情 报 学 报 ISSN 1000 - 0135
第35卷第12期1305 - 1312,2016年12月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000 - 0135
Vol. 35 No. 12, 1305 - 1312 December 2016

doi:10.3772/j.issn.1000-0135.2016.012.008

基于二分网络分析方法的学术文献关键词自动抽取方法研究[1)]

李树青 曹 杰 庄光光 陈俊鹏

(南京财经大学信息工程学院,南京 210046)

摘要 学术文献关键词的自动抽取在各种数字图书馆工作中有着广泛的应用,可以极大减少人工标注的成本和改善文献标注的准确度。本文在候选关键词的选择基础上,提出了一种利用文献结构中标题关键词和摘要关键词共现特征构造二分网络的方法,并详细说明了网络节点和边的权值设定依据和方法,据此实现了基于二分网络权值扩散方法的关键词自动抽取方法。最后,文章利用图情类文献数据集合的实验分析,对相关方法和评价结果都做了详细说明。

关键词 关键词抽取 二分网络 迭代算法 学术研究

The Automatic Extraction of Keywords from Academic Literature based on Bipartite Network Analysis Method

Li Shuqing, Cao Jie, Zhuang Guangguang and Chen Junpeng

(*College of Information Engineering, Nanjing University of Finance & Economics, Nanjing* 210046)

Abstract The automatic extraction of keywords from academic literature has been wildly used in all kinds of applications in digital libraries, which can distinctly reduce the cost of manual indexing and improve the accuracy of indexing. Based on the choosing of candidate keywords, this paper introduces how to compose the bipartite network with the co-occurrence of keywords in titles and abstracts of academic literature. The gauge and methods of weight assignment for nodes and edges are also proposed. Then an automatic extraction method of keywords based on the iteration algorithm of bipartite network is discussed in detail. Finally, we report the experiments results of algorithm and evaluation with the analysis in the collection of academic literature of Library and Information Science.

Keywords extraction of keyword, bipartite network, iteration algorithm, academic research

1 引 言

关键词是可以表征全文主题内容信息的词语,一般通过人工标注或计算机自动抽取方法来获取。准确有效的发现自然语言语句中的关键词语,进行关键词抽取,也一直是知识挖掘和文本分析的重要研究内容。该项技术在现代数字图书馆的各种应用中有着广泛的应用领域。

不同领域的自然语言语句往往各有特点,从而

收稿日期:2016 年 7 月 12 日

作者简介:**李树青**,男,1976 年生,教授,博士,硕士生导师,主要研究方向:个性化服务、信息检索和 Web 挖掘,E-mail:leeshuqing@163.com。**曹杰**,男,1969 年生,教授,博士,硕士生导师,主要研究方向:Web 挖掘和大数据处理。**庄光光**,男,1991 年生,硕士研究生在读,主要研究方向:个性化服务和信息检索。**陈俊鹏**,男,1979 年生,讲师,博士,主要研究方向:文本分析和数据挖掘。

1) 本文系江苏省高校哲学社会科学基金项目"个性化信息服务中用户兴趣演变机理研究"(编号:2014SJB144)、科技部科技支撑项目"外贸行业电子商务服务技术研究与应用"(项目编号:BAH29F01)。

决定了相应的分析方法也各有不同。聚焦到学术文献分析领域,存在于文献标题和摘要甚至全文中的语句都是蕴藏大量原始知识内容的集合体。单纯依赖于文献关键词存在着明显的问题,如以典型的学术文献为例,按照我国科学论文编写格式标准GB7713-87规定,每篇论文应选取3~8个词作为关键词,但通常一篇文献只具有4个左右的关键词,这些有限数量的词语很难全面准确的表达现有文献的研究主题范围,即使考虑到人工标注主题词方法,也不可避免受到人工主观判断的影响,甚至会产生文献原始关键词和后续人工标注关键词具有明显的差异和区别。需要说明的是,自动分词也是一种较为常见的关键词抽取方法,虽然处理速度和分词效果现在已经取得突破性的进展,然而对于专业领域的关键词,尤其一些较长的关键词术语,关键词的多词汇组合特征使其不适合采用基于分词方法的传统词频统计方法进行抽取,自动分词往往识别效果并不理想,常常会产生在不同的文献中对相同词语序列产生不同的分词结果,一致性难以保证[1]。

