# 科技期刊热词评估指标构建及知识服务[1]

常宗强[1,2)]，刘蔚[1,2)]，侯春梅[1,2)]，叶喜艳[1,2)*]，张静辉[1,2)]，陶华[1,2)]

1）中国科学院西北生态环境资源研究院，甘肃 兰州，730000

2）甘肃省知识计算与决策智能重点实验室，甘肃 兰州，730000

**摘要** 【目的】构建期刊热词评估指标，并通过热词排行榜服务，帮助用户快速直观地了解期刊研究的前沿领域和方向。【方法】利用网络文献调研法分析当前智能分词词库析出词条特征；利用参数化、标准化唯一标记等方法提取与热词关联的计算参数，并进行统计分析，构建数学模型，进行多维度排序。【结果】构建了期刊有效词条析出模型，对有效词条进行筛选和标准化标记，对带有参数信息的有效词条通过逻辑计算构建了热词评估指标数学模型，并以排行榜的形式给出热词指标知识服务的一个应用示例。【结论】标准化唯一标记方法可提升分词词库的词条识别能力，使分词结果更加专业、可靠；期刊热词排行榜服务，可帮助用户快速、直观地了解期刊研究的前沿领域和方向。

**关键词：**科技期刊；评估指标；知识服务；智能分词



大数据时代，期刊的发展离不开数据库的支撑，期刊联动数据库的数字化出版模式已成为期刊出版发展的趋势之一。随着数字化出版技术的发展，用户获取知识的效率不断提升，必然地，其对期刊的知识服务能力也要求越来越高。期刊热词作为一种反映期刊发展动态的文本要素，可帮助用户快速获取当前期刊发展前沿的热点信息。因而能否有效提取期刊热词，并进行知识服务应用，是值得研究的课题。

目前，关于热词提取分析的研究主要集中在社会网络信息方面，主要涉及社会热点话题[1]、价值观分析[2-3] 、伦理研究[4-5]等方面，而涉及行业领域的研究也主要是针对大众网络信息的研究，如基于农业网络信息分类的热词自动提取方法[6]，而对科技期刊领域的热词提取研究相对较少。值得一提的是，关于期刊评价

**资助项目：**2023-2024 年度中国科学技术期刊编辑学会基金项目（cessp-2023-B-05）；中国科学院自然科学期刊编辑研究会研究课题（YJH202325）

**作者简介：**常宗强，博士，编审，E-mail: cahngzq@lzb.ac.cn；刘蔚，博士，研究员；侯春梅，硕士，编审；张静辉，硕士，编辑；陶华，硕士，副编审。

***通信作者：**叶喜艳(ORCID: 0000－0001－6466－0428)，硕士，编辑，E-mail: yexy@lzb.ac.cn

的指标有多种，如期刊的影响因子、总被引频次、来源文献量、基金论文比[7]等等，然而关于期刊热词的评价指标，却鲜为人知。另外，在期刊热词的数字化服务技术方面，欧阳柳波等[8]采用一种基于混合判定模型的复合概念抽取方法，通过文本分词、词条降噪和同义词合并等处理，实现了多重复合概念抽取，提升了复合概念词条抽取的准确率和效率；傅士光等[9]提出了一种基于词库的结合词频、词性、中文文法规则和未登录词识别规则的分词算法，该分词算法能够在很大程度上消除歧义划分,提高未登录词的识别概率；中国专利[10]报道了一种面向出版的智能模板模型的建立方法，利用该方法可实现针对异种排版软件的基于模板技术数据的一次整理多次出版功能；中国专利[11]报道了一种数字出版资源语义增强描述系统及其方法，该发明得到的数字出版资源语义增强描述可标识出数字出版资源的基础版权点和语义表述点。可见上述方法均停留在出版功能的实现和智能分词与文本识别方面，并没有从用户的角度出发，挖掘词条的知识服务功能。知识服务广义定义为基于知识资源或知识产品，根据用户的需求和使用场景，融入用户解决问题的过程中，提供能够有效支持知识应用和创新的行为[12,13]，是对已有知识的二次加工和多次衍生。目前，知识服务是科技期刊转型发展的主要方向[14]，因此，基于期刊热词评估指标，进一步挖掘其知识服务功能，可更好地发挥科技期刊的知识服务能力，从而扩大并提升期刊的品牌影响力。

