文章编号：1003-0077(2019)09-0069-10

融合图结构与节点关联的关键词提取方法

马慧芳[1,2]，王 双[1]，李 苗[1]，李 宁[3]

(1. 西北师范大学 计算机科学与工程学院，甘肃 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；
3. 中国科学院 信息工程研究所，北京 100093)

摘 要：单篇文本的关键词提取可应用于网页检索、知识理解与文本分类等众多领域。该文提出一种融合图结构与节点关联的关键词提取方法，能够在脱离外部语料库的情况下发现单篇文本的关键词。首先，挖掘文本的频繁封闭项集并生成强关联规则集合；其次，取出强关联规则集合中的规则头与规则体作为节点，节点之间有边当且仅当彼此之间存在强关联规则时，边权重定义为关联规则的关联度，将强关联规则集合建模成关联图；再次，综合考虑节点的图结构属性、语义信息和彼此的关联性，设计一种新的随机游走算法计算节点的重要性分数；最后，为了避免抽取的词项之间有语义包含关系，对节点进行语义聚类并选取每个类的类中心作为关键词提取结果。通过设计关联图模型参数的选取、关键词的提取规模、不同算法对比 3 个实验，在具有代表性的中英文数据上证明了该方法能够有效提升关键词提取的效果。

关键词：关键词提取；随机游走；节点属性；语义信息；节点关联

中图分类号：TP391　　**文献标识码**：A

A Keywords Extraction Method via Graph Structure and Nodes Association

MA Huifang[1,2], WANG Shuang[1], LI Miao[1], LI Ning[3]

(1. College of Computer Science and Engineering, Northwest Normal University, Lanzhou, Gansu 730070, China; 2. Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, Guangxi 541004, China;
3. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China)

Abstract: Keywords extraction is an important technique for web page retrieval, knowledge comprehension, and document classification, etc. In this paper, a novel keywords extraction method of combining graph structure with nodes association(GSNA) is proposed, which is able to locate keywords without a corpus. Firstly, the frequent closed itemset are exploited and the strong association rules are generated. Secondly, an association graph is constructed based on association rules, where the head and the body of the rules represent nodes, and an edge exists if and only if there is a strong association rule between two nodes and value of lift are adopted to represent weight. Thirdly, three node factors (i. e. graph structure, node semantics and associations) are unified under the same keyword extraction framework for random walking. Finally, a trustworthy sematic clustering algorithm is employed to avoid the semantic overlapping among terms. Three experiments conducted on the Chinese and English data sets show that GSNA is effective for keywords extraction.

Keywords: keywords extraction; random walk; node attribution; semantic information; node association

0 引言

随着网络技术的普及，网页新闻与各类电子文档快速地融入人们的生活，用户如何从海量文档中获取有价值的信息，文本关键词提取技术显得至关重要。在大多数的文本挖掘任务中，关键词提取均表现为根据词项对文本内容的相关程度对其排序，

收稿日期：2018-12-17　**定稿日期**：2019-03-11

基金项目：国家自然科学基金(61762078,61802404,61363058)；广西可信软件重点实验室研究课题(kx201705)

所以各种单篇文本的关键词提取算法也随之而生。

单篇文本的关键词提取技术依赖于词项所处的上下文语境，在去除停用词并按照某种特定规则生成关键词的候选集后，可采用基于统计的方法[1-4]、基于潜在语义分析的关键词提取方法[5-6]、基于图的关键词排序方法[7-10]从候选集中选取关键词。基于统计的方法关注词项在文本的内部统计特征和在语料库的外部统计特征，例如，词频—逆文档频率[1](term frequency-inverse document frequency，TF-IDF)、JSD[2](Jensen-Shannon Divergence)等算法关注词频与词项位置等内部特征统计信息，基于Query logdoe的抽取关键词方法[3]、语义关联抽取关键词方法[4]等结合搜索引擎记录关注词项在语料库中的词频、链接次数等外部特征统计信息，通过计算权值选出高权重的词项作为关键词。从潜在语义而言，PLSA[5]，LDA[6]采用文档主题生成模型技术来识别大规模文档集中潜在的语义信息，其重点关注建立词项、主题和文档三层结构，并得到词项与文档之间的隐藏语义关系。基于图对关键词排序算法的基本来源为PageRank[7]。PageRank模仿上网者以一定概率在各个页面游走，经过有限次游走可到达一个稳定状态的概率分布，该分布即为节点的排名依据。TextRank[8]将PageRank应用于文本的关键字和关键句提取，可自动生成文本摘要。GraphSum[9]认为节点之间的关联性联系有强弱之分，负关联关系的节点之间投票时应降低相应的PageRank分数。AttriRank[10]指出计算节点的重要性时不仅要考虑图的结构，更要关注节点属性之类的外部信息。

本文综合考虑节点的图结构属性、语义信息与节点间的关联性特征，提出了一种融合图结构与节点关联的关键词提取方法(A Keywords Extraction Method via Combining Graph Structure with Nodes Association，GSNA)。首先，为了避免挖掘冗余的频繁项集造成执行效率低下，本文挖掘文档的频繁封闭项集并生成强关联规则集合；其次，将强关联规则集合中不重复的规则头与规则体作为节点，节点之间有边当且仅当彼此存在强关联规则时，以关联规则的关联强度作为边权重构建文档的关联图；然后，使用GSNA在关联图上随机游走，迭代计算每个节点的重要性分数；最后，为了避免提取的关键词之间有语义包含关系，对结果降序排序，并选取前若干个节点聚类，取出所有的类中心作为文档的关键词提取结果。基本技术流程如图1所示。

