学科代码：520.2020

融合句义结构模型的微博话题摘要算法研究

林萌[1]，罗森林[1]，韩磊[1]，潘丽敏[1]

（1. 北京理工大学 信息与电子学院，北京 100081）

摘要：为了更快地从海量微博中获取话题的核心内容，本文提出一种融合句义结构模型的微博话题摘要方法。该方法首先利用句义结构模型抽取句子的语义格得到句子的语义特征，然后基于 LDA 主题模型使用句义结构计算句子两两之间的语义相似度，构建相似度矩阵，划分子主题类，得到句子的关联特征。最后，融合句子的语义特征和关联特征，选取子主题内信息量最大的句子作为摘要结果。在压缩比为 1.5%的条件下，ROUGE-1 值达到 51.30%，ROUGE-SU*达到 25.268%。实验结果表明，综合考虑句子语义特征和关联特征的句子权重计算方法能丰富句子的特征表示，融合句义结构模型能深化句子的语义分析层次，从而提升摘要与话题的相关度。

关键词：微博；话题摘要；句义结构模型；自然语言处理

中图分类号：TP391　　**文献标识码：**　　**文章编号：**

Research on microblog topics summarization merging sentential semantic structure model

LIN Meng[1], LUO Sen-lin[1], HAN Lei[1],PAN Li-min[1]

（1.*School of Information & Electronics, Beijing Institute of Technology, Beijing* 100081, *China* ）

Abstract: It is desirable to provide concise summarization to help users to quickly grasp the essence of topics. In this paper, a new microblog summarization framework based on sentential semantic structure model is proposed. Sentential semantic features are firstly extracted by sentential semantic structure model. Then LDA topic model fusing sentential semantic structure model is used to calculate the pairwise sentence similarities and construct the similarity matrix. Based on the similarity matrix, sentences are clustered into several subtopics and the sentential relationship features are obtained. At last, combining both sentential semantic features and relationship features, the most informative sentences are extracted from each subtopic. Experimental results demonstrate the improvement of our proposed framework, ROUGE-1 value is 51.30% while ROUGE-SU* value is 25.268% on compress ratio at 1.5%. The results indicate that introducing sentential semantic structure model can better understand sentential semantic and using both sentential semantic features and relationship features can also enrich the features representation.

Key words: microblog; topic summarization; sentential semantic structure model; natural language processing

微博的出现深刻改变了人们的信息交流方式。以新浪微博为例，截至 2014 年 6 月，微

收稿日期：2014-11-17

基金项目：国家 242 项目（2005C48），北京理工大学科技创新计划重大项目培育专项（2011CX01015）

作者简介：林萌（1991-），女，湖南，硕士研究生，从事中文信息处理的研究. E-mail:lemon0919@bit.edu.cn 通信联系人：罗森林(1968－)，男，教授，研究方向为中文信息处理、信息安全、数据挖掘、媒体计算等，E-mail：luosenlin@bit.edu.cn

博月均活跃用户数为 1.565 亿人，微博平均日活跃用户数为 6970 万人[2]。热门话题是微博中正在热议的新鲜话题，用户查询一个热门话题，得到的是按照热门程度或发表时间排序的所有相关微博。然而，由于微博数量庞大，用户得到的信息经常是不完整的，甚至是不相关的或者是重复的，信息获取的效率很低。因此对微博进行摘要能大大提高信息获取的效率。

多文档自动摘要技术其处理对象为结构完整书写规范条理清楚的长文本。微博篇幅短小（140 字以内）用词不规范，缺失长文档的结构信息，并包含大量垃圾内容和垃圾用户。直接利用已有的多文档摘要技术对其摘要存在严重的特征稀疏和结构缺失问题，这些问题导致抽取特征不足以准确描述文本内容，抽取的句子与话题的中心发生漂移，生成的摘要与主题相关度下降，大大影响摘要生成的效果[1]。所以解决微博话题摘要方法生成摘要与话题相关度差的问题具有十分重要的意义。

1 相关工作

多文档自动文摘技术经过多年发展已经出现了很多方法和技术。代表性的方法有基于词频的方法，比如 SumBasic[2]和 MEAD[3]；基于概率浅层语义分析（Probabilistic Latent Semantic Analysis，PLSA）[4]和浅层狄利克雷分布（Latent Dirichlet Allocation，LDA）[5]的方法；基于图的方法，例如 LexPageRank[6]算法，这种方法已经成功应用到了 Google PageRank 中；另外还有基于机器学习[7]的方法等等。

然而，这些方法大都是针对长文本的词项特征进行统计分析处理。微博篇幅短小，单条微博的关键词一般只有十几个甚至几个，关键词特征稀疏，并且单条微博内关键词的重复率不明显，缺失长文档的结构信息，因此传统自动摘要技术将失去原有效果，需要结合微博特性与传统自动摘要技术的优点来进行微博话题摘要。

近些年来，以 Twitter 为代表的英文社会化短文本摘要逐渐获得科研人员的关注，也取得了一些成果。2010 年 Sharifi[8]等提出将包含主题词的最常使用词汇链作为摘要，这种方法获得的摘要只是包含主题词的一句话，信息并不全面。2011 年，Harabagiu 和 Hickl [9]通过构建复杂事件的发展结构模型和用户行为模型来生成微博复杂事件的摘要。 Deepayan[10]使用隐马尔可夫模型学习微博事件的隐藏状态，对高度结构化重复出现的话题，如运动赛事进行摘要。Inouye 等[11] 在 Sharifi 研究的基础上提出一种基于聚类的 Hybird TF-IDF 摘要方法，这种方法计算词的 TF（Term Frequency）值是该词在语料库中出现的次数除以语料库中出现的词数，而计算 IDF（Inverse Document Frequency）值时，又将每篇微博作为一个单独的文档对待，计算方法为出现该词的微博总篇数除以语料中的微博总数。实验证明 Hybrid TF-IDF 取得的效果比一些主流的文摘方法 MEAD，LexRank[12],TextRank[13]要好。中文的微博摘要处于刚刚起步阶段，可以查阅的资料较少。2011 年，武汉大学的何炎祥[1]等提出一种轻巧新颖的 LN 算法，以树的形式将话题以摘要的方式展现给用户，但不能形成可读性文摘。2013 年，Bian[14]等引入微博文本的配图作为新的特征，提出一种新的概率生成模型 MMLDA 来发现微博话题的子主题并进行摘要。

