DOI:10.11992/tis.201706023
网络出版地址:http://kns.cnki.net/kcms/detail/23.1538.TP.20170831.1058.012.html

基于时空域联合建模的领域知识演化脉络分析

金晨,谢振平,任立园,刘渊
(1.江南大学 数字媒体学院,江苏 无锡 214122; 2.江苏省媒体设计与软件技术重点实验室, 江苏 无锡 214122)

摘 要:同一领域不同知识概念之间存在演化关系,分析演化关系能有效地梳理领域知识的发展脉络,然而网络知识的碎片化、无序性、大规模等特性使得用户很难准确地分析并获取知识之间的这种关系。针对该问题,本文提出一种基于时空域联合建模的领域知识演化脉络分析方法,该方法首先考虑将知识系统以时空域联合知识网络的形式进行表达,随后采用骨架聚类方法提取历年知识网络演化路径,并按知识概念的发展进行演化路径衔接及路径分析。以数字媒体领域知识为例的实验分析表明,该方法能有效提取按年份发展的领域知识演化路径,对于辅助用户进行领域知识的理解与学习,以及个性化推荐具有显著的价值。

关键词:知识演化;演化路径;知识网络;知识系统;时空域联合;骨架聚类;数字媒体知识
中图分类号:TP181 **文献标志码**:A **文章编号**:1673-4785(2017)05-0735-10

中文引用格式:金晨,谢振平,任立园,等.基于时空域联合建模的领域知识演化脉络分析[J]. 智能系统学报, 2017, 12(5): 735-744.
英文引用格式:JIN Chen, XIE Zhenping, REN Liyuan, et al. Evolutionary path mining of domain knowledge by joint modeling in space-time correlation[J]. CAAI transactions on intelligent systems, 2017, 12(5): 735-744.

Evolutionary path mining of domain knowledge by joint modeling in space-time domain

JIN Chen, XIE Zhenping, REN Liyuan, LIU Yuan
(1. School of Digital Media, Jiangnan University, Wuxi 214122, China; 2. Jiangsu Key Laboratory of Media Design and Software Technology, Wuxi 214122, China)

Abstract: In special technology fields, there might be evolutionary relationships between various knowledge concepts, and these evolutionary relationship can be used to depict the developmental venation of the corresponding technology field. However, the characteristics of fragmentation, disorder, and large scale in domain knowledge systems make it difficult for users to accurately identify these knowledge relationships. To address this problem, in this paper, we propose an evolutionary path mining method based on skeleton clustering and the joint modeling of domain knowledge with respect to the space-time correlation. In this method, first we express the knowledge system as a knowledge network with joint space-time correlations, then we adopt the skeleton clustering method to extract the evolutionary path of the knowledge network. In addition, we analyze the connection between the evolutionary paths based on the development of the knowledge concept. An experimental analysis of the digital media domain shows that the proposed method can effectively extract the evolutionary path of domain knowledge, which has significant value for knowledge learning and personalized recommendation.

Keywords: knowledge evolution; evolution path; knowledge network; knowledge systems; space-time domain combination; skeleton clustering; digital media knowledge

领域知识是一个随时间扩展的体系,那些重要的理论定律不断被引用,新颖的思想和观点不断产生,新旧知识之间始终保持动态的知识增长。在这个过程中,学科领域逐步细化,知识框架也将发生改变,但科学知识始终保持一个整体,这其中体现了知识的演化。知识之间存在一种建构的关系[1-2],任何新知识不可能凭空产生,必然基于现有

收稿日期:2017-06-07. 网络出版日期:2017-08-31.
基金项目:江苏省自然科学基金项目(BK20130161);国家自然科学基金项目(61572236);国家科技支撑计划项目(2015BAH54F01).
通信作者:谢振平.E-mail :xiezhenping@ hotmail.com.

的知识经验,可以说,新知识是现有知识的演化和创新。知识演化体现了知识之间传承和发展的关系,提取知识间的这种演化关系具有十分重要的意义。一方面,科学知识的增长,知识数量的膨胀,给用户准确有效地获取所需知识带来了巨大的挑战,知识演化分析[3-4]能够帮助用户有效地梳理复杂的知识关系,获悉领域研究热点及发展动向。另一方面,目前网络个性化知识服务已相当成熟,然而能体现时空上演化的知识服务却少有研究,设计一种自动提取领域知识演化关系的方法能够为用户生成具有时间上连续的演化知识序列,对个性化知识服务的改进具有一定的价值。

一对演化关系由两个实体概念组成,演化路径则是演化关系的连续序列,包括演化起点、演化终点及演化中间点。例如,1990 年数字媒体领域的一条演化路径“电视广播—电视教育—电教媒体—远距离教育—电化教学—计算机技术”,“电视广播”表示知识演化起点,“计算机技术”表示知识演化终点。可以看出,1990 年数字媒体领域热点话题围绕传统媒体,并将传统媒体广泛应用于教学,整体的演化趋势从传统媒体趋向计算机技术。

为了较好地挖掘知识间的这种演化关系,研究者通常采用知识网络来表示不同形式知识单元之间的联系,并设计自动化的知识关系抽取模型,进而获取有效的知识信息。根据知识单元的不同表现形式,常见的知识网络包括引证网络[5-6]、合作网络[7-9]、共词网络[10-12]等。此外,可视化文献分析软件也广泛应用在研究学科领域的发展趋势与动向。例如,马费成等[4]在引文网络的基础上采用网络分析软件 Citespace,以生物医学领域为例进行了领域主题聚类、关键路径提取、核心文献分析等研究,实验分析结果为学科发展提供了较好的理解。但 Citespace 只有在文献引文网络分析中有较好表现,并且对文献格式等有一定要求。

