



面向科技文献的多模态语义关联 特征提取与表达体系研究¹

□王春生 刘 珍

摘要 科技文献资源是一种多模态数据，除文本信息外，还包含丰富的图像、表格、公式、音频、视频等多种模态的信息，有利于用户充分理解科技文献资源中的知识。本文把多模态思想引入科技文献的语义表示方面，对科技文献中的图像、表格和公式信息进行语义分析，与文本信息共同表示文献语义内容，通过科技文献中多种模态信息的语义表示规则相互关系完善科技文献内容的语义化表示，发展对科技文献对多模态的表达体系。

关键词 多模态 科技文献 语义相关性 语义关联特征

1 引言

科技文献资源是一种多模态数据，具有多模态，往往包含着丰富的图像、表格、公式、音频、视频等多种模态的信息，这些多模态的信息与占主导地位的文字信息相互说明，互为补充，帮助用户充分理解科技文献资源中的知识。

具体来说，用户通过“阅读”图像来理解所表达内容的速度比单纯阅读文字来理解相同内容的速度快，而且在某些情况下，图像可以将通过文字所不能表达的内容：表格和公式是科技文献中不可缺少的部分，可以使内容的表达更加直观、严谨；音频和视频则能将科技文献资源中表达的知识具体化、可视化，有助于用户的充分理解。

在科技文献中，分析单模态信息与综合多模态信息所产生的语义理解之间可能会存在偏差，因此可以利用不同模态的相容互补性，对图像等多种模态的信息进行语义表示，发现不同模态的语义特征的潜在语义相关性，对于理解仅考虑单模态难以明确的语义可以起到积极的促进作用。因此，如何充分有效地对科技文献中的多模态信息加以关联利用，就成为了一个亟待解决的问题。

2 多模态的概念

2.1 多模态的基本概念

多模态(multi-modal)的概念是相对于单模态(unimodal or single-modality)而言的。多模态的研究一般指使用两个或两个以上不同模态的信息来解决一个特定的问题，目前还没有形成一个明确的广义上的定义。最早出现的关于多模态的文献是 1968 年关于模式识别中的多模态测试的研究^[1]，之后在 1970 年关于多信号检测的功能研究的论文中相对应于单信号提出了双信号的概念，即多模态的信号检测^[2]，同时期也出现于医学多模态治疗方法、生物系统中的多模态学习等领域。20 世纪 90 年代中后期，有关多模态的研究逐渐增多，应用领域也更加广泛。

2.2 “多模态”与“多媒体”的比较

与多模态相似的是“多媒体”(multi-media)的概念，多媒体是指组合两种或两种以上媒体的一种人机交互式信息交流和传播媒体。使用的媒体包含连贯的媒体数据(如视频、音频等)和离散的媒体数据(如文本、图形、图像等)^[3]。它不只是各种媒体的简单复合，而是一种把文字、图形、图像、动画和声音等形式的信息结合在一起，并通过计算机进行综合处

¹ 本文得到国家科技支撑计划项目(2011BAH11B04)、国家社科基金项目(C12BTQ006)、中国科学技术信息研究所项目基金(TT-201105)的资助。



理和控制,而支持完成一系列交互式操作的信息技术。目前,多模态技术在科学数据检索和处理、商业应用、教育和职业培训、娱乐等领域有广泛的应用。其研究大都是围绕着如何为用户提供更好的多模体信息服务,也就是广泛的视听觉服务展开的^[2]。而多模态的研究则侧重于通过对同一个目标的不同特征或同一特征的不同信息的对比和融合来解决一个特定的问题。目前主要应用于图像、音频、视频的处理和检索、医学图像配准与融合、生物特征身份识别、语义分析等领域。

3 国内外相关领域的研究现状及动态

3.1 多模态研究成果及动态

多模态是一个较新的研究领域,由于强调的是运用不同模态的信息解决问题的方法,并不局限于某一个学科领域,因此国内外目前的研究涉及许多不同的模态因素,例如图像、视频、音频、生物特征、语义表达等,研究领域较多,主要包括以下几个方面。

3.1.1 多模态图像自动标注和检索

图像对于人们理解信息有重要的补充作用,对于图像的标注和检索可以追溯到 20 世纪 70 年代中期,人们对图像库中的每张图像进行关键字的标注,然后利用人工标注的文本信息来检索图像。随着科学技术的发展,根据图像内容进行自动标注和检索的技术已越来越受到关注。

