软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(12):2777−2789 [doi: 10.13328/j.cnki.jos.004729]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于改进语义距离的网络评论聚类研究*

杨 震, 王来涛, 赖英旭

(北京工业大学 计算机学院,北京 100124)

通讯作者: 杨震, E-mail: yangzhen@bjut.edu.cn

摘 要: 针对在线评论,提出了一种短文本语义距离计算模型,将文本距离看成是形式距离和单元语义距离的综合.首先,在对变异短文本进行预处理的基础上,以中文词语为单位,利用词典进行语义扩展,计算短文本间最大匹配距离,将其作为衡量短文本间形式距离的指标;其次,基于短文本中的实义单元和非实义单元的不同作用,利用改进的编辑距离算法计算短文本的单元语义距离;最后,利用加权的方法将形式距离和单元语义距离综合为文本距离,并将其应用于网络在线评论的聚类分析.特别地,为了缓解短文本长度差异所造成的计算误差,提出利用词表长度对距离进行惩罚,并根据 Zipf's Law 和 Heap's Law 的对应关系,给出了一种文本词表长度的估计方法,并阐明了文本 Zipf 指数α对长度惩罚的关键性作用机制.实验结果表明,改进算法优于传统方法,聚类性能显著提升.

关键词: 文本聚类;在线评论;语义距离;长度惩罚;Heap's Law;Zipf's Law

中图法分类号: TP181

中文引用格式: 杨震,王来涛,赖英旭.基于改进语义距离的网络评论聚类研究.软件学报,2014,25(12):2777−2789. http://www.jos.org.cn/1000-9825/4729.htm

英文引用格式: Yang Z, Wang LT, Lai YX. Online comment clustering based on an improved semantic distance. Ruan Jian Xue Bao/Journal of Software, 2014,25(12):2777−2789 (in Chinese). http://www.jos.org.cn/1000-9825/4729.htm

Online Comment Clustering Based on an Improved Semantic Distance

YANG Zhen, WANG Lai-Tao, LAI Ying-Xu

(College of Computer Science, Beijing University of Technology, Beijing 100124, China)

Corresponding author: YANG Zhen, E-mail: yangzhen@bjut.edu.cn

Abstract: An improved semantic distance for short text is proposed. The new method calculates the semantic distance between two word strings as balance of the extent of word sequence alignment and the meaning matching between word strings. First, after linguistic preprocessing, the extent of word sequence alignment is computed by the structural distance which measures the maximum matching based on the HIT-CIR Tongyici Cilin (extended edition). Then the meaning matching between word strings is computed by an improved edit distance which allocates each word a weight according to its word type. Finally, the semantic distance between the word strings is measured as a balance of structural distance and word meaning matching distance. In addition, in order to eliminate the influence of the sentence length, the proposed semantic distance is adjusted using the distinct word count estimated by the Heap's law and Zipf law. Experimental results show that the presented methods are more efficient than the classical edit distance models.

Key words: text clustering; online comment; semantic distance; length penalty; Heap's law; Zipf's law

移动业务的普及,催生出一种新的网络语言表达范式——短文本范式.据《中国互联网络发展状况统计报告(2014 年 7 月 21 日版)》(http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/201407/P020140721507223212132.pdf)

* 基金项目: 国家自然科学基金(61001178); 国家软科学研究计划(2010GXQ5D317); 北京市优秀人才计划; 北京市属高等学校青年拔尖人才计划(CITTCD201404052); 北京市教育委员会科技计划(KM201210005024); 北京工业大学基础研究基金; 可信计算北京市重点实验室开放课题

收稿时间: 2014-05-05; 修改时间: 2014-08-21; 定稿时间: 2014-09-30

最新公布数据,截至 2014 年 6 月,中国网民规模达 6.32 亿,其中,手机网民规模 5.27 亿,互联网普及率达到 46.9%.网民上网设备中,手机使用率达 83.4%,首次超越传统 PC 整体 80.9%的使用率,手机作为第一大上网终端的地位更加巩固.手机媒体作为现代信息的重要载体,已被一些学者誉为继报纸、广播、电视、网络之后的第五大媒体,其使用已渗透到社会的各个领域(http://newspaper.jfdaily.com/jfrb/html/2010-02/19/content_282432.htm).相比过去,用户倾向用更简洁、更个性的方式表达个人观点,这样的变革,给中文信息处理领域的研究者带来了新的挑战和机遇.

以本文的研究对象——网络评论聚类分析为例,其即为一种典型的面向短文本的应用.网络评论通常开始于某个公共事件或热点话题,表达内容具有较强的主观性,反映出大众对公共事件的态度[1],其典型来源有微博、短信、BBS 社区等.在线评论具有快速传播及影响广泛的特点,不仅表达了评论者自身的观点,也会影响其他参与者的观点[2].通过对用户发布的在线评论进行基于内容(语义)的聚类分析,可以及时掌握人们对各种热点话题的观点和立场,对于国家、企业和社会都具有重要意义,引起了学术界的高度重视[3].

研究者一开始并没有充分考虑到短文本处理的特殊性,希望沿用处理传统长文本的成功方法处理短文本.比较典型的有:

1) 基于贝叶斯和朴素贝叶斯的方法.

Kyosuke 等人[4]在 Twitter 的信息过滤中使用基于 *n*-gram 模型的朴素贝叶斯模型,将输入 Twitter 流的长度作为变量,利用动平均方法计算词条件概率,较好地平衡了文本长度的影响.Fan 等人[5]提出基于组合朴素贝叶斯和 *K* 近邻分类器的两步中文短文本分类措施.

2) 利用全体文本集或上下文关系将短文本整体考虑来降低文本长度的影响.

Yang 等人[6]尝试借由时间、空间、联系等要素挖掘文本间隐含的关联关系,重构文本上下文范畴,解决短文本情感极性分类问题.Sriram 等人[7]针对 Twitter 数据,分析了 Bag-of-words 类处理方式在短文本分类中的局限性,提出了从用户信息(user profile)上下文进行限定领域的特征抽取方法.Mihalcea[8]提出了一种基于语料集-知识集的短文本语义近似度测量方法,充分考虑短文本本在大规模语料中表现的特性,在识别转述问题中取得较好的性能.Qiang 等人[9]使用 LSA+ICA 整体提取短文本的语义,从而回避了文本长度较短的问题.

但这些在处理长文本时取得成功的方法,对短文本的处理效果与现实要求还有巨大的差距.追本溯源,正是因为文本对象特征(即语言)本身所固有的多义性,即存在一词多义(polysemy)和一义多词(synonymy)的特点——面对短文本特征的稀疏性和上下文缺失的情况,造成语义难以明辨,理解偏差无法消解,最终形成短文本底层特征和高层语义之间巨大的语义鸿沟.以文本表示环节为例,长文本常用的文本表示方法(如 VSM 模型)通常会遇到的特征稀疏性问题,会因为文本较短而进一步加剧.如在线评论、微博、短信等短文本,稀疏性可达到 95%~99%,使得短文本处理在文本表示环节就遭遇极大的困难[6].

另外,一些在长文本处理中由于计算复杂度过高而难以大规模展开的方法,如基于文本语义和形式比较的方法,因短文本特性又重新受到研究者的重视.目前比较流行的方法包括:

1) 基于文本语义相似度的方法,利用本体(ontology)或其他语义资源计算短文本的语义相似度.

