



融合 BERT 语义加权与网络图的关键词抽取方法

李 俊, 吕学强

(北京信息科技大学 网络文化与数字传播北京市重点实验室, 北京 100101)

摘 要: 结合文档本身的结构信息与外部词语的语义信息, 提出一种融合 BERT 词向量与 TextRank 的关键词抽取方法。在基于网络图的 TextRank 方法基础上, 引入语义差异性并利用 BERT 词向量加权方式优化 TextRank 转移概率矩阵计算过程, 同时通过迭代运算对文档中的词语进行综合影响力得分排序, 最终提取得分最高的 TopN 个词语作为关键词。实验结果表明, 当选取 Top3、Top5、Top7 和 Top10 个关键词时, 与基于词向量聚类质心与 TextRank 加权的关键词抽取方法相比, 该方法的平均 F 值提升了 2.5%, 关键词抽取效率更高。

关键词: 关键词抽取; 语义关系; 词向量; TextRank 方法; 基于 Transformer 的双向编码器表示

开放科学(资源服务)标志码(OSID):



中文引用格式: 李俊, 吕学强. 融合 BERT 语义加权与网络图的关键词抽取方法[J]. 计算机工程, 2020, 46(9): 89-94.

英文引用格式: LI Jun, LÜ Xueqiang. Keyword extraction method based on BERT semantic weighting and network graph[J]. Computer Engineering, 2020, 46(9): 89-94.

Keyword Extraction Method Based on BERT Semantic Weighting and Network Graph

LI Jun, LÜ Xueqiang

(Beijing Key Laboratory of Internet Culture and Digital Dissemination Research,
Beijing Information Science and Technology University, Beijing 100101, China)

[Abstract] Based on the structural information of the document and the semantic information of external words, this paper proposes a keyword extraction method based on Bidirectional Encoder Representation from Transformer (BERT) word vectors and TextRank. Using network graph-based TextRank, this method introduces the semantic difference and uses BERT word vector weighting to optimize the calculation process of the transfer possibility matrix of TextRank. At the same time, the overall influence scores of words in the document are sorted by iteration, and the words with the TopN scores are selected as keywords. Experimental results show that when keywords are selected Top3, Top5, Top7 and Top10 words, the average F value of the proposed method is 2.5% higher than that of the keyword extraction method based on word vector clustering centroid and TextRank weighting. The proposed method can improve the efficiency of keyword extraction.

[Key words] keyword extraction; semantic relation; word vector; TextRank method; Bidirectional Encoder Representation from Transformer (BERT)

DOI: 10.19678/j.issn.1000-3428.0055368

0 概述

随着计算机信息技术和网络技术的快速发展, 各行各业每天产生并积累大量数据, 从海量数据中提取对人们有价值的信息已成为急需解决的问题。关键词抽取是在对象文本中自动抽取能够体现文本内容的中心概念或者重要词语, 可帮助人们快速定位所需文档, 因此其在自然语言处理、图书馆学和情

报学等领域得到广泛应用^[1]。

目前, 关键词抽取方法主要分为有监督和无监督两种。有监督的关键词抽取方法通过二分类思想确定文档中的候选词是否为关键词。该方法将已标注的关键词数据作为训练语料库, 通过语料库训练关键词判别模型, 并利用该模型对待处理文本进行关键词提取, 但是该方式需要人工提前标注大量语料, 并且若在标注过程中存在误差, 则会直接影响模

基金项目: 国家自然科学基金(61671070); 国家语委重点科研项目(ZD1135-53)。

作者简介: 李 俊(1994—), 男, 硕士研究生, 主研方向为自然语言处理; 吕学强, 教授、博士。

收稿日期: 2019-07-03 修回日期: 2019-09-18 E-mail: lijun60@qq.com

型性能。无监督的关键词抽取方法无须事先标注训练语料,通过关键词重要性排序实现关键词抽取。该方法利用关键词权重等量化指标进行权重计算与排序,选出综合影响得分较高的若干词作为关键词。无监督的关键词抽取方法近年来受到学者们的广泛关注,其中的 TextRank 方法^[2]在构建网络图时主要利用文档本身的结构信息,但缺少外部语义知识的支持,而基于 Transformer 的双向编码器表示(Bidirectional Encoder Representation from Transformer, BERT)语言模型能将词语映射成高维的向量,并保留其语义上的相似关系。

