北京大学互联网金融研究中心
Institute of Internet Finance, Peking University

北京大学互联网金融情绪指数
（2016 年 9 月）

王靖一[①] 窦笑添[②] [③]

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。为了能够科学、准确、量化地刻画互联网金融情绪发展变化的脉络，我们利用近 1500 万条新闻全文数据，借助自然语言处理、深度学习等方法，编制了一套覆盖 2013 年 1 月至 2016 年 9 月的互联网金融情绪指数，指数包含了对于互联网金融整体与 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，互联网金融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。而互联网金融各子类，在关注程度与正负情感态度上，则有着较大分化。

关键词：互联网金融、情绪指数、主题模型、词向量模型

2016 年 10 月

① 王靖一，北京大学国家发展研究院博士研究生
② 窦笑添，康奈尔大学硕士研究生，北京大学互联网金融研究中心助研
③ 本课题为北京大学互联网金融研究中心课题《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、任洁、王旭、曹琦、杨雨成、予象、周伊敏、王勋、苟琴、纪洋、傅秋子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））。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。

然而，截至目前，虽然互联网金融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互联网金融情绪指数（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 2016 年 9 月 30 日，1477 万余条新闻数据，原始数据规模逾 500GB，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，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，并将其归类到互联网金融几个子类之中。

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

[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%。

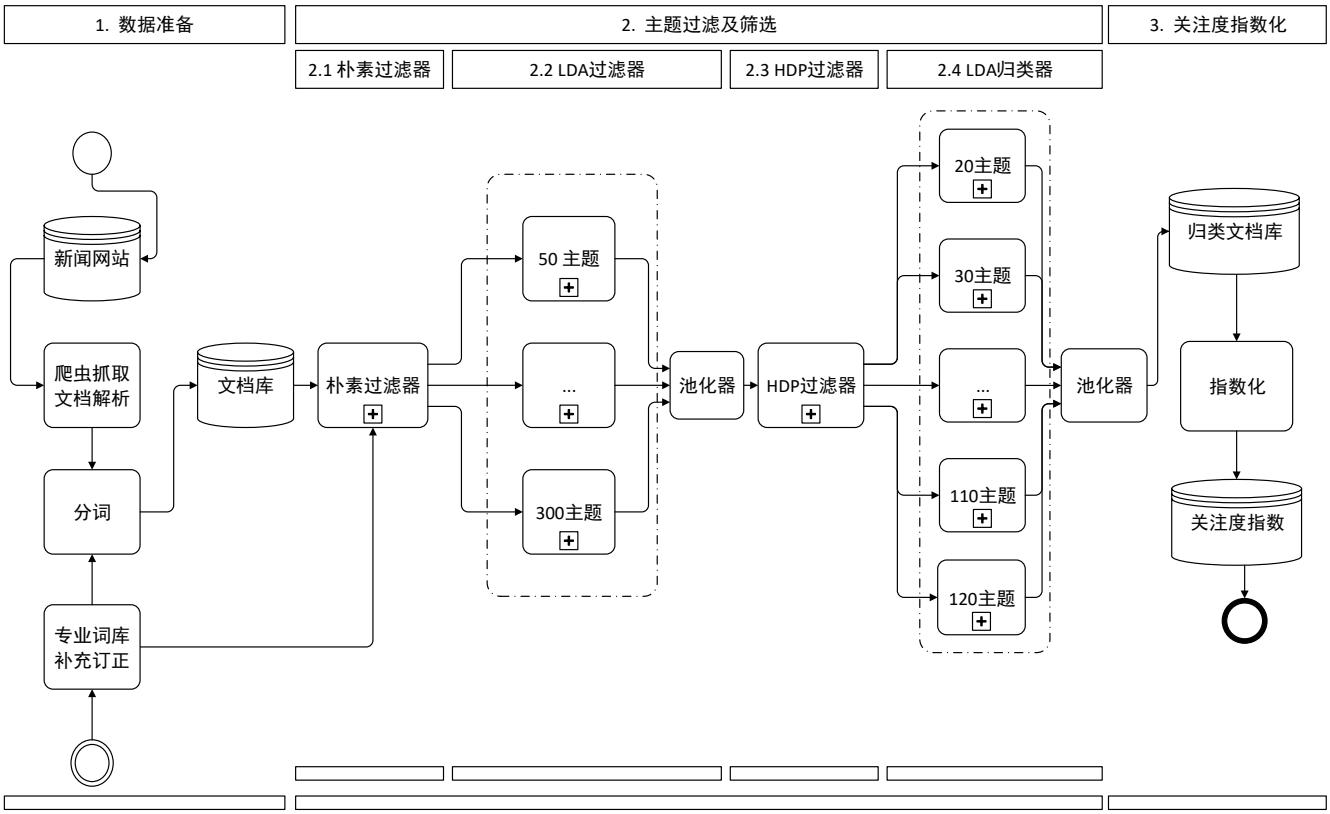
找到一个较为适宜的平衡。随着本文所使用开源工具 Gensim[①]的发展，未来还会引入动态主题模型（DTM）。

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，是以度量互联网金融及其重要组成部分，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 1400 万条新闻中，寻找互联网金融相关的新闻；第二步，将这些新闻归类至互联网金融的不同子类中；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，在近 1500 万各色新闻中寻找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。



图表 1 关注度指数计算流程图

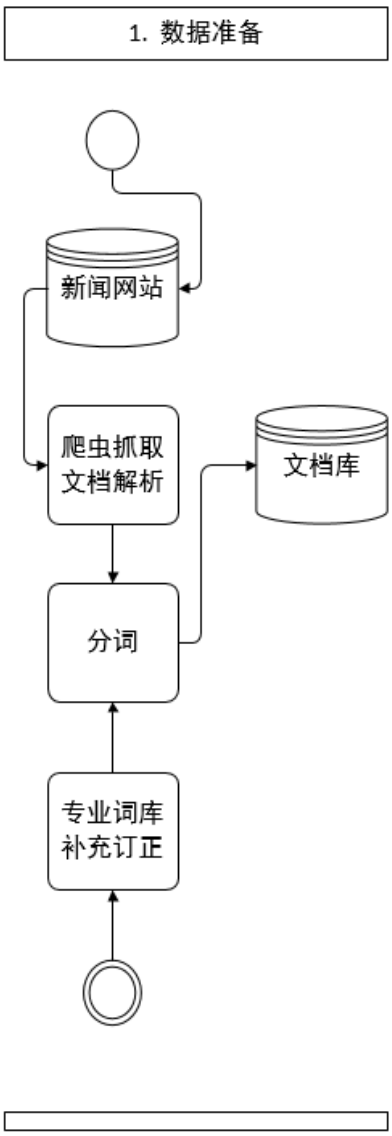
① http://radimrehurek.com/gensim/

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，互联网金融专业词库补充订正，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互联网金融相关新闻，我们并不能做出相关推断。



图表 2 数据准备阶段流程

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们“有礼貌的爬虫”影响到了网站的正常运作。文

档解析则使用 Beautifulsoup[①]，在本地对网页进行操作。

分词算法，我们使用较为准确易用的 jieba 分词[②]。专业词库的打造及对部分词义的补充修订，则是数据准备阶段较具特色的一环。分词一直是中文自然语言处理相当重要的一环，虽然算法经历了多年发展，已经相当成熟，但如果我们能够先验地给出一个分词词典，仍能极大提高算法效率与准确度。而互联网金融作为一个新兴的、发展迅速的领域，既有分词词典并不能完整包含；加之，分词词典往往包含着该词出现的先验概率以解决中文分词中的歧义问题，但是我们针对的文档集中在互联网及金融领域，某些关键词出现概率应当有所调节。词库扩充中，我们首先增补了互联网子类和其主要从业者的公司名和代表人物名字，例如，我们的自定义词库包含了黄益平等（2016）中纳入考量的 3600 余家 P2P 网贷平台平台名、公司名；其次，我们继而在一个较为纯净、稠密的互联网金融语料库——《互联网金融十二讲》使用 jieba 分词的搜索引擎模式，自主识别新词，对分词结果进行逐一人工筛查，判别其是否属于互联网金融专业分词词典、正负情感词词典；最后，在前两步的基础上，我们对 2013 年 1 月至 2016 年 6 月和讯网归类为互联网金融的文章进行切词，并对出现频率大于 20 的词汇进行人工筛查，判别是否应纳入分词专业词库与正负情感词词典。

