融合耦合距离区分度和强类别特征的短文本相似度计算方法

马慧芳[1,2,3],刘 文[1],李志欣[3],蔺想红[1]
(1. 西北师范大学计算机科学与工程学院,甘肃兰州 730000; 2. 桂林电子科技大学广西可信软件重点实验室,广西桂林 541004; 3. 广西师范大学广西多源信息挖掘与安全重点实验室,广西桂林 541004)

摘 要: 短文本相似度计算在社会网络、文本挖掘和自然语言处理等领域中起着至关重要的作用.针对短文本内容简短、特征稀疏等特点,以及传统的短文本相似度计算忽略类别信息等问题,提出一种融合耦合距离区分度和强类别特征的短文本相似度计算方法.一方面,在整个短文本语料库中利用两个共现词之间的距离计算词项共现距离相关度,并以此来对词项加权从而捕获词项间内联和外联关系,得到短文本的耦合距离区分度相似度;另一方面,基于少量带类别标签的监督数据提取每类中强类别区分能力的特征项作为强类别特征集合,并利用词项的上下文来对强类别特征语义消歧,然后基于文本间包含相同类别的强类别特征数量来衡量文本间的相似度.最后,本文结合耦合距离区分度和强类别特征来衡量短文本的相似度.经实验证明本文提出的方法能够提高短文本相似度计算的准确率.

关键词: 文本挖掘; 自然语言处理; 文本聚类; 社会网络; 耦合关系; 特征提取; 语义消歧; 相似度计算
中图分类号: TP393.092 **文献标识码:** A **文章编号:** 0372-2112 (2019)06-1331-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.06.021

Combining Coupled Distance Discrimination and Strong Classification Features for Short Text Similarity Calculation

MA Hui-fang[1,2,3], LIU Wen[1], LI Zhi-xin[3], LIN Xiang-hong[1]
(1. *College of Computer Science and Engineering, Northwest Normal University, Lanzhou, Gansu* 730000, *China*;
2. *Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, Guangxi* 541004, *China*;
3. *Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin, Guangxi* 541004, *China*)

Abstract: Text similarity measures play a vital role in text related applications in tasks such as social networks, text mining, natural language processing, and others. The typical characteristics of short texts demonstrate severe sparseness and high dimension while the traditional short texts similarity calculation always ignores category information. A coupled distance discrimination and strong classification features based approach for short text similarity calculation, CDDCF, is presented. On the one hand, co-occurrence distance between terms are considered in each text to determine the co-occurrence distance correlation, based on which the weight for each term can be determined and the intra and inter relations between words are established. The similarity of coupling distance discrimination on short text can be captured. On the other hand, strong classification features are extracted via labeled texts. The similarity between two short texts is measured by using the common number of strong discrimination features with the same context. Finally, the distance discrimination and strong classification features are unified into a joint framework to measure the similarity of short texts. Experimental results show that CDDCF performs better compared to baseline algorithms in term of its performance and efficiency of similarity computation.

Key words: text mining; natural language processing; text clustering; social network; coupling relation; feature extraction; word sense disambiguation; similarity computation

1 引言

随着互联网技术的飞速发展,微博,微信,手机短信凭借开放性和便捷性等优势,已发展成为人们社交和娱乐的主流媒体,是人们了解时事动态和发表观点和评论的主要平台[1].面对这些应用产生的超大规模短

收稿日期:2018-01-30;修回日期:2019-03-11;责任编辑:蓝红杰
基金项目:国家自然科学基金(No. 61762078, No. 61363058, No. 61663004);广西多源信息挖掘与安全重点实验室开放基金项目(No. MIMS18-08);广西可信软件重点实验室研究课题(No. KX201705)

文本数据,怎样挖掘隐藏在数据中的巨大的潜在价值是研究的热点和难点. 而短文本相似度的计算的优劣对于挖掘数据隐藏的价值起着至关重要的作用,被大量用于文本聚类,舆情分析,兴趣推荐等多个领域.

当前短文本相似度的计算方法主要分为两大类. 第一类方法常见的是在向量空间模型的基础上计算向量间的相似度,典型的工作有:Song 等人[2]利用共现词项的概率相关度来计算词项在文本中的权重改进了相似度计算方法. 但是该类方法并未很好的描述词项间更深层次的关系. 第二类基于外部语料库的方法,常见的方法有:Li 等人[3]利用大规模语义网络 Probase 将两个词项映射到概念空间中,并在聚类后的概念空间中计算词项的相似度. 基于语料库的方法的局限在于:只能处理语料库中的词项,不能处理语料库中未出现的词项.