基于上述分析，本文主要聚焦于期刊热词评估指标的构建，根据期刊有效词条的析出特征，提取影响词条热度的信息变量，并进行参数化和标准化标记，通过设计底层计算逻辑构建数学模型，提出一种多维度衡量期刊热词的指标，为期刊热词的评价提供一种新的思路，进一步挖掘其知识服务功能， 帮助用户快速、直观地了解期刊研究的前沿领域和方向。此外，热词指标的应用也可为编辑人员提升期刊的知识服务能力提供路径借鉴，可为出版行业技术人员提升热词的专业性和可靠性提供模型参考。

## 1 研究方法

主要利用网络文献调研法分析目前智能分词词库析出词条特征，如词条的有效性、专业性、完整性等，基于目前存在的词条析出的问题，通过与定期维护更新的专业基础词条库对比匹配的方法，提取有效词条；再根据热度参数调研情况，利用参数化、标准化唯一标记等方法提取与热词热度相关的计算参数，并对数据

统一进行标准化处理和统计分析。在此基础上，通过分析各参数与期刊热词热度的相关性，利用指标构建法构建期刊热词评估指标数学模型，最后根据各参数和指标特征，对期刊热词从不同维度进行大小值排序，并以排行榜的形式给出热词指标知识服务的一个应用示例。

## 2 期刊热词评估指标构建

### 2.1 期刊有效词条析出模型

目前，智能分词技术已较为成熟，分词效果也越来越好，但针对学科术语、新词等特色化分词功能还有待进一步提升。本文主要针对科技期刊进行专业化、学术化，因而需要对智能分词技术析出的词条进行筛选，提取有效词条。为满足这一条件，我们主要构建了由基层专业词条库、词条筛选通道、中层智能分词词库、有效词条输出通道、顶层析出词库构成的期刊热词析出模型，如图 1 所示。词条筛选通道关联基层专业词条库和中层智能分词词库；有效词条输出通道关联中层智能分词词库和顶层析出词库。

在上述模型中，基层专业词条库是指内置于期刊平台的与期刊定位高度相关的专业词库；期刊可对该词库进行修改、删除、增加等更新操作，是期刊平台的基础词条库。其中，值得注意的是，期刊需要主动对基础词条库进行及时的增补维护，如新文章析出的新词条、学术界出现的新概念、新术语等，以确保基础词条库的时效性。中层智能分词词库是指利用现有的智能语义分词技术将期刊论文进行语义分词，并对这些语义分词进行临时存储的词库。顶层析出词库是指存储带有标准化唯一标记的有效词条的词库。

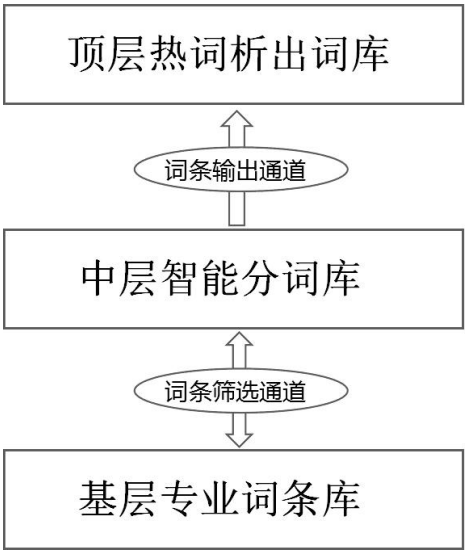

图 1 期刊有效词条析出模型示意图

## 2.2 有效词条的筛选与标准化标记

筛选有效词条主要是为了在众多词条中筛选出满足期刊发展需要的词条。主要通过对比匹配的方法进行筛选：先将中层智能分词词库中临时存储的语义分词与基层专业词条库中的词条进行比对；然后针对不同的比对结果进行不同的标记操作即可。若比对词条 A 在基层专业词条库中存在，则对词条进行标准化唯一标记；若不存在，则不标记，以此达到清洗无效词条的目的。

