doi:10.3969/j.issn.1001-893x.2017.01.013
引用格式:黄贤英,李沁东,刘英涛.结合词性的短文本相似度算法及其在文本分类中的应用[J].电讯技术,2017,57(1):78-82.[HUANG Xianying, LI Qindong, LIU Yingtao. A grammatical category-combined short-text similarity algorithm and its application in text categorization [J]. Telecommunication Engineering, 2017, 57(1):78-82.]

结合词性的短文本相似度算法及其在文本分类中的应用*

黄贤英,李沁东**,刘英涛

(重庆理工大学 计算机科学与工程学院,重庆 400054)

摘　要:针对基于语义的短文本相似度计算方法在短文本分类中准确率较低这一问题,提出了结合词性的短文本相似度算法(GCSSA)。该方法在基于 hownet("知网")语义的短文本相似度计算方法的基础上,结合类别特征词并添加关键词词性分析,对类别特征词和其他关键词的词性信息给定不同关键词以不同的权值系数,以此区别各种贡献度词项在短文本相似度计算中的重要程度。实验表明,该算法进行文本相似度计算后应用于短文本分类中较基于 hownet 的短文本分类算法在准确率宏平均和微平均上提升 4% 左右,有效提高了短文本分类的准确性。

关键词:短文本分类;短文本相似度;词性;hownet 语义;分类准确率

中图分类号:TP391.1　　**文献标志码**:A　　**文章编号**:1001-893X(2017)01-0078-05

A Grammatical Category-combined Short-text Similarity Algorithm and Its Application in Text Categorization

HUANG Xianying, LI Qindong, LIU Yingtao

(College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China)

Abstract: To address the problem that the categorization accuracy of hownet-based short-text similarity calculation method in short-text is low, a grammatical category-combined short-text similarity algorithm (GCSSA) is proposed. Based on short-text hownet semantic similarity calculation method and combing with categorized features words, this method adds keywords grammatical category analysis, targets at categorized features words and the grammatical category information of keywords, gives different weights for different keywords, so as to differentiate the importance of various items' contribution in the text similarity calculation of short-texts. Experiments show that compared with hownet-based short-text categorization algorithm, the proposed method improves the macro-average and micro-average accuracy by 4% in short-text categorization, and improves the short-text categorization accuracy effectively.

Key words: short text categorization; short-text similarity; grammatical category; hownet semantic; categorization accuracy

1　引　言

随着微博、短信等文本交流方式的兴起,越来越多的短文本充斥着现代网络和人们的日常生活,短文本的处理研究越来越成为文本处理的主流方向。

*　**收稿日期**:2016-05-05;**修回日期**:2016-07-12　　Received date: 2016-05-05; Revised date: 2016-07-12
基金项目:国家自然科学基金资助项目(11547148);重庆市教委科技计划项目(16SKGH133);重庆市社会科学规划博士项目(2015BS059)
**　**通信作者**:18883844511@163.com　　**Corresponding author**: 18883844511@163.com

由于短文本具有篇幅短小、组成文本的特征词较少、关键词之间的关联性较弱等不同于长文本的特点,使得传统的长文本相似度计算方式很难直接应用于短文本相似度计算。

文本关键词是指文本经停用词过滤、噪声数据去除等处理后,用于表达文本主题内容的词。近年来,基于语义的关键词自动提取方法得到了广泛应用,这类方法主要运用语义词典(如 WordNet、"同义词词林"、"知网"等)获取词汇间的语义知识进而提取文本关键词。

针对提取得到的关键词进行文本相似度计算,许多学者进行了研究。文献[1]使用 WordNet 和 Brown 文集计算文本中的概念相似性,基于此提出了一种综合考虑概念、句法等信息来计算短文本语义相似性的新方法,但此方法未考虑语境及上下文等信息,出现场景可能不准确的情况。文献[2]针对传统的句子相似度计算方法只关注句子某个特征,导致召回率和准确率不均衡的问题,提出一种基于多特征的句子相似度计算方法。文献[3]提出了一种基于树结构本体的句子相似度计算方法,利用本体概念与句子中关键词之间建立的语义索引,构建句子与本体间的语义联系,形成描述句子的语义向量后计算句子间的语义相似度。文献[4]按照不同特征粒度抽取训练集中各类别的高频词和核心词作为领域关键词,利用概率主题模型提取文本的主题概率分布后进行文本分类,但短文本有许多"知网"未收录特征词而无法计算相似度进而影响分类效果。文献[5]基于知识库 WordNet 以及词法库 The Brown Corpus 的相似性算法,提出了一种新的短文本语义距离计算算法。

