基于社会媒体内容和网络拓扑的特定话题推特摘要研究

贺瑞芳 段兴义 张雪菲 赵文丽

（天津大学智能与计算学部 天津 300350）

（天津市认知计算与应用重点实验室 天津 300350）

摘 要 推特摘要旨在从话题相关的社会媒体短文本中提炼概要的推文集，以获取有效信息，可用于舆情监控、竞争情报分析及电子商务等. 然而社会媒体的海量、嘈杂及不规范性使得仅依赖纯文本的传统摘要方法难以直接迁移到社交媒体情景中；而现有的推特摘要方法很少考虑数据稀疏性和社会网络传播带来的强冗余性，鲜有通过挖掘推文之间潜在的社会网络结构关系进行文摘内容选择，忽略了信息可以沿着社交网络进行传播. 受压缩感知及社会学理论的启发，该文提出基于社会网络和稀疏重构的推特摘要方法（SNSR）以更好地融合社会媒体内容和结构信息. 首先，挖掘推文中隐含的摘要模式，将其建模为组稀疏正则项，以捕捉代表性的推特摘要组合；其次，建模社会网络中表达一致性与表达传染性为社会化正则项，以探索推文之间的潜在网络结构关系在推特摘要中的作用；再次，建模社会媒体信息传播带来的强冗余性为多样性正则项，进而将这些约束整合到稀疏重构的推特摘要框架中；最后，提出基于 Nesterov 加速梯度下降的推特摘要算法，以解决推特摘要优化框架中的覆盖性、稀疏性以及多样性等问题. 同时，由于推特摘要标准语料的缺乏，作者建设了 12 个话题的评测数据集. 相关的实验结果证明了文中提出方法的有效性.

关键词 推特摘要；稀疏重构；网络拓扑；社会学理论；Nesterov 加速梯度下降算法

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2019.01174

Topic Oriented Twitter Summarization Based on Social Media Content and Network Topology

HE Rui-Fang DUAN Xing-Yi ZHANG Xue-Fei ZHAO Wen-Li

(*College of Intelligence and Computing, Tianjin University, Tianjin* 300350)

(*Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin* 300350)

Abstract Social media platforms, such as Twitter, provide us a very convenient way to access information, through which amounts of users can freely produce content (called tweets) on their interested topics. Therefore, it becomes one of the most popular social network. Fast increasing posts make people lost in the ocean of fragmented texts. Twitter summarization aims to extract the core and concise tweet summary from topic relevant short texts in social media so as to quickly acquire essential information. It can be used in opinion monitoring, competitive intelligence analysis and electronic commerce, especially in some emergencies, which helps to aid agencies monitor crisis progress so as to assist recovery and provide disaster relief. Yet traditional summarization methods only consider text information, which is insufficient in social media situation with the large scale, noisy and informal messages. Previous existing Twitter summarization approaches

收稿日期：2018-01-18；在线出版日期：2018-10-15. 本课题得到国家自然科学基金面上项目（61472277）和天津市自然科学基金一般项目（18JCYBJC15500）资助. **贺瑞芳**（通信作者），博士，副教授，主要研究方向为自然语言处理、社会媒体挖掘及机器学习. E-mail：rfhe@tju.edu.cn. **段兴义**，硕士，主要研究方向为自然语言处理、多文本自动摘要. **张雪菲**，硕士，主要研究方向为自然语言处理、社会计算. **赵文丽**，硕士研究生，主要研究方向为自然语言处理.

usually regard tweets as sentences, and adopt traditional summarization methods, such as SumBasic, Centroid, LexRank to validate the relevant performance on microblogging posts. However, it is not clear whether adding the complexity of methods will improve the system performance. Some other researches explore to utilize static social features except for textual content, such as number of replies, number of retweets, number of likes, author popularity and temporal signals. These methods rarely consider the data sparsity, the strong social redundancy and the potential social relations between tweets explicitly, ignoring the fact that information can spread along the social network. Inspired by compressive sensing and social theories, we propose a novel approach for Twitter summarization based on Social Network and Sparse Reconstruction (SNSR) for integrating social media content and structure information. The social analysis indicates that the members in a social network often exhibit correlated behaviors, sentiment and topic can be diffused through network. Consistency means that social behaviors conducted by the same person keep consistent in a short period of time. Contagion means that friends can influence each other. We explore whether social relations (expression consistency and expression contagion) can help Twitter summarization under a given topic, modeling relations between tweets described as the social regularization and integrating it into the group sparse optimization framework. It conducts a sparse reconstruction process by selecting tweets that can best reconstruct the original tweets, with considering coverage and sparsity. We simultaneously design the diversity regularization to remove the strong redundancy brought by social information propagation. In particular, we present a mathematical optimization formulation and develop an efficient Twitter summarization algorithm with Nesterov's accelerated gradient descent. Meanwhile, due to the lack of public corpus, we construct the gold standard tweet summary datasets for 12 different topics by asking 24 volunteers to manually select the most informative tweets, all in 48 expert summaries. Experimental results on this datasets show the effectiveness of our approach for handling the large scale short and noisy messages in social media. It suggests that integrating social network information into the proposed sparse reconstruction framework helps improve Twitter summarization. Mining the group sparsity patterns of salient tweets and designing the diversity regularization in terms of redundancy brought by social network are also effective.

Keywords　Twitter summarization; sparse reconstruction; network topology; social theories; Nesterov's accelerated gradient descent algorithm

1　引　言

社会媒体的繁荣改变和影响了人们获取和发布信息的方式.本文研究面向特定话题的推特摘要,旨在从事件相关的社会媒体短文本中提炼简洁、核心的推文集,以捕捉有效信息,可用于竞争情报分析、电子商务等;同时,也可协助政府监管危机事件,从而降低灾难损失、给出有益的反馈,并把控舆情方向.

尽管传统的文本摘要技术发展了很多年,但是新兴的社会媒体平台产生了大规模嘈杂且不规范的碎片化短文本,为社会媒体摘要研究带来了诸多挑战,然而也带来了新的机遇.现有的推特摘要方法通常将推文看作句子,对其进行重要性打分并筛选推文集.包括:(1)利用传统的文本摘要方法[1],即仅考虑文本信息,这些方法包括SumBasic[2]、Centroid[3]、LexRank[4]和TextRank[5]等;(2)利用社交媒体平台的静态特性[6-7],包括推文转发数、回复数、点赞数、用户权威特性(粉丝数、关注数等)、时间特性、地理特性等;但这些方法忽略了社会媒体短文本是网络互联的;(3)利用社交媒体平台的动态特性[8-9],即社会网络结构信息,包括转发关系、回复(Reply)关系、关注(Follow)关系等.但该类研究主要是从用户层次考虑网络结构,一般认为高权威度用户所发的推文同样具有很高的重要性.然而,通过用户之间的社会网络连接可以推测,推文之间也存在潜在的

网络结构.不同于通过计算推文相似度使得推文之间互相关联的传统方法,该方法仅仅利用了纯文本信息.通过社会网络结构构建推文层面的相互关联网络结构可能包含更多的语义线索.因此,本文需要探索一种建模推文层面网络信息互联的新方法,以进行推特摘要.

社会学理论揭示了互联信息的这种相互影响的现象.人们在短时间内更倾向于保持一致的情感、爱好,这种现象称之为一致性.除此之外,人们通过一系列交互和反馈行为在彼此之间建立了联系,这层联系对彼此产生的影响是微妙的,可以对一个人的爱好、说话方式或者表达内容产生重大的影响.人们渐渐会和好友在某个话题上保持相似的观点,甚至以相似的语调和用词来表达这些观点,这种现象称之为传染性.受到这两种社会学理论的启发,本文将进一步探索如何利用这两种理论做推特摘要.

近年来,基于数据重构的摘要方法被提出[10-12],并且在传统评测任务 DUC/TAC 上表现出色,但其并不能直接迁移到社会媒体情景中.也正是由于之前提到的社会学理论可与基于数据重构的方法无缝结合,本文从压缩感知、稀疏重构角度出发,将推文看作一种信号,提出了整合社交网络结构信息统一的推特摘要优化框架.其综合考虑了一个好的推特摘要应该具备的几个特性:(1)覆盖性(Coverage),即一个好的摘要应该尽可能包含整个语料谈论话题的各个方面;(2)稀疏性(Sparsity),即假设信号是稀疏可压缩的,摘要只是原推特语料的一部分.假设每个句子可以由所有其他句子通过非负线性组合来表示,那么不是所有句子在表示该句子的过程中都占有很大比重.摘要句子即是这样一组句子基,通过这组句子基张成的子空间,可以表示整个语料的其它句子,从而以尽可能小的误差重构原始语料;(3)多样性(Diversity),即保证摘要句子之间的冗余度尽可能小.主要贡献如下:

① 从统计学角度验证了两种社会学理论的存在,即表达一致性和表达传染性;形式化地定义了整合社会网络结构的推特摘要框架.

② 建模了推文层面的网络结构信息,并作为社会项正则整合到基于稀疏重构的优化框架中.

③ 引入组稀疏正则可以从语料层面选择重要的推文,引入多样性正则可以缓解由于社交网络的引入而带来的更加严峻的冗余度问题.

④ 构建了推特摘要语料,包括 12 个特定话题数据集以及每个数据集对应的四个专家摘要.

本文第 2 节综合分析和讨论相关工作;第 3 节给出问题阐述和数据分析,并进行社会学理论的验证;第 4 节详细论述本文提出的基于稀疏重构和社会网络结构的社会媒体推特摘要方法;第 5 节给出基于 Nesterov 加速梯度下降的摘要优化算法;第 6 节介绍真实数据集上推特摘要的标准语料制作方案,并在此基础上验证本文方法的有效性;第 7 节进行总结和展望.

2 相关工作

社会媒体平台的产生、成长经历近 10 年时间.它的繁荣催生了以推特摘要为代表的社会媒体摘要研究,其部分地传承了传统文本摘要方法.现有的自动摘要方式一般可分为两大类:抽取式和理解式.(1)抽取式摘要从原始语料中抽取一部分句子形成摘要,可以保证摘要句子的可读性,但是摘要句子之间以及摘要句子内部会产生冗余信息;(2)理解式摘要通常采用句子压缩、融合、改写等自然语言处理技术实现,在技术难度上比较大.故当前文本摘要研究大多数还是基于抽取式的研究路线.同时,由于推特文本的碎片化、不规范性以及大量噪声,使得理解式摘要方法中的语法分析、句法分析等底层技术难以发挥作用.因此,本文采取抽取式摘要路线,相关工作的调研也围绕此展开.