在筛选出有效词条之后，中层智能分词词库将对有效词条进行标准化唯一标记，该标准化唯一标记的有效词条通过有效词条输出通道输出至顶层析出词库。其中，中层智能分词词库对有效词条进行标准化唯一标记的方法是指利用具有定量化特征的多个参数来标记筛选出的有效词条的方法。比如：

以词条 A 为例，其析出的期刊论文出版年份记为 Y，在这篇论文中出现的频次记为 F，设词条 A 是第 1 次与基层专业词库进行比对，则将此次筛选的词条 A 按年份标记为 ${}^{Y}A_{F1}$，按频次标记为 ${}^{F1}A_{1}$，出现篇次记为 $1({}^{Y}A)$。若为第 2 次比对，则分别标记为 ${}^{Y}A_{F2}$、${}^{F2}A_{2}$ 和 $2({}^{Y}A)$。以此类推，若为第 n 次比对，则将其年份、频次、篇次分别标记为 ${}^{Y}A_{Fn}$、${}^{Fn}A_{n}$ 和 $n({}^{Y}A)$。

## 2.3 有效词条参数信息的逻辑计算

从上节可知，用来标记有效词条 A 的参数有年份、频次、篇次，且已经给出某一出版年单次比对标记的方法。然而词条 A 有可能在同一出版年析出多次，也可在不同的出版年析出，因此需要分别作标记。若顶层热词析出词库中在同一出版年 Y 析出了多个带有年份标记的词条 A，即 ${}^{Y}A_{F1}$，${}^{Y}A_{F2}$，${}^{Y}A_{F3}$，…${}^{Y}A_{Fn}$，则 Y 年词条 A 的出现篇次为 $n({}^{Y}A)$，总出现频次 F(Y)=F1+F2+F3+…Fn，则将带有年份、频次和篇次信息的词条 A 标记为 ${}^{Y}{}_{F(Y)}A_{n}$，如图 2 所示。如 2020 年共有 8 篇论文出现了词条 A，8 篇论文中词条 A 的出现频次分别记为 F1，F2，F3，…，$F_8$，则 2020 年词条 A 的出现篇次为 8，总出现频次 F(Y)=F1+F2+F3+…$F_8$，记为 ${}^{2020}{}_{F(2020)}A_{8}$。

在所有年份中词条 A 的出现篇次记为 $N_A$，出现频次记为 $F_A$，则 $N_A=n1+n2+n3+\ldots+ni$，其中 n1，n2，n3，…，ni 分别为不同年份词条 A 出现的篇次；$F_A=F(Y1)+F(Y2)+F(Y3)+\ldots+F(Yi)$，其中 F(Y1)，F(Y2)，F(Y3)，…，F(Yi)分别为不同年份中词条 A 出现的总频次。

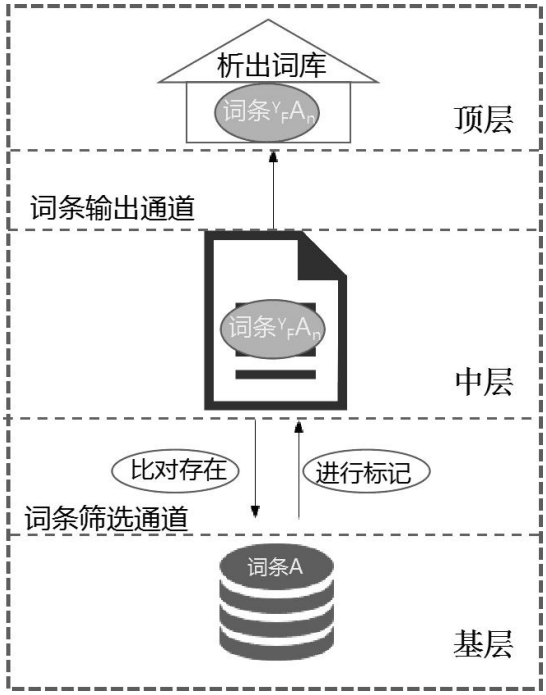


图 2 以词条 A 为例的热词析出模型示意图

## 2.4 词条热度评估指标数学模型

### （1）权重评分评估指标

权重评分评估指标 $S_A$ 主要针对词条 A 在所有年份的出现篇次 $N_A$ 和出现频次 $F_A$ 进行权重赋值计算，设其权重组合为（x，y），即

$$S_A=xN_A+yF_A$$

这里的权重组合（x，y）可根据期刊学科特征灵活调试，本文中，暂且根据二八法则，取 x=0.8，y=0.2。则有:

$$S_A=0.8N_A+0.2F_A$$

式中：$N_A=n1+n2+n3+\ldots+ni$，其中 n1，n2，n3，…，ni 分别为不同年份词条 A 出现的篇次；$F_A=F(Y1)+F(Y2)+F(Y3)+\ldots+F(Yi)$，其中 F(Y1)，F(Y2)，F(Y3)，…，F(Yi)分别为不同年份中词条 A 出现的总频次。

### （2）热度评估指标

考虑到热词随时间的冷却问题，即在权重评分的基础上，结合热词冷却系数R，进一步构建热词热度评估指标。即热词热度主要与权重评分 $S_A$ 和冷却系数 R 有关，根据其相关关系，构建热度初始数学模型为

$$H_A=\alpha S_A/ R$$

式中：α为热度适用的调试参数，根据经验计算，文中取α=8.76。冷却系数 R 的确定遵循如下逻辑：随时间的推移，热词的热度是衰减的。本文令时间推移量$\Delta T= i*b-Y_i$，其中：b 为当前年份；$Y_i$ 为词条出现的所有年份之和，即$Y_i=Y1+Y2+Y3+\ldots+Yn$；i=1，2，3，…，n。则

$$R=(i*b-Y_i+C)^{\beta}$$

式中：C 为常数，确保ΔT=0 时，热度 $H_A$ 仍有效，文中取 C=2。参数β表示热词随时间推移的衰减程度，可根据具体情况进行调试，本文取β=1.2。则最终构建的热度评估指标数学模型为：

$$H_A=8.76S_A/（i*b-Y_i+2）^{1.2}$$

文章主要以年份为时间尺度，为期刊的热词指标构建提供了一种参考思路。虽然有效词条可以设置实时动态更新，但如果按照年份尺度来计算，一些新词则可能不会及时呈现。因此，在热词的基础上，可增加期刊新词模块。如前文所述，底层基础专业词条库需要及时进行增补更新维护。另外，热词的时间尺度也可按照月或者季度来统计，如果不具备统计意义，则可将增补的新词呈现在期刊新词模块。

## 3 基于期刊热词评估指标的知识服务

在得出热词参数信息和评估指标之后，为进一步挖掘其应用价值，我们主要通过排序的方法，以排行榜的形式给出热词指标知识服务的应用示例。这里排序的方法是指按年份、频次和篇次排序，按权重评分排序，按词条热度排序中的任意一种。这 3 种排序方式在顶层热词析出词库中是同时存在的，在具体排序时，参与排序的项目内容可以同时呈现，但是排序方式只能选择其中的一种。比如词条 A 按频次来排序，那么它就不能同时按权重评分或者热度来排序，但可以同时显示权重评分或者热度的值。

### 3.1 有效词条基础参数排行榜

有效词条的排序可按年份、频次和篇次排序，年份区间（a，b）可自由选择，其中 b 为当前年份，a 为 b 之前的年份，a≤b。由此排行榜可查询同一年份（表 1）或者某年份区间中（表 1）不同词条的出现频次 F(Y)和出现篇次 n；在选定年份区间的基础上，可进一步按词条频次或篇次排序。

也可不设定年份，直接按顶层析出词条库中所有词条的出现频次 F 和出现篇次 N 来排序，据此可查询期刊频次和篇次维度的 TOP10，TOP20，TOP50 等的热词排行榜，如表 3 所示。

