第 46 卷 第 7 期 计 算 机 工 程 2020 年 7 月
Vol. 46 No. 7 Computer Engineering July 2020

·人工智能与模式识别· 文章编号：1000-3428(2020)07-0078-06 文献标志码：A 中图分类号：TP18

基于通配符模式与随机游走的关键词提取方法

马慧芳[1,2]，李 苗[1]，童海斌[1]，詹子俊[1]

（1. 西北师范大学 计算机科学与工程学院，兰州 730070；2. 桂林电子科技大学 广西可信软件重点实验室，广西 桂林 541004）

摘 要： 结合通配符模式与引入先验信息的随机游走算法，提出一种改进的关键词提取方法。使用通配符约束捕获词语之间的语义关系，提取满足间隙约束和一次性条件的顺序模式以计算模式支持度，并在模式支持度大于等于最小支持度阈值时建立节点关联图。将维基百科知识库中词语间的相似度作为先验信息，利用基于先验信息的 PageRank 算法在关联图上进行随机游走直至其排名分数趋于稳定，选取排名前 Top K 个词语作为关键词。实验结果表明，与 TextRank、GraphSum 算法相比，该方法具有更高的提取准确率及稳定性。

关键词： 关键词提取；通配符模式；随机游走；间隙约束；PageRank 算法

开放科学（资源服务）标志码（OSID）：

中文引用格式： 马慧芳，李苗，童海斌，等. 基于通配符模式与随机游走的关键词提取方法[J]. 计算机工程，2020，46(7)：78-83.

英文引用格式： MA Huifang，LI Miao，TONG Haibin，et al. Keyword extraction method based on wildcard pattern and random walk[J]. Computer Engineering，2020，46(7)：78-83.

Keyword Extraction Method Based on Wildcard Pattern and Random Walk

MA Huifang[1,2]，LI Miao[1]，TONG Haibin[1]，ZHAN Zijun[1]

（1. College of Computer Science and Engineering，Northwest Normal University，Lanzhou 730070，China；
2. Guangxi Key Laboratory of Trusted Software，Guilin University of Electronic Technology，Guilin，Guangxi 541004，China）

【Abstract】 Based on the wildcard patterns and the random walk algorithm with prior information，this paper proposes an improved keyword extraction algorithm. The algorithm uses wildcard constraint to capture the semantic information between words，and extracts the sequential pattern that satisfies the gap constraint and the one-time condition in order to calculate the pattern support degree. When the pattern support degree is not lower than the threshold of minimum support degree，the node association graph is established. The similarity between words in the Wikipedia knowledge base is taken as priori information，and random walks are performed on the association graph by using the PageRank algorithm based on priori information，until the ranking scores stabilize. The Top K words are selected as keywords. Experimental results show that the proposed method has higher extraction accuracy and stability than TextRank，GraphSum and other algorithms.

【Key words】 keyword extraction；wildcard pattern；random walk；gap constraint；PageRank algorithm

DOI： 10.19678/j.issn.1000-3428.0054895

0 概述

随着计算机网络技术的发展与上网用户的增加，网页新闻与各类电子文档逐渐融入人们的生活中，文本关键词提取技术可以帮助用户从海量文档中获取有价值的信息，快速理解文档的核心内容。关键词提取在文本挖掘中主要是根据词项对文本内容的相关程度进行排序，因此单篇文档的关键词提取算法应运而生，并且广泛应用于推荐系统[1-2]、网络广告[3-5]及语义导航[6]等技术中。

关键词提取方法主要分为基于特征的关键词提取方法[7-9]、基于图的关键词提取方法[10-11]和基于语义的关键词提取方法[12-13]。基于特征的关键词提取方法使用句子位置、长度、签名词等度量为每个句子分配一个分数，如文献[8]研究频率对关键词提取的影响。基于图的关键词提取方法将文

基金项目： 国家自然科学基金（61762078，61363058）；甘肃省高等学校创新基金（2020B-089）；广西可信软件重点实验室研究课题（kx202003）；西北师范大学青年教师科研能力提升计划（NWNU-LKQN2019-2）。

