基于加权超图随机游走的文献关键词提取算法

马慧芳,刘 芳,夏 琴,郝占军
(西北师范大学计算机科学与工程学院,甘肃兰州 730070)

摘 要: 针对科技文献类标题短文本关键词提取时,已有自然语言处理算法难以建模文献时间与权威性且短文本词语较少建模往往存在高维稀疏问题,本文提出了一个综合实时性以及权威性的关键词提取算法为研究者进行相关推荐.该方法将文献标题视为超边,将标题中不同词项视为超点来构建超图,并对超图中的超边与超点同时加权,进而设计一种基于加权超图随机游走的关键词提取算法对文献标题的词项进行提取.该模型通过对文献来源,发表年份以及被引次数建模来对超边进行加权,根据节点之间的关联度以及每对节点在特定标题中的共现距离对超点加权.最后,通过超图上的随机游走计算出节点的重要性进而确立可推荐的关键词.实验表明,与三种基准短文本关键词提取算法相比,本文算法在精确率和召回率方面均有所提高.

关键词: 加权超图; 加权策略; 关键词推荐; 随机游走; 自然语言处理; 数据挖掘
中图分类号: TP393.092 **文献标识码:** A **文章编号:** 0372-2112 (2018)06-1410-05
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.06.020

Keywords Extraction Algorithm Based on Weighted Hypergraph Random Walk

MA Hui-fang, LIU Fang, XIA Qin, HAO Zhan-jun
(*College of Computer Science and Engineering, Northwest Normal University, Lanzhou, Gansu* 730070, *China*)

Abstract: It is difficult for the existing natural language processing algorithms to model the time and authority of short texts such as paper titles of scientific literature. Besides, the short texts always tend to have fewer words and thus suffer from high dimension and sparsity. A keyword extraction method involving both real-time and authoritativeness is presented. A weighted hyper-graph is constructed where vertexes represent weighted terms and weighted hyper-edges measure the semantic relatedness of both binary relations and nary relations among terms. On one hand, the source of the documents, the year of publication and number of citations are considered for weighting hyper-edges, on the other hand, the degree of association between the nodes and co-occurrence distance for each pair of nodes in particular title are calculated for weighting hyper-vertexes. The random walk approach is performed on the weighted hyper-graph to obtain the recommended keywords. Experimental results demonstrated that compared with three baseline algorithms, the proposed approach is able to extract keywords with higher precision and recall.

Key words: weighted hypergraph; weighting strategy; keywords extraction; random walk; natural language processing; data mining

1 引言

如何选择合理关键词进行检索是科技文献检索过程中经常遇到的问题.对文献而言,其标题本身是对文献的高度性总结,也是知识概念表达和交流的主要形式.本文所选定的研究对象为文献标题,已有的研究主要涉及以下两个研究领域:短文本关键词抽取和基于超图的应用.部分研究者把关键词抽取看作是分类问

收稿日期:2017-01-23;修回日期:2017-07-20;责任编辑:覃怀银
基金项目:国家自然科学基金(No. 61762078, No. 61363058, No. 61762079, No. 61762080);中科院智能信息处理重点实验室开放课题(No. IIP2014-4);广西可信软件重点实验室研究课题(No. kx201705)

题[1,2],该方法需要事先标注高质量训练数据,人工预处理代价较高. 无监督关键词抽取的主流方法可归纳为三种: 基于统计特征的关键词抽取、基于主题模型的关键词抽取和基于图的关键词抽取. 基于统计特征关键词抽取方法考虑词项统计信息[3],但忽略了重要的低频词和文档主题分布语义特征. Song 等[4]考虑了词与词之间的共现度和关联度等因素,对传统模型做了进一步改进. Hua 等[5]提出了共现距离概念,用来惩罚那些共同出现但间隔较远的词对. 基于主题模型的关键词抽取方法在近年来得到了重视,主题模型中基于潜在狄立克雷分配(Latent Dirichlet Allocation,LDA) 的关键词抽取方法应用最为广泛[6,7],其关键词抽取的效果与训练数据的主题分布关系密切. 基于图的关键词抽取方法所建立的图模型皆为普通图. 文献[8]中提出的方法应用最邻近耦合图构造图,结合向心率以及区域位置因子来衡量词语的重要性. 然而基于图模型所提出的关键词抽取算法仅仅考虑了词与词之间的二元共现关系,忽略了文档本身所携带的社会属性因素,这就需要建立一个能全面揭示文档-词,词-词之间的高阶关系的模型,如超图模型. Zhou 等[9]已提出基于超图的随机游走方法,基于 Zhou 等人提出的方法其他研究人员提出了基于超图的半监督关键词排序算法的定义[10].

