DOI:10.11992/tis.201608010

基于分类词典的文本相似性度量方法

李海林[1],邹金串[2]
(1.华侨大学 信息管理系,福建 泉州 362021; 2.华侨大学 现代应用统计与大数据研究中心,福建 厦门 361021)

摘 要:针对现有基于语义知识规则分析的文本相似性度量方法存在时间复杂度高的局限性,提出基于分类词典的文本相似性度量方法。利用汉语词法分析系统 ICTCLAS 对文本分词,运用 TF×IDF 方法提取文本关键词,遍历分类词典获取关键词编码,通过计算文本关键词编码的近似性来衡量原始文本之间的相似度。选取基于语义知识规则和基于统计两个类别的相似性度量方法作为对比方法,通过传统聚类与 KNN 分类分别对相似性度量方法进行效果验证。数值实验结果表明,新方法在聚类与分类实验中均能取得较好的实验结果,相较于其他基于语义分析的相似性度量方法还具有良好的时间效率。

关键词:文本挖掘;语义分析;分类词典;关键词提取;词语编码;相似性度量;聚类;分类
中图分类号:TP301 **文献标志码**:A **文章编号**:1673-4785(2017)04-0556-07

中文引用格式:李海林,邹金串.基于分类词典的文本相似性度量方法[J]. 智能系统学报, 2017, 12(4): 556-562.
英文引用格式:LI Hailin, ZOU Jinchuan. Text similarity measure method based on classified dictionary[J]. CAAI transactions on intelligent systems, 2017, 12(4): 556-562.

Text similarity measure method based on classified dictionary

LI Hailin[1], ZOU Jinchuan[2]
(1. Department of Information Systems, Huaqiao University, Quanzhou 362021, China; 2.Research Center of Applied Statistics and Big Data, Huaqiao University, Xiamen 361021, China)

Abstract: Existing text-similarity measurement methods based on the semantic knowledge rules analysis have the limitation of high time complexity. In this paper, we propose a text-similarity measurement method based on the Classified Dictionary. First, we segmented texts using the Chinese Lexical Analysis System. Then, we extracted text keywords using the term frequency-inverse document frequency (tf * idf) method and performed keywords coding by traversing the dictionary. By calculating the coding similarity of the text keywords, we can determine the similarity of the original texts. As our two comparison methods, we selected similarity measurement methods based on semantic knowledge rules and statistics. We verified our similarity measurement results using traditional clustering algorithms and the k-nearest neighbors classification method. Our numerical results show that our proposed method can obtain relatively good results in clustering and classification experiments. In addition, compared with other semantic analysis measurement methods, this method has better time efficiency.

Keywords: data mining; semantic analysis; classified dictionary; keywords extraction; encoder; similarity measure; clustering; classification

大数据时代,相似性度量方法通常作为数据挖掘任务的基础,使得相应的算法和技术能够在复杂数据中发现具有潜在价值的信息与知识[1-2],文本挖掘技术与方法通常用于处理与分析非结构化文本数据,其中相似性度量质量的好坏将很大程度上影响文本挖掘质量和效率,与文本相关的数据挖掘任务结合,也广泛存在于现实应用中,例如聚类与分类、信息检索、机器学习、网络信息认定[3]与人工智能等文本信息处理。

根据文献[4]中提到的概念层次理论,文本相似性度量建立在句子相似性度量之上,句子相似性

收稿日期:2016-08-30.
基金项目:国家自然科学基金项目(61300139);福建省自然科学基金项目(2015J01581);华侨大学中青年教师科研提升计划项目(ZQN-PY220);华侨大学研究生科研创新能力培育计划项目(1511307006).
通信作者:邹金串.E-mail:Zou_jinchuan@163.com.

