基于关键词抽取算法的隐喻研究趋势分析*

张冬瑜[1] 顾 丰[1] 崔紫娟[2] 胡绍翔[1] 张 伟[1] 林鸿飞[3]

[1]（大连理工大学软件学院 大连 116620）

[2]（大连理工大学国际合作与交流处 大连 116024）

[3]（大连理工大学计算机科学与技术学院 大连 116023）

摘要：**【目的】**针对中国近40年隐喻研究的发展和演变规律进行梳理和定量分析，旨在为相关语言学家和计算语言学家提供参考，缩小中国隐喻研究与国外先进水平的差距。**【方法】**采用关键词抽取算法将隐喻文献映射为关键词集合，结合科学计量学原理筛选出6个有效特征作为回归模型参数预测下一年的热点词频度，对隐喻发展进行历时和共时分析。**【结果】**对比5种回归模型预测结果，发现拟合程度最好的梯度上升回归树模型对下一年度的关键词预测精度最高，特征消融实验的结果也证实所选的每一个特征均有效。**【局限】**关键词抽取算法的准确性有待进一步提高。**【结论】**隐喻研究正在向多领域、多学科交叉的方向发展。本文特征选择的方法可以为隐喻自动识别研究提供参考。

关键词：隐喻 热点预测 回归模型 历时分析 共时分析

分类号：TP393

DOI：10.11925/infotech.2096-3467.2019.0893

引用本文：张冬瑜，顾丰，崔紫娟等. 基于关键词抽取算法的隐喻研究趋势分析[J]. 数据分析与知识发现，2022，6（4）：130-138.(Zhang Dongyu, Gu Feng, Cui Zijuan, et al. Reviewing Metaphor Research Based on Keyword Extraction Algorithm[J]. Data Analysis and Knowledge Discovery, 2022, 6(4): 130-138.)

1 引 言

隐喻不仅是一种生动、形象的语言现象，而且是人类认知世界的方式[1]。它在人类语言中使用的频率非常高，日常交流中每三句话就可能有一句隐喻出现[2-4]。人们在词汇与相关知识缺乏的情况下经常使用隐喻，即用简单、具体、熟悉的事物去描述和理解复杂、抽象、未知的事物。

近年来，微博、论坛、贴吧等社交媒体提供了分享个人观点、发表评论、在线评价等越来越多的功能，相关的中文隐喻数量随之激增，得到了更广泛的关注。导致隐喻研究成果也急剧增加，因此迫切需要利用文本挖掘方法对隐喻研究的发展过程进行详细梳理，科学、客观地反映隐喻研究的现状、规律和趋势等问题，为相关语言学家和计算语言学家提供参考。在自然语言处理领域，处理海量文本文件最关键的是要把可以表达整个文本主题思想的几个关键词提取出来，因此，关键词提取在文本挖掘领域是一个很重要的部分。

为此，本文提出一种采用文本挖掘技术对中文隐喻研究进行分析和处理的方法，揭示隐喻研究的发展历程和演变规律。以期为隐喻研究这一领域的研究者提供进一步的理论指导和实践参考。

本文从知网上下载1980年-2018年所有标题和摘要中包含关键词“隐喻”的期刊论文。采用关键词抽取算法中的TextRank方法将文献映射为关键词集合，通过提取文档中已经存在的关键词和句子，根据

通讯作者（Corresponding author）：林鸿飞（Lin Hongfei），ORCID：0000-0003-0872-7688，E-mail：hflin@dlut.edu.cn。

*本文系辽宁省社会科学规划基金项目（项目编号：L20BYY023）的研究成果之一。

The work is supported by Liaoning Province Social Science Planning Fund Project (Grant No. L20BYY023).

词之间的共现关系构造无向有权网络，从而形成关键词集合，并选取 Top-100 的关键词作为候选关键词。结合科学计量学的基本原理，考虑到当年的关键字频数、关键字的存在年限、当年增加的新作者数量、当年的基金资助论文数量、跨学科提及数量等，利用 5 种回归模型分别预测下一年的关键字频度，获得候选的热点关键字，并比较 5 种回归方法对于隐喻热点的预测。最后，分别从历时和共时两个维度进行分析，对于历时维度，对隐喻的研究历程大致划分为沉默期和增长期两个阶段；对于共时维度，利用主题分析思想，根据当年热点词的占比变化将研究内容分解为不同研究主题，观察主题的演化过程。

2　研究现状

西方隐喻研究最早可以追溯到古希腊时期，英文“Metaphor”（隐喻）一词来源于希腊语“Metaphorá”。隐喻在这一时期被认为是一种修辞方式。直到 20 世纪 80 年代，随着现代认知科学的迅速发展，隐喻研究进入到一个崭新时期：认知阶段。其中以 Johnson 等[1]的概念隐喻理论影响最为深远，建立了隐喻认知本质的理论基础，推动了隐喻研究从语义到认知角度的转变，带动了包括语言学[5-6]、心理学[7-8]、神经科学[9-10]、计算机科学[11-12]等各学科隐喻研究的蓬勃发展。

