DOI:10.13530/j.cnki.jlis.170019

面向网络信息资源聚合搜索的细粒度聚合单元元数据研究*

曹树金　李洁娜　王志红

摘　要　由于相关信息片段分散分布在海量且复杂多样的网络信息资源中,用户往往需要花费大量时间浏览、查询和收集所需信息。面向聚合搜索的细粒度聚合单元元数据可以深入揭示信息特征及其关联关系,促进知识发现并提升知识服务效率。因此,有必要构建细粒度聚合单元的元数据描述框架。本文以图书情报领域开放获取期刊论文、在线百科、博客等网络信息资源为数据源,采用逻辑结构分析和形式结构分析方法建立聚合单元划分框架,包括篇章层级的标题、著者等外部特征,以及节段、句群、图表单元中的话语意图和语义功能等特征;通过分析聚合单元的属性特征及复用 DC、LOM 元数据元素,构建描述聚合单元访问信息、物理信息和语义信息的元数据框架;设计检索数据库并采用实验法对聚合单元元数据框架进行验证。实验表明,该元数据框架可支持多类型网络信息资源、各层级细粒度聚合单元的检索,可为细粒度信息聚合与搜索提供理论基础与实践指导。图 7。表 6。参考文献 58。

关键词　网络信息资源　信息聚合　细粒度　聚合单元　体裁分析　元数据

分类号　G250.7

Research on the Meta-data Schema for Fine-grained Aggregation Units of Internet Resources

CAO Shujin, LI Jiena & WANG Zhihong

ABSTRACT

In the big data era, the Internet is increasingly indispensable for people to access academic or work related information. However, facing with decentralized distribution of Internet resources and lacking of in-depth description and correlation of their contents and relationships, people have to spend massive time to look through the whole search results returned and assemble the relevant information from different sources. Therefore, this paper aims to develop a meta-data schema for fine-grained aggregate units of Internet resources to reveal deeply and correlate the scattered and various kinds of information snippets, so as to meet the complex information needs of users, improve the effectiveness of retrieval and support better knowledge services.

First and foremost, this paper firstly extracted three types of free Internet resources in the field of Library

* 本文系国家社会科学基金重大项目“基于特定领域的网络资源知识组织与导航机制研究”(编号:12&ZD222)的研究成果之一。(This article is an outcome of the major project “Research on the Knowledge Organization and Navigation Mechanism of Network Resources in Specific Domains” (No.12&ZD222) supported by the National Social Science Foundation of China.)

通信作者:王志红,Email:wangzh629@163.com,ORCID:0000-0003-0651-0901 (Correspondence should be addressed to WANG Zhihong, Email:wangzh629@163.com, ORCID:0000-0003-0651-0901)

and Information Science, including OA papers, online encyclopedia, and blogs. Then, a general framework to split these resources was developed from the perspectives of logical structure and formal structure of text manually. In the aspect of logical structure analysis, it was divided into four levels: chapter level which is a whole document, section level based on the chapter title given by authors, sentence group level including macro analysis and micro analysis and chart level. The components of the whole document were fragmented by macro analysis based on the genre theory. And the information snippets revealing rhetorical intentions and semantic functions were identified using micro analysis further. The relationships between aggregate units of different levels were analyzed. Moreover, characteristics and attributes of aggregate units were depicted and classified, including 14 elements of access attributes, 3 elements of physical attributes and 2 elements of semantic attributes. Corresponding to the categories, a metadata schema was developed. Lastly, to examine the effectiveness of metadata schema, Access 2013 was used to design and develop a database, and five search tasks from genre level, section level, sentence group level and chart level were set up.

The research results conclude that the logical structures which are implications of the author's intention, have some similarities among different types of Internet resources if they have the same topics. It is feasible to apply the logical structures of the journal papers to other Internet genres. DC and LOM metadata frameworks can be reused in the metadata schema for fine-grained aggregate units of Internet resources, while there are special characteristics needed to be revealed. More importantly, search experiments implicate that it is effective to reveal and correlate aggregate units scattered in various sources and different granular when using the aggregated search database based on the metadata framework proposed in this paper. Aggregated search can support information aggregation and maintain at the same time the whole context of entire piece of information. Therefore, users can judge the relevance of search results more quickly and find the required content more effectively.

Via apreliminary study of metadata schema of fine-grained aggregation units, this research is a useful attempt to apply linguistic theories and methods to organization of Internet resources, and also a significant step toward the rising interdisciplinary research field.

The future researches are to improve the fine-grained aggregation units framework and metadata schema through analyzing other emerging Internet genres. Furthermore, vocabulary and syntactic features of aggregated units need to be analyzed so as to implement fine-grained aggregation search intelligently and construct knowledge repository automatically. 7 figs. 6 tabs. 58 refs.

KEY WORDS

Internet resources. Information aggregation. Fine-grained. Aggregation unit. Genre analysis. Meta-data.

0　引言

大数据时代，互联网已经成为人们日常生活、工作或学术研究中获取所需信息资源不可或缺的一种渠道。已有调查显示，利用网络信息资源的学术论文数量及网络引文的数量均呈现出较大幅度的增长[1-5]，人们对于互联网这一重要信息渠道的依赖性不断增加。然而，面对海量且类型复杂的网络信息资源分散分布的现状，由于缺乏对其内容的深度揭示及关联关系的挖掘与组织[6]，“信息孤岛”现象仍然普遍存

在,这严重阻碍了用户对多来源和细粒度相关信息的有效获取与利用。为此,有必要对来源分散的信息片段进行深入揭示和关联,实现网络信息资源的细粒度聚合,以有效满足用户的复杂信息需求,提升知识服务的能力。元数据正是一种对信息资源进行描述、表示、管理和使用的有效方案,通过建立网络信息资源细粒度聚合单元的元数据描述框架,既可以根据网络信息资源内容逻辑进行分解与重组,促进智能检索、知识发现、自动综述等一系列应用的有效实现,还可以通过深入揭示信息特征及其关联,帮助用户便捷、高效地获取和利用所需信息。

信息资源聚合被认为是网络环境下知识组织的一种新模式,以此种组织方式为基础的聚合搜索成为继布尔检索和联邦检索之后一种新的信息检索范式[7]。早在2008年,信息科学国际会议信息检索特别兴趣小组组建了聚合搜索专题研讨会,正式提出"聚合搜索"(Aggregated Search)的概念,认为聚合搜索是指搜寻并且重组各种来源的信息,并在一个统一的界面展示[8]。即在聚合搜索的模式下,搜索引擎为用户展示的是重新整合后的信息,而不是返回按相关性排序的完整文档列表,如查询一个乐队时,返回的结果是该乐队的描述、一般资料、相关图片、视频等[7]。2012年,欧洲信息检索会议同样组织了主题为"任务与聚合搜索"的研讨会[9]。此后,信息聚合及聚合搜索成为信息组织与检索领域的重要议题,陆续出现了大量相关的研究成果,主要包括信息聚合相关概念[10-12]和理论[13-14]、聚合方法和技术[15-17]、聚合结果呈现与评估[18-19]、信息聚合的应用与服务[20-21]等方面。从细粒度及深入内容层面描述与揭示信息资源的特征是信息聚合及聚合搜索中最基本的问题。目前细粒度聚合单元相关研究主要包括知识元[22-27]、多粒度划分与分析[28-30]、关联数据[31-35]等方面,为细粒度聚合单元的解析、抽取、分析与关联提供相应的理论基础。但是在细粒度聚合单元划分方面,已有研究多是从形式结构的角度展开[36-38],且对于如何利用元数据对细粒度聚合单元进行充分揭示与规范,除了教育学领域对于学习对象元数据及相关标准的研究[39-40]之外,鲜有研究从逻辑结构的角度划分细粒度聚合单元,并依此构建相应的元数据框架。

