

学校编码 : 10384

学号 : 15420131152025

厦门大学

硕士 学位 论文

时态文本关联规则及其应用——基于网络舆情的  
统计分析研究

Temporal Text Association Rule and Its  
Applications in the Statistical Analysis of  
Network Public Opinion

骆翔宇

指导教师: 朱建平

专业名称: 统计学

答辩日期: 2016年4月

## 厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下, 独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果, 均在文中以适当方式明确标明, 并符合法律规范和《厦门大学研究生学术活动规范(试行)》。

另外, 该学位论文为( )课题(组)的研究成果  
, 获得( )课题(组)经费或实验室的资助, 在(  
)实验室完成。(请在以上括号内填写课题或课题组负责人或  
实验室名称, 未有此项声明内容的, 可以不作特别声明。)

声明人(签名) :

年 月 日

## 厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文(包括纸质版和电子版)，允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

( ) 1. 经厦门大学保密委员会审查核定的保密学位论文，于年月日解密，解密后适用上述授权。

( ) 2. 不保密，适用上述授权。

(请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。)

声明人(签名)：

年 月 日

## 摘要

自古以来，社会舆情便是维系社会经济发展与进步的重要一环。随着社会科技的进步与互联网的广泛应用，网络工具正逐步成为民众舆情传播与获取的关键手段之一，网络舆情已经成为社会民众舆情的主要组成部分。作为网络舆情内容的主要构成部分，半结构化的文本数据一直是传统舆情分析中主要的分析对象之一。与此同时，网络舆情信息的实时性特征也凸显了时间属性在其数据结构中的重要地位。随着统计学与数据挖掘研究的不断进展，文本挖掘与时态数据挖掘技术越来越受到学界的关注，将二者结合运用于现代网络舆情的系统性分析框架中，既是网络舆情分析发展的中的重要一环，也是必经之路。

在对文本挖掘及关联规则重要概念整理与统计学阐述的基础上，本文对文本挖掘中的时态文本关联规则方法应用于网络舆情分析的理论可行性进行了相关论证。根据网络舆情本身的特征，本文提出了适用于舆情文本事件化分析的向量空间存储形式，并结合网络舆情事件的一般发展规律，基于费希尔最优分割思想制定了可用于网络舆情信息的时态划分方法，同时从网络舆情分析的实际需要出发，对对应的时态关联规则生成算法的运算逻辑与输出特征进行了研究。

在此基础上，本文以2015年10月份至12月份“央行双降影响楼市”主题舆情分析为案例，针对我国房地产的代表性网络舆情进行了时态文本关联规则的挖掘与分析，并结合关联网络变化情况，发现了目标网络舆情在不同时态阶段中明显的阶段性特征，进一步验证了时态文本关联规则在网络舆情分析应用中的合理性。

**关键词：**网络舆情；文本；时态关联

## Abstract

As we known, social public opinion is an important link to maintain social and economic development and progress. With the wide application of the progress of social technology and the Internet, network tool is gradually becoming one of the key ways that public opinion spread around or be got through, which means that network public opinion has become a main part of social public opinion. As the main part of the content of network public opinion, text data, which is semi structured has always been one of the main analysis objects in the analysis of the traditional public opinion. At the same time, the characteristics of the network public opinion information also highlights the important status of the time attribute in its data structure. With the continuous progress of statistics and research of data mining, text mining and temporal data mining technology get more and more attention in academia, taking both them in the modern network public opinion system analysis framework is important and inevitable for the development of network public opinion analysis.

According to the elaboration of the important concept in text mining and association rules, in this paper we apply the text mining of temporal text association rules method in the network public opinion analysis. With the characteristics of the network public opinion itself, this paper puts forward a vector space storage mode which is suitable for public opinion event analysis, and combined with the Fisher optimal segmentation theory, we present a method that can be used for the temporal partitioning of network public opinion information. At the same time, from the needs of network public opinion analysis, this paper make a study on the temporal association rule generation algorithm which can be used in web text and network public opinion.

October 2015 to December 2015, the Chinese central bank cut both interest rates and reserve ratio, which certainly have impact on the real estate market in China.

To this incident as a case, this paper did a temporal text mining and analysis on China's real estate representative network public opinion. With the observation correlation network change, we found the target network public opinion in different temporal phases of obvious stage characteristics, which further verify the rationality of the application of temporal text association rule in network public opinion analysis.

