**二十世纪中文图书中的历史：**

**基于谷歌中文图书语料库的分析**

**摘要：**

本研究使用谷歌中文图书语料库数据，依据《知网》情感分析用词语表对历年积极情感词语和消极情感词语的使用频次进行统计并计算历年的情感分值，发现情感分值的两个峰值分别对应中国近代历史的两个重要历史事件。采用层次聚类法对情感词语的频次矩阵进行聚类分析，发现聚类得出的五个聚集分别大致对应于近代历史的五个独特时期。以及构建一个基于情感词使用频次的分类树模型，可用于对特定年份所属历史时期进行判断。

**关键词：N-gram，情感分析，层次聚类，分类树，数据再利用**

**1引言：**

记录着人类行为的数据正在数据量和数据可获取性两方面飞速地发展着，这样的大数据正在改变着人们对社会和文化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] 这种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，即将原本的数据按照原本的研究意图再次使用一次，而且也是可再利用的，即将原本的数据用于新的研究意图。

各个领域中新出现将已有数据再利用于新问题的案例中，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，词语的使用频次都扮演了重要的角色。[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 ，用于预测流感疫情。[3] Bollen等人对微博上情感词语的使用频次进行统计，发现统计结果可用于预测当天股市的涨跌。[4] Erez Aiden等人利用谷歌英文图书N-gram语料库中“United States is”和“United States are”两个用语的使用频次随时间的变化，解释了美国是在哪个历史时期开始在人民心中成为一个完整的国家。[5] Acerbi等人对谷歌英文图书N-gram语料库中表示积极情感和消极情感词语的使用频次，对每一年图书中所揭示的总体情感值进行打分，发现该情感值随时间变化的规律与美国历史上一些重大事件之间有明确的相关性。[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，根据各个作家作品中不揭示内容的词汇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]

本研究中使用谷歌中文图书Ngram数据库，尝试去发现二十世纪中文图书中情感表达的变化趋势，并根据各年不同情感词语的使用频次对年份进行聚类分析，最后针对各个重要历史时期构建分类树模型，探索哪些情感词语的使用频次对历史时期的归属起着关键作用。

**2 结果**

本研究主要包括三个结果。首先，如图1所示，我们在数据中区分出若干明显的积极情感和消极情感时间段，两个消极情感的主要时间段（即标准化的总体情感分值小于零）为：1901年至1920年以及1969年至1985年，总体情感值在1975年前后达到底谷,并在之后逐年回升，该底谷大致对应于文革时期的结束；两个积极情感（即标准化的总体情感分值大于零）的主要时间段为：1921年至1968年以及1986年至2000年，总体情感值在1949年达到全局最高值，该峰值对应于新中国的成立。

其次，如图4和图5所示，根据情感词语使用频次对年份进行层次聚类，发现5个明显聚集：

|  |  |  |
| --- | --- | --- |
| 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

最后，如图6所示，若我们用某年情感词语使用频次预测该年属于哪一个聚集，使用L1正则化的多项回归模型[9]，采用10折交叉确认法发现，只需要5个词语的使用频次即可达到最高预测准确率，使用3个词至13个词均在1个标准差之内。采用5个情感词语构建一个分类树模型[10]用于预测年份所属聚集，结果发现采用“好”、“惊讶”、“愿意”、“爱护”和“关注”5个词语所构建的分类树模型达到99%的准确率。利用该分类树模型对2001年至2009年的数据进行分类判断，发现其中7年均属于聚集5。

**3 讨论**

本研究发现，在没有任何历史知识的情况下，仅仅通过谷歌中文图书语料库数据中情感词语的频次分布，研究者就可以找出历史时期的关键时间节点，并对历史时期进行有意义地划分。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，并且不需要任何人为的主观判断，关键历史时间节点的发现和历史时期的划分都由大数据本身提供。

利用图书中的词频统计数据研究社会和历史，有很多成功的案例，但也有其不可忽视的缺陷。首先，作者将对社会和历史的看法和意见撰写并出版成书的过程需要时间，即从图书中反映出的社会变化相对于真实的历史时间存在一个滞后。其次，图书中反映的社会变化可能存在偏差，图书作者没有义务将对社会变革的观察如实地付诸文字，并且作者群体也不能完美地代表总体人群。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。

**4 数据与方法**

**4.1 数据来源**

谷歌中文图书Ngram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频次。[10]该语料库被频繁用于语言学和文化变革的定性研究。[11]本研究采用了中文1gram数据中年份在1900年至2000年之间的一个子集作为研究数据，其中包括661281个不同的词语，所有词语的总出现频次为72622729574。

在统计每年表达积极情感与消极情感的词语和用语使用频次时，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。

**4.2 情感分值计算**

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{c1,…cn}，则这种情感在该年的情感分值为词语频次正则化后的平均值，即：

为了使总体情感分值（即积极的情感分值和消极的情感分值之差）有意义，需要对两种情感分值标准化处理为Z值：

其中和分别为该种情感分值的平均值和标准差。

**4.3 层次聚类**

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。本研究使用层次聚类法对历史年份进行聚类,在聚类过程中采用了欧几里德距离作为聚类距离，即两个聚集和之间的距离为：