本文针对上述缺陷提出一种融合耦合距离区分度和强类别特征的短文本相似度计算方法,首先利用词项的关联权重计算词项的内联关系,接着利用链接词产生的路径的共享熵来表征外联关系,然后耦合这两种关系得到耦合距离区分度. 此外本文利用有监督的方法来衡量文本间的相似度,即利用加类标数据得到每个类别的强类别特征集合,并利用强类别特征词项的上下文信息进行语义消歧,基于文本包含每个类的强类别特征越多则越相似,得到强类别特征相似度. 最后通过平衡因子来调节两种相似度来得到最终的短文本相似度.

2 基于耦合距离区分度的短文本相似度计算

2.1 共现距离区分度

由给定短文本集合 $D=\{d_1,d_2,\cdots,d_m\}$ 和词项集 $T=\{t_1,t_2,\cdots,t_n\}$,词项 t_i 与词项 t_j 在特定短文本 d_s 中的共现距离的计算式如下:

$$\mathrm{cor}_{d_s}(t_i,t_j)=\begin{cases}\mathrm{e}^{-\mathrm{dist}_{d_s}(t_i,t_j)}, & t_i\in d_s \text{ 且 } t_j\in d_s\\ 0, & t_i\notin d_s \text{ 或 } t_j\notin d_s\end{cases} \tag{1}$$

其中 $\mathrm{dist}_{d_s}(t_i,t_j)$ 为词项 t_i 与 t_j 在短文本 d_s 中的距离,可由这两个词项之间相隔的词数得到,由式(1)可知若两个词之间相隔词数越多则它的共现距离越低. 词项共现区分度定义如下:

$$\mathrm{cor}_D(t_i\mid t_j)=\frac{\sum\limits_{d_s\in D}\mathrm{cor}_{d_s}(t_i,t_j)}{\sum\limits_{d_s\in D}\sum\limits_{t_k\in T}\mathrm{cor}_{d_s}(t_i,t_k)}\times\log_2\frac{n}{\mathrm{Nei}(t_j)} \tag{2}$$

其中,$\mathrm{cor}_D(t_i\mid t_j)$ 反应了词项 t_i 与 t_j 的共现情况,$N_{\mathrm{nei}}(t_j)$ 表示与词项 t_j 共现的词的个数. 对称化处理后有:

$$\mathrm{cor}(t_i,t_j)=\frac{\mathrm{cor}_D(t_i\mid t_j)+\mathrm{cor}_D(t_j\mid t_i)}{2} \tag{3}$$

在特定文本 d_s 中,基于共现距离相关度来对词项 t_i 加权如下:

$$\mathrm{cow}_{d_s}(t_i)=w_{t_i}+\frac{\sum\limits_{t_j\in d_s}w_{t_j}\times\mathrm{cor}(t_i,t_j)}{|d_s|} \tag{4}$$

其中,w_{t_i} 可由词项 t_i 在短文本 d_s 中的词频计算得到. 为了使词项特征尽可能的具有区分性,将词项权重和逆文档频率(Inverse Document Frequency,IDF)结合起来,词项的关联权重可进一步表示为:

$$w_{d_s}(t_i)=\mathrm{cow}_{d_s}(t_i)\times\mathrm{IDF}=\mathrm{cow}_{d_s}(t_i)\times\log_2\frac{m}{n(t_i)} \tag{5}$$

式(5)中 $n(t_i)$ 表示包含词项 t_i 的文本个数. 因为考虑到词项在整个文本的逆文档频率,所以式(5)计算得到的权重结果优于式(4).

2.2 耦合共现距离区分度的相似度

2.2.1 内联关系

若两个词项频繁共现,它们之间就有显式的关系. 此外,如果它们自身有很高的权重,那词项间关系则更强. 文献[4]将这种关系定义为内联关系. 本节对内联关系进行了改进,即使用词项的关联权重来代替 tf-idf 来捕获更加丰富的语义信息,内联关系的定义如下.