本文提出一种基于加权超图随机游走的关键词提取算法,将文献标题视为超边,将标题中不同词项视为超点来构建了超图模型. 通过考虑文献自身所携带的社会信息来衡量文献自身的重要性,如文献来源,文献被引次数和文献发表时间. 同时定义词项之间的关联度,共现度,以及每对词项在特定标题中的共现距离等完成超图的超边和超点加权. 最后将随机游走的方法在超图上进行推广.

2　超边与超点的加权策略

具体地,将某标题 d_i 视为一个由不同的关键词 $d_i=\{v_1,v_2,\cdots,v_s\}$ 所组成的词袋模型,而这些标题的集合 $D=\{d_1,d_2,\cdots,d_m\}$ 即为本文所定义的词汇超图,且标题中所有关键词集合为 $T=d_1\cup d_2\cup\cdots\cup d_m=\{v_1,v_2,\cdots,v_n\}$.

2.1　加权超图模型构建

设 HG(V,E) 表示普通超图,其中 V 为超点集合,E 为超边集合. 超边 e 实质上是超点集合的子集且 $\cup_{e\in E}e=V$. 当 $v\in e$ 时,称超边 e 指向顶点 v. 一个普通超图可以用指示矩阵 $\boldsymbol{H}$ 来表示,若 $v\in e$,$\boldsymbol{H}$ 中的元素 $h(v,e)=1$,否则 $h(v,e)=0$.

设 WHG(V,E,$w(e)$,$w(v_e)$) 为加权超图,其中 $w(e):e\rightarrow R^+$ 代表超边 e 的权重,$w(v_e):v_e\rightarrow R^+$ 代表超点 v 在特定超边 e 上的权重. 带权重的超图指示矩阵 $\boldsymbol{H}_w$ 中的元素定义如式(1) 所示.

$$h_w(v,e)=\begin{cases}w(v_e), & v\in e\\ 0, & v\notin e\end{cases} \tag{1}$$

在加权超图中超点的度 $d(v)$ 与超边的度 $d(e)$ 定义分别如式(2) 和式(3) 所示.

$$d(v)=\sum_{e\in E}w(e)\,h(v,e) \tag{2}$$

$$d(e)=\sum_{v\in V}w(v_e)\,h(v,e) \tag{3}$$

与普通超图相比,加权超图中超边和超点皆有权重,本节详细介绍对超边和超点加权的具体策略,文中提及到的 d_i 与 e 同义,皆代表某条特定的超边.

2.1.1　超边加权策略

对科学研究而言,所选定研究对象的权威性,实时性以及它对该领域的贡献率是极为重要. 本文根据文献的来源来确定其权威性,即 $R_{\text{paper-rank}}(d_i)$. 依据中国计算机学会(http://www.ccf.org.cn) 给出的文献重要性分类信息,分别选取 A、B、C 三类的文献为实验数据,其对应的 $R_{\text{paper-rank}}(d_i)$ 值如式(4) 所示.

$$R_{\text{paper-rank}(d_l)}=\begin{cases}1, & d_l\in A\\ 2/3, & d_l\in B\\ 1/3, & d_l\in C\end{cases} \tag{4}$$

构建函数 $R_{\text{time-quote}}$来表征文献的实时性与被引次数. 近期所发表的文献,实时性越强,其值就越大; 文献的被引次数越多,该文献对相关领域贡献越大,其值相应也越大. 具体函数如下:

$$R_{\text{time-quote}}(d_i)=\mathrm{e}^{-\frac{(c-y_i)+1}{k+1}} \tag{5}$$

c 和 y_i 分别代表当前时间与该文献的出版时间,以年为单位,k 为被引次数.

