文章编号: 1000-8020(2012)06-0895-05 ·论著·

中国总膳食研究食物聚类自动化探索

达能营养中心
DANONE INSTITUTE CHINA
青年科学工作者论坛
Young Scientists' Forum

李万庆 沈雯捷 潘俊霞 闵捷[1] 李筱薇[1,2]
东南大学公共卫生学院流行病与卫生统计学系,南京 210009

摘要: 目的 实现中国总膳食研究食物聚类自动化,提高食物聚类计算的质量和效率。**方法** 通过对中国食物成分表中食物编码特点研究,以及总膳食研究食物聚类的原则进行分析,构建能够让计算机语言识别这些特点和原则的算法,实现食物聚类自动化。**结果** 以第五次中国总膳食研究某省膳食调查数据为例,将 292 种食物聚为 53 种聚类食物,聚类结果符合总膳食研究聚类要求。**结论** 该方法能够有效地实现中国总膳食研究食物聚类自动化计算,提高了中国总膳食研究中食物聚类的质量和效率。

关键词: 总膳食研究 食物聚类 自动化计算
中图分类号: R151.42 **文献标识码:** A

Exploration of the food cluster automation in the China Total Diet Study

LI Wanqing, SHEN Wenjie, PAN Junxia, MIN Jie, LI Xiaowei
Department of Epidemiology and Biostatistics, School of Public Health, Southeast University, Nanjing 210009, China

Abstract: Objective To achieve the food cluster automation of China Total Diet Study and improve the quality and efficiency of the food cluster calculation. **Method** The food coding features were studied in the Chinese Food Composition Table. After analyzing the principles of food clustering on Total Diet Study, constructing an algorithm allows the computer language to identify these characteristics and principles to achieve automatic food cluster. **Results** The instance was selected from the data of a province in the fifth Chinese Total Diet Study. 292 food items were clustered into 53 kinds of cluster food by computer program, and the results are corresponded to the cluster requirements of total diet study. **Conclusion** The automatic calculation of the food cluster can be achieved by computer program more reliable and effectively in the Total Diet Study.

Key words: Total Diet Study, food cluster, automatic calculation

总膳食研究(total diet study,TDS),也称"市场菜篮子研究"(market basket study),是目前国际上公认的评价一个国家或地区大规模人群膳食中化学污染物和营养素摄入量的通用的最好方法[1-2]。TDS 研究步骤包括:(1)膳食调查;(2)食物聚类;(3)采样;(4)烹调及样品制备;(5)样品测定;(6)计算成年男子每人每日从各类食物摄入的污染物及营养素的量[3]。食物聚类是非常重要的一个环节,由于我国居民食用膳食种类繁多,因此按照统一的聚类原则聚成品种较少且具有代表性的食物,既能减少工作量又能达到研究

作者简介: 李万庆,男,硕士研究生,E-mail: liwanqingseu@163.com
1 通讯作者: 闵捷,E-mail: minjie93@163.com; 李筱薇,E-mail: eveline73@vip.sina.com
2 国家食品安全风险评估中心

目的[4]。TDS 虽然提出了食物聚类的一些原则性要求，但具体计算却没有统一模式，以往各省自行聚类时所花时间较多，聚类结果也因为没有统一的模式而达不到满意的效果。因此在第五次中国总膳食研究开展之际，作者对 TDS 聚类原则和食物编码规则进行了深入研究，构建了能够让计算机语言识别并实现食物聚类自动化的算法，提高食物聚类的质量和效率。

1 方法

1.1 食物聚类原则

按照中国总膳食研究要求，食物聚类后一般分为 13 大类，即粮谷类及其制品、豆和坚果类及其制品、薯类及其制品、肉类及其制品、蛋类及其制品、水产类及其制品、乳类及其制品、蔬菜类及其制品、水果类及其制品、糖类、饮料与水、含酒精饮料、调味品与烹调用油。各聚类食物需按照如下原则进行聚类：(1) 按所属类别聚类；(2) 可食部分和水分相当；(3) 营养素含量相近；(4) 根据具体情况而定。总的原则是聚类后的食物品种既能反映当地居民的饮食习惯又能减少采样和烹调时的工作量[5-6]。

