基于文本大数据的企业信用风险评估

杨扬[1]，周一懋[2]，周宗放[3]
1. 西南财经大学经济数学学院，四川 成都 611130；
2. 江苏汇誉通数据科技有限公司，江苏 苏州 215123；
3. 电子科技大学经济与管理学院，四川 成都 611731

摘要
相较于个人信用风险评估，银行关于企业信用风险评估的资料具有更多的文本信息且标准化程度更高，然而文本信息中蕴含的大量关于企业信用风险的增量信息却没有得到有效利用。通过对我国上市公司年报以及网民评论的文本进行分析，探索了文本大数据对企业信用风险评估的作用。研究发现，年报管理层讨论与分析的风险提示内容、独立董事意见中关于关联交易的非正式感情表达以及网民关于公司风险的评论对于企业信用风险评估都具有明显的增量信息。

关键词
文本；大数据；信用风险；评估

中图分类号：F830 文献标识码：A doi: 10.11959/j.issn.2096-0271.2017006

Credit risk evaluation with text big data from text

YANG Yang[1], ZHOU Yimao[2], ZHOU Zongfang[3]
1. School of Economic Mathematics, Southwestern University of Finance and Economics, Chengdu 611130, China
2. Jiangsu HYT Data Technology Co., Ltd., Suzhou 215123, China
3. School of Economics and Management, University of Electronic Science and Technology of China, Chengdu 611731, China

Abstract
Banks have more text information pertaining to credit risk of enterprises. This information is more complete and standardized compared with that of individuals. Nevertheless, this incremental information of enterprise credit risk, which is contained in documents has not been effectively utilized by banks. The role of text big data on credit risk evaluation through annual reports and commentaries of crowds on networks was explored. It found that both the annual reports and crowds' commentaries significantly contain the incremental information on credit risk of enterprises.

Key words
text, big data, credit risk, evaluation

1 引言

随着我国市场化水平的不断提升、信用金融体系的逐步形成，信用风险更加凸显为各大金融机构所面临的主要风险，而银行信用风险管理也面临着前所未有的挑战，这种挑战在当前经济下行和结构调整深化的宏观背景下显得尤为突出[1]。另一方面，互联网以及移动互联网的发展和普遍应用为银行提供了通过大数据进行信用风险管理并挖掘新业务增长点的重要机会。根据Syrus关于风险管理的定义，银行信用风险管理的流程可以分为3个步骤：风险数据的收集、风险的度量和风险的模拟评估[2]。而大数据的应用无论是对银行信用风险数据的收集、信用风险的度量，还是对信用风险的模拟评估，都提供了崭新的视角和方法。相较于银行传统信用风险管理手段，通过对大数据进行整合和分析，降低了成本，提高了效率。这不但可以加强经济波动周期下的信用风险控制，同时也提升了复杂环境下交叉风险的管理能力[3]。事实上，大多数金融机构已经看到了大数据产业化蕴含着的商机，美国花旗银行通过对客户信用卡大数据的分析，对客户进行了有效的信息推送，实现了二次精准营销；美国富国银行则通过客户交易大数据的管理，有效识别了欺诈性交易行为，提升了自身风险控制水平；在我国，阿里巴巴、腾讯等电子支付通道运营商也通过对客户社交活动的大数据分析，纷纷推出“微粒贷”等金融产品，不断挤压金融机构终端零售市场，同时扩宽自身新的利益增长点[4]。

对于大多数银行而言，企业信用风险评估一直以来都是其核心和市场基点。大量的参考文献探讨了在P2P环境下大数据应用对个人信用风险评估的影响。例如Greiner M E和艾金娣等人[5,6]对P2P个人信用风险的数据和信用风险进行了分析；Carlos S C等人[7]考察了个人财富评级对其信用风险的影响；而Larrimore L等人[8]则对Prosper.com的数据进一步分析指出，一些非结构化信息的利用可以有效降低投资者面临的信用风险；周宗放等人[9]对个人信用风险评估的理论和方法进行了拓展性研究。相较于个人信用风险的评估，一方面银行所持有的有关企业信用风险评估的资料包含更多的文本信息（如财务报表、审计报告等），也更加标准化；另一方面，这些文本信息往往并没有被有效利用，许多隐藏在文本中的信息被银行忽略，这大大损失了银行企业信用风险评估的效率。一些学者，如Yang Y等人[10]已经注意到了充分利用文本信息能大大提高银行对企业信用风险评估的效率，而本文则进一步提出了文本大数据应用于企业信用风险评估的3个基本视角，希望能借此为大数据下的企业信用风险评估提供有益启示。