度量进一步以词语的相似性为前提。因此,词语相似性度量结果的好坏直接影响文本相似性度量以及文本聚类、分类等后续文本挖掘任务与工作的质量。

词语相似度指在不同位置,词语可以互相替换使用的程度,文本相似性度量通常分为基于语义知识规则的相似性度量和基于统计的相似性度量。基于语义知识规则的文本相似度计算主要建立在基于 Wordnet[5-6]、MindNet[7]、FramNet[8] 等语义知识库的基础上。20 世纪 90 年代开始,涌现出大量基于 Wordnet 的语义相似度计算算法,主要针对外文长文本的语义相似度计算[9]。现有基于语义分析的中文文本相似性度量方法主要依托于同义词词林[10]与知网[11]。刘群等[12]以知网为依托,将词语相似性度量分为义原相似性度量、概念相似性度量和词语相似性度量 3 个步骤,并提出了基于知网的词语相似度计算方法(ZW_Sim)。由于该方法的适用性和有效性,部分学者在此基础上对该相似度计算方法进行改进。林丽等[13]在基于知网的词语相似度计算中引入弱义原的概念,即通过计算除区分能力弱的第一基本义原外的其他义原来计算词语相似度,以减少计算时间和提高计算精度;王小林[14]在原始基于知网方法的基础上,改进不同类别义原在词语相似度计算中所占权重的计算方法提高计算精度,通过义项词性判断降低相似性计算复杂度;张亮等[15]利用知网,从义项的主类义原、主类义原框架和义项特性描述三方面综合分析词语相似度,并从语义特征相似度和句法特征相似度两方面来描述词语相似度;田久乐等[16]提出基于同义词词林的词语相似度计算方法(CL_Sim),并通过人工测试、非人工测试以及与 ZW_Sim 方法进行比较,证明了方法的可行性;徐庆等[17]在此基础上对词语相似度计算公式进行改进,并将该方法应用于中文实体关系抽取,取得了较好的实验结果;郑红艳等[18]将词林与 TF×IDF 方法相结合,过滤同义词并对词语权重赋值进行文本特征提取,新的方法具有更好的特征提取结果。各位学者将基于知网与词林的相似性度量方法在参数与时间复杂度方面不断完善,使方法的准确性与时间效率都一定程度上有所提高。基于语义知识库的相似性度量方法均需要对语义知识库多次遍历,各位学者虽不同程度提高了方法的时间效率,但时间复杂度高的问题依然存在。

苏新春教授编写的《现代汉语分类词典》[19]与《同义词词林》在结构上具有相似性,但《现代汉语分类词典》对词语分类更细,词语间相似度只需通过两个词语编码进行计算比较,相较于 ZW_Sim 方法,不需要对词语相似度进行分层计算,时间复杂度大大降低。基于距离的语义相似度计算主要包括语义重合度(共同祖先节点数)、语义深度、语义密度、语义距离等 4 个方面的度量。多级分类体系使得基于分类词典的相似性度量结果可以直接反映两个词语在语义树中的重合度、深度与距离。在此基础上,本文提出一种基于现代汉语分类词典的文本相似性度量方法(Similarity measure based on Cidian, CD_Sim)。CD_Sim 方法通过中科院研制出的汉语词法分析系统对待分析文档进行分词等一系列基本处理,统计词语与文档间的词频矩阵,结合 TF×IDF 算法构建词语文档的向量空间模型[20],对向量空间模型进行标准化处理、排序等操作实现对文档的特征提取。通过 AP 聚类[21]、Kmeans 聚类[22]、谱聚类[23]3 种聚类算法以及 KNN 分类[24]方法对方法计算结果进行检验分析。方法理论简单、易于应用,对降低同义词、同类词导致的误差有一定作用,在短文本相似度量应用中相较于基于统计学的方法可以降低度量误差,相较于基于知识库的方法简单易行。数值实验结果表明,CD_Sim 方法在聚类与分类实验中均能取得较好的实验结果,证明了方法的可行性与度量效果。

1　相关理论基础

1.1　现代汉语分类词典

我国现代汉语首部分类词典是《同义词词林》,按照词语的概义来对词语进行分类编排。但现在《同义词词林》一定程度上不能很好地反映当前语言现状。《现代汉语分类词典》在吸收前人成果的基础上,收录了 8.3 万条通用性词语,较《同义词词林》新增常用词 2.9 万条,按五级语义层编排,包含 9 个一级类,62 个二级类,508 个三级类,2 057 个四级类,12 659 个五级类。

