基于矩阵分解和子模最大化的微博新闻摘要方法*

刘彼洋，孙　锐，姬东鸿
（武汉大学 计算机学院，武汉 430072）

摘　要：针对面向微博的中文新闻摘要的主要挑战，提出了一种将矩阵分解与子模最大化相结合的新闻自动摘要方法。该方法首先利用正交矩阵分解模型得到新闻文本潜语义向量，解决了短文本信息稀疏问题，并使投影方向近似正交以减少冗余；然后从相关性和多样性等方面评估新闻语句集合，该评估函数由多个单调子模函数和一个评估语句不相似度的非子模函数组成；最后设计贪心算法生成最终摘要。在 NLPCC2015 数据集上的实验结果表明，该方法能有效提高面向微博的新闻自动摘要质量，ROUGE 得分超过其他基线系统。

关键词：子模属性；正交矩阵分解；新闻摘要；抽取式摘要；微博
中图分类号：TP391.1　　**文献标志码**：A　　**文章编号**：1001-3695(2017)10-2892-05
doi:10.3969/j.issn.1001-3695.2017.10.003

Weibo-oriented news summarization based on matrix factorization and submodular maximization

Liu Biyang, Sun Rui, Ji Donghong
(*School of Computer*, *Wuhan University*, *Wuhan* 430072, *China*)

Abstract: This paper presented a novel method for Weibo-oriented Chinese new summarization which combined matrix factorization and submodular maximization. It used the orthogonal matrix factorization(OrMF) model to solve the information sparsity issue of short texts and the information redundancy problem in the projection procedure, and obtained robust latent vectors for news sentences. Moreover, it evaluated news sentences for its relevance and diversity. The objective function included several submodular functions and a non-submodular function that evaluated sentence dissimilarities. Finally, it designed a greedy algorithm to select summary sentences. Experimental results on NLPCC2015 datasets show that the ROUGE scores of the proposed method outweigh other baseline systems and that the quality of Weibo-oriented news summaries is improved effectively.
Key words: submodularity; orthogonal matrix factorization; news summarization; extractive summarization; Weibo

0　引言

社交媒体的兴起与发展改变了人们传播新闻的方式，一个新闻事件发生后，用户倾向于使用微博这类注重时效性和随意性的社交媒体形式来分享新闻内容。由于社交媒体消息的长度限制（如微博文本内容不得超过 140 字），传统新闻文本难以在社交媒体上直接发布，而人工生成新闻摘要对用户来说既费时又难以保证摘要质量，所以，面向微博等社交媒体的新闻摘要在文本摘要领域的重要性不断增加。

面向微博的新闻摘要任务是对给定的单个中文新闻文本自动生成一篇短摘要，本文利用单文本抽取式摘要技术完成这一任务。抽取式文本摘要过程分为三步[1]：a）抓取文本关键信息并生成文本的中间形式；b）在中间形式的基础上对语句（集）评分；c）选取高分语句（集）形成摘要。

主流的生成文本中间形式的方法包括词频（TF）、词频—逆文本频率（TF-ID）[2,3]和潜语义分析（LSA）[4,5]等。前两者虽然计算方便，但是其并未处理词语与语句上下文之间的语义交互，因此在词语歧义的处理方面存在缺陷，并且没有充分利用词语的同现信息。LSA 虽然可以通过在低维空间同时对词和语句的语义进行建模来解决 TF-IDF 的缺陷，但是其对文本缺失词汇和出现词汇同等对待，导致缺失词汇对模型影响过大，生成的潜语义向量在计算语句相似度，特别是信息稀疏的短文本语句相似度方面效果不佳[6]。本文利用正交矩阵分解（orthogonal matrix factorization，OrMF）模型[7]解决这一问题，在潜语义挖掘过程中区别对待缺失词汇，获得高质量文本潜语义向量用于语句相似度的计算。

对语句（集）的评分与选取可以定义为一个 NP-hard 的组合优化问题[8]，Lin 等人[9~11]提出用子模函数最大化的方式快速得到问题的近似最优解。相关学者在此方向进行了大量研究工作[12~17]，但他们的研究大多关注通用题材的多文本、面向查询和用户生成内容等的摘要任务，所设计的函数和方法并非完全适用于单文本新闻摘要；并且，单纯使用子模函数的方法也缺少对语句间不相似度的评估。

本文在评分时考虑了新闻的写作特点，利用 OrMF 模型生成的潜语义向量计算文本语句相似度，并结合语句位置、词频等信息，从相关性和多样性角度设计多个子模函数及一个基于

收稿日期：2016-07-12；**修回日期**：2016-08-31　　**基金项目**：国家社科重大招标计划资助项目（11&ZD189）；国家自然科学基金面上资助项目（61373108）

作者简介：刘彼洋（1992-），男，湖北武汉人，硕士研究生，主要研究方向为自然语言处理、数据挖掘（liubiyangwhu@126.com）；孙锐（1977-），男，博士研究生，主要研究方向为自然语言处理、深度学习等；姬东鸿（1967-），男，教授，博士，主要研究方向为自然语言处理、语义网技术、机器学习与智能交互、数据挖掘与分析等.