随着西方隐喻理论进入中国，国内学者开展隐喻的本质、分类、认知机制、中西对比等多方面定性的研究[13-14]，也出现了隐喻定量研究的成果和论文[15-16]。可见，隐喻研究受到极大关注。

虽然隐喻文献急剧增加，但是目前利用文本挖掘方法对隐喻进行梳理的工作还是很少，对于隐喻研究演化和发展趋势主要采用人工梳理以及使用计量学工具两种方法。人工梳理的方法在科学性、客观性上也有一定主观性的偏差。此外，有一些研究采用科学计量学工具对隐喻研究进行统计，例如孙亚等[17]利用文献分析工具 SATI 和网络分析软件 Ucinet 对 Chinese Social Sciences Citation Index（CSSCI）期刊中带有“Metaphor”的论文进行检索和梳理。但是这种方法对于模型的运用灵活性不足，数据采集量也不充足。另外，这些工具针对英语开发，对于汉语很难适用。

综上所述，目前隐喻研究的梳理工作存在以下两方面的不足。

（1）定性分析居多，根据专家经验进行预测。这种方法依赖于专家经验，缺乏对隐喻整体发展的把握。

（2）定量分析模型呈定式，依赖于现有的软件系统，对于模型的运用灵活性不够，数据采集量不足。

3　数据采集

从 CNKI 上下载 1980 年−2018 年所有篇名以及摘要中含有“隐喻”的文献，作为实验的初始数据，获取其摘要、作者、年份、出版刊物等信息。数据具体细节示例如表 1 所示。最终共查询期刊数 2 206，得到文献数量 14 644 篇。

表 1　关键词对比的文章示例

Table 1　Examples of Articles with Keyword Comparisons

摘要	候选关键词	文章关键词
特别是在**词汇**多义研究领域，认知**语言**学致力于……概念**隐喻**理论……英语教师的**教学**策略……[19]。	隐喻；语言；教学；词汇	原则性多义网络；英语介词；中国英语学习者；习得
根据弗雷德里克·詹姆逊的看法，**文学**文本是……利用家庭关系作为社会秩序的**隐喻**……在**空间**流动中面对诱惑……具有特殊的象征**功能**[20]。	隐喻；空间；功能；文学	乔纳森·弗兰岑；伦理责任；新教伦理；自由市场
他精辟地指出……“不合身的长袍”的**隐喻**……喻义层面看……现实**政治**是所指[21]。	隐喻；意义；政治；分析	伊格尔顿；《麦克白》；名分；能指；所指
由于中外在语言、**思维**以及文化之间存在诸多差异……通过字词**结构**和音韵的组合……探讨**小说**中蕴含的民间信仰以及文化**隐喻**的翻译[22]。	隐喻；小说；结构；思维	法译；格非；《人面桃花》；古典性
纳博科夫在**转喻**与**隐喻**的强烈**对比**中展示着故乡的美好与现今的漂泊感[23]。	隐喻；转喻；诗歌；对比	纳博科夫；爱情；自我意识

根据得到的文献筛选出候选关键词，本文使用一种用于文本的基于图的排序算法TextRank[18]选出候选关键词。

为验证使用算法所得关键词作为本实验研究的基础比文章作者所提供的关键词更合理，增加对比实验验证：从使用算法得到的候选关键词中随机选择4个作为检索词，在中国知网上搜索文章，为保证所得文章为实验所采用的文章，限定搜索范围为“摘要”，时间限定“1980-2018”。关键词对比的文章示例如表1所示，文章所给关键词与实验的目的隐喻趋势研究无关，但是算法抽出来的关键词对研究趋势分析任务有价值。

由于文章作者提供的关键词是反映作者该篇文章主观想要表达和说明的主题，而本实验与文章本身内容无关，用抽取算法得到的关键词才能客观反映整篇文章的统计信息。比较使用抽取算法得到的关键词和文章原有的关键词可知，抽取算法找的关键词用于隐喻趋势分析更合理。

