词语位置加权 TextRank 的关键词抽取研究*

夏　天
(中国人民大学数据工程与知识工程教育部重点实验室　北京 100872)
(中国人民大学信息资源管理学院　北京 100872)

【摘要】把关键词抽取问题看作是构成文档词语的重要性排序问题,基于 TextRank 基本思想,构建候选关键词图,引入覆盖影响力、位置影响力和频度影响力用于计算词语之间的影响力概率转移矩阵,通过迭代法实现候选关键词分值计算,并挑选前 N 个作为关键词抽取结果。实验结果表明,对词语位置加权的 TextRank 方法优于传统的 TextRank 方法和基于 LDA 主题模型的关键词抽取方法。

【关键词】关键词抽取　词排序　TextRank　图模型　LDA
【分类号】G350

Study on Keyword Extraction Using Word Position Weighted TextRank

Xia Tian
(Key Laboratory of Data Engineering and Knowledge Engineering of
Ministry of Education, Renmin University of China, Beijing 100872, China)
(School of Information Resource Management, Renmin University of China, Beijing 100872, China)

【Abstract】The keyword extraction problem is taken as a word importance ranking problem. In this paper, candidate keyword graph is constructed based on TextRank, and the influences of word coverage, location and frequency are used to calculate the probability transition matrix, then, the word score is calculated by iterative method, and the top N candidate keywords are picked as the final results. Experimental results show that the proposed word position weighted TextRank method is better than the traditional TextRank method and LDA topic model method.

【Keywords】Keyword extraction　Word rank　TextRank　Graph model　LDA

1　引　言

关键词是表达一个文档核心意义的最小单元,人工抽取关键词耗时费力,结果因人而异,因此,实现自动抽取具有重要意义。关键词抽取的任务就是从一段给定的文本中自动抽取出若干有意义的词语或词组,抽取方法既可以通过训练语料构建模型实现,也可以借助于词语之间的关系直接从文本本身抽取,后者因无需训练过程,应用较为方便,TextRank[1] 是其中的典型代表。

经典的 TextRank 算法在应用于关键词抽取时,未探讨词语对相邻结点影响力强弱不同时的处理方法。本文贡献在于指出词语本身的重要性差异会影响相邻结点的影响力传递结果,并通过分析影响词语重要性的主要因素,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,并归并计算形成候选关键词的整体概率转移矩阵,结合抽税机制实现词语排序和关键词抽取,取得了较好的实验结果。

收稿日期: 2013 - 07 - 01
收修改稿日期: 2013 - 08 - 19
* 本文系国家社会科学基金项目"Web2.0 环境下的网络舆情采集与分析"(项目编号:09CTQ027)和国家社会科学基金重大项目"云计算环境下的信息资源集成与服务研究"(项目编号:12&ZD220)的研究成果之一。

2　研究背景

针对关键词抽取问题已经开展了大量研究，部分研究者把关键词抽取看作是一个分类问题[2,3]，通过训练数据构建学习模型，进而判断词语是归属于关键词类别还是非关键词类别，属于典型的有指导学习方法。有指导学习需要事先标注高质量的训练数据，人工预处理的代价较高。

本文关注的重点是无指导关键词抽取方法的研究，主流方法可归纳为三种：基于 TF－IDF 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词图模型的关键词抽取方法。其中，基于 TF－IDF 统计特征进行关键词抽取是一种简单易行的常用方法，但这种方法忽略了重要的低频词语和文档内部的主题分布语义特征，因此，基于主题模型的关键词抽取方法在近年来得到了人们的重视。主题模型中以基于 LDA 的关键词抽取方法应用最为广泛[4-6]，LDA 是一种无指导机器学习技术[7]，通过大量已知的“词语－文档”矩阵和一系列训练，推理出隐藏在内部的“文档－主题”分布和“主题－词语”分布，出现在文档中主要主题中的主要词语更有可能被识别为关键词。主题模型需要对数据进行训练得到，关键词抽取的效果与训练数据的主题分布关系密切。

