Computer Science and Application 计算机科学与应用, 2016, 6(11), 755-760
Published Online November 2016 in Hans. http://www.hanspub.org/journal/csa
http://dx.doi.org/10.12677/csa.2016.611090

Hans汉斯

The Improvement of the Computational Method of Words' Similarity Based on Flexible Logic

Chengxia Liu[1,2]

[1]Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information and Technology University, Beijing
[2]Computer School, Beijing Information and Technology University, Beijing
Email: cecilia7812@163.com

Received: Nov. 12th, 2016; accepted: Nov. 26th, 2016; published: Nov. 29th, 2016

Abstract

In order to improve the text mining efficiency, the computational method of words' similarity is studied and analyzed in this article. Flexible logic is used to improve the adaptability of the words' similarity computational method. The flexible factor can be changed according to the different requirements but the computational method is consistent. This is the base and the further study will be carried out in future research.

Keywords

Text Mining, The Words' Similarity, Flexible Logic

基于柔性逻辑的词语相似度计算方法的改进

刘城霞[1,2]

[1]北京信息科技大学网络文化与数字传播北京市重点实验室，北京
[2]北京信息科技大学计算机学院，北京
Email: cecilia7812@163.com

收稿日期：2016年11月12日；录用日期：2016年11月26日；发布日期：2016年11月29日

文章引用: 刘城霞. 基于柔性逻辑的词语相似度计算方法的改进[J]. 计算机科学与应用, 2016, 6(11): 755-760.
http://dx.doi.org/10.12677/csa.2016.611090

摘　要

为了更好的进行文本挖掘，本文对词语间的相似度计算方法进行了研究及分析，并对其进行了柔性化改进，使之能适应不同的环境需求。具体为引入泛非柔性运算因子，在不同情况下使用不同的柔性参数，可以用统一的运算模型适应不同的应用需求，并对将来进一步的研究打下基础。

关键词

文本挖掘，词语相似度，柔性逻辑

1. 引言

在数据挖掘中除了针对数据库中数据的挖掘研究外还有对文本的挖掘研究。在互联网成为生活必不可少的组成部分的今天，网络信息充斥生活，而海量数据中哪些是有用的、需要的，而哪些又是垃圾信息，需要剔除的，如何能自动轻松的完成筛选？本文中研究的就是基于 Web 的文本挖掘。在进行 web 挖掘的时候，如何计算词语的相似度会影响到最终挖掘的效果。传统的基于知网的词语相似度算法有它的局限性，很多学者对其也进行了改进，比如文献[1] [2]中提到的。本文从柔性逻辑的角度改进该词语相似度算法，使之能根据需要自适应的改变。

2. 柔性逻辑基础

20 世纪以来，数十种适用于不同的背景的逻辑学被先后提出，用于满足各新兴学科的不同需要。为了满足智能应用的需要，在刚性逻辑中引入现实世界的柔性因子，去补充刚性的不足，本世纪初何华灿教授提出了一种新的柔性逻辑－泛逻辑学[3]。泛逻辑理论利用二值逻辑、多值逻辑和模糊逻辑的理论体系去研究人工智能领域中的不确定性、不完全性以及模糊性。其中它对命题的真值域、命题连接词、量词等都进行了柔性化[4] [5] [6]，可以全面反映命题真值的不确定性、真值误差的不确定性、命题之间相关关系的不确定性和相对权重的不确定性对逻辑推理的影响，使之更适合于现实世界的推理规则。目前对泛逻辑的研究已经建立并证明了各级泛运算的模型和算子并进行了证明和应用，形成了一套形式推理系统。在本文中只应用了泛逻辑中的泛非运算，下面就着重介绍一下泛非运算的逻辑公式。

2.1. 泛非命题连接词及其逻辑公式

在泛逻辑中，泛逻辑学中非运算(算子)以三角范数理论中的 N 范数作为其数学原型。在三角范数研究中很早就研究了模糊非算子，称为 N 范数，文献[7] [8] [9] [10] [11]对其进行了研究，得到的结果不尽相同。

1. N 范数的一般定义[3]

设一元运算$N(x)$是$[0,1]\to[0,1]$的，则关于$N(x)$有以下条件:

边界条件N1　$N(0)=1$，　$N(1)=0$；

单调性N2　$\forall x,y\in[0,1]$，若$x<y$，则$N(x)\geq N(y)$；

严格单调性N2’　$\forall x,y\in[0,1]$，若$x<y$，则$N(x)>N(y)$；

连续性N3　$\forall x\in[0,1]$，　$N(x^-)=N(x)=N(x^+)$，　x^-,x^+是x的左右邻元；

逆等性N4　$\forall x\in[0,1]$，　$N(x)=N^{-1}(x)$，即$N^{-1}(x)$是$N(x)$的逆。

定义 1：满足条件 N1 和 N2 的 $N(x)$称为弱 N 范数(Weak n-norm)；如果满足条件 N1、N2 和 N4，则称为 N 范数；如果满足条件 N1、N2 和 N3，则称为连续(弱)N 范数；如果满足条件 N1、N2 和 N2’，则称为严格单调(弱)N 范数；

例如$1-x^2$，$(1-x)^2$，$\mathbf{N}_2$，$\mathbf{N}_1$和 Sugeno 算子簇都是严格单调连续弱 N 范数(簇)。而 $\mathbf{N}_3$和 $\mathbf{N}_0$中都存在间断点，不是连续 N 范数，只是弱 N 范数(簇)。一般情况下讨论的都是连续的严格单调 N 范数。

2. N 性生成元及其性质

在特征空间 $\boldsymbol{E}$ 中，当每一个模糊测度 $m(X)$可以精确得到时，$m(X)+m(\neg X)=x+N(x)=1$，$N(x)=1-x$，中心非算子成立，它是泛非运算的基模型。但当得到的模糊测度 $m(X)$的值 x 不精确时，设 $m(X)=x^*$，$m(X)+m(\neg X)=x^*+N(x^*)\neq 1$，$N(x^*)\neq 1-x^*$，若需要在一定约束条件下对 $N(x^*)$进行估计，则一般约束条件如下

$$\phi(x^*)+\phi(N(x^*))=1,\ N(x^*)=\phi^{-1}(1-\phi(x^*)) \tag{1}$$

其中 $\phi(x^*)$为 N 性生成元，它是连续的严格单调增函数，$\phi(0)=0$，$\phi(1)=1$，作用是修正误差对模糊测度值 x^*的影响。$\phi(x)=x$ 是特殊的 N 性生成元，它表示模糊测度是精确的。

3. N 范数完整簇及广义自相关系数

有了 N 性生成元及生成方法，下面来研究用它们来生成 N 范数完整簇。

广义自相关系数 k 值：N 范数 $N(x)=\phi^{-1}(1-f(x))$是连续的严格单调减的，它的不动点即广义自相关系数 $k=\phi^{-1}(0.5)$。

研究表明，N 性生成元完整簇的模型有无穷多种，它们与误差分布的形式有关。因而由 N 性生成元完整簇生成的 N 范数完整簇也有无穷多种，最常用的是多项式模型和指数模型。下面用 $\delta(x,k)$表示误差分布函数完整簇，$\Phi(x,k)$表示 N 性生成元完整簇，$N(x,k)$表示 N 范数完整簇.

定义 2：设 N 性生成元簇 $\Phi(x,k)$，$k\in[0,1]$，对某个特定的 $k_1\in[0,1]$，$\phi(x)=\Phi(x,k_1)$是一个 N 性生成元，若 $\Phi(x,k)$满足：

1) $\Phi(x,k)$可随 k 连续的严格单调减地变化；
2) $k=\Phi^{-1}(0.5,k)$，且当 $k=0.5$ 时 $\Phi(x,k)=\boldsymbol{\Phi}_1=x$；
3) 当 $k\to 1$ 时 $\Phi(x,k)\to\boldsymbol{\Phi}_3$，当 $k\to 0$ 时 $\Phi(x,k)\to\boldsymbol{\Phi}_0$；
4) 对 $k_1,k_2\in[0,1]$，$\exists k_{21}\in[0,1]$，使 $\Phi(x,k_{21})=\Phi(\Phi(x,k_1),k_2)$；
5) 对 $k_1\in[0,1]$，$\exists k_1'\in[0,1]$，使 $\Phi^{-1}(x,k_1)=\Phi(x,k_1')$.

