

基于带权文本矩阵分解-信息熵模型的新闻评论摘要

国玉静*, 姬东鸿

(武汉大学 计算机学院, 武汉 430072)

(* 通信作者电子邮箱 yujing.guo@whu.edu.cn)

摘 要:针对新闻的评论摘要的抽取问题,提出了一种将带权文本矩阵分解(WTMF)与信息熵结合的社交媒体评论自动抽取方法。该方法对微博(tweets)和news信息构建基于异质图的WTMF模型,解决短文本特征稀疏问题,保障信息的相似性;根据tweet的特征分布,构建基于特征的二元信息熵和连续信息熵,保证信息的多样性。最后依据子模属性,设计基于贪心的抽样算法,获取优化问题近似最优解。实验结果表明,WTMF与信息熵结合的方法能有效提高社交媒体摘要性评论抽取的性能,在ROUGE2上召回率和F1值分别达到0.40074和0.27330。与潜在狄利克雷分配(LDA)扩展模型——基于位的主体模型(BTM)相比,分别提高了0.05和0.03,有效地提高了新闻评论摘要质量。

关键词:优化问题;带权文本矩阵分解模型;异质图模型;信息熵;子模属性

中图分类号: TP391.1 **文献标志码:** A

Summary extraction of news comments based on weighed textual matrix factorization and information entropy model

GUO Yujing*, JI Donghong

(School of Computer Science, Wuhan University, Wuhan, Hubei 430072, China)

Abstract: This paper addressed to select the most interesting and useful comments for an online news article. In summary of comments for news extraction problem, a new way was introduced, and it was proved to be effective in the social media comments automatic extraction with the combination of Weighed Textual Matrix Factorization (WTMF) and information entropy. The construction of information for tweets and news was based on heterogeneous graph WTMF model which solved the sparse problems of short text and maintained the similarity of information. Meanwhile, according to tweet character distribution, binary entropy and continuous entropy were built to guarantee the diversity of information. Last, according to the characteristics of submodularity, a greedy algorithm was designed to get an approximate optimal solution for the optimization problems. The experimental results show that, the method with combination of WTMF and information entropy can improve the extraction performance of summary of comments for social media effectively. The recall rate and F1 value on the ROUGE2 respectively reaches 0.40074 and 0.27330, which is increased by 0.05 and 0.03 in comparison of the Latent Dirichlet Allocation (LDA) extended model—Biterm Topic Model (BTM). The proposed model improves the quality of news summary of comments effectively.

Key words: optimization problem; Weighed Textual Matrix Factorization (WTMF) model; heterogeneous graph model; information entropy; submodularity

0 引言

新闻事件爆发,人们希望了解到其客观情况以及普遍评价。然而,传统新闻报道普遍较官方、正统,并不能完整反映网民们的观点。微博(tweets)作为一种轻便式的社交媒体,使得新闻事件在微博上的传播相比传统媒体更加实时、快速、客观。

但是微博数据量庞大,根据相关公开数据,目前Twitter上,每天更新的微博总数已超过5千万。如此庞大的数据中存在着大量冗余、无用信息,从其中抽取针对某一新闻事件的评论摘要,成了自动文摘领域一个炙手可热的研究课题。

但是由于微博的文本内容过短(平均长度只有14个

字),信息表述不尽完整,以词组、短语为特征的向量十分稀疏,导致很多传统的自然语言处理(Natural Language Processing, NLP)方法在微博这种短文本上并不适用。例如在语义分析领域,文献[1]表明,在长影评的数据集上,可达到87.5%的准确率,而文献[2]表明在句子级,准确率下降至75%。更甚者在一些NLP领域,短文本不能生成任何结果。例如下面的微博信息:

Pray for WHUer ...

在常规的事件抽取系统中,此条微博缺少关键性内容信息,其并不能被发现为“武汉大学樱花节”话题。为使传统NLP工具适用于微博信息,本文对微博信息,进行更深层的语义挖掘,将短文本向量进行语义扩充,从而解决数据稀疏问

收稿日期:2014-04-17;修回日期:2014-06-18。

基金项目:国家自然科学基金重点项目(61133012);国家自然科学基金面上项目(61173062)。

作者简介:国玉静(1989-),女,天津人,硕士研究生,主要研究方向:自然语言处理、数据挖掘;姬东鸿(1966-),男,北京人,教授,博士,主要研究方向:自然语言处理、数据挖掘、智能信息处理、搜索技术、机器学习、生物信息处理、词汇语义学、现代语言学、认知语言学。

