基于词汇相似度的医学分类体系映射研究与实现*

单连慧 赵迎光 钱 庆

（中国医学科学院医学信息研究所 北京 100020）

〔摘要〕 采用编辑距离法计算《学科分类与代码》（医学类目）与《医学专业分类表》分类体系类目词汇相似度，通过计算机辅助映射结合人工判断类目间的映射关系，建立《学科分类与代码》（医学类目）与《医学专业分类表》类目之间的映射关系表，以期满足不同类型、不同层次的用户需求。

〔关键词〕 学科分类与代码；医学专业分类表；知识组织系统；词汇相似度；编辑距离

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2016.11.012

Research and Implementation on the Mapping of Classification Systems Based on Similarity of Words *SHAN Lian－hui, ZHAO Ying－guang, QIAN Qing, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, 100020, China*

〔**Abstract**〕 Using edit distance method, the paper calculates the similarity of words between categories of *Classification and Code of Disciplines* (*Medical Category*) and *Special Classification for Medicine*. By combining computer－aided mapping with human judgment, mapping table of the categories is established. It is expected to meet the demands of users in different types and levels.

〔**Keywords**〕 Classification and code of discipline; Special classification for medicine; Knowledge organization system; Similarity of words; Edit distance

1 引言

知识组织（Knowledge Organization）一词最早于 1929 年由英国著名图书馆学家分类法专家布利斯（H. E. Bliss，1870－1955）提出并于当年出版了《知识组织和科学系统》一书，从文献分类的角度阐述了知识组织的思想[1]。知识组织系统是一种机器可理解的系统，可以被计算机系统所识别、读取和理解，既包括各种词表、分类法等传统信息组织技术，也包括诸如语体网络、本体等现代信息和知识组织技术。知识组织系统的互操作是知识组织工具发展的重要特征，对于传统的图书文献资源和现代网络信息资源来说，分类法和主题法均为当前主

〔修回日期〕 2016－05－25

〔作者简介〕 单连慧，硕士，助理研究员，发表论文 10 余篇；通讯作者：钱庆，研究员，发表论文 40 余篇。

〔基金项目〕 中央高校基本科研业务费专项资金和协和青年基金资助项目“基于 TOPSIS－Benchmarking 模型的医学学科科技影响力评价研究”（项目编号 3332015084）；中国医学科学院中央公益性基本科研业务费课题“中国医院科技影响力评价中的关键问题研究”（项目编号 15R0111）。

要的信息组织方式。不同分类法在体系结构、列类标准、类目划分和设置等方面存在较大差异，因此，确定不同分类法类目之间的对应关系较为困难。但是大多数分类法、大类展开形成的类目都表达一定的主题概念，类目表达概念存在很大程度的相似性，而这种相似性是实现不同分类法映射的理论基础。

由于应用目的不同，我国产生了不同的学科分类体系，如《学科分类与代码》、《中国图书馆分类法》（简称《中图法》）以及国家自然科学基金委员会的学科分类名称和代码等。《学科分类与代码》与《中国图书馆分类法》是两种不同的分类体系，但是在教育和科研行业中，经常会遇到交叉使用这两种分类体系的需求，例如进行学科文献统计或者以学科为中心开展评价工作。如果能建立二者之间的映射关系，进而建立起两种分类体系之间通用的互操作关系，则不仅可以满足不同类型、不同层次的用户需求[2]，还可以更好地开展分类科学评价工作等。映射是实现知识组织系统互操作的一种重要方法，分类表之间互操作的主要方法是建立类表之间的映射。在医学领域，开展分类表之间互操作的研究还较少，迫切需要建立医学分类表映射关系，将不同来源的数据映射到统一的学科分类体系中，以解决医学文献统计和科学评价过程中的数据整合统一问题。

目前，分类表映射包括基于词汇相似度[3]和语义相似度[4]的映射方法。本研究为医学领域分类法互操作研究初探，鉴于缺少能够反映词汇语义的背景信息，故采用没有背景信息的方法。现有的词汇相似度计算方法，比较简单的有基于词汇字面匹配的方法，如编辑距离法、字符串匹配法和压缩距离法等。其中，编辑距离法用两个字符串由一个转成另一个所需的最少编辑操作代价来衡量词汇相似度，这里的编辑操作包括替换、插入和删除；字符串匹配法以字符串字面匹配的程度来衡量词汇的相似度；压缩距离法将词汇用二进制编码表示，然后根据二进制编码的柯氏复杂度计算词汇相似度[5]。在此，本研究主要介绍基于编辑距离法的类目词汇相似度计算。

