情报工程 ISSN 2095-915X
第2卷 第6期 60-68 2016年12月

TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN 2095-915X Vol.2 No.6 60-68 Dec 2016

doi:10.3772/j.issn.2095-915x.2016.06.007

知识组织系统自适应构建关键问题研究

张运良[1]，李辉[2]，缪建明[3]

（1. 中国科学技术信息研究所 北京 100038；2. 北京市科学技术情报研究所 北京 100048；
3. 北方导航科技集团有限公司 北京 100070）

摘要：为了统一解决特定领域信息资源筛选和知识组织系统丰富的自动化问题，本文提出了知识组织系统自适应构建理论，明确了其技术组成，设计了构建过程迭代的工程化思路，并在医学影像小领域范围和简化流程情况下进行了实证研究。相关结果证明经过迭代后筛选的信息资源和构建知识组织系统的质量更高，知识组织系统自适应构建在理论和工程上是可行的。

关键词：知识组织系统，自适应构建，迭代，聚类分析，契合度评估

中图分类号：G254.2，TP391

Research on the Key Problems in Adaptive Construction of Knowledge Organization Systems

ZHANG YunLiang[1], LI Hui[2], MIAO JianMing[3]

(1. Institute of Scientific and Technical Information of China, Beijing 100038, China; 2. Beijing Institute of Science & Technology Information, Beijing 100048, China; 3. North Navigation Technology Group Corporation Ltd., Beijing 100070, China)

Abstract: In this paper, a theory about adaptive construction of knowledge organization systems was proposed to synchronously resolve both the problems of domain information resources filtering and knowledge organization system enrichment, and the necessary technologies were also listed. An iteration engineering process was

基金项目：本文受国家科技支撑计划项目：面向科技情报分析的信息服务资源开发与支撑技术研究（2015BAH25F01），中国工程科技知识中心建设项目：知识组织体系建设（CKCEST-2016-2-10），国家留学基金（201605360003）资助。
作者简介：张运良（1979-），男，博士，副研究员，研究方向：知识工程、知识管理、知识组织，Email:zhangyl@istic.ac.cn；李辉（1975-）女，硕士，副研究员，研究方向：情报学、知识组织；缪建明（1977-），男，博士，高级工程师，研究方向：数据处理。

designed and certified in the medical imaging domain with simplified processing methods. The results indicated that the quality of the knowledge organization system and the filtered information resources were increased with the iteration processes, which proved that both the theory and application of adaptive construction of knowledge organization systems were feasible.

Keywords: Knowledge organization systems, adaptive construction, iteration, clustering analysis, kos-resources fit evaluation

1 引言

知识组织系统作为一类集中体现领域概念及概念内在关联关系的知识工具，已经在信息资源组织、科技情报分析、新闻出版[1-2]、工程科技[3-4]等领域的知识服务中发挥重要作用。除了传统的文献标引和检索、知识导航和图谱展示等之外，知识组织系统还能够支持文本的碎片化、按需重组、精准问答和实时情报分析等个性化服务。但是由于以人工为主的知识组织系统建设通常成本较高，耗时较长，造成了不断扩大的需求与难以接受的建设成本之间的矛盾。为了解决这个问题，就需要知识组织系统能够快速构建，这种快速构建需要少量来自专家的判断之外，本身更需要大量现实的基础支撑，这些基础包括已经建成的部分知识组织系统、来自用户的数据以及真实的文献文本资源（或称为语料）[5-6]。

同时各类知识服务实际上都离不开特定信息资源集合以及面向这些资源的工具方法和分析结果的支撑，因此面向特定主题信息资源集合本身的界定也是知识服务的最重要的基础性工作之一。信息资源往往体量庞大，增长迅速，但是针对具体的知识分析和服务特定主题而言，这些资源并不都是必要的。用户可以构造出非常复杂的检索策略，但是由于全文检索、合取关系引入以及一词多义的影响，很难得到完美匹配预期目标的文本集合。而对于文本挖掘、频率统计等绝大多数分析工作，最终的分析处理结果通常会受到特定领域知识文本的质量的极大影响，也就是所谓的GIGO（Garbage in, Garbage out!，废进废出）。因此，对信息资源集合范围的界定尤为关键，而集合边界也不是一成不变的，会随着技术发展和信息资源的增长而不断调整。在实践工作中，为了得到高质量的分析研究输入，很多情况下要清洗掉一半甚至更多的资源文本。在多数情报分析的研究论文中都会出现“人工审阅”、“人工校对”、“筛选”、“剔除”等字样，一些严谨的论文更是具体给出了原始数据集及筛选后的条目数量。但是对于大体量的数据集，如达到10^4或以上的情况，用人工进行清洗筛选已经基本上不可行，不仅工作量大，而且时效性和一致性难以保障。

