第35卷 第10期 计 算 机 工 程 2009年5月
Vol.35 No.10 Computer Engineering May 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)10—0201—02 文献标识码：A 中图分类号：TP311

基于语义的高质量中文短信文本聚类算法

刘金岭
(淮阴工学院计算机工程系，淮安 223003)

摘 要：现有数据聚类方法在处理文本数据时，没有考虑词之间潜在的相似信息，导致聚类效果不理想。针对中文短信文本聚类提出一种基于语义的聚类算法。给出中文概念、词和中文短信文本的相似度度量方法，通过向下连锁裂变和向上两两归并完成中文短信文本聚类。实验结果表明，该算法的聚类质量高于传统算法。

关键词：短信文本；语义；概念相似度

High Quality Algorithm for Chinese Short Messages Text Clustering Based on Semantic

LIU Jin-ling
(Department of Computer, Huaiyin Institute of Technology, Huaian 223003)

【Abstract】 Existing data clustering method lacks considering of latent similar information existing among words, and it leads to unsatisfactory clustering result. Aiming at Chinese short message text clustering, this paper proposes a clustering algorithm based on semantic. It offers Chinese concept, and the measuring methods to calculate the similarity degree about words and Chinese short message text. It completes the clustering of Chinese short messages text through fission downwards and mergence of twos upwards. Experimental results show that this algorithm has better clustering quality than traditional algorithm.

【Key words】 short messages text; semantic; concept similarity

1 概述

短信信息在舆论导向和传播上扮演着越来越重要的角色，被一些学者誉为继报纸、广播、电视、网络后的第5大媒体。由于传统基于关键词集的文本聚类技术存在不足，因此基于概念的文本聚类技术成为新的研究热点。目前，国内对中文文本聚类的研究主要包括利用概率统计的方法[1]、基于训练学习的方法来生成概念空间[2]，或通过自定义模糊概念图[3]描述概念空间、采用潜在语义分析技术[1]等。采用传统数据挖掘方法处理文本数据前，必须先将文本转换为向量空间模型或后缀树模型等。此类模型从不同角度使用不同方法处理特征加权、类别学习等问题，其中，向量空间模型是最有效的模型之一。

Gerard Salton于20世纪60年代提出采用向量空间模型进行文本特征表达，用TFIDF(Term-Frequency Inverse-Document-Frequency)将文档转化为向量形式，并在向量空间中计算文本相似度。在基于TFIDF的向量空间模型中，因为没有考虑词之间存在的概念相似情况，所以影响了数据聚类的准确性，尤其对于中文短信的文本聚类，得到的相似度与实际情况偏离严重。因此，最终的聚类结果与人们的直观感受相差很大。文献[4]基于知网模型，提出一种相似度计算方法，但由于该方法只能用于词和概念的相似度计算，没有提供文本相似计算分析，且缺少对计算公式的理论分析，因此难以用于聚类分析。

本文利用《知网》中对每个概念进行描述时的丰富语义信息，得到与人的直觉较符合的结果，进而利用词语相似度值进行较细致的刻划。

2 基于《知网》的中文短信相似度计算

《知网》是一个以汉语和英语词语代表的概念为描述对象，以揭示概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。《知网》通过义原的组合标注各种单纯的或复杂的概念，以及各个概念之间、概念的属性和属性之间的关系。相对而言，新词虽然不断增加，但义原的增加极少。在《知网》中，词义定义为各种义原的组合。在向量空间模型中引入《知网》作为背景知识，将关键词映射到义原，可以在一定程度上解决同义词替换的问题，使相同主题、包含不同同义词和近义词的文档能更好地聚集在一起。

由于《知网》中对一个词的语义采用一种多维知识表示形式，因此给词语相似度的计算带来了麻烦。在WordNet和《同义词词林》中，所有同类的语义项(WordNet的synset或《同义词词林》的词群)构成一个树状结构，要计算语义项之间的距离，只要计算树状结构中相应节点的距离即可。而在《知网》中，词语语义描述具有如下特性：