2.1 传统文本摘要

产生文本摘要的过程通常可以描述为:句子重要性打分、句子筛选、摘要句子排序这三个过程.如何对句子进行重要性打分是摘要研究的重点.已有的方法包括:(1)基于特征的方法,例如 Centroid 和 SumBasic[2-3],这些方法考虑了词频和句子位置信息来计算句子的权重;(2)基于图的方法采用了类似 PageRank 的算法,例如 LexRank 和 TextRank[4-5],以句子或词作为节点,句子或词之间的相似度作为构建图中边权重的依据,利用随机游走的思想最终得到句子或词的重要性.然而,该类方法面临去冗余的问题.一些研究者开始利用(3)基于聚类的思想来保持摘要的多样性[13-18].其主要采用主题建模、聚类算法以及矩阵分解的思想来产生覆盖性更高的摘要.最近,(4)基于数据重构的摘要方法的出现[10-12],为解决摘要研究中存在的经典问题,即覆盖性、重要性及多样性,带来了新的可能性.然而由于社交媒体中大规模的文本具有简短、嘈杂及其附带的社会特性,使得这些传统方法不能很好地发挥作用.

2.2　推特摘要

社交媒体的蓬勃发展使得人们不断探索传统摘要方法在类似推特平台上的应用. 这些方法包括(1) Hybrid TF-IDF[1]，其针对推特短文本语料对经典的 TF-IDF 模型进行变种，在计算 IDF(Inverse Document Frequency)的时候是把每个帖子看成一个文档，在计算 TF(Term Frequency)的时候是把所有帖子看成一个大文档；(2) 短语强化算法[19-20]，其通过不断选择使用频率最高的短语，最终生成摘要句子. 这些方法仅考虑了文本信息. 然而，社交媒体平台还包含除文本信息之外的大量丰富信息，比如推文转发数、回复数、用户粉丝数、关注数、时间、地理信息，以及社交网络结构等信息；(3) 基于社会网络静态信息和用户层面网络结构的推特摘要，Duan 等人[9]考查了推文内容质量、用户发文数、粉丝数、粉丝数与关注数的比率等信息，以及通过关注(Follow)关系构建的用户网络结构；Liu 等人[7]考查了推文转发数、用户粉丝数以及推文可读性三方面的特性. 以上两个工作均采用基于 PageRank 的扩展模型. Alsaedi 等人[21]提出了用于事件摘要的三个方法，考查了时间和转发数两个特性. Chang 等人[6,8]把推特摘要研究看成有监督的分类任务，即判断每条推文可否被选择为一个摘要句子. 并通过充分挖掘社交媒体中存在的一些特性，包括推文转发数、回复数、点赞数、内容相关度、用户粉丝数、关注数、权威度、时间间隔等，作为分类器的输入特征，从而选择摘要推文.

以上方法主要利用了社会网络的静态信息或者用户层面的网络结构，并没有考虑到推文层面的网络结构，而诸如情感、话题、内容等推特信息是可以沿着潜在的网络结构进行传播，本文通过研究这种传播现象，以期获得更多潜在的语义线索进行推特摘要研究.

2.3　结合社会网络结构的探索

社会学理论以及社会网络分析为我们做推特摘要研究提供了新思路，即如何结合拓扑结构和文本内容做推特摘要. 社会网络传播，或者说社交影响力在多个领域都有研究. 比如，(1) 情感分析[22]认为人们在短时间内对某个话题或事物保持情感一致性，具有朋友关系的两个人在情感上更容易互相影响，并称之为情绪传染性；(2) 话题识别[23]认为人们在短时间内对专注的话题会保持一致的偏好，除此之外，具有朋友关系的两个人更有可能对同一话题感兴趣，并称之为社会传染性；(3) 话题具体的影响力分析[24]识别特定话题下具有影响力的用户，并对用户关注(Follow)关系中的这种现象进行建模：一部分用户关注其他人是由于对共同话题感兴趣，故粉丝所发内容是会受到关注者所发内容的影响；而一部分用户关注其他用户只是由于热度(粉丝数等)，他们所发内容很少受到关注者所发内容的影响. 通过将这种现象建模到主题模型中，既可以识别特定话题下具有影响力的用户，同时也可以提升话题检测的性能；以及(4) 网络推断和话题模型的联合建模[25]等. 根据这些研究可知，情感和话题是可以沿着网络传播的. 本文将深入探索作为情感和话题的载体——表达内容是否可以沿着网络进行传播，以及如何影响推特摘要.

3　问题陈述与数据分析

本文面向特定话题进行推特摘要研究，即输入与某个话题相关的推特文本集，输出若干条重要推文形成摘要并可描述该话题的主要内容. 本节首先给出一些符号定义，并正式描述本文推特摘要的整个流程；其次由于缺乏推特摘要的公开语料，本节将介绍语料建设中的数据准备环节，其中专家摘要的制作过程放在实验部分；同时，重新定义两种社会学理论(表达一致性、表达传染性)，并在我们的数据集上验证其存在性.

3.1　问题陈述

本文约定加粗的大写字符表示矩阵(例如 $\boldsymbol{M}$)，加粗的小写字符表示向量(例如 $\boldsymbol{m}$)，小写字符表示标量(例如 m). $\boldsymbol{M}_{i*}$ 和 $\boldsymbol{M}_{*j}$ 分别表示矩阵 $\boldsymbol{M}$ 的第 i 行和第 j 列. $\boldsymbol{M}_{ij}$ 表示矩阵 $\boldsymbol{M}$ 在第 i 行第 j 列的值. $\|\boldsymbol{M}\|_{\mathrm{F}}$ 表示矩阵的 Frobenius 范数，$\|\boldsymbol{M}\|_{2,1}$ 表示矩阵的 $\ell_{2,1}$ 范数. 特别地，

$$\|\boldsymbol{M}\|_{\mathrm{F}}=\sqrt{\sum_{i}^{m}\sum_{j}^{n}\boldsymbol{M}_{ij}^{2}},$$

$$\|\boldsymbol{M}\|_{2,1}=\sum_{i}^{m}\|\boldsymbol{M}_{i*}\|_{2}=\sum_{i}^{m}\sqrt{\sum_{j}^{n}\boldsymbol{M}_{ij}^{2}}.$$

给定特定话题的推特语料，该语料可表示为 TF-ITF(Term Frequency Inverse Tweet Frequency)矩阵，即 $\boldsymbol{S}=[\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_n]\in\mathbb{R}^{m\times n}$，其中 m 表示词汇表大小，n 表示推文数量. $\boldsymbol{S}$ 矩阵的每一列 $\boldsymbol{t}_i$ 为单个推文的向量表示. $\boldsymbol{U}\in\mathbb{R}^{d\times n}$ 表示用户-推文矩阵，其中，d 为用户数，$\boldsymbol{U}_{ij}=1$ 表示第 j 条推文是由第 i 个用户发布. 本文根据 Follow 关系构建用户-用户矩阵 $\boldsymbol{F}\in\mathbb{R}^{d\times d}$，其中 $\boldsymbol{F}_{ij}=1$ 表示第 i 个与第 j 个用户是有联系的.

依据上面给定的符号，本文的推特摘要任务可

定义为:给定描述特定话题的推特语料 C,可以获得文本内容矩阵 $\boldsymbol{S}$、用户-推文矩阵 $\boldsymbol{U}$ 以及用户-用户矩阵 $\boldsymbol{F}$,我们的目标是通过优化模型得到重构系数矩阵 $\boldsymbol{W}$,并根据 $\boldsymbol{W}$ 按一定压缩比自动生成摘要 $Summary \approx \boldsymbol{SW}$.

3.2 数据描述

使用公开的推特语料作为原始数据集,其最初由伊利诺伊大学①的一个研究团队所收集.由于集中在 5,6,7 三个月的数据最多(数据收集方式所致),我们统计每个月的 Hashtag 频数,选择那些频数较大且对应于某个具体事件(一方面可以查看包含该 Hashtag 的推文内容是否描述某个事件,一方面可以在浏览器上通过检索时间及 Hashtag,确认是否发生了某个事件)的 Hashtag 作为话题标签,利用这些话题标签收集话题数据集.除此之外,每个话题不止一个话题标签,比如"#osama"和"#osamabinladen"描述的是一个话题.有些推文内容虽然不包含标签信息,但是包含类似"osama"这样的关键词,因此,我们主要根据推文内容是否包含 Hashtag 或者除掉"#"号后得到的关键词来收集话题.

结合时间信息(每条推文都包含发布时间这一信息)以及上述处理过程,可以得到某个话题的推文数量随时间的变化,通过观察这种时序变化趋势,大致把话题分为热点话题(如图 1)和突发话题(如图 2)两种.同时考虑到社会学理论中一致性和传染性的短时间效应.进一步做如下处理:若该话题是热点话题,则收集该话题发生前后共五天时间内的推特数据作为该热点话题的数据集(图 1);若该话题是突发话题,则收集该话题发生后五天时间内的推特数据作为该突发话题的数据集(图2).最后筛选得到 12 个话题,这些话题涉及政治、科技、体育、自然灾害、恐怖袭击和娱乐八卦等领域.得到特定话题的数据集后,需进一步做数据清洗,把满足以下条件之一的推文过滤掉,最终得到 12 个话题的统计信息参见表 1:

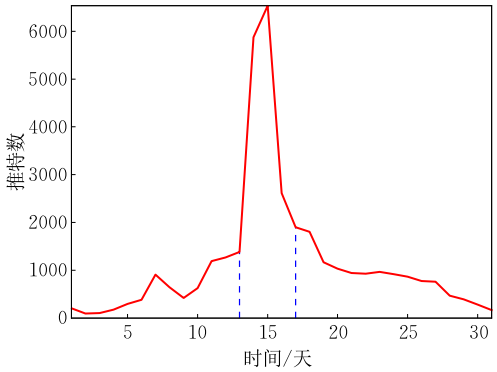


图 1 哈利波特上映

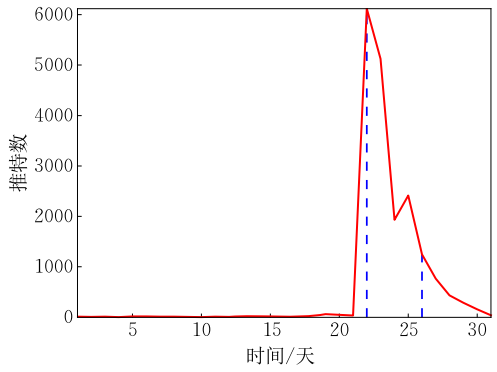


图 2 挪威恐怖袭击

(1)重复多次的推文(只保留一次).

(2)除 Hashtag、关键词、@、URL 以及停用词外,单词数少于 3 的推文(对于特定话题,几乎所有的推文都包含一致的 Hashtag 或关键词.除掉以上这些信息后几乎没有其他内容的推文,我们认为其增量信息比较少).

(3)对应用户在数据集中属于孤立点的推文(保证网络结构的稠密性).