特定领域的信息资源既是构建领域知识组织系统稳定而客观的支撑，又是进一步依托知识组织系统提供知识服务的基础。无论是领域信息资源还是领域知识组织系统，往往一开始都是不完备的。本文希望通过分析两者的紧密耦合关系，找到将两者循环起来的可能，在少量人工干预下半自动构建知识组织系统和选择信息资源范围，从而实现知识组织系统的丰富和信息资源的筛选优化，使得服务于整个知识服务过程的信息资源及用于组织的知识系统日益精准完善，相关的研究即为自适应知识组织系统构建。

2　理论框架

知识组织系统自适应构建的理论框架如图 1 所示，两类重要的研究对象分别是特定领域的信息资源和对应的知识组织系统。技术的核心是将信息资源和知识组织系统紧密结合，使得信息资源能够利用知识组织系统加以筛选，而知识组织系统可以基于信息资源计算基础上进行丰富完善。检索策略变化、领域信息资源本身增加以及知识组织系统变化都能触发新一轮循环的开始，如果变化不大，可以采用增量的方式循环演进，如果变化较大，则可以采用全量重新计算的方式。为了保证整个流程的顺利进行，需要采用自动处理的关键技术并对过程进行人工干预。其中人工干预是个可选的过程，具体包括资源样本的研判、知识组织系统（KOS）的内容审核、检索策略评估等，其中资源样本研判主要判断信息资源是否符合数据集的构建目标，KOS 内容审核主要对选词和构建的关系进行部分或全部的审核，检索策略评估主要是评估当前建设的知识组织系统和资源的契合程度。如果不用人工干预，则可以把一些经验参数作为阈值加入系统，让系统自动运行，而针对具体领域人工干预一定程度上可以加快迭代过程。自动处理的关键技术包含基于知识组织系统的资源筛选、基于信息资源的知识组织系统丰富、检索策略的自动重构等。