《现代汉语分类词典》用 5 层编码代表分类词典的 5 层结构,例如“B03Cc04”是“灰浆”和“砂浆”的编码,示例编码中各层编码意义如表 1,表示“灰浆”和“砂浆”均是“具体物”类别下“材料”类别中“建筑材料”范畴内“水泥石灰沥青”小类中的“灰浆”类别。若两个词语各级编码均相同,则二者是

同义词,相似度为 1。

表 1 分类词典编码方式示例

Table 1 Example of coding method of classified dictionary

编码位	符号举例	类别名	级别
1	B	具体物	第一级
2	3	材料	第二级
3	C	建筑材料	第三级
4	c	水泥石灰沥青	第四级
5	4	灰浆	第五级

1.2 向量空间模型

向量空间模型是当前使用较多的文本表示方法,向量空间矩阵为待分析文本样本词语-文档权重矩阵。假设待分析样本 $\boldsymbol{D}$ 中有 n 个文档 $d_j(j=1,2,\cdots,n)$,用 m 个词语 $t_i(i=1,2,\cdots,m)$ 在文档中出现的频数组成的向量对一篇文档进行向量表示,根据词语在该文档中出现的概率及在整个样本中出现的概率对该特征词的重要性赋值权重 w_{ij},则样本 $\boldsymbol{D}$ 表示为

$$\boldsymbol{D}' = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1j} \\ w_{21} & w_{22} & \cdots & w_{2j} \\ \vdots & \vdots & & \vdots \\ w_{i1} & w_{i2} & \cdots & w_{ij} \end{bmatrix} \tag{1}$$

式中:w_{ij}表示第 i 个词语在第 j 篇文档中重要程度的权值。

词语权重的计算方法有多种,经典权重计算方法如 TF×IDF 算法:

$$w_{ij} = \mathrm{TF}_{ij} \times \mathrm{IDF}_i \tag{2}$$

式中:TF_{ij}指特征词 t_i 在文档 d_j 中出现的次数 p_{ij} 占文档 d_j 中总词数 p_j 的比重:

$$\mathrm{TF}_{ij} = \frac{p_{ij}}{p_j} \tag{3}$$

IDF_i 为逆文档频率,计算公式为

$$\mathrm{IDF}_i = \log\left(\frac{N}{n_i}\right) \tag{4}$$

式中:N 为样本中文档总数,n_i 为样本中出现过特征词 t_i 的文档数。

2 文本相似度计算

针对目前基于语义知识规则的文本相似性度量方法存在计算过程中多次遍历语义知识库导致方法时间复杂度高的局限性,提出了基于现代汉语分类词典的文本相似性度量方法(Similarity measure based on Cidian, CD_Sim)。方法侧重于词语相似度量方法的改进,最终应用于文本相似度量,且度量方法较基于统计学的方法可以一定程度降低同义词、同类词导致的误差,故方法效果通过文本相似度量结果进行对比衡量。方法以《现代汉语分类词典》作为语义知识库,以基于 TF×IDF 方法的向量空间模型作为文本关键词提取依据,文本相似性度量过程包括词语编码获取、词语相似度计算和文本相似度计算 3 个步骤。

2.1 词语相似度计算

基于语义知识库的词语相似度通常通过计算义原相似度(ZW_Sim 方法)或者词语编码相似度(CL_Sim 方法)来计算。CD_Sim 方法通过遍历分类词典,在分类词典中搜索关键词,用该关键词在分类词典中对应的编码替换关键词进行关键词相似度计算。样本 D 中各文档以关键词编码集的形式表示。

分类词典中每一个大类均可以看做一棵语义树,同一个节点下的叶子节点为同义词,且同义词编码相同。通常词语相似性通过其在语义树中的位置进行度量计算,包括语义密度、语义深度、语义重合度、语义距离四方面衡量。分类词典对所有词语均采用 5 级分类,即所有词语语义深度相同,语义重合度与语义距离可通过公式计算互换(见式(8)),故可仅取其中一种衡量方式进行计算(涉及时间复杂度,语义密度暂不考虑)。

