情报工程 ISSN 2095-915X
第2卷 第1期 109-113 2016年2月

TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN 2095-915X Vol.2 No.1 109-113 Feb 2016

doi:10.3772/j.issn.2095-915x.2016.01.014

海量数据的组织与管理方法研究

曾文，李颖，韩红旗，张运良，徐红姣，翟娟华

（中国科学技术信息研究所 北京 100038）

摘要：随着信息技术的发展，需要存储和传播的信息数据量越来越大，数据的种类和形式越来越丰富，数据资源呈现规模大、多源性、多语言等特点，使得海量数据资源的组织和管理面临极大的挑战。本文分析和阐述海量数据资源在组织与管理等方面的问题和方法，并介绍在相关领域的研究工作和体会。

关键词：海量数据，数据资源，词表，组织工具

分类号：C289

Study on the Method of Data Organization and Management

ZENG wen, LI ying, HAN hongqi, ZHANG yunliang, XU hongjiao, ZHAI juanhua

(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: With the development of information technology, the need of data storage and transmission of information is getting bigger than ever. Data types and forms are richer. Data resources present scale, origin, language, etc, which makes the organization and management of the huge data resources face great challenges. This paper analyzes and expounds the difficulty and key technical problems, and introduced the related research and experience.

Keywords**:** Huge data, data resources, thesaurus, organizational tools

基金支持：本研究得到国家科技支撑计划项目（2015BAH25F00）和国家社会科学基金项目（14BTQ038）的支持。

1 引言

随着信息技术的发展，数据信息具有海量性，来源分布广泛，数据信息质量良莠不齐、数据信息内容深度千差万别。以国家科技数字图书馆的科技文献数据为例，科技文献数据资源的规模截止2014年2月，已拥有西文期刊21215307条，外文会议6477409条，外文学位论文343266条，国外科技报告1264440条，英文专利3899542条，国外标准113960条，中文期刊46657150条，中文会议1746867条，中文学位论文2456211条，中国专利4617220条，中国标准30631条。面对如此庞大的数据规模，以及日益递增的用户对数字化数据资源的服务需求，都要求数据的服务质量和模式进行相应的改变，开发和探索多元服务模式和技术思路，而海量数据的组织与管理是提供良好的海量数据服务的基本前提之一。

2 海量数据资源组织与管理存在的主要问题和挑战

信息化，网络化时代的数据规模大，数据来源和内容比较复杂、不同来源的数据，其行文结构和语言也不同，数据信息存储和组织方式也不一致。以科技文献数据资源为例，科技文献数据资源包括中外文期刊文献、中外文专利文献、中外文行业技术标准、中外文科技报告等等。如何对这些数据资源进行有效的组织和管理是最终实现智能化数据分析的前提。由于不同数据的信息采集、存储方式和组织标准不同，对于传统的信息组织工具，如叙词表如何进行改造和更新，以实现海量数据资源的组织和管理是关键性的问题和挑战之一。以科技文献数据资源为例：①科技文献数据资源的数字化加工和获取方式不尽相同。例如：国家科技数字图书馆的国内科技文献数字化加工是由各成员单位负责，各单位对过去的纸质文献和现有的电子版文献进行数字化加工可以采用不同的方式，这就使得现有的科技文献数据资源存在数据结构和描述不统一，文献题录数据错误等实际问题。外文纸质科技文献的数字化处理同样存在相似的问题，而对于购买的外文科技文献数据库，无论是外方提供的光盘数据，还是提供直接链接访问外文数据库网站方式得到的文献数据，都无法直接用于数据的信息组织和分析。因此，如何针对现有的不理想的数据、如何对外文数据，进行数据资源的组织、数据内容的挖掘和分析是一个挑战性的难题；②数据资源的数量是逐月逐年、海量递增的，数据资源存储量日益庞大，快速有效提供数据的基础是数据资源的有效组织。已有和新兴的大批新术语，在现有的数据信息组织工具——词表中未有出现，这无疑对词表的适用性提出了严峻的挑战，因此，更新和改造词表使之适用于数据资源的组织，保证海量数据的有效利用，是一个难点问题。因此本文将重点以词表为主要研究内容，介绍相关研究工作。

3 海量数据组织与管理存在的主要问题

目前，涉及数据资源问题的研究工作主要针对文本数据内容的挖掘和分析，如何挖掘海量数据背后隐含的知识和技术信息、数据之间关联信息等，是当前情报学研究领域开展研究的技术热点。但是开展这些研究面临的首要问题均是数据资源的组织和管理问题。例如：对于文献数据，由于国内现有的比较成熟的数字化文献数据加工方式，基本上是采用本地加工和外包加工方式[1]，使用TPI、TBS、TRS、DIPS等数字资源加工系统，这些系统实现已有和现有纸质文献的基本加工过程，将文献资源制作成为数字化文献数据资源，