图的用以评估语句不相似度的非子模函数,对语句集合进行评分。针对 140 字的微博长度限制,设计基于贪心的迭代抽取算法求目标函数近似最优解,生成摘要集合。在 NLPCC2015(http://tcci.ccf.org.cn/conference/2015/)任务数据集上的实验表明本文方法能有效提高新闻摘要质量。

1　研究现状

抽取式文本摘要在近年来被诸多研究人员研究并取得了巨大的进步,TF-IDF 和 LSA 等方法是研究中主流的生成文本中间形式的方法。García-Hernández 等人[2]在向量空间模型中评估了使用最大频繁词序列作为特征的抽取式摘要方法,Wei 等人[3]则是利用 TF-IDF 计算权重将子模函数方法用于演说类文本摘要。在 LSA 的应用方面,Gong 等人[4]提出用 LSA 方法识别在语义上重要的语句形成抽取式文摘,Davis 等人[5]提出的 OCCAMS_V 系统则是利用 LSA 方法求词语的权重并设计了语句选择算法。

在利用子模函数进行抽取式摘要领域,相关研究成果也十分突出。Li 等人[12]基于词语覆盖和文本单元相似度设计子模函数并在四类摘要任务中取得较好效果;Sipos 等人[13]提出了一种有监督学习方法训练子模打分函数并在训练过程中利用大间隔方法直接针对性能评价指标进行优化;Morita 等人[14]利用子模最大化从文本簇中抽取依赖子树,同时对文本语句进行抽取和压缩并保证可读性;Dasgupta 等人[15]将子模属性和离散属性结合进行摘要抽取;Vigneshwaran 等人[16]利用结构信息构造单调子模打分函数,生成连贯性与可理解性高的文本摘要;Wang 等人[17]则是将子模函数应用在面向查询的意见摘要领域。

近年来,新闻文本的自动摘要技术逐渐成为研究的热点方向。Svore 等人[18]利用神经网络模型和大量外部输入数据(如维基百科)对 CNN 新闻进行自动摘要;Kastner 等人[19]利用句法、语义和普通统计学特征识别新闻类文章中最重要的语句形成摘要;Chen 等人[20]提出使用递归神经网络语言模型对广播新闻进行摘要,考虑了大跨度的结构特征和词语同现关系;Wang 等人[21]的研究还有莫鹏等人[22]提出的基于超图的协同抽取模型 HCBE 则是专注于中文新闻的摘要。

2　模型描述

本文的任务是给定一篇中文新闻文本,自动生成一篇短摘要。因为是面向微博的摘要,所以摘要中语句总长度小于 140 字。为了完成此任务,本文使用 OrMF 模型获得新闻的文本向量,用于计算新闻文本语句之间的余弦相似度;然后设计目标函数从相关性和多样性两方面评估摘要质量,将任务转换为目标函数极值优化问题;最后利用贪心算法求得目标函数的近似最优解。模型目标函数表示为

$$f(S)=R(S)+D(S) \tag{1}$$

其中:$R(S)$用于衡量摘要的文本相关性,由评估标题相似度、语句位置、词语覆盖度和内容覆盖度的子模函数组成;$D(S)$用于衡量摘要的文本多样性,由评估内容多样性的子模函数及评估语句不相似度的非子模函数组成。

为了下文表示与理解方便,记 U 为新闻文本中所有语句的集合,S 为生成的目标摘要文本集合。对任意语句 $u,v\in U$,记 u,v 的余弦相似度为 $\mathrm{sim}(u,v)$。

本文的自动文摘模型主要流程如图 1 所示,在文本处理模块主要做了分词、去停用词和去低频词的工作。本文使用了 Stanford 的中文分词工具对语料进行分词,然后使用停用词表进行去停用词。另外语料中出现总次数小于 3 的词语是低频词语,也将其从语料中移除。

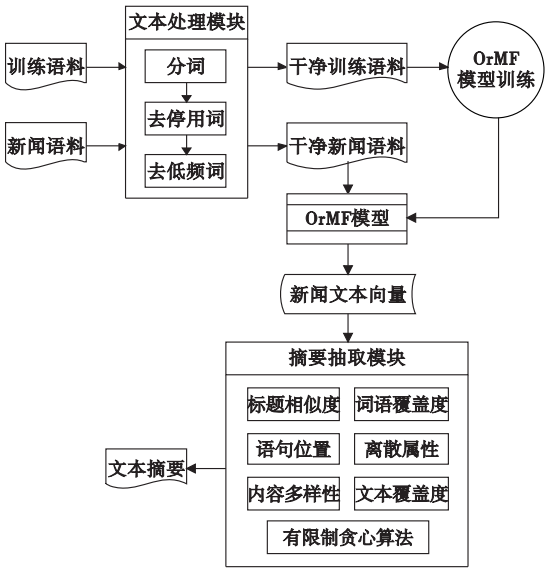

图1　模型主要流程

2.1　文本中间形式

为了解决新闻文本中短句的语义缺失问题和冗余信息问题,本文采用 OrMF 模型生成新闻文本中间形式,获取潜语义向量。OrMF 模型是对带权文本矩阵分解(weighed textual matrix factorization,WTMF)模型[6]的改进,在其基础上减少了冗余信息。WTMF 模型将文本外的缺失词汇也作为文本特征进行训练,主要解决潜语义模型计算语句相似度时因为语句稀疏导致模型效果下降的问题。当文本词汇稀疏导致正相关潜在语义不能准确表达文本含义时,对缺失词汇的训练可以挖掘文本的负相关潜在语义,表达文本与哪些概念无关。WTMF 模型同时利用文本中的出现词汇和缺失词汇构造语义空间,表达文本的完整语义。