2 文献回顾

关键词抽取在国内外的研究都比较早,早期的方法多采用一些经典文本挖掘算法来实现,如遗传算法、决策树机器学习[2]和朴素贝叶斯技术[3]等,这些方法都利用大量的样本训练以获取模型的权值。同时由于中文分词的特殊问题,国内学者对此的研究更有针对性,主要分为利用基于分词的传统统计方法和基于词序列的语言学方法等。如早期的利用互信息[4]和最大熵模型[5]的方法,由于特征的选择以及估计特征参数时不够准确,会极大的限制这些方法进一步的应用。有学者通过识别由一系列相关词语组成词汇链,并将一篇文本中的词汇按照它们的词义相似度构建多个词汇链,进而从中挑选出能够代表文本主题的关键词,该方法受到收录词汇种类的影响、文章题目本身的内容和分词算法的影响都比较大[6]。还有学者提出用于自动标引的主题关键词抽取方法,但仅限于从已标引的结构化语料库的元数据中抽取关键词[7]。还有学者提出利用经典的TF-IDF方法来构建每篇文献的词语向量,并根据向量单元项值的权值大小来自动抽取最为相关的关键词,使用的分词方法极大的限制了获取关键词的稳定性[8]。

在较新的研究中,越来越多的学者将各种有效的机器学习方法和文本挖掘方法应用于关键词抽取问题的研究。相对于传统的TF-IDF方法而言,该方法考虑了重要的低频词语和文档内部的主题分布语义特征,因此近年来逐渐得到了学者的重视[9]。如有学者提出了将条件随机场序列标注机器学习算法引入到关键词抽取中,建立基于字角色标注的中文关键词标引模型,提出了关键词角色空间模型和综合利用字序列上下文特征的设计思路,但该方法需要人工进行合理的角色设定和必要的机器学习训练[1]。还有学者提出一种反映词主题代表能力的特征计算方法,利用LDA主题模型中各个主题下词的分布情况来计算词的主题特征并据此实现关键词抽取,实验表明该方法可以有效提高系统的性能,但性能稳定性还有很大的改进空间[10]。

在这些研究中,利用基于词图模型的复杂网络分析方法来进行关键词抽取的研究不断得到更多学者的关注,此类方法非常适用于非规范文本信息的抽取,如网络信息中标签内容的抽取等[11]。相对于上述方法而言,这种方法是一种无需指导学习过程的方法,人工预处理的代价很少,不依赖于训练数据样本的准确性,优势比较明显。这些方法的基础思想都建立词语共现关系这一基本假设之上,也就是说,词语之间的共现表达了一种相互推荐关系,与重要词语共现的其他词语也应该较为重要[12]。实验结果也表明,由于词项排序结果是基于稳定的收敛数值,因此算法效果和结果的稳定性也优于以词频为基础的传统方法[13]。比较著名的TextRank是较早出现的方法,但是该方法在设计词语链接关系时没有很好的考虑到词语之间边权值的设置和权值扩散的不同系数,因此关键词权值计算结果倾向于文档频率较高的词项,而文档频率较低的有效专业词汇却往往难以获得较高权值[14]。后续的研究还证明进一步引入用户信息或者领域知识还可以进一步提高关键词抽取效果[15]。在较新的研究中,有学者提出从词语的覆盖影响力、位置影响力和频度影响力三个方面加权计算邻接词语所传递的影响力[16]。还有学者通过限定关联跨度长度形成语言词汇网络,并通过基于复杂网络特征的中文文档关键词抽取算法实现了关键词抽取,平均准确率高于经典TF-IDF方法所获得的平均准确率,但当抽取的关键词数目增多时,那些词频相对较低但却在文章中起重要作用的词也开始更多的被TF-IDF关键词抽取方法所抽取,因而所提出的关键词抽取方法效果逐渐变低[17]。

