报刊新闻量化舆情数据库

2019

数行者科技
DATAGO TECHNOLOGY LIMITED

数行者科技有限公司（Datago Tech. Ltd.）

目录

数行者 DATAGO

目录

数行者 DATAGO

基于大数据的报刊传媒研究

- 2014年，经济学家Matthew Gentzkow因其在利用文本大数据技术开展传媒经济学研究中的建树获得了美国经济学会克拉克奖（俗称“小诺贝尔经济学奖”）
- 特别地，Matthew Gentzkow最具影响力的研究之一则是通过研究美国429家报纸和1000个报道用语分析指出，报纸会投读者所好，有针对的提供有倾性行的新闻和消息[1]
- 此外， Matthew Gentzkow 还利用美国的报刊媒体数据研究了报刊媒体对政治的影响、以及执政党对美国新闻业的影响等[2-3]

[1] Gentzkow, Matthew, and Jesse M. Shapiro. "What drives media slant? Evidence from US daily newspapers." Econometrica 78.1 (2010): 35-71.

[2] Gentzkow, Matthew, Jesse M. Shapiro, and Michael Sinkinson. "The effect of newspaper entry and exit on electoral politics." The American Economic Review 101.7 (2011): 2980-3018.

[3] Gentzkow, Matthew, et al. "Do newspapers serve the state? Incumbent party influence on the US press, 1869–1928." Journal of the European Economic Association 13.1 (2015): 29-61.

报刊媒体数据的广泛应用

- 同时，管理学、金融学等不同学科领域基于报刊媒体大数据的研究也方兴未艾，已经成为各个学科的国际学术热点（相关文献见附录一），主要包括：
  - 经济学
  - 金融学
  - 管理学
  - 会计学
  - 市场营销
  - 社会学
  - 政治学
  - 传媒学

  ……

中国报刊媒体研究现状

- 然而，受限于中文报刊新闻数据整理难和中文分析技术要求高等障碍，基于中国报刊媒体的相关研究在近年来才刚刚起步
  - You, J., B. Zhang, and L. Zhang (2017). “Who captures the power of pen?” *Review of Financial Studies, Forthcoming.*
  - Piotroski, D. J., TJ Wong and T. Zhang. (2017). "Political Bias of Corporate News in China: Role of Commercialization and Conglomeration Reforms" *Journal of Law and Economics*
  - Wu, Y. (2015). "Authority, Incentives, and Performance: Evidence from a Chinese Newspaper." *The Review of Economics and Statistics* 99(1): 16-31.
  - Jin, L., et al. (2014). "The Effect Of Politician Career Concern On Media Slant And Market Return: Evidence From China." *Working Paper.*
- 因此，本报刊新闻量化舆情数据库致力于构建国内领先的中文报刊媒体数据库，帮助相关研究人员逾越技术障碍、迅速开展相关前沿研究。

目录

数行者 DATAGO

产品概述

- 以香港中文大学、美国斯坦福大学和南加州大学等高校商学院教授的相关学术研究成果为指导

  Piotroski, Joseph D. and Wong, T.J. and Zhang, Tianyu, Political Bias of Corporate News in China: Role of Commercialization and Conglomeration. Journal of Law and Economics, 60 (2017): 173-207.

- 采用香港中文大学授权使用的文本分析技术和相关数据，并基于业界领先的机器学习技术进行拓展完善

- 采集并量化分析了自1998年以来，国内外报社所刊登的中文财经新闻的情感倾向性、相关的A股（及B股）上市公司、以及与历史新闻的内容相似性等指标

数据表简介

- 本数据库主要包括了五张数据表（如表1所示），用户可以通过自定义不同的条件，关联检索出更复杂细致的信息以支持不同的研究应用。

表1 数据库总体结构表

表名	字段数	区间	内容说明	记录数（截止2018-12-31）
NEWS_SENTIMENT	18	1998-04-02~	新闻的情感分析结果	550万＋
NEWS_SIMI	4	1998-04-02~	新闻与历史新闻的内容相似度分析结果	5300万＋
NEWS_MENT_STKS	9	1998-04-02~	新闻提及的上市公司及相关信息	1300万＋
PRESS_INFO	6	1998-04-02~	报社基本信息	300＋
NEWS_BASIC_INFO	20	1998-04-02~	新闻基本信息表，包括提及最多的公司，情感倾向评分，发布日期和报社等	690万＋

