基于 Isomap 的中文短信文本聚类算法

刘金岭
LIU Jin-ling

淮阴工学院,江苏 淮安 223001
Huaiyin Institute of Technology,Huai'an,Jiangsu 223001,China
E-main:liujinlingg@126.com

LIU Jin-ling.Chinese short messages text clustering algorithm based on Isomap.Computer Engineering and Applications, 2009,45(34):144-146.

Abstract: The calculating way proposed in this paper is to calculate the likeness degree of Chinese message and a message is gottern which embedded in the semantic space by using the Isomap method.This paper analyzes the messages according to the different clustering types in low-dimensional embedding.This algorithm has overcome difficulties in analyzing messages of traditional clustering types on different layers,and it has also overcome weakness of word frequency statistics which can not gather the similar meaning messages together.Experimental result indicates the algorithm is effective.

Key words: short messages clustering;Isomap algorithm;semantic space

摘 要:给出的算法思想是首先计算出中文短信的相似度,再通过使用 Isomap 方法得到短信在语义空间中的嵌入情况,然后将短信在低维嵌入上进行聚类分析。该算法克服了短信的传统聚类分析在表示层次上遇到的困难,也克服了词频统计法不能将内容意思相似的短信聚集在一起的缺点,实验表明该算法是行之有效的。

关键词:短信聚类;Isomap 算法;语义空间
DOI:10.3778/j.issn.1002-8331.2009.34.044 **文章编号**:1002-8331(2009)34-0144-03 **文献标识码**:A **中图分类号**:TP311

目前,短信信息在舆论导向和传播上扮演着越来越重要的角色,短信信息已被一些学者誉为继报纸、广播、电视、网络之后的第四大媒体。短信的使用已渗透到社会的各个领域,与此同时,通过短信传播非法、色情以及垃圾信息的现象也随之增多,且其带来的损失也在不断地增大。因此,进行海量短信信息的研究分析,建立有效、准确的舆情预测模式,就显得十分重要。短信具有其自身的特点:(1)长度比较短,通常出现的是不会超过 140 个字符的短信用语;(2)频繁使用网络语言以及缩略语;(3)由于群发、转发、下载短信行为频繁,短信库中存在大量的重复短信。因此,基于短信的文本处理必须面对这些特点,开发适应性技术,才能应对现实中纷繁复杂的应用需求。目前,国内中文文本聚类方面的研究中,主要有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],或者通过自定义模糊概念图[3]来描述概念空间。在文本特征选择方面,提出了词频-倒文档频度法(Term Frequency-Inverse Document Frequency,TF-IDF)、信息增益法(Information Gain,IG)、CHI 统计量法、互信息法(Mutual Information,MI)等[4]专门的方法,同时还将主成分分析、线性鉴别分析和奇异值分解的方法应用于文本特征选择,衍生出了潜在语义索引[5](Latent Semantic Index,LSI)的重要概念。由于短信自身的特点,使得传统的聚类分析方法在短信表示层次上就遇到了极大的困难,无论是用传统的文本表示模型,还是用现在一些新兴的文本表示模型,都无法良好地表示。总会遇到特征向量稀疏性的问题,最终使得短信的聚类的变为简单层次上"词重现"一级的短信聚集。而该文提出了一种基于《知网》的中文文本聚类的方法,绕开了文本表示的问题,通过直接计算短信词块之间的相似度的办法计算短信之间的相似性,最后利用发现其内蕴的流形结构,再在此流形结构上进行聚类分析,因而能够取得比传统方法优异的结果。

1 中文短信在语义空间中的流形结构

语义空间特指文本分类空间。用语义空间这样的表述是为了说明,所做的研究已经脱离原先语法的层次,向更易于为人所理解的层次进了一步。借助流形学习的方法,发现中文短信在语义空间分类空间中低维的流形嵌入情况。需要说明的是,中文短信在嵌入空间的坐标是有意义属性的,例如不同的距离代表着文本之间不同的语义联系。类似于 Roweis S.和 Saul L.[6]在 Lee D.和 Seung H.[7]工作的基础上,对 Grolier's Encyclopedia[7]中 31 000 篇文档空间中的流形结构的研究,很容易得到中文短信在语义空间的流形结构。

