doi:10.3772/j.issn.2095-915x.2022.03.001

语义知识驱动的论文摘要关键词抽取方法

段建勇[1,2,3] 鲁朝阳[1,3] 王昊[1,2,3] 李欣[1,3] 何丽[1,3]

1. 北方工业大学信息学院 北京 100144；
2. 富媒体数字出版内容组织与知识服务重点实验室 北京 100038；
3. 北方工业大学 CNONIX 国家标准应用与推广实验室 北京 100144

开放科学（资源服务）标识码（OSID）

摘要：[目的 / 意义] 关键词抽取技术可以帮助用户从海量文本中快速定位核心内容，对情报收集工作有着重要意义。目前，关键词抽取主要依靠词频和共现关系，忽视了知识库对关键词抽取的指导作用。[方法 / 过程] 本文提供了一种融合知识的关键词抽取方法，首先基于义原和词林构建词汇知识图谱，其次结合词语的共现关系，生成新的概率转移矩阵，最后实现关键词抽取。[结果 / 结论] 基于海量摘要数据集的实验表明，融合知识的关键词抽取方法，能有效提高现有关键词抽取方法的性能。

关键词：关键词抽取；融合知识；义原；词林

中图分类号：G35；TP391

A Semantic Knowledge-Driven Keyword Extraction Method for Paper Abstracts

DUAN Jianyong[1,2,3] LU Zhaoyang[1,3] WANG Hao[1,2,3] LI Xin[1,3] HE Li[1,3]

1. School of information, North China University of Technology, Beijing 100144, China;
2. The Key Laboratory of Rich-Media Knowledge Organization and Service of Digital Publishing Content, Beijing 100036, China;
3. CNONIX National Standard and Promotion Laboratory, North China University of Technology, Beijing 100144, China

Abstract: [Objective/Significance] Keyword extraction technology can help users quickly locate core content from massive short texts, which is of great significance to intelligence collection. At present, keyword extraction mainly relies on word frequency and co-occurrence relationship, ignoring the guiding role of the knowledge base in keyword extraction. [Methods/Process] This

基金项目 国家自然科学基金项目“基于多源特征学习的中文查询纠错方法研究”（61672040）；“面向新闻事件的查询时效性计算模型研究”（61972003）；富媒体数字出版内容组织与知识服务重点实验室开放基金“垂直领域知识图谱构建关键词技术研究”（ZD2021-11/05）；北京市教育委员会科学研究计划项目资助（KM202210009002）。

作者简介 段建勇（1978-），博士，教授，研究方向为人工智能（自然语言理解）、知识图谱、文本挖掘；鲁朝阳（1996-），硕士研究生，研究方向为自然语言处理、知识图谱；王昊（1980-），博士，副教授，研究方向为自然语言处理、知识图谱，E-mail: wanghaomails@gmail.com；李欣（1991-），博士，讲师，研究方向为高维非均匀偏移数据特征提取、优先采样学习算法、知识抽取；何丽（1976-），副教授，研究方向为数据仓库及数据挖掘、知识工程、数字版权服务等。

引用格式 段建勇 , 鲁朝阳 , 王昊 , 等 . 语义知识驱动的论文摘要关键词抽取方法 [J]. 情报工程 , 2022, 8(3): 3-12.

doi:10.3772/j.issn.2095-915x.2022.03.001

article provides a method of keyword extraction that integrates knowledge. First, build a vocabulary knowledge graph based on the original meaning and the word forest, and then combine the co-occurrence relationship of the words to generate a new probability transition matrix, and finally realize the keyword extraction. [Results /Conclusions] Experiments based on massive abstract data sets show that the keyword extraction method based on fusion knowledge can effectively improve the performance of existing keyword extraction methods.