Yang 等人[10]通过利用 WordNet 计算短文本字、词、句的相似性,引入 Isomap 流形降维研究中文词汇在语义空间(分类空间)的分布聚集情况.刘宇鹏等人[11]基于 WordNet 进行词义消歧来指导混淆网络对齐,句子的相似性计算使用了二分图的最大匹配算法.李彬等人[12]根据语言学知识,通过分析语句中各成分之间的依存关系揭示其句式结构,通过对有效搭配对的加权计算,得到语句的相似度.彭京等人[13]基于语义内积空间模型进行文本聚类,利用内积空间建立了针对中文概念、词和文本的相似度计算方法.通过一个两阶段处理过程,即向下分裂和向上聚合,完成文本聚类.李素建等人[14]利用同义词词林和 HowNet 两种语义资源计算词语间的相似度与相关度,通过加权计算,得到语句间的相关度.

2) 基于文本形式相似性的方法,通过比较文本包含的词和词序来计算短文本间语义相似度.

Eric 等人[15]提出一种将概率模型应用于计算字符编辑距离的方法.穗志方等人[16]借助于少量的语法、语义信息,建立了一个基于骨架依存树的语句相似度计算模型.Niladri[17]提出了一种基于线性模型的方法,综合考虑

实用、语法和语义相似性,通过利用预先按照某一个原则设计好的句对,对调查对象进行测试来实现模型估计.

我们认为:要对短文本间相似性进行合理化度量,文本形式相似性和文本单元语义两个方面不可偏废,必须综合考虑.而现有的工作较少地考虑两者之间的结合,或只是两种方法的简单加和,没有考虑其内在联系.有鉴于此,本文提出了一种改进的短文本语义距离,将文本距离看成是形式距离和单元语义距离的综合:首先,在对变异短文本进行预处理的基础上,以中文词语为单位,利用词典进行语义扩展,计算短文本间最大匹配距离,将其作为衡量短文本间形式距离的指标;其次,基于短文本中的实义单元和非实义单元的不同作用,利用改进的编辑距离算法计算短文本的单元语义距离;最后,利用加权的方法将形式距离和单元语义距离综合为文本距离.最重要的是,为了消除不同文本长度对文本距离的影响,提出利用词表长度对长度进行惩罚的理论,并通过 Zipf's Law 和 Heap's Law 的对价关系,给出一种词表长度的估计方式.

第 1 节首先给出一种改进的短文本间语义距离度量模型,并证明语义计算方法在短文本相似性计算中的作用强于长文本.第 2 节利用词表长度对长度进行惩罚,并给出一种词表长度的估计方式.阐明文本 Zipf 指数α对长度惩罚的关键性作用机制.第 3 节在介绍在线评论聚合系统的系统框架的基础上,给出实验结果分析.最后对工作进行总结与展望.

1 基于改进语义距离的文本距离计算

要精确定义文本语义,是一个涉及计算机、人工智能、心理学和认知科学等众多学科的复杂问题.任何定义都可能是片面且富有争议的.但是从统计语言学的角度来看,我们只能计算那些能够用统计特性描述的文本差异.

假设Σ是单词表,$\Phi\in\Sigma$是空字符,Σ^*是由Σ组成的句子集合.给定句子 $R,S\in\Sigma^*$,表示为 $R=r_1r_2\ldots r_l,S=s_1s_2\ldots s_m$,其中,$r_i,s_i$分别表示句子中的第 i 个单词.那么文本距离可以定义成形式距离和单元语义距离的综合,具体计算方法见表 1.

Table 1 Improved short text similarity measure

表 1 改进的短文本距离计算方法

输入:	短文本 $R=r_1r_2\ldots r_l,S=s_1s_2\ldots s_m$;
步骤 1:	比较短文本 R,S 的长度,依据所有词对间的语义相似度①,以较短文本向较长文本进行一对一不重复最大匹配对齐②,使其形式结构与较长文本形式达到最大相似性,记录对齐操作次数 d_1.
步骤 2:	在上一步对齐的基础上,区分标记文本中的实义词单元和非实义单元③.
步骤 3:	通过"插入"、"删除"和"替换"这 3 种编辑操作,并根据语义,为不同的编辑操作赋以不同的权值,记录以词为单元,计算将一个句子变换为另一个句子所需要的最少编辑操作数 $d_2$④.
步骤 4:	将 d_1 和 d_2 归一化的基础上⑤, 将短文本距离 $DIS(R,S)$定义成形式距离 d_1 和单元语义距离 d_2 的综合⑥.
输出:	短文本 R,S 距离 $DIS(R,S)$.

① 实验中的词语间相似度是使用《同义词词林扩展版》[14,18]计算得到的.《同义词林扩展版》是在《同义词词林》[19]的基础上,按照人民日报语料库中词语的出现频度,剔除罕用词和非常用词完成的.基于此,词语间相似度可以如下定义:

定义 1. 假设为 r_i,s_j 两词各自在《同义词林扩展版》中的义项集合为 $M(r_i)$和 $M(s_j)$,义项 $a\in M(r_i)$和 $b\in M(s_j)$间的相似度定义为

$$sim(a,b)=n/(N+1) \tag{1}$$

n 为义项代码开始不同的级数,N 为编码的位数.那么 r_i,s_j 两词的相似度定义为:

$$sim(r_i,s_j)=\max_{a\in M(r_i),b\in M(s_j)} sim(a,b) \tag{2}$$

② 语义对齐是指通过计算两个短文本任意词语的相似度,根据词语最大相似度进行一对一不重复匹配,调整短句的词语顺序,使短句的形式结构与长句形式达到最大的形式相似度[11−16].

③ 根据语言学的相关知识,句子的语义由实义单元(主、谓、宾等)和修饰词(定、状、补等)组成.实义单元

表达了句子的主要意义,修饰词起了次要修饰作用.因此,为了区分贡献度,给实义和非实义单元赋以不同权值是自然选择.在不引起歧义的基础上,简化处理将短文本中的全部名词、代词、动词、形容词作为中文实义单元,其他词性的词语如数词、量词、副词等作为非实义单元.

④ 与传统以字为单位的编辑操作不同,这里以词为单位进行编辑操作.从汉语的分析来看,使用以字为单位计算编辑距离的方法,得到的结果只能是形式上的相似度量.采用传统的插入、删除和替换这 3 种编辑操作,但是根据操作对象词的属性(是否为实义单元)和语义相似度为不同的编辑操作赋以不同的权值.具体计算方式如下:

定义 2. ω_1/ω_2 为插入或删除实义/非实义单元的操作权值,γ_1/γ_2 为替换实义/非实义单元的操作权值,θ为对近义词的替换权值.$a_1/a_2,b_1/b_2,c$ 为对应的编辑操作次数.那么,以词为单元将一个句子变换为另一个句子所需要的最少编辑操作数 d_2 可如下计算:

$$d_2=\omega_1\times a_1+\omega_2\times a_2+\gamma_1\times b_1+\gamma_2\times b_2+\theta\times c \tag{3}$$

为不同编辑操作赋权值是考虑到同样的编辑操作作用于不同词语对象,效果不同.例如:

句 1:我喜欢水果;

句 2:我喜欢西瓜;

句 3:我喜欢电脑.