优势总结

- **规范严谨的数据处理流程** 数据平台的设计开发以前沿学术研究为指导，保证了数据处理过程的规范性和数据库字段的实用性。
- **权威先进的分析技术** 基于由香港中文大学授权使用的文本分析技术及训练数据，并结合业界领先的机器学习和自然语言处理技术进行新闻内容和情感分析。
- **全面细致的数据采集及清理** 考虑历年新上市或者退市的上市公司，结合公司独立整理的上市公司完整的更名记录，易混公司名称记录等，对所采集超过1500万篇原始新闻进行清理筛查。
- **丰富易用的数据字段** 借鉴RavenPack和Thomson Reuters等关注英文媒体和欧美证券市场的海外成熟数据平台，本数据平台提供了丰富易用的字段供研究人员开展广泛研究。

目录

数行者 DATAGO

数据采集与清理—报社清理

- 采集了由1998年以来由1154家海内外报刊媒体刊登的逾1600万篇新闻
- 关注四大证券报，人民日报在内的304个重要报社（排除周/月刊、公信力或者重要性偏低、刊登新闻数较少的报社）刊登的1200余万篇新闻

数据采集与清理—新闻清理

- 关注约1080万篇事件驱动的财经新闻（清除广告、天气预报、招聘启事、公司公告、荐股评述，纯粹的市场行情等新闻）

《苏宁“万人空巷抢彩电”大促销八月来袭》 ✗

《全国主要旅游区天气预报》 ✗

《武汉武商集团股份有限公司1999年年度报告摘要》 ✗

《湖北长江电气有限公司招聘简章》 ✗

《A股上网股市下跌是自行规律的结果》 ✗

《荐神：建行博反弹首选》 ✗

《浦发行高位企稳带大盘回升》 ✗

《苏宁海尔昨突然宣布联手组建销售公司》 ✓

数据采集与清理—内容清理

- 繁简转换，统一规范格式

“而香港華潤集團今年中收購深萬佳百貨，屆時兩家的年銷售額加起來將近40億元，其目標就是要進軍全國零售業的三甲。”

“而香港华润集团今年中收购深万佳百货，届时两家的年销售额加起来将近40亿元，其目标就是要进军全国零售业的三甲。”

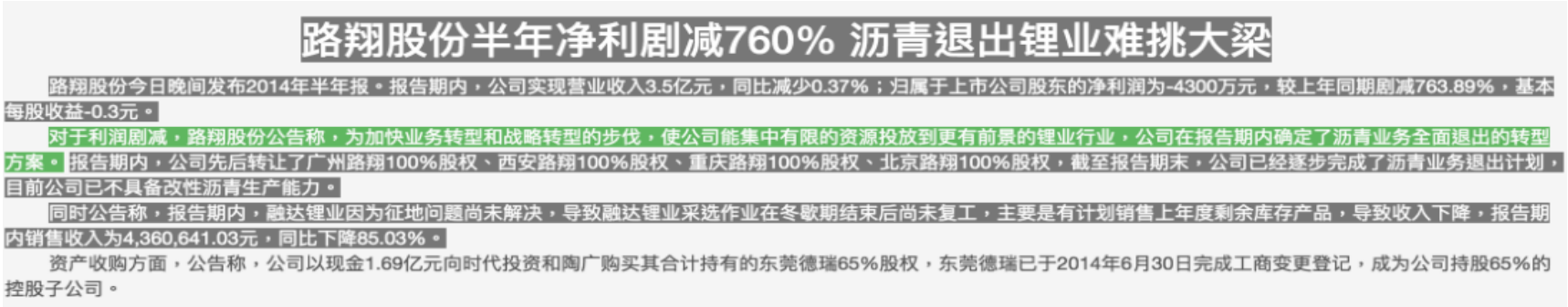
- 清楚冗余内容

“【商报专讯】万达集团<a href="http://zh.wikipedia.org/万达集团" target="_blank">董事长王健林则透露，双方合作绝不仅限于万达广场周边土地的开发，如果接下来的谈判能达成一致，其规模可能超过千亿。（更多报道见A5版）”

“万达集团董事长王健林则透露，双方合作绝不仅限于万达广场周边土地的开发，如果接下来的谈判能达成一致，其规模可能超过千亿。”

新闻情感倾向性分析

- 丰富准确的训练数据：训练数据经多达10名以上香港中文大学商学院学生标注及交叉验证
- 采用机器学习模型在**句子层面**进行标注及分析，更准确细致[1]，在全景网[2]等媒体公开的标注数据集上情感判别精度达90%以上，高于业界领先模型[3]



[1] Zhang, Changli, et al. "Sentiment analysis of Chinese documents: From sentence to document level." *Journal of the Association for Information Science and Technology*, 2009