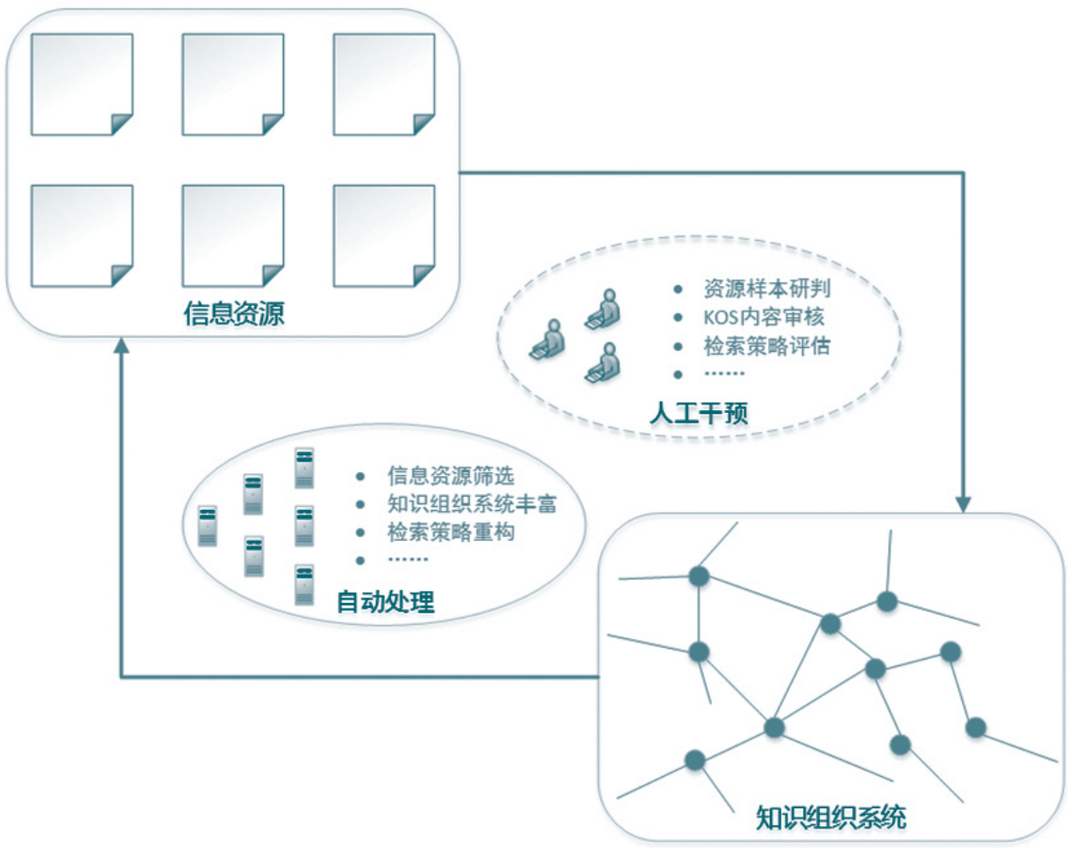

图 1　知识组织系统自适应构建理论框架

3 研究现状

3.1 信息资源筛选

对信息资源筛选本身的研究比较少，这一工作往往结合在情报分析之中，但是有几个研究方向与之密切相关。分别是信息过滤（信息检索）、数据清洗（数据质量）、语料库选择等，本文将从这些工作和本体等知识组织系统关系的角度进行阐述。

3.1.1 信息过滤

信息资源的筛选与信息过滤（检索）或信息抽取有非常相似的地方，即需要将符合用户特定需求的资源或文本取出，以供后续使用。但是两者的后续处理过程不同，信息资源筛选的结果多用于工具分析，而信息过滤的结果多用于人的分析和理解，相对而言，信息过滤由于后续有人的选择性参与，对结果的精度要求没有那么高。

一般来讲，从信息过滤的角度，需要用与文本和文本集合有关的一些参数，如词频 TF，文档频率 DF 以及它们的一些加权综合。在利用知识组织系统方面，Malhotra 等研发了一个关于多发性硬化（Multiple Sclerosis，MS）医学领域的本体，并且将这一本体应用于医学的科学文献和电子病历的检索。在实践中，MS 本体和疾病 / 并发症、基因 / 蛋白质、通路和药物等其它本体及词典共同发挥作用。通过引入这些本体，检索系统的 F 测度值较高，并且可以揭示新的发病机理的看法以及新的临床信息 [7]。Vicente-L ó pez 等提出了一种个性化搜索的方式，在该方式中用户的描绘以本体的形式来表达，主要通过对用户的兴趣以领域本体作为基础概念候选进行打分来描述，从而实现了用户描绘的简化和归一。作者采用来自开放目录项目（Open Directory Project，ODP）的分类目录作为本体进行个性化搜索的匹配。论文还采用了一种扩展激活程序来维护用户的兴趣，结果表明根据用户描绘的结果是非常有效的 [8]。此外 Bansal 等还做了关于本体技术在信息检索系统中应用的一个综述，综述中作者强调本体中用到的一些对象属性，也即关联关系 [9]，这一部分也是大多数关联知识组织系统提供的核心知识内容。

3.1.2 数据清洗

信息资源筛选和数据清洗的目的都是为了得到高质量的后续分析处理的输入资源，但是数据清洗一般针对结构化的数据，并且存在对数据字段的补齐等操作，而数据资源筛选主要用于一些非结构化的文本数据，尤其是文本的内容主体而非其元数据。