本文将文档信息与 BERT 词语语义信息同时融入基于网络图的关键词抽取模型中,通过词向量进行语义表示并利用 BERT 词向量加权方式计算 TextRank 中词节点的概率转移矩阵,以提升关键词抽取效果。

1 相关研究

文献[3]利用词图中的度中心性、接近中心性等中心性指标,加权计算邻接词语所传递的影响力概率转移矩阵,提升关键词抽取效果。文献[4]通过词语覆盖范围权重、位置权重和频度权重以及 TextRank 实现关键词自动提取。文献[5]提出针对中文文档的关键词抽取算法 TextRank-CM。文献[6]将 TextRank 与隐含狄利克雷分布(Latent Dirichlet Allocation, LDA)^[7]主题模型相结合,通过融合文档整体的主题信息及单篇文档的结构信息来提高关键词抽取效果。文献[8]通过训练隐马尔可夫模型挖掘主题信息和关键词,并在测试语料上取得了较好的挖掘效果。文献[9]通过构建语义网络图并融合图路径转化量、词聚类系数以及词窗口信息提出综合计算指标,且在专利数据集中具有较好的性能表现。文献[10]利用各种图节点缩减算法重构语言网络图,筛选出文档中影响得分高的节点并将其作为关键词。文献[11]提出在 TextRank 转移概率计算过程中融合词图的边和点信息来提升关键词抽取效果。

近年来,随着 Word2Vec^[12]、GloVe^[13]等词向量模型和语言模型的发展,学者们开始利用词向量模型训练文本库生成词向量获得词汇语义关系,并结合传统 TextRank 方法进行关键词提取。文献[14]将词向量融入候选词中以增强关键词抽取的语义关系。文献[15]通过词语的语义距离计算实现词语的主题聚类,并依据聚类结果选取中心词为关键词。文献[16]利用文本局部结构信息和文本整体的词向量语义信息抽取关键词。文献[17]通过词向量计算词语的相似性,再根据词聚类算法实现关键词抽取。文献[18]利用 Word2Vec 词向量实现相似词语的聚类,通过计算距离质心最远的词来更新概率转移矩

阵,并将其引入到 TextRank 词图的迭代计算过程中优化关键词抽取效率。

综上所述,在 TextRank 权值分配计算中如何融合外部语义信息是 TextRank 方法优化的关键。词位置分布加权及 LDA 主题模型加权等方法均需要对待提取文本进行预处理,但对于不同数据集效果差异较大,而词向量训练与待抽取关键词的文档无关,若利用包含外部语义信息的词向量对 TextRank 方法进行优化,则可以更好地解决关键词抽取问题。因此,本文采用基于网络图的关键词抽取方法,将词向量计算的语义信息和文本信息融入 TextRank 计算过程中,先利用 BERT 模型^[19]获取词向量,再使用词向量加权方式优化 TextRank 中词节点的转移概率矩阵计算,提升关键词抽取效果。

2 融合 BERT 与 TextRank 的关键词抽取

在单篇文章中通常具有多个关键词,而这些关键词一般不属于同一个主题,一些学者通过 LDA 主题聚类进行关键词抽取^[6,20],因此结合理论分析和实际应用可知,不同的主题表明这些关键词在语义角度存在明显差异。传统关键词抽取方法通过挖掘词语的共现关系构建词的图模型,并对文档中词语进行综合影响力得分排序实现关键词抽取,从而选择相对重要的词语。该方法很容易将高频率的词语当作关键词,由于多数情况下一篇文档中的某些关键词的词频很低,因此此类关键词容易被遗漏。为此,本文在 TextRank 方法的基础上,引入关键词的语义差异性优化词节点间的概率转移矩阵计算,并经过迭代计算获取词语在文本中的重要程度,从而完成关键词的综合影响力排序及抽取。

2.1 候选关键词的词图构建

基于 TextRank 思想将一篇文档转换成词图模型,先把所有已出现的词语去重并作为单独的节点,通过词语的共现窗口决定各个词节点之间的边并构成词图。单篇文档的词图构建过程如下:

1) 对文档 D 进行分句,则 D 由 n 个句子组成,即 $D = [s_1, s_2, \dots, s_n]$ 。

2) 对 $s_i \in D$ 进行分词、去停用词和保留重要词性等预处理,生成候选关键词序列 $s_i = [w_1, w_2, \dots, w_n]$ 。

3) 对关键词序列进行词图构建 $G = (V, E)$,其中: V 为候选的关键词节点集合, $V = \{v_1, v_2, \dots, v_n\}$; E 为候选关键词之间的链接集合, E 中的边由词的共现关系决定,例如 w_i, w_j 在词窗口内共现时会在词图中新增两条有向链接边,即 $v_i \rightarrow v_j$ 和 $v_j \rightarrow v_i$ 。在生成词图后,可利用式(1)计算节点分数:

$$S(v_i) = (1 - d) + d \times \sum_{v_j \in \text{In}(v_i)} \frac{W_{ji}}{\sum_{v_k \in \text{Out}(v_j)} W_{jk}} S(v_j) \quad (1)$$

其中: $\text{In}(v_i)$ 是其他节点到词节点 v_i 的节点集合; $\text{Out}(v_j)$ 是词节点 v_j 所指向的集合; w_{ji}, w_{jk} 是两词节点所形成边的权值; $S(v_i)$ 是节点 v_i 的得分权重; d 是平滑因子, 其实际意义是词语转移到其他词语的概率, 并且可以保证式(1)在迭代计算时能够稳定传递并达到收敛, 通常设置为 0.85。

利用迭代计算式(1)完成对候选关键词的重要性排序, 该过程是一个马尔可夫过程, 因此最终结果与词节点的最初权值及边的权值无关, 仅与文档中词节点的跳转矩阵相关。传统 TextRank 方法使用相同的跳转概率表示相邻节点之间的比重。令 $P(v_i, v_j)$ 代表词节点 v_i 到词节点 v_j 的跳转概率, 利用式(2)计算词节点的转移概率^[2]:

$$P(v_i, v_j) = \begin{cases} \frac{1}{\deg(v_i)}, & (v_i \rightarrow v_j) \in E \\ 0, & \text{其他} \end{cases} \quad (2)$$

其中, $\deg(v_i)$ 代表词节点 v_i 的度。本文提出一种融合 BERT 向量的语义信息计算方法, 优化 TextRank 中的权值计算过程。

2.2 BERT 词向量的语义加权

词向量是使用向量的形式来表达词语, 此类方法中应用较广泛的为 Word2Vec 模型^[12] 和 BERT 模型^[19]。Word2Vec 模型利用浅层神经网络进行模型学习, 将词语映射到相应的高维空间中得到词向量。BERT 模型本质是一种可微调的双向 Transformer^[21] 编码器, 其摒弃了循环神经网络 (Recurrent Neural Network, RNN) 结构, 将 Transformer 编码器作为模型的主体结构, 主要利用注意力机制对句子进行建模。现有的 Word2Vec、GloVe 等词向量模型均不能较好地处理一词多义的情况, 而 BERT 语言模型不仅可以生成词向量, 而且可以解决一词多义问题。BERT 模型结构如图 1 所示, 其中, E 表示训练向量, Trm 表示 Transformer 编码器。

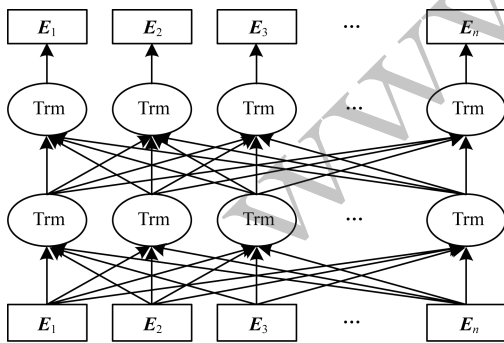


图1 BERT 模型结构

Fig.1 Structure of BERT model

BERT 将两个句子序列相连并作为模型输入部分, 在每个句子的开始和结束位置打上标记符号。对于每个单词, BERT 分别进行单词位置信息编码、

单词 Word2Vec 编码和句子整体编码 3 种嵌入操作。将这 3 种嵌入结果向量进行拼接获得 BERT 词向量。相对已有语言模型, BERT 是百层左右的深度神经网络模型, 其利用大规模语料进行模型参数学习, 因此在 BERT 词向量中融入了更多的语法、词法以及语义信息, 同时 BERT 以字为单位进行训练, 在一定程度上解决了 Word2Vec 面临的未登录词问题。

