编辑距离算法在科研基金名称数据分析中的应用

赵胜钢 李军莲 陈颖
（中国医学科学院医学信息研究所，北京 100020）

摘要：通过对科研基金名称数据特点和文本数据聚类方法的分析，提出并实现了基于编辑距离算法（Levenshtein Distance）的科研基金名称数据分析方法，该算法首先通过设定相似度方式对科研基金名称数据进行聚类形成数据分组，再对分组数据进行二次聚类计算出组的相似度之和，并据此判定数据聚类中心。该方法已经成功应用于中国医学科学院医学信息研究所的医学文献基金数据处理。

关键词：文本挖掘；聚类算法；科研基金；编辑距离算法
分类号：TP391
DOI：10.3772/j.issn.1673—2286.2014.05.009

1 引言

科研基金是为支持特定科学研究而设立的立项资助，如国家自然科学基金、国家社会科学基金等。这些科学研究所形成的论文、报告等科技文献，一般会在首页页脚标注出其所依托的科研基金。对科技文献中包含的科研基金数据进行采集、整理以及规范化加工，可以实现按科研基金对科技文献进行检索、统计和分析，为评价科研基金的作用提供客观依据。

科研基金有政府、高校、研究所、企业等多个不同来源，目前尚没有一个统一的科研基金标准名称目录；不同的期刊对基金内容的著录格式没有统一的标准，作者在标注基金时存在很大的随意性；在用外文标注基金时存在着基金名称翻译的多样性；在对文献中的基金数据进行采集过程中，也会存在OCR识别、数据录入等操作产生的错误。上述原因导致了从文献中采集的基金数据中会存在同一基金名称有多种表达形式等问题。如何对不同表达形式的基金名称数据进行归一化处理是实现基金数据分析和评价中需要解决的重要问题。

从科技文献中采集来的科研基金数据一般是半结构或无结构的自由文本数据。从自由文本中提取有意义的信息和知识是从属于机器学习和数挖掘领域的问题，而文本数据挖掘已经变成了数据挖掘领域中最活跃的研究课题之一[1]。对基金数据进行加工的目标是从这些自由文本数据中挖掘提取出其中包含的基金名称、基金支持的项目名称和项目编号等信息，并依据该基金的通用名称（后文中称为“基金标准名称”）对基金数据进行标识、分类，为进一步的查询分析提供支持。考虑到这些文本数据的半结构化或无结构化的特点，基金数据挖掘的基本思路是：在对基金数据进行清洗的基础之上，先对清洗出的基金名称数据按相似度进行聚类，形成聚类组；再从聚类组中分析、提取出可能的基金标准名称以及这些基金标准名称的别名，最后再对这些基金标准名称及别名进行分析、验证，从而形成一套基金名称目录体系。这套目录体系可以成为对基金数据进一步标识分类的依据。

本文在分析了科研基金名称数据的特点以及对文本聚类方法进行分析比较的基础上，提出了基于编辑距离算法的基金名称数据分析方法和处理流程。文章提出的方法已经在医学文献基金数据中进行了成功应用。通过实践，作者进一步提出了对编辑距离算法进行优化以提高基金数据聚类的效率和准确度的思路。

2 基金名称数据特点及聚类方法选择

2.1 原始基金数据和基金名称数据

科研基金数据的最初来源是科技文献（论文、报告等）的作者在文献特定位置（一般为文献首页的脚注）

对基金进行的标注。这些标注信息经人工采集后形成原始基金数据。原始基金数据以自由文本形式存在，格式内容都不规范，必须经过清洗，才能形成可供分析的基金名称数据。表1示例了一条原始基金数据及其清洗后形成的基金名称数据。

表1 原始基金数据清洗结果示例

原始基金数据	基金名称数据
国家自然科学基金（No.30560186）；广西科技厅广西科学基金项目（No.0728128）；广西科技厅广西科学基金项目（No.0991147）	国家自然科学基金 广西科技厅广西科学基金 广西科技厅广西科学基金

表中左列为从一篇文献中提取的原始基金数据，标注出了支持该研究的基金项目及项目编号。右列为对该条数据经过清洗后提出的基金名称数据（尚未去重）。基金清洗过程并非本文研究目标，在此略去。