最终计算超边 d_i 权重如下:

$$w(d_i)=\lambda R_{\text{paper-rank}}(d_i)+(1-\lambda)R_{\text{time-quote}}(d_i) \tag{6}$$

λ 为值域在[0,1]之间的平滑因子,其值越大,代表更注重文献来源; 相反,其值越小,更注重文献实时性与被引次数. 根据 λ 的不同取值来调节超边加权,进而选定最优 λ 值. 实验中发现随着 λ 取值增大算法性能先提升后下降,且当 $\lambda=0.7$ 时推荐精确率达到峰值,因而设定 λ 值为 0.7.

2.1.2　超点加权策略

通过超点之间的共现度,关联度以及在特定超边共现距离对超点在特定超边中加权. 对于超边 d_l,给定超点 v_i 与 v_j,v_i,$v_j\in d_l$,它们在该超边的共现度 $\text{co\_}d_l(v_i,v_j)$ 如式(7) 所示.

$$\text{co\_}d_l(v_i,v_j)=n(d_l)\times\mathrm{e}^{-\text{dist}_{d_l}(v_i,v_j)} \tag{7}$$

其中共现距离 $\text{dist}_{d_l}(v_i,v_j)$ 即超点 v_i 与 v_j 在 d_l 中间隔的单词个数.

任意两个超点 v_i 与 v_j 之间的共现度即为所有超边 $co\_d_l(v_i,v_j)$ 的和值,定义如下:

$$co(v_i,v_j) = \sum_{l=1}^{m} co\_d_l(v_i,v_j) \tag{8}$$

v_i 与 v_j 的单边关联度定义如下:

$$ucor_i(v_i,v_j) = \frac{co(v_i,v_j)}{\sum_{q=1}^{n} co(v_i,v_q)} \times \log_2 \frac{n}{N_{nei}(v_j)} \tag{9}$$

式(9)右侧前者体现了观测到 v_i 联想到 v_j 的概率,后者惩罚那些和很多词都共现过的 v_j,$N_{nei}(v_j)$ 代表与 v_j 共现过的所有词的个数.

同理可求得 v_j 单边关联度并定义 v_i 与 v_j 的关联度为 v_i 与 v_j 的单边关联度的均值,如式(10)所示.

$$cor(v_i,v_j) = \frac{ucor_i(v_i,v_j) + ucor_j(v_j,v_i)}{2} \tag{10}$$

定义某特定标题所求的某个词的关联权重 $cow(v_i,d_l)$ 反映在特定超边 d_l 中超点 v_i 的主题指示性,v_i 的关联权重越高意味着当 v_i 出现时,超边 d_l 中其他顶点 v_j 随之出现的概率也就越高. 具体地,定义超点初始权重 $iw(v_i,d_l)$,结合关联度与初始权重,进一步求出某个词的关联权重如式(11)所示.

$$cow(v_i,d_l) = iw(v_i,d_l) + \frac{\sum_{j=1}^{|d_l|} iw(v_j,d_l) \times cor(v_i,v_j)}{|d_l|} \tag{11}$$

将关联权重与加权体系中的全局统计权重相结合,最终超点 v_i 在超边 d_l 中加权计算公式如式(12)所示.

$$w(v_i,d_l) = cow(v_i,d_l) \times idf(v_i) = cow(v_i,d_l) \times \log_2 \frac{m}{df(v_i)} \tag{12}$$

3 加权超图随机游走

超图的一条超边所包含的超点个数往往不止两个($\delta(e)>2$),对于超图而言需要一种更为普遍的随机游走方法. Bellaachia 等人将随机游走的方法在超图上进行了推广[14]. 其随机游走过程如下:选定起始顶点 u,依超边权重 $w(e)$ 大小概率成正比的选择一条包含当前顶点 u 的特定超边 e;然后,在已经选中的超边中,根据顶点权重大小概率成正比的选择转移顶点 v. 设 $\boldsymbol{P}$ 为随机游走的转移概率矩阵,其中元素计算方法如式(13)所示.

$$P(u,v) = \sum_{e\in E} w(e) \frac{h(u,e)}{\sum_{\hat{e}\in\varepsilon(u)} w(\hat{e})} \frac{h_w(v,e)}{\sum_{\hat{v}\in e} h_w(\hat{v},e)} \tag{13}$$

