一种基于《知网》的文本语义相似度的计算方法*

孙 滨 刘 林
(郑州华信学院信息工程系 郑州 451100)

摘 要 论文提出一个基于语义的文本间的相似度算法,以文本的特征词相似度为基础,来计算文本间的相似度,利用聚类算法对文本簇进行聚类。实验结果证明基于知网的文本语义相似度方法在对文本相似度计算以及文本聚类方面,能有效提高聚类的效果。

关键词 文本聚类;义原相似度;语义相似度
中图分类号 TP391 **DOI**:10.3969/j.issn1672-9722.2014.02.003

A Method of Computing the Semantic Similarity of Sentences Based on HowNet

SUN Bin LIU Lin
(Department of Information Engineering, Zhengzhou Huaxin University, Zhengzhou 451100)

Abstract A similarity algorithm based on semantic similarity is proposed, which calculates the similarity of texts according to feature words of the text and makes text clusters by employing clustering algorithm. The experimental results prove that the method of text semantic similarity based on CNKI is very efficient in text similarity calculation and text clustering, which can effectively improve the effect of clustering.

Key Words text clustering, primitive similarity, semantic similarity
Class Number TP391

1 引言

词语相似度[1]指的是两个词语在不同的上下文中可以相互替换使用而不改变文本的句法语义结构的程度。二者相似度程度越高,可替换的可能性越大。义原相似度是词语相似度及基于语义的文本相似度的基础,义原相似度的精确度的提高,有利于提高文本聚类的效果。文本聚类在文本挖掘和信息检索发挥着重要作用,它得到许多人的研究,取得了不少成果,但是文本聚类也面临着各方面的挑战[2]:1) 非常高的数据维数:要求聚类算法能够处理稀疏矩阵,或者对矩阵降维。2) 数据库规模可能非常大(例如万维网):因此聚类算法对大型数据库也要有很高的分析效率。3) 可以理解的聚簇描述:聚簇描述引导用户浏览聚簇,因此,聚簇描述对于非专业用户也应该是可以理解的。

目前的文本表示方法和聚类方法[3~5]存在以下缺点:1) 文本的特征词的维数非常用高,并且文本常用 VSM 表示成向量空间,构成了稀疏矩阵,造成了文本向量的表示空间难以有效地降维。2) 由于不同的文本可能采用不同的词汇来表示相同概念,"一对一"的匹配方法在处理时就显得无能为力了[6]。特别是同义词和近义词不能识别,造成了聚类的误差。例如:文本 1:土豆盛产于中国。文本 2:地瓜在中国产量丰富。这里的土豆和地瓜是同义,而在以前文本表示中作为两个不同的词,如果按以前文本的计算方法,则它们的相似度较低。实际上它们有较高的相似性。

针对以上限制提出了基于《知网》的语义相似度来计算文本相似度—MMTS(Max Min Text

* **收稿日期**:2013 年 8 月 8 日,**修回日期**:2013 年 9 月 25 日
基金项目:河南省教育厅科学技术研究重点项目(编号:12B520063)资助。
作者简介:孙滨,男,硕士,讲师,研究方向:云计算与知识发现。刘林,女,助教,研究方向:语义 Web。

Similarity),通过计算文本的特征的相似度为基础,来计算文本间的相似度,减少了数据维数,有效地提高了文本之间的聚类效果。

2 词语的相似性计算

2.1 改进的义原相似度计算

义原相似度的计算是词语相似性的基础,它的数值的准确性影响词语相似性的准确性。刘群[1]在基于知网的义原的相似度计算采用式(1)

$$Sim(p_1,p_2)=\frac{a}{d+a} \tag{1}$$

其中 p_1,p_2表示两个义原,d 是 p_1,p_2在义原层次体系中的路径长度,a 是一个可调节的参数。此计算只是简单考虑了两个义原的距离长度,而忽视了义原在不同层次中的深度,导致了计算结果比较粗略。例如:动物和植物,哺乳动物和爬行动物,这两对概念间的路径长度都是2,但它们位于语义树的深度不同,动物和植物位于语义树上层,而哺乳动物和爬行动物位于较深层,因而哺乳动物和爬行动物的相似性值比动物和植物相似性值大、更相似。并且式(1)还需要跟据实际的需要调节 a 的取值,给义原相似度计算带来不稳定性,也会影响义原相似度的值。为克服上述缺点,在原有基于 HowNet 的义原相似度计算的基本上,本文采用了式(2)来计算义原间的相似度。在式(1)的基础上既考虑了义原间的距离又考虑到了义原间的深度,同时减少调节参数,进一步提高了义原相似度值的精确性[7~8]。

$$Sim(p_1,p_2)=\frac{|p_1\cap p_2|}{|p_1|+|p_2|-|p_1\cap p_2|} \tag{2}$$

其中$|p_1|$、$|p_2|$表示从根节点到 p_1、p_2经过的节点的个数(包括 p_1,p_2)$|p_1\cap p_2|$表示从根节点到 p_1、p_2时,两路径重叠的节点数。经过实验证明该方法能有效提高相似度计算的准确性。