传统编辑距离得到句 1 和句 2,句 1 和句 3 的距离一样.但如果考虑到“水果”和“西瓜”之间较相似,而“水果”和“电脑”相似度较低,以此为出发点,依据语义相似度(方法如操作①)计算编辑操作代价是合理的.而为实义单元和非实义单元赋不同权值,是考虑到其在句子表达中的不同作用.

不同权值的设置体现对语义不同的理解,实验中,权值设置按照以下原则:

- 语义大于形式,既有词义的对齐代价小于插入删除的代价,$\omega_1>\lambda,\omega_2>\lambda$;
- 既有语义的替换大于语义的增减,$\gamma_1>\omega_1,\gamma_2>\omega_2$;
- 实义单元的操作大于非实义单元,$\omega_1>\omega_2,\gamma_1>\gamma_2$;
- 近义词的操作代价小于非近义词的代价,$\lambda>\theta,\omega_1>\theta,\omega_2>\theta,\gamma_1>\theta,\gamma_2>\theta$;
- 形式距离、单元语义距离的权值归一化处理,$\lambda+\omega_1+\omega_2+\gamma_1+\gamma_2+\theta=1$.

⑤ 归一化是指将操作次数利用短文本长度进行归一化处理.假设 d 为操作次数,$|R|,|S|$分别为文本 R,S 的文本长度,则操作次数 d 的归一化数定义为

$$\tilde{d}=\frac{2\times d}{|R|+|S|} \tag{4}$$

⑥ $DIS(R,S)$定义成形式距离 d_1 和单元语义距离 d_2 的综合:

定义 3. 短文本 R,S 的语义距离定义为形式距离 $d_1(R,S)$和单元语义距离 $d_2(R,S)$的综合:

$$D(R,S)=d_1(R,S)+d_2(R,S) \tag{5}$$

给定句子 $R,S\in\Sigma^*$,$|R|,|S|$表示相应句长,那么经典编辑距离的时间计算复杂度为 $O(|R|\times|S|)$,即 $O(n^2)$[15].本文改进的距离的计算流程如图 1 所示,算法时间计算开销要由两个部分组成:a) 形式距离计算开销 A;b) 单元语义距离计算开销 B.其中,

A 表示依据所有词对间的语义相似度,进行句子 R,S 对齐的操作数,其计算时间复杂度可以表示为 $O(|R|\times|S|\times T)$,即 $O(Tn^2)$.其中,T 是基于同义词词林计算两个词的语义相识度的时间复杂度;

B 表示在区分实义单元和非实义单元并根据语义为不同的编辑操作赋以不同权值基础上,对对齐后句子进行编辑操作的时间复杂度,因而可近似表示为 $O(|R|\times|S|)$,即 $O(n^2)$.

综上所述,改进距离测度的时间复杂度为 $O(Tn^2)+O(n^2)$.同时,考虑到语义计算 T 的高复杂度,本文所提出的改进语义距离时间复杂度高于经典编辑距离的时间复杂度 $O(n^2)$.但值得注意的是,改进算法主要面对短文本,即 n 比较小的情况下,整体时间开销能够满足实际应用要求.这一点也在我们后续的实验比较中得到验证.正如我们在论文中讨论的,一些在长文本处理中由于计算复杂度过高而难以大规模展开的方法,如基于文本语义和

形式比较的方法,因短文本特性又重新受到研究者的重视.

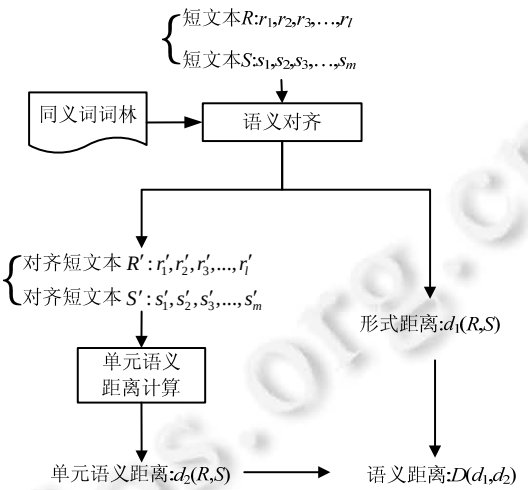

Fig.1 Flow of text distance computation

图 1 文本距离计算流程

2 基于词表长度惩罚的短文本距离计算方法

2.1 句子长度对文本距离度量的影响

第 1 节给出了一种改进的语义距离度量,虽然简单,但充分融合了形式距离和单元语义距离.而现有的工作较少地考虑两者之间的结合,或只是两种方法的简单加和,没有考虑其内在联系.主要原因之一是在处理长文本时,引入语义度量方法所带来的计算量过于巨大,而当处理任务比较简单的时候,引入这样的高复杂度操作是不恰当的.而短文本由于文本较短,却给了基于语义的方法实际应用的机会.此外,用户处理目的日趋复杂,也使得研究者开始考虑基于语义的方法.

实际上,除了计算复杂度的考虑,接下来我们的分析也将说明:引入语义度量方法所带来的文本差异性,随着文本长度的增加而逐渐弱化.而只有当文本长度较短时,这样的差异性才是显著的.这也从另外一方面支持了文献[11–17]的结论,即基于语义的方法没有在长文本分析中表现出明显的优势.同时我们也发现文本长度在短文本距离计算中起到的决定性作用,特别是当两个短文本长度不一致的时候,那么句子间语义相似度将由长句子所主导.这毫无疑问会给计算带来误差,这也是第 2.2 节基于词表长度的短文本距离惩罚算法的出发点.

命题 1. 语义计算方法在短文本相似性计算中的作用强于长文本.

证明:假设Σ是单词表,$\Phi\in\Sigma$是空字符,Σ^*是由Σ组成的句子集合.给定句子 $R,S\in\Sigma^*$,表示为 $R=r_1r_2\ldots r_l$,$S=s_1s_2\ldots s_m$,句子长度分别为 l 和 m,其中,r_i,s_i 分别表示句子中的第 i 个单词.假设文本按单词码本Σ以随机方式产生,即任意字符 $c_i\in\Sigma$以概率 P_{c_i} 出现在句子中,且字符出现统计独立,那么,随机文本 R,S 中同时出现字符 c_i 的概率为

$$P(c_i)=\sum_{k=1}^{l}C_l^kP_{c_i}^k(1-P_{c_i})^{l-k}\sum_{r=1}^{m}C_m^rP_{c_i}^r(1-P_{c_i})^{m-r} \tag{6}$$

假设字符 c_i 的语义等价类定义为

$$c_i+=\{c_j|c_j\in\Sigma,sim(c_i,c_j)\geqslant\varepsilon\} \tag{7}$$

表示与字符 c_i 在语义上类似的字符集合,其中,$sim(\cdot,\cdot)$表示语义近似度量.假设文本按单词码本Σ以随机方式产生,即任意字符 c_i 的等价类 c_i+以概率 Q_{c_i} 出现在句子中,那么随机文本 R,S 中同时出现等价类 c_i+的概率为

$$P(c_i+)=\sum_{k=1}^{l}C_l^kQ_{c_i}^k(1-Q_{c_i})^{l-k}\sum_{r=1}^{m}C_m^rQ_{c_i}^r(1-Q_{c_i})^{m-r} \tag{8}$$