定义 关键词 A 的编码为"$a_1a_2a_3a_4a_5$",关键词 B 的编码为"$b_1b_2b_3b_4b_5$",两关键词语义重合度计算公式:

$$k_1 = A \otimes B = \sum_{i=1}^{5} a_i \otimes b_i \tag{5}$$

$i=1$ 时,

$$a_i \otimes b_i = \begin{cases} 1, & a_i = b_i \\ 0, & a_i \neq b_i \end{cases} \tag{6}$$

$i>1$ 时,

$$a_i \otimes b_i = \begin{cases} 1, & a_i = b_i \text{ and } a_{i-1} \otimes b_{i-1} = 1 \\ 0, & a_i \neq b_i \end{cases} \tag{7}$$

任意两个编码(假设两编码前三位相同,后两位不同)的语义重合度与语义距离在编码中可表示为式(8)形式:

$$\begin{array}{ccccc} & & & \overset{\leftarrow}{} & \\ a_1 & a_2 & a_3 & a_4 & a_5 \\ \Updownarrow & \Updownarrow & \Updownarrow & \downarrow & \\ b_1 & b_2 & b_3 & b_4 & b_5 \\ & & & \underset{\rightarrow}{} & \end{array} \tag{8}$$

则根据 $a_1 \Leftrightarrow b_1, a_2 \Leftrightarrow b_2, a_3 \Leftrightarrow b_3$ 前三对编码位相同,语义重合度(即相同父节点数)记为 3,语义距离(即从末位编码开始向上遍历编码位,经过第一共同编码位再到另一编码末位编码所经过的不同编码位的路径数)表示为 $a_5 \rightarrow a_4 \rightarrow b_4 \rightarrow b_5$,记为 3。根据语义重合度和语义距离的概念与计算规则,通过换算,得到任意两编码语义距离公式为

$$k_2 = \begin{cases} 9 - 2 \times k_1, & k_1 < 5 \\ 0, & k_1 = 5 \end{cases} \tag{9}$$

根据编码语义重合度和语义距离的计算公式,列出 3 个编码,分别求两两编码的语义重合度和语义距离,验证计算公式的正确性与可行性。二者换算示例如表 2。

表 2　语义重合度与语义距离换算示例

Table 2　Example of conversion between coincidence and distance of semantic

编码	深度	重合度/距离		
		B03Cc04	B03Dc03	C02Cb01
B03Cc04	5	5/0	2/5	0/9
B03Dc03	5	2/5	5/0	0/9
C02Cb01	5	0/9	0/9	5/0

考虑到语义重合度与语义距离可互相换算,CD_Sim 方法中词语相似度均采用语义重合度进行计算,将语义重合度标准化公式:

$$\mathrm{Sim}(A,B) = \frac{k_1}{5} \tag{10}$$

将关键词转化为编码可以更加直观表示关键词在词典中所属类别,在关键词相似度计算过程中直接通过编码计算,不需要多次访问语义知识库,提高了计算的时间效率。

2.2　相似度计算

文本相似度计算建立在词语相似性度量之上,每个关键词与对比文档中关键词的距离取该关键词与对比文档中所有关键词相似度的最大值。设文档 $d_1(t_1,t_2,\cdots,t_p)(p=1,2,\cdots,x)$ 有 x 个关键词,文档 $d_2(t_1,t_2,\cdots,t_q)(q=1,2,\cdots,y)$ 有 y 个关键词,计算 d_1 与 d_2 中所有关键词的相似度矩阵

$$\mathbf{Sim} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1q} \\ s_{21} & s_{22} & \cdots & s_{2q} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pq} \end{bmatrix} \tag{11}$$

式中:s_{pq} 表示文档 d_1 中的第 p 个关键词与文档 d_2 中的第 q 个关键词的相似度。根据两文本关键词相似度矩阵可求文本相似度为

$$\mathrm{SIM}(d_1,d_2) = \frac{\sum_{p=1}^{x} \max(s_{p1},s_{p2},\cdots,s_{py}) + \sum_{q=1}^{y} \max(s_{1q},s_{2q},\cdots,s_{xq})}{x+y} \tag{12}$$

