第 14 卷 第 28 期 2014 年 10 月
1671—1815(2014)28-0081-05

科 学 技 术 与 工 程
Science Technology and Engineering

Vol. 14 No. 28 Oct. 2014
ⓒ 2014 Sci. Tech. Engrg.

一种混合型的句子语义相似度计算方法

翟继友

(南京工程学院计算机工程学院,南京 211167;河海大学计算机与信息学院,南京 210098)

摘 要 句子相似度计算是智能答疑等应用领域的关键技术之一,当前相似度计算方法大都对语义因素考虑不全,使得计算的结果准确度差,为此提出了一种混合型句子语义相似度计算方法。首先,通过语义资源得到词间语义相似度;其次,以词之间的相似度为基础,结合词的权重和句型的相似度得到了混合的句子间的语义相似度,避免了单一角度度量相似度带来的偏差;最后,通过对比实验验证了本文提出的混合型计算方法相对传统方法获得了更好的测试结果。

关键词 智能答疑 句子相似度 语义

中图法分类号 TP391.1; **文献标志码** A

语句相似度计算是对语句之间的相似性的一种度量方法[1],在自然语言处理、智能检索[2]和机器翻译[3]等领域都有广泛的应用,也是智能答疑系统中的一个关键技术[4],在答疑过程中一方面是把用户提出的问题跟系统中的问题库进行相似度计算得到匹配,另一方面是利用问题和答案库的候选答案之间的相似度来返回结果。

计算句子相似度的算法有很多,总体来说可以分为下面两大类:一类是基于表层特征的相似度计算方法,比如基于词统计特征的方法、字面匹配方法等;另一类是基于语义层的相似度计算方法,比如语义分析、语义扩展等[5]。

基于词统计特征的方法一般通过考虑词频、词性等因素来度量句子相似度,如基于向量空间的词频逆向文档频率(term frequency-inverse document frequency,TF-IDF)方法,把句子表示成 TF-IDF 向量的形式,通过计算两向量的余弦值得到相似度[6]。该方法只考虑句子的表层信息,基本上没有使用语义知识、句子结构等深入特征,致使效果难以提升。

基于词语义特征的方法一般依赖于知网(HowNet)及同义词词林等词汇语义词典[7]。该方法考虑了词的语义特征,可以将表面信息不同而语义信息相同的词进行关联,但该方法依赖语义词典的完整性,词典不全将直接影响准确性,存在一定的局限性。李彬等[8]提出的基于句子结构特征的相似度计算方法,考虑了句式结构,但语义依存信息考虑不足,最终影响句子相似度的准确性。

上述方法对因素的综合考虑方面存在不足。近年来的研究表明,综合考虑多种特征来计算相似性的效果要好于只基于某一特征的方法[9],因此,现提出一种综合考虑词义、词在句子中的权重以及句式结构等因素的混合型语义相关性计算方法。

1 基于向量空间模型的相似度计算

基于向量空间的 TF-IDF 模型目前广泛应用在答疑系统和信息检索等领域,主要用于计算文本之间的相似度。该方法相似度计算可描述为:假设有 n 个索引项,对于查询 Q 与文档 D 计算其中每个词的 TF-IDF 值,可以得到对应的向量分别为 $Q=(Q_1, Q_2,\cdots,Q_n)$ 和 $D=(D_1, D_2,\cdots,D_n)$,其中 $Q_i(1\leqslant i\leqslant n)$ 表示第 i 个索引项在查询语句 Q 中的权重,$D_i(1\leqslant i\leqslant n)$ 表示第 i 个索引项在文档 D 中的权重,通过计算这两个向量的夹角的余弦值即得到句子间的相似度,如式(1)所示。

$$\mathrm{Sim}(Q,D)=\frac{\sum_i (Q_i D_i)}{\sqrt{\sum_i Q_i^2 \sum_i D_i^2}} \tag{1}$$

