2022.4.19 星期二
责编 白云 吴丽敏 实习编辑 孙少杰 联系电话:8222133
太原晚报 22版

大数据分析

- 唐代诗人白居易作品量第一,但影响力排10名之外。
- 宋词名篇中收录词作最多的不是苏轼、辛弃疾而是周邦彦。

大数据研究唐诗宋词 结论很颠覆

用大数据分析唐宋诗词结论超出你想象

用大数据分析唐宋诗词,结论可能超出你的想象——唐代诗人中作品量居第一的白居易,影响力排在十名之外;宋词名篇中收录词作最多的不是苏轼、辛弃疾而是周邦彦;综合影响指数表明,杜甫高于李白、辛弃疾强于苏轼……

以上新发现,是国家社科基金重大项目"唐宋文学编年系地信息平台建设"首席专家、四川大学文学与新闻学院讲席教授王兆鹏分析所得。

唐诗是中国诗歌史上第一座高峰。唐代诗歌5万多首,诗人3000余位,诗人和诗作都达到前所未有的量级。宋代词人近1500位,词作超21000阕。

从个体诗人来看,唐诗宋词里谁的作品最多?王兆鹏的大数据显示,白居易名列唐诗作品量的榜首,诗作近3000;杜甫和李白紧随其后,都超过千首大关。宋词中辛弃疾的词作量位居第一,有600余阕,其次是苏轼、刘辰翁。宋诗的篇数,则由陆游称雄,凡9000多首,其次是刘克庄和杨万里。

根据综合影响指数排名,唐代诗人影响力第一的是杜甫,其次为李白、王维,而作品量居第一的白居易,影响力排在10名之外。宋代词人作品量和影响力第一的都是辛弃疾,苏轼和周邦彦分别居第二、第三。高居宋诗影响力榜首的是苏轼,作品量雄居榜首的陆游紧随其后。

提到唐诗宋词的名家,人们习称"李杜""苏辛",似乎李优于杜、苏胜于辛。但综合影响指数表明,杜甫高于李白、辛弃疾强于苏轼。更令人意外的是,最受追捧的词人不是苏、辛而是周邦彦。在100首和300首宋词名篇中,周邦彦各占15首和40首,占有率远高于苏、辛。

用客观的数据去衡量、分析颇为主观的诗词鉴赏,是否科学、能否可行?在接受记者专访时,王兆鹏强调,虽然数据能在一定程度上描述显示文学史的发展面貌和进程,但也有明显的局限性。

30年前开始研究积累了上百万条数据

问:《大数据里的唐宋诗词世界》课题的初衷是什么?

答:我1992年就开始做唐宋诗词的定量分析了。初衷是每人心目中都有自己的唐宋诗词名篇。究竟历史上哪些唐宋诗词被视为名篇,我想用统计数据来分析衡量。

问:那您是如何利用大数据来衡量唐宋诗词质量的呢?这些数据是如何统计出来的?

答:唐宋诗词作品的质量,目前还没找到有效数据来评估、衡量。我目前正在努力尝试构建文学作品质量的评价指标体系,以便搜集数据。这需要比较长的过程。此外,个人建立的评价指标体系,需要学界的认同和共识。

问:关于文学指标体系,学界目前的研究现状怎样?

答:大数据时代的文学数据,需要分类分层建立起文学史数据的指标体系,以确保数据的信度和效度。但目前用大数据来做唐诗宋词研究的学者不多,为学界共享的唐宋诗词大数据也相当有限。

从1992年到现在,我虽然积累了100多万条和唐诗宋词有关的数据,但还不完备、不均衡。有的时段数据多,有的时段数据少;有的这一类数据多,那一类数据少;有的诗人数据多,有的诗人数据少。我们常感慨"书到用时方恨少",数据更是这样。全方位分析唐诗宋词时,常常觉得数据不够用。