[2] 全景网舆情频道.http://www.p5w.net/yuqing/

[3] fastText. https://github.com/facebookresearch/fastText

新闻情感倾向性分析

- 汇总新闻整体情感倾向性评分时，有区别地赋予标题，正文以及正文中的首尾段句子和段首尾句子不同权重[1]
- 提供连续可比较新闻整体情感倾向性评分（见表2），而非简单的三级量化情感（正面，中性和负面）

表2 部分正面新闻的情感倾向评分比较

新闻标题	连续可比评分（本数据库）	三级评分
《岭南园林上半年利润增长近一倍》	0.92	正面
《联通集团混改方案获发改委批准》	0.64	正面
《全新好拟改用现金收购港澳资讯》	0.16	正面

[1] Lin, Chin-Yew, and Eduard Hovy. "Identifying topics by position." Proceedings of the fifth conference on Applied natural language processing. Association for Computational Linguistics, 1997

新闻相关公司分析

- 收集核验了上万条上市公司的全称及简称更名记录以准确匹配相关公司
- 避免传统的全文检索匹配带来的相关公司分析误差
- 处理上百个易混淆的公司简称，如"老百姓"，"太平洋"和 "机器人"等

"东南汽车力图销量重回历史高点，与三菱汽车合作加深发展自主品牌开拓海外市场。"

深发展(SZ000001)

"经济形势进入良性循环轨道 老百姓就业机会增多"

老百姓(SH603883)

老百姓拟募集加码主业

老百姓(SH603883)

新闻相关公司分析

- 量化分析了新闻与相关上市公司的相关程度，包括：该公司是否在标题中提及，正文中被提及次数，句子数及首次提及的位置等

任性停牌遭监管"祭旗" 厦华电子首家被公开谴责

7月31日晚间，上交所官网发布了对厦华电子进行公开谴责的监管函。监管函表示，厦华电子办理停牌事项不审慎、重大事项进展披露和风险揭示不及时、不充分等细节曝光，同时上交所对公司及时任公司董事长王玲玲予以公开谴责。上交所表示，对于上述纪律处分，上交所将通报中国证监会和厦门市人民政府，并将记入上市公司诚信档案。厦华电子（600870.SH）也由此成为了沪深两市首家因“任性停牌”被监管“祭旗”的公司 ……

……

据Wind统计数据显示，截至7月31日，A股市场上有45家上市公司的连续停牌天数超过了100天，其中更有包括*ST新亿（600145.SH）、*ST华泽（000693.SZ）、中环股份（002129.SZ）、*ST爱富（600636.SH）在内的4家上市公司停牌时间均已超过300天，上述4家公司连续分别停牌了403天、348天、310天、301天……

最相关公司 → 厦华电子 (SH600870)

新闻相似性分析

- 量化分析每篇文章与长达4周（含当天）内历史新闻内容的相似性
- 可构建新闻原创性及热度，并追踪新闻事件的发展演进等

【商报北京专讯】新组成的北京市第10届政协委员会与上届相比，又有14位港澳人士被增补为政协委员，令港澳人士总数达45人。

据介绍，新增补的14位港澳委员中，突出了专业人士和部分香港知名人士，如香港财经小说作家梁凤仪……

相似度0.91

【新华社北京卅一日电】新组成的政协北京市十届委员会与九届相比，又有十四位港澳人士被新增补为北京市政协委员，十届北京市政协港澳委员总数已达四十五人.

据介绍，新增补的十四位港澳委员中，突出了专业人士和部分香港知名人士事业继承者，如香港畅销财经小说作家梁凤仪……

数行者 DATAGO

目录

数行者 DATAGO

典型客户

高校	采购方式
清华大学	订购3年，季度更新
上海财经大学	同
上海交通大学	同
南京大学	同
南开大学	同

数行者 DATAGO

本科 | 博士 | 硕士 | MBA | EMBA | 高管教育 | 职业发展 | 校友 | English Version | 经管学院 | 联系我们 | 清华

清华经管学院 Tsinghua SEM

清华大学经济管理学院
Tsinghua University School of Economics and Management

招生 | 教师与研究 | 研究中心 | 职业发展 | 校友 | 新闻 | 学院概况 | 建院30周年

- 图书馆
- 最新发布
- 通知公告
- 学院学术成果展示
- 图书馆资源利用
- 读者服务
- 管理规定
- 本馆概况
- 联系我们

图书馆新资源：报刊新闻量化舆情数据库

发布日期：2017-10-23

更多

近年来，利用大数据分析技术采集并分析报刊等权威渠道的新闻资讯，并根据量化分析后的新闻舆情等结果指导金融领域的研究与实务操作，已经得到海外学者和金融从业者的广泛接受。《报刊新闻量化舆情数据库》（新闻报刊量化舆情库说明.docx）借鉴RavenPack和Thomson Reuters等海外成熟数据平台设计， 将分析结果存储到丰富易用的字段，旨在帮助金融领域的学者及业界专业人士逾越大数据和机器学习等技术障碍，较快地利用中文报刊类新闻的量化分析结果开展高质量的前沿研究和金融实务工作。

