第41卷 第6A期 计 算 机 科 学 Vol. 41 No. 6A
2014年6月 Computer Science June 2014

基于 Relative-IDF 的医药数据相似度算法研究

向林泓 张 炬 孙启龙 赵学良
(中国科学院重庆绿色智能技术研究院 重庆 404100)

摘 要 医药数据相似度计算在药物信息处理中具有重要的作用。传统的文本相似度计算在医药领域并不能取得很好的效果。针对医药数据文本的特殊性,提出基于 Relative-IDF 的医药数据相似度计算算法。实验结果表明:相比传统 TF-IDF、编辑距离等计算方法,基于 Relative-IDF 的医药数据相似度计算在效率和准确性上都有了很大的提升。

关键词 医药数据相似度,编辑距离,Relative-IDF,TF-IDF
中图法分类号 TP311.1 **文献标识码** A

Medical Data Similarity Algorithm Analysis Based on Relative-IDF

XIANG Lin-hong ZHANG Ju SUN Qi-long ZHAO Xue-ling
(Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 404100, China)

Abstract Medical data similarity calculation plays an important role in drug information treatment. Traditional text similarity measurement in the field of medicine and can't get good results. Particularity for the pharmaceutical data text proposed based on Relative-IDF similarity calculation algorithm of medical data. Experimental results show that compared to traditional TF-IDF, edit distance calculation method, based on Relative-IDF medical data similarity measurement in efficiency and accuracy has been greatly improved.

Keywords Medical data similarity, Edit distance, Relative-IDF, TF-IDF

1 引言

文本聚类是指自动地将文本集合分为不同的类别,同一个类别中的文本相似度高,不同类别之间的文本则不相似[1]。文本聚类过程关键在于文本之间相似度的计算。文本的相似度是一个在语言学、心理学和信息理论等领域内被广泛研究的重要话题。文本相似度是表示两个或者多个文本之间匹配程度的一个度量参数,相似度越大,文本相似度越高,反之亦然。文本相似度在许多领域有着广泛的应用:在信息检索领域,文本相似度被认为是改进信息检索效果的最好方法之一;[2]在图像检索领域,利用图像周围的文本可以大大提高图片检索精度[3];此外,文本相似度还广泛应用于文本的重复检测[4]、文本摘要自动生成[5]、文本分类等领域[6]。

随着医药医学领域的迅速发展,越来越多的学者将分词技术、文本聚类、文本分类等自然语言处理技术应用到医药医学领域。国家科技部中药基础数据库项目组建立了"中医药学语言系统"[7],该系统提高了计算机程序理解生物医学词汇语义的能力,并利用这种理解帮助用户检索和获取相关信息。重庆推出了"药品在线交易服务平台"[8],进一步推进了自然语言处理技术在医药医学领域的发展。由于医药数据文本的特殊性,传统的文本相似度计算并没有起到很好的效果。本文针对传统文本相似度计算在医药领域存在的缺陷,提出了一种新型的文本相似度计算方法,并且能够在较为广泛的领域使用。

2 相关工作

目前,国内外有很多学者在研究文本相似度计算问题,并提出了一些解决方法。Gerard Salton 和 McGill 于 1969 年提出基于向量空间模型 VSM(Vector Space Model)计算相似度,把文档简化为以特征项的权重向量表示,通过词频统计和向量降维计算相似度,从而简化了文本中关键词之间的复杂关系[9]。挪威 Agder 大学的 Vladimir Oleshchuk 等人提出基于 Ontology 的文本相似度比较方法,其将本体论引入文本相似度计算[10]。Chris H. Q. Ding 提出基于隐性语义索引模型 LSI(Latent Semantic Indexing)计算文本相似度,该方法将全部文档生成文档矩阵,然后一直进行分解,最后通过标准化内积计算来计算向量之间的余弦夹角,从而获得相似度值[11]。在国内,刘群、李素建提出基于《知网》的词汇相似度计算,以义原的相似度计算方法、集合和特征结构相似度计算方法为基础,利用《知网》进行相似度计算[12]。晋耀红提出基于语境框架的文本相似度计算方法,把文本内容抽象成领域、情景、背景 3 个侧面,实现文本间语义相似程度的量化[13]。颜瑞武、成晓、甘利人提出基于领域本体和概念向量的中文文本相似性测度研究,通过 12 种抽象型关系连接,构成领域本体的网状结构计算相似度[14]。穗志方、俞士汶提出基于骨架依存树的语句相似度计算模型,即通过基于骨架依存树的语句相

