 主题地图。

干细胞早知道⑪

间充质干细胞的临床试验及其前景展望

■韩忠朝

间充质干细胞的独特生物学特性使其临床移植成为继药物治疗、手术治疗之后的又一种治疗新技术,并在国际范围内开展临床试验及应用。截至 2011 年 11 月 20 日,美国国立卫生研究院的临床试验登记网站(www.clinical trials.gov)数据表明,已经登记的 MSC 临床试验共有 203 项,其中骨髓 MSC 临床试验 103 项,脐带胎盘 MSC 临床试验 37 项,脐带 MSC 临床试验 20 项,脂肪 MSC 临床试验 16 项,涉及几十种疾病,包括血液病、骨与软骨相关疾病、心血管性疾病、神经系统疾病、遗传性疾病、皮肤疾病、肺病、肾病、肝病、床上修复、自身免疫性疾病、糖尿病、眼科疾病等。以下仅列举间充质干细胞在血液病和糖尿病中的治疗作用。

间充质干细胞在血液病中的应用

造血干细胞移植(HSCT)是采用大剂量放化疗和(或)其他免疫抑制预处理,清除受者体内肿瘤细胞或异常克隆细胞,然后将体内或异体造血干细胞移植入受者体内使其重建正常的造血和免疫功能的一种细胞治疗手段。这种方法广泛应用于良恶性血液病、遗传性疾病和某些实体瘤。移植物排斥、植入失败与移植物抗宿主病是影响 HSCT 疗效的主要并发症。

目前有人尝试将 MSC 输注作为植入失败的挽救治疗,希望通过改善骨髓微环境来刺激残存 HSC 增殖分化。植入失败的传统治疗包括造血干细胞的二次回输、供者淋巴细胞输注、应用 G-CSF 和二次移植。这些治疗方式均有增加 GVHD 和 TRM 的潜在风险。Meuleman 对 6 例造血恢复失败的恶性血液病患者,在未进行预处理和无 HSC 支持的条件下,单纯输注 MSC 观察其对造血重建的作用。所有患者输注 MSC 前均为完全供者嵌合造血,原发病为缓解状态,但骨髓增生低下,不能脱离血小板和红细胞输注。患者分别在输注 MSC 后 12 天和 21 天造血恢复。有学者推荐将 MSC 作为植入失败的一线治疗。

间充质干细胞移植在糖尿病中的应用

近几年,各种来源的 MSC 已在糖尿病中进行临床研究和试验。截至 2011 年 11 月底,在 www.clinicals.gov 网站上注册的临床试验有 19 项,包括Ⅰ型和Ⅱ型糖尿病。

干细胞治疗疾病的原理一直是科学家想弄清楚的问题,近几年这个领域的研究虽然发展得很快,但总的来说,人类对干细胞的了解才刚刚开始,对其疾病的治疗原理也正在探索中。根据已有的文献报道,干细胞治疗糖尿病的原理大概包括几个方面:①干细胞定植于胰腺受损部位诱导分化成胰岛 β 细胞;②干细胞分泌的因子对胰岛 β 细胞的修复起重要作用;③干细胞降低胰岛素抵抗的作用;④干细胞的免疫调控作用;⑤部分干细胞分化内皮血管细胞,改善胰腺血液供应。

从以上干细胞的基础研究和临床应用情况来看,有许多可喜的进展。建立健全规范的法律法规及统一的生产和质量标准是发展干细胞技术的关键。只有通过严格的临床试验审查,在严格监管下才能保证干细胞技术应用的安全性和有效性。通过建立统一的生产和质量标准,对组织来源、细胞筛选、制备过程和最后的产品进行严格的检验,获得符合标准的干细胞才具有质量上的可靠性,才可以让干细胞的研究和应用健康发展。