将语料表示为一个 $D\times N$ 的 TF-IDF 向量矩阵 $\boldsymbol{X}$,行代表 D 个不同的词,列代表 N 条语句,其中单元 X_{ij} 是词 w_i 在语句 s_j 中的 TF-IDF 值。WTMF 模型方法同 SVD 十分相似,不同之处在于其可以对矩阵中的每一个单元 X_{ij} 进行控制。WTMF 模型将原始矩阵 $\boldsymbol{X}$ 分解为两个矩阵 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$,使得 $\boldsymbol{X}\approx\boldsymbol{P}\times\boldsymbol{Q}^{\mathrm{T}}$。其中:$\boldsymbol{P}$ 为 $D\times K$ 的矩阵,$\boldsymbol{P}_{i,\cdot}$ 表示词 w_i 的 K 维潜语义向量;同样地,$\boldsymbol{Q}$ 为 $N\times K$ 的矩阵,$\boldsymbol{Q}_{j,\cdot}$ 是语句 s_j 的 K 维潜语义向量。

WTMF 模型中对矩阵 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$ 进行最优化的目标函数可以表示为

$$\min\sum_i\sum_j W_{ij}(\boldsymbol{P}_{i,\cdot}^{\mathrm{T}}\times\boldsymbol{Q}_{j,\cdot}-X_{ij})^2+\alpha\|\boldsymbol{P}\|_2^2+\alpha\|\boldsymbol{Q}\|_2^2 \tag{2}$$

其中:α 是一个自由正则化因子,$\boldsymbol{W}$ 是一个权重矩阵,对矩阵 $\boldsymbol{X}$ 中的每一个单元 X_{ij} 分配对应权重。$\boldsymbol{W}$ 对缺失词汇赋予较小的权重 w_m(如 0.01),在考虑缺失词汇的同时也保证文本中出现的词汇对潜语义的影响程度,即

$$W_{ij}=\begin{cases}1 & X_{ij}\neq 0\\ w_m & X_{ij}=0\end{cases} \tag{3}$$

对 WTMF 模型进行求解的过程中，首先随机初始化 P 和 Q，然后根据式(4)(5)进行迭代计算使 P 和 Q 矩阵趋于稳定。

$$P_{i,\cdot} = (Q^{\mathrm{T}} \tilde{W}^{(i)} Q + \alpha I)^{-1} Q^{\mathrm{T}} \tilde{W}^{(i)} X^{\mathrm{T}}_{i,\cdot} \tag{4}$$

$$Q_{j,\cdot} = (P^{\mathrm{T}} \tilde{W}^{(j)} P + \alpha I)^{-1} P^{\mathrm{T}} \tilde{W}^{(j)} X^{\mathrm{T}}_{\cdot,j} \tag{5}$$

其中：$\tilde{W}^{(i)} = \mathrm{diag}(W_{i,\cdot})$ 是一个包含了权重矩阵 W 第 i 行的 $N \times N$ 的对角矩阵；同理可知 $\tilde{W}^{(j)} = \mathrm{diag}(W_{\cdot,j})$ 是包含了 W 第 j 列的 $D \times D$ 对角矩阵。最终 Q 矩阵就是所求的文本潜语义向量矩阵。

需要指出的是，对于 $D \times K$ 的矩阵 P 的含义存在两种解释。在前文中已提到 $P_{i,\cdot}$ 表示了词 w_i 的 K 维潜语义向量，而 P 的列 $P_{\cdot,k}$ 则表示了投影向量，这与 LSA 方法的特征向量相似，也是 OrMF 模型的主要研究对象。在 WTMF 模型的求解过程中，P 和 Q 被迭代计算优化，这种方式较好地保存了文本中词语的相关性，但在降维过程中可能向投影向量 $P_{\cdot,k}$ 中编入了重复信息。例如：第一个维度 $P_{\cdot,1}$ 可能与 80% 的音乐主题与 20% 的政治主题相关，而第二个维度 $P_{\cdot,2}$ 与 90% 的政治主题与 10% 的文化主题相关。但理想情况下，各个维度之间应互不关联，从而使得从数据中抓取的主题更加清晰。达成这个目标的一种方法就是使矩阵 P 正交，即 $P^{\mathrm{T}}P = I$。这表明如果 $k \neq j$，$P^{\mathrm{T}}_{\cdot,j} \times P_{\cdot,k} = 0$。