短文本相似度计算常用于短文本分类中,对于短文本分类国内众多学者也进行了深入研究。文献[6]用支持向量描述训练得到各类样本的最小超球体,使间隔最大化,提出了一种最大分类间隔 SVDD 的多类文本分类算法。文献[7]借助维基百科抽取相关概念,对短文本的特征向量进行扩展,提出了基于维基百科的中文短文本分类方法。文献[8]挖掘具有相同类别倾向的频繁词集作为文本特征扩展知识库,提出了一种基于频繁词集的特征扩展方法。

基于语义词典的短文本相似度计算方法增加了语义信息,提高了相似度计算的准确性,但由于太过依赖语义词典而存在局限性。在组成短文本的词中对文本起类别判别作用的词只占很小一部分,这些词称为类别特征词。本文在此基础上提出一种结合词性的短文本相似度算法,采用基于卡方统计对训练空间每个类别提取类别特征词,针对测试集中关键词是类别特征词或非类别特征词的不同词性给定关键词不同权重系数后进行文本相似度计算,提高了短文本相似度计算的准确性,实现了短文本的有效分类。

2　基于 hownet 的短文本相似度算法

传统的基于 hownet 的短文本相似度算法是在基于 hownet 词语语义相似度计算的基础上,根据短文本中关键词词项来计算短文本的相似度值。通过计算两个短文本关键词对应词项相似度值的平均值作为两个短文本的相似度值,是依赖 hownet 词典来计算两个短文本中关键词的相似度值得到短文本的相似程度。

假设两个短文本 d_1 和 d_2 预处理后的词项向量为

$$d_1=(w_{11},w_{12},\cdots,w_{1m}),$$
$$d_2=(w_{21},w_{22},\cdots,w_{2n}),$$

则 d_1 和 d_2 的语义相似度计算公式如下:

$$sim(d_1,d_2)=\frac{\sum_{k=1i\in[1,m]}^{mi}\max_{j\in[1,n]}(sim(w_{1i},w_{2j}))+delta\times(ma-mi)}{ma}。\tag{1}$$

式中:m 表示短文本 d_1 中关键词个数;n 表示短文本 d_2 中关键词个数;$mi=\min(m,n)$;$ma=\max(m,n)$;$\sum_{k=1i\in[1,m]}^{mi}\max_{j\in[1,n]}(sim(w_{1i},w_{2j}))$ 表示对两个文本中的关键词依次选取相似度最大的关键词对的相似度之和,直至关键词少的文本内的关键词全部取完;$delta$ 为一个调节参数,定义为一个非空关键词项与空关键词项的相似度,一般取值为 0.2;$sim(w_{1i},w_{2j})$ 为 w_{1i} 与 w_{2j} 基于 hownet 词典的词语语义相似度值。文本的词语语义相似度计算方法选取李峰改进后的中文词语语义相似度算法[9],公式如下:

$$sim(w_1,w_2)=\frac{\alpha\times\min(depth_{w_1},depth_{w_2})}{\alpha\times\min(depth_{w_1},depth_{w_2})+distance(w_1,w_2)}。\tag{2}$$

式中:$\min(depth_{w_1},depth_{w_2})$ 表示 w_1 与 w_2 在"知网"中的最小深度;$distance(w_1,w_2)$ 表示 w_1 与 w_2 在"知网"中的路径长度;α 为调节参数,表示词语相似度为 0.5 时路径长度。

3　结合词性的短文本相似度算法

通过公式(1)计算所有的关键词处于不同类别

或不同词性的权值都是一样的,但在实际的分类过程中,不同的关键词对短文本分类的贡献度是不同的。类别特征词由于其更能代表包含该词的短文本的类别,对短文本分类类别的贡献更大,应该具有更高的权值。同时,像名词、动词等更能表征句子主题,对分类类别具有较高的贡献度,而形容词、副词次之。其他词对于短文本分类类别的表征很弱,对于判定分类类别的贡献也很小,如果给这些表征较弱的词很大的权值,由于基于 hownet 词语语义计算时冗余特征的添加,反而会使得短文本分类的准确性降低。基于此,我们提出一种结合词性的短文本相似度算法(Grammatical Category-combined Short-text Similarity Algorithm,GCSSA)来提高短文本相似度计算的准确性。文本相似度计算公式如下:

$$sim(d_1,d_2)=\frac{\sum_{i=1}^{n}\rho\times\max_{j}(sim(w_{1i},w_{2j}))}{n},\quad(3)$$

$$\rho=\begin{cases}\alpha,\text{当 } w_{1i}\text{为名词或者动词}\\ \beta,\text{当 } w_{1i}\text{为形容词或者副词}\\ \gamma,\text{当 } w_{1i}\text{为其他词性词项}\end{cases}。$$