本文提出一种基于时空域联合建模的领域知识演化脉络分析方法,在传统知识网络分析技术的基础上引入骨架聚类技术[13-15],针对网络结构中的最短路径进行骨架聚类分析,骨架聚类效果最优的最短路径视为该知识网络的演化路径,并根据时空上连续的网络结构进行演化路径衔接,形成连续年份的知识演化脉络。考虑到近年来数字媒体领域发展之快,影响面之广,本文实验以 CNKI 在数字媒体领域发表的期刊文献作为数据来源,按年份发展逐年构建知识网络并采用骨架聚类提取演化路径,进而对数字媒体领域的发展历程进行研究分析。

1 模型框架

本文提出的基于时空域联合建模的领域知识演化脉络分析方法着重考虑两个问题:如何表示知识概念之间的演化关系;如何从复杂的演化关系中提取演化路径。针对上述问题,本文设计的模型框架由两部分组成:1)采用知识网络来表示知识概念之间的演化关系,网络节点表示知识概念,网络边表示连接两个知识概念存在知识演化关系;2)采用“局部聚合,整体关联”的思想进行网络骨架聚类分析。“局部聚合”指骨架节点能够作为邻近节点的聚类中心,形成局部稠密子图;“整体关联”指各骨架节点在网络图中是连通的,并且整条骨架上的各节点聚类系数之和最小,则该骨架认为是网络图的一条最优知识脉络。

1.1 知识网络模块

在知识图谱领域,知识网络是研究知识发展的重要工具[16-20]。知识网络由节点和边组成,节点表示知识实体单元,边表示实体单元之间的知识关联。按实体单元不同,节点可以是论文、专利、书籍、关键词等;按知识关联不同,边可以是引证关系、共现关系、合作关系等。本文采用的知识网络是一种改进的共词网络,以领域关键词为节点,以演化关系权重作为边。相对于引证网络,共词网络能够更加直观有效地体现实体概念在网络结构上的演化脉络。

本文构建的知识网络是一种加权无向网络,目前对该类网络的研究主要包括两方面:1)网络节点在网络图中的重要性评价,评价指标主要有节点词频,节点度(无向图中出度入度相等,统称为节点度),中介中心度等;2)基于节点连线的网络路径分析,包括最短路径、关键路径、平均路径长度等。网络节点分析常用于获取网络主题分布,而网络路径分析则用于预测领域知识发展方向、发现研究热点等。

本文构建知识网络的过程主要包括以下 3 个步骤。1)获取网络节点:自动抽取数字媒体文献的关键词,进行关键词筛选和统计,获取具有代表性的领域关键词作为网络节点。2)提取节点关系:统计关键词在文档中的共现频率,基于共现频率计算关键词对的演化权重,以演化权重作为节点关系。3)根据获取的网络节点以及节点关系逐年构建知识网络,形成相邻年份网络结构关联的时空域联合知识网络。

1.1.1 知识术语抽取

随着自然语言处理领域的快速发展,领域术语

抽取技术已取得显著的成绩[21-23],并且涌现出了一批成熟的术语抽取系统[24],其中最著名的是中科院汉语分词系统 NLPIR。本节主要介绍如何使用 NLPIR 工具进行文档术语抽取及统计工作。术语抽取的具体流程如图 1 所示:首先搜集指定领域相关的文本语料,然后调用 NLPIR 系统的 KeyExtract_GetKeyWords 方法进行单篇文档术语抽取,并将获取的关键词以键值对的形式存入 HashMap 中,Key 表示关键词,Value 表示关键词出现的次数,从而实现关键词次数统计。统计过程首先提取文档 t_i 的关键词集合 K,如果关键词首次出现则存入 HashMap,并将 Key 值设为 1;如果关键词在 HashMap 中已存在,则将关键词对应的 Value 值累加 1;直到统计完该年所有文档中的关键词。最终按 Value 值对关键词进行降序排序,获取频次较高的前 N 个关键词作为该领域术语集合。

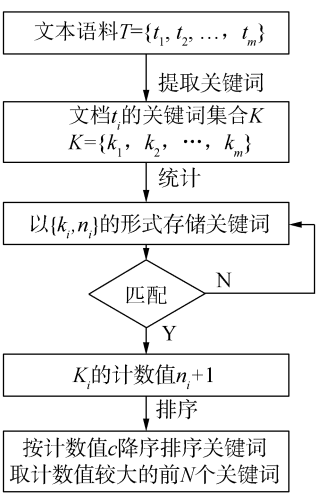


图 1　术语抽取流程

Fig.1　Term extraction process

1.1.2　时空域联合知识网络构建

时空域联合知识网络旨在通过构建空间上连续层面的知识网络来表现知识在时间上的连续演化关系。联合知识网络构建过程可分为两步:首先逐年创建知识网络,然后根据相邻年份重复的网络节点自动形成空间上连续的网络结构。

