doi:10.3772/j.issn.2095-915x.2017.03.012

基于主题的 SE-TextRank 情感摘要方法

1. 南华大学计算机科学与技术学院 衡阳 421001；
2. 国家行政学院电子政务研究中心 北京 100089

刘志明[1] 于波[1] 欧阳纯萍[1] 余颖[1] 阳小华[1] 翟云[2]

摘 要 文本情感摘要技术的目的是以简洁的形式准确表达文章的核心情感内容。为解决不同的文档结构及内容特征等问题对摘要结果的影响，提出了一种基于主题的 SE-TextRank 情感摘要方法。通过 LDA 模型自动获取收敛后的文本主题，利用余弦距离算法进行主题句子分组，使用传统多特征融合以及 SE-TextRank 情感摘要算法对组内中心句抽取，最终获取目的摘要。实验表明，采用此方法能够更为高效的获取新闻文本摘要结果。

关键词： 文本摘要 , LDA 模型 , 余弦距离 , SE-TextRank, 特征融合

中图分类号： G35

SE-TextRank Opinion Summarization Method Based On Topic Model

1. School of Computer Science and Technology, University of South China, Hengyang 421001, China;
2. E-Government Research Center, Chinese Academy of Governance, Beijing 100089, China

LIU ZhiMing[1] YU Bo[1] OUYANG ChunPing[1] YU Ying[1] YANG XiaoHua[1] ZHAI Yun[2]

Abstract The purpose of the text sentiment summarization is to express the content of the article in a concise form. A topic-based SE-TextRank emotional abstract method was proposed in this study to solve the influence of different document structure and content characteristics on abstract results. This study

基金项目： 本文受国家自然科学基金（61402220，61672178），湖南省哲学社会科学基金（14YBA335,16YBA323），湖南省教育厅科研项目（15C1186）的资助。

作者简介： 刘志明(1972-)，博士，教授，硕士生导师，研究方向：大数据检索与挖掘、舆情监测；于波(1991-)，硕士研究生，研究方向：自然语言处理，Email: yuzhibbob@163.com；欧阳纯萍(1979-)，博士，副教授，研究方向：语义Web、情感分析；余颖(1980-)，硕士，讲师，研究方向：知识发现、信息检索；阳小华(1963-)，博士，教授，研究方向：信息检索、舆情分析；翟云（1979-），博士，副教授，研究方向：大数据、互联网+、政府治理创新。

doi:10.3772/j.issn.2095-915x.2017.03.012

obtained the convergent text theme automatically through LDA model, grouped the sentence topic through the cosine distance algorithm, applied traditional multi feature fusion and SE-TextRank sentiment summary algorithm to extract the central sentence within the group, and ultimately get the purpose summary. Results of the experiment showed that this method can be used to obtain the results of news text more efficiently.

Keywords：Text summarization, LDA model, cosine distance algorithm, SE-TextRank, feature fusion

1 概述

随着大数据时代的到来，对海量数据的处理成为自然语言处理领域中的关键一环。当前网络上积累了海量的新闻文本数据，对新闻文本的情感摘要提取可以满足用户迅速掌握当前事件主观情感的需要，从而使用户对资讯做出快速反应。因此，如何对其中的关键情感内容进行高效自动处理成为重中之重。

文本情感摘要技术的主要目的是以综合文本中各方面情感的文字信息准确反应文本表达的核心情感内容。依照文本情感摘要的特点，我们可以将它分为文本摘要和情感分析两个子问题。国内外的学者从这两种方向对文本情感摘要进行了细致的研究。

