融合 LDA 与 TextRank 的关键词抽取研究*

顾益军[1]　夏　天[2,3]
[1](中国人民公安大学网络安全保卫学院　北京 100038)
[2](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)
[3](中国人民大学信息资源管理学院　北京 100872)

摘要: 【目的】通过将单一文档内部的结构信息和文档整体的主题信息融合到一起进行关键词抽取。【方法】利用 LDA 对文档集进行主题建模和候选关键词的主题影响力计算，进而对 TextRank 算法进行改进，将候选关键词的重要性按照主题影响力和邻接关系进行非均匀传递，并构建新的概率转移矩阵用于词图迭代计算和关键词抽取。【结果】实现 LDA 与 TextRank 的有效融合，当数据集呈现较强的主题分布时，可以显著改善关键词抽取效果。【局限】融合方法需要进行代价较高的多文档主题分析。【结论】关键词既与文档本身相关，也与文档所在的文档集合相关，二者结合是改进关键词抽取结果的有效途径。

关键词: 关键词抽取　LDA　TextRank　图模型
分类号: TP393

1　引　言

关键词抽取是指从给定文本中快速获取有意义的词或短语的过程，在信息检索、自动摘要、知识发现与挖掘等领域发挥着重要作用，更是图书情报领域实现自动标引的关键方法。

完整的关键词抽取一般分为两个步骤：候选关键词选取和从候选关键词中推荐关键词，候选关键词的选取可以结合词性和 N 元词串实现，相对较为简单，因此，如何从众多的候选关键词中挑选出目标关键词，成为关键词抽取的关键。关键词抽取既可以通过统计方法获取，也可以利用词图方法计算得到，其中以 TextRank[1]为典型代表的基于词图的无监督关键词抽取方法应用效果较好，引起了研究领域的广泛关注。

TextRank 及其改进算法[2]仅利用了文档本身的信息，文档中每个词语的重要性默认会被均匀传递到所有相邻节点，为了能够充分利用文档本身的信息和文档集合所提供的外部主题信息，本文将 LDA 与 TextRank 融合到一起，通过 LDA 预先统计学习候选词语的主题重要性，进而对 TextRank 算法进行改进，将候选词语节点的重要性按照主题分布进行非均匀转移，通过迭代计算为每个候选词语进行重要性赋值，并通过排序输出关键词抽取结果，取得了较为理想的实验结果。

2　研究背景

根据是否需要标注训练语料，可以把关键词抽取方法分为两大类：有监督关键词抽取和无监督关键词抽取。典型的有监督抽取方法把关键词抽取看作是一个是否为关键词的二分类问题[3,4]，这种方法能够利用更多的已有信息，效果相对较好，但由于需要事先标注高质量的训练数据，人工预处理的代价较高。因此，现有的关键词抽取研究主要集中在无监督关键词抽取方面，主流方法又可归纳为三种：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

收稿日期: 2014-02-07
收修改稿日期: 2014-03-05
*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)和北京高等学校青年英才计划项目“基于链接和主题分析的微博社区挖掘研究”(项目编号: YETP0215)的研究成果之一。

词图模型的关键词抽取。

基于 TF-IDF 统计特征进行关键词抽取简单易行，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 LDA[5]隐含主题模型的关键词抽取方法在近年来得到了人们的重视[6-8]。LDA 的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。

词图模型把文档看作是由词组成的网络，以 PageRank[9]链接分析理论为基础对词语的重要性进行迭代计算，无需数据集训练处理，仅利用文档本身的信息即可实现候选关键词重要性计算和关键词抽取，成为目前无监督关键词抽取的主流方法。Litvak 和 Last 将链接分析的另一方法 HITS[10]应用于候选关键词排序，在关键词抽取性能上与 TextRank 表现相似[11]。同时，TextRank 比 HITS 更为简洁和易于处理，因此本文采用 TextRank 作为词图计算的基本算法。文献[2]在 TextRank 基础上，分析提出了词语本身的重要性差异会影响相邻节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，结果表明根据位置对词语进行加权处理可以有效提高关键词抽取效果。文献[6]直接应用 LDA 模型进行主题词抽取，不能满足单文档抽取要求；文献[7]利用主题模型计算词语的主题特征，进而用装袋决策树方法构建关键词抽取分类模型，同样需要对多个文档进行统计处理。相比而言，文献[8]对关键词抽取问题进行了比较系统的研究，综合利用隐含主题模型和 PageRank 算法实现了关键词抽取，和本文研究更为接近。