3.1.2 多模态医学图像配准与融合

随着医学影像学和计算机技术的完善,医学图像在应用中的地位越来越重要。但是,从单一的图像中无法得到全面的诊断信息,人为的空间构想又会影响结果的准确性,因此多模态医学图像配准与融合的研究得到了广泛的的关注,成为目前生物医学工程中的一个热点问题^[3]。研究主要集中对同一名患者在不同时间、不同传感器或不同条件下获取的两幅或多幅图像进行配准和融合的方法和关键技术的讨论和改进方面,涉及数字图像处理、计算机图形学和医学领域的知识,是计算机图形学和图像处理在生物工程领域中的重要应用^[4]。

3.1.3 多模态身份识别

多模态身份识别方面的研究包括多模态生物特征识别和音视频,视频中的发言人身份识别。对于多模态生物特征识别的研究在 2000 年之前处于起步阶段,之后便开始迅速发展。该研究通过结合多种

生物特征(如指纹、虹膜、人脸、掌纹、静脉等)来进行对于个人身份的鉴定,提高了识别的准确性,是生物特征识别技术研究领域的热点之一^[5]。音视频中的发言人身份识别通过分析发言人的音调等语言特征,结合同步画面中的面部特征来确定发言人的身份^[6],在视频会议等应用中有重要的作用^[7]。

3.1.4 多模态视觉信息的分类与检索

该研究主要应用于对于视频事件的检测、分类和检索方面。视频是没有结构的数据流,主要包括图像、音频和文本三种视觉数据,具有复杂性和随机性,因此用单一特征进行查询得到的结果并不能令人满意。多模态信息融合可以针对视频的多种视觉信息分别查询,再进行有效的融合,能够取得较好的效果。早期的视频检索是根据视频的底层视觉特征(如图像纹理特征等)进行分类和处理的^[8],随后逐渐发展到基于概念的视频检索^[9]。

此外,多模态研究还包括多模态人机交互系统研究^[10]、多模态语义分析^[11]、机器人口识别^[12]、多模态情感识别^[13]、多模态信息融合的一般功能模型设计^[14]等方面。

3.2 图像语义特征提取与表示研究成就

科技文献资源中涉及大量的图像信息,对于图像与文本信息的语义关联特征提取与表示是研究的一个重点。目前国内外对于图像语义特征的研究主要包括图像自动标注和图像检索。

图像自动标注的目的是让计算机自动用关键字等文本信息进行图像标注,通过标注在图像的底层视觉特征与高层语义特征之间搭起一座桥梁。目前大多数图像自动标注系统是结合统计方法来确定图像视觉特征和文本之间的关系,在一个训练集中对图像进行标注,之后该训练集中已标注过的视觉特征和文本之间的关系就可以用来标注该集以外的新的图像^[15]。目前计算机提取的视觉特征主要包括颜色特征、纹理特征和形状特征等,研究主要集中在于对更有效的自动标注方法和模型的开发方面^{[16][17]}。

早期的图像检索使用的是基于文本的检索方式,起源于 20 世纪 70 年代,当时图像数量相对较少,图像的标注工作可以完全由人工进行。但随着数字摄影技术和互联网技术的高速发展,手工标注所耗费的人力和时间太大,而且对于图像的不同理解可能带来不同的标注,因此基于文本的检索方式



已经不能很好地适应庞大图像库图像检索。为了解决这一问题，20世纪90年代初，研究人员提出了基于内容的图像检索方式。对于基于内容的图像检索的讨论起源于1993年美国国家科学基金会(NSF)组织的研讨会，会议认为可视化信息管理系统可在科学、工业、医学、环境、教育、娱乐等多方面得到应用，应成为研究人员的主要研究领域。之后，美国伊利诺伊大学的NCSA(National Center for Supercomputing Applications)组织在1993年发表了第一个可以显示图片的Mosaic浏览器^[2]。基于内容的图像检索是利用图像的视觉特征信息进行检索，用户根据自己的检索需求提供一连串查询图像，系统从该图像中提取出视觉特征，再在图像库中检索与视觉特征相似的图像返回给用户。此外，图像检索还涉及图像相似度的度量和学习问题，即如何判断图像库中的图像与用户查询图像之间的相似度。对于相似度度量的方法也是研究的一个热点问题，已提出的方法包括基于区域的相似度学习^[3]、多模态相似性传播方法^[4]、基于区域的模糊特定匹配方法等^[5]。目前已有的多模态图像检索系统包括 QBIC 图像检索系统、Vimage 图像搜索引擎、Retrieval-Ware 图像检索工具、Photoshot 图像检索工具和 VisualSEEK 图像检索工具等。