**Keywords:** Network Public Opinion; Text; Temporal Association

## 参考资料

- [1]程显毅,朱倩.文本挖掘原理[M].北京 : 科学出版社,2010.
- [2]Srikant R., Agrwal R. Mining Generalized Association Rules [J]. Future Generation Computer Systems, 1997, 13(2-3):161-18
- [3]朱建平.应用多元统计分析[M].北京 : 科学出版社,2011.
- [4]Evfimievski A., Srikant R., Agrawal R., Gehrke J.. Privacy Preserving Mining of Association Rules[J]. Information Systems, 2004, (29):343-364.
- [5]Serkan Altuntas, Turkyay Dereli, Andrew Kusiak. Analysis of Patent Documents with Weighted Association Rules [J]. Technological Forecasting & Social Change, 2015, (92):249-262.
- [6]朱建平,谢邦昌.数据挖掘中关联规则的提升及其应用[J].统计研究,2004,(12):4-9
- [7]A.A. Lopes, R. Pinho, F.V. Paulovich, R. Minghim. Visual Text Mining Using Association Rules [J]. Computers & Graphics, 2007, (31):316-326.
- [8]Rajeev Rastogi, Kyuseok Shim. Mining Optimized Support Rules for Numeric Attributes [J]. Information Systems, 2001, (6):425-444.
- [9]朱建平,来升强.时态数据挖掘在手机用户消费行为中的应用[J].数理统计与管理,2008, 27(1):42-53.
- [10]来升强,朱建平.数据挖掘中关联规则算法的考察[J].统计与信息论坛,2005,(1),106-109.
- [11]Agrawal R., Imilienski T., Swami A.. Mining Association Rules Betweensets of Items in Large Databases[C]. proc. of the ACM SIGMOD International Conference on Management of Data, 1993.207-216.
- [12]王爱平,王占凤,陶嗣干,燕飞飞. 数据挖掘中常用关联规则挖掘算法[J]. 计算机技术与发展, 2010, 20(4):105-108
- [13]方匡南,谢邦昌.基于聚类关联规则的缺失数据处理研究[J].统计研究, 2011, 28(2):87-92
- [14]李弼程,林琛,周杰,王允. 网络舆情态势分析模式研究[J]. 情报科学, 2010, 28(7):1083-1088
- [15]董祥军,宋瀚涛,姜合,陆玉昌. 时态关联规则的研究[J]. 计算机工程, 2005, 31(15):24-26
- [16]毕宏音.网民的网络舆情主体特征研究.广西社会科学,2008,(7):166-169
- [17]王国华,冯伟,王雅蕾. 基于网络舆情分类的舆情应对研究[J]. 情报杂志, 2013, 32(5):1-4
- [18]Ya-ping Ma, Xue-ming Shu, Shi-fei Shen, Jiang Song, Gang Li, Quan-yi Liu. Study on Network Public Opinion Dissemination and Coping Strategies in Large Fire Disasters[J]. Procedia Engineering, 2014, 71:616-621
- [19]屈启兴,齐佳音. 基于微博的企业网络舆情热度趋势分析. 情报杂志, 2014, 33(6):133-137
- [20]李雯静,许鑫,陈正权. 网络舆情指标体系设计与分析. 情报科学, 2009, 27(7):986-991
- [21]曾润喜,杜换霞,王君泽. 网络舆情指标体系、方法与模型比较研究[J]. 情报杂志, 2009, 33(4):96-101
- [22]林琛. 基于网络舆论形成过程的舆情指标体系构建研究. 情报科学, 2015, 33(1):146-161
- [23]Ronen Feldman, Ido Dagan. Knowledge Discovery in Textual Databases (KDT)[R]. in: proc. of KDD, 1995.112-117
- [24]G. Salton, A. Wong, C. S. Yang . A Vector Space Model for Automatic Indexing [J]. Communications of the ACM, 1975, 18(11):613-620
- [25]姜宁,宫秀君,史忠植. 高维特征空间中文本聚类研究[J]. 计算机工程与应用, 2002, 10:63-67
- [26]王明文,付剑波,罗远胜,陆旭. 基于协同聚类的两阶段文本聚类方法[J]. 模式识别与人工智能, 2009, 22(6):848-852
- [27]唐晓波,房小可. 基于文本聚类与LDA相融合的微博主题检索模型研究[J]. 情报理论与实践, 2013, 36(8):85-90
- [28]王鹏,高铖,陈晓美. 基于LDA模型的文本聚类研究[J]. 情报科学, 2015, 33(1):63-68
- [29]李荣陆,王建会,陈晓云,陶晓鹏,胡运发. 使用最大熵模型进行中文文本分类[J]. 计算机研究与发展, 2005, 42(1):94-101
- [30]陈晓云,陈祎,王雷,李荣陆,胡运发. 基于分类规则树的频繁模式文本分类[J]. 软件学报, 2006, 17(5):1017-1025
- [31]姚全珠,宋志理,彭程. 基于LDA模型的文本分类研究[J]. 计算机工程与应用, 2011, 47(13):150-153