式中:w_{1i}表示短文本 d_1 的第 i 个关键词;w_{2j}表示短文本 d_2 的第 j 个关键词;ρ 为权重系数,且满足条件 $1>\alpha>\beta>\gamma>0$。

结合词性的短文本相似度算法的算法描述如下:

输入:短文本 $d_1=(w_{11},w_{12},\cdots,w_{1n})$、$d_2=(w_{21},w_{22},\cdots,w_{2m})$,类别特征词集合 $s=\{t_1,t_2,\cdots,t_k\}$,n 表示短文本 d_1 所包含的关键词数量,m 表示短文本 d_2 所包含的关键词数量,k 表示类别特征词集合 s 所包含的类别特征词数量。

输出:短文本 d_1 与短文本 d_2 的相似度值 $sim(d_1,d_2)$。

步骤 1 提取短文本 d_1、d_2 中的关键词 w_{1i}、w_{2j}。

步骤 2 判断关键词 w_{1i}是否是类别特征词项集合 s 中的元素,若是,则根据类别特征词对权重系数 ρ 取值并转入步骤 4;若不是,则转入步骤 3。

步骤 3 根据关键词 w_{1i}的词性对权重系数 ρ 进行取值。当关键词 w_{1i}是名词或者动词,权重系数 ρ 取值为 α;当关键词 w_{1i}是形容词或者副词,权重系数 ρ 取值为 β;当关键词 w_{1i}是其他词性词项,权重系数 ρ 取值为 γ。

步骤 4 根据式(3)计算短文本 d_2 中关键词 w_{2i}与短文本 d_1 关键词 w_{1i}的词语语义相似度,记录其中最大相似度值,转入步骤 1 直至计算完短文本 d_1 的所有关键词。

步骤 5 将步骤 4 记录的所有相似度求平均值,即为短文本 d_1 与短文本 d_2 的相似度值 $sim(d_1,d_2)$。

4 GCSSA 算法在短文本分类中的应用

4.1 研究数据

实验数据选取从数据堂下载的搜狗新闻语料库,以其中的新闻标题作为数据源,从中提取出可确定类别的数据,剔除类别中数据条数小于 10 条的类别及数据,对于剩下的数据,对每条数据只提取其中的标题及所属类别。数据的提取主要是基于卡方统计(Chi-square,CHI),即通过计算关键词与类别之间的 CHI 值得到两者之间的相关度,相关度越大,关键字就越能表征类别。

提取后的数据包含 auto、culture 等 15 个类别。对提取得到的数据按照 9∶1 的比例提取训练集和测试集,并进行标题去重处理。提取后的训练空间分布情况如表 1 所示。

表 1 训练空间数据分布

Tab. 1 The data distribution of training space

类别	数据条数	类别	数据条数
Auto	22 286	Lady	19 513
Culture	2 152	Local	2 099
Edu	25 444	Military	9 701
Ent	68 500	Olympics	55 376
Finance	153 368	Society	1 734
Health	2 185	Sports	236 011
House	6 134	World	2 026
IT	15 171		

4.2 GCSSA 算法的权重系数确定

GCSSA 算法准确率受权重系数的影响较大。为得到最优的短文本相似度算法,需要确定 GCSSA 算法中的最优权重系数。确定权重系数时,选取测试文本数量为 100,当设置不同权重系数时,GCSSA 算法在 KNN 短文本分类算法中的准确率不同,选取准确率最大时的系数值作为 GCSSA 算法的权重系数。

文献[10]指出,中文短文本主要依靠名词、动词、形容词、副词 4 种词性进行表达。文中给出了各种词性的关键词对短文本信息表征能力的强弱关系,其中动词和名词对短文本信息的表征能力最强,因此实验中将名词或者动词的权重系数 α 从 0.5 开