文献[2]在衡量词语重要性时，采用的是相对简单的经验赋值方法，例如，对于在标题中出现的词语赋予较高的权重，实际上，词语的重要性与其主题密切关联，并可以通过对文档集合的主题学习获取。文献[8]利用主题信息对 PageRank 的随机跳转概率进行变更，在计算不同主题的 PageRank 结果后，合并排序输出关键词。与以上方法不同，本文保持了 PageRank 的均匀随机跳转假设，采用 LDA 隐含主题分析计算词语的整体影响力，结合词语之间的邻接关系用于 TextRank 中概率转移矩阵的计算，以较为简洁的方式改善了关键词抽取效果。

3 基于 LDA 的词语主题影响力计算

LDA(Latent Dirichlet Allocation)于 2003 年首次由 Blei 等[5]提出并用于文档主题建模。在 LDA 中，每篇文档被表示为 K 个隐含主题的混合分布，而每个主题又是在 W 个词语上的多项分布，其概率图如图 1 所示:

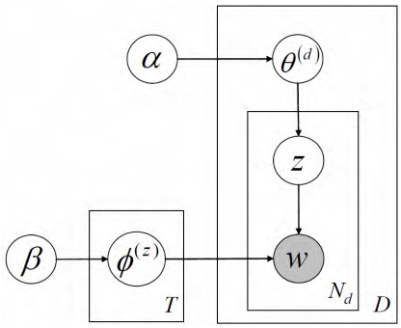

图 1 隐含主题模型 LDA 的概率图表示[12]

其中，φ表示 LDA 中主题–词语的概率分布，θ表示文档–主题的概率分布，α和β分别表示θ和φ所服从的 Dirichlet 先验分布的超参数，空心圆圈表示隐含变量，实心圆圈表示可观察到的变量，即词语。

对于文档 d 中的词语 w 来说，令 ti(w | d) 表示该词语在文档 d 中的主题影响力(Topic Influence, TI)，并假设文档 d 由 t 个隐含主题组成，可以认为，w 出现在一个主题 z 中的概率越大，则该词语相对于主题 z 而言越重要；若 w 对应的主题 z 在 d 中的出现概率越大，则表明主题 z 相对于文档 d 越重要，因而，w 也越重要。基于以上分析，令 φ_w^z 表示词语 w 在主题 z 中的概率，$\vartheta_z^{(d)}$ 表示指定文档 d 中的主题 z 的出现概率，笔者把词语 w 的主题影响力定义为:

$$ti(w \mid d) = \sum_{j=1}^{t} (\vartheta_z^{(d)} \times \varphi_w^{(z=j)}) \quad (1)$$

LDA 需要推断出两个参数，即每篇文档的“文档–主题”分布 θ 和每个主题的“主题–词语”分布 φ，通常利用 Dirichlet 分布和多项式分布之间的共轭性质通过 Gibbs 采样完成，此时，LDA 模型中会维护词语–主题的同现矩阵和文档–主题的同现矩阵[12]，并有:

$$\vartheta_{z=j}^{(d)} = \frac{C1(d, j)}{\sum_{k=1}^{t} C1(d, k) + t \times \alpha} \quad (2)$$

$$\varphi_w^{(z=j)} = \frac{C2(w, j)}{\sum_{k=1}^{N} C2(k, j) + N \times \beta} \quad (3)$$

其中, C1(d, j) 表示文档 d 中的词被赋给主题 j 的次数, C2(w, j) 表示在训练语料库中词语 w 被赋给主题 j 的次数, N 为词汇表的大小。借助于公式(2)和公式(3), 即可求解公式(1), 从而计算出一个词语在文档中的主题影响力。