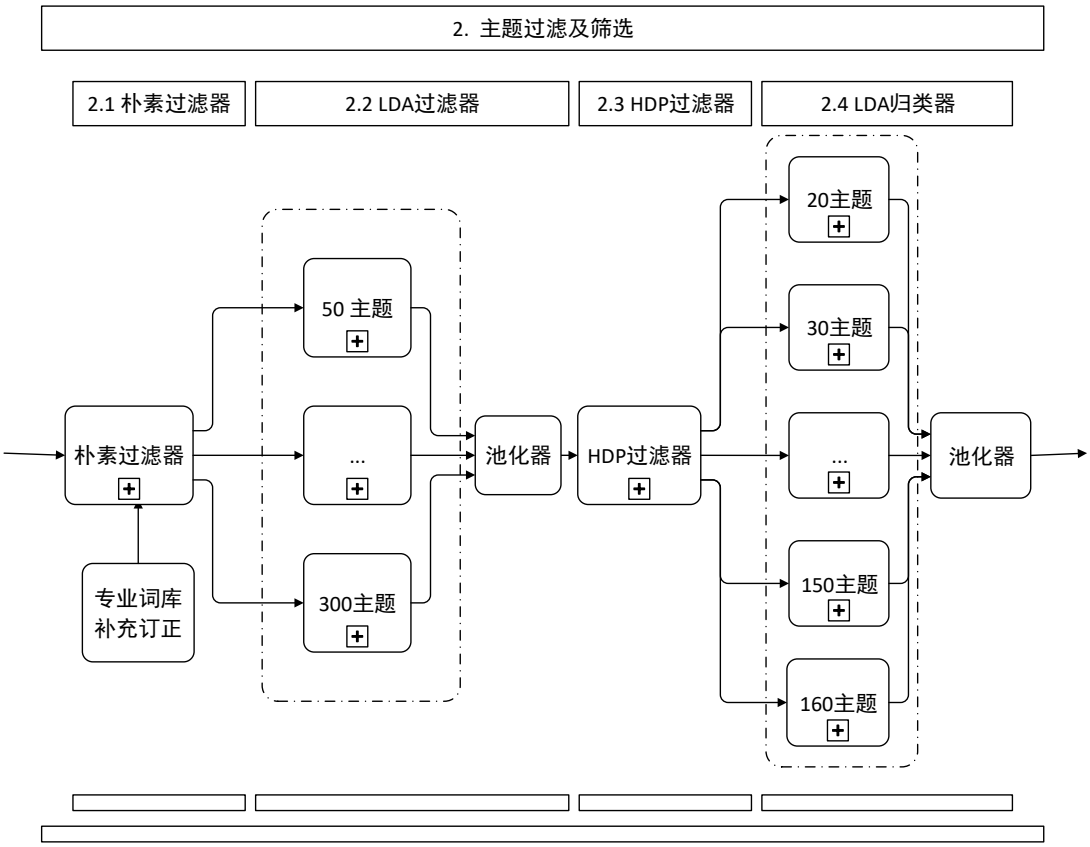
如是，我们完成了数据准备阶段，构建了用于后续分析的文档库。

2.2 主题过滤及筛选

主题过滤及筛选为整体流程中最为重要的一环，其输入为数据准备阶段完成的文档库，而输出则是分类完毕的文档，这个环节实际上决定了那些文档是否属于互联网金融范围，以及具体属于互联网金融的那个子领域。主要可以细分为四个子环节：朴素过滤器、LDA 过滤器、HDP 过滤器、LDA 归类器，类似于其命名方式，前三个环节主要完成了从文档集中过滤出可能与互联网金融相关的子集，在前三部得到一个较为纯净的文档集的基础上，最后一步归类器计算文档归属于某个主题的对应概率。

其主要流程如图表 3 所示：

① https://www.crummy.com/software/BeautifulSoup/
② https://github.com/fxsjy/jieba



图表 3 主题过滤及筛选流程

2.2.1 朴素过滤器

朴素过滤器为本流程的第一个环节，其所面对的处理量是较大的，近 1500 万份文档，如果使用后文的 LDA 主题模型进行处理，一方面计算资源和时间的消耗是很惊人的，另一方面，如此大量的语料库[①]，所应预设的包含主题也是相当大的，即便 LDA 模型能够计算成功，对其千余乃至数千主题进行人工筛选，也是惊人的工作量。

所以我们借鉴了 Baker et al.(2015)的方法，我们定义了一系列的关键词，并对文档集中的每篇文档进行筛查，如果这篇文章没有包含任何关键词，那我们便认为这篇文章与互联网金融有关的概率很低，可以从文档集中剔除，以便于在后续流程中使用一些更为复杂的方法。这些关键词包含一些非常宽泛的词汇，比如“金融”、“互联网”，所以这个筛选实际上是较为宽松的。需要指出的是，Baker et al.(2015)的筛选逻辑是同时包含四个列表里的关键词，便识别为该文档属于他们感兴趣的主题；而我们的逻辑是，如果一篇文章没有包含任何我们列表里的

[①] 语料库为自然语言处理常用术语，在这里可认为是所有文档的集合

关键词，便不属于我们感兴趣的主题；产生这样的不同，主要是由于和讯网相较于 Baker et al.(2015)使用的南华早报包含更多更杂的新闻种类，所以我们只能在这一步完成一些粗筛，而更加细致、可信的筛选则由后续部分完成。

2.2.2 LDA 过滤器

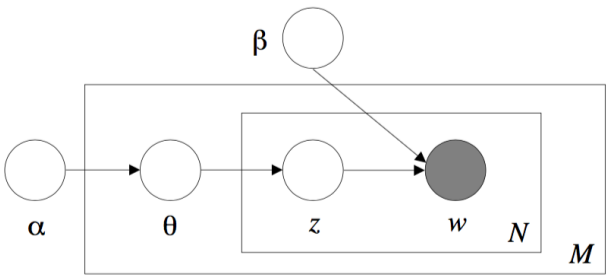
LDA 过滤器实际上是将上一步朴素过滤器的输出，输入至一系列预设主题数不同的 LDA 模型之中，继而将这些不同主题的 LDA 模型的输出输入到一个池化（pooling）器中，池化器的作用实际上是综合每个 LDA 模型的结果，给出每篇文章的主题归属。

为了阐释 LDA 过滤器有效的原因，我们必须对 LDA：隐狄利克雷分布进行一个简单的介绍。LDA 是主题模型中最为经典的算法之一，而主题模型的目的则是将语料库中一系列的文档，归结为不同的主题，而每个主题则可以用一系列词语表述，它认为词语、文档、主题三者之间存在一个如下的概率联系：

$$\mathrm{p}(\text{词语}|\text{文档}) = \sum_{\text{主题}} p(\text{词语}|\text{主题}) \times \mathrm{p}(\text{主题}|\text{文档})$$

这个公式实际上陈述的了一个朴素的推定，一篇文档包含很多词语这个直观事实，背后隐含着一篇文章包含多个主题，而每个主题则又对应着若干词语，则“词语|文档”的概率分布其实是“词语|主题”、“主题|文档”两个分布的联合概率。所以，当有足够多的“词语|文档”的训练集，我们便能够推算出隐含在其中的主题。

Blei et al.(2003)提出的 LDA 是对这一思想的一个很好实践，其主要算法的主要流程使用“盘子表示法”（plate notation）如图表 4



图表 4 LDA 模型示意[①]

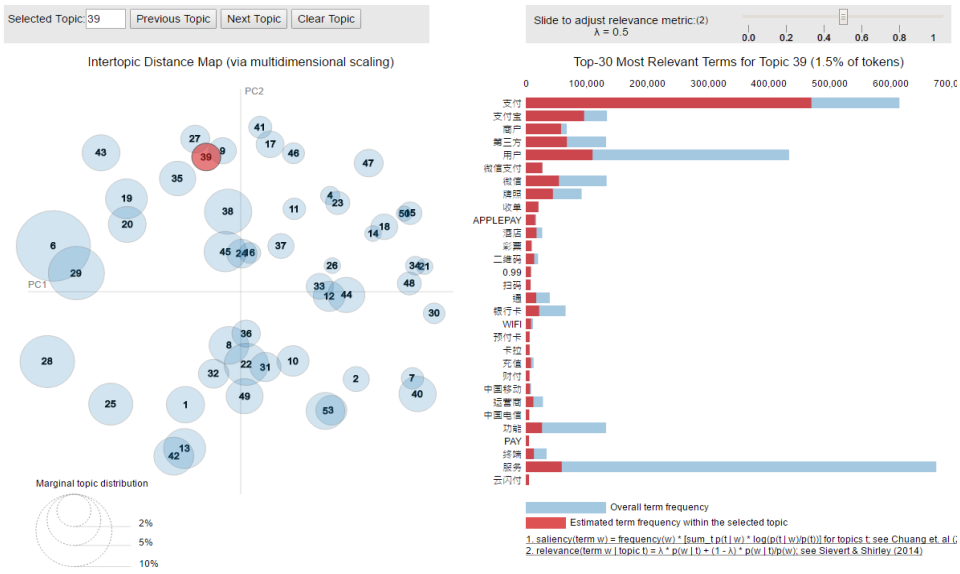
① 图引自 Blei et al.(2003)