1.2 食物聚类自动化计算方法

1.2.1 计算每种食物每标准人日消费量 从个人信息调查表中提取每个人的性别、年龄、生理状态、劳动强度，根据标准人能量消耗表可以得到每个人对应的能量消耗，再除以 18 岁成年男子轻体力劳动状态的能量消耗值，得到每个被调查对象相应的标准人数，再提取每个被调查对象的标准日，相乘得到每个被调查者的标准人日数，累加计算总的标准人日数，每种食物总消费量除以总标准人日数得到每种食物每标准人日消费量[7]。

1.2.2 聚类食物类别定义 中国总膳食研究采用的食物编码为中国疾病预防控制中心营养与食品安全所编写的中国食物成分表。该编码系统包含加工食物、原料食物，共有 21 大类。每条食物编码由 6 位数字构成，前 2 位数字是食物的类别编码如蔬菜类及其制品，第 3 位数字是食物的亚类编码如蔬菜类及其制品下的根菜类、鲜豆类，最后 3 位数字是食物在亚类的排列序号[8]，且第 4 位数字往往表示该亚类下不同属、不同加工方法等。该编码系统的 21 个类别与总膳食研究食物聚类的 13 大类存在错位关系，如食物编码中的某一类可能对应到总膳食研究 13 大类中的好几个大类，反之亦然。因此，不能单纯提取食物编码的前 2 位定义聚类食物的大类，部分要提取前 3 位甚至前 4 位定义聚类食物的大类。例如，食品编码系统前 4 位为 '0471' 和 '0472' 的食物在总膳食研究中定义为薯类及其制品，前 4 位为 '0473' 的食物在总膳食研究中定义为第 13 大类调味品与烹调用油，其余前 2 位为 '04' 的食物在总膳食研究中定义为第 3 大类蔬菜类及其制品。总之，首先要将 2000 多条食物编码根据其编码特点和总膳食研究食物分类标准定义为对应的 13 大类。

1.2.3 定位聚类食物 对 13 大类中的每一大类食物，分别提取食物编码的前 4 位、前 3 位、前 2 位定义三个层次的小类类别，以第一层次为例，首先提取了所有食物编码的前 4 位作为其小类类别，并计算得到每小类中消费量最大的食物对应的食品编码，以此作为该前 4 位相同的食物对应的聚类食物。由食物编码前 3 位、前 2 位定义的小类类别及其聚类食物的计算类似。

1.2.4 初步聚类 首先进行第一层次聚类，计算每种食物消费量占其所在大类食物消费量之比，如果该比值 ≥ 0.05，则将此食物留下作为第一层次聚类食物，如果该比值 < 0.05，则聚为大类相同且前 4 位食物编码相同的食物中消费量最大的食物，该食物已经在上一步定义。重新计算第一层次聚类食物消费量之和占该大类食物消费量之比，进入第二层次聚类，如果该比值 ≥ 0.05，则保留该食物为第二层次聚类食物，如果该比值 < 0.05，则聚为大类相同且前 3 位食物编码相同的食物中消费量最大的食物作为第二层次聚类食物。继续计算第二层次聚类食物消费量之和占该大类食物消费量之比，进入第三层次聚类，如果该比值 ≥ 0.05，则保留该食物作为第三层次聚类食物，如果该比值 < 0.05，则聚为大类相同且前 2 位食物编码相同的食物中消费量最大的食物。另外，每一个层次聚类食物消费量都要根据相对应食物的可食部分和水分系数进行消费量折算。经过三层次的聚类，从细到粗，层层递进，完成了初步聚类，并且符合相同类别属性聚类、营养元素、可食部水分比例相近和聚少留多的原则。文中所述界值是人为选择，经对 0.01、0.02、0.05 和 0.1 多次运行后，选取了聚类结果最好的比值 0.05。

1.2.5 聚类调整 初步聚类之后，还有一些食物聚类需要调整，分为 3 种情况。第一，如果某些大类想多保留一些聚类食物，可以将该大类食物在第一层次或第二层次聚类时即可停止，即聚为前 4 位或前 3 位相同的消费量最大的食物；如果某些大类想少保留一些聚类食物，可以将该大类食物直接进入第三层次聚类，即聚为前 2 位相同的