本文所采用的TextRank算法是抽取式自动文摘的重要方法，通过提取文档中已存在的关键词和句子，根据词之间的共现关系构造无向有权边网络，从而形成摘要。TextRank文本排名算法和其他算法有一个明显区别：其他算法都要基于现成的文本库，而文本排名算法可以脱离语料库，仅仅对单篇文章进行分析即能提取该文档的关键词。因此本文采取该算法为实验筛选候选关键词，计算方法如公式(1)所示。V_i，V_j分别表示句子中两个独立的单词，d为阻尼系数常量，$out()$，$in()$为文本的有向有权图中，节点的出度和入度，ω_{ji}为连接两个节点的边的权重。

$$WS(V_i)=(1-d)+d\times\sum_{V_j\in in(V_i)}\frac{\omega_{ji}}{\sum_{V_k\in out(V_j)}\omega_{jk}}WS(V_j) \quad (1)$$

(1)将每篇文献映射为关键词集合，主要处理方法：将给定文本按照整句进行分割，构成句子集合T；对于每个T中的句子进行分词和词性标注，剔除停用词，保留如名词、动词、形容词等构成词集合S；构建词图G，其节点全部为词集合S中的元素，采用共现关系构造任意两个节点之间的边，迭代计算公式(1)直至收敛得到各个节点的权重。

(2)将所有关键词集合映射为一个文本表示。对节点关键词的权重进行倒序排序，实验中选取Top-100作为候选关键词的集合。在原始文本中标记得到的关键词，若相邻，则作为关键词组提取出来。

4 隐喻研究的趋势分析

4.1 特征选择

本文采用历时和共时的分析方法以及回归预测模型，其中特征选择的有效性直接影响模型的预测精度，从而影响本实验中隐喻研究热点预测的准确率。

为得到有效的特征作为自变量带入计算，根据科学计量学原理[24]，基于文献做出如下假设：

(1)热点应该来自论文的关键字，而不是一般的普通词汇，因为概括了论文的研究方向、内容；

(2)热点应该来自交叉学科的研究，而不只局限于一门学科的研究，所有学科都在研究的课题理应为最热的研究方向；

(3)热点应该来自新作者的介入，如果第一次发表论文就为该课题的研究，可见该课题研究数正在增加；

(4)热点应该来自基金的支持，及热点应是有研究价值的方向；

(5)热点应该与存在年限呈现反比趋势，一般情况下，不再被提及的关键词不应该是刚刚流行的词；

(6)热点应该体现出前瞻性。

综上所述，实验最终确定6个特征作为隐喻热点预测分析的主要变量，如表2所示。

表2 隐喻研究趋势的主要特征

Table 2 Key Features of Metaphor Research Trends

特征	特征说明
F1	候选关键词t在第i年作为关键词的频数
F2	候选关键词t在第i年出现在文档摘要中的频数
F3	候选关键词t截止到第i年出现过的存在年限
F4	候选关键词t在第i年中新作者的论文提及数
F5	候选关键词t在第i年中的重点基金论文提及数
F6	候选关键词t在第i年中期刊提及数

其中，F3候选关键词t截止到第i年出现过的存在年限是指第i年度到词t第一次被提及的总年数；F4指在第i年中，新作者的论文里提及该候选关键词的篇数，F5为第i年中提及该关键词的重点基金

论文数；F6表示第i年中，文献摘要中包含关键词的期刊数；新作者指在当前年第一次发表论文的作者；重点基金论文定义为获得“国家社会科学基金”和“教育部人文社会科学基金”资助的论文。由于隐喻研究在国内自然科学领域起步较晚，相关的国家自然科学基金类项目数量非常少，所以在定义重点基金论文时，没有包括自然科学基金类项目。

后续实验会对这6个特征选择做有效性检测，以验证本文的假设。

4.2　回归模型

本实验的基本思想是通过选择出的第i年的6个特征作为自变量，计算并预测出第i+1年的候选关键词作为关键词的频数y。在回归预测模型的自变量确定后，接下来要选择一个预测精度最高的回归算法。分别比较了5种回归算法的实验结果，其中有常规的回归方法[25]：逻辑回归（Logistic Regression，LR），决策树回归（Decision Tree Regression，DTR），K近邻回归（K-Neighbors Regression，KNR），以及集成方法：随机森林回归（Random Forest Regression，RFR），梯度上升回归树（Gradient Boosting Regression Tree，GBR）。

为使各个算法都达到最优，本文尝试不同的参数，最后设置了各个回归算法效果最佳的参数。参数选择也是基于不同模型的特点：逻辑回归为概率型非线性回归模型，是研究二分类观察结果y与一些影响因素$(x_1, x_2, \cdots, x_n)$之间关系的一种多变量分析方法。假设待求解的目标值为$h_\theta(x)$，影响目标值的因素分别表示为θ_i，则逻辑回归的假设函数如公式(2)-公式(3)所示。

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}} \tag{2}$$

$$g(z) = \frac{1}{1 + e^{-z}} \tag{3}$$

其中，$\theta^T x = \sum_{i=0}^{n} \theta_i x_i = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$。