定义 1(内联关系) 若两个词项至少在一个文本 d_s 中共现,则两个词项间有内联关系,定义为:

$$\mathrm{CoR}(t_i,t_j)=\frac{1}{|H|}\sum_{d_s\in H}\frac{w_{d_s}(t_i)\times w_{d_s}(t_j)}{w_{d_s}(t_i)+w_{d_s}(t_j)-w_{d_s}(t_i)\times w_{d_s}(t_j)} \tag{6}$$

$w_{d_s}(t_i)$ 和 $w_{d_s}(t_j)$ 分别表示为词项 t_i 和 t_j 在文本 d_s 中的关联权重,可由式(5)得到. 其中 $H=\{d_s\mid(w_{d_s}(t_i)\neq0\vee w_{d_s}(t_j)\neq0)\}$,特别地,当 $H=\varnothing$,定义 $\mathrm{CoR}(t_i,t_j)=0$. 对式(6)归一化如下:

$$\mathrm{UIaR}(t_i\mid t_j)=\begin{cases}1, & i=j\\ \dfrac{\mathrm{CoR}(t_i,t_j)}{\sum\limits_{j=1}^{n}\mathrm{CoR}(t_i,t_j)}, & i\neq j\end{cases} \tag{7}$$

$\mathrm{UIaR}(t_i\mid t_j)$ 表示词项 t_i 和 t_j 的内联关系占词项 t_i 与词典中的其他词项内联关系和值的比重,即看到词项 t_i 时 t_j 共同出现的概率. 内联关系是不对称的,对式(7)对称化处理如下:

$$\mathrm{IaR}(t_i,t_j)=\frac{\mathrm{UIaR}(t_i\mid t_j)+\mathrm{UIaR}(t_j\mid t_i)}{2} \tag{8}$$

2.2.2 外联关系

内联关系仅仅捕获了两个共现词项的显性关系,但是并未考虑两个词项未在文本中共现过,但通过一些其他词项形成链接路径而产生一定的关联关系. 本节中引入外联关系来建模词项间的隐含关系. 通过考虑这种以

其它词项为路径的词对,构建词对外联关系图(Inter-Coupling Graph,ICG),其中顶点为词项,边表示词项间的关系,当且仅当词对在文本中共现则结点存在连边[5].

定义 2(外联路径) 对于任意两个词项 t_i 和 t_j,存在一条或多条从词项 t_i 开始,且有序的链接多个词项后以 t_j 结束的词项序列称为路径,则词项间的外联路径被形式化的定义为:

$$\begin{aligned}\mathrm{Path}(t_i\rightarrow t_j)=\{&T^{p}_{i\rightarrow j},E^{P}_{i\rightarrow j}\mid t_i,t_{l_1},\cdots,t_{l_g},t_j\in T^{P}_{i\rightarrow j},\\&e_{il_1},e_{l_1l_2},\cdots,e_{l_gj}\in E^{P}_{i\rightarrow j},t_i\neq t_j,\\&T^{P}_{i\rightarrow j}\subset T,E^{P}_{i\rightarrow j}\subset E,g\in[1,\theta]\}\end{aligned}\tag{9}$$

其中词项 t_i 为起始点,t_j 为终止点,$\mathrm{Path}(t_i\rightarrow t_j)$ 代表路径 $\mathrm{Path}(t_i\rightarrow t_j)$ 上的词项,即链接词,g 是路径中链接词的个数,$T^{P}_{i\rightarrow j}$ 为 $\mathrm{Path}(t_i\rightarrow t_j)$ 上特定路径 P 上所有点的集合,e_{ij} 表示两个点之间有边,E 为所有边的集合,$E^{P}_{i\rightarrow j}$ 为 $\mathrm{Path}(t_i\rightarrow t_j)$ 第 P 条路径上所有经过边的集合. θ 是用户为限制 $\mathrm{Path}(t_i\rightarrow t_j)$ 数量(即链接词个数)所定义的阈值. 本文设计新的方法来量化路径上多个词之间的高阶关系. 将词对的某路径上包含的词项集合定义为:

定义 3(链接词集) 词项 t_i 到 t_j 的路径 $\mathrm{Path}(t_i\rightarrow t_j)$ 上任一路径 p 上的所有词项的集合为链接词项集 $T^{\text{p-link}}$,路径长度为 h,公式如下:

$$T^{\text{p-link}}=\{t_{l_h}\mid t_{l_h}\in T^{P}_{i\rightarrow j},T^{P}_{i\rightarrow j}\in\mathrm{Path}(t_i\rightarrow t_j)\}\tag{10}$$

链接词集表示某条路径上包含起止点的所有词项的集合. 因此,对于任意词对,可以通过考虑其所有可能路径上的词项间的高阶关系,来挖掘词项之间的语义关联. 在模式识别任务中定义了共享熵用来量化多模态特征之间的关联程度[6]. 而本文利用共享熵来衡量两个词项在特定路径上的关联程度,定义如下:

定义 4(链接词集共享熵) 词项 t_i 到 t_j 的第 p 条路径的链接词集 $T^{\text{p-link}}$ 上词对间的关联程度,用该条路径的链接词集共享熵表示:

$$\begin{aligned}S_p(T^{\text{p-link}})=&(-1)^0\sum_{t_{l_h}\in T^{\text{p-link}}}J(t_{l_h})+\\&(-1)^1\sum_{1\leqslant i<j\leqslant h}J(t_{l_i},t_{l_j})+\cdots+\\&(-1)^{h-1}J(T^{\text{p-link}})\end{aligned}\tag{11}$$

其中,$J(\cdot)$ 为该路径上链接词间的联合熵公式如下:

$$\begin{aligned}J(T^{\text{p-link}})=-&\sum_{t_i,t_{l_1},t_{l_2},\cdots,t_{l_{h-1}},t_j}P(t_i,t_{l_1},t_{l_2},\cdots,t_{l_{h-1}},t_j)\\&\times\log_2P(t_i,t_{l_1},t_{l_2},\cdots,t_{l_{h-1}},t_j)\end{aligned}\tag{12}$$

其中 $P(t_i,t_{l_1},t_{l_2},\cdots,t_{l_{h-1}},t_j)$ 是一条路径上词项的联合概率. 随着路径长度的增加,计算共享熵所消耗的代价越大,本文通过实验限定了路径的最大长度,即 $3\leqslant h\leqslant 8$. 任意词对 t_i 与 t_j 某条路径上对应的的链接词集共享熵越大,词对间的关联性就越强. 通过归一化词对在第 p 条路径上的共享熵来表示第 p 条路径的外联关系,定义如下:

$$\mathrm{IeR}_p(t_i,t_j)=\frac{S_p(T^{\text{p-link}})}{\sum_{p=1}^{q}S_p(T^{\text{p-link}})}\tag{13}$$

其中 q 为该词对间路径的总数,计算词对在第 p 条路径上的共享熵占所有路径上的比例来表示第 p 条路径的关联程度. 最后,本文选取词对所有路径中共享熵最大值来表征词对 t_i 与 t_j 间的外联关系的强弱.

$$\mathrm{IeR}(t_i,t_j)=\max\{\mathrm{IeR}_p(t_i,t_j)\}\tag{14}$$

2.2.3 耦合距离区分度的相似度计算

为了更好地捕获两个词项的隐藏关系,本文同时结合内联和外联关系来定义两个词项的耦合关系. 基于式(8)和(14)定义了两个词项的耦合距离区分度. 给定 T 中的任意词项 t_i 和 t_j,本文定义耦合距离区分度为:

$$\mathrm{CR}(t_i,t_j)=\begin{cases}1, & i=j\\ \alpha\times\mathrm{IaR}(t_i,t_j)+(1-\alpha)\times\mathrm{IeR}(t_i,t_j), & i\neq j\end{cases}\tag{15}$$

其中,$\alpha\in[0,1]$,表示为调节内联关系和外联关系权重的调节因子. $\mathrm{CR}(t_i,t_j)$ 表示为词项 t_i 和 t_j 之间的相似度,若取值越趋于为 0,表明两个词项关联性越弱,反之则两个词项关联关系越强. 选取 $\mathrm{CR}(t_i,t_j)$ 的词对 (t_i,t_j) 放入集合 M 中,根据经验,本文中 $\eta=0.3$. 利用上面的知识设计了基于耦合距离区分度的文本相似度算法(Couple distance-differentiation Relation for short text similarity measure,CR)为:

$$\begin{aligned}S_{\mathrm{CR}}(d_1,d_2)=&\frac{1}{\|d_1\|\|d_2\|}\sum_{t_i\in d_1}\prod w_{d1}(t_i)\\&\times w_{d2}h(t_i)\times\mathrm{CR}(t_i,h(t_i))\end{aligned}\tag{16}$$

其中 $h(t_i)=\{t_j\mid t_j\in d_2\wedge(t_i,t_j)\in M\}$,$h(t_i)$ 表示为与词项 t_i 相似度值大于阈值 η 的词项的集合. 该方法避免了余弦相似度必须同维且必须一对一的局限,同时大大减少了相似度计算的代价.

3 基于强类别特征的相似度计算

3.1 改进的期望交叉熵对强类别特征的提取

对于有监督的特征提取方法,如何得到每个类中最能代表该类的特征对于计算相似度至关重要. 本文利用改进的期望交叉熵[7]的方法来提取强类别特征词. 改进的期望交叉熵(Improved Expected Cross Entropy,IECE)计算方法详见文献[7].