表 1 数据集统计信息

统计信息	时间	推文数	用户数	用户最大度	用户最小度	用户最大推文数	用户最小推文数	用户平均推文数	P-value(一致性)	P-value(传染性)
Osama	0501	4680	1309	69	1	42	2	3.65	4.78E-125	1.82E-33
Jopin	0522	2896	1082	68	1	93	1	2.68	2.10E-98	6.60E-09
Mavs	0612	3859	1780	76	1	92	1	2.18	9.01E-211	8.09E-08
Weinergate	0616	1278	885	52	1	11	1	1.45	4.62E-19	5.17E-10
Betawards	0626	787	200	16	1	57	1	3.94	2.10E-27	3.62E-05
Casey-Anthony	0705	6241	1318	74	1	180	2	4.74	5.66E-97	1.59E-27
Asobama	0706	4888	2009	142	1	63	1	2.43	5.32E-99	9.81E-06
Atlantis	0708	2515	712	47	1	21	2	3.53	1.01E-72	1.44E-14
Harrypotter	0715	2760	865	37	1	26	2	3.19	3.41E-89	2.24E-10
WWC	0717	3642	2103	219	1	25	1	1.73	3.08E-54	1.05E-08
Oslo	0722	4571	1026	77	1	56	2	4.46	2.62E-131	4.98E-19
SDCC	0722	5817	442	81	2	161	2	13.16	3.21E-143	1.31E-11

3.3 社会学理论验证分析

社会学理论,尤其是一致性[26]和传染性[27-28],

① https://wiki.engr.illinois.edu/display/forward/Dataset-UDI-TwitterCrawl-Aug2012

已经在很多社交媒体任务中被证明是有用的.社会学理论指出社会网络中成员之间通常会展现出相关的行为,情感和话题都会随着网络进行传播.一致性一般认为,同一个人在短时间内表现出的社会行为具有一致性;传染性一般认为,具有朋友关系的两个人可以对彼此产生影响.本节我们主要考查对于每个话题集,社会学理论是否存在,并且给出验证方法.首先对于我们的任务,重新定义了一致性和传染性:

(1)表达一致性.同一用户所发的两个推文在内容上是否比随机选择的两个推文更相似?

(2)表达传染性.具有朋友关系的两个用户所发的推文在内容上是否比随机选择的两个推文更相似?

为了验证这两个问题,我们给出计算两个推文距离的公式 $\boldsymbol{A}_{ij}=\|\boldsymbol{t}_i-\boldsymbol{t}_j\|_2$,其中 $\boldsymbol{t}_i$,$\boldsymbol{t}_j$ 分别为第 i,j 条推文的向量表示.两个推文越相似,$\boldsymbol{A}_{ij}$ 越接近 0.对于第一个问题,我们构建两个维度一致的向量 $\boldsymbol{cons}_c$ 和 $\boldsymbol{cons}_r$.第一个向量的每一维是通过计算同一用户所发两条推文的距离得到,第二个向量的每一维是通过计算两条随机选择推文的距离得到.然后对这两个向量做双样本 T 检验,并设置空假设为,两个向量并无很大差异,即 $H_0:\boldsymbol{cons}_c=\boldsymbol{cons}_r$;备择假设为,同一用户所发的两条推文在距离上比随机选择的两条推文更小,即 $H_1:\boldsymbol{cons}_c<\boldsymbol{cons}_r$.

类似地,为了验证第二个问题,我们构建了两个维度一致的向量 $\boldsymbol{cont}_c$ 和 $\boldsymbol{cont}_r$.第一个向量的每一维是通过计算具有朋友关系的用户所发两条推文的距离得到,第二个向量的每一维是通过计算随机选择的两条推文的距离得到.我们同样在这两个向量上做双样本 T 检验.设置空假设为 $H_0:\boldsymbol{cont}_c=\boldsymbol{cont}_r$,即两个朋友关系的用户所发推文的距离与随机选择两条推文的距离并无很大差异.备择假设为 $H_1:\boldsymbol{cont}_c<\boldsymbol{cont}_r$,表示两个朋友关系的用户所发的推文在距离上比随机选择的两条推文更小.

对于所有的话题集,一致性空假设和传染性空假设都以置信度 $\alpha=0.01$(两种社会学理论在所有数据集中均以超过 99%的概率存在)的水平被排斥,其中 P-values 在表 1 的最后两列呈现.该验证分析说明,两种社会学理论在数据集中是真实存在的,这为在推特摘要优化建模中融入社会媒体信息传播的一致性和传染性奠定了基础.

4　SNSR 总体框架

4.1　推特摘要的组稀疏模式挖掘

压缩感知理论认为,自然信号一般是稀疏且可压缩的.稀疏重构的思想与稀疏编码(Sparse Coding)类似,来源于压缩感知理论.即原始信号可由一组基向量来表示,通过约束最小化重构误差,找到最具有代表性的基向量,其张开的子空间可以很好地表示原始信号的空间.这种思想被广泛应用于信号处理、图像或视频压缩等领域,比如针对图像压缩,一方面通过观察基像素点即可了解原始图像的大致内容;另一方面通过保存基像素点和重构矩阵,即可最大地还原原始图像,使得对于超大规模图像的保存,可以大大节省空间消耗.推特摘要任务的目标与稀疏重构的思想不谋而合.从推文集中抽取简洁、核心的代表信息形成摘要,通过阅读摘要即可了解原始数据集的概要内容,也相当于是对原始文本集的一种压缩处理.

特别地,对于抽取式推特摘要方法,可以把原始推文(句子)集看作信号,那么推特摘要的任务就是从原信号中寻找能最好地重构其的样本子空间,即一组推文基向量,使得这组推特摘要句子可以最大化地重构原始推特文本集.

对于给定的推特语料 C,其可以表示为文本矩阵 $\boldsymbol{S}\in\mathbb{R}^{m\times n}$,对于第 i 个推文 $\boldsymbol{t}_i\in\boldsymbol{S}$,可以通过其他推文的线性组合形式化表示为

$$\boldsymbol{t}_i=\sum_{j=1}^{n}\boldsymbol{c}(j)\boldsymbol{W}_{ji}\boldsymbol{t}_j \tag{1}$$

为了更好地解释文摘句选择的物理含义,式(1)中 $\boldsymbol{c}(j)=0$ 表示第 j 个推文 $\boldsymbol{t}_j$ 不是摘要推文,$\boldsymbol{c}(j)=1$ 表示第 j 个推文 $\boldsymbol{t}_j$ 最终被选为摘要推文.对于抽取式摘要,假设我们最终需要抽取 k 条推文形成摘要,则有 $\sum_{j=1}^{n}\boldsymbol{c}(j)=k$.$\boldsymbol{W}_{ji}$ 表示推文 $\boldsymbol{t}_j$ 在重构推文 $\boldsymbol{t}_i$ 时的权重,其值越大表示在重构推文 $\boldsymbol{t}_i$ 时所占的比重越大.由于每一个推文向量 $\boldsymbol{t}_i$ 是通过计算 TF-ITF 得到的,每一个维度均为非负值,故需要对 $\boldsymbol{W}_{ji}$ 增加非负约束.除此之外,我们需要增加额外的约束 $\boldsymbol{W}_{ii}=0$ 来避免句子自身重构自身的现象,否则会导致其重构系数接近于 0,$\boldsymbol{W}_{ii}$ 接近于 1,以至于失去稀疏重构原本的意义.因此,基于稀疏重构的推特摘要方法的目标函数可以表示为

$$\min \frac{1}{2}\sum_{i=1}^{n}\left\|\boldsymbol{t}_i-\sum_{j=1}^{n}\boldsymbol{c}(j)\boldsymbol{W}_{ji}\boldsymbol{t}_j\right\|_2^2$$

$$\text{满足 } \boldsymbol{c}\in\{0,1\}^n,\ \sum_{j=1}^{n}\boldsymbol{c}(j)=k \tag{2}$$

$$\forall_{i\in\{1,2,\cdots,n\}}\boldsymbol{W}_{ii}=0$$

$$\forall_{i,j\in\{1,2,\cdots,n\}}\boldsymbol{W}_{ji}\geq 0$$

式(2)可以进一步用矩阵形式来表示：

$$\min \frac{1}{2}\left\|\boldsymbol{S}-\boldsymbol{S}\boldsymbol{D}(\boldsymbol{c})\boldsymbol{W}\right\|_{\mathrm{F}}^2$$

$$\text{满足 } \boldsymbol{c}\in\{0,1\}^n,\ \sum_{j=1}^{n}\boldsymbol{c}(j)=k \tag{3}$$

$$diag(\boldsymbol{W})=0,\ \boldsymbol{W}\geq 0$$

式(3)中 $\boldsymbol{S}$ 表示文本矩阵，$\boldsymbol{D}(\boldsymbol{c})$ 是一个对角矩阵，第 i 行对角元素的值对应于 $\boldsymbol{c}(i)$ 的取值，$\boldsymbol{W}=[\boldsymbol{W}_{*1},\boldsymbol{W}_{*2},\cdots,\boldsymbol{W}_{*n}]\in\mathbb{R}^{n\times n}$ 是一个重构系数矩阵，每一列 $\boldsymbol{W}_{*j}=[\boldsymbol{W}_{1j},\boldsymbol{W}_{2j},\cdots,\boldsymbol{W}_{nj}]$ 是重构推文 $\boldsymbol{t}_j$ 的系数向量. $\boldsymbol{W}\geq 0$ 保证矩阵元素非负，再加上约束 $diag(\boldsymbol{W})=0$，即可保证对角线的每一个元素为 0，即 $\boldsymbol{W}_{ii}=0$.

由于 $\boldsymbol{c}$ 向量的约束，使得目标函数式(3)的优化是一个混合线性规划问题，求解非常困难. 鉴于 $\boldsymbol{D}(\boldsymbol{c})$ 的对角线只有有限多个 1，而多数取值为 0，使 $\boldsymbol{D}(\boldsymbol{c})\boldsymbol{W}$ 所得矩阵中会有很多整行为 0 的情况. 令 $\boldsymbol{W}=\boldsymbol{D}(\boldsymbol{c})\boldsymbol{W}$，并通过对 $\boldsymbol{W}$ 添加 $\ell_{2,1}$ 范数约束，即组稀疏正则项，可以确保 $\boldsymbol{W}$ 的行稀疏性，从而近似模拟 $\boldsymbol{D}(\boldsymbol{c})\boldsymbol{W}$ 的行选择过程. 由于 $\boldsymbol{W}_{i*}=[\boldsymbol{W}_{i1},\boldsymbol{W}_{i2},\cdots,\boldsymbol{W}_{in}]$ 中的每一维表示第 i 条推文重构其他推文时的权重，当第 i 行元素全部为 0 时，表示该推文在重构整个语料中的重要性比较低，也就很大概率不会被选择为摘要推文，所以对 $\boldsymbol{W}$ 的行选择可以认为是对推文的选择. 实际上，若是去掉 $\boldsymbol{D}(\boldsymbol{c})$ 的相关表达，直接加组稀疏约束建模效果是等价的，对运算过程没有影响. 式(3)可以重新改写为

$$\min_{\boldsymbol{W}}\frac{1}{2}\left\|\boldsymbol{S}-\boldsymbol{S}\boldsymbol{W}\right\|_{\mathrm{F}}^2+\lambda\left\|\boldsymbol{W}\right\|_{2,1} \tag{4}$$

$$\text{满足 } diag(\boldsymbol{W})=0,\ \boldsymbol{W}\geq 0$$

式(4)中 λ 为组稀疏正则项参数，

$$\left\|\boldsymbol{W}\right\|_{2,1}=\sum_{i=1}^{n}\left\|\boldsymbol{W}_{i*}\right\|_2,$$

$$\left\|\boldsymbol{W}_{i*}\right\|_2=\sqrt{\sum_{j=1}^{n}\boldsymbol{W}_{ij}^2}.$$

由此通过组稀疏学习的约束可以实现挖掘推文集中的摘要推文组的潜在模式，使得摘要推文从全局角度保证了一定的非冗余性.