消费量最大的食物。第二，在总膳食研究当中，要求某些食物即使消费量特别少，但是其所含污染物的量较高，对于今后所要进行的风险评估意义重大的食物，也要保留采样，对于具有这种性质的食物从食物编码表中总结归纳，按大类不同分别定义其对应的小类，最后选取该小类下消费量最大的食物作为聚类食物。第三，对于 13 大类某些食品编码前 2 位或前 3 位或前 4 位不同、但是在总膳食研究中希望将它们聚在一起的食物，根据其属性相同重新定义小类，再寻找该小类消费量最大的食物作为聚类食物。经过上述所有步骤，就完成了食物聚类，整个聚类调整过程要根据实际工作中需要重点关注的污染物品种的不同确定不同的聚类调整过程，并在相关专家的指导之下完成。

2　结果

依据上述自动化计算方法编写程序，利用 SAS 编程计算 2011 年西北某省总膳食研究的食物聚类，调查共获得 292 种食物，经过聚类为 53 种代表食物，结果以 excel 表单形式输出。现以谷类食物聚类过程结果再次说明自动化聚类方法（见表 1）。

在表 1 中，食物编码前四位为 '0114' 的食物显然属于同一类别，其消费量占大类食物消费量之比均小于 0.05，则全部聚类为第一层次聚类食物即这些食物中消费量最大的食物 '011404'，并根据每个食物的可食部和水分系数将其消费量转化为聚类食物的消费量，计算聚类食物消费量之和及其大类食物消费量占比，该比值 0.11（大于 0.05），聚类到此结束，第二层次和第三层次的聚类食物维持不变，但是第三层次聚类结束后需要对部分食品进行调整。油饼、烧饼、烙饼、油条均为含油面食，在总膳食研究中作为特殊检测对象须纳入聚类结果中，因此将这几个食物作为一个小类聚为该类中消费量最大的烙饼。并且在考虑含油面食时，食物编码表中 '011502' 油面筋需要调整，可以看到第一层次聚类中聚为 '011502'，因为食物编码前 4 位为 '0115' 的食物只有一个，其消费量占大类比值小于 0.05，进入第二层次聚类，聚为前 3 位相同的小类中消费量最大的小麦面粉 '011206'，由于同时还有其他很多食物被聚为 '011206'，因此 '011206' 的消费量占大类消费量之比 0.32（大于 0.05），被聚为 '011206' 的食物到该层次结束，第三层次聚类维持不变，但是 '011502' 由于是含油面食需要调整聚类结果，与食品编码前 4 位为 '0114' 的含油食物作为一类，聚为该类中食物消费量最大的烙饼。前 2 位为 '14' 和 '15' 的谷类加工食物作为一类，经过多次调整聚类为 '152201' 方便面。

通过上述谷类食物的聚类过程，可以清楚的看到大部分食物在三层次的初步聚类中，分别寻找其前 4 位相同、前 3 位相同、前 2 位相同的小类类别中消费量最大的食物，并结合占大类消费量之比，最终找到合适的聚类食物，其余少部分食物需要根据总膳食研究专家的意见进行保留，每大类食物中都有相应的特殊食物在食物聚类中给予保留，比如谷类食物中含油食物，肉类中的内脏，腌制食物等都必须在聚类结果中保留，在编写程序中需要给予考虑。如果利用 excel 表单计算食物聚类，结果和程序运行结果理论上应是一致的，但是 excel 在计算总标准人和不同食物间消费量折算时，有一定困难，极易出现计算错误，更难以体现专家的聚类思想。利用本文所述方法编写程序，可将专家的聚类思想固定在程序中，运行就可以直接得到食物聚类结果，实现总膳食研究食物聚类自动化、标准化。

3　讨论

我国 1981 年加入全球环境规划监测体系/食品分部（GEMS/Food），该系统要求每个会员依据本国国情进行食品污染物的监测工作，收集相关的污染水平数据，并通过电子网页或者电子文档的形式上报 GEMS/Food 相关组织，在实验室质量控制和数据收集方面都有一系列标准，对上报数据进行分级，确保数据的质量和可靠性。在数据收集方面，GEMS/Food 于 1996 年建立了污染物数据库，其中包括一般食品污染物数据库和总膳食数据库[9]。总膳食研究作为世界卫生组织广泛推广的开展大规模人群最经济有效的膳食暴露评估方法，在国家食品污染物风险监测以及国际食品安全风险评估中都发挥着重大作用。长期以来，我国在食品污染物监测工作中存在重实验室质量控制而轻数据收集处理方面控制的问题，从一定程度上制约了我国食品污染物监测数据的质量和可靠性。在第 5 次中国总膳食研究中，对数据处理标准化研究非常重视，希望能建立一套具有中国总膳食研究特色的膳食调查数据录入、食物聚类、采样烹调信息、暴露评估的标准化自动化体系。总膳食研究数据处理中计算最费时费力且最不易进行质量控制的在食物聚类这一环节，我国各省区居民膳食消费种类平均在四五百种，如

