**二十世纪中文图书中的历史：**

**基于谷歌中文图书语料库的研究**

**摘要：**

本研究依据《知网》情感分析用词语表对谷歌中文图书语料库数据进行情感分析。通过对积极情感用语和消极情感用语的使用频次进行统计，计算出各年的情感分值，发现该情感分值的走势与近代中国历史事件和时期相互照应：情感分值的上升和下降阶段对应于明确的历史时期，情感分值的两个峰值准确地对应着新中国成立和文革结束两个重要的时间节点。采用层次聚类法对情感用语的频次矩阵进行聚类分析，得出的对应于近代历史的五个独特时期的聚集。基于情感用语的频次矩阵构建了分类树模型，对2001年至2008年的数据进行判断，根据该模型预测结果表明二十一世纪以来为上一时间段的延续没有出现转折。

**关键词：N-gram，情感分析，层次聚类，分类树，数据再利用**

**1引言：**

记录人类行为数据正在飞速增长着，这些数据也正日益开放，可供研究者方便地获取。这些大数据正在改变着人们对社会文化现象和变革进行研究的方式，使得更多人类社会的未解之谜有望能科学地被人们理解。[1] 数据密集型的科研方式被称为科研的“第四范式”。[2] 在这种研究范式下，数据不仅是可重用的，即将数据按照原本的研究意图再次使用一次，还也是可再利用的，即将原本的数据用于新的研究意图。

在各个领域涌现出的数据再利用研究案例中，既有针对短时间和少量数据的研究，也有针对长时间跨度和大量数据的研究，但在两种不同类型的研究中，有很多研究都涉及了词语的使用频次。[3~7] 谷歌的研究者利用搜索引擎的检索数据，统计人们检索与流感有关的词语的频次，发现该频次与流感的爆发有相关性，讲结果用于对流感疫情的预测。[3] 通过对微博上股民们所发微博中的情感词语使用频次进行统计，研究者发现该统计结果可用于判断当天股市的涨跌。[4] 这种研究方法也对社会学上的一些难题，提出了可能的解答，Erez Aiden等人利用谷歌英文图书语料库中“United States is”和“United States are”两个用语的使用频次变化，判断出美国是在什么时候开始在其人民心中成为一个完整国家的。[5] 在众多运用谷歌图书语料库的研究中，Acerbi等人根据WordNet情感词表，对语料库中积极情感词语和消极情感词语的使用频次进行了统计，对各年进行情感值的打分，发现该情感值随时间的变化趋势与美国历史上一些重大事件之间有显著的相关性。[6] Hughes等人利用古腾堡电子图书馆项目所数字化的电子书数据，根据各个作品中无意义用语的使用频次对作家进行聚类分析，聚类结果发现相似的作家通常都属于同一种文学风格。[7]

既然美国图书中情感词语的使用频次与美国历史事件之间有密切的联系，不免让人好奇，中文图书中是否也蕴含了这样的生动画面。既然无意义用语的使用频次可以用来判断一部小说归属于那种文学风格，那么情感用语的使用频次又能告诉我们怎样的故事？本研究使用谷歌中文图书语料库和《知网》情感分析用词语表,尝试去发现二十世纪中文图书中情感表达的变化，并根据情感词语的使用频次对年份进行聚类分析，尝试发现中国近代历史上的独特历史时期，最后通过构建的分类树模型展望二十一世纪。

**2 结果**

本研究有四个主要分析结果。结果一：情感用语使用频次的变化趋势。情感用语总使用频次是一年中情感用语的总使用次数与该年所有用语使用次数之比，可以被视为该年出版图书的“情绪化”程度。如图1所示：1980年以前情感用语总使用频率上下起伏较大，但在1950年后开始趋于稳定，总体呈现出上升趋势，而在1980年后开始出现明显下降趋势。1950年前的上下起伏可能预示着新中国成立前社会的动荡，而80年的下降趋势可能与科学、教材等所含情感用语较少的书籍出版量上升有关。