这样一来,$P(c_i)$可以用来度量不引入语义计算时,句子 R,S 相似的概率;$P(c_i+)$可以用度量来引入语义计算时,句子 R,S 相似的概率.显然, $Q_{c_i}>P_{c_i}$,计 $Q_{c_i}=P_{c_i}+\Delta$,那么 $P(c_i+)$和 $P(c_i)$的差异为

$$\delta=P(c_i+)-P(c_i)=[1-(1-P_{c_i}-\Delta)^l][1-(1-P_{c_i}-\Delta)^m]-[1-(1-P_{c_i})^l][1-(1-P_{c_i})^m] \tag{9}$$

首先,我们考察引入语义计算产生的句子差异δ和Δ间极限、偏导和弹性系数关系:

$$\lim_{\Delta\to\infty}\delta=0 \tag{10}$$

$$\frac{\partial\delta}{\partial\Delta}=2m[1-(1-p-\Delta)^m][1-p-\Delta]^{m-1}>0 \tag{11}$$

$$\rho=\frac{\partial\delta}{\partial\Delta}.\frac{\Delta}{\delta}\approx1>0 \tag{12}$$

在这里,为了简化分析,假设 $l=m$,阶乘用斯特林公式 $n!\approx\sqrt{2\pi n}\left(\frac{n}{\mathrm{e}}\right)^n$ 近似.从公式(10)~公式(12)可知,δ是Δ的单调增函数,且弹性系数$\rho>0$ 表示δ和Δ存在正相关性.通常,实际语言中Δ较小.这一点不难理解,若Δ较大,表明语言本身存在较大歧义性.

其次,当句子长度 l 和 m 较大时,即,当 R,S 为长文本时:

$$\lim_{l,m\to\infty}\delta=0 \tag{13}$$

$$\frac{\partial\delta}{\partial m}=2\log(1-p-\Delta)[(1-p-\Delta)^m-1](1-p-\Delta)^m-2\log(1-p)[(1-p)^m-1](1-p)^m \tag{14}$$

从公式(13)、公式(14)可知:虽然句子长度 l 和 m 与语义差异δ间不存在单调关系,但是无论 P_{c_i} 和Δ的取值如何,随着文本长度 l 和 m 的增加,$\delta\to0$.说明对于长文本来说,随着文本长度的增加,由语义计算带来的差异性逐渐被文本长度消弭.特别是对于实际语言, P_{c_i} 和Δ都相对较小,这样的趋势尤为显著.

再次,当句子长 l 和 m 长度差异性较大时,即当 R,S 一个为长文本,另外一个为短文本时,文本间语义差异由较长的文本主导,$\delta\to0$,即:

$$\lim_{l\to\infty}\delta=0,\lim_{m\to\infty}\delta=0 \tag{15}$$

而只有当 R,S 为短文本时,由Δ主导差异性δ.即在短文本相似性计算中,语义度量方法作用较为显著.

综上可知,语义计算方法在短文本相似性计算中的作用强于长文本. □

2.2 基于词表长度惩罚的短文本距离计算方法

上一节说明了短文本范式给了语义计算方法新的机会,这不仅仅是因为计算复杂度的原因,更主要是因为语义计算的方法会随着文本长度的增加,其作用不断弱化.实际上,从定理 1 可以简单推断:语义计算受到文本长度的影响,特别是当文本长度不一致的时候,较长的文本起的作用更大.从信息论的角度来看,较长的文本天然地比稍短的文本蕴含更多的信息,如果不对句子长度进行处理,势必会造成各种基于语义的距离度量方法退化成结构形式比较.对于传统长文本来说,由于平均句长较长,这样的影响还不明显;但对于短文本来说,本来平均句长就较短,句子长度所造成的影响就非常显著了.聚类结果按句长分布,长句和长句在一起,短句和短句在一起,而忽视句子的意义.这毫无疑问会给计算带来误差.

为了解决上述问题,研究者们最直接地想到使用文本原始长度对距离进行惩罚,希望借此消弭句子长度不同所带来的影响.具体做法如下:

定义 3. 假设Σ是单词表,$\Phi\in\Sigma$是空字符,Σ^*是由Σ组成的句子集合.给定句子 $R,S,T\in\Sigma^*$,表示为 $R=r_1r_2\ldots r_l$, $S=s_1s_2\ldots s_m$,$T=t_1t_2\ldots t_n$,其中,r_i,s_i,t_i 分别表示句子中的第 i 个单词.$|\cdot|$表示文本的长度,即文本中所有包含的单词数目,$\|\cdot\|$表示文本的词表长度(distinct words count),即文本中所有不相同单词数量.

定义 4. 假设 $DIS(\cdot,\cdot)$是定义在Σ^*所张成的空间的二元短文本语义距离测度,其满足:

1. $DIS(R,S)\geqslant 0$;
2. 当且仅当 $R=S, DIS(R,S)=0$;
3. $DIS(R,S)=DIS(S,R)$.

定义 5. $LDIS(\cdot,\cdot)$表示基于句长惩罚的语义距离测度,定义为

$$LDIS(R,S)=DIS(R,S)/\max(|R|,|S|) \tag{16}$$

其满足:

1. $LDIS(R,S)\geqslant 0$;
2. 当且仅当 $R=S, LDIS(R,S)=0$;
3. $LDIS(R,S)=LDIS(S,R)$.

虽然研究者尚未彻底弄清句长的作用机制,但这种简单的方法在实际应用中却表现出惊人的有效性,被应用在各种领域的应用当中[20].但研究者也在怀疑是否基于句长的惩罚方法是最优解决方案.实际上,引入句长惩罚的初衷是为了解决不同长度文本所承载的信息容量不同的问题.如果从这个角度来看,混同考虑所有单词而不区分文本的具体内容是不合适的,特别是不应忽视文本中重复出现的单词所造成的影响.正是考虑到重复单词出现的影响,我们提出了一类基于词表长度惩罚的语义测度,定义如下:

定义 6. $NDIS(\cdot,\cdot)$表示基于词表长度惩罚的语义距离测度,定义为

$$\begin{aligned} NDIS\_\mathrm{I}(R,S) &= \left(\frac{\max(|R|,|S|)}{\max(\|R\|,\|S\|)}\right)\cdot\left(\frac{DIS(R,S)}{\max(|R|,|S|)}\right) \\ &= \left(\frac{DIS(R,S)}{\max(\|R\|,\|S\|)}\right) \\ &= \left(\frac{\max(|R|,|S|)}{\max(\|R\|,\|S\|)}\right)\cdot LDIS \\ &= \beta\cdot LDIS \end{aligned} \tag{17}$$

$$\begin{aligned} NDIS\_\mathrm{II}(R,S) &= \left(\frac{\max(\|R\|,\|S\|)}{\max(|R|,|S|)}\right)\cdot\left(\frac{DIS(R,S)}{\max(|R|,|S|)}\right) \\ &= \left(\frac{\max(\|R\|,\|S\|)}{\max(|R|,|S|)}\right)\cdot LDIS \\ &= \frac{1}{\beta}\cdot LDIS \end{aligned} \tag{18}$$

其中, $\beta=\left(\frac{\max(|R|,|S|)}{\max(\|R\|,\|S\|)}\right)$,且满足:

1. $NDIS(R,S)\geqslant 0$;
2. 当且仅当 $R=S, NDIS(R,S)=0$;
3. $NDIS(R,S)=NDIS(S,R)$.