2 数据说明

为了使文本信息更加标准化，同时考虑到数据的真实性和可获取性，本文选取2009—2015年沪深A股被“特殊对待（special treatment，ST）”的上市公司样本，借鉴参考文献[11]的如下配对准则，为这些样本选取“非ST”上市公司的配对样本。准则如下：

- 时间一致性，即“ST”样本和“非ST”样本为同一年度数据；
- 配对样本行业类型相同或相近；
- 配对样本资产总规模相当；
- 配对样本上市时间接近；
- 剔除数据缺失或两年内被“ST”的上市公司；

● 剔除因其他异常情况被“ST”的上市公司。

为了方便研究，本文将样本和配对样本的文本信息进行了梳理，将年报中“管理层讨论与分析”不含“风险分析”“风险提示”的样本删除；进一步，将网络评论少于30条的样本删去。最终得到配对的“ST”样本和“非ST”样本，一共600对。

2.1 信用风险度量

本文选取Logistic回归[12]作为样本的信用风险度量。Logistic回归是一种广泛应用于经济管理的二分类（成功或失败）或者多次序排序的评估（预测）方法。其核心思想是假设X_i为第i个企业的信用风险评估的变量向量，企业出现信用风险违约的概率P_i和X_i之间的关系如下：

$$P_i=1/[1+e^{-(\alpha+\Sigma\beta_i X_i)}] \quad (1)$$

其中，α和β_i为Logistic回归的系数。

通过求最大对数似然函数$\ln(\alpha,\beta)$，就可以得出式（1）中的参数，然后通过外生性设定临界值来判断企业的信用风险情况，对企业信用风险进行评估。

本文将样本分为训练样本和检验样本，其中训练样本400个，检验样本200个。

表1 控制变量

变量	释义
X_1：流动比率	流动资产/流动负债
X_2：速动比率	(流动资产－存货)/流动负债
X_3：资产负债率	负债总额/资产总额
X_4：资产回报率	税后净利润/资产总额
X_5：净利润率	净利润/主营业务收入
X_6：营业利润率	营业利润/全部营业收入
X_7：每股收益率	税后净利润/发行在外的普通股股份数
X_8：存货周转率	营业收入/存货平均余额
X_9：资本周转率	销售收入/股东权益平均余额
X_{10}：销售净现率	经营现金净额/销售收入

在利用训练样本对模型进行训练的过程中，当样本为“ST”样本时，响应值赋值为1；当样本为“非ST”样本时，响应值赋值为0。在确定了系数后，再用训练好的Logistic模型对检验样本的评估精准性进行检验。在本文中，选取0.4作为信用风险评估的临界值。当式（1）中P_i<0.4时，评估样本为“非ST”样本，而当$P_i\geqslant$0.4时，评估样本为“ST”样本。

2.2 控制变量

财务数据是各大金融机构现用于企业信用风险评估的最为重要的数据。本文借鉴已有关于企业信用风险评估、财务预警等相关文献[13]，选取财务变量作为控制变量。考虑到文本信息部分表征了财务数据中蕴含的信息，因此，通过内生性检验后，本文将**表1**中的变量作为控制变量。

3 信用风险评估

本文将考虑两类文本（规范性文本和非规范性文本）中蕴含的有关企业信用风险的信息增量。规范性文本来源于企业的年度财务报表和独立董事意见。年度财务报表和独立董事意见，特别是本文中样本上市公司的年度财务报表和独立董事意见具有高度的标准化特征。而非规范性的文本则来源于巨潮资讯网、新浪财经网等广大股民对企业的评价，体现了文本的多样性特征。

3.1 规范性文本

规范性文本的分析对象在本文中分为两个部分，即企业年度财务报表和企业关于关联交易的独立董事意见。

第一，企业年度财务报表。财务年度报表是各大金融机构用以对其进行信用风险评估时最为重要的文件。然而企业的财务年度报表中除了财务数据以外，大量的文本信息并未得到有效的利用。由于本文要对企业信用风险进行评估，因此，选取企业财务年度报表中“管理层讨论与分析”的风险提示相关内容作为文本分析的对象。这部分文本内容大多对企业未来可能面临的宏观环境、行业竞争以及市场结构等风险进行定性论述。基于此，本文借鉴了Morinaga S等人[14]的词频分析准则，并对风险提示部分相关内容进行了修改。