(1)每个词的语义描述由多个义原组成；

(2)在词语的语义描述中，各个义原并不是平等的，它们之间有着复杂的关系，通过一种专门的知识描述语言来表示。

2.1 概念相似度计算

与传统语义词典不同，《知网》采用了1 500多个义原，通过一种知识描述语言对每个概念进行描述。义原是描述概念的最基本单位，它们相互之间存在复杂的关系。在《知网》

作者简介：刘金岭(1958－)，男，教授，主研方向：数据仓库，数据挖掘

收稿日期：2008-12-05 **E-mail：**liujinlingg@126.com

中，共描述了义原之间的 8 种关系：上下位关系，同义关系，反义关系，对义关系，属性-宿主关系，部件-整体关系，材料-成品关系，事件-角色关系。可以看出，义原之间组成了一个复杂的网状结构，而不是一个单纯的树状结构。义原关系中最重要的是上下位关系。根据义原的上下位关系，所有基本义原组成了一个义原层次体系，该义原层次体系是一个树状结构，可以得到一棵义原概念树，它是本文进行语义相似度计算的基础。

设义原集合为 M，义原数量表示为$|M|$，义原用 p_i 表示，i=1,2,⋯,$|M|$。

设 L_i 为义原 p_i 在概念树中的深度，y 为距离初始阈值，x 为满足不等式 max(L)<y/x 的一个正实数，则 p_i 与其父节点的距离定义为

$$d(p_i, parent(p_i)) = y - L_i \cdot x \tag{1}$$

任意 2 个义原 p_i, p_j 之间的距离定义为

$$d(p_i, p_j) = \omega_k \cdot [y - \max(L_i, L_j) \cdot x] \tag{2}$$

其中，ω_k 表示第 k 种关系对应的权重，通常取 $\omega_k \geq 1$。

可以验证，上述定义符合对距离函数的数学要求。

从式(1)、式(2)可以看出，义原在义原层次树中的分类深度越深，它们之间的距离越小，即越相似，这与人的直观印象一致。

根据《知网》中的定义，可以得到义原上下位关系集合 P 和其他关系的集合 Q。

算法 1 求 2 个义原最小距离的算法

输入 p_i, p_j

输出 p_i, p_j 最短距离 d_{min}

任取 p_i, p_j 间的一条路径 L

d_{min}=0

For $\forall p_i', p_j' \in L$

利用式(1)、式(2)计算出 $d(p_i', p_j')$ ；

$d_{min} = d_{min} + d(p_i', p_j')$

End for

从而得到 p_i, p_j 间的距离 d_{min} ；

对于 p_i, p_j 间的任意一条路径 t，重复(3)~(6)可得 $d_t(p_i, p_j)$；

If($d_t(p_i, p_j)$<= d_{min} then $d_{min} = d_t(p_i, p_j)$;

/* 得到 p_i, p_j 间的最小距离*/

返回义原 p_i, p_j 的最小距离 d_{min}.

定义 1 利用义原最短距离定义义原 p_i, p_j 间的相似度

$$Sim(p_i, p_j) = \frac{\alpha}{d_{\min}(p_i, p_j) + \alpha} \tag{3}$$

其中，α 是一个可调节的参数。

董振东先生在描述《知网》的结构时，强调概念是对词汇语义的一种描述，每个词可以表达为几个概念。概念用一种知识表示语言来描述，该知识表示语言所用的词汇称为义原。义原是用于描述一个概念的最小语义单位，用一系列义原对每个概念进行描述，概念相似度的计算是基于义原间的相似度。

定义 2 设概念 x 和 y 分别由义原组(p_{x1},p_{x2},⋯,p_{xn})和(p_{y1}, p_{y2},⋯,p_{ym})表示，记

$$(x, y) = \sum_{i=1}^{n} \sum_{j=1}^{m} Sim(p_{xi}, p_{yj})$$

定义概念 x, y 的相似度为

$$Sim(x, y) = \frac{(x, y)}{\sqrt{(x, x) \cdot (y, y)}} \tag{4}$$

2.2 词语相似度计算

度量 2 个词语关系的一个重要指标是词语的相似度，因为每个词可以由 1 个或多个概念组成，所以可以把 2 个词语之间的相似度问题归结到 2 个概念之间的相似度问题。