2 基于编辑距离法的类目词汇相似度计算

2.1 编辑距离法

编辑距离法是由俄国科学家 Vladimir Levenshtein 首先提出的，故又称为 Levenshtein Distance，是一种字符串之间相似度计算的方法[6]。给定两个字符串 P、T，将 P 转换成 T 所需要的删除、插入、替换操作的数量称为 P 到 T 的编辑路径，而最短的编辑路径就称为符串 P 和 T 的编辑距离，即编辑距离越小，相似度越高。举例来说：P = “eeba”，T = “abac”。可以按照下述步骤转变：（1）将 P 中的第 1 个 e 变成 a。（2）删除 P 中的第 2 个 e。（3）在 P 中最后添加一个 c。那么 P 到 T 的编辑路径就等于 3。当然，这种变换并不是唯一的，但如果 3 是所有变换中最小值，那么就可以说 P 和 T 的编辑距离等于 3。

动态规划是解决编辑距离的一种主要手段，其基本思路为：将一个复杂的最优解问题分解成一系列较为简单的最优解问题，再将较为简单的最优解问题进一步分解，直到可一眼看出最优解为止。动态规划算法是解决复杂问题最优解的重要算法，其核心是进行递归运算，算法本身比较简单，关键是应用算法时应符合动态规划的思想。动态规划公式如下：

如果 $i=0$ 且 $j=0$，edit（0，0）$=1$；

如果 $i=0$ 且 $j>0$，edit（0，j）$=$edit（0，$j-1$）$+1$；

如果 $i>0$ 且 $j=0$，edit（i，0）$=$edit（$i-1$，0）$+1$；

如果 $i>0$ 且 $j>0$，edit（i，j）$=\min$［edit（i－1，j）$+1$，edit（i，$j-1$）$+1$，edit（$i-1$，$j-1$）$+$f（i，j）］；

其中：edit（i，j）表示 P 中［0…i］的子串 p_i 到 T 中［0…j］的子串 t_1 的编辑距离；f（i，j）表示 P 中第 i 个字符 s（i）转换到 T 中第 j 个字符 s（j）所需要的操作次数，如果 s（i）$=s$（j），则不需要任何操作，f（i，j）$=0$；否则，需要替换操作，f（i，j）$=1$。这就是将长字符串间的编辑距离问题一步一步转换成短字符串间的编辑距离问题，直至只有 1 个字符的串间编辑距离为 1。

2.2 分类表类目分析

《学科分类与代码》建立的学科分类体系是直接为科技政策和科技发展规划以及科研项目、科研成果统计和管理服务的，依据科学性、实用性、简明性、兼容性、扩延性和唯一性原则分类，科技部和卫计委对学科领域的划分目前多采用《学科分类与代码》。在医学领域，在《中国图书馆分类法》

编委会主持下，中国医学科学院医学信息研究所/图书馆组织编制了《中国图书馆分类法·医学专业分类表》（以下简称《医学专业分类表》）。

本文对《学科分类与代码》（医学类目）、《医学专业分类表》的编制原则、体系结构及类目描述方式进行差异研究，综合不同医学分类体系特点，基于类目相似度，建立它们之间的类目映射关系。《医学专业分类表》中，一级学科为 R 后 1 位或 2 位阿拉伯数字，二级学科为 R 后 2 位或 3 位阿拉伯数字，或者 4 位阿拉伯数字，格式为 R × × ×. ×，例如 R54 代表“心血管（循环系）疾病”，R540.4 代表“心血管（循环系）疾病诊断学”；《学科分类与代码》划分为一、二、三级学科层次，用阿拉伯数字表示，例如，320 24 10 代表“临床医学 内科学 心血管病学”。采集《医学专业分类表》（以下简写成 YF）中共 6 913 个学科类目，《学科分类与代码》（以下简写成 GB）中一级学科及各一级学科下二级和三级共 253 个学科类目作为样本数据，即匹配范围涉及《医学专业分类表》中类目等级 YFGRADE = Y2 或 YFGRADE = Y3 及以下的类目，涉及《学科分类与代码》（医学）中等级 GBGRADE = G2 或者 GBGRADE = G3 的类目。

