一种基于 Opinosis 图和马尔科夫随机游走模型的多文本情感摘要框架

康世泽,马 宏,黄瑞阳
(国家数字交换系统工程技术研究中心,河南郑州 450002)

摘 要: 针对在线文本情感摘要生成问题,本文提出了一种基于 Opinosis 图和马尔科夫随机游走模型的情感摘要框架. 首先,该框架将原始文本转化为 Opinosis 图,并利用其挖掘出文本中的特征词,这些特征词可以用来对原始文本的句子进行分类;其次本文在基于聚类的条件马尔科夫随机游走模型的基础上增加了情感层,改进后的模型可以判断同一聚类中各句子的情感倾向是否具有代表性并结合情感和聚类信息对句子进行排序. 实验结果表明,本文提出的方法与基准算法相比在 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)值上具有明显提高.

关键词: Opinosis 图; 马尔科夫随机游走模型; 情感摘要
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2017)12-3005-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.12.024

An Opinosis and MRW Based Sentiment Summarization Framework

KANG Shi-ze, MA Hong, HUANG Rui-yang
(*National Digital Switching System Engineering & Technological R&D Center, Zhengzhou, Henan* 450002, *China*)

Abstract: In order to produce summaries of online comment text, this paper presents a novel sentiment summarization framework which can produce abstractive summary based on Opinosis graph and Markov random walk model. This framework first convert the original text into Opinosis graph and use the Opinosis graph to mine the features of the original text, which can be used to classify the sentences. And then this paper adds a sentiment layer upon the cluster-based conditional Markov random walk nodel, and this improved model can judge which sentiment polar of the sentences in the same cluster is representative and select the proper sentence to produce abstractive summary based on the factors of sentiment and cluster. Experimental results show that this framework has achieved better results in ROUGE(Recall-Oriented Understudy for Gisting Evaluation) value compared to the baseline algorithm.

Key words: Opinosis graph; Markov random walk model; sentiment summarization

1 引言

互联网上蕴含大量针对不同评价对象的在线评论文本. 相同的评价对象属于同一个领域,而同一领域中评价对象的各种属性都称为该领域的特征. 可以利用情感摘要对在线评论文本进行理解,情感摘要旨在从关于某一特定领域的包含观点的文本集中抽取一系列可以传达原文主要情感倾向和观点的句子组成摘要[1,2]. 已有的大部分关于情感摘要的工作都集中在给某个领域特征预测情感倾向[3,4]或为该领域的各个特征评定等级上[5,6]. 这些类型的摘要虽然也能提供一定的信息量,但当用户想要了解某个领域的各种特征细节时,就必须阅读大量关于该特征的高度冗余的句子. 因此,为了帮助用户进一步理解与各个领域特征相关的信息,有必要生成一种基于特征的简洁情感摘要.

与传统的抽取式多文本摘要[7]相比,基于特征的多文本情感摘要主要面临两个挑战,一是需要识别出领域文本集中的特征词[8,9]. 一组关于某特定领域的文本集通常会包含大量关于该领域的特征词,因此需要一种有效的方法对蕴含其中的特征词进行识别. 基于特征的情感摘要面对的另一个挑战就是需要判断所抽

收稿日期:2015-07-13;修回日期:2015-12-23;责任编辑:梅志强
基金项目:国家自然基金(No.61171108);国家重点基础研究发展计划("973"计划)基金(No.2012CB315901,No.2012CB315905)

取的关于某一领域特征的句子的情感倾向是否具有代表性,因为一些“垃圾”评论的情感倾向会与主流的情感倾向相悖[1]. 因此需要一种有效的判断句子的情感倾向是否具有代表性的方法.

识别文本中特征词的工作通常是与对特征词的情感极性分析同时进行的,使用特征级别的情感分析方法进行特征识别取得了较好的效果,但是这些方法普遍工作量较大[10],而本文的主要工作还是侧重于摘要生成,因此本文在识别特征词时略去了情感分析的工作,采用了 Ganesan 和 Zhai 等[11]提出的 Opinosis 图来进行特征识别. Opinosis 图用单词来表示每个节点的同时使用有向边来表示句子. Ganesan 和 Zhai 等在 Opinosis 图的基础上构建了可以生成摘要的 Opinosis 框架,本文没有使用 Opinosis 框架生成摘要,而是采用 Opinosis 图来挖掘文本集中较短的路径作为特征词. 当原文本集是关于某一特定领域的语料时,为了识别出文本集中只和该领域相关的特征词,本文将该领域与其它领域的语料进行对比,将几个领域语料中共同频繁出现的特征词去除,从而使保留下的特征词具有更好的领域相关性.