构建知识网络的核心工作在于提取网络节点之间的关系权重,本文将知识概念之间的演化关系视为网络边权重。演化关系可认为是实体关系[25-26]的一种,这种关系是由知识概念在文档中的语义距离和共现频率决定的,距离越小频率越高则演化强度越大。本文针对演化关系给出如下定义:对于给定的文档 T,文档知识概念实体序列表示为 $S=\{s_1,s_2,s_3,\cdots\}$,两个实体概念 s_i 和 s_j 在序列 S 中的语义距离计算如(1)式:

$$\mathrm{dis}(s_is_j)=\sum_{s_is_j\in S}|j-i|/n \tag{1}$$

式中:i 和 j 表示知识概念在序列中出现的位置,n 表示知识对在序列中出现的次数。语义距离越小表明实体概念 s_i 和 s_j 之间的演化强度越大。在实验过程中本文设定语义距离阈值 ε,当知识对在共现序列中位置差小于给定阈值时认为两个知识存在演化关系,否则认为没有关系。如果节点对 s_i 和 s_j 之间存在演化关系,则节点对在知识网络中必然存在一条关联路径。演化距离的定义如(2)式所示:

$$\mathrm{evo}(s_is_j)=\exp[(\sum_{s_is_j\in\bar{T}}\mathrm{dis}(s_is_j))^2/(2m^2\delta^2)]/m^2 \tag{2}$$

式中:$\bar{T}$ 表示实体概念 s_i 和 s_j 共现的文档集合,m 表示共现文档数,$\mathrm{evo}(s_is_j)$ 值越小表明从概念 s_i 演化至 s_j 越容易。

提取演化关系具体流程如图 2 所示:将提取的关键词导入 NLPIR 分词工具,作为用户自定义词典,使分词工具能够实现粒度较大的分词。对单篇文档进行分词,筛选分词结果中的用户自定义词,初步得到文档关键词序列 S,合并序列中相邻重复出现的关键词,得到相邻关键词不重复的新序列 S'。在此基础上,进一步统计序列 S' 中两两关键词对的关系。例如,s_i 和 s_j 是 S' 中的两个关键词,按 $\{s_{ij},d_{ij},n_{ij}\}$ 的格式进行存储,s_{ij} 表示关系对,d_{ij} 表示关系对在文档中的语义距离,n_{ij} 表示关系对出现的次数。进一步,统计所有文档中出现的关系对,对于重复出现的关系对,d_{ij} 值累加,n_{ij} 值累加。最终得到每一对关系的平均语义距离及出现的次数。根据式(2)计算每一对关系的演化距离,作为知识网络边的权重。

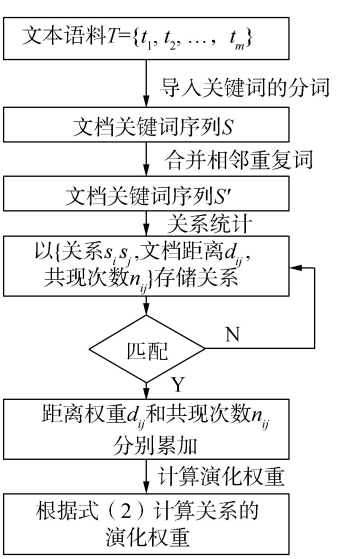


图 2　演化关系抽取流程

Fig.2　Evolutionary relationship extraction process

按照上述方法,我们以关键词作为知识网络节点,以演化距离作为知识网络边的权重,逐年构建知识网络,并根据相邻年份重复节点自动形成时空域联合知识网络。图 3 为连续 3 年的时空域联合知识网络结构,圆点表示知识概念,圆点半径越大表明该知识在网络中的重要性越强;节点间的连线表示演化关系,权重越小则节点距离越近,表明两个知识之间演化强度越大。虚线表示相邻年份知识网络之间存在重叠的知识概念,通过这些重复的知识概念来建立连续年份之间的知识演化关系。

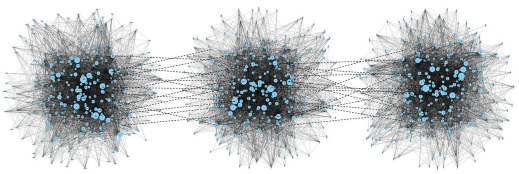


图 3 时空域联合知识网络结构

Fig. 3 Space-time domain joint knowledge network structure

1.2 骨架聚类分析

基于给定知识网络,如何从该知识网络中提取理想的演化路径是本节主要讨论的问题。一条理想的演化路径可看作若干条网络结构骨架的连接,骨架是用于支撑网络结构或轮廓的支架,一条理想的骨架应具有中心性、连通性等特性。本文提出一种骨架聚类的方法提取知识网络中的演化路径。骨架聚类方法的整体思想是“局部聚合,整体关联”。“局部聚合”的目的是将知识网络进行聚类划分,每一个类可认为是一个知识主题,骨架节点应尽可能地分布在不同的知识主题中,并且该骨架节点能够作为主题的一个聚类中心,使得主题聚类效果最优。“整体关联”的目的是将所有的骨架节点进行连接,整合成一条完整的骨架,理论上整条骨架应尽可能全面地覆盖知识网络,并且使得骨架节点的主题聚类效果之和最优。

图 4 为知识网络演化路径示意图,圆点表示知识节点,连线表示演化关系。图 4 展示了 3 个连续年份的知识网络 i, j, k, 其中 A—B—C—D 和 O—P—Q—R 表示两条完整的演化路径。每个知识网络中提取两条骨架路径,如年份 i 知识网络中实线 A—B 和 O—P 所示。相邻知识网络间的虚线连接表示上一年演化路径的终点与下一年演化路径的起点为同一个知识,如图 4 中 B—B、C—C 等。

两个知识之间存在许多种可能的演化路径,其中最短路径认为是最优的演化路径。最短路径通常用于计算网络图中一个节点到其他节点的最小路径代价,在知识演化网络图中,最短路径可认为