本文受国家科技支撑计划课题:药品在线交易服务技术与平台研发及示范应用(2012BAH19F01)资助。
向林泓(1985—),男,助理研究员,主要研究方向为自然语言处理、数据挖掘、机器学习;张 炬(1964—),男,研究员,主要研究方向为数据挖掘、计算机数学。

似度计算模型 SBCM 进行文本相似度计算[15]。

TF-IDF 方法也被广泛应用到文本相似度计算中，该方法将文本表示为文中出现的 n 个加权词项组成的向量[16]。TF-IDF 的主要思想是：如果某个词或短语在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。IDF 的主要思想是：如果包含词条的文档越少，则说明词条具有很好的类别区分能力[17]。那么根据上述概念可以得到每一个词项 w_i 的 TF-IDF 值：

$$TF\text{-}IDF(w_i)=tf_j(w_i)*\log(N/df(w_i)) \quad (1)$$

式中，$tf_j(w_i)$ 表示当前词项 w_i 在文本 j 中出现的频率，N 表示文本集合中所有文本的综述，$df(w_i)$ 表示文本集合中有多少篇文本出现了当前词项 w_i。通过对文本集合中的每一个词项都进行上诉分析，得到每一篇文本中每一个词项的 TF-IDF 值，然后再利用这些 TF-IDF 值为每一篇文本建立一个向量模型，通过计算向量间的余弦相似度来确定文本之间的相似度。

由于医学医药领域文本的特殊性，传统的 TF-IDF 的文本相似度方法并没有取得很好的效果。本文在传统 TF-IDF 的基础之上，结合 Relative-IDF 相似度计算方法。首先选取文本中的关键词项，有效地降低文本模型的维度，为文本相似度计算提供一个合适的表征模型。其次，通过分析关键词项的相对语义，给出文本相似度的定义。

3 Relative-IDF

传统 IDF 是一种统计方法，用以评估一个词项相对一个文件集或一个语料库的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库出现的频率成反比下降。Relative-IDF 则在于评估医药词周围词项的重要程度，它去除了文档边界，将所有的文档或者一个语料库中当作整体，以医药关键词作为分界符，并以此为基准计算周围词语的相对于医药关键词的权重，医药关键词则充当了传统 IDF 里文档边界的作用。于是词项 w 的 Relative-IDF 值的计算为：

$$RIDF_i^T(w)=(1-\delta)\cdot\log(|T|/KTF_i^T(w))+\delta\cdot\log(\mathrm{Max}(NTF_i^T(w))/NTF_i^T(w)) \quad (2)$$

式中，T 代表了语料中出现的医药关键字，类似于 IDF 中所有文档数量；i 指定了医药关键字的统计宽度，如果 i 为 5，则代表了统计医药关键字前 5 个和后 5 个词项；NTF 代表了以 i 为宽度的词项中 w 出现的次数；KTF 则代表了词项 w 在医药关键词前面出现的次数；δ 是一个常量。

4 医药数据相似度算法分析和设计

4.1 药品数据简介

药品数据主要由药品通用名、生产企业、药品规格、剂型、和转换系数 5 个参数确定。药品数据参数都是上下文无关短语，并且具有“人为因素主导、标点符号辅助”的特点。在进行药品数据相似度计算时，会分为两组数据。一组数据是标准数据，标准数据是以《国家药品标准化学药说明书》为基础，符合化学药品国家标准说明书的数据。原始数据是各个地方单位、药品制造企业根据药品规格、剂型、转换系数，人为编写的数据，数据的编写具有很大的自由度，并且各个地方对于同一种药品的描述也不尽相同，从而加大了原始数据和标准数据之间匹配的难度。表 1 展示了一组标准数据和原始数据的常见情况。