为了判断所抽取的关于某一特征的句子的情感倾向是否具有代表性,本文将每个句子的情感倾向都作为一个元素加入到句子的词向量中作为一个附加的元素,并提出了基于情感和聚类的条件马尔科夫随机游走模型,该模型同时融合句子的聚类、情感信息对句子进行排序. 马尔科夫模型之前已被成功应用于多文本摘要[12],它通过构建表明句间关系的有向图并利用特定的图排序算法对句子进行排序. 本文在 wan[12] 构建的两层马尔科夫随机游走模型的基础之上加入了情感层使其可以生成情感摘要.

2 相关工作

2.1 Opinosis 图

2.1.1 Opinosis 图简介

算法 1 概述了构建 Opinosis 图的步骤. 从一组文本集中所有的句子入手,将这些句子定义为 $Z=\{z_i\}_{i=1}^{n}$,其中每个 z_i 代表一个包含词性标注标签的句子. 每个 $z_i \in Z$ 又被细分成了一组词单元,其中每个单元 w_j 包含一个单词和单词所对应的词性标签(例如{service:nn},{good:adj}). 每个单元 w_j 形成 Opinosis 图中的一个节点 v_j,w_j 是该节点的标签. 同时,由于每个节点仅包含一个单词,该节点使用句子标识(SID)来记录它所在句子的编号和它在该句子中所出现位置的编号(PID). 因此每个节点会携带一个位置参考信息组(PRI),该信息组是由表示节点位置信息的{SID:PID}元组对组成的列表. 一个句子的原始结构是利用有向边记录的.

算法 1 Opinosis 图生成算法

输入:句子集合 $Z=\{z_i\}_{i=1}^{n}$
输出:$G=(V,E)$

```
for i = 1:n
    w←Tokenize(z_i)
    sent_size←SizeOf(w)
    for j = 1:sent_size
        LABEL←w_j
        PID←j
        PID←j
        if ExistsNode(G,LABEL)
            v_j←GetExistingNode(G,LABEL)
            PRI_vj←PRI_vj ∪ (SID,PID)
        else
            v_j←CreateNewNode(G,LABEL)
            PRI_vj←(SID,PID)
        end if
        if not ExistingEdge(v_j←v_j-1,G)
            AddEdge(v_j←v_j-1,G)
        endif
    endfor
endfor
```

2.1.2 利用 Opinosis 图计算路径冗余性

定义(路径冗余性) 冗余就是重复的意思,路径 R 的路径冗余性 $r(q,s)$ 是被该路径覆盖的重叠句子个数,某个路径的冗余性越高说明该路径越具有代表性,

$$r(q,s)=n_q \overline{\cap} n_{q+1} \cdots \overline{\cap} n_s \tag{1}$$

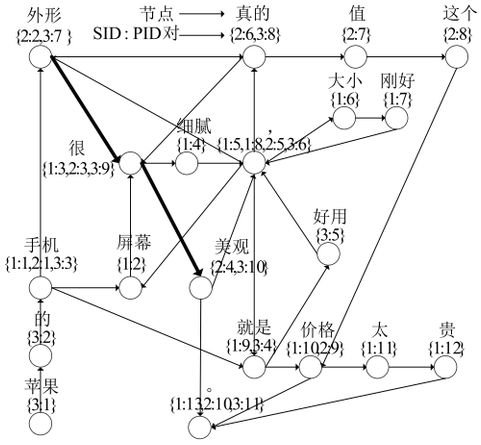
其中 $n_i=\mathrm{PRI}_{v_i}$,$\overline{\cap}$ 是两组 SIDs 的交集,并且要求两组 SIDs 对应的 PIDs 的差距小于参数 σ_{gap},σ_{gap} 满足 $\sigma_{\mathrm{gap}}>0$.

路径冗余性表明在路径的每个点有多少句子在阐明相似的内容,而参数 σ_{gap} 控制着在挖掘冗余时允许的最大差距. 因此,对于一个位于 v_q 和 v_r 之间的常规路径 R,在 $(\mathrm{PID}_{v_{iR}}-\mathrm{PID}_{qR}) \leqslant \sigma_{\mathrm{gap}}$ 时会被视为一个有效的交集.

图 1 表示一个基于四个中文句子的 Opinosis 图. 如图 1 所示,红线所连接的三个词构成一个突显路径. 所谓的突显路径就是路径冗余度相对其它路径较高的子路径.

3 总体框架

本文所提出框架的总体流程如下图 2 所示,框架首先将文本集合转化为 Opinosis 图,再利用 Opinosis 图挖掘出特征标签对句子进行聚类,最终利用本文提出的基于情感和聚类的条件马尔科夫随机游走模型(Sentiment-ClusterCMRW)结合聚类和情感因素生成摘要.



