面向营养成分检索的中国菜谱搜索引擎构建*

徐晓巍 王 举 李 姣

（中国医学科学院医学信息研究所/图书馆 北京 100020）

〔摘要〕 以《中国居民膳食营养素参考摄入量（2013 版）》中富含钙、磷、钾、钠、镁 5 种营养素的常见食材作为权威参考数据，结合从互联网获取到的成品菜数据，利用搜索引擎、数据库存储等计算机技术对数据进行计算，由此获得富含某种营养素的成品菜列表。实现营养素——>成品菜的计算，为指导居民针对性营养素补充膳食提供解决方案。

〔关键词〕 营养素；Elasticsearch（ES）；菜谱；搜索引擎

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2017.08.017

Construction of Nutrient Content Retrieval Oriented Search Engine for Chinese Recipes *XU Xiao－wei, WANG Ju, LI Jiao, Institute of Medical Information & Library, Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 The paper takes common food materials that contain 5 nutrients of calcium, phosphorus, potassium, sodium and magnesium in the *Chinese Dietary Reference Intakes* (2013) as authoritative reference data, combines the data of finished dishes obtained from the Internet, and calculates the data by making use of the search engine, database storage and other computer technologies, in order to obtain the list of finished dishes containing certain nutrients, achieve the calculation from nutrients to finished dishes, and provide solutions to the guidance for targeted nutrient supplement diet of residents.

〔**Keywords**〕 Nutrient; Elasticsearch (ES); Recipes; Search engine

1 引言

随着营养学研究的深入发展，人们发现微量营养素在人体内的平衡状态已和许多疾病联系在一起，由国家卫生计生委编写的《中国居民营养与慢性病状况报告（2015）》[1]显示，与 2002 年[2]相比，我国城乡居民钙、铁、维生素 A、维生素 D 等营养素缺乏状况仍然存在。多项研究也表明营养素与慢性病之间的密切关系，如提高膳食钾的摄入量有助于预防高血压等慢性病[3-4]；而钙缺乏除与骨健康相关外，还可能与糖尿病、心血管病、高血压等慢性疾病相关[5]。正因如此，国家对居民膳食的指导愈加重视，在国家卫生计生委日前发布的《中国居民膳食指南 2016》[6]（以下简称“膳食指南”）中，根据不同类型人群特点，提出了不同的膳食营养补充指导条目。如针对普通人群的推荐条目中，其中一、三、四条都与具体吃什么食物相关，其目的在

〔修回日期〕 2017－04－12

〔作者简介〕 徐晓巍，硕士，实习研究员；通讯作者：李姣，博士。

〔基金项目〕 协和青年基金“多元自我量化在健康促进中的应用研究”（项目编号：3332015083）。

于通过不同种类食物的摄入满足人体对多种营养素的需求；又如由于备孕妇女对铁元素的需求升高，指南提出了“一日三餐应该有瘦畜肉 50 ~ 100 克”的膳食建议；再如考虑到学龄前儿童对钙的需求量，指南建议“2 ~ 5 岁儿童每天饮用 300 ~ 400 毫升奶或相当量奶制品”。

居民膳食营养管理广受各界关注，互联网公司以国家相关部门（如农业部等）开放的权威数据为基础，结合其自身在大数据计算、线下数据积累等方面的优势，开发了膳食营养计算及个人膳食营养记录相关产品。食谱搜索网站 Yummly[7] 以美国农业部发布的《营养标准数据库》（National Nutrient Database for Standard Reference，NDSR）[8] 为基准数据库，对从网络上搜集的菜谱进行二次计算获取其中各项营养成分，同时该公司也通过与线下餐厅数据的连通实现了在餐厅点餐就可以获得各项营养成分摄入的数据，从而实现对个体营养素摄入情况的记录；中国的健康减肥管理网站薄荷网[9] 收录了超过 30 万种食物的营养成分数据，其中包括各种基本的食材、烹饪好的菜肴等，同时以图片的方式可视化地呈现出食物分量的变化，只要用户主动地输入所食用相应分量的食物，该网站就可以记录该用户摄入各项营养成分的数据。