TF-IDF 算法考虑词在问句中的出现频率和在整个答案库中对不同句子的区分能力,在信息检索等领域能产生好的效果,但 TF-IDF 是具有统计特性的方法,一般只涉及词统计方面的特性,对词本身的语义信息欠缺考虑,比如,“马铃薯什么味道?”和“土豆什么味道”的意思相同,因为“马铃薯”和“土豆”语义相同,TF-IDF 算法不考虑此类语义信息,具有局限性[6]。为此,现引入基于语义的相似度计算方法。

2014 年 5 月 17 日收到 南京工程学院青年基金(QKJB201211)资助
作者简介:翟继友(1978—),男,博士研究生。研究方向:信息检索,软件工程。E-mail:jiyou1018@126.com。

2 混合型句子语义相似度计算

2.1 词之间语义相似度计算

在计算词语之间相似度之前先计算词间的语义距离,语义的相似度计算还需要语义知识资源作为支持,现采用知网(hownet)作为语义知识资源,知网中存在 11 棵义原树,但有些义原树里面的义原没有父子关系,不体现词之间的上下位关系,无法使用,因此采用秦兵等[10]从知网义原树中选取的 Entitv、Attribute 等 6 棵义原树来计算词之间的语义距离,把语义距离定义为两个词对应的义原在义原树中的最短距离,并约定如果两个词中有一个词的义原无法在六棵义原树中找到,或者说两个词的义原分别处在不同的义原树上,则用∞表示这两个词之间的语义距离,形式化表示方式如下。

设 d 表示词 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 之间的语义距离,那么 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的相似度可以用式(2)表示。

$$\begin{cases} \mathrm{WordSim}(\boldsymbol{A},\boldsymbol{B}) = 0 & (d = \infty) \\ \max - \dfrac{d(\max - \min)}{H}, & (d \neq \infty) \end{cases} \tag{2}$$

式(2)中,min 和 max 分别表示两个词之间相似度可能取得的最小值和最大值。在实验阶段令 min = 0, max = 1。H 表示 $\boldsymbol{A}$、$\boldsymbol{B}$ 所在的义原树的中两个义原的语义距离可能取得的最大值。也就是说,某义原树中深度最大的两个义原深度分别为 H_1、H_1,那么这棵语义树的语义距离 $H = H_1 + H_1$。当 d ≠ ∞ 时,$\boldsymbol{A}$、$\boldsymbol{B}$ 的义原必定在同一棵义原树上,所以 H 的定义合理。

2.2 加权的语句之间语义相似度计算

有词间相似度作为基础,下面构建句子间语义相似度计算模型。为使得算出的语义相似度更加合理,在算法中加入了词语的权重,权值由词语在库中出现的频次计算得来。

设 $\boldsymbol{C}$ 和 $\boldsymbol{D}$ 两个向量代表两个句子。

$\boldsymbol{C} = (C_1, C_2, C_3, \cdots, C_m)$;

$\boldsymbol{D} = (D_1, D_2, D_3, \cdots, D_n)$。

则可以构造词 $C_i(1 \leqslant i \leqslant m)$ 和 $D_j(1 \leqslant j \leqslant n)$ 之间的相似度矩阵 $\boldsymbol{M}(\boldsymbol{C},\boldsymbol{D})$ 为

$$M(\boldsymbol{C},\boldsymbol{D}) = \begin{pmatrix} C_1D_1 & \cdots & C_1D_n \\ \vdots & \ddots & \vdots \\ C_mD_1 & \cdots & C_mD_n \end{pmatrix}。$$

式中,C_mD_n 表示词 $C_i(1 \leqslant i \leqslant m)$ 和词 $D_j(1 \leqslant j \leqslant n)$ 的语义相似度;w_i 分别表示词语 C_i 的权重,将上述矩阵转换为一维,则可以得到 $\boldsymbol{C},\boldsymbol{D}$ 两个句子之间的加权语义相似度。

$$\mathrm{SenSim}(\boldsymbol{C},\boldsymbol{D}) = \frac{\sum_i (w_i \times \max C_iD_j, j \in [1,n])}{\sum_i w_i} \tag{3}$$