与基于 TF－IDF 和 LDA 的关键词抽取方法不同，以 TextRank 为典型代表的基于词图模型的关键词抽取算法[1]，不需要事先对多篇文档进行学习训练，因其简洁有效而得到了广泛应用。TextRank 的思想来源于 PageRank[8]，通过把文本分割成若干组成单元并建立图模型，利用投票机制对文本中的重要成分进行排序，仅利用单篇文档本身的信息即可实现关键词抽取。

TextRank 的迭代模型在理论上支持带权运算，但文献[1]在应用 TextRank 方法进行英文关键词抽取时，仅考虑了词语的词性信息，采用的是词语结点影响力均分的无权图模型，以 A、B 和 C 三个词语为例，假设 A 与 B 相邻接，A 与 C 相邻接，传统 TextRank 算法会把 A 的影响力分值以 50% 的比例均匀传递到 B 和 C，而与 B 和 C 本身的差异无关。然而，根据马太效应，与 A 相邻的重要词语应获取更多的由 A 所传递的分值，非重要词语所吸收的分值相应减少。基于这一假设，本文分析提出了词语结点影响力的相关因素，并给出了影响力分值的具体迭代计算方法，从而显著提高了单文档的关键词抽取效果。

3　候选关键词图的构建

简单说来，可以将关键词的抽取问题转换为构成文档词语的重要性排序问题，把词语按照表达文档意图的强度递减排序，前 n 个词语即可作为文档的关键词。如果能够将构成文档的词语及其关系组织成为一张图，即可利用图模型的有关理论对结点进行排序，进而实现关键词抽取，为此，在文献[1]基础上，笔者针对中文文本采用如下步骤构建候选关键词图：

(1) 把给定文本 T 按照完整句子进行分割，即 $T=[S_1,S_2,\cdots,S_m]$；

(2) 对于每一个句子 $S_i \in T$，进行中文分词和词性标注处理，仅保留切分后的重要词语，如名词、动词、形容词，即 $S_i=[t_{i,1},t_{i,2},\cdots,t_{i,n}]$，其中，$t_{i,j} \in S_i$ 是保留后的候选关键词；

(3) 构建候选关键词图 $G=(V,E)$，其中，V 为结点集，由步骤(2)生成的候选关键词组成，$E \subseteq V \times V$，对于每一个 $t_{i,j} \in S_i$，$t_{i,j+1} \in S_i$，有 $<t_{i,j},t_{i,j+1}> \in E$。

为便于形式化描述，令候选关键词图 $G=(V,E)$ 是由结点集 V 和边集 E 组成的一个有向图，对于任意的一个结点 v_i，令：

$$In(v_i)=\{v_j \mid <v_j,v_i> \in E\}$$

$$Out(v_i)=\{v_j \mid <v_i,v_j> \in E\}$$

即 $In(v_i)$ 表示指向结点 v_i 的结点集合，$Out(v_i)$ 表示结点 v_i 所指向的结点的集合。

进一步，令 w_{ij} 表示由结点 v_i 指向 v_j 的边的权重，则结点 v_i 的分值可通过如下公式计算[1]：

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

公式(1)即为 TextRank 的递归公式，与最早提出的 PageRank 算法完全一致，通过结点之间的投票或推荐机制，实现重要性排序，一个结点链入的结点集表示其投票支持者，投票者越重要、数量越多，则被投票者的排名越靠前。其中的 d 为阻尼系数(Damping Factor)，一般取值为 0.85，其在 PageRank 中的原始意义表示在任意时刻，用户到达某页面后并继续向后浏览的概率值[8]。

TextRank 的主要贡献在于将 PageRank 的思想引入到文本组成单元的重要性排序领域，TextRank 算法

应用于关键词抽取时，构建的是一种无向无权图，每一个结点被赋予一个初始值 1，然后迭代计算权重。直观看来，可以根据某种策略对部分重要的结点赋予较高的初值，以改善排序结果，但这种方法无法奏效，实际上，排序的结果对应转移矩阵的特征向量，和结点赋予的初值无关，而是由连接结点的边的权重所决定。