2.3 类目词汇相似度计算

以表 1 列出的类目词汇为例，采用编辑距离法计算类目词汇相似度具体实现过程为：首先将 G1 和 Y1 两组字符串按表 2 结构进行构造，其中 A1、B1、C1、D1、E1 等均为位置符号，这些位置数值均需进行计算。A1 处数值与其左边、上边和左上角 3 处数值有关，即左边 L = 1，上边 U = 1，左上角 LU = 0。按照 Levenshtein Distance 规则，L 和 U 的值均需要加 1，即 L = 1 + 1 = 2，U = 1 + 1 = 2；若 A1 处所对应的两个字符相同，则 LU = 0 + 0 = 0，否则，LU = 0 + 1 = 1。此时取 L、U 和 LU 中数值最小的值作为 A1 处数值，即 A1 处数值 = Min（L，U，LU） = LU = 0。按照上述方法分别计算，即可得出其他处数值，见表 3。G1 和 Y1 的编辑距离为 5，两个字符串长度最大值为 8（即 Y1 的字符串长度），则两个字符串的相似度等于 1 − 5/8 = 0.375。利用上述方法计算 G1 与 Y2、Y3、Y4 和 Y5 的相似度，分别为 0.2、0.2、0.1 和 0.125，G1 与 Y1 的相似度最高，见表 4。

表 1 类目词汇示例

GB 类目	YF 类目
G1 心血管病学	Y1 心血管循环系疾病
	Y2 症状诊断学
	Y3 临床医学
	Y4 内分泌腺疾病及代谢病
	Y5 头部及神经外科学

表 2 字符串结构

字符串	心血管病学	心	血	管	病	学
心血管循环系疾病	0	1	2	3	4	5
心	1	A1	B1	C1	D1	E1
血	2	A2	B2	C2	D2	E2
管	3	A3	B3	C3	D3	E3
循	4	A4	B4	C4	D4	E4
环	5	A5	B5	C5	D5	E5
系	6	A6	B6	C6	D6	E6
疾	7	A7	B7	C7	D7	E7
病	8	A8	B8	C8	D8	E8

表 3 位置数值计算结果

字符串	心血管病学	心	血 *	管	病	学
心血管循环系疾病	0	1	2	3	4	5
心	1	0	1	2	3	4
血	2	1	0	1	2	3
管	3	2	1	0	1	2
循	4	3	2	1	1	2
环	5	4	3	2	2	2
系	6	5	4	3	3	3
疾	7	6	5	4	4	4
病	8	7	6	5	4	5

表 4 类目词汇相似度计算结果

GBNAME	YFNAME	相似度
心血管病学	心血管循环系疾病	0.375
	症状诊断学	0.200
	临床医学	0.200
	内分泌腺疾病及代谢病	0.100
	头部及神经外科学	0.125

3 基于类目词汇相似度的分类表映射

3.1 映射步骤

基于类目词汇相似度的分类表映射实现的基本步骤是：首先采集《医学专业分类表》与《学科分类与代码》中所有学科类目，作为计算机辅助映射

中的处理数据；然后利用编辑距离法计算两种分类体系类目词汇相似度，根据类目相似度值确定类目间的映射关系类型，辅以人工判断；最后建立《学科分类与代码》（医学类目）与《医学专业分类表》类目之间的映射关系表。

3.2 映射原则

拟定相关映射原则：（1）以《医学专业分类表》作为通用分类法，建立《医学专业分类表》与《学科分类与代码》的类目映射关系，保证尽量专指。（2）通过词汇字面相似度进行匹配，设定匹配相似度为不同比例，比较映射关系，自定义合适的匹配相似度。（3）映射关系可分为完全对等、大部分重合概念关系、小部分重合概念关系（根据词汇相似度阈值确定）。（4）从《医学专业分类表》中的类目到《学科分类与代码》中的类目单向进行映射，可以为一对多或者多对一的关系。