表 1

	标准数据	标准数据	原始数据
药品名	化痰消咳片	化痰消咳片	化痰消咳片
生产企业	广东逸舒制药有限公司	广东逸舒制药有限公司	广东逸舒制药有限公司
规格	300mg：3mg（亚硫酸氢钾苯丙酮）-20mg（鱼腥草素钠）	0.5g（含鱼腥草素钠 40mg：亚硫酸氢钾苯丙酮 6mg）	0.3g（含鱼腥草素钠 20mg、亚硫酸氢钾苯丙酮 3mg）
剂型	片剂（薄膜衣）	薄膜衣片	薄膜衣片
转换系数	60	60	60

从表 1 可以发现，标准数据和原始数据之间的差异重点在于药品的“规格”参数，而剂型和转换系数参数起到的作用相对较小，于是将两组数据的相似度计算转换成为两组数据的“规格”参数数据的差异计算。药品数据规格参数具有“人为因素主导、标点符号辅助”的特点。参数的填写是因人而异的，并且地方区域差异明显，填写同一种规格参数都会有不一样的表达方式。虽然表达方式多样化，但是他们通常都会以标点符号作为辅助工具对文本进行辅助性解释。比如表 1 中广东逸舒制药有限公司生产的化痰消咳片的其中一种参数规格为“300mg：3mg（亚硫酸氢钾苯丙酮）-20mg（鱼腥草素钠）”，那么双括号“（）”就可以理解为对 300mg：3mg 的补充说明，而“-”则可以理解为并行符号，前面一段和后面一段的权重相等。基于这种思路就可以将药品数据规格参数转换成为结构化层次数据，从而进行结构化数据比较，而不仅仅是字符串数据匹配流程，从而大大提高了数据相似度计算的精确度。

4.2 药品规格数据结构化表示

药品数据规格参数具有“人为因素主导、标点符号辅助”的特点。经研究发现，规格参数里面的数据经过标点符号规整之后，总是会由药品名、数量、单位 3 个因素组成，于是就可以将规格参数转换成为“分层链表结构”的数据结构进行表示。比如表 1 中化痰消咳片的其中一种规格为“300mg：3mg（亚硫酸氢钾苯丙酮）-20mg（鱼腥草素钠）”，就可以将其转换成为图 1 所示的分层链表结构。

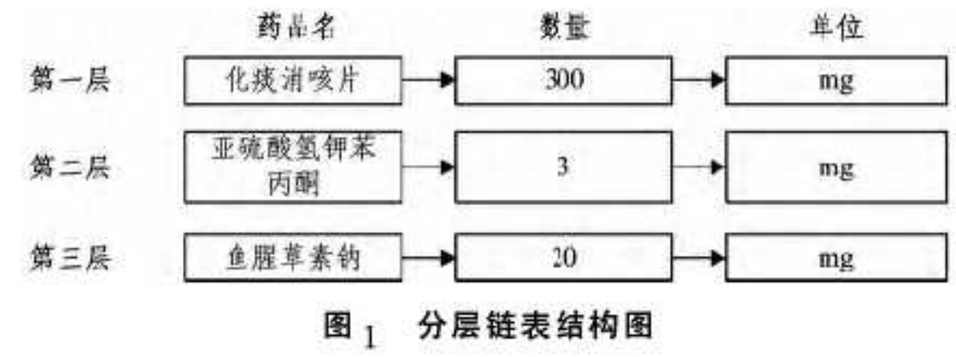


图 1 分层链表结构图

4.3 药品规格数据匹配算法设计

首先需要将标准数据和采集数据的规格参数都转换成为分层链表数据结构，再计算分层链表结构所有元素之间的两两之间的相似度，然后再进行加权平均。于是采用下面的方法进行集合的相似度计算：