2.2 基金名称表现形式的多样性

对于同一个基金名称，原始基金数据可能有多种表现形式，如别名、错名、简称、外文翻译等。如对于“国家重点基础研究发展计划”基金，其官网上的描述[7]是“1997年6月4日，原国家科技领导小组第三次会议决定要制定和实施《国家重点基础研究发展规划》，随后由科技部组织实施了国家重点基础研究发展计划（亦称973计划）”，由此可确定该基金的标准名为“国家重点基础研究发展计划”。实际上在对清洗后的基金名称数据经过分析后，针对该基金提取出了超过300种表现形式，典型名称如表2所示。

表2 基金别名示例

基金标准名	基金名称数据
国家重点基础研究发展计划	国家重点基础研究发展计划 国家重点研究发展基础计划 国家重点基础研完发展规划 国家重点基础研究规划项目 “973”项目 973项目 973 计划课题 973 国家项目 973基金资助 国家九七三课题 Grants from China National 973 Ministry of Science and Technology 973 program Major State Basic Research Development Program The China National Basic Research Program The Chinese National Basic Research Program

表2中示例了由于本身存在的别名、作者标注的不规范性、随意性、英文翻译的多样性和识别错误等原因形成的同一个基金名称的多种表现形式。

基金名称表达形式的多样性给按基金检索分析文献带来了困扰，必须对这些不同表达方式进行归一化处理才能保证基金检索结果的准确性和完整性。

2.3 基金名称数据聚类需求

通过对基金名称数据表达方式多样性的特点和形成的原因的分析，对其进行归一化处理的可选方案是先将可能属于同一个基金的基金名称数据进行分组，以找出这些数据背后包含的基金标准名称和别名，建立起基金标准名和别名之间的映射关系。文本聚类方法是解决上述问题的有效工具。

聚类（clustering）是一个将数据集划分为若干组或类的过程。聚类的目的是使同一个组内的数据对象具有较高的相似度，而不同组中的数据对象尽量不相似[2]。把数据库中的对象分类是数据挖掘的基本操作，其准则是使属于同一类的个体间距离尽可能小，而不同类个体间距离尽可能大。为了找到效率高、通用性强的聚类方法，人们从不同角度提出了近百种聚类方法，这些聚类算法适用于特定的问题及用户[3]。

文本聚类作为数据挖掘的研究分支，在对半结构化和非结构化数据提取有效规律和规则方面有着明显的优势。在处理不同数据集时，应根据数据集的维度和组织情况选择最适用的挖掘分类方法[4]。

依据著名的聚类假设：同类的文本相似度较大，不同类的文本相似度较小，在对科研基金名称数据进行聚类分析时，其基本需求简单而直接，就是“依据相似度”对基金名称数据进行聚类。因此对于聚类算法的要求就是输入“相似度”一个参数，相似度以百分比表示，100%意味着两条基金记录内容完全相同，90%意味着两条基金记录每10个字符中有9个字符相同。通过不断地降低相似度，可以逐步把相同、相似、基本相似等基金名称记录聚类在一起。

2.4 编辑距离法选择依据

文本聚类是一种“无教师”的机器学习方法，它从给定的文本本身出发，根据文档特征词向量，将相关者聚为一类[4]。文本聚类方法包括基于划分的聚类、基于层次的聚类、基于密度的聚类、基于网格的聚类、基于模型的聚类以及基于字符串相似度的聚类。基于划分的聚类方法主要包括K-means、X-means、K-medoid和ISODATA；基于层次的聚类主要包括Birch Cluster、Cure Cluster、Single Link Cluster、Complete Link Cluster和Average Link Cluster；基于密度的聚类主要包括DBScan和Optics；基于网格的聚类主要包括Sting Cluster和Clique Cluster，Cobweb 属于基于模型的聚类[5]。这些方法各自有不同的特点，应用于不同的文本数据集。