2 流形学习及 Isomap 算法

近年来,流形学习领域产生了大量的研究成果[8],LLE[9]和

作者简介:刘金岭(1958-),男,教授,主要研究方向:数据仓库、数据挖掘。
收稿日期:2008-08-22 **修回日期**:2008-11-10

Isomap[10]是两种具有代表性的非线性降维方法。Roweis 和 Saul 提出的 LLE 算法能够实现高维输入数据点映射到一个全局低维坐标系,同时保留了邻接点之间的关系,这样固有的几何结构就能够得到保留。此算法不仅能够有效地发现数据的非线性结构,同时还具有平移、旋转等不变特性。Tenen-baum 等人提出的 Isomap 算法是建立在多维尺度变换[10](MDS)的基础上,力求保持数据点的内在几何性质,即保持两点间的测地距离。Isomap 算法的关键是利用样本向量之间的欧式距离 $d_x(i,j)$ 计算出样本之间的测地距离 $d_c(i,j)$,真实地再现高维数据内在的非线性几何结构。然后使用经典 MDS 算法构造一个新的 d 维空间 Y(d 是降维后空间的维数),最大限度地保持样本之间的欧式距离 $d_Y(i,j)$ 与 $d_c(i,j)$ 误差最小,以达到降维的目的,算法描述如下:

输入:输入样本 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\in R^N\}$,样本的降维的维数 d,邻域参数 $k(\varepsilon$ 邻域);

输出:低维嵌入 $\boldsymbol{Y}=\{y_1,y_2,\cdots,y_n\in R^d\}$;

算法:计算每个点 x_i 的近邻点(k 邻域),$1\leqslant i\leqslant n$;

/* 构造近邻图 */

若两点 x_i,x_j 互为近邻点,则相应的边值设为欧式距离 $d_x(i,j)$,否则为 ∞;

/* 求得最短路径矩阵 $\boldsymbol{D}_G=\{d_G(i,j)\}$,其中两点间的最短距离可采用 Dijkstra 算法 */

计算任意两点 x_i、x_j 间的最短路径 $d_G(i,j)$;

/* 用 MDS 求低维嵌入流形 */

$S=(S_{ij})=(D_{2ij})$,$H=(H_{ij})=(\delta_{ij}-1/N)$,$\tau(D)=-HSH/2$,低维嵌入是 $\tau(D)$最小的第 2 个到第 d+1 个特征值对应的特征向量。

3 基于语义的中文短信相似度计算

3.1 基于语义的相似度计算方法

与传统的语义词典不同,在《知网》(http://www.keenage.com,2008 年 1 月 18 日)中,并不是将每一个概念对应于一个树状概念层次体系中的一个结点,而是通过用一系列的义原,利用某种知识描述语言来描述一个概念。而这些义原通过上下位关系组织成一个树状义原层次体系。目标是要找到一种方法,对用这种知识描述语言表示的两个语义表达式进行相似度计算。

利用《知网》计算语义相似度一个最简单的方法就是直接使用词语语义表达式中的第一独立义原,把词语相似度等价于第一独立义原的相似度。这种方法好处是计算简单,但没有利用知网语义表达式中其他部分丰富的语义信息。利用 Li Sujian,et al.提出的一种词语语义相似度的计算方法[11],该计算过程不仅综合利用了《知网》的义原相似度计算,还考虑了义原之间的上下文关系和义原之间的其他关系,记义原 s_i,s_j 的相似度为 $Lsim(s_i,s_j)$。

3.2 词语相似度计算

对于两个汉语词语 W_1 和 W_2,如果 W_1 有 n 个义项(概念):$S_{11},S_{12},\cdots,S_{1n}$,$W_2$ 有 m 个义项(概念):$S_{21},S_{22},\cdots,S_{2m}$,规定,$W_1$ 和 W_2 的相似度等于各个概念的相似度之最大值(当然可以用其他方法来定义),也就是说:

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

这样,就把两个词语之间的相似度问题归结到了两个概念之间的相似度问题。当然,这里考虑的是孤立的两个词语的相似度。如果是在一定上下文之中的两个词语,最好是先进行词义排岐,将词语标注为概念,然后再对概念计算相似度。

3.3 基于语义的中文短信相似度计算

假设短信 $SM_i,SM_j,i=1,2,\cdots,m,j=1,2,\cdots,n$,分别由词语:$\{T_1^{SM_i},T_2^{SM_i},\cdots,T_m^{SM_i}\}$和$\{T_1^{SM_j},T_2^{SM_j},\cdots,T_n^{SM_j}\}$表示。