Keywords: Keyword extraction; fusion of knowledge; sememe; cilin

引言

当今社会信息高速发达，每天发表在社交平台上的内容不计其数，其中绝大部分都是短文本。对于时间有限的人们来说，想要快速检索自己感兴趣的内容就成了一个亟待解决的问题。关键词可以帮助读者快速了解一篇文档的主题内容，通过关键词抽取技术可以在海量的文本中将人们最关心的问题提取出来。关键词通常为一个或多个能够描述文档主题信息的词语或者词组[1]。早期文章的关键词主要靠的是人工标注，这导致关键词标注工作既费时又费力。随着计算机技术的发展，越来越多的机构和个人开始研究关键词抽取技术，已经有不少方法在关键词抽取领域取得了较好的效果。随着在自然语言处理领域对关键词抽取方法研究的逐步深入，在多项文本挖掘任务例如文本摘要、文本分类中都发挥了重要的作用。比如从最近一天所有用户发表的微博中提取出关键词，就可以知道当天人们最关心的问题。但是现有关键词抽取方法的性能依然较差，距离实际应用还有很长一段路要走。

传统的关键词抽取方法通常只关注单词出现的频率，迭代计算过程给与高频词较高的权重。但在短文本中，一些关键词低频较低，这导致了关键词的丢失。本文所采用的数据集为大量论文摘要，论文摘要符合一般短文本特征，并且对应作者所标注的关键词基本都能准确表达摘要以及文章的核心含义，因此本文以大量论文摘要作为数据集进行关键词抽取技术的研究。本文基于知网提供义原树结合词林知识构建知识图谱，并将知识融入到 TextRank 方法所采用的词图模型中，使得构建的词图不仅包含词语之间的共现信息，还融入了语义知识。实验结果证明，本文所提出的方法相比传统 TextRank、TD-IDF、Word2Vec 方法有一定提升。

1 相关工作

1.1 关键词抽取研究

早期的关键词抽取研究主要是基于统计的方法，对候选词的一些特征进行统计，然后根据统计的结果对候选词进行排序。包括以 N-gram、TF-IDF 等指标来评价候选词在文档中的重要性。TextRank 是基于统计的图模型中最为典型的代表，首先通过词性标签筛选出文本中的候选词，其次为在同个窗口中出现的候选词之间建立边，最后赋予每个节点相同的初始值并运行 PageRank 算法直至收敛[2]。SGRank 使用单词的首次出现位置、词长等统计指标为

候选词的边赋值[3]。

随着关键词抽取技术在文本挖掘等领域的深入应用，越来越多的学者开始从事研究相关工作。Zhang 等[4]利用“全局上下文信息”，提出基于支持向量机的任务执行方法从文档中提取关键词。Beliga 等[5]提出了一种新的基于选择性的关键字提取方法，该方法从源文本中提取以网络表示的关键字。通过加权网络计算节点选择值，将其作为权重分布在单个节点链路上的平均值，用于关键词候选排序和提取过程。Biswas 等[6]基于图模型，提出一种无监督关键词抽取方法，该方法通过综合各种影响参数来确定关键字的重要性。闫强等[7]将词语的语义信息引入到 TextRank 算法中，改进了关键词抽取效果。还有学者对经典的 TF-IDF 加权公式进行改进，构建一个综合考虑多种影响因素的候选关键词评分加权公式；对 SharpICTCLAS 分词进行改进，增加位置标注；选择评分较高的词作为候选关键词，利用词的位置标注进行关键词抽取优化操作，将“切碎”的候选关键词进行组配，形成正式抽取的关键词[8]。

回顾最近的工作，Zhang 等[9]提出了一个关键词提取框架。该框架有 2 个模块，分别是对话上下文编码器和关键词标记器。对话上下文编码器从他们的对话上下文中捕获指示性表示并将该表示输入关键词标记器，关键词标记器从目标帖子中提取显着词。这两个模块经过联合训练，以优化对话上下文编码和关键词提取过程。胡少虎等[10]通过对关键词提取方法，尤其是关键词生成方法进行总结，阐明了关键词提取方法的研究重心从特征转向数据的趋势与原因，并指出现有关键词提取评价体系所存在的缺陷。