盘子表示法是图模型的一种常用表示法，图表 4 中涉及的相关语法有两单需要阐明，其一，两个矩形框代表着重复，外层的表示对于语料库中 M 篇文档重复使用框外元素，内层的则表示对每篇文档中 N 个单词套用框外元素；其二，图中阴影表示的圆圈表示可见（每篇文章中每个词语 w 是可见的），而空白的圆圈则表示隐含（一系列参数）

图表 4 中各个参数的含义如下，w 是每篇文档中的每个词语，z 为 w 所属于的主题，θ则是在文档层面，对于每个词的主题 z 的分布，而α、β分别是语料库层面，对θ和 w 的先验狄利克雷分布的参数。至于狄利克雷分布，这里不做展开，我们只需要知道它是一族连续多元概率分布，在贝叶斯统计中常被用来作为分类分布（categorical distribution）和多元正态分布的先验。所以狄利克雷分布在这里起到一个寻找合理先验起点的作用。对于一篇有 N 个词$\mathbf{w} = (w_1, w_2 \dots w_N)$的文章而言，其主题θ的概率分布为：

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta)$$

这里的算法略显晦涩，而其结果却较为直观，我们只需要给出一个先验的主题数量 T，即我们认为语料库中的诸多文章应当归属于多少个主题，之后算法收敛后会返回这 T 个主题，每个主题用一系列关键词及其代表当前主题的概率。得益于 pyLDAvis[①]，我们能够相当直观地了解到每个主题的含义及相关性。



图表 5 一个 LDA 模型的结果示例

① https://github.com/bmabey/pyLDAvis

图表 5 即为 LDA 过滤器阶段，我们使用的 50 主题 LDA 得出的语料库的 50 个主题，每个主题即是图表 5 左侧的一个圆圈，这五十个主题按照前两个维度的主成分在二维坐标轴分布[①]，红色的为当前选择的主题，这时画面右侧出现了这个主题对应的一系列关键词，和每个关键词对于该主题形成的影响力。LDAvis[②]对于 LDA 的直接输出结果一大改进便是，有了更为科学的影响力算法：LDA 对于每个主题会直接给出一系列代表主题词和概率，但是这里主题词的概率收到其在全语料库的分布的强烈正向影响，即一个词在全语料库出现的越多，它越可能成为描述某个主题的主题词。所以 LDAvis 的作者 Sievert & Shirley(2014)在 Chuang et al.(2012)的基础上给出了一种更为避免这一影响的算法，并给出了一个可调节的超参数λ，λ越靠近 1，便越受到改词在在语料库的总体概率影响，越靠近 0，则越剥离了这种影响。在该篇文章中，作者建议使用$\lambda = 0.5$，本文后续模型如无特殊说明，均按照这一数值进行筛选。

接下来，对于每一个主题，我们进行手工筛选，结合每个主题对应的关键词和其各自的影响力，标定每个主题是否属于互联网金融相关；加之我们又能取得每一篇文档归属于各个主题的概率，这样，我们实际上通过 LDA 确定了一篇文章从属于互联网金融领域的概率。

至此，我们完成了对于 LDA 的介绍。LDA 过滤器则是使用主题数分别为 30、50、100、150、200、300 的 LDA 进行计算，并将它们输出的结果进行池化。所谓池化，这里是指对不同主题数 LDA 的结果进行综合汇总：对于每一篇文章，我们有其在不同主题数 LDA 的主题归属，池化将这些不同模型进行综合，给出这篇文章能否进入下一阶段的判断。这里的池化策略为，对每一篇文章，其是否能进入下一阶段，取决于它在各个主题数的 LDA 模型中的结果，是否存在某 LDA 的结果，概率最大的两个主题有否属于互联网金融领域。逻辑表达式为[③]：

$$\bigcup_{n}(\text{最高主题 } \epsilon T_{IF,n} \cup \text{次高主题 } \epsilon T_{IF,n})$$

① 主成分分析是一种有效、常用的降维手段，可以简单地将每个主题看做一个高维向量，我们分析所有这些向量，找到两个其投影方差最大的维度，然后将这些高维向量投影在这两维组成的平面上，可以粗糙地认为，两个圆如果离得越近，那么它们对应的主题越有更大概率相似。比如上图中 9 号圆圈对应的是征信，而 27 号圆圈对应的则是社交营销。

② https://github.com/cpsievert/LDAvis LDAvis 的最初版本为 R 语言，前文提及的 pyLDAvis 为 LDAvis 的 python 语言版本。本文所实际使用的是 pyLDAvis，但相关算法的提出者实际上是 LDAvis 的作者

③ 特殊的，对于 n=30，即 30 主题的 LDA 模型，我们只取其最高主题

其中，n为 LDA 的主题数，$T_{IF,n}$即为主题数为n的 LDA 计算出所有属于互联网金融的主题中。

至此我们完成了 LDA 过滤器的介绍。

2.2.3 讨论：为什么不将 LDA 的结果直接输出作为关注度指数

在进行后续介绍之前，我们先要回答一个问题，既然 LDA 模型能够将不同文档分配至不同的主题，为什么不将此时的输出作为最终的关注度指数结果？

之所以将这个问题单独列为一个小节，是因为这个问题的答案对整个流程构成有着很重要的指导意义：为什么要叠加多层过滤器？为什么要对于过滤器，筛选的尺度较为宽松？

事实上，本指数的早期版本（发布于外滩峰会）就是将这部分 150 主题 LDA 输出直接作为指数。但之后的复核中发现，因为只经历过一个朴素筛选器，语料库中存在着大量的和互联网金融不相干的文档，这些文档的存在，让 LDA 的主题数选择存在了较大问题：如果主题数太少，则无法识别互联网金融的不同子类——比如 P2P 网贷和支付混在了一起，毕竟这二者的关系要比它们和黄金的关系近的多——而如果主题数太多，则会造成最后主题输出的结果过于稀疏，反映在指数上的结果便是许多主题的指数在大段时间缺少变化，维持在低位。

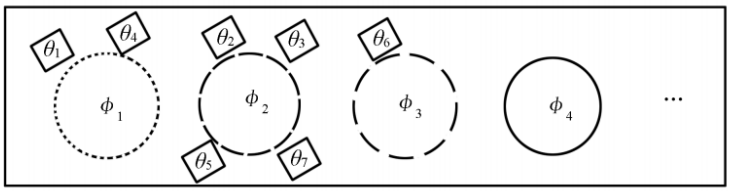
故而，我们对方法进行了一系列改进，首先，叠加多层过滤器，让最后参与主题归类的文档尽可能地纯粹，这样我们通过叠加过滤器的方式，层层筛；其次，由于过滤器叠加层数的增多，为了保证不错杀过多文档，我们将池化条件放松。最后，在主题归类环节，我们依然采用了类似的池化策略，使得主题不单一地为某特定主题数的 LDA 决定。

2.2.4 HDP 过滤器介绍

LDA 的一个局限性就是，我们需要人为地给出一个主题数量，而不同主题数量的选择会对主题的生成和文档的归类产生较强的影响。一种克服它的方法便是 Teh et al.(2006)提出的 HDP：层次狄利克雷过程，它可以自主地选择语料库应被归结的主题数量。这种方法可以帮助我们进一步地过滤文档。

为了阐明其作用原理，我们需要先对于 DP，即狄利克雷过程进行简介，DP 是一种应用于非参数贝叶斯模型中的随机过程，由 Ferguson(1973)给出定义并

证明其以概率 1 离散。关于其构造过程，有多种描述，其中一种较为形象的是，中国餐厅过程（Chinese Restaurant Process）



图表 6 中国餐厅过程[①]