1）将药品规格数据形成语料库，用 Relative-IDF 在此语料库上进行计算，将计算结果保存在数据库中。

2）将要匹配的标准数据和采集数据转换成为分层链表结构，计算所有分层结构的两两之间的相似度，每一层的相似度计算以 Relative-IDF 为基础，如果匹配数据中医药关键字不

同，则认为不匹配；如果医药关键字相同，则以医药关键字周围词项的 Relative-IDF 差值作为相似度值，将相似度值最高的那两层配对，并从集合里面剔除。

3）重复步骤 1），直到标准数据和采集数据所有分层两两之间配对完毕。

4）没有取得配对的层与空层相对应。

5）利用权值计算平均相似度，返回。

上述算法首先根据相似度的取值建立两个集合元素的一一对应关系，然后计算两个集合的相似度。集合的相似度等于其元素对的相似度的加权平均，因为集合的元素之间都是平等的，所以将所有的权值取成相同的。于是集合的相似度等于其元素对的相似度的算术平均。

5 实验结果及分析

本次实验选取 2012 年 8 月份左右的数据进行药品相似度比较。表 2 是在该实验数据上进行 Relative-IDF 的计算结果。

表 2

词项	频率	方差	RIDF
含	3907	154.39	13.05
口服	1144	2.97	14.69
和	11174	2184.88	11.17
服用	1704	10.08	14.16
与	10732	2360.29	11.52
使用	1683	10.46	14.02
注射	921	8.41	17.73
加入	794	5.43	19.13
复方	614	2.45	18.02
甲基	470	3.02	24.12
含有	1445	20.06	13.65
盐酸	128	0.49	26.39

从表 2 可以看到，在医学医疗领域表达同一个意思的词汇的 RIDF 值非常接近。比如，“含有”和“含”的 RIDF 分别为 13.65 和 13.05；“口服”、“服用”和“使用”的 RIDF 值分别是 14.69、14.16 和 14.02；而“盐酸”和“甲基”这类复合医药构成词的 RIDF 值就大了很多。

为了验证 Relative-IDF 在药品数据相似度计算的有效性和准确性，将进行以下 3 组实验。

实验 1　选取采集数据库和标准数据库中已经成功匹配的一组数据，该组数据的相似度理论值应该为 100%。表 3 显示了该组数据。

表 3

	采集数据	标准数据
药品名	注射用阿莫西林钠舒巴坦钠	注射用阿莫西林钠舒巴坦钠
规格	1.5g：500mg（包含舒巴坦）—含有阿莫西林 1g	1.5g（舒巴坦：500mg）—（阿莫西林：1000mg）

分别用余弦定理、编辑距离和 Relative-IDF 3 种方法进行相似度计算。表 4 显示了该组数据和相似度计算结果。

表 4

	TF-IDF	编辑距离	RIDF
相似度值	64.2%	83.3%	91.4%

可以看到，利用 Relative-IDF 大大提高了药品相似度计算的准确度。

实验 2　选取数据库中 12 万条已经匹配完成的数据，这些数据的相似度理论值应该为 100%。对这些数据分别用 TF-IDF、编辑距离和 Relative-IDF 计算相似度，并统计相似度值分布情况。表 5 和图 2 显示了结果。

表 5

相似度值区间	0～50%	50%～80%	80～100%
TF-IDF	23.1%	66.3%	10.6%
编辑距离	33.4%	53.8%	12.8%
RIDF	9.87%	25.9%	64.23%