关键词与比较文本关键词相似度取该关键词与比较文本所有关键词相似度最大值,即对关键词相似度矩阵每行每列均取最大值,平均值即为两文本相似度。

基于现代汉语分类词典的文本相似性度量算法 $Z=\mathrm{CD\_Sim}(\boldsymbol{D})$:

输入　待分析样本 $\boldsymbol{D}$;

输出　样本 $\boldsymbol{D}$ 中所有文本间相似度集合 Z。

1)对样本 $\boldsymbol{D}$ 中所有文档进行分词、过滤停用词处理;

2)对处理后的结果构建词语-文档频数矩阵,并结合 TF×IDF 方法构建样本的向量空间模型 $\boldsymbol{D}'$;

3)根据向量空间模型 $\boldsymbol{D}'$ 对每篇文档按照一定的规则进行关键词提取;

4)for $i=1:\mathrm{size}(\boldsymbol{D},1)-1$

①for $j=i+1:\mathrm{size}(\boldsymbol{D},1)$

a)根据式(10)计算文档 i 和文档 j 中所有关键词相似度,并按式(11)将计算结果存入相似度矩阵 **Sim**;

b)将相似度矩阵 **Sim** 按式(12)进行计算,得到文档 i 和文档 j 的相似度 $\mathrm{SIM}(d_i,d_j)$;

②End

5)End

6)得出样本 $\boldsymbol{D}$ 中所有文本间相似度集合 Z。

根据方法介绍,CD_Sim 方法与 CL_Sim 方法时间复杂度均为 $O(n^2)$,ZW_Sim 方法时间复杂度为 $O^3(n^2)$。

3　仿真实验

为检验 CD_Sim 方法的结果在应用中的准确性与时间效率,从搜狗分类语料库[25]中随机选择 5 类数据作为实验样本,采用中科院分词软件对样本进行预处理,通过 TF×IDF 方法对处理结果进行关键词提取,选择基于语义知识规则和基于统计两类词语相似性度量方法作为对比方法,用聚类与分类两种方法对相似性度量结果进行检验。文中文本相似性度量方法仿真实验对每篇文档取词语权值排序前 15 位词语作为文本关键词进行数值实验。

3.1 实验数据与实验设计

实验语料数据选自搜狗实验室提供的搜狗分类语料库,该语料库包含了环境、计算机、交通、教育、经济、军事、体育、医药、艺术和政治 10 个类别文本文档。

数值实验选取了环境、交通、政治、教育、体育 5 个类别,每个类别随机选取 20 个文本文档共 100 个文本文档进行实验。实验中通过 TF_IDF 特征选择方法在 100 个文本中分别选择 15 个关键词进行相似性度量,其中,由于基于统计方法的特殊性,该类方法采用整个词语-文档权重矩阵进行相似度计算。

实验选择基于 LSA 的文本相似性度量方法、基于词林的文本相似性度量方法和基于知网的语义相似性度量方法作为对比方法,分别采用 AP 聚类、Kmeans 聚类、谱聚类以及 KNN 分类对相似性度量结果进行检验。

3.2 聚类分析

相似性度量结果的好坏直接影响文本聚类算法的精度,在已知文档类别的样本中,聚类精度可以反过来检验文本相似性度量结果的好坏。比较经典的基于距离矩阵的聚类算法有 Kmeans,AP 聚类及后来发展起来的谱聚类算法等。Kmeans 与谱聚类算法均是给定聚类数目的聚类算法,时间复杂度低,聚类准确度高;在聚类数目未知的情况下,上述两种方法聚类结果会产生较大的偏差。AP 聚类没有事先给定聚类数目,根据数据自身的特性进行聚类,聚类结果与聚类对象特征更加吻合。将相似性度量方法实验结果做聚类分析,数值实验结果如表 3。