由此可见，对从食物中摄入的营养成分吸引了各方的注意力，如何能够从日常膳食中获取相应的营养成分也越来越受到重视。各国均结合其国人体质发布有权威的膳食营养素参考摄入量，如由美国医学研究所发布的针对欧美人的膳食参考摄入量[10]（DRIs）、由中国营养学会编著的《中国居民膳食营养素参考摄入量（2013 版）》[12] 在对每种营养素描述后都会总结日常可获取的食物中有哪些富含该营养素，为普通居民的膳食营养补充提供了权威参考；程蜀琳等通过为期两年的对 860 个青春期早期女孩的随机对照试验，从其日常膳食中分离出钙元素摄入量、奶制品消耗量及维生素 D 补充剂摄入量评估该阶段女孩膳食情况及其对该群体的骨量及身体构成的影响[11]；同时该团队还通过对糖尿病前期非酒精脂肪肝的绝经妇女和中年男子进行低碳水的饮食干预确定该方法对此类人群的存在影响[23]。特定营养素层面上的膳食不但是大众需求，对于患有慢性疾病的群体来说，用营养素指导日常膳食变得愈发重要。

综上可以看出，人们已经认识到日常膳食摄入营养素的重要性，但目前无论是商业领域还是科研领域对膳食相关的研究主要集中在记录日常摄入食品，根据膳食记录计算其摄入的营养成分，即从给定的菜品推导其所含营养成分，从而实现营养素摄入量的测算。然而，在实际应用场景中，常遇到需要查询富含某种营养成分的菜品（例如有研究表明，镁在预防高血压和一些相关慢性病方面有重要作用，适量补充镁有利于预防高血压[13]，所以会有相关人群查询富含镁元素的菜品），即输入营养素输出富含该营养素的菜品，对于这样信息的检索与展示、底层数据存储与管理使用还鲜有研究。鉴于此，本研究以具体的营养素作为输入，以权威数据和网络获取数据作为数据支撑，利用搜索引擎、网络爬虫等计算机相关技术，实现富含该营养素的成品菜的计算过程并输出，以此实现营养素——→成品菜的计算。

2 研究路线

为了实现输入营养素名称、输出富含该营养素的成品菜的功能，满足居民从营养素角度搜索菜谱的需求，本研究分两步进行。首先，将《中国居民膳食营养素参考摄入量（2013 版）》[14] 中常见食物各项营养素含量较高的食材存储为计算机可识别的格式，利用开源搜索引擎框架 Elasticsearch[15]（以下简称 ES）实现营养素——→食材的搜索，搜索结果按照营养素的含量降序排列；同时，通过计算机识别 + 人工校验的方式建立测试用成品菜数据库，再次利用 ES，将上一步获得的食材作为检索词搜索其作为主料的成品菜，从而实现食材——→成品菜的搜索。以上实现营养素——→成品菜的计算。本研究的总体框架，见图 1。

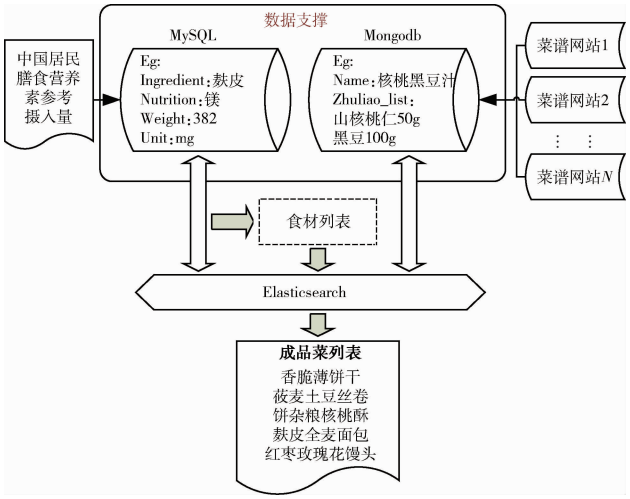

图 1 面向营养成分检索的中国菜谱搜索引擎架构

3 数据来源及存储

3.1 常见食物中常量营养元素含量

该部分数据以《中国居民膳食营养素参考摄入量（2013 版）》[14] 为主要依据，选取钙、磷、钾、钠、镁 5 种常量营养元素，将 5 种常量营养元素含量高的食物存储为计算机容易处理的数据格式[16]。该部分数据结构化较好，选用 MySQL 进行存储，各字段定义，见表 1。目前该表中共存有 5 种常量营养元素，共计 146 条常见食材相应营养元素的含量值。

表 1 常见食物中常量营养元素含量

Id	原料（Ingredient）	营养（Nutrition）	重量（Weight）	单位（Unit）
1	黄豆	蛋白质	35.00	g

3.2 菜谱数据

为了获取测试数据构建菜谱库，本研究通过编写网络爬虫对菜谱网站数据进行收集并定义统一的字段对收集到的数据进行存储。通过对各菜谱网站的分析，确定使用 Mongodb[17] 作为存储菜谱数据的数据库。Mongodb 支持的数据结构非常松散，使用类 json 的 bson 格式可以存储比较复杂的数据类型。通过对菜品数据的分析，结合膳食营养素记录的需求，确定菜谱数据的存储格式如下：

```
{
    “name”：“蜜汁叉烧饭”，
“type”：“粤菜”，
    “info_ list”：[
{ “cooking_ style”：“烤”，
“taste”：“甜”
“number”：“1”}
]
“zhuliao_ list”：[……]
    “fuliao_ list”：[……]
  }
```