1.2 知识图谱嵌入

近年来，知识图谱（Knowledge Graph，KG）作为一种新的知识表示方法，在问题回答、信息检索以及自然语言处理等领域有着重要的应用。知识图谱是结构化的语义知识库，用于以符号形式描述物理世界中的概念及其相互关系；其基本组成单位是“实体—关系—实体”三元组，以及实体及其相关属性键值对，实体间通过关系相互联结，构成网状的知识结构[11]。知识图谱嵌入（Knowledge Graph Embedding，KGE）是一种新的研究方向，其基本思想是将包含实体和关系的 KG 组件嵌入到连续的向量空间中，从而在保持 KG 固有结构的同时简化操作[12]。

Guo 等[13]首先提出了一个将由实体和关系构成的知识图谱嵌入到低纬稠密向量空间中的方法。后来他们又提出了一种新的知识图谱分布式表示学习方法——规则引导嵌入（Rule-Guided Embedding，RUGE），其借助软规则的迭代引导完成知识图谱表示学习[14]。TransE[15]是一种经典的知识表示学习方法，可以对知识图谱进行补全。其通过对头实体、尾实体及对应关系进行建模，将实体和关系都表示为同一空间中的向量，能够通过训练得到不错的低维嵌入向量。但是，它在处理一对多、多对多等关系方面存在缺陷。TransH[16]是对 TransE 的改进，在一定程度上缓解了 TransE 不能很好地处理一对多、多对一等关系属性的问题，在预测精度方面有了显著的改进。知识图谱嵌入已经可以有效处理各种下游任务，例如链路预测（Link Prediction, LP）[17]、关系抽取[18]

doi:10.3772/j.issn.2095-915x.2022.03.001

和推荐系统[19]。

2 词汇知识库

2.1 义原简介

董振东花费数十年时间建立了一个汉语常识库——知网，一个可用于自然语言处理的知识系统，能解释词语概念和属性间关系的知识库。义原是知识库中不能再分割的最小的单位，在知网的知识库中每一个词语都可以使用若干个义原表示。

在知网中，并不是将每一个概念对应于一个树状概念层次体系中的一个结点，而是通过用一系列的义原，利用某种知识描述语言来描述一个概念。而这些义原通过上下位关系组织成一个树状义原层次体系。如图 1 所示，知网中词语“联想”有两种意思，第一个“联想”是由电脑（computer）、样式值（PatternValue）、能（able）、携带（bring）、特定牌子（SpeBrand）组成；第二个“联想”表示精神（Mental）。知网定义了约 2000 个义原，并且用这些义原表示了约 10 万个中文和英文单词。

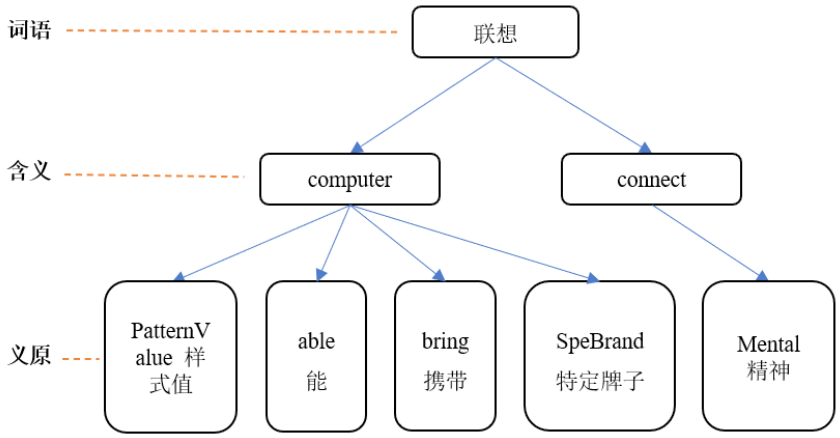

图 1 联想在知网中的描述

2.2 词林简介

同义词词林是由梅家驹等编撰的汉语词库，里面归类了汉语词语的同义词和同类词。同义词词林经过哈尔滨工业大学信息检索研究室的扩展后，共有 100093 个词语。如图 2 所示，同义词词林的扩展板具有五层树状结构，上面四层的节点代表抽象的类别，最底层的叶子节点是具体的词条，词条的编码不唯一，可能在不同的类别中同时存在。词林根据汉语的特点和使用方式，将词语分为十二个大类。其中第一类至第三大类大多数是名词，数词和量词在第四大类中，第五类一般是形容词，第六类至第十类一般是动词，第十类多数是虚词，第十二类是其他类别词语。大类和中类的排序遵守从具体概念到抽象概念的原则。