NDIS_I 和 NDIS_II 型语义距离测度的不同在于,句子中重复词起了不同的作用.NDIS_I 假设句子中重复词的出现拉大了句子间的语义距离,而 NDIS_II 假设句子中重复词的出现缩短了句子间的语义距离.由于 NDIS_I 和 NDIS_II 型语义距离测度更加精细地反映了句子重复成分的影响,因而可以预计,其性能应优于传统的 LDIS 距离测度.

令人惊讶的是,我们在实验中发现,NDIS 和 LDIS 在大量语料集上竟然表现出一致的特性.进一步的深入研究发现:当文本 Zipf 指数α<1 时,文本长度和词表长度呈线性关系,NDIS_I,NDIS_II 型语义距离测度和 LDIS 等价;当 Zipf 指数α>1 时,文本长度和词表长度呈非线性关系,基于词表长度的距离加权方法显著优于基于文本长度加权的方法.

命题 2. 当文本 Zipf 指数α<1 时,文本长度和词表长度呈线性关系,基于词表长度惩罚的语义距离测度

NDIS 线性等价与基于句长惩罚的语义距离测度 LDIS;当文本 Zipf 指数α>1 时,文本长度和词表长度呈非线性关系,基于句长惩罚的语义距离测度 LDIS 不线性等价与基于词表长度惩罚的语义距离测度 NDIS.

证明:

1) 首先,根据 Heaps 提出的 Heaps 定律[21],即在自然语言中,文本的词表长度(distinct words count,文本中所有不相同单词的数量)并不随着文档规模的增长呈线性增长,而是亚线性增长关系,即:

$$N(t)\sim t^{\lambda},\lambda<1 \tag{19}$$

其中,λ称为 Heap 指数,t 为文档长度,$N(t)$为文档中词表长度.

2) 其次,根据 Zipf 提出的 Zipf's 定律[22],即自然语言词频的分布符合一个确定的实验数学模型,具体表现为:在自然语言中,如果把单词出现的频率按倒序排列,单词的词频 Z 与排名 r 之间满足指数公式,且它们之间成反比的关系:

$$z(r)=z_{\max}r^{-\alpha} \tag{20}$$

其中,α称为 Zipf 指数,r 为单词 w 在词频表中的排名,$z(r)$为单词 w 在词频表中的词频.Zipf 定律和 Heaps 定律都是实验定律,而非理论定律,在自然界许多复杂系统中都存在 Zipf 定律和 Heaps 定律,只是不同的环境下存在不同的指数.

3) Zhou 等人[23]证明,对 Heap 指数与 Zipf 指数存在以下对价关系:

$$\lambda=\begin{cases}1/\alpha, & \alpha>1\\ 1, & \alpha<1\end{cases} \tag{21}$$

那么,通过公式(21),将公式(19)中 $N(t)$与 Heap 指数的关系转化为 $N(t)$与 Zipf 指数的关系:

$$N(t)=\begin{cases}(\alpha-1)^{1/\alpha}t^{1/\alpha}, & \alpha>1\\ (1-\alpha)t, & \alpha<1\end{cases} \tag{22}$$

当文本 Zipf 指数α<1 时,文本长度和词表长度呈线性关系,这样一来:

$$NDIS\_\mathrm{I}(R,S)=\frac{1}{1-\alpha}\cdot LDIS \tag{23}$$

$$NDIS\_\mathrm{II}(R,S)=(1-\alpha)\cdot LDIS \tag{24}$$

基于句长惩罚的语义距离测度 LDIS 线性等效于基于词表长度惩罚的语义距离测度 NDIS.当文本 Zipf 指数α>1 时,文本长度和词表长度呈非线性关系,基于句长惩罚的语义距离测度 LDIS 不线性等效于基于词表长度惩罚的语义距离测度 NDIS. □

综上所述,利用文本长度对文本距离进行惩罚加权已经成为了研究者们的共识,但我们认为:不区分对待句子中不同的成分,特别是忽视句子中重复单词的作用是不应该的.因此本节中,我们提出了基于词表长度加权惩罚的语义距离测度 NDIS,并证明了:当文本 Zipf 指数α<1,NIDS 线性等价与 LDIS,两者只相差一个比例系数,这样的差别可以通过类似归一化处理进行消除;而当文本 Zipf 指数α>1 时,NIDS 不等价与 LDIS,后续实验也表明了,基于词表长度的距离加权方法显著优于基于文本长度加权的方法.因此我们有理由认为:文本语义距离应和词表长度成比例关系,只不过现实文本大多 Zipf 指数α<1,从而表现为和文本长度成比例关系.

实际上,公式(19)~公式(24)还给出 Zipf 指数α计算 NDIS 语义距离的简便方法,即:根据 Zipf's Law 和 Heap's Law 的对价关系,通过估算样本集的 Zipf 指数α和文本的原始长度计算文本的词表长度.这样一来,可以省去为每一个文本统计词表的工作.

3 实验结果与分析

3.1 在线评论聚类分析系统框架

在线评论聚类系统框架由 4 部分组成(如图 2 所示):在线评论提取模块、数据预处理模块、文本距离计算模块、在线评论聚合模块.利用在线评论提取模块从 Internet 获取在线评论;通过数据预处理模块对在线评论进行变异短文本处理;文本距离计算模块用于计算经过规范化处理的文本之间的距离,得到文本距离矩阵;利用在

线评论聚合模块对文本距离矩阵进行聚合分析.通过聚合分析,最终将在线评论按照主题簇的方式呈现给用户.

具体模块功能如下:

(1) 在线评论提取模块.利用 push 和 pull 两种方式完成在线评论获取.同时也为脱机数据预留处理接口;

(2) 数据预处理模块.在线评论表达形式自由随意,需要通过统一字符编码对原始数据进行了预处理,消除不同编码的字符对语义距离计算结果的影响;

(3) 文本距离计算模块.在线评论聚合系统的关键问题是短文本间距离的度量,通过综合分析文本间的形式距离和单元语义距离;

(4) 在线评论聚合模块.通过文本距离计算模块得到文本距离矩阵,利用聚类算法对矩阵进行聚类分析,得到不同的主题簇.

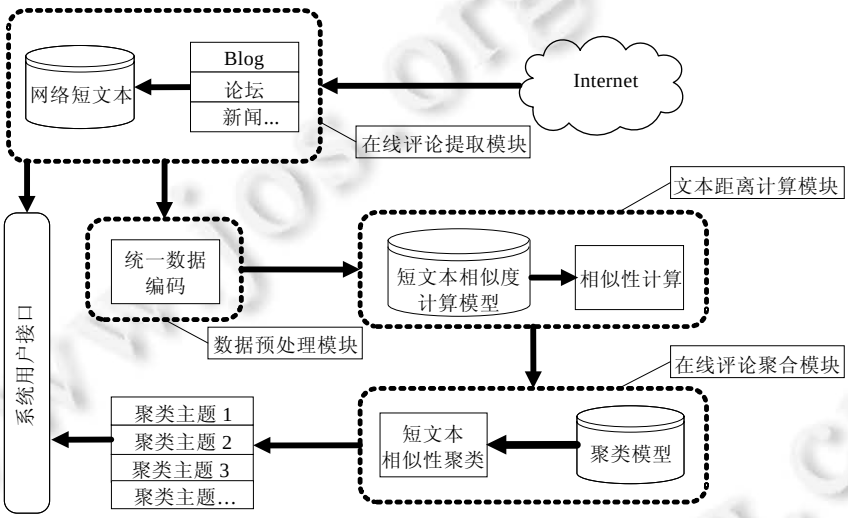

Fig.2 Diagram of online comment clustering framework

图 2 在线评论聚合系统框架图

3.2 实验设计

实验中采用的语料集合包括:

(1) 在线评论数据集 Comment:是通过在线评论提取模块得到的在线短文本评论数据,其中,一类为新浪微博(http://weibo.com)中关于网易邮箱的微博评论信息(时间跨度 2011 年 6 月 14 日~15 日);另一类为战网游戏论坛魔兽世界版块(http://www.battlenet.com.cn/wow/zh/forum/)中关于魔兽世界的评论(时间跨度 2011 年 6 月 14 日~15 日).实验中,分别随机选择两类各 220 条文本,共 440 条短文本作为实验语料.通过计算,得到其文本 Zipf 指数α=0.716<1.