想象一个有无限张桌子的中国餐厅，每个桌子上有一道菜，坐在同一张桌子上的客人共享桌子上的菜。对于某个新来的客人θ_i，他落座桌子ϕ_k的概率正比于该桌子业已存在的人数m_k，以正比于α_0的概率增加一个新的桌子，即总桌子数量 K 增加 1。而$\phi_K \sim G_0, \theta_i = \phi_k$。由此我们完成了以集中参数为$\alpha_0$的基分布$G_0$的构建，而此时餐厅里每张有人坐的桌子，即可被认为是在基分布G_0的基础上构建的狄利克雷过程$G \sim DP(\alpha_0, G_0)$。

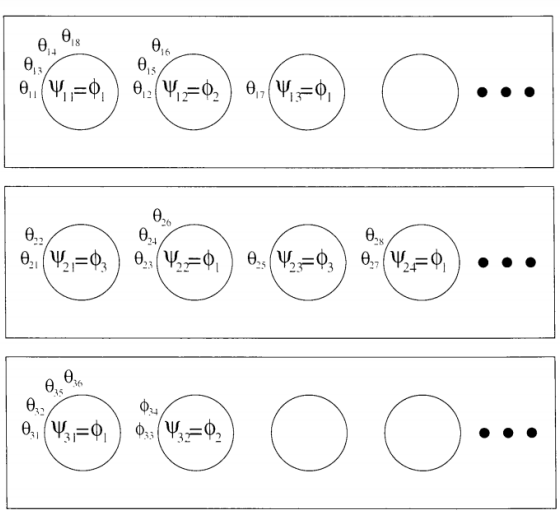
从上面的过程中我们不难发现，DP 过程天然地具有将一系列数据θ_i聚类为若干集合ϕ_k的趋势，而且，这一聚类过程是完全非参的，我们不需要预设这群人要占满多少桌。一个不甚严谨的类比，LDA 模型相当于设定了桌子数量的中国餐厅，即使有人不爱吃某桌子上的菜，也因为没有办法开新桌而被迫坐在不中意的的桌子上（预设主题数小于语料库需要的）同理，当设定的桌子太多时，本来兴趣相投的顾客也会被强拆成几桌，而且他们点的菜十分类似（主题表达能力区分度较低。）

而 HDP，如其名字所示，是对于层次化的表达，扩展前文类比于中国餐厅的构造过程，现在我们经营一家中国餐厅集团，即我们有多家餐厅，这些餐厅共享菜谱，每家餐厅依然有无限张桌子，依然有很多顾客光顾每家店，坐在不同的桌子上。但这时，坐在一张桌子上的第一个人依然可以点菜谱上的任意一道菜，即使这道菜已经在相同或者不同的餐厅的某张桌子上点过。

正式地，θ_{ji}为第 j 家店光顾的第 i 个顾客；ψ_{jt}表示第 j 家店第 t 张桌子上（由第一个人）点的菜；各家店菜单统一，每道菜用ϕ_k表示。每个人坐在某张餐桌上的概率正比于这张餐桌上已经落座的人数；而每个桌子所点的菜正比于

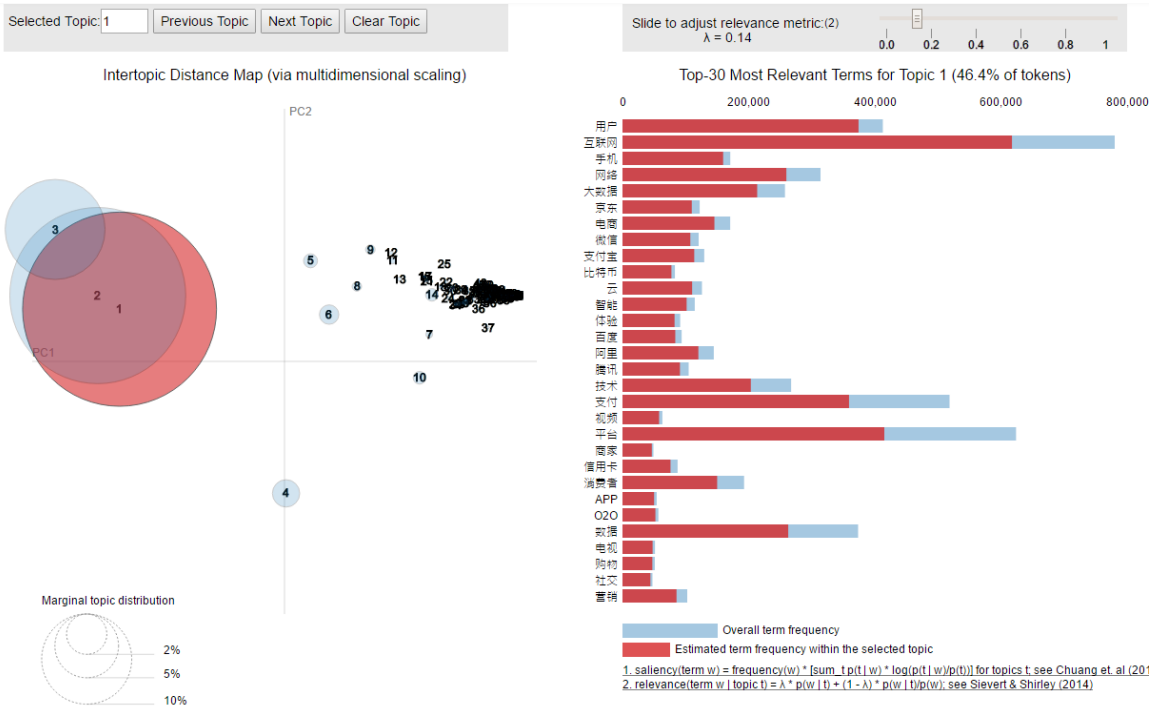
[①] 图片引自周建英等（2011）

已经点这道菜的桌子的数量（不论其处于哪个餐厅）。



图表 7 中国餐厅集团过程[①]

我们使用的算法，实际上使用的是 Wang et al.(2011)的在线（online）学习版本，其在对每篇文章所属主题进行推断时，实际上是将 HDP 的结果，近似于 LDA，所以如图表 8 所示，我们看见了若干极度靠近的主题（同时他们包含了大部分参与计算的文档）：



图表 8 HDP 结果

我们达到了目的：朴素过滤器后的文档库依然包括繁杂的主题，而我们预设

① 图片引自 Teh et al.(2006)

的主题数不足以将它们区分出来（不能再开新的桌子，导致有不喜欢桌子上菜的客人坐在了桌子上），所以在 HDP 的结果中我们看见，游离于我们所需主题之外，散布着大量含义不明的小主题，它们与所需主题较远，而且对应文档数量较少。HDP 过滤器的用途便是清除它们。