(1)将$\{T_1^{SM_i},T_2^{SM_i},\cdots,T_m^{SM_i}\}$和$\{T_1^{SM_j},T_2^{SM_j},\cdots,T_n^{SM_j}\}$之间的相似度矩阵利用式(1)定义为:

$$\begin{bmatrix} Sim(T_1^{SM_i},T_1^{SM_j}) & \cdots & Sim(T_1^{SM_i},T_n^{SM_j}) \\ Sim(T_2^{SM_i},T_1^{SM_j}) & \cdots & Sim(T_2^{SM_i},T_n^{SM_j}) \\ \cdots & & \cdots \\ Sim(T_m^{SM_i},T_1^{SM_j}) & \cdots & Sim(T_m^{SM_i},T_n^{SM_j}) \end{bmatrix} \tag{2}$$

(2)利用 Hungarian 算法[12]去找出$(T_1^{SM_i},T_2^{SM_i},\cdots,T_m^{SM_i})$和$(T_1^{SM_j},T_2^{SM_j},\cdots,T_n^{SM_j})$之间的最大匹配(即将其视为二部图的最大匹配问题,两个词之间的连接权重可以看成其之间的相似度)。设 SM_i 在 SM_j 中的最大匹配是$\{T_{j_1}^{SM_j},T_{j_2}^{SM_j},\cdots,T_{j_m}^{SM_j}\}$,$j_k\in\{1,2,\cdots,n\}$,$k=1,2,\cdots,m$。$SM_j$ 在 SM_i 中的最大匹配是$\{T_{j_1}^{SM_i},T_{j_2}^{SM_i},\cdots,T_{j_n}^{SM_i}\}$,$j_k\in\{1,2,\cdots,m\}$,$k=1,2,\cdots,n$。

(3)短信 SM_i,SM_j 之间的相似度可以如下定义为[13]:

$$Distance(SM_i,SM_j)=\frac{m+n}{A\times B} \tag{3}$$

其中:

$$A=sum\{Sim(T_1^{SM_i},T_{j_1}^{SM_j}),\cdots,Sim(T_m^{SM_i},T_{j_m}^{SM_j})\} \tag{4}$$

$$B=sum\{Sim(T_1^{SM_j},T_{j_1}^{SM_i}),\cdots,Sim(T_n^{SM_j},T_{j_n}^{SM_i})\} \tag{5}$$

这样就可以计算两条短信之间的语义距离,它比较适用于所计算文本比较短的情况,因为无论 Hungarian 算法还是查《知网》词典都是计算复杂度相当高的操作。而短信正好具备了这样的特点,因而特别适用于该方法。

4 基于 Isomap 的中文短信聚类的算法

该算法分为如下几步处理:

(1)预处理:对于短信 $SM_i,i=1,2,\cdots,k$,首先将其分散成 m 个词语的列表,即 $SM_i=\{T_{i1},T_{i2},\cdots,T_{im}\}$,其中每个 T_{ik} 表示一个词语;

(2)对词语进行词性消歧和词义消岐[14];

(3)计算短信 SM_i 和 SM_j 之间的相似度;

(4)使用 Isomap 流形学习算法发现短信在语义空间中的内嵌低维结构;

(5)对降维后的短信使用 K-means 算法对短信文档集进行聚类。

5 实验及分析

实验算法用 Visual Basic 6.0 实现,在内存为 2.0 GB,主频

为酷睿双核2.0 GHz,操作系统为 Windows XP 的方正计算机上进行实验。为了程序的简单,实验短信库中的 1 000 条短信样本是由日常生活交流中短信的大概规律筛选的。首先计算词语的相似度,再计算出短信的相似度,然后使用 Isomap 方法(该实验取近邻数为 6)实现降维。为了直观性,实验选择降维至二维。聚类中所使用的算法是 *K*-means,将其聚为 5 个类:'*'表示 A 类短信;'O'表示 B 类短信;'+'表示 C 类短信;'△'表示 D 类短信;'·'表示 E 类短信;'★'表示聚类中心,结果如表 1 及图 1。

表 1 1 000 条短信的聚类结果表

类别	A	B	C	D	E
数量	68	193	103	504	132

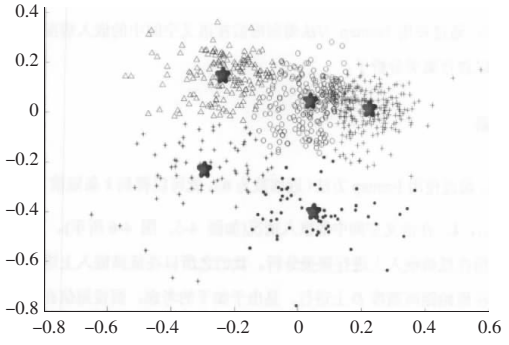

图 1 短信在连续语义空间的排列情况

实验结果分析:

(1)和以前基于词频统计的聚类方法类似,该算法能保证形式上一致的短信(如 E 类短信中有三条重复短信:短信库中的原始标号为 33、47、56;A 类短信中有两条重复短信:短信库中的原始标号 29、43),聚在一起(映射为一点)。

(2)基于语义的短信聚类能够将内容意思相似的短信聚集在一起,这是其他聚类算法(包括词频统计的聚类方法)做不到的,如 B 类短信基本上都是情感交流类的短信。

(3)由于短信库自身的限制:数量比较少,而且来源单一,基本上以人们的日常交流为主,这为聚类分析带来了困难。从聚类的结果来看, 所有聚类类别之间的差异性不是特别显著,如 B、C、D 类基本上都是以情感交流类的短信为主;而 A、E 类短信就涉及到日常工作和生活相关内容。

(4)由于该算法的第 4 步实现了降维,减少了计算量,使用该算法对上述 1 000 条短信聚类比直接用 *K*-means 算法效率提升了约 0.8%。

6 结论

基于 Isomap 的中文短信聚类的处理方法是一种基于语义的短信聚类,能够将内容意思相似的短信聚集在一起,这是其他聚类算法不能达到的,另一方面,在处理手段上利用了流形学习领域的非线性降维方法,以实现高维的问题降到低维来解决,正因为如此,才可以给出基于 Isomap 的中文短信聚类二维甚至三维的直观表示。由于短信之间的相似度可以由其邻近程度来表示,这样,能够很容易地找到和某一条短信最相似的若干条短信,通过简单的邻域就可以确定,这将为进一步研究基于内容的检索工作打下良好的基础。

对于此种短信聚类的方法,至少在两个方面需要做进一步的研究:

(1)寻找更有效的计算短信相似度的方法;

(2)寻找更有效的流形学习算法。

参考文献:

[1] 宫秀军,史忠植.基于 Bayes 潜在语义模型的半监督 Web 挖掘[J].软件学报,2002,13(8).

[2] 傅伟鹏,吴斌,何清,等.一种概念空间自生成方法[J].计算机工程与应用,2002,38(7):63-65.

[3] 陈宁,陈安,周龙骧,等.基于模糊概念图的文档聚类及其在 Web 中的应用[J].软件学报,2002,13(8).

[4] 刘涛.用于文本分类和文本聚类的特征选择和特征抽取方法的研究[D].天津:南开大学,2004.

[5] Dumais S T,Furnas G W,Landauer T K,et al.Using latent semantic analysis to improve information retrieval[C]//Proc of the Int Conf on Human Factors in Computing,1988:281-285.

[6] Roweis S,Saul L.Nonlinear dimensionality reduction by locally linear embedding[J].Science,2000:2323-2326.

[7] Lee D,Seung H.Learning the parts of objects by non-negative matrix factiorization[J].Nature,1999:788-791.

[8] Roweis S,Saul L.Nonlinear dimensionality reduction by locally linear embedding[J].Science,2000,290(5500):2323-2326.

[9] Tenenbaum J,Silva D D,Langford J.A global geometric framework for nonlinear dimensionality reduction[J].Science,2000,290(5500):2319-2323.

[10] Borg I,Groenen P.Modern multidimensional scaling:Theory and application[M].New York:Springer-Verlag,1997.

[11] Li Su-jian,Zhang Jian,Huang Xiong,et al.Semantic computation in Chinese question-answering system[J].Journal of Computer Science and Technology,2002.

[12] Alsuwaiyel M.算法设计技巧与分析(影印版)[M].北京:电子工业出版社,2003:237-248.

[13] Resnik P.Semantic similarity in a taxonomy:An information-based measure and its application to problems of ambiguity in natural language[J].Journal of Artificial Intelligence Research,1999(11):95-130.

[14] 杨尔弘,张国清,张永奎.基于义原同现频率的汉语词义排歧方法[J].计算机研究与发展,2001,38(7).

[15] 张军平,曹存根.神经网络及其应用[M].北京:清华大学出版社,2004:228-236.