表 1　同一年份不同词条的当年频次和篇次排序示例

| 年份 | 词条 | Y 年频次 | Y 年篇次 |
|---|---|---|---|
| Y | A | ${}^{Fn}A_n$ | $n({}^{Y}A)$ |
| Y | B | ${}^{Fn}B_n$ | $n({}^{Y}B)$ |
| Y | C | ${}^{Fn}C_n$ | $n({}^{Y}C)$ |
| … | … | … | … |

表 2　不同年份同一词条年频次和年篇次排序示例

| 年份 | 词条 | 年频次 | 年篇次 |
|---|---|---|---|
| Y1 | A | F(Y1) | n1 |
| Y2 | A | F(Y2) | n2 |
| Y3 | A | F(Y3) | n3 |
| … | … | … | … |

表 3　不同词条的总频次和总篇次排序示例

| 词条 | 总频次 F | 总篇次 N |
|---|---|---|
| A | $F_A$ | $N_A$ |
| B | $F_B$ | $N_B$ |
| C | $F_C$ | $N_C$ |
| … | … | … |

### 3.2 热词权重评分和热度排行榜

主要显示不同词条的权重评分 S 和热度 H，可分别按 S 和 H 排序，据此可查询期刊权重评分和热度维度的 TOP10，TOP20，TOP50 等的热词排行榜，如表 4 所示。

表 4　　不同词条的权重评分和热度排行榜示例

| 词条 | 权重评分 | 热度 |
|---|---|---|
| A | $S_A$ | $H_A$ |
| B | $S_B$ | $H_B$ |
| C | $S_C$ | $H_C$ |
| ... | ... | ... |

综上可知，文中共给出了 4 种排行榜形式，从多个角度分别去呈现热词在某一特定维度下的状态，各维度可以互相辅助参考，以尽量做到客观有效。其中，热度指标综合考虑了频次、篇次、权重、时间衰减等要素， 是反映词条热度的重要指标，但受限于实证分析，因而后期还需进一步订正。另外，热词的呈现形式除过排行榜，还可以基于排行榜底层的数据进行绘图，以图件的形式进行呈现。

## 4 优势分析

（1）本文将顶层热词析出词库中带有标记的词条进行相关基础计算，来获得期刊论文中专业词条的频次和篇次，在此基础上，根据词条总频次和总篇次，按照预设公式对词条进行权重评分，再根据权重评分和年份计算词条热度，然后通过频次、篇次、权重评分和热度不同维度的词条排行榜，帮助用户快速、直观地了解期刊的研究动态。

（2）通过构建期刊热词析出模型，引用基层专业词条库和中层智能分词词库，将中层智能分词词库中临时存储的语义分词与基层专业词条库中的词条进行比对，然后针对不同的比对结果进行不同的标记操作，以清洗无效词条，筛选出与本期刊定位相关的专业词条，提高中层智能分词的专业性和有效性。

（3）本文对中层智能分词词库析出的词条进行定量化标记，即利用具有定量化特征的多个参数来标记筛选出的有效词条，从而使筛选出的有效词条具有唯

一标识符特征，提升中层智能分词词库的词条识别能力，使分词结果更加可靠。

（4）期刊热词排行榜主要包括频次、篇次、权重评分和热度不同维度的词条排行榜，可为用户从多个维度呈现期刊热词的变化动态，帮助用户快速掌握期刊研究方向，从而提升期刊对用户的知识服务能力。

## 5 结束语

文章基于内嵌的专业词条库，主要通过参数化方案提取出具有统计学意义的参数变量，再将这些参数变量进行相关性分析，设计底层计算逻辑，构建基础数学模型，最后是调试常量系数，进行模型优化，达到预期效果。文章进一步开展了热词评估指标的转化应用，推出热词排行榜知识服务产品。回顾热词热度评估指标的构建过程，其主要的难点在于参数化方案的确定和底层逻辑的设计。由于该指标缺乏具体案例的实证检验，且不同学科之间可能存在差异，因而下一步计划将热词评估指标进行实证研究，并在不同的学科之间进行对比分析，进一步调参、订正，提升指标的可靠性，并研究其普适性问题。在此基础上，进一步在时间维度上，分析热词之间的相关性，深入挖掘有相关性的热词在研究热度上的转移趋势，有效提升期刊的知识服务能力。

## 参考文献

[1]黄菊. 面向微博热点话题的热词抽取与情感分析[D].合肥：安徽理工大学,2022.