栏目主持:赵广立

趣味科学

开香门:让树木结出沉沉的香

■本报记者 王剑

“袅袅沈水烟,乌啼夜阑景。”这是唐代诗人李贺在诗中描述一位公子在夜阑人静之时点燃沉香,思念情人的情景。

植物香,自古以来就被赋予了诸多浪漫情怀和神秘色彩。为了寻找香料,早在 15 世纪初人类就开启了探险时代。可以说哥伦布“发现”美洲都是探索香料、寻找香料的副产物。

在众香料中,沉香作为中国传统名贵香料,以其常温下香气淡雅、燃之浓郁甘甜,温和醇厚,且历久不散,加之成香时间漫长,稀少难得,因此自古为世人所推崇。

近日,《中国科学报》记者就香料在生活中的作用、人们对香料的探索历史以及沉香形成过程等问题,采访了中国科学院华南植物园研究员廖景平。

“实际上,香料在人类认识自然的过程中发挥了非常重要的作用。如果没有对香料植物的探索,就没有对热带植物的发现,也就没有我们今天赖以生存的粮食作物的发现和驯化。”廖景平在谈到人类对香料最初的探索阶段,向记者介绍:“从 15 世纪起,由于荷兰人对香料的垄断,横跨亚欧大陆的土耳其奥斯曼帝国阻断了香料贸易的陆上交通,于是就出现了包括葡萄牙、西班牙以及后来法国、英国等探索自然、寻求香料的热潮。”

香料当中,沉香是上好的,同时也是非常名贵的中药材。它原本只是看上去黑漆漆的一段朽木,但因能散发出素雅悠远的香而为世人所重视。

廖景平告诉记者,沉香是密实的固态凝聚物,混合了树胶、树脂、挥发油、木质等多种成分。天然香树一般要经年累月才有可能形成“香结”。而形成“香结”之后,还要经过漫长的时间才能真正“成熟”。有的香树寿命长达几百岁以至上千岁,其倒伏后留存的沉香往往也有几百岁以上的“寿命”,廖景平感慨道:“所以就有了古人对沉香‘集千百年天地灵气’的赞叹。”

沉香的形成过程(也称结香)很奇妙也很复杂。廖景平介绍,就目前的研究,若树干由于虫蛀、外伤等多种原因,有真菌侵入,则常常引起香树的系列变化,使树胶、树脂、挥发油等成分聚集沉积,形成“香结”。香树倒伏后,再经过长期的腐蚀、分解,在剩余的“香结”处即可形成沉香。《本草纲目》曾引前人经验解之曰:沉香“其积年老木,长年其外皮俱朽,木心与枝节不坏,坚黑沉水者,即沉香也”。用现在的话讲,就是由于树干损伤后被真菌侵入,在菌体作用下,使木薄壁细胞产生一系列变化,形成香脂,经多年沉积而得。沉香树开始分泌树脂后,原本宽松材质自此开始变硬,成长阶段真菌“侵入”心材,进而产生共生变化,树心颜色变深,硬度密度逐渐增加,此时输送养分的组织受阻而生长顿止,树干因无法支撑重量而倒伏,自然分解成各种不同形状。不同沉香有不同的沉香气味,因每棵沉香树的生长时间、过程、地理环境及品种的不同,受伤结香的条件各异,而酝酿着不同的油脂、香韵及颜色。这也是沉香独具魅力之处。

廖景平向记者介绍,能形成沉香的植物主要是瑞香科沉香属的 8 种树木,包括印度沉香树、厚沉香树(又称奇楠沉香树)、马来沉香树(又称容水沉香树)、白木香树(又称莞香树)等。

为了获得沉香,民间会把香树的茎干人为地刀砍打孔,使虫蛀、病腐后,使其感染真菌,并把伤口封闭起来,而后在菌丝所分泌的酶类作用下,产生一系列变化,最后形成香脂。这个通过人工干预促进香树结香的过程被称为“开香门”。廖景平说,首次得到的香叫作“初香”,但是它的品质不一定最好。而每一次开香门后结香的品质都有不同,真菌的种类和数量越多,其混合香就越复杂,在树与真菌共同的作用下,越往后的结香品质就越好。

沉香种类多样而珍贵,不仅是上等药材极品,同时也是供佛修持的圣品。古今中外概莫能外。

英国民谣《绿袖子》有一句歌词:“我燃心香,寄语上苍。我心犹炽,不灭不伤。”想必歌中所唱的那位暴戾的亨利八世在思慕绿衣少女时所燃心香也是袅袅沉香吧?

科普问答

“寻踪”暗物质

■本报记者 郝俊

前不久,中国暗物质实验合作组(CDEX)的最新研究成果在《物理评论 D》发表。据媒体报道,该成果将暗物质探测灵敏度提升了大约 10 倍。

有不少物理爱好者非常关注暗物质的研究进展。为此,记者采访到多年从事暗物质研究的中国科学院高能物理所研究员毕效军,请他以较为通俗、生动的方式,向大家介绍暗物质研究的背景和此次研究进展的意义。

《中国科学报》:首先请介绍一下什么是暗物质?