为了达成这一目标，模型采用了梯度下降的方法，在每次的迭代过程中都使 P 向 $(P^{\mathrm{T}}P - I)^2$ 的负梯度方向逼近。但 $(P^{\mathrm{T}}P - I)^2$ 要求投影向量 $P_{\cdot,k}$ 是一个单位向量，当矩阵 X 中非零值较大时无法达成，所以模型将矩阵 I 乘以一个系数 β，β 是当前迭代代数中 $P^{\mathrm{T}}P$ 对角线的平均值，即在最终的 WTMF 迭代计算公式中要加上式(6)和(7)。

$$\beta = \mathrm{mean}(\mathrm{diag}(P^{\mathrm{T}}P)) \tag{6}$$

$$P = P - \theta P(P^{\mathrm{T}}P - \beta I) \tag{7}$$

在计算过程中 P 的量级未受影响，θ 为步长大小，将修改后的模型称为 OrMF 模型。

本文首先利用训练语料训练 OrMF 模型，迭代计算得到矩阵 P，然后结合新闻文本的 TF-IDF 向量矩阵 X 计算文本的潜语义向量矩阵 Q，用于新闻语句间余弦相似度的获取。例如对语句 s_j 和 s_h 两者之间的相似度可以由其潜语义向量 $Q_{j,\cdot}$ 和 $Q_{h,\cdot}$ 的余弦距离得出。所得语句相似度用于后文中对标题相似度、内容覆盖度和语句分离度三个特征的评估。

2.2 相关性评估

2.2.1 标题相似度

新闻标题是以最简练的文字将新闻中最重要、最有价值的内容展示给读者，不仅能向读者提示新闻内容，而且可以帮助读者进一步理解新闻的内容和意义。因此新闻文本中与新闻标题相似度高的语句与新闻有更大的相关性，应有较高可能性被选入摘要集合 S。本文使用式(8)计算摘要集合 S 的标题相似度得分。

$$\mathrm{Tsim}(S) = \sum_{u \in S} \mathrm{sim}(u, \mathrm{title}) \tag{8}$$

其中：title 为新闻标题语句，$u \in U$。由于标题相似函数 Tsim (S) 为线性增长，所以 Tsim(S) 符合单调子模属性。

2.2.2 语句位置

考虑到新闻写作通常使用倒金字塔结构，即将最重要和最具概括能力的事实写在新闻最前面，写作时按事实重要程度和读者关注程度的主次顺序安排。本文认为语句位置是对新闻中语句的文本相关性评估的一个重要指标，书写位置越靠前的语句对新闻的代表程度越高。对于单条语句 $u \in U$，本文用式(9)评估 u 的位置得分。

$$\mathrm{pos}_u = \frac{n - p_u}{n} \tag{9}$$

其中：n 为新闻文本中的语句总数，p_u 表示语句 u 在新闻中的语句序号，即 u 为新闻中的第 p_u 条语句。在此基础上，对于摘要集合 S，位置函数 pos(S) 定义如下：

$$\mathrm{pos}(S) = \sum_{u \in S} \mathrm{pos}_u \tag{10}$$

类似标题相似度函数，位置函数 pos(S) 同为线性增长，所以其同样符合单调子模属性。

2.2.3 词语覆盖度

因为抽取的摘要集合中语句的总长度不超过 140 字，所以在摘要文本的选取过程中笔者倾向于选取长度较短并使摘要质量获得较大提升的语句，而词频是能表现这一特点的最直接的统计特征。对任意语句集合 S'，记 S' 中语句所包含的词语集合为 $T(S')$。本文中摘要集合 S 的词语覆盖度函数如下：

$$\mathrm{frq}(S) = \sum_{t \in T(S)} tf_t \tag{11}$$

其中：t 为 $T(S)$ 中的词，tf_t 为 t 所对应的词频。因为 frq(S) 是一个累加函数，所以候选语句对 frq(S) 的增益非负。另外，对任意两个语句集合 S_1 和 S_2，假设 $S_1 \subseteq S_2$，则有 $T(S_1) \subseteq T(S_2)$，故对任意候选语句 $u \in U$，有 $T(\{u\}) \setminus T(S_1) \supseteq T(\{u\}) \setminus T(S_2)$。frq$(S)$ 是对 $T(S)$ 中所有词语词频的累加，所以 S_1 中加入候选语句 u 的函数增益一定大于等于 S_2 中加入 u 的函数增益。因此证明了词语覆盖函数 frq(S) 的单调子模属性。

2.2.4 内容覆盖度

内容覆盖度评估了摘要集合 S 对语句集合 U 的覆盖程度，覆盖度函数记为 cov(S)。覆盖度函数首先为单调函数，因为一个较大摘要集合对新闻的覆盖度肯定更大；同时覆盖度函数还要具有子模属性，例如有两个语句集合，其中一个为另一个的子集，向较小集合中加入语句所获得的覆盖度函数增益应大于向较大集合中加入语句的增益，因为新加入语句所包含的信息可能已经被较大集合中的语句所覆盖，而未被较小集合中的语句覆盖。在本文中，定义覆盖度函数为

$$\mathrm{cov}(S) = \sum_{u \in U} \min(C_u(S), 0.25 C_u(U)) \tag{12}$$

其中：

$$C_u(S') = \sum_{v \in S'} \mathrm{sim}(u, v) \tag{13}$$

S' 为任一语句集合，$C_u(S)$ 衡量了摘要集合 S 对语句 u 的覆盖程度，而 $C_u(U)$ 是 $C_u(S)$ 所能达到的最大值。当 $\min(C_u(S), 0.25\ C_u(U)) = 0.25\ C_u(U)$ 时，可以认为 S 中语句对语句 u 的覆盖程度已足够大，达到了饱和状态。此时按照覆盖度函数的公式，假设新增的语句为 v，无论 sim(u, v) 为多少，语句 v 都不会再增加 S 对 u 的覆盖程度，贪心算法就会倾向于选取那些 S 对其覆盖程度未达到饱和状态的语句。因此保证了最终摘要集合能更好地覆盖新闻文本信息。