4.2 建模推文层次的网络结构

为了减少重构误差，并且在重构过程中做出纠正，我们使用社会学理论建模推文层次的网络结构信息，并作为社会正则项整合到稀疏重构的优化框架中. 源于 Graph Lasso[29] 思想的启发，也就是说，对于两条相关连的推文，由于其本来距离就很接近，需要让它们在重构过程中依旧保持相似.

为了利用社交网络结构做推特摘要，我们使用之前提到的社会学理论来构建推文层次的网络结构. 具体地，需要把给定的用户-推文矩阵 $\boldsymbol{U}$ 和用户-用户矩阵 $\boldsymbol{F}$ 转换为推文-推文矩阵：(1)通过表达一致性理论构建的推文-推文关联矩阵定义为 $\boldsymbol{T}_{\mathrm{cons}}=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}$，其中 $\boldsymbol{T}_{\mathrm{cons}}=1$ 表示两条推文是同一用户所发；(2)通过表达传染性理论构造的推文-推文关联矩阵被定义为 $\boldsymbol{T}_{\mathrm{cont}}=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{U}$，其中 $\boldsymbol{T}_{\mathrm{cont}}=1$ 表示两条推文是具有朋友关系的用户所发. 然后，我们通过线性组合这两种矩阵，最终得到推文-推文关联矩阵为 $\boldsymbol{T}=\boldsymbol{T}_{\mathrm{cons}}+b\boldsymbol{T}_{\mathrm{cont}}$，其中 b 是这两种矩阵的平衡参数. 理论上，b 值越大，说明传染性的影响越大，两个具有朋友关系的用户所发的推特越接近，在公式上表现为尽可能拉近推特的距离，避免重构误差(亦即本来距离较近的两条推特，在重构后距离拉大，加上该约束可以拉近距离，纠正重构偏差). b 值越小，说明传染性的影响越小，具有朋友关系的用户所发推特越容易避免被强制拉近. 至于 b 取大取小，取决于数据集本身的网络特性(传染性)，是一个可调节的参数. 实验时我们简单取 $b=1$，当然也可以对 b 的不同取值作分析. $\boldsymbol{T}_{ij}=1$ 表示两条推文是有关联的，否则没有关联. 我们定义 $\boldsymbol{S}$ 的重构矩阵为 $\hat{\boldsymbol{S}}=\boldsymbol{S}\boldsymbol{W}$，那么 Graph Lasso 惩罚项，即社会正则项可以表示为

$$\begin{aligned}\Omega_{\mathrm{graph}}&=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{T}_{ij}\left\|\hat{\boldsymbol{S}}_{*i}-\hat{\boldsymbol{S}}_{*j}\right\|_2^2\\&=\sum_{i=1}^{m}\hat{\boldsymbol{S}}_{i*}\left\|\boldsymbol{D}-\boldsymbol{T}\right\|\hat{\boldsymbol{S}}_{i*}^{\mathrm{T}}\\&=tr(\boldsymbol{S}\boldsymbol{W}\boldsymbol{L}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}^{\mathrm{T}})\end{aligned} \tag{5}$$

其中，$tr(\cdot)$ 表示矩阵的迹，$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{T}$ 是拉普拉斯矩阵，$\boldsymbol{D}\in\mathbb{R}^{n\times n}$ 是对角矩阵，而且 $\boldsymbol{D}_{ii}=\sum_{j=1}^{n}\boldsymbol{T}_{ij}$，每一个对角元素表示推文节点在图中的度.

通过整合式(5)到式(4)，可以得到：

$$\min_{\boldsymbol{W}}\frac{1}{2}\left\|\boldsymbol{S}-\boldsymbol{S}\boldsymbol{W}\right\|_{\mathrm{F}}^2+\frac{\alpha}{2}tr(\boldsymbol{S}\boldsymbol{W}\boldsymbol{L}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}^{\mathrm{T}})+\lambda\left\|\boldsymbol{W}\right\|_{2,1} \tag{6}$$

$$\text{满足 } diag(\boldsymbol{W})=0,\ \boldsymbol{W}\geq 0$$

其中，α 是社会正则项参数，式(6)产生了将社会媒

体内容与网络拓扑结构相融合的基于稀疏重构的推特摘要框架.

4.3 社会网络传播带来的强冗余信息建模

去冗余一直是摘要研究的重点.社会研究表明[28],社会网络中的互惠关系以及某些三元结构大大增加了社会传染性,这将导致在某个特定的网络结构中,会带来更多的冗余以及缺乏新颖性的信息.因此,相较于传统摘要研究,推特摘要面临更严峻的去冗余问题.目前存在一些考虑多样性来选择摘要的研究.最具代表性的方法称为最大边缘相关性(Maximal Marginal Relevance,MMR).该方法一般是在推文排序后使用,通过综合考虑相关度与冗余度进行句子选择,其作为额外的步骤来实施,而不是模型的一部分.基于话题的方法可以发现语料中的子话题,同时在某种程度上解决了多样性问题.但是这类方法存在一些关键挑战:评估推文在每个子话题的重要性,评估每个子话题在整个语料的重要性,子话题之间的去冗余(比如,由于划分粒度较小,使得两个子话题很类似).

基于稀疏重构的摘要方法倾向于从语料层面选择重要的句子,但是没有明显的倾向会包含语料的各个方面.这类工作已有的针对多样性的研究包括:Liu 等人[11]引入一个相关性项来控制多样性,但是他们的模型求解过程比较复杂;Yao 等人[12]引入了一个不相似度矩阵,大大地降低了计算复杂度.然而他们计算该不相似度矩阵的方法并不适用于推特语料,一方面是由于推特语料的嘈杂、不规范;一方面他们使用句子长度或者词汇库大小来计算每个单词的编码损失,这种计算方法使得不相似度矩阵中的每个元素都很大,导致 $\boldsymbol{W}$ 的每个元素都接近于 0,不容易区分句子的重要性.

受到该不相似度矩阵的启发,本文提出相对简单但却很有效的余弦相似度矩阵 ∇ 对社会媒体传播引发的强冗余信息建模.对于每个元素 $\nabla_{ij}\in[0,1]$ 表示推文 $\boldsymbol{t}_i$ 和推文 $\boldsymbol{t}_j$ 的余弦相似度.在稀疏重构的过程中,我们添加约束 $diag(\boldsymbol{W})=0$ 来避免自身重构自身的现象,基于这种认识,我们有理由避免推文被那些与其极为相似的推文重构.例如下面这个例子:

(1) Tweet1: the mood was solemn at the garden of reflection in lower makefield following the death of osama bin laden. video: http://fb.me/tof3pqok

(2) Tweet2: the mood was solemn at the garden of reflection in lower makefield following osama bin laden's death. video: http://bit.ly/l9tvdw

显然这两条推文很相似,这会导致重构系数 $\boldsymbol{W}_{12}$ 和 $\boldsymbol{W}_{21}$ 都接近于 1,从而导致提高了这两条推文在整个语料的重要性的问题.通过初步的实验,我们观察到,倘若不加多样性约束,在生成的最终摘要中会存在很多这种相似推文.为了更好地避免这种"相似"的重构现象,我们重新计算 ∇ 为

$$\nabla_{ij}=\begin{cases}1, & \text{若}\nabla_{ij}\geq\theta\\ 0, & \text{否则}\end{cases} \tag{7}$$

式(7)中 θ 是用来区分相似推文对和常规推文对的阈值.由此进一步建模推特摘要的多样性,并提出多样性正则项的表达形式:

$$tr(\nabla^{\mathrm{T}}\boldsymbol{W})=\sum_{i=1}^{n}\sum_{j=1}^{n}\nabla_{ij}\boldsymbol{W}_{ij}.$$

并整合到式(6)中得到最终的目标函数为

$$\min_{\boldsymbol{W}}\frac{1}{2}\|\boldsymbol{S}-\boldsymbol{S}\boldsymbol{W}\|_{\mathrm{F}}^{2}+\frac{\alpha}{2}tr(\boldsymbol{S}\boldsymbol{W}\boldsymbol{L}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}^{\mathrm{T}})+\gamma tr(\nabla^{\mathrm{T}}\boldsymbol{W})+\lambda\|\boldsymbol{W}\|_{2,1}$$
$$\text{满足}\ diag(\boldsymbol{W})=0,\ \boldsymbol{W}\geq 0 \tag{8}$$

其中,γ 是多样性正则项参数.通过优化目标函数式(8),可以得到重构的系数矩阵 $\boldsymbol{W}$.每条推文的重要性分数可以通过下式得到:

$$Score(\boldsymbol{t}_i)=\|\boldsymbol{W}_{i*}\|_2 \tag{9}$$

依据重要性分数对推文进行排序,最后筛选前几条推文形成最终摘要.

5 优化的推特摘要算法

5.1 算　法

由于 $\|\boldsymbol{W}\|_{2,1}$ 不可导,所以目标函数式(8)是非平滑的.受到前人工作[22,30-31]的启发,本文提出基于 Nesterov 加速梯度下降的摘要优化算法.针对该非平滑优化问题的解决方法,目标函数可以被等价的表示为

$$\min_{\boldsymbol{W}}f(\boldsymbol{W})=\frac{1}{2}\|\boldsymbol{S}-\boldsymbol{S}\boldsymbol{W}\|_{\mathrm{F}}^{2}+\frac{\alpha}{2}tr(\boldsymbol{S}\boldsymbol{W}\boldsymbol{L}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}^{\mathrm{T}})+\gamma tr(\nabla^{\mathrm{T}}\boldsymbol{W})$$
$$\text{满足}\ Z=\{\boldsymbol{W}\,|\,\boldsymbol{W}\geq 0,diag(\boldsymbol{W})=0,\|\boldsymbol{W}\|_{2,1}\leq z\} \tag{10}$$

式(10)中 z 是 $\ell_{2,1}$ 球的半径,并且 λ 和 z 具有一对一的映射关系.