毕效军:简单来讲,暗物质就是“在宇宙中还没有被我们‘看到’的物质”。这个解释马上会引起两个疑问:何为“看到”?既然没有“看到”,那凭什么知道宇宙中存在这样的物质?

在天文学上,技术发展最充分、设备最为先进、能力也最强大的“看”的办法,就是探测来自太空不同物质发射的各种波段的电磁波。没有“看到”的物质,也就是指在全部的电磁波段,我们都没有看到这类物质的电磁辐射,它们对于我们来讲是“黑暗”的。

另一方面,虽然我们从来没有直接“看到”宇宙中存在暗物质,但我们却发现了由于暗物质的引力作用对其他可见的物质运动的影响,这就是我们断定宇宙中存在暗物质的理由。目前从星系到宇宙学尺度的观测,都发现可观测物体运动的异常现象,由此我们推断宇宙中可能存在我们还没有“看到”的物质,即暗物质,它们通过引力效应影响了可见物体的运动。

《中国科学报》:既然如此,科学家如何对暗物质进行探测?

毕效军:由于暗物质粒子不参与电磁相互作用,通常的天文观测方法很难发现它的踪迹,为了了解其本质,目前的探测方法大致分为三类。

首先是在对撞机上将暗物质粒子“创造”出来,从而研究其物理特性。由于暗物质粒子不能被探测器直接观察到,因而,它们表现为一部分能量丢失了,从丢失的“能量”和分布可以推测暗物质的某些性质。

第二种方法是直接探测法。该方法是直接探测暗物质粒子和原子核碰撞所产生的信号。由于发生碰撞的概率很小,产生的信号也很“微弱”。为了降低由宇宙线与探测器碰撞造成的本底噪声信号,通常需要把探测器放置在很深的地下。暗物质直接探测实验,是目前寻找暗物质粒子最重要的探测方式。目前的实验精度下,我们只可能探测到弱作用重粒子(WIMP)的信号,而更弱的信号,如轴子、超对称引力子是无法用这种方法探测的。中国暗物质实验合作组(CDEX)的研究,就是通过这种方法探测的。

第三种办法称为暗物质的间接探测法。间接法是观测暗物质粒子衰变或相互碰撞并湮灭后产生的稳定粒子如伽玛射线、正电子、反质子等。根据目前的理论模型,暗物质粒子衰变或湮灭后,可以产生稳定的高能粒子,如果我们能够精确测量这些粒子的能谱,可能会发现暗物质粒子留下的蛛丝马迹。

《中国科学报》:探测暗物质的科学意义是什么?

毕效军:我们认为,暗物质是在宇宙极早期产生的,并且遗留到了今天,当时的宇宙温度非常高,远高于目前对撞机所能产生的能量。探测宇宙中的暗物质并了解其性质,就相当于我们直接在观测宇宙的超大对撞机给我们产生出来的新粒子,这对理解超出粒子物理标准模型的新物理、理解物质的基本相互作用有重大科学意义。

《中国科学报》:CDEX 将暗物质探测的灵敏度提升大约 10 倍是如何做到的?

毕效军:主要是通过增加新的设备以压低本底、设计了新的数据分析方法等手段实现的。

《中国科学报》:有研究者将探测暗物质比喻为“追捕”,好比在茫茫大海上捕鱼。这是否意味着,目前科学家们的工作主要集中于对“追捕区域范围”的确定?

毕效军:我感觉,找暗物质有点像找今年 3 月“马航”的失联飞机,通过一块海域、一块海域的排除,最后找到飞机所在的地点。

我们不知道暗物质在哪个区域,就像不知道这个飞机在哪儿一样,只能一块、一块地排除完了才能确定。当然,也有可能,就是把怀疑的海域都排除了,仍然没有找到飞机。那说明,我们一开始的推断可能有问题,也许飞机没有坠海,而是在其他地方。

暗物质探测也一样。我们也是在搜索一块高度怀疑的“海域”,但目前仍然没有迹象。有可能在下一块海域就找到了,也有可能搜索了一遍仍然没有找到。

《中国科学报》:暗物质研究领域未来几年的重点方向是什么?科学家们能否预计,大概何时能够“捕获”暗物质的踪影?

毕效军:未来几年的重点仍然是增大探测器体积、减小本底以提高灵敏度,首先力争把我们怀疑的区域都搜索完。因而,何时“捕获”是完全无法预料的,运气好的话,也许明天就让我们碰上了,运气不好,也许搜索完了也没能发现。