文本摘要是一种用简单文字表达一个或多个文档中心思想的技术[1]。根据采用的摘要方法的不同，主要分为基于统计的自动摘要和基于理解的自动摘要两种。基于统计的自动摘要技术主要是通过抽取原文中部分关键句子获得摘要结果。Luhn 等[2]提出了一种基于词频来进行自动文本摘要的方法，利用词频的高低抽取句子获得目的摘要。Conory 等[3]在 2001 年使用隐马尔科夫模型进行文本摘要，在模型算法的基础上融合了句子特征对摘要的影响。Wang 等[4]提出一种加权一致性文本摘要方法，采用聚合汇总的方式对句子排序。莫鹏等[5]在 2015 年提出一种基于超图的文本摘要方法，利用超图模型及词间高阶信息来生成摘要和关键词。近年，基于理解的摘要技术成为流行趋势。基于理解的摘要技术主要是利用句子语义和文本句式，通过文本内容分析来自动生成摘要的方法。林鸿飞等[6]2001 年提出了一种基于潜在语义索引的文本摘要方法，利用潜在语义索引在语义空间进行相似度计算来生成文本摘要。章芝青[7]于 2010 年提出一种基于语义的单文档自动摘要算法，将语义相似度与改进的 K-Medoids 聚类算法相融合。林萌等[8]采用融合句义结构模型的摘要方法，通过深化句子的语义分析层次提高句子语义表达。

文本情感分析是对文本的主观情感进行分析和归纳的过程[9]，具有较大的研究价值和应用价值。Kamps J 等[10]采用点互信和 WordNet 方法对形容词和情感词关联度计算来识别评价词语，避免了词汇的多义性。柳位平等[11]通过构建基础领域情感词典的方式对词语的语义情感倾向性分析，有效提高了情感分类效果。林莉媛等[12]在 2014 年提出一种基于 PageRank 的中文多文档文本情感摘要方法，利用主题与情感词关系来生成文本情感摘要。实际上，文本中句子之间语义及情感的关联性也会对情感摘要产生较大影响。本文通过主题收敛，融合了

句子语义相似度与句子情感相似度，提出了一种基于主题的 SE-TextRank 情感摘要方法。

2 基于主题的SE-TextRank情感摘要方法

新闻文本情感摘要的生成需要基于各个相关方面的集成判断。本文在情感摘要过程中，需要对新闻文本句子进行词汇化处理，将文本语料中的句子转化为一组词的集合形式，经过去停用词、去冗余、特殊标点符号等处理之后作为输入语料；然后利用 LDA 主题模型对输入语料进行主题抽取，采用余弦相似度算法获取 Topic-Sentence 关联分组；同时融合传统的文本特征进行分析；使用 SE-TextRank 文本情感排序算法获取句子语义情感排序权值；最终获取新闻文本情感摘要结果。

2.1 LDA主题抽取与句子分类

LDA 模型是一种三层贝叶斯概率主题生成模型，是在 PLSA(Probailistic latent semantic analysis) 模型的基础之上的贝叶斯化模型，在文本建模与检索等领域获得了广泛的应用。

在文本数据集的主题信息建模过程中，LDA 模型将每个文档表示为主题的集合：$T=\{t_1,t_2,..,t_n\}$。其中，每个主题都是文档词 w 的概率分布。LDA 模型采用了 Dirichlet-Multinomial 共轭结构来获取当前文档中各词的主题概率分布，同时也采用此共轭结构获取各主题中词汇的生成概率分布。我们利用 MCMC 算法中的Gibbs Sampling对联合分布进行采样。最终得到 LDA 模型的 Gibbs 抽样公式：

$$P(t_i=k\mid t_{\neg i},w)\propto\frac{n^{(k)}_{m,\neg i}+\alpha_k}{\sum_{k=1}^{K}(n^{(t)}_{m,\neg i}+\alpha_k)} \quad (公式\ 1)$$

其中，i 为第 i 个主题；为文本 m 中不包括第 i 项出现的主题 k 的次数；为主题 k 的先验参数；为文档词 t 的先验参数。

通过 LDA 模型的主题抽取过程，我们基于主题与文本语料中句子的关联关系，采用余弦距离算法来对文本语料中的句子进行主题分类。余弦相似度算法是一种用空间向量中两个向量夹角余弦值来计算彼此间差异的方法，本文使用该算法来计算主题与语料中各句子的相似度，进而得到 Topic-Sentence 关联分组集。