文献[7]定义词项 t_i 在类别 C_r 中的整体权重值为 $\mathrm{IECE}_{C_r}(t_i)$. 该值越大,说明词项 t_i 对类别 C_r 的指示性越强. 类别 C_r 中词项 t_i 的最终权重计算式如下:

$$W_{C_r}(t_i)=\mathrm{IECE}_{C_r}(t_i)\times\mathrm{idf}(t_i)=\mathrm{IECE}_{C_r}(t_i)\times\log_2\frac{|C_r|}{n_{C_r(t_i)}}\tag{17}$$

最后对类别 C_r 中的词项按 $W_{C_r}(t_i)$ 值进行降序排列,取每个类中前 L 个词项构成强类别特征集合 $S=\{s_1,s_2,\cdots,s_{kL}\}$.

3.2 基于强类别特征的相似度计算

强类别特征词项若是多义词,那么同时包含该词项并不一定使得这两个文本相似,原因是该词项在两个文本表示完全不同的含义. 因此本文设计一种新的策略利用强类别特征词在两个文本中的上下文相似度对词项进行语义消歧.

对于任意两个文本 d_1 和 d_2,词项满足特定条件的 t_i,即 $t_i\in s(t)=\{t_j|t_j\in d_1,t_j\in d_2,t_j\in S\}$. 词项 t_i 分别与文本 d_1 和 d_2 共现过的窗口内的词构成该词项的上下文,即为 $K_{d_1}(t_i)$ 和 $K_{d_1}(t_i)$,定义如下:

$$K_{d_1}(t_i)=\{t_j|t_j\in d_1,\mathrm{dis}_{d_1}(t_i,t_j)\leqslant\partial\};\quad(18)$$

$$K_{d_2}(t_i)=\{t_j|t_j\in d_2,\mathrm{dis}_{d_2}(t_i,t_j)\leqslant\partial\}$$

其中$\mathrm{dist}_{d_1}(t_i,t_j)$为词项 t_i 与 t_j 在短文本 d_1 中的相隔的词数,∂为一个控制窗口大小的阈值,本文中∂的值设置为2,表示与词项 t_i 共现且前后间隔词数为 2 的词组成词项 t_i 的上下文. 接下来通过计算上下文的相似度来确定强类别特征词 t_i 是否有歧义. 计算公式如下:

$$S_{\mathrm{context}}(t_i)=\frac{\sum_{t_1\in K_{d_1}(t_i)}\sum_{t_2\in K_{d_2}(t_i)}\mathrm{CR}(t_1,t_2)}{|K_{d_1}(t_i)|\times|K_{d_2}(t_i)|}\quad(19)$$

利用强类别特征词 t_i 的上下文相似度来得到一个指示函数 $I(t_i)$ 来表示词项 t_i 是否表征同一个含义.

$$I(t_i)=\begin{cases}1, & S_{\mathrm{context}}(t_i)\geqslant 0.5\\ 0, & \text{otherwise}\end{cases}\quad(20)$$

此外综合考虑词项在文本的关联权重和在类中的排名情况,重新定义强类别特征词项 t_i 在 d_1 的权重为:

$$w'_{d_1}(t_i)=w_{d_1}(t_i)\times w_{C_r}(t_i)\quad(21)$$

同理可得到强类别特征词项 t_i 在文本 d_2 的权重 $w'_{d_2}(t_i)$. $w_{d_1}(t_i)$ 可由式(5)计算得到,而 $w_{C_r}(t_i)$ 可由式(17)得到. 因此利用两个文本包含相似含义的强类别特征的情况来计算两个文本的强类别特征相似度(Strong Classification Features similarity,SCF). 计算公式如下:

$$S_{\mathrm{SCF}}(d_1,d_2)=\sum_{r=1}^{k}\sum_{t_i\in C_r}(\min(w'_{d_1}(t_i),w'_{d_2}(t_i))\times I(t_i))\quad(22)$$

利用 $I(t_i)$ 可以对强类别特征词 t_i 语义消歧. $S_{\mathrm{CF}}(d_1,d_2)$ 越大,d_1 和 d_2 包含同一类别的强类别特征词越多,则属于同一类的概率越大. 定义归一化的强类别特征相似度为:

$$S'_{\mathrm{SCF}}(d_1,d_2)=\frac{S_{\mathrm{SCF}}(d_1,d_2)}{\sum_{r=1}^{k}\sum_{t_i\in C_r}(\max(w_{d_1}(t_i),w_{d_2}(t_i))\times I(t_i))}\quad(23)$$