作者简介： 马慧芳（1981—），女，教授、博士，主研方向为数据挖掘、机器学习；李 苗、童海斌、詹子俊，本科生。

收稿日期： 2019-05-13 **修回日期：** 2019-07-10 **E-mail：** mahuifang@yeah.net

计算机工程 www.ecice06.com

档表示为密集连接的图,其中每个节点表示一个句子,边连接两个句子,边的权重值表示两个句点之间的相似性,然后使用 PageRank[14] 等图算法对句子进行重要性评分。基于语义的关键词提取方法通过考虑文档内容的潜在语义,生成准确的关键词,如文献[12]利用词汇链表示文本的词汇连贯结构,将提取出包含高出现频次的链词的句子作为文档关键词。本文提出一种基于通配符模式和随机游走算法的关键词提取方法,利用深度优先模式搜索策略发现具有通配符模式的所有实例,通过数据结构层次实例图将模式支持度计算嵌入深度优先模式搜索过程中。

1　通配符模式

序列 S 是有序的项目列表,用 $S=s_1s_2\cdots s_n$ 表示,Σ 是序列 S 中所有可能项的集合,通配符是一种能够将 Σ 中的任意项进行匹配的符号,$g[N,M]$ 表示具有最小间隙 N 和最大间隙 M 的间隙。模式 $P=p_1g[N,M]p_2g[N,M]\cdots g[N,M]p_d$ 是项目和间隙的序列,其中,模式长度用 $|P|$ 表示,为 P 中的项目数。

定义 1(模式出现和实例)　给定模式 $P=p_1p_2\cdots p_d$,序列 $S=s_1s_2\cdots s_n$ 和间隙约束 $g=[N,M]$,如果存在位置序列 $1\leqslant l_1<l_2<\cdots<l_d\leqslant n$,使得 $s_{l_j}=p_j(1\leqslant j\leqslant d)$ 且 $N\leqslant l_j-l_{j-1}-1\leqslant M(2\leqslant j\leqslant d)$,则 $(l_1,l_2,\cdots,l_d)$ 是序列 S 中出现的模式 P。给定模式 P 和序列 S,如果 $(l_1,l_2,\cdots,l_d)$ 是 S 中出现的模式 P,那么 $(l_1,l_2,\cdots,l_d)$ 被认为是序列 S 的 P 实例。

定义 2(一次性条件)　假设 $\mathrm{occ}=(l_1,l_2,\cdots,l_d)$ 和 $\mathrm{occ}'=(l'_1,l'_2,\cdots,l'_d)$ 是序列 S 中模式 P 的 2 个实例。如果对于所有 $1\leqslant p\leqslant d,1\leqslant q\leqslant d$,有 $l_p\neq l'_q$,则这两个实例满足一次性条件。

定义 3(支持度和支持集)　序列 S 中模式 P 的支持度被定义为所有可能的实例集的最大值,其中任何两个实例都满足一次性条件。用 $\mathrm{Sup}(P)$ 表示 P 的支持度,具有 $\mathrm{Sup}(P)$ 大小的实例集被称为 P 的支持集。

定义 4(非重复模式)　给定模式 $P=p_1p_2\cdots p_d$,如果 $\forall 1\leqslant i,j\leqslant d,p_i\neq p_j$,则将 P 称为非重复模式。

定义 5(子模式与超模式)　对于两种模式 $P=p_1p_2\cdots p_m$ 和 $P'=p'_1p'_2\cdots p'_n(n\geqslant m)$,如果存在位置序列 $1\leqslant i_1<i_2<\cdots<i_m\leqslant n$,对于所有 $1\leqslant j\leqslant m$,有 $p'_{i_j}=p_j$,那么 P 被认为是 P' 的子模式,用 $P\subseteq P'$ 表示 P' 是 P 的超级模式。只有在没有超级模式 P 的情况下,模式 P' 才是封闭模式,此时 $\mathrm{Sup}(P)=\mathrm{Sup}(P')$。