Kedad 等将数据清洗问题聚焦于术语系统的差异性，并提出了一种通过领域本体提供领域知识的解决方案，这一方法主要解决数据实例中的矛盾冲突问题 [10]。Bruggemann 等提出了一种基于本体的自动的数据清洗规则生成算法，该算法的使用国际疾病分类 (International Classification of Diseases，ICD)，这使得计算机能够找到和替换数据集合中无效的知识元组 [11]。Patuelli 等直接提出了数据质量本体 (Data Quality Ontology， DQO) 并用于一些领域（如糖尿病）的知识管理，他们也介绍了开发这一特殊本体的架构以及具体的步骤 [12]。

3.1.3 语料库选择

筛选的信息资源大多数情况下是一个文本的集合，从这一点和语料库是一致的。但是语料库的作用可能有多种，有一些是用来做语言现象分析的，可能和筛选的信息资源是一致的，但是也有相当一部分语料库是用来做机器学习参数训练使用的。

为了完成统计机器翻译的领域适应任务，Axelrod 等提出了从非常大的和目标领域相关联的通用领域并行语料库中抽取语句样例的方法。这些句子的选择和抽取的依据主要是三种基于交叉

熵（cross-entropy）的方法，其中双语交叉熵差分法（Bilingual Cross-Entropy Difference Method）效果最好，从而实现从大语料库中提取 1% 规模的子语料库能够获得和整个语料库训练一样好的效果 [13]。Kapralova 等也提及到交叉熵，但是他们通过利用可信度评分（confidence score）结合一些其他的方法（如“transcription flattening”）产生的经验阈值选择训练语料库 [14]。

3.1.4 人工筛选策略

以上研究为信息资源筛选奠定了基础，信息资源筛选核心要解决的问题是信息资源本身的差异性问题，而且其本身的处理尽管技术上可行，但是实践中还是需要人工做审阅和校对，尤其是在前期。这就面临大体量的信息资源和有限的人工精力的矛盾问题，为此本文提出如下策略。首先对初步的信息资源集合进行聚类分析，这样将原来大量的文本转变为少量的类。对所有的类，根据其内容进行统计分析，得到关于这一类的向量表示，通过这些向量，根据词频，可以初步确定该类的核心内容，为了便于进一步确认，还可以审阅这一个聚类簇的若干实例，进一步确认。确认分为三种情况，第一种是确认该类符合领域预期，则整类文档都保留，第二种是整类基本上都不符合领域预期，则整类文档都可以删除，第三类是综合的情况，可能感觉部分相关，部分无关，则需要对该类下的信息资源进行再聚类，可以通过聚类筛选出若干子集是符合要求的，如果各个子集仍然很难确定，则可以整体放弃。然后将相关信息资源集合，交给知识组织系统丰富模块。此外，还可以将知识组织系统中已有的词条作为指示信息来匹配信息资源，根据匹配度的高低对大类和类中的具体文档进行排序，当然可以预见，在知识组织系统内容不太完整的情况下，刚开始可能整体的匹配度偏低，匹配度差异性比较小，但是随着知识组织系统日益完善，则这种情况会得到改善。

3.2 知识组织系统丰富

基于信息资源的知识组织系统丰富主要包括两个方面，一是词汇的丰富，二是关系等知识的丰富。词汇的丰富主要是利用术语度等指标，也可以利用一些统计信息，如词频、词性、词长等，关系丰富主要利用关系等知识识别和抽取算法等。丰富工作可以在整理后领域信息资源集合基础上进行，选取适当指标范围的词条集合，在人工参与下进行筛选，选出符合知识组织系统选词标准的结果，如果有必要也可以人工补充关联知识。对筛选出的词条进行共现等分析，可以得到待构建关系集合，可以对集合中的候选关系按照支持度从高到低排序，根据情况采用高阈值自动认定和低阈值人工辅助确定等方式，确定知识组织系统关系。由于知识组织系统丰富方面的相关研究较多 [5]，且相对比较成熟，本文不再详细论述。