本文的回归树模型具有可解释性以及分类速度快的优点，但是容易过拟合，为解决这个现象，需要对决策树进行修剪，决策树的剪枝要考虑全局最小选择。随机森林是在决策树的基础上综合多个决策的结果，可提高分类的灵敏度和准确度。由于每棵树都会完整成长而不会剪枝，将生成的多棵分类树组成随机森林。通过不断尝试后，设定20个决策树作为RFR的最佳参数。除了随机森林算法外，梯度上升回归树也是多决策树组合模型。这是一种迭代的决策树算法，它的每一棵树是从之前所有树的残差中来学习，所有树的结论累加起来做最终答案并加入了Boosting这一项以防止过拟合。GBR的优点是可处理不同类型的数据，预测能力强，通过健壮的Loss函数对空间外的异常点处理很好。GBRT的缺点是扩展性不好，因为Boosting天然就是顺序执行的，很难并行化。在不断测试后，最终设定100个决策树作为GBR的参数。

4.3　评价指标

为衡量比较哪种算法的预测精度最高，引用平均绝对值误差（Mean Absolute Error，MAE）和R^2决定系数（R^2-score）两种评价指标作为衡量标准，计算方法如公式(4)-公式(5)所示。

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \tag{4}$$

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{sample}-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{sample}-1}(y_i - \bar{y})^2} \tag{5}$$

平均绝对误差是指预测值与真实值之间平均相差多大。其中，f为预测值，y为真实值$e = |f - y|$即是绝对误差。

R^2衡量的是回归方程整体的拟合度，表达因变量与所有自变量之间的总体关系。公式(5)中，分母理解为原始数据的离散程度，分子为预测数据和原始数据的误差，二者相除可以消除原始数据离散程度的影响。由于R^2是无量纲系数，有确定的取值范围(0-1)，便于对不同回归模型拟合优度进行比较，很适合本文实验的研究分析。

4.4　实验结果

候选关键词每年抽取100条，共抽取近37年文献，实验数据为3 700条。从实验数据中随机选择80%作为训练集，20%作为测试集，并做5折交叉验证。

至此已经准备好所有的实验条件，对于每一个候选关键词，都要执行上述回归模型。同时加强检验，利用历史数据作为已知数据，检验模型的预测精度，计算平均绝对值误差和R^2相关系数。不同回归

模型的实验结果如表3所示。

表3　不同回归模型的实验效果

Table 3　Experimental Results of Different Regression Models

模型	*MAE*	R^2决定系数
LR	14.785	0.790
DTR	9.862	0.909
KNR	8.711	0.938
RFR	8.072	0.952
GBR	**7.803**	**0.964**

从表3中可以明显地比较出GBR回归模型具有显著优势。其*MAE*误差值最小为7.803，R^2决定系数最高为0.964，拟合程度最好。

此外，为体现所选特征的有效性，在最好的回归模型即梯度上升回归树模型的基础上进行特征消融实验，具体实验结果如表4所示。其中Baseline方法使用F1与F2特征。表4显示结果和假设一致，每一个所选特征都是有效的，实验的合理性也得以验证。

表4　特征消融实验结果

Table 4　Feature Ablation Experiment Results

特征选择	*MAE*	R^2决定系数
Baseline	8.735	0.859
Baseline+F3	8.247	0.913
Baseline+F3+F4	7.959	0.894
Baseline+F3+F4+F5	8.091	0.950
Baseline+F3+F4+F5+F6	**7.803**	**0.964**

5　隐喻热点的历时分析

分析隐喻热点的历时性，即研究热点词自身是如何发展，如何随历史的变化而进化。为方便分析不同类型频数的演变，除去F3的年限变化。为更具有普遍性，从Top-100候选关键词中随机选取4个关键词分别为“隐喻”、“小说”、“情感”、“语料库”绘制从1980年-2017年的其余5个特征变化曲线，展示关键词在这一时间段内的被提及次数以及作为关键词摘要的时间变化趋势，这5个特征可以体现关键词的热度变化。由于关键词的频数F1的变化最能体现关键词的热度变化，但是Top1-4的F1特征远超过其他候选关键词，为更明显地体现关键词的历时变化，图5是Top5-35候选关键词F1特征随时间变化趋势集合图。本文实验结果如图1–图4所示。