在挖掘序列模式时,本文引入 SPMW 算法[15-16]。给定序列 $S=s_1s_2\cdots s_n$,模式 $P=p_1p_2\cdots p_m$,间隙约束 $g[N,M]$,模式 P 的水平实例图表示为二元组 $<V,E>$,其中,V 是节点集,E 是边集。将节点集划分为 m 层,第 $i(1\leqslant i\leqslant m)$ 层节点对应于 p_i 的位置。假设 A 和 B 是 p_i 和 p_{i+1} 的两个节点,如果 A 和 B 的位置在同一个序列上,且满足间隙约束,则有一个从 A 到 B 的父子边和一个从 B 到 A 的子父边。图 1 是一个水平实例图,其中,实线表示父子边缘,虚线表示子父边缘,并且使用虚线连接相同的级别节点。

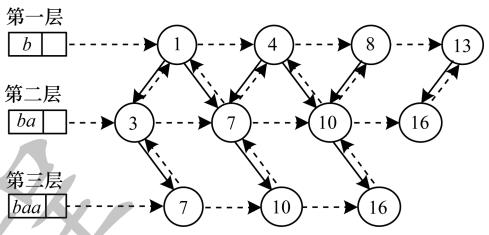

图 1　水平实例图

Fig. 1　Horizontal example graph

在图 1 中,已知序列 $S=bcabccabcacdbdda$ 和间隙约束 $g[0,5]$。模式"b"在序列 S 中出现 4 次,分别在节点 1、4、8 和 13 处。第二层节点是项目 a 的 4 个位置,在满足间隙约束 $g[0,5]$ 时,连接第一层节点。由前 2 个层级节点组成的图是模式 $Q=ba$ 的实例图。类似地,模式 $R=baa$ 的实例图由 3 个层级节点组成。实例图第二层中的节点 16 没有子节点。当计算模式 R 的支持度时,通过对节点 7、10 和 16 的深度优先遍历策略扫描实例图,其中 $<1,3,7>$ 和 $<4,10,16>$ 的出现次数为 2。

2　关键词提取方法

2.1　关联图生成

模式 $P=p_1p_2\cdots p_d$ 表示 d 个词满足间隙约束的关系,d 取正整数。在构建关联图时,以词语之间存在的语义关系来确定节点之间的关系,因为只需要已知项目两两之间的关系,所以模式长度 d 取 2。间隙约束为 $g[N,M]$,则 $P=p_1g[N,M]p_2$。给定序列 S,最小支持度阈值 min_sup,计算出所有满足间隙约束和一次性条件的序列模式集合,并且计算出每个模式的支持度 $\mathrm{Sup}(P)$。将模式 P 中的所有不重复的项作为图的节点,当且仅当模式 P 的支持度 $\mathrm{Sup}(P)$ 大于等于最小支持度阈值 min_sup 时节点之间有边,边的权重为支持度的值,由于可能存在 $p_1g[N,M]p_2\neq p_2g[N,M]p_1$,因此节点之间的边是有向边。

已知序列 $S=bcabccabcacdbdda$,间隙约束 $g[0,2]$,最小支持度阈值 min_sup = 2,所有可能项的集合 $\Sigma=\{a,b,c,d\}$,计算 Σ 集合中任意两项的模式支持度如表 1 所示。将模式中所有不重复的项作为图的节点,图的节点集合为 $\{a,b,c,d\}$,当支持度大于等于 2 时节点之间有边,生成的节点关联图如图 2 所示。

计算机工程 www.ecice06.com

表 1 模式支持度

Table 1 Pattern support degree

节点	a	b	c	d
a	0	3	3	0
b	3	1	2	1
c	2	2	2	0
d	1	1	0	1

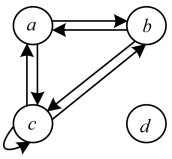

图 2 节点关联图

Fig. 2 Node association diagram

2.2 引入先验信息的随机游走算法

PageRank 是一种计算网页重要程度的算法,该算法认为如果一个网页被很多其他网页链连到,则说明该网页比较重要。模型定义如式(1)所示:

$$\mathrm{PR}(N_i)=\alpha\cdot\sum_{k=1}^{e}\frac{\mathrm{PR}(N_k)}{C(N_k)}+(1-\alpha)\bar{P} \tag{1}$$