关于词条的编码如表 1 所示。前七位编码可以唯一确定一条编码，第八位编码只有三种情况：“=”代表同义；“#”代表同类；“@”代表独立，自我封闭，它在字典中既没有同义词，也没有同类词。本文所使用的词林为《哈工大信息检索研究同义词词林扩展版》1.0 版本。

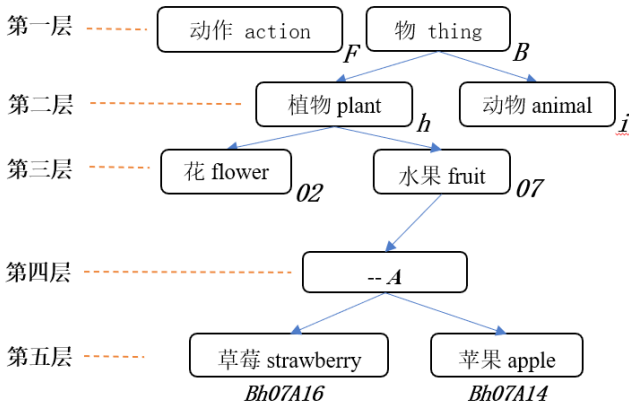

图 2 词林结构

表 1 词林中词语的编码结构

编码位	1	2	3	4	5	6	7	8
符号举例	A	a	0	1	B	0	2	=/#/@
性质	大类	中类	小类		词群	原子词群		
级别	第一层	第二层	第三层		第四层	第五层		

表 2 词林中词语的编码结构

编码位	符号举例	性质	级别
1	A	大类	第一层
2	a	中类	第二层
3	0	小类	第三层
4	1		
5	B	词群	第四层
6	0	原子词群	第五层
7	2		
8	=/#/@		

3 融合知识的关键词抽取方法

3.1 词汇图谱的构建

受知识表示学习方法的启发，本文基于知网提供的义原树和词林知识，利用 TransH 模型[20]训练并建立了具有三种关系的知识图谱(HowNet and CilinE Knowledge Graph，HCKG)。三种关系分别是“同义”“同类”“是义原”。前两种关系是基于词林的表示词语和词语之间的关系，第三种关系是基于知网的表示词语和义原之间的关系。如图 3 所示，“眼光(look)”和“目光(eye)”是同义关系，它们有着相同的义原。“急性 (acute)”和“慢性 (chronic)”是同类关系，它们有一部分相同的义原。