笔者将基于候选关键词图，讨论如何引入边的权重以改进排序效果，实现关键词抽取。

4 权重赋值与关键词抽取

假设候选关键词图中与词语 v 相关的结点和边如图 1 所示：

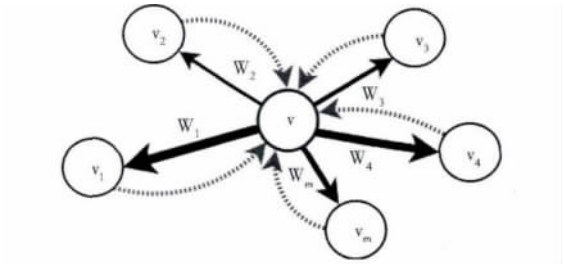

图 1 候选关键词图示例

其中带箭头实线表示结点 v 到所指向结点 v_i 的转移概率，线条的粗细表示转移概率的大小；虚线则表示由 v_i 结点跳转到结点 v 的转移概率。

如果用权重的大小代表转移概率的大小，构建两个词语之间的权重就成为关键词抽取的关键。通过观察分析，发现结点之间的权重主要与以下因素密切相关：

(1) 覆盖重要性：如果指向候选关键词 t 的不同词语数量越多，则 t 越重要，即 t 所拥有的不同投票者越多，其重要性越高。

(2) 位置重要性：候选关键词 t 出现在重要的位置，例如标题、摘要、段落的首句中，则 t 越重要。

(3) 频度重要性：候选关键词 t 出现的频度越高，则 t 越重要。

对于图中的任一结点 v 来说，其重要性得分由其相邻结点的贡献组成（如图 1 中相邻结点：$v_1, v_2, \cdots, v_m$），而其本身的得分也将被转移到相邻结点，即影响力的转移。根据以上假设，一个结点对相邻结点集的影响力可以分解为三个组成部分：覆盖重要性的影响力、词语重要性的影响力和频度重要性的影响力，令 W 表示结点的整体影响力权重，α, β, γ 分别表示这三类不同的影响力所占的比重，则有：

$$W = \alpha + \beta + \gamma = 1 \quad (2)$$

对于任意两个结点 v_i 和 v_j，结点 v_i 对 v_j 的影响力通过其有向边 $e = <v_i, v_j>$ 传递，边的权重决定了 v_j 最终所获得 v_i 部分的分值大小。针对不同类型的影响力权重，分别给出其计算方法：

(1) 令 $w\alpha(v_i, v_j)$ 表示结点 v_i 的覆盖影响力传递到结点 v_j 的权重，计算公式如下：

$$w\alpha(v_i, v_j) = \frac{1}{|Out(V_i)|} \quad (3)$$

其中，$|Out(V_i)|$ 表示 v_i 所指向的结点集合的元素数量，公式(3)说明结点 v_i 的覆盖影响力将会被均匀传递到相邻结点。

(2) 令 $w\beta(v_i, v_j)$ 表示结点 v_i 的词语重要性影响力传递到结点 v_j 的权重，计算公式如下：

$$w\beta(v_i, v_j) = \frac{I(v_j)}{\sum_{v_k \in Out(v_i)} I(v_k)} \quad (4)$$

其中，$I(v_j)$ 表示结点 v_j 的重要性取值，重要性赋值可根据实际需要采取不同的策略，在本文实验中，笔者仅处理了词语位置影响力，即采用如下方式对词语重要性进行赋值：

$$I(v) = \begin{cases} \lambda & v \text{ 所对应的词语在标题中出现} \\ 1 & \text{其他情况} \end{cases} \quad (5)$$