已有的网络信息资源元数据描述框架以粗粒度揭示为主,对于面向网络信息资源的细粒度聚合单元还缺乏统一的元数据描述标准。为此,本文主要探索如下问题:①对不同类型的网络信息资源,如何划分不同粒度的聚合单元;②为了构建细粒度聚合单元元数据框架,应该如何描述、揭示和关联不同类型和粒度的网络信息资源聚合单元;③细粒度聚合单元元数据框架在信息检索中的效果如何。为回答上述问题,本文选取图书情报领域不同类型的网络信息资源,包括OA论文与题录、在线百科、博客,在复用DC、LOM元数据元素的基础上,深入分析网络信息资源细粒度聚合单元的属性特征,构建细粒度聚合单元元数据框架,并利用基于该元数据框架设计与开发的数据库,设置相应的检索任务进行检验,为网络信息资源的细粒度聚合提供坚实的理论基础与实践指导。

1 聚合单元划分的依据与方法

1999年,Bishop[41]提出了解构和重构期刊文献单元的相关问题,在其另一个研究中指出[42],文献组件(Component)是指一篇学术期刊论文的任一逻辑部分,包括文献题名、段落标题、副标题、表格、图片、说明、参考文献、摘要、关键词、作者、作者机构、作者联系信息、脚注、尾注、附录、节段、句子、词组、单词和其他与文献关联的信息(如数据集、附加分析等),并发现读者会从期刊文献中抽取出独立的组件重新组合来撰写自己的文章。Sandusky[43]进一步划分学术期刊文献,认为学术期刊文献包括两种类型的结构,一种是文献的形式结构,比如摘要、正文、图表和参考文献等;另一种是文献的逻辑结构,即将文献组织成一个个叙述部分,提供从

文献综述到方法、结果和讨论等文献构思的整个线索。据此，本文以逻辑结构和形式结构作为聚合单元划分的重要依据。

1.1　逻辑结构分析

逻辑结构针对的是篇章形式结构中的正文部分。文献的逻辑结构包括两个方面，一个是作者根据行文框架与逻辑对整篇文档内容的分割，即节段单元；另一个为具有一定交际意图和修辞目的的语篇结构，即句群单元。节段单元往往是句群单元的宏观反映，因此需要先划分节段单元，然后根据相应的逻辑结构划分句群单元。

1.1.1　节段单元

由于期刊论文各级标题清晰地反映了论文的研究思路和结构，因此利用标题标识的节段单元可以认为是有价值且可操作的细粒度单元，通过这种标引和描述有助于用户迅速了解相应的内容以及所属文献的层级位置，从而更加有效地获取所需内容，如研究方法或结论部分。划分并描述节段单元的优点有三：①可以让用户快速检索并定位所需的节段单元内容，从而节省浏览和查找其他不相关信息的时间；②帮助用户根据各级标题把握文档的整体结构，结合其需求和所处情境判断该资源的相关性；③可以利用标题的中心语作为节段单元主题维度的描述和标引，有助于实现主题关联与聚合。如以《国外网络引文研究的现状与展望》一文为例，根据论文内部各个部分的标题，可以划分为前言、P—P 网络引文的研究、P—W 网络引文的研究、W—P 网络引文与传统引文的比较、W—W 网络引文的研究、网络引文研究的展望六个节段单元；P—P 网络引文的研究部分可以进一步划分为网络环境下引文的变化、引文数据库的对比分析两个节段单元。

与期刊论文相似的是，大多数网络信息资源的内容也有一定的行文框架与逻辑结构，需要通过文档内部标题使其更加清晰可读，如百度百科人物类词条的结构包括人物基本信息、主要经历及成就、个人生活、评价或影响等，只是各级标题不如期刊论文那样严谨和规范，甚至可能会缺失标题标识的节段单元。因此，对于网络信息资源节段单元的揭示和描述，如能明确识别节段单元就利用其标题进行标引，否则按照后文提出的体裁分析提取并存储该节段单元的话语意图。

1.1.2　句群单元

句群单元的划分以语言学领域中的体裁（Genre）及体裁分析作为基础。体裁是语言使用者共同遵守的、程式的社会交往工具[44]，被视为社会交际活动或交际事件的一种分类，如新闻报道、期刊论文、学位论文、法律文件、百科等，这些体裁还可以进行细分，如期刊论文可以分为实证型和非实证型，百科可以分为术语定义型和人物介绍型等。体裁分析是从语篇体裁角度解析特定语篇所具有的特定认知结构，对表达话语意图的宏观结构和基于交际功能的微观结构进行深层解释，最为显著的特点就在于它的解释性[45]。不同类型体裁分析的结果可能具有一定的差异性，因此需要从话语意图和交际功能的角度分析不同类型的体裁。

本研究所选的体裁类型主要包括 OA 论文及题录、在线百科、博客，OA 论文进一步分为实证型和非实证型两种，在线百科以百度百科词条建议结构为依据分为人物类词条和非人物类词条，博客以科学网对博客类型的分类为参考主要分为科研笔记类、观点评述类、人物记事类等。对于 OA 论文及题录的划分，首先在句群单元宏观分析层面，实证型论文以 Swales 提出的 IMRD 模型[46]，即介绍（Introduction）、方法（Methods）、结果（Results）和讨论（Discussion）四个组件为基础；非实证型论文则借鉴杨瑞英提出的介绍（Introduction）、理论基础（Theoretical Basis）、论证（Argumentation）和结论（Conclusion）四个组件[47]。其次，在句群单元微观分析层面，主要以 CARS（Creating a research space，创建研究空间）模型及其修正模型[48-50]为基础，采用语轮/语步（Moves/Steps）的方法进

行划分。为了避免不同研究领域或主题在体裁类型上的差异,本文选取图书情报学领域“引文分析”主题的 OA 论文及题录进行试划分,总结归纳出句群单元宏观和微观层次的描述和标识框架。如以《引文分析可视化现状》一文为例,句群单元宏观分析的组件包括介绍、理论基础、论证、结论等部分,介绍部分的微观分析和语轮/语步划分如表 1 所示。

表 1 《引文分析可视化现状》一文介绍部分微观分析及语轮/语步划分

语轮	语步	句群单元
语轮 1:提出某研究领域的论题	语步 1:提出定义	可视化技术指的是:……理论、方法和技术。
	语步 2:归纳问题相关客观知识	可视化技术包含了……数据可视化。
	语步 3:阐述对象发展的历史	可视化技术最早运用于计算科学中,……研究与应用正在逐步扩大。
	语步 4:收窄论题	因此考虑将专利引文分析和专利文本挖掘方法……的方法体系。
语轮 2:提出已有研究或知识体系的不足	语步 1:提供开展研究的理由	由于引文分析处理的是大量的抽象数据……促进引文分析相关研究的发展。
	语步 2:指出以往研究的贡献	可视化技术的一系列算法也应用到了引文分析领域……分析作品间的相似性。

考虑到学术界对在线百科、博客等新兴互联网体裁的研究较少,缺少可供参考的理论框架,且主题的一致性一定程度上可以保证文本语言功能的相似性,因此本文对于博客、在线百科等网络信息资源句群单元的划分,借鉴期刊论文的划分方法和框架,宏观分析和微观分析分别对应网络信息资源句群单元的话语意图与语义功能,表 2 部分展示了博客的观点评述类和在线百科的人物类句群单元划分后的话语意图与语义功能及其对应关系。