doi:10.3772/j.issn.2095-915x.2022.03.001

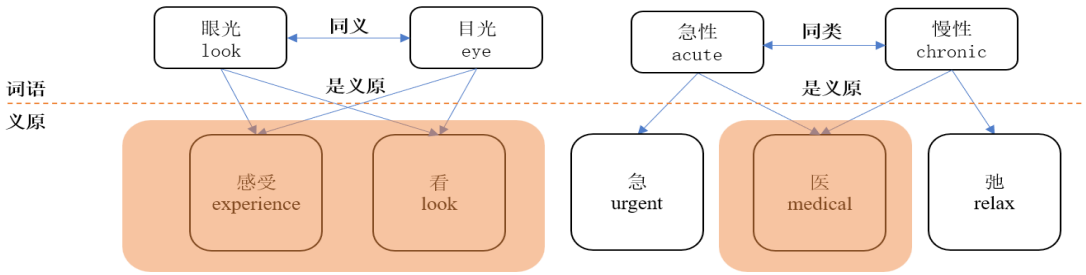

图 3 知识图谱内的关系示例

HCKG 里面的所有关系为一个集合 $R=\{R_{sym}, R_{sim}, R_{sem}\}$，$R_{sym}$ 和 R_{sim} 分别代表词语之间的“同义”和“同类”关系，R_{sem} 代表词语和义原之间的“是义原”关系，例如在图 3 中，“医（medical）”是词语“慢性（chronic）”的义原。图谱 $\{E_{word}, E_{sem}\}$ 里面的每个关系都由一个三元组（h, r, t）表示，h 是属于 E_{word} 的一个头结点，t 是属于 $E_{word} \cup E_{sem}$ 的一个尾结点，r 代表关系。损失函数定义如下：

$$L_1 = \sum_{(h,r,t)\in R}\left[\varepsilon + \left\|h_\perp + r - t_\perp\right\| - \left\|h_\perp + r - t_\perp\right\|^2\right]_+ \quad (1)$$

其中，t' 是属于 t 的一个负例，ε 取 4，$[x]_+$ 代表 $\max(0, x)h_\perp$ 和 $t_\perp$ 分别代表 h 和 t 在超平面上实体表示的投影向量，$h_\perp$ 和 $t_\perp$d 定义如下，其中 w 代表超平面的法向量：

$$h_\perp = h - w_r^T h w_r,\ t_\perp = t - w_r^T t w_r \quad (2)$$

图中的节点具有具体的含义，文本引入一个正则化的术语表示这些语义约束，任何一个词语的含义都可以由一组义原组成：

$$L_2 = \left\| w + r_e - \sum_{s\in S_w} s \right\| \quad (3)$$

其中 r_e 是一个向量，表示相等的关系，S_w 代表词语 w 的义原集合。最终损失函数定义如下：

$$L=\lambda_L L_1+(1-\lambda_L)L_2 \quad (4)$$

λ_L 是超参数，用来设置两个损失函数的权重。

3.2 融合知识的关键词抽取方法

本文基于 HCKG 提供的知识表示来计算词语之间的语义信息。通过词语之间的语义信息构建语义词图，生成语义矩阵 ω_{sim}，结合词语之间的共现信息生成的共现矩阵 ω_{co}，分别赋予两者相应权重，得到新的概率转移矩阵 P：

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix} \quad (5)$$

$$P_{ij} = \alpha \times \omega_{co}(i,j) + (1-\alpha) \times \omega_{sim}(i,j) \quad (6)$$

p_{ij} 代表节点 j 跳转至节点 i 的概率，α 是超参数，用来设置语义矩阵和共现矩阵的权重。再进行矩阵运算：

$$(W)^n = \left[P(W)^{n-1}, 1\right]\begin{bmatrix} d \\ 1-d \end{bmatrix} \quad (7)$$

其中 W 是候选词的权重矩阵，矩阵大小为 $n\times1$，初始 W 设置为 $\left[\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right]^T$。$d$ 为阻尼系数，类似机器学习中目标函数的正则项，阻尼系数的加入可以让整个计算更平滑。通过不断迭代计算，当结果收敛时，就可以得到各个候选词的权重，根据权重大小对候选词进行排序，最终排名靠前的候选词就是要抽取的关键词。

4 实验研究

4.1 实验环境和参数设置

4.1.1 实验环境

本次实验是在windows11操作系统上进行，处理器型号 AMD Ryzen 7 5800H @3.20 GHz，内存 16G，Python 版本 3.7.11。

4.1.2 参数设置

在 HCKG 中，词语、义原和关系的嵌入维度设置为 800，采用 Adam 作为优化器，学习率设置为 0.05。由于存在未登录词的现象，如果实验中存在未登录词，将它与其他词语之间的相似度设为 η。本次实验所取参数均在实验中取得了最好的效果。根据多次尝试得到的经验，公式（5）的参数 α 设置为 0.85，η=0.4。