本文致力于脱离语料库的单篇长文本关键词提取，依据节点的图结构属性、语义信息和彼此间的关联性对词项排序，主要贡献如下：

(1) 通过挖掘频繁封闭项集生成强关联规则用于构建关联图，得到的强关联规则规模小、速度快且无信息损耗，避免了重复扫描事务集与提取冗余的频繁项集导致图结构过于复杂的问题，提高了算法的整体执行效率。

(2) 综合考虑关联图中节点的图结构属性与语义信息，计算节点在关联图中的相似度，增大与关联图相似程度更大的节点被选中的概率，解决了节点之间等概率跳转的缺陷。

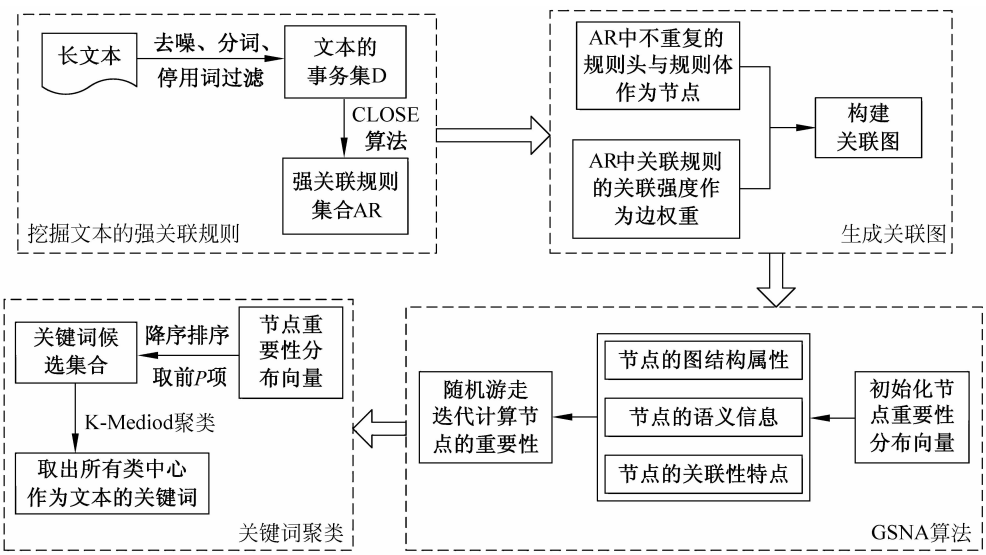

图1 基本技术流程图

(3) 考虑节点间的关联性事实，在关联图上随机游走时，节点间投票有正负关联之分，加强节点间的正关联投票分数，降低节点间的负关联投票分数，使得节点得分更加合理。

1　相关理论

文档可视为由句子构成的事务数据库，文档的关联图刻画了节点所代表词项之间的关联关系。本文的关键词提取算法作用于关联图之上，首先将文本处理为一个事务集，然后将 CLOSE[11] 算法作用于该事务集，挖掘强关联规则集合，最后将强关联规则集合建模成关联图。

1.1　文本预处理

文本在经过数据去噪、分词、停用词过滤等预处理步骤后，可视为由一组停止标记(“。”，“?”，“!”，“……”)分隔的句子集合 $S=\{s_1,s_2,\cdots,s_n\}$。s_i 是由一组不重复的词项序列构成的句子，w_{iq} 是 s_i 的第 q 个词。假设 s_i 存在与之对应的事务 t_i，那么 w_{iq} 可视为事务 t_i 的第 q 个词项。所以，由 S 中每个句子对应的事务就构成了文档的事务集 $T=\{t_1,t_2,\cdots,t_n\}$。表 1 是一篇文档的原始内容，经过预处理后，得到该文本的事务集 T，如表 2 所示。

表 1　原始文档

这是一篇关于文本分析与文本挖掘的文章。特别之处在于，它能够分析出隐藏在数据中的文本信息！通过文本分析，我们能够加强对文本内容的理解。

表 2　文档的事务集

事务 ID	项集
t_1	文本，分析，挖掘
t_2	分析，隐藏，数据，文本，信息
t_3	文本，分析，加强，内容，理解

1.2　CLOSE 算法

关联规则[11]反映了事务中不同项集之间的关联关系，基本形式为 $A\rightarrow B$，其中 $A\cap B=\varnothing$，$A\neq\varnothing$ 且 $B\neq\varnothing$。关联规则的挖掘建立在事务集之上，文档的事务集为 T，词典为 $I=\{w_1,w_2,\cdots,w_m\}$，基本概念定义如下：

定义 1(k-项集)　一个项集中项的数目称为项集的长度，长度为 k 的项集称为 k-项集。

定义 2(项集的支持度)　令 T_A 表示 T 中包含项集 A 的事务集合，则 A 在事务集的支持度如式(1)所示。

$$\sup(A)=\frac{|T_A|}{|T|} \tag{1}$$

定义 3(关联规则的支持度)　对于关联规则 $A\rightarrow B$，$A\subseteq I$，$B\subseteq I$，$A\cap B=\varnothing$，令 $T_{A\cup B}$ 表示 T 中包含 $A\cup B$ 的事务集，则关联规则 $A\rightarrow B$ 的支持度如式(2)所示。

$$\sup(A\rightarrow B)=\frac{|T_{A\cup B}|}{|T|} \tag{2}$$

定义 4(关联度)　对于关联规则 $A\rightarrow B$，$A\subseteq I$，$B\subseteq I$，$A\cap B=\varnothing$，则 $A\rightarrow B$ 的关联度如式(3)所示。

$$\mathrm{lift}(A\rightarrow B)=\frac{\sup(A\rightarrow B)}{\sup(A)\times\sup(B)} \tag{3}$$