始取值,各权值系数的增大步长均为 0.1。按照 $1>\alpha>\beta>\gamma>0$ 的规则,采用三重循环依次增大 α、β、γ 的值。部分实验结果如表 2 所示。

表 2　准确率与权值系数关系

Tab. 2 The relationship between accuracy and weight coefficient

准确率	α	β	γ	准确率	α	β	γ
0.69	0.5	0.2	0.1	0.75	0.6	0.6	0.6
0.71	0.5	0.3	0.2	0.71	0.7	0.2	0.2
0.73	0.5	0.4	0.3	0.73	0.7	0.3	0.1
0.75	0.5	0.5	0.4	0.73	0.7	0.5	0.4
0.74	0.5	0.5	0.5	0.73	0.7	0.7	0.4
0.69	0.6	0.1	0.1	0.72	0.8	0.1	0.1
0.73	0.6	0.4	0.2	0.71	0.8	0.3	0.1
0.75	0.6	0.5	0.2	0.73	0.8	0.5	0.1
0.73	0.6	0.6	0.3	0.72	0.8	0.6	0.4

实验结果表示在 α、β、γ 分别取(0.5、0.5、0.4)、(0.6、0.5、0.2)、(0.6、0.6、0.6)时,算法的准确率取得最大值 0.75。根据取值规则 $1>\alpha>\beta>\gamma>0$,只有当 α、β、γ 取 0.6、0.5、0.2 时满足要求。根据 4.1 中经卡方提取后得到类别特征词对于短文本表示具有极其重要的作用,因此当关键词为类别特征词时,权重系数 ρ 取值为 1.0。

4.3　类别特征词项数量的选定

类别特征词的数量对算法的准确性有很大影响,确定最优类别特征词数量对于改进的 KNN 短文本分类算法的性能有着重要的意义。最优类别特征词数量的选定要在保证改进的 KNN 短文本分类算法效率不受影响或影响在一定范围内时,选取使得改进的 KNN 短文本分类算法取得最高准确率时的类别特征词项数量。类别特征词项数量初始选取为 100,依次增大类别特征词数量来进行实验,查看不同类别特征词数量对改进算法的效率影响情况。

表 3 是改进的 KNN 短文本分类算法选取不同类别特征词数量时的分类准确率情况。类别特征词数量小于 300 时,改进的 KNN 短文本分类算法由于类别特征词数量较小,训练空间样本拆分粗糙,重构后的训练集与原测试文本的相关度较低造成准确率相对较低。类别特征词数量在 300 ~ 600 之间时,算法的分类准确率相对较高,当类别特征词项数量在 600 时改进的 KNN 短文本分类算法的分类准确率最高,所以本文选取的类别特征词数量为 600。

表 3　不同特征词数量下算法准确率

Tab. 3 Algorithm accuracy of different number of feature word

类别特征词项个数	准确率/%	类别特征词项个数	准确率/%
100	54	700	62
200	58	800	59
300	63	900	60
400	64	1 000	58
500	63	1 500	59
600	65	2 000	62

4.4　应用分类结果比较与分析

确定 GCSSA 算法的最佳权重系数和类别特征词数量后,将改进的 KNN 短文本分类算法与基于语义的 KNN 短文本分类算法在准确率上进行对比实验。表 4 和表 5 分别表示不同数量测试文本时基于语义的 KNN 短文本分类算法与改进的 KNN 短文本分类算法准确率宏平均和微平均的对比情况。从表 4 可以看出,改进的 KNN 短文本分类算法的准确率宏平均值都要高于基于类别特征的 KNN 短文本分类算法。从表 5 可以看出,改进的 KNN 短文本分类算法的准确率微平均值都要高于基于语义的 KNN 短文本分类算法。总体来讲,GCSSA 算法使得 KNN 短文本分类算法具有更高的准确率。

表 4　准确率宏平均比较

Tab. 4 The comparison of accuracy of macro-average

测试数据量	准确率宏平均/%	
	基于语义的 KNN 短文本分类算法	改进的 KNN 短文本分类算法
100	66.00	67.00
300	63.67	68.33
500	62.80	70.00
800	69.38	71.25
1 100	70.27	72.18
1 500	67.47	72.13
2 000	67.10	71.00
4 000	66.68	70.07

表 5　准确率微平均比较

Tab. 5 The comparison of accuracy of micro-average

测试数据量	准确率微平均/%	
	基于语义的 KNN 短文本分类算法	改进的 KNN 短文本分类算法
100	32.66	45.75
300	31.56	35.67
500	33.85	47.98
800	42.07	44.48
1 100	38.11	40.71
1 500	38.24	41.56
2 000	38.89	42.45
4 000	37.92	41.39