作为复杂网络方法的一种，二分网络（Bipartite Network）方法作为一种有效的网络分析算法近年来也逐渐受到越来越多学者的关注[18]。该方法需要定义一个特殊的二分网络结构，即所有的节点都被分为两个不同的层次，链接关系只发生在不同层次间的节点之间[19]。最为广泛的应用是协同推荐研究和意见网络研究[20]，近年来的研究范围也逐渐扩大，如在算法上解决无向图结构中向量的最大优先度匹配（Maximum Priority Matching）问题[21]，在应用上利用读者和图书借阅关系形成的二分网络特征来改进现有个性化文献服务[22]等。值得注意的是，已有很多研究开始进一步关注在二分网络分析基础之上的三分网络[23]或者四分网络[24]结构，从而结合更多可以获取的语义信息来改进单纯利用网络结构信息的传统方法。这些研究都广泛证实了该方法在利用网络节点权值计算方面的有效性，也构成了本文研究的主要内容。本文在二分网络分析方法的基础上，主要研究如何利用文献自身的结构特点来构造反映词语共现关系的二分网络结构，并据此实现对文献关键词的自动识别，研究重点在于如何有效的获取最能反映文献主题的重要关键词，从中我们可以看出二分网络分析方法在文献关键词识别工作的有效作用。

3 候选关键词的选择

3.1 候选关键词集合的确定

利用文献已有的关键词，可以抽取出来构成完整的关键词集合。考虑到实际数据样本中可能产生的数据错误，需要首先对这些关键词进行必要的检验处理。一般而言，词频过滤是一种较为简单有效的方法，然而在电子文献集合中，典型的关键词错误往往是没有正确进行分词，从而形成一些多个关键词的组合词语，真正意义上由于错字产生的错误关键词数量极为有限，同时矛盾的地方也在于由于专业关键词自身的特点，一般较长的常见关键词更能准确反映研究主题。

本文采取了一种多指标集成的迭代过滤方式。定义的两个指标分别是词语长度和词频，这两个指标需要同时综合考虑。一般大于 1 次即可绝大多数的过滤掉由于错字产生的错误关键词和一些没有正确分词形成的较长关键词序列。由于这个长度阈值的确定主观性较强，因此我们设计了一种迭代判定方法，具体方法的伪代码如下所述：

```
int N = 3; //词语长度
int cur_count = 0, pre_count = 0; //当前过滤后得到的关键词数量，上次过滤后得到的关键词数量
while ((cur_count-pre_count)/ cur_count < THRESHOLD)
{
  pre_count = cur_count;
  cur_count = filter(df > 1 and len(cikeywords) >= N) OR len(cikeywords) < N and len(cikeywords) > 1); //filter 函数返回指定参数条件下过滤得到的关键词数量
  N++;
}
```

此时随着词语长度 N 的不断变大，过滤得到的关键词数量不断增多，但增幅不断递减，因此利用合适的阈值参数 THRESHOLD 设定，即可停止迭代过滤判断过程。

3.2 标题和摘要中抽取候选关键词

如果已有文献的标题和摘要中含有现有关键词集合中的关键词，即可从中抽取所含的关键词。但是由于关键词之间可能存在相互包含（如“个性化推荐”和“个性化”）和相互衔接（如“个性化推荐”和“推荐技术”）等问题，直接抽取会抽取出大量无用的子关键词和对较长关键词产生不一致的切分。如对于相互包含问题，“个性化”就是“个性化推荐”关键词的子关键词，显然较长的后者更能准确反映当前研究主题；再如相互衔接问题，对于“个性化推荐技术”词语序列而言，就存在着“个性化推荐——技术”和“个性化——推荐技术”等多种切分方法。

本文采用基于从后往前最长字符串匹配的切分算法和基于从前往后最长字符串匹配的切分算法相结合的方法，以文献标题和文摘中的每一个句子为基本单位，切分出其中所含的有效关键词，并对每个基本单位，汇总归类相同的关键词。

3.3 基于二分网络权值扩散方法的关键词抽取方法

二分网络构造的关键在于网络节点和边的选择。对于文献关键词集合而言，可以采用多种方法来设计。由于文献标题和摘要中都存在除了已有给定的关键词外其他重要关键词，所以我们采取了如

下设计方案：

设二分网络表示形式为：G =(V(V1,V2);E)，对于每一篇文献而言，以文献标题抽取的关键词构成二分网络的第一层(V1)，以文献摘要抽取的关键词构成二分网络的第二层(V2)。通过权值迭代计算，我们可以获取到每一层中关键词的权值。关于两层之间的边选择(E)，我们提出了基于共现原理的假设1：