这里需要解释两个可能产生的问题：

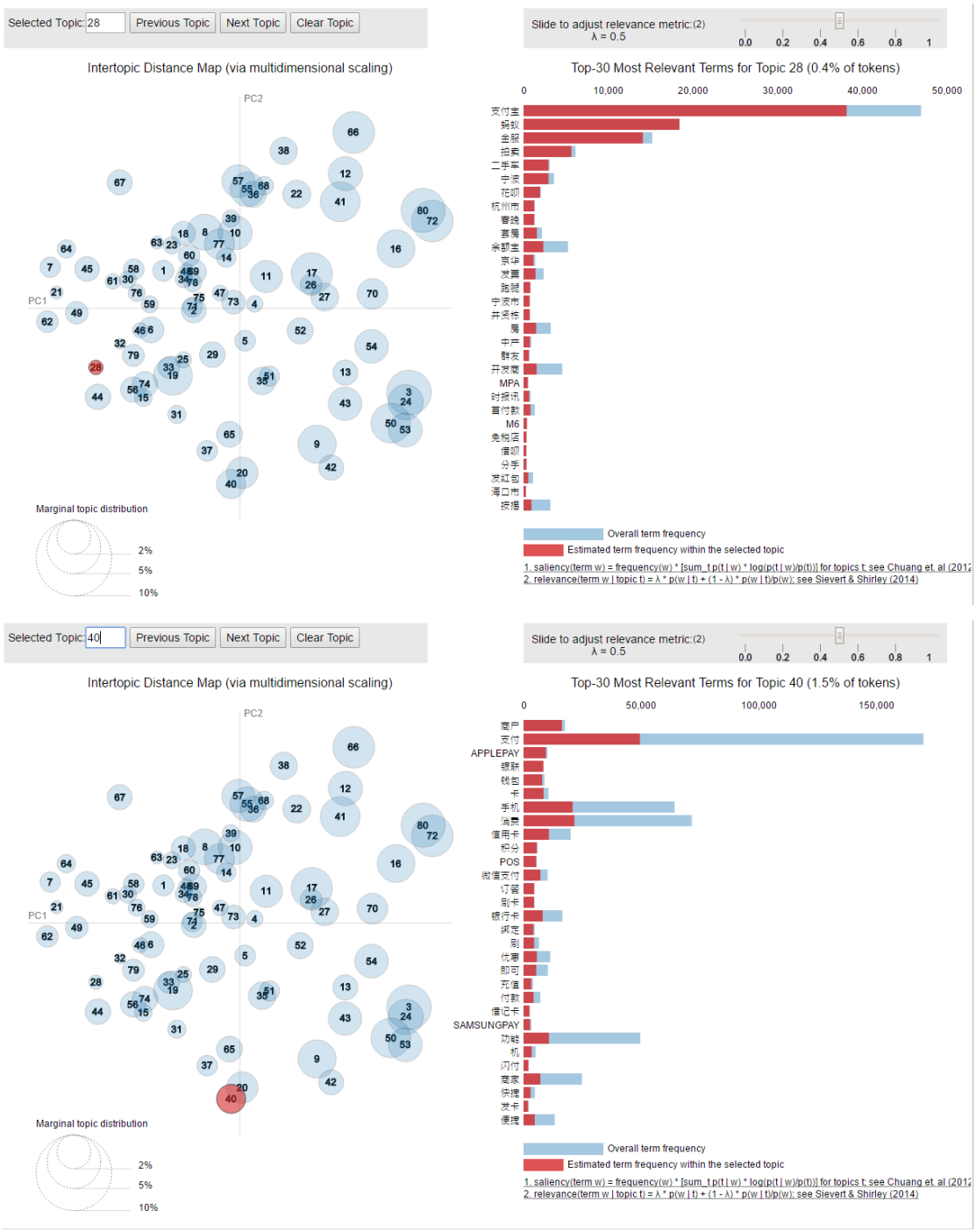
为什么不丢弃 LDA 过滤器直接使用 HDP 过滤器？主要出于两点考虑，首先是计算的复杂度。我们使用的计算环境为 i7-6800k@3.4GHz，训练一个 LDA 模型，基本上只需要不超过 12 个小时，而在计算应用于同等数据规模的 HDP 时，程序运行超过 72 个小时没有得出结果。其次，是朴素过滤器筛选过的结果仍旧包含太多的非互联网金融文章，直接使用 HDP 的效果并不理想（我们在单个月的文档集上尝试过）。

为什么不将 HDP 的结果作为输出作为关注度指数？这一点是由于我们所用程序实现的算法在推断文章主题时，实际使用的是 LDA 进行近似，所以最后输出的结果比较含糊，如图表 8 所示。

2.2.5 LDA 归类器

在经历了数道筛选之后，我们得出了一个较为纯净的互联网金融相关的文档集。之后的操作类似于 LDA 过滤器，但由于这时文档的纯净度已经大幅度提升，我们不需要尝试过多的主题数量。于是我们将主题数设置为区间[20,120]的所有偶数进行训练，但由于人工精力有限，我们只分析了整十的主题数量。对于不同数量的输出，我们使用了不同于之前 LDA 过滤器的池化策略。

因为此时，每一篇文档可以被归结为 n=20…120 个主题里，对于这些主题进行人工分析，我们又能够将不同主题归结到互联网金融的不同子类里面，这个时候，由于文档集的噪音的减少，当所选主题数量较多时，互联网金融子类被进一步拆分。如图表 9 所示，支付宝和银联、微信支付现在被划分为两个不同的主题，虽然此时微信支付和银联仍然混杂在一起，但是我们有理由相信，随着预设主题的增多，二者也将被区分开

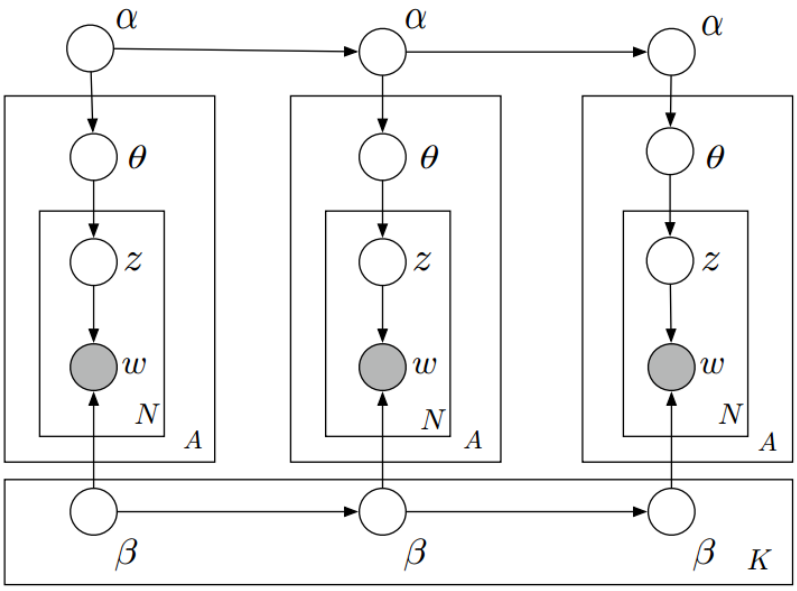


图表 9 LDA 归类器识别了支付子类下的不同主题

对于池化策略，我们使用了如下流程：首先，对于每个主题数量的 LDA 下，将每篇文章对应不同主题的概率，按照主题-子类的方式进行加总，得到文章-子类的概率分布；其次，将每篇文章在不同主题数 LDA 的文章-子类概率取最大值、求平均，即不同主题概率最大值和不同概率平均值我们都有所尝试；最后，我们便得到了每篇文章对应到不同子类的概率分布（两套，最大值和平均值）。

2.2.6 未来扩展：动态主题模型（DTM）

LDA 归类器的一个弊端是，我们将 2013 年 1 月至 2016 年 9 月 45 个月跨度的数据统一分配主题。而事实上，我们很难否认这些主题没有发生一些演进变化。但是，如果我们针对不同季度的数据，分别使用 LDA 也是不妥的，因为这一行为忽视了相同主题在不同时间段的内在联系。2013 年一季度的支付相关主题和 2015 年一季度的支付主题不应视作两个相似的主题，而应该是一个主题在不同时间点的演进。故而，比较理想的应该是使用 Blei & Laffery(2006)的动态主题模型（DTM）。



图表 10 动态主题模型[①]

DTM 是 LDA 的一种扩展，通过允许先验分布参数α,β在不同时间片进行变动，保证了不同时间段内所得主题的演进与内在一致。

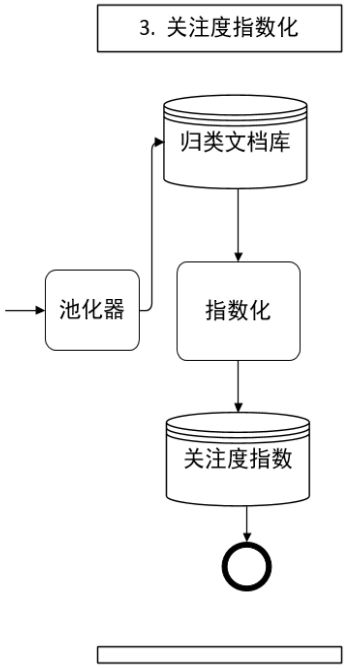
但是遗憾的是，我们所使用的开源工具 Gensim 的动态主题模型目前还处于开发阶段，我们尝试了其初期版本，训练了一个星期仍旧没有获得结果。但是鉴于这篇论文在计算机顶级会议 ICML2016 中被评为最具时间价值奖（近十年影响力最大的论文），相关工具的开发应该会进入一个较为活跃的阶段，相信在可预期的时间内，我们可以寻找到合适的工具，更新流程与结果。

2.3 关注度指数化

① 图片引自 Blei & Laffery(2006)