2.2 传统文本特征

（1）句子的位置特征。在结构性文章中，根据句子的位置特征可以判断句子的重要程度。我们知道在新闻文本中，越接近句首或句尾的句子其重要性程度越高。以此为判断标准得到句子的位置特征权重计算公式如下所示：

$$P(s_i)=\frac{1}{Min(i,N-i+1)} \quad (公式\ 2)$$

其中 i 为文本中句子 s_i 所在的位置，且 $i=\{1,....,N\}$，N 为整篇文本中句子的总数。

（2）关键字特征。由于句子是由词构成的序列，因此句子中关键字权重对句子的重要性有着重大影响。本文使用 Jieba 分词工具提取文本语料中的关键词数量，将获取的关键词与句子中的词组进行比对，统计句子中共现的关键字。关键字权值计算公式为：

$$G(s_i)=\frac{m}{M} \quad (公式\ 3)$$

其中 m 为当前句子中关键词的数量，M 为当前句子中词的总数量。

doi:10.3772/j.issn.2095-915x.2017.03.012

（3）句子长度特征。为了使摘要句子不受太短或太长的影响，需要为句子的长度添加权重值，定义句子长度特征权重计算公式为：

$$L(s_i)=1-\frac{|x-\mu|}{\mu}$$ （公式 4）

其中，μ 是文档中句子的平均长度，x 是当前句子的长度。

经过多次训练分别获取 η、β、λ 加权参数，将（7）、（8）、（9）、（10）式进行特征融合获取最终的句子传统文本特征权值：

$$W(s_i)=\eta\cdot POS(s_i)+\beta\cdot G(s_i)+\varepsilon\cdot L(s_i)$$ （公式 5）

按各组的句子摘要权值的重要度排序，得到每个 Topic-Sentence 关联分组的摘要句，最终获取新闻文本摘要。

2.3 融合语义情感关系的SE-TextRank情感排序算法

文本中的语义情感关系涉及语义相似度与情感相似度两个方面。语义相似度可以准确表述文档中句子词汇之间的语义关联，例如：“书籍”和“报纸”两个词语之间具有明显的概念相似性，因为两者都包含有相似的语义内容。文档中句子间情感相似度则是以句子情感义原为依托的情感综合体现，情感义原是以 < 中心词，情感词极性 > 二元结构来表征句子情感语义的基本单位。因此，本文采用句子情感义原的共现率来描述句子之间情感相似度。将句子语义相似度与情感相似度通过改进的 SE-TextRank 文本情感排序算法相关联，最终获取文本中各句子的语义情感排序权值。

2.3.1 语义相似度

本文采用《知网》语义相似度计算方法来计算词汇之间的语义相似度[13]。《知网》中有两个主要概念：“概念”与“义原”。“概念”是对词汇语义的一种表示，“义原”则是用于描述“概念”的最小意义单位。每个汉语词汇 p 的概念由一个或者多个义原进行描述，词语间义原的相似度计算采用如下公式：

$$sim(p_1,p_2)=\frac{\alpha}{\alpha+d}$$ （公式 6）

其中，p_1 和 p_2 为两个义原，α 为可变参数，d 为义原间的路径长度。

义原相似度为词汇间的相似度计算建立了基础，因此计算两个词汇语义相似度的过程就是计算词汇之间义原相似度最大值的过程：

$$wsim(w_1,w_2)=\max_{i=1..2,j=1..m} sim(m_{1i},m_{2j})$$ （公式 7）

对于句子 $s_a=\{w_{a1},w_{a2},...,w_{an}\}$ 和 $s_b=\{w_{b1},w_{b2},...,w_{bn}\}$，将句子 s_a 与句子 s_b 中各个单词进行两两相似度计算，取各词汇计算过程中的最大值作为本词汇的相似度权值。通过句子间词汇的相似度求和并标准化，最终得到句子语义相似度公式：

$$S(s_a,s_b)=\frac{\sum_{\substack{k=1....n\\l=1...m}}\max wsim(w_{ak},w_{bl})}{\log(|s_a|)+\log(|s_b|)}$$ （公式 8）