3.3 检索策略自动重构

检索策略的重构是自动生成包含各种逻辑符号（与、或、非和括号）的检索词组合，由于情报工作中，检索策略多为手工构建，自动构建研究较少。

张进早在上个世纪 80 年代末期，提出了人工智能技术对检索策略的影响问题，并提出通过专家系统和知识库建设，使得依赖人工智能技术的人机交互系统能够智能的构建检索策略，同时作者也强调了通过模式识别、机器翻译、自然语言理解等技术形成可以兼容多种人机交互形式的检索策略服务接口 [15]。金瑞红提出在智能检索系统的之中，可调用用户兴趣模型，自动修正检索策略并可依用户兴趣将检索结果迅速聚类和分类，但是还停留在一个设想 [16]。

很多研究针对不同的需求，提出具体的检索策略的研究，如黄玉基等针对基于事例推理系统中检索问题，提出最近相邻算法相结合的检索策略（L&NCBR），改善了检索效果 [17]，但是这

是一种通用的检索策略，而非智能的检索策略。Pubmed 提供使用 MESH 转换表、短语表等对输入的检索词进行扩展检索 [18]，这种检索策略的重构需要人工驱动，构建的方法也比较单一。乔丽等提出基于改进的 K-means 聚类，通过人工介入以及优化匹配过程，可以实现对输入的案例匹配更加近似的案例 [19] 的获取，但严格来讲检索策略是不变的，是检索算法的优化。

从实践上，将知识组织系统与信息资源进行关联，即对科技信息资源进行语义标注，利用这些关联进行进一步聚类分析，对资源进行再次筛选。这一过程实际上是检索策略的重构，知识组织系统的变化可以促进检索策略的完善和重新制定。

4 实证研究

4.1 数据集准备

本研究以项目组采集的 183 条医学影像专题下的期刊二次文献为基础进行实验验证，这些数据包含多个字段，但由于本文的研究思路是从文本内容出发，作者、机构、期刊、年代等外部数据等对计算关联不大，因此本研究中只选取中文标题、关键词、摘要和首段文字（可能为空）等字段作为处理的对象。

4.2 简化流程策略

由于目前使用的工具还比较分散，开发语言也不一致，没有集成，采用部分自动程序处理，部分人工处理的方式，将流程贯通。在这一流程中，核心的程序包括聚类、词频分析、标引等。整个流程通过聚类基础上的人工抽样识别以及文献标引可以实现信息资源集合的缩小，以及知识组织系统与资源集合的契合程度的评估。简化的流程图如图 2 所示，在流程中主要简化包括：

（1）检索策略重构的简化：在流程中，实际上没有严格实施检索策略的重构，而是以筛选出来的核心词表示检索策略，通过标引后的覆盖度来判断是否达标。

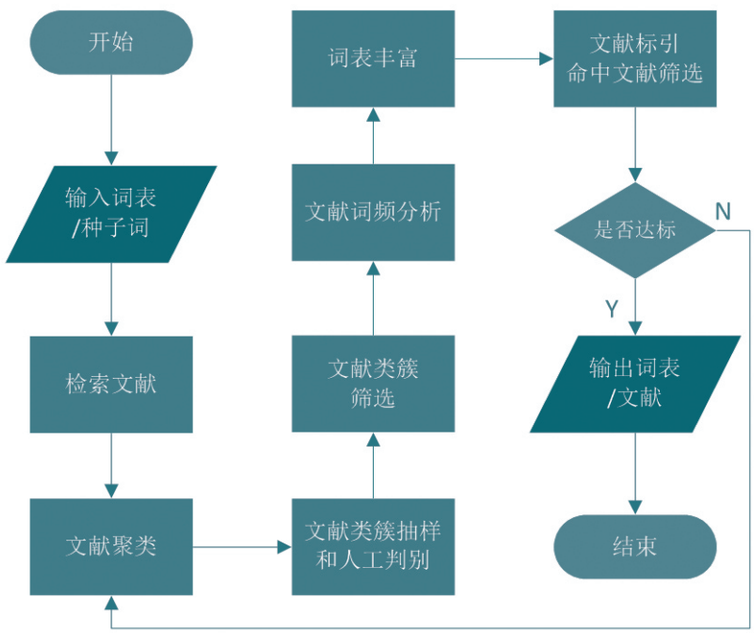

图 2　简化处理流程图

（2）信息资源集合与知识组织系统匹配度的简化：实际上采用标引的方式，将信息资源集合用知识组织系统中的词条来简化表示，通过简化后的词条的量和原文的长度进行对比，得到标引比例，依据比例进行取舍。