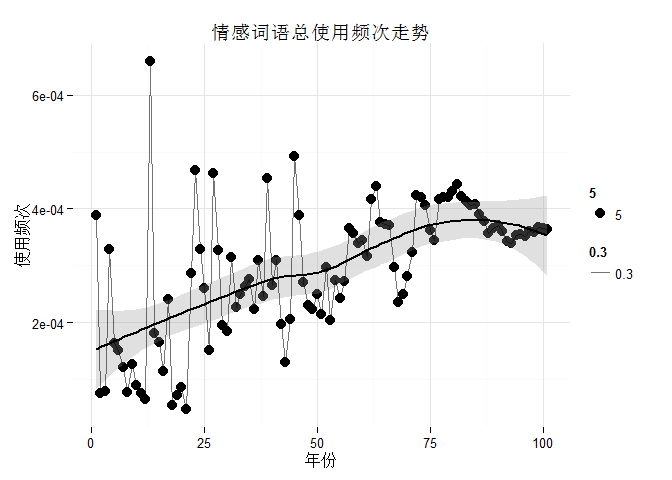


图1情感用语总使用频次变化趋势

结果二：积极、消极和总情感值的变化趋势。这里我们分别计算了积极情感值、消极情感值和总情感值的变化趋势。

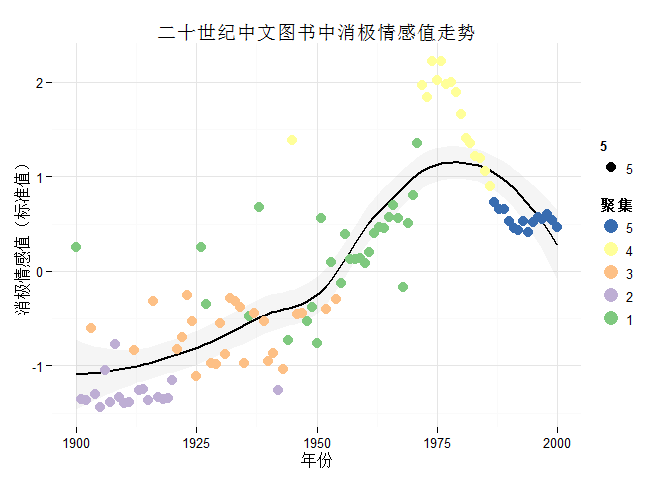
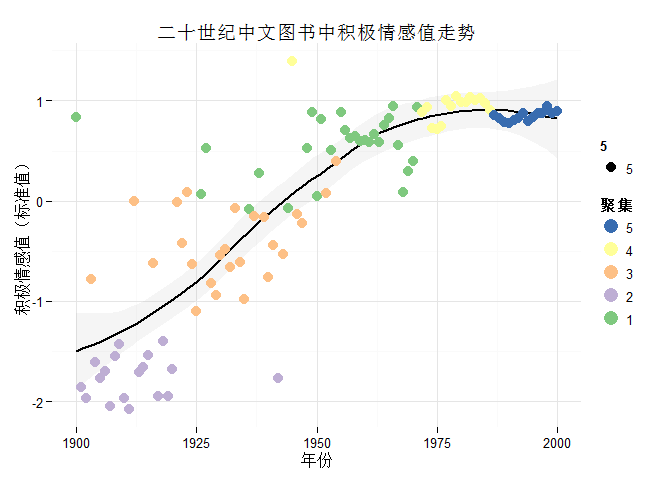


图2 积极情感值的变化趋势 图3 消极情感值的变化趋势

积极情感值和消极情感值分别是两种情感用语使用频次正则化后的标准值。如图2图3所示，积极情感值逐年上升于七十年代趋于稳定，消极情感值在七十年代出现峰值随后逐年下降。

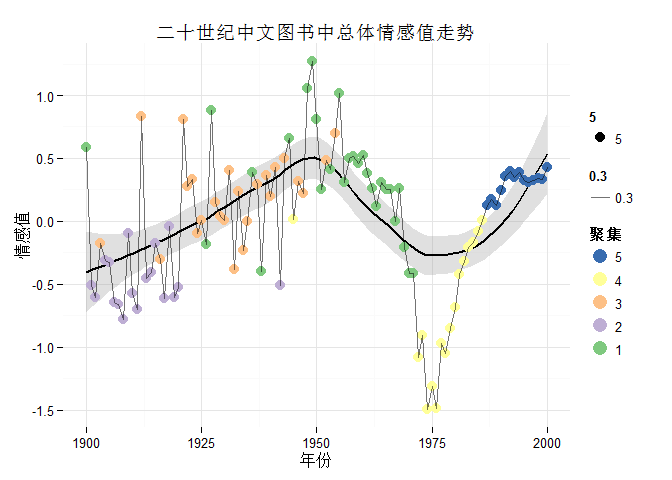


图4 总情感分值的变化趋势

总情感值是通过对积极情感用语使用频次正则值与消极情感用语使用频次正则值之差进行标准化后所得。如图4所示，总情感值在1950年前呈现上升趋势，之后在经历过70年代的急剧下降后处于稳定的上升中。总情感值变化趋势中可以明显地区分出几个时间段，即：两个消极情感主要时间段（总情感分值小于零），分别为1901年至1920年以及1969年至1985年；两个积极情感主要时间段（总情感值大于零），分别为1921年至1968年以及1986年至2000年。可以看出，总情感值中所体现的趋势与人们对于中国近代历史的认识之间是吻合的。