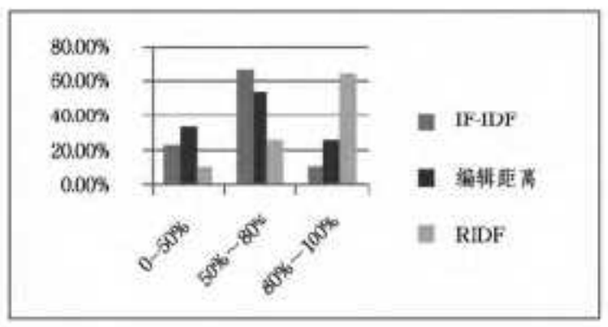


图 2

从表 5 可以看到，TF-IDF 和编辑距离由于没有考虑到药品数据的特殊性，只是传统的字符串比较，因此得到的相似度值会比较低，相似度值主要分布在 80% 以下。而 RIDF 更能够准确地反映医药数据的真实匹配情况。

实验 3　选取数据库中 12 万条已经匹配完成的数据，再随机挑选 1000 条数据作为干扰数据，分别对这些数据用 TF-IDF、编辑距离和 Relative-IDF 3 种方法进行相似度计算，检验在干扰数据下，各种方法在数据匹配的表现，本文选取 P 值作为测量标准。表 6 显示了统计结果。

P = 正确匹配的医药数据对 / 所有匹配的医药数据对

表 6

	正确匹配数量	P
TF-IDF	89345	74.45%
编辑距离	95320	79.43%
RIDF	115400	96.17%

从表 6 可以看出，由于药品数据的特殊性，利用 TF-IDF 和编辑距离在进行相似度计算的时候并没有取到很好地效果。TF-IDF 表现得最为糟糕，只有 89345 条数据正确匹配，正确率只有 74.45%，而利用 Relative-IDF 则有 115400 条数据正确匹配，正确率达到了 96.71%。从实验结果可以看到，利用 Relative-IDF 很好地解决了药品数据相似度匹配问题。

结束语　本文在传统 TF-IDF 的基础之上，利用 Relative-IDF 解决了医药数据相似度计算的问题。与传统的基于 TF-IDF 算法不同，本文所提算法通过寻找合适的边界和关键字周围文本的阈值，有效地降低了传统向量模型表示文本所带来的高维影响，很好地解决了传统文本相似度在药品数据相似度中的缺陷，大大提高了药品相似度计算的精准度。

本文后续研究将在现有的 Relative-IDF 基础之上，进一步探讨如何将 Relative-IDF 应用在文本聚类、文本分类、自动摘要中，进一步深入研究文本相似度所蕴含的语义相似特征，更好地提高文本相似度效果，推动语义网的应用发展。

参 考 文 献

[1] Fung B C M, Wang K, Ester M. Hierarchicaldocument clustering Wang John ed[C]// The Encyclopedia of Data Ware housing and Mining. IdeaGroup-2005: 970-975

[2] Hall P, Dowling G. Approximatestring matching[J]. Computing Survey, 1980, 12(4): 381-402

[3] Coelho T,Calado P,Souza L,et al. Image retrieval using multiple evidence ranking[J]. IEEE Transactions on Knowledge and Data Engineering,2009,16(4):408-417

[4] Theobald M,Siddharth J,Paepcke A. SpotSigs:Robust and efficient near duplicate detection in large Web collections[C]//Proceedings of the 31st Annual International ACMSIGIR Conference on Research and Development in Information Retrieval. Singapore,2008:563-570

[5] Erkan G,Radev D. Lexrank:Graphbased lexical centrality as salienceinin text summarization[J]. Journal of Artificial Intelligence Research,2009,22(7):457-479

[6] Ko Y,Park J,Seo J. Improving text categorizationusingthe importance of sentences[J]. Information Processing and Management,2010,40(1):6579

[7] 中医药学语言系统 Wi-ki[EB/OL]. http://www.cintcm.com/yuyan/index.htm,2013-05-01

[8] 医药在线交易服务平台 Wi-ki[EB/OL]. http://www.yaol.cn/,2013-07-01

[9] VSMWi-ki[EB/OL]. http://en.wikipedia.org/wiki/VSM,2012-03-19

[10] Oleshchuk V. Ontology based semantic similarity comparison of documents [C]//14th International Workshop on Database and Expert Systems Applications,2003. 2003,1