目前图像语义研究的重点主要是语义鸿沟问题，即由于计算机获取的底层视觉信息与用户对图像理解的高层语义信息不一致而导致的底层特征提取和高层检索需求之间的距离。这是图像语义理解面临的根本障碍，其根源之一就是图像本身所固有的多义性。研究人员试图从不同的角度来解决这一问题，例如在图像的更小区域内进行特征匹配，进行相关反馈来改善图像检索效果^[6]等。

3.3 表格语义分析与表示研究

3.3.1 基于图像的表格识别与处理研究

现代社会中，信息资源迅速膨胀，除了数字化信息资源外，还存在着大量的纸质文档资源。为了更好地利用和管理这些纸质文物资源，就需要利用计算机对大量的纸质文档资料进行数字化处理和存储，由此产生了光学字符识别技术(OCR)。OCR技术通过扫描和摄像等光学输入方式读取纸张上的文字图像信息，然后利用各种模式识别算法分析文字形态特征，判断出汉字的标准编码，并按照通用格式存储在文本文件中^[7]。表格识别是光学字符识别技

术重要的应用领域之一，由于扫描而来的图像中存在的都是像素点，因此最初这种表格识别技术是基于图像的^[8]。目前，对于基于图像的表格识别的研究主要集中在对关键技术及系统的讨论和改进方面。例如，文献^[9]重点讨论了表格识别预处理技术与表格字符串提取算法；文献^[10]研究并实现了一种手写表格识别系统，可以对纸质手写表格图像进行扫描，设置输出规则，进行表格识别处理。

3.3.2 电子文档中表格式信息的抽取

表格式信息抽取一般包括表格检测、表格分解与处理两个方面的工作。值得注意的是，表格检测与处理的一个关键问题在于输入的格式。我们可以把电子文档中的表格分为两类^[11]，一类是原始文本表格，使用ASCII等宽字符文本，用空格或特殊字符作为分隔符。另一类是多格式文本表格，包括基于LaTeX、PDF、HTML等格式的文本。目前大多数的研究是针对基于HTML格式的表格展开的^{[12][13]}。

然而，大量的科技文献是以PDF格式存在的，因此也有研究者对基于PDF格式的表格识别和数据抽取技术进行了研究，但此方面研究尚处于起步阶段。PDF中的表格是基于视觉的，具有独特的结构，被称为“文字流”表格。用户一般只能直观地从展示结果看到表格，而无法直接从文档格式中获取表格信息^[14]，因此其处理相比其他格式的表格处理更为困难。此方面的研究集中在对关键技术的探讨和改进方面，例如文献^[15]提出了一种PDF表格的元数据抽取的算法，即基于定位分析和关键词匹配技术，确定表格单元内容，识别表格结构的方法；文献^[16]提出了一种通过图像线检测PDF中表格位置的算法，通过对PDF文档页面中的“稀疏线”(sparse line)进行探测，来判断内容中标题、表格、脚注等具有“稀疏线”特征的文字的布局信息。也有研究者提出了先将PDF文档通过pdf2html工具(<http://pdf2html.sourceforge.net>)转换为HTML或XML格式，再进行表格的识别和解析^[17]。

3.4 公式语义分析与表示研究

1968年，Anderson在博士论文中首次提出了公式识别的问题^[18]，之后公式处理的研究进展比较缓慢，进入20世纪90年代，相关的研究才逐渐增多。前文提到的OCR系统对手写、印刷体文本都有很高的识别率，已经广泛应用于办公自动化、快速录入等



领域,但对于分析公式结构、识别出文档中的数学公式还设有很好的效果^[3]。目前数学公式图像处理方面的研究较多,提出了一些公式图像识别系统,如MathReader^[4],它可以处理包含数学公式的文档图像,实现公式定位、识别、分析、输出的全过程。