目前的研究大多基于词形词频等统计信息进行特征抽取，忽略了句子的句义成分以及成分之间的关系特征，对微博内容挖掘深度不够，因此导致仅仅基于词形匹配的相似度计算方法无法准确计算句子的内容相似性；同时在选择句子时，没有抽取句子之间的隐藏语义联系，未充分利用句子所处的子主题信息，因而导致抽取的句子与主题的相关性较差。本文针对以上问题提出一种融合句义结构模型[15]的微博话题摘要方法。

[2] http://en.wikipedia.org/wiki/Sina_Weibo

2 句义结构模型及句义分析

句义结构模型以现代汉语语义学为基础，从句义角度研究句子的句义成分以及成分之间关系的句义结构化表示模型，将抽象的句义表示成计算机可处理的结构化数据。模型将句义结构分为句型层、描述层、对象层和细节层四个层次，包含的句义成分有句义类型、话题、述题、谓词和项等。句义成分中的项又分为基本项与一般项，其功能用语义格表示，对应的语义格又分为 7 个基本格和 12 个一般格。模型的基本形式[16]如图 1 所示。

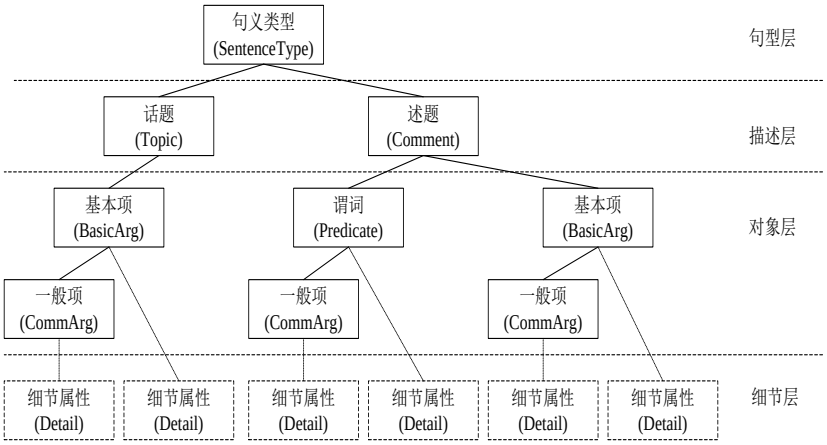


图 1 句义结构模型的基本形式

Fig.1 Basic form of sentential semantic model

句义分析是通过句义结构模型分析句子结构信息和语义信息，抽取能够表述句子语义的特征，这些特征能够表达人物实体重要信息是文本强特征。句义分析的具体方法是根据句义结构模型基本框架，分别处理不同语义格的对象成分以及语义格结构信息，主要语义格类型说明如表 1 所示。

表 1 主要语义格类型说明

Tab.1 Description of Semantic case

	类型	说明
基本格	施事格	变化、动作、行为的发起者
	受事格	变化、动作、行为的承受者
	遭遇格	变化、动作等的非自主发起者
	结果格	变化、动作等所产生的结果
	主事格	谓词所描述性质、状态等的对象
	说明格	协同谓词对主事格作出说明
一般格	属格	表示一个对象为另一个对象所有
	描写格	起修饰作用的项
	空间格	行为等出现、发生的处所等
	范围格	动作、行为等的领域范围
	根由格	动作等的依据、原因、目的
	同位格	与被其修饰项指同一对象
	基准格	动作、性质等的参照标准
	时间格	行为等发生或持续的时间

3 算法原理

针对现有方法生成摘要内容冗余度高的问题，本文从准确计算句子内容相似性的角度出发，利用句义结构模型分析语义项和项之间的依存关系抽取句子的句义特征，扩充句子的语义维度，利用句义特征准确表达语句信息及句子内容的相似性，抽取句子时根据句子内容相似性有效控制文摘冗余度。针对现有摘要方法抽取的句子与子主题相关性差的问题，本文从挖掘句子之间的隐藏语义联系及子主题信息的角度出发，提出一种抽取句子关联特征的方法。关联特征表示句子与话题的语义联系度，利用关联特征增强相似语句的语义联系。最后综合加权句子的语义特征和关联特征，抽取子主题内的关键句子，得到话题的摘要。

本方法首先将文档集合分句，句子清洗，分词和词性标注得到预处理结果，然后对预处理结果分别计算语义权值和关联权值。计算语义权值时，统计预处理结果中所有实词出现的句子频率，按句子频率从大到小排序，选择前 N 个词作为主题词的种子词，加入哈工大同义词林扩展版（HIT IR-Lab Tongyici Cilin (Extended)）进行扩展，得到扩展后的主题词表。结合主题词表，分析句子的语义特征，包括词性词法特征，以及句义结构特征，对各个特征线性加权，得到句子的语义权值。计算关联权值时，需要先对预处理结果进行句子相似度计算，得到句子两两之间的语义相似值，构建相似度矩阵，划分子主题类。利用句子两两之间的语义相似值，计算句子与类内及类外其他句子的语义相似度，得到句子的关联权值。最后对句子的语义权值和关联权值综合加权，得到句子的最终权值。最后依次选择子主题内权重最大的句子作为文摘句。算法原理图如图 2 所示。

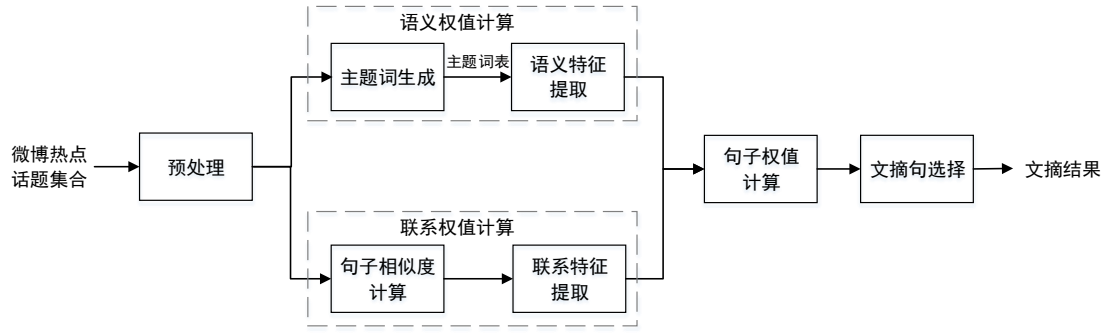
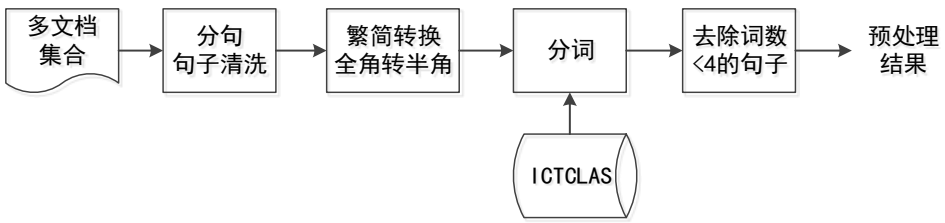