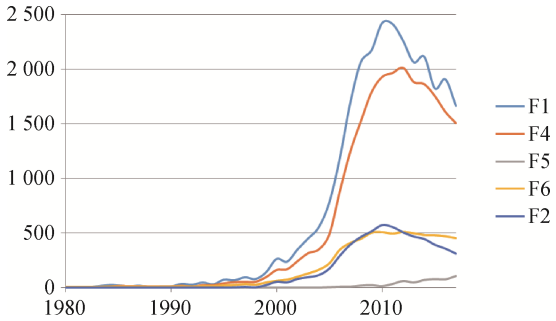

图1　Top-1 隐喻

Fig.1　Top-1 Metaphor

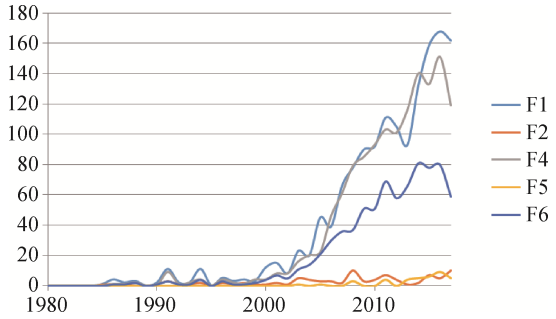

图2　Top-13 小说

Fig.2　Top-13 Novel

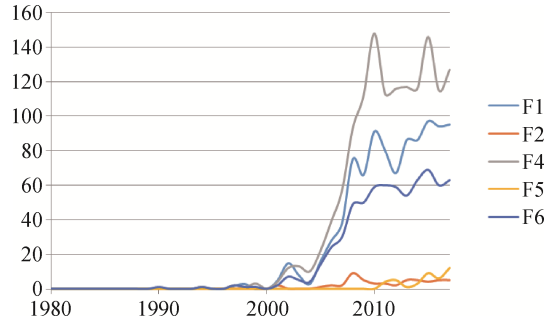

图3　Top-43 情感

Fig.3　Top-43 Sentiment

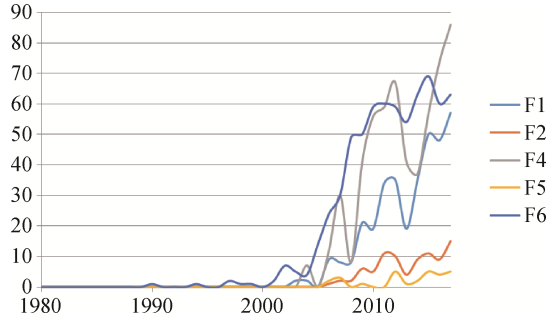

图4　Top-75 语料库

Fig.4　Top-75 Corpus

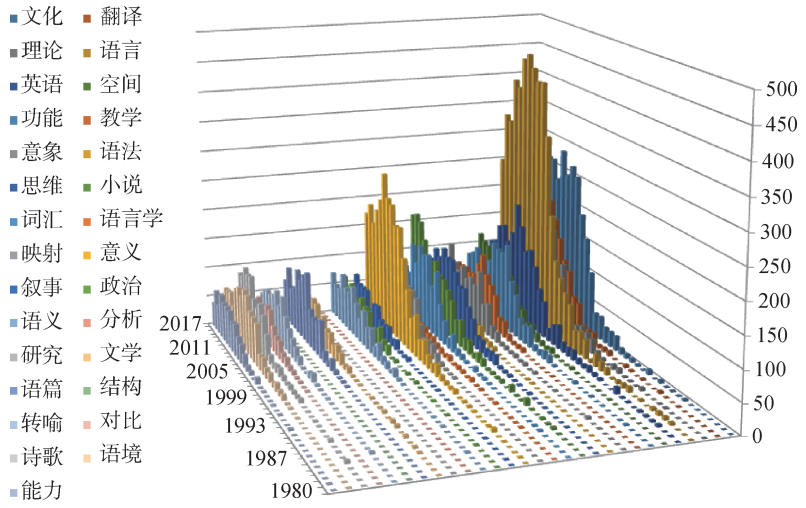

图 5　Top5-35 候选关键词 F1 特征随时间变化趋势集合

Fig.5　Change Trend Collection of Top5-35 Candidate Keyword F1 Features Over Time

整体来看，隐喻研究的历程可以以 2000 年为界限划分为沉默期和增长期两个时间段。从图 5 可以看出，从 1980 年-2000 年所有关键词的特征增长都十分缓慢且基数小，研究活跃度十分低。但是从 2000 年以后，F1 关键词在关键词 i 中出现的频数和 F4 热点词 i 的新作者提及数的增长趋势尤为显著，说明热点词的热度迅速增加，已经进入增长期。这一结果说明隐喻在这一时期已经成为研究热点并被广大学者所重视。其结果与本文所选取热点词的目的一致。

本文还对热点词的历时变化进行了分析。图 1 “隐喻”这一热点词虽然在数量上远远超过其他热点词，在 2010 年左右 F1 特征“隐喻”作为关键词的频数高达 2 420 次，随后呈现下降趋势但总体还是热度最高的词语。与之不同的是图 4“语料库”这一热点词虽然总体频数低，但其热度始终呈波动上升趋势，而且从 F4 和 F6 特征曲线上升趋势可见“语料库”越来越受新作者喜爱并且在跨学科论文中提及数也越来越多。由此可预测在隐喻研究中不只局限于“隐喻”的语言学领地，其研究正在逐渐向其他领域扩展。