题。

另一方面,对于新闻相关微博的选取,除了要求微博信息与新闻的高相关性,还需要保证微博信息间的多样性、代表性。本文将信息抽取问题转化为极值优化问题,而多样性的衡量则采用信息熵的概念,保证数据信息量最大化。

1 研究现状

目前,相关学者在社交媒体的文摘抽取方面作了诸多研究,其目的是对于给的主题或新闻,抽取出一系列有代表性的信息。对于微博这种短文本摘要抽取,其稀疏性问题的解决,相关学者也作了诸多尝试。在文献[3]中,借助新闻信息对 tweet 信息进行辅助,生成新闻和 tweet 的联合摘要,并没有生成独立 tweet 摘要。在文献[4]中,借助语料库 WordNet 对短文本进行上下文扩展。此种方法效率较低,并且对外界信息依赖较大,只适用于离线系统,难以满足对于实时性要求较高的系统需求。文献[5-6]借助社交网络信息,如用户个人信息,兴趣爱好、朋友圈、历史微博等信息对 tweet 信息进行辅助,对数据集的数据量要求较高,并且涉及到用户隐私问题。

潜语义模型,其突破了表面文本的限制,对文本库进行建模,挖掘文本内部的潜在语义信息,将短文本表示成一个主题分布,从而弱化了数据稀疏问题的影响。潜语义分析(Latent Semantic Analysis, LSA)是通过矩阵进行奇异值分解(Singular Value Decomposition, SVD),生成文本特征向量^[7]。而概率潜语义分析(probabilistic Latent Semantic Analysis, pLSA)^[8]则在 LSA 的基础上,引入概率模型。其隐含的多项式分布假设,相对 LSA 的高斯分布假设,更加符合文本特征。主题模型——潜在狄利克雷分配(Latent Dirichlet Allocation, LDA)模型^[9]是一种生成模型,其主要思想为:每篇文档都由一些主题做构造的概率分布,而每个主题又由单词的概率分布构成。此类算法在长文本主题分析上取得了卓越成绩,然而在短文本上并不奏效。本文根据 tweet 的特有特征:图片信息、话题标签、命名实体、时间戳,构建 tweet news 间的异质图。将异质图与带权文本矩阵分解(Weighted Textual Matrix Factorization, WTMF)算法与扩展短文本语义相结合,此算法由 Guo 等^[10]在 2012 年首次提出,其将缺失词汇作为短文本的特征,增加文本的信息量,进而通过语义分析得到向量表示。实验结果证明其在短文本上效果优于 LDA 的算法。

由于 tweet 的信息较为短小,加之转发等行为的盛行,信息冗余度大,因此要产生一篇新闻的最佳 tweet 子集,除了要求信息精准相关,还需要内容多样。相关学者采用候选句子间进行相似度比对,此种方法仅对信息去重,并不能保证信息多样性。另一些学者采用聚类,分别抽取各个类别的子集,此种方式将不同的类别无差别看待,选取的句子不能保证高相关性。本文采用信息熵的概念,即选取的句子集合信息熵最大化,同时要求每一个元素都对候选集合的信息熵有正向贡献,存在信息增益,从而保证了信息低重复、多样化。

由于全局求解优化问题计算难度大,本文以子模属性为理论依据,构建基于贪心的迭代抽样算法,从而使问题快速得到近似最优解。

2 算法描述

本文的任务是给定一篇新闻 news 和 tweet 集合 T ,发现与新闻组相关的容量为 k 的 tweet 子集合 $S, S \in T$ 。为了证明算

法有效性,本文将多篇新闻和其相关 tweet 混合在一起,则每篇新闻除了包含与自己相关的评价 tweet,还包含了大量的间接相关的 tweet 评价以及噪声信息。

本文将每条 tweet 作为一个发言者的观点,发言者 p_i 持有一个观点 t_i ,关于此事件的讨论会存在 n 个观点,则原任务可以理解为一个事件搜集发言者的普遍并且多样的观点。如图 1 所示。