此外,还有数学公式检索方面的研究。目前大致有两类检索数学公式的方法:一种是首先生成公式的字符串表示,然后运用普通的信息检索方法来检索;另一种是利用内容表示中内在的结构进行检索^[5]。

4 主要研究内容与研究方案

4.1 主要研究内容

4.1.1 多模态信息语义分析理论和方法研究

分析国内外多模态研究方向的主要研究范围,研究范围和发展方向,整合相关领域的研究思想和方法,强调多学科交叉融合,突出原始创新的推动作用。

4.1.2 多模态异构特征的内容理解及语义相关性研究

探讨多模态与语义的关系,挖掘不同模态特征之间的语义相关性,构建面向科技文献内容理解的以实体、关系和事件为核心的结构化语义描述体系,实现其内容表示的语义化。

4.1.3 单模态信息解析与特征提取技术研究

研究基于语义表达的多种模态的解析与提取技术的集成方法,实现科技文献中广泛涉及的图谱、表格、公式、文本等多种模态的有效解析与提取。

4.1.4 多模态语义关联特征的提取与表达研究

研究多模态语义关联特征在内容特征上潜在的统计关系,建立多模态特征的共生矩阵,以生成包含不同类型数据的同构子空间来反映其关联,最终实现多模态语义特征之间关联关系的表达。

4.1.5 基于上下文关联的多模态融合与表达模型构建与实现研究

研究适用于多模态信息的融合机制和多模态协同分析的学习算法,实现基于上下文关联的多模态语义特征融合,建立多模态高维异构数据的特征提取与描述的理论和方法。

4.2 研究方案

研究旨在利用以自然语言处理技术为主的各种语体处理理论与方法,结合基于单结构化文本捕获

构建本体的技术与语义,针对科技文献中广泛涉及的文字、图谱、表格、公式等多模态数据,提出基于上下文更晚的多模态语义特征融合与表达的思路及方法,研究开发多种模态的解析与特征提取技术,挖掘多模态异构特征的内在规律,探索多模态数据间的相容互补性,构建面向科技文献内容理解的以实体、关系和事件为核心的结构化语义描述体系,建立基于语义分析的多模态数据的特征提取与描述的理论和方法,为内容理解及知识服务提供理论与技术支持。

本文的研究方案及整体技术路线如图1所示。

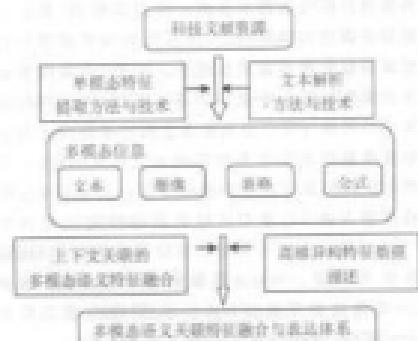


图1 研究方案及整体技术路线

5 成果与展望

我们的研究旨在结合科技文献中的文本、图像、表格及公式等不同模态的语义特征来完善对科技文献的语义理解,因此需要在各个模态的语义特征提取和表示方面分别开展研究,然后将不同模态的语义特征有效地关联起来,构成一个完整的表达体系。目前,课题组在文本语义特征的分析方面取得了一定的进展^[6-11]。我们针对医学领域,基于现有的语言分析技术及语义资源,结合科学技术文献的特征,研究了构建专业领域语义资源的关键技术、理论及方法,建立了理解科技文献文本内容的多语言数据资源库。在这项研究中,我们建立了一个多级别的、全方位的语义标注系统,具体来说,就是在科技文献的词、句、章三个层面上对文档中的多层次语义关系进行标记,并开发相关的标注工具来帮助实现对文本中多层次语义关系的标注。其中,多层次的标注不仅需



要标注主题词，还要对主题词之间或主题词与常用词之间的关系进行标注。在语句层面的标注中，我们分析了语句在章节中的重要性，通过语义角色和谓词类型的分析结果来进行标注。在章节层面的标注中，需要分析并标注句子之间的语义关系，再进行规范化处理和标注工作。该研究可以支持我们对科技文献的文本内容进行深入的语义理解和分析，为下一步分析图像、表格及公式等概念信息语义特征的工作提供基础。