4.2 基准方法和数据预处理

4.2.1 基准方法

实验所采用的基准方法是 TextRank，阻尼系数 d 取 0.85，当候选词权重迭代前后变化小于 0.0001，停止迭代。

$$S(V_i)=(1-d)+d\times\sum_{v_j\in In(v_i)}\frac{1}{|out(v_j)|}S(V_j) \quad (8)$$

在上式中，$In(Vi)$ 代表指向节点 i 的节点集合，而 $Out(Vj)$ 表示节点 j 指向的节点集合。d 为阻尼系数，阻尼系数使得每个节点都有跳转至随机顶点的概率，从而避免节点无法跳出的情况，取值范围 (0, 1)。

4.2.2 数据预处理

实验所使用的数据集为课题组搜集的33192篇期刊论文摘要，以及对应作者所标注的关键词，摘要一般在 120 字数左右，关键词个数一般在 4 个左右。相关论文所属领域主要为自然工程类，其中大部分属于工程科技类，少部分属于信息科技类和社会科学类。本文首先对期刊论文的标题和摘要进行拼接，组成一个新的短文本。拼接完成后，使用中文分词库 jieba 对摘要进行分词。分词完成后剔除不需要的词，剩下的词语就是候选词。在研究了大量人工标注的关键词词性后，本次实验留下来的候选词类型有名词、动词、名动词、人名、地名、机构团体和其他专名。

4.3 对比实验

Word2Vec 方法是李跃鹏等 [21] 提出的一种基于深度学习工具 Word2Vec 关键词提取算法，该算法首先使用 Word2Vec 模型将所有词语映射到一个更抽象的词向量空间中，其次基于词向量计算词语之间的相似度，最后通过词语聚类得到文章关键词。TD-IDF 算法 [22] 采用 TF 值和文本逆频率 IDF 进行加权，根据候选词权值大小提取关键词。

4.4 评价标准

本文使用准确率 P、召回率 R 和 F1 值作为指标来评价各种方法的效果。设 w_1 为自动抽取的关键词集合，w_2 为人工标注的关键词集合，则评价指标如公式（9）、公式（10）和公式（11）所示。其中 TP 代表正类被判定为正类，FP 代表负类被判定为正类，FN 代表正类被判定为负类，Num 代表个数。

$$P=\frac{TP}{TP+FP}=\frac{Num(w_1\cap w_2)}{Num(w_1\cap w_2)+Num(w_1-w_1\cap w_2)} \quad (9)$$

doi:10.3772/j.issn.2095-915x.2022.03.001

$$R=\frac{TP}{FP+FN}=\frac{Num(w_1+w_2)}{Num(w_1+w_2)+Num(w_2-w_1\cap w_2)} \quad (10)$$

$$F1=\frac{2\times P\times R}{P+R} \quad (11)$$

4.5 实验结果及分析

本节主要展示使用各种方法进行关键词抽取时得到的结果。如表 2 所示，在抽取 4 个关键词时，本文所提出的方法相比 TextRank 方法 F1 值提升 5.9%，比 TD-IDF 方法 F1 值提升 14.2%，比 Word2Vec 方法 F1 值提升 65.2%。

表 2 抽取 4 个关键词时实验结果

抽取方法	准确率P	召回率R	F1值
Word2Vec	0.111	0.113	0.112
TD-IDF	0.161	0.164	0.162
TextRank	0.172	0.177	0.175
TextRank+ HCKG	0.182	0.188	0.185

如表 3 所示，在抽取 5 个关键词时，本文所提出的方法相比 TextRank 方法 F1 值提升 2.9%，比 TF-IDF 方法 F1 值提升 11.4%，比 Word2Vec 方法 F1 值提升 52.9%。

表 3 抽取 5 个关键词时实验结果

抽取方法	准确率P	召回率R	F1值
Word2Vec	0.107	0.137	0.121
TF-IDF	0.149	0.187	0.166
TextRank	0.160	0.206	0.180
TextRank+ HCKG	0.164	0.212	0.185