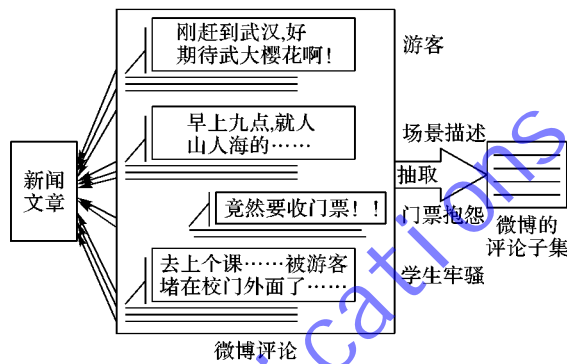


图1 任务描述

为了完成此任务,本文构建了一个层次信息评估模型,将其转化为优化问题进行求解。函数 $R()$ 用于衡量函数的 tweet 信息的高相关性。函数 $H_0()$ 用于衡量候选 tweet 信息的高多样性。文本表述为一个极值优化问题,将两者统一起来。

$$g(S) = \alpha R(S) + (1 - \alpha) H_0(S)$$

$$\text{其中: } R(S) = \sum_{q_i \in S} \text{sim}(\text{new}, q_i).$$

相关性评价由 WTMF 模型构建的潜语义,计算两者间的相似性,信息多样性由信息熵模型实现。

2.1 相关性得分

本文为了解决短文本语义缺失问题,采用 WTMF 模型,此模型为非监督算法,不需要额外的标注语料。WTMF 为弥补 tweet 文本短的缺陷,将文本外的缺失词汇也作为文本特征进行训练。缺失词汇一方面可以利用相似文本,相同特征的传递性,确定短文本的上下文信息,并扩展出短文本的正相关潜在语义;另一方面,缺失词汇也表征了与文本的负相关性,摒弃噪声信息。如图 2 所示。

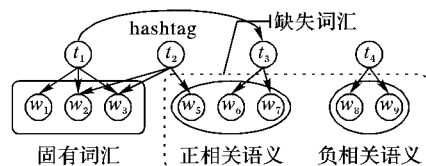


图2 WTMF 缺失词汇诠释

WTMF 将一篇语料表示为 TFIDF 的统计向量 X ,行代表 word,列代表 doc,其中 X_{ij} 是第 i 篇文档的第 j 个词的 TFIDF 值。假设 X 是大小为 $M * N$ 的向量,则将 X 分解为 P 和 Q 向量:

$$X = P^T \times Q$$

其中: P 为 $K * M$ 的向量, $P_{*,i}$ 表示 $word_i$ 在各篇文档中的语义权重; Q 为 $K * N$ 的向量, $Q_{*,j}$ 表示 doc_j 的各个语义向量。

WTMF 算法是基于 LSA 的改进版本,与 LSA, pLSA 相比, LDA 更加适合于短文本的潜语义分析。这一类的算法的核心思想都是,基于 word 与 doc 的同现。WTMF 的算法主要改进在缺失词汇的处理。LSA 算法同样考虑了缺失词汇,但将其缺失词汇和文本中固有词汇等同对待,由于短文本固有词汇较少,99% 都是缺失词汇,此时导致缺失词汇对潜在语义的

分析占主导地位,而固有词汇的影响微乎其微,导致最终的潜语义分析结果出现较大误差。如表1所示。

表1 信息偏差统计

特征编号 N	特征	原始向量	LSA 向量	信息偏差
1	WHU	1	0.98	$= \sqrt{0.0013} \approx 0.035$
2	pray	1	0.86	
3	Wuhan	0	0.05	
4	cherry	0	0.04	$\approx \sqrt{0.0004 * N} = 4.472$
5	university	0	0.02	
\vdots	\vdots	0	0.01	
50000	crowd	0	0.01	

表1中,信息偏差 R 的计算如公式所示:

$$R = \min \sqrt{\sum_i (\hat{X}_{ij} - X_{ij})^2} = \min \sqrt{\sum_{o \in O} (\hat{X}_{oj} - X_{oj})^2 + \sum_{m \in M} (\hat{X}_{mj} - X_{mj})^2}$$

其中: $o \in O$ 表示此词汇来源于文本的固有词汇, $m \in M$ 表示缺失词汇,由表达式可知,当 $n(M) \geq n(O)$ 的时候,潜语义向量的主要贡献在于缺失词汇,如表所示 $R_o = 0.035$,而 $R_m = 4.47$ 。固有词汇的信息丢失量为 $R_o - R_m \approx -R_m$ 。而WTMF将缺失词汇给以小的权值,保证固有词汇对潜语义的绝对贡献。 $R_o - WR_m$,其中 W 为较小值,如 $W = 0.01$ 。WTMF模型的最终表述为:

$$\hat{X} = P^T \times Q$$

优化问题的目标函数为:

$$\min \sum_i \sum_j W_{ij} (\hat{X}_{ij} - X_{ij})^2 + \gamma \|P\|_2^2 + \gamma \|Q\|_2^2$$

其中:

$$W_{ij} = \begin{cases} 1, & X_{ij} \neq 0 \\ w_m, & X_{ij} = 0 \end{cases}$$

WTMF主要利用 word-doc 之间的特征同现与传递,为了充分挖掘出文本的潜语义,利用 tweet 的四个特有特征构建基于图的 WTMF,即 doc-doc 的文本间特征。如图3所示。