结果三：发现了5个独特历史时期。通过对情感用语使用频次进行层次聚类，发现了5个明显的聚集，各个聚集所历史时期如表1所示:

|  |  |  |
| --- | --- | --- |
| 聚集 | 时间段 | 对应历史时期 |
| 1 | 1948年至1970年 | 新中国成立至文革期间 |
| 2 | 1900年至1920年 | 中国共产党成立以前 |
| 3 | 1921年至1947年 | 中国共产党成立至新中国成立 |
| 4 | 1971年至1985年 | 文革期间至改革开放初期 |
| 5 | 1986年至2000年 | 改革开放初期至20世纪末 |

表1 情感词语使用频次聚类结果

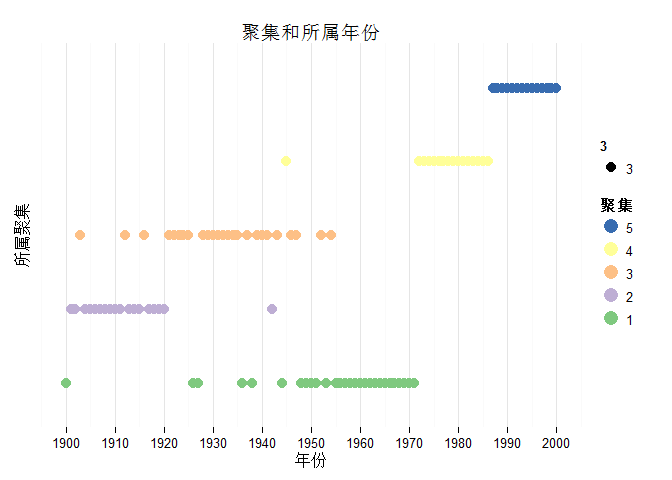
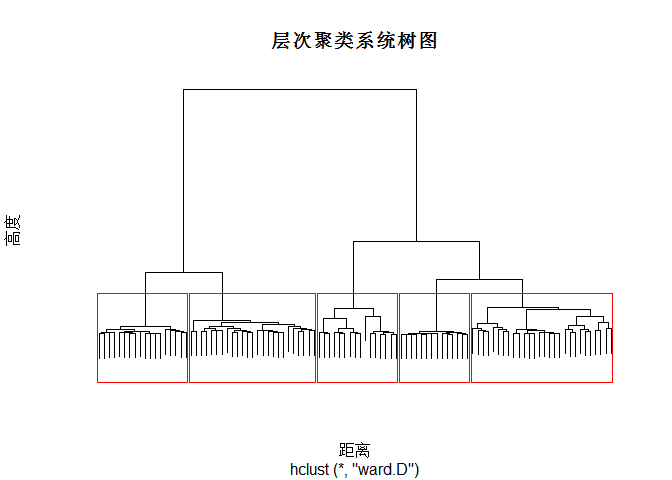
如图5、图6所示，虽然有聚类的结果又少许误差，但是5个聚集之间基本不存在交叉。这个聚类结果表明，在没有任何历史知识的参与下，仅靠对数据中所揭示的数据便可以对历史进行较为有意义的划分。

图5层次聚类法系统树图 图6各个聚集所包含年份

结果四：构建了一个分类树模型。如果我们想知道二十一世纪初与二十世纪中哪个时间段更为相似，会否出现急剧的上升或下降？则可以从二十世纪的数据中提取模型，用于回答这样的问题。根据情感用语的使用频次，采用L1正则化的多项回归模型[9]和10折交叉确认法，我们发现仅需要5个词语作为分支节点即可达到最高预测准确率（图7）。因此，本研究中，我们构建了一个包含5个节点的分类树模型[10]用于预测年份所属聚集，模型仅使用了“好”、“惊讶”、“愿意”、“爱护”和“关注”五个词语（图8），在训练数据上的分类准确率为99%。使用该分类树模型对2001年至2008年的数据进行预测，结果这八年均被分类为聚集5，表明二十一世纪数据是二十世纪末以来的延续。



图7 Lasso多项回归模型交叉确认误差 图8分类树模型

**3 数据与方法**

**3.1 数据来源**

谷歌中文图书语料库由谷歌所数字化处理的30多万册中文图书所构成，包含词语和短语在各年中的使用次数。[10]该语料库被频繁用于语言学和文化变革的定性研究。[11]本研究采用了1900年至2000年之间的中文1gram数据作为研究对象，其中包括66万多个不同的用语，所有用语的总出现次数为720多亿。

在统计积极情感词语与消极情感词语的使用频次时，本研究采用了《知网》情感分析用词语集中包含的两个词表。[12]该词表被广泛应用于中文句子以及文本初步情感分析。

**3 讨论**

本研究的结果表明，在没有任何领域知识的前提情况下，仅通过谷歌中文图书语料库中情感词语的频次分布进行分析，研究者就可以找出中国近代历史的关键时间节点，并对中国近代历史进行有意义地划分。本研究采用方法的优点在于：分析过程中不需要任何领域知识，并无任何的主观判断，关键历史时间节点的发现和历史时期的划分均为大数据本身所揭示。