1.2 形式结构分析

形式结构分析中将篇章视为整体、其各个部分视为组成要素。本文以顾小清等[51]提出的学习对象的划分原则为基础,通过分析期刊论文的形式结构,拆分期刊论文的不同组成部分,将期刊论文的外部特征,包括标题、著者、机构、摘要、关键词、参考文献和附录等,作为期刊论文的元数据元素信息加以标注和存储;此外,通过分析期刊论文的正文部分,提取图片和表格信息,与逻辑结构分析后得到的节段单元和句群单元共同构成正文中的细粒度聚合单元(见图 1)。由于图片和表格往往是对信息的高度概括或是对观点的形象展示,是阅读中关注的焦点,因此提取图片和表格分别予以描述和揭示对聚合搜索具有重要的价值。值得注意的是,图表单元需要额外的文字描述作为理解的情境,而句群单元表示的内容具有一定的主题性,因此通过将图题或表头与解释该图表的句群单元(通常以“如图”或“见表”等表示)进行匹配关联,可为图表单元提供相应的情境信息。

对于其他类型的网络信息资源,虽然在形式结构上不如期刊论文那么严谨和统一,但也有类似之处,都可以划分为正文和其他描述信息,并且可以独立出图表单元,如在线百科的形式结构一般是:标题→名片→介绍→目录→正文→参考资源,正文中包含图表,标题、名片和参考资源中可以提取相关的元数据信息;博客的形式结构一般是:标题→作者信息和更新日期→正文→参考资源→评论,正文中包含图表和链接。

表 2　网络信息资源句群单元的话语意图与语义功能及其对应关系

体裁	话语意图	语义功能	句群单元
博客观点评述类——以科学网“特征因子(Eigenfactor)的背景及部分性质”一文为例	介绍	阐述对象的发展历史	特征因子(Eigenfactor,中文名由任胜利老师提出)可能是09年以来最受关注的期刊评价新指标。一方面因为特征因子创新性地将学界热切期望的“引文质量”纳入测评范围,颇具理论价值。另一方面,汤姆森公司(SCI体系的掌控者)09年初已经将其作为新版JCR的指标,正式开始实际应用。
		收窄论题	下面就特征因子的发展背景及部分性质进行一点讨论。
		指出以往研究的不足或贡献	Garfield提出的期刊影响因子,已对科学发展产生了巨大影响。但随着研究的深入,影响因子也暴露出诸多缺陷:如易被人为操纵、统计错误、不能跨学科比较、选源标准问题以及对非英文期刊不公平等。另外,期刊影响因子隐含假设:在剔除论文数量因素后,期刊越多地被引用则其影响力越高。显然,此假设成立需要满足“所有引文重要性等价”的条件。
		……	
	论证	下文提要	一些后续的推测和讨论:
		回顾以往研究	国外有学者发现国际医学期刊的特征因子与总被引存在对数变换后的高度Pearson相关,因此认为两者并没有太大的排序差异。
		提出论点	个人觉得有待讨论,Pearson相关性强仅代表宏观总体的数值同向变化,并不说明指标在微观上可相互替代或排序差异不大。
		支持论点	当样本足够时,某一样本的位序改变会导致整个序列不同程度的“联动”。精确检验排序是否存在显著差异采用非参数检验的秩和类方法可能更适宜。另一方面,总被引次数和特征因子都是基于引文的期刊的正向评价方法,自然不会有结果上的较大差异,若存在这样的差异,则说明其中一种方法可能有误。
		……	
	……	……	
在线百科人物类——以百度百科“尤金·加菲尔德”词条为例	介绍	介绍身份信息	(Eugene Garfield, September 16, 1925~) 美国信息学家。
		介绍生平事迹	1925年9月16日生于纽约。1949年毕业于美国哥伦比亚大学化学系,1954年获该校图书馆学硕士学位,1961年获宾夕法尼亚大学博士学位。1955年提出编制引文索引的设想,然后进行了数年小规模试验。60年代初创科学信息服务社(ISI),并开始编制《科学引文索引》,1963年编成出版,成为文献检索和引文分析的重要工具,为文献计量学和科学学的发展做出了重要贡献。
		……	
	贡献	介绍人物获奖或荣誉情况	1975年以来曾先后荣获美国信息产业协会名人奖、美国信息学会最佳著作奖和荣誉奖、美国化学会赫尔曼·斯考尔尼克奖。
		介绍成果,如著作等	著有《一个信息学家的论文集》和《引文索引:它的理论及在科学技术与人文科学中的应用》等。
		……	
	……	……	

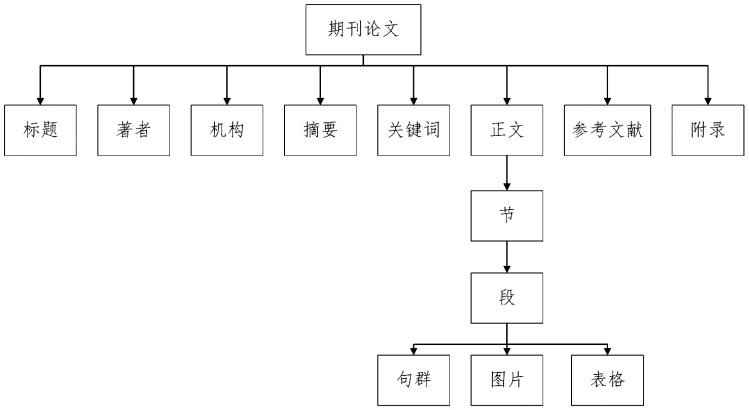

图 1　期刊论文形式结构和逻辑结构分析框架

1.3　不同层级聚合单元之间的关系

根据上文分析的结果,网络信息资源的细粒度聚合单元共包括篇章单元、节段单元、句群单元(即功能单元)和图表单元四个层级,不同层级聚合单元间的对应关系及划分依据与说明分别如图 2、表 3 所示。篇章单元包含图表单元、节段单元和功能单元,且都是一对多的关系;图表单元需要篇章单元和具有相对完整意义的相关句群单元进行解释。因此,图表单元需要与提及该图或表的句群单元相关联,由于可能存在不止一个句群单元提及图或表的情况,句群单元也可能不只提到一个图或表,所以图表单元与功能单元是多对多的关系;从形式结构上看,句群单元包含于节段单元之中,节段单元与句群单元是一对多的关系,节段单元可以指示句群单元所在的物理和逻辑结构位置。

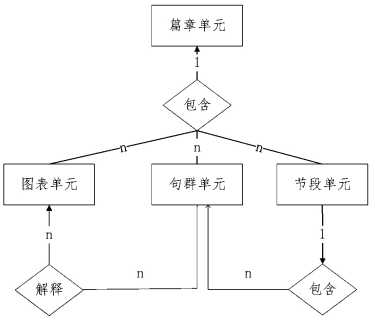

图 2　不同层级聚合单元 ER 图

表 3　聚合单元的不同层级划分依据与说明

聚合单元层级	划分依据	说明
篇章单元	完整的信息资源	一篇 OA 论文、一篇论文题录、一篇在线百科、一篇博客等。
节段单元	作者自身划分信息资源逻辑结构后的整节或整段信息资源	如百度百科“影响因子”第一章为查询,第一章的内容就是一个节段单元。
句群单元/功能单元	依据体裁分析方法划分得到的具有一定语义功能的最小粒度单元	如期刊论文《学科交叉的测度可视化研究及应用》第一段前三句“当前主流的知识组织方式是……以体现学科发展的专业化”的语义功能是归纳问题相关客观知识,这一句群单元为一个功能单元。
图表单元	依照信息资源形式结构划分得到的图表	期刊论文、在线百科或博客中的图片或表格。