2.3.2 情感相似度

中文句子的情感语义需要依托句中词间关系和情感词极性共同表征，本文采用拥有 < 中心词，情感词极性 > 二元结构的情感义原作为描述句子情感的基本单位。首先对句子进行词性标注及句法分析来获取句子中依赖关系，然后将抽取的实词与主题特征词匹配来得到情感义原的中心词。使用《知网》的 HowNet 情感词典来对句子进行情感词抽取并对其情感极性进行识别。最后将句子中获取的中心词与情感

词极性两两组合共同构成的句子的情感义原集合 $s=\{et_1,et_2,...,et_k\}$。如果句子 s_i 和 s_j 之间具有情感相似性，那么我们就认为句子之间存在情感义原的共现，以此来计算句子间情感相似度值：

$$E(s_i,s_j)=\frac{\left|\{et_k \mid et_k \in s_i \ \&\ et_k \in s_j\}\right|}{\log(|s_i|)+\log(|s_j|)} \quad (公式\ 9)$$

其中，分子为情感义原在句子 S_i 和句子 S_j 中的共现值，分母为句子中情感义原归一化数量值之和。

2.3.3 基于SE-TextRank的文本情感排序算法

TextRank 是用于自然语言处理的一种基于图排序的文本处理模型 [14]。传统文本情感摘要方法仅从句子之间情感词与共现词两个方面来计算句子间情感相似性，忽略了句子间语义情感关系对句子间情感相似度的影响。在本文中，我们将文本中句子的语义相似度和情感语义相似度分别进行图计算并将二者结果相融合，最终得到句子的情感排序结果集。图 1 为融合语义情感关系的 SE-TextRank 图模型。

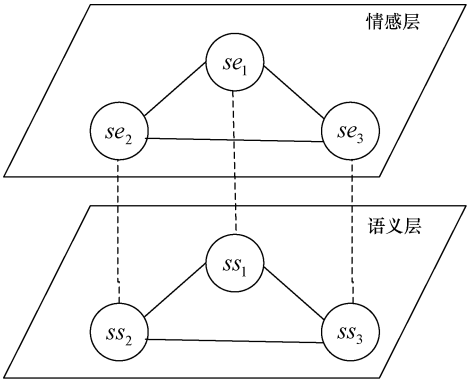

图1 融合语义情感关系的SE-TextRank图模型

其中，上层为情感层图模型：$G'=\langle v,e,w\rangle$，其中 v 为文本句子集合，e 为句子间的情感关系，w 为对应的情感相似度权值，se 为句子节点的情感权值。下层为语义层图模型：$G=\langle v,s,m\rangle$，其中 s 为句子间的语义关系，m 为对应的语义相似度权值，ss 为句子节点的语义权值。情感层与语义层相关联来计算文本句子情感排序。

基于 SE-TextRank 图模型，我们对传统的 TextRank 文本排序算法进行修改，继而得到改进后的 SE-TextRank 公式：

$$Q(v_i)=(1-d)+d\cdot\sum_{v_j\in In(v_i)}\left(\frac{\omega\cdot S_{ij}}{\sum_{v_k\in Out(v_j)}S_{jk}}+\frac{\lambda\cdot E_{ij}}{\sum_{v_k\in Out(v_j)}E_{jk}}\right)\cdot Q(v_j) \quad (公式\ 10)$$