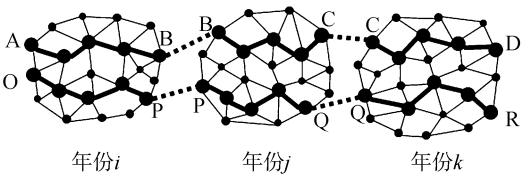


图 4 联合时空域知识网络演化路径

Fig.4 Evolutionary paths in joint space-time knowledge network

是知识间演化的必然趋势。不同的演化起点和演化终点对应不同的最短路径,因此需进一步通过骨架聚类的方法来分析不同最短路径对整个网络结构的演化重要性。

骨架由骨架节点组成,一个理想的骨架节点应具有较好的中心性,相邻节点簇以骨架节点为聚类中心构成一个知识主题。节点聚类系数具体计算公式如(3)式所示:

$$\mathrm{CH}(s)=\sum_{c_i\in C}\mathrm{djs}(c_i,s)/C_n \tag{3}$$

式中:$\mathrm{CH}(s)$ 表示骨架节点 s 的聚类系数;C 表示 s 所对应的知识主题;C_n 表示主题 C 包含的节点数;$\mathrm{dis}(*,*)$ 表示节点间的最短路径。如果 $\mathrm{CH}(s)$ 聚类系数值最小,则节点 s 被认为主题聚类中心,即骨架节点。进一步计算整条骨架的主题聚类系数,根据骨架节点聚类系数平均值来选取最优的骨架。具体计算公式如(4)式所示:

$$\mathrm{SH}(S)=\sum_{s_i\in S}\mathrm{CH}(s_i)/S_n \tag{4}$$

式中 S_n 表示骨架 S 包含的骨架节点数。如果骨架 S 的聚类系数平均值 $\mathrm{SH}(S)$ 最小,则认为该最短路径对应的骨架是一条理想演化路径。

1.3 演化路径抽取

进一步对演化路径抽取流程进行具体描述,算法详细描述了 1990—2016 年间的演化路径抽取流程。算法中步骤 2)~4),首先根据每一年的文本语料生成该年的知识网络 G,然后获取相邻年份知识交集 $\bar{G}$ 作为相邻年份演化路径的衔接知识,使得上一年演化路径的终点为下一年演化路径的起点,如算法中 5)~10)从 1990 年开始逐年提取演化路径。演化第一年以当前年份知识网络 G 中任意节点为演化起点,以相邻年份知识网络交集 $\bar{G}$ 中的节点为演化终点,进而提取所有可能的最短路径集合 S'作为该年候选的演化路径。从演化第二年开始,上一年所提取的 top-k 条最优演化路径的演化终点 VT 作为下一年演化起点。演化最后一年,由于不存在与下一年知识网络知识交集,故演化终点即为该年知识网络中的任意节点。获取候选演化路径之后

通过骨架聚类来分析最优的演化路径,如算法中11)~13)。对于 S' 中的任何一条最短路径,以该路径上的节点作为网络的聚类中心,路径包含的节点数作为聚类数,根据式(3)、(4)计算每一条最短路径的聚类系数 C_v,然后根据 C_v 值对 S' 中的所有路径进行排序,选择聚类系数最小的 k(实验中 $k=10$)条路径作为该年演化路径。完整的演化路径需将相邻年份的演化路径进行连接,形成一条连续的,覆盖该领域所有年份的演化脉络。步骤6)、8)、10)已经保证了相邻年份演化路径的演化终点和演化起点相同,如算法中步骤15)所示,只需将相邻年份的演化路径根据相同节点进行连接即可形成连续年份的知识演化脉络。

连续年份知识演化路径提取算法详细描述如下:

1) for literature $l_i \in L \mid 1990 \leqslant i \leqslant 2016$
2) $G_i = \text{GetKnowledgeGraph}(l_i)$
3) if $1990 \leqslant i \leqslant 2015$
4) $\bar{G}_i = G_i \cap G_{i+1}$;
5) if $i = 1990$
6) $S' = \{\text{djs}(V_x, V_y) \mid V_x \in G_i, V_y \in \bar{G}_i\}$;
7) if $1991 \leqslant i \leqslant 2015$
8) $S' = \{\text{djs}(V_x, V_y) \mid V_x \in \text{VT}_{i-1}, V_y \in \bar{G}_i\}$;
9) if $i = 2016$
10) $S' = \{\text{djs}(V_x, V_y) \mid V_x \in \text{VT}_{i-1}, V_y \in G_i\}$;
11) Let $C_v = \text{SH}(s) \mid s \in S'$;
12) Sort S' by C_v;
13) get top-k items of S';
14) end
15) $S = \text{Link}(S_i, S_{i+1})$;

2　实验研究

2.1　实验数据

考虑领域的发展现状及研究热点,本文以数字媒体领域作为实验研究对象。搜集和处理数据的步骤如下:首先数据来源选择 CNKI 中国知网,分别以“媒体”和“数字媒体”作为检索输入,以“关键词”和“摘要”作为检索项,检索 1990~2016 年期间发表的期刊文章。再按年份下载 CAJ 格式论文,并以“1990-01”的格式保存在相应年份的文件夹下。如果某一年发表的文章数量较多,则根据文章的下载量和被引量择优选择 300~500 篇。然后,采用 CAJViewer 自带的“另存为”功能将 CAJ 格式转化成 TXT 格式,便于 Java 程序进行处理。由于早期发表的部分文章均采用图片格式保存,导致格式转换出现乱码,需通过程序进行筛选,去除无效数据。最终,获取 1990~2016 年间数字媒体领域发表的部分具有代表性的学术文章,共计5 420篇,其中 1990 年文章数量最少仅有 11 篇,2016 年最多 514 篇。这些文章基本能够代表数字媒体领域的发展动态及研究成果。