2　聚合单元的元数据框架

依据已有的元数据方案设计思想[52-54]，本文在开发细粒度聚合单元元数据方案时遵循如下五个流程：①资源分析与文献调研；②建立模型；③属性提取与元素精炼；④规范控制、限定规则与著录规则；⑤系统实现。前两个阶段见上述小节，本节着重论述③和④两个步骤。

2.1　聚合单元属性特征

各个层级聚合单元属性特征及元素包括复用自 DC、LOM 元数据元素和分析其特有属性新增的元素。

（1）DC 元数据复用。篇章层级的网络信息资源聚合单元就是传统的粗粒度信息单元，节段、句群和图表层级的细粒度聚合单元同样具备与之类似的一般物理形态和结构体例，因此通用 DC 元数据对本研究具有一定的可移植性。复用的 DC 元数据元素包括标识、题名、关键词（主题）、作者、其他贡献者、出版者、日期、类型、语言、关系、来源。其中对题名、来源、关系、其他贡献者、日期、类型元素进行了相应的限定，其他元素直接复用。

（2）LOM 元数据复用。对于网络信息资源聚合单元的描述需要深入揭示细粒度信息单元的特征，且不同层级之间细粒度聚合单元具有相互包含和关联的复杂特征，因此在复用通用 DC 元数据元素之外，还需利用教育学领域以描述可重用、易组合的教学资源构件为对象的 LOM 元数据。该元数据描述的学习对象以用于教学的知识元为主，粗粒度学习对象可以自由拆分及重新组合，与本研究的目的相一致。复用 LOM 元数据元素为聚合层级、图表类型、存储路径，主要体现在物理属性方面。

（3）新增元素。复用 DC 和 LOM 元数据之后，仍然无法充分描述各层级聚合单元的属性特征，因此必须通过分析其独有特征新增相应的元素。首先，对于篇章单元层级新增“相关信息”和“资源体裁”两种元素。“相关信息”元素主要是指描述网络信息资源的一些附加基本信息，如 OA 论文可能包含会议信息、基金信息、项目信息等。已有研究[55]发现，按照综述、报告、评论等分类的数据库可以帮助用户更快地找到所需的学术信息资源，因此有必要增加“资源体裁”元素来描述网络信息资源的体裁特征。其次，由于节段单元往往具有多个层级，如“节”“小节”或“段落”，且不同节段单元之间的句群单元可能具有相同的话语意图和语义功能，为了标明节段单元和句群单元所在文档内部的逻辑位置，增强文档可读性以及揭示和聚合相同主题不同文档细粒度聚合单元的关联关系，新增“聚合层级”元素予以描述。最后，新增“话语意图”和“语义功能”两个元素进一步深入揭示句群单元这一最小粒度单元。根据宏观分析和微观分析结果，发现句群单元具有话语意图和语义功能（对应于修辞目的的描述）两个重要语言学特征，话语意图是基于文本的社会文化属性并隐含文本表示中作者所表达的交际意图，语义功能则体现为在推进交际意图实现上起到的作用，如厘清定义、回顾以往研究等，两者的描述可以从句群单元层级揭示作者使用特定文字时所赋予的交际意图和修辞目的，从而帮助用户捕捉作者的意图，有效搭建作者与用户之间沟通的桥梁。

根据已有 DC 元数据元素类型体系[56]，细粒度聚合单元属性分为访问属性、物理属性、语义属性三种类型，前两类分别来自 DC 元数据和 LOM 元数据，第三类为新增的属性类型，表 4 为网络信息资源聚合单元属性及其定义。

2.2　聚合单元元数据框架

对应于聚合单元三种属性及其子属性，聚合单元元数据框架分为访问元数据、物理元数据和语义元数据（见表 5）。访问元数据用于聚合单元外部特征的描述与管理，物理元数据用于聚合单元物理形态的描述，语义元数据用于聚合单元内容特征的描述。元数据元素包括核心元素、资源类核心元素与个别元素三种基本结构。核心元素指在各类资源对象的描述中都

表 4　网络信息资源聚合单元属性及其定义

属性	子属性	定义
访问属性	资源类型	描述信息资源的类型(如 OA 论文、学位论文、会议论文、博客、百科、课件、论坛帖子、微博等)
	资源体裁	描述该信息资源体裁类型(如 OA 论文包括实证研究、理论研究、研究综述等,博客包括科研笔记、观点评述、人物记事等)
	标题	篇章单元、节段单元、图表单元的标题
	摘要	信息资源的摘要
	关键词	概括聚合单元内容的若干词语
	来源	聚合单元的来源
	时间	信息资源发表、更新等相关时间信息
	作者	信息资源的作者
	作者单位	作者来源单位或发表该信息资源的机构
	贡献者	除作者外,对该信息资源有贡献的人或机构
	语言	语言,英语、中文或其他
	相关资源	描述信息资源参考的其他资源的信息
	相关信息	描述信息资源的其他附加信息,如会议信息、基金信息等
	图表描述	图表单元的描述信息
物理属性	存储路径	聚合单元存储的位置
	聚合层级	聚合单元的层级,篇章/节/小节/段/句群单元/图片/表格
	图表类型	图片类型:框架图/界面图/流程图/模型图等
语义属性	话语意图	按体裁分析划分,所属的宏观结构
	语义功能	按语轮/语步分析划分,句群单元归属的语步

表 5　聚合单元元数据框架

访问元数据	核心元素	标识符,关键词,来源
	资源类核心元素	标题
	篇章单元个别元素	主要责任者,其他贡献者,日期,语种,资源类型,资源体裁,相关资源,相关信息,分类
	图表单元个别元素	描述
物理元数据	核心元素	聚合层级,存储路径
	节段单元个别元素	节段单元层级
	图表单元个别元素	图表类型
语义元数据	资源类核心元素	话语意图
	句群单元个别元素	语义功能

通用的元素；资源类核心元素是指类似资源共用的元素，例如节段单元与篇章单元拥有相同的资源类核心元素“标题”等；个别元素是指仅用于描述某类资源对象的特定属性，例如图表单元中的“描述”元素。

对元数据框架中的所有元数据元素进行著录，包括元数据名称、标签、定义、注释、元素修饰词和规范档六个方面。以下分别从访问元数据、物理元数据和语义元数据中选择一个元数据元素，介绍具体的著录方式。

（1）来源（访问元数据）

元数据名称：来源

标签：Source

定义：聚合单元的来源

注释：著录来源信息

元素修饰词：资源来源；篇章单元来源；节段单元来源

规范档：无

著录细则：①资源来源：此项为篇章单元著录网络信息资源的来源信息，如 OA 论文来源期刊、在线百科、博客来源网站 URL 等；②篇章单元来源：此项为节段单元、句群单元、图表单元著录来源篇章单元的标识符，以建立聚合单元之间的关联，如多个不同节段、句群、图表均来源于同一篇论文，则为该篇论文指定的唯一标识符就可以作为这三类聚合单元所属篇章单元的来源；③节段单元来源：此项为句群单元著录其所在节段单元标识符，以描述其所在节段的位置，建立节段单元与句群单元之间的关联，如多个不同句群、图表均来源于同一篇论文中的同一个节段单元，则为该节段单元指定的唯一标识符就可以作为这两类聚合单元所属节段单元的来源。