根据对基金名称数据表达方式多样性的特点和其聚类需求的分析，通过对不同文本数据聚类方法特点的研究比较，作者认为采用一种基于字符串相似匹配技术的算法比较适于科研基金名称聚类。现有的计算字符串相似度的方法按照计算所依据的特征的不同，可以划分为3种基本方法：基于字面相似的方法、基于统计关联的方法、基于语义相似的方法，以及综合3种基本方法的多层特征方法。基于字面相似的计算方法主要有基于编辑距离的计算方法和基于相同字或词的方法。其中编辑距离法（Levenshtein Distance，LD）应用广泛，计算方法相对成熟[6]。

2.5 编辑距离算法

编辑距离算法（Levenshtein Distance，LD）由Levenshtein于1966年中提出，其原理是对源字符串通过“插入一个字符”、“删除一个字符”和“替换一个字符”这三种“编辑”操作，将源字符串变换为目标字符串。编辑距离是指由源字符串变化到目标字符串所需要的最小编辑操作的数量，而相似度可以定义为编辑距离与源字符串和目标字符串这两个字符串长度的最大值的比值[6]。

假设源字符串S与目标字符串T长度的最大值为Lmax（S,T），编辑距离为L，相似度为SA，那么

$SA=（1 - L /L_{max}（S,T））*100\%$

L=0意味着不需要任何编辑操作两个字符串就相同，计算出的相似度为

SA=（1-0）*100%=100%

$L=L_{max}（S,T）$意味着对S字符串中的每个字符都进行一次编辑才能变成T。假设S字符串的长度为L_s，T字符串的长度为L_t，则分为$L_s=L_t$、$L_s>L_t$和$L_s<L_t$这3种情况：

（1）如果$L_s= L_t$，则$L_{max}（S,T）=Max（L_s,L_t）=L_s= L_t$，需要对S字符串进行 L_t次“替换”操作（即S中每个字符都换成T中对应的字符），S就转换成了T，此时$L=L_{max}（S,T）=L_s=L_t$, $SA=（1-L_{max}/L_{max}）*100\% =0\%$，意味着S和T完全没有相似性

（2）如果$L_s>L_t$，则$L_{max}（S,T）=Max（L_s,L_t）=L_s$，需要先对S字符串进行 L_t次“替换”操作，再进行（L_s-L_t）次“删除”操作，S就转换成了T，此时 $L=L_t+（L_s-L_t）= L_s= L_{max}（S,T）$，SA=0%

（3）同理如果$L_s<L_t$，则$L_{max}=Max（L_s,L_t）=L_t$，需要先对S字符串进行L_s次“替换”操作，再进行（L_t-L_s）次“插入”操作，S就转换成了T，此时 $L=L_s+（L_t-L_s）= L_t = L_{max}（S,T）$，SA =0%

因此可推断出$0<=L<= L_{max}$，即0%<= SA<=100%.意味着只要给定一个介于0%和100%之间的相似度参数，就可以对S和T两个字符串进行比较后对其是否符合给定的相似度进行判定。

编辑距离是一个比较成熟的算法，有许多开源的实现方式。

3 编辑距离算法在基金名称数据聚类中的应用

由于编辑距离算法可以计算出任意两个字符串的相似度，也可以根据一个表示相似度的参数就可以对字符串进行分组，符合基金名称数据聚类的要求，因此可以应用于其聚类分析。虽然原理比较简单，但具体实现上要考虑很多影响因素。

3.1 应用编辑距离算法进行聚类分组

假设原始基金数据已经清洗完毕并按拼音字母顺序排序，放入数据库DS（其中的数据已经去掉了噪音数据，并经过拆分，保证一条记录里只包含一条基金名称数据，并且这些基金名称数据已经去重），建立数据库DD，存放聚类形成的数据组，聚类分组基本流程如图1所示。