假设1：如果摘要中的出现了标题中的关键词，则同一句中的其他关键词应该和当前标题关键词具有一定的语义联系

基于该假设，我们可以在两层节点间定义边的连接。如文献标题有三个关键词：*K*1，*K*2 和 *K*3，摘要有三句分别为：

句1：*k*1，*k*3，*k*4

句2：*k*5，*k*6，*k*7

句3：*k*2，*k*4，*k*6，*k*7

句4：*k*3，*k*7

则最终构建的二分网络如图1所示：

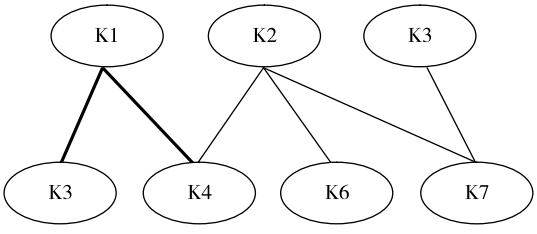

图1 利用文献标题关键词在摘要句子中与其他关键词的共现关系构造的二分网络结构

（去除了没有边连接的孤立节点，边的粗细反映了关键词之间的联系）

对于该二分网络结构，进一步计算节点权值就需要对权值扩散方法进行定义。不同于一般的节点信息，文献中的关键词分布遵循着明显的齐普夫法则，因此很多高频出现的关键词往往都是语义价值较低、区分度很差的一般词语，如“方法”、“原理”等，相反真正反映文摘主题的关键词往往由于研究面较窄而具有较低的词频。词频的高低直接决定了与其他词语共现的概率，所以只按照节点出度来计算的传统二分网络权值迭代算法将不可避免的有利于高频一般词语。

所以，我们提出了基于倒文档频率的假设2：

假设2：文献关键词二分网络中节点的重要性与倒文档频率成正比，边的重要性与相连接节点的重要性成正比

显然，连接着标题重要关键词和摘要句子重要关键词的边反映了两个高价值关键词之间存在的语义联系。据此，可以实现对迭代计算中节点权值的计算方法，如式(1)所示

$$R^i(Node_k) = \alpha + (1-\alpha) * \sum_{v \in V_{Nodei}} R^{i-1} \| Node_v) * E_{k,v} \quad (1)$$

R^i 表示第 i 次迭代计算中节点权值，$E_{i,j}$ 表示节点 $Node_i$ 和 $Node_j$ 之间边的权值，V 代表节点相连的另一层其他节点集合，α 为实现收敛的衰减系数。

其中，E 权值的计算方法如式(2)所示：

$$E_{i,j} = Normalize\left(Log\left(\frac{N}{DF_{Nodei}}\right)\right) * Normalize\left(Log\left(\frac{N}{DF_{Nodej}}\right)\right) \quad (2)$$

DF 为节点 i 的文档频率，$Normalize$ 函数为数值规范化函数，本文采用了与最大值相除的计算方法。

以图1为例，考虑标题中三个关键词的权值计算结果。初始节点权值设为1/3，并假设 *k*1 关键词与 *k*3、*k*4 关键词边权值为其他边权值的3倍，表1展示了迭代10次的计算结果：

表1 图1所示例子的10次迭代计算结果

迭代次数	关键词	权值	迭代次数	关键词	权值
第1次	*k*1	0.6558139	第6次	*k*1	0.7733470
	*k*2	0.2527132		*k*2	0.1892030
	*k*3	0.0914729		*k*3	0.0374499
第2次	*k*1	0.7502162	第7次	*k*1	0.7733677
	*k*2	0.2024658		*k*2	0.1891910
	*k*3	0.0473180		*k*3	0.0374412
第3次	*k*1	0.7691189	第8次	*k*1	0.7733715
	*k*2	0.1916478		*k*2	0.1891889
	*k*3	0.0392333		*k*3	0.0374397
第4次	*k*1	0.7726015	第9次	*k*1	0.7733721
	*k*2	0.1896346		*k*2	0.1891885
	*k*3	0.0377639		*k*3	0.0374394
第5次	*k*1	0.7732330	第10次	*k*1	0.7733723
	*k*2	0.1892691		*k*2	0.1891884
	*k*3	0.0374980		*k*3	0.0374393