元素设置意义：复用自 DC 元数据的来源元素，但原元素只适用于描述一篇完整的文档，对于不同层级聚合单元，本研究新增元素修饰词，包括资源来源、篇章单元来源和节段单元来源，从而使所著录的元数据不仅能描述整篇文档来源，还能够描述节段单元的篇章来源以及句群单元和图表单元的篇章单元与节段单元来源，用户可以清晰地获知不同层级聚合单元之间的联系，更好地进行定位。

（2）聚合层级（物理元数据）

元数据名称：聚合层级

标签：Aggregated Level

定义：聚合单元划分时物理上的粒度层级

注释：著录聚合单元的粒度层级

元素修饰词：无

规范档：无

著录细则：备选项（篇章、节段、句群、图表），分别对应篇章单元、节段单元、句群单元和图表单元。

元素设置意义：复用自 LOM 元数据的聚合层级元素，对每一个层级聚合单元进行著录，使用户能了解聚合单元所在的层级，同时，也是系统根据层级进行聚合的必要条件。

（3）语义功能（语义元数据）

元数据名称：语义功能

标签：Semantic Function

定义：有关句群单元内容的语义功能描述

注释：著录反映句群单元内容语义功能的受控词汇

元素修饰词：无

规范档：参见表 2 中所示的话语意图与语义功能对应关系中的功能列

著录细则：利用体裁微观分析的结果作为受控词汇进行著录。

元素设置意义：本文新添元素，著录语义功能有助于用户了解该聚合单元（句群单元）在整篇文章的写作中所起的作用，也有助于系统基于语义功能聚合不同层级的聚合单元。

3　基于细粒度聚合单元元数据框架的聚合搜索效果检验

基于所构建的细粒度聚合单元元数据框架，利用 Access 2013 设计并实现了数据库聚合搜索系统，数据库功能包括细粒度聚合及聚合

单元描述,细粒度聚合单元检索与获取,细粒度聚合单元管理。实验数据来源选用课题组成员共同采集、划分和构建的图书情报学领域网络信息资源数据集,设置五个检索实例分别检验细粒度聚合单元元数据框架用于聚合搜索的效果(见表6)。

表6 聚合搜索的检验内容及对应的检索实例

编号	检验内容		检索实例
1	支持基于体裁的聚合搜索,满足特定体裁网络信息资源搜索的需求	对体裁的聚合	检索引文分析综述类OA论文
2	支持不同粒度信息单元的聚合与检索,满足细粒度聚合单元搜索的需求	对句群单元的聚合	检索网络引文定义的句群单元
3		对图表单元的聚合	检索文献老化规律相关的图片
4		对节段单元的聚合	检索介绍引文分析发展历史的节段单元
5	综合细粒度聚合单元元数据描述框架的多重揭示功能,支持多维度细粒度聚合搜索(如作者、主题、关键词等以及节段单元话语意图、语义功能维度)	基于作者写作意图聚合句群单元	检索实证研究的引文分析的OA论文中作者对方法进行描述的句段

(1)支持基于体裁的聚合搜索,满足特定体裁网络信息资源搜索的需求

相较于传统检索系统如全文检索等,聚合搜索可以借助对资源类型和体裁等的多维度描述和揭示,帮助用户通过一个统一的平台快速便捷、一站式地检索获取所需资源。以下以检索引文分析综述类OA论文为实例进行说明。在数据库系统中选择篇章单元检索功能,设置检索条件为:资源类型="OA论文",资源体裁="研究综述",资源关键词="引文分析",图3显示检索结果(左)及篇章单元具体信息(右)。检索结果中每一个条目即一个篇章层级的聚合单元,点击"查看篇章单元具体信息",可以获得篇章单元所有元数据信息,从而进一步了解该聚合单元。正是由于在元数据框架中描述了网络信息资源的类型和体裁,才可以很快地检索到体裁为综述类的OA论文的资源,表明该元数据框架能够支持从网络信息资源类型和体裁角度进行聚合和检索。

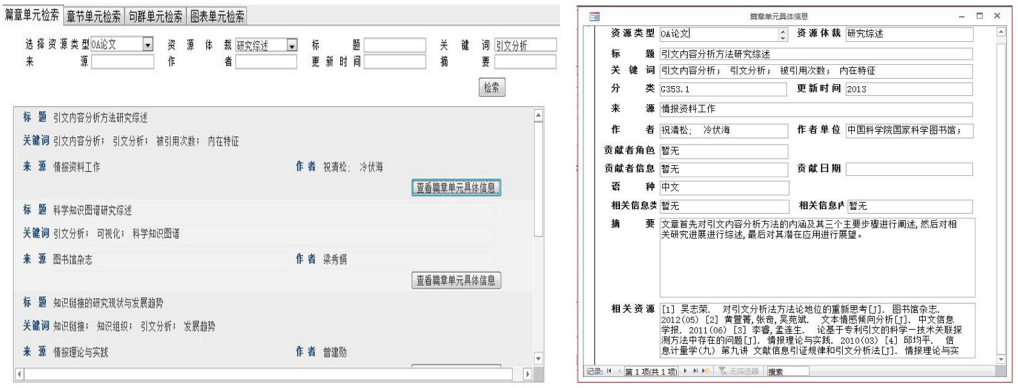

图3 检索引文分析综述类OA论文的结果

(2)支持不同粒度信息单元的聚合与检索，满足细粒度聚合单元搜索的需求

为用户聚合网络信息资源各个层级的细粒度单元是聚合搜索区别于传统检索的重要特征。相较于传统检索系统中只返回整篇文档而言，面向细粒度聚合单元的检索可以帮助用户快速定位所需资源，并通过元数据描述辅助用户判断信息的相关性，为用户节省浏览、判断及筛选相关信息的时间并减轻他们的认知负担，提高信息搜寻的效率。下面针对不同层级聚合单元分别进行举例说明。

① 检索引文分析主题下介绍分析方法的句群单元

检索条件：选择句群单元检索功能，设置网络信息资源关键词="引文分析"，句群单元话语意图="方法"，语义功能="介绍分析方法"。检索结果如图4所示。可以通过功能单元具体内容选项查看相关原文。

图4　检索引文分析主题下介绍分析方法的句群单元

可以发现，尽管句群单元脱离整体的网络信息资源而独立存在，但是由于标明了所属节段单元标题、句群功能和句群关键词，用户可以快速理解和判断句群单元的内容。与传统的检索系统相比较，以"引文分析的分析方法"为关键词检索超星发现系统，检索结果只能根据资源类型和主题维度进行聚合；使用相同的关键词在百度搜索引擎中进行检索，结果也只能根据资源主题维度进行聚合，并显示来源网址和所在关键词的上下文信息。可见，传统检索系统一方面由于没有对句群单元的属性特征进行描述，用户仅通过显示的相关片段很难进行相关性判断；另一方面，用户无法直接获得该句群单元的完整内容信息，只能费时费力地通过链接查找原网络信息资源，再从原资源中搜寻到相关内容。聚合搜索不仅可以克服上述问题，而且由于聚合单元之间关联关系的揭示，搜索结果还会返回网络信息资源的类型及体裁等相关信息，如可以再限制检索条件"源资源类型""源资源体裁"等对结果进一步聚合。

② 检索文献老化规律相关的图片

检索条件：选择图表单元检索功能，设置来源资源关键词="文献老化"。检索结果如图5所示。由图5可见，检索结果聚合并呈现了与文献老化规律相关的图表，用户可以通过预览每一条检索结果，并根据所显示的各项元数据元素进行相关性判断。同样如果图表单元条目中有"描述"这一元素，则说明该元素表示的是与该图表相关的文本中的句群单元内容，通过点击"查看相关描述信息"用户可以继续了解图表