第二，企业关联交易的独立董事意见。关联交易被普遍认为是影响企业信用风险的重要因素[15]，中国证券监督管理委员会要求上市公司独立董事针对关联交易给出独立意见。然而我国独立董事受限于客户压力和失位风险，很少发表否定性意见。尽管如此，从独立董事意见里仍然可以找出倾向性的感情表达。例如**表2**中两种不同类型的表达就带有明显的感情倾向[16]。

鉴于此，本文借鉴了中文文本情感分析相关方法，对样本独立董事关于关联交易的意见进行分析，将独立董事就关联交易的意见肯定程度分为0~5这6个标度，并作为Logistic模型的判别变量。

表3为在企业信用风险评估的3类分析方法下的第一类错误率和第二类错误率以及评估的准确率①。可以看到：纳入企业财务报表“风险提示”相关内容和独立董事意见以后，第一类和第二类错误率都明显降低，评估准确率有所提高。而**图1**则显示了将检验样本分组后，随着检验样本的增加，评估准确性率变化。

由**表3**和**图1**可知，无论是企业年度财务报表，还是独立董事意见文本信息，均蕴含着关于企业信用风险的信息增量。将文本分析纳入企业信用风险评估能显著提高评估的准确性。此外，随着检验样本量的增加，文本评估准确性呈单调上升趋势。相较于企业年度财务报表中“风险提示”等相关文本信息，独立董事关于关联交易的感情倾向更能帮助银行分析企业的信用风险状况。这可能是由于企业年度财务报表的“风险提示”相关文档过分标准化，很多企业对这部分内容的撰写甚至前后两年的修改程度很小，属于“例行公事”的模版化处理，这也弱化了这部分文本信息的有效性。

3.2 非规范性文本

巨潮资讯网和新浪财经网作为中国最大的财经资讯网，集聚了众多网友针对我国

① 第一类错误率是指将信用好的样本误判为信用差样本的比例；第二类错误率是指将信用差的本样误判为信用好样本的比率；评估准确率是指正确判别的样本占总样本的比率

表 2 企业关于关联交易所发表的独立董事意见

独立董事意见一	独立董事意见二
关联交易不会对公司未来财务状况造成风险，不存在损害公司或投资人利益的情形	关联交易有利于公司降低经营风险，对公司未来的发展具有积极意义

表 3 规范性文本信用风险评估的准确率

分析方法	第一类错误率	第二类错误率	评估准确率
传统Logistic模型	20%	30%	75%
企业年度财务报表	16%	24.5%	79.75%
独立董事关联交易意见	10%	25.5%	82.25%