利用图书中的词频统计数据研究社会和历史已有许多成功的案例，但也有其不可忽视的缺陷。首先，作者将对社会的看法和意见撰写并出版成书的过程通常需要不短的时间，即从图书中反映出的社会现象相对于真实的历史时间存在一个时间上的滞后。其次，图书中反映的社会变化可能存在偏差，因为作者没有义务将对社会的观察如实地付诸文字，而且作者群体也不能无偏地代表总体人群。最后，图书数据不可能完整，没有组织或机构能将历史上存在过的所有图书均数据化，本研究所采用的谷歌图书语料库，也只包含百分之三左右的出版图书。

**3.2 方法**

在计算积极或消极情感分值时，对于给定年份y, 计算当年所有词语使用的总次数S以及相应词表中情感词的使用频次，则这该年这种情感值为词语频次正则化后的平均值，即：

其中和分别为第y年的积极情感分值和消极情感分值，和分别为第y年各个用语的出现频次，m和n分别为消极词表和积极情感词表中的用语数目。

为了使总体情感分值有意义，需要对两种情感分值标准化处理为标准值：

则总体情感分值即为积极的情感分值和消极的情感分值之差：

在发现独特的历史时期时，采用了层次聚类法。层次聚类法是广为采用的非监督学习方法，聚类的过程中反复将距离最近的聚集进行合并直至满足特定的中止条件。本研究在R语言中使用是Ward最小方差法对历史年份进行层次聚类，该方法通常用于发现数据中紧密相连的数据点构建聚集。

分类树模型是一种通过迭代地寻找能够最好地将数据进行分类的节点变量而构建的模型，模型可以以树的形式表达，具有较好的解读性。本研究采在R语言中采用rpart包裹构建的分类树模型。

**4 讨论**

本研究通过对谷歌中文图书语料库数据进行再利用，通过词频统计的方法对其进行情感分析，得到了若干有趣的结果。本研究的结果表明，在没有任何领域知识的前提情况下，仅通过书语料库中情感词语的频次分布进行分析，研究者就可以找出中国近代历史的关键时间节点，并对中国近代历史进行有意义地划分。研究采用方法的优点在于：分析过程中不需要任何领域知识，也不夹杂人为的主观判断，关键历史时间节点的发现和历史时期的划分均为大数据本身所揭示。

利用图书中的词频统计数据研究社会和历史已有许多成功的案例，但也有其不可忽视的缺陷。首先，作者将对社会的看法和意见撰写并出版成书的过程通常需要不短的时间，即从图书中反映出的社会现象相对于真实的历史时间存在一个时间上的滞后。其次，图书中反映的社会变化可能存在偏差，因为作者没有义务将对社会的观察如实地付诸文字，而且作者群体也不能无偏地代表总体人群。最后，图书数据不可能完整，没有组织或机构能将历史上存在过的所有图书均数据化，本研究所采用的谷歌图书语料库，也只包含百分之三左右的出版图书。

**参考文献：**

1. Wcinberger D. (2011) The machine that would predict the future[J]. Scientific American 305(6): 52-57.
2. Gray J. On eScience: A Transformed Scientific Mehod[M]. The Fourth Paradigm: Data – Intensive Scientific Discovery.
3. Ginsberg J., Mohebbi M. H., Patel R. S. et al. (2009) Detecting influenza epidemics using search engine query data[J]. Nature 457, 1012-1014.
4. Bollen J, Mao H, Zeng X-J. (2011) Twitter mood predicts the stock market[J]. Journal of Computational Science 2:1-8.
5. Aiden E., Michel J-B. Uncharted: Big Data as a Lens on Human Culture. Penguin 2013.
6. Acerbi A. Lampos V. Garnett P. et al. (2013) The Expression of Emotions in 20th Century Books[J]. PLOS ONE.
7. Hughes JM, Foti NJ, Krakaucer DC. et al. (2012) Quantitative patterns of stylistic inuence in the evolution of literature[]. Proceedings of the National Academy USA 109:7682-7686.
8. Hastie T., Tibshirani R., Friedman J. H. (2003) The Elements of Statistical Learning.[M] Springer
9. Breiman L., Friedman J. H., Olshen R. A., Stone, C. J. (1984) Classification and Regression Trees. Wadsworth.
10. Books.google.com/ngrams [DB/OL]
11. Michel J-B., Shen Y.K., Aiden A.P. et al (2011) Quantitative analysis of culture using millions of digitized books[J]. Science.
12. 董振东等. 《知网》情感分析用词语集(beta版)[DB/OL] <http://www.keenage.com/>