表 3 基于聚类检验方法的数据实验结果

Table 3 Experiment results based on clustering method

方法	AP 聚类			Spectral 聚类		Kmeans 聚类	
	NUM	熵值	纯净度	熵值	纯净度	熵值	纯净度
CL_Sim	14	0.96	0.74	1.76	0.47	1.84	0.41
ZW_Sim	9	2.13	0.28	1.84	0.41	2.22	0.24
CD_Sim	18	0.33	0.90	0.90	0.82	1.26	0.66
LSA_Sim	18	0.60	0.85	1.60	0.50	1.68	0.51

数值实验中,聚类结果通过熵值和纯净度来度量。聚类结果熵值越低、纯净度越高,则聚类结果越好。NUM 记录了将各相似性度量方法结果进行 AP 聚类所得聚类类别数。基于 LSA 的相似性度量算法,K 值取[10,20,…,100]这 10 组数据值进行实验,每种聚类检验方法中均取熵值最小且纯净度最高的实验结果作为基于 LSA 的相似性度量算法的实验结果。

根据聚类实验结果分析,对 4 种相似性度量方法进行比较。AP 聚类中,CD_Sim 方法聚类结果最好,但数值实验样本仅包含 5 类文档,CD_Sim 方法聚类数目达 18 种,存在一定的不合理性。在谱聚类算法中,CD_Sim 方法聚类检验结果明显优于其他相似性度量方法,在 4 种相似性度量方法中,熵值最小,纯净度最高。Kmeans 聚类算法中,CD_Sim 方法实验结果纯净度较低、熵值较大,但结果仍优于其他相似性度量方法。

根据实验结果,对 3 种基于语义知识规则的相似性度量方法聚类实验结果进行比较分析,CD_Sim 方法实验结果优于 CL_Sim 方法和 ZW_Sim 方法,聚类熵值最小、纯净度最高。

3.3 分类实验

分类检验采用 KNN 算法进行分析,算法从每个类别样本中均选取一半作为已知类别样本,剩下一半作为实验集,检验结果以分类准确率进行度量,分类算法 K 值分别取[1,2,…,10],得出 10 组不同 K 值下的 KNN 分类结果并取平均值 mean。采用不的相似性度量方法作为文本之间近似性度量方法,结合 KNN 方法进行数值实验, 其实验结果如表 4 所示。

表 4 基于分类检验方法的数据实验结果

Table 4 Experiment results based on classified method

方法	分类										mean
	1	2	3	4	5	6	7	8	9	10	
CL_Sim	0.60	0.58	0.64	0.64	0.72	0.68	0.70	0.62	0.70	0.78	0.67
ZW_Sim	0.22	0.24	0.24	0.26	0.24	0.26	0.26	0.24	0.26	0.28	0.25
CD_Sim	0.80	0.84	0.84	0.90	0.90	0.84	0.88	0.90	0.86	0.84	0.86
LSA_Sim	0.82	0.84	0.84	0.84	0.84	0.84	0.86	0.76	0.78	0.80	0.82

数值实验结果表明,4 种相似性度量方法中,CD_Sim 方法分类实验结果最好,分类准确率最高,LSA_Sim 方法实验结果次之,优于其他方法分类实验结果。3 种基于语义知识规则的相似性度量方法分类检验结果进行比较,CD_Sim 方法分类实验结果优于 CL_Sim 方法和 ZW_Sim 方法,分类准确度最高。

3.4 时间复杂度分析

实验中方法的时间复杂度是除准确性外方法

可行性的重要指标,实验过程中对各方法 100 个文档的相似度矩阵计算时间计时,结果如表 5。

表 5　相似性度量方法时间复杂度

Table 5　Time complexity of similarity measure method

方法	CL_Sim	ZW_Sim	CD_Sim	LSA_Sim
时间/s	10 266	1 014 682	5 734	10.6

根据表 5 实验数据,4 种文本相似性度量方法中,基于统计的文本相似性度量方法时间效率较高,基于语义知识规则的文本相似性度量方法较基于统计的方法时间效率较低。在 3 种基于语义知识规则的文本相似性度量方法中,CD_Sim 方法时间效率最高,CL_Sim 方法时间效率次之,ZW_Sim 方法时间效率最低。CD_Sim 方法遍历知识库的次数为样本中所有文档关键词的个数 m,CL_Sim 方法遍历知识库次数为$(m+O(n^2))$,ZW_Sim 方法遍历知识库次数为 m。综合文本相似性度量方法时间复杂度与遍历知识库的次数,CD_Sim 方法在 3 种基于语义知识规则的文本相似性度量方法中时间效率最高。