图 2 算法原理图

Fig.2 Schematic Diagram

3.1 预处理

分句、句子清洗是预处理的第一步。将微博中内嵌链接 URL、表情符号和@后的用户名从原始语料库中删除。之后采用中科院提供的中文分词软件“ICTCLAS”[3]按照北京大学词性标注规范对数据集进行分词。将有效词数（名词、动词、形容词、数词、时间词等实词）小于 4 的句子去除。最后，对分词后的结果去除中英文停顿词（stopwords）。预处理原理图如图 3 所示。



[3] http://ictclas.cn/index.html

图 3 预处理原理图

Fig.3 Schematic Diagram of Preprocessing Module

3.2 语义权值计算

3.2.1 主题词生成

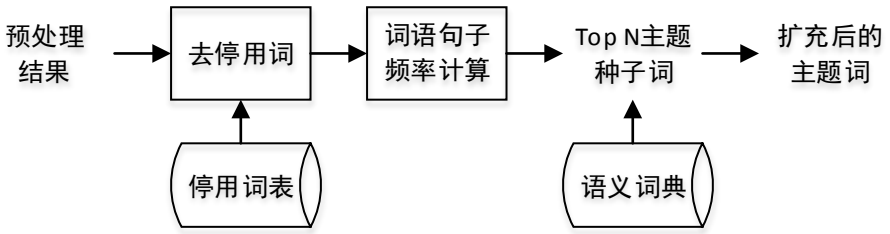


图 4 主题词生成原理图

Fig.4 Schematic Diagram of Topic Generation Module

主题词生成原理图如图 4 所示。将所有实词按句子频率从大到小排序，选择前 N 个词作为主题词的种子词，加入哈工大同义词林扩展版进行扩展，得到扩展后的种子词。

扩展版的同义词林包含 77,343 条词语，按照五层树形结构组织到一起。有研究证明[4]，对词义进行有效扩展， 或者对关键词做同义词替换可以明显改善信息检索、 文本分类和自动问答系统的性能。本文利用第 5 级分类对种子词进行扩展，分别按照词义相等，词义相关两大类别扩展种子词。

3.2.2 语义特征提取

分析句子的语义特征，是计算句子内容重要性的关键步骤。现有研究一般只用到句子的词法和句法特征，对句子内容的挖掘仅限于词、句法层次。本文不仅使用了传统的词法句法特征，还加入了句子的句义结构特征。句义结构特征增加了句子的分析深度，能够更好的表达句子的深层含义，对句子内容的挖掘更有效。

句义结构模型是对句子语义层次的分析，是句义的形式化表达。句义结构模型中的话题、谓词、述题等信息可以体现一个句子的核心内容，此外句义结构模型中各个句义成分之间的关系对句子的语义表达也很有意义。本文使用的语义特征如表 2 所示。

编号 1,2 的特征项分析句子有效词的统计特征。一般认为名词（noun）、动词（verb）比其他词性更重要，赋权重为 2，其余词性权重为 1。话题、谓词、述题特征是句子的核心内容，若该句的以上特征在主题词表内，则说明该句的核心内容跟主题相关，出现的词数越多则该句与主题联系越紧密，越能表达主题中心的意义。一般项的句义功能是描述基本项和谓词，对其表达的内容做进一步说明和补充。因此，本文也将句子一般格中包含的主题词选为特征，作为对一般项和谓词的补充。句子的语义权重值计算方法如式 1 所示。

$$p_{con}(S_k)=\sum_{i=1}^{6}\mu_i * F_i \tag{1}$$

$p_{con}(S_k)$ 是句子 S_k 的语义权重值， F_i 和 μ_i 分别代表语义特征的值和该特征的加权系数。

[4] http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm

表 2 句子语义特征

Tab.2 Semantic Features of Sentences

特征代码	特征说明
F_TFIDFPOS	句子有效词的平均 TFIDF*POS 词权重
F_KEYWORDS	句子中的主题词在所有主题词中的覆盖率
F_TOPB	句子话题基本格包含主题词的个数/句子有效词数
F_PRED	句子谓词包含主题词的个数/句子有效词数
F_COMB	句子述题基本格包含主题词的个数/句子有效词数
F_COMM	句子一般格包含主题词的个数/句子有效词数

3.3 联系权值计算

3.3.1 句子相似度计算

由于句子长度的限制，单个句子的关键词一般只有几个，特征尤其稀疏，仅仅基于词形匹配的方法无法准确衡量句子内容的相似度。本文在句义结构的基础上，使用 LDA 主题模型，对单个句子的关键词进行扩充，从而解决由于句子长度限制特征严重缺失所带来的无法计算句子相似度的问题，并在句义层面计算句子的内容相似度。

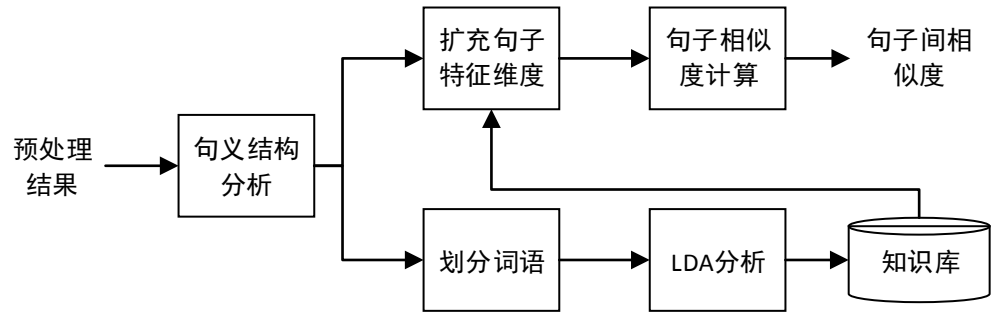


图 5 句子相似度计算原理图

Fig.5 Schematic Diagram of Sentence Similarity Module

句子相似度计算的原理图如图 5 所示。输入是预处理后的所有句子，输出是句子两两之间的相似值。其中，句义结构分析模块利用 BFS-CSA[5]分析句子得到句义结构；划分词语模块是根据句义结构中的成分，将词语划分成基本格、一般格和谓词；LDA 分析模块对划分好的语义格计算得到知识库；扩充句子维度模块使用知识库的信息对句子中的格进行扩充得到新的表示向量；句子相似度计算模块对扩充后的句子向量计算余弦相似度，得到两个句子间的相似值。

