一种事件粒度的抽取式话题简短表示生成方法

王伟玉[1,2],史存会[1],俞晓明[1],刘 悦[1],程学旗[1]
1. 中国科学院计算技术研究所，中国科学院网络数据科学与技术重点实验室
2. 中国科学院大学

论文摘要

- 用简洁明了的文字表征事件粒度的话题，可以帮助用户迅速了解话题大意
- 目前自动生成的话题简短表示效果欠佳
- 该文利用事件报道描述内容高度相似的特点，提出一种抽取式话题简短表示生成方法，把事件文档集中的标题作为处理对象，从不同的标题中抽取出保留原有语序的共性信息，并进一步融合这些共性信息，生成事件粒度的话题简短表示
- 在来自搜索引擎中的事件数据上，实验结果表明该方法能生成精练准确、语义明确完整且可读性好的话题简短表示

系统模型

抽取式话题简短表示生成方法（Extractive Topic Brief Representation method，ETBR）

- 难点分析
  - 同一事件下的多篇文档是相关的，都在报道同一件事，但由于用语习惯、表达方式等不同，有不同的描述
  - 如何从文档集不同的描述中获得该事件的主要内容，并保证生成的话题简短表示可读性好
  - 通过对同一事件的文档集研究发现，通常具有以下特征：
    - 标题概括了文档主要内容。事件的文档集主要内容一致，同一事件文档集的标题内容往往高度相似
    - 在文档集的标题中重复出现的信息通常是该事件主要内容的反映
    - 标题本身有好的可读性，其原有语序表达流畅。将标题中的部分信息按原序结合组织，会继承一定程度上好的可读性
  - 本文采用最长公共子序列的方法提取事件文档标题集的共性信息
    - 这些共性信息能保留原有语序，满足关键信息的显著性和语言表达的流畅性

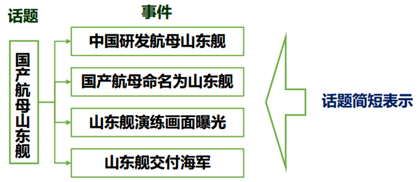
事件内的相关报道样例

事件：清华大学发现古墓
报道文章的标题
清华大学校内发现近百座古墓，暂未发现陪葬品
清华历史系师生热议发现古墓群：期待、兴奋
清华大学发现古墓：几乎无陪葬品，墓主或为平民
清华大学校园内发现95座古墓 下面到底有什么?
据考古人员认定 清华大学发现的古墓属于明清朝代

最长公共子序列示例

论文简介

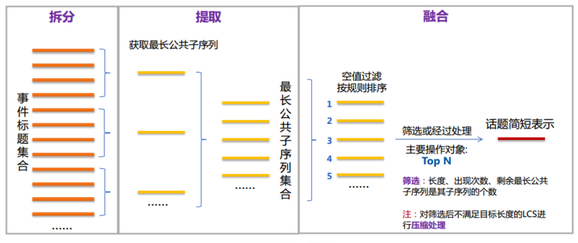
- 由多文档描述的话题需要以简洁明了的文字来表征（即话题简短表示）
- 目前生成的话题简短表示→可读性差、表意不明、不准确、以偏概全等问题
- 本文研究生成的话题简短表示形式如搜索引擎中的热搜词，比标题和摘要简洁
- 人们平时关注和讨论的话题大多是事件粒度的话题



算法原理

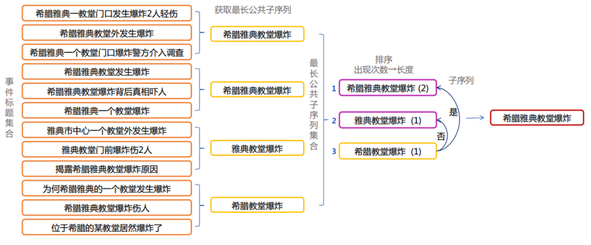
事件粒度的抽取式话题简短表示生成方法（ETBR）

- 利用事件内容相似、重复度高的特征，基于抽取出的事件文档标题集中保留原有语序的共性信息生成事件的话题简短表示
- 该算法主要包含3部分：数据的拆分、提取、融合
  - 在第2步提取出的信息还存在噪音等问题，需要在融合阶段进行筛选甚至压缩等处理