本文采用 BERT 向量计算得到外部语义关系并将其融入文档关键词提取中。根据统计发现文档中的多个关键词不一定具有很强的关联性, 一篇文档的关键词通常代表不同的文档主题, 用于概括文档中心内容, 例如利用 LDA 聚类的关键词抽取方法就是针对一篇文章的多个主题提取关键词并取得了较好的效果。因此, 考虑到关键词所属不同主题导致的语义差异性, 本文假设在 TextRank 词语的权值分配计算中, 若相邻词节点集中两词节点的语义差异越大, 则赋予更高的转移概率且具有更高的跳转权重。本文选用余弦距离表征词语的语义距离, 由式(3)计算得到:

$$\text{sim}(a_i, a_j) = \frac{a_i \cdot a_j}{\|a_i\| \cdot \|a_j\|} \quad (3)$$

其中, a_i, a_j 表示候选关键词词节点 v_i, v_j 的词向量。由于语义差异越大, 转移概率越高, 因此使用式(4)计算节点 v_i 到节点 v_j 的跳转概率:

$$P_{\text{sim}}(v_i, v_j) = k - \text{sim}(a_i, a_j) \quad (4)$$

其中, k 为实验参数, 实验中需对 $\text{sim}(a_i, a_j)$ 进行归一化处理, 使得 $\text{sim}(a_i, a_j) \in (0, 1)$, 因此令 $k = 1$ 。

2.3 转移概率矩阵的计算

根据马尔可夫过程可知, 节点的重要性得分与候选关键词图的转移矩阵有关。在 TextRank 节点影响力得分计算中, 某个节点对其相邻节点的权重计算主要分为覆盖范围、位置和频度权重三部分^[4], 令 W 表示词节点的综合影响力权重, α, β, γ 分别表示这三部分权重所占的比重, 计算公式如式(5)所示:

$$W = \alpha + \beta + \gamma = 1 \quad (5)$$

在本文实验中的参数设置参考文献[4], 令 $\alpha = 0.33, \beta = 0.34, \gamma = 0.33$ 。

借鉴传统 TextRank 方法, 通过式(2)计算得到覆盖范围影响力 P_{range} , 而节点位置影响力 P_{loc} 由式(6)计算得到^[4]:

$$P_{\text{loc}}(v_i, v_j) = \frac{I(v_j)}{\sum_{v_k \in \text{Out}(v_i)} I(v_k)} \quad (6)$$

其中, $I(v_j)$ 表示词语 v_j 在文档中的位置重要性权重, 根据文献[4]可知, 如果 v_j 在标题中出现时, 则 $I(v_j) = 30$, 否则 $I(v_j) = 1$ 。由于本文实验语料为新

闻文体,若考虑新闻中导语位置的重要性,则实验效果将得到显著改善,因此新增权重条件,若 v_j 出现在导语中时,则令 $I(v_j) = 10$ 。

根据上文词向量语义加权影响力的定义,将一个节点对相邻节点的权重计算优化为词覆盖范围、词位置和词语义加权影响力三部分。因此,利用式(7)计算得到词节点综合跳转概率:

$$P(v_i, v_j) = \alpha \cdot P_{\text{range}} + \beta \cdot P_{\text{loc}}(v_i, v_j) + \gamma \cdot P_{\text{sim}}(v_i, v_j) \quad (7)$$

改进权重转移矩阵 M 的计算公式为:

$$M = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix} \quad (8)$$

假设矩阵 M 中的 j 值代表第 j 个词节点 v_j 跳转到其他词节点时的比重,例如 p_{ij} 表示 v_j 跳转到第 i 个词节点 v_i 的比重,其可通过式(7)计算得到,而矩阵 M 的稳定值则可通过式(9)迭代计算进行确定。

$$B_i = (1 - d) + d \times B_{i-1} \times M \quad (9)$$

其中, B_i 是第 i 次迭代操作结束时所有节点的综合得分。迭代次数的上限为30,当连续两次计算结果的收敛误差为0.000 1时停止,而每个词的综合得分就是其在关键词词图中的节点影响力得分,根据分值高低对所有词节点进行降序排序,并选取其中前 N 个词节点作为关键词抽取结果。