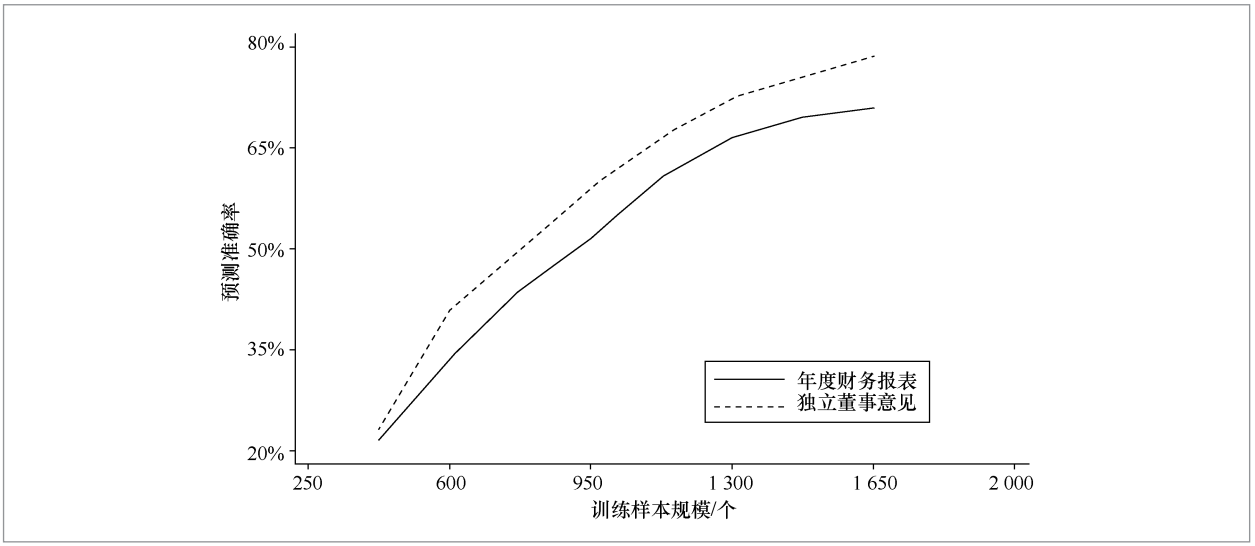


图1 规范性文本评估（预测）准确率

上市公司各个方面发表的看法和观点。这些信息具有典型的非结构化特点。为了简便，在本文中，仅仅利用Python抓取了其中的纯文本部分，由于不同的帖子具有不同的表达习惯，因此，本文借鉴了Tetlock P C等人和Loughran T等人[17,18]在处理社交网络文本时应用的否定词频表方法，并进一步利用Chen H等人[19]的模型，结合Logistic模型对企业信用风险进行检验。**表4**为通过非规范性文本进行信用风险评估的结果。

从**表4**可以看出，对网络媒体文本信息的分析，显著提高了企业信用风险评估的准确率，甚至远远高于利用规范文本信用风险评估的准确率。这说明来自网络媒体的文本信息蕴藏着更多的关于企业信用风险的增量信息。这可能一方面是由于网络媒体的自适应性，众多发表帖子的网民都显然将自己认为“正确”的信息发表在了网络媒体；而另一方面可能与网络大数据有关。对于规范性文本而言，本文的文本样本量为2 400个；而对于非规范性文本而言，本文的文本样本量增加至12 840个，显著增加的文本数据量可能大大提高了信用风险评估的准确率。

4 结束语

银行在进行企业信用风险评估时，往往都要求企业提供诸如年度财务报表、审计报告以及公司所提供抵押担保等大量的文本材料。相较于个人信用风险评估，这些文本材料标准化程度更高。然而银行对这些文本信息的利用效率却并不高，大量隐藏在文本中的有关企业信用风险的增量信息都被浪费，并未能形成银行贷款策略的决策支持。为此，本文利用我国上市公司年报标准文本数据以及在线数据，从公司年报管理层讨论与分析中“风险提示”的内容、公司年报中关联交易“独立董事意见”

表4 非规范文本信用风险评估准确率

分析方法	第一类错误率	第二类错误率	评估准确率
传统Logistic模型	20%	30%	75%
网络媒体	8.5%	18.5%	86.75%

的感情以及众多网友网上回帖内容的多样性等方面着手，探索了文本大数据对企业信用风险评估的影响。研究发现，现有企业文本信息中仍然具有大量的关于企业信用风险的增量信息没有被银行有效利用；进一步利用好文本大数据，从中挖掘出有益信息对银行针对企业信用风险的评估和控制具有重要意义。

总体而言，本文尝试了利用文本大数据对企业信用风险的评估，然而这种研究仍然是探索性的。所用的方法和对数据的处理还比较粗糙，对文本信息的利用也缺乏精细度。相信对企业文本更加细致的挖掘和分析将能更深刻地认识企业的信用风险，进一步提高银行关于企业信用风险评估的准确度，也为银行信用风险管控提供更多的帮助。希望本文不但能对银行利用文本大数据进行企业信用风险评估起到启示作用，更能起到抛砖引玉的作用，为更多的大数据情景下银行信用风险评估的研究提供借鉴。

参考文献:

[1] 庞淑娟. 大数据在银行信用风险管理中的应用[J]. 征信, 2005, 33(3): 12-15.
PANG S J. The application of big data to bank credit risk management[J]. Credit Reference, 2005, 33(3): 12-15.

[2] DUFFIE D, SINGLETON K. Credit risk[M]. Princeton: Princeton University Press, 2003.

[3] 宋长龙. 大数据背景下的银行信用风险管理[J]. 商业经济, 2015(10): 108-109.
SONG C L. Credit risk management of banks in the circumstance of big data[J]. Business Economy, 2015(10): 108-109.

[4] 曾伟, 孔新川, 陈威, 等. 大数据发现银行贷款风险[J]. 大数据, 2014024.
ZENG W, KONG X C, CHEN W, et al. Uncovering the risk in bank loans by big data[J]. Big Data Research, 2014024.

