肿瘤领域关键词共现网络聚类方法研究*

宋培彦 李丹丹

（中国科学技术信息研究所 北京 100038） （首都经济贸易大学信息学院 北京 100070）

〔摘要〕 在共词计算基础上引入两步聚类算法，设计术语遴选、两步聚类和评价迭代聚类流程。以肿瘤领域术语为例构建共现矩阵和关系网络，采用两步聚类法进行自动聚类和实证研究，对其区分度进行评测。实验结果表明该方法聚类效果较好。

〔关键词〕 知识组织；自动聚类；术语；知识网络

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2018.08.011

Study on the Clustering Method of Keyword Co－occurrence Network in the Field of Tumor *SONG Pei－yan, Institute of Scientific and Technical Information of China, Beijing 100038, China; LI Dan－dan, School of Information, Capital University of Economics and Business, Beijing 100070, China*

〔**Abstract**〕 To introduce the two－step clustering algorithm based on the co－word calculation, the terminology selection, two－step clustering and the clustering processes of evaluation iteration are designed. Taking the terminology in the area of tumor as an example, the co－occurrence matrix and relational network is built. To adopt the two－step clustering method to carry out the automatic clustering and empirical study, the discrimination is evaluated. The results of the tests show that the clustering effect of the method is good.

〔**Keywords**〕 Knowledge organizaiton; Automatic clustering; Terminology; Knowledge network

1 引言

自动聚类（Clustering）是大数据环境下知识组织的重要支撑技术。事先并不依赖于既定的分类框架，而是按照语义关联性对术语和词间关系进行语义计算，以更细的颗粒度动态揭示知识的主题相关性与局部关联性，其语义关联性、自适应能力以及可移植性更高，比较适合大数据时代知识的有效组织，提升信息检索、信息推送、知识导航等智能化水平，具有重要理论和应用价值。

自动聚类的核心是语义相似度计算，主要依赖于计算模型和语料数据。当前图书情报界通过共词计算等方法构建共现关系网络，对特定领域的术语及词间关系进行关联和分析，已经较为成熟，有助于形成更深层次、更有针对性的知识组织方式。然而共现网络呈现松耦合关系，语义关联性不足，对开放领域的大规模知识关联仍有待深入探索。以开放领域的关键词共现关系网络为基础，引入聚类算法，对共现网络中的术语进行自动聚类，更精细、

〔修回日期〕 2018－05－14

〔作者简介〕 宋培彦，副研究馆员，硕士生导师，发表论文24 篇。

〔基金项目〕 2016 国家社会科学基金项目“基于知识组织的科研项目评审专家发现研究”（项目编号：16BTQ079）；2017 年度中国科学技术信息研究所创新研究基金面上项目“面向国家科技大数据的知识图谱动态构建方法研究”（项目编号：MS2017－06）。