3.3 融合耦合距离相似度和强类别特征的相似度计算

综合考虑文本间耦合距离区分度和强类别特征的相似度方法,最终设计了融合耦合距离相似度和强类别特征的相似度计算方法(combining Coupled Distance-Discrimination and strong Classification Features for short text similarity calculation,CDDCF).

$$S_{\mathrm{CR\text{-}CF}}(d_1,d_2)=\beta\times S_{\mathrm{CR}}(d_1,d_2)+(1-\beta)\times S'_{\mathrm{SCF}}(d_1,d_2)\quad(24)$$

其中β为偏好因子,介于[0,1]之间,用来调节两种不同的相似度计算方法,该相似度即考虑词项的耦合距离共现关系,又考虑到带类别标签信息的文本间的相似度,更能够体现文本间隐含的关系,使相似度值更加精确. 而当β取值分别为0和1时则分别退化为SCF算法和CR算法.

4 实验与性能分析

4.1 实验数据分析

本文在接下来的实验中使用三个数据集,包括DBLP[8]数据集,搜狗语料库数据集[9]和20Newsgroups[10]数据集来验证本文方法的有效性. 其中DBLP是计算机领域内英文文献的集成数据库系统,通过爬取了DBLP中数据挖掘,人工智能,自然语言处理等10个领域的文章标题作为实验数据,每个领域包含1000篇文章标题. 搜狗语料库来源于Sohu新闻网站保存的经过手工整理与分类的新闻语料,符合短文本场景本文只抽取新闻标题作为实验数据,搜狗语料库包含汽车,财经,IT等11个类别,每个类别选取1000篇新闻标题作为实验数据. 而20newsgroups数据集是用于文本领域研究的国际标准数据集之一,数据集收集了大约20,000左右的新闻组文档,均匀分为20个不同的主题,本文主要聚焦于短文本,因此实验只选取文章标题作为实验数据.

4.2 评价指标

在文本聚类中,相似度的计算的优劣决定着聚类算法的性能. 本文采用划分聚类中的 k-means 聚类算法. 使用 k-means 算法来对文本聚类时,通过观察聚类结果来衡量相似度计算的效果,实验中 k 值分别被设置为三个数据集类别的个数(即DBLP为10,Sougou为11,20newsgroups为20). 本文将采用以下两个经典的聚类指标来评价算法的性能:兰德指数(Rand Index,RI)和F值(F-measure).

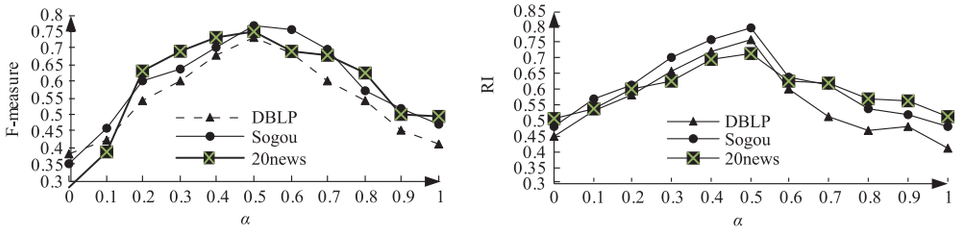
4.3 实验结果与分析

为了验证本文方法的有效性设计了两个实验. 一是对本文中的三个重要参数 α,L,β 的分析;二是比较本文提出的三种方法的聚类性能和比较本文的方法和已存在的相似度计算方法的聚类性能的对比.

4.3.1 参数对算法的影响

在本节中,通过一系列实验来分析参数 α,L,β 对算法性能的影响. 其中参数 α 用于调节耦合距离区分度中的内联和外联关系的相对重要性,L 用来调节每个类中强类别特征的个数,β 用来调节耦合距离相似度和强类别特征相似度间的相对重要性.

参数 α 的取值以 0.1 为步长从[0,1]之间改变 α 的取值,分析 CR 在不同数据集上两种评价指标的变化趋势. 通过图 1 观察到,即随着 α 取值逐渐增大,RI 和 F-measure 值也随着递增且当 $\alpha=0.5$ 时达到峰值,之后随着 α 的增大,RI 和 F-measure 的值减小. 这是因为随着 α 的增大,外联关系可以提升聚类性能,也就意味这外联关系对内联关系有促进作用,外联关系考虑到了词项即使不共现也可以通过外联路径产生关联性,当 $\alpha>0.5$ 时,外联关系会影响聚类性能的提升. 因此在接下来的实验中本文统一选取 $\alpha=0.5$ 作为最优的实验参数.