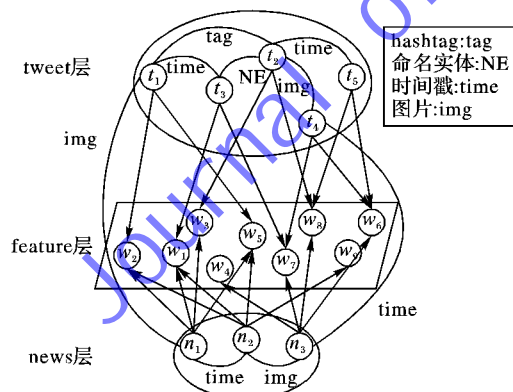


图3 基于图的 WTMF 算法示意图

2.1.1 话题标签

在 tweet 中用 # 开头来表示话题标签,一般称为 hashtag,其揭示了 tweet 的主要内容。若两篇 tweet 包含同样的 hashtag,则两篇 tweet 讨论的内容有高相似性,在两 tweet 之间建立连接。

经统计发现,只有 40% 的 tweet 包含 hashtag,倘若在中文的新浪微博中这个概率会低至 15% 左右。

因此,本文将所有的 hashtag 形成集合 U ,对于没有 hashtag 的 tweet 若包含任一集合 U 中的元素,则将其赋以间接 hashtag。经过扩展后,hashtag 的覆盖率达到 95%。为表达两者的强弱,将直接边赋值为 1,间接边为 0.6。

$$U_{tag} = \{X_{tag}^1, X_{tag}^2, \dots, X_{tag}^i, \dots, X_{tag}^n\}$$

$$R_i = \begin{cases} 1, & X_i \text{ 包含 hashtag} \\ 0, & \text{其他} \end{cases}$$

$$V_{ij} = \begin{cases} 1, & R_i \otimes R_j = 1 \\ 0.6, & R_i \otimes R_j = 0, X_i, X_j \in U_{tag} \end{cases}$$

2.1.2 命名实体

专有名词具有一定的指向性。由于 tweet 的文本较短,因此含有相同命名实体的 tweet 具有高相似性。本文借助现有工具 Stanford Named Entity Recognition (<http://nlp.stanford.edu/software/CRF-NER.shtml>) 实现,其采用条件随机场分类器训练而得。经过实验抽取,数据集集中有 70% 的 tweet 可以识别出至少一个命名实体。本文着重关注人物、组织和地点类实体。

2.1.3 多媒体信息

多媒体信息主要指视频、音频和图片,本文主要使用图片信息。一些 tweet 包含很少的文本信息,附加一张图片,此时文本不能提取出任何有用信息,图片便显得尤为重要。例如在图1的例子中,若配一张图片,如图4所示。



图4 武大樱花节拥挤情况

当其他 tweet 也引用这张图片,并指明“武大樱花节拥挤”,那么通过此图片的信息传递,即使原 tweet 中只有“Pray for WHUer...”,也同样可知此文本是关于樱花节拥挤事件。

由于图片完全相同判断局限性较大,本文利用 Simhash 算法来判别图片相似性,其相似度 $Score_{sim} \geq 90\%$,便将两条 tweet 建立连接。

2.1.4 时间戳

根据对 tweet 的观察,当事件爆发,在 tweet 上会迅速传播并且时效性很强,事件过后又迅速冷却,故时间因素对于微博的关联性有很大帮助。另一方面,由于 tweet 时间间隔低,若全部建立相似连接,将间接降低实体关联的有效性。因此,本文将发送的时间间隔小于 24 h 的前 k 条微博建立连接。本文取 $k = 4$ 。

2.2 基于异质图的 WTMF 模型

将以上四种特征移植到新闻信息构建实体相似。只有时间戳和图片信息适用,由于新闻信息篇幅较长,主题标签和命名实体会抽取大量信息,其中噪声信息过多,会将有用信息淹没。

根据以上四种特征构建实体相似图模型,此时的 WTMF

需要加入一项因子,即目标函数变为:

$$\min \sum_i \sum_j W_{ij} (\hat{X}_{ij} - X_{ij})^2 + \gamma \|P\|_2^2 + \gamma \|Q\|_2^2 + \sigma \left(\frac{Q_{*,j1} \cdot Q_{*,j2}}{|Q_{*,j1}| |Q_{*,j2}|} - 1 \right)^2$$