4　引入主题影响力的转移矩阵构建与关键词抽取

根据链接分析理论, 可以把关键词抽取问题转换为构成文档词语的重要性排序问题, 词语的重要性则可以通过文档本身内部词语之间的结构关系推举产生。为构建具有相互联系的候选关键词图, 基于文献[2], 笔者对文本按照句子分割, 进行分词和词性标注, 并保留切分后的名词、动词、形容词等重要词语, 作为词图的节点, 根据词语的相邻关系构建词图的边, 形成候选关键词图。

TextRank 的基本思想是一个节点的重要性取决于有多少个相邻节点指向, 令 V 表示节点集, E 表示边集, $OD(v_i)$ 表示节点 v_i 的出度, 则 v_i 的 TextRank 值计算公式为:

$$R(v_i)=d\sum_{j:v_j\to v_i}\frac{1}{OD(v_j)}R(v_j)+(1-d)\frac{1}{|V|} \tag{4}$$

其中, $d\in[0,1]$ 为阻尼系数(Damping Factor), 表示每个节点均有 1–d 的概率随机跳转到其他节点, 而不要求两个节点之间一定存在边的连接, 从而确保算法在任意图上均可以收敛。

传统 TextRank 中每个节点的重要性会根据迭代计算不断改变, 为把词语的主题影响力纳入到计算过程中, 将任一词图节点 v_i 的关键属性分为两部分, 一部分为节点的当前重要性分值, 代表节点在文档内部结构中的权威性, 默认值为 1, 并在迭代过程中按照相邻节点的分值进行调整, 记为 $TR(v_i)$; 另一部分为节点本身的主题影响力分值, 通过公式(1)计算得到, 代表文档外部信息在节点上的重要性, 在词图计算过程中保持不变, 记为 $TI(v_i)$。例如, 图 2 为由{A, B, C, D, E, F}6 个节点组成的候选关键词图, 圆圈的左半部分表示节点的名称, 右上部分表示节点的重要性, 右下部分表示节点的影响力, 影响力一旦由主题分析确定, 就不再改变。

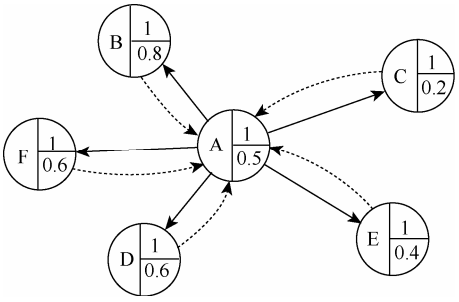

图 2　候选关键词图示例

在图 2 中, 传统算法会由节点 A 以相等概率向另外 5 个节点跳转, 即各有 20%的概率分别到达 B 节点至 F 节点。然而, 更为合理的方式应是按照节点的影响力和节点之间的投票结果综合得分进行跳转, 即向高影响力节点转移的概率要高于向低影响力节点转移的概率, 为此, 定义新的节点重要性迭代计算公式如下:

$$TR(v_i)=d\left(\alpha\sum_{j:v_j\to v_i}\frac{TI(v_i)}{OTI(v_j)}TR(v_j)+\beta\sum_{j:v_j\to v_i}\frac{1}{OD(v_j)}R(v_j)\right)+(1-d)\frac{1}{|v|} \tag{5}$$

其中, $OTI(v_i)=\sum_{j:v_i\to v_j}TI(v_j)$, α 和 β 为大于 0 的权重因子, 且有 $\alpha+\beta=1$, 实验中取值均为 0.5, 即节点的影响力和投票贡献程度均等。

仍以图 2 为例, 在新的计算方式下, 节点 A 向节点 B 新的转移概率变为:

$$0.5\times\frac{0.8}{0.8+0.2+0.6+0.4+0.6}+0.5\times\frac{1}{5}\approx 25.38\%$$

假设文档 d 共包含 n 个候选关键词, 即候选关键词图由 n 个节点组成, 则所有节点的初始分值向量 B_0 定义如下:

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

为迭代计算需要, 需构建词语之间的概率转移矩阵 M:

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

其中, 元素 w_{ij} 表示词语节点 v_j 转移到第 i 个词语 v_i 的概率, 并且每一列的数值之和为 1。根据公式(5), w_{ij} 可通过以下公式计算得到:

$$w_{ij} = \alpha \frac{TI(v_i)}{OTI(v_j)} + \beta \frac{1}{OD(v_j)} \quad (6)$$

在计算得到引入主题影响力的概率转移矩阵 M 之后，令 B_i 表示一次迭代结果，即所有词语当前的重要性分值所形成的向量，则迭代公式可由公式(4)进一步转换为:

$$B_i = d \times M \times B_{i-1} + (1-d) \times e/n \quad (7)$$

其中，e 为一个所有分量为 1、维数为 n 的向量。通过公式(7)进行迭代计算，当两次迭代运算的结果差异较小时计算结束，此时，进一步对向量 B 的分量进行降序排序，选择最前面指定数量的节点所对应的词语作为关键词输出结果。

简要说来，整个融合过程分为两步：基于 LDA 主题模型对文档集合进行文本建模，利用公式(1)实现词语的主题影响力计算；把主题影响力带入公式(6)，通过迭代计算实现节点的重要性计算，进而得到词语的重要性列表和关键词抽取结果。

5 实验结果

本文所提算法采用 Java1.6 予以实现，并选取公开的互联网网页作为测试文档，与学术文献类文档不同，网页文档虽容易获取，但语言规范性相对较弱，因此，关键词抽取的难度更大。为测试算法效果，实验采用文献[2]中所使用的通过链接自动抽取[13]和正文自动抽取[14]所构建的数据集，共由 1 000 篇网页文档组成。并定义准确率 P、召回率 R 和宏平均 F 值如下:

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

其中，KA 表示数据集本身提供的关键词集合，KB 表示自动提取的关键词集合。

实验中使用开源的分词和词性标注工具 ANSJ[15] 和机器学习包 Mallet[16]，并通过停用词表和词性过滤形成候选词关键词集合，过滤的词性集合为{代词，量词，数词，介词，方位词，副词，时间词，标点符号}，同时，实验中针对数据集手工添加形成的停用词表集合为{记者，报道，搜狐，时间，体育讯，虽然，责任编辑，编辑，直接，可能}。实验对比了三种不同的方法，分别是:

(1) LDA+TextRank：本文提出的通过 LDA 计算词语影响力，进而通过 TextRank 抽取关键词的方法;

(2) TextRank：基于 PageRank 原理的传统 TextRank 方法;

(3) LDA：通过本文所给出的基于 LDA 计算词语影响力的方法，按照词语影响力高低抽取关键词。

实验中，$d = 0.85$，每种方法抽取前 5 个词语构成关键词抽取结果，TextRank 迭代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。LDA 中，设 Gibbs 抽样的主题数目为 k，则超参数 $\alpha = 50/k$，$\beta = 0.01$，迭代次数为 1 500。当主题数目 k 从 2 到 180 变化时，三种方法的抽取结果如图 3 所示:

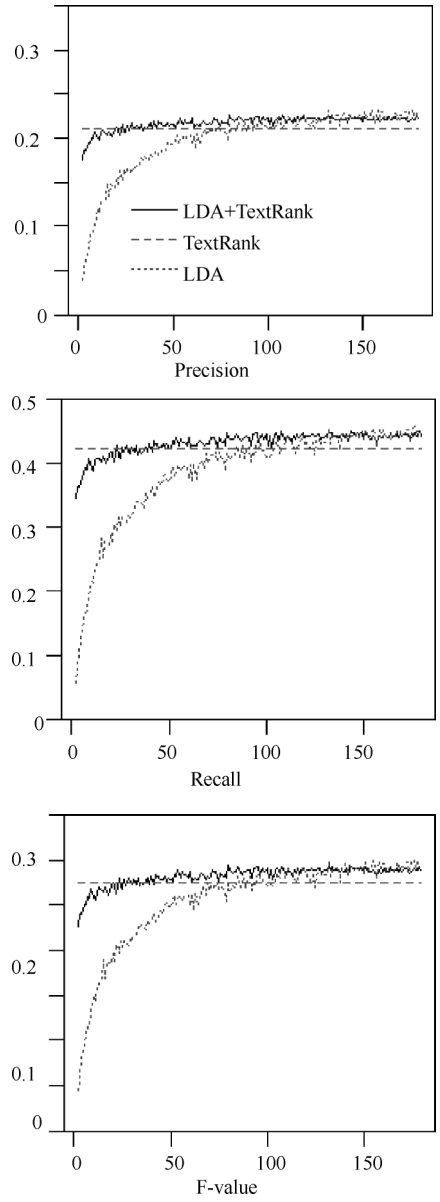

图 3 不同方法的关键词抽取结果对比

从图 3 中看出，TextRank 算法仅使用文档本身的信息即可实现关键词抽取，因此其结果和主题数量无关；方法 1 在主题数量超过 30 之后，优于 TextRank 方法和 LDA 方法；方法 3 抽取结果不够稳定，并且在主题数量较少时，效果较差。但三种方法在该数据集下的抽取结果总体差异不大，其 F 值平均分别为 28.24%、27.39%、25.61%，总体而言，方法 1 效果略优于其他方法。

由于该数据集的主题较为分散，为测试在主题性较强的情况下，三种方法在关键词抽取方面的差异，实验进一步挑选了文献[17]中所提供的文档集进行测试，该文档集共有 953 篇文档组成，但未提供关键词列表，实验中主题数量取值为 10，其他参数同上。为考察方法差异，将不同方法之间所抽取的关键词结果进行集合比对，任意两种方法抽取结果中共现的词语数量分布情况如图 4 所示:

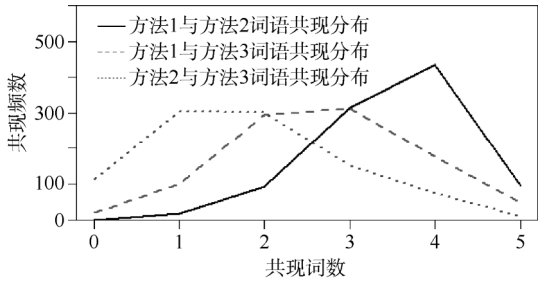

图 4　不同方法抽取结果的词语共现情况

图 4 中，折线后部分线段占得比重越高，则说明两种方法共现词语的数量越多，算法越接近。其中，方法 1 和方法 2 的对比结果里，5 个关键词中有 3 个、4 个和 5 个相同词语的文档数量分别为 315、435 和 96，三者占文档总数量的 88.77%，反映出方法 1 在很大程度上保留了 TextRank 可以充分利用文档本身结构信息的特点。方法 1 和方法 3 中，共现词语数量为 3、4、5 的文档数量分别为 314、179 和 47，共计 56.67%，说明方法 1 在一定程度上利用了文档集合的主题信息。而方法 2 和方法 3 的折线分布呈先凸后凹形状，二者之间存在明显差异。

另外，为观察本文所提方法与 TextRank 和 LDA 方法的差异，把有显著差异的抽取结果，即抽取结果完全不同或仅有一个词语相同的情况，按文档编号顺序将前 5 组结果列于表 1 和表 2 之中。

表 1　方法 1 与方法 2 抽取结果对比

文档编号	抽取方法	关键词				
		1	2	3	4	5
0ac20d05 fcdf9e72	1	利物浦	比赛	北京	英超	对手
	2	利物浦	红军	库伊特	gt	茶余饭后
07e0d395 fcdf9e72	1	冠军	比赛	选手	跳水	获得
	2	洛加尼斯	跳水	跳台	运动员	熊倪
06279765 fcdf9e72	1	图巴	比赛	机会	队员	继续
	2	图巴	苑维玮	换人	尼古拉	泰山
0de2fc45 fcdf9e72	1	钟诚	北京	比赛	冠军	来源
	2	钟诚	清华附中	篮球	男篮	职业
09098945 fcdf9e72	1	中国女足	女足	联队	明星	比赛
	2	中国女足	复兴	久违	比分	世界