单元所在句群单元的具体内容。相比而言,以“文献老化规律”为关键词在中国知网学术图片库进行检索,鼠标滑到对应图片时可显示图片标题和文献来源,并提供颜色、发表年度、关键词、图片类别的分面功能,点击具体图片之后也能显示图片所在的上下文信息,可见目前已较好地实现了图表单元的检索,不足之处在于还未深入描述与揭示图表单元与其他层级聚合单元之间的关联关系。

图 5　检索文献老化规律相关的图片

③ 检索介绍引文分析发展历史的节段单元

检索条件:节段单元关键词=“引文分析”and“发展历史”。检索结果如图 6 所示。通过查看节段单元具体内容可获取节段单元原文。通过从节段单元层级聚合关键词为“引文分析”和“发展历史”的信息片段,用户可以直接获得关于引文分析发展历史的内容,而无需先搜寻可能包含引文分析发展历史的多篇网络信息资源,再逐一浏览筛选之后获取与需求相关的部分内容。

图 6　检索介绍引文分析发展历史的节段单元

以上三个例子说明本文提出的元数据框架能够支持各层级细粒度单元的聚合，满足特定的信息需求。

(3)综合细粒度聚合单元元数据框架的多重揭示功能，支持多维度细粒度聚合搜索

主题只是细粒度聚合单元的一个维度，更重要的是综合细粒度聚合单元元数据描述框架的多重揭示功能，尤其是新增的话语意图和语义功能维度，从多维度实现面向细粒度聚合单元的搜索。因此，除提供基于作者、主题、关键词、引文等传统检索功能之外，细粒度聚合单元检索数据库还可依据体裁、话语意图、语义功能来聚合节段单元。例如，当需要查找引文分析主题文档中“提出以往研究不足或贡献”的句群单元时，首先选择句群单元检索功能，设置检索条件为：来源资源关键词=“引文分析”，话语意图=“介绍”，语义功能=“提出以往研究的不足或贡献”。检索结果如图 7 所示。

图 7　检索引文分析主题“提出以往研究不足或贡献”的句群单元

一般情况下，用户在查找所需信息时并非是要获得整篇文档，更多的是多个相互关联的信息片段或知识元，然而由于这些信息片段的分散性且未获得充分描述与揭示，用户需要查找、浏览、提取并整合所需相关信息，这一过程不仅费时费力，也比较容易在信息搜寻过程中产生“信息迷航”现象[57]。此外，已有研究也表明[58]，对于不同粒度的聚合单元，希望以段为聚合单元的被调查者所占比例最大，达 30%以上。面对用户信息认知负担不断加重，且对细粒度聚合单元具有较高需求的情况下，对信息单元进行多维度、细粒度聚合就显得尤为重要。根据搜索结果，用户可以通过限定话语意图和语义功能，获取与检索主题相关的表达某一写作意图与功能的句群单元。因此，说明基于细粒度聚合单元元数据框架的聚合搜索数据库，可以有效综合该元数据的多重揭示功能，支持多维度细粒度聚合单元的搜索，更加高效地满足用户多样化的信息需求。

4　讨论与结论

为了回答本文提出的研究问题，笔者充分借鉴语言学逻辑结构和形式结构分析的相关理论与基础，探索了网络信息资源细粒度聚合单元划分方法，并依据其属性特征构建了相应的元数据框架，通过实验验证了该框架的有效性。以下根据本文所提出的三个研究问题分别讨论研究结果。

对于第一个研究问题，主要从逻辑结构和形式结构两个方面进行划分。在逻辑结构分析方面，依据语言学体裁分析和修辞结构理论，将各类型网络信息资源划分为篇章单元、节段单元、句群单元和图表单元四个层级，划分方法与

依据主要为文档内部标题以及利用体裁分析中宏观和微观层面获得的文档逻辑结构和形式结构,从而构建了不同类型的网络信息资源不同层级聚合单元的划分框架,如期刊论文逻辑结构包括阐述对象的发展历史、收窄论题、指出以往研究的不足或贡献、下文提要等,形式结构包括标题、著者、机构、摘要、关键词、参考文献、附录、正文及正文中的图片和表格。进而深入分析了四个层级聚合单元之间的对应关系,如篇章单元包含图表单元、节段单元和功能(句群)单元,均为一对多关系。研究发现,尽管不同类型的网络信息资源具有不同的表达方式和结构特征,同时由于各种类型互联网体裁正在不断产生与兴起,与此相关的研究相对较少,但是对于相同主题的信息资源而言,文本片段中所隐含的作者意图即逻辑结构往往存在一定的相似之处,以语言学体裁分析和修辞结构理论为基础的期刊论文逻辑结构划分方法,对于网络信息资源的聚合单元划分具有一定的适用性,如博客观点评述类文章可以以非实证型期刊论文句群单元的语步分析框架为依据。

对于第二个研究问题,本文通过复用 DC 和 LOM 元数据元素以及分析网络信息资源不同层级聚合单元的自有特征,建立细粒度聚合单元属性分类体系,包括 14 个访问属性、3 个物理属性和 2 个语义属性,并以此为基础构建了不同层级聚合单元的元数据框架,该框架包括聚合单元访问元数据、物理元数据和语义元数据,访问元数据用于聚合单元外部特征的描述与管理,物理元数据用于聚合单元物理形态的描述,语义元数据则用于聚合单元内容特征的描述,每一类元数据的基本结构可包括核心元素、资源类核心元素与个别元素三大类,分别对应通用元素、类似资源共用元素和特定层级资源适用元素。

对于第三个研究问题,本文利用 Access2013 数据库开发工具,依据所构建的元数据框架设计并实现了聚合搜索数据库,通过从体裁、句群单元、图表单元、节段单元以及基于作者写作意图与文本交际功能的句群单元五个聚合层次,分别设置了相应的检索任务进行实验。实验结果表明,该聚合搜索数据库能够有效支持从多个维度搜索多种体裁、不同层级的细粒度聚合单元,同时也能够通过描述节段单元话语意图、语义功能等,在保留文档整体及上下文情境的情况下深入揭示细粒度聚合单元的特征。实验证明,本研究构建的细粒度聚合单元元数据框架,能够从多维度、多层级深入揭示和关联分散在各类型网络信息资源中的细粒度聚合单元,帮助用户更加快速地判断检索结果的相关性,从而更为有效地发现和获取所需的内容。此外,研究发现语言学相关理论能够为细粒度知识组织提供重要的理论基础与指导,从语言学角度揭示作者写作意图和文本功能与修辞特征,可以有效搭建作者与读者之间沟通的桥梁,减少语义的模糊性和歧义,同时还可以迅速揭示文本的特征以及文本之间的关联关系,并以此为依据将表示相同主题或具有相同功能和修辞特征的文本进行关联和聚合。因而,从语言学角度对知识组织、聚合单元等进行深入分析与研究将是未来的重要发展方向之一,有待信息组织领域的相关学者进行深入探索与研究。

本研究是对细粒度聚合单元元数据方案的初步探索,是将语言学相关理论与方法应用于网络信息组织的一种有益尝试。未来可通过对其他新兴互联网体裁的研究,进一步完善细粒度聚合单元的划分框架及元数据描述框架;同时可以深入分析细粒度聚合单元的词汇和句法特征,通过开发具体算法实现细粒度聚合单元的智能搜索与知识库的自动构建。

参考文献

[1] Moghaddam A I M K, Saberi M A, Esmaeel S M. Availability and half-life of Web references cited in information

research journal: a citation study[J]. International Journal of Information Science & Management, 2010, 8(2): 57-75.