准确地提高知识的内聚性和关联性，是本文的主要研究目标。

2 相关研究

在知识网络中节点一般代表专业概念的术语；边表示知识之间的关联关系，如词语之间的相关关系。由于知识网络是一个复杂、抽象的动态网络，因此需要按照逻辑关系对术语进行有效聚类才能删繁就简、提高知识的可读性和可用性。Chen 等[4]利用语义相关的对数似然比和 K－means 方法对文献资料搜索结果涉及的概念进行提取和聚类，同时通过聚类和引文耦合实现搜索结果的组织和可视化呈现。Xu 等[5]研究挖掘文本中实体、关系的语法和语义模型、关系的上下文句子、关系背景知识图、关系出现的背景区域等识别方法，对实体的不同语义关系进行挖掘，从而探讨实体中不同语义关系随时间的演化规律。Ibekwe－Sanjuan F 等从科技语料库中挖掘低频术语，用于科技监测[6]。Sanjuan E 从术语监测角度提出基于图结构的术语聚类方法[7]。Saason 等提出基于共词分析方法扩展概念图的研究模型，使用网页计量网络计数来改进相似性度量[8]。总之，对大规模的专业知识进行聚类、构建领域知识网络是实现知识有序化的重要工作。图书情报界也对词语聚类进行较为深入的研究。马张华等对自动聚类的特点和方法体系进行梳理，阐释自动聚类对知识组织的意义[9]。章成志等基于多语言社会化标签进行自动聚类，用于发现潜在社会关系网络[10]。杜坤等利用维基百科知识库计算语义相关度，结合特征项在文本中的表示权重，进行 K－means 文本聚类[11]。李秀霞等采用“密度－距离”快速搜索聚类算法自动确定聚类中心的类名、类团的数目等[12]。李慧宗等基于潜在犹利克雷分布（Latent Dirichlet Allocation，LDA）的社会化标签综合聚类方法，将大量歧义、不受控制的标签进行自动聚类[13－14]。刘伟等将术语间的词义关系转化为图结构，初步实现词义自动聚类[15]。崔家旺等尝试利用关联数据揭示主题词间语义关系，弥补传统共词聚类分析在语义方面的不足[16]。李慧等通过推理得出每个文档的主题分布并进行聚类，进而创建 Web 服务发现机制[17]。软光册等将 LDA 主题模型与 K－means 算法相结合，利用 LDA 模型实现文本潜在语义计算[18]。刘勘等提出改进词频－逆文件频率（Term Frequency－Inverse Document Frequency，TF－IDF）特征词加权算法，用于科技文献聚类[19]。孙海霞等采用 K－means 算法，提出一种面向大规模机构名称归一化处理应用的机构聚类方法，效果良好[20]。总体来看上述聚类方法都各具特色并取得积极进展，特别是在知识标引、个性化推荐、智能检索等领域展现出重要的应用前景。

从知识组织的角度来看，面向专业领域的术语聚类仍需要进一步研究，因为大数据环境下用户对知识的内在关联性有更多的要求，以满足个性化、精准化知识需求。知识组织本身已经成为计算机知识库，是智能检索、信息推荐、文本挖掘等不可或缺的知识资源。因此研究面向计算机的聚类算法，建立动态化、个性化的知识网络结构，形成适合大规模真实场景的专业术语聚类方法并提高其普适性。本文将从术语聚类着手，以两步聚类方法为基础，优化聚类分析方法，探索开放式、网络化、动态更新的知识聚类方法，最终用于提高知识组织工具效率、提升知识服务水平。

3 聚类计算模型

3.1 总体流程

采用两步聚类和层次聚类法对术语关键词间的关联性进行判定，包括 3 个主要阶段：（1）预处理阶段。实现对关键词数据的提取、高频词抽取并形成计算矩阵。（2）聚类阶段。通过两步聚类法，迭代进行聚类并优化，是最关键的步骤。（3）输出阶段。对领域知识网络进行可视化分析，进行动态更新。具体实现流程，见图 1。

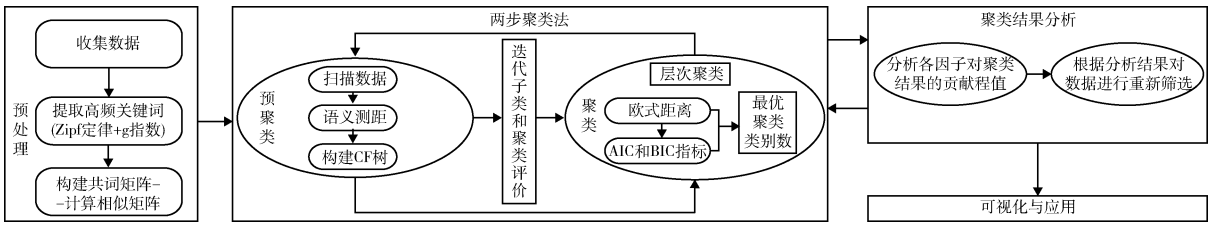


图1 领域知识网络两步聚类法总体流程

3.2 词语遴选计算模型

3.2.1 概述 选择关键词构建共词矩阵是共词分析的第1个关键步骤。根据关键词出现的词频进行确定。如果选择的关键词过少，可获得的信息量太少，则不能全面反映学科领域的构成；如果选择的关键词过多，又会造成聚类图太复杂而影响对结果的正确解读，给共词分析过程带来不必要的干扰。本文以高频词为主构建初步知识网络，然后低频词向高频词近似挂靠，逐步实现知识网络的扩展。对于高频词阈值的设定主要有两种方法：一种是基于经验判定法；另一种是结合齐普夫定律——低频词分布定律和高低频词分界公式判定高频词的阈值。