由于 $\mathrm{lift}(A\rightarrow B)=\mathrm{lift}(B\rightarrow A)$，所以 lift 满足对称性。由条件概率分析可知，当 $\mathrm{lift}(A\rightarrow B)=1$ 时，A 与 B 服从独立概率分布，表明 A 与 B 无关联关系；当 $\mathrm{lift}(A\rightarrow B)>1$ 时，表示引入 A 后 B 的后验概率高于 B 的先验概率，说明 A 与 B 之间存在正关联关系；反之，当 $\mathrm{lift}(A\rightarrow B)<1$ 时，表示 A 与 B 之间存在负关联关系。

定义 5(强关联规则)　对于关联规则 $A\rightarrow B$，$A\subseteq I$，$B\subseteq I$，$A\cap B=\varnothing$，若满足 $\sup(A\rightarrow B)\geqslant\min\text{-}\sup$ 且 $\mathrm{lift}(A\rightarrow B)\leqslant\max^{-\mathrm{lift}}$ 或 $\mathrm{lift}(A\rightarrow B)\geqslant\min^{+\mathrm{lift}}$，则 $A\rightarrow B$ 为强关联规则。其中，minsup 为关联规则的最小支持度阈值，$\max^{-\mathrm{lift}}$ 为最大负关联度阈值，$\min^{+\mathrm{lift}}$ 为最小正关联度阈值。

定义 6(频繁封闭项集)　项集 X 为频繁封闭项集当且仅当 $\sup(X)\geqslant\mathrm{minsup}$ 且 $S(X)=X$。其中，(X) 为 X 的闭包运算[12]，minsup 为 X 在事务集中的最小支持度。

为了快速地从文档事务集中提取有效的强关联规则，本文应用 CLOSE 算法挖掘频繁封闭项集，进而生成强关联规则集合，该算法使用闭包运算，结合广度优先策略搜索项集空间，实现非频繁项集与非封闭项集的快速剪枝。相比于传统的关联规则挖掘，CLOSE 算法提高了执行效率。在获取到不同长度的频繁项集集合 $L=\{L_1,L_2,\cdots,L_k\}$ 后，由 L_i 中的 i-频繁项集生成 m-项集($1\leqslant m\leqslant i\text{-}1$)的规则头与 $(i\text{-}m)$-项集的规则体；然后，根据 L 计算 lift(m-项集$\rightarrow(i\text{-}m)$-项集)，对 CLOSE 算法中抽取强关联规则的步骤改进；最后，取出 $\mathrm{lift}\leqslant\max^{-\mathrm{lift}}$ 或 $\mathrm{lift}\geqslant\min^{+\mathrm{lift}}$ 对应的强关联规则。本文对 CLOSE 算法中抽取强

关联规则的改进步骤如表3所示。

表3 强关联规则生成步骤

输入	不同长度的频繁项集集合 $L=\{L_1,L_2,\cdots,L_k\}$，$\max^{-\text{lift}}$，$\min^{+\text{lift}}$
输出	强关联规则集合 AR
1)	AR=∅；
2)	for c in L_i where $i\geqslant 2$ begin //对 $L_2\sim L_k$ 循环每个 i-频繁项集 c
3)	$Hl=\{l\text{-itemset} \mid l\text{-itemset}\subseteq c, 1\leqslant l\leqslant i-1\}$//取出 c 的所有1-项集～(i-1)-项集作为关联规则的规则头
4)	for hl in Hl begin //将每个 hl-项集作为规则头，(c-hl)-项集作为规则体
5)	lift(hl-itemset，(c-hl)-itemset)=support(c)/(support(hl-itemset)×support((c-hl)-itemset))；
6)	if (lift $\leqslant\max^{-\text{lift}}$ or lift $\geqslant\min^{+\text{lift}}$) then
7)	AR=AR∪{hl-itemset→(c-hl)-itemset}；
	end
	end
8)	return AR

1.3 关联图构建

从文档事务集中挖掘得到的强关联规则集合为AR，将AR中所有不重复的规则头与规则体作为关联图的节点集合 N。关联图中的任意两个节点之间有边当且仅当彼此存在满足 lift$\leqslant\max^{-\text{lift}}$或 lift$\geqslant\min^{+\text{lift}}$的强关联规则，由于 lift 具有对称性，所以节点之间只要有强关联关系就一定存在双向边，且边的权重定义为 lift 值。

表4为预处理后的文档事务集样例，假设 minsup=0.6，$\min^{+\text{lift}}=1.2$，$\max^{-\text{lift}}=0.95$，由 CLOSE 算法得到该事务集的频繁封闭项集为 FC={{a,c},{b,e},{c},{b,c,e}}，导出强关联规则集合 AR={$a\leftrightarrow c$,$c\leftrightarrow be$,$c\leftrightarrow b$,$c\leftrightarrow e$,$b\leftrightarrow ce$,$b\leftrightarrow e$,$e\leftrightarrow bc$}，抽取出AR中不重复的规则头与规则体作为关联图节点 N={{b,c},{b,e},{c,e},{a},{b},{c},{e}}，以强关联规则的关联强度作为边权重，构建得到如图2所示的无向加权关联图。