接下来，便是将 LDA 归类器的输出结果进行指数化，完成关注度的指数指数化；同时，生成的归类文档库，作为下一阶段正负情感分析的输入数据。

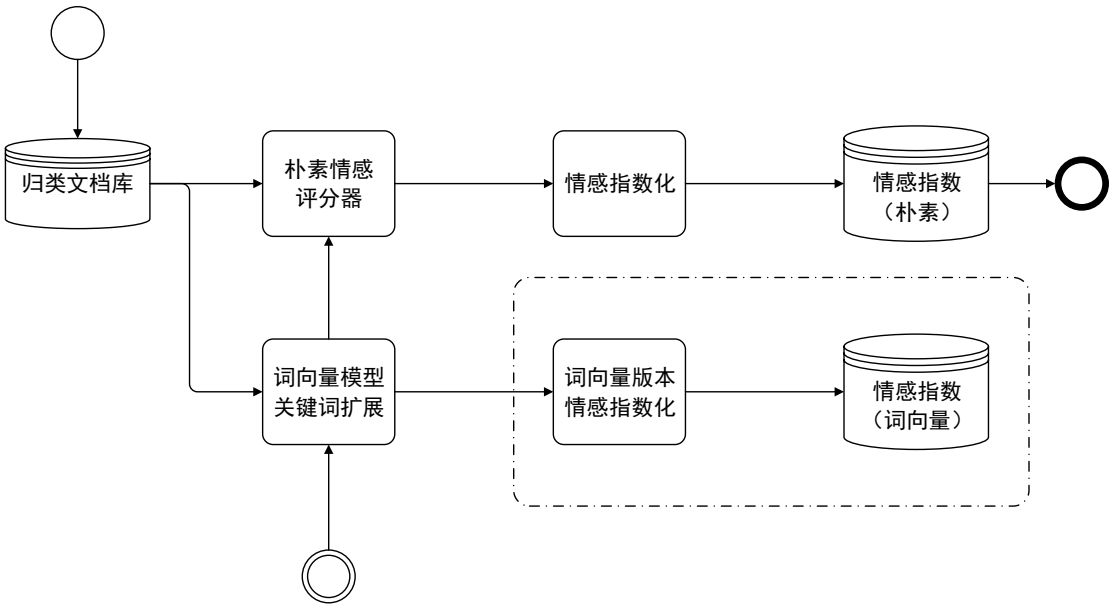
指数化的流程如图表 11 所示，首先，根据 LDA 归类器的结果，根据文档-子类的概率，将它们分配到归属于不同子类的归类文档库，准备下一阶段计算正负情感时使用；之后，根据各个文档所在的日期，将它们归属于每个子类的概率按照月份进行加总，形成当月指数，即某月互联网支付的关注度指数，即为当月所有文章归属于互联网支付这个子类的概率的加总。虽然我们有每篇文章精确至秒的发布时间，构建日度、周度频率的指数没有技术上的困难，但为了与北京大学互联网金融研究中心其他指数产品保持一致，我们暂时只公布月度指数；最后，为了剔除月度总体数量对于指数的影响，我们使用每个月的总体新闻（所有抓取的数据，不论其分类）数量进行平减，计算每个月加总值除以总体新闻数量的值，并将 2013 年 1 月总指数的结果标准化为 100，其他时间、子类的指数值做出相应调整。



图表 11 关注度指数化

3. 正负情感指数构建

在完成将所有文档划分至互联网金融不同子类，构建关注度指数之后，接下来的问题便是衡量这些文档对于互联网金融及其各个子类的正负情感态度。这个任务是较为困难的：首先，我们评价的对象是一篇完整的文章，不同于常见的衡量一个评论，文章的长度与复杂度远高于评论，而且又因为是新闻文章，同一篇里可能包含多种情感，而每种情感的表达也远不如评论来的直接强烈；其次，我们面对文档集是一个无标签的数据集，即我们没有精力去人工标定足够数量的文本，来训练一个有监督的方法；最后，类似于关注度指数编制中遇见的问题，在互联网金融的情感分析中，一方面出现了大量和行业紧密相联的专业词汇包含强烈情感，比如 P2P 网贷平台中的“自融”，包含很强的负面色彩，另一方面，一些词汇也在互联网金融语境中发生了异变，比如“雷了”是投资人常用的表达某 P2P 网贷平台爆发问题，拥有强烈的负面情感。



图表 12 情感指数构建流程

情感指数的构建流程如图表 12 所示：数据基于关注度指数生成的归类文档库；较为重要的一环是，通过词向量模型，扩展我们人工挑选的有限的正负情感关键词，从而在一定程度上解决情感词异变的问题；而对于情感的指数化，我们现在还使用的是一个朴素的方法，将在下文情感指数化一节中展开介绍；同时，受到词向量模型启发，我们也在尝试使用词向量方法，从一个新的角度度量互联网金融情感的变化，并以关键词迭代这样一个公众接受程度高的方式呈现。相较于关注度指数，我们自明正负情感的衡量还处于一个较初级的阶段，我们十分期待更为了解这一领域的学者可以提供宝贵意见或者加入研究！

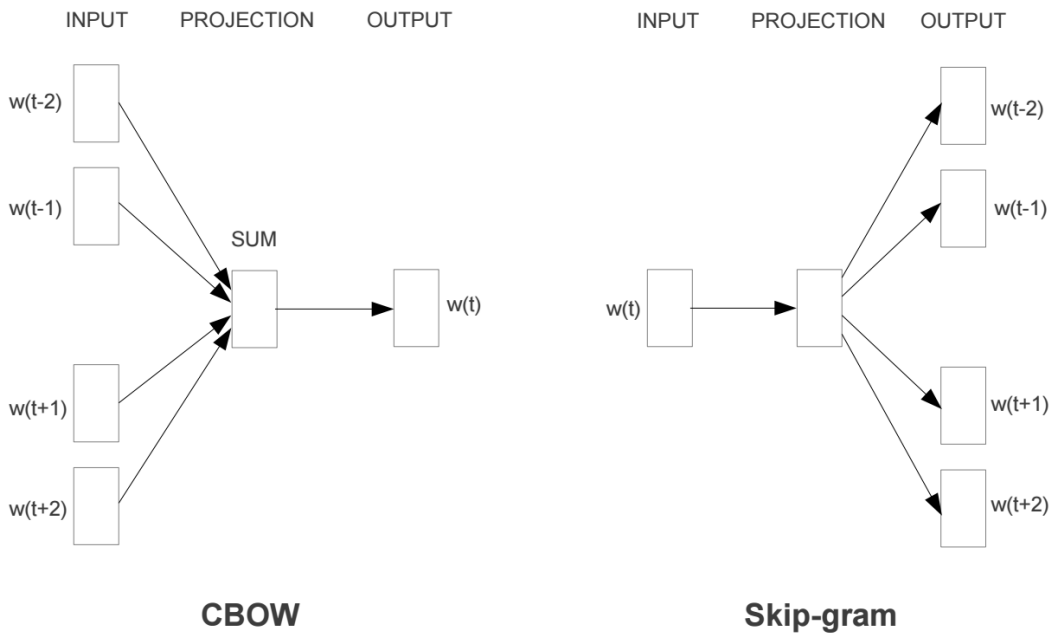
3.1 词向量模型关键词拓展

词向量模型（word2vec[①]）是一种适用于大规模样本集、简单有效的自然语言处理方法，对于语料库中的每个词，计算一个向量来描述这个词。这之后，任何两个词的关系就可以通过二者所对应的向量进行计算得出。比如词 A 和词 B 分别对应着向量 a 和向量 b，那么计算词 A 和词 B 语义的相近性，我们只需要计算 a, b 两个向量的相似性（比如，最简单的，计算两个向量的余弦值）。

Mikolov et al.(2013)提出了两种易行且精确的词向量模型训练方法，使得原来无法扩展到大数据集的神经网络方法在处理十亿规模的语料库时依然可行。这两种方法分别是连续词包 CBOW（continuous bag of words）和 Skip-gram，

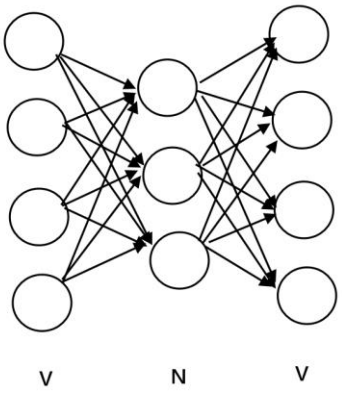
① https://code.google.com/archive/p/word2vec/

其算法思路如图表 13。



图表 13 CBOW 和 Skip-gram 模型示意图[①]