在抽取 6 个关键词时，结果见表 4，本文所提出的方法相比 TextRank 方法 F1 值提升 2.2%，比 TF-IDF 方法 F1 值提升 11.0%，比 Word2Vec 方法提升 46.0%。

表 4 抽取 6 个关键词时实验结果

抽取方法	准确率P	召回率R	F1值
Word2Vec	0.103	0.158	0.124
TD-IDF	0.136	0.203	0.163
TextRank	0.147	0.227	0.178
TextRank+ HCKG	0.149	0.231	0.181

本文所提出的方法在抽取不同数量的关键词时效果均为最好。究其原因，论文摘要的关键词不一定为高频词汇，如果只考虑词语之间的共现信息，最终抽取结果会偏向高频词而忽略低频词。本文提出基于 TextRank 算法的改进方法，在概率转移矩阵里面融入词语语义知识，赋予某些出现频率较低，重要程度较高的词语更高权重，实验证明该方法抽取效果更佳。Word2Vec 方法基于维基百科中文语料使用 Word2Vec 模型训练生成词向量模型，再从中抽取候选关键词的词向量作为 K-means 聚类模型的输入，选择聚类中心作为关键词。但是维基百科中文语料可能未包含某些专业或者生僻的词语，加上选择聚类中心作为文本的关键词聚类方法时，选择聚类中心作为文本的关键词本身就是不太准确的，因此使用这种方法得到的效果不佳。TF-IDF 结构比较简单，不能有效反映词语的重要程度，对于专业性稍微强一点的文本例如论文摘要抽取效果较差，所以 TF-IDF 方法的效果也不是很理想。

4.6 相关实例

为了更详细地展现出本文所提出方法的关键词抽取效果，表 5 展示了一些示例在使用不

同方法时的抽取结果示例。在例一中，本文所提出的 TextRank+ HCKG 方法提取出了“试验台”和“喷油泵”关键词，其他方法只提取出了“试验台”关键词。例二同样只有 TextRank+ HCKG 方法提取出了“柴油机”和“增压”关键词，其他方法只提取出了“柴油机”关键词。这也证明本文所提出的方法相比其他方法，可以取得更好的论文摘要关键词抽取效果。

表 5 不同关键词抽取方法效果对比

摘要	抽取方法	关键词（人工标注）	抽取关键词
“柴油机喷油泵试验台数据采集处理系统的开发。介绍了柴油机喷油泵性能试验台数据采集处理系统的总体设计、基于IPC的硬件设计、面向对象软件设计、数据采集以及在LabVIEW环境下的实现。”	Word2Vec	柴油机、喷油泵、试验台、数据采集	软件设计、总体设计、面向对象、试验台
	TF-IDF		采集、试验台、处理、面向对象
	TextRank		采集、数据、试验台、处理
	TextRank+HCKG		采集、数据、试验台、喷油泵
“12V24OZJ型柴油机的研制。简要介绍了12V240ZJ型柴油机的研制过程，重点说明在开发设计中采用两种增压配套方案的对比，对该柴油机的性能试验结果进行了分析和总结，指出柴油机进一步提高功率和改善性能指标应做的工作。”	Word2Vec	柴油机、增压、性能、试验、	柴油机、研制、设计、重点
	TF-IDF		柴油机、研制、配套方案、性能指标
	TextRank		柴油机、研制、设计、重点
	TextRank +HCKG		柴油机、研制、增压、性能指标

5 总结

本文基于义原树和词林知识，利用 TransH 模型训练并构造了知识图谱 HCKG。利用知识图谱提供的知识表示计算词语的语义信息，构建语义词图，结合词语之间共现词图，生成新的概率转移矩阵进行迭代计算得到各候选词权重，最终根据权重大小提取关键词。实验证明，本文所提出的方法比传统 TextRank、TD-IDF 和 Word2Vec 方法有一定提升。

观察实验结果，有些常用词也出现在了其中。例如基于摘要“柴油机喷油泵试验台…在 LabVIEW 环境下的实现”抽取出的关键词中，词语“采集”和“数据”被当成关键词抽取了出来。这也是课题组下一步要做的工作，根据研究目标特点建立特定词库，提高分词效率和关键词抽取效果。

参 考 文 献

[1] Turney P D. Learning Algorithms for Keyphrase Extraction[J]. Information Retrieval, 2000, 2(4):303-336.