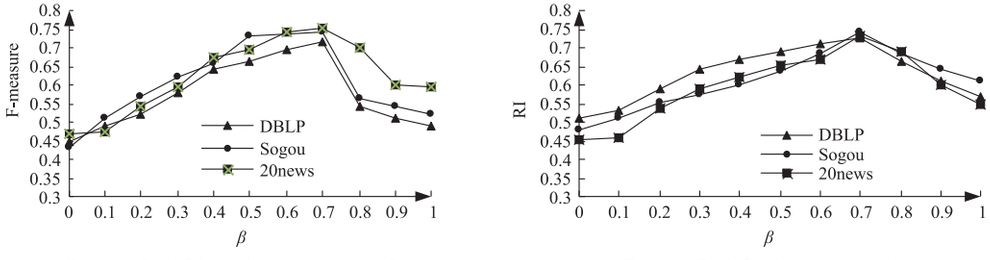
(*a*) 使用*k*-means聚类后参数α变化对F-measure的影响　(*b*) 使用*k*-means聚类后参数α变化对RI的影响

图1　α改变在不同数据集上对聚类性能的影响

表 1　L 改变对聚类性能的影响

	$L=50$	$L=100$	$L=150$	$L=200$	$L=250$	$L=300$	$L=350$
DBLP	0.35	0.425	0.512	0.627	0.575	0.53	0.495
Sogou	0.39	0.477	0.553	0.696	0.599	0.573	0.532
20news	0.412	0.459	0.586	0.692	0.654	0.543	0.52

通过调节 L 的取值来观察对 SCF 方法的聚类性能的影响,使用 k-means 聚类选取使得 F-measure 最高的 L 值. 通过表 1 可以看到随着 L 值增加,F-measure 和 RI 也随着增加,当 L 等于 200 时达到峰值,紧接着随着 L 的增加 F-measure 反而变小,最后趋于稳定. 分析原因是 L 的取值过小,会导致该类中强类别特征不足代表该类的类别信息,而 L 值过大则会导致一些不太重要的词项作为强类别特征来看待,使得每个类的类别信息含有噪声,导致相似度计算结果不精确. 通过实验结果的分析,最后选取 $L=200$ 为实验最优参数.



(*a*) 使用*k*-means聚类后参数β变化对F-measure的影响　(*b*) 使用*k*-means聚类后参数β变化对RI的影响

图2　β改变在不同数据集上对聚类性能的影响

β 的取值以 0.1 为步长在[0,1]之间逐渐递增,且根据前面的实验参数 α 选择为 0.5,L 的取值选为 200. 实验结果如图 2 所示,当 β 逐渐递增时,RI 和 F-measure 值随之递增,当 $\beta=0.7$ 时,CDDCF 方法的 RI 和 F-measure 达到峰值. 这是因为耦合距离相似度对于整个相似度的计算更重要,原因在于 CR 方法考虑到词项间的更全面的耦合关系,使得计算结果更加精确. 当 $\beta>0.7$ 时,之后随着 β 的递增,RI 和 F-measure 值反而减小,且当 $\beta=1$ 时,CDDCF 方法退化为耦合距离区分度的相似度.

4.3.2 聚类性能的评估

实验通过设置本文方法(CDDCF)与三个基准方法的实验结果对比来验证本文方法的有效性. 三种方法为:融合共现距离和区分度的短文本相似度计算方法(CDPC)[11],耦合词项关系模型(CRM)[4]和强类别近邻传播聚类算法(SCFAP)[12]和词分布表示. 实验结果如表 2 所示.

观察表 2,可以看到本文的方法在聚类性能上优于其他四种基准方法. 分析实验结果,CDPC 方法仅仅利用了词项间的共现和距离关系,因此缺少了词项间的更丰富的语义关系导致性能最差. 而 CRM 不仅考虑到了词项的共现关系(内联关系),而且考虑到了外联关系,但其忽略了词项间的高阶语义关系. SCFAP 方法的性能最差,这是因为 SCFAP 对于文本的语境和词项间的关系没有考虑到,因此 SCFAP 聚类性能劣与 CDPC. 而本文的 CDDCF 方法不仅考虑到了词项的类别信息,

而且将词项的内联关系和外联关系都考虑到了.

