北京大学数字金融研究中心
Institute of Digital Finance, Peking University

北京大学金融科技情绪指数
（2017 年 9 月）

王靖一[①] [②]

摘要：互联网金融，自其作为一个概念被提出，其发展便伴随着媒体的不同声音；近年来，随着“互联网金融”这一名词逐渐污名化，金融科技，或 FinTech 因其更好的概念包装与国际沟通遍历，逐渐成为业界与媒体描述的主要表达。为了能够科学、准确、量化地刻画互联其发展变化脉络，我们利用逾 1700 万条新闻全文数据，借助自然语言处理、机器学习等方法，编制了一套覆盖 2013 年 1 月至 2017 年 9 月的互联网金融情绪指数，指数包含了对于金融科技整体与 P2P 网络借贷、互联网支付等 12 个子类的关注度与正负情感的度量。指数表明，金融科技的整体关注情况呈现出波动上扬的趋势，而对其整体的正负情感态度，则振动较为剧烈。而金融科技各子类，在关注程度与正负情感态度上，则有着较大分化。

关键词：金融科技、情绪指数、主题模型、词向量模型

2017 年 11 月

[①] 王靖一，北京大学国家发展研究院博士研究生
[②] 本课题为北京大学数字金融研究中心课题《北京大学互联网金融情绪指数》资助下的阶段性成果；作者感谢黄益平、沈艳、黄卓、谢绚丽、孔涛、王海明、郭峰、鄂维南、窦笑添、任洁、王旭、曹琦、杨雨成、予象、周伊敏在指数编制过程中的建议与帮助。

目录

图表目录

1.引言

互联网金融，自作为一个独立概念，在四十人论坛 2012 年年会被谢平提出，其发展过程始终伴随着来自不同源头、秉持不同态度的声音。互联网金融得益于信息技术，其发展速度远超传统金融，据北京大学互联网金融发展指数度量，在 2014 年 1 月至 2016 年 3 月期间，增长了 4.3 倍（郭峰等（2016））；而同时，截止至 2015 年 11 月，累计爆发问题的 P2P 网贷平台较 2012 年之前的数字增长了 72.31 倍,而《网络借贷信息中介机构业务活动管理暂行办法》中提出的监管框架似不能有效解决 P2P 网贷所 面临的问题（黄益平等（2016））。2015 年 12 月至 2017 年 9 月间，据网贷之家数据统计，爆发问题或停业退出平台数量达到 2442 家，是之前累积问题平台数量的 1.65 倍。这些负面新闻的密集出现，则令公众对于互联网金融产生了质疑，甚至大有“污名化”之势。另一方面，曾建光（2015）的研究则发现，公众可以有效地通过信息化手段，感知网络安全风险，而公众对于风险的规避，则影响了互联网金融资产的价格。互联网金融的发展情况，与对应的新闻报道的舆论情绪间的相关分析，对于学术界、政府与业界均有较高的价值。

伴随本土概念“互联网金融”遭遇污名化，舶来词 FinTech 及其中文译名金融科技受到业界与媒体的宠爱。虽然指数算法具有较强的稳定性，在 2016 年 6 月的初版中，我们的算法便能够将金融科技或 FinTech 的最近义词识别为互联网金融；但是由于 FinTech 一词在国际交流上的便利，以及规避互联网金融伴随的天然化负向情绪，自 2017 年 9 月，本指数更名为金融科技指数。然而，笔者仍认为，在当前中国，金融科技与互联网金融并无本质的区别，只是一件事物出于不同原因获得的两个名字。

然而，截至目前，虽然金融科技发展情况有大量的结构化数据与指数可以度量，但对于新闻报道这种非结构化信息，尚无一个有效的量化分析。故此，我们编制了本北京大学金融科技情绪指数（下简称情绪指数），以资后续研究。

为使所得数据具有足够的覆盖广度与稳健性，我们收集了 2013 年 1 月 1 日，至 2017 年 9 月 30 日，1702 万余条新闻数据，原始数据规模逾 700GB，数据来源为和讯网[①]。虽然和讯网自身对于新闻有所分类，并且“互联网金融”单独成类，

[①] 作者本人与所在单位与和讯网无合作关系或直接利益关系，选择其作为数据来源，是综合考虑新闻覆盖

但数据收集整理过程中我们发现，这一分类存在着较大的遗漏，例如在 2013 年 10 月 25 日之前，“互联网金融”类目下不存在任何新闻，我们分析中的一个重要环节，便是重新在全部新闻中寻找“互联网金融”相关新闻，并将其归类到金融科技几个子类之中。