2.2 词语间的相似度计算

刘群教授在文献[1]中提出两个词语间的相似度计算最终归结到两个概念之间的相似度计算[5]。词语 w_1有 n 个义项(概念):s_{11},s_{12},…,s_{1n},词语 w_2有 m 个义项(概念):s_{21},s_{22},…,s_{2n},则 w_1和 w_2的相似度等于各个要念的相似度之最大值,如式(3)所示

$$Sim(w_1,w_2)=\max_{i=1,\cdots,n,\ j=1,\cdots,m} Sim(s_{1i},s_{2j}) \tag{3}$$

两个义项(概念)语义相似度公式,如式(4)所示

$$Sim(s_1,s_2)=\sum_{i=1}^{4}\beta_i\prod_{j=i}^{i}Sim(s_1,s_2) \tag{4}$$

其中:$\beta_i(1\leq i\leq 4)$是可调查节的参数,且有 $\beta_1+\beta_2+\beta_3+\beta_4=1$,$\beta_1\geq\beta_2\geq\beta_3\geq\beta_4\geq 0$。$Sim_1(s_1,s_2)$是第一独立义原描述式,$Sim_2(s_1,s_2)$是其他独立义原描述式,$Sim_3(s_1,s_2)$是关系义原描述式,$Sim_4(s_1,s_2)$是符号义原描述式。

3 文本的相似度计算

3.1 文本的表示

通过特征词[9]提取算法,从文本中提取特征词,为克服现有的文本表示维数过多现象,本文采用以文本实际的特征词来表示文本,表示形式如式(5)~式(6)所示

$$A=\{(f_1,w_1),(f_2,w_2),\cdots,(f_n,w_n)\} \tag{5}$$

$$f_i=\frac{d_i}{\sqrt{\sum_{j=1}^{n}d_j^2}} \tag{6}$$

式中:w_i表示在文本 A 中出现的特征词,f_i表示 w_i在文本中出现的频率,并采用式(6)对其进行归一化,d_i表示 w_i在文本中出现的次数。通过式(5)把文本表示成结构化的列表。

3.2 改进文本间的相似度计算(MMTS)

基于语义的文本相似性[10]就通过计算不同文本特征词的相似性,来计算文本间主题或内容的接近程度。本文将文本表示成(5)的向量形式。克服了字面上差异,而是从字意上来理解文章。计算两文本的公式如式(7)所示。

$$Sim(A_i,A_u)=\frac{\sum_{j=1}^{n}\max(sim(w_{ij},A_u))*\min(f_{ij},f_{ut})+\sum_{q=1}^{n}\max(sim(A_i,w_{up}))*\min(f_{iv},f_{uq})}{m+n} \tag{7}$$

式中:$\max(sim(w_{ij},A_u)$表示计算文本 A_i的 w_{ij}特征词与 A_u文本向量特征词相似度的最大值,f_{ut}表示在 A_u文本中使得 $\max(sim(w_{ij},A_u)$取最大值时的特征词的频率,$\min(f_{ij},f_{ut})$表示 f_{ij},f_{ut}的最小值。

算法思想是:首先计算文本的每个特征词与另一篇文本的所有特征词的相似度,取出最大值,同时该值乘以在最大值处的两个特征词的最小频率值,得出该特征词的贡献度,依次求出每一个特征的贡献度,最后求出贡献度的平均值即两个文本的