6　隐喻热点的共时分析

研究隐喻这一课题，还应该在某一特定的时间中去考察它们部分与部分间的关系，部分与整体间的关系，所以隐喻的“现时”关系尤为重要。因此除了历时性分析，还应考虑隐喻的共时性分析，从而更好地了解当今隐喻的发展状况和预测隐喻研究热点。将每年不同热点词的 F1 特征绘制为隐喻热点共时分析示意图如图 6 所示。

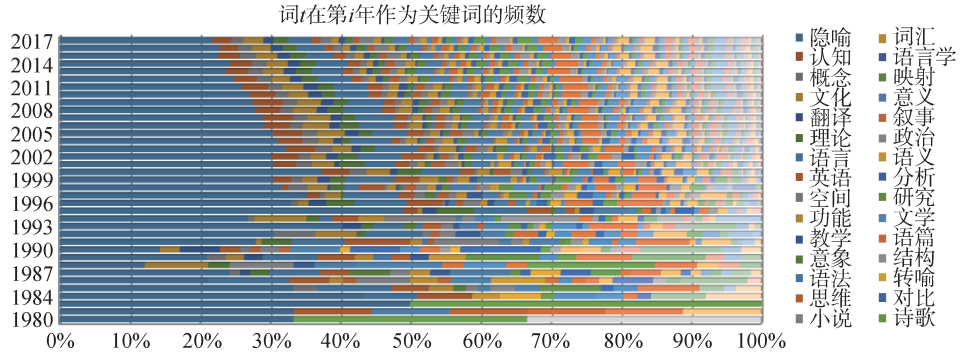

图 6　隐喻热点共时分析示意

Fig.6　Synchronic Analysis of Metaphorical Hotspots

图6中显示的不同颜色对应该年份出现的热点词,因此年份中显示的颜色种类越多,该年份的热点词越多。在1980年隐喻研究刚刚起步时,除了隐喻本身这一热点词以外,图中仅显示两种不同颜色。因此根据这两种颜色所代表的热点词将隐喻研究划分为两个主题:影视和新闻。由此可见,此时隐喻研究中作为关键词的热点候选词较少,并且由于互联网和新媒体尚未兴起,其研究热点与电视、广播、报纸等当时占主流的传统媒体[26]联系十分紧密。然而,这种情况在2017年显示出显著的不同,与1980年仅仅出现两个热点词不同,在2017年的隐喻研究中有150个候选关键词出现在关键词中。这一结果可以验证在历时分析中的假设,即隐喻研究不再局限于个别主题和领域,正在向多种学科以及多个领域扩展。

7 结 语

不同于目前人工梳理以及依赖现有软件分析预测隐喻研究的趋势,本文采用一种文本挖掘技术下的关键词抽取算法对隐喻研究论文进行分析和处理,提出6种有效的特征选择参数,揭示隐喻研究的发展历程和演变规律。对比5种回归模型不同预测结果,并从历时和共时维度对隐喻研究的热点预测做出分析。从历时性分析中可以看出隐喻研究的热点词在不同阶段的变化趋势,即国内隐喻研究始于20世纪80年代初,此后发展缓慢,直到进入21世纪后,无论是新作者的提及数还是出现在关键词中的频数、关键词的热度呈现突增的趋势,至今仍持续增加。对共时性分析中,实验结果表明,在隐喻研究早期,整个研究课题中出现的热点词比较集中。但是当下的隐喻研究中热点词彼此区分度不大,隐喻研究正在向多领域、多学科发展。

参考文献:

[1] Johnson M, Lakoff G. Metaphors We Live by[M]. Chicago: University of Chicago Press, 1980.

[2] Cameron L. Metaphor in Educational Discourse[M]. London: Continuum, 2003.

[3] Steen G, Dorst A G, Herrmann J B, et al. A Method for Linguistic Metaphor Identification: From MIP to MIPVU[M]. John Benjamins Publishing, 2010.

[4] Shutova E, Teufel S. Metaphor Corpus Annotated for Source-Target Domain Mappings[C]// Proceedings of the 7th Conference on International Language Resources and Evaluation. 2010: 3255-3261.

[5] Fainsilber L, Ortony A. Metaphorical Uses of Language in the Expression of Emotions[J]. Metaphor and Symbolic Activity, 1987, 2(4): 239-250.

[6] Kovecses Z. Metaphor: A Practical Introduction[M]. New York: Oxford University Press, 2010.