分析方法上，我们主要使用了 Baker et al. (2015)构建经济政策不确定性指数时使用的关键词查找法，自然语言处理中较为经典的隐含狄利克雷分布（LDA）和层次狄利克雷过程。综合使用这三种算法，我们在数据处理能力和算法精度间找到一个较为适宜的平衡。随着本文所使用开源工具 Gensim[①]的发展，未来还会引入动态主题模型（DTM）。

文章后续安排如下，第二节介绍指数的指标构建方法，第三节汇报指数的主要结果，并做出初步分析。

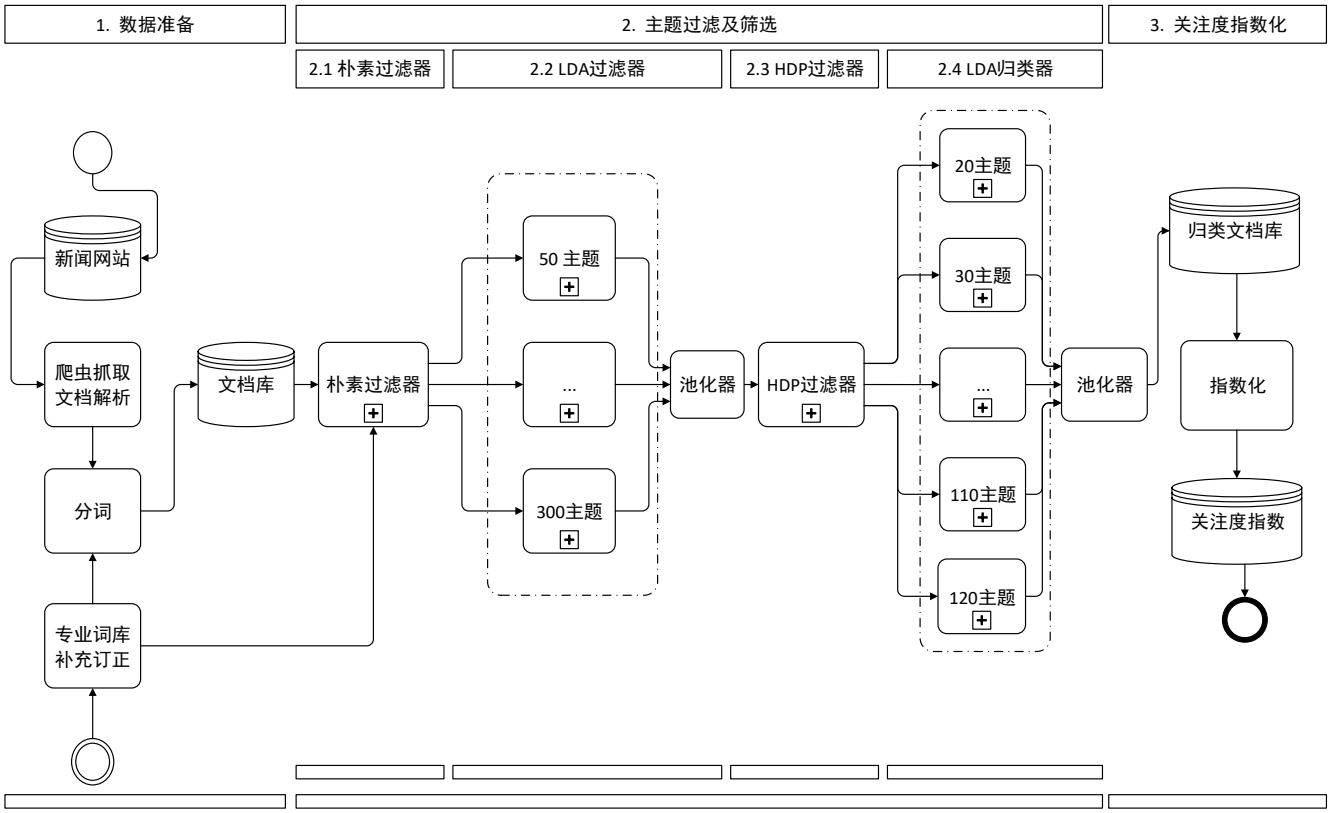
2.关注度指数构建方法

情绪指数的目的，是以度量金融科技及其重要组成部分，在不同时期的受关注情况；同时，描绘新闻媒体对于它们的正负评价情况。那么，构建工作其实可以分为三个步骤，第一步，从 1700 万条新闻中，寻找金融科技相关的新闻；第二步，将这些新闻归类至金融科技的不同子类中；第三步，构建对新闻的正负情感的量化描述。