其中,N_i 表示第 i 个节点,$\mathrm{PR}(N_i)$ 表示 N_i 的 PageRank 分数,e 表示 N_i 的入边条数,$C(N_i)$ 表示 N_i 的出边条数,α 表示一个节点跳转至非邻居节点的衰减系数,$\bar{P}$ 表示所有节点的先验分数。

在式(1)中,节点之间的跳转概率是相同的,而理论上相似节点之间跳转的可能性会更大。节点之间边的权值越大,节点之间跳转的概率越大。在构建关联图时,将节点之间的支持度作为边的权重,因为支持度的取值为正整数,不利于随机游走计算,所以将支持度进行归一化处理,节点 N_i 跳转到节点 N_k 的概率如式(2)所示:

$$\mathrm{Nor}_{i,k}=\frac{\mathrm{Sup}(N_kN_i)}{\sum_{j=1}^{C(N_k)}\mathrm{Sup}(N_kN_j)} \tag{2}$$

式(1)中的 $\bar{P}$ 通常被初始化为 $1/|C|$,其中 $|C|$ 表示关联图中的节点个数。但是实际情况通常希望更重要的节点分配更高的先验分数,而不太重要的节点的先验分数变低,本文利用知识库给出先验排名分数。文献[16-17]提出的监督随机游走被用于预测社交网络中链接的出现,即监督信息引入到标准的无监督随机游走模型,并将知识库的语义关系融入监督信息。维基百科或 YAGO 的知识库提供内容信息以及内容单元之间的链接信息,在这些链接中含有大量语义信息。维基百科包含超过 400 万篇文章,并提供比其他知识库更多的文字覆盖,因此本文使用工具包 Wikipedia-Miner 得到维基百科知识库中的链接信息。对于 2 个节点 N_i 和 N_j,它们之间的语义相似度计算如式(3)所示:

$$\mathrm{sim}(N_i,N_j)=1-\frac{\log_a(\max(|E_i|,|E_j|))-\log_a(|E_i|\cap|E_j|)}{\log_a(|W|)-\log_a(\min(|E_i|,|E_j|))} \tag{3}$$

其中,E_i 是连接到 N_i 的文档集,$|W|$ 是文件总数。

式(3)表示 2 个节点 N_i 和 N_j 之间的语义相似度,在随机游走时需要节点 N_i 的先验分数,所以分别计算节点 N_i 与其他节点的相似度。为了更形式化地度量一个节点的先验分数,对节点 N_i 的先验分数进行归一化,如式(4)所示:

$$\begin{cases}r_i=\dfrac{1}{z}\sum_{j\in C}\mathrm{sim}(N_i,N_j)\\ z=\sum_{i\in C}\sum_{j\in C}\mathrm{sim}(N_i,N_j)\end{cases} \tag{4}$$

其中,r_i 表示节点 N_i 的先验分数,$r=(r_1,r_2,\cdots,r_{|N|})$ 即为关联图中所有节点的先验概率分布。式(2)修正了节点之间的跳转概率,式(4)引入了知识库中的先验信息。结合式(2)和式(4)修正 PageRank 公式,如式(5)所示:

$$\mathrm{PR}(N_i)=\alpha\times\sum_{k=1}^{e}(\mathrm{PR}(N_k)\times\mathrm{Nor}_{i,k})+(1-\alpha)\times r_i \tag{5}$$

根据式(5)在关联图上随机游走,迭代计算每个节点的 PR 值,直至满足式(6)[18-19],使节点分数达到收敛状态,其中 δ 为随机游走终止阈值,并且使用图中的排名分数 PR 对关键词进行排序,将排名 Top K 个词作为关键词。

$$\sum_{i=1}^{|C|}|\mathrm{PR}^{t+1}(N_i)-\mathrm{PR}^{t}(N_i)|\leqslant\delta \tag{6}$$