Input :
SID:1. 手机屏幕很细腻 ， 大小刚好 ， 就是价格太贵 。
SID:2. 手机外形很美观 ， 真的值这个价格 。
SID:3. 苹果的手机就是好用 ， 外形真的很美观 。

图1　Opinosis图示例，粗线表示突显路径

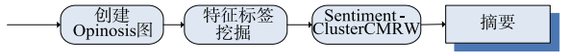


图2　总体框架图

3.1　特征标签挖掘

挖掘特征标签是受 twitter 中 hashtag 的启发,hashtag 是以符号“#”开头置于关键字或短语前方的标签,twitter 将这些 hashtag 作为粗粒度的主题对 tweet 进行分类. 基于类似的思想,本文利用 Opinosis 图挖掘出文本中涉及的高频词语作为特征标签来对文本进行分类.

经过中文分词后大多数可以表达某个含义的中文短语都不会超过三个词,因此本文限制特征标签中所包含的词语个数不超过 3 个. 为了挖掘出原文本集中的特征标签,首先应将原文转化为 Opinosis 图,之后设定阈值 σ 并将 σ_{gap} 设置为 2 从 Opinosis 图中挖掘出所有路径长度不超过 3 并且路径冗余性值大于 σ 的路径作为特征标签存入特征标签集合 $H_1 = \{(h_1,w_1),(h_2,w_2),\cdots,(h_n,w_n)\}$ 中,其中 h_i 为特征标签,w_i 为该标签的路径冗余度值也是它的权重. 注意,如果一条长度超过 3 的路径其冗余性值也超过了 σ,则该路径内部长度不超过 3 的子路径也不要加入到 H_1 中,因为其子路径不一定具有独立的含义.

挖掘出的初始特征标签集合 H_1 去除停用词后集合内大致可以分为领域特征词和非领域特征词,本文为了将两种特征词加以区分设置了参考语料.

本文使用从亚马逊中文网收集的 20 个主题的中文产品评论,当使用其中一种主题的产品评论作为语料集时,将其它主题的语料分别作为参照,几个主题都涉及到的高路径冗余度值的特征标签可以视为非领域特征词,例如如果几个主题都使用了网络用语“伤不起”,该标签就可以视为非领域特征词. 最终本文在 H_1 集合中仅保留了领域相关性强的特征词从而生成了最终的集合 H.

3.1.1　特征标签聚类

H 中可能包含一些内容相似的标签,将它们归为一类可以提高句子分类的准确性并降低生成摘要的复杂度.

为了将 H 中的特征标签聚类,首先按照算法 2 的方法生成一个初始类:

算法 2　初始类生成算法

输入:$H = \{(h_1,w_1),(h_2,w_2),\cdots,(h_n,w_n)\}$
输出:$C_1 = \{c_1,c_2,\cdots,c_m\}$($m$ 为初始类的个数)
0. $C_1 = \{h_1\}$
1. for h_i in H
2. if 对于 C_1 中的任意 c_i 有 $\text{sim}(\boldsymbol{h}_i,\boldsymbol{c}_i) < \sigma_2$($\sigma_2$ 为本文设置的相似度阈值)
3. $C_1 = C_1 + \{h_i\}$
4. end
5. end

之后使用 Kmeans 算法对特征标签进行聚类,其描述如下:

首先将 C_1 中的标签作为聚类中心,之后将 H 中的每个标签划入和其最近的聚类中心所在的那一类,前提是该标签和聚类中心的相似度值大于阈值 σ_2,相似度采用如式(2)所示的余弦测度计算,其中 $\boldsymbol{h}_i$ 和 $\boldsymbol{c}_j$ 分别 h_i 为 c_j 和的词向量. 聚类完毕后重新计算选择各个类的新聚类中心,重复迭代数次直到聚类中心不再改变为止. 如果某个标签和 C_1 中的所有聚类中心相似度值都小于阈值,则将其单独作为一类.

$$\text{sim}(\boldsymbol{h}_i,\boldsymbol{c}_j) = \text{cosine}(h_i,c_j) \tag{2}$$

最终可以将 H 中的标签聚类得到特征类标签集合 C,对于 C 中的任一类,其权重为类中各标签权重的均值.

3.1.2　句子聚类

对于包含特征词的句子,当其中包含了多个特征词时将句子归入权重最大的领域特征词代表的那一特征类中.

至此,所有包含特征词的句子都被划入了特征类当中,这些特征类将作为 3.2 中本文所提出 Sentiment-ClusterCMRW 模型的最上层.