矩阵 $\boldsymbol{P}$ 的计算方法如下:

$$\boldsymbol{P} = \boldsymbol{D}_v^{-1}\boldsymbol{H}\boldsymbol{W}_e\boldsymbol{D}_e^{-1}\boldsymbol{H}_w^{\mathrm{T}}$$

其中,$h_w(v,e)$ 是 v 在超边 e 中的权重,$\boldsymbol{D}_v$ 是超点度的对角线矩阵;$\boldsymbol{H}$ 是普通超图指示矩阵;$\boldsymbol{W}_e$ 为超边权重的对角线矩阵;$\boldsymbol{D}_e$ 是超边度的对角线矩阵;$\boldsymbol{H}_w$ 是加权超图的指示矩阵.

随机游走过程刚开始时,初始分布向量 $\boldsymbol{v}^0 \in R^{|V|\times 1}$ 等概率. 随机游走过程在经过若干步后,若已将所有节点遍历,则概率分布向量 $\boldsymbol{v}$ 不再发生变化. 使用类似 PageRank 算法来实现随机游走过程,该算法加入了心灵转移的思想,依经验[11]参数 α 设置为 0.85.

$$\boldsymbol{v}^{i+1} = \alpha\boldsymbol{P}^{\mathrm{T}}\boldsymbol{v}^i + (1-\alpha)\boldsymbol{e}/n \tag{14}$$

当随机游走的迭代过程停止,即向量 $\boldsymbol{v}$ 不再发生变化时,对向量 $\boldsymbol{v}$ 中各顶点的权重依照由大到小的顺序排序. 最后,只需选取 $\boldsymbol{v}$ 中对应的前 Top-K 个词项作为所选取出的关键词集.

4 实验性能与分析

为了验证本文算法的有效性,设计实验进行验证. 首先对实验数据进行描述,其次提出相应实验评价指标并对本文的方法进行验证,最后对实验结果做出进一步分析.

4.1 数据描述

从 CCF 推荐排名的 *A*、*B*、*C* 三类文献中,选取 10 个领域的英文数据进行验证,每个类别包含 1000 条文章标题作为实验数据. 中文数据集以 CSCD 数据库中核心期刊为数据来源,在这些期刊中抓取文章标题作为实验数据,每个期刊选取 500 篇文章标题,总计 5000 篇文章标题作为实验数据集. 首先对实验数据进行预处理操作,即去除停用词,大小写转换等,最终得到每个文献的标题,并标识其发表年份、期刊等级与被引次数.

4.2 实验评价指标

考虑到对某特定领域没有完全“正确”的推荐关键词词集,邀请三位专业研究人员从现有语料库中针对特定主题分别选择不定个数关键词,并对平均推荐结果进行相关分析. 采用外部评价指标精确率 Precision(简记为 Pr),召回率 Recall(简记为 Re)以及 F-measure 值对该算法进行评估:

$$Pr = \frac{|\{\text{相关关键词词集}\} \cap \{\text{返回的关键词词集}\}|}{|\{\text{返回的关键词词集}\}|} \tag{15}$$

$$Re = \frac{|\{\text{相关关键词词集}\} \cap \{\text{返回的关键词词集}\}|}{|\{\text{相关关键词词集}\}|} \tag{16}$$

$$F\text{-measure} = \frac{2\times Pr\times Re}{Pr+Re} \tag{17}$$

Pr 所涉及到的{相关关键词词集}只需在三位研究人员给出的关键词列表中至少出现一次;而计算 Re 时的{相关关键词词集}须在三位专业研究人员给出的关键

词列表中同时出现;{返回的关键词词集}表示每次返回的关键词词集,本文将该集合大小设为 10.