3 实验结果与分析

3.1 实验数据与评价指标

为验证本文方法的有效性,本文选取《物种起源》作为中文实验数据。经过预处理操作后有 71 923个词项。选取 MEHRI 与 DAROONEH 合著的“The role of entropy in word ranking”文献(MD' paper)作为英文实验数据。经过预处理操作后有 1 180 个词项。选取维基百科知识库作为先验信息,使用工具包 Wikipedia-Miner 获得词语相似度。根据《物种起源》重要词汇注解表,选取 15 个重要词项作为评价关键词提取是否有效的基准,提取 MD' paper 中的 9 个重要词项作为评价英文关键词提取是否有效的基准。

综合平均精度均值(Mean Average Precision,MAP)、召回率(R)和 F_β 作为关键短语提取的性能指标。设 M_{ret} 表示本文关键词提取结果序列,M_{rel} 表示真实词汇表序列,MAP 定义如式(7)所示:

计算机工程 www.ecice06.com

$$P(i)=\frac{1}{i}\sum_{j=1}^{i}g(M_{\mathrm{ret}}(j),M_{\mathrm{rel}})$$

$$\mathrm{AP}(i)=\frac{\sum_{j=1}^{i}P(j)\times g(M_{\mathrm{ret}}(j),M_{\mathrm{rel}})}{\sum_{j=1}^{i}g(M_{\mathrm{ret}}(j),M_{\mathrm{rel}})}$$

$$\mathrm{MAP}=\frac{1}{M_{\mathrm{ret}}}\sum_{i=1}^{M_{\mathrm{ret}}}\mathrm{AP}(i) \tag{7}$$

其中,$M_{\mathrm{ret}}(j)$表示关键词返回序列 M_{ret}的第 j 个词项,$g(t,M_{\mathrm{rel}})$表示指示函数,若词项 t 出现在原词汇表序列 M_{rel} 中则返回 1,否则返回 0,$P(i)$与 $\mathrm{AP}(i)$分别表示M_{ret}中前 i 个词项的准确率与平均准确率。

F_β准确率是由 MAP 和 R 相结合计算得到,其中 β 取值为 0.5。F_β定义如式(8)所示:

$$R=\frac{|M_{\mathrm{ret}}\cap M_{\mathrm{rel}}|}{|M_{\mathrm{ret}}|}$$

$$F_\beta=\frac{(1+\beta^2)\times\mathrm{MAP}\times R}{\beta^2\times\mathrm{MAP}+R} \tag{8}$$

3.2　关联图模型参数分析

本节对最大间隙 M、最小支持度阈值 min_sup 和衰减系数 α 这 3 个参数进行分析。在分析 min_sup 时,最大间隙 M 取 3,在中文数据上的准确率如图 3 所示,在英文数据上的准确率如图 4 所示。图 3 结果表明,在 min_sup 取 5 时,F_β 具有最高的准确率,所以在中文数据上 min_sup 的最优取值为 5。

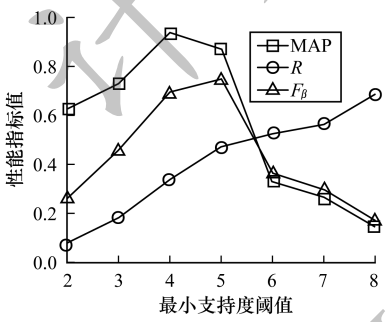

图 3　最小支持度阈值对准确率的影响(中文)

Fig. 3　Effect of minimum support degree threshold on accuracy(Chinese)

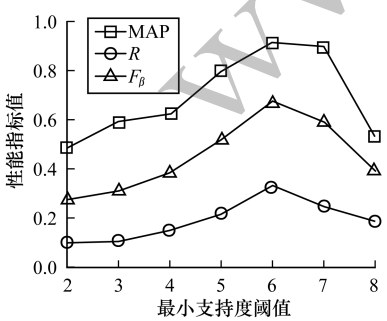

图 4　最小支持度阈值对准确率的影响(英文)