[2] 张翠英,安美荣,王建芳,等. Web 引文数量探析[J].情报学报,2004,23(5):566-570.(Zhang Cuiying, An Meirong, Wang Jianfang, et al. Quantitative exploration of Web citation[J]. Journal of the China Society for Scientific and Technical Information, 2004, 23(5): 566-570.)

[3] 丁敬达. 国内档案学期刊网络引文的类型和相关特征分析——以 2002—2011 年《档案学通讯》、《档案学研究》为例[J].档案学通讯,2012(6):8-11.(Ding Jingda. Analyzing on the types and characteristics of Web citations of Chinese journals in Archive Science: taking *Archives Science Bulletin* and *Chinese Archives Science Study* between 2002-2011 as the examples[J]. Archives Science Bulletin, 2012(6): 8-11.)

[4] 曹树金,李洁娜. 我国图书情报领域研究者对网络信息资源的利用分析[J]. 情报学报,2014(9):994-1008.(Cao Shujin, Li Jiena. The analysis of researchers' utilization of Web resources in the field of Library and Information Science[J]. Journal of The China Society for Scientific and Technical Information, 2014(9): 994-1008.)

[5] 王志红. 我国图情领域期刊论文在线百科的利用特征探析[J]. 图书情报工作,2016,60(19):99-107.(Wang Zhihong. Research on characteristics of online encyclopedia cited by LIS journal articles in China[J]. Library and Information Service, 2016, 60(19): 99-107.)

[6] 索传军.网络信息资源组织研究的新视角[J].图书情报工作,2013(7):5-1.(Suo Chuanjun. The new research perspective of network information resources organization[J]. Library and Information Service, 2013(7): 5-1.)

[7] Kopliku A, Pinel-Sauvagnat K, Boughanem M. Aggregated search: a new information retrieval paradigm[J]. Acm Computing Surveys, 2014, 46(3): 41.

[8] Murdock V, Lalmas M. Workshop on aggregated search[J]. Acm Sigir Forum, 2008, 42(2): 80-83.

[9] Larsen B, Lioma C, Vries A D. Report on TBAS 2012: workshop on task-based and aggregated search[J]. Acm Sigir Forum, 2012, 46(1): 71-77.

[10] 邱均平,王菲菲. 基于共现与耦合的馆藏文献资源深度聚合研究探析[J]. 中国图书馆学报,2013(3):25-33.(Qiu Junping, Wang Feifei. An exploration of in-depth aggregation of library document resources based on co-occurrence and coupling[J]. Journal of Library Science in China, 2013(3): 25-33.)

[11] Žumer M, O'Neill E T. Modeling aggregates in FRBR[J]. Cataloging & Classification Quarterly, 2012, 50(5-7): 456-472.

[12] Lalmas M. Aggregated search[G]//Massimo M, Baeza-Yates R. Advanced topics in information retrieval. Berlin: Springer, 2011: 109-123.

[13] 毕强.尹长余.滕广青,等. 数字资源聚合的理论基础及其方法体系建构[J].情报科学,2015,33(1):9-14,2.(Bi Qiang, Yin Changyu, Teng Guangqing, et al. Theoretical basis and methods system construction of digital resources integration[J]. Information Science, 2015, 33(1): 9-14, 2.)

[14] 马翠嫦,曹树金,郑建瑜.多学科领域视角下网络聚合单元概念框架构建研究[J]. 情报科学,2015,33(10):16-22.(Ma Cuichang, Cao Shujin, Zheng Jianyu. Multi-discipline perspective on conceptualizing aggregation unit on Internet resources [J]. Information Science, 2015, 33(10): 16-22.)

[15] Chuklin A, Schuth A, Zhou K, et al. A comparative analysis of interleaving methods for aggregated search[J]. Acm Transactions on Information Systems, 2015, 33(5): 5.

[16] Essaimechmache F Z, Alimazighi Z. Aggregated search in XML documents[J]. Journal of Emerging Technologies

in Web Intelligence,2012,4(2):181-188.

[17] 王学东,胡宋敏,谢辉,等.多模态网络主题资源聚合与实证研究[J]. 情报科学,2014,32(7):9-13.(Wang Xuedong,Hu Songmin,Xie Hui,et al. Integration and empirical study of multi-modal Web theme resources[J]. Information Science,2014,32(7):9-13.)

[18] 张玉峰,何超.馆藏资源聚合结果的层次可视化方法研究[J].情报理论与实践,2013,36(8):41-44.(Zhang Yufeng,He Chao. Research on the hierarchical information visualization methods to display the aggregation results of library resources[J]. Information Studies:Theory & Application,2013,36(8):41-44.)

[19] Santos R L T,Macdonald C,Ounis I. Aggregated search result diversification[C]//International Conference on Advances in Information Retrieval Theory. Berlin:Springer,2011:250-261.

[20] 杜晖,邱均平.领域专家库系统构建研究[J].情报学报,2014,33(10):1022-1032.(Du Hui,Qiu Junping. Research on construction of domain expert database system[J]. Journal of The China Society for Scientific and Technical Information,2014,33(10):1022-1032.)

[21] 胡昌平,胡吉明,邓胜利.基于社会化群体作用的信息聚合服务[J].中国图书馆学报,2010,36(3):51-56.(Hu Changping,Hu Jiming,Deng Shengli. Information aggregation service based on the role of socialization groups [J].Journal of Library Science in China,2010,36(3):51-56.)

[22] 温有奎,徐国华. 知识元链接理论[J]. 情报学报,2004,22(6):665-670.(Wen Youkui,Xu Guohua. Knowledge element linking theory[J]. Journal of The China Society for Scientific and Technical Information,2004,22(6):665-670.)

[23] 周宁,余肖生,刘玮,等. 基于XML平台的知识元表示与抽取研究[J]. 中国图书馆学报,2006,32(3):41-45.(Zhou Ning,Yu Xiaosheng,Liu Wei,et al. Research on representation and extraction of knowledge unit based on XML platforms[J].Journal of Library Science in China,2006,32(3):41-45.)

[24] 苏牧,肖人彬. 基于语句聚类识别的知识动态提取方法研究[J]. 计算机学报,2001(5):487-495.(Su Mu,Xiao Renbin. A dynamic knowledge extraction method based on sentence-clustering recognition[J]. Chinese Journal of Computers,2001(5):487-495.)

[25] Zhang C,Jiang P. Automatic extraction of definitions[C]//2009 2nd IEEE International Conference on Computer Science and Information Technology. NewYork:IEEE,2009:364-368.

[26] 温有奎,温浩,徐端颐,等. 基于知识元语义网格平台的知识发现研究[J]. 计算机工程与应用,2006,42(4):4-6.(Wen Youkui,Wen Hao,Xu Duanyi. Knowledge discovery of semantic grid platform on knowledge element[J].Computer Engineering and Application,2006,42(4):4-6.)

[27] 温有奎. 基于"知识元"的知识组织与检索[J]. 计算机工程与应用,2005,41(1):55-57.(Wen Youkui. The Knowledge organization and searches with the knowledge element [J].Computer Engineering and Application,2005,41(1):55-57.)

[28] 刘平峰,余文艳,游怀杰,等. 基于模糊等价关系的文本多粒度划分方法[J].情报学报,2012,31(6):589-594.(Liu Pingfeng,Yu Wenyan,You Huaijie,et al. The method of dividing documents into multi-level granules based on fuzzy equivalence relationship[J]. Journal of The China Society for Scientific and Technical Information,2012,31(6):589-594.)