2.2　实验分析

实验部分主要基于知识网络展开分析,首先整合数字媒体领域历年的期刊文献,构建一个整体的领域知识网络,根据网络的词频、节点度来整体分析数字媒体领域的核心知识和研究热点;然后,针对历年数字媒体知识网络进行知识演化分析,并提取演化路径来展示数字媒体领域的发展历程。

2.2.1　网络节点分析

首先采用 NLPIR 分词工具进行数字媒体领域关键词提取,实验从每一篇文档中择优提取 10 个关键词,并统计5 420篇文档中所有关键词及其相应的词频,最终筛选获取词频最高的 953 个关键词作为数字媒体领域的知识术语。表 1 为出现频数最高的 Top10 关键词,表中展示的“数字媒体”、“媒体”、“传统媒体”等关键词都是数字媒体领域非常有代表性的知识术语,这在一定程度上展示了关键词提取的有效性。

表 1　数字媒体领域整合词频前十关键词列表

Table1　Most frequent ten key words in digital media knowledge domain

序号	关键词	出现频数
1	数字媒体	836
2	媒体	708
3	传统媒体	667
4	信息	542
5	新闻	409
6	传播	356
7	数字电视	336
8	网络	325
9	广告	313
10	电视	298

进一步整合数字媒体 1990～2016 年所有的期刊文献,构建一个涵盖 27 年知识发展的整体知识网络并分析网络节点度。以获取的 953 个知识术语作为数字媒体知识库,从 5 420 篇期刊中提取这些关键词在文档中的序列,并根据式(2)计算序列中关键词对的演化距离,以关键词作为网络节点,演化距离作为网络边构建知识网络。

节点度表示知识网络中节点拥有的关系数量,关系数量越大表明该关键词的重要性越强。图 5 为 953 个关键词所拥有的 116 274 对知识关系,关键词度数服从长尾分布,表明知识网络内部拥有小部分节点度较大的核心知识,大部分节点度较小的边缘知识,核心知识在网络中起到“桥梁”的作用,为边缘知识建立知识关联。

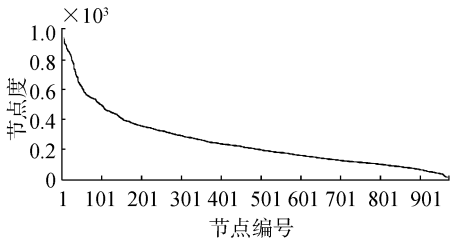


图 5 知识网络节点度分布曲线

Fig.5 Knowledge network node degree distribution

表 2 为部分年份知识网络节点度前 20 关键词列表,分析列表数据可知,1990 年主要以“电视”、“广播”、“电化教学”等传统媒体关键词为主,2000 年以后“网络”、“互联网”、“手机”等关键词开始涌现,而具有领域广泛代表性的“媒体”、“电视”等关键词在各年份都高频出现,这在一定程度上体现了数字媒体领域伴随年份的演化特征。

表 2 部分年份知识网络节点度前 20 关键词列表

Table 2 First 20 key works list of knowledge networks in some years

序号	1990 年	1995 年	2000 年	2005 年	2010 年	2015 年
1	电视	媒体	网络	媒体	媒体	媒体
2	广播	多媒体	电视	网络	网络	网络
3	媒体	计算机	媒体	数字	电视	互联网
4	计算机	电视	数字	电视	数字	数字
5	视听教材	图像	图像	数字化	互联网	视频
6	图像	网络	计算机	广播	数字化	电视
7	电大	多媒体技术	广播	多媒体	视频	数字化
8	视听教学	数字	数字化	视频	手机	数字技术
9	广播电视	软件	多媒体	计算机	多媒体	传统媒体
10	远距离教育	广播	软件	互联网	广告	手机
11	软件	视频	视频	图像	广播	软件
12	电影	电子	互联网	软件	计算机	网站
13	电视教育	光盘	上网	广告	传媒	传媒
14	电视教学	程序	电子	数字电视	数字技术	广告
15	电化教学	动画	广播电视	数字媒体	数字媒体	数字媒体
16	程序	多媒体信息	数字技术	电子	软件	媒体时代
17	动画	数据库	电脑	广播电视	图像	多媒体
18	数字媒体	多媒体系统	光盘	数字技术	网站	计算机
19	磁带	电脑	网站	电影	传统媒体	广播
20	录像机	电影	数据库	网站	电子	微博

2.2.2　演化脉络分析

实验给出了数字媒体领域 1990—2016 年 10 条最优的演化路径。首先以 1990 年作为知识演化起始年份,从中提取了 10 条聚类效果最优的演化路径,并以该年的 10 个演化终点知识作为下一年的知识演化起点,以此获取 10 条连续的涵盖数字媒体领域 27 年的知识演化脉络。需要指出的是,由于知识网络是一个无向图,某一年的演化路径无法体现演化的方向性,演化方向主要体现在连续年份上知识的发展。例如,某一年存在两条演化路径 *A*—*B*—*C* 和 *C*—*B*—*A*, 演化的下一年将分别以节点 *C* 和节点 *A* 作为演化起点,因此在连续年份的知识演化上这两条路径的知识演化方向是完全不同的。