定义 3 对于 2 个汉语词语 W_1 和 W_2，如果 W_1 有 n 个概念 S_{11},S_{12},⋯,S_{1n}，W_2 有 m 个概念 S_{21},S_{22},⋯,S_{2m}，则规定 W_1 和 W_2 的相似度等于各个概念相似度的最大值，即

$$Sim(W_1, W_2) = \max_{i=1...n, j=1...m} Sim(S_{1i}, S_{2j}) \tag{5}$$

本文考虑孤立的 2 个词语的相似度。如果是在一定上下文中的 2 个词语，则最好先进行词义排岐[5]，将词语标注为概念后，再对概念计算相似度。

2.3 中文短信文本相似度计算

中文短信文本可以用 TFIDF 将文档转化为词语向量形式，并利用词语相似度计算文本相似度。

定义 4 设中文短信文本 SM_i, SM_j 分别可以由词语组(W_{i1}:t_{i1}, W_{i2}:t_{i2},⋯,W_{in}:t_{in})和(W_{j1}:t_{j1}, W_{j2}:t_{j2},⋯,W_{jm}:t_{jm})表示，其中，(W_{i1},W_{i2},⋯,W_{in})和(W_{j1},W_{j2},⋯,W_{jm})分别表示中文短信文本 SM_i, SM_j 的词向量，而(t_{i1},t_{i2},⋯,t_{in})和(t_{j1},t_{j2},⋯,t_{jm})表示相应词的权重，记

$$(SM_i, SM_j) = \sum_{k=1}^{n} \sum_{l=1}^{m} t_{ik} \cdot t_{jl} \cdot Sim(W_{ik}, W_{jl})$$

定义 SM_i，SM_j 的相似度为

$$Sim(SM_i, SM_j) = \frac{(SM_i, SM_j)}{\sqrt{(SM_i, SM_i) \cdot (SM_j, SM_j)}} \tag{6}$$

3 基于语义的中文短信高质量聚类算法

基于语义的中文短信高质量聚类算法的思想如下：先将整个中文短信文本集合连锁二分裂变为内部相似，以相互之间互异的多个较小的集合为元素构成结果集 S，S 的质量由稠密度函数

$$Density(S) = \sqrt{\sum_{SM, SM' \in S} Sim^2(SM, SM') / |S|} \tag{7}$$

判定，然后对给定的阈值，两两归并成符合条件的子集，直至结果完成。

算法 2 基于语义的中文短信高质量聚类算法

输入 中文短信文本集合 Set，二分裂变阈值 t，两两归并阈值 g

输出 聚类结果集 Set={S_1,S_2,⋯,S_n}

(1) 初始化聚类集 Set；

/*先将 Set 看成一个类集*/

(2) for ∀ 子类 S ∈ Set；

(3) 利用式(7)，计算 S 中的每个子类的质量；

(4) 如果 Density(S)>t

(5) 从 S 中任选 2 个中文短信文本做 2 个裂变小聚类 S_i 和 S_j 的中心点 SM_i, SM_j；

(6) do while changNum<>0 or IterNum<maxNum

/*循环条件 changNum<>0，还有小聚类需要裂变；maxNum 是限制最大循环次数*/

(7) 对 ∀ SM ∈ S，根据 SM 与 S_i 和 S_j 的相似度，将 SM 分配到相应的小聚类中；

(8) 如果 SM 的类别发生了改变，则 changNum++；changNum=0；

(9) 重新计算小聚类 S_i 和 S_j 的中心点并替换中心点，得新的中心点

SM_i,SM_j；

(10) end while

(11) end for

(下转第 205 页)