[29] 王玉林,王忠义.细粒度语义共词分析方法研究[J]. 图书情报工作,2014,58(21):73-80.(Wang Yulin,Wang Zhongyi. Research on fine-grained semantic co-word analysis method [J].Library and Information Service,2014,58(21):73-80.)

[30] The Reusable Learning Project. Types of metadata[EB/OL].[2016-12-28]. http://www.reusablelearning.org/

metadata/types.html.

[31] Tummarello G, Delbru R, Oren E. Sindice. com: weaving the open linked data[G]//Aberer K, et al. The semantic Web: lecture notes in Computer Science. Berlin: Springer, 2007: 552-565.

[32] Harting O. Langegger A. A database perspective on consuming linked data on the Web [J]. Datenbank Spektrum, 2010, 10(2): 57-66.

[33] 沈志宏，张晓林.关联数据及其应用现状综述[J].现代图书情报技术，2010，26(11)：1-9.(Shen Zhihong, Zhang Xiaolin. Linked data and its applications: an overview[J]. New Technology of Library and Information Service, 2010, 26(11): 1-9.)

[34] 李成龙.科技报告中粒度关联数据的创建与发布研究[D].武汉：华中师范大学，2014.(Li Chenglong. The creation and publishing middle linked data of scientific and technical report[D]. Wuhan: Central China Normal University, 2014.)

[35] 王忠义，夏立新，石义金，等. 数字图书馆中层关联数据的创建与发布[J].现代图书情报技术，2013(5)：28-33.(Wang Zhongyi, Xia Lixin, Shi Yijin, et al. The creation and publishing of middle linked data in digital library[J]. New Technology of Library and Information Service, 2013(5): 28-33.)

[36] 李波. 专题数据库开发中的文献粒度问题研究——以《桂西北少数民族历史、文化资源数据库》为例[J].新世纪图书馆，2014(6)：57-60.(Li Bo. Granularity problems in the development of special database: taking *The Guangxi Northwest Minority History & Cultural Resources Database* as an example[J]. New Century Library, 2014(6): 57-60.)

[37] 化柏林. 针对学术文献的句子级知识抽取研究[D]. 南京：南京大学，2013.(Hua Bolin. Research on sentence-level knowledge extraction of academic literature [D]. Nanjing: Nanjing University, 2013.)

[38] 郑彦宁，化柏林. 句子级知识抽取在情报学中的应用分析[J]. 情报理论与实践，2011 (12)：1-4.(Zheng Yanning, Hua Bolin. An analysis of the application of sentence-level knowledge extraction in Information Science [J]. Information Studies: Theory & Application, 2011 (12): 1-4.)

[39] 郑雯，吴开华，赵阳.国外教育资源元数据标准比较研究[J].图书情报工作，2005，49(1)：107-111.(Zheng Wen, Wu Kaihua, Zhao Yang. Comparison of foreign educational metadata standards[J]. Library and Information Service, 2005, 49(1): 107-111.)

[40] 王丽霞. 基于学习对象元数据的教学资源管理系统的研究与实现[D]. 呼和浩特：内蒙古大学，2008.(Wang Lixia. Research and realization of the management system of teaching resources based on the learning object metadata[D]. Hohhot: Inner Mongolia University, 2008.)

[41] Bishop A P. Document structure and digital libraries: how researchers mobilize information in journal articles[J]. Information Processing & Management, 1999, 35(3): 255-279.

[42] Bishop A P. Digital libraries and knowledge disaggregation: the use of journal article components[C]//ACM International Conference on Digital Libraries. Pittsburgh: DBLP, 1998: 29-39.

[43] Sandusky R J, Tenopir C. Finding and using journal-article components: impacts of disaggregation on teaching and research practice[J]. Journal of the American Society for Information Science and Technology, 2008, 59(6): 970-982.

[44] 王水莲. 体裁、体裁分析与体裁教学法[J]. 外语教学，2001，22(4)：91-94.(Wang Shuilian. Genre, genre analysis and genre-based teaching approaches[J]. Foreign Language Education, 2001, 22(4): 91-94.)

[45] 邓炼. 功能语法指导下英语语篇体裁分析的运用[J].当代教育理论与实践，2012(6)：117-119.(Deng Lian. Application of English discourse genre analysis under the guidance of functional grammar[J].Theory and Practice

of Contemporary Education,2012(6):117-119.)

[46] Swales J M. Genre analysis:English in academic and research settings[M]. Shanghai:Shanghai Foreign Language Education Press,2001:127-137.

[47] 杨瑞英. 体裁分析的应用:应用语言学学术文章结构分析[J].外语与外语教学,2006(10):29-34.(Yang Ruiying. Genre analysis in action:the structure of applied linguistics research articles[J].Foreign Languages and Their Teaching,2006(10):29-34.)

[48] Kanoksilapatham B. Rhetorical structure of biochemistry research articles[J]. English for specific purposes,2005,24(3):269-292.

[49] Zhang L. A study of functional units for information use of scholarly journal articles[D]. Vancouver:The university of British Columbia,2011.

[50] Nwogu K N. The medical research paper:structure and functions[J]. English for specific purposes,1997,16(2):119-138.

[51] 顾小清. 终身学习视野下的微型移动学习资源建设[M].上海:华东师范大学出版社,2011:78-80.(Gu Xiaoqing. Construction of micro-mobile learning resources from the perspective of lifelong learning[M].Shanghai:East China Normal University Press,2011:78-80.)

[52] 张晓林.元数据开发应用的标准化框架[J].现代图书情报技术,2001,17(2):9-11.(Zhang Xiaolin. The standardization framework of metadata[J].New Technology of Library and Information Service,2001,17(2):9-11.)

[53] 肖珑,陈凌,冯项云,等. 中文元数据标准框架及其应用[J]. 大学图书馆学报,2001,19(5):29-35.(Xiao Long,Chen Ling,Feng Xiangyun,et al. Chinese metadata standard framework and its applications[J].Journal of Academic Libraries,2001,19(5):29-35.)

[54] 张春景. 信息系统元数据规范应用研究[D].上海:华东师范大学,2004.(Zhang Chunjing. Research on metadata schema applying to information system[D].Shanghai:East China Normal University,2004.)

[55] 李洁娜.图书情报研究生网络学术信息搜寻行为研究——以中山大学研究生为例[J].中山大学研究生学刊(社会科学版),2014,35(2):54-63.(Li Jiena. Research on scholar information seeking behavior of graduate students of Library and Information Science:take the graduate students of Sun Yat-sen University for example[J]. Journal of the Graduate of Sun Yat-sen University(Social Sciences),2014,35(2):54-63.)

[56] Caplan P,Guenther R. Metadata for Internet resources[J]. Cataloging & Classification Quarterly,2009,22(3):43-58.

[57] 王焕景,张海燕. 网络阅读中迷航现象的认知解析[J]. 图书馆学研究,2008(10):98-100.(Wang Huanjing,Zhang Haiyan. Cognitive analysis and strategies on solving confusion of reading in Web-based environment[J]. Researches in Library Science,2008(10):98-100.)

[58] 陈飞飞.学科领域网络信息资源深度聚合的用户需求研究[D].广州:中山大学,2014.(Chen Feifei. Research on user needs for disciplines depth aggregation of network information resources[D].Guangzhou:Sun Yat-sen University,2014.)

曹树金 中山大学资讯管理学院教授。广东 广州 510006。

李洁娜 华为技术有限公司工程师。广东 东莞 523000。

王志红 中山大学资讯管理学院博士研究生。广东 广州 510006。

(收稿日期:2017-03-16)