根据句义结构理论，句义包括话题和述题，话题是被描述的成分，述题是语义表达的描述成分，同时考虑到句子的主干（基本格）和句子的修饰成分（一般格），我们将知识库分为三类：话题（基本格）知识库、述题（基本格和谓词）知识库和一般格知识库。话题知识库中的词语来源于文集中句子话题下的基本格，用于对句子中话题下的基本格词语进行扩充，述题知识库中的词语来源于文本集中句子述题下的基本格和谓词，用于对述题下的基本格词语和谓词进行扩充，一般格知识库中的词语来源于句子中的一般格，用于对句子中所有一般格词语进行扩充。

按照 Blei[17]的理论，LDA 主题模型计算得到同一 Topic 下的词语具有相似的属性或意

[5] http://www.isclab.org/csa/bfs-csa.php

义，因此本文利用 LDA 对三组不同的词语集合分别计算不同 Topic 下的概率，最后将句子中话题（基本格）、述题（基本格和谓词）和一般格下的词语分别选择对应知识库所在 Topic 下的其他词语作为特征向量上该词的维度扩充，扩充维度的取值计算公式如下。

$$V = n * w / w_c \tag{2}$$

V 是扩充词语的取值，n 是待扩充词在句子中出现的次数，w 是待扩充词在相应 Topic 下的概率取值，w_c 是扩充词语在相应 Topic 下的概率取值。

对句子的话题和述题分别进行扩充，得到句子的话题向量和述题向量，分别计算句子的话题相似度和述题相似度，对两个相似度进行加权得到最终的句子相似度，计算公式如式 3 所示。

$$S(S_A, S_B) = \omega * \frac{\overrightarrow{S_{At}} \cdot \overrightarrow{S_{Bt}}}{|\overrightarrow{S_{At}}||\overrightarrow{S_{Bt}}|} + (1-\omega) * \frac{\overrightarrow{S_{Ac}} \cdot \overrightarrow{S_{Bc}}}{|\overrightarrow{S_{Ac}}||\overrightarrow{S_{Bc}}|} \tag{3}$$

S_A 和 S_B 代表语料库中的任意两个句子，$\overrightarrow{S_{At}}$ 和 $\overrightarrow{S_{Bt}}$ 表示扩充后的句子话题向量，$\overrightarrow{S_{Ac}}$ 和 $\overrightarrow{S_{Bc}}$ 表示扩充后的句子述题向量，参数 ω 调整话题和述题的权重。

以 0.1 为步进值调整 ω，得到实验结果如表 3 所示。

表 3 参数 ω 实验结果

Tab.3 Parameter Selection Experiments of ω

ω	ROUGE-1	ROUGE-2	ROUGE-w	ROUGE-SU*
0	0.4825	0.1837	0.1373	0.2250
0. 1	0.4977	0.1771	0.1438	0.2389
0. 2	0.5018	0.2243	0.1398	0.2458
0. 3	0.4521	0.1516	0.1225	0.1937
0. 4	0.5387	0.2683	0.1616	0.2655
0. 5	0.5457	0.2548	0.1628	0.2780
0. 6	0.4764	0.1447	0.1281	0.2196
0. 7	0.5206	0.1968	0.1519	0.2575
0. 8	0.4994	0.1534	0.1345	0.2390
0. 9	0.4933	0.1644	0.1357	0.2376
1	0.5067	0.1533	0.1355	0.2455

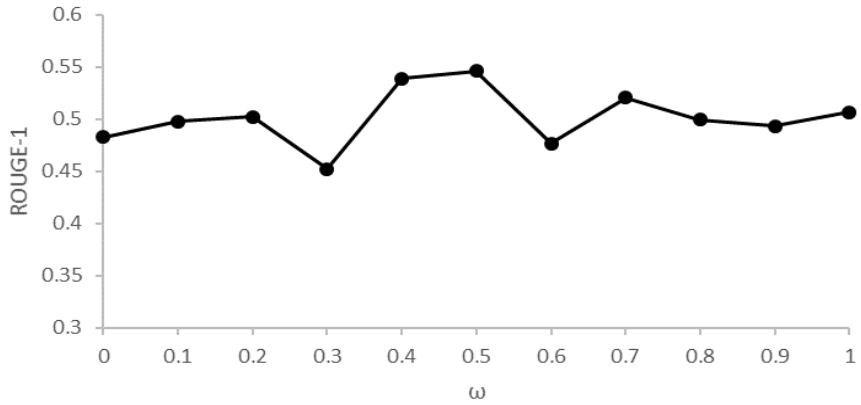


图 6 ROUGE-1 值随ω的变化趋势图

Fig.6 Tendency Gragh of ROUGE-1

由表 3 可得，当ω为 0.5 时，即话题和述题的权重相等时，ROUGE 评价指标得分最高。以 ROUGE-1 指标为例，观察图 6 当ω从 0.5 开始向 0（1）方向减小（增大）时，ROUGE-1 指标为逐步减小的趋势，说明话题和述题是综合衡量句子意义的两个方面，偏向于任何一方，句子相似度的计算值都不能完全表达句子的意义。由实验结果可知，本文ω最佳取值为 0.5。

3.3.2 关联特征提取

通过句子相似度计算出句子两两之间的语义相似值，构建句子的 n 维空间向量表示，形式如式 4 所示，空间中的每一维$w_{k,j}$是句子S_k对S_j的相似度值。

$$V(S_k)=\{w_{k,1},w_{k,2},...,w_{k,n}\} \tag{4}$$

子主题是围绕中心主题发生的现象、后果以及原因等的说明，是对中心主题不同侧面的描述。利用构建的句子特征空间对语料中所有的句子进行 kmeans 聚类，划分子主题。对于本文所使用的语料库，每一个话题下的子主题数目一般不多于 10 个。因此，设定初始聚类中心为 10，并将类内句子数量小于总量 5%的类作为噪音去除，剩余的类作为子主题划分结果。