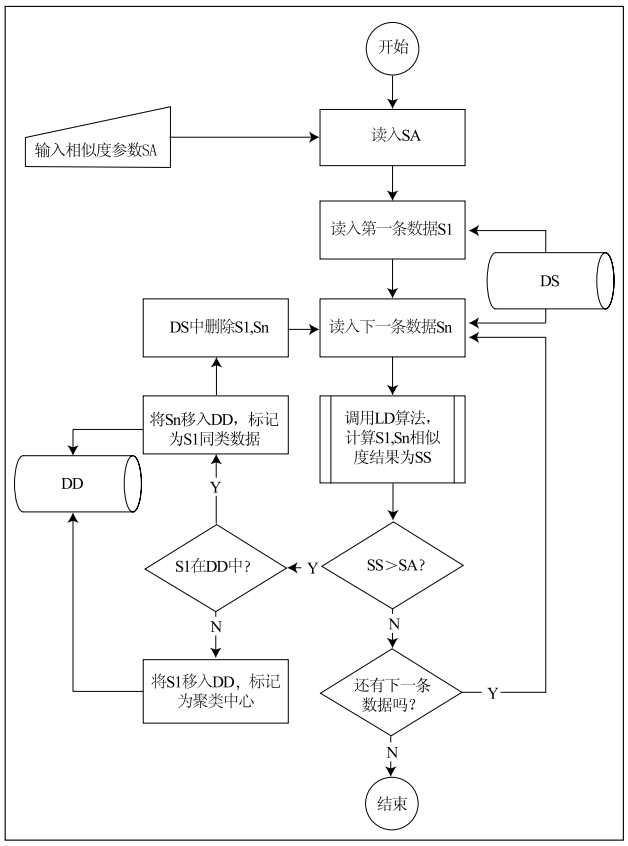

图1 应用LD算法聚类基本流程

在图1所示流程中，先给出相似度的判断参数SA，然后读入第一条数据S1作为聚类中心数据，接着读入下一条数据Sn并与S1进行比较，如果相似度大于等于SA，则这两条数据都从DS数据库中移动到DD中，否则什么也不做，读入下一条数据再跟S1进行比较，直到DD中所有的数据完成比较。

完成这个流程后，可能有两个结果，一种是DD中为空，DS中数据没有变化，这种结果意味着没有形成以S1为中心的相似度不小于SA的聚类数据；一种是DD中保存了一组S1为中心的相似度不小于SA的聚类数据，同时所有保存在DD中的数据都已经从DS中删除。这样就可以从DS中剩下的数据中再选取第一个数据作为“S1”进行聚类。通过对DS中的数据进行多次循环，最终在DD中保存了所有满足SA条件的多个基金数据聚类组，不满足条件的数据保留在DS中。

在此流程中SA是数据能否进入DD的门槛。通过调整SA的值，可以调整DD中的组数和每个组包含的成员数。极端情况下，如果SA=100%，则只有完全相同的数据才能进入DD，在DS中的数据已经去重的前提下，DD中必然为空；如果SA=0%，则所有的数据都会进入DD，形成一个以S1为聚类中心的聚类组，当然这种条件的聚类结果并无意义。

以表2所示的数据为样本数据，通过调整SA经过几轮循环，在DD中可以形成3个聚类组，如图2所示。

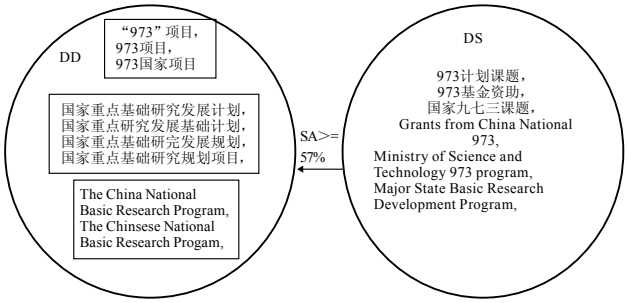

图2 样本数据聚类结果

图2中经过调整SA到 57%，在DD中形成了3个聚类组，而没有通过聚类相似度标准的数据留在了DS中。

3.2 应用2次聚类形成聚类中心

考虑在DD中的数据，假设目前DD中只保留了一组数据（为说明问题，取样本数据为：S1=国家重点基础研究发展计划，S2=国家重点研究发展基础计划，S3=国家重点基础研究发展规划，S4=国家重点基础研究规划项目），这组数据以S1为聚类中心，包含了S2、S3、S4这3个同组数据。由于S1是在第一次聚类中按字顺顺序挑选出来的，现在的问题是S1是不是这组4个数据中最合适的聚类中心呢？这个问题的意义在于，合适的聚类中心可能代表了该组数据的最通用表达方式。解决这个问题的依据是，尽管同一个基金的基金名称有各种表达方式，但可以认为各种表达方式都与基金标准表达方式最相似，而两个非标准表达方式之间的相似度较小。这个判断为计算机自动查找合适的聚类中心提供了依据。