表 1 谷类食物聚类自动化计算过程

Table 1 The food cluster automation calculation of cereals

食物编码	食物名称	每人每日消费量(g)	大类占比	第一层次聚类食物	聚类食物大类占比	第二层次聚类食物	聚类食物大类占比	第三层次聚类食物	聚类食物大类占比
11201	小麦粉(标准粉)	19.6	0.045	11206	0.267	11206	0.324	11206	0.288
11101	小麦	2.11	0.005	11101	0.005	11206	0.324	11206	0.288
11206	小麦面粉[标准粉]	100.02	0.230	11206	0.267	11206	0.324	11206	0.288
11302	挂面(标准粉)	0.98	0.002	11305	0.046	11206	0.324	11305	0.049
11304	挂面(精制龙须面)	0.1	0.000	11305	0.046	11206	0.324	11305	0.049
11306	面条(标准粉,切面)	4.36	0.010	11305	0.046	11206	0.324	11305	0.049
11313	挂面(富强粉)	0.53	0.001	11305	0.046	11206	0.324	11305	0.049
11301	挂面(x)	1.16	0.003	11305	0.046	11206	0.324	11305	0.049
11305	面条(x)	12.79	0.029	11305	0.046	11206	0.324	11305	0.049
11408	油饼	6.81	0.016	11404	0.111	11404	0.116	11403	0.049
11407	烧饼(加糖)	1.48	0.003	11404	0.111	11404	0.116	11403	0.049
11502	油面筋	0.97	0.002	11502	0.002	11206	0.324	11403	0.049
11403	烙饼(标准粉)	7.74	0.018	11404	0.111	11404	0.116	11403	0.049
11409	油条	1.22	0.003	11404	0.111	11404	0.116	11403	0.049
11405	馒头(标准粉)	3.72	0.009	11404	0.111	11404	0.116	11404	0.066
11401	花卷	4.14	0.010	11404	0.111	11404	0.116	11404	0.066
11404	馒头(x)	15.71	0.036	11404	0.111	11404	0.116	11404	0.066
11411	馒头(富强粉)	2.01	0.005	11404	0.111	11404	0.116	11404	0.066
11410	花卷(加牛奶)	1.66	0.004	11404	0.111	11404	0.116	11404	0.066
12401	米饭(蒸)(x)	21.13	0.049	12401	0.048	12001	0.503	12001	0.510
12301	糯米[江米](x)	0.3	0.001	12301	0.001	12001	0.503	12001	0.510
12215	香米	16.11	0.037	12215	0.042	12001	0.503	12001	0.510
12001	稻米(x)	188.68	0.433	12001	0.424	12001	0.503	12001	0.510
12213	香大米	2.66	0.006	12215	0.042	12001	0.503	12001	0.510
15104	小米(黄)	0.05	0.000	15101	0.005	15101	0.005	15101	0.006
15101	小米	2.15	0.005	15101	0.005	15101	0.005	15101	0.006
13110	玉米糁(黄)	0.09	0.000	13110	0.000	13110	0.000	15101	0.006
15103	小米粥	0.13	0.000	15101	0.005	15101	0.005	15101	0.006
142332	沙琪玛蛋酥	0.17	0.000	142321	0.005	142101	0.011	152201	0.032
142321	麻花	1.1	0.003	142321	0.005	142101	0.011	152201	0.032
142311	混糖糕点	0.99	0.002	142321	0.005	142101	0.011	152201	0.032
142101	蛋糕(x)	1.97	0.005	142101	0.004	142101	0.011	152201	0.032
141039	黑芝麻汤圆	0.33	0.001	141013	0.020	141013	0.021	152201	0.032
141023	酿皮	0.88	0.002	141013	0.020	141013	0.021	152201	0.032
141022	年糕	0.61	0.001	141013	0.020	141013	0.021	152201	0.032
141013	凉粉	1.51	0.004	141013	0.020	141013	0.021	152201	0.032
152301	面包(x)	0.91	0.002	152301	0.002	152201	0.014	152201	0.032
152206	红烧牛肉方便面	0.86	0.002	152201	0.009	152201	0.014	152201	0.032
152221	冬菜虾仁馄饨	0.95	0.002	152201	0.009	152201	0.014	152201	0.032
152202	鸡汁味干脆面	0.33	0.001	152201	0.009	152201	0.014	152201	0.032
152114	麦片(原味香奶)	0.04	0.000	152114	0.000	152201	0.014	152201	0.032
152201	方便面	2.2	0.005	152201	0.009	152201	0.014	152201	0.032
153203	锅巴(小米)	0.73	0.002	153202	0.005	153202	0.006	152201	0.032
153202	锅巴(豆香)	1.59	0.004	153202	0.005	153202	0.006	152201	0.032
153109	洋葱圈	0.12	0.000	153109	0.000	153202	0.006	152201	0.032
152421	早茶饼	1.37	0.003	152421	0.003	152201	0.014	152201	0.032
152416	饼干(夹心)	0.04	0.000	152421	0.003	152201	0.014	152201	0.032
152406	钙奶饼干	0.13	0.000	152421	0.003	152201	0.014	152201	0.032
152304	法式配餐面包	0.06	0.000	152301	0.002	152201	0.014	152201	0.032

