［论文名称］大数据时代的古典文学研究　——以数据分析、数据挖掘与图像检索为中心

[作者]：刘京臣，青年人文社会科学研究中心,中国社会科学院文学研究所

［摘要］新世纪以来，随着大数据、云计算、图像检索等技术的发展，古典文学信息化的重点应当由数据检索向数据分析、数据挖掘和图像检索转型。

现阶段，在“文本”领域，最值得重视的便是大数据基础上的数据挖掘。面对数据挖掘，有两个思路：要么是人工制订足够多的规则，使非结构化文本被拆分成机器能理解的半结构化或结构化文本；要么就是通过深度学习，使机器越来越智能，这样就不必对非结构化文本进行人工拆分，这当然是科技发展的大势所趋。在这个大趋势中，做为研究者，我们也应以积极的态度对不同文体展开深入研究。在“图像检索”领域，主要探讨两个问题：一是疑难文字的 OCR，二是基于文本与图像内容相结合的图像检索。后一种虽是一种理想状态，但却是未来的发展方向和趋势。