再进行人工操作实现分类存储，组织和管理，人工录入题录信息、二次文献信息等。而对于加工后的结构化数据，由于存储格式不一，特殊字符等导致的数据加工错误，如果未经发现则会将错误的数据排序和内容全部存储于数据库中，这些诸如操作过程中出现的人为错误和文献内容本身字符问题带来的误存储错误，随着数据累积的增多，带来的问题显而易见。尽管国内外很多研究人员正从事对文本数据挖掘和分析的研究工作，并取得相应的研究成果。但这些研究成果基本是建立在数据规整、数量规模有限的文本数据实验基础之上的。通过我们对科技数字图书馆现有数据资源进行的实际调研可以发现，已加工的数据资源，其质量和规范程度距离现有在文本实验数据基础上取得的技术成果真正实现实用化还有很大的差距。对于网络数据，由于其内容来源、数据结构和格式等相对科技文献数据更为复杂，在数据采集阶段面临的难度相对更大。 主要存在以下几方面的问题，需要智能化的手段进行分析处理：

（1）现有数据的存储内容存在加工或录入错误，这些错误的存在对于海量的大数据集来说，人工识别和解决都是相当困难的。

（2）数据存储的结构、描述等不尽相同，数据需要进行结构映射和结构描述的归一化处理。

（3）网络数据的采集方式、数据的过滤、数据价值的确定等是亟待研究解决的问题。

4 数据资源组织工具和方法研究

4.1 数据资源组织工具——词表研究

海量数据资源的有效组织是提供数据快速检索和保证数据分析的重要手段。词表作为代表数据资源内容特征的、最能说明问题的、起关键作用的词集，是实现对数据资源有效、快速检索的基础。以叙词表为例，叙词表通过建立术语集，以及术语之间的用、代、属、分、参等关系，实现对数据信息资源的标引和组织。叙词表已有几十年的研究历史，积累了丰富的理论和实践经验。我国有一百多部综合或专业中文叙词表，国内外共有一千多部叙词表。由于技术的快速发展和外文文献数据的增加，使得国内的叙词表在内容上、领域上，语言上需要实现词表间的集成构建和互操作，以及术语的更新和维护。同样，完全依赖领域专家人工实现叙词表的集成及术语的更新构建是不符合现实需要和技术潮流的。但是，现有的叙词表已经无法满足不同分类、不同语言数据信息的有效组织需求，客观上需要对叙词表进行更新和改造，更好的实现对数据的标引和组织。

国外发达国家正在积极开展对多源多语言数据资源有效组织和利用的研究工作，特别是叙词表方面的研究尤为突出[2-5]。例如，欧洲共同体已经集成构建名为 Eurovoc 的叙词表（Euro vocabulary thesaurus），它可以支持欧盟 22 种官方语言。美国 DIALOG 的 DIALINDEX、BRS 的 CROSS、SDC 的 Database Index 以及 ESA 的 Quest Index 等是词表集成的典型应用。多语言标题表、多语言叙词表的互操作问题在欧洲很受重视，相关的研究项目有：MACS、HEREIN、AGROVOC、GRMET、Merimee、LIMBER 等。欧洲四国国家图书馆主办及赞助的 MACS 项目 (Multilingual Access to Subject，多语言存取主题) 是多语言标题表互操作的典型项目。我国借鉴国外的研究经验相继开展了一些研究工作，主要有中国医科院医学信息研究所研制的“医学分类主题一体化系统建设”和“统一的中国医学语言系统”、中国中医研究院中医药信息研究所研制的“中医药一体化语言系统”等，但自动化水平和应用性差。与国外研究现状相比，我国对于词表

集成构建、多语言叙词表互操作技术的研究尚未有实质性涉及和深入的研究工作。例如，我国在2005年完成的国际合作项目——联合国粮农组织（FAO）的农业多语言叙词表是以英语为源语言，汉语为目标语言的人工中文翻译和维护工作。总体上看，目前的数据资源组织工具——词表的研究还存在以下问题：

（1）国内在词表集成构建方面的研究工作起步较晚，存在自动化水平低、资源共享性差、实践成果不多、缺乏方法的深入探讨等问题；

（2）国外已有的多语言叙词表研究工作是针对隶属于同一语系，特别是非汉藏语系（如：英语）之间的多语言叙词表构建及互操作技术研究为主，而对于非同一语系之间（特别是与汉语之间）的叙词表构建及互操作技术研究涉及甚少；