3.2.2 齐普夫第一定律

$$\frac{I_n}{I_1} = \frac{2}{n(n+1)}$$

其中 I_n 表示出现 n 次的词量，I_1 为出现1次的词的数量。齐普夫第二定律——高频词分界公式：

$$T = \frac{-1 + \sqrt{1 + 8I_1}}{2}$$

T 为高频词和低频词的分界频次，I_1 为出现1次的词的数量。如 $T=100$，则出现次数大于100的为高频词，其余为低频词。

3.2.3 低频词处理 用高频词代表领域整体的学科研究存在以偏概全的可能性，而低频词有助于获取一些隐含主题或新兴主题的信息，因此在获得一定数量的高频词后，借助共现计算可以挖掘专指性较强且含有大量重要共现关系的中、低频词对其进行补充，两者结合形成新的共词矩阵进行分析，这样在兼顾数据处理与计算代价的同时，也能够最大程度反映领域知识全貌。

3.3 建立高频关键词共词矩阵

两两统计不同关键词在同一篇文章中共同出现的次数，形成共词矩阵。为消除频次间的差距对分析结果造成影响，将共词矩阵的数据转化成相关矩阵。引入 Ochiia 相似系数法进行计算，将共词矩阵转换成相关矩阵。计算公式为：

$$\text{Ochiia 系数} = \frac{N_{ij}}{\sqrt{N_i * N_i}}$$

其中 N_i 和 N_j 分别代表关键词 i 和 j 出现的次数，N_{ij} 指关键词 i 和 j 共现的次数。在所得的相关矩阵中由于0值过多，进行统计分析时易造成较大误差，为方便处理用“1”减去相关矩阵中的每个数据，得到表示两词间相异程度的相异矩阵。

3.4 两步聚类算法

3.4.1 概述 两步聚类法分为两个步骤。第1步是预聚类，即对案例进行初步归类（允许的最大类别数由使用者指定）；第2步是正式聚类，此时将对第1步得到的初步类别进行再聚类并确定最终的聚类方案，在这个步骤中会根据实际需求确定聚类的类别数量与层级深度。

3.4.2 预聚类 通过构建和修改聚类特征树完成。聚类特征树包含许多层的节点，每一节点包含若干案例。与树模型类似，聚类特征树也将节点区分为分枝节点与叶节点。每个叶节点代表1个子类。针对每个案例，都要从根开始进入聚类特征树，依照节点条目信息指引找到最接近的子节点，直到到达叶节点为止。如果这一案例与叶节点中条目的距离小于临界值，则进入该节点，各节点的聚类特征都会更新，反之该案例会重新生成1个叶节点。如果这时叶节点的数目大于指定的最大聚类数

量，则聚类特征树会通过调整距离临界值进行重新构建。当所有案例都通过上述方式进入聚类特征树，预聚类过程结束。

3.4.3 正式聚类 将第1步得到的预聚类结果作为输入，对其进行再聚类。由于这个阶段所需处理的类别已经远远小于原始数据的数量，所以可以直接采用传统的聚类方法进行处理。一般采用合并型层次聚类法进行。其中层次聚类方法是指集群不断融合的过程，直到1个集群组将所有的记录全部覆盖。这个过程始于为每个子集定义1个初始集群。然后所有集群进行比较并且集群之间距离最小的两个集群会合并成1个集群。这个过程一直迭代执行，直到所有集群已经合并。因此它能够简单、快速地比较不同数量的集群并进行聚类。采用层次关系对整个分析过程进行计算，每一步中完成的合并或者分割都可以用二维图形，即树状图来表示。计算集群间的距离可以使用欧氏距离和对数似然距离。欧氏距离适用于连续变量的情况，对数似然距离既可以分析连续变量也可以计算离散变量。其中欧式距离的计算方法为：空间中任意两个点 A,B，对应的坐标分别为 $A(x_1,y_1)$，则其对应的平方欧式距离计算公式为：

$$|AB| = \sqrt{(x_1-x_2)^2 + (y_1-y_2)^2}$$

在层次聚类的每个阶段中都会计算反映现有分类是否适合现有数据的统计指标：赤池信息准则（Akaike Information Criterion，AIC）或者贝叶斯信息准则（Baysian Information Criterion，BIC），这两个指标越小，说明聚类效果越好，两步聚类算法会根据AIC和BIC的大小以及类间最短距离的变化情况来自动确定最优的聚类类别和数量，实现所有词语各入其类。