3.3 映射过程

本文基于编辑距离理论，开发了词汇相似度计算软件 WordSimilarity，相似度计算代码如下。利用该软件辅助进行《医学专业分类表》与《学科分类与代码》的类目相似度计算。

```
float __fastcall TForm1::levenshtein(unsigned short *sCompare1,unsigned short *sCompare2, int sLen1, int sLen2)
{
aAnsiString sStr;
  int temp,dif[1024][1024];

  //计算两个字符串的长度。
  int len1 = sLen1;
  int len2 = sLen2;
  unsigned short *str1_array = sCompare1;
  unsigned short *str2_array = sCompare2;

  //赋初值,步骤B。
  for (int a = 0; a <= len1; a++)  dif[a][0] = a;
  for (int a = 0; a <= len2; a++)  dif[0][a] = a;

  //计算两个字符是否一样,计算左上的值
  for (int i = 1; i <= len1; i++)
  {
    for (int j = 1; j <= len2; j++)
    {
      if (str1_array[i - 1] == str2_array[j - 1])
      {
        temp = 0;
      }
      else
      {
        temp = 1;
      }

      //取三个值中最小的
      dif[i][j] = min(dif[i - 1][j - 1] + temp, dif[i][j - 1] + 1, dif[i - 1][j] + 1);

    }
  }

  //计算相似度
  float similarity = 1 - (float) dif[len1][len2] / max(sLen1, sLen2);
  return (similarity);
}
```

首先将《医学专业分类表》的“YFNAME”字段和《学科分类与代码》（医学）的“GBNAME”字段进行匹配，将相似度阈值为 1.00 的条目的“YFNUMBER”字段和“YFNAME”字段存放在《学科分类与代码》（医学）对应的“YFNUMBER”字段和“YFNAME”字段中。《学科分类与代码》的 253 个学科类目中，有 79 个条目与《医学专业分类表》中相应类目完全对等，见图 1。其次，将没有匹配上的类目（即《学科分类与代码》中“YFNUMBER”字段和“YFNAME”字段为空的类目）再进行匹配，降低阈值，将相似度阈值为 0.80～1.00 的条目的“YFNUMBER”字段和“YFNAME”字段存放在《学科分类与代码》（医学）对应的“YFNUMBER”字段和“YFNAME”字段中。相似度阈值大于等于 0.8 的涉及《学科分类与代码》的 107 个条目，需要结合人工干预确定映射关系，建立的映射关系为大部分重合概念关系，见图 2。以此类推，针对未匹配上的类目，逐渐降低阈值为 0.7→0.6→0.5→0.4，同时结合人工判断类目间的映射关系，直至最后建立完整的《医学专业分类表》与《学科分类与代码》的类目映射关系。对于计算机辅助计算相似度较低或映射效果差的类目，主要采用人工方法确定映射关系；对于无法匹配的类目，YF 中的某些类目若无确切 GB 类目可对应时，采用上位归类或者靠类的方法对应在相关类目之下。