从中可以看出关键词权值计算已经取得了较好的收敛效果。

4 实 验

我们对万方和CSSCI两大中文期刊数据库进行了文献数据获取,抽取了图书情报方向核心期刊共52种,时间跨度为1998年1月到2015年6月,总共获得179 438篇有效文献。其中,文献原始关键词总数为108 962,平均关键词个数为3.67。

4.1 关键词抽取实验

按照本文3.1节所述方法,阈值THRESHOLD设定为0.02,迭代计算确定的关键词长度N为10,最终获取了106 237个有效候选关键词,其中最大长度22。进一步按照3.2节所述方法,具体的计算结果如表2所示:

表2 文献标题和摘要的关键词识别数量结果

计算对象	算法	识别关键词数量	汇总后关键词数量
标题	从后往前最长匹配算法	718 609	840 880
	从前往后最长匹配算法	716 877	
摘要	从后往前最长匹配算法	4 317 843	4 598 139
	从前往后最长匹配算法	4 312 005	

4.2 重要关键词识别实验

以下面一篇文献为例,标题为《基于加权关键词共现时间元的个性化学术研究时序路径发现及其可视化呈现方法》,摘要内容为"个性化学术研究时序路径发现方法和可视化呈现技术/可以帮助用户了解相关学术研究点的演化规律和掌握学术研究的发展趋势/本文首先提出了加权关键词共现时间元的基本方法/并据此介绍了关键词时序路径和关键词时序网络结构的表达方法/其次/文章在说明扩展关键词数据获取方法关键词权值和关键词共现权值设定方法的基础上/重点介绍了个性化学术研究时序路径的发现方法/最后/文章对以学者发文为数据集合的个性化学术研究时序路径实验/可视化界面设计和实验结果/以及用户满意度评价实验都做了必要的说明",原始关键词为"时序路径、可视化、学术研究、词语共现"。

本文方法总共识别了9个标题关键词和31个摘要关键词,表3展示了部分权值最高的关键词权值计算结果:

表3 权值最高的前5个文献标题关键词

关键词	权值
时序路径	0.1875
学术研究	0.1212
关键词共现	0.1183
可视化呈现	0.0957
方法	0.0952

从中可以看出,大部分已有原始关键词都被识别为最高权值关键词。

我们对所有文献进行了数据处理,由于数据量很大,我们采用了Hadoop为基础的分布式处理框架,完成了全部文献的关键词识别工作,从标题中总共抽取了468 582个关键词,从摘要中总共抽取了2 120 034个关键词。

4.3 方法评价

由于方法所用的关键词来源于文献原始关键词,所以在最终计算结果中,只要文献标题和摘要出现这些原始关键词,这些词语就会参与迭代计算。鉴于人工评价方法存在主观性强的不利特点,为此我们利用文献关键词原始数据为基准,设计了如下的评价方法。

因为长度原因,每篇文献的标题关键词总数并不是很高,因此我们只考虑权值最高的2位关键词。文献原始关键词能够反映文献主要主题的关键词,本文方法可以从标题中识别出更多主要关键词,因此好的方法应该至少将已有的原始文献关键词设定为较高的权值。为此,对于每篇文献,评价方法有效性的指标为计算在已有的权值结果中,排名前N位关键词中出现多少个原始文献关键词,N为当前文献的原始关键词数量,如式3所示:

$$\text{validation1}_i = \frac{\text{CountofKeywordsinResulsandArticlei}}{\text{CountofKeywordsinArticlei}} \tag{3}$$

最终汇总每篇文献的指标平均值即可得到评价指标validation1。在实验数据中,并非所有文献都有原始关键词,具有原始关键词的文献数量为124 004篇,同时也并非所有文献标题和摘要都有原始关键词,其中标题已经含有原始关键词的文献数

量为94 931篇,摘要含有原始关键词的文献数量为89 374篇。因此我们分别统计了不同文献范围内的具体指标,并做了对比实验。实验中每个抽取标题的关键词按照标准TF-IDF权值倒序排列,由于标题信息较短,在计算中我们假设TF都为1,因此我们主要利用关键词的IDF权值。最终统计validation1指标值,对比结果如表4所示。