用这种方式计算出的语句之间的语义相似度的准确性对词之间的语义相似度存在一定依赖性。

2.3 结合句式结构的混合语义相似度计算

考虑到问句除了由词构成以外,还具有一定的句式结构特征。而问句类型也能反映问句的结构特征,因此可以作为问句相似度的判断依据。问句间的结构相似度用式(4)表示。

$$\mathrm{SrucSim}(\boldsymbol{E},\boldsymbol{F}) = \begin{cases} 0, & E,F \text{ 问句类型不同} \\ 1, & E,F \text{ 问句类型相同} \end{cases} \tag{4}$$

式(4)中,$\boldsymbol{E}$、$\boldsymbol{F}$ 表示两个问句,上述的三种相似度计算方法是从不同侧面来描述语义相似性,具有不同的重要性。因此,将这三种相似度计算进行带权线性组合,得到一种混合相似度计算算法,见式(5)。

$$\mathrm{MixSim}(\boldsymbol{Q},\boldsymbol{D}) = \alpha\mathrm{WordSim}(\boldsymbol{A},\boldsymbol{B}) + \beta\mathrm{SenSim}(\boldsymbol{C},\boldsymbol{D}) + \lambda\mathrm{SrucSim}(\boldsymbol{E},\boldsymbol{F}) \tag{5}$$

式(5)中,$\mathrm{MixSim}(\boldsymbol{Q},\boldsymbol{D})$ 表示问题和文档或答案集的相似度,三个不同参数 $\alpha > 0$、$\beta > 0$、$\lambda > 0$ 满足约束条件 $\alpha + \beta + \lambda = 1$,用于调节词间语义相似度、句子间语义相似度和问句类型相似度所占的权重。实际设置时,由于问句类型相似度语义信息相对单一,一般 λ 的值设置得相对小一些。至此,混合相似度计算方法考虑到了问题和答案集的多样性,从不同角度来评估语义相似度,这种结合多种特征的计算方法从理论上来讲能避免单一角度判别相似度带来的偏差,能更全面、更准确地评估语义相似度。

3 实验与结果分析

3.1 实验准备

目前还没有关于中文句子相似度计算的公共测试集,在基于实例的答疑系统应用背景下,实验语料数据来自某公司开发的税务答疑系统,收集了该系统 2012 年 9 月到 2013 年 9 月期间的所有问题,并进行了分类。在实验之初,首先确定式(5)中的各参数的合理取值,实验选取答疑系统中前 100 条非重复的待回答问题。

实验过程首先在同一输入的情况下分别计算 TF-IDF 和本文算法的值,进行结果对比分析;接着为进一步评估本文算法,引入召回率和准确率评价指标,进行关键词检索和语义检索的查准率和查全率进行比较。准确率(Precision)是指检索出的相似

文档总数与检索出的总文档数的比率;召回率(Recall)是指检索出的相似文档总数和系统中存在相似文档的总数的比率[11],相应的计算公式定义如下。

$$Precision = \frac{G_1 \cap G_2}{G_1} \times 100\%;$$

$$Recall = \frac{G_1 \cap G_2}{G_2} \times 100\%。$$

式中,G_1 表示检索出的文档总数,G_2 表示系统中存在相似文档的总数,$G_1 \cap G_2$ 表示检索出的相似文档总数。召回率和准确率分别评价算法查找相似句子对的查全性和查准性,两者互相制约形成互补的关系,在准确率高的情况下很难达到高召回率。为了更好地对整体性能进行评价,这里用 F 值来反映召回率和准确率的综合性能。

$$F_{\sigma\text{-score}} = \frac{(\sigma^2 + 1)Precision \times Recall}{\sigma^2 Precision + Recall}。$$