表 2　方法 1 与方法 3 抽取结果对比

文档编号	抽取方法	关键词				
		1	2	3	4	5
03e62ea5 fcdf9e72	1	比赛	彭帅	中国	佩内塔	意大利
	3	比赛	体育	新闻	决赛	获得
0413dd15 fcdf9e72	1	比赛	决赛	种子	结束	以及
	3	比赛	冠军	中国	来源	北京
0260b2e5 fcdf9e72	1	邵佳一	比赛	不过	桑德尔	情况
	3	比赛	球队	联赛	新闻	赛季
0c279415 fcdf9e72	1	郑智	比赛	表现	机会	查尔顿
	3	比赛	主场	联赛	客场	裁判
09c577b5 fcdf9e72	1	比赛	孙继海	进攻	右前卫	开始
	3	比赛	联赛	机会	本场	进球

从对比列表中可以看出，TextRank 无法利用文档整体的信息，对于部分主题标识性强的词语，往往难以出现在抽取结果中；而单纯的 LDA 方法又无法充分利用文档内部的结构信息，主题性词语经常占据抽取结果的主要部分。 LDA+TextRank 方法综合利用了文档的内部和外部信息，效果相对较好，但抽取结果中依然出现了“以及”、“来源”等不适宜作为关键词的词语，该问题一方面和数据集包含噪音数据有关，另一方面和词法分析有关，可以通过词性过滤和完善停用词表解决。

6　结　语

文档本身的结构组成和文档之间所蕴含的主题信

息是关键词抽取的重要依据。本文基于 LDA 主题分析方法预先计算词语的主题影响力，进而改进 TextRank 算法的词语重要性迭代计算公式，把词语主题影响力纳入到概率转移矩阵的构建过程中，通过迭代计算实现候选关键词的排序和关键词抽取，进而在主题分布不明显和具有明显主题特征的两组数据集上分别进行了实验和对比分析。

实验结果表明，当主题分布不明显时，本文所提方法略优于 TextRank 和 LDA 方法，当数据集呈现出较强的主题分布特性时，融合 LDA 与 TextRank 能够取得更为理想的关键词抽取结果。同时，词法分析的准确率对关键词抽取结果也有较大影响，因此，进一步优化词表的构造与更新，并为词表中的词条设置不同的权重以提高关键词抽取效果，将是本研究的后续工作之一。

参考文献:

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

[2] 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. (Xia Tian. Study on Keyword Extraction Using Word Position Weighted TextRank [J]. New Technology of Library and Information Service, 2013(9): 30-34.)

[3] 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. San Francisco: Morgan Kaufmann Publishers Inc., 1999: 668-673.

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

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

[6] 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.)

[7] 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.)

[8] 刘知远. 基于文档主题结构的关键词抽取方法研究[D]. 北京：清华大学, 2011. (Liu Zhiyuan. Research on Keyword Extraction Using Document Topical Structure [D]. Beijing: Tsinghua University, 2011.)

[9] Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking: Bringing Order to the Web [R]. Stanford InfoLab, 1999.

[10] Kleinberg J M. Authoritative Sources in a Hyperlinked Environment[J]. Journal of the ACM, 1999, 46(5): 604-632.

[11] Litvak M, Last M. Graph-Based Keyword Extraction for Single-Document Summarization [C]. In: Proceedings of Workshop Multi-source Multilingual Information Extraction and Summarization (MMIES'08). Stroudsburg: Association for Computational Linguistics, 2008: 17-24.