[7] Averill J R. Inner Feelings, Works of the Flesh, the Beast Within, Diseases of the Mind, Driving Force, and Putting on a Show: Six Metaphors of Emotion and Their Theoretical Extensions[J]. Metaphors in the History of Psychology, 1990, 2: 104-132.

[8] Thibodeau P H, Boroditsky L. Metaphors We Think with: The Role of Metaphor in Reasoning[J]. PLoS One, 2011, 6(2): e16782.

[9] Malinowski J E, Horton C L. Metaphor and Hyperassociativity: The Imagination Mechanisms Behind Emotion Assimilation in Sleep and Dreaming[J]. Frontiers in Psychology, 2015, 6: 1132.

[10] Jabbi M, Bastiaansen J, Keysers C. A Common Anterior Insula Representation of Disgust Observation, Experience and Imagination Shows Divergent Functional Connectivity Pathways [J]. PLoS One, 2008, 3(8): e2939.

[11] 林鸿飞,张冬瑜,杨亮,等. 情感隐喻计算及其应用研究[J]. 大连理工大学学报, 2015, 55(6): 661-670. (Lin Hongfei, Zhang Dongyu, Yang Liang, et al. Computational Processing of Affective Metaphors and Its Applications[J]. Journal of Dalian University of Technology, 2015, 55(6): 661-670.)

[12] 田嘉, 苏畅, 陈怡疆. 隐喻计算研究进展[J]. 软件学报, 2015, 26(1): 40-51. (Tian Jia, Su Chang, Chen Yijiang. Computational Metaphor Processing[J]. Journal of Software, 2015, 26(1):40-51.)

[13] 束定芳. 隐喻学研究[M]. 上海: 上海外语教育出版社, 2000. (Shu Dingfang. Studies in Metaphor[M]. Shanghai: Shanghai Foreign Language Education Press, 2000.)

[14] 林书武. 隐喻研究的基本现状、焦点及趋势[J]. 外国语, 2002(1): 38-45. (Lin Shuwu. Studies on Metaphor: State of Arts, Focuses, and Trend[J]. Journal of Foreign Languages, 2002(1): 38-45.)

[15] 贾玉祥, 俞士汶. 基于词典的名词性隐喻识别[J]. 中文信息学报, 2011, 25(2): 99-105. (Jia Yuxiang, Yu Shiwen. Nominal Metaphor Recognition Based on Lexicons[J]. Journal of Chinese Information Processing, 2011, 25(2): 99-105.)

[16] 张冬瑜, 杨亮, 郑朴琪, 等. 情感隐喻语料库构建与应用[J]. 中国科学: 信息科学, 2015, 45(12): 1574-1587. (Zhang Dongyu, Yang Liang, Zheng Puqi, et al. Construction and Application of Affective Metaphor Corpus[J]. Scientia Sinica: Informationis, 2015, 45(12): 1574-1587.)

[17] 孙亚, 钱玉彬, 马婷. 国外隐喻研究现状及发展趋势[J]. 现代外语, 2017, 40(5): 695-704. (Sun Ya, Qian Yubin, Ma Ting. The

Current Status and Development Trends in Metaphor Studies[J]. Modern Foreign Languages, 2017, 40(5): 695-704.)

[18] 方俊伟, 崔浩冉, 贺国秀, 等. 基于先验知识TextRank的学术文本关键词抽取[J]. 情报科学, 2019, 37(3): 75-80.(Fang Junwei, Cui Haoran, He Guoxiu, et al. Keyword Extraction of Academic Text with TextRank Model Based on Prior Knowledge[J]. Information Science, 2019, 37(3): 75-80.)

[19] 李亚平. 基于原则性多义模式的中国英语学习者空间介词Under, Below, Beneath习得研究[D]. 临汾: 山西师范大学, 2017. (Li Yaping. A Study of the Acquisition of English Prepositions Under, Below and Beneath by Chinese EFL Learners Based on Principled-polysemy Model[D]. Linfen: Shanxi Normal University, 2017.)

[20] 陆赟. 乔纳森·弗兰岑小说中的伦理责任主题研究[D]. 南京: 南京大学, 2017. (Lu Yun. Ethical Responsibility in Jonathan Franzen's Fiction[D]. Nanjing: Nanjing University, 2017.)

[21] 张薇. 伊格尔顿对《麦克白》的政治符号学解读[J]. 国外文学, 2017(4): 41-50, 154.(Zhang Wei. Eagleton's Political Semiotic Interpretation of Macbeth[J]. Foreign Literatures, 2017(4): 41-50, 154.)

[22] 甘露. 格非《人面桃花》法译本研究[D]. 南京: 南京大学, 2016. (Gan Lu. A Study of the French Version of Ren Mian Tao Hua [D]. Nanjing: Nanjing University, 2016.)