考虑召回率和精准率同等重要,综合评价指标中的参数 σ 取 1,得到 F_1(下文中用 F 表示)的指标。

3.2　结果与分析

在答疑系统的应用背景下,对不同相似度计算方法进行比较,比较之前首先通过实验的方式确当参数的选择,这里使用 TREC 会议的 MRR(mean reciprocal rank)[12]方法,得到 MRR 值见表 1。

表 1　不同参数取值对应的 MRR 值

Table 1　Different parameter values corresponding MRR value

α	β	λ	MRR
1	0	0	0.361
0	1	0	0.198
0.40	0.40	0.20	0.357
0.50	0.35	0.15	0.439
0.55	0.35	0.10	0.482
0.60	0.25	0.15	0.415
0.65	0.25	0.10	0.391

表 1 给出了混合语义算法性能随着不同参数组合的变化规律。当取值为 $\alpha = 1, \beta = 0, \lambda = 0$ 时,表示只采用词间语义相似度计算方法;$\alpha = 0, \beta = 1, \lambda = 0$,表示只采用加权的句子相似度计算方法;问句类型相似度只用来判断结构是否相似,不单独进行相似度评估。从表 1 的结果可以看出,混合相似度计算模型中的三种相似度的重要程度不同,词间相似度最重要,语句相似度次之,结构相似度最低,这与常理基本吻合。结果同时表明采用多种相似度融合的方法能更全面的评估相似度。根据表 1 中的取值结果,下面的对比实验中参数设定为 $\alpha = 0.55$, $\beta = 0.35, \lambda = 0.1$。

实验 1:基于词频的算法和本文算法对比

答疑系统中的问句库中有如下两个语句:语句 1 为“个人所得税有哪些免税规定?”和语句 2 为“私售发票有什么处罚措施?”。分别计算这两个句子跟用户输入的问句“个税有什么优惠措施?”之间的相似度,结果如表 2。

表 2　两种相似度计算结果比较/%

Table 2　Comparison of two kinds of similarity calculation results/%

采用方法	问句与语句 1 相似度	问句与语句 2 相似度
TF-IDF 方法	13.692	31.658
本文方法	90.836	6.157

从表 2 的计算结果可以看出,语句 1 和问句之间相同的词很少,利用 TF-IDF 方法计算得出的相似度仅为 13.692%。依据常理,这两句问句在语义上是近乎相同,计算结果就不符合应用的要求了。对比利用 TF-IDF 方法算出的两个值,语句 2 和问句之间的相似度大于语句 1 和问句之间的相似度,明显是错误的。利用本文算法得到语句 1 和问句之间的相似度为 90.836%,说明问句和语句 1 这两个句子在语义上非常接近,而问句和语句 2 之间的语义相似度很小,跟实际情况相吻合。

通过本实验可以看出,本文的混合语义相似度计算方法在处理两个问句中词的语义几乎相同但字面不相同的情况下,明显比 TF-IDF 方法效果好。

实验 2:算法的召回率、准确率等指标对比

第二个实验对 4 种算法进行对比评估,包括前文提到的 TF-IDF 算法、改进的 TF-IDF 算法(new TF-IDF)[13]、包含潜在语义的 LSK 算法[14]及本文提出的算法,结果见表 3。

表 3　四种算法实验结果对比表/%

Table 3　The results of four algorithms comparison table/%

采用方法	Precision	Recall	F
TF-IDF 算法	76.26	81.31	78.70
new TF-IDF	81.47	86.95	84.12
LSK 算法	70.75	93.23	80.45
本文算法	84.82	92.98	88.71