相似度。此算法考虑到由于每个特征词都对文本的组成部分，对文本的相似度都有自己的贡献度，并且取两个特征词的最小频率，避免了频率较大特征词的影响，使相似度更精确。

文本的相似度算法描述如下：

输入：A_i，A 文本向量

输出：文本相似度 $Sim(A_i,A_u)$

```
Begin
参数：total,sim,max,c;//计算：
∑_{j=1}^{n} max(sim(w_ij,A_u)) * min(f_ij,f_ut)
For(取 A_i 中的每个特征词){
    For(取 A_u 中的每个特征词){
        if(两特征词词性是否相同)
    {
    Sim=sim(w_ij,w_ux);
          If(Sim<max)
          {
              Sim=max;
              c=x
      }
    }
    }
    total=total+sim * min(f_ij,f_ux);
}
//计算：∑_{q=1}^{n} max(sim(A_i,w_up)) * min(f_it,f_uq)
For(取 A_u 中的每个特征词){
    For(取 A_i 中的每个特征词){
        if(两特征词词性是否相同)
    {
          Sim=sim(A_i,w_up);
        If(Sim<max)
        {
            Sim=max;
            c=x;
    }
    }
    }
    total=total+sim * min(f_ic,f_uq);
}
retrun total/(A_i 和 A_u 特征词的个数之和);
end
```

4　实验结果与分析

4.1　实验环境

本实验的编程工具是 Vistual Studio 2008，编程语言是 C#。本实验的程序是原有的基于《知网》词语相度计算的程序上，依据本文提出的义原，文本相似度计算理论上实现的编写而成的。语料库选自搜狗 2008 版分类语料库，从其中选取：汽车，财经，IT，健康，体育，旅游，教育，文化，军事的语料各 50 篇，共计 450 篇作为测试。

4.2　文本聚类应用

本实验的目的考察本文采用的文本表示模型、基于语义的文本间相似度计算的方法在信息检索与数据挖掘常用的文档聚类中的精度。分别将文本用 VSM 模型和本文的文本表示模型进行表示，采用了基础的聚类算法：k-means，DB-SCAN，层次聚合算法。文本间相似度分别采用式(8)和式(7)进行比较。$k=4,6,8,12$，表中的正确率结果为它们的平均值。

$$Sim(A_i,A_u)=\frac{\sum_{k=1}^{n} f_{ik}\cdot f_{uk}}{\sqrt{(\sum_{k=1}^{n} f_{ik}^2)\cdot(\sum_{k=1}^{n} f_{uk}^2)}} \tag{8}$$

4.3　结果评价与分析

表 1　三种聚类方法比较表

	VSM 模型	本文文本表示模型
k-means	0.453	0.685
DB-SCAN	0.526	0.621
层次聚合算法	0.652	0.864

通过表 1 的数据分析可知，采用本文文本表示的模型，在实现文本聚类方面有显著的提高，聚类效果明显。采用 VSM 模型表示的文本在聚类效果上明显不如本文中提出的文本的表示模型，这主要是由于本文文本表示模型通过特征词提取算法，从文本中提取特征词，克服现有的文本表示维数过多现象，采用以文本实际的特征词来表示文本。

5　结语

本文改进了基于《知网》的义原相似度的算法，提高词语的相度值的精确度。克服了 VSM 文档表示模型数据维数过高和聚族难以描述的缺点。提出一种基于语义的文本相似度的计算方法—MMTS。实验结果证明，该算法提高了聚类的效果。

参 考 文 献

[1] 刘群，李素建. 基于《知网》的词汇语义相似度的计算[C]//第三届汉语词汇语义学研讨会. 中国台北，2002.

[2] Pandya A, Bhattacharyya P. Text similarity measurement using concept representation of texts[C]//Proceedings of First International Conference on Pattern Recognition and Machine Intelligence. Berlin, Germany: Springer, 2005: 678-689.

（下转第 209 页）

化[J]. 兵工自动化,2007,26(2):20-22.
ZHANG Huibin. Optimization of Number of Service Teams Based on Queuing Theory in Vehicles Service Safeguard[J]. Ordnance Industry Automation,2007,26(2):20-22.

[2] 巴威,等. 基于排队论的车辆装备维修力量需求预测研究[J]. 军事交通学院学报,2009,11(6):89-91.
BA Wei, et al. Study on Requirement Forecast of Vehicle Equipment Maintenance and Repairing Force Based on the Queuing Theory[J]. Journal of Academy of Military Transportation,2009,11(6):89-91.

[3] 周鑫磊,孟祥印,解学参,等. 基于排队论的大型舰船甲板弹射器数量研究[J]. 中国舰船研究,2011,6(2):15-18.
ZHOU Xinlei, MENG Xiangyin, XIE Xueshen, et al. Investigation into Deck Catapults Quantity of Large Ship by Using Queuing Theory[J]. China Journal of Ship Research,2011,6(2):15-18.