表 3 给出了实验提取的 10 条最优演化路径,由于路径包含大量演化节点,表中仅展示了每一年演化路径的演化起点和演化终点。例如,1990 年演化起点包括“报纸媒体”、“大众传媒”、“广告”、“数据库”、“软件”,演化终点包括“数据库”、“广告”、“大众传媒”、“软件”、“电视信号”,由于中间节点的不同,这些起始节点总共组成了 10 条演化路径。1991 年演化起点包含 5 个节点,演化终点包含 4 个节点,其中“大众传播”和“电视信号”均演化为“远距离教学”,总路径数为 5 条。直到 2010 年所有的演化路径合并为一条,演化终点为“现代传媒”。进入“现代传媒”时代之后,数字化技术开始盛行,包括“数字广播”、“数字影音”、“数字游戏”等,整个过程体现了从“传统媒体”至“现代传媒”的一条演化脉络。10 条演化路径演化趋势基本一致,表明了知识演化脉络的可靠性。

表 3　数字媒体领域 10 条最优演化路径

Table 3　Ten key evolutionary paths in digital media knowledge domain

年份	演化起点	演化终点	路径数
1990	报纸媒体、大众传媒、广告、数据库、软件	数据库、广告、大众传播、软件、电视信号	10
1991	数据库、广告、大众传播、软件、电视信号	无线电广播、动画片、远距离教学、电视广播	5
1992	无线电广播、动画片、远距离教学、电视广播	信息服务、新闻传播	4
1993	信息服务、新闻传播	报纸广告、人机交互	2
1994	报纸广告、人机交互	微软、卫星直播	2
1995	微软、卫星直播	视听教学、电子技术	2
1996	视听教学、电子技术	数字化处理、电视信号	2
1997	数字化处理、电视信号	杂志广告、软件工程	2
1998	杂志广告、软件工程	微处理芯片、计算机图形	2
1999	微处理芯片、计算机图形	媒体资源、电脑	2
2000	媒体资源、电脑	通信技术、网络广告	2
2001	通信技术、网络广告	互联网用户、卫星数字电视	2
2002	互联网用户、卫星数字电视	智能卡、媒体广告	2
2003	智能卡、媒体广告	网络电视、电影电视	2
2004	网络电视、电影电视	传统媒介、媒体业务	2
2005	传统媒介、媒体业务	客户端软件、楼宇电视	2
2006	客户端软件、楼宇电视	电视新闻媒体、信息社会	2
2007	电视新闻媒体、信息社会	电视广告、新闻媒体	2
2008	电视广告、新闻媒体	音频广播、移动电视	2
2009	音频广播、移动电视	游戏产业、大数据	2
2010	游戏产业、大数据	现代传媒	2
2011	现代传媒	音乐产业	1
2012	音乐产业	网络运营商	1
2013	网络运营商	网络资源	1
2014	网络资源	家庭影院	1
2015	家庭影院	互动体验	1
2016	互动体验	数字期刊	1

为了更进一步地分析演化的细节,表 4 展示了 1990—2016 年一条具有代表性的完整的演化路径。分析表 4 可知,数字媒体领域发展日新月异,新的事物新的概念不断涌现。1990—1993 年数字媒体领域主要以传统媒体为主,包括电视、广播、报纸等,并且将传统媒体广泛应用于教育事业,出现了关键词“电化教学”、“教学媒体”、“远距离教育”等;1994 年演化终点出现了关键词“微软”,这是推动数字媒体领域发展最重要的企业,这也标志着计算机技术与数字媒体的融合。1995—2000 年,计算机技术得到更广泛的应用,包括数字化处理、图像处理等,同时“笔记本电脑”、“互联网用户”等关键词的出现标志着互联网技术也越来越成熟;2000 年开始数字媒体正式进入“数字时代”,“数字广播”、“数字电视”、“数字音乐”、“数字信息”等大量出现在人们的视野中;2010 年开始,数字媒体领域呈现了多方趋势。“游戏产业”、“网络游戏”等关键词体现了游戏行业的繁荣发展;“虚拟世界”、“互动体验” “家庭影院”、“智能移动终端”等关键词体现了数字媒体的发展将越来越贴近人们的生活,标志着生活智能化和艺术平民化时代的到来。路径整体演化趋势与 10 条路径综合演化的整体趋势基本一致,进一步表明演化脉络的可靠性。