Fig. 4　Effect of minimum support degree threshold on accuracy(English)

在分析最大间隙 M 时,min_sup 取 5,在中文数据上的准确率如图 5 所示,在英文数据上的准确率如图 6 所示。图 5 结果表明 MAP 和 F_β都在 M 取 4 时达到峰值。图 6 结果表明,当 M 取 3 时,MAP 具有最高的平均精度均值。造成中英文数据上 M 的最优取值不同的原因是中文数据《物种起源》是一个较长的文本,而 MD'paper 英文数据相比于中文数据来说较短。

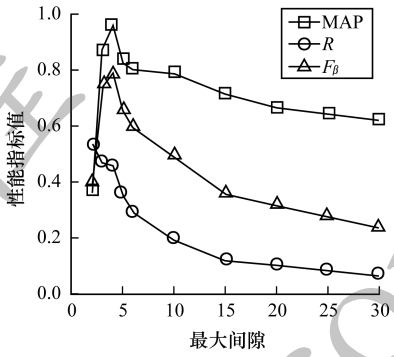

图 5　最大间隙对准确率的影响(中文)

Fig. 5　Effect of maximum clearance on accuracy(Chinese)

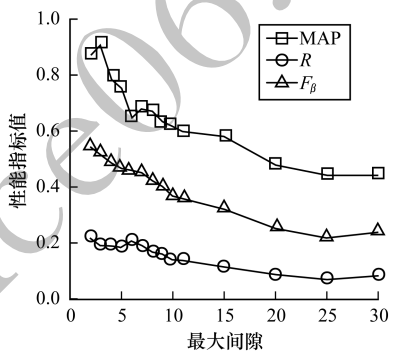

图 6　最大间隙对准确率的影响(英文)

Fig. 6　Effect of maximum clearance on accuracy(English)

在分析衰减系数 α 时,采用中文数据,当 min_sup 取 5、最大间隙 M 取 4 时,得出的 MAP 如图 7 所示。当 α 取 0.55 时,MAP 达到峰值。

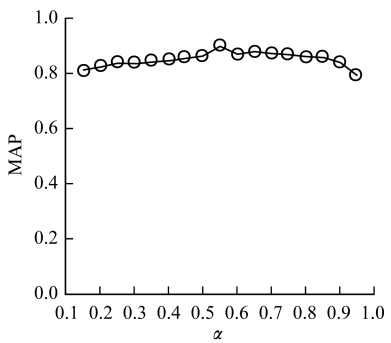

图 7　衰减系数 α 对 MAP 的影响

Fig. 7　Effect of attenuation coefficient α on MAP

计算机工程 www.ecice06.com

3.3 关键词提取性能对比及分析

为验证本文关键词提取算法的准确性与优越性,选取 TextRank、GraphSum 算法与本文算法进行比较分析。实验结果如表 2、表 3 所示,其中阴影部分表示该算法提取出的关键词在关键词参考集中不存在。可见,在中文数据及英文数据上,GraphSum 算法得到的关键词与 TextRank 算法得到的关键词相比更为准确。然而,与本文方法相比,使用 GraphSum、TextRank 算法得到的结果均有所欠缺。3 种方式在中文数据和英文数据进行关键词提取时得到的 MAP、R 和 F_β 如表 4 所示。本文方法在中文数据和英文数据上进行关键词提取得到的 MAP 均大于其他两种算法。

表 2 3 种关键词提取方式的结果对比(中文)

Table 2 Result comparison of three keyword extraction modes(Chinese)

词序	词汇	本文方法	TextRank 算法	GraphSum 算法
1	物种	物种	物种	物种
2	变异	选择	类型	变异
3	遗传	遗传	动物	生活
4	本能	变异	变化	灭绝
5	自然选择	变种	差异	遗传
6	不育 杂交	植物	植物	分化
7	隔离	性状	生物	性状
8	灭绝	家养	变异	变种
9	性状	生物	生活	祖先
10	地理 分布	斗争	后代	亲缘
11	生存 斗争	杂交	种	化石
12	祖先	种群	构造	差异
13	后代	祖先	性	后代
14	化石	后代	自然选择	自然选择
15	种群	灭绝	变种	植物