[4] 李如琦,苏浩益. 基于排队论的电动汽车充电设施优化配置[J]. 电力系统自动化,2011,35(14):58-60.
LI Ruqi, SU Haoyi. Optimal Allocation of Charging Facilities for Electric Vehicles Based on Queuing Theory[J]. Automation of Electric Power Systems,2011,35(14):58-60.

[5] 毛德耀,林广积,何舍炳,等. 基于排队论的舰船装备维修保障模型[J]. 兵工自动化,2012,31(6):35-37.
MAO Deyao, LIN Guangji, HE Shebing, et al. Warship Maintenance Support Model Based on Queuing Theory[J]. Ordnance Industry Automation, 2012, 31(6):35-37.

[6] 张波,于永利,徐英,等. 基于排队论的维修资源需求分析方法[J]. 军械工程学院学报,2011,23(4):9-12.
ZHANG Bo, YU Yongli, XU Ying, et al. Method to Analyze Maintenance Resources Requirement Based on Queuing Theory[J]. Journal of Ordnance Engineering College,2011,23(4):9-12.

[7] 甘应爱,等. 运筹学[M]. 北京:清华大学出版社,1995,6.
GAN Ying'ai, et al. Operation Research[M]. Beijing: Qinghua University Press,1995:6.

[8] 董肇君. 系统工程与运筹学[M]. 北京:国防工业出版社,2011,8.
DONG Zhaojun. Systems Engineering and Operation Research[M]. Beijing: National Defense Industry Press,2011:8.

[9] 李明,等. 基于排队论的战术装备维修力量分配问题研究[J]. 计算机与数字工程,2011,39(2):23-25.
LI Ming, et al. Assignment Model of Tactical Maintenance Force in Wartime Based on Queuing Theory[J]. Computer and Digital Engineering,2011,39(2):23-25.

[10] 肖慧鑫,等. 基于排队论的装备保障最优化[J]. 火力与指挥控制,2008,33(2):142-144.
XIAO Huixin, et al. Research on Equipment Support Optimization Based on Queuing Theory[J]. Fire Control and Command Control,2008,33(2):142-144.

(上接第 189 页)

[3] Rodriguez M A, Egenhofer M J. Determining Semantic Similarity Among Entity Classes from Different Ontologies[J]. IEEE Trans. on Knowledge and Data Engineering,2003,15(2):442-456.

[4] Budanitsky A, Hirst G. Evaluating Word Net-based Measures of Lexical Semantic Relatedness[J]. Computational Linguistics,2006,32(1):13-47.

[5] Giunhiglia F, Shvaiko P, Yatskevich M. Semantic Schema Matching[R]. Trento, Italy: University of Trento,2005.

[6] 王晓东,郭雷,方俊,等. 一种基于 EMD 的文本语义相似性度量[J]. 电子与信息学报,2008,30(9):2156-2161.
WANG Xiaodong, GUO Lei, FANG Jun, et al. An EMD-Based Metric for Document Semantic Similarity [J]. Journal of Electronics & Information Technology, 2008,30(9):2156-2161.

[7] 江敏,肖诗斌,王弘蔚,等. 一种改进的基于《知网》的词语语义相似度计算[J]. 中文信息学报,2008(5):265-267.
JIANG Min, XIAO Shibin, WANG Hongwei, et al. An Improved Word Similarity Computing Method Based on HowNet[J]. Journal of Chinese Information Processing,2008(5):265-267.

[8] 吴雅娟,陈尧,尚福华. 一种新的基于相似度计算的本体映射算法[J]. 计算机应用研究,2009(3):230-234.
WU Yajuan, CHEN Rao, SHANG Fuhua. New Similarity-Based Approach for Ontology Mapping[J]. Application Research of Computers,2009(3):230-234.

[9] 徐茜,彭进业,李展. 本体映射中一种综合的概念相似度计算方法[J]. 计算机工程与应用,2010(24):354-359.
XU Qian, PENG Jinye, LI Zhan. Integrated Concept Similarity Computing Method in Ontology Mapping [J]. Computer Engineering and Applications, 2010, 46(24):34-36.

[10] 高炜,梁立. 一种改进的基于相似度的本体映射方法[J]. 甘肃联合大学学报(自然科学版),2009(5):125-129.
GAO Wei, LIANG Li. Improved Ontology Mapping Method Base on Semantic Similarity[J]. Journal of Gansu Lianhe University(Natural Sciences),2009(5): 125-129.