3.5　方差分析

方法的稳定性也是方法可行性的重要指标。通常方差用来检验数据的稳定性,方差值越小,数据越稳定。分别对 4 种方法的 4 个实验结果准确率求方差,来验证 4 种相似性度量方法实验稳定性:

根据表 6,ZW_Sim 方法实验结果最稳定,CD_Sim 方法次之,LSA_Sim 方法方差最大,实验结果稳定性相对较差。

表 6　相似性度量方法方差

Table 6　Variance of similarity measure method

方法	CL_Sim	ZW_Sim	CD_Sim	LSA_Sim
方差	0.019	0.005	0.008	0.027

综合聚类实验、分类实验、时间复杂度和稳定性,CD_Sim 方法准确性优于对比方法,稳定性优于大部分对比方法,时间效率优于其他基于语义知识规则的对比方法,对比基于统计的方法时间效率仍有差距。

4　结束语

文本相似性计算的关键在于关键词相似度计算,文本可以看作词语的集合,关键词根据其提取方法认为是不同筛选程度下文本中能够区别于其他文本的词语,各位学者的语义方法均是在不同程度关键词筛选结果的基础上进行。文章提出了基于分类词典的文本相似性度量方法,对样本进行分词、计算词语权重、提取文本关键词等一系列基本处理,定义基于关键词编码的词语相似度计算公式,构建文档关键词相似度矩阵,根据关键词相似度矩阵计算文档相似度。通过聚类与分类实验对相似性度量结果进行验证,验证了该方法的合理性。方法采用分类词典作为知识库,分类词典相较于词林和知网收录了更多的词语,词语编码匹配成功概率更高,对文本相似性度量影响较小;计算过程中仅在词语编码匹配一个阶段访问知识库,提高了基于语义知识库方法的时间效率;提出了新的词语相似度计算方法,计算结果优于其他基于语义知识库的方法。由于各领域的发展都会不断产生新的词语,文本实验过程中出现部分分类词典中未收录的词语,这部分词语不参加文本相似度计算,一定程度上会导致实验结果的误差;相较于基于统计的相似度计算方法,方法的时间效率有待提高。在保证方法准确度的前提下提高时间效率是 CD_Sim 方法未来的研究方向。

参考文献:

[1]李海林,郭韧,万校基.基于特征矩阵的多元时间序列最小距离度量方法[J].智能系统学报, 2015, 10(3): 442-447, 2015.
LI Hailin, GUO Ren, WAN Xiaoji. A minimum distance measurement method for a multivariate time series based on the feature matrix [J]. CAAI transactions on intelligent systems, 2015, 10(3): 442-447.

[2]XU R, WUNSCH D. Survey of clustering algorithms[J]. IEEE transactions on neural networks, 2005, 16(3): 645-678.

[3]CHEN Wei, HUO Junge. Judicial determination of malicious forwarding cyber false information[J]. Journal of Chongqing university: social science edition,2017(5) : 103-113.

[4]苗传江.HNC(概念层次网络理论)引导[M]. 北京:清华大学出版社,2005.

[5]PARK E K, RA D Y, JANG M G. Techniques for improving web retrieval effectiveness[J]. Information processing and management, 2005, 41(5): 1207-1223.

[6]WordNet Documentation [EB/OL]. [2010 - 10 - 27]. http://wordnet.princeton.edu/wordnet/documentation/.

[7]RICHARDSON S D, DOLAN W B. VANDERWENDE L. MindNet: Acquiring and structuring semantic information from text [C]//Proceeding of the 17th International Conference on Computer Linguistics Volume 2. Stroudsburg: Association for Computational Linguistics, 1998: 1098-1102.

[8]BAKER C F, FILLMORE C J, LOWE J B. The Berkeley framenet project [C]//Proceeding of the 36th Annual

Meeting of the Association for Computational Linguistics and 17th International Conference on Computer Linguistics Volume 1. Stroudsburg: Association for Computational Linguistics, 1998: 86-90.