以S1、S2、S3、S4这组数据为例，如何判断出哪个数据最可能是标准名称数据呢？可以再次应用编辑距离算法，轮流以每个数据为聚类中心，对这组数据进行2次聚类，算出以不同数据为聚类中心的该组数据的总体相似度，再对聚类结果进行分析比较。算法是每次聚类，可以在组中先选定第一个数据作为聚类中心，其他数据与聚类中心数据进行比较，算出各自的与聚类中心

的相似度，最后再将该组每个数据的相似度相加，形成以选定数据作为聚类中心的相似度之和；对组中每个数据依次进行聚类，就可以计算出组中以每个数据为聚类中心得出的该组相似度之和数据。如图3所。

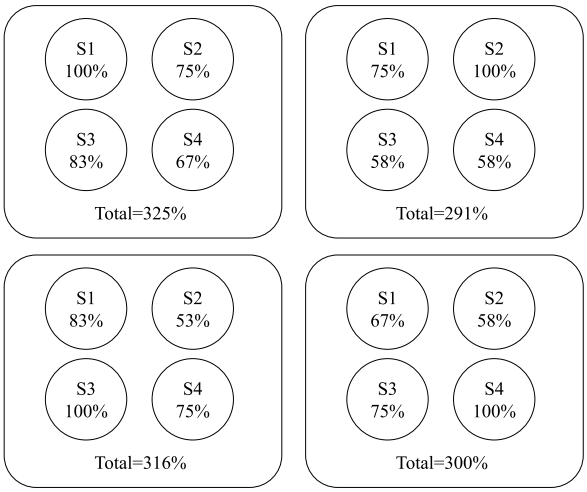

图3 二次聚类后相似度分析

在图3中，依次以S1、S2、S3、S4为聚类中心，计算其他同组数据与聚类中心的相似度，形成4组数据，这样可以算出每组的相似度的总和。以第一组为例，以S1为聚类中心，S1与其自身的相似度为100%，S2、S2、S4与S1的相似度分别为75%、83%、67%。在上例中，分别以S1、S2、S3、S4为聚类中心的组相似度总和分别为325%、291%、316%、300%，由此可以判断出此组数据以S1为聚类中心更为合适。

在这种算法中，如果出现多个相似度之和相同的情况，任选其中一个数据作为聚类中心即可。

3.3 聚类算法的优化

通过对聚类算法的分析和应用测试，发现该算法存在以下两个需要优化的地方，一是该算法需要大量循环运算，随着参加聚类的数据量线性增加，其完成运算需要的时间呈指数性增长。二是目前聚类算法都是围绕初始S1为中心进行，由于S1按字顺产生，排序方法对聚类结果产生干扰。必须对其算法进行优化，以提高数据处理的效率和聚类结果的准确性。

目前优化的思路有以下几种：

（1）数据分批次处理

考虑到需要完成的时间与参加聚类的数据量呈指数增长的特点，将数据按数量分成不同批次处理，以节省处理时间，提高处理效率。以处理100万条数据为例，100万条数据聚类一次处理所需要的时间，要比10万条一次，处理10次的时间要长很多。具体差别因数据的复杂程度有所区别，一次测试录得的结果是100万条聚类一次，耗时近96个小时，而10万条处理一次需要近一个小时，分批处理总耗时10个小时左右。

然而分批次处理的结果，由于每次聚类覆盖的数据面不全，其得出的聚类中心与聚类组需要进一步处理组合，方法就是下面提到的“建立标准数据聚类中心”的方式。

（2）建立标准数据聚类中心

建立标准数据聚类中心的含义，就是将通过聚类已经挑选出的每个组中作为聚类中心的数据形成一个集合，加入到未处理的数据集合中，并作为指定的聚类中心，再次进行聚类操作，然后以聚类中心数据为依据将新形成的聚类组与已经形成的聚类组进行合并。对于已经人工确定为基金标准名称数据的数据，指定为聚类组的确定聚类中心，不再进行2次聚类。这种方法解决了数据分批次聚类带来的问题，同时又对人工干预聚类形成的结果进行了保存与再利用。