表4 文档事务集样例

事务 ID	项集
tr_1	a,c,d
tr_2	b,c,e
tr_3	a,b,c,e
tr_4	b,e
tr_5	a,b,c,e

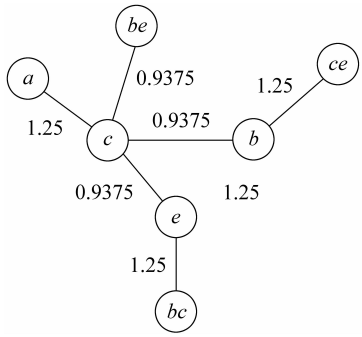

图2 文档的关联图

1.4 PageRank 算法

PageRank[7]为网页排名算法，其认为节点 A 有边指向 B 可看作 A 对 B 投票，节点的入边权重总和越高则该节点的重要性越高。PageRank 模仿上网者，能够在关联图上的相邻节点之间移动，也可以从当前节点跳转至非邻居节点。在经过有限足够多次游走后，各点达到稳定状态，即节点的重要性分数趋于定值。每经过一次游走迭代，就会产生一次节点排序，该排序值即为 PageRank 分数。节点 N_i 为关联图中的第 i 个节点，e 表示 N_i 的入边条数，PR(N_i)表示 N_i 的 PageRank 分数，$C(N_i)$表示 N_i 的出边条数，d 用于衡量一个节点跳转至另一个节点的衰减系数，N_i 的 PageRank 分数定义如式(4)所示。

$$\mathrm{PR}(N_i)=(1-d)\times\frac{1}{|N|}+d\times\sum_{k=1}^{e}\frac{\mathrm{PR}(N_k)}{C(N_k)} \tag{4}$$

其中，节点间的跳转为等概率事件即 $1/|N|$，而理论上，相似节点之间跳转的可能性会更大。另外，节点之间还存在正关联关系和负关联关系，所谓正关联关系为一个节点的出现后往往伴随着另一个节点出现，而负关联关系为一个节点出现后使得另一个节点出现的概率降低。对于节点间的正关联关系，需要增大 PR 投票分数，而对于负关联关系，应降低 PR 的投票分数。

2 融合图结构与节点关联的关键词提取

关联图中不同的节点对文本的重要性往往是不同的，GSNA 算法分别从节点的图结构相似性与关联性对 PageRank 加以改进。衡量节点的相似性时，GSNA 考虑了节点在关联图中的结构属性与隐含的语义信息；衡量节点的关联性时，需要增强正关

联传播，降低负关联传播。GSNA 合理地调节 PR 分数，使得节点排名更准确。

2.1　节点间的图结构与语义相似性

从图结构形式而言，若两节点与其余节点的连接形式相似，则它们极可能为图结构形式相似节点。

为了量化节点的图结构属性，本文使用表 5 所示的 6 个指标。其中，attr1～attr3 直观地表示了一个节点与其余节点的关联能力，attr4 体现了一个节点与周围邻居节点的相似匹配程度，attr5、attr6 定量地说明了一个节点在关联图中的传播能力。对于节点 N_i，可将其形式化表示为 $\mathrm{Attr}_i=(\mathrm{attr1},\mathrm{attr2},\mathrm{attr3},\mathrm{attr4},\mathrm{attr5},\mathrm{attr6})$。

表 5　节点的结构属性表

属性	定义
attr1	距离为 1 的节点个数
attr2	距离为 2 的节点个数
attr3	距离为 3 的节点个数
attr4	同配性(attr1/邻居节点度的平均值)
attr5	attr2/attr1
attr6	attr3/attr2

从节点的语义信息而言，若两节点与词典中词项的共现分布情况相似，则它们有可能为语义相似节点。预处理后文档的词典为 $I=\{w_1,w_2,\cdots,w_m\}$，节点 N_i 与 I 的共现分布为 $P=\{p_1,p_2,\cdots,p_m\}$，p_j 表示同时包含 N_i 与 w_j 的句子在事务集中的归一化值，N_j 与 I 的共现分布为 $Q=\{q_1,q_2,\cdots,q_m\}$，则 N_i 与 N_j 的语义距离[3]如式(5)所示。

$$\begin{cases}\mathrm{JSD}(N_i,N_j)=\dfrac{1}{2}\left(\displaystyle\sum_{i=1}^{m}p_i\times\log_2\frac{p_i}{m_i}+\sum_{i=1}^{m}q_i\times\log_2\frac{q_i}{m_i}\right)\\ m_i=\dfrac{p_i+q_i}{2}\end{cases}\tag{5}$$

综合节点的图结构属性与语义信息，节点 N_i 与 N_j 的相似度衡量如式(6)所示。当 N_i 与 N_j 的图结构属性或语义分布差异越大时，相似度 s_{ij} 越低，反之当 N_i 与 N_j 的结构属性或语义分布越接近时，s_{ij} 相应越高。

$$\begin{cases}s_{ij}=\dfrac{1}{\mathrm{JSD}(N_i,N_j)}\times\mathrm{e}^{-\|\mathrm{Attr}_i-\mathrm{Attr}_j\|_2^2}\\ \|\mathrm{Attr}_i-\mathrm{Attr}_j\|_2^2=\|\mathrm{Attr}_i\|^2+\|\mathrm{Attr}_j\|^2-2\mathrm{Attr}_i^T\cdot\mathrm{Attr}_j\end{cases}\tag{6}$$

关联图中存在强关联关系的节点之间有相似度，无强关联关系的节点之间相似度为 0，则可形成关联图的相似度矩阵 $S_{|N|\times|N|}$。为了更形式化地度量一个节点与所有节点的相似程度，本文将每个节点在关联图中的相似度计算归一化，如式(7)所示。

$$\begin{cases}r_i=\dfrac{1}{z}\displaystyle\sum_{j\in V}s_{ij}\\ z=\displaystyle\sum_{i\in V}\sum_{j\in V}s_{ij}\end{cases}\tag{7}$$

其中，r_i 表示 N_i 与图中所有节点的相似程度，使用 r_i 修正 PageRank 公式，使得节点不再服从等概率跳转，而是使与关联图相似程度更大的节点被选中的概率更高，如式(8)所示。

$$\mathrm{PR}(N_i)=(1-d)\times r_i+d\times\sum_{k=1}^{e}\frac{\mathrm{PR}(N_k)}{C(N_k)}\tag{8}$$