“name”字段存储菜名，“type”字段存储所属菜系，“info_ list”字段存储菜肴所使用的烹饪方式、菜品口味及该道菜对应的食用人数，“zhuliao_ list”存储该菜品所含主料及对应用量，“fuliao_ list”存储该菜品所用到的各种辅料及调料的名称及对应用量。

4 应用 ES 实现搜索功能

4.1 环境部署

ES 采用 Java 开发，服务器只需安装 Java 运行环境即可部署 ES。本研究中使用操作系统为 Ubuntu 14.04.1，Java 版本为 1.8.0。使用的是 ES 2.3.1。

4.2 数据同步

本研究使用了两部分数据：一部分存储营养素——→食材的数据采用 MySQL 数据库；另一部分存储食材——→成品菜的数据采用 Mongodb 数据库。因此，需使用不同的工具分别对两部分数据与 ES 进行同步。MySQL 与 ES 数据同步：本研究选择了 elasticsearch - jdbc[18] 作为 ES 和 Mysql 的数据同步插件，该插件开发社区活跃度高、持续更新，获取、安装都非常容易。Mongodb 与 ES 数据同步：选取的解决方案为 mongo - connector[19]，这是 Mongodb 官方用 Python 实现的同步工具，目前支持将 Mongodb 的数据同步到 Solr、ES、Mongodb 中。该同步工具可以将 Mongodb 端数据的变化自动同步到 ES 端。

4.3 索引配置

各个字段被映射后的字段，见表 2。

表 2 映射字段

字段名称	(Id)	原料 (Ingredient)	营养 (Nutrition)	重量 (Weight)	单位 (Unit)
例子	1	黄豆	蛋白质	35.00	g
ES 映射后数据类型	Long	String	String	Double	String

分词器：将一个文本块标记为适用于倒排索引的单独的词，因为本研究内容主要为中国菜品，所以使用了当前应用广泛的 ik 分词器[20]。如“西红柿炒鸡蛋”经过该分词器则可切分为“西红柿”、“炒”、“鸡蛋”。

4.4 搜索实现

ES 提供了丰富的查询语句，通过这些语句的搭配使用，就可以完成营养素——>原材料——>成品菜的搜索。如上所述，该搜索分为两次搜索实现：营养素——>食材的搜索和食材——>成品菜的搜索。

营养素——>食材搜索实现：以“常见食物中常量营养元素含量表”为文档库实现营养素——>食材的搜索，将营养元素作为检索词，默认情况下，结果集会按照相关性进行排序，即相关性越高，排名越靠前。为了查找含有同一种营养元素含量较高的前 10 种食材，此处添加了 sort 参数对结果集进行排序，即将结果集按照“weight”字段值降序排列。以搜索营养素“钾”为例，搜索语句为：

```
{“query”:{“match”:{“nutrition”:u“钾”}},
“sort”:{“weight”:{“order”:“desc”}}}
```

搜索结果集为：[‘黄豆’，‘赤小豆’，‘绿豆’，‘海带（干）’，‘金针菜’，‘花生（炒）’，‘羊肉（瘦）’，‘马铃薯’，‘羊肉（瘦）’，‘芭蕉’]（此处仅取搜索结果前 10 位）与《中国居民膳食营养素参考摄入量（2013 版）》常见食物中钾含量表格中含量最高的前 10 种食物相符。

原材料——>成品菜搜索实现：以来自互联网的菜谱数据作为文档库实现原材料——>成品菜的搜索，以上一步营养素——>原材料的搜索获得的搜索集作为检索词，此处匹配条件为只要某种钾含量高的原材料出现在菜谱数据的“主料”字段中，则该成品菜为钾含量高的菜品。此处假设：成品菜中主料种类不低于两种。若某种成品菜主料中只含有一种主料，则该成品菜较单调或只是调料类，以“黄豆”为例，将该食材作为检索词查询成品菜时出现的第一条检索结果为“黄豆酱”，但是甚少有人会将黄豆酱作为一道菜食用。由此可保证成品菜种类多样性和营养素含量丰富。

5 搜索结果分析

以营养素“镁”作为检索词，共得到常见食物镁含量较高的食材 10 种，见表 3。

表 3 常见食物中镁含量最高的 10 种食材（mg/100g 可食部）

食材名称	麸皮	南瓜子	山核桃	黑芝麻	葵花子仁	杏仁	虾皮	荞麦	黑豆	莲子
含量	382	376	306	290	287	275	265	258	243	242