3 实验结果与分析

3.1 实验数据

为保证测试数据的客观性和测试结果的可重现性,同时便于对不同关键词抽取方法进行实验对比,本文实验使用搜狐校园算法大赛提供的来自搜狐网站的新闻语料,解析其中的新闻标题和正文内容并将其作为文档集,将事先标记的关键词标签作为文档对应的人工标注关键词组成测试数据集,共选取1 000篇文档数据。本文选择搜狐校园算法大赛数据的主要原因为:1)数据由搜狐提供,保证了真

实性;2)搜狐新闻的新闻文章关键词通常经过人工筛选,具有参考性。

本文提出的关键词自动抽取方法采用Python实现,使用Jieba开源工具作为分词和词性分析工具。由于BERT模型对训练条件的要求较高,因此使用Google提供的BERT模型及中文预训练模型文件(词向量维度为768)。

3.2 结果分析

实验使用准确率(P)、召回率(R)以及F值(F)来评价关键词抽取效果并进行统计对比,3种指标的计算方法如式(10)~式(12)所示:

$$P = \frac{| \text{关键词抽取集合} \cap \text{人工标注集合} |}{| \text{关键词抽取集合} |} \quad (10)$$

$$R = \frac{| \text{关键词抽取集合} \cap \text{人工标注集合} |}{| \text{人工标注集合} |} \quad (11)$$

$$F = \frac{2 \times P \times R}{P + R} \quad (12)$$

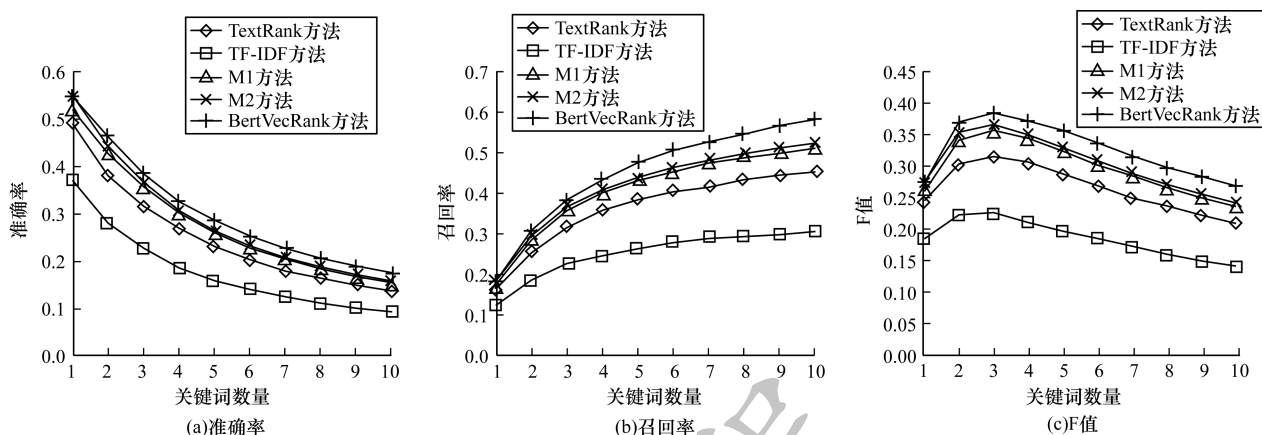
实验分别抽取 N (N 取1~10)个关键词作为自动抽取的关键词与数据集中人工标注的关键词进行对比。实验对比方法有:1)TF-IDF,传统词频逆文本频率关键词抽取方法;2)TextRank,传统TextRank关键词抽取方法^[2];3)M1,利用Word2Vec进行词向量聚类的关键词抽取方法^[17];4)M2,基于词向量聚类质心与TextRank加权的关键词抽取方法^[18];5) BertVecRank,本文提出的关键词抽取方法。