在接下来的工作中，我们会在现有的相关研究的基础上，对科技文献资源中图像、表格及公式等概念信息的语义特征进行分析和提取，同时利用文本语义分析技术来辅助研究，例如结合图像、表格或公式的标题、上下文中的相关描述、脚注等文本信息来完善对非文本概念信息的语义表示，然后将不同概念的语义特征有效地关联起来，构成一个完整的科技文献多模态语义关联特征表达体系，完善对科技文献资源的语义理解与分析。

参考文献

- Cappo Daniel, Becker Robert, Ramsey Craig. Improvement of recognition on a multi-modal pattern discrimination test. *Perception and Motor skills*, 2008, 107(2), 481–484.
- Fidell Sanford. Sensory function in multidimensional signal detection. *Journal of the acoustical society of America*, 1970, 47(4B), 1009–1012.
- 董丽. 基于视觉语义概念的多模态服务的技术综述. *计算机应用研究*, 2008, 25(10), 8–10.
- 董丽等. 多媒体语义概念研究综述. *计算机科学*, 2010, 37(11), 1–17.
- 董丽. 基于概念语义的图像概念识别的研究及评价体系的构建 [博士学位论文]. 华中科技大学, 2008.
- 刘晓波. 多概念语义原像的配准与融合 [硕士学位论文]. 山东大学, 2009.
- A. K. Jain, A. Ross. Multimodal Systems. *Communications of the ACM*, Special Issue in Multimodal Interfaces, 2004, 47(1), 34–40.
- Ivan Arivic, Roger Vilagut, Jean-Philippe Thiran. Automatic extraction of geometric lip features with application to multi-modal speaker identification. *IEEE international conference on multimedia and expo*, Toronto, 2004.
- Kammerer P, Bass M. A human perception model for multi-model feedback in telepresence systems. *IEEE international conference on systems, man and cybernetics*, Japan, 1999.
- 万华丽. 图像纹理特征及其在CBIR中的应用. *计算机辅助设计与图形学学报*, 2002, 15(2), 185–188.
- Gu J, Jing H F, Ngou C W, et al. Distribution-based concept selection for concept-based video retrieval. *Proceedings of ACM International Conference on Multimedia*, Beijing, 2009.
- Hideo Shimura, Toshiaki Takemoto. Multi-Model Method: a design method for building multi-model systems. *Proceedings of the 16th conference on computational linguistics*, 1994.
- 宋生. 多模态语义分析的研究基础与研究方法. *外语学习*, 2007(1), 82–86.
- C. Marin Christodoulou, Evangelos Urtasun, Mathieu Salzmann, Trevor Darrell. Learning to recognize objects from unseen modalities. *Lecture notes in computer science*, 2010, 6321, 677–681.
- Ze-Jing Chang, Cheng-Hsien Wu. Multi-Model emotion recognition from speech and text. *Computational linguistics and Chinese language processing*, 2004.
- 谭海平. 多模态信息融合的一般框架暨设计——基于融合功能与信息流观. *计算机工程与应用*, 2008, 44(20), 27–30.
- Vassilios Stavropoulos, Iasonas Fotiou, Jorma Jääskeläinen. Semantic relationship in multidimensional graphs for automatic image annotation. *Lecture notes in computer science*, 2008, 5054, 180–187.
- 王长庚. 从单模态融合到大规模多模态语义分析、检索和自动标注的研究 [博士学位论文]. 北京: 中国科学院大学, 2009.
- 吴天乐. 相似度、匹配度、相似度、基于图示语义概念的图像标注和检索. *计算机科学*, 2008, 35(8), 223–231.
- Sanderson A, W. M., Worring M., Smeulders A, et al. Content-based image retrieval at the end of the early years. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 2003, 25(12), 1349–1360.
- Audhkun S, Barakat L. Window-based region-based image retrieval using wavelets. In: *The Tenth International Workshop on Database and Expert Systems Applications*. 1999.
- Wang X-J, Si W-Y, Xie G-B, et al. Multi-model similarity propagation and its application for Web image retrieval. *Proceedings of the 12th ACM International Conference on Multi-media*, 2004.
- T. Chen, J. Z. Wang. A region-based fuzzy feature matching approach to content-based image retrieval. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2004, 26(10).
- S. Yu, D. Cai, J. R. Wen, et al. Improving pseudo-relevance feedback in web information retrieval using web page segmentation. *Proceedings of the 12th World Wide Web Conference*, 2003.
- 董丽. 图像语义通路分析. 北京: 清华大学出版社, 2005.
- Mori S, Saito C, Y. Tomono R. Historical review of OCR research and development. *Proceedings of the IEEE*, 1992, 80(7), 1029–1046.
- 董丽. 基于语义通路处理直连与表格字符串抽取的研究 [博士学位论文]. 广州: 中山大学, 2003.
- 王长庚. 相似度、匹配度、平均距离度量函数的实验研究. *计算机科学*, 2008, 35(8), 268–271.
- Wen Wang, David Martens, Lawrence Carin. Extraction of Named Entities from Tables in Game Mutation Literature. *Proceeding of the Workshop in Current Trends in Non-Standard Neural Language Processing*, 2005.
- 董小丽等. 基于构义元多中语义通路的技术研究. *微计算机信息*, 2008, 24(18).
- W. W. Cohen, M. Herter, L. S. Jensen. A flexible learning system for wrapping tables and lists in legal documents. *Proceeding of the 11th International Conference on World Wide*