表 3 3 种关键词提取方式的结果对比(英文)

Table 3 Result comparison of three keyword extraction modes(English)

词序	词汇	本文方法	TextRank 算法	GraphSum 算法
1	entropy	entropy	keyword	word rank
2	keyword	keyword	text	keyword
3	complexity	word	entropy	entropy
4	word rank	text	rank	text entropy
5	text mine	information	retrieve	information
6	information	maximum	system	extraction
7	complex system	system	frequence	metric measure
8	systematic	distance	standard	text keyword
9	statistical	node	extraction	retrieve

表 4 3 种关键词提取方式的性能对比结果

Table 4 Performance comparison results of three keyword extraction modes(English)

性能指标	中文			英文		
	本文方法	TextRank 算法	GraphSum 算法	本文方法	TextRank 算法	GraphSum 算法
MAP	0.890 7	0.749 0	0.861 4	0.912 7	0.259 3	0.870 3
R	0.458 1	0.266 6	0.600 0	0.190 5	0.111 1	0.444 4
F_β	0.749 2	0.550 1	0.792 4	0.519 1	0.204 7	0.730 1

为验证本文方法运行效率,在 Celeron 1.40 GHz处理器的 Windows 10 操作系统下,给出本文方法在不同词量规模下的运行时间,如图 8所示。由于本文考虑了语义信息和先验信息,因此本文方法的执行效率会比其他算法更低。但随着词量规模的扩大,本文方法的运行时间接近于线性增长。

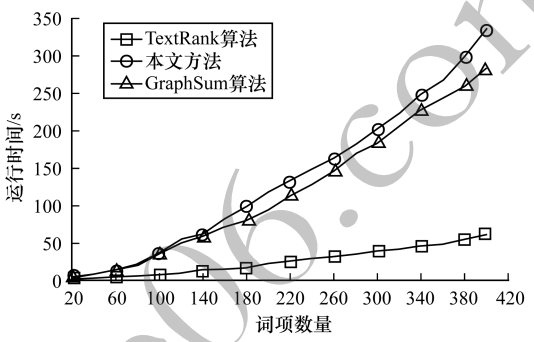

图 8 3 种方式的运行时间比较

Fig. 8 Comparison of operation time of three modes

4 结束语

本文提出一种基于通配符模式和随机游走算法的关键词提取方法。该方法基于通配符约束和一次性条件来挖掘序列模式,使用深度优先搜索策略计算模式支持度,挖掘出具有间隙约束的所有模式实例,并在模式支持度大于等于最小支持度阈值时构建关联图,同时通过引入先验信息的 PageRank 算法获取排名前 Top K 个词语作为关键词。实验结果表明,本文方法相比传统关键词提取算法具有更高的准确率和稳定性。后续可将句子位置、签名词、图结构等信息引入到随机游走算法中,进一步降低关键词提取算法复杂度并提高执行效率。

参考文献

[1] JI Ke, SHEN Hong. Addressing cold-start: scalable recommendation with tags and keywords [J]. Knowledge-Based Systems, 2015, 83: 42-50.

[2] WEN Junhao, YUAN Peilei, ZENG Jun, et al. Research on collaborative filtering recommendation algorithm based on topic of tags[J]. Computer Engineering, 2017, 43(1): 247-252. (in Chinese)