其中和分别为两个聚集中词语的使用频次。对于包含多个年份的聚集，采用平均词频作为聚集的词频。

**4.4 分类树模型**

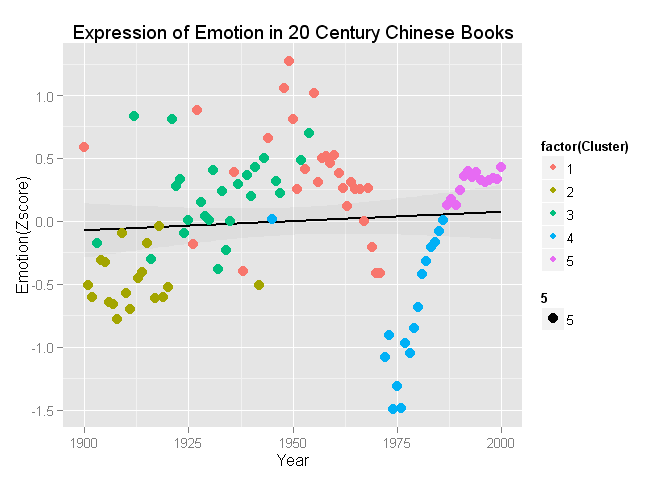


图1 标准化总体情感分值随年份的变化趋势

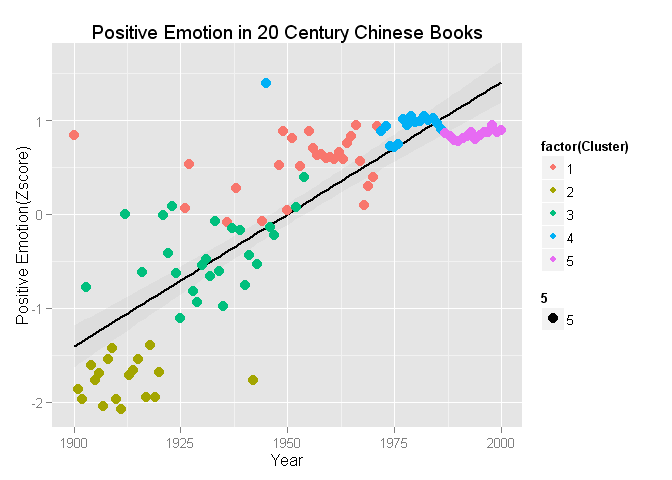
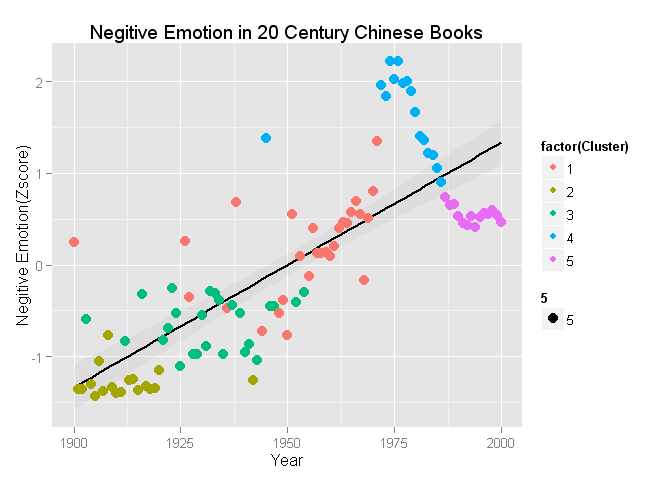
 

图2积极情感分值随年份的变化趋势 图3 消极情感分值随年份的变化趋势

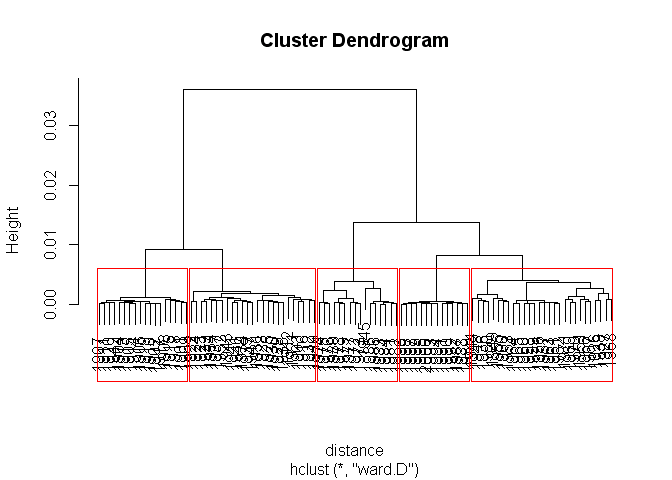
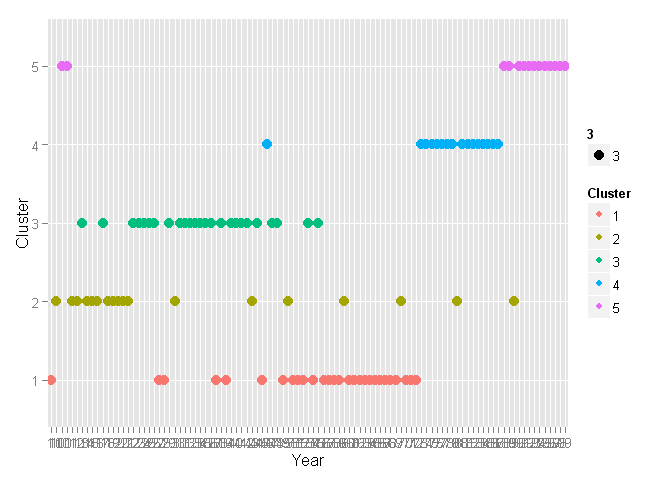
 

图4层次聚类法系统树图 图5 各个聚集所包含年份



图6 Lasso多项回归模型交叉确认误差 图7 分类树模型

**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/>