表4 不同文献范围内的对比评价结果

文献类别	数量	平均标题关键词个数	标题validation1	对比validation1	提高率
所有文献	179 438	3.05	0.3299	0.3251	1.48%
所有具有原始关键词的文献	124 004	3.21	0.4774	0.4704	1.49%
标题含有至少1个原始关键词的文献	97 283	3.37	0.6085	0.5655	7.6%
标题含有1个原始关键词的文献	48 636	2.80	0.4340	0.4332	0.18%
摘要含有至少1个原始关键词的文献	95 847	3.50	0.5479	0.5054	8.41%
摘要含有1个原始关键词的文献	32 155	2.94	0.3553	0.3667	-3.11%
标题和摘要都含有至少1个原始关键词的文献	85 069	3.62	0.6173	0.5608	10.07%
标题和摘要都含有1个原始关键词的文献	22 250	3.01	0.4253	0.4176	1.84%
标题或者摘要都含有至少1个原始关键词的文献	108 061	3.29	0.5478	0.5159	6.18%
标题或者摘要都含有1个原始关键词的文献	58 541	2.79	0.3940	0.4026	-2.14%
标题含有至少2个原始关键词的文献	48 647	3.95	0.7830	0.6979	12.19%
标题含有2个原始关键词的文献	34 896	3.62	0.7594	0.6643	14.32%
摘要含有至少2个原始关键词的文献	63 692	3.78	0.6451	0.5754	12.11%
摘要含有2个原始关键词的文献	35 484	3.48	0.5985	0.5364	11.58%
标题和摘要都含有至少2个原始关键词的文献	43 690	4.06	0.7832	0.6946	12.76%
标题和摘要都含有2个原始关键词的文献	19 875	3.67	0.7557	0.6671	13.28%
标题或者摘要都含有至少2个原始关键词的文献	68 649	3.72	0.6549	0.5863	11.7%
标题或者摘要都含有2个原始关键词的文献	50 505	3.50	0.6478	0.5733	12.99%
标题含有至少3个原始关键词的文献	13 751	4.77	0.8430	0.7829	7.68%
标题含有3个原始关键词的文献	11 508	4.58	0.8376	0.7728	8.39%
摘要含有至少3个原始关键词的文献	28 208	4.16	0.7038	0.6245	12.7%
摘要含有3个原始关键词的文献	20 059	4.01	0.6969	0.6137	13.56%
标题和摘要都含有至少3个原始关键词的文献	11 864	4.91	0.8399	0.7780	7.96%
标题和摘要都含有3个原始关键词的文献	7 519	4.65	0.8358	0.7722	8.24%
标题或者摘要都含有至少3个原始关键词的文献	30 095	4.15	0.7137	0.6364	12.15%
标题或者摘要都含有3个原始关键词的文献	24 048	4.08	0.7208	0.6403	12.57%

从对比实验可以明显看出,本文方法比传统的基于 IDF 关键词权值方法具有较高的有效性,尤其对于标题和摘要至少含有 1 个原始关键词的文献而言,总体有效性大都在 10% 以上。值得说明的是,在“所有文献”和“所有具有原始关键词的文献”中指标值较低,主要原因在于并非全部文献的原始关键词都存在于对应文献的标题中。同时,在文献标题和摘要分别含有 1 个原始关键词时的各种组合可能中,指标低也都非常低,究其原因主要在于我们最少选择 2 个最高权值的标题关键词,所以此时最好的匹配情况也只有 50%。

5 总结与展望

本文利用文献自身的结构特点构造出反映词语共现关系的二分网络结构,并提出了一种基于二分网络分析方法的学术文献关键词自动抽取方法,初步实现了预期的设计目标,不过该方法存在很多需要进一步研究和改进的地方,主要有以下几点:①目前的研究主要侧重于对于文献标题关键词的自动抽取研究,文献摘要和全文中存在的更多关键词也可以成为进一步获取关键词的基础,这方面的相关研究正在展开;②现有的二分网络结构设计方法是在多次实验选择的基础上,结合文献结构特征而提出,如何能根据不同的文档形式提出更为通用的二分网络结构设计方案,对于进一步推广该方法在关键词自动抽取方面的应用更为重要。这些都构成了我们下一步的研究目标。

参 考 文 献

[1] 邓三鸿,王昊,秦嘉杭,等. 基于字角色标注的中文书目关键词标引研究[J]. 中国图书馆学报,2012,38(2):38-49.