由于任意范式都定义了一个凸集,故 Z 是一个封闭的凸集.由此,我们的问题转换为定义域为封闭凸集、目标函数为凸函数的凸优化问题.

接下来,我们阐述本文基于 Nesterov 加速投影

梯度下降的优化算法，可以用来解决式(10)中带约束的凸优化问题.

首选不考虑式(10)中带约束 $\boldsymbol{W}\in Z$ 的优化问题：

$$\min_{\boldsymbol{W}} f(\boldsymbol{W}).$$

我们知道，通过梯度下降，$\boldsymbol{W}_{t+1}$ 可以通过式(11)更新：

$$\boldsymbol{W}_{t+1}=\boldsymbol{W}_t-\frac{1}{lr}f'(\boldsymbol{W}_t) \tag{11}$$

其中，lr 表示学习率，lr 的值根据 Armijo-Goldstein[32] 规则通过线搜索(Line Search)方法得到. $f'(\boldsymbol{W})$ 表示目标函数 $f(\boldsymbol{W})$ 对 $\boldsymbol{W}$ 的求导：

$$f'(\boldsymbol{W})=\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{W}-\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}+\gamma\nabla+\alpha\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{W}\boldsymbol{L} \tag{12}$$

优化问题中的平滑部分可以等价地改写为线性函数 $f(\boldsymbol{W})$ 在 $\boldsymbol{W}_t$ 处的近端正则，表示为

$$\boldsymbol{W}_{t+1}=\arg\min_{\boldsymbol{W}} G_{lr,\boldsymbol{W}_t}(\boldsymbol{W}),$$

其中，

$$G_{lr,\boldsymbol{W}_t}(\boldsymbol{W})=f(\boldsymbol{W}_t)+\langle f'(\boldsymbol{W}_t),\boldsymbol{W}-\boldsymbol{W}_t\rangle+\frac{lr}{2}\|\boldsymbol{W}-\boldsymbol{W}_t\|_{\mathrm{F}}^2 \tag{13}$$

考虑到我们优化问题的等价形式以及约束项 Z，我们可以通过以下迭代公式得到最终解：

$$\boldsymbol{W}_{t+1}=\arg\min_{\boldsymbol{W}\in Z} G_{lr,\boldsymbol{W}_t}(\boldsymbol{W}) \tag{14}$$

通过忽视式(13)中独立于 $\boldsymbol{W}$ 的项，式(14)可以归约为

$$\boldsymbol{W}_{t+1}=\min_{\boldsymbol{W}\in Z}\frac{1}{2}\|\boldsymbol{W}-\boldsymbol{U}_t\|_{\mathrm{F}}^2 \tag{15}$$

其中，$\boldsymbol{U}_t=\boldsymbol{W}_t-1/lr\,f'(\boldsymbol{W})$，则 $\boldsymbol{W}$ 表示 $\boldsymbol{U}$ 在凸集 Z 上的欧几里得投影(Euclidean projection)：

式(15)可以分解为 n 个子问题：

$$\boldsymbol{w}_{t+1}^j=\min_{\boldsymbol{w}^j\in z^j}\frac{1}{2}\|\boldsymbol{w}^j-\boldsymbol{u}_t^j\|_2^2 \tag{16}$$

其中，$\boldsymbol{u}_t^j,\boldsymbol{w}^j,\boldsymbol{w}_t^j$ 分别表示矩阵 $\boldsymbol{U}_t,\boldsymbol{W},\boldsymbol{W}_t$ 的第 j 行. 给定 λ，通过欧几里得投影得到的解的形式为

$$\boldsymbol{w}_t^j=\begin{cases}\left(1-\dfrac{\lambda}{lr\|\boldsymbol{u}_t^j\|}\right)\boldsymbol{u}_t^j, & \text{若}\|\boldsymbol{u}_t^j\|\geq\dfrac{\lambda}{lr}\\ 0, & \text{否则}\end{cases} \tag{17}$$

上述方法的收敛速度为 $O(1/k)$，而 Nesterov 方法加速了该优化过程，使收敛速度达到 $O(1/k^2)$，其中 k 表示迭代次数. Nesterov 方法基于两个序列 $\{\boldsymbol{W}_t\}$ 和 $\{\boldsymbol{V}_t\}$，其中 $\{\boldsymbol{W}_t\}$ 是近似解序列，$\{\boldsymbol{V}_t\}$ 是搜索点序列. $\{\boldsymbol{V}_t\}$ 是 $\{\boldsymbol{W}_t\}$ 和 $\{\boldsymbol{W}_{t-1}\}$ 的结合：

$$\boldsymbol{V}_t=\boldsymbol{W}_t+\varsigma(\boldsymbol{W}_t-\boldsymbol{W}_{t-1}),$$

其中，ς 是结合系数. $\boldsymbol{U}_t$ 可以由 $\{\boldsymbol{V}_t\}$ 通过类似“梯度”更新的方法计算得到，所以 $\boldsymbol{U}_t$ 可以重新计算为

$$\boldsymbol{U}_t=\boldsymbol{V}_t-\frac{1}{lr}f'(\boldsymbol{W}).$$

详细的算法过程见算法 1.

算法 1. 基于 NAG 的模型优化算法.

输入：$\boldsymbol{S},\boldsymbol{U},\boldsymbol{F},\nabla,\boldsymbol{W}_0,\alpha,\gamma,\lambda,\theta,\varepsilon$

输出：$\boldsymbol{W}$

1. 初始化 $\mu_0=0,\mu_1=1,\boldsymbol{W}_1=\boldsymbol{W}_0,lr=0.1$
2. $\boldsymbol{T}=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{U}+\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U},\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{T},\nabla=\nabla\geq\theta$
3. FOR $iter=0,1,2,\cdots$, DO
4. $\quad\boldsymbol{V}=\boldsymbol{W}_1+(\mu_0-1)(\boldsymbol{W}_1-\boldsymbol{W}_0)/\mu_1$
5. $\quad f'(\boldsymbol{W})=\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{W}_1-\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}+\gamma\nabla+\alpha\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{W}_1\boldsymbol{L}$
6. $\quad$LOOP
7. $\quad\boldsymbol{U}=\boldsymbol{V}-1/lr\,f'(\boldsymbol{W})$
8. $\quad$FOR each row $\boldsymbol{U}_{i*}$ of $\boldsymbol{U}$ DO
9. $\quad\boldsymbol{W}_{i*}=S_{\lambda/lr}(\boldsymbol{U}_{i*})$//使用式(17)解决
10. $\quad$ENDFOR
11. $\quad\boldsymbol{W}=\boldsymbol{W}-diag(\boldsymbol{W}),\boldsymbol{W}=\max(\boldsymbol{W},0)$
12. $\quad$IF $f(\boldsymbol{W})\leq G_{lr,\boldsymbol{V}}(\boldsymbol{W})$ THEN
13. $\quad\quad$break
14. $\quad$ENDIF
15. $\quad lr=2\times lr$
16. $\quad$ENDLOOP
17. $\quad$Set $funVal(iter)=f(\boldsymbol{W})+\lambda\|\boldsymbol{W}\|_{2,1}$
18. $\quad$IF $|funVal(iter)-funVal(iter-1)|\leq\varepsilon$ THEN
19. $\quad\quad$break
20. $\quad$ENDIF
21. $\quad\boldsymbol{W}_0=\boldsymbol{W}_1$
22. $\quad\boldsymbol{W}_1=\boldsymbol{W}$
23. $\quad\mu_0=\mu_1$
24. $\quad\mu_1=(1+\sqrt{1+4\mu_1^2})/2$
25. ENDFOR

算法中第 3 到第 25 行描述了用 Nesterov 方法解决优化问题式(8)，第 6 到第 16 行描述了用 Armijo-Goldstein 规则通过线搜索得到学习率 lr，第 24 行 μ_1 的值依据[33]所提到的方法得到，这里，μ_0 和 μ_1 均为 Nesterov 梯度下降法中计算步长的辅助变量. 通过该算法可以解决我们模型的优化问题，并通过式(9)计算每条推文的重要性以形成摘要.

5.2 收敛性及时间复杂度分析

收敛性：对于 Nesterov 加速梯度下降算法，通过在搜索点 $\boldsymbol{V}_t$ 执行梯度下降而不是在近似点 $\boldsymbol{W}_t$ 执行梯度下降，收敛率可以达到 $O(1/k^2)$，其中 k 为迭代次数，同时可以得到算法的理论迭代次数为 $O(1/\sqrt{\varepsilon})$，其中 ε 表示收敛阈值. 该结论可以通过如

下理论得到.

理论 1.　根据文献[32]，假设$\{\boldsymbol{W}_t\}$通过算法 1 得到，那么对于任意 $t\geq 1$ 都有：

$$f(\boldsymbol{W}_{t+1})-f(\boldsymbol{W}^*)\leq\frac{2\hat{L}_f\|\boldsymbol{W}^*-\boldsymbol{W}_1\|_{\mathrm{F}}^2}{(t+1)^2},$$

其中，$\hat{L}_f=\max(2L_f,L_0)$，L_0 是 $f(\boldsymbol{W})$的 Lipschitz 连续梯度 L_f的初始值. 该理论的详细证明部分可以参考文献[32].

时间复杂度：基于上述理论，算法会在 $1/\sqrt{\varepsilon}$ 次迭代后收敛. 接下来只需讨论每次迭代需要花费的时间. 目标函数式(8)包含四个部分，给定 n 条推特，m 个特征(即词表)，对于目标函数中的平方损失即 $\|\boldsymbol{S}-\boldsymbol{SW}\|_{\mathrm{F}}^2$，在计算目标函数值和相应的梯度值时，需要进行 $O(n^2m)$次浮点运算；同样对于社会正则项即 $tr(\boldsymbol{SWLW}^{\mathrm{T}}\boldsymbol{S}^{\mathrm{T}})$，需要进行 $O(n^2m)$次浮点运算；对于多样性正则项即 $tr(\nabla^{\mathrm{T}}\boldsymbol{W})$，需要进行 $O(n^2)$次浮点运算；对于 $\ell_{2,1}$范数正则，通过欧几里得投影方法可以达到 $O(n^2)$的时间复杂度. 综合以上四个部分，整个算法的时间复杂度为

$$O\left(\frac{1}{\sqrt{\varepsilon}}(n^2m+n^2m+n^2+n^2)\right)\approx O\left(\frac{1}{\sqrt{\varepsilon}}(n^2m)\right).$$

6　实验结果与分析

6.1　专家摘要制作过程

为了评估提出的模型产生摘要的内容质量，我们邀请 24 位志愿者为 3.2 节中 12 个话题集制作了专家摘要. 对于每个话题集，请四位志愿者在我们所给出的《Twitter 话题专家摘要制作指南》(以下简称《指南》)指引下筛选 25 条推文形成专家摘要，这样每个话题集最终拥有四个专家摘要. 其中《指南》内容大致如下：

(1) 每个话题集需要抽取 25 条推文形成摘要，这些推文能够尽可能覆盖话题包含的各个方面，并且避免冗余信息.