4 数据实验

4.1 数据预处理

考虑到期刊出版和收录具有一定的时滞性，部分期刊可能受出版和收录周期影响，2017年的数据并不完整。因此本实验选择2000－2016年的数据，以保证数据的完整性。首先对万方期刊库肿瘤领域2000－2016年核心期刊中用户关键词进行统计，得到154 535个原始关键词。根据关键词的齐普夫分布规律，按出现频次选择前10%的词语作为候选，然后由人工判断，删去其中无实际意义的关键词，对剩余的同义关键词进行合并，最后实际确定15 260个关键词。将这些关键词按照出现频率由高到低进行排序，作为最终聚类的候选数据集。

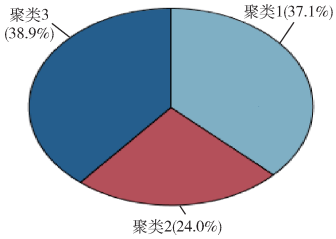
4.2 高频词选定

为选取到合适的高频词汇进行后续聚类分析，通过齐普夫第二定律进行高频词选取，得到高频词的阈值为199，取前837个关键词作为高频关键词进行后续的聚类分析。为便于计算和辨识对每个关键词加编号，以避免词语歧义问题并作为种子类别使用。

4.3 两步聚类

聚类的目的是将数据聚集成类，使得不同类间的相似性最小，而同一类中的相似性尽可能大。对选取的高频关键词得到的共现矩阵进行两步聚类分析，将共词矩阵导入到SPSS19中进行共词聚类分析，选择“分析”⟶“分类”中的“两步聚类”，计算结果，见图2。可看出这些关键词共被分为3类，每类所占比例分别为38.9%、37.1%和24%，分布比较均匀。为进一步考察聚类结果的详细信息，给出模型概要中的详细聚类信息，见图3。首先表格中给出各变量的主要分布特征；其次在进行聚类分析时需要考虑用于进行聚类分析的变量区分度。如果有变量的重要性比较低，可以考虑剔除这些变量，再重新进行聚类分析。按照关键词对聚类结果的贡献度降序排列，可以识别出核心类别，见表1。图3是以编号152和390两个词语的贡献值进行可视化展示，发现这两个词语呈现出明显的差异。可以看出关键词15“胶质瘤”的重要性最高，学科区分对比较明显，对聚类分析结果贡献度较大；其次是152“肝细胞癌”、218“胃癌”和498“膀胱癌”，在医学中属于不同的类别，对肿瘤学的学科分类也有相当程度的区分度（>0.8）；重要性最低的是390“喉肿瘤”、719“癌”、669“癌，鳞状细胞”、288“X线计算机”，属于肿瘤中的常见词或非规范词，出现次数多但对聚类分析的