句子的关联特征表示句子与话题的语义联系度，通过加权计算该句与不同子主题中其他句子的语义重合度得出。句子S_k对S_j的语义重合度$C(S_k,S_j)$定义为句子S_j的语义权重值$P_{con}(S_j)$与S_j对S_k的句子相似度值$S(S_k,S_j)$的乘积，如公式 5 所示。

$$C(S_k,S_j)=P_{con}(S_j)*S(S_k,S_j) \tag{5}$$

构建无向图 *G(S, E)*，图中的每个节点 *S* 对应一个语句，边$E(S_i,S_k)$表示语句S_i与S_k的句子相似度值。节点 S 的度 d 是与 S 相连的边的数目，反映了 S 包含信息的重要程度: d 越大，则对应语句所关联的语句数目越多，那么这个句子所包含的信息越重要；反之亦成立。另一方面，如果一个节点的度比较大，那么与之相关联的语句也相应地比较重要。令节点 S 的初始值为句子的内容权重值，通过计算其他句子对该句的语义重合度得到句子的联系权重值。考虑到同一个子主题下句子联系紧密，设加权系数为 1，不同子主题下句子的加权系数由子主题的平均句子内容权重得出。计算公式如式 6 所示。

$$P_{rel}(S_k)=\sum_{i=1}^{m}S(S_k,S_i)*P_{con}(S_i)+\sum_{j=m+1}^{n}\frac{C_j}{C_j+C_k}S(S_k,S_j)*P_{con}(S_j) \tag{6}$$

$P_{rel}(S_k)$是句子S_k的关联权重值，$S(S_k,S_i)$是句子S_k对S_i的句子相似度值，$P_{con}(S_i)$是句子S_i的语义权重值。若S_k和S_i属于同一个子主题，加权系数为 1；若S_k和S_j分属不同子主题，加权系数为$\frac{C_j}{C_j+C_k}$，其中C_j，C_k分别代表句子所属子主题的句子平均语义权重值，

m 表示句子 S_k 所属子主题的句子个数，*n* 表示语料库中所有句子的个数。

3.4 句子权值计算

现有方法计算句子重要性大部分都仅仅偏重于挖掘句子本身的内容，而忽略了句子所处“环境”的影响。一个好的文摘句，内容上不仅要紧扣主题，同时也应该与语料库中的其他句子联系紧密。本文所使用的句子权重计算方法不仅考虑了句子的语义信息同时也考虑了句子的关联特征。句子权值的计算公式如下：

$$P(S_i)=\alpha * P_{con}(S_i)+\beta * P_{rel}(S_i) \qquad \alpha+\beta=1 \tag{7}$$

$P(S_i)$ 是句子 S_i 的最终权值，$P_{con}(S_i)$ 是句子 S_i 的语义权值，$P_{rel}(S_i)$ 是句子 S_i 的联系权值，参数 α 调整语义权值和联系权值的权重。为了得到选择参数 α 的最佳取值，α 从 0 开始以 0.1 为步进变化到 1，得到压缩比为 1.5%时，ROUGE-1 的取值变化如图 7 所示。

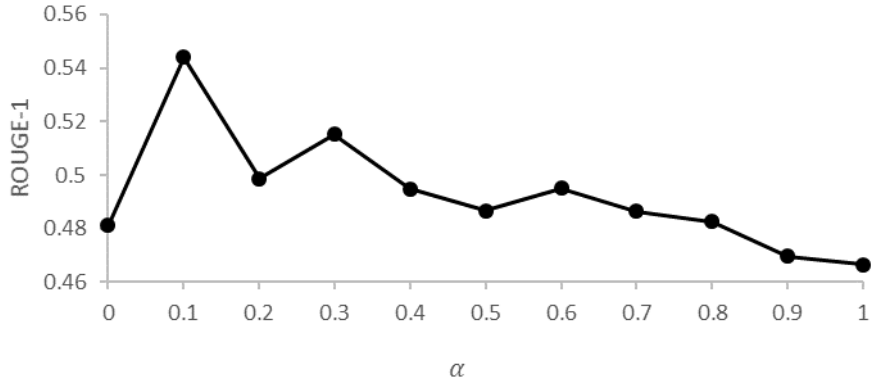


图 7 参数 α 选择实验

Fig.7 Parameter Selection Experiments of α

由图 7 可知，当 α=0.1，β=0.9 时 ROUGE-1 的取值最高。

3.5 文摘句选择

句子选择模块首先根据子主题的重要程度从高到低对子主题排序，确定子主题的抽取顺序和抽取句子数。然后根据句子的重要性和冗余度在子主题内抽取文摘句。子主题的重要性与两方面因素有关：子主题包含的句子数目，句子数目越多说明该子主题在文档集合中出现的频率越高；子主题包含句子的重要程度，子主题中平均句子权重越大则表示该子主题越重要。子主题打分策略如式 8 所示。

$$S(C_i)=\theta\frac{P_{ave}(C_i)}{\sum_{j=1}^{k}P_{ave}(C_j)}+(1-\theta)\frac{N_i}{\sum_{j=1}^{k}N_j} \tag{8}$$

其中 $S(C_i)$ 表示子主题 C_i 的得分，$P_{ave}(C_i)$ 表示子主题 C_i 的句子平均权值，*k* 表示子主题个数，N_i 表示子主题 C_i 包含的句子个数。参数 θ 调整子主题内句子平均权值和句子数目

的权重，一般认为两者同样重要，本文θ取 0.5。

根据压缩比，从不同子主题内抽取相应数量的句子生成摘要。子主题句子抽取个数由该子主题的重要程度决定，计算公式如式 9 所示。

$$T_{C_i} = \frac{S(C_i)}{\sum_{j=1}^{k} S(C_j)} * R * \sum_{j=1}^{k} N_j \tag{9}$$

其中T_{C_i}表示子主题C_i的句子抽取个数，$S(C_i)$表示子主题C_i的得分，k表示子主题个数，R代表压缩比的值，N_j表示不同子主题内句子的数目。

选择文摘句时，选择的句子不仅要与主题的相关度高，也要保证该句与已选文摘句之间的冗余度尽可能小，从而避免包含同一条重要信息的句子反复出现在文摘里。句子选择的具体过程如下。

输入：原始句子、句子权值、句子相似值及子主题句子抽取个数

输出：文摘结果

定义：k = 子主题个数

T_{c_i} = 子主题C_i的句子抽取个数

$P(S_j)$ = 句子S_j的权值

$S_{ave}(S_j, S_{doc})$ = 句子S_j与已选文摘句的平均句子相似度

Begin:

for i = 0 to i < k

  for j = 0 to j < T_{c_i}

    选择子主题内权重最大的句子

    for 子主题内剩余每条候选句子

      计算候选句子与已选文摘句的平均句子相似度值，更新句子权重

      $P(S_j) := P(S_j) * (1 - S_{ave}(S_j, S_{doc}))$

    end for

  end for

end for

4 实验及分析

4.1 数据源

实验数据采用 NLP&&CC 会议 2013 年中文微博观点要素抽取评测语料[6]。该语料包含 2013 年 3 月的微博话题，类别样本不平衡，实验数据的具体描述如表 4 所示。其中，文本有效长度是指经过分词去除停用词后，每篇微博包含的词的个数。由北京理工大学信息系统