则称 $\Phi(x,k)$是N性生成元完整簇，简称N元簇(N-generate cluster).

这里 $\Phi^{-1}(x,k)$表示以x为变元对 $\Phi(x,k)$求逆。

2.2. 泛非命题连接词的相关性质

泛非命题连接词即具有一级不确定性的泛非运算，由于模糊测度的不精确性导致了它的不确定性，这种不确定性由认识偏差或测量误差引起，用广义自相关系数也称误差系数 $k\in[0,1]$来表示。使用一级泛非运算的条件是命题和它的非命题都服从相同的误差分布 $\delta(x,k)$，并有相同的误差水平 k。

泛非命题连接词的运算模型是一个N范数完整簇 $N(x,k)$，其中位置标志参数 k 是 $N(x,k)$的不动点，也是非运算中的阈元，它代表否定中的风险程度。$N(x,k)$是一个可在其存在域内随 k 连续变化的非算子完整簇，它的存在域是：$[0,1]\times[0,1]$，最大非算子是 $\mathbf{N}_3=N(x,1)$，中心非算子是 $\mathbf{N}_1=N(x,0.5)$，最小非算子是 $\mathbf{N}_0=N(x,0)$。

N范数完整簇 $N(x,k)$由泛非运算模型的生成基 $N(x)=1-x$ 和N性生成元完整簇 $\Phi(x,k)$相互作用生

成 $N(x,k)=\Phi^{-1}(1-\Phi(x,k),k)$。

其中泛非运算模型的生成基 $N(x)=1-x$ 是精确命题真值的非运算即中心非算子。N 性生成元完整簇 $\Phi(x,k)$ 的逻辑意义是修正模糊测度误差对命题真值的影响，它与模糊测度的误差分布函数簇 $\delta(x,k)$ 有关。$\delta(x,k)$ 簇有无限多种，故 $\Phi(x,k)$ 簇也有无限多种。一个逻辑推理系统中一般只需要使用同一个 $\Phi(x,k)$ 簇和 $N(x,k)$ 簇。常用的是多项式模型和指数模型。

多项式模型:

$$\Phi_1(x,k)=x(1+\lambda)^{1/2}\Big/\Big(1+\big((1+\lambda)^{1/2}-1\big)x\Big),\ \lambda=(1-2k)/k^2 \tag{2}$$

$$N_1(x,k)=(1-x)/(1+\lambda x),\ k=\Big(\big(1+\lambda\big)^{1/2}-1\Big)\Big/\lambda \tag{3}$$

指数模型:

$$\Phi_2(x,k)=x^n,\ n=-1/\log_2 k \tag{4}$$

$$N_2(x,k)=\left(1-x^n\right)^{1/n},\ k=2^{-1/n} \tag{5}$$

由于泛非命题连接词是由 N 范数完整簇定义的，所以 N 范数和 N 范数完整簇的性质就是线序连续值逻辑泛非命题连接词的性质，归纳起来有：

封闭性 $N(x,k)\in[0,1]$：命题 p 的泛非命题 $\sim_k p$ 仍是命题。

对合律 $\sim_k\sim_k p=p$：命题经过 2 (偶数)次相同误差水平 k 的泛非运算后回到原命题。

泛非性如果 $k\Rightarrow p$，则 $\sim_k p\Rightarrow k$;如果 $p\Rightarrow k$，则 $k\Rightarrow\sim_k p$：不假命题的泛非命题一定不真；不真命题的泛非命题一定不假。

对偶律 $\sim_{k_2}\sim_{k_1}\sim_{k_2}p=\sim_k p$，$k=N(k_1,k_2)$，$\sim\sim_k\sim p=\sim_{1-k}p$：泛非运算模型簇 $N(x,k)$ 满足对偶律，它以中心非算子 $N(x)$为中心，零级对偶和一级对偶都在簇中。