结合表1与图2可以看出,对文档进行关键词抽取时,TextRank方法明显优于TF-IDF方法,抽取效果更稳定。从F值可以看出,直接利用词向量聚类进行关键词抽取的M1方法相比M2方法效果略差,而将距离词向量质心越远、权重越高的词作为关键词的M2方法的抽取效果相对更好,表明关键词差异性有助于提高关键词抽取效率,但由于其计算聚类中心时受到外部词向量计算的影响较大,因此聚类效果与BertVecRank方法存在一定差距。本文使用词节点及其邻接节点直接进行差异比较,利用BERT词向量加权方式计算概率转移矩阵,以减少质心计算误差对聚类结果的影响,并且增加了不同主题词间的跳转概率,具有较好的关键词抽取效果。

表1 5种关键词抽取方法的性能对比

Table 1 Performance comparison of five keyword extraction methods

方法	$N=3$			$N=5$			$N=7$			$N=10$		
	准确率	召回率	F值	准确率	召回率	F值	准确率	召回率	F值	准确率	召回率	F值
TF-IDF	0.226	0.226	0.226	0.158	0.263	0.197	0.124	0.288	0.173	0.092	0.306	0.141
TextRank	0.297	0.297	0.297	0.221	0.326	0.263	0.176	0.363	0.237	0.137	0.403	0.204
M1	0.354	0.354	0.354	0.258	0.430	0.322	0.203	0.474	0.285	0.153	0.511	0.236
M2	0.366	0.366	0.366	0.263	0.439	0.329	0.206	0.481	0.289	0.157	0.523	0.242
BertVecRank	0.384	0.384	0.384	0.285	0.475	0.357	0.226	0.526	0.316	0.175	0.582	0.269

图2 N 取值为1~10时5种关键词抽取方法的准确率、召回率和F值对比Fig. 2 Comparison of precision, recall and F-value of five keyword extraction methods when N is 1~10

在表1中,当 $N=3$ 时,不同实验方法的准确率、召回率和F值基本相同,在实验过程发现由于抽取语料中人工提取的关键词平均个数为3,因此导致关键词为Top3时的准确率、召回率和F值基本一致。当关键词为Top3时,BertVecRank方法与M2方法的F值均为最优,BertVecRank方法比M2方法的F值提高1.8%。当 N 取3、5、7和10时,BertVecRank方法的平均F值比M2方法提升2.5%,并结合图2中F值可知,当BertVecRank方法抽取的关键词数量大于Top3并不断增加时,F值与其他方法相比具有明显优势,说明BertVecRank方法抽取出的关键词整体排序靠前,改进效果明显。由图2可看出,当关键词为Top1~Top10时所有方法的准确率、召回率和F值变化情况,其中BertVecRank方法的准确率整体高于其他方法,并且其召回率与其他方法的差距不断增加。可见,本文利用BERT词向量获取外部语义信息,并结合关键词间的差异性加权明显提升了重要关键词的抽取效率,因此BertVecRank方法的整体抽取效果最佳。

4 结束语

关键词抽取是快速获取文档核心语义的重要技术,是自然语言处理和信息检索等领域的重要组成部分,具有较高的理论和应用价值。本文提出一种融合BERT语义加权与网络图的关键词抽取方法,利用BERT词向量获取外部语义信息,并结合关键词间的差异性加权提升重要关键词的抽取效率。实验结果表明,当关键词为Top1~Top10时,本文方法的抽取准确率整体高于TF-IDF、TextRank、M1和M2这4种对比方法。后续将利用神经网络方法提取文档的结构信息特征,进一步优化关键词抽取效率。