（3）国内研究工作严重滞后，与国外相比，我国的词表互操作技术标准与规范还不成熟，缺乏相关的理论研究，具有影响的词表互操作技术研究成果还未见报道。

海量数据的组织和管理是海量数据分析的重要基础工作，本文认为词表仍是不可替代的重要组织和管理工具之一，目前中国科学技术信息研究所采用机器辅助翻译和人工翻译校正相结合的方法，已经完成对英语EI（Engineering Index工程索引）叙词表和日本JST（Japan Science and Technology Agency）叙词表进行了汉化研究工作，汉化的研究工作主要包括：借助专业领域翻译词典，通过与词典词条的匹配进行英语、日语的自动汉化翻译；借助机器翻译软件进行词表的翻译汉化；通过汉化人员进行人工汉化，完成自动翻译叙词表中未实现的术语和词条的翻译工作。为了节省人力、物力和财力，我们研究和开发了叙词表辅助汉化平台辅助人工进行术语汉化，以呈现叙词表中包含的概念和语义信息，保证数据的安全性、完整性和汉化结果的可恢复性。叙词表辅助汉化平台主要功能是实现术语汉化，它借助于叙词表本身的信息及各种外部资源，为汉化工作人员提供多来源的辅助汉化信息。其中，辅助汉化信息包括两类：一类是叙词表本身提供的有助于确定叙词准确含义的信息，这部分信息主要通过叙词表展示模块中提示的信息来获取；另一类是各种翻译资源提供的翻译参考信息，包括翻译词典匹配的翻译结果，机器翻译系统翻译结果等。此外，平台提供词表的浏览、搜索和修改等管理功能。通过随机进行术语抽样，人工评价的准确率结果分别为98.8%和89%。

4.2 数据资源的组织方法

词表研究的重要价值在于如何实现词表向实用化的转变，即如何快速的对领域性的海量数据集进行词表构建，利用词表进行术语标引，进而实现海量数据的有效组织和管理，例如：多语主题词表在多语言数据组织中的主要作用是揭示不同语种数据信息的主题内容并加以标识，从而将不同语种的文档纳入到统一的知识系统，揭示不同语种文档的相关关系，实现多语言文档集合的有序化。利用多语词表进行多语言数据资源组织的实现方式主要有以下两种：

（1）分别单独为每种语言的文档进行基于词表的单语言赋词标引，利用多语主题词表中提供的不同语种间的语义关系，间接实现多语言数据信息的标引。

（2）将不同语种的文档映射到一个与语言无关的词典或概念空间中，或者某种中间语言，使得获取的任意一种语种的术语标引“模式”能运用于其他语种的文档标引。

传统的基于词表的数据资源标引由于主要是依靠手工进行，费时费力。为了克服人工标引效率偏低且不能满足一致性要求的缺陷，研究人员

开始研究自动标引技术。基于多语词表的自动赋词标引方法大致可以分为三类：基于语言分析的方法、基于统计的方法及混合方法。基于语言学的方法，其主要利用词形还原[6]、复合词分解[7]、去除停用词、短语 / 组块识别[8]等方法分别对待标引的文档和词表中的术语进行处理，将处理后的文档中的词汇和词表中术语进行匹配，为文档赋予标引词。此方法不足之处在于未考虑文档和主题词表中术语的语义关联性。基于统计的方法是分析现有的已经标引好的文档资源，从中获取标引模型并将其应用于新文档的自动标引，标引模型的获取可以通过简单的统计算法来实现。目前我们针对海量数据的组织方法和技术的研究还不成熟，数量规模不大或者数据格式统一的数据可以借助于成熟的数据库技术来辅助数据的组织和管理，但对于数量巨大、数据格式不同和语言不同数据的进行有效组织和管理方法仍有待于探索和研究。

5 结语

海量数据的组织和管理研究是实现数据分析的基础和依据，脱离真实有效的数据环境进行数据的组织和管理研究不具应用价值，如前文所述，现有的研究工作仅仅是起步阶段，距离已有研究成果的实用化实现还有很大的差距，需要我们开展更加深入和具体的科研工作，未来重点需要在词表基础上拓展相关的应用研究，需要我们进一步生产有实际应用价值的数据组织和管理体系，更需要智能化的技术辅助。

参考文献

[1] 曹艳红 . 论图书馆开展数字化加工工作的必要性及对策 [J]. 科技情报开发与经济 ,2010,20（28）:111-113.

[2] Losee R M. Decisions in thesaurus construction and use[J]. Information Processing & Management,2007(4):958-968.

[3] Soergel D. Organizing Information: Principles of Data Base and Retrieval Systems (Library and Information Science[M].London: ACADEMIC PRESS, 1985.

[4]ISO5964:Guidelines for the establishment and development of multilingual thesauri [S/OL].（2009-09-00）[2011-09-05].http://www.iso.org/iso/catalogue_detail.htm?csnumber-12159.

[5] 陆定军 . 大数据时代图书馆数字资源的组织与建设 [J]. 科技情报开发与经济 ,2015,25（11）:38-40.

[6] Saric F, Snajder J, Dalbelo D, et al. Enhanced thesaurus terms extraction for document indexing[C] // Proceedings of the 27th International Conference on Information Technology Interfaces, 2005: 227-232.

[7] Markó K G, Hahn U, Schulz S, et al. Interlingual Indexing across Different Languages.[C]// Computer-Assisted Information Retrieval. 2004:82-99.

[8] Daudaravicius V. The influence of collocation segmentation and top 10 items to keyword assignment performance [C] // 11th International Conference, CICLing, 2010: 648-660.