[11] Ding C H Q. Research on Optimize Technology in Latent Semantic Indexing Based on Semantic Block[C]//Chinese Conference on Pattern Recognition,2009(CCPR 2009). 2009

[12] 刘群,李素建. 基于《知网》的词汇语义相似度计算[C]//第三届汉语词汇语义学研讨会,2002

[13] 晋耀红. 基于语境框架的文本相似度计算[J]. 计算机工程与应用,2004(16)

[14] 颜端武,成晓,甘利人. 基于领域本体和概念向量的中文文本相似性测度研究[J]. 中国图书馆学报,2007,33(6)

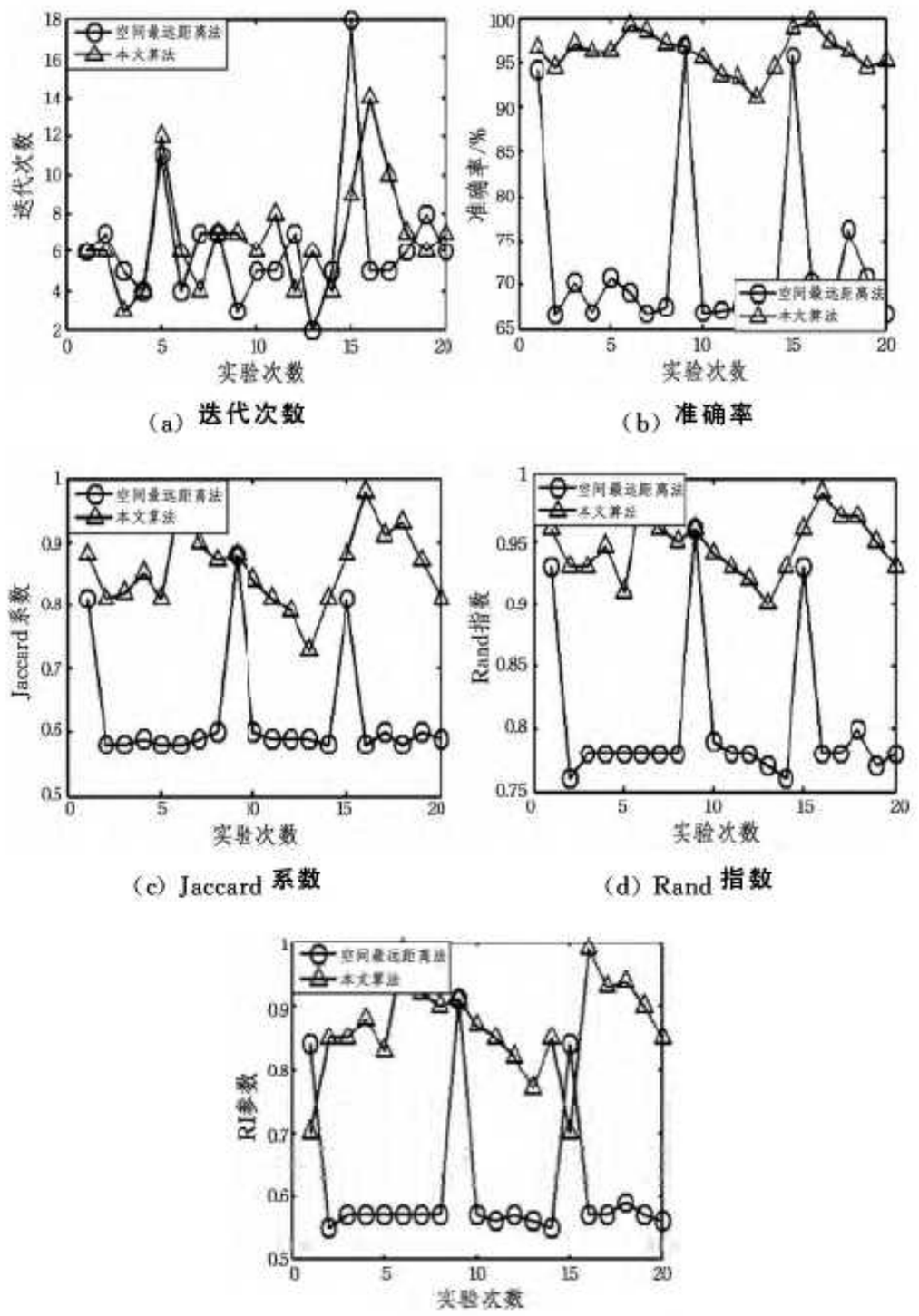
[15] 穗志方,俞士汶. 基于骨架依存树的语句相似度计算模型[C]//中文信息处理国际会议. 1998