其中，d 是规范化因子（一般取 0.85），ω 与 λ 分别为平分语义权值与平分情感权值的比例参数。S 是句子语义相似度权值，E 代表句子情感相似度权值。

3 实验设置与结果分析

本文所采用的实验语料为 NLPCC 2015 微博导向中文新闻文摘评测任务中的 250 篇新闻文本。在每个文本文件中，第一行是新闻标题（没有被选入摘要），其余为新闻文本内容。按评测要求进行人工摘要处理，并在每篇新闻文本中抽取 20 个关键词。对预处理结果进行一致性检查后，将最终得到的文本语料作为评测数据。我们使用 ROUGE-1.5.5 工具对摘要结果进行评测，采用 ROUGE-1、ROUGE-2、ROUGE-3 和 ROUGE-W 作为评价指标。

doi:10.3772/j.issn.2095-915x.2017.03.012

在试验中，我们通过设置多种方法对新闻文本进行基于多特征融合的摘要实验，对不同方法的权重大小进行综合分析。我们采用ROUGE-SU* 来作为评价主题数量占文本句子集比例对文本摘要影响的标准。通过预处理实验，得到主题数量占文本句子比例为 30% 时 ROUGE-SU* 值最优。不同主题数量比例下的 ROUGE-SU* 值变化如图 2 所示。

结合句子的位置权重、关键字权重以及长度权重等传统文本特征对文本摘要的影响，需要对 Topic-Sentence 关联分组集进行参数化分析。通过多次实验训练，我们采用的各个特征的加权参数 η、β、ε 分别为：0.4、0.4、0.2（$\eta+\beta+\varepsilon=1.0$）时，各项评测指标结果最优或接近最优。

使用 HowNet 情感词典获取句子中的情感词及其情感极性，同时利用哈工大的 LTP 工具对句子进行依存句法分析得到实词性词语。将获取的实词性词语与获取的主题词进行匹配，匹配得到的实词性主题特征词汇作为中心词与情感词极性组成 < 中心词，情感词极性 > 结构的情感义原。利用《知网》语义相似度相计算方法对主题句子分组中句子义原相似度权值进行计算。最后，通过 SE-TextRank 对情感义原与语义相似度进行训练，得到句子的最优语义特征值与最优情感特征值的加权参数 ω、λ 分别为 0.6、0.4（$\omega+\lambda=1.0$）。实验过程如表 1 所示。

表1 不同特征值参数下的摘要评测对比

ω	λ	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-W
0.1	0.9	0.29541	0.08121	0.01923	0.10149
0.2	0.8	0.32742	0.09382	0.02342	0.12415
0.3	0.7	0.34957	0.10207	0.03695	0.13342
0.4	0.6	0.38131	0.11828	0.05732	0.16592
0.5	0.5	0.41962	0.12031	0.05864	0.17139
0.6	**0.4**	**0.42109**	**0.12765**	**0.06121**	**0.17781**
0.7	0.3	0.41767	0.11471	0.05398	0.16485
0.8	0.2	0.40807	0.11056	0.04966	0.16037
0.9	0.1	0.36758	0.10466	0.03857	0.15854

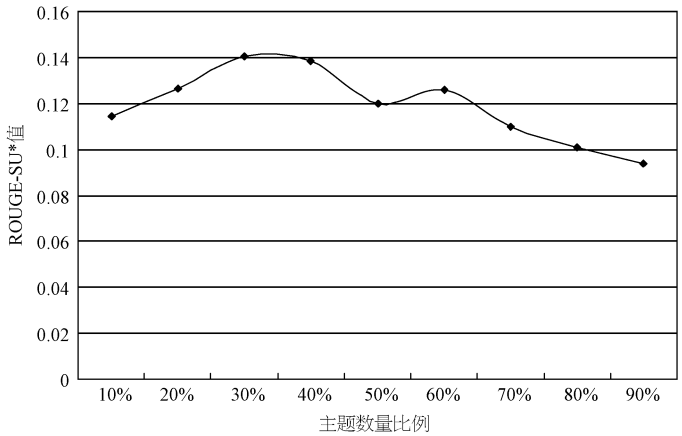

图2 不同主题比例下的ROUGE-SU*值

我们将传统的多特征文本摘要方法：采用传统的 TF-IDF 方法的多特征摘要方法和使用 PageRank 算法得到的摘要与采用本文的基于主题的 SE-TextRank 情感摘要方法进行比对，对比结果如表 2 所示。