(2) 标注者可以通过关键词以及时间线索在网上(百度或谷歌等)搜索相应话题，加深对话题的了解，辅助专家摘要的选择，比如 2011 年 5 月 1 日，本拉登被击毙. 另外，为避免遗漏重要信息，需要标注者在选择摘要推文时能够浏览话题集中的全部推文.

(3) 对于两条描述内容大致相同的推文，选择包含信息更多的推文，比如都是描述“本拉登被葬于大海”的推文，选择包含时间、地点等信息更丰富的推文；同时，选择表达更规范的推文，包括语法是否正确、拼写是否有误等.

专家摘要质量的好坏直接影响了性能的评价，一个低质量的专家摘要甚至会降低我们模型效果的可信度，所以对于专家摘要的评价是必要的环节. 我们让三位没有参与专家摘要选择的志愿者分别对所有专家摘要作评价，并且依据全面性、多样性、书写质量这三个方面进行综合打分，打分范围为[1,5]. 在 25 条推文中，若只有 0～6 条推文是令人满意的，则打分为 1；若有 7～12 条推文是令人满意的，则打分为 2，依次类推，13～18 打分为 3，19～24 打分为 4；若所有推文质量都很高，则打分为 5. 分数越高说明摘要质量越高. 我们保留那些至少两位志愿者打分在 3 分及以上的专家摘要，并要求对那些低质量的专家摘要做修改，直到符合要求. 同时，与国际评测中专家摘要水平的上限相比，6.3 节中表 2 专家摘要互评的平均结果显示了本文的数据标注具有一定的可参考性.

6.2　评价方法

本文使用 ROUGE[34] 作为评价指标，该方法主要度量系统摘要与专家摘要的 N-gram 覆盖率. 其中使用 ROUGE-N 特别是 ROUGE-1 和 ROUGE-2，以及 ROUGE-SU4 作为评价指标. 由于每种评价指标都包含准确率 P(Precise)、召回率 R(Recall)和 F 值(F-Measure)，鉴于 F-Measure 综合考虑了准确率和召回率，本文在实验结果中只展示 F-Measure. 下面我们给出 ROUGE-N 的计算公式：

$$P=\frac{\sum_{t\in S_{\mathrm{ref}}}\sum_{gram_n\in t}Count_{\mathrm{match}}(gram_n)}{\sum_{gram_n\in G}Count(gram_n)},$$

$$R=\frac{\sum_{t\in S_{\mathrm{ref}}}\sum_{gram_n\in t}Count_{\mathrm{match}}(gram_n)}{\sum_{t\in S_{\mathrm{ref}}}\sum_{gram_n\in t}Count(gram_n)},$$

$$F=\frac{2PR}{P+R}.$$

其中，n 是指 N-gram 中的单词数，S_{ref}是指专家摘要集，G是指我们模型产生的系统摘要，$Count_{\mathrm{match}}(gram_n)$是指系统摘要和专家摘要匹配到的 N-gram 数量，$Count(gram_n)$是指摘要中的 N-gram 数量.

ROUGE-SU4 是 ROUGE-S 的扩展版本，ROUGE-S 主要计算系统摘要和专家摘要 Skip-Bigram 的覆盖率，其中 Skip-Bigram 可以指任意单词对，而且单词

对中单词的相对位置与原句子中的相对位置保持一致.为了避免匹配到“the the”、“of in”类似的单词对,可以设置最大跳跃距离.由于系统摘要与专家摘要可能没有共同出现的 Skip-Bigram,所以增加了平滑措施,即在计算 ROUGE-S 时同时考察 unigram 的覆盖率,此时称为 ROUGE-SU.倘若设置最大跳跃距离为 4,则 ROUGE-SU 又可以转换为 ROUGE-SU4.

参考标准评测任务 DUC/TAC 的评估机制,这里,采用 ROUGE 标准选项①来设置 ROUGE 评价参数.

6.3 对比实验设计与实验结果分析

由于本文提出的摘要模型属于抽取式、无监督的方法,这里,只与相同类型的摘要方法作对比.在所有对比实验中,系统输出摘要长度均为 25 条推文.表 2 中展示了专家摘要互评(人类摘要的平均水平)、所有对比实验、本文模型的一些退化模型以及本文模型的性能评价结果,其中:

(1) Expert 表示专家摘要互评的平均结果;

(2) Random 随机选择 25 条推文形成摘要;

(3) Centroid[3] 通过计算每条推文和整个话题集伪质心的相似度作为对每条推文的重要性评估,其中伪质心计算公式为式(18),相似度的计算公式为式(19);

$$\boldsymbol{c}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{t}_i \tag{18}$$

$$Similarity(\boldsymbol{c},\boldsymbol{t}_i)=\frac{\boldsymbol{c}\cdot\boldsymbol{t}_i}{|\boldsymbol{c}|\cdot|\boldsymbol{t}_i|} \tag{19}$$

其中,$\boldsymbol{c}$ 表示伪质心向量,$\boldsymbol{t}_i$ 为每条推文的向量表示,n 是特定话题集的推文数目.

(4) LexRank[4] 构建了一个无向图,节点为推文,根据推文之间的相似度确定是否建边,以及边权的数值,具体见式(20).其中 δ 为判断是否建边的阈值,大于这个阈值,则推文之间建立一条边,否则不建边.并利用类似 PageRank 的思想更新推文的分数,直到收敛,参见式(21);

$$Matrix_{ij}=\begin{cases}1, & 若\ Similarity(\boldsymbol{t}_i,\boldsymbol{t}_j)\geqslant\delta\\ 0, & 否则\end{cases} \tag{20}$$

$$Score(\boldsymbol{t}_i)=(1-p)\frac{1}{n}+p\sum_{j=1}^{n}Score(\boldsymbol{t}_j)\frac{Matrix_{ji}}{\sum_{k=1}^{n}Matrix_{jk}} \tag{21}$$

本文取 $\delta=0.1$,$p=0.85$,$Matrix_{ij}=1$ 表示推文 $\boldsymbol{t}_i$ 和推文 $\boldsymbol{t}_j$ 有边相连,$Similarity(\boldsymbol{t}_i,\boldsymbol{t}_j)$ 表示推文 $\boldsymbol{t}_i$ 和推文 $\boldsymbol{t}_j$ 的余弦相似度.

(5) LSA[35] 利用奇异值分解方法(Singular Value Decomposition,SVD)分解文本矩阵 $\boldsymbol{S}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}$,其中 $\boldsymbol{U}\in\mathbb{R}^{m\times k}$,$\boldsymbol{\Sigma}\in\mathbb{R}^{k\times k}$,$\boldsymbol{V}\in\mathbb{R}^{k\times n}$,$k$ 在本文取 25.利用 SVD 抽取摘要的过程如下:① 取最大奇异值所对应的右特征向量 $\boldsymbol{V}_{i*}$,取该特征向量中最大值 $\boldsymbol{V}_{ij}$ 所对应的推文 $\boldsymbol{t}_j$ 为一个摘要推文;② 按照奇异值从大到小依次选择相应的推文.

(6) NNMF[36] 对文本矩阵 $\boldsymbol{S}$ 执行非负矩阵分解,即 $\boldsymbol{S}=\boldsymbol{P}\boldsymbol{T}$,其中 $\boldsymbol{P}\in\mathbb{R}^{m\times k}$,$\boldsymbol{T}\in\mathbb{R}^{k\times n}$ 接下来选取摘要的方法类似于 SVD,相当于为每个话题选择一条与该话题相关度最大的推文以形成摘要.

(7) DSDR[10] 和 MDS-Sparse[11] 属于两种基于稀疏重构的摘要方法.

(8) SNSR-div,SNSR-sparse,SNSR-social 是我们提出模型 SNSR 的三个退化模型,“-”表示从模型中去掉相应的多样性正则项、组稀疏性正则项、社会性正则项.图 3 中“-”的设置规则与此类似.

表 2 所有方法在推特数据集的实验性能

模型	ROUGE-1	ROUGE-2	ROUGE-SU4
Expert	0.47814	0.16337	0.20389
Random	0.41701	0.09439	0.14231
Centroid	0.38190	0.12384	0.15668
LexRank	0.42046	0.13273	0.17366
LSA	0.43474	0.13023	0.16625
NNMF	0.43784	0.13321	0.17433
DSDR	0.43236	0.12946	0.16521
MDS-Sparse	0.42240	0.10060	0.14666
SNSR-div	0.40191	0.12940	0.15894
SNSR-sparse	0.43327	0.13692	0.17749
SNSR-social	0.43236	0.10271	0.15379
SNSR	0.44887	0.13882	0.18147

通过观察表 2 可以得到以下结论:

(1) 除了比专家摘要互评结果低一些,我们模型的性能比其他对比实验都要高.然而值得注意的是,所有模型的性能普遍比较高,特别是对于 ROUGE-1 这个性能指标.导致这种现象的原因可能有两点:① 我们的任务是针对特定话题的推特摘要研究,而且是根据 Hashtag 或关键词筛选的推文,这就导致推文之间难免会保持一致的内容;② 经过 3.2 节数据清洗过程,数据集相对比较干净,进一步说明我们执行了一个比较有效的预处理过程.

(2) 在所有的对比实验中,基于矩阵分解的方法,特别是 NNMF,显示出比较不错的性能.其中的原因可能有两个方面:① NNMF 同样可以被看成是基于稀疏重构的方法,而且 NNMF 方法所体现的

① ROUGE-1.5.5.pl -e data -n 4 -2 4 -u -c 95 -r 1000 -a -d -m -f A -p 0.5 -t 0

思想类似于 Li 等人[37]所做的工作，其主要思想是利用方面向量(aspect vector)来重构原始词向量空间，其中的“方面”可以理解为话题，“方面向量”可以理解为每个话题的向量表示，在 NNMF 中对应于 $\boldsymbol{P}$ 矩阵的列向量；② 通过挖掘子话题，在一定程度上解决了全面性和多样性问题.