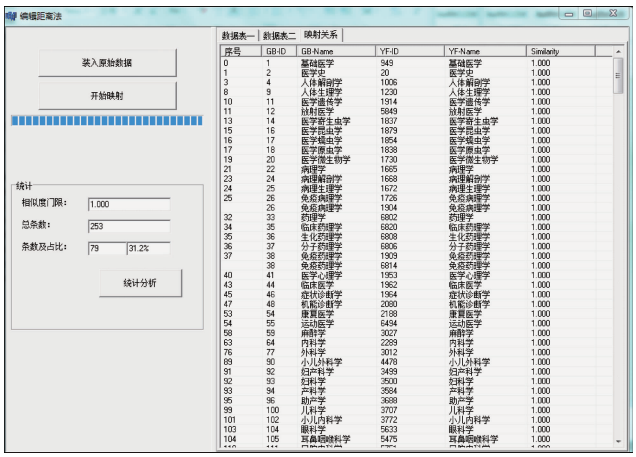

图 1 词汇相似度计算（阈值 1.00）

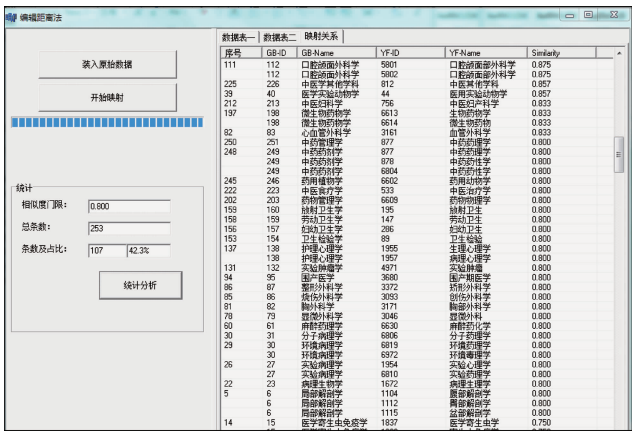

图 2　词汇相似度计算（阈值 0.80）

3.4　映射结果

采用计算机辅助映射，结合人工判断和专家咨询，确定了《医学专业分类表》与《学科分类与代码》类目间的映射关系，图 3 节选基础医学部分类目来展示映射结果，其中 3 级类目 G3 中对应的 YF 类目后续将自动归类到上位 2 级类目 G2 中，同理 G2 归类到 G1 中。通过建立这两种分类体系之间通用的互操作关系，希望能满足不同类型、不同层次的用户需求。

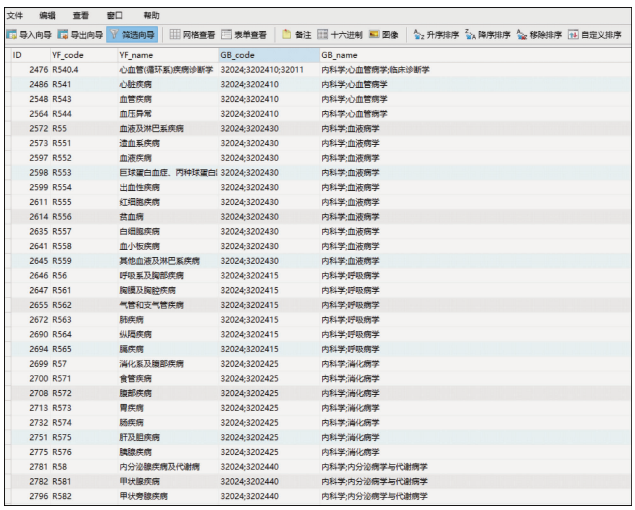

图 3　《医学专业分类表》与
《学科分类与代码》类目映射关系（节选）

4　结语

《学科分类与代码》是中国科学评价领域较好的分类体系，解决医学类分类体系间不匹配问题，可以为科学评价活动中以学科为中心开展分类评价提供支持，进行学科数据例如文献数据的专业统计。本文通过计算机辅助映射，结合人工映射方法，最终实现了《医学专业分类表》与《学科分类与代码》类目间映射关系的建立，可以在一定程度上解决医学文献统计和科学评价过程中不同来源数据整合统一问题。本文为医学类专业分类表之间互操作的研究初探，因缺少可反映词汇语义的背景信息，所以选择了相对比较简单的基于词汇字面匹配的方法编辑距离法，对于相似度较低的类目，需加入较多人为判断。而基于词汇字面匹配的方法认为词汇之间相互独立，事实上许多语义上相似的词汇在字面上并不相似，故用基于词汇字面匹配的方法可能并不能正确计算出词汇的相似度，所以，本研究后续将发掘并利用各种背景信息，例如语料库、医学语义词典等来帮助进行相似度计算，以更好地揭示词汇间的语义关系，建立更准确全面的医学学科分类间映射关系。

参考文献

1　Bliss HE. The Organization of Knowledge and the System of the Sciences［M］. New York: H Holt and Company, 1929.

2　詹萌. 学科（专业）分类与文献分类之间的映射关系研究［J］. 情报理论与实践, 2013, 36（10）: 40－43, 35.

3　周林志, 齐建东, 王建新等. 基于词汇相似度的 IPC 与 CLC 映射［J］. 计算机工程, 2010, 36（23）: 274－276, 279.

4　胡绍波. 基于语义相似度的本体映射方法研究［D］. 昆明: 云南师范大学, 2008.

5　刘萍, 陈烨. 词汇相似度研究进展综述［J］. 现代图书情报技术, 2012,（Z1）: 82－89.

6　Levenshtein VL. Binary Codes Capable of Correcting Deletions, Insertions and Reversals［J］. Soviet Physics－doklady, 1966, 10（8）: 707－710.