表2 三种方法的摘要对比

Method	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-W
TF-IDF	0.37541	0.10211	0.03522	0.13497
PageRank	0.40742	0.12458	0.05654	0.16415
SE-TextRank	0.42109	0.12765	0.06121	0.17781

实验评测结果比对显示，本文所采用的一种基于主题的 SE-TextRank 情感摘要方法的各项指标都接近最好值，说明通过主题句子分组，采用 SE-TextRank 算法获取句子相似度特征，且加以传统多特征融合后获取的摘要切实可行。但各项指标值普遍不高，可能出现的原因主要有：(1) 在预处理过程中部分词语分词及去除无意义冗余时出现了误差；(2)LDA 模型获取主题的过程中出现了主题遗漏。(3) 句子情感极性获取时未考虑评价词、程度副词等对情感表征的影响，导致句子情感权值的计算出现误差；

4 结束语

本文提出了一种基于主题的 SE-TextRank 情感摘要方法， 充分考虑了新闻文本中主题、语义及情感三者间关联关系对摘要内容的影响。实验结果表明采用此方法的摘要结果更为稳定、高效，在方法上的可操作性也更强。接下来我们将在不同的领域及短评论等不同类型新闻文本进行摘要分析，同时将文章句法结构纳入文本情感摘要之中，使摘要方法的效率获得进一步的提高。

参考文献

[1] 尹存燕 , 戴新宇 , 陈家骏 . Internet 上文本的自动摘要技术 [J]. 计算机工程 , 2006, 32(3): 88-90.

[2] Luhn H P. The Automatic Creation of Literature Abstracts[J]. LBM Journal of Research & Development, 1958, 2(2): 159-165.

[3] Conroy J M, O'Leary D P. Text Summarization via Hidden Markov models[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2001: 406-407.

[4] Wang D, Li T. Weighted Consensus Multi-document Summarization[J]. Information Processing & Management, 2012, 48(3): 513-523.

[5] 莫鹏 , 胡珀 , 黄湘冀 , 等 . 基于超图的文本摘要与关键词协同抽取研究 [J]. 中文信息学报 , 2015, 29(6): 135-140.

[6] 林鸿飞 , 高仁璟 . 基于潜在语义索引的文本摘要方法 [J]. 大连理工大学学报 , 2001, 41(6): 744-748.

[7] 章芝青 . 基于语义的单文档自动摘要算法 [J]. 计算机应用 , 2010, 30(6): 1673-1675.

[8] 林萌 , 罗森林 , 贾丛飞 , 等 . 融合句义结构模型的微博话题摘要算法 [J]. 浙江大学学报工学版 , 2015,

doi:10.3772/j.issn.2095-915x.2017.03.012

49(12): 2316-2325.

[9] Hatzivassiloglou V, Mckeown K R. Predicting the Semantic Orientation of Adjectives[J]. Proceedings of the ACL, 2002: 174-181.

[10] Kamps J, Marx M, Mokken R J. Using WordNet to Measure Semantic Orientation of Adjectives. In: Calzolari N, et al., eds. Proc. Of the LREC. 2004. 1115-1118.

[11] 柳位平 , 朱艳辉 , 栗春亮 , 等 . 中文基础情感词词典构建方法研究 [J]. 计算机应用 , 2009, 29(10): 2875-2877.

[12] 林莉媛 , 王中卿 , 李寿山 , 等 . 基于 PageRank 的中文多文档文本情感摘要 [J]. 中文信息学报 , 2014, 28(2): 85-90.

[13] 董振东 , 董强 . 知网 [EB/OL]. [2008-10-10]. http://www.keenage.com.

[14] Mihalcea R, Tarau P. TextRank: Bringing Order into Texts[J]. Unt Scholarly Works, 2004: 404-411.