www.ecice06.com 计算机工程

文俊浩,袁培雷,曾骏,等. 基于标签主题的协同过滤推荐算法研究[J]. 计算机工程,2017,43(1):247-252.
[3] XU Guangong,WU Zongda,LI Guiling,et al. Improving contextual advertising matching by using Wikipedia thesaurus knowledge[J]. Knowledge and Information Systems,2015,43(3):599-631.
[4] MISHRA R,KUMAR P,BHASKER B. A Web recommendation system considering sequential information[J]. Decision Support Systems,2015,75:1-10.
[5] YOU W, FONTAINE D, BARTS, J P. An automatic keyphrase extraction system for scientific documents[J]. Knowledge and Information Systems,2013,34(3):691-724.
[6] YE Ning,LIANG Zuopeng,DONG Yisheng. A Web site navigation based on ant colony algorithm[J]. Journal of Applied Sciences,2003,21(4):357-361. (in Chinese)
业宁,梁作鹏,董逸生. 基于蚁群算法的 Web 站点导航[J]. 应用科学学报,2003,21(4):357-361.
[7] LIU Jun, ZOU Dongsheng, XING Xinlai, et al. Keyphrase extraction based on topic feature[J]. Application Research of Computers, 2012, 29(11): 4224-4227. (in Chinese)
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.
[8] ZHANG Qingguo,XUE Dejun,ZHANG Zhenhai,et al. Automatic keyword extraction from massive data sets based on feature combination[J]. Journal of the China Society for Scientific and Technical Information,2006,25(5):587-593. (in Chinese)
张庆国,薛德军,张振海,等. 海量数据集上基于特征组合的关键词自动抽取[J]. 情报学报,2006,25(5):587-593.
[9] CHANG Yaocheng, ZHANG Yuxiang, WANG Hong, et al. Features oriented survey of state-of-the-art keyphrase extraction algorithms[J]. Journal of Software, 2018, 29(7):224-248. (in Chinese)
常耀成,张宇翔,王红,等. 特征驱动的关键词提取算法综述[J]. 软件学报,2018,29(7):224-248.
[10] LIU Xiaojian. XIE Fei, WU Xindong. Graph based keyphrase extraction using LDA topic model[J]. Journal of the China Society for Scientific and Technical Information,2016,35(6):664-672. (in Chinese)
刘啸剑,谢飞,吴信东. 基于图和 LDA 主题模型的关键词抽取算法[J]. 情报学报,2016,35(6):664-672.
[11] XIE Wei, SHEN Yi, MA Yongzheng. Recommendation system for paper reviewing based on graph computing[J]. Application Research of Computers,2016,33(3):798-801. (in Chinese)
谢玮,沈一,马永征. 基于图计算的论文审稿自动推荐系统[J]. 计算机应用研究,2016,33(3):798-801.
[12] SUO Hongguang,LIU Yushu,CAO Shuying. A keyword selection method based on lexical chains[J]. Journal of Chinese Information Processing,2006,20(6):25-30. (in Chinese)
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):25-30.
[13] HOFMANN T. Unsupervised learning by probabilistic latent semantic analysis[J]. Machine Learning, 2001, 42(1):177-196.
[14] BRIN S, PAGE L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer Networks and ISDN Systems,1998,30(1):107-117.
[15] XIE F, WU X, ZHU X. Efficient sequential pattern mining with wildcards for keyphrase extraction[J]. Knowledge-Based Systems,2017,115:27-39.
[16] TURNEY P D. Learning algorithms for keyphrase extraction[J]. Information Retrieval, 2002, 2(4): 303-336.
[17] NIU Xinzheng,NIU Jiajun. Community detection based on weighted content-structural network and random walks[J]. Acta Electronica Sinica,2017,45(9):2135-2142. (in Chinese)
牛新征,牛嘉郡. 基于加权内容-结构网络和随机游走的社团划分算法[J]. 电子学报, 2017, 45(9): 2135-2142.
[18] HSU C C,LAI Y A,CHEN W H,et al. Unsupervised ranking using graph structures and node attributes[C]// Proceedings of the 20th ACM International Conference on Web Search and Data Mining. New York,USA:ACM Press,2017:771-779.
[19] ZHU Liang, LU Jingya, ZUO Wanli. Query-doc association mining based on user search behavior[J]. Acta Automatica Sinica, 2014, 40(8): 1654-1666. (in Chinese)
朱亮,陆静雅,左万利. 基于用户搜索行为的 query-doc 关联挖掘[J]. 自动化学报,2014,40(8):1654-1666.

编辑　陆燕菲

计算机工程 www.ecice06.com