[5] GREINER M E, WANG H. The role of social capital in people-to-people lending market place[C]// The 2009 International Conference on Information Systems, December 15-18, 2009, Arizona, USA. [S.l.:s.n.], 2009: 1-17.

[6] 艾金娣. P2P网络借贷平台风险防范[J]. 中国金融, 2012(14): 79-81.
AI J D. Risk prevention of online P2P lending platform [J]. China Finance, 2012(14): 79-81.

[7] CARLOS S C, BEGONA G N. The use of profit scoring as an alternative to credit scoring systems in P2P lending[J]. Decision Support Systems, 2016, 9(89): 113-122.

[8] LARRIMORE L,JIANG L,MARKOWITZ D, et al. Peer to peer lending: the relationship between language features, trustworthiness, and persuasion success[J]. Journal of Applied Communication Research, 2011, 1(39): 19-37.

[9] 周宗放, 帅理, 周一懋. 个人信用风险评估理论和方法：拓展性研究[M]. 北京：中国金融出版社, 2015.
ZHOU Z F, SHUAI L, ZHOU Y M. Personal credit assessment theories and methods：extension research[M]. Beijing: China Financial Publishing House, 2015.

[10] YANG Y, JING G, ZHOU Z F. Credit risk evaluation based on social media[J]. Environmental Research, 2016, 7(148): 582-585.

[11] 赖娟, 肖珉, 周宗放. 我国集团上市公司财务危机预测建模与实证[J]. 管理学家, 2010(9): 18-24.
LAI J, XIAO M, ZHOU Z F. Empirical study on forecasting financial distress of listed enterprise groups[J]. An Academic Edition of ManaMaga, 2010(9): 18-24.

[12] MARTIN K, HIRSHLEIFER D, TEOH S H. Investor psychology in capital markets: evidence and policy implications[J]. Journal of Monetary Economics, 2002, 49(1): 139-209.

[13] 刘新文. 制造业上市公司信用风险实证研究——基于逻辑回归方法分析[J]. 金融经济, 2016(7): 113-115.

LIU X W. Empirical study on credit risk of listed manufacturings [J]. Finance & Economy, 2016(7): 113-115.

[14] MORINAGA S, YAMANISHI K, TATEISHI K. Mining product reputations on the Web[C]//The 8th ACM International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada. New York: ACM Press, 2002: 341-349.

[15] DOWNS D H, OOI J T L, WONG W C, et al. Related party transactions and firm value: evidence from property markets in hong kong malaysia and singapore[J]. Journal of Real Estate Finance and Economics, 2016, 52(4): 408-427.

[16] 赵子夜. “无过”和“有功”:独立董事意见中的文字信号[J]. 管理世界, 2014(5): 131-141.

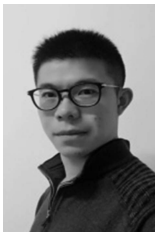
ZHAO Z Y. “No mistake” and “having meritorious deed”: the text information in independent directors’ opinions [J]. Management World, 2014(5): 131-141.

[17] TETLOCK P C, SAAR-TSECHANSHY M, MACSKASSY S. More than words: quantifying language to measure firms’ fundamental[J]. Journal of Finance, 2008, 43(1): 1437-1467.

[18] LOUGHRAN T, MCDONALD B. When is a liability not a liability? textual analysis, dictionaries, and 10-ks[J]. Journal of Finance, 2011, 66(1): 35-65.

[19] CHEN H, PRABUDDHA D, HU Y J, et al. Wisdom of crowds: the value of stock opinions transmitted through social media[J]. Review of Financial Studies, 2014, 5(27): 1367-1403.

作者简介



杨扬(1987-),男,博士,西南财经大学经济数学学院讲师、硕士生导师,主要研究方向为信用风险管理、管理信息系统、集团管控、风险投资。

周一懋(1982-),男,江苏汇誉通数据科技有限公司大数据事业部总监、工程师,主要研究方向为数据挖掘、信用评价、项目管理、系统平台设计等。

周宗放(1950-),男,电子科技大学经济与管理学院教授、博士生导师,风险分析与数据科学研究中心主任,中国科学院/汇誉通大数据联合实验室特聘研究员,主要研究方向为信用风险管理、信用评估、优化理论与方法等。

收稿日期: 2016-12-14

基金项目: 国家自然科学基金项目(No.71271043)

Foundation Item: The National Natural Science Foundation of China(No. 71271043)