[12] Steyvers M, Griffiths T. Probabilistic Topic Models [A]. // Landauer T, McNamara S D, Kintsch W. Handbook of Latent Semantic Analysis: A Road to Meaning [M]. Lawrence Erlbaum, 2007: 424-440.

[13] 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.)

[14] 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.)

[15] GitHub. ANSJ [EB/OL]. [2014-03-05]. https://github.com/ansjsun/ansj_seg.

[16] Mallet. Topic Modeling[EB/OL]. [2014-03-05]. http://mallet.cs.umass.edu/topics.php.

[17] Wang C, Zhang M, Ma S, et al. Automatic Online News Issue Construction in Web Environment [C]. In: Proceedings of the 17th International Conference on World Wide Web(WWW'08). New York: ACM, 2008: 457-466.

作者贡献声明:

顾益军：共同提出研究思路，共同设计研究方案，共同起草论文，负责最终版本修订；

夏天：共同提出研究思路，共同设计研究方案，共同起草论文，负责数据收集和实验。

(通讯作者：夏天 E-mail: xiatian1119@gmail.com)

Study on Keyword Extraction with LDA and TextRank Combination

Gu Yijun[1]　Xia Tian[2, 3]
[1](Schools of Cyber Security, People's Public Security University of China, Beijing 100038, China)
[2](Key Laboratory of Data Engineering and Knowledge Engineering, MOE, Renmin University of China, Beijing 100872, China)
[3](School of Information Resource Management, Renmin University of China, Beijing 100872, China)

Abstract: [**Objective**] Realize keyword extraction through the merger of the internal structure information of single document and the topic information among documents. [**Methods**] LDA is used for topic modeling and influence calculation of candidate keywords, then, the TextRank algorithm is improved and the importance of the candidate words is uneven transferred by topic influences and word adjacency relations. Furthermore, the probability transition matrix for iterative calculation is built and used to extract keywords. [**Results**] The effective combination of LDA and TextRank is achieved, and the keyword extraction results are improved significantly when the data set presents strong topic distribution. [**Limitations**] High-cost multi-document topic analysis is required for combination method. [**Conclusions**] Document keywords are associated with document itself and the related documents collection, combination of these two aspects is an effective way to improve the results of keyword extraction.

Keywords: Keyword extraction　LDA　TextRank　Graph model

EBSCO 为其开放元数据共享和技术协作政策增加 50 个数据库

EBSCO 信息服务(EBSCO)为其元数据共享和与发现服务提供商技术合作政策增加了数据库的数量。目前, EBSCO 的所有元数据(如果协议允许, 还包括全文)可用于 179 个 EBSCO 全文数据库, 也可用于所有 74 个 EBSCO 数字历史档案(含全文)和全部 55 万多本电子书。

2014 年 4 月, EBSCO 宣布了一项数据共享的开放政策, 明确了与其他发现服务提供商进行相互合作的关键问题。从该政策可以看出 EBSCO 致力于与合作供应商交换元数据、进行技术集成, 为双方用户提供更好的发现服务体验。除了元数据共享之外, 该政策还指出 EBSCO 需要为所有 EBSCO 全文数据库中的全文提供链接技术支持。

该政策的出台是为了鼓励系统集成服务供应商、发现服务提供商以及内容供应商进行更多的合作, 并进一步明晰共同的目标。为了达到这一目标, 该政策需要发现服务提供商(同时也是系统集成服务供应商)在 EBSCO 发现服务(EDS)中启用读者目录功能, 还需要双方进行其他的技术集成。

EBSCO 执行副总裁 Sam Brooks 表示本次最新增加的 50 个数据库是 EBSCO 考虑用户的利益, 为提供开放数据而不断努力的一个证明: "在政策出台的初期, 我们就立即提供了比其他内容供应商更加多的学术内容发现服务。期待与整个图书馆界进一步合作, 始终将用户受益放在第一位。"

完整的政策内容可参见: http://www.ebscohost.com/metadata-sharing-policy。

(编译自: http://www2.ebsco.com/en-us/NewsCenter/Pages/ViewArticle.aspx?QSID=700)

(本刊讯)