[2]邱红霞,王习胜.近年来“网络热词”的价值观分析[J].牡丹江师范学院学报(哲学社会科学版),2017(05):5-10.

[3]方婷. 网络矛盾热词与受众的社会认同[D].合肥：安徽大学,2019.

[4]周思源. 中国网络热词的伦理研究[D].石家庄：河北经贸大学,2021.

[5]胡青青. 网络热词的伦理研究[D].长沙：湖南师范大学,2015.

[6]段青玲,张璐,刘怡然，等.基于农业网络信息分类的热词自动提取方法[J].农业机械学报,2018,49(7):160-167.

[7]孙晓红,闫立娟,张改侠,等.《地球学报》2005—2018 年主要期刊评价指标变化趋势分析[J].地球学报,2021,42(1):124-128.

[8]欧阳柳波,邹北骥,刘丽杰.一种基于混合判定模型的复合概念抽取方法[J].电子学报,2013,41(3):488-495.

[9]傅士光,林友芳,万怀宇，等. 一种基于规则的中文分词算法[C]//中国中文信息学会,新加坡

中文与东方语言信息处理学会,武汉大学语言与信息研究中心.中国计算技术与语言问题研究——第七届中文信息处理国际会议论文集.电子工业出版社,2007:52-56.

[10] 王荣,李平立,龚健. 一种面向出版的智能模板模型的建立方法[P]. 北京市：CN100392654C,2008-06-04.

[11]陈琳,谢冰,卢朋，等. 一种数字出版资源语义增强描述系统及其方法[P]. 北京市：CN102999487B,2015-06-24.

[12] 新闻出版知识服务知识资源建设与服务基础术语：GB/T 38377—2019 [S].北京：中国标准出版社，2019

[13] 新闻出版知识服务知识资源建设与服务工作指南：GB/T 38382—2019 [S].北京：中国标准出版社，2019

[14]郭雨梅,景勇,郭晓亮等.开放科学形势下科技期刊知识服务平台运营模式探析[J].编辑学报,2023,35(3):273-278.

作者贡献声明：

常宗强，叶喜艳：提出研究方向，调研整理文献，撰写论文；

刘蔚，侯春梅：指导撰写思路，审核、修订论文；

张静辉，陶华：修订论文。

**Construction and application of evaluation indicators for hot keywords in journals based on term banks**

CHANG Zongqiang[1,2)], LIU Wei[1,2)], HOU Chunmei[1,2)], YE Xiyan[1,2)*], ZHANG Jinghui[1,2)], TAO Hua[1,2)]

1) Northwest Institute of Eco-Environment and Resources， Chinese Academy of Sciences，Lanzhou 730000，China

2) Key Laboratory of Knowledge Computing and Intelligent Decision ， Gansu Province，Lanzhou 730000，China

**Abstract:** [**Purposes**] To construct evaluation indicators for hot words in journals, and through the hot word ranking service, help users quickly and intuitively understand the cutting-edge fields and directions of journal research. [**Methods**] Using online literature research method to analyze the features of extracted entries in the current intelligent word segmentation lexicon; Using methods such as parameterization and standardized unique labeling to extract computational parameters associated with hot words, conducting statistical analysis, constructing mathematical

models, and conducting multidimensional sorting. [**Findings**] A journal effective entry extraction model was constructed, and effective entries were screened and standardized. A mathematical model for hot word evaluation indicators was constructed through logical calculation for effective entries with parameter information. An application example of hot word indicator knowledge service was presented in the form of a ranking list. [**Conclusions**] The standardized unique labeling method can improve the entry recognition ability of the segmentation lexicon, making the segmentation results more professional and reliable; The journal hot word ranking service can help users quickly and intuitively understand the cutting-edge fields and directions of journal research.

**Keywords:** Science and technology journals; Evaluation indicators; Knowledge services; Intelligent segmentation