其中，λ 为一个大于 1 的参数。

(3) 令 $w\gamma(v_i, v_j)$ 表示结点 v_i 的频度影响力传递到结点 v_j 的权重，计算公式如下：

$$w\gamma(v_i, v_j) = \frac{C(v)}{\sum_{v_k \in Out(v_i)} C(v_k)} \quad (6)$$

其中，$C(v)$ 表示结点 v 在文本中出现的次数，公式(6)体现出高频度的词语将从连接点获得更高的影响力权重。

假设文本 T 共包含 n 个候选关键词，即 T 所对应的候选关键词图由 n 个结点组成，所有结点的初始分值可以通过一个分量为 1、维数为 n 的向量 B_0 表示，即：

$$B_0 = (\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n})$$

同时，构建词语之间的影响力转移矩阵 M 如下：

$$M = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix}$$

其中，M 中的第 j 列元素表示第 j 个候选关键词 v_j 的影响力转移到其他词语时的权重分配情况，每一列数值之和为 1。元素 w_{ij} 表示 v_j 的影响力转移到第 i 个词语 v_i 的权重比重，可通过下式计算得到：

$$w_{ij} = \alpha \cdot w\alpha(v_j, v_i) + \beta \cdot w\beta(v_j, v_i) + \gamma \cdot w\gamma(v_j, v_i) \quad (7)$$

给定一个结点 v_i，令 $S(v_i)$ 表示每次迭代计算中从其入点集合 $In(v_i)$ 中所计算得到的分值，结合上述公式，$S(v_i)$ 可通过下式计算得到：

$$S(v_i) = \sum_{v_j \in In(v_i)} S(v_j) \times w_{ji} \quad (8)$$

所有结点的分值在经过转移矩阵的第一次转移之后，其新分值 B_1 可通过 $M \times B_0$ 计算得到，但是，这种迭代计算方式无法处理非连通候选关键词图，为此，实验中引入了链接分析中的抽税机制[9]，允许每个结点能够以一个较小的概率随机转移到任一个结点，此时，迭代公式更改为：

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

其中，e 为一个所有分量为 1、维数为 n 的向量。

基于公式(9)进行迭代运算，B_i 最终会收敛，当两次迭代结果 B_i 和 B_{i-1} 之间的差异非常小，趋近于零时，停止迭代运算。设最终迭代结果为 B，进一步对 B 按分值大小降序排序，并取前 σ 个候选关键词作为关键词抽取结果。

5　实验结果

本文提出的关键词抽取算法采用 Java1.6 予以实现，分词和词性标注使用中国科学院计算技术研究所 ICTCLAS 的 Java 开源实现 ANSJ①。

为测试算法的有效性，笔者选择基于 LDA 的关键词抽取算法进行对比，基于文献[7]中对 LDA 模型的描述，令 φ 表示 LDA 中主题－词语的概率分布，φ_w^z 表示词语 w 在主题 z 中的概率，θ 表示文档－主题的概率分布，$\theta_z^{(d)}$ 表示指定文档 d 中的主题 z 的概率。则一个词语 w 在文档 d 中面向 t 个主题的概率可通过以下公式计算：

$$p(w|d) = \sum_{j=1}^{t} \varphi_w^{(z=j)} \theta_{z=j}^{(d)}$$

该数值的大小可以反映词语在文档中面向主题的重要性，按照大小顺序挑选前 σ 个词语作为关键词，即可实现关键词的抽取。

鉴于目前缺少权威的面向关键词抽取的中文公开测试数据集，为保证测试数据的客观性和测试结果的可重现性，实验针对门户网站所提供的较为规范的新闻报道，采用自主研发的主题链接自动抽取算法[10]和网页正文自动抽取算法[11]，自动提取网页包含的标题、内容以及 META 中的关键词，去除关键词为空、关键词为文章标题的无效报道，并人工审核抽取结果，保证所采用的基准关键词与文章主题内容一致，最终保留 1 000 篇报道形成测试数据集合②，每篇测试文档所拥有的关键词数量平均为 2.73 个。为评估关键词抽取效果，用 KA 表示文章本身所提供的关键词集合，KB 表示算法抽取出的关键词集合，则准确率 P、召回率 R 和常用的宏平均 F 值可通过如下公式计算：

$$P = \frac{|KA \cap KB|}{|KB|} \quad R = \frac{|KA \cap KB|}{|KA|} \quad F = \frac{2 \times P \times R}{P + R}$$