（3）人工审阅的简化：由于实验文献较少，聚类后每一大类数量都不会太多，没有进行向量化表示，审阅的时候没有抽样，而是对标题全部进行浏览，对摘要和正文首段进行随机浏览。

（4）当前对知识组织系统关系自动构建简化：当前暂不进行关系自动构建，对文献资源集合进行标引利用知识组织系统，不需要关系也可以进行，因此暂不构建关系。

4.3 处理结果及分析

4.3.1 信息资源过滤结果分析

采用 K-means 聚类，将初始数据集聚为 10 类，然后人工分析去掉一些相关性不强的类，得到 135 篇。其中 5、6、8、9 四个类相关性不强，如表 1 所示。

表 1 聚类分析结果

类别	数量	总结的含义
5	13	会议、人等有关的记叙
6	6	教育相关
8	16	相关设备保养维护
9	13	涉及多方面是管理机制方面的内容（似乎部分内容可以保留）

4.3.2 知识组织系统丰富结果分析

从上述 135 篇文献中提取关键词 456 条，删除重复后剩余 320 条，以此作为分词词典，以中信所开发的词频工具分析得到 2441 条词条，选择其中 n 和 vn，以词长大于等于 4，词频大于等于 5 得到 74 个词条，前 10 位的是医学影像、数字化、医学影像设备、质量控制、医学影像学、放射科、计算机、影像诊断、影像设备和影像科。

4.3.3 知识组织系统和筛选资源契合度分析

以上述 74 个词作为标引词典，到全文（这里包括标题、摘要、关键词和正文首段）进行标引，标引过程中有多个词的，多次计数，并且有包含关系的词典词重复计数，如全文出现 2 次医学影像设备，则医学影像设备和医学影像分别计数 2 次，将标引的结果的长度和全文长度做对比。进行筛选，可以看到该值比较低的大部分是与我们的领域无关的，但是存在值比较低却是有关的情况，如“基于逆向技术的骨盆环三维重构与结构分析”与本领域相关，但该论文讨论的话题，其他论文涉及不多，导致选词有偏颇，同时权重低，但是如果确实是少量研讨，对于知识组织本身的构建意义确实不大。而相对的，如果权重比较高的，确实比较相关。通过本轮计算不相关的 27 条，相关的 108 条，匹配度量为 108/135=0.8，如果阈值设为 0.8，则符合要求，如果阈值更高，可以进一步迭代，由于本文研究一些工作手工完成，就不继续迭代了。

4.3.4 综合评估

通过上述分析，文献信息资源得到进一步精简，而且得到文献更加与主题吻合，可以在以上的 108 篇文献基础上，进一步聚类和选词，得到相应的选词词条，对前 50 条作比较如表 2 所示。

通过分析经过一次迭代后选择的词顺序有所变化，也出现了 12 个词不一样的情况，但是发现这 12 个不一样的词条，可能都有一定的合理性，但是相对而言，经过迭代，“组成部分”、“科学技术”、“数据库”、“准确性”等比较通用

表 2　迭代前后选词对比表

迭代前相比迭代后的差异词	迭代后相比迭代前的差异词
医学影像技术	核医学
组成部分	数字水印
医学物理	质量管理
科学技术	成像原理
疾病诊断	放射剂量
数据库	医用磁共振成像系统
成像设备	经皮冠状动脉介入
数字减影	计算机化
准确性	算法平台
生物医学工程	医疗辐射
医学影像仪器	金融危机
分子影像学	正电子

的词汇已经去除了，新增的词条专业性更强。

5　结论与展望

知识组织系统自适应构建是应对大规模情报分析和知识服务的关键技术，本文从理论和实践两个方面探索了其内涵和实现方法。从理论上，知识组织系统自适应构建由基于知识组织系统的信息资源筛选、基于信息资源的知识组织系统丰富以及关联信息资源和知识组织系统的检索策略重构三个部分组成，通过信息资源和知识组织系统的契合程度评估，实现了流程的循环迭代的流程。本文还以医学影像专题下的小规模文本为例，通过一个简化的流程，实现了对资源的筛选以及知识组织系统的丰富，经过一次迭代，改善效果已经较为明显。当然本文研究如果需要进一步工程化，一方面需要利用统一的平台，将全部的参数配置、自动处理和人工干预过程整合到一起；另一方面面对大规模的数据，在并行处理和算法效率上也需要提升；此外，完全自动化处理的经验参数也需要通过多次实验得到。