何根据消费量和消费种类聚为数十种食物进行采样烹调和污染物检测并能够准确有效地反映当地居民的饮食习惯较为困难，另外要求能够快速聚类并即时采样烹调，因为很多新鲜蔬菜水果具有季节性，时间错过很难开展采样工作。本研究通过对中国总膳食研究食物聚类的原则和食物编码特点的分析，构建了能够被计算机语言识别并实现聚类自动化的算法，使得食物聚类能够自动化、标准化。现在已经根据本文所述方法对总膳食调查数据从数据的导出、核查、每种食物每标准人日消费量的计算、食物聚类等数据处理过程编写了 sas 程序，结果得到总膳食研究专家的认可，并在开展第 5 次中国总膳食研究中得到广泛应用，取得较好的工作效果。

在本研究中利用了中国疾病预防控制中心营养与食品安全所编写的食物编码表，能够从食物属性不同区分大类亚类的的特点为食物聚类自动化计算奠定了基础，但是食物编码中前 2 位为‘14’和‘15’的食物作为加工食品，其编码顺序稍显混乱，如果能够有效区分出该加工食品的主要原材料的不同则会更好。同理，本文所述食物聚类自动化算法具有一定的延展性，任何一种食物编码系统只要能有效区分出食物种类即可利用该算法进行食物聚类自动化编程。如果将来采用国际食品法典委员会的食物编码方式和规则编写中国的食物编码表，作为膳食调查食物数据录入编码系统，那么食物聚类、膳食消费量数据上报等都可以与国际接轨，更加有利于国际交流。

参考文献

1 陈君石，高俊全，王绪卿. 中国总膳食研究 [J]. 中国肿瘤，1994，11(3)：22-23.

2 FAO. Assuring food safety and quality: Guidelines for strengthening national food control system [R]. FAO Food and Nutritional Paper，2003：76.

3 李筱薇，吴永宁，陈君石. 中国总膳食研究二十年的发展演变 [J]. 中华流行病学杂志，2011，32(5)：456-458.

4 李筱薇，高俊全. 总膳食研究：一种使食品更安全的方法 [J]. 中国食品卫生杂志，2006，18(3)：254-255.

5 中国疾病预防控制中心营养与食品安全所. 2010 年中国总膳食研究工作手册 [M]. 2010：5-6.

6 李筱薇，高俊全，王永芳，等. 2000 年中国总膳食研究：膳食砷摄入量 [J]. 卫生研究，2006，36(1)：63-65.

7 张磊，高俊全，李筱薇. 2000 年中国总膳食研究：不同性别年龄组人群膳食铅摄入量 [J]. 卫生研究，2007，36(4)：459-467.

8 杨月欣，王光亚，潘兴昌. 中国食物成分表 2002 [M]. 北京：北京大学医学出版社，2002：4.

9 杨杰，樊永祥，杨大进，等. 国际食品污染物监测体系理化指标监测介绍及思考 [J]. 中国食品卫生杂志，2009，21(2)：161-168.

收稿日期：2012-06-29