3.2　摘要生成:基于情感和聚类的条件马尔科夫随机游走模型(Sentiment-ClusterCMRW)

为了同时融入特征聚类信息、句间关系以及情感信息,本文在 Wan[12] 提出的基于聚类的条件马尔科夫随机游走模型(ClusterCMRW)的基础上融入了情感信

息提出了基于情感和聚类的条件马尔科夫随机游走模型(Sentiment-ClusterCMRW). 该模型如图 3 所示,中间的一层为传统的表示句间关系的马尔科夫随机游走模型,最下面的一层表示各个句子的情感极性,最上面的一层表示特征类. 上层与中层的虚线表示句子和其所在类的相互影响.

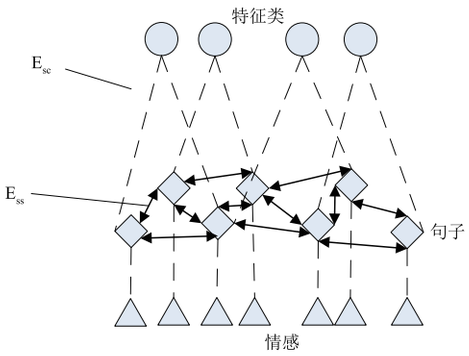


图3 基于情感和聚类条件马尔科夫随机游走模型

最终该三层图模型可以表示为 $G^* = \langle V_s, V_d, V_c, E_{ss}, E_{sc} \rangle$,其中 $V_s = V = \{v_i\}$ 是句子组,$V_c = C = \{c_j\}$ 是特征聚类组,$V_d = D = \{p_j\}$ 是情感极性组,其中任一句子的情感极性 $p_j = \{-1, 0, 1\}$(-1 代表消极,0 代表中性,1 代表积极)将乘以一个情感极性影响因子 θ 融入到其对应句子的词向量 $\boldsymbol{v}_i$ 中作为单独的一项. $E_{ss} = E = \{e_{ij} | v_i, v_j \in V_s\}$ 表示所有句子之间的链接,$E_{sc} = \{ e_{ij} | v \in V_s, c_j \in V_c \text{ and } c_j = \text{clus}(v_i)\}$ 表示一个句子和其所在类之间的关联. $\text{clus}(v_i)$ 表示包含句子 v_i 的类,$\pi(\text{clus}(v_i)) \in [0, 1]$ 表示类 $\text{clus}(v_i)$ 在文本集 D 中的重要性,让 $\omega(v_i, \text{clus}(v_i)) \in [0,1]$ 表示句子 v_i 和其所在类之间的关联强度. $\text{sp}(v_i)$ 表示句子 v_i 的情感极性.

当把聚类情感信息都融入后句子 v_i 到句子 v_j 的过渡概率可以表示如下:

$$
\begin{aligned}
&p(i \to j \mid \text{clus}(v_i), \text{clus}(v_j), \text{sp}(v_i), \text{sp}(v_j)) \\
&= \begin{cases} \dfrac{f(i \to j \mid \text{clus}(v_i), \text{clus}(v_j), \text{sp}(v_i), \text{sp}(v_j))}{\sum_{k=1}^{|V|} f(i \to k \mid \text{clus}(v_i), \text{clus}(v_k), \text{sp}(v_i), \text{sp}(v_j))} \\ \quad ,\text{if } \sum f \neq 0 \\ 0, \text{otherwise} \end{cases} \\
&= \begin{cases} \dfrac{f(i^\theta \to j^\theta \mid \text{clus}(v_i^\theta), \text{clus}(v_j^\theta))}{\sum_{k=1}^{|V|} f(i^\theta \to k^\theta \mid \text{clus}(v_i^\theta), \text{clus}(v_k^\theta))} , \text{if } \sum f \neq 0 \\ 0, \text{otherwise} \end{cases}
\end{aligned}
\tag{3}
$$

$f(i \to j \mid \text{clus}(v_i), \text{clus}(v_j))$ 表示句子 v_i 和句子 v_j 之间在考虑包含两个句子情感聚类情况下的吸引权重. $f(i^\theta \to j^\theta \mid \text{clus}(v_i^\theta), \text{clus}(v_j^\theta))$ 表示将情感因素融入了句子的词向量中. 利用 Wan 的方法线性结合原类($f(i^\theta \to j^\theta \mid \text{clus}(v_i^\theta))$)和目标类($f(i^\theta \to j^\theta \mid \text{clus}(v_j^\theta))$)的吸引权重来计算条件吸引权重如下:

$$
\begin{aligned}
&f(i^\theta \to j^\theta \mid \text{clus}(v_i^\theta), \text{clus}(v_j^\theta)) \\
&= \lambda \cdot f(i^\theta \to j^\theta \mid \text{clus}(v_i^\theta)) + (1-\lambda) \cdot f(i^\theta \to j^\theta \mid \text{clus}(v_j^\theta)) \\
&= \lambda \cdot f(i^\theta \to j^\theta) \cdot \pi(\text{clus}(v_i^\theta)) \cdot \omega(v_i^\theta, \text{clus}(v_i^\theta)) \\
&+ (1-\lambda) \cdot f(i^\theta \to j^\theta) \cdot \pi(\text{clus}(v_j^\theta)) \cdot \omega(v_j^\theta, \text{clus}(v_j^\theta)) \\
&= f(i^\theta \to j^\theta) \cdot (\lambda \cdot \pi(\text{clus}(v_i^\theta)) \cdot \omega(v_i^\theta, \text{clus}(v_i^\theta)) \\
&+ (1-\lambda) \cdot \pi(\text{clus}(v_j^\theta)) \cdot \omega(v_j^\theta, \text{clus}(v_j^\theta)))
\end{aligned}
\tag{4}
$$

其中 $\lambda \in [0,1]$ 是控制原类到目标类相对分布的结合权重. 为了对类的重要性以及句子-聚类相关强度进行测量,本文采用余弦测度的方法.

$f(i^\theta \to j^\theta)$ 表示融入情感因素后句子 v_i 和句子 v_j 之间的连接权重,该权重使用两权重之间的标准余弦测度计算.

$$
f(i^\theta \to j^\theta) = \text{sim}_{\text{cosine}}(\boldsymbol{v}_i^\theta, \boldsymbol{v}_i^\theta) = \frac{\boldsymbol{v}_i^\theta \cdot \boldsymbol{v}_j^\theta}{|\boldsymbol{v}_i^\theta| \times |\boldsymbol{v}_j^\theta|} \tag{5}
$$

$\pi(\text{clus}(v_i^\theta))$ 用于衡量类 $\text{clus}(v_i^\theta)$ 在文本集合 D 中的重要性,衡量的方法就是计算类和整个文本的余弦相似性.

$$
\pi(\text{clus}(v_i^\theta)) = \text{sim}_{\text{cosine}}(\text{clus}(v_i^\theta), D) \tag{6}
$$

$\omega(v_i^\theta, \text{clus}(v_i^\theta))$ 用于衡量句子 v_i 和它的类 $\text{clus}(v_i)$ 之间的连接强度,衡量的方法也是计算二者的余弦相似性.

$$
\omega(v_i^\theta, \text{clus}(v_i^\theta)) = \text{sim}_{\text{cosine}}(v_i^\theta, \text{clus}(v_i^\theta)) \tag{7}
$$

最终新的行归一化矩阵 $\boldsymbol{M}^*$ 定义如下:

$$
\boldsymbol{M}_{i,j}^* = p(i \to j \mid \text{clus}(v_i^\theta), \text{clus}(v_j^\theta)) \tag{8}
$$

为了使 $\boldsymbol{M}$ 成为随机矩阵,所有元素都为 0 的行将被替换为所有元素都为 $1/|V|$ 的平滑向量.

基于矩阵 $\boldsymbol{M}$,句子的排序分数可由其它与之链接的句子得到并可以表达为与 PageRank 算法类似的递归形式.

$$
\text{SenScore}(v_i) = \mu \cdot \sum_{j \neq i} \text{SenScore}(v_j) \cdot \boldsymbol{M}_{j,i} + \frac{(1-\mu)}{|V|} \tag{9}
$$

它的矩阵形式为:

$$
\boldsymbol{\lambda} = \mu \boldsymbol{M}^{\mathrm{T}} \boldsymbol{\lambda} + \frac{(1-\mu)}{|V|} \boldsymbol{e} \tag{10}
$$

其中 $\boldsymbol{\lambda} = [\text{SenScore}(v_i)]_{|V| \times 1}$ 是所有句子的分数向量. $\boldsymbol{e}$ 是所有元素等于 1 的列向量. μ 是阻尼系数,设为 0.85.

最终的马尔科夫链过渡矩阵表示为 $\boldsymbol{A} = \mu \boldsymbol{M}^{\mathrm{T}} + \frac{(1-\mu)}{|V|} \boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}$ 并且句子的分数是通过计算过渡矩阵 $\boldsymbol{A}$ 的特征向量得到的.

在执行阶段,所有句子的初始分数都设为 1 并利用

式(10)来迭代计算各个句子的分数. 通常两次连续迭代各个句子的差值都小于一个固定阈值时算法收敛. 最终句子被按分数排序并根据摘要所规定的压缩比选入摘要. 本文规定的压缩比为1%.