致谢 本文在实验数据方面得到了中国科学技术信息研究所张英杰副研究馆员的帮助，特此致谢。

参考文献

[1] 张新新 . 出版机构知识服务转型的思考与构想 [J]. 中国出版 , 2015(24):23-26.

[2] 李征 , 李亚芬 , 李茂 . 数字出版资源库中的知识组织 [J]. 计算机工程与设计 , 2016, 37(5):1405-1410.

[3] 苏新宁 . 面向知识服务的知识组织 [J]. 情报资料工作 , 2015(1):5-5.

[4] 徐绪堪 , 房道伟 , 苏新宁 , 等 . 面向知识服务的水利工程知识组织模型构建 [J]. 情报杂志 , 2014(3):150-155.

[5] 张运良 , 张兆锋 , 闫莹莹 , 等 . 知识组织系统构建中对既有资源的利用方式分析 [J]. 数字图书馆论坛 , 2013(11):27-32.

[6]Aussenac-Gilles N, Bi é bow B, Szulman S. Revisiting Ontology Design: A Method Based on Corpus Analysis[J]. Lecture Notes in Computer Science. 1999, 1937: 27-66.

[7]Malhotra A, G ü ndel M, Rajput A M, et al. Knowledge Retrieval from PubMed Abstracts and Electronic Medical Records with the Multiple Sclerosis Ontology.[J]. Plos One. 2015, 10(2):

e0116718.

[8]Vicente-López E, Campos L M D, Fernández-Luna J M, et al. An Automatic Methodology to Evaluate Personalized Information Retrieval Systems[J]. User Modeling and User-Adapted Interaction. 2015, 25(1): 1-37.

[9]Bansal M, Arora J. A Review on Ontology Based Information Retrieval System[J]. International Journal of Engineering Development and Research.2016,4(2):263-265

[10]Kedad Z, Métais E. Ontology-Based Data Cleaning[J]. Lecture Notes in Computer Science. 2002, 2553(2553): 137-149.

[11]Brüggemann S. Rule Mining for Automatic Ontology Based Data Cleaning[M]. Progress in WWW Research and Development. Springer Berlin Heidelberg, 2008:522-527.

[12]Patuelli R, Reggiani A, Nijkamp P. Development of a Methodological Approach for Data Quality Ontology in Diabetes Management[J]. International Journal of E-Health and Medical Communications. 2014, 5(3): 58-77.

[13]Axelrod A, He X, Gao J. Domain adaptation via pseudo in-domain data selection[C]// Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John Mcintyre Conference Centre, Edinburgh, Uk, A Meeting of Sigdat, A Special Interest Group of the ACL. 2011:355-362.

[14]Kapralova O, Alex J, Weinstein E, et al. A big data approach to acoustic model training corpus selection[C]// INTERSPEECH 2014, Conference of the International Speech Communication Association. 2014:2083-2087.

[15] 张进 . 人工智能技术对检索策略的影响 [J]. 技术与市场 , 1988(2):53-56.

[16] 金瑞红 . 基于数据挖掘的数字档案管理 [J]. 科学时代 , 2015(6):186.

[17] 黄玉基 , 魏伟杰 , 曾文 . 基于事例推理系统中检索策略的分析与研究 [J]. 东北大学学报 (自然科学版), 2006, 27(1):33-36.

[18] 乔丽 , 姜慧霖 , 贾世杰 . 基于改进 K-means 聚类的案例检索策略 [J]. 计算机工程 , 2011, 37(5):193-195.

[19] 吴晓萍 . 探讨 PubMed 检索系统词汇的自动匹配功能 [J]. 医学信息 , 2007, 20(9):1561-1565.