偶等性 $\sim_k\sim_k\sim_k p=\sim_k p$：任何泛非运算的自对偶仍然是自己。

3. 词语相似度计算及其改进

3.1. 词语相似度计算

设有词语 W_1 和 W_2，如果 W_1 用 $\{S_{11},S_{12},\cdots,S_{1n}\}$ n 个概念来描述，W_2 用 $\{S_{21},S_{22},\cdots,S_{2m}\}$ m 个概念描述，则计算词语间的相似度即计算概念集合间的相似度。

1. 义原相似度的计算

要计算两组概念的相似度首先要计算义原间的相似度，因为所有概念都是用义原来表示的。文献[12]中采用通过语义距离来计算义原结点间相似度的办法，即假设两个义原在此层次体系中的路径距离为 d，则这两个义原间的语义相似度可由

$$Sim(W_1,W_2)=\frac{\alpha}{Dis(W_1,W_2)+\alpha} \tag{6}$$

2. 虚词概念的相似度计算

基于“知网”的知识描述语言的，虚词用“{句法义原}”或“{关系义原}”描述，所以虚词的相似度就可以通过对应的“句法义原”或“关系义原”间的相似度计算来得到。

3. 实词概念的相似度计算

实词概念采用的相似度计算方法是通过部分相似度的合成来代替整体的相似度。首先要将两个整体

的各个部分之间建立起一一对应的关系，随后计算各个配对间的相似度，加权求和。若某一部分的对应为空时其相似度定义为一个比较小的常数 δ，和具体词与义原的相似度定义为同一级别。

得到词语的概念集合后，建立起两个集合元素的一一对应关系，之后集合的相似度计算就等于其元素对相似度的算术平均值。具体算法在文献[13]中曾有详细描述，这里不再重复。

3.2. 义原相似度的计算及其改进

要计算两组概念的相似度首先要计算义原间的相似度，因为所有概念都是用义原来表示的。文献[12]中采用通过语义距离来计算义原结点间相似度的办法，即假设两个义原在此层次体系中的路径距离为 d，则这两个义原间的语义相似度可由式(6)计算得来，具体化为：

$$Sim\left(P_1,P_2\right)=\frac{\alpha}{d+\alpha} \tag{7}$$

其中 a 为可调节参数，通常 a 是指相似度为 0.5 时的词语距离值。p_1 和 p_2 为两个不同的义原，d 是在层次树中 p_1 和 p_2 间路径的长度，为一正整数。具体的计算方法：找到 p_1 和 p_2 的最近共同双亲结点 cp，如果 cp 存在，则 $d=dis\left(p_1,cp\right)+dis\left(p_2,cp\right)$；否则 $d=20$。

在柔性逻辑中，泛非运算模型为 $N\left(x,k\right)=\Phi^{-1}\left(1-\Phi\left(x,k\right),k\right)$，常用的多项式模型为 $N_1\left(x,k\right)=\left(1-x\right)/\left(1+\lambda x\right)$，$k=\left(\left(1+\lambda\right)^{1/2}-1\right)/\lambda$，$k$ 为广义自相关系数。

这里采用柔性逻辑来分析义原相似度，由原义原相似度计算公式

$$Sim\left(P_1,P_2\right)=Sim\left(d,\alpha\right)=\frac{\alpha}{d+\alpha} \tag{8}$$

其中 d 是 p_1 和 p_2 在层次树中的路径长度。

将泛非运算因子引入，即将泛非运算模型和式(8)结合，对 d 进行归一化处理，令 $x=\frac{d}{d+\left(1+\lambda\right)\alpha}$，则 $d=\frac{\left(1+\lambda\right)\alpha x}{1-x}$，代入式(7)得到

$$Sim\left(d,\alpha\right)=Sim\left(x,\lambda\right)=\frac{\alpha}{\frac{\left(1+\lambda\right)\alpha x}{1-x}+\alpha}=\frac{1-x}{\left(1+\lambda\right)x+\left(1-x\right)}=\frac{1-x}{1+\lambda x}=N_1\left(x,k\right) \tag{9}$$