CBOW 是一个三层的神经网络，输入层和输出层均为 V 个神经元，V 是整个语料库词语的数量；隐藏层则有 N 个神经元，各个层之间全连接（如图表 14 所示）。每个词语对应一个$V \times 1$的向量，这个向量有且只有一个分量为 1，其余为 0，表示着当前单词是哪个；将这个向量作为输入，即输入层 V 个神经元只有一个为 1，其余为 0。而隐藏层 N 个神经元，可以视作我们对这个词语有着 N 个维度的解释，N 是自定义的超参数，一般根据 V 的大小进行选择。而输出层则使用 softmax 的方法，输出结果也是一个$V \times 1$的向量，有且只有一个分量为 1，其余为 0，也就是说，这个向量代表着词语库里的一个词。而 CBOW 的训练方式便是，将一句话中，某个词语 w(t)的前若干个词 w(t-1),w(t-2)…和后若干个词 w(t+1),w(t+2)…作为输入，而将这个词本身 w(t)作为输出的训练目标。对于所有词 w(t)及其前后文，这个三层神经网络共享权值，所以我们可以形象地理解，我们预设了 N 个维度来让机器学习 V 个词语，机器以句子为单位阅读



图表 14 三层神经网络示意

[①] 图片引自 Mikolov et al.(2013)

这些词，最后找到了一个最优的 V 到 N 的映射，来理解这些词语。而这个$V \times N$的映射，左乘一个代表某词语的$1 \times V$向量，便成为了一个$1 \times N$的向量，这个向量便代表着这个词语在该语料库中含义对应的向量。而 Skip-gram 算法和 CBOW 类似，不同的则是它的输入输出的数据和 CBOW 相反，输入为某个单词，而输出的目标则是该单词的上下文，但共享权值以及将词语表达为 N 维向量的内核则是一致的。

词向量模型在本指数中的应用有着直观的效果，正如前文所说，我们只能手工挑选出一些较为明显的情感词，而有更多的情感词，因为其专业性和在特殊语境下的含义变化让我们难以事先察觉，所以我们使用词向量模型，寻找我们手工挑选的词库里每个词和其含义最为接近的 10 个词，比如，我们对手工词库里“庞氏骗局”这个词寻找近义词，模型返回的结果如图表 15：

排序	1	2	3	4	5	6	7	8	9	10
词语	骗局	自融	旁氏	拆东墙补西墙	拆标	传销	击鼓	阴谋	谎言	圈钱
相似度	0.564	0.536	0.524	0.509	0.500	0.494	0.482	0.480	0.478	0.472

图表 15 词向量模型，“庞氏骗局”近义词输出结果

于是，我们对每个手工词库中的关键词的前十个近义词进行人工复查，选取其中符合要求的，构成新的情感词库。这样便能很大程度上弥补纯人工构建词库的主观性和局限性。

3.2 情感指数的计算

到目前为止，我们业已获得了一个正向情感词表、一个负向情感词表，一个经过筛选的文档集和它们分属于互联网金融和各个子类的概率。接下来，我们对于每一个正向情感词表里的每一个词语W_k，对其在每一篇文章D_i的出现次数进行计数（以$Count(W_k, D_i)$表示）文章D_i属于互联网金融领域的概率为$Prob_i$反向运算出一个参数Pos_k使得以下等式成立：

$$Pos_k = \frac{100000^{①}}{K \times \sum_i Count(W_k, D_i) \times Prob_i}$$

对于负向情感列表里的词，我们采用类似的计算的方法计算参数Neg_m，只是将 100000 替换成-100000。

上述计算公式事实上遵循了两个思路：其一，我们将样本期间：2013 年 1 月

① 100,000 这里可以是任何常数，类似于关注度指数中对初期指数标准化为 100，

至 2016 年 9 月所有互联网金融相关新闻的平均情感设为 0；其二，我们认为那些出现频次较低的情感词，其单次出现造成的影响将更大一些。

这样，对于每个子类S_j在一个时间段 t（比如，自然月）内的正负情感指数E_{jt}即是：

$$E_{jt} = \sum_{i}^{D_i \in C_t} Prob_{ij} \times \left(\sum_{k=1}^{K_{pos}} Pos_k \times Count(\mathrm{W}_k, D_i) + \sum_{m=1}^{M_{Neg}} Neg_m \times Count(\mathrm{W}_m, D_i) \right)$$

上式中W_k、W_m分别是正向情感词和负向情感词列表中的词，$Prob_{ij}$表示文档 i 属于子类 j 的概率（在关注度指数的计算中得出）。C_t表示在时间段 t 内所有文档的集合K_{pos}与M_{Neg}分别是正向情感词和负向情感词列表的总数。

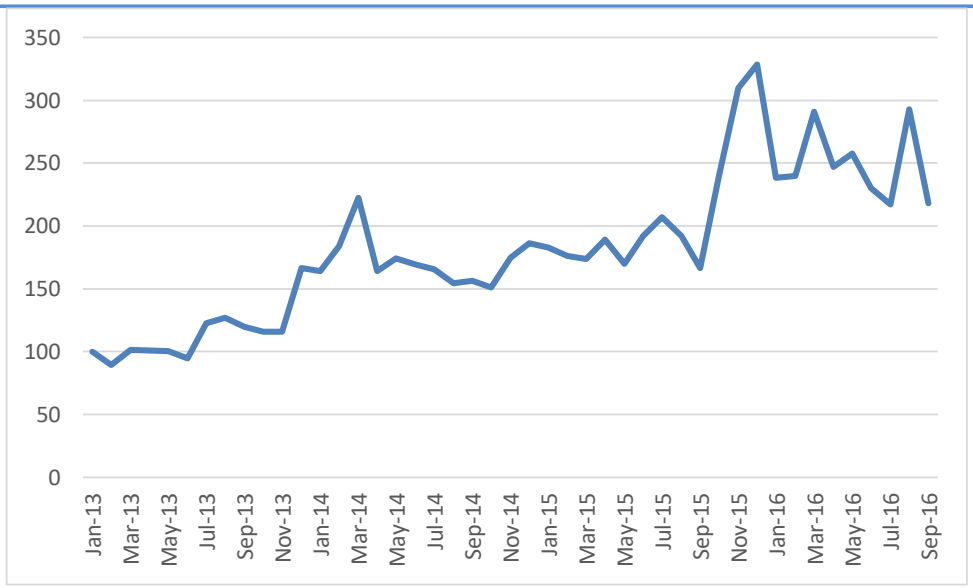
3.3 词向量版本的情感描述

词向量模型在帮助扩充情感词库的同时，为我们提供了新的角度来描述互联网金融及其子类的关注点与情感变化。比如，我们可以以不同季度的文档集作为样本进行训练，这样，对于每个季度，我们可以计算一些关键词的近义词，以度量在该季度舆论环境下，这些关键词受到关注的方面；同时，我们也可以计算在不同季度两个关键词的相似性，比如“P2P”和“监管”，“P2P”和“诈骗”，这是对于情感态度的一个较为直接表达，但将其量化为一个可使用的指数，仍需要更多的尝试。受限于纸面媒介有限的传递方式，我们将在中心网站搭建成熟之后，以交互的方式呈现，即我们可以为感兴趣的人提供任意关键词的查询，返回在每个季度中该关键词含义最接近的几个；或者查询两个关键词在不同季度之间的变化。