ETBR算法框架

算法应用实例



实验仿真

数据集

- 构建来源
  - 人为筛选出事件粒度的话题热搜词，爬取这些事件热搜词及搜索引擎中该事件文档集中各个文档的标题
  - 对爬取的标题预处理

- 规模

数据来源	标题数	事件数
搜索引擎	11945	123

- 样例

事件数据集样例

事件样例
云南怒江突降大雪
全国人大今日开幕
QQ将实现注销功能
运营商回应提速降费
量子摩尔定律问世
5G手机将上市
大兴机场今年运行

事件热搜词	标题
云南怒江突降大雪	春雪又来了云南怒江突降大雪 这是新年的第几场雪
	云南怒江风雪丫口突降大雪百余人几十辆车被困
	云南怒江突降大雪 147名被困人员人及36辆车均已获救
	支槽正式发威云南怒江突降大雪接下来湖南江西要成暴雨
	云南怒江突降鹅毛大雪 3小时救援被困的147人
	云南怒江海拔三千米丫口突降大雪 再次致百余人受困
	怒江多地突降大雪别怕我们的安危由他们守护

实验结果

不同算法的ROUGE评测结果

Model	ROUGE-1			ROUGE-2			ROUGE-L		
	F1	P	R	F1	P	R	F1	P	R
TextRank_weight_pro	0.75	0.71	0.82	0.51	0.48	0.56	0.48	0.47	0.53
TextRank_order_pro	0.75	0.70	0.82	0.62	0.58	0.68	0.72	0.69	0.80
TextRank_permutation_pro	0.75	0.71	0.82	0.62	0.59	0.68	0.73	0.71	0.81
TF-IDF_weight_pro	0.75	0.71	0.80	0.52	0.49	0.57	0.56	0.54	0.61
TF-IDF_weight	0.83	0.79	0.88	0.60	0.57	0.64	0.62	0.61	0.67
TF-IDF_order_pro	0.74	0.70	0.80	0.61	0.57	0.67	0.71	0.68	0.79
TF-IDF_order	0.83	0.79	0.88	0.71	0.68	0.76	0.80	0.78	0.87
TF-IDF_permutation_pro	0.74	0.70	0.80	0.61	0.57	0.67	0.71	0.68	0.79
TF-IDF_permutation	0.83	0.79	0.88	0.71	0.67	0.76	0.80	0.78	0.87
ETBR	**0.94**	**0.93**	**0.96**	**0.91**	**0.90**	**0.93**	**0.93**	**0.93**	**0.96**

不同算法的人工评测结果（6人）

Model	语义明确完整性评分	可读性评分	准确性评分	综合评分
TextRank_weight_pro	1.68	1.66	1.55	1.63
TextRank_order_pro	2.82	2.99	2.57	2.80
TextRank_permutation_pro	3.05	3.30	2.78	3.04
TF-IDF_weight_pro	2.24	2.18	1.88	2.10
TF-IDF_weight	2.54	2.46	2.17	2.39
TF-IDF_order_pro	2.96	3.07	2.64	2.89
TF-IDF_order	3.29	3.40	3.01	3.23
TF-IDF_permutation_pro	3.09	3.29	2.75	3.04
TF-IDF_permutation	3.30	3.47	3.06	3.28
ETBR	**4.58**	**4.69**	**4.47**	**4.58**

ETBR算法生成结果样例

标准话题简短表示	生成话题简短表示	标题数
运营商回应提速降费	运营商回应提速降费	95
18.6万元手机丢了	18.6万元手机丢了	97
凉山木里火场复燃	四川凉山木里火场复燃	108

论文结论

- 本文提出了一种事件粒度的抽取式话题简短表示生成方法ETBR，把事件文档集中的标题作为处理对象，该方法能抽取出同一事件下文档标题集中反映主要内容的共性信息，并将这些共性信息按原有语序组合，基于上述信息，生成其话题简短表示
- 实验结果表明该方法针对事件能较好地生成精练准确、语义明确完整且可读性好的话题简短表示

西安电子科技大学 XIDIAN UNIVERSITY