以表 3 中的 10 种食材作为检索词在菜谱库中进行查询，结果集中共有 21 道成品菜，截取排名前 10 条数据，见表 4。从表 4 中可以看出，标黄的为镁元素含量高的食材，成品菜搜索结果包含有 8 种原材料，没有出现以“荞麦”、“虾皮”为食材的成品菜。而且结果集涵盖了主食、饮品、菜品、小吃等，餐食类型覆盖范围较广，同时镁元素含量丰富，可满足不同人群补充镁元素的需求。

表 4 成品菜搜索结果

序号	成品菜名称	食材列表
1	香脆薄饼干	荞麦面粉、燕麦薄片、燕麦麸皮、小麦麸皮、鼠尾草籽
2	核桃黑豆汁	山核桃仁、黑豆
3	山楂薏米核桃酪	山楂、薏米、核桃
4	香脆什锦麦片	燕麦、蔓越莓干、葵花瓜子仁、榛子仁、杏仁片

续表 4

5	糯米黏糕	糯米面、榛子仁、杏仁、核桃仁、葵花子、芝麻、葡萄干、南瓜子
6	五仁二豆豆浆	熟核桃仁、熟花生仁、葵瓜子仁、南瓜子仁、黑芝麻、黄豆、黑豆
7	瓜子仁蛋挞	南瓜子仁、葵花子仁、淡奶油、牛奶、蛋黄
8	杏仁蛋羹	鸡蛋、杏仁
9	黑米黑豆粥	黑米、黑豆
10	莲子炒莲藕片	莲子、莲藕

此外，由于成品菜数据来自于各菜谱网站，通过对来源网址的点击追踪，用户可以获得相应成品菜的烹饪步骤、烹饪方式、所需时间等其他影响因素，从而不但能够从营养元素的输入获得富含该营养元素的成品菜，也能够掌握该成品菜的烹饪方法。具体示例，见表 5。

表 5 富含镁元素菜品做法示例

项目	黑米黑豆粥	杏仁蛋羹
来源网址	http：//home. meishichina. com/recipe - 213066. html	http：//home. meishichina. com/recipe - 53205. html
烹饪方式	煮	蒸
所需时间	1 小时	10 分钟
步骤	（1）黑米、黑豆，加糯米和 3 个大红枣； （2）黑米、糯米、黑豆淘洗干净，放到煮锅里加水浸泡 2 小时； （3）红枣切下枣肉，炉灶开小火慢慢熬煮； （4）大约 1 小时即可	（1）打 3 个鸡蛋在碗里搅匀； （2）杏仁捣碎，葱花切碎，辣椒切丝； （3）把捣好的杏仁、葱花放入鸡蛋碗里，放入盐、味精、酱油搅拌均匀，然后撒入辣椒丝； （4）取保鲜膜把碗口收住，用筷子扎几个小洞，放入蒸锅按火候蒸 20 分钟左右即可

6 结语

本研究以《中国居民膳食营养素参考摄入量（2013 版）》中常见食物营养素含量为权威参考数据，结合网络获取到的成品菜数据，利用计算机存储、搜索等技术初步实现了营养素——>成品菜的计算，完成了输入某种营养素，输出富含该营养素的成品菜的功能，同时通过对数据来源网址的追踪，可获得该成品菜的烹饪方法，实现指导菜品制作的功能。该研究中所用到的计算机工具、插件等皆为开源软件，可以非常方便地进行移植等操作，同时也可根据自己的需求对输入输出以及所用算法进行调整，使其可以便捷地移植到各种面向健康管理、膳食营养记录的应用中。本研究的不足之处为系统收集的成品菜数据为从菜谱网站获取，虽然经过人工校验，但也不能保证所有数据的质量。未来需要对成品菜数据进行二次筛选，保证其内容的科学性与合理性；此外，在具体的食材——>成品菜的计算中，所得成品菜列表排序完全依靠搜索引擎内置相关性计算方法确定，并未将食材因素考虑在内，如对于两种都富含某种营养素的食材出现在同一道成品菜中的搜索结果并未给予其更高的权重，这也是本研究后续需要探索的方向。

参考文献

1 国家卫生计生委. 图解：中国居民营养与慢性病状况报告（2015 年）［EB/OL］.［2016 - 09 - 22］. http：//www. nhfpc. gov. cn/jkj/s5879/201506/4505528e65f3460fb88685081ff158a2. shtml.