4.3 实验结果与相关分析

选取 TF × IDF 方法,LDA 方法[6],基于普通图的方法 $TKG_2 \mid W^{1/F} \mid C^E$[8],超点加权不考虑共现距离方法(COW-dist) × IDF 作为参照,分别观测实验结果. 以上四种无监督关键词提取算法均是主流的无监督关键词提取算法. 其中,LDA 关键词抽取方法将主题个数设为 10,超参数 $\alpha = 0.1, \beta = 0.01$ 通过 Gibbs Sampling 算法学习得出;基于普通图的抽取方法中 TKG_2 代表边的构建方法,$W^{1/F}$ 代表边的权重,C^E 代表顶点离心率. 因篇幅限制,分别对网络与信息安全与 Artificial Intelligence 的中英文实验结果进行分析. 表 1 和表 2 给出了不同算法抽取得到的关键词. 其中加粗的关键词代表至少在三位研究人员选取的关键词的并集中出现一次,灰色背景的关键词则属于三位研究人员选取的关键词交集. 前者加粗的关键词用来计算 Pr,后者带有灰色背景的关键词用来计算 Re.

表 1　中文数据集上不同算法抽取出的关键词

	TF × IDF	LDA	$TKG_2 \mid W^{1/F} \mid C^E$	(COW-dist) × IDF	COW × IDF
1	加密	安全	检测	检测	检测
2	解密	加密	恢复	加密	加密
3	算法	解密	漏洞	解密	云
4	安全性	网络	加密	云	量子
5	防御	模型	解密	保护	入侵
6	改进	漏洞	漏洞	信任关系	动态
7	密钥	协议	动态	动态	防御
8	问题	算法	公钥	防御	传感器
9	优化	身份	云	漏洞	分布式
10	分析	伪造	故障	预测	优化
Pr	40%	30%	60%	60%	70%
Re	44.44%	22.22%	44.44%	55.56%	77.78%
F-measure	41.90%	25.29%	51.06%	57.93%	73.78%

表 2　英文数据集上不同算法抽取出的关键词

	TF × IDF	LDA	$TKG_2 \mid W^{1/F} \mid C^E$	(COW-dist) × IDF	COW × IDF
1	algorithm	classification	recommendation	pattern recognition	pattern recognition
2	key word	clustering	optimization	solve	recommendation
3	clustering	data	model	theme	solve
4	model	network	classification	recommendation	image
5	classification	subject	clustering	local	data mining
6	sign	feature selection	vector	feature	network
7	improvement	recommendation	algorithm	sparse	natural language
8	utilize	language	deep learning	optimization	clustering
9	recommendation	distinguish	space	modeling	information
10	data	learn	detect	natural language	deep learning
Pr	40%	30%	50%	60%	70%
Re	50%	50%	66.67%	50%	83%
F-measure	44.44%	37.5%	57.14%	54.55%	75.95%

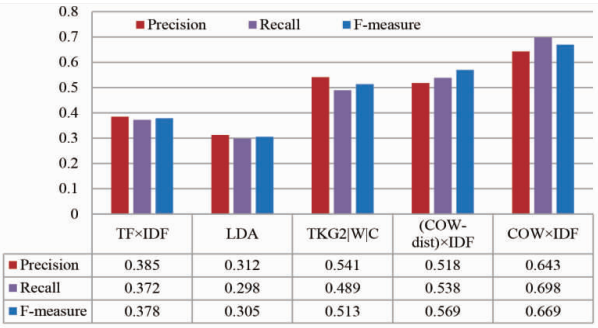

图1　中文数据集上不同算法性能对比

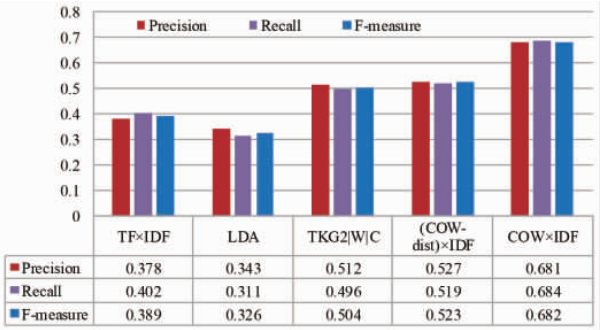

图2　英文数据集上不同算法性能对比

图 1 与图 2 为不同算法在整个中英文实验数据集上所求得的均值. 因传统的 TF × IDF 加权方法较适应于长文本,本文所选取的实验数据为文献标题,故在特定的标题中各词项出现次数趋于相同,该方法的效率因此降低. LDA 方法尽管在长文本上效果不错,但本文数据太短不利于 LDA 模型训练,故该方法效率不佳. 基于普通图的方法虽已将文本以图的形式表示,但图中的边仍是简单依据点与点在特定窗口范围内的共现关系来确定,忽略了文档本身所携带的社会属性因素. 本文算法通过全面揭示了文档-词,词-词之间的高阶关系,并综合考虑了词项共现距离,共现度,关联权重以及全局统计权重,从图中可以看出: 本文方法相比上述四种关键词提取方法在精确率、召回率以及 F-measure 值这三方面都更优异.