由于关联图为无向图，所以式(8)中 $C(N_k)$ 表示节点 N_k 的度，e 表示 N_k 的邻居个数。

2.2　节点间的关联性

在关联图中，两节点之间有边当且仅当它们之间存在强关联规则。为了加强正关联节点间的 PR 投票分数，降低负关联节点间的投票分数，本文使用 lift 值对式(8)改进得到式(9)。

$$\mathrm{PR}(N_i)=(1-d)\times r_i+d\times\sum_{k=1}^{e}\frac{\mathrm{PR}(N_k)\times\mathrm{lift}(N_i,N_k)}{C(N_k)}\tag{9}$$

式(9)融合了节点的图结构属性、语义信息与节点之间的关联性，使得节点重要性排名更加合理。$\mathrm{lift}(N_i,N_k)$表示 N_i 与 N_k 的关联强度，当 N_i 与 N_k 为负关联关系时 $\mathrm{lift}<1$，为正关联关系时 $\mathrm{lift}>1$，所以 lift 能够根据正负关联性恰当地放缩 $\mathrm{PR}(N_k)$的大小，实现了加强正关联传播、降低负关联传播的目的。

根据式(9)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(10)，使节点分数达到收敛状态，其中 δ 为随机游走终止阈值，往往取为 10^{-4}。

$$\sum_{i=1}^{|N|}|\mathrm{PR}^{t+1}(N_i)-\mathrm{PR}^{t}(N_i)|\leqslant\delta\tag{10}$$

2.3　关键词语义聚类

为了避免提取的关键词之间有语义包含关系，对词项的 PageRank 分数降序排序后进行语义聚类，最后取出所有的类中心作为文档的关键词提取结果。

关联图的节点是根据不同长度的频繁封闭项集取子集形成的，k-项集一共含有 2^k-1 个非空子集。例如，1-项集一定是$(k-1)$-项集的子集，这种现象可视为$(k-1)$-项集对 1-项集有语义包含关系。为了防止得到的关键词之间存在语义包含关系，本文将排序后前 P 个节点进行 K-Medoid 聚类[12]并抽取所有的类中心作为最终的关键词。K-medoid 利用语义距离作为类中心选择的基准，保证类中心的语义能够兼收该类中其他词项的语义。

本文抽取出文本的强关联规则集合 AR 后，依据 AR 构建文本的关联图，使用 GSNA 在关联图上随机游走计算每个节点的 PR 分数，并对节点聚类得到文档的关键词。首先，初始化每个节点的 PR 分数为 $1/|N|$；其次，计算相似度矩阵 $S_{|N|\times|N|}$，并归一化得到每个节点在关联图中的固定相似度值 r_i；然后，在关联图上随机游走，迭代计算每个节点的 PR 值直至收敛；最后，对前 P 个词项聚类，取所有的类中心作为文档的关键词。

3 实验与结果分析

为了验证本文算法的可行性，首先选取中英文数据和恰当的评价指标，其次寻找最佳关联图模型对应的输入参数，然后在最佳关联图的情形下对关键词的抽取规模做出讨论，最后选取 4 种关键词提取算法与 GSNA 对比，验证 GSNA 的优越性。

3.1 实验数据与评价指标

为了验证 GSNA 算法的有效性，本文选取达尔文的汉译版著作《物种起源》[13]作为中文实验数据，选取 Li 编写的文献[14](Li'paper)作为英文实验数据。以《物种起源》为代表的大规模数据和以 Li'paper 为代表的小规模数据，已被许多学者用于验证各种新型关键词提取算法的准确性[2-8]。在经过数据去噪、中文文本分词、英文数据大小写转换与词干还原、停用词过滤和事务词项去重等预处理步骤后，《物种起源》共含有 39 599 个词项、3 495 个句子，分布在 14 个章节的各个段落；Li'paper 共含有 3 315 个词项，402 个句子。此外，本文分别选取《物种起源》的重要词汇注解表中的 15 个重要词项与专家对 Li'paper 列举的 7 个关键词作为评价中文与英文关键词提取是否有效的基准。

本文将提取的关键词序列标记为 M_{ret}，将词汇表序列标记为 M_{rel}，所涉及的评价指标包括 MAP[15]（Mean Average Precision）、召回率 Recall[3]和 F_β。其中，MAP 定义如式(11)所示。

$$\begin{aligned} P(i) &= \frac{1}{i}\sum_{j=1}^{i} g(M_{\text{ret}}(j), M_{\text{rel}}),\\ \text{AP}(i) &= \frac{\sum_{j=1}^{i} P(j)\times g(M_{\text{ret}}(j), M_{\text{rel}})}{\sum_{j=1}^{i} g(M_{\text{ret}}(j), M_{\text{rel}})},\\ \text{MAP} &= \frac{1}{M_{\text{ret}}}\sum_{i=1}^{M_{\text{ret}}} \text{AP}(i) \end{aligned} \tag{11}$$

式(11)中，$M_{\text{ret}}(j)$表示关键词返回序列 M_{ret} 的第 j 个词项；$g(t, M_{\text{rel}})$为指示函数，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 $\text{AP}(i)$分别表示 M_{ret} 中前 i 个词项的准确率与平均准确率。

本文期望检索到的关键词不仅准确率高而且排名尽可能居前，所以将 MAP 引入 F_β 的计算。相比于召回率，本文更侧重于提升准确率，故 F_β 指标中 β 取值为 0.5，其计算如式(12)所示。

$$\begin{aligned} R &= \frac{|M_{\text{ret}} \cap M_{\text{rel}}|}{|M_{\text{ret}}|},\\ F_\beta &= \frac{(1+\beta^2)\times \text{MAP}\times R}{\beta^2\times \text{MAP} + R} \end{aligned} \tag{12}$$