2.3 多样性评估

2.3.1 语句分离度

尽管子模属性模型较为优秀，但其在表达自动文摘的特征方面还存在缺陷：子模属性模型无法从文本语句之间的不相似度方面去衡量摘要的多样性和冗余性。对文本摘要而言，一个天然要求是摘要中语句间不相似度的和尽可能大，但这一特征

无法通过子模函数表现,本文将其称为语句分离度。

本文将新闻文本构造为图模型评估语句分离度,令 $d'(u,v)=1-\mathrm{sim}(u,v)$,$d'(u,v)$是图中 u、v 两点间边的权重,并且定义 $d(u,v)$为图中 u 到 v 的最短距离。本文定义的语句分离度函数 $\mathrm{sep}(S)$通过衡量摘要集合 S 中所有语句对之间的最短距离的和来评价 S 的文本多样性,即

$$\mathrm{sep}(S)=\sum_{u,v\in S}d(u,v) \tag{14}$$

S 中所有语句对的最短距离和值越大,表明 S 中语句的多样性程度越高。很容易构筑例子证明 $\mathrm{sep}(S)$不具有子模属性,但是当目标函数 $f(S)$由一组子模属性函数和一个非子模属性的函数 $\mathrm{sep}(S)$组成时,仍可以用贪心算法在多项式时间内求得近似最优解[23]。并且记 O 为函数 g 的最优解,$\tilde{S}$为求得的近似最优解,有 $g(\tilde{S})\geqslant g(O)/2$。

2.3.2　内容多样性

此函数通过对摘要内容的多样性进行奖励来提升摘要文本的多样性。首先利用随机算法将新闻文本的语句集合 U 分为 n 个簇,记为 $C=\{C_1,C_2,\cdots,C_n\}$(C_i 之间互不相交),则相应的目标函数为

$$\mathrm{conD}(S)=\sum_{i=1}^{n}\log(1+|C_i\cap S|) \tag{15}$$

由函数 $\log(1+x)$的性质可知,如果某簇没有语句在摘要集合 S 中,选取此簇的语句可以使 $\mathrm{conD}(S)$得到最大提升。而随着某簇被选入 S 的语句增多,此簇中其他语句对 $\mathrm{conD}(S)$增益就会逐渐减少,即内容多样性函数 $\mathrm{conD}(S)$会对摘要内容的多样性进行奖励。假设语句 $u_1,u_2\in C_1,u_3\in C_2,u_1$ 已经在 S 中,选择将 u_2 加入 S 后 $\mathrm{conD}(S)=\log 3$,而选择加入 u_3 后 $\mathrm{conD}(S)=2$,则贪心算法选取语句时会将 u_3 加入 S 中,因为 u_3 对目标函数 $\mathrm{conD}(S)$有更大的提升。这样从不同簇中选取语句的特性可以使摘要 S 的内容多样性得到提升。

2.4　模型求解

最终的目标函数 $f(S)$可以表示为 $\lambda_1\mathrm{Tsim}(S)+\lambda_2\mathrm{pos}(S)+\lambda_3\mathrm{frq}(S)+\lambda_4\mathrm{sep}(S)+\lambda_5\mathrm{conD}(S)+\lambda_6\mathrm{cov}(S)$的形式,$\lambda_1\sim\lambda_6$ 均为非负系数。本文使用网格搜索的方式对函数系数进行最优化,先以粗粒度搜索最优系数比例的大致区间,再以细粒度对区间内各系数比例组合分别进行实验,最终选取实验效果最优的系数组合。所有的文本相关性函数和内容多样性函数是单调子模函数,然而语句分离度函数 $\mathrm{sep}(S)$是一个非子模属性的函数,因此 $f(S)$不具有子模属性。求 $f(S)$的最大值是一个 NP-hard 问题,但仍可以利用贪心算法在多项式时间内求得 $f(S)$近似最优解[23]。考虑微博内容的 140 字长度限制,设计了贪心抽样算法对 $f(S)$求解。其中 L 为 140,x 为比例因子,本文中取 x 为 1/3。

算法　贪心抽样算法

输入:新闻文本语句集合 U,长度限制 L。
输出:摘要集合 S。

$S\leftarrow\varnothing$
while $U\neq\varnothing$ do
　$c\leftarrow\arg\max_{u\in U}\dfrac{f(U\cup\{u\})-f(U)}{(l_u)^x}$
　if $\sum_{v\in S}l_v+l_c\leqslant L$ and $f(S\cup\{c\})-f(S)>0$ do
　　$S\leftarrow S\cup\{c\}$
　$U\leftarrow U\backslash\{c\}$
end while
return S

3　实验

3.1　评价标准及数据集

本文将所提模型在 NLPCC2015 面向微博的中文新闻摘要任务数据集上进行评估。此数据集包含 250 篇新闻文本并且已进行分句,每篇新闻都有对应的两篇参照摘要作为实验结果对比,所有摘要字数均不超过 140 的微博长度限制。