- [32]Fang Li,Qunxiong Zhu. Study on Association Rules Mining Based Chinese Text Representation [C]. proc. of 2008 IEEE First International Conference on Intelligent Networks and Intelligent Systems,2008,725-728
- [33]Feiyue Ye, Jiannan Xiong, Lingyu Xu. A Text Association Rules Mining Method Based on Concept Algebra [C]. proc. of 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber,Physical and Social Computing,2013,2153-2158
- [34]张保稳,何华灿.时态数据挖掘研究进展[J].计算机科学,2002,29 (2):25-31.
- [35]孟志青.一类相邻时态关联规则的知识发现问题[J].模式识别与人工智能,2001,14(4):495-462
- [36]Yang J., Wang W., P.S.Yu. Mining Asynchronous Periodic Patterns in Time Series Data[J]. IEEE Transactions Knowledge and Data Engineering, 2003,15(3):613-628
- [37]潘定,沈均毅.时态数据挖掘的相似性发现技术[J].软件学报,2007,18(2):246-258
- [38]黄晓斌,赵超.文本挖掘在网络舆情信息分析中的应用[J].情报科学,2009,27(1):94-99
- [39]唐涛.基于大数据的网络舆情分析方法研究[J].现代情报,2014,34(3):3-11
- [40]周亚东,孙钦东,管晓宏,李卫,陶敬.流量内容词语相关度的网络热点话题提取.西安交通大学学报,2007,41(10):1142-1150
- [41]林萍,董卫东.基于LDA模型的网络舆情事件话题演化分析[J].情报杂志,2013,32(12):26-32
- [42]张寿华,刘振鹏.网络舆情热点话题聚类方法研究.小型微型计算机系统,2013,34(3):471-474
- [43]石彭辉.基于社会网络分析的网络舆情实证研究.现代情报,2013,33(2):27-31
- [44]王来华.论网络舆情与舆论的转换及其影响[J].天津社会科学,2008,(4):1-4
- [45]高承实,陈越,荣星,邬江兴.网络舆情几个基本问题的探讨[J].情报杂志,2011,30(11):52-56
- [46]SrikantR.,AgrwalR.Mining Quantitative Association Rules in Large Relational Table[C]. proc.of ACM SIGMOD,1996.
- [47]张朝晖,陆玉昌,张钹.发掘多值属性的关联规则 [J]. 软件学报,1998,9(11):801-805
- [48]苑森淼,程晓青.数量关联规则发现中的聚类方法研究[J].计算机学报,2000,23(8):866-871
- [49]Tung A.,Lu H.,Han J.,Feng L.,Breaking the Barrier of Transactions:Mining Inter-Transaction Association Rules[C].proc.of KDD,1999.
- [50]Lu H.,Feng L.,Han J. Beyond Intratransaction Association Analysis: Mining Multidimensional Intertransaction Association Rules[J].ACM Transactionson Information Systems,2000,18(4):423-454
- [51]Agrawal R.,SrikantR..Fast Algorithm for Mining Association Rules in Large Databases[C]. Proc. of the 1994 International Conference on VLDB,1994.487 – 99.
- [52]Jiawei Han,Jian Pei,Yiwen Yin.Mining Frequent Patterns without Candidate Generation[C].Proc. Conf. on the Management of Data (SIGMOD'00),2000.1-12
- [53]Ramesh C. Agarwal, Charu C. Aggarwal, V.V.V. Prasad. A Tree Projection Algorithm For Generation of Frequent Itemsets[J]. Journal of Parallel and Distributed Computing,2001,61(3):350-371
- [54]Savasere A., Omieinski E., Navathe S.An Efficient Algorithm for Mining Association Rules in Large Databases[C]. Proceedings of the 21st International Conference on Very Large Databases,1995.432-443 .
- [55]Brin S., Motwani R., Ullman J D.Dynamic Itemset counting and implication rules for market basket data[C].Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data.1997.255-264.
- [56]Toivonen H.Sampling Large Databases for Association Rules [C] .Proc. of 1996 International Conference on Very Large Databases(VLDB' 96),1996.134-145.
- [57]欧阳为民,蔡庆生.在数据库中发现具有时态约束的关联规则[J].软件学报,1999,10(5):527-532.
- [58]张春燕,孟志青,袁沛.文本挖掘的时态文本关联规则算法研究[J].计算机科学,2013,40(6):219-224
- [59]杨毅,赵国浩,秦爱民.面板数据的有序聚类分析及其应用——以全球气候变化聚类分析为例.统计与信息论坛,2012,27(7):13-18
- [60]Yiyong Xiao, Yun Tian, QiuHong Zhao.Optimizing Frequent Time-window Selection for Association Rules Mining in Temporal Database Using a Variable Neighbourhood Search[J] . Computers & Operations Research,2014, (52):241-250.
- [61]Michael Hahsler. Rstudio-Visualizing Association Rules Introduction to the R-extension Package arulesViz [Z] . 2010-12-16. Michael Hahsler, Sudheer Chelluboina.

Degree papers are in the “[Xiamen University Electronic Theses and Dissertations Database](#)”.

Fulltexts are available in the following ways:

1. If your library is a CALIS member libraries, please log on <http://etd.calis.edu.cn/> and submit requests online, or consult the interlibrary loan department in your library.
2. For users of non-CALIS member libraries, please mail to [etd@xmu.edu.cn](mailto:etd@xmu.edu.cn) for delivery details.