贡献度较低（<0.17），所以在后续聚类分析中可以归入下位类。



最小聚类大小	40(24%)
最大聚类大小	65(38.9%)
大小比率: 最大聚类比最小聚类	1.62

图 2 聚类结果概要

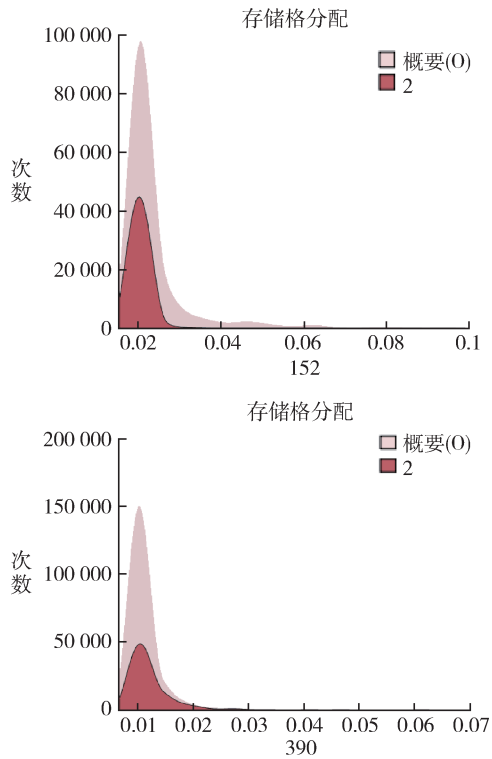


图 3 关键词对聚类结果的贡献值

表 1 关键词对聚类结果的重要性

关键词词号	重要性	关键词词号	重要性
15	1.00	610	0.17
152	0.92	288	0.17
218	0.84	669	0.16
498	0.81	719	0.16
57	0.81	390	0.15

4.4 聚类结果分析

根据上述聚类结果，从两步聚类结果中对聚类结果不重要的关键词进行归并或舍弃，如“诊断”、“化疗”、“基因”、“人类”等。将剩余词语重新计算，得到新的共词相关性矩阵，经过人工判断后最终得到 61 条核心关键词的相似度矩阵，对矩阵采用系统聚类方法进行重新聚类分析，将数据导入 SPSS19 中点击“分析”⟶“分类”⟶“系统聚类”，生成树状图，聚类过程，见表 2。可以看出“结合的群集”给出在某一步骤中参与合并的对象，在第 1 步中关键词 35 和关键词 50 合并，第 2 步关键词 21 和关键词 32 合并，第 3 步关键词 22 和关键词 60 合并。依此类推直到所有变量全部被合为一类。“系数”列给出每一步聚类的聚类系数，该数值表示被合并的两个类别之间的距离大小，即按照组间平均连接法计算出的两类间平均平方欧式距离。“阶段群集第 1 次出现”表示参与合并的对象最早出现在第几步，0 代表第 1 次出现。“下一个位置”表示在第几步中与其他类再进行合并。聚类结果树状图，见图 4。可以看出“胶质瘤”、“神经胶质瘤”、“骨肉瘤”、“鼻咽肿瘤”、“非霍奇金淋巴瘤”、“恶性肿瘤”、“子宫肌瘤”、“颅内动脉瘤”等首先被聚为一类，“预后”和“免疫组织化学”被单独分为一类，“体层摄影术”、“X 线计算机”和“误诊”被分为一类，这也说明不同类型的肿瘤病情之间的相关性比较大，同一类别的内聚性比较强，聚类效果较为理想。

表 2 关键词聚类归并计算过程

阶段	结合的群集		系数	阶段群集第一次出现		下一个位置
	群集 1	群集 2		群集 1	群集 2	
1	35	50	0.000	0	0	14
2	21	32	0.000	0	0	5
3	22	60	0.000	0	0	9
4	26	40	0.000	0	0	32
5	21	51	0.000	2	0	7
…	…	…	…	…	..	…
31	52	55	0.002	0	0	44
32	26	30	0.002	4	19	34
33	3	11	0.002	26	28	34
…	…	…	…	…	…	…
58	3	56	0.016	57	0	60
59	1	2	0.020	0	0	60
60	1	3	0.045	59	58	0