本数据平台包含从1998年起由报纸媒体发布的与A股（及B股）全部上市公司相

安泰经管学院 ANTAI COLLEGE Economics·Management

教职员登录 | 学生登录 | 交大主页 | English

社交媒体 | LinkedIn

安泰首页 | 安泰概况 | 教学项目 | 师资和研究 | 行业研究 | 国际交流 | 校友 | 更多

首页 > 图书馆 > 数据库

- 首页
- 概况
- 图书
- 论文
- 数字资源
  - 数字图书
  - 数据库
  - 资源下载

报刊新闻量化舆情数据库

发布者：学科与科研办公室　　发布时间：2018-03-27

报刊新闻量化舆情数据库是以香港中文大学、美国斯坦福大学和南加州大学等高校教授的相关学术研究为指导，基于由香港中文大学授权使用的分析技术和相关数据，并结合业界领先的机器学习和自然语言处理技术设计开发。报刊新闻量化舆情数据库包含从1998年到目前为止由报纸媒体发布的与A股（及B股）全部上市公司相关的新闻的文本分析结果，约300家报纸媒体，500万+条舆情信息，季度更新，涉及新闻来源、新闻的情感倾向性、相关上市公司以及与其它历史新闻的内容相似性等指标。

数据库访问链接：http://datago.com.hk

账号：sjtu　密码：sjtu_datago2017

上海财经大学 Shanghai University of Finance and Economics 会计与财务研究院

快速索引 请输入关键字

首页 | 概况 | 研究人员 | 学术科研 | 社会服务 | 制度与简报 | 学术期刊 | 研究中心

数行者报刊新闻量化舆情数据库

发布时间：2017-12-07

一、数据库简介

报刊新闻量化舆情数据库关注从1998年起由报纸媒体发布的与A股及B股上市公司相关新闻，以前沿学术论文为指导，采用由香港中文大学授权使用的技术及数据对上千万篇原始新闻进行了严谨的清理分析及核验。该库分析并提供了每篇新闻的相关上市公司，全文情感量化评分，以及新闻间的相似性等指标。该库可帮助学者快速开展相关实证研究，同时也可以指导金融机构制定投资和风险控制策略。

二、访问入口

会计学院在职教师和在读博士生申请。

【登陆平台网址】http://datago.com.hk

【账号申请】请点击以下链接填写数据库申请表：https://www.wjx.top/jq/18948258.aspx

【使用地点】上海财经大学校区内

【使用须知】

1、教师账号最大并发访问数15人，博士账号最大并发访问数5人；

南京大學圖書館 NANJING UNIVERSITY LIBRARY

登录我的图书馆

首页 | 电子资源 | 热门服务 | 智慧图书馆 | 新闻 | 特色资源 | 图书馆活动 | 学习&研究支持 | 学术交流 | 院系分馆 | 关于图书馆

当前位置：首页 — 电子资源 — 数据库导航 — 正文

- 电子资源
- 资源动态
- 数据库导航
- 版权说明

报刊新闻量化舆情数据库

发布时间：2019.03.12 来源：浏览次数：465

报刊新闻量化舆情数据库

·数据平台地址：http://datago.com.hk/auth/login

·用户名：nju

·密码: nju_datago

·授权时限：2018.05.30至2021.05.31

登陆方式也可通过CSMAR数据库网站的滚动图第三页点击登陆

咨询电话：89683242**电子邮箱：**tsgzxb@nju.edu.cn

目录

数行者 DATAGO

竞品分析—RavenPack

- 关注来自英文媒体所发布的英文新闻，并覆盖欧美证券市场中的上市公司
- 提供连续可比较的新闻情感倾向性评分，原创性评分以及与提及上市公司的相关程度评分（取值范围均为0到100分）
- 建议用户选取70及以上的相关程度评分以及较高（低）的情感评分范围来降低噪声干扰从而获得稳定的分析结果

竞品分析—THOMSON REUTERS

- 关注来自英文媒体所发布的英文新闻，并覆盖欧美证券市场中的上市公司
- 提供主要的新闻情感倾向性及具体分布，与提及上市公司的相关程度评分（取值范围均为0到1分），首次提及的句子数，相似的新闻等