表 4 一条完整的代表性数字媒体知识演化路径

Table 4 Typical complete evolutionary path in digital media knowledge domain

年份	演化路径
1990	报纸媒体-广告-媒体-电教媒体-磁带-现代化教学媒体-媒体传播-图像-电视信号-软件-数据库
1991	数据库-图像-电视信号-录像机-电影-教学媒体-动画制作-电视-无线电广播
1992	无线电广播-大众传播媒体-远距离教育-电视媒体-印刷教材-广播-报纸-信息网络-信息服务
1993	信息服务-广告-音像出版社-电化教学-电子-大众传播媒介-传媒-电视新闻-报纸广告
1994	报纸广告-报纸-广播电视-录像机-录音机-游戏-微软
1995	微软-电子出版物-多媒体应用-计算机应用-多媒体系统-录音机-视听教学
1996	视听教学-视听教育-超媒体-电子-数字信号-信息技术-通信技术-数字化处理
1997	数字化处理-通信技术-电视传播-图像处理-平面媒体-媒体-杂志广告
1998	杂志广告-报纸广告-彩电-多媒体-信息内容-光盘-记录媒体-计算机-网络-微处理芯片
1999	微处理芯片-电子计算机-计算机-广播电视教育-电视-电视会议-电脑
2000	电脑-数字媒体-互联网-信息产业-网络-通信技术
2001	通信技术-数字信息-电子信息-信息技术-远程教育-笔记本电脑-移动电话-互联网用户
2002	互联网用户-多媒体数据-计算机-数字技术-数字化时代-数字时代-媒体广告
2003	媒体广告-数字时代-广播电视媒体-广播电视-因特网-音视频-广播-电影电视
2004	电影电视-广播-网络艺术-电子-宽带网络-网络媒体-媒体环境-互动媒体-传统艺术-传统媒介
2005	传统媒介-电视行业-有线电视-终端设备-计算机-软件开发-游戏-信息咨询-电子邮件-客户端软件
2006	客户端软件-软件-电视业-电影-网站-通信技术-广播-电视-电视新闻媒体
2007	电视新闻媒体-电视-广播-电子信息-移动电话-计算机-通信技术-新闻媒体
2008	新闻媒体-远程教育-电子商务-电子-数码-数字音频-音频广播
2009	音频广播-移动电视-网络游戏-网络广告-门户网站-视频广告-电视-游戏产业
2010	游戏产业-音乐产业-网络游戏-数据库-现代传媒
2011	现代传媒-网络-媒体-传统媒体-网络媒体-数字广告-文化产业-音乐产业
2012	音乐产业-信息网络传播-移动互联网-国际互联网-数字技术-网络运营商
2013	网络运营商-信息服务-博客-微博-媒体-网络资源
2014	网络资源-数字音频-传统电视节目-传统媒体-媒体-网络-通信技术-计算机-家庭影院
2015	家庭影院-电脑-虚拟世界-互联网-媒体-传统媒体-大众媒体-媒体时代-移动智能终端-互动体验
2016	互动体验-图像-印刷技术-数字媒体-媒体-广播-报纸-数字期刊

3　结束语

本文提出了一种基于时空域联合建模的领域知识演化脉络分析方法,并对 1990—2016 年间 5 420篇数字媒体领域期刊文献进行了研究分析。首先,构建了一个数字媒体领域的整体知识网络,从节点词频、节点度等分析领域的核心知识及知识结构。进而,构建了一个时空域联合知识网络,并根据骨架聚类算法提取相应年份的网络骨架,连接形成连续年份的演化脉络,并根据获取的演化脉络对数字媒体领域的发展进行深入分析。研究表明,数字媒体领域的发展可以概括为,从 20 世纪 90 年代初期的“电视”、“广播”、“报纸”等传统媒体到 2000 年正式进入现代传媒,各种传统媒体都向数字化转型,并由此又衍生出多个重点领域,包括“数字游戏”、“数字动漫”、“数字影音”、“数字出版”、“数字学习”等。

综合分析可知,本文方法是领域知识建模分析的一种新颖手段,不仅具备良好的技术参考价值,而且对个性化知识推荐与学习具有显著实用价值。

参考文献:

[1]BODNER G M. Constructivism: a theory of knowledge [J]. Journal of chemical education, 1985, 63(10): 873-878.

[2]MCCOURT D M. Practice theory and relationalism as the new constructivism[J]. International studies quarterly, 2016, 60(3): 475-485.

[3]高俊平, 张晖, 赵旭剑, 等. 面向维基百科的领域知识演化关系抽[J]. 计算机学报, 2016, 39(10): 2088-2101.
GAO Junping, CHEN Hui, ZHAO Xujian. Evolutionary relation extraction for domain knowledge in Wikipedia[J]. Chinese journal of computers, 2016, 39(10): 2088-2101.

[4]马费成, 陈潇俊, 刘向. 基于科学知识图谱分析的知识演化研究—以生物医学为例[J]. 情报科学, 2012, 30(1):1-7.
MA Feicheng, CHEN Xiaojun, LIU Xiang. Study on the knowledge evolution based on mapping scientific domain—a case of the biomedicine field [J]. Information science, 2012, 30(1): 1-7.

[5]刘向, 马费成. 科学知识网络的演化与动力——基于科学引证网络的分析[J]. 管理科学学报, 2012, 15(1): 87-94.
LIU Xiang, MA Feicheng. Evolution and dynamics of scientific knowledge network: Based on the study of scientific citation network[J]. Journal of management sciences in China, 2012, 15(1): 87-94.

[6]许琦, 冯羽静. 一种基于专利引证网络的知识流提取方法:随机行走中的聚合效应[J]. 情报理论与实践, 2015, 38(12): 98-103.
XU Qi, FENG Yujing. A method of knowledge flow extraction based on patent citation network: aggregation effect in random walk[J]. Information theory and practice, 2015, 38(12): 98-103.

[7]黄玮强, 庄新田, 姚爽. 产业集群广义创新合作网络演化[J]. 东北大学学报自然科学版, 2012, 33(4): 592-596.
HUANG Weiqiang, ZHUANG Xintian, YAO Shuang. Evolution of generalized innovation network in industry clusters [J]. Journal of northeastern university: natural science, 2012, 33(4): 592-596.

[8]关世杰, 赵海. 互联网技术领域科研合作网络分析[J]. 东北大学学报:自然科学版, 2013, 34(4): 509-511.
GUAN Shijie, ZHAO Hai. Analysis of scientific research cooperation network in internet technology [J]. Journal of northeastern university: natural science, 2013, 34(4): 509-511.

[9]陆浩, 王飞跃, 刘德荣, 等. 基于科研知识图谱的近年国内外自动化学科发展综述[J]. 自动化学报, 2014, 40(5): 994-1015.
LU Hao, WANG Feiyue, LIU Derong, et al. A summary of development of automation discipline at home and abroad in recent years based on scientific research knowledge [J]. Acta automatica sinica, 2014, 40(5): 994-1015.