为直观展现几种算法在查全率、查准率和综合性能的表现,下面给出几种算法的对比图,如图 1 所示。

图 1 的对比结果可以做如下分析:TF-IDF 算法把句子中的词语看做相互独立的特征,忽略了句子序列关系、位置关系对句子语义的影响;改进后的

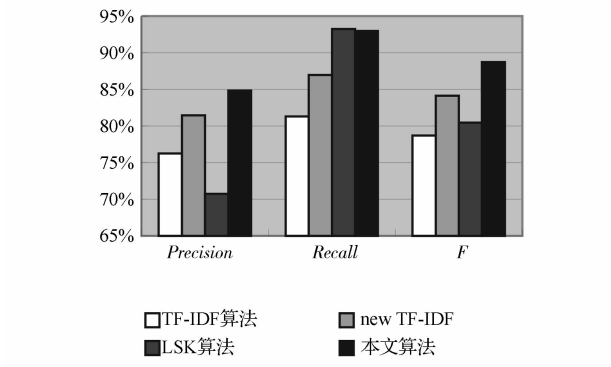


图 1 不同算法得到的结果对比

Fig. 1 The results obtained by comparing different algorithms

TF-IDF 比在准确率和召回率方面都有了小幅提高，说明进行同义词扩展也有利于句子之间相似度的比较，弥补了原来模型的不足；LSK 方法考虑到不同词间的潜在语义关系，召回率和综合指标明显提升，说明在相关类中语义特征起到主要作用，而词特征起到次要作用，但其精准率在 4 个算法中表现最差。从表 3 可以看出本文算法得出的结果大于其他三个算法在对应问题集的计算精准率，综合指标 F_1 达到 91.38%，比 TF-IDF 算法、改进的 TF-IDF 算法和 LSK 算法分别提高了 10.01%、4.59% 和 8.26%，由此表明本文方法的可行性。

4 结 语

首先对常用的度量相似度的 TF-IDF 算法进行分析，指出其在语义方面存在的欠缺，之后提出一种新的基于多特征、多角度的混合型相关性计算方法，包括了词间语义相似度、词在句子中的权重和句式结构相似度等，从不同角度来评估语义相似度，对句子的语义考虑更全面，避免只考虑单一特征带来的偏差，能更全面、更准确地评估语义相似度。通过不同算法实验对比结果可以看出本文提出的方法在相似性的综合性能方面比其他算法取得了更好的效果。下一步将借助机器学习的相关算法来优化参数的取值，更好地应用到在线答疑、在线互动等平台。

参 考 文 献

1 李 茹，王智强，李双红，等. 基于框架语义分析的汉语句子相似度计算. 计算机研究与发展，2013；50(8)：1728—1736
Li R，Wang Z Q，Li S H，*et al*. Chinese sentence similarity computing based on frame semantic parsing. Journal of Computer Research and Development，2013；50(8)：1728—1736

2 唐歆瑜，乐文忠，李志成，等. 基于知网语义相似度计算的特征降维方法研究. 科学技术与工程，2006；6(21)：3442—3446
Tang X Y，Li Z C，Le W Z，*et al*. The research on reduced feature dimension based on hownet similarity computing. Science Technology and Engineering，2006；6(21)：3442—3446

3 刘树杰，李志灏，李 沐，等. 一种面向统计机器翻译的协同权重训练方法. 软件学报，2012；23(12)：3101—3114
Liu S J，Li Z H，Li M，*et al*. Co-training framework for feature weight optimization of statistic machine translation. Journal of Software，2012；23(12)：3101—3114

4 黄 新，徐小娟. 基于 ontology 的智能答疑系统的研究. 科学技术与工程，2007；7(12)：3001—3003
Huang X，Xu X J. Study of automatic answering system based on ontology. Science Technology and Engineering，2007；7(12)：3001—3003

5 毛媛媛. 基于语义扩展的中文信息检索系统设计与实现. 成都：电子科技大学，2013
Mao Y Y. Based on the semantic extension in chinese information retrieval system design and implementation. Cheng Du：University of Electronic Science and Technology of China，2013

6 Goossen F，Jntema W，Frasincar F，*et al*. News personalization using the CF-IDF semantic recommender. Proc of the International Conference on Web Intelligence，Mining and Semantics，2011

7 郑 诚，李 清，刘福君. 改进的 VSM 算法及其在 FAQ 中的应用. 计算机工程，2012；38(17)：201—204
Li C，Li Q，Liu F J. Improved VSM algorithm and its application in FAQ. Computer Engineering，2012；38(17)：201—204