实验中，$d = 0.85$；$\sigma = 5$，矩阵迭代运算的终止条件为两次迭代结果的差异值小于等于 0.005，或者迭代次数超过 20。为确定参数 λ 的合理取值，选定两组较有代表性的 α，β，γ 参数，分别计算 λ 从 1 到 100 时的 F 值变化情况，结果如图 2 所示：

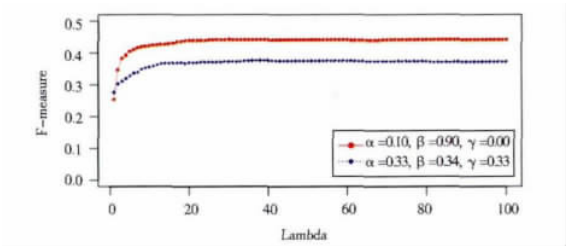

图 2　不同取值对应的 F 值分布曲线

对于 $\alpha = 0.33$，$\beta = 0.34$，$\gamma = 0.33$ 的情况，在 λ 取值为 36 到 41 之中的任一值时，F 值达到最大值 0.3761；当 $\alpha = 0.1$，$\beta = 0.9$，$\gamma = 0.0$ 时，λ 取值为 30 时达到最大值 0.4429。通过观察分析可得，λ 超过 20 之后，不同取值对于准确率的影响基本一致，在 30 附近已经比较稳定，因此，实验选定 $\lambda = 30$，进一步针对 α、β、γ 不同的参数组合，分别计算每一篇报道的关键词抽取结果，并取其均值，结果如表 1 所示：

表 1　实验结果

α	β	γ	P(%)	R(%)	F(%)
1.00	0.00	0.00	19.52	39.06	25.40
0.00	1.00	0.00	34.72	65.93	44.22
0.00	0.00	1.00	21.34	42.59	27.75
0.33	0.34	0.33	29.00	56.56	37.32
0.10	0.90	0.00	34.80	65.97	45.56

当参数取 $\alpha = 1$，$\beta = 0$，$\gamma = 0$ 时，算法即转化为传统的基于 TextRank 的关键词抽取算法，从表 1 可以看出，传统方式在实验数据中的 F 值仅有 25.40%，是实

① https://github.com/ansjsun/ansj_seg.

② 测试数据集已发布到如下地址：https://github.com/iamxiatian/data/tree/master/sohu-dataset.

验中不同参数组合中 F 值最低的情况，而均匀分配参数权重时，F 值提高到 37.32%。从数据中还可以看出，以位置信息为主的词语本身的重要性起了最为重要的作用，而词语频度信息所起作用较小。

同时，对于实验中所对比的 LDA 算法，文档的内容由文档标题拼接文档正文组合而成，词汇的概率分布通过 Gibbs 抽样获取，Gibbs 抽样的主题数目记为 K，超参数 $\alpha=50/K$，$\beta=0.01$，每次迭代次数为 1 000。由于实验数据的内容主题较为分散，分别统计了主题数目 K 从 2 到 180 变化时，关键词抽取结果的变化情况，如图 3 所示：