5 结束语

本文提出了一个综合实时性以及权威性的关键词提取算法为研究者进行相关推荐. 将文献标题视为超边,标题中不同词项视为超点来构建超图,进而设计了一种基于加权超图随机游走的关键词提取算法. 该模型在超边加权过程中考虑了文献自身所携带的社会信息来衡量文献自身的重要性,超点加权阶段利用节点之间的关联度,共现度以及每对节点在特定标题中的共现距离来实现对超点的加权. 最后,通过超图上的随机游走计算出节点的重要性进而确立可推荐的关键词. 实验部分评估并对比了本文方法的有效性. 在今后的研究中,未来工作考虑进一步将本文的方法用于更大的数据集,进一步优化本方法的计算效率,缩短运行时间.

参考文献

[1] Ma Hui-fang, Xing Yu-ying, et al. Leveraging term co-occurrence distance and strong classification features for short text feature selection [A]. Proceedings of the 10th International Conference on Knowledge Science, Engineering and Management [C]. Melbourne, Australia: Springer, 2017. 303 - 310.

[2] 刘喜平,万常选等. 空间关键词搜索研究综述 [J]. 软件学报,2016,27(2):329 - 347.
Liu Xi-ping, Wan Chang-xuan, et al. Survey on spatial keyword search [J]. Journal of Software, 2016, 27(2): 329 - 347. (in Chinese)

[3] Ugo Erra, Sabrina Senatore, et al. Approximate TF-IDF based on topic extraction from massive message stream using the GPU [J]. Information Sciences, 2015, 292(20): 143 - 161.

[4] Song Shao-xu, Zhu Han, et al, Probabilistic correlation-based similarity measure on text records [J]. Information Sciences, 2014, 289(1): 8 - 24.

[5] Hua Wen, Wang Zhong-yuan, et al . Short text understanding through lexical-semantic analysis [A]. Proceedings of the 31st International Conference on Data Engineering [C]. Seoul, South Korea: IEEE, 2015. 495 - 506.

[6] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(1): 993 - 1022.

[7] Rahim Saeidi, Ramon Fernandez Astudillo, et al. Uncertain LDA: Including observation uncertainties in discriminative transforms [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(7): 1479 - 1488.

[8] Willyan D Abilhoa, Leandro N de Castro. A keyword extraction method from twitter messages represented as graphs [J]. Applied Mathematics and Computation, 2014, 240(4): 308 - 325.

[9] Zhou Deng-yong, Huang Jia-yuan, et al. Learning with hypergraphs: clustering, classification, and embedding [A]. Proceedings of the 20th International Conference on Neural Information Processing Systems [C]. Vancouver, Canada: MIT Press, 2006. 1601 - 1608.

[10] Li De-cong, Li Su-jian. Hypergraph-based inductive learning for generating implicit key phrases [A]. Proceedings of the 20th International Conference on World Wide Web [C]. Hyderabad, India: Springer, 2011. 77 - 78.

[11] Bellaachia A, Al-Dhelaan M. HG-Rank: A hypergraph-based keyphrase extraction for short documents in dynamic genre [A]. Proceedings of the 4th Workshop on Making Sense of Microposts [C]. Seoul, Korea: CEUR-WS, 2014. 42 - 49.

作者简介

马慧芳 女,1981 年 7 月出生,甘肃兰州人. 博士,硕士生导师,现为西北师范大学计算机科学与工程学院副教授. 研究领域为人工智能,数据挖掘与机器学习.
E-mail: mahuifang@yeah.net

刘 芳 女,1996 年 12 月出生,甘肃平凉人. 现为西北师范大学计算机科学与工程学院本科生. 研究方向为文本挖掘.
E-mail: 1554216090@qq.com