- Web, 2002.
- 12 陈伟. 基于 PDF X-12 的表格识别和本体的研究 [博士学位论文]. 上海: 上海交通大学, 2005.
- 13 Ying Liu, Praveen Mittal, C. Lee Giles, et al. Automatic extraction of table metadata from digital documents. Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries, 2004.
- 14 Ying Liu, Praveen Mittal, C. Lee Giles. Identifying table boundaries in digital documents via sparse line detection. Proceedings of the 17th ACM conference on information and knowledge management, 2008.
- 15 Barry Vihre, Katharina Kaiser, Silvia Miksch. A method to extract table information from PDF files. Proceedings of the 1st India International Conference on Artificial Intelligence, 2001.
- 16 R. H. Anderson. Syntax-directed recognition of hand-printed two-dimensional mathematics. Interactive Systems for Experimental Applied Mathematics. Academic Press, 1988.
- 17 陈伟. 基于特征字典的表格公式识别研究 [博士学位论文]. 上海: 上海交通大学, 2005.
- 18 陈雷等. 数学公式识别系统. *MathReader*. 计算机学报, 2004, 25(11).
- 19 Michael Kohlhase, Loek A. Sonst. A search engine for mathematical formulae. *Computer Science*, 2008, 1120:2008, 241–253.
- 20 Yan Liu, Yuhua Zhao, Zhilong Su. Research on Automatic Construction of Medical Ontology Based on a Multidimensional Model. *Journal of Computational Information Systems*, 2009, 5 (3), 1713–1720.
- 21 陈雷等. 中医药本体概念语义表示的研究. *现代图书出版技术*, 2009(5):23–28.
- 22 Yan Liu, Yuhua Zhao. Research on Ancient Literature Corpus Creation and Development of Chinese Traditional Medicine. *E&C Express Letters – An Int. J. of Research and Survey*, 2009, 3(10), 1227–1232.
- 23 陈雷等. 基于内部与形式语义的图书各版面数据语义关联方法研究. *图书馆理论与实践*, 2010(11), 101–107.

作者单位：中国科学院文献情报中心，北京，100033

收稿日期：2011年11月21日

Study on the Feature Extraction and Expression System of Multi-Modal Semantic Information for Scientific and Technical Literature

王瑞娟 刘耀

Abstract Scientific and technical literature contains images, tables, formulas, audio and video files besides the common text format, which will help the users to fully understand the knowledge presented in the literature. So the resource of scientific and technical literature can be taken as a kind of multi-modal information. This paper adopts the multi-modal approach to make the semantic presentation of the scientific and technical literature. To be specific, it analyzes the texts, images, tables and formulas in the literature on the semantic level, builds a system to present the semantic multi-modal features in the literature, and optimizes the semantic presentation of the literature with the semantic features and the relations between them.

Keywords Multi-Modal; Scientific and Technical Literature; Semantic Dependency; Semantic Feature

(插图 48 图)

On the Idea of “Universal Library” and Its Practice

王兴强

Abstract A universal library is a library with universal collections. This article, with the clue of time, reviews the different ideas and practices of universal library given by modern librarians, scientists and people in the context of contemporary information technology. It manifests the significance of universal library as a phenomenon of librarianship to the development of library science, libraries and the full progress of society.

Keywords Universal Library; World Encyclopedia; World Brain