生成的结果摘要将利用 ROUGE 自动评测工具与参照摘要进行对比。本实验中共使用了 ROUGE 提供的五类评测分数,ROUGE 工具对每类评测分数分别计算了准确率、召回率和 F 值,本文实验评测中使用 F 值作为评测指标。

在训练 OrMF 模型时,本文使用了 2012 年的全网新闻数据和搜狐新闻数据作为语料进行建模,训练过程中 OrMF 参数为 $\alpha=20$、$W_m=0.001$、$\theta=0.0001$,生成的文本潜语义向量维度 $k=100$,模型迭代计算的次数为 20 次。最终语料有 6 200 742 条文本数据,其中包含 329 414 个单词。

3.2　特征组合分析

为了研究本文所使用的六个评估相关性和多样性特征的函数对实验结果的影响,选择性地组合了这些函数分别进行实验分析,实验过程中对函数的系数均进行参数最优化,实验结果如表 1 所示。以单独使用词语覆盖度特征所得结果为基础结果,先评估了四个相关性方面特征。从表 1 中可以发现,单独使用词语覆盖度并不足以让本文模型产生令人满意的文摘,而加入标题相似度或语句位置特征均可以使实验结果有所提升。其中,加入语句位置特征后实验效果提升较为明显,ROUGE-1 的 F 值提高了近七个百分点。表明针对新闻写作特点引入的语句位置特征促进了新闻摘要质量的提升,原因可能在于距离新闻标题最近的语句中包含了新闻最重要且最相关的信息,更适合作为摘要内容,位置函数通过奖励这一部分语句使其更容易被选入摘要。而加入内容覆盖度函数对实验效果的提升几乎可以忽略,原因可能在于面向微博的摘要具有 140 的长度限制,在摘要较短的情况下摘要集合中语句较少,使得集合中不同语句对新闻文本内容覆盖度的区分度较低。

表 1　模型参数效果对比

评估函数	ROUGE-1	ROUGE-2	ROUGE-SU4
frq	0.423 99	0.260 15	0.248 79
frq,Tsim	0.432 48	0.270 52	0.259 26
frq,pos	0.502 75	0.368 11	0.352 42
frq,cov	0.420 94	0.260 83	0.250 34
frq,pos,Tsim	0.505 08	0.371 06	0.355 34
frq,pos,Tsim,cov	0.506 06	0.372 51	0.356 55
frq,pos,Tsim, cov, conD	0.509 54	0.376 17	0.360 51
frq,pos,Tsim,conD,cov,sep	0.519 03	0.387 36	0.371 60

在加入了四个相关性特征的基础上,本文在实验中首先加入内容多样性特征进行实验。实验结果表明,内容多样性可以使摘要质量有所提升,但是提升幅度有限。而加入语句分离度后,实验效果相较单独使用子模属性函数进一步提高,在 ROUGE-1 的 F 值已达 0.509 的情况下增加了近 1 个百分点。这表明使用语句分离度函数对语句间的不相似度进行建模,在多样性方面的确对仅使用子模函数的方法形成了较好的补充。综合语句分离度函数与子模函数能产生更高质量的摘要。最

终结果表明,使用全部六个函数对摘要集合进行评估时模型的效果达到最优。

3.3 文本中间形式对比

本文使用 OrMF 模型产生文本中间形式,获得潜语义向量,并完成语句相似度计算。为了对比 OrMF 模型与传统生成文本中间形式方法效果的差异,本文使用 TF-IDF 和 LSA 模型分别获得文本向量并与 OrMF 进行对比实验。为了排除干扰项影响,在对比时仅使用基于语句相似度计算的标题相似度、内容覆盖度和语句分离度三个特征进行实验,三种方式的实验结果如图 2 所示。

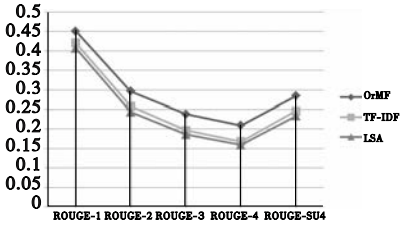

图2 文本向量生成方式对比结果

从图 2 中可以看出,使用 TF-IDF 和 LSA 模型生成文本向量的实验结果均小于使用 OrMF 模型,其中 LSA 的效果略逊于 TF-IDF。表明使用 OrMF 模型所生成的文本潜语义向量在实验中能获得更好的效果。原因可能在于实验数据集的新闻文本中存在部分短文本,OrMF 模型使用缺失词汇扩充文本信息去解决短文本语句稀疏问题,从而获得更高质量潜语义向量,提升其在语句相似度计算方面的效果。而 TF-IDF 和 LSA 模型并未对短文本进行合适的处理,在出现短文本的新闻语料中并不适合。

3.4 模型对比

为了验证本文所提模型的有效性,本文将实验结果与参加了 NLPCC2015 面向微博中文新闻文本摘要任务的参赛组的官方评测结果,以及莫鹏等人提出的 HCBE 模型[22]、Lin 等人[9]的方法和 Li 等人[12]提出的 MSSF 模型在同样数据集上的实验结果进行对比。在 ROUGE-1 和 ROUGE-2 上的对比结果如图 3、4 所示。