Date/Time	Comp	Source	Item Type	Item Genre	Headline	Relv	Prev Sentmt	Pos	Neut	Neg	1st Ment Loc	Total Sent
01/05/2009 10:42:00.699	C	RTRS	ARTICLE	INTERVIEW	INTERVIEW-Philippines seeks underwriters for bond issue	0.35	-1	0.39	0.16	0.46	9	12
01/05/2009 13:02:12.042	C	RTRS	ALERT	NOT DEFINED	Deutsche Bank DEUTSCHE BANK CUTS CITIGROUP <C.N> 2010 SHR VIEW BY $0.40 TO $0.75	1	-1	0.06	0.13	0.82	1	1
01/08/2009 21:12:49.798	C	RTRS	ARTICLE	US STOCKS SNAPSH	US STOCKS SNAPSHOT-S&P 500, Nasdaq up; Dow off on Wal-Mart	0.20	-1	0.08	0.13	0.79	2	13

竞品分析—总结

	Datago	RavenPack	Reuters
学术研究指导支撑	✓	✓	N/A
全文连续量化情感评分	✓	✓	×
公司相关程度量化分析	✓	✓	✓
新闻热度、原创性评分	✓	✓	✓
历史新闻相似性时间窗	**4周**	N/A	N/A
新闻数(截止2018-12-31)	**550万+**	N/A	N/A
主要关注市场	**A股**	欧美股市	欧美股市
公开数据集准确率	**90%**	-	-

数行者 DATAGO

目录

数行者 DATAGO

附录一 相关学术热点(1)

- 经济学
  - Reuter, J. And E. Zitzewitz (2006). “Do Ads Influence Editors? Advertising And Bias In The Financial Media.” Quarterly Journal of Economics 121(1): 197-227.
  - Dellavigna, S. And E. Kaplan (2007). "The Fox News Effect: Media Bias And Voting." The Quarterly Journal Of Economics,: 1187-1234.
  - Enikolopov, R., Et Al. (2011). "Media and Political Persuasion: Evidence From Russia." The American Economic Review.
- 金融学
  - Ahern, K. R. And D. Sosyura (2014). “Who Writes The News? Corporate Press Releases During Merger Negotiations.” The Journal of Finance 69(1): 241-291.
  - Tetlock P., 2007. Giving Content to Investor Sentiment: The Role Of Media In The Stock Market. The Journal of Finance 62, 1139-1167。
  - Peress, J. (2014). "The Media And The Diffusion Of Information In Financial Markets: Evidence From Newspaper Strikes." The Journal of Finance 69(5): 2007-2043.

附录一 相关学术热点(2)

- 会计学
  - Bushee, B., Core, J., Guay, W., And Hamm, S.. (2010). "The Role Of The Business Press As An Information Intermediary." *Journal Of Accounting Research* 48.
  - Miller, G. “The Press As A Watchdog For Accounting Fraud.” *Journal Of Accounting Research* 44 (2006): 1001–33.
- 管理学
  - Baron, D. P. (2005). "Competing for The Public Through The News Media." *Journal Of Economics & Management Strategy* 14(2): 339-376.
  - Pollock, T.G. and Rindova, V. P., 2003. Media Legitimation Effects in the Market for Initial Public Offerings Academy of Management Journal 46(5): 631-642.
- 市场营销学
  - Lisette de Vries, Sonja Gensler, Peter S.H. Leeflang, 2012. Popularity of Brand Posts on Brand Fan Pages: An Investigation of the Effects of Social Media Marketing, In Journal of Interactive Marketing 26 (2): 83-91

附录一 相关学术热点(3)

- 社会学
  - Sutter, D. (2002). “Advertising and Political Bias in The Media: The Market For Criticism Of The Market Economy.” *American Journal of Economics And Sociology* 61(3): 726-745.
- 政治学
  - Andina-Di´Az, A. N. (2006). "Political Competition When Media Create Candidates’ Charisma." *Public Choice* 127: 353-374.
  - Stockmann, D. (2011). "Race To The Bottom: Media Marketization And Increasing Negativity Toward the United States in China." *Political Communication* 28(3): 268-290.
  - King, G., et al. (2013). "How Censorship in China Allows Government Criticism but Silences Collective Expression." American Political Science Review **107**(02): 326-343
- 传媒研究
  - Beam, R., 2003. Content Differences between Daily Newspapers with Strong and Weak Market Orientations. Journalism and Mass Communication Quarterly 80 (2): 368-390.