参考文献

- [1] ZHAO Jingsheng, ZHU Qiaoming, ZHOU Guodong, et al. Review of research in automatic keyword extraction[J]. Journal of Software, 2017, 28(9): 2431-2449. (in Chinese)
赵京胜,朱巧明,周国栋,等.自动关键词抽取研究综述[J].软件学报,2017,28(9):2431-2449.
- [2] MIHALCEA R, TARAU P. TextRank: bringing order into texts[C]//Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing. Philadelphia, USA: Association for Computational Linguistics, 2004: 404-411.
- [3] HABIBI M, POPESCU-BELIS A. Diverse keyword extraction from conversations[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA: Association for Computational Linguistics, 2013: 651-657.
- [4] XIA Tian. Study on keyword extraction using word position weighted TextRank[J]. New Technology of Library and Information Service, 2013(9): 30-34. (in Chinese)
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(9): 30-34.
- [5] ZHANG Lijing, LI Yeli, ZENG Qingtao, et al. Keyword extraction algorithm based on improved TextRank[J]. Journal of Beijing Institute of Graphic Communication, 2016, 24(4): 51-55. (in Chinese)
张莉婧,李业丽,曾庆涛,等.基于改进TextRank的关键词抽取算法[J].北京印刷学院学报,2016,24(4): 51-55.
- [6] GU Yijun, XIA Tian. Study on keyword extraction with LDA and TextRank combination[J]. New Technology of Library and Information Service, 2014(7): 41-47. (in Chinese)
顾益军,夏天.融合LDA与TextRank的关键词抽取研究[J].现代图书情报技术,2014(7):41-47.
- [7] BLEI D M, NG A Y, JORDAN M I, et al. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

- [8] SIU M H, GISH H, CHAN A, et al. Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery[J]. Computer Speech and Language, 2014, 28(1): 210-223.
- [9] LI Junfeng, LÜ Xueqiang, ZHOU Shaojun. Patent keyword indexing based on weighted complex graph model[J]. New Technology of Library and Information Service, 2015(3): 26-32. (in Chinese)
李军锋, 吕学强, 周绍钧. 带权复杂图模型的专利关键词标引研究[J]. 现代图书情报技术, 2015(3): 26-32.
- [10] TIXIER A, MALLIAROS F, VAZIRGIANNIS M. A graph degeneracy-based approach to keyword extraction [C]// Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing. Philadelphia, USA: Association for Computational Linguistics, 2016: 1860-1870.
- [11] ZHANG Yuxiang, CHANG Yaocheng, LIU Xiaoqing, et al. MIKE: keyphrase extraction by integrating multidimensional information[C]// Proceedings of 2017 ACM Conference on Information and Knowledge Management. New York, USA: ACM Press, 2017: 1349-1358.
- [12] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. [2019-05-14]. <https://arxiv.org/abs/1301.3781>.
- [13] PENNINGTON J, SOCHER R, MANNING C. GloVe: global vectors for word representation[C]// Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Philadelphia, USA: Association for Computational Linguistics, 2014: 1532-1543.
- [14] WANG R, LIU W, MCDONALD C. Corpus-independent generic keyphrase extraction using word embedding vectors [EB/OL]. [2019-05-14]. <http://www.oalib.com/paper/4057741>.
- [15] JIANG Fang, LI Guohe, YUE Xiang. Semantic-based keyword extraction method for document[J]. Application Research of Computers, 2015, 32(1): 142-145. (in Chinese)
姜芳, 李国和, 岳翔. 基于语义的文档关键词提取方法[J]. 计算机应用研究, 2015, 32(1): 142-145.
- [16] NING Jianfei, LIU Jiangzhen. Using Word2vec with TextRank to extract keywords[J]. New Technology of Library and Information Service, 2016(6): 20-27. (in Chinese)
宁建飞, 刘降珍. 融合 Word2vec 与 TextRank 的关键词抽取研究[J]. 现代图书情报技术, 2016(6): 20-27.
- [17] LI Yuepeng, JIN Cui, JI Junchuan. A keyword extraction algorithm based on Word2vec[J]. E-science Technology & Application, 2015(4): 54-59. (in Chinese)
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.
- [18] XIA Tian. Extracting keywords with modified TextRank model[J]. Data Analysis and Knowledge Discovery, 2017(2): 28-34. (in Chinese)
夏天. 词向量聚类加权 TextRank 的关键词抽取[J]. 数据分析与知识发现, 2017(2): 28-34.
- [19] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2019-05-14]. <https://arxiv.org/abs/1810.04805>.
- [20] WEI Hongxi, GAO Guanglai, SU Xiangdong. LDA-based word image representation for keyword spotting on historical mongolian documents [C]// Proceedings of International Conference on Neural Information Processing. Berlin, Germany: Springer, 2016: 432-441.
- [21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing. New York, USA: ACM Press, 2017: 6000-6010.

编辑 陆燕菲