[10]张斌. 共词网络的结构与演化:概念与理论进展[J]. 情报杂志, 2014, 33(7):103-109.
ZHANG Bin. The structure and evolution of co-word networks: concept and theoretical progress[J]. Journal of intelligence, 2014, 33(7): 103-109.

[11]张豪锋, 李海龙. 我国教育技术学研究前沿探讨——基于核心期刊关键词的共词网络与聚类分析[J]. 电化教育研究, 2011(10): 26-29.
ZHANG Haofeng, LI Hailong. Frontier study of educational technology research in China-Co-word network and cluster analysis based on keywords in core journals [J]. Eeducation research, 2011(10): 26-29.

[12]吴建南, 郑烨, 张攀,等. 基于共词网络分析的国内创新驱动研究热点与趋势[J]. 中国科技论坛, 2014(6): 17-23.
WU Jiannan, ZHENG Ye, ZHANG Pan, et al. Research focus and trend of domestic innovation driven research based on co-word network analysis[J]. China science and technology forum, 2014(6): 17-23.

[13]宗瑜, 李明楚, 江贺. 近似骨架导向的归约聚类算法[J]. 电子与信息学报, 2009, 31(12): 2953-2957.
ZONG Yu, LI Mingchu, JIANG He. Approximation of skeleton-oriented reduction clustering algorithm [J]. Journal of electronics and information technology, 2009,

31(12): 2953-2957.

[14]金萍, 宗瑜, 屈世超,等. 面向不确定数据的近似骨架启发式聚类算法[J]. 南京大学学报自然科学, 2015, 51(1): 197-205.
JIN Ping, ZONG Yu, QU Shichao, et al. Approximate skeleton heuristic clustering algorithm for uncertain data[J]. Journal of Nanjing university: natural sciences, 2015, 51(1): 197-205.

[15]LU Z, SUN X, WEN Y, et al. Skeleton construction in mobile social networks: algorithms and applications[C]// Eleventh IEEE International Conference on Sensing, Communication, and Networking. Singapore, Singapore, 2014: 477-485.

[16]刘向, 马费成, 王晓光. 知识网络的结构及过程模型[J]. 系统工程理论与实践, 2013, 33(7): 1836-1844.
LIU Xiang, MA Feicheng, WANG Xiaoguang. The structure and process model of knowledge network[J]. System engineering theory and practice, 2013, 33(7): 1836-1844.

[17]马费成, 刘向. 科学知识网络的演化模型[J]. 系统工程理论与实践, 2013, 33(2): 437-443.
MA Feicheng, LIU Xiang. Evolution model of scientific knowledge network[J]. System engineering theory and practice, 2013, 33(2): 437-443.

[18]刘向,马费成,陈潇俊,等. 知识网络的结构与演化——概念与理论进展[J]. 情报科学,2011(06): 801-809.
LIU Xiang, MA Feicheng, CHEN Xiaojun, et al. The structure and evolution of knowledge network-concept and theory progress[J]. Information science, 2011(06): 801-809.

[19]PFEIFFER J J, MORENO S, FOND T L, et al. Attributed graph models: modeling network structure with correlated attributes[C]// The International World Wide Web Conference. Seoul, Korea, 2014: 831-842.

[20]CHOI J, YI S, LEE K C. Analysis of keyword networks in MIS research and implications for predicting knowledge evolution[J]. Information and management, 2011, 48(8):371-381.

[21]袁劲松, 张小明, 李舟军. 术语自动抽取方法研究综述[J]. 计算机科学, 2015, 42(8): 7-12.
YUAN Jinsong, ZHANG Xiaoming, LI Zhoujun. A summary of the study on automatic extraction of terminology[J]. Computer science, 2015, 42(8): 7-12.

[22]GUAN A, WANG Y, YANG L. Automatic term extraction for chinese opera domain ontology[C]//International Conference on Fuzzy Systems and Knowledge Discovery. Zhangjiajie, China, 2015: 1372-1376.

[23]TAO L, WANG X L, GUAN Y, et al. Domain-specific term extraction and its application in text classification[J]. Acta electronica sinica, 2007, 35(2): 328-332.

[24]MTC Castellví, RE Bagot, JV Palatresi. Automatic term detection: a review of current systems[J]. Recent advances in computational terminology, 2008, 52(1): 53-88.

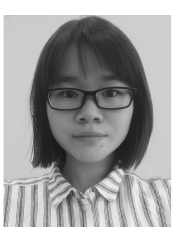
[25]黄勋, 游宏梁, 于洋. 关系抽取技术研究综述[J]. 现代图书情报技术,2013, 29(11): 30-39.
HUANG Xun, YOU Hongliang, YU Yang. A summary of research on relational extraction technology[J]. New technology of library and information service, 2013, 29(11): 30-39.

[26]DEY L, ABULAISH M, SHARMA G. Text Mining through Entity-Relationship Based Information Extraction[C]// International Conferences on Web Intelligence and Intelligent Agent Technology-Workshops. Silicon Valley, USA, 2007:177-180.

作者简介:

金晨,男,1991 年生,硕士研究生,主要研究方向为人工智能、机器学习、知识网络。

谢振平,男,1979 年生,副教授,CCF 会员,博士,主要研究方向为演化认知、知识网络、机器视觉。



任立园,女,1990 年生,硕士研究生,主要研究方向为机器学习、数据挖掘。