3.2 关联图模型参数分析

关联图的形式结构直接决定了关键词提取的效果，不同参数设置可能会对结果带来不同程度的影响。CLOSE 算法中的 minsup 设置过高则会丢弃大量有意义的词汇，设置过低则会包含众多冗余词汇，$\min^{+\text{lift}}$与 $\max^{-\text{lift}}$的变化会直接影响关联图的形式结构，同时 GSNA 算法在随机游走时衰减系数 d 的调节也会影响关键词的提取效果。

在预实验过程中发现，相比于节点的图结构属性，节点的关联性对关键词的提取效果影响更大，所以本文设置 minsup=0.8%，d=0.65。令降序排序后的关键词序列为 M_{v}，为了避免过多参数变化导致分析过程复杂，暂不考虑关键词的提取数量，在调节模型参数时，均抽取每次所得关键词序列 M_{v} 的前 $|M_{\text{rel}}|$ 个词项作为提取的关键词序列 M_{ret}，并绘制关联图模型的输入参数与 MAP 的变化曲线。

对于中文数据，图 3 和图 4 显示当 $\max^{-\text{lift}}$与 $\min^{+\text{lift}}$分别在[0.55,0.75]和[2,14]区间内变化时，GSNA 算法的平均准确率呈上升趋势，而 $\max^{-\text{lift}}$与 $\min^{+\text{lift}}$的轻微增加就会引起 MAP 的急速下降，这是因为数据中重要词汇彼此间的负关联度与正关联度更多地集中于[0.65,0.75]与[11,14]范围内。所

以，当 $\min^{+\text{lift}}=14$，$\max^{-\text{lift}}=0.7$ 时，在中文数据上构建的关联图模型能取得最佳提取效果。

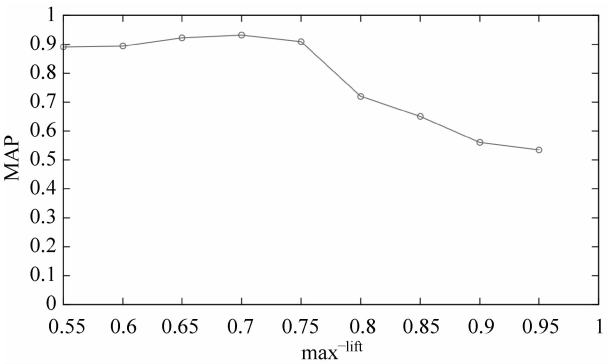

图 3　$\max^{-\text{lift}}$ 对关键词提取效果的影响[中文数据]

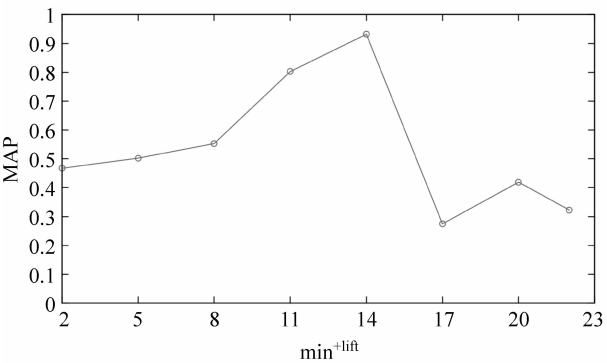

图 4　$\min^{+\text{lift}}$ 对关键词提取效果的影响[中文数据]

由于英文数据规模相对较小，本文设置 $\max^{-\text{lift}}$ 以 0.4 为基数，以步长为 0.1 的速度增长；$\min^{+\text{lift}}$ 初始值为 2，以步长为 2 的速度增长。图 5 和图 6 表明，当 $\min^{+\text{lift}}=6$，$\max^{-\text{lift}}=0.7$ 时，MAP 达到峰值 0.904 8。

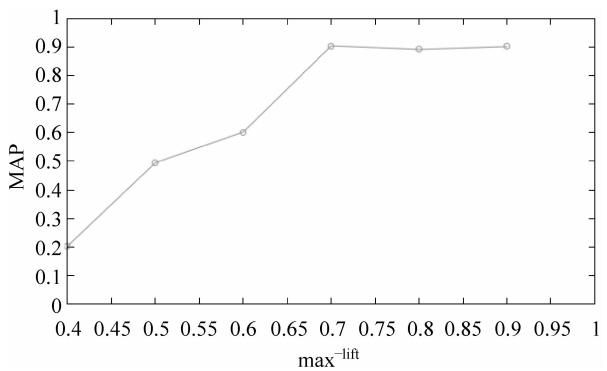

图 5　$\max^{-\text{lift}}$ 对关键词提取效果的影响[英文数据]

由于两种语料库的规模不同，所以效果最好时的 $\min^{+\text{lift}}$ 相差较大。$\min^{+\text{lift}}$ 的取值越大就会去除越多的关联词项，而小规模语料库的关键词候选量较少，为了更准确地提取关键词，所以较中文数据而言，英文数据的 $\min^{+\text{lift}}$ 取值较小。

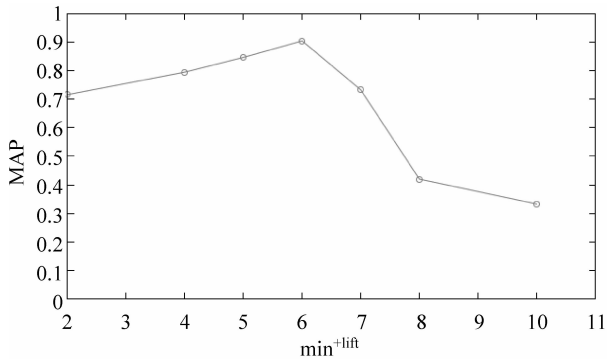

图 6　$\min^{+\text{lift}}$ 对关键词提取效果的影响[英文数据]