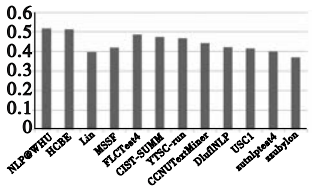

图3 ROUGE-1评测结果

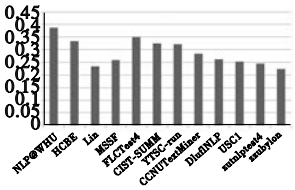

图4 ROUGE-2评测结果

其中本文的实验结果以 NLP@ WHU 标志,从图 3、4 中可以看出,本文所提模型在各项 ROUGE 评分上的效果均优于 NPCC2015 参赛组的结果和其他基线模型。其中,ROUGE-1 的 F 值比第二的 HCBE 方法高出约 0.005;而在二元语言模型上的优势更为突出,ROUGE-2 的评分较第二的 FLCTest4 的结果提高近 0.04。表明本文所使用的基于子模最大化进行文本摘要抽取方法相较于其他方法更为有效,本文针对新闻文本所设计的评估函数也相对合理。

在与同样使用了子模函数框架进行自动摘要的 Lin 的方法和 MSSF 模型的比较中发现,本文所提方法对中文新闻文本摘要的效果明显优于这两者。可能是因为他们的方法更加注重对所有文本类型及任务类型的通用性,Lin 的方法针对文本覆盖程度和语句多样性设计子模函数,MSSF 模型则是从词语覆盖和语句相似度的层面考虑摘要与文本的相关性,他们的方法在多文本摘要和面向查询摘要等领域均有较好的效果。而本文所使用的特征与函数则更多针对新闻类文本的单文本摘要且同时考虑了相关性和多样性两个方面,评估的特征更具指向性;此外,本文还考虑了语句分离度的评估,而另外两者仅使用子模函数建模,这也可能是本文效果更好的原因之一。

4 结束语

本文利用矩阵分解和子模最大化方法获得面向微博的中文新闻文本摘要,以解决社交媒体中新闻传播受限于文本长度的问题。利用正交矩阵分解模型获取文本向量计算文本相似度,扩展短文本信息量并避免信息冗余;同时考虑新闻写作特点,从相关性与多样性两方面设计目标函数选取摘要集合。利用子模最大化方法将文摘任务转换为极值优化问题并设计贪心算法求解。实验结果表明本文模型能有效提高中文新闻文摘质量,效果较为明显。

未来研究将从以下几个方向进行更深入的扩展:

a)将在语法、词性、依存关系、社会背景和时效性等本文未曾涉及的方面引入更多中文新闻摘要特征。

b)尝试将基于事件的生成式自动文摘技术引入本文模型中,研究可行性与效果。

c)将单文本新闻摘要研究扩展到多文本领域和面向查询的新闻自动文摘领域,改进模型以提高通用性。

参考文献:

[1] Nenkova A, McKeown K. A survey of text summarization techniques [M]. New York: Springer US, 2012: 43-76.

[2] García-Hernández R A, Ledeneva Y. Word sequence models for single text summarization[C]//Proc of the 2nd International Conferences on Advances in Computer-Human Interactions. 2009: 44-48.

[3] Wei Kai, Liu Yuzong, Kirchhoff K, *et al*. Using document summarization techniques for speech data subset selection [C]//Proc of NAACL Conference. 2013: 721-726.

[4] Gong Yihong, Liu Xin. Generic text summarization using relevance measure and latent semantic analysis[C]//Proc of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2001: 19-25.

[5] Davis S T, Conroy J M, Schlesinger J D. OCCAMS: an optimal combinatorial covering algorithm for multi-document summarization[C]//Proc of the 12th International Conference on Data Mining. Washington DC: IEEE Computer Society, 2012: 454-463.

[6] Guo Weiwei, Diab M. Modeling sentences in the latent space[C]//Proc of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2012: 864-872.

[7] Guo Weiwei, Liu Wei, Diab M T. Fast tweet retrieval with compact binary codes[C]//Proc of the 25th International Conference on Computational Linguistics. 2014: 486-496.

[8] McDonald R. A study of global inference algorithms in multi-document summarization[C]//Proc of the 29th the European Conference on Information Retrieval. Berlin: Springer, 2007: 557-564.

[9] Lin Hui, Bilmes J. A class of submodular functions for document summarization[C]//Proc of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2011: 510-520.

(下转第 2928 页)

频繁项集都会减少,所用的时间也就相应地降低。相比于经典 Apriori 和 FIUT-Stream 算法,AMST 算法需要更少的运行时间和更小的内存空间,这是因为 Apriori 和 FIUT-Stream 算法在频繁项集产生的过程中,需要进行大量的连接和剪枝操作,因而算法的时间效率不高。而 AMST 算法只需对事务数据库进行一次扫描,避免产生大量的候选子树集,也省去了连接和剪枝的操作,而是在找出全部频繁子树后再进行剪枝,提高了挖掘的时间效率,减少了内存的消耗。

通过 AMST 算法在所处理的事务数和不同的支持度 S 下的运行时间的实验以及 AMST 算法与其他算法在时间和内存使用比较实验中可以看出,AMST 算法能稳定、快速准确地挖掘出结构二叉树中具有层次关联关系的频繁子树。

4 结束语

最近几年中,数据流中挖掘频繁项集成为重要的研究课题,然而还没有将数据流转换成结构二叉树进而挖掘出频繁模式中具有层次关联关系的频繁项集相关方面的研究。本文提出了一种数据流中结构二叉树的挖掘算法(AMST),算法将数据流转换成结构二叉树,然后将二叉树中的子树用数据流矩阵表示,最后挖掘出具有层次关联关系的频繁子树。实验表明,AMST 算法性能稳定,能够快速准确地进行频繁子树的挖掘。但是本文所提出的算法只适用于对确定数据流的处理,下一步工作将继续优化 AMST 算法,使之能够应用在不确定数据流的处理中。

参考文献:

[1] Leung C K, Joseph K W. Sports data mining: predicting results for the college football games[J]. Procedia Computer Science, 2014, 35: 710-719.