（3）基于分词的聚类

传统的聚类数据处理以字符为单位进行处理，如S=“国家自然科学基金”，T=“国家自然基金”，传统的方式是将S分成 8个字，T分成 6个字，运行LD方法进行聚类，此时Lmax=8，相似距离L=2（S删除两个字符转换为T）。如果先对S和T进行分词操作，将S分为由“国家”“自然”“科学”“基金”四个词组成的字串，T分为“国家”“自然”“基金”三个词组成的字串，再以词为单位运用LD方法进行聚类，此时Lmax=4，L=1 （S删除一个词组转换为T），考虑到对字符串进行分词是个线性操作，两个字符串的LD比较是一个矩阵操作，进行分词操作所需要的时间会远小于由于减少矩阵运算维度所节省的时间，基于分词的聚类运行效率会大幅提高。

（4）随机选择S1进行聚类

为消除通过人为指定排序方式选定S1对聚类结果的影响，可通过随机数转换为数据顺序号的方式挑选S1。此种方式虽然规避了人为排序的影响，但可能会带来类心漂移等新问题，其可行性需要进一步评估。

4 结语

本文讨论的聚类方法已经在中国医学科学院医学信息研究所进行的医学文献基金数据处理中得到了应用，取得了较好的结果。但如果进一步提高文本聚类的效率与准确度，还有很多工作要做。如文中提到的4种优化方法，其分批次处理法和建立标准数据聚类中心处理法已经在研制的基金处理系统中实现，但基于分词的聚类方法和随机产生初始聚类中心的方法目前只是一个设想，有待于进一步研究和验证。同时能否使用其他更有效的文本数据聚类方法进一步提高基金数据处理的效率和准确度也有待于深入研究。

聚类分析是文本类数据挖掘的一个有力工具，而编辑距离方法是其中的一个常用方法，编辑距离方法适合于对聚类结果要求相对简单的需求，应用起来比较方便。本文讨论的将编辑距离算法应用于科研基金数据的处理方法对其他文本类数据的处理挖掘也具有参考价值。

参考文献

[1] ZHOU X Z, PENG Y H, LIU B Y. Text mining for traditional Chinese medical knowledge discovery: A survey [J]. Journal of Biomedical Informatics, 2010, 43(4): 650-660.

[2] 徐东亮,董开坤,李斌,等.基于文本挖掘的聚类算法研究[J].微计算机信息,2011(2):168-169.

[3] 张红云,刘向东,段晓东,等.数据挖掘中聚类算法比较研究[J].计算机应用与软件,2003(2):5-6.

[4] 张晓艳,华英.文本挖掘的方法及应用研究[J].电脑与电信,2011(12):68-69.

[5] 张雯雯,许鑫.文本挖掘工具述评[J].图书情报工作,2012(8):26-31.

[6] 赵作鹏,尹志民,王潜平,等.一种改进的编辑距离算法及其在数据处理中的应用[J].计算机应用,2009(2):424-426.

[7] 国家重点基础研究发展计划官方网站[EB/OL][2014-5-14]. http://www.973.gov.cn/AreaAppl.aspx.

作者简介

赵胜钢，男，1967年生，博士，中国医学科学院医学信息研究所，研究方向：信息系统、知识管理、文献检索。E-mail：zhao.shenggang@imicams.ac.cn。

Using Levenshtein Distance Algorithm in the Name Data Analysis on the Scientific Research Fund

ZHAO ShengGang, LI JunLian, CHEN Ying
（Institute of Medical Information & Library, CAMS, Beijing 100020, China）

Abstract: Based on the analysis of the clustering method of text data and the characteristics of the scientific research funds name data, the method of applying the Levenshtein Distance algorithm twice in the scientific research fund name data clustering to identify the clustering center of fund name data automatically has been discussed.

Keywords: Text mining; Data clustering; Scientific research fund; Levenshtein distance

（收稿日期：2014-03-28）