3.3　关键词提取规模

关键词提取注重挑选出与文本最相关的词汇，所以选取关键词的数量规模显得至关重要。本文引入关键词抽取规模参数 $\text{scale}=M_{\text{ret}}/M_{\text{v}}$，图 7 展现了 F_β 在两种数据上随着 scale 的增长而变化的曲线。

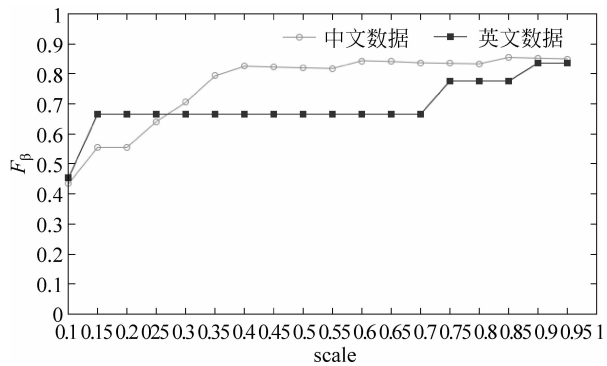

图 7　scale 对关键词提取效果 F_β 的影响

对于《物种起源》而言，在最佳关联图的情形下，经过语义聚类后得到 $M_{\text{v}}=23$ 个重要词汇。当 scale=85%时，F_β 取得最大值 0.854 8，实验表明可选取规模为 $M_{\text{ret}}=0.85\times M_{\text{v}}$ 的词序列作为该数据的关键词提取结果。对于 Li'paper，聚类后返回 $M_{\text{v}}=7$ 个重要词项，当 scale=80%时，F_β 取得最大值 0.810 2。

3.4　算法对比及分析

为了验证本文算法的优越性，将 TextRank[8]、GraphSum[9]、考虑图结构而不考虑节点关联性的关键词提取方法(Key Words Extraction Method by Nodes Association，NA)、考虑节点关联性而不考虑图结构的关键词提取方法(Key Words Extraction Method by Graph Structure，GS)与本文算法进行对比。

表6和表7列举出5种关键词提取算法分别作用于中文与英文数据上关于MAP、R、F_β指标上的对比结果。其中,最优指标值用粗字体标出,带有灰色背景的关键词表示未出现在Glossay的词项。TextRank在中文数据上的词项共现窗口长度为20,在英文数据上的词项共现窗口长度设为5,GraphSum、NA与GS算法均在GSNA所得的最佳关联图上迭代计算节点的重要性分数。

均未考虑节点的图结构属性与关联性的TextRank算法得到的关键词效果最差。值得注意的是,

表6 中文数据集上不同关键词提取算法的结果对比分析

词序	Glossary	GSNA	TextRank	GraphSum	NA	GS
1	物种	物种	物种	物种	物种	物种
2	变异	变异	类型	变异	变异	遗传
3	遗传	自然选择	动物	生活	栖息	习性
4	本能	生活 环境	变化	灭绝	遗传	变异
5	自然选择	遗传	差异	遗传	亲缘	灭绝
6	不育 杂交	地理 分布	植物	分化	性状	祖先
7	隔离	性状	生物	性状	自然选择	自然选择
8	灭绝	化石	变异	变种	杂种	本能
9	性状	不育 杂交	生活	祖先	本能	性状
10	地理 分布	生存 斗争	后代	亲缘	不育 杂交	亲缘
11	生存 斗争	地质	种	化石	化石	种群
12	祖先	种群	构造	差异	祖先	生物
13	后代	灭绝	性	后代	变种	杂交
14	化石	家养	自然选择	自然选择	灭绝	分布
15	种群	哺乳类 两栖类 陆栖类	变种	植物	后代	化石
MAP	—	**0.933 4**	0.749 0	0.861 4	0.852 5	0.900 3
R	—	**0.733 3**	0.266 6	0.6	**0.733 3**	0.666 6
F_β	—	**0.885 1**	0.550 1	0.792 4	0.825 7	0.841 3

表7 英文数据集上不同关键词提取算法的结果对比分析

词序	Glossary	GSNA	TextRank	GraphSum	NA	GS
1	term similarity	term similarity	term	semantic similarity	term similarity	term similarity
2	multi-word expression	semantic network	similarity	term similarity	semantic similarity	context similarity
3	clustering	concept-entity	clustering	pruning	muti-word expression	semantic network
4	semantic network	semantic similarity	concept	textual context	concept-entity	probase
5	probase	pruning	entity	context similarity	sense disambiguation	clustering
6	word sense disambiguation	multi-word expression	corpora	concept-entity	textual context	concept-entity
7	pruning algorithm	probase	probability	clustering	probability	pruning
MAP	—	**0.904 8**	0.653 1	0.622 4	0.693 9	0.903 7
R	—	**0.571 4**	0.142 9	0.285 7	0.142 9	**0.571 4**
F_β	—	**0.810 2**	0.381 1	0.503 7	0.391 7	0.809 6

分别只考虑节点关联性与图结构属性的 NA 与 GS 算法的 MAP 和 R 指标明显高于 GraphSum 和 Text-Rank，说明 NA 和 GS 能检索到更为重要的关键词。GSNA 算法从 F_β 上比 NA 和 GS 的效果更加可观，说明关键词提取过程中节点的图结构属性与关联性均不容忽视。相对于 GSNA 算法，GraphSum 缺少词项的语义聚类环节，导致提取结果存在语义冗余的词项，故其 MAP 和 R 指标均低于 GSNA。

值得关注的是，5 种方法在以《物种起源》为代表的大规模数据集上的性能均高于在以“Li'paper”为代表的小规模数据集上的性能。这是因为大规模的数据在构建关联图时能够形成更为复杂的图结构形式，同时更能充分地揭示大量节点间的关联关系与词项间的语义关系，所以文本规模越大，GSNA 的效果会越好。