(2) 短信数据集合 SMS[24]:精细标注的短信子集合(共 4 486 条精细标注短信数据,数据集中的短信根据不同的主题被分为 5 个类别).实验中,分别随机选择 10%的样本计 440 条短信作为实验样本,样本中 5 个类别的样本数呈均匀分布,保证了实验效果的有效性.通过计算,得到其文本 Zipf 指数α=0.861<1.

(3) 863 文本分类评测数据集:2004 年~2006 年 863 文本分类评测提供的标准分类数据集,该数据集按中图分类法体系进行划分.实验中,将文档标题截取出来作为评测数据.我们使用两组子集数据 AK,ADK.其中,AK 数据集为分别在 A,K 两类文本中各随机抽取 100 条文本共 200 条短文本组成 AK 数据集,Zipf 指数α=1.07>1;ADK 数据集为分别在 A,D,K 这 3 类文本中各随机抽取 100 条文本共 300 条文本组成 ADK 数据集,Zipf 指数α=1.05>1.其中,A 分类表示马克思主义、列宁主义、毛泽东思想、邓小平理论,D 分类表示政治、法律,K 分类表示历史、地理.

(4) TanCorp60-Raw 数据集(http://www.searchforum.org.cn/tansongbo/software.htm):共收集了财经、地域、人才、艺术等 12 个一级类别、60 个二级类别,共计 14 150 条文本.经过统计分析,语料中的二级类别人才创业、艺术古董两类样本的统计特性α=1.05>1,因此,将两类样本作为实验的测试样本.在两类中分别随机选择 200 条样本共 400 条样本作为测试集.

此外,短文本聚类模型采用 Affinity Propagation 算法[25]对文本进行聚类,AP 算法不需要用户预先提供目标类数,在聚类的过程中,通过计算类内密度与类间密度的差别进行自动类别确定.

聚类性能评价指标采用 E 熵[26],这是在聚类数据标签未知的情况下,评价聚类性能较好的方法[27].给定一个聚类簇 S_r 以及簇样本数 n_r,则此簇的熵表示为

$$E(S_r)=-\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r} \tag{25}$$

其中,$E(S_r)$为簇的熵,q 为测试数据的原始类别数,n_r^i 为簇 S_r 中第 i 类的样本数,则聚类结果的 E 熵定义为所有簇的熵值与簇样本数加权平均:

$$Entropy=\sum_{r=1}^{k}\frac{n_r}{n}E(S_r) \tag{26}$$

其中,$E(S_r)$为簇的熵,n 为测试数据的样本总数,n_r 为簇样本数,k 为聚类结果的簇数.一个好的聚类结果是每个簇中只包含某一类样本,此时熵值为 0.熵值越小,说明聚类性能越好.

3.3 实验结果分析

实验比较了本文提出的改进语义距离 ISD(improved semantic distance)和编辑距离 Edit(edit distance)采用不同句长惩罚策略的性能,共 10 种距离计算模型在 5 组测试集上的聚类性能.实验结果见表 2,其中,Edit,Edit/LDIS,Edit/NDIS_I 和 Edit/NDIS_II 分别表示简单编辑距离、基于句长惩罚的编辑距离、基于词表长度 NDIS_I 和词表长度 NDIS_II 惩罚的编辑距离.由于 ISD 综合考虑句子之间形式结构和单元语义上的差别程度,通过调节权值的大小来调节形式结构和单元语义在文本内容(语义)方面的权重.ISD_I 设置单元语义的信息贡献度要大于形式结构信息,因此,单元语义的操作权值大于形式结构的操作权值.结合第 1 节权值选择原则和实验中的经验值,其权值如下:λ=0.04,ω_1=0.27,ω_2=0.05,γ_1=0.54,γ_2=0.09,θ=0.01.ISD_II 是参数设置的一种极端情况,参数设置为 λ=0,ω_1=0,ω_2=0,γ_1=0,γ_2=0,θ=1,这种情况只考虑文本中的近义词,忽略了文本之间差别.ISD_I(II)/LDIS,ISD_I(II)/NDIS_I 和 ISD_I(II)/NDIS_II 分别表示基于句长惩罚的改进语义距离、基于词表长度 NDIS_I 和词表长度 NDIS_II 惩罚的改进语义距离.

Table 2 Clustering entropy comparison

表 2 实验结果聚类熵性能比较

Data	Edit	Edit/ LDIS	Edit/ NDIS_I	Edit/ NDIS_II	ISD_I/ LDIS	ISD_I/ NDIS_I	ISD_I/ NDIS_II	ISD_II/ LDIS	ISD_II/ NDIS_I	ISD_II/ NDIS_II
Comment α=0.71,<1 平均句长=24	.339713 ± .0234455	.322022 ± .0244534	/	/	.249995 ± .0218901	/	/	.258674 ± .0225912	/	/
SMS α=0.86,<1 平均句长=27.7	.565637 ± .0145590	.556349 ± .0123944	/	/	.441248 ± .0136696	/	/	.442597 ± .0124655	/	/
AK α=1.07,>1 平均句长=6.38	.810145 ± .032444	.756264 ± .029680	.71417 ± .03767	.727619 ± .036779	.619925 ± .0423806	.620711 ± .0424544	.604157 ± .0445348	.627691 ± .0431302	.625863 ± .0457003	.611244 ± .0423936
ADK α=1.05,>1 平均句长=7.32	.707030 ± .0263162	.650199 ± .026593	.61964 ± .02956	.626530 ± .030322	.569507 ± .0283713	.555950 ± .0284528	.556869 ± .0294188	.565615 ± .0353589	.556180 ± .0293197	.555222 ± .0296977
TanCorp60-Raw α=1.05,>1 平均句长=45.6	.885703 ± .018970	.715268 ± .024734	.756296 ± .025581	.657141 ± .033997	.411847 ± .032498	.459219 ± .027552	.407863 ± .032698	.406041 ± .027629	.452665 ± .027511	.403785 ± .030853

从实验结果可知:

首先,对于本文提出的改进语义距离,相比编辑距离,表现出了稳定一致的优势,且距离惩罚能显著提升模型性能.ISD_I 比 ISD_II 算法性能更优,主要是因为 ISD_I 模型同时考虑了词义以及结构信息,而 ISD_II 模型偏重考虑语义距离.