[6] http://tcci.ccf.org.cn/conference/2013/pages/page04_evares.html

及安全对抗实验中心的博士、硕士对每个话题生成压缩比为 0.5%，1%，1.5%的三篇标准摘要。其生成过程如下：首先每 3 人对同一话题文本集提取不同压缩比的人工摘要，然后由自然语言处理小组的 10 名博士、硕士对 3 份人工摘要进行评价并计算平均得分，将平均得分最高的摘要作为标准摘要放入标准摘要集，如果得分相同则都放入标准摘要集中。

表 4 实验数据

Tab.4 Experimental Data

类别	文本数	文本有效长度			
		Max	Min	Mean	SD
毒玩具	960	89	5	31	±4.05
查韦斯	1000	87	2	26	±3.83
曼联 vs 皇马	1000	79	9	28	±3.78
王语嫣	1000	66	4	19	±3.29
锤子 ROM	1990	80	5	29	±3.88
中国方言式英语	2470	81	6	33	±4.25

4.2 评价方法

本文采用多文档摘要的通用评价方法 ROUGE[18] toolkit（v1.5.5）作为评价标准。ROUGE 方法通过计算候选摘要与标准摘要的词单元重合度来区分候选摘要的质量，计算的值包括 ROUGE-N，ROUGE-W（本文 w 取 1.2）和 ROUGE-SU 等。具体计算公式如下：

$$R_{ROUGE-N}=\frac{\sum_{S\in\{ref\}}\sum_{gram_n\in S}Count_{match}(gram_n)}{\sum_{S\in\{ref\}}\sum_{gram_n\in S}Count(gram_n)} \tag{10}$$

$$R_{ROUGE-N}=\frac{\sum_{S\in\{candi\}}\sum_{gram_n\in S}Count_{match}(gram_n)}{\sum_{S\in\{candi\}}\sum_{gram_n\in S}Count(gram_n)} \tag{11}$$

$$F_{ROUGE-N}=\frac{2*P_{ROUGE-N}R_{ROUGE-N}}{P_{ROUGE-N}+R_{ROUGE-N}} \tag{12}$$

其中 n 代表 *n-gram* 的长度，S 表示文档，其中下标 *ref* 表示文档属于标准摘要，下标 *candi* 表示文档属于待评价摘要，$Count_{match}(gram_n)$ 表示同时出现在待评价摘要和标准摘要的 *n-gram* 的个数，$Count(gram_n)$ 为标准文摘中的 *n-gram* 个数。

4.3 实验结果及分析

4.3.1 关联特征验证实验

为了验证引入句子关联特征对摘要结果的提升，在压缩比为 1.5%的条件下，采用单因变量法，令 $P_{con}(S_i)$ 加权系数 α 为 0.1 保持不变，从 0 开始以 0.05 为步进调整 $P_{rel}(S_i)$ 加权

系数β的值，以加权系数的比值$\frac{\beta}{\alpha+\beta}$为横坐标，得到 ROUGE-1 的取值变化如图 8 所示。

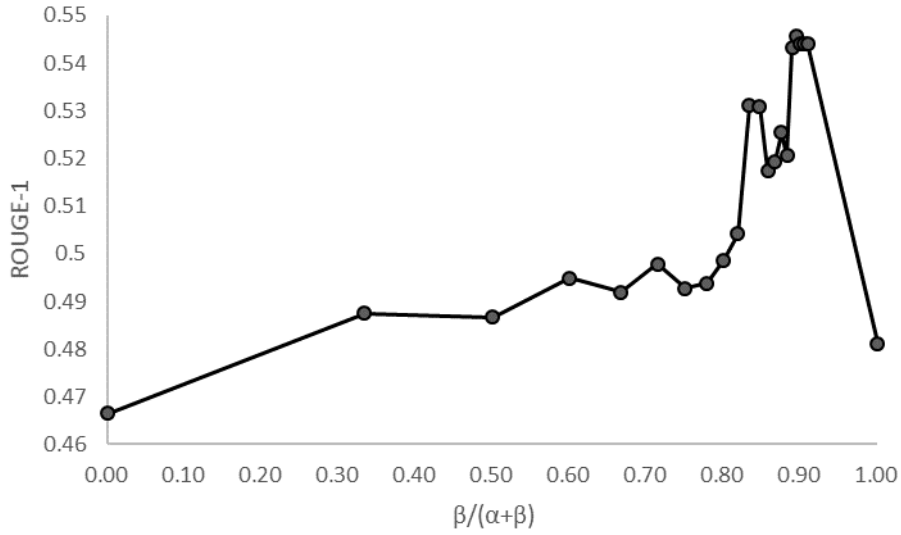


图 8 ROUGE-1 值变化趋势图

Fig.8 Tendency Gragh of ROUGE-1

由图 8 所示可知，当$\frac{\beta}{\alpha+\beta}$=0 时，即不考虑句子的关联权重，只考虑句子本身的语义权重，ROUGE-1 为 0.46658，加入句子关联权重特征，ROUGE-1 值有明显的改善。随着β值的增加，在比值小于 0.9 的条件下，ROUGE-1 值呈现上升趋势，且均明显优于仅考虑语义权重的 ROUGE-1 取值，比值大于 0.9 后，继续增加关联权重的值，ROUGE-1 值为下降趋势，当$\frac{\beta}{\alpha+\beta}$=1，即关联权重占比远大于语义权重时，ROUGE-1 取值为 0.48125。实验结果说明本文提出的加入句子关联权重特征的句子权值计算方法，在深入理解句子本身语义的基础上，有效地量化了该句与语料库中其他句子之间的语义联系。综合考虑句子内外部特征的权值计算方法，丰富了句子的特征维度，准确描述了文本内容与话题的相关度，合理利用了句子内外部语义特征，使同类数据内聚性增强，噪音影响减弱，对于选择关键文摘句以及减少文摘的冗余度都很有意义。