2 国家卫生计生委. 中国居民营养与健康现状［EB/OL］.［2016 - 09 - 22］. http：//www. moh. gov. cn/wsb/pzcjd/200804/21290. shtml.

3 牟建军，刘治全，刘富强，等. 食盐中添加钾、钙对青少年及其家庭成员高血压一级预防的随机对照试验［J］. 中华高血压杂志，2009，（6）：502 - 506.

4 王崇琴. Ⅰ级高血压患者低钠高钾饮食对血压的影响（附 70 例观察）［J］. 中国临床研究，2007，20（5）：477 - 477.

5 Pittas A G, Lau J, Hu F B, et al. The Role of Vitamin D and Calcium in Type 2 Diabetes. A Systematic Review and Meta – analysis. [J]. Journal of Clinical Endocrinology & Metabolism, 2007, 92 (6): 2017 – 29.

6 中国营养学会. 中国居民膳食指南 2016 [M]. 北京: 人民卫生出版社, 2016: 59 – 69.

7 Yummly: 个性化菜谱推荐和搜索引擎 [EB/OL]. [2016 – 09 – 21]. http: //www. yummly. com/.

8 US Department of Agriculture, Agricultural Research Service, Nutrient Data Laboratory. USDA National Nutrient Database for Standard Reference, Release 28. Version Current: September 2015 [EB/OL]. [2016 – 09 – 22]: http: //www. ars. usda. gov/ba/bhnrc/ndl.

9 薄荷网: 健康减肥网站 [EB/OL]. [2016 – 09 – 22]. http: //www. boohee. com/.

10 美国农业部. 美国膳食营养素参考摄入量 [EB/OL]. [2016 – 09 – 22]. https: //www. nal. usda. gov/fnic/dietary – reference – intakes.

11 Cheng S, Lyytikäinen A, Kröger H, et al. Effect of Calcium, Dairy product, and Vitamin D Supplementation on Bone Mass Accrual and Body Composition in 10 – 12 – y – old Girls: a 2 – y randomized trial [J]. Am J Clin Nutr, 2005, (82): 1115 – 1126.

12 Liu WY, Lu DJ, Du XM, et al. Effect of Aerobic Exercise and Low Carbohydrate Diet on Pre – diabetic Non – alcoholic Fatty Liver Disease in Postmenopausal Women and Middle Aged Men – the Role of Gut Microbiota Composition: study protocol for the AELC randomized controlled trial [J]. BMC Public Health, 2014, (14): 48 – 59.

13 Rosanoff A. Magnesium Supplements May Enhance the Effect of Antihypertensive Medications in Stage 1 Hypertensive Subjects [J]. Magnesium Research Official Organ of the International Society for the Development of Research on Magnesium, 2010, 23 (1): 27 – 40.

14 中国营养学会. 中国居民膳食营养素参考摄入量 (2013 版) [M]. 北京: 科学出版社, 2014.

15 Elastic. Elastic Search2. 3. 1 下载 [EB/OL]. [2016 – 09 – 22]. https: //www. elastic. co/downloads/past – releases/elasticsearch – 2 – 3 – 1.

16 徐晓巍, 郭臻, 王举, 等. 面向健康管理的食物营养成分数据表达方法研究 [J]. 中国食物与营养, 2016, 22 (12): 5 – 9.

17 MongoDB. MongoDB 数据库 [EB/OL]. [2016 – 09 – 21]. https: //www. mongodb. com/.

18 GitHub. elasticsearch 与 MySQL 同步工具 [EB/OL]. [2016 – 09 – 21]. https: //github. com/jprante/elasticsearch – jdbc.

19 Python Software Foundation. elasticsearch 与 MongoDB 同步工具 [EB/OL]. [2016 – 09 – 21]. https: //pypi. python. org/pypi/mongo – connector

20 GitHub. Elasticsearch 分词插件 ik [EB/OL]. [2016 – 09 – 21]. https: //github. com/medcl/elasticsearch – analysis – ik.

关于《医学信息学杂志》启用“科技期刊学术不端文献检测系统”的启事

为了提高编辑部对于学术不端文献的辨别能力，端正学风，维护作者权益，《医学信息学杂志》已正式启用“科技期刊学术不端文献检测系统”，对来稿进行逐篇检查。该系统以《中国学术文献网络出版总库》为全文比对数据库，可检测抄袭与剽窃、伪造、篡改、不当署名、一稿多投等学术不端文献。如查出作者所投稿件存在上述学术不端行为，本刊将立即做退稿处理并予以警告。希望广大作者在论文撰写中保持严谨、谨慎、端正的态度，自觉抵制任何有损学术声誉的行为。

《医学信息学杂志》编辑部