基于图的 WTMF 模型中,生成边情况如表 2 所示。

表 2 实体相似图模型边数统计表

特征	连接类型		
	tweet 间	news 间	tweet 与 news 间
话题标签	33 240	—	—
命名实体	24 184	—	—
图片信息	2 829	2 427	1 209
时间戳	69 776	16 880	50 816
合计	130 029	19 367	52 025

2.3 多样化函数——最大熵

给定一个集合 Q ,从中选取一个大小为 k 的子集 S ,使得子集 S 的信息量最大。本文采用最大熵的概念来评估信息多样化。即信息越丰富多样,则熵越大。若子集特征个数为 N ,当信息出现概率区域平均时,熵达到极值状态 $\lg N$ 。本文采用两种方式构建熵:一种为二元离散模型,另一种为连续模型。

2.3.1 二元离散特征熵模型

经 WTMF 模型训练后,得到的向量中的值的取值范围为 $q \in [-1, 1]$,其中 $q > 0$ 表示此特征与文本正相关;反之负相关。在抽取任务中,只关心正相关特征。在衡量信息熵时,正向相关特征越多,并且分布越均匀,信息量越大。因此将特征表示为二元值:

$$m_j = \begin{cases} 1, & q_j > 0 \\ 0, & q_j \leq 0 \end{cases}$$

则:

$$P(m_j) = \frac{\sum_{k \in S} m_{jk}}{n(S)}$$

则二元特征熵为:

$$H_0(S) = \frac{\sum_{j=1}^N -p(m_j) \lg(p(m_j))}{\lg N}$$

其中: N 为特征总数, S 为候选子集, $n(S)$ 为候选元素总数。 $\lg N$ 是对熵进行归一化,使多样性熵得分在 $[0, 1]$ 内。

2.3.2 连续特征熵模型

在二元特征中,并没有体现出元素的重要性,在连续模型中,将 m_j 表述为 $[0, 1]$ 的 WTMF 训练出的固有权重:

$$m_j = \begin{cases} q_j, & q_j > 0 \\ 0, & q_j \leq 0 \end{cases}$$

概率及熵的表达不变。

3 模型求解

对于模型的求解,需分两步进行:首先对基于异质图的 WTMF 求解,从而得到每条 tweet 的潜语义向量表示,然后全局选取最佳 tweet 子集。

3.1 基于异质图的 WTMF 模型求解

由于 WTMF 模型为非严格矩阵分解,对于极值问题,本文对目标函数求偏导,并迭代优化直至稳定。迭代表达式为:

$$\begin{cases} P_{*,i} = (QW_i Q^T + \gamma I)^{-1} QW_i X_{*,i} \\ Q_{*,j} = (PW_j P^T + \gamma I + \delta L_j Q_{*,s(j)} \text{diag}(L_{s(j)}^2) Q_{*,s(j)}^T)^{-1} \\ \quad (PW_j X_{j,*}^T + \delta L_j Q_{*,s(j)} L_{n(j)}) \end{cases}$$

对上述偏导数进行迭代,得到稳定的 P 和 Q ,其中 Q 为短文本的潜语义向量。用余弦距离衡量两 tweet 的相似度:

$$\text{sim}(i, j) = \frac{Q_{*,i} \times Q_{*,j}}{|Q_{*,i}| |Q_{*,j}|}$$

3.2 全局优化模型的求解

计算此优化问题的确切数值解,则需要搜索解空间内的任一子集,此计算量十分庞大,因此本文采用贪心算法,求解近似最优解。经过观察,发现 $g(S)$ 符合子模属性:随着输入的增大,函数的增长幅度减小。即:

$$\forall X, Y \in U, X \subseteq Y, \forall x \in U - Y$$

$$f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y)$$

在 $g(S) = \alpha R(S) + (1 - \alpha) H_0(S)$ 中,由于相似函数为线性增长,即 $R(X \cup \{x\}) = R(X) + R(\{x\})$ 符合子模属性,而熵在文献[11]中早已被证明有子模属性,因此, $g(S)$ 符合子模属性。