也就是说，原来的义原相似度定义可以由泛非运算来定义。如此以来使相似度的计算更加的柔性化，可以随着不同的要求来改变使用的相似度计算方法，在簇中选取适合计算的范数。比如当前的义原相似度计算中只考虑了义原在层次树中的深度及相互的路径长度，并没有考虑区域密度等影响因素，而实际使用中密度的影响会更大，所以就需要修改当前的相似度计算方法，比如在文献[14]中修改了其义原距离 d 计算公式为

$$d=\delta\frac{dis\left(p_1,cp\right)+dis\left(p_2,cp\right)}{con\left(p_1\right)+con\left(p_2\right)} \tag{10}$$

其中 $con\left(p\right)=\gamma deep\left(p\right)+\mu density\left(p\right)$，$\gamma<\mu$，$\gamma+\mu=1$，$\delta$ 为可调节参数。

该定义中考虑了深度及区域密度的共同影响，并且设定了 δ 为可调节参数，更符合实际。但这样的定义的距离和原来定义的距离不同，需要全部重新计算，而且该定义假设义原的深度和区域密度对义原相似度的贡献是独立的，深度对相似度的影响比密度对相似度的影响要小，但实际深度和密度是有相互关联的，如此又需要改进该计算方法。

4. 总结与展望

使用泛非算子后，可以通过调节广义相关系数 k 来体现不同因素对相似度计算的影响。广义相关系数原意是修正测度误差的，在这里可以通过不同的 k 值得到不同的泛非运算模型，也即新相似度计算模型，如此使得相似度计算能在不同的情况下有不同的计算结果。如此可为文本挖掘的研究提供了新的思路和方向，以期能更方便有效的进行相似度的计算，帮助企业及用户更有效的挖掘需要的数据。

基金项目

本项目得到网络文化与数字传播北京市重点实验室开放课题资助(ICDD201610)；2015 课程建设“数据结构教学方式改革的研究项目”资助。

参考文献 (References)

[1] 江敏, 肖诗斌, 王弘蔚, 施水才. 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, 22(5): 84-89.

[2] 王小林, 王义. 改进的基于知网的词语相似度算法[J]. 计算机应用, 2011, 31(11): 3075-3090.

[3] 何华灿, 王华, 刘永怀, 等. 泛逻辑学原理[M]. 北京: 科学出版社, 2001.

[4] 马盈仓. 命题泛逻辑的演算理论及推理研究[D]: [博士学位论文]. 西安: 西北工业大学, 2005.

[5] 王华. 命题泛逻辑学的包容性研究[D]: [硕士学位论文]. 西安: 西北工业大学, 2004.

[6] 徐章艳, 汤服成, 李凡. 泛逻辑中的泛或和泛与的逻辑运算公式[J]. 模糊系统与数学, 2002(16): 152-155.

[7] Luo, M.X. and He, H.C. (2010) Algebraic Method to Study Universal Logic. Science Press, Beijing.

[8] Esteva, F., Trillas, E. and Domingo, X. (1981) Weak and Strong Negation Functions for Fuzzy Set Theory. System on Multiple-Valued Logic.

[9] Ovchinnikov, S.V. (1983) General Negations in Fuzzy Set Theory. *Journal of Mathematical Analysis and Applications*, **92**, 234-239. https://doi.org/10.1016/0022-247X(83)90282-2

[10] Weber, S. (1983) A General Concept of Fuzzy Connectives: Negations and Implications Based on t-Norms and t-Conorms. *Fuzzy Sets and Systems*, **11**, 115-134. https://doi.org/10.1016/S0165-0114(83)80073-6

[11] Mizumoto, M. (1989) Pictorial Representations of Fuzzy Connectives, Part I: Cases of T-Norms, T-Conorms and Averaging Operators. *Fuzzy Sets and Systems*, **31**, 217-242. https://doi.org/10.1016/0165-0114(89)90005-5

[12] 刘群, 李素建. 基于《知网》的词汇语义相似度计算[J]. 中文计算语言学, 2002, 7(2): 59-76.

[13] 刘城霞, 吴菲滢. 基于关键词相似度的 Web 用户挖掘研究与实现[J]. 计算机科学与应用, 2013, 3(4): 232-238.

[14] 袁晓峰. 《知网》义原相似度计算的研究[J]. 辽宁大学学报, 2011, 38(4): 358-361.