3.5　算法效率分析

为了观测 GSNA 的运行效率，本节在 Celeron 1.40GHz 处理器的 Windows 操作系统下，比较了 TextRank、NA、GS、GSNA 与 GraphSum 这 5 种方法在不同词量规模下的执行时间，如图 8 所示。由于 GSNA 考虑了图结构、语义信息、节点间的关联性信息，所以在同等词量规模下，GSNA 的执行效率会比其他方法更低。但随着词量规模的扩大，GSNA的运行时间接近于线性增长，当词项规模不超过 40 000 时，GSNA 可在 10s 内完成关键词提取的过程。

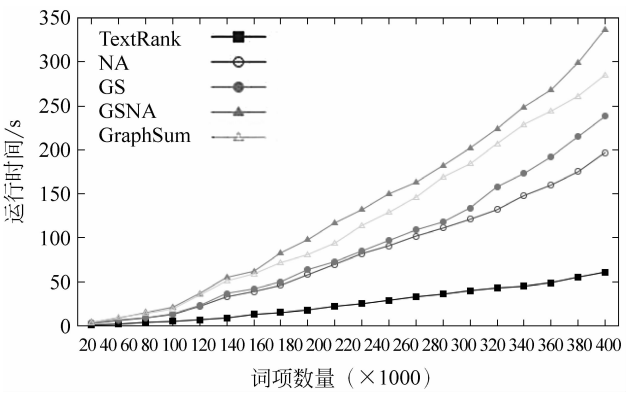

图 8　5 种方法在不同词量规模下的执行时间

4　总结

本文提出了一种融合图结构与节点关联的关键词提取方法，能够在脱离外部语料库的情况下检索到单篇文本的关键词。首先将挖掘到的强关联规则建模成关联图，然后利用 GSNA 在关联图上随机游走计算节点的重要性分数，最后对结果进行降序排序与语义聚类。实验选取中、英文验证数据，分别探索了关联图模型参数的影响、关键词的提取规模、5 种关键词提取算法的性能，综合分析，本文算法性能最佳。未来工作将致力于构建新型关联图模型，优化 GSNA 算法，提高整体执行效率。

参考文献

[1]　黄九鸣，吴泉源，张圣栋，等. 基于 AC-Trie 的在线社交网络文本流热点短语挖掘[J]. 电子学报，2016，44(10)：2466-2470.

[2]　Mehri A，Jamaati M，Mehri H. Word ranking in a single document by Jensen-Shannon divergence[J]. Physics Letters A，2015，379(28-29)：1627-1632.

[3]　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.

[4]　冯冲，廖纯，刘至润，等. 基于词汇语义和句法依存的情感关键句识别[J]. 电子学报，2016，44(10)：2471-2476.

[5]　Vorontsov K，Potapenko A. Additive regularization of topic models[J]. Machine Learning，2015，101(1-3)：303-323.

[6]　Md. Akmal Haidar，Mikko Kurimo. LDA-based context dependent recurrent neural network language model using document-based topic distribution of words[C]// Proceedings of ICASSP 2017. IEEE，2017：5730-5734.

[7]　Celik T. Spatial mutual information and PageRank-based contrast enhancement[J]. IEEE Transactions on Image Processing，2016，25(10)：4719-4728.

[8]　Li W，Zhao J. TextRank algorithm by exploiting Wikipedia for short text keywords extraction[C]//Proceedings of 2016 3rd International Conference on Information Science and Control Engineering. IEEE，2016：683-686.

[9]　Baralis E，Cagliero L，Mahoto N，et al. GraphSum：Discovering correlations among multiple terms for graph-based summarization[J]. Information Sciences，2013，249(16)：96-109.

[10]　Hsu C C，Lai Y A，Chen W H，et al. Unsupervised ranking using graph structures and node attributes [C]//Proceedings of the 10th ACM International Conference on Web Search and Data Mining. ACM，

2017：771-779.

[11] JiaweiHan，MichelineKamber，JianPei，等. 数据挖掘：概念与技术[M]. 北京：机械工业出版社，2012.

[12] Pasquier N，Bastide Y，Taouil R，et al. Efficient mining of association rules using closed itemset lattices [J]. Information Systems，1999，24(1)：25-46.

[13] 《物种起源》数据源 [EB/OL]. [2018-12-17]. http://vdisk. weibo. com/s/uheNHb1stTh6u.

[14] Li P，Wang H，Zhu K Q，et al. A large probabilistic semantic network based approach to compute term similarity[J]. IEEE Transactions on Knowledge & Data Engineering，2015，27(10)：2604-2617.

[15] Li K，Huang Z，Cheng Y C，et al. A maximal figure-of-merit learning approach to maximizing mean average precision with deep neural network based classifiers[C]//Proceedings of the 2014 IEEE International Conference on Acoustics，Speech and Signal Processing. IEEE，2014：4503-4507.

马慧芳(1981—)，博士，教授，主要研究领域为机器学习与文本挖掘。
E-mail：mahuifang@yeah. net

李苗(1997—)，硕士研究生，主要研究领域为自然语言处理。
E-mail：1607518663@qq. com

王双(1995—)，硕士研究生，主要研究领域为自然语言处理。
E-mail：1817349368@qq. com

(上接第 59 页)

刘倩(1993—)，通信作者，硕士研究生，主要研究领域为文档信息处理。
E-mail：lq199309@126. com

田英爱(1975—)，博士研究生，副教授，主要研究领域为文档信息处理。
E-mail：tianyingai@bistu. edu. cn

李宁(1964—)，博士，教授，主要研究领域为文档信息处理、信息技术标准化。
E-mail：ningli. ok@163. com