[23] 刘德伟. 纳博科夫小说爱情主题初探[D]. 济南: 山东师范大学, 2008.(Liu Dewei. On the Theme of Love in Nabokov's Novels [D]. Ji'nan: Shandong Normal University, 2008.)

[24] 郭涵宁. 新兴研究领域识别计量[M]. 北京: 科学出版社, 2017. (Guo Hanning. Measurement of Identifying Emerging Research Areas[M]. Beijing: Science Press, 2017.)

[25] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.(Zhou Zhihua. Machine Learning[M]. Beijing: Tsinghua University Press, 2016.)

[26] 郭战江. 新媒体:角色和路径——从互联网的迅猛发展思考传统广播电视台的转型[J]. 中国广播, 2015(1): 32-34.(Guo Zhanjiang. New Media: Roles and Paths——Thinking About the Transformation of Traditional Broadcasting Stations from the Rapid Development of the Internet[J]. Chinese Broadcasts, 2015 (1): 32-34.)

作者贡献声明:

张冬瑜,林鸿飞:提出研究思路,设计研究方案;
胡绍翔,崔紫娟:进行实验;
顾丰,胡绍翔:采集、清洗和分析数据;
张冬瑜,崔紫娟,张伟:论文起草;
顾丰,林鸿飞:论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储,E-mail: zhangdongyu@dlut.edu.cn。
[1] 林鸿飞, 张冬瑜. metaphoressay.xlsx. CNKI隐喻论文数据.
[2] 林鸿飞, 张冬瑜. keywords.txt. CNKI论文关键词数据.

收稿日期:2019-07-30
收修改稿日期:2019-11-23

Reviewing Metaphor Research Based on Keyword Extraction Algorithm

Zhang Dongyu[1] Gu Feng[1] Cui Zijuan[2] Hu Shaoxiang[1] Zhang Wei[1] Lin Hongfei[3]

[1](School of Software, Dalian University of Technology, Dalian 116620, China)

[2](International Office, Dalian University of Technology, Dalian 116024, China)

[3](School of Computer Science and Technology, Dalian University of Technology, Dalian 116023, China)

Abstract: [**Objective**] This paper reviews the developments of metaphor research in China in the past 40 years, aiming to provide references for linguists and then narrow the gaps between Chinese and foreign researchers. [**Methods**] First, we used keywords extraction algorithms to map metaphor documents into keyword sets. Then, we chose effective features as parameters for the regression models, which helped us predict the frequency of trending words in the next year. Finally, we analyzed the developments of metaphor research diachronically and synchronously. [**Results**] Our study compared the results of five regression models. Among them, the GBR model with the best fitting degree had the highest prediction accuracy for next year's trending words. The feature ablation experiment also confirmed that our selected features were effective. [**Limitations**] The accuracy of keyword extraction algorithm could be optimized. [**Conclusions**] Metaphor research is developing towards the direction of cross-domains and inter-disciplines. The method of feature selection provides more references for research in prediction models.

Keywords: Metaphor Trend Prediction Regression Model Diachronic Analysis Synchronic Analysis

在不屏蔽偏见的情况下如何放大社交媒体上值得信赖的新闻内容

社交媒体网站往往会放大并发酵错误信息和阴谋论。来自南佛罗里达大学的跨学科团队开发了一种用户内容推荐算法，能够让社交媒体用户接触到更可靠的新闻来源。研究人员没有根据用户数量和浏览量来衡量参与度，而是着眼于新闻源中放大了哪些内容，重点关注新闻来源的可靠性得分和受众的政治多样性。

“低质量的内容很有吸引力，因为其符合人们已经知道和喜欢的内容，无论它是否准确，”研究人员说：“因此，错误信息和阴谋论经常在大众中快速传播。现有的内容推荐算法选择了错误的信号并继续推广这些错误信息和阴谋论。为了打破这一循环，应该寻找高质量的引人入胜的内容，应该考虑多样化的受众群体。”

研究人员使用由在线民意调查公司YouGov提供的6 890名用户的党派关系数据和网络流量数据设计了一种新算法，这些用户来自美国不同性别、不同种族和不同政治派别。此外，研究人员还根据NewGuard可靠性指数审查了3 765个新闻来源的可靠性得分。结果发现，结合新闻受众的党派多样性可以提高推荐来源的可靠性，同时仍然保持推荐内容的相关性。“这对于社交媒体平台来说是好消息，社交媒体一直不愿对推荐算法进行更改，因为害怕被批评存有党派偏见。”研究人员建议社交媒体平台采用本研究提出的这种新策略，以防止错误信息的广泛传播。

（编译自：https://www.sciencedaily.com/releases/2022/02/220203122858.htm）

（本刊讯）