表 2 不同相似度算法在聚类性能上的性能对比

Methods	DBLP		Sogou		20newsgroups	
	RI	F-measure	RI	F-measure	RI	F-measure
SCFAP	0.478	0.378	0.358	0.325	0.466	0.421
CDPC	0.502	0.482	0.542	0.512	0.497	0.43
CRM	0.534	0.512	0.425	0.394	0.568	0.501
CDDCF	0.724	0.665	0.759	0.732	0.714	0.688

5 结束语

针对短文本特征稀疏,本文提出了一种融合耦合距离区分度和强类别特征的短文本相似度计算方法. 首先计算考虑耦合词项的内联和外联关系得到文本的相似度计算方法,然后提取带类别信息的文本的强类别特征和语义消歧来计算文本强类别特征相似度,最后本文结合两种方法得到最终的相似度计算方法. 在今后研究中,考虑将本文的方法与深度学习相结合,得到语义信息和语境信息更丰富的文本相似度的计算方法.

参考文献

[1] 曹玖新,陈高君,吴江林,等. 基于多维特征分析的社交网络意见领袖挖掘[J]. 电子学报,2016,44(4):898 - 905.
Cao J X,Chen G J,Wu J L,et. al. Multi feature based opinion leader mining in social networks. [J]. Acta Electronica Sinica,2016,44(4):898 - 905. (in Chinese)

[2] Song S,Zhu H,Chen L. Probabilistic correlation-based similarity measure on text records[J]. Information Sciences,2014,289(1):8 - 24.

[3] Li P,Wang H,Zhu K Q,et al. Alarge probabilistic semantic network based approach to compute term similarity[J]. IEEE Transactions on Knowledge & Data Engineering,2015,27(10):2604 - 2617.

[4] Chen Q,Hu L,Xu J,et al. Document similarity analysis via involving both explicit and implicit semantic couplings [A]. IEEE International Conference on Data Science and Advanced Analytics[C]. Montreal,Canada:IEEE Computer Society,2016. 1 - 10.

[5] Cheng X,Miao D,Wang C,et al. Coupled term-term relation analysis for document clustering[A]. International Joint Conference on Neural Networks[C]. Dallas,Texas,USA:IEEE Computer Society,2013. 1 - 8.

[6] Zhang L,Gao Y,Hong C,et al. Feature correlation hypergraph:exploiting high-order potentials for multimodal recognition[J]. IEEE Transactions on Cybernetics,2014,44(8):1408 - 1419.

[7] Ma H,Xing Y,Wang S,et al. Leveraging term co-occurrence distance and strong classification features for short text feature extraction[A]. International Conference on Knowledge Science,Engineering and Management[C]. Australia:Springer,2017. 67 - 75.

[8] Michael Ley,DBLP Dataset[EB/OL]. http://dblp.uni-trier.de/xml/,2016-04-20.

[9] 搜狗实验室. 文本分类语料库[EB/OL]. http://www.sogou.com/labs/dl/c.html,2012-04-30/2012-09-01.

[10] Ken Lang,20 Newsgroups Dataset[EB/OL]. http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.data.html,2009-12-9.

[11] 刘文,马慧芳,脱婷,等. 融合共现距离和区分度的短文本相似度计算方法[J]. 计算机工程与科学,2017,29(3):52 - 53.
Liu W,Ma H,Tuo T,Chen H B. Co-occurrence distance and discrimination based similarity measure on short Text [J]. Computer Engineering and Science,2018,40(7):1281 - 1286. (in Chinese)

[12] Wen H,Xiao N. A semi-supervised text clustering based on strong classification features affinity propagation[J]. Pattern Recognition and Artificial Intelligence,2014,27(7):646 - 654.

作者简介

马慧芳(通信作者) 女,1981 年 7 月出生,甘肃兰州人. 博士,硕士生导师,现为西北师范大学计算机科学与工程学院教授. 研究领域为数据挖掘与机器学习.
E-mail:mahuifang@yeah.net

刘 文 男,1994 年 8 月出生,甘肃定西人. 西北师范大学计算机科学与工程学院硕士研究生. 研究方向为机器学习.
E-mail:nwnuliuw@yeah.net

李志欣 男,1971 年 10 月出生,广西桂林人. 博士,博士生导师,现为广西师范大学计算机科学与信息工程学院教授. 研究方向为图像理解、机器学习.
E-mail:lizx@gxnu.edu.cn

蔺想红 男,1976 年 1 月出生,甘肃天水人. 2009 年获哈尔滨工业大学计算机应用技术专业博士学位,现任西北师范大学计算机科学与工程学院教授,硕士生导师. 研究方向为神经网络与深度学习、大数据分析.
E-mail:linxh@nwnu.edu.cn