[9]翟延东,王康平. 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. An algorithm for semantic similarity of short text based on WordNet[J]. Acta electronica sinica, 2012, 40(3): 617-620.

[10]梅家驹,竺一鸣,高蕴琦,等.同义词词林[M].上海:上海辞书出版社,1996.

[11]董振东, 董强. 知网简[EB/OL]. http://www.keenage.com.

[12]刘群,李素建. 基于"知网"的词汇语义相似度计算[C]//第三届汉语词汇语义学研究会论文集.台北,中国, 2002: 59-76.

[13]林丽,薛方,任仲晟. 一种改进的基于知网的词语相似度计算方法[J].计算机应用,2009, 29(1): 217-220.
LIN Li, XUE Fang, REN Zhongsheng. Modified word similarity computation approach based on Hownet[J]. Journal of computer applications, 2009, 29(1): 217-220.

[14]王小林,杨林,王东. 基于知网的新词语相似度算法研究[J]. 情报科学, 2015, 33(2): 67-71.
WANG Xiaolin, YANG Lin, WANG Dong. New word similarity algorithm research based on HowNet[J]. Information science, 2015, 33(2): 67-71.

[15]张亮,尹存燕.基于语义树的中文词语相似度计算与分析[J]. 中文信息学报, 2007, 21(3): 99-105.
ZHANG Liang, YIN Cunyan. Chinese word similarity computing based on semantic tree[J]. Journal of Chinese information processing, 2007, 21(3): 99-105.

[16]田久乐,赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报: 信息科学版,2010, 26(6): 602-608.
TIAN Jiule, ZHAO Wei. Word similarity algorithm based on Yongyici Cilin in Semantic Web adaptive learning system[J]. Journal of Jilin university: information science edition, 2010, 26(6): 602-608.

[17]徐庆,段利国.基于实体语义相似度的中文实体关系抽取[J]. 山东大学学报:工学版, 2015, 45(6): 7-14.
XU Qing, DUAN Liguo. Chinese entity relation extraction based on entity semantic similarity[J]. Journal of Shandong university: engineering science, 2015, 45(6): 7-14.

[18]郑红艳,张东站.基于同义词词林的文本特征选择方法[J]. 厦门大学学报:自然科学版, 2012, 5(2): 200-203.
ZHENG Hongyan, ZHANG Dongzhan. A text feature selection method based on TongYiCi CiLin[J]. Journal of Xiamen University: Natural Science, 2012, 5(2): 200-203.

[19]苏新春. 现代汉语分类词典[M]. 上海:商务印书馆, 2013.

[20]SALTON G. The transformation analysis and retrival of information by computer[M]. Wesley Reading Massach-uetts, 1989.

[21]FREY B J, DUECK D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976.

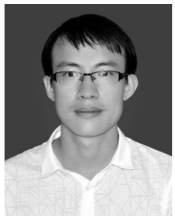
[22]FORGY E W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications[J]. Biometric, 1965, 21: 768-769.

[23]丁世飞,贾洪杰.基于自适应 Nystrom 采样的大数据谱聚类算法[J]. 软件学报, 2014, 25(9): 2037-2049.
DING Shifei, JIA Hongjie. Spectral clustering algorit-hm based on adaptive nystrom sampling for big data analysis[J]. Journal of software, 2014, 25(9): 2037-2049.

[24]WU Xindong, KUMAR V, QUINLAN J R, et al. Top 10 algorit-hms in data mining[J]. Knowledge and information systems, 2008, 14(1): 1-37.

[25]搜狗实验室语料[EB/OL]. http://www.sogou.com/labs/resource/list_yuliao.php.

作者简介:



李海林,男,1982 年生,副教授,博士,主要研究方向为数据挖掘与决策支持,主持国家自然科学基金 1 项和省部级基金 2 项,发表学术论文 40 余篇,其中被 SCI 检索 11 篇,EI 检索 20 余篇。

邹金串,女,1993 年生,硕士研究生,主要研究方向为文本挖掘。