4　实验

本文使用从亚马逊中文网收集的20个主题的中文产品评论来对本文提出的框架及其基准算法的性能进行评测,每个主题有150篇评论,与每个主题相关的文本的平均规模为13000词. 本文请了3名专家对这些语料进行标注. 每名专家会独立地从相关主题抽取8至12个句子(对句子个数的限制是为了满足压缩比的要求). 抽取句子的原则是这几个句子可以还原相关主题各个特征的真实情况并且代表大部分人对该主题的观点. 最终在对算法生成的摘要质量进行评价时,分别将3名专家标注的摘要作为参考标准,并将评价结果取均值. 同时设置评价结果的方差阈值,当3个结果的方差大于该阈值则舍弃该主题的实验结果. 实验结果表明20个主题有2个的主题的实验结果方差过大并且已将其舍弃. 专家所合成摘要的平均压缩比为0.95%,与使用本文方法抽取句子所限制的1%比较接近.

本文用于中文分词和中文句子情感分析的工具是Python的SnowNlp库.

本文使用ROUGE(Recall-Oriented Understudy for Gisting Evaluation)来定量衡量机器生成摘要和人工合成摘要之间的一致性. ROUGE基于机器摘要和人工摘要之间的n元共现,并且是广泛用于评估摘要质量的标准. ROUGE-N为如下所示的n元召回措施:

$$\text{ROUGE-}N = \frac{\sum_{S\in\{\text{Ref Sum}\}}\sum_{n\text{-gram}\in S}\text{Count}_{\text{match}}(n\text{-gram})}{\sum_{S\in\{\text{Ref Sum}\}}\sum_{n\text{-gram}\in S}\text{Count}(n\text{-gram})} \tag{11}$$

其中n表示n元的长度,$\text{Count}_{\text{match}}(n\text{-gram})$表示被评估摘要和参考摘要之间共现$n$元子串的最大数量.

在本文的实验中,使用ROUGE-1,ROUGE-2和ROUGE-W. 研究表明[13] ROUGE-1和ROUGE-2与人工合成摘要之间的关联性较大,同时更高阶的ROUGE-N分数适用于评估摘要的流畅性. ROUGE-W是基于加权最长公共子序列的.

4.1　对比实验

4.1.1　各个算法的性能对比

本文将PageRank、MRW以及ClusterCMRW作为基准算法. 其中PageRank[14]也是一种应用广泛的随机游走模型[14],它同样通过构建图表示句子之间的关系,因此本文选择PageRank作为基准算法. 而MRW和ClusterCMRW都是构建Sentiment-ClusterCMRW的基础,因此也被选作了基准算法. 实验效果如下表1所示:

表1　各个算法的性能对比

算法	ROUGE-1	ROUGE-2	ROUGE-W
PageRank	0.3534	0.0536	0.0841
MRW	0.3512	0.0549	0.0828
ClusterCMRW(Kmeans)	0.3654	0.0897	0.0872
Sentiment-ClusterCMRW(Opinosis)40	0.3789	0.0931	0.1032

从表1给出的实验效果可以看出,MRW和PageRank由于都是通过构建图表示句子之间的关系并且都是随机游走模型因此实验效果比较相近. 而ClusterCMRW相比较MRW和PageRank多考虑了句间的聚类关系及其各个聚类的重要性,因此实验效果相对MRW和PageRank较好. 本文提出的Sentiment-ClusterCMRW由于考虑了句子在各个聚类下的情感倾向是否具有代表性,因此实验效果与ClusterCMRW相比有所提高,这当然也与专家在合成参考摘要时的句子选择倾向有关,因为本文要求专家在抽取句子时选择那些可以还原相关主题各个特征的真实情况并且代表大部分人对该主题的观点的句子.

4.1.2　路径冗余性σ对结果的影响

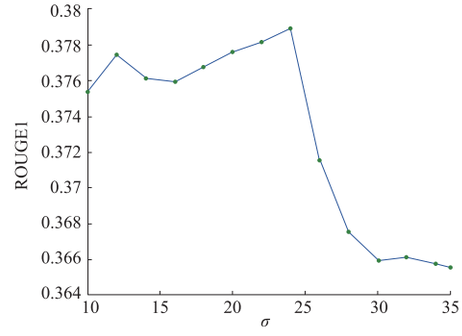
图4表示了路径冗余性σ取不同的值时对实验评价指标ROUGE1值和ROUGE2值的影响. 从实验效果图可以看出,当σ取值15~25左右时实验效果较好. 当σ取值较小时实验效果稍差,可能的原因是虽然当σ较小时H_1中分成的特征类会比较多,但是由于特征标签聚类的过程以及参考语料的设置对最终的C集合影响不大,但是会造成实验运行时间的增长;而当σ过大时实验效果较差,原因是σ值过小会导致许多应该出现的特征类的缺失,从而使生成的摘要信息量变小.