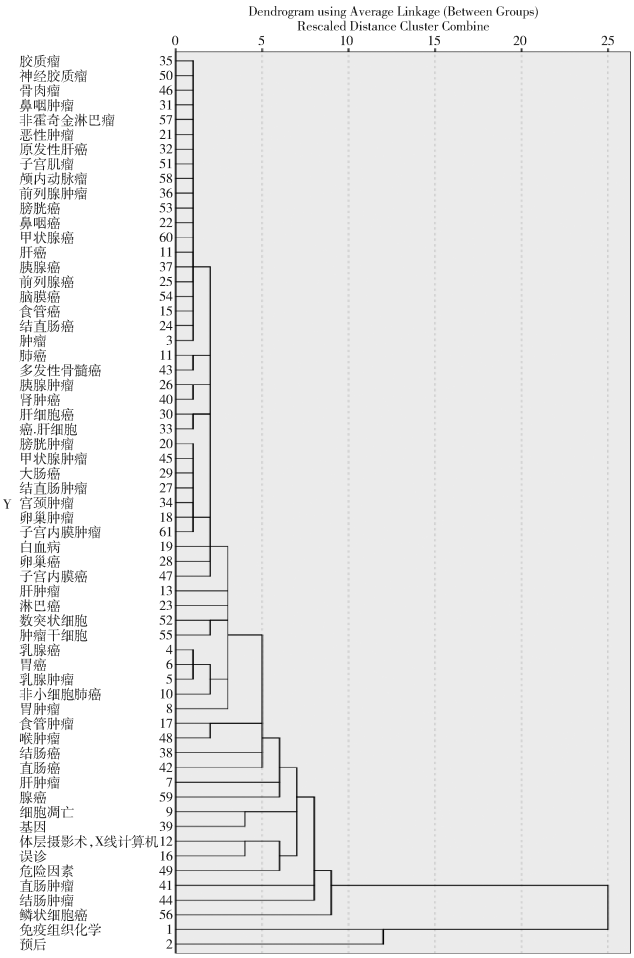


图4 聚类分析树状图

实验表明与以往的聚类方法相比，两步聚类法具有较为突出的特点。首先，用于聚类的变量可以是连续变量也可以是离散变量，能够处理不同类型的数据，比较适合于文本处理与形式化计算；其次，两步聚类法占用内存资源少，运算速度较快，适合对大量数据的聚类；最后，它是利用统计量作为距离指标辅助进行聚类决策，同时又可根据一定的统计标准来自动地评价甚至确定最佳类别数量，逐步达到较高的准确率和召回率。本实验局限性在于聚类方法本身是对语义关系的概率性统计，具有较大的动态性和主观性，因此目前对聚类结果的评价主要以定性分析为主，如何在可对比的数据集上对聚类结果进行定量评价仍需进一步研究。这也是今后要着力改进之处。

5 结语

以当前社会关注度较高、文献数据基础较丰富的肿瘤领域为例开展聚类研究，不仅有助于提升肿瘤领域知识管理和服务，而且有望形成具有一定学科普适性的知识组织新方法。对术语进行自动聚类，本质上是根据评价函数逐步逼近现实类别体系的过程，最终实现“物以类聚”，满足用户的知识导航、智能检索、术语服务等需求。两步聚类算法的计算量较小，能自动判断最佳类别数与类别层级，同时又能发掘类别间的复杂联系，比较适合处理一定规模专业领域的科技术语聚类问题。共词分析中的关键词对聚类效果也有较大影响[21-22]，聚类算法需要多次迭代，低频词的噪音干扰还比较大，将低频词（其中不少是代表新知识的新术语）进行准确聚类，特别是将关键词与现有的规范词表进行融合，将有助于提高知识聚类的准确性。

参考文献

1 Berners-Lee T, Hendler J, Lassila O. The Semantic Web [J]. Scientific American Magazine, 2008, 23 (1): 1-4.

2 Singhal A. Introducing the Knowledge Graph: things, not strings [EB/OL]. [2017-08-28]. http://52opencourse.com/186/google-knowledge-graph.

3 Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology [J]. Nucleic Acids Research, 2004, 32 (Database issue): 267-270.

4 Chen S Y, Chang C N, Nien Y H, et al. Concept Extraction and Clustering for Search Result Organization and Virtual Community Construction [J]. Computer Science and Information Systems, 2012, 9 (1): 323-354.

5 Xu Z, Luo X F. Mining Temporal Explicit and Implicit Semantic Relations between Entities Using Web Search Engines [J]. Future Generation Computer Systems, 2014, 37 (6): 468-477.

6 Ibekwe-Sanjuan F, Sanjuan E. Mining Textual Data through Term Variant Clustering: the TermWatch system [EB/OL]. [2018-07-24]. https://www.research-

gate. net/publication/37759670_ Mining_ Textual_ Data _ through_ Term_ Variant_ Clustering_ the_ TermWatch _ System.

7 Sanjuan E. TermWatch II: unsupervised terminology graph extraction and decomposition [J]. Communications in Computer & Information Science, 2013, (348): 185 - 199.

8 Saason, Ravid, Nava, et al. Improving Similarity Measures of Relatedness Proximity: toward augmented concept maps [J]. Journal of Informetics, 2015, 9 (3): 1751 - 1577.