因此针对子模属性,本文设计贪心的抽样算法。在算法中为保证集合中每个元素的信息熵贡献,要求每个元素相对集合的信息增益为正。

算法:基于贪心的迭代抽样算法。

输入:tweet 的潜语义向量 Q ,抽样集大小 k 。

输出:抽样集合 S 。

$$S \leftarrow \arg \max_{q_i \in Q} (r(q_i))$$

$$C \leftarrow Q - S$$

while $|S| < k$ do:

$$q^* \leftarrow \arg \max_{q_i \in C} (g(S \cup \{q_i\}))$$

$$C \leftarrow C - q^*$$

$$G(q_i, S \cup \{q_i\}) = H(S \cup \{q_i\}) - H(S)$$

$$\text{if } G(q_i, S \cup \{q_i\}) > 0$$

$$S \leftarrow S \cup \{q_i\}$$

End

$$S^* \leftarrow S$$

4 实验

4.1 数据集

本文借助 Guo 等^[12]的数据集来做实验,其中包括了从 2013-06-11 至 2013-06-27 的全部 tweet 和 CNN、NYTIMES 的新闻数据,经过预处理,过滤掉其中不包含任何新闻 url 连接的 tweet 和从未被提到的新闻数据,最终形成的数据集如表 3 所示。

表 3 数据集信息统计表

数据	数量	tweet 中新闻评论数范围	数量
tweet	34 888	(100, ∞)	42
news	12 704	(50, 100]	120
		(30, 50]	260
		(10, 30]	842

本文将新闻评论定义为在 tweet 中包含此新闻的 url。tweet 包含发布时间、作者、文本内容、新闻 url、图片 url。新闻包含发布时间、标题、摘要信息、url 连接、图片链接。由于新闻信息过长,本文只抽取新闻的标题和摘要信息作为新闻内

容进行建模。

在训练 WTMF 模型时,本文同样借助维基百科的布朗语料, Wiki 语料以及 WorkNet 语料集进行建模,将其中文本按照句子级别进行拆分,最终生成 441 258 个短文本数据集,其中包含 5 149 122 个单词。

4.2 算法比较

为了验证算法的有效性,本文将信息检索 (Information Retrieval, IR) 作为算法的 baseline,即字面上的词语匹配。同时将潜语义分析的代表算法 LDA,以及文献[13]提出针对短文本对 LDA 进行改进的 BTM 算法作为对比算法。

LDA 的训练参数为 $\alpha = 0.05, \beta = 0.05$, 迭代次数为 5 000; LDA-BTM 算法的参数为: $\alpha = 1, \beta = 0.01$, 迭代次数为 1 000; WTMF 的参数为 $\gamma = 20, \delta = 3$, 迭代次数为 20, 同时两者的信息熵权重 $\alpha = 0.5$ 。

算法全部抽取与指定新闻最相关的 10 条 tweet 生成评价信息。计算新闻与 tweet 集合的信息匹配度,其在评测工具 ROUGE (ROUGE1 表示应用一元语言模型, ROUGE2 表示应用二元语言模型) 的召回率得分如表 4 所示。

表 4 算法信息指标统计

算法	召回率得分		Cover
	ROUGE1	ROUGE2	
IR	0.538 10	0.280 3	0.780 2
LDA	0.639 40	0.241 7	0.628 5
LDA-BTM	0.663 40	0.355 4	0.881 9
WTMF	0.660 80	0.362 3	0.942 9
WTMF + H2	0.705 79	0.394 9	0.912 9
WTMF + H	0.719 47	0.396 0	0.916 6

其中: H2 是指的结合二元信息熵的 WTMF; H 是指的结合连续信息熵的 WTMF; Cover 表示新闻连接覆盖度,即抽取的 tweet 中附带新闻连接的比例。由表 4 看到一个奇怪的现象: LDA 的效果不如 IR 的效果好,原始 LDA 在短文本领域并不适合。针对短文本改进 LDA 的 BTM 相比 LDA 有很大提高,但是其运算效率明显变低。有实验结果看到 WTMF + H 在 ROUGE1 上和 LDA-BTM 相差无几,但是在 ROUGE2 和 Cover 的结果上都有明显优于其他结果。

从图 5 的评分结果可以看出,进入熵模型后,在原始仅考虑相似性的基础上有了较大提升,两个熵模型在 ROUGE1、ROUGE2 召回率上均提高了 0.3 以上(从表 4 可知)。但两者的 Cover 值普遍有所降低,即新闻直接评论覆盖率降低,而信息吻合度反而增加。因此证明信息熵的引入促进了潜语义信息的被挖掘与抽取。

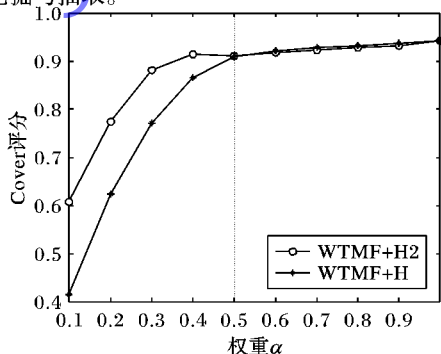


图 5 随 α 变化两种熵模型的 Cover 评分结果

为进一步证明本文的猜想,本文进一步实验观察权重 α 发生变化时的 Cover 评分情况,结果如图 5 所示,当 $\alpha < 0.5$ 时, Cover 评分增长明显;当 $\alpha \geq 0.5$, Cover 增长极为缓慢,但整体趋势一直在增,即纯粹的 WTMF 算法能得到最高的 Cover 得分。但是后期的 Cover 已经不能够保证提供更多的信息量。故此时加入信息熵,更有效地挖掘潜语义信息。