4.3.2 对比实验

为了验证本文方法的有效性，建立了两个对照方法与本文方法进行对比实验。

Hybird TF-IDF[11]是 Inouye 于 2011 年提出的一种基于聚类的微博话题摘要方法，并已被证明比一些主流的多文档摘要方法效果要好。SumBasic[2]是经典的多文档摘要方法，在 DUC06 测评大会上按代表性指标排序排名第三，并已取得实用。在压缩比分别为 0.5%，1%，1.5%的条件下，三组系统的 F 值实验结果如表 5、表 6、表 7 所示。

表 5 压缩比为 0.5%的对比实验结果

Tab.5 Contrast Experiments Results with Compress Ratio at 0.5%

0.5%	ROUGE-1	ROUGE-2	ROUGE-W	ROUGE-SU*
Hybird TF-IDF	0.31451	0.09527	0.09659	0.08696
SumBasic	0.32510	0.10937	0.10179	0.10163
本文方法	0.42893	0.14419	0.14415	0.17585

表 6 压缩比为 1%的对比实验结果

Tab.6 Contrast Experiments Results with Compress Ratio at 1%

1.00%	ROUGE-1	ROUGE-2	ROUGE-W	ROUGE-SU*
Hybird TF-IDF	0.38075	0.12438	0.10649	0.14109
SumBasic	0.39127	0.12510	0.10738	0.14305
本文方法	0.47986	0.16883	0.14433	0.22125

表 7 压缩比为 1.5%的对比实验结果

Tab.7 Contrast Experiments Results with Compress Ratio at 1.5%

1.50%	ROUGE-1	ROUGE-2	ROUGE-W	ROUGE-SU*
Hybird TF-IDF	0.43845	0.15848	0.11761	0.18786
SumBasic	0.41575	0.14456	0.10942	0.15968
本文方法	0.51303	0.18820	0.14701	0.25268

由表 5 至表 7 可知，本文提出的微博话题摘要方法在 ROUGE-1，ROUGE-2，ROUGE-W，ROUGE-SU*的评价指标下平均表现最优，四个指标的值均有明显提高。相比于 Hybrid TF-IDF，SumBasic 等基于词形词频的短文本摘要方法，本文提出的融合句义结构模型的微博话题摘要方法生成的摘要在兼顾冗余度的同时与话题更相关，综合表现 F 值更高。这表明分析句子的句义结构，提取句义特征项和项之间的依存关系可以深入挖掘句子的语义信息，深化了句子分析层次，提取的句义特征增强了语义特征的表达能力，有效避免了信息丢失；构建相似度矩阵划分子主题的方法使类内语义相关性增大，同类数据内聚性增强，有效降低了噪声的影响；综合考虑句子内部语义特征和外部关联特征的句子权重计算方法，不仅丰富了句子的特征表示，更全面的考虑句子的语义环境，从而提升了摘要与话题的相关度。

同时，对比压缩比为 0.5%，1%，1.5%的摘要结果，在一定范围内压缩比越大系统的性能越好，原因在于人工抽取标准摘要的随机性比较大，而压缩比提高、数据量变大在一定程度上克服了这种随机性，使得最终得到的摘要更加合理而使评价效果有所提高。

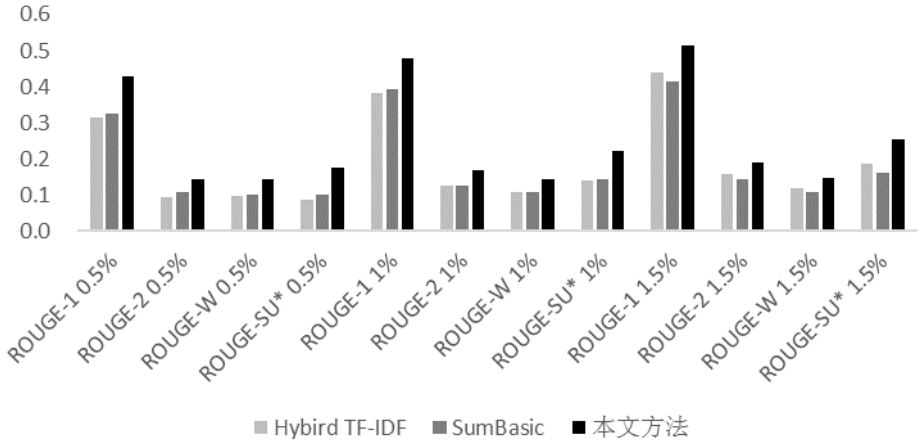


图 9 对比实验结果柱状图

Fig.9 Histogram of Contrast Experiments Results

4.3.3 泛化能力实验

对语料中的 6 个话题进行摘要实验，在压缩比 0.5%，1%，1.5%的条件下计算系统的 ROUGE 评价指标，因篇幅所限，图 10 仅展示压缩比为 1.5%的实验结果。

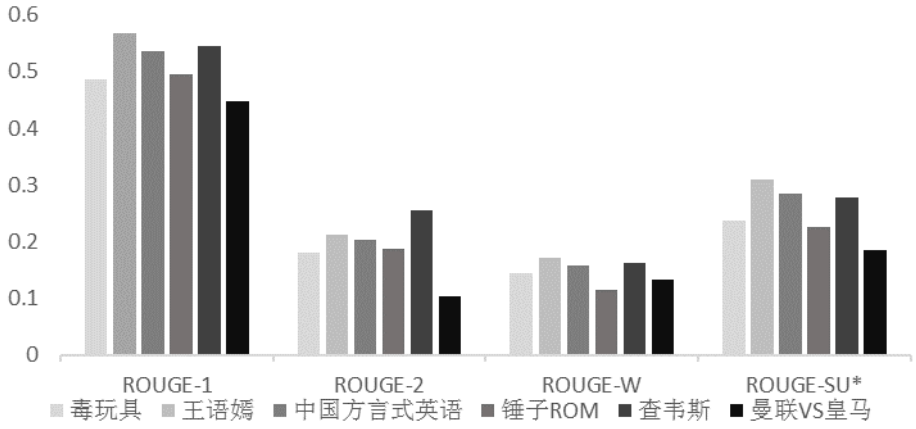


图 10 系统泛化能力实验结果

Fig.10 System Performance on Different Topics

由图 10 可知，系统在不同话题下的评价结果会有一定的差异,一方面是因为人工抽取标准摘要的随机性，另一方面是因为不同话题子主题的结构不同。由 ROUGE 评价指标来看，六个话题的 ROUGE-1 值均在 0.45 以上，ROUGE-2、ROUGE-W 均在 0.1 以上，ROUGE-SU* 均在 0.15 以上，因此本文方法处理不同话题的泛化能力较优，适用广泛。