8 李 彬，刘 挺，秦 兵，等. 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
Li B，Liu T，Qin B，*et al*. Chinese sentence similarity computing based on semantic dependency relationship analysis. Application Research of Computers，2003；20(12)：15—17

9 赵 臻，吴 宁，宋盼盼. 基于多特征融合的句子语义相似度计算. 计算机工程，2012；38(1)：171—173
Zhao Z，Wu N，Song P P. Sentence semantic similarity calculation based on multi-feature fusion. Computer Engineering，2012；38(1)：171—173

10 秦 兵，刘 挺，王 洋，等. 基于常问问题集的中文问答系统研究. 哈尔滨工业大学学报，2003；35(10)：1179 — 1182
Qin B，Liu T，Wang Y，*et al*. Question answering system based on frequently asked questions. Journal of Harbin University of Civil Engineering and Architecture，2003；35(10)：1179 — 1182

11 李文超，杨妮妮. 基于本体的语义相似性研究. 科学技术与工程，2012；12(21)：5328—5330，5336
Li W C，Yang N N. Three-tier information fusion system research based on multi-agent technology. Science Technology and Engineering，2012；12(21)：5328—5330，5336

12 王 磊，黄广君. 结合概念语义空间的语义扩展技术研究. 计算机工程与应用，2012；48(35)：106—109，193
Wang L，Huang G J. Research of semantic query expansion related to concept semantic space. Computer Engineering and Applications，2012；48(35)：106—109，193

13 路永和，李焰锋. 改进 TF-IDF 算法的文本特征项权值计算方法. 图书情报工作，2013；57(3)：91—95
Lu Y H，Li X F. Improvement of text feature weighting method based on TF-IDF algorithm. Library and Information Service，2013；57(2)：91—95

14 Cristianini N，Shawe-Taylor J，Lodhi H. Latent semantic kernels. Journal of Intelligent Information Systems，2002；18(23)：127—152

A Hybrid Measurement for Sentence Similarity Based on Semantic

ZHAI Ji-you
(School of Computer Engineering, Nanjing Institute of Technology, Nanjing 211167, P. R. China;
College of Computer and Information, Hohai University, Nanjing 210098, P. R. China)

[**Abstract**] Sentence similarity computation is one of the key technologies of intelligent question answering etc. The current method to calculate the similarity are not completely in terms of semantics, resulting the lack of accuracy. Therefore a hybrid type of sentence semantic similarity calculation method is proposed. First of all, word semantic similarity are gotten by semantic resources, and then the semantic similarity between sentences are gotten by mixedsimilarity with the word similarity and word weighting and sentence patterns, avoiding bias caused by single factor. Finally, the contrast experiment shows the proposed algorithm has better results compared with the traditional method.

[**Key words**] intelligent question answering　　sentence similarity　　semantic

(上接第 80 页)

A Software System for Calculating the Resistance in Liquid Crystal Display Panel Design

ZHAO Chen-su[1,2], YU Wen-jian[2*]
(School of Information Engineering, ShandongYingcai University[1], Jinan 250104, P. R. China;
Department of Computer Science and Technology, Tsinghua University[2], Beijing 100084, P. R. China)

[**Abstract**] Aiming at the resistance calculation in liquid crystal display(LCD) panel design, an accurate, efficient and with friendly interface software system is proposed. The presented system developed by MATLAB development environment is an interactive user interface which can display the LCD Layout writing, calculate resistance, input and output data. Fast boundary element algorithm bas been used in resistance calculation. The algorithm has been implemented in C language. Experiments show that the presented system has advantages of accuracy, fast speed, friendly interactivity, scalability, and easy to use. At present, the presented system has been used in computer-aided design of high-performance LCD panel.

[**Key words**] liquid crystal display panel　　wiring resistance calculation　　computer aided design　　graphical user interface(GUI)　　integrated system