4.3 模型参数比对

为了分析参数对模型的影响,本文对模型参数 α 进行深层分析。观察信息熵模型对潜语义信息挖掘有怎样积极的作用,其实验结果如图 6~7 所示。其中: R 为召回率; F 为召回率和准确率的综合评分。

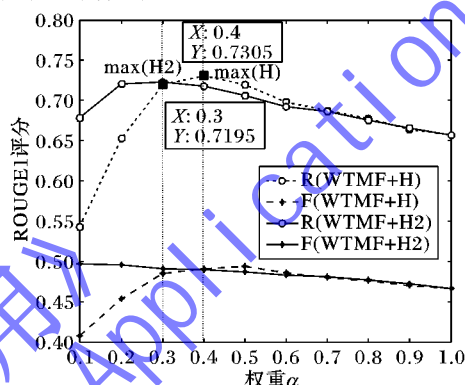


图 6 随 α 变化 WTMF + H 的 ROUGE1 评测结果

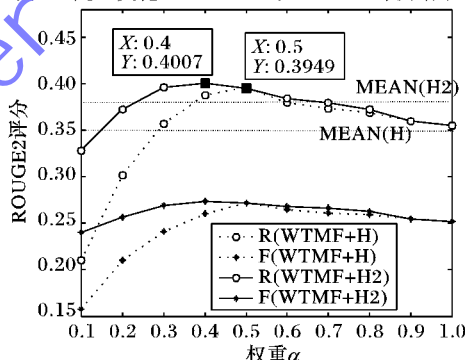


图 7 随 α 变化 WTMF + H 的 ROUGE2 评测结果

当模型权重 α 发生变化时,由图 6,7 可看出,两种熵模型所得到的最优结果类似, ROUGE1 的召回率平衡在 0.72 ~ 0.73, ROUGE2 召回率在 0.395 ~ 0.4, 这说明信息熵对模型的有效性。当 $\alpha = 1$ 时,模型退化为 WTMF 直接抽取高相似句子模型;当 $\alpha = 0$ 时,则只考虑信息多样性,此时根据贪心法的设计,只有选取的第一个 tweet 与信息相关。

但是基于连续特征熵变化浮动较大,当 α 过大或过小时,其评测结果低于 Baseline 的几种算法。而基于二元特征熵评测结果稳定, ROUGE1、ROUGE2 的召回率均值分别高达 0.6924, 0.3745。

由此可见二元熵在本数据集上更为有效。

5 结语

本文试图用优化模型来解决社交媒体的信息选取问题。将 tweet 信息作为新闻的评论信息,弥补了新闻评论信息量稀少的问题。构建新闻与 tweet 间的异质图,结合 WTMF 算法对短文本向量进行语义扩展,计算高相似短文本,同时利用信息熵来衡量候选短文本集合的信息多样性,利用信息增益保证候选个体对整体摘要的贡献性。对于极值优化模型,本文

根据子模性质,设计贪心算法进行模型求解,快速求解其高近似解。实验证明本文的思路是可行的,效果显著。

本文着眼社交媒体信息的抽取,仅仅利用了文本和简单的图片信息,很多方面都可以进行更加深入的扩展,主要有以下几个方面:

1) 本文抽取的 tweet 只考虑了信息的高相关性和多样性。利用用户的社交网络信息,还可以增加信息的权威度以及大众性信息。

2) 本文对图片只进行了简单的关联分析,除了借助文本网络的关联信息,图片本身也表达了一些信息,可抽取整个互联网环境此图片的周边文本,作为图片的文本表述,从而增加 tweet 短文本的信息量,缓解数据稀疏问题。

参考文献:

- [1] MAAS A L, DALY R E, PHAM P T, *et al.* Learning word vectors for sentiment analysis [C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2011, 1: 142 – 150.
- [2] LI H, CHEN Y, JI H, *et al.* Combining social cognitive theories with linguistic features for multi-genre sentiment analysis [EB/OL]. [2014-02-06]. <https://aclweb.org/anthology/Y/Y12/Y12-1013.pdf>
- [3] GAO W, LI P, DARWISH K. Joint topic modeling for event summarization across news and social media streams [C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York: ACM, 2012: 1173 – 1182.
- [4] ZHAI Y, WANG K, ZHANG D, *et al.* An algorithm for semantic similarity of short text based on WordNet [J]. Acta Electronica Sinica, 2012, 40(3): 617 – 620. (翟延冬, 王康平, 张东娜, 等. 一种基于 WordNet 的短文本语义相似性算法 [J]. 电子学报, 2012, 40(3): 617 – 620.)
- [5] YANG Z, CAI K, TANG J, *et al.* Social context summarization [C]// Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2011: 255 – 264.
- [6] SRIRAM B, FUHRY D, DEMIR E, *et al.* Short text classification in twitter to improve information filtering [C]// Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2010: 841 – 842.
- [7] LANDAUER T K, FOLTZ P W, LAHAM D. An introduction to latent semantic analysis [J]. Discourse Processes, 1998, 25(2/3): 259 – 284.
- [8] HOFMANN T. Probabilistic latent semantic indexing [C] // Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1999: 50 – 57.
- [9] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(3): 993 – 1022.
- [10] GUO W, DIAD M. Modeling sentences in the latent space [C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2012, 1: 864 – 872.
- [11] KO C W, LEE J, QUERRYANNE M. An exact algorithm for maximum entropy sampling [J]. Operations Research, 1995, 43(4): 684 – 691.
- [12] GUO W, LI H, JI H, *et al.* Linking tweets to news: a framework to enrich short text data in social media [EB/OL]. [2014-02-16]. <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=96887CAC6EC4C2A8A4203B9D1A80CE3A?doi=10.1.1.361.4604&rep=rep1&type=pdf>.
- [13] YAN X, GUO J, LAN Y, *et al.* A biterm topic model for short texts [C]// Proceedings of the 22nd International Conference on World Wide Web. Geneva: International World Wide Web Conferences Steering Committee, 2013: 1445 – 1456.
- [14] STAJNER T, THOME E B, POPESCU A, *et al.* Automatic selection of social media responses to news [C]// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2013: 50 – 58.
- [15] SREBRO N, JAAKKOLA T. Weighted low-rank approximations [C]// Proceedings of the 20th International Conference on Machine Learning. Palo Alto: AAAI Press, 2003: 720 – 727.
- [16] YAN R, WAN X, LAPATA M, *et al.* Visualizing timelines: evolutionary summarization via iterative reinforcement between text and image streams [C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York: ACM, 2012: 275 – 284.
- [17] ABU-JBARA A, RADEV D. Coherent citation-based summarization of scientific papers [C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2011, 1: 500 – 509.
- [18] NEMHAUSER G L, WOLSEY L A, FISHER M L. An analysis of approximations for maximizing submodular set functions—I [J]. Mathematical Programming, 1978, 14(1): 265 – 294.
- [45] HAZON N, MIELI F, KAMINKA G A. Towards robust on-line multi-robot coverage [C]// Proceedings 2006 IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2006: 1710 – 1715.
- [46] HAZON N, KAMINKA G A. On redundancy, efficiency, and robustness in coverage for multiple robots [J]. Robotics and Autonomous Systems, 2008, 56(12): 1102 – 1114.
- [47] PARLAKTUNA O, SIPAHIOGLU A, KIRLIK G, *et al.* Multi-robot sensor-based coverage path planning using capacitated arc routing approach [C]// Proceedings of the 2009 IEEE Conference on Control Applications. Piscataway: IEEE Press, 2009: 1146 – 1151.
- [48] SIPAHIOGLU A, KIRLIK G, PARLAKTUNA O, *et al.* Energy constrained multi-robot sensor-based coverage path planning using capacitated arc routing approach [J]. Robotics and Autonomous Systems, 2010, 58(5): 529 – 538.
- [49] HSU P-M, LIN C-L, YANG M-Y. On the complete coverage path planning for mobile robots [J]. Journal of Intelligent and Robotic Systems, 2014, 74(3/4): 945 – 963.
- [50] MAO Y, DOU L, CHEN J, *et al.* Combined complete coverage path planning for autonomous mobile robot in indoor environment [C]// Proceedings of the 2009 7th Asian Control Conference. Piscataway: IEEE Press, 2009: 1468 – 1473.
- [51] YAN M, ZHU D, YANG S. Complete coverage path planning in an unknown underwater environment based on D-S data fusion real-time map building [EB/OL]. [2014-03-05]. <http://www.hindawi.com/journals/ijdsn/2012/567959/>.

(上接第 2849 页)