4.1.3　情感因子θ对结果的影响

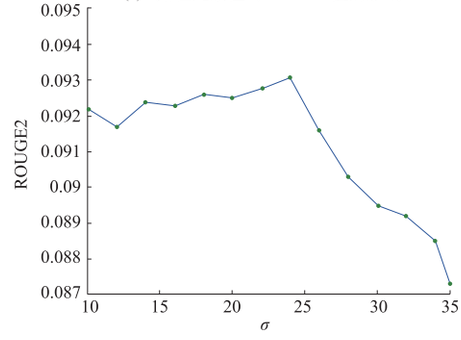
图5表示了情感因子θ取不同的值时对实验评价指标ROUGE1值和ROUGE2值的影响. 在取值为0时本模型结果仍然比ClusterCMRW模型稍好,原因是本文所提出的基于Opinosis图的特征聚类方法相比ClusterCMRW模型所使用的kmeans聚类方法更加面向特征. 在θ取值较小时实验效果基本会随着θ值的升高而改善,可能的原因是随着θ值的升高使具有代表性观点的句子被抽取的几率更大;但是当θ取值过高时由于情感因素占的比重过大而降低了聚类的影响则会导致实验效果的降低.

4.1.4　控制参数λ对结果的影响

图6表示了控制参数λ取不同的值时对实验效果的影响. 从图中可以看出λ的取值在0和1之间某些值时由于使吸引权重$f(i\to j|\text{clus}(v_i),\text{clus}(v_j))$同时融合

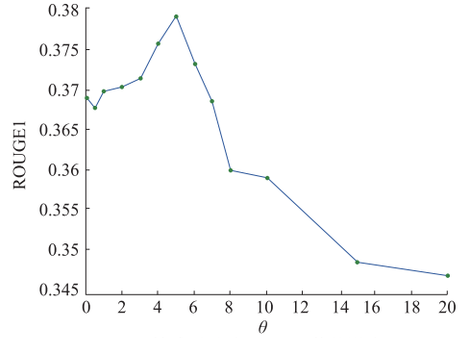


(a) 路径冗余性对ROUGE1值的影响

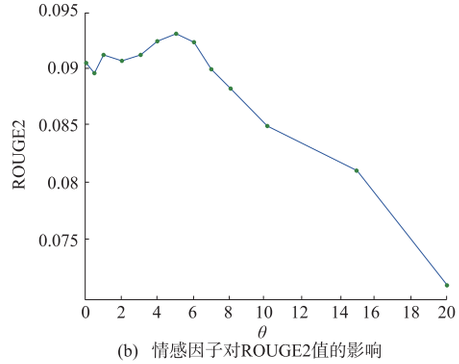


(b) 路径冗余性对ROUGE2值的影响

图4 路径冗余性对结果的影响



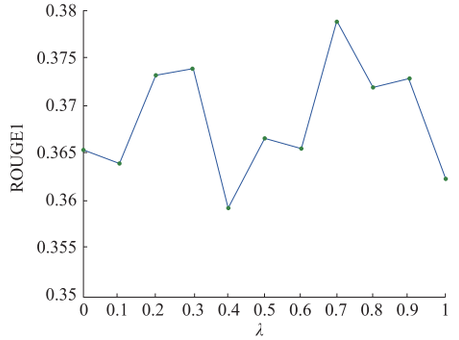
(a) 情感因子对ROUGE1值的影响



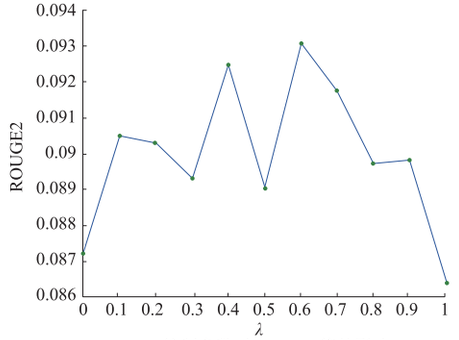
(b) 情感因子对ROUGE2值的影响

图5 情感因子对结果的影响

了原类和目标类实验效果更好,而当 λ 取 0 或 1 时吸引权重将会仅倾向于原类或目标类,达不到最佳的实验效果.



(a) 控制参数对ROUGE1值的影响



(b) 控制参数对ROUGE2值的影响

图6 情感因子对结果的影响

5 结束语

本文提出了一种基于 Opinosis 图和马尔科夫随机游走模型的情感摘要框架. 框架首先利用 Opinosis 图挖掘文本中的特征词,同时设置了参考语料用于增强所挖掘特征词的领域相关性,该方法具有较小的领域依赖性. 其次框架利用本文提出的 Sentiment-ClusterCMRW 模型同时融入情感和分类因素对句子进行排序. 实验结果表明,本文提出的方法与基准算法相比在 ROUGE 值上具有明显提高而且本文提出的模型具有一定的鲁棒性. 在接下去的工作中我们将考虑如何通过对所抽取句子更合理的排序使生成的摘要具有更好的可读性.