[2] Mihalcea R, Tarau P. TextRank:Bringing Order into Text[C]. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain. Association for Computational Linguistics. 2004:404-411.

[3] Danesh S, Sumner T, Martin J H. SGRank:Combining Statistical and Graphical Methods to Improve te State of the Art in Unsupervised Keyphrase Extraction[C]. Proceedings of the 4th Joint Conference on Lexical and Computational Semantics, Colorado, USA. 2015:117-126.

[4] Zhang K, Xu H, Tang J, et al. Keyword extraction using support vector machine[C]. International conference on web-age information management.

doi:10.3772/j.issn.2095-915x.2022.03.001

Springer, Berlin, Heidelberg. 2006:85-96.

[5] Beliga S, Meštrović A, Martinčić-Ipšić S. Selectivity-based keyword extraction method[J]. International Journal on Semantic Web and Information Systems (IJSWIS), 2016, 12(3):1-26.

[6] Biswas S K, Bordoloi M, Shreya J. A graph based keyword extraction model using collective node weight[J]. Expert Systems with Applications, 2018(97):51-59.

[7] 闫强 , 张笑妍 , 周思敏 . 基于义原相似度的关键词抽取方法 [J]. 数据分析与知识发现 , 2020, 5(4):80-89.

[8] 钱爱兵 , 江岚 . 基于改进 TF-IDF 的中文网页关键词抽取——以新闻网页为例 [J]. 情报理论与实践 , 2008(6):945-950.

[9] Zhang Y, Zhang C, Li J. Joint modeling of characters, words, and conversation contexts for microblog keyphrase extraction[J]. Journal of the Association for Information Science and Technology, 2020, 71(5):553-567.

[10] 胡少虎 , 张颖怡 , 章成志 . 关键词提取研究综述 [J]. 数据分析与知识发现 , 2020, 5(3):45-59.

[11] 刘峤 , 李杨 , 段宏 , 等 . 知识图谱构建技术综述 [J]. 计算机研究与发展 , 2016, 53(3):582.

[12] Wang Q, Mao Z, Wang B, et al. Knowledge graph embedding:A survey of approaches and applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(12): 2724-2743.

[13] Guo S, Wang Q, Wang B, et al. Semantically smooth knowledge graph embedding[C]. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1:Long Papers). 2015:84-94.

[14] Guo S, Wang Q, Wang L, et al. Knowledge graph embedding with iterative guidance from soft rules[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[15] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[J]. Advances in Neural Information Processing Systems, 2013, 26.

[16] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes[C]. Twenty-Eighth AAAI conference on artificial intelligence. 2014

[17] Zhang Q, Sun Z, Hu W, et al. Multi-view knowledge graph embedding for entity alignment[J]. arXiv preprint arXiv:1906. 02390, 2019.

[18] Weston J, Bordes A, Yakhnenko O, et al. Connecting language and knowledge bases with embedding models for relation extraction[J]. arXiv preprint arXiv: 1307. 7973, 2013.

[19] Sun Z, Yang J, Zhang J, et al. Recurrent knowledge graph embedding for effective recommendation[C]. Proceedings of the 12th ACM Conference on Recommender Systems. 2018: 297-305.

[20] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2014.

[21] 李跃鹏 , 金翠 , 及俊川 . 基于 word2vec 的关键词提取算法 [J]. 科研信息化技术与应用 , 2015(4):54-59.

[22] Witten I H, Paynter G W, Frank E, et al. Kea:Practical automated keyphrase extraction[M]. Design and Usability of Digital Libraries: Case Studies in the Asia Pacific. IGI global. 2005:129-152.