其次,对于本文提出的距离惩罚机制来说,基于词表长度进行距离惩罚优于基于文本长度惩罚.对于α<1 的文本,文本长度和词表长度呈线性关系,两者只相差一个比例系数.通过类似归一化处理以后,NDIS_I,NDIS_II 型语义距离模型和 LDIS 模型等价,可以直接利用文本的原始长度代替词表长度对文本距离进行惩罚.因此,此时基于 LDIS,NDIS_I 和 NDIS_II 的惩罚性能一致,在表中只填写了 LDIS 性能,其他两列用斜线划去不填.当α>1 时,文本长度和词表长度呈非线性关系,NDIS_II>NDIS_I>LDIS,即,利用词表长度对距离进行处理优于利用文本长度处理.

同时,我们还比较了算法的时间复杂度,如图 3 所示,比较了 10 种距离在 5 个数据集上计算时间.由于 Edit/NDIS_I 和 Edit/NDIS_II,ISD_I/NDIS_I 和 ISD_I/NDIS_II,ISD_II/NDIS_I 和 ISD_II/NDIS_II 只是归一化时所除以的分母不同,而计算时间具有一致性,为了节省画图空间,我们仅绘制了 Edit,Edit/LDIS,Edit/NDIS_I,ISD_I/LDIS,ISD_I/NDIS_I,ISD_II/LDIS,ISD_II/NDIS_I 的性能.从图中可以看出,改进算法在时间复杂度高于经典的编辑距离.这一点也与前面的时间复杂度分析一致.因此,在数据规模特别巨大、实时性要求高的应用场景下,应选择使用经典编辑距离及其改进,如 Edit,Edit/NDIS_I 和 Edit/NDIS_II 等,以优先满足实时性要求.但在实时性要求较低的应用场景,例如网络评论分析等,可以脱机处理大量数据的应用,也可以考虑使用 ISD_I/NDIS_I 和 ISD_I/NDIS_II,ISD_II/NDIS_I 和 ISD_II/NDIS_II 等改进型语义计算方法.考虑到改进算法带来的性能的提升,在这样的应用场景下,对计算时间的略微牺牲在一定程度上是可接受的.

再次,为了深入分析所提算法的性能,进一步将算法和 8 种短文本距离测度进行了比较:1) 改进型编辑距离,包括 Damerau-Levenshtein 距离[28]、加权编辑距离[29];2) 基于 Bag-of-word 文本表示模型的距离改进[30],即用向量空间模型表示短文本后,尝试使用不用的距离测度衡量短文本间相似度,包括 Canberra 距离、Minkowski 距离(p=0.5)、Chi square 距离、Manhattan 距离、cosine 距离、欧式距离.实验结果如图 4 所示,从中可以看出:算法性能在各语料集上保持一致性,同时,改进算法性能均优于所对比算法.

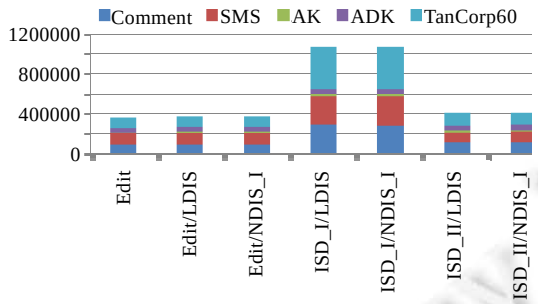

Fig.3 Comparison of CPU computation time (ms)

图 3 算法 CPU 计算时间比较(毫秒)

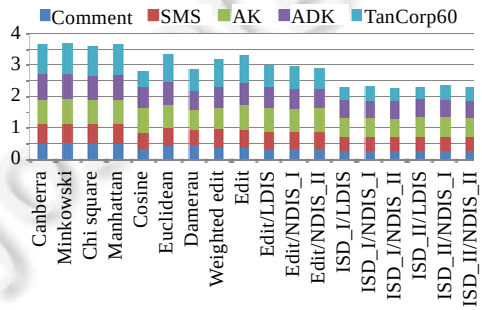

Fig.4 Clustering entropy comparison of performance

图 4 算法性能聚类熵比较

最后,表 3 从不同语料集的角度分析了算法性能.虽然实验用语料集来源不同,但有可能包含相同语义性质的内容.如果能从语义内容的角度观察算法性能,将能够更加明确算法的适用范围,从而对实际应用进行指导.因此,我们尝试从语义角度重新组织实验分析.表 3 中,Beitzel[31]在对约 13%美国本土网络在线查询日志的逐小时分析后发现,网络上最受用户欢迎的查询主题内容比例分别是购物(shopping)13%、娱乐(entertainment)13%、色情(porn)10%、计算技术(computing)9%、研究学习(research & learn)9%、健康(health)5%、旅行 5%、游戏(games)5%、家庭(home)5%、体育(sports)3%、个人经济计划(personal finance)3%、假日(holidays)1%以及其他.经过分析实验所用语料特性,科研看出实验语料集合覆盖了大部分网络用户的热门查询主题,从而一定程度上

说明本文的方法有可能适用于网络真实应用场景.

Table 3 Semantic analysis of corpus

表 3 语料集语义特性分析

语义主题	百分比(%)	语料集
Shopping	13	Comment
Entertainment	13	Comment, TanCorp60-Raw
Porn	10	–
Computing	9	短信集 SMS
Research & learn	9	863 语料
Health	5	短信集 SMS
Travel	5	短信集 SMS
Games	5	Comment
Home	5	短信集 SMS
Sports	3	短信集 SMS
Personal finance	3	短信集 SMS, TanCorp60-Raw
Holidays	1	短信集 SMS
US sites	3	–
Other	16	–

References:

[1] Yang F, Peng QK, Xu T. Sentiment classification for online comments based on random network theory. Acta Automatica Sinica, 2010,36(6):837−844 (in Chinese with English abstract). [doi: 10.3724/SP.J.1004.2010.00837]

[2] Balog K, Mishne G, de Rijke M. Why are they excited? Identifying and explaining spikes in blog mood levels. In: Proc. of the ECACL 2006. 2006. 207−210.

[3] Turney PD, Littman ML. Measuring praise and criticism: Inference of semantic orientation from association. ACM Trans. on Information Systems, 2003,21(4):315−346.

[4] Kyosuke N, Takahide H, Ko F. Improving tweet stream classification by detecting changes in word probability. In: Proc. of the ACM SIGIR 2012. 2012. 971−980. [doi: 10.1145/2348283.2348412]

[5] Fan XH, Wang P. Chinese short text classification in two steps. Journal of Dalian Maritime University, 2008,34(3):121−124 (in Chinese with English abstract).

[6] Yang Z, Lai YX, Duan LJ, Li YJ. Short text sentiment classification based on context reconstruction. Acta Automatica Sinica, 2012, 38(1):55−67 (in Chinese with English abstract). [doi: 10.3724/SP.J.1004.2012.00055]

[7] Sriram B, Fuhry D, Demir E, Ferhatosmanoglu H, Demirbas M. Short text classification in twitter to improve information filtering. In: Proc. of the ACM SIGIR 2010. 2010. 841−842. [doi: 10.1145/1835449.1835643]

[8] Mihalcea R, Courtney C, Strapparava C. Corpus-Based and knowledge-based measures of text semantic similarity. In: Proc. of the AAAI 2006. 2006. 775−780.