参考文献

[1] 林莉媛,王中卿,李寿山,等. 基于 PageRank 的中文多文档文本情感摘要[J]. 中文信息学报,2014,28(2):85-90.
Lin Li-yuan, Wang Zhong-qing, Li Shoushan et al. Chinese multi-document opinion summarization via PageRank[J]. Journal of Chinese Information Processing, 2014, 28(2): 85-90. (in Chinese)

[2] 荀静,刘培玉,杨玉珍,张艳辉. 基于潜在狄利克雷分布模型的多文档情感摘要[J]. 计算机应用,2014,34(6):1636-1640.
XUN Jing, LIU Peiyu, YANG Yuzhen, ZHANG Yanhui.

Multi-document sentiment summarization based on latent Dirichlet allocation model[J]. Journal of Computer Applications,2014,34(6):1636 – 1640.

[3] Jiang L,Yu M,Zhou M,et al. Target-dependent twitter sentiment classification[A]. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics[C]. Portland, Oregon: ACM, 2011. 151 – 160.

[4] Singh V K,Piryani R,Uddin A,et al. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification[A]. Automation,Computing, Communication, Control and Compressed Sensing (iMac4s),2013 International Multi-Conference onAutomation,Computing,Communication,Control and Compressed Sensing[C]. India,Kottayam:IEEE,2013. 712 – 717.

[5] Di Fabbrizio G,Aker A,Gaizauskas R. Summarizing online reviews using aspect rating distributions and language modeling[J]. IEEE Intelligent Systems,2013,28(3):28 – 37.

[6] Yu J,Zha Z J,Wang M,et al. Aspect ranking:identifying important product aspects from online consumer reviews [A]. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics[C]. Stroudsburg,PA,USA: ACM,2011. 1496 – 1505.

[7] Wang D,Zhu S,Li T,et al. Multi-document summarization using sentence-based topic models. [A]. Meeting of the Association for Computational Linguistics[C]. Suntec,Singapore:ACM,2009. 297 – 300.

[8] Zhao W X,Jiang J,Yan H,et al. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid[A]. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics[C]. Massachusetts,USA:ACM,2010. 56 – 65.

[9] Jo Y,Oh A H. Aspect and sentiment unification model for online review analysis[A]. Proceedings of the fourth ACM international conference on Web search and data mining [C]. Hong Kong,China:ACM,2011. 815 – 824.

[10] 欧阳继红,刘燕辉,李熙铭,等. 基于 LDA 的多粒度主题情感混合模型[J]. 电子学报,2015,43(9):1875 – 1880.
Ouyang Ji-hong, Liu Yan-hui, Li Xi-ming, et al. Multi-grain sentiment/topic model based on LDA[J]. Acta Electronica Sinica,2015,43(9):1875 – 1880.

[11] Ganesan K,Zhai CX,Han J. Opinosis:A graph-based approach to abstractive summarization of highly redundant opinions[A]. Proceedings of the 23rd International Conference on Computational Linguistics. Asisocation for Computational Linguistics [C]. Beijing: ACM, 2010. 340 – 348.

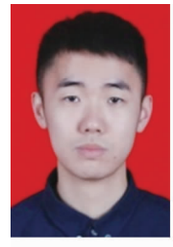
[12] Wan X,Yang J. Multi-document summarization using cluster-based link analysis[A]. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. New York,NY, USA:ACM,2008. 299 – 306.

[13] Lin C Y,Hovy E. Automatic evaluation of summaries using n-gram co-occurrence statistics[A]. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics [C]. Stroudsburg, PA, USA: ACM, 2003. 71 – 78.

[14] Gustavsson P,Jönsson A. Text summarization using random indexing and pagerank[A]. Proceedings of the third Swedish Language Technology Conference (SLTC-2010) [C]. Linköping,Sweden:Research Gate,2010. pp – pp

[15] Li F,Tang Y,Huang M,et al. Answering opinion questions with random walks on graphs[A]. Joint Conference of the, Meeting of the ACL and the, International Joint Conference on Natural Language Processing of the Afnlp: Volume. Association for ComputationalLinguistics[C]. Suntec,Singapore:ACM2009:737 – 745.

作者简介



康世泽　男,1991 年 5 月出生,内蒙古呼伦贝尔人. 目前在国家数字交换系统工程技术研究中心攻读硕士学位,主要研究方向为数据挖掘、自然语言处理.
E-mail:xdkangshize@163.com

马　宏　男,1968 年出生,江苏东台人. 国家数字交换系统工程技术研究中心研究员,主要研究方向为数据挖掘、电信网信息关防.

黄瑞阳　男,1986 年出生,福建漳州人,博士. 国家数字交换系统工程技术研究中心助理研究员,主要研究方向为文本挖掘、图挖掘.