4. 主要指数结果汇报

在这里，我们只对指数结果做一个简单的汇报。更多的分析以及学术应用将在后续报告及论文中呈现。同时我们也欢迎感兴趣的研究者使用指数进行更多角度、更加深入的研究。

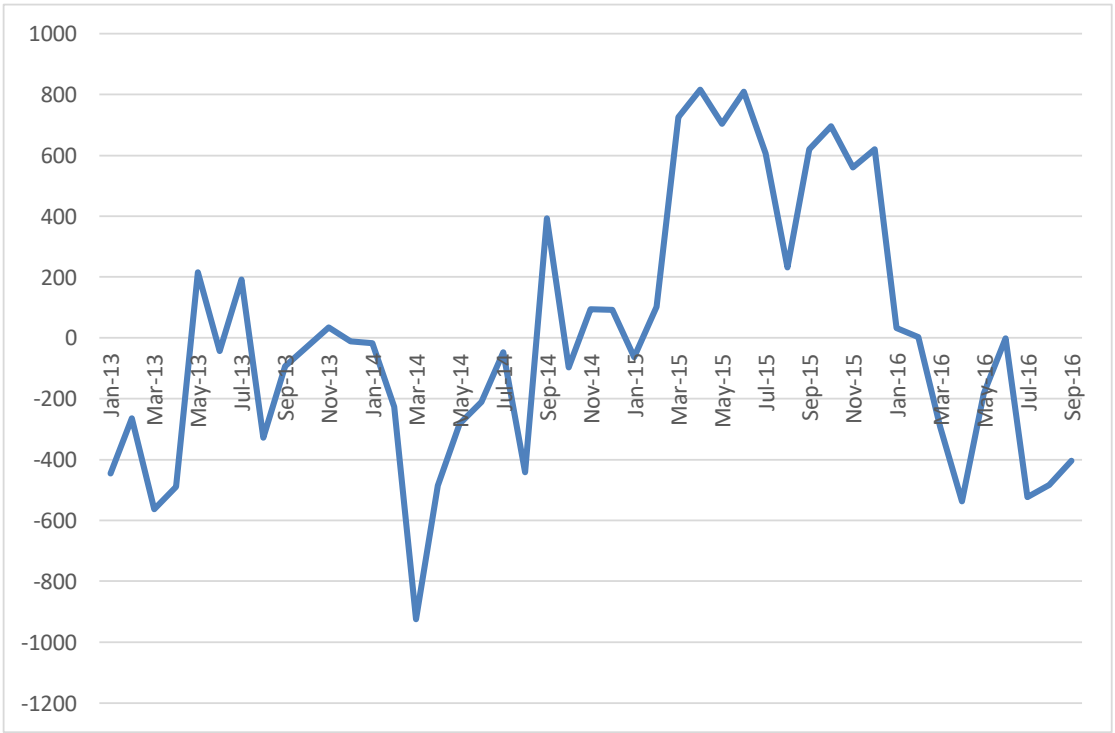
4.1 关注度指数



图表 16 互联网金融情绪指数-关注度指数

图表 16 即为北京大学互联网金融情绪指数-关注度指数的结果。从图中我们可以发现两个重要的时间点：其一，从 2013 年 6 月，至 2014 年 3 月，互联网金融关注度逐步攀升，2013 年 6 月是余额宝推出的时间，而 2014 年 3 月，互联网金融第一次写入政府工作报告；其二，2015 年 12 月，随着“e 租宝”事件的爆发与发酵扩大。

4.2 正负情感指数



图表 17 互联网金融情绪指数-正负情感指数

图表 17 即为北京大学互联网金融情绪指数-正负情感指数的结果，从中我们

仍能发现两个重要的时间点：其一，从 2014 年 3 月，互联网金融写入政府工作报告，互联网金融的正负情感进入一个上升通道；其二，2015 年 12 月，“e 租宝事件”及人们对于 P2P 网贷的恐慌使得互联网金融的正负情感出现断崖式下落。

5. 展望与扩展：开源

我们决定将北京大学互联网金融情绪指数开源，即我们会将编制指数的所有源代码[①]公布，欢迎任何对这一主题有兴趣的人参与其中。

我们将其开源主要是基于三点考虑：

其一，虽然互联网金融的本质是金融，但我们也希望为其带来更多积极的互联网色彩。我们在编制指数过程中，先后使用了 Beautifulsoup、Jieba、Gensim、pyLDAvis 等开源项目的产品，在这里再次感谢。

其二，我们明白现阶段的指数编制手段并不完善，只是应用了较为成熟的算法，受精力所限，我们无法探索那些处于学术前沿的方法。将其开源，我们希望能够吸引一些更专业的人员来参与其中。毕竟，我们兴趣与专长更多地是在指数结果的分析而不是指数编制。

其三，我们也希望类似的分析方法应用于其他领域，为经济学研究带来更多的素材。在指数编制过程中，我们发现了其他一些有意思的话题，比如对于“人民币汇率”的情绪，比如对“网约车”的情绪，这些问题无疑是有意思的，但受限于精力我们不能尽数去做。将代码公开，可以让不甚熟悉相关程序的研究者较为容易地去指数化这些内容。将这套方法应用于其他数据我们是欢迎的，但请遵守以下两个基本准则，其一，请合规引用本文；其二，请将指数结果公开，为大家提供更多的研究素材。

[①] 因为我们希望分析的对象是一般化的文档集，爬虫和网页分析的代码只是针对特定网站，加之，将之公布可能对和讯网产生不可预估的影响，所以我们并不会公布数据准备阶段的代码

参考文献

Baker, S. R., Bloom, N., & Davis, S. J. (2015). Measuring economic policy uncertainty (No. w21633). National Bureau of Economic Research.

Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning (pp. 113-120). ACM.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Chuang, J., Manning, C. D., & Heer, J. (2012, May). Termite: Visualization techniques for assessing textual topic models. In Proceedings of the International Working Conference on Advanced Visual Interfaces (pp. 74-77). ACM.

Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.

Hinton, G. E. (1986, August). Learning distributed representations of concepts. In Proceedings of the eighth annual conference of the cognitive science society (Vol. 1, p. 12).

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. In advances in neural information processing systems (pp. 856-864).

Huang, Y., Shen, Y. and Wang, J. (2016), Analyses and thoughts on individual online lending and its regulatory framework, Comparative Studies, [in Chinese], 2(2016).

Mikolov, T., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J.

(2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Mikolov, T., Yih, W. T., & Zweig, G. (2013, June). Linguistic Regularities in Continuous Space Word Representations. In HLT-NAACL (Vol. 13, pp. 746-751).

Rong, X. (2014). word2vec parameter learning explained. arXiv preprint arXiv:1411.2738.

Sievert, Carson, and Kenneth E. Shirley. "LDAvis: A method for visualizing and interpreting topics." Proceedings of the workshop on interactive language learning, visualization, and interfaces. 2014.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes. Journal of the American Statistical Association, 101(476), 1566-1581.

Wang, Chong, John William Paisley, and David M. Blei. "Online Variational Inference for the Hierarchical Dirichlet Process." AISTATS. Vol. 2. No. 3. 2011.

北京大学互联网金融研究中心课题组（郭峰、孔涛、王靖一、程志云、阮方圆、邵根富、王芳、杨静），2016，《互联网金融发展指数的编制与分析》，《新金融评论》，第 1 期，第 101-129 页。

曾建光. (2015). 网络安全风险感知与互联网金融的资产定价. 经济研究 (07), 131-145.

黄益平, 王海明, 沈艳, &黄卓. (2016). 互联网金融十二讲 中国人民大学出版社.

谢平, & 邹传伟. (2012). 互联网金融模式研究. 金融研究, 12(11), 1.

周建英, 王飞跃, & 曾大军. (2011). 分层 Dirichlet 过程及其应用综述. 自动化学报, 37(4), 389-407.

北京大学互联网金融研究中心简介

北京大学互联网金融研究中（Institute of Internet Finance, Peking University）心，是由上海新金融研究院、北京大学中国社会科学调查中心和蚂蚁金服集团共同发起，2015年10月经北京大学校长办公会批准正式成立的研究机构，中心目前挂靠北京大学国家发展研究院。中心致力于开展对互联网金融、金融科技、普惠金融、金融改革等领域的学术研究，向社会提供权威的科研成果，为政府决策提供参考，服务于金融行业的发展和监管。中心施行理事会领导下的主任负责制，首届理事长由北京大学社科调查研究中心主任李强教授担任，理事会成员包括蚂蚁金服集团首席战略官陈龙、上海新金融研究院执行院长王海明和北京大学国家发展研究院院长姚洋。首任主任由北京大学国家发展研究院副院长黄益平教授担任，王海明、黄卓、沈艳担任副主任。

中心的研究团队包括黄卓、黄益平、孔涛、吕晓慧、沈艳、谢绚丽、徐建国、薛兆丰、朱家祥等北京大学国家发展研究院、中国社会科学调查中心的教授，和中国工商银行原行长杨凯生、证通股份有限公司董事长万建华、中南财经政法大学教授龚强、最高人民法院法官吴景丽等特约高级研究员，以及多位专职博士后和博士研究生。自成立以来，中心研究人员已经独立或联合开发了三个互联网金融方面的指数指数。此外，中心开展了关于个人征信体系建设、商业银行应对互联网金融转型策略、个体网络借贷平台的风险、大数据及普惠金融等多项课题研究，发表学术论文数十篇。中心还在北京大学国家发展研究院开设了互联网金融讲座课，并在此基础上出版了《互联网金融十二讲》。除了大量的企业调研，中心研究人员也积极参与政策研究与咨询，包括网络借贷平台监管和G20框架下的数字普惠金融发展等。