[2] Leung C K, MacKinnon R K, Wang Yang. A machine learning approach for stock price prediction[C]//Proc of the 18th International Database Engineering and Application Symposium. New York: ACM Press, 2014: 274-277.

[3] Tanbeer S K, Leung C K, Cameron J J. Interactive mining of strong friends from social networks and its applications in e-commerce[J]. Journal of Organizational Computing and Electronic Commerce, 2014, 24(2-3): 157-173.

[4] 王建华. 制造物联海量数据流处理方法研究[D]. 广州: 广东工业大学, 2015.

[5] 娄建楼, 邹伟, 王玲, 等. 社交网络大数据下贪婪式实时网站推荐算法[J]. 计算机应用研究, 2015, 32(5): 1361-1364.

[6] 李楠. 基于关联数据的知识发现研究[D]. 北京: 中国农业科学院, 2012.

[7] 王亚琴. 道路交通流数据挖掘研究[D]. 上海: 复旦大学, 2007.

[8] 陈鹏. 数据流关联规则挖掘研究及其应用[D]. 杭州: 浙江大学, 2011.

[9] Leung C K, Khan Q I. DSTree: a tree structure for the mining of frequent sets from data streams[C]//Proc of the 6th IEEE International Conference on Data Mining. 2006: 928-932.

[10] Gurmeet S, Rajeev M. Approximate frequent counts over data streams [C]//Proc of the 28th VLDB Conference. 2002: 346-357.

[11] Leung C K, Jiang F, Hayduk Y. A landmark-model based system for mining frequent patterns from uncertain data streams[C]//Proc of the 15th Symposium on International Database Engineering and Applications. New York: ACM Press, 2011: 249-250.

[12] 李爱国, 厍向阳. 数据挖掘原理、算法及应用[M]. 西安: 西安电子科技大学出版社, 2012.

[13] 寇香霞, 任永功, 宋奎勇. 一种基于滑动窗口的数据流频繁项集挖掘算法[J]. 计算机应用与软件, 2013, 30(1): 143-146.

(上接第 2896 页)

[10] Lin Hui, Bilmes J. Multi-document summarization via budgeted maximization of submodular functions[C]//Proc of Annual Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010: 912-920.

[11] Lin Hui, Bilmes J, Xie Shasha. Graph-based submodular selection for extractive summarization[C]//Proc of IEEE Workshop on Automatic Speech Recognition & Understanding. 2009: 381-386.

[12] Li Jingxuan, Li Lei, Li Tao. Multi-document summarization via submodularity[J]. Applied Intelligence, 2012, 37(3): 420-430.

[13] Sipos R, Shivaswamy P, Joachims T. Large-margin learning of submodular summarization models[C]//Proc of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2012: 224-233.

[14] Morita H, Sasano R, Takamura H, *et al*. Subtree extractive summarization via submodular maximization[C]//Proc of ACL. 2013: 1023-1032.

[15] Dasgupta A, Kumar R, Ravi S. Summarization through submodularity and dispersion[C]//Proc of the 51st Meeting on Association for Computational Linguistics. 2013: 1014-1022.

[16] Vigneshwaran L J K P M, Sharma M V V D M. Non-decreasing submodular function for comprehensible summarization[C]//Proc of the 15th Annual Conference of the North American Chapter of the Association for Compututional Linguistics. 2016: 94-101.

[17] Wang Lu, Raghavan H, Cardie C, *et al*. Query-focused opinion summarization for user-generated content[C]//Proc of International Conference on Computational Linguistics. 2014: 1660-1669.

[18] Svore K M, Vanderwende L, Burges C J C. Enhancing single-document summarization by combining RankNet and third-party sources [C]//Proc of EMNLP-CoNLL Conference. 2007: 448-457.

[19] Kastner I, Monz C. Automatic single-document key fact extraction from newswire articles[C]//Proc of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2009: 415-423.

[20] Chen Kuanyu, Liu Shihhung, Chen B, *et al*. Extractive broadcast news summarization leveraging recurrent neural network language modeling techniques[J]. IEEE/ACM Trans on Audio, Speech, and Language Processing, 2015, 23(8): 1322-1334.

[21] Wang J H, Yang J Y. Statistical single-document summarization for Chinese news articles[C]//Proc of the 26th International Conference on Advanced Information Networking and Applications. Washington DC: IEEE Computer Society, 2012: 183-188.

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

[23] Borodin A, Lee H C, Ye Yuli. Max-sum diversification, monotone submodular functions and dynamic updates[C]//Proc of the 31st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems. New York: ACM Press, 2012: 155-166.