9 马张华. 论中文信息动态自动聚类的特点和方法体系 [J]. 中国图书馆学报, 2006, 32 (6): 73 - 78.

10 章成志, 汤丽娟. 基于多语言社会化标签聚类的潜在社会关系网络发现 [J]. 情报理论与实践, 2013, 36 (9): 67 - 71.

11 杜坤, 刘怀亮, 王帮金. 基于语义相关度的中文文本聚类方法研究 [J]. 情报理论与实践, 2016, 39 (2): 129 - 133.

12 李秀霞, 邵作运. "密度 - 距离" 快速搜索聚类算法及其在共词聚类中的应用 [J]. 情报学报, 2016, 35 (4): 380 - 388.

13 李慧宗, 胡学钢, 杨恒宇, 等. 基于 LDA 的社会化标签综合聚类方法 [J]. 情报学报, 2015, 34 (2): 146 - 155.

14 李慧宗, 胡学钢, 何伟, 等. 社会化标注环境下的标签共现谱聚类方法 [J]. 图书情报工作, 2014, 58 (23): 129 - 135.

15 刘伟. 互联网同义词搜索中的词义聚类问题研究 [J]. 图书情报工作, 2013, 57 (16): 15 - 19.

16 崔家旺, 李春旺. 基于关联数据的类簇语义揭示模型研究 [J]. 数据分析与知识发现, 2017, 1 (4): 57 - 66.

17 李慧, 胡云凤. 基于主题模型的 Web 服务聚类与发现机制 [J]. 现代图书情报技术, 2016, 32 (5): 30 - 37.

18 阮光册, 夏磊. 基于主题模型的检索结果聚类应用研究 [J]. 情报杂志, 2017, 36 (3): 179 - 184.

19 刘勘, 周丽红, 陈譞. 基于关键词的科技文献聚类研究 [J]. 图书情报工作, 2012, 56 (4): 6 - 11.

20 孙海霞, 李军莲, 吴英杰. 基于 K - means 的机构归一化研究 [J]. 医学信息学杂志, 2013, 34 (7): 41 - 44, 71.

21 胡昌平, 陈果. 共词分析中的词语贡献度特征选择研究 [J]. 现代图书情报技术, 2013, 34 (7): 89 - 93.

22 胡昌平, 陈果. 科技论文关键词特征及其对共词分析的影响 [J]. 情报学报, 2014, 33 (1): 23 - 32.

(上接第 50 页)

信息平台，既包括公共卫生服务也包括医疗服务，对信息标准化的要求涵盖区域卫生信息和医院诊疗信息两方面。因此在 PHIS 标准化实践中区域卫生信息和医院信息标准化提供十分有益的借鉴和参考。

6 结语

随着人口健康信息共享数据以及业务信息系统建设的深入与细化，我国卫生信息标准体系将得到持续的更新与维护。通过 PHIS 信息标准体系的研制方法，从标准的顶层向下梳理，有助于理清思路、整体收集、协调一致、补充缺项、融合应用，为 PHIS 提供有力的标准化支撑。

参考文献

1 张黎黎, 汤学军, 刘丹红, 等. 卫生信息互联互通数据标准化方法研究 [J]. 中国卫生信息管理杂志, 2016, 13 (5): 467 - 472.

2 李小华. 医疗卫生信息标准化技术与应用 [M]. 北京: 人民卫生出版社, 2016.

3 唐辉. 物流公共信息平台标准体系解析 [M]. 北京: 电子工业出版社, 2016.

4 孟群. 我国卫生信息标准体系建设 [J]. 中国卫生标准管理, 2012, 4 (12): 24 - 31.

5 原国家卫生和计划生育委员会. 电子病历基本数据集 (WS 455 - 2014) [S]. 2017 - 05 - 30.

6 李春田. 标准化概述 (第六版) [M]. 北京: 中国人民大学出版社, 2014.

7 陈修. 互操作性概念的演变历程及在医疗卫生信息化领域的实践 [J]. 中国数字医学, 2016, 11 (5): 6 - 9.

8 汤学军, 董方杰, 张黎黎. 我国医疗健康信息标准体系建设实践与思考 [J]. 中国卫生信息管理杂志, 2016, 13 (1): 31 - 36.