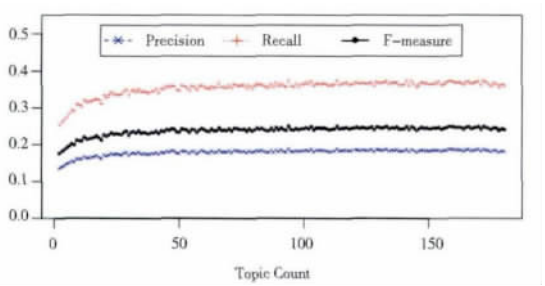

图 3　不同主题数量的关键词抽取结果

针对基于 LDA 的关键词抽取方法，由图 3 可得出两点结论：主题数量参数的设置对关键词抽取结果有一定影响；基于 LDA 的关键词抽取算法在主题分布不明显的情况下，准确率不高，实验中最高的 F 值仅为 25.14%（此时主题数量 K = 113），与传统的 TextRank 算法抽取结果相似。究其原因，LDA 算法基于词袋假设，未考虑词语的位置信息，适用于文档集背后存在较有规律的主题分布的情况。可见，本文所提方法充分利用了文档本身的信息，抽取效果显著优于主题模型的关键词抽取方法和传统的 TextRank 抽取方法。

6　结　语

本文介绍了 TextRank 用于关键词抽取的基本思想和候选关键词图的构建过程，提出引入覆盖影响力、位置影响力和频度影响力用于计算候选关键词之间的转移概率，进而构建词语之间的影响力概率转移矩阵，通过迭代法计算候选关键词的重要性分值，并挑选高分值的候选关键词作为文本的关键词抽取结果。实验结果表明，对于单文档的无指导关键词抽取来说，词语所在的位置对于正确抽取关键词的贡献最大，不同投票构成的覆盖影响力次之，而频度的重要性较低。同时，由于 LDA 基于词袋假设，不考虑文档中词语的排列顺序和位置，对于主题数量不确定的数据集，其抽取的准确率显著低于改进的 TextRank 抽取方法。

下一步研究工作包括：考虑词语组合情况，实现由多词构成的关键词抽取；考虑词语位置因素、词语在主题分布中的重要性因素，优化词语重要性的参数设置，并进行试验分析。

参考文献：

[1] Mihalcea R, Tarau P. TextRank: Bringing Order into Texts [C]. In: *Proceedings of Empirical Methods in Natural Language Processing*, Barcelona, Spain. 2004: 404 – 411.

[2] Frank E, Paynter G W, Witten I H, et al. Domain – Specific Keyphrase Extraction [C]. In: *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden. 1999: 668 – 673.

[3] Turney P D. Learning Algorithms for Keyphrase Extraction [J]. *Information Retrieval*, 2000, 2(4): 303 – 336.

[4] Pasquier C. Task 5: Single Document Keyphrase Extraction Using Sentence Clustering and Latent Dirichlet Allocation [C]. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010: 154 – 157.

[5] 石晶，李万龙. 基于 LDA 模型的主题词抽取方法 [J]. *计算机工程*, 2010, 36(19): 81 – 83. (Shi Jing, Li Wanlong. Topic Words Extraction Method Based on LDA Model [J]. *Computer Engineering*, 2010, 36(19): 81 – 83.)

[6] 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. (Liu Jun, Zou Dongsheng, Xing Xinlai, et al. Keyphrase Extraction Based on Topic Feature [J]. *Application Research of Computers*, 2012, 29(11): 4224 – 4227.)

[7] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. *Journal of Machine Learning Research*, 2003, 3: 993 – 1022.

[8] Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking: Bringing Order to the Web [R]. Stanford Digital Library Technologies Project, 1998.

[9] Rajaraman A, Ullman J D. Mining of Massive Datasets [M]. Cambridge University Press, 2012: 171 – 173.

[10] 夏天. 中心网页中主题网页链接的自动抽取 [J]. *山东大学学报：理学版*, 2012, 47(5): 25 – 31. (Xia Tian. Automatic Extracting Topic Page Links from Hub Page [J]. *Journal of Shandong University: Natural Science*, 2012, 47(5): 25 – 31.)

[11] 夏天. 基于扩展标记树的网页正文抽取 [J]. *广西师范大学学报：自然科学版*, 2011, 29(1): 133 – 137. (Xia Tian. Content Extraction of Web Page Based on Extended Label Tree [J]. *Journal of Guangxi Normal University: Natural Science Edition*, 2011, 29(1): 133 – 137.)

（作者 E – mail: xiat@ruc.edu.cn）