其中，前两步对于指数的正确性有着很重要的影响，在逾 1700 万各色新闻中寻找金融科技（互联网金融）这样一个不算主流的主题，并进一步区分至各个子主题，要求算法一方面能够高效处理大量数据，另一方面在一定规模的数据量的计算中，收敛至一个较为精确的结果，为此我们设计了一套如图表 1 所示的流程。

广度、报道专业性、收集处理可行性的结果，数据获得方式为友好、无欺诈的爬虫。作者仅保证对于和讯网目前公开、正常网页的完整准确采集，而对于和讯网收集过程中的完整、准确则无法做出相应承诺。采集时间为 2016 年 6-7 月，此间部分过去时间的网页已无法正常访问，对于这部分网页的缺失原因与缺失带来的影响，作者无法准确度量，但缺失数量小于样本总体的 0.1%。

① http://radimrehurek.com/gensim/



图表 1 关注度指数计算流程图

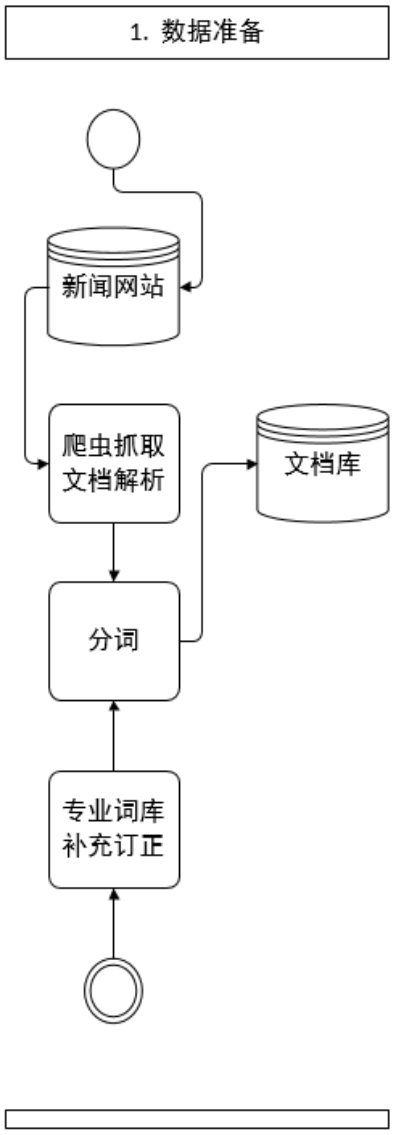
按照功能划分，该流程可以被视为三个部分：1.对数据进行准备，从网页抓取到生成分词完毕的待处理数据；2.对数据进行主题过滤及其结果的筛选，这一部分为该流程的核心，完成了识别文章主题并正确归类的任务；3.对归类完毕的文档进行指数化，刻画关注度指数。下文将具体地对每个流程进行详细介绍。

2.1 数据准备

本部分完成了从采集网络数据到构建筛选用文档数据库的工作，主要包括新闻网站的选取，爬虫抓取与文档解析，金融科技专业词库补充订正，分词几个部分。

新闻网站的选取，需要综合考察三个方面，第一是网站所覆盖的广度，是否能够较为全面的将媒体的声音容纳；第二是网站的专业性，我们不希望数据库中充斥着大量重复、无用的报道，特别是这些报道集中在那些我们不关注的领域，比如娱乐、体育；第三是网站的数据易抓取和解析性，对爬虫友好、网页模板清晰统一的网站，可以节约我们大量时间与计算资源。综合以上三点，我们最终选取了和讯网作为数据来源，这里需要再次强调的是，作者本人和所在单位与和讯网并无任何合作关系，我们做出这样的选择，是基于上述三个标准的最优选择，而我们所能保证的也只是在数据采集期间，和讯网可采集的数据的完整性，而对于和讯网是否全面包含所有金融科技相关新闻，我们并不能做出相关推断。

爬虫抓取和文档解析并无特别的地方，且受网站结构、网页模板限制没有什么可扩展的余地，所以这部分省略。唯一需要说明的是，因为和讯网对于爬虫有较高的包容度，所以我们并不需要进行欺诈等“灰色操作”，采集时程序有所限速，并没有直接证据表明我



图表 2 数据准备阶段流程