4.3.4 实例分析

以话题“查韦斯”为例，采用 Hybird TF-IDF 方法和本文方法分别进行摘要实验，在压缩比为 5%的条件下得到摘要结果如表 8 所示。

表 8 话题“查韦斯”摘要结果展示

Tab.8 Generated summaries from topic “Chávez”

Hybrid TF-IDF	本文方法
2013 年 3 月 5 日，委内瑞拉副总统马杜罗宣布，总统查韦斯当天逝世。委内瑞拉总统查韦斯于当地时间 5 日下午逝世，享年 58 岁。 委内瑞拉总统查韦斯患癌病逝终年 58 岁。委内瑞拉政府确认总统查韦斯去世。沉痛哀悼：伟大的反美斗士，伟大的中国人民的朋友－－委内瑞拉查韦斯总统逝世。 从此，世界又少了一位反美斗士。网易新闻“反美斗士”一路走好。中国人民的老朋友查韦斯，遗憾的是世界又少了一位反美强人。	国际新闻：委内瑞拉总统查韦斯患癌病逝终年 58 岁当地时间 5 日，委内瑞拉副总统马杜罗发表电视讲话，称委总统查韦斯在与癌症抗争两年后去世，结束了对委内瑞拉 14 年的执政。 对作为病人的查韦斯之亡故表示同情，但对作为邪恶统治者的查韦斯之死掉表示快意。 查韦斯没有安排好继承人离去，其实知道自己癌症后，马上安排马杜罗参选总统，让自己的思想传承下去，自己在台下做个旁观者不更伟大，已经连任四届总统，获得权利其乐无穷！

白俄罗斯总统新闻局公告中说，由于委内瑞拉总统查韦斯去世，白俄罗斯宣布哀悼三天。委内瑞拉总统查韦斯逝世欧盟首脑致电表哀悼。 长春偷车犯掐死婴儿市民悼念；委内瑞拉政府确认查韦斯去世。	美国人想让萨达姆死萨达姆死了，美国人想让本拉登死本拉登死了，美国人想让卡扎菲死卡扎菲死了，然后美国人想让查韦斯死。 查韦斯的去世不禁让我们发问，食品安全到底是什么？

由表 8 可知，对比两种方法，Hybird TF-IDF 生成的摘要包含子主题较少，内容比较片面，摘要的冗余度也较大。本文方法生成的摘要覆盖了话题的多个子主题，内容较全面，摘要冗余度较小，因而本文方法在语义上生成的摘要效果更优。

5 结论

本文提出一种融合句义结构模型的微博话题摘要方法。本方法首先利用句义结构模型分析句子的语义特征得到语义权值，然后基于 LDA 主题模型使用句义结构计算句子内容相似值，构建相似度矩阵划分子主题，将子主题信息及句子相似度值加权得到关联特征。最后综合加权句子内部语义特征和外部关联特征抽取类内最重要的句子生成摘要。实验证明利用句义结构模型深化了句子分析层次，提取的句义特征增强了语义特征的表达能力，有效避免信息丢失，同时综合加权句子内部语义特征和外部关联特征的句子权重计算方法使同类数据内聚性增强，语义相关性增大，有效降低了噪声的影响，从而使得生成的摘要与话题相关度更高。此外，本文方法处理不同话题的泛化能力较优，适用广泛。

下一步研究的重点是引入句子结构项之间的依存关系作为特征，完善句义结构模型的特征体系，提高文摘句抽取效果，从而生成更高质量的微博话题摘要。

参考文献（References）

[1] He Y, Suwen, Tian Y, et al. Summarizing Microblogs on Network Hot Topics[C]. Piscataway, NJ, USA: IEEE, 2011.

[2] Vanderwende L, Suzuki H, Brockett C, et al. Beyond SumBasic: task-focused summarization with sentence simplification and lexical expansion[J]. Information Processing and Management. 2007, 43(6): 1606-1618.

[3] Radev D R, Jing H, Stys M, et al. Centroid-based summarization of multiple documents[J]. Information Processing and Management. 2004, 40(6): 919-938.

[4] Singh M, Khan F U. Effect of Incremental EM on Document Summarization using Probabilistic Latent Semantic Analysis[C]. Hong Kong, China: Newswood Limited, 2012.

[5] Gao D, Li W, Ouyang Y, et al. LDA-based topic formation and topic-sentence reinforcement for graph-based multi-document summarization[C]. Tianjin, China: Springer Verlag, 2012.

[6] Binti Zahri N A H, Fukumoto F, Matsuyoshi S. Link Analysis Based on Rhetorical Relations for Multi-Document Summarization[J]. IEICE Transactions on Information and Systems. 2013, E96-D(5): 1182-1191.

[7] Sujatha C, Chivate A R, Ganihar S A, et al. Time driven video summarization using GMM[C]. Piscataway, NJ, USA: IEEE, 2013.

[8] Sharifi B, Hutton M, Kalita J. Summarizing microblogs automatically[C]. Los Angeles, CA, United states: Association for Computational Linguistics (ACL), 2010.

[9] Harabagiu S M, Hickl A. Relevance Modeling for Microblog Summarization.[C]. 2011.

[10] Chakrabarti D, Punera K. Event Summarization Using Tweets.[C]. 2011.

[11] Inouye D, Kalita J K. Comparing Twitter Summarization Algorithms for Multiple Post Summaries[C]. Los Alamitos, CA, USA: IEEE Computer Society, 2011.

[12] Erkan G, Radev D R. LexRank: graph-based lexical centrality as salience in text summarization[J]. Journal of Artificial Intelligence Research. 2004, 22: 457-479.

[13] Mihalcea R, Tarau P. TextRank: Bringing order into texts[C]. Association for Computational Linguistics, 2004.

[14] Bian J, Yang Y, Chua T. Multimedia summarization for trending topics in microblogs[C]. San Francisco, CA, United states: Association for Computing Machinery, 2013.

[15] 罗森林, 韩磊, 潘丽敏, 等. 汉语句义结构模型及其验证[J]. 北京理工大学学报, 2013, 33(2): 166-171.

[16] 罗森林, 刘盈盈, 冯扬, 等. BFS-CTC 汉语句义结构标注语料库构建方法[J]. 北京理工大学学报, 2012, 32(3): 311-315.

[17] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research. 2003, 3(4-5): 993-1022.

[18] Lin C Y. Rouge: A package for automatic evaluation of summaries[C]//Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. 2004: 74-81.