[2] Turney. Learning to extract keyphrases from text[R]. National Research Council, Institute for Information Technology, Canada, ERB21057, 1999.

[3] Witten I, Paynter G, Frank E, et al. KEA: Practical automatic keyphrase extraction[C]//Proceedings of the fourth ACM conference on Digital libraries, ACM, 1999 (8): 254-255.

[4] Yang W. Chinese keyword extraction based on max-duplicated strings of the documents[C]//Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, 2002: 439-440.

[5] 李素建,王厚峰,俞士汶,等. 关键词自动标引的最大熵模型应用研究[J]. 计算机学报,2004,27(9):1192-1197.

[6] 索红光,刘玉树,曹淑英. 一种基于词汇链的关键词抽取方法[J]. 中文信息学报,2006,20(6):27-32.

[7] 王军. 词表的自动丰富——从元数据中提取关键词及其定位[J]. 中文信息学报,2005,19(6):38-45.

[8] 徐文海,温有奎. 一种基于 TFIDF 方法的中文关键词抽取算法[J]. 情报理论与实践,2008,31(2):298-302.

[9] Pasquier C. Task 5: Single document key-phrase extraction using sentence clustering and Latent Dirichlet Allocation [C]//Proceedings of the 5th international workshop on semantic evaluation, Association for Computational Linguistics, 2010(7): 154-157.

[10] 刘俊,邹东升,邢欣来,等. 基于主题特征的关键词抽取[J]. 计算机应用研究,2012,29(11):4224-4227.

[11] Liu Z, Chen X, Sun M. Mining the interests of Chinese microbloggers via keyword extraction[J]. Frontiers of Computer Science, 2012, 6(1): 76-87.

[12] 李鹏,王斌,石志伟,等. Tag-TextRank:一种基于 Tag 的网页关键词抽取方法[J]. 计算机研究与发展,2012,49(11):2344-2351.

[13] Rahman M M, Roy C K. TextRank based search term identification for software change tasks[C]//Software Analysis, Evolution and Reengineering (SANER), 2015 IEEE 22nd International Conference on IEEE, 540-544.

[14] Mihalcea R, Tarau P. Textrank: Bringing order into texts [C]//Proceedings of EMNLP, Barcelona, Spain, Association for Computational Linguistics, 2004: 404-411.

[15] 姜霖,王东波. 采用连续词袋模型(CBOW)的领域术语自动抽取研究[J]. 现代图书情报技术,2016,32(2):9-14.

[16] 夏天. 词语位置加权 TextRank 的关键词抽取研究[J]. 现代图书情报技术,2013,29(9):30-34.

[17] 赵鹏,蔡庆生,王清毅,等. 一种基于复杂网络特征的中文文档关键词抽取算法[J]. 模式识别与人工智能,2008,20(6):827-831.

[18] Serratosa F. Speeding up fast bipartite graph matching through a new cost matrix[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2015, 29(02): 1550010.

[19] Wang P, Pattison P, Robins G. Exponential random graph model specifications for bipartite networks—A dependence hierarchy[J]. Social networks, 2013, 35(2): 211-222.

[20] Zhou T, Ren J, Medo M, et al. Bipartite network projection and personal recommendation[J]. Physical Review E,

2007,76(4):046115.

[21] Turner J. Faster Maximium Priority Matchings in Bipartite Graphs[EB/OL]. arXiv preprint arXiv:1512.09349, 2015.

[22] 李树青,徐侠,许敏佳. 基于读者借阅二分网络的图书可推荐质量测度方法及个性化图书推荐服务[J]. 中国图书馆学报,2013,39(3):83-95.

[23] Zhang Z K,Zhou T,Zhang Y C. Personalized recommendation via integrated diffusion on user-item-tag tripartite graphs [J]. Physica A: Statistical Mechanics and its Applications,2010,389(1):179-186.

[24] 刘彤,倪维健,柳梅. 面向搜索引擎查询日志的领域术语自动识别方法[J]. 现代图书情报技术,2016,32(2):25-33.

(责任编辑 马 兰)