传统数据与大数据的融合

刘 涛雄和徐晓飞(2015)¨。使用了PC端百度搜索指数 讨论了互联网搜索行为对宏观经济预测产生的影 响，得出非结构化数据有助于提高预测宏观经济的 准确性，但不能替代政府统计数据，且要使用合适的 预测模型，由此提出了“两步法”，即先使用政府统 计数据进行初步预测，再加入百度搜索指数。

本文选取的在线大数据不仅 包括了代表Pc端和移动端网民对碳价关心程度的 百度整体搜索指数，还包括了反映网络媒体对碳价 关注程度的媒体指数，力求用更全面的数据信息预 测碳价。在借鉴“两步法”思想的同时，不限制新增 变量的个数，旨在选出最优模型。

媒体数据

非结构化信息采用了百度整体搜索指数和媒体指数。百度搜索指数是以网民在百度的搜索量为数 据基础，以关键词为统计对象，计算出的每个关键词在百度网页搜索中搜索频次的加权和。百度媒体指数是以各大互联网媒体报道的新闻中，与关键词相关的、被百度新闻频道收录的数量，采用新闻标题包 含关键词的统计标准，与百度搜索指数无直接关系，两者均可在百度指数网站获得。虽然经过标准化处理，但本质上是非结构化的，属于非结构化数据¨

交易数据

随着互联网应用的不断普及，证券、期货、大宗商品等金融交易已高度网络 化，基于这些在线交易所产生的数据具有很强的挖掘价值。在证券交易行业，券 商对于大数据的重要性已经有了高度共识，特别是在市场行情预测等方面，起步 非常早。在期货市场以及大宗商品等场外交易市场，对交易数据的综合分析也得 到越来越多企业的重视。我国的大宗商品场外交易市场经过十多年的发展，已形 成一个新兴行业。在一些重点领域，如棉花、钢材等，大宗商品交易市场的交易 量已经占到全国总交易量的一半以上。一些互联网公司同样聚集了该领域的大量 数据，如金网安泰公司为全国 370 多家大宗商品交易市场（约占到国内近 2/3 合 规市场）提供了大宗商品交易平台软件，从而积累了大量原始交易数据。对这些 数据源的后续挖掘和分析，对于宏观经济监测预测也具有重要意义。