[16] 黄慧,印鉴,侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报,2011(5):856-864

[17] TF-IDFWi-ki[EB/OL]. http://zh.wikipedia.org/wiki/TF-IDF,2013-05-01

(上接第 408 页)

类准确率达到 90% 以上,具有优良的抗噪性能。



(a) 迭代次数 (b) 准确率

(c) Jaccard 系数 (d) Rand 指数

(e) RI 参数

图 3 空间距离法与本文算法在噪声数据集上的分类效果比较

结束语 本文针对 K-means 聚类算法中初始质心的选择,改进基于空间最远距离方法存在的不足,采用空间距离差法寻找初始聚类质心,从而降低噪声点对初始聚类质心选择的影响,以提高 K-means 聚类分析的准确性。通过使用 UCI 数据库中的数据集和带有噪声点的人工随机数据测试,本文所提算法效果稳定,处理噪声数据集的能力得到提升,能够用于实际采集的含有噪声大数据集的分析和处理。

参 考 文 献

[1] 欧陈委. K 均值聚类算法的研究与改进[D]. 长沙:长沙理工大学,2011

[2] 张俊生. 数据挖掘中的聚类方法及其应用研究[D]. 天津:天津理工大学,2012

[3] Anil K J. Data clustering:50 year beyond K-Means[J]. Pattern Recognition Letters,2010,31(08):651-666

[4] 吴晓蓉. K-均值聚类算法初始质心选取相关问题的研究[D]. 长沙:湖南大学,2008

[5] Fayyad U,Reina C,Bradley P S. Initialization of Iterative Refinement Clustering Algorithms[C]//Proc of the Fourth International Conference on Knowledge Discovery and Data Mining. 1998:194-198

[6] Adil M,Bafirov,Julien U. Fast modified global k-means algorithm for incremental cluster construction[J]. Pattern Recognition,2011,44(4):866-876

[7] 曹志宇,张忠林,李元韬. 快速查找初始聚类质心的 K-means 算法 [J]. 兰州交通大学学报,2009,28(6):15-18

[8] 张真,任贺宇. 一种基于动态网格技术的 K-means 初始质心选取算法[J]. 微电子学与计算机,2013,30(6):101-104

[9] 谢娟英,蒋帅,王春霞. 一种改进的全局 K-均值聚类算法[J]. 陕西师范大学学报,2010,38(2):18-22

[10] 谢娟英,郭文娟,谢维信. 基于样本空间分布密度的初始聚类中心优化 K-均值算法[J]. 计算机应用研究,2012,29(3):888-892

[11] 杨燕,靳蕃,Kamel M. 聚类有效性评价综述[J]. 计算机应用研究,2008,25(6):1631-1632

[12] 张惟皎,刘春煌,李芳玉. 聚类质量的评价方法[J]. 计算机工程,2005,31(20):10-12

[13] Hubert L,Arabie P. Comparing partitions [J]. Journal of Classification,1985,2(1):193-218

[14] 王纵虎,刘志镜子,陈东辉. 基于粒子群优化的模糊 C-均值聚类算法研究[J]. 计算机科学,2012,39(9):166-169