在我看来,文学评价指标体系应该以作品为中心来建立。作家的影响力是以作品的影响力为前提。而作品评价,可分两个维度,一是相对稳定作品的内在文学价值,二是动态不居作品的外在影响力。其文学价值,可考虑从内容和形式两个层面来评估。

作品影响力则从创作者、评论者、普通读者三个层面来衡量。一是对创作者的影响,包括引用、化用、仿效、改编、翻译等,体现出作品的典范性和吸引力;二是对评点者的评论和学者的研究,反映出作品在文学批评、学术研究层面的美誉度和关注度;三是在普通读者中的传阅度和知晓率。确定作品的价值、影响的基本要素和结构后,再构建计算模型,然后由计算机在相关资源库、语料库和网络运行,挖掘提取相关数据,最后计算出每篇作品的得分。

数据无法测度艺术含量和审美价值高低

问:您在课题里提到,据统计,东汉到隋末近600年,诗歌总共才5000余首,而到唐代,诗歌第一次超过万首并直接跨越到5万多首。唐诗比之前的八代诗增加了七倍多,诗人由600余位增加到3000余位,诗人和诗作都达到前所未有的量级。这个数据从何而来,有参考哪些重要的文献资料吗?

答:数据来源于我的老朋友尚永亮教授的两篇论文:《八代诗歌分布情形与发展态势的定量分析》和《唐知名诗人之层级分布与代群发展的定量分析》。

问:白居易的诗数量最多,影响力却在前10名开外,这是如何判定的?

答:用数据衡定的。我们用了多种数据,对唐代诗人影响力进行排名。白居易的影响力,在现当代大于古代。他的综合影响力,远不如李白、杜甫。

问:那您通过大数据判定唐诗宋词质量的依据是什么?

答:目前只能用大数据衡量唐诗宋词影响力的大小——包括对后代词人创作的吸引力,在后代词评家中的美誉度,在普通作者中的知名度等等。目前暂时还不能用数据测度唐诗宋词艺术含量和审美价值的高低。

文学中心在北宋初就完全移到南方

问:用大数据研究唐宋诗词有无遇到一些学术上的困难,又是如何克服的?

答:文学研究从来没有数据意识,困难不仅在于到哪儿去找数据,更在于找什么样的数据。究竟什么样的数据有用有效,既需要理论的支撑,也需要在实践中检验。理论上,我们不断探求,从统计学、计量信息学和计量历史学中去寻找理论和方法的启示;实践上,反复试验,失败了重来。最痛苦的是,数据库建好了,文章也写完了,忽然发现数据来源不全,只好从头补齐数据,写好的论文又推倒重来。

问:您在大数据具体研究中还有哪些新发现呢?

答:数据的意义,既能确证传统的结论,也会修正传统的结论,更能发现新问题,改变传统的认知。比如,中国文化地理有一个著名的结论,中国文化中心,是由北方中原逐步向南方移动,第一次南移是东晋永嘉之乱,第二次南移是唐代安史之乱,第三次南移是宋代靖康之乱。3次战乱推动了文化中心的南移,靖康之乱后,文化中心就彻底移到南方。我们的大数据发现,文学中心在北宋初就完全移到南方,南方作者的数量全面超越北方,无需等到靖康之乱后。而且,战争不是推动文化中心南移的唯一因素。

我们还发现,宋代的文学中心,是逐步向东南沿海移动。按今天的地市级行政区划来统计,宋代福建南平的作者人数最多,名列第一,福州居第二,这很让人惊讶。与此相关的是,宋代进士人数福州第一,南平第二。可见当时南平、福州教育发达,进士多,诗词作者也多。教育与文学是高度正相互的。

此外,我们还发现苏东坡词的创作高峰是在黄州,他三分之一的词是在贬谪黄州期间写的,他的名篇佳作一半是在黄州写的。比如宋词的第一名篇《念奴娇·赤壁怀古》就是在黄州写的。黄州成就了苏轼词作的辉煌。

据《北京青年报》