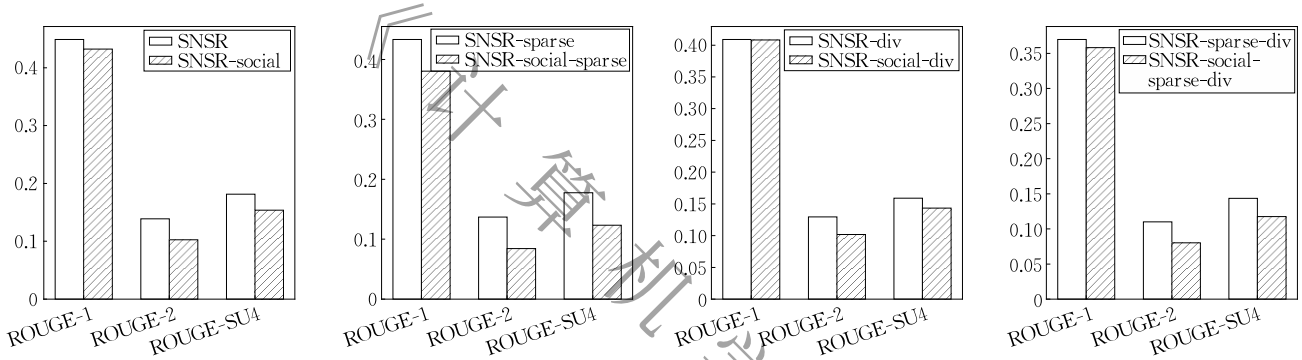
(3) 特别地，虽然我们模型和 NNMF 都能在一定程度上解决全面性和多样性问题，但是我们模型的效果更好，可能的原因有两点：① 针对推特语料情景，我们模型通过引入社会项正则及多样性正则项，一定程度上弥补了推特文本简短、嘈杂、不规范的问题；② 通过引入组稀疏正则项，使得我们的模型是从语料级别直接筛选得到摘要，而不是从话题级别即先选择每个话题的重要推文，再组合成摘要.

(4) 通过和三个退化模型相比，进一步说明了社会项正则和多样性正则对于我们的模型是有效的. 而且观察可知，社会项正则的引入有利于提高 ROUGE-2 和 ROUGE-SU4 评价指标，多样性正则有利于提高 ROUGE-1 评价指标.

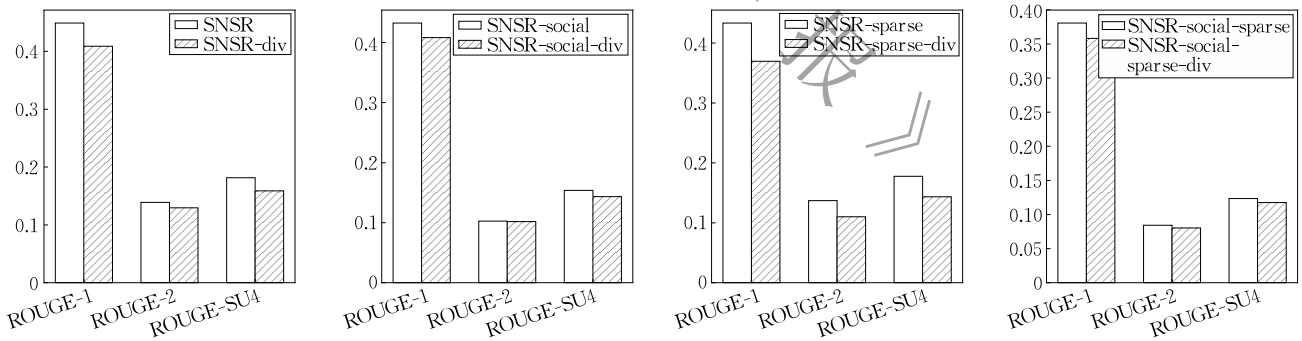
6.4　不同正则项对实验性能的影响

为了进一步评估社会项正则的效果，我们设计四组对比实验，参见图 3(a)，每组对比实验呈现两个模型，分别表示加社会正则项与不加社会正则项的情况. 除此之外，我们采用类似的方法来评价多样性正则和组稀疏性正则的作用，参见图 3(b)和图 3(c). 通过分析实验结果，我们得出以下几点结论：

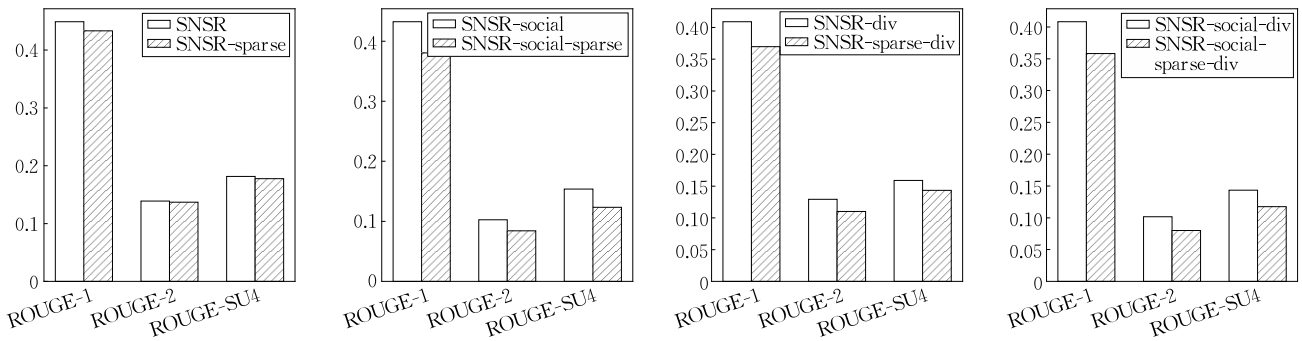
(1) 对于每一项正则，通过分析每一组对比实验，我们可以计算增加相应正则项对于 ROUGE-1、



(a) 社会项正则的影响



(b) 多样性正则的影响



(c) 组稀疏性正则的影响

图 3　三种正则项对推特摘要实验性能的影响

ROUGE-2、ROUGE-SU4 的平均增长百分比. 通过观察表 3 可知,不同正则项对三个评价指标的影响,社会正则项影响最大,其次是组稀疏性正则,最后是多样性正则. 特别地,对于 ROUGE-2 和 ROUGE-SU4 这两个评价指标,增加社会正则项会分别提高 46.01%和 23.69%,说明增加社会正则项更倾向于选择那些接近专家摘要的推文.

表 3 不同正则项相对于退化模型的平均增长百分点

正则项	ROUGE-1/%	ROUGE-2/%	ROUGE-SU4/%
Social	4.84	46.01	23.69
Sparse	9.98	21.03	14.87
Diversity	10.27	6.34	12.51

(2)值得注意的是,多样性正则对 ROUGE-1 的影响要大于其它两项正则,这种现象说明,去冗余确实可以减少重复的单词,并且增加了同专家摘要的单词覆盖率.

6.5 参数设置与调节

本文实验中主要需分析四个参数,分别为社会正则 α、多样性正则 γ、组稀疏正则 λ 以及多样性阈值参数 θ. 本文首先在[0,1]范围内设置步长 0.1 来调节每个参数,然后观察变化趋势,再通过不断调节步长以及区间范围来尽可能寻找最优参数. 比如对于参数 α,其在[0,1]范围内呈现出[0,0.1]内快速上升而[0.1,1]内平稳变化的趋势,然后通过在[0,0.1]内设置步长 0.01(图 4 显示 ROUGE-1, ROUGE-2 和 ROUGE-SU4 随 α 的变化趋势),可知当 $\alpha=0.03$ 时取值最大. 对于参数 λ,其在[0,1]内呈现上升的趋势,在 $\lambda=1$ 时取到最大值,然后扩大区间范围为[0,2],步长设置为 0.1(图 5 显示三个评价指标随 λ 的变化趋势),最后本文设置 $\lambda=1$. 类似的,设置 $\gamma=1$.

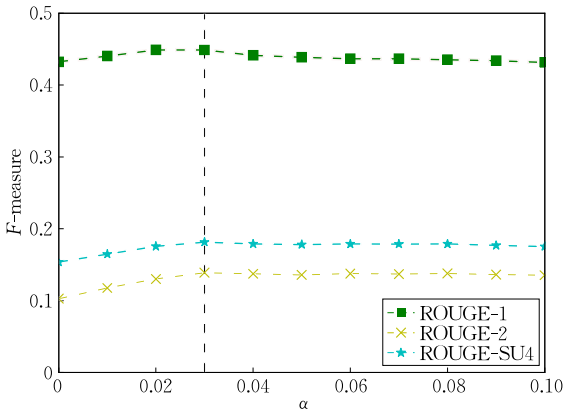
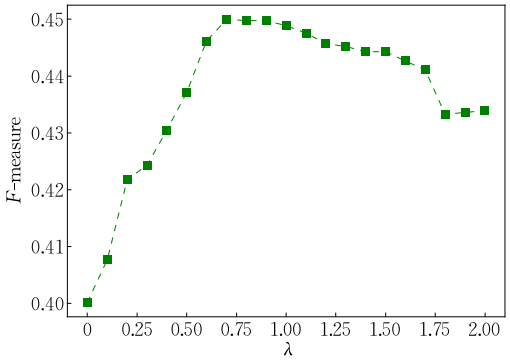
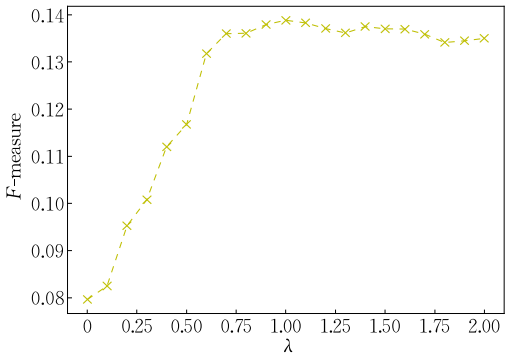


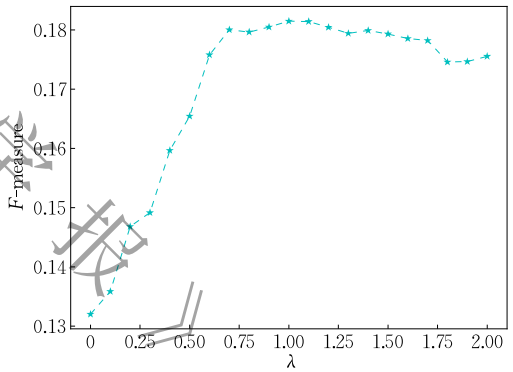
图 4 α 对实验性能的影响



(a) ROUGE-1



(b) ROUGE-2



(c) ROUGE-SU4

图 5 λ 对三个评价指标的影响

通过统计数据中余弦相似度值的分布,发现大多数的推文对相似度分布在[0,0.1]区间内,极少部分分布在(0.1,1]区间内. 而且在实验中通过设置步长,即在[0,1]范围内,设置步长为 0.1,可以得到当 $\theta=0.1$ 时性能最好,这也进一步验证了本文对于数据分布的观察与分析.