5 结束语

本文在基于知网的 KNN 分类算法的基础上,结合类别特征词及关键词词性信息,给定不同关键词以不同的权重系数来表征短文本中不同关键词对文本的重要程度,实现了在 KNN 短文本分类算法中更为准确的短文本相似度计算。与基于语义的 KNN 短文本分类算法相比,该算法具有更好的准确性并能广泛应用于短文本分类中。

但本文算法在短文本类别特征词的提取上还有待提高,在提取过程中加入语义信息,保证特征词提取的精度将对短文本相似度算法的准确性有进一步提高。

参考文献:

[1] 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.
ZHAI Yandong, WANG Kangping, ZHANG Dongna, et al. An algorithm for semantic similarity of short text based on Wordnet[J]. Acta Electronica Sinica, 2012, 40(3):617-620. (in Chinese)

[2] 赵臻,吴宁,宋盼盼. 基于多特征融合的句子语义相似度计算[J]. 计算机工程,2012,38(1):171-173.
ZHAO Zhen, WU Ning, SONG Panpan. Sentence semantic similarity calculation based on multi-feature fusion[J]. Computer Engineering, 2012, 38(1):171-173. (in Chinese)

[3] 刘宏哲. 一种基于本体的句子相似度计算方法[J]. 计算机科学,2013,40(1):251-256.
LIU Hongzhe. Ontology based sentence similarity measurement. computer science[J]. Computer Science, 2013, 40(1):251-256. (in Chinese)

[4] 李湘东,曹环,丁丛,等. 利用《知网》和领域关键词扩展方法的短文本分类研究[J]. 现代图书情报技术,2015(2):31-38.
LI Xiangdong, CAO Huan, DING Cong, et al. Short-text classification based on hownet and domain keyword set extension[J]. New Technology of Library and Information Service, 2015(2):31-38. (in Chinese)

[5] 张东娜. 基于 WordNet 的短文本语义相似性计算研究[D]. 长春:吉林大学,2010.
ZHANG Dongna. The research of semantic similarity between short text based on WordNet[D]. Changchun: Jilin University, 2010. (in Chinese)

[6] 罗琦. 一种最大分类间隔 SVDD 的多类文本分类算法[J]. 电讯技术,2014,54(4):496-499.
LUO Qi. A multi-class text categorization algorithm based on maximal classification margin SVDD[J]. Telecommunication Engineering, 2014, 54(4):496-499. (in Chinese)

[7] 范云杰,刘怀亮. 基于维基百科的中文短文本分类研究[J]. 现代图书情报技术,2012(3):47-52.
FAN Yunjie, LIU Huailiang. Research on chinese short text classification based on wikipedia[J]. New Technology of Library and Information Service, 2012(3):47-52. (in Chinese)

[8] 袁满,欧阳元新,熊璋,等. 一种基于频繁词集的短文本特征扩展方法[J]. 东南大学学报(自然科学版),2014(2):256-260.
YUAN Man, OUYANG Yuanxin, XIONG Zhang, et al. Short text feature extension method based on frequent term sets[J]. Journal of Southeast University (Natural Sciences), 2014(2):256-260. (in Chinese)

[9] 李峰,李芳. 中文词语语义相似度计算——基于《知网》2000[J]. 中文信息学报,2007(3):99-105.
LI Feng, LI Fang. A new approach measuring semantic similarity in hownet 2000[J]. Journal of Chinese Information Processing, 2007(3):99-105. (in Chinese)

[10] 黄贤英,张金鹏,刘英涛,等. 基于词项语义映射的短文本相似度算法[J]. 计算机工程与设计,2015,36(6):1514-1518,1534.
HUANG Xianying, ZHANG Jinpeng, LIU Yingtao, et al. Short text similarity algorithm based on term mapping with semantic[J]. Computer Engineering and Design, 2015, 36(6):1514-1518, 1534. (in Chinese)

作者简介:

黄贤英(1967—),女,重庆人,教授、硕士生导师,主要研究方向为信息检索、移动计算;

李沁东(1992—),男,四川巴中人,硕士研究生,主要研究方向为信息检索;

Email:18883844511@163.com

刘英涛(1988—),男,山东烟台人,硕士研究生,主要研究方向为自然语言处理。