[9] Pu Q, Yang GW. Short-Text classification based on ICA and LSA. In: Proc. of the ISNN 2006. LNCS 3972, Heidelberg: Springer-Verlag, 2006. 265−270. [doi: 10.1007/11760023_39]

[10] Yang Z, Fan KF, Lei JJ, Guo J. Text manifold based on semantic analysis. Acta Electronica Sinica, 2009,37(3):557−561 (in Chinese with English abstract). [doi: 10.3321/j.issn:0372-2112.2009.03.024]

[11] Liu YP, Li S, Zhao TJ. System combination based on wsd using WordNet. Acta Automatica Sinica, 2010,36(11):1575−1580 (in Chinese with English abstract). [doi: 10.3724/SP.J.1004.2010.01575]

[12] Li B, Liu T, Qin B, Li S. Chinese sentence similarity computing based on semantic dependency relationship analysis. Application Research of Computers, 2003,20(12):15−17 (in Chinese with English abstract). [doi: 10.3969/j.issn.1001-3695.2003.12.005]

[13] Peng J, Yang DQ, Tang SW, Fu Y, Jiang HK. A novel text clustering algorithm based on inner product space model of semantic. Chinese Journal of Computers, 2007,30(8):1354−1363 (in Chinese with English abstract). [doi: 10.3321/j.issn:0254-4164.2007.08.017]

[14] Li SJ. Research of relevancy between sentences based on semantic computation. Computer Engineering and Applications, 2002, 38(7):75−76 (in Chinese with English abstract). [doi: 10.3321/j.issn:1002-8331.2002.07.025]

[15] Ristud ES, Yianilos PN. Learning string-edit distance. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1998,20(5): 522−532. [doi: 10.1109/34.682181]

[16] Sui ZF, Yu SW. The skeletal-dependency-tree-based computational model for the sentence similarity. In: Proc. of the Conf. of Chinese Information Processing. 1998. 458−465 (in Chinese with English abstract).

[17] Chatterjee N. A statistical approach for similarity measurement between sentences for EBMT. In: Proc. of Symp. on Translation Support Systems. 2001.

[18] Che WX, Li ZH, Liu T. LTP: A Chinese language technology platform. In: Proc. of the Coling 2010. 2010. 13−16.

[19] Mei JJ, Zhu YM, Gao YQ, Yin HX. TongYiCi CiLin. 2nd ed., Shanghai: Shanghai Lexicographic Publishing House, 1996 (in Chinese).

[20] Li YJ, Liu B. A normalized levenshtein distance metric. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2007,29(6): 1091−1095. [doi: 10.1109/TPAMI.2007.1078]
[21] Heaps HS. Information Retrieval: Computational and Theoretical Aspects. Academic Press, 1978.
[22] Zipf GK. Human Behavior and the Principle of Least Effort. Oxford: Addison-Wesley Press, 1949.
[23] Lü LY, Zhang ZK, Zhou T. Zipf's law leads to heaps' law: Analyzing their relation in finite-size systems, Plot Ones, 2010,5(12):1−11.
[24] Ma X, Xu WR, Guo J, Hu RL. SMS-2008: An annotated Chinese short messages corpus. Journal of Chinese Information Processing, 2009,23(4):22−26 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2009.04.004]
[25] Frey BJ, Dueck D. Clustering by passing messages between data points. Science, 2007,315(5814):972−976. [doi: 10.1126/science.1136800]
[26] Zhao Y, Karypis G. Empirical and theoretical comparisons of selected criterion functions for document clustering. Machine Learning, 2004,55(3):311−331. [doi: 10.1023/B:MACH.0000027785.44527.d6]
[27] Papapetrou O, Siberski W, Fuhr N. Decentralized probabilistic text clustering. IEEE Trans. on Knowledge and Data Engineering, 2012,24(10):1848−1861. [doi: 10.1109/TKDE.2011.120]
[28] Damerau FJ. A technique for computer detection and correction of spelling errors. Communications of the ACM, 1964,7(3): 171−176. [doi: 10.1145/363958.363994]
[29] Schauerte B, Fink GA. Focusing computational visual attention in multi-modal human-robot interaction, In: Proc. of the Int'l Conf. on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction. 2010. [doi: 10.1145/1891903.1891912]
[30] Perlibakas V. Distance measures for PCA-based face recognition. Pattern Recognition Letters, 2004,25(6):711−724. [doi: 10.1016/j.patrec.2004.01.011]
[31] Beitzel SM, Jensen EC, Chowdhury A, Grossman D, Frieder O. Hourly analysis of a very large topically categorized Web query log. In: Proc. of the ACM SIGIR 2004. 2004. 321−328. [doi: 10.1145/1008992.1009048]

附中文参考文献:

[1] 杨锋,彭勤科,徐涛.基于随机网络的在线评论情绪倾向性分类.自动化学报,2010,36(6):837−844. [doi: 10.3724/SP.J.1004.2010.00837]
[5] 樊兴华,王鹏.基于两步策略的中文短文本分类研究.大连海事大学学报,2008,34(3):121−124.
[6] 杨震,赖英旭,段立娟,李玉鑑.基于上下文重构的短文本情感极性判别研究.自动化学报,2012,38(1):55−67. [doi: 10.3724/SP.J.1004.2012.00055]
[10] 杨震,范科峰,雷建军,郭军.基于语义的文本流形研究.电子学报,2009,37(3):557−561. [doi: 10.3321/j.issn:0372-2112.2009.03.024]
[11] 刘宇鹏,李生,赵铁军.基于 WordNet 词义消歧的系统融合.自动化学报,2010,36(11):1575−1580. [doi: 10.3724/SP.J.1004.2010.01575]
[12] 李彬,刘挺,秦兵,李生.基于语义依存的汉语句子相似度计算.计算机应用研究,2003,20(12):15−17. [doi: 10.3969/j.issn.1001-3695.2003.12.005]
[13] 彭京,杨冬青,唐世渭,付艳,蒋汉奎.一种基于语义内积空间模型的文本聚类算法.计算机学报,2007,30(8):1354−1363. [doi: 10.3321/j.issn:0254-4164.2007.08.017]
[14] 李素建.基于语义计算的语句相关度研究.计算机工程与应用.2002,38(7):75−76. [doi: 10.3321/j.issn:1002-8331.2002.07.025]
[16] 穗志方,俞士汶.基于骨架依存树的语句相似度计算模型.中文信息处理国际会议论文集.1998.458−465.
[19] 梅家驹,竺一鸣,高蕴琦,殷鸿翔.同义词词林.第 2 版,上海:上海辞书出版社,1996.
[24] 马旭,徐蔚然,郭军,胡日勒.SMS-2008 标注中文短信息库.中文信息学报,2009,23(4):22−26. [doi: 10.3969/j.issn.1003-0077.2009.04.004]

杨震(1979－),男,贵州六盘水人,博士,副教授,主要研究领域为数据挖掘,机器学习,可信计算,内容安全.
E-mail: yangzhen@bjut.edu.cn

赖英旭(1973－),女,副教授,主要研究领域为网络安全,可信计算.
E-mail: laiyingxu@bjut.edu.cn

王来涛(1988－),男,硕士,主要研究领域为数据挖掘,机器学习.
E-mail: wltao123@163.com