6.6 时间复杂度分析

从理论层面讲,本文方法 SNSR 的时间复杂度约为 $O\left(\frac{1}{\sqrt{\epsilon}}(n^2m)\right)$. 对比实验的时间复杂度不再赘述,详情可以参见具体的引文. 为了进一步对比具体

的时间开销，我们在 CPU 为 Intel(R) Core(TM) i7-6700、主频 3.4GHz、内存 16GB、操作系统 Win10、编程平台 matlab2014a 的环境下布置运行了对比实验. 由于机器的浮点运算每次都有一定的差异，本文列出了三次的实验运行时间 *bb*(以秒为单位)，如表 4 所示，可以观察到在基于稀疏重构系列的对比实验(DSDR 及 MDS-Sparse)中，SNSR 算法的运行效率具有优势.

表 4　对比实验的时间消耗分析

批次	模型							
	Random	Centroid	LexRank	LSA	NNMF	DSDR	MDS-Sparse	SNSR
1	0.0010	5.0817	18.1822	4.1769	257.3272	1.2054E+4	2.0966E+4	805.8405
2	0.0011	4.9196	17.9158	3.7036	249.9049	1.0501E+4	2.0024E+4	653.7737
3	0.0009	4.9174	17.6496	3.8883	249.4944	1.1564E+4	2.1029E+4	724.7493

7　总结与展望

本文针对特定话题的社会媒体文本信息做推特摘要研究，受压缩感知和社会学理论的启发，提出了基于稀疏重构和社会网络拓扑结构的推特摘要方法. 探索社会网络推特层面的动态信息，通过社会学理论，即表达一致性和表达传染性，把用户之间的网络结构转换为推文之间的网络结构，进而捕捉了推文之间更多潜在的语义线索. 并将其建模为社会正则项，整合到基于稀疏重构的推特摘要优化框架中；其次，建模由于社会网络传播带来的强冗余信息为多样性正则；挖掘了推特摘要的隐含组模式，建模为组稀疏正则项. 在稀疏重构的原则下，从三个不同角度诠释我们的模型：面向话题推特摘要内容的覆盖性、稀疏性及多样性. 同时，提出基于 Nesterov 加速梯度下降的推特摘要算法. 另外，由于缺乏标准评测语料，本文构建了 12 个话题的推特摘要评测数据集，在其上的实验证明了我们方法的有效性. 通过进一步针对不同正则项设计对比实验，验证了每一个正则项都是有效果的，尤其是社会正则项影响最大.

未来工作中，我们将尽可能挖掘社交网络的其它特有属性来辅助摘要研究. 通过设置对比实验，我们发现基于矩阵分解的摘要方法，尤其是 NNMF 显示出了不错的性能. 这种现象启发我们，或许话题检测和文本摘要同时进行，可以互相促进各自任务的性能. 因此，如何结合话题检测、文本摘要、社交网络特性这三点，是未来工作的一个研究方向.

致　谢　感谢各位审稿专家对本文工作的指导！

参　考　文　献

[1] Inouye D, Kalita J K. Comparing Twitter summarization algorithms for multiple post summaries//Proceedings of the Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom). Boston, USA, 2011: 298-306

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

[3] Radev D R, Blair-Goldensohn S, Zhang Z. Experiments in single and multi-document summarization using MEAD//Proceedings of the Document Understanding Conference Workshop. Ann Arbor, USA, 2001: 12-20

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

[5] Mihalcea R, Tarau P. TextRank: Bringing order into text//Proceedings of the Empirical Methods in Natural Language Processing. Barcelona, Spain, 2004: 404-411

[6] Chang Y, Tang J, Yin D, et al. Timeline summarization from social media with life cycle models//Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI). New York, USA, 2016: 3698-3704

[7] Liu X, Li Y, Wei F, et al. Graph-based multi-tweet summarization using social signals//Proceedings of the International Conference on Computational Linguistics(COLING). Mumbai, India, 2012: 1699-1714

[8] Chang Y, Wang X, Mei Q, et al. Towards Twitter context summarization with user influence models//Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. Rome, Italy, 2013: 527-536

[9] Duan Y, Chen Z, Wei F, et al. Twitter topic summarization by ranking tweets using social influence and content quality//Proceedings of the International Conference on Computational Linguistics(COLING). Mumbai, India, 2012: 763-780

[10] He Z, Chen C, Bu J, et al. Document summarization based on data reconstruction//Proceedings of the AAAI Conference on Artificial Intelligence. Toronto, Canada, 2012: 620-626

[11] Liu H, Yu H, Deng Z-H. Multi-document summarization based on two-level sparse representation model//Proceedings of the AAAI Conference on Artificial Intelligence. Austin, USA, 2015: 196-202

[12] Yao J-G, Wan X, Xiao J. Compressive document summarization via sparse optimization//Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI). Buenos Aires, Argentina, 2015: 1376-1382

[13] Cai X, Li W, Ouyang Y, et al. Simultaneous ranking and clustering of sentences: A reinforcement approach to multi-document summarization//Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, China, 2010: 134-142

[14] Wang D, Zhu S, Li T, et al. Integrating document clustering and multi-document summarization. ACM Transactions on Knowledge Discovery from Data, 2011, 5(3): 1-26

[15] Shen C, Li T, Ding C H. Integrating clustering and multi-document summarization by bi-mixture probabilistic latent semantic analysis (PLSA) with sentence bases//Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco, USA, 2011: 914-920

[16] Wang D, Zhu S, Li T, et al. Multi-document summarization using sentence-based topic models//Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. Singapore, 2009: 297-300

[17] Gao D, Li W, You O, et al. LDA-based topic formation and topic-sentence reinforcement for graph-based multi-document summarization//Proceedings of the Asia Information Retrieval Symposium (AIRS). Tianjin, China, 2012: 376-385

[18] Litvak M, Vanetik N, Liu C, et al. Improving summarization quality with topic modeling//Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. Melbourne, Australia, 2015: 39-47

[19] Sharifi B, Hutton M-A, Kalita J. Summarizing microblogs automatically//Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL). Los Angeles, USA, 2010: 685-688

[20] Nichols J, Mahmud J, Drews C. Summarizing sporting events using Twitter//Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces. Lisbon, Portugal, 2012: 189-198

[21] Alsaedi N, Burnap P, Rana O. Automatic summarization of real world events using Twitter//Proceedings of the 10th International AAAI Conference on Web and Social Media. Munich, Germany, 2016: 511-514

[22] Hu X, Tang L, Tang J, et al. Exploiting social relations for sentiment analysis in microblogging//Proceedings of the 6th ACM International Conference on Web Search and Data Mining. Rome, Italy, 2013: 537-546

[23] Wang X, Wang Y, Zuo W, et al. Exploring social context for topic identification in short and noisy texts//Proceedings of the AAAI Conference on Artificial Intelligence. Austin, USA, 2015:1868-1874

[24] Bi B, Tian Y, Sismanis Y, et al. Scalable topic-specific influence analysis on microblogs//Proceedings of the 7th ACM International Conference on Web Search and Data Mining. New York, USA, 2014: 513-522

[25] He X, Rekatsinas T, Foulds J, et al. Hawkes topic: A joint model for network inference and topic modeling from text-based cascades//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 871-880

[26] Abelson R P. Whatever became of consistency theory? Personality and Social Psychology Bulletin, 1983, 9(1): 37-54

[27] Shalizi C R, Thomas A C. Homophily and contagion are generically confounded in observational social network studies. Sociological Methods and Research, 2011, 40(2): 211-239

[28] Harrigan N, Achananuparp P, Lim E-P. Influentials, novelty, and social contagion: The viral power of average friends, close communities, and old news. Social Networks, 2012, 34(4): 470-480

[29] Ye J, Liu J. Sparse methods for biomedical data. ACM SIGKDD Explorations Newsletter, 2012, 14(1): 4-15

[30] Ji S, Ye J. An accelerated gradient method for trace norm minimization//Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada, 2009: 457-464

[31] Nesterov Y. Introductory Lectures on Convex Optimization: A basic course. Louvain-la-Neuve, Belgium: Université Catholique de Louvain (UCL), 2004

[32] Boyd S, Vandenberghe L. Convex Optimization. Cambridge, UK: Cambridge University Press, 2004

[33] Liu J, Ji S, Ye J. Multi-task feature learning via efficient $l_{2,1}$-norm minimization//Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence. Montreal, Canada, 2009: 339-348

[34] Lin C-Y. ROUGE: A package for automatic evaluation of summaries//Proceedings of the ACL-04 Workshop on Text Summarization Branches Out. Barcelona, Spain, 2004

[35] Gong Y, Liu X. Generic text summarization using relevance measure and latent semantic analysis//Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New Orleans, USA, 2001: 19-25

[36] Park S, Lee J-H, Kim D-H, et al. Multi-document summarization based on cluster using non-negative matrix factorization //Proceedings of the SOFSEM 2007: Theory and Practice of Computer Science, 2007: 761-770

[37] Li P, Wang Z, Lam W, et al. Salience estimation via variational auto-encoders for multi-document summarization//Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 3497-3503

HE Rui-Fang, Ph. D., associate professor. Her main research interests include natural language processing, social media mining, and machine learning etc.

DUAN Xing-Yi, M. S. His main research interests include natural language processing, automatic document summarization.

ZHANG Xue-Fei, M. S. Her main research interests are natural language processing and social computing.

ZHAO Wen-Li, M. S. candidate. Her main research interest is natural language processing.

Background

Twitter summarization belongs to the field of natural language processing and social media text mining. It aims to help people quickly grasp key information from a volume of noisy, informal and short messages. Existing Twitter summarization methods are mainly divided into three categories: (1) pure text based; (2) static social property based and (3) dynamic social property based. Some researches exploit the user-level social network, and generally suppose that a high authority user posts a high quality tweet. However, none of the existing methods mine the latent tweet-level network, ignoring that information can flow along the network.

The previous methods rarely consider the data sparsity, the strong social redundancy and relations between tweets explicitly, ignoring that information can spread along the social network. Inspired by compressive sensing and social theories, we propose a novel approach for Twitter summarization by integrating Social Network and Sparse Reconstruction (SNSR). We explore whether social relations (expression consistency and expression contagion) can help Twitter summarization, modeling relations between tweets described as the social regularization and integrating it into the group sparse optimization framework. It conducts a sparse reconstruction process by selecting tweets that can best reconstruct the original tweets, with considering coverage and sparsity. We simultaneously design the diversity regularization to remove the strong social redundancy. In particular, we present a mathematical optimization formulation and develop an efficient Twitter summarization algorithm with Nesterov's accelerated gradient descent.

Meanwhile, due to the lack of public corpus, we construct the gold standard Twitter summary datasets for 12 different topics. Experimental results on this datasets show the effectiveness of our approach is effective. The experimental results show that our model is effective.

Our research group has devoted a lot of effort in summarization. We have some papers published in the respectable journals, such as *Journal of Software*, *Information Science and World Wide Web Journal* and so on.

This work is supported by the National Natural Science Foundation of China (61472277) and the Tianjin Natural Science Foundation (18JCYBJC15500).