信息系统学报
第 18 辑：38-48

China Journal of Information Systems
38-48

面向用户兴趣的知识关联挖掘模型研究*

应　璇　孙济庆
（华东理工大学 商学院，上海 200237）

摘　要　用户兴趣的挖掘分析是实现高效精准知识服务的重要基石。本文在总结现有领域内数据挖掘分析方法的基础上，试图通过融合共现统计关联、时序统计以及语义关联等知识挖掘手段，从知识视角构建全面的用户知识兴趣关联模型。以用户知识的空间结构为基准，有效精准地发现用户知识需求热点，为信息资源服务机构的知识服务提供有力支撑。

关键词　用户兴趣模型，语义关联，共现关联，时序关联，知识结构

中图分类号　G353.1

1　引言

近年来，用户兴趣模型的构建研究在各种信息服务领域中有着广泛的应用，其目的在于从用户的各类历史行为文本数据中挖掘出与用户兴趣需求相关的价值信息，从而有效支持数据分析、决策制定等行为，对个性化信息推荐等诸多情报分析、信息组织、知识服务流程起着基础性的支持作用。值得指出的是用户的个性化知识需求是相对稳定的、具有一定时序规律的知识需求，与那些随机的、临时的信息需求有明显区别，因此处理用户知识需求角度的挖掘方法必须以用户知识结构的自身属性为基准，能够从知识产生时间的纵向发展和空间的横向渗透角度，有效测度反映用户主要知识兴趣特征的术语及其术语相关性，在形式化表达客观知识体系的基础上，深入研究知识结构及其蕴含的规律，并通过逻辑推理获取术语间的潜在语义信息，进而能够多维度扩展知识网络空间，同时也能为知识组织、信息扩展、文本挖掘、自然语言处理等一系列的用户研究提供技术基础。

目前，由于各类文本文件中知识点分散、混沌、无序，数据爆炸和知识相对匮乏的矛盾日益突出，为面向知识发现、知识服务等的兴趣挖掘与管理方式带来了新的机遇和挑战。要解决这一问题，需要从传统物理层次的信息组织转向认知层次的知识组织，从提供信息服务转向提供知识服务[1]。因此，本文旨在从现有数据间的关联基础上重新探索有效的知识兴趣提取方式，探索以大数据信息分析提取整体目标数据的知识关联，力图构建相对完善的知识空间体系，通过有序发散的知识架构表达研究对象的时序知识路径，为目标用户知识需求提供更精准的个性化知识服务。

2　现有研究评述

用户兴趣的知识结构指的是研究目标（学科主题领域、研究学者、科技文献、行为文本等）内

* 基金项目：面向知识服务的学科领域术语语义分析及应用研究（项目编号：13BTQ053）、国家社会科学基金项目。
通信作者：应璇，华东理工大学商学院，博士研究生，E-mail：nhwhzyh2006@sina.com。

部隐性知识的关联性提取及挖掘方法。随着人类认知观的发展，信息处理技术表现出明显的数据—信息—知识—情报转换驱动效应，价值信息的获取在各个行业、产业领域获得了重视。

目前，学术界对用户知识兴趣的关联挖掘方法已经形成了不同层面的研究成果。传统的挖掘方法为结合文献计量学指标的方法、基于科技文献数据库的统计分析方法、机器学习方法和基于文献共引聚类网络分析的方法等[2]，这些分析方法针对用户兴趣的研究被直接应用于个性化信息检索和 Web 概念识别等网络信息服务中[3]。在较新的兴趣挖掘相关研究中，其主要分为几个方面：苏雪阳等研究了基于本体和模式进行的网络用户兴趣挖掘，该方法从用户搜索日志中获取访问行为元素，并借助通用本体中的概念进行关联，进而描述网页所体现的用户个体兴趣[4]；任沁和刘伟采用了叙词表改造本体的方法介绍用户兴趣模型的构建，构建过程中应用到了初始本体、领域本体、用户本体和参考本体[5]；周红卫和周宏印构建了层次向量空间用户兴趣模型，从而形成基于向量空间用户兴趣模型的态势情报信息分发机制，通过词频–逆向文件频率（term frequency-inverse document frequency，TF-IDF）分类法进行分析和个性化推荐[6]；陈冬玲等为了在个性化搜索过程中能够准确地挖掘到用户的潜在兴趣并进行相应的聚类分析，提出采用潜语义空间的 Zipf 分布的特性，找到文档的潜在语义空间，在此空间中对用户的兴趣进行聚类，并建立用户兴趣层次树[7]；龚卫华等构造了用户兴趣特征与主题类间的二部图关系，并在此基础上提出了一种基于主题的用户兴趣聚类算法[8]；王金龙等使用概率图模型对文献中的主题知识信息进行了挖掘研究，该方法利用主题模型来获取时间文本的主题及其强度信息，并利用时间序列的逐段线性表示方法去除噪声，最终得到有效的趋势信息[9]。王冰怡等通过分析用户行为数据，从兴趣广度、兴趣深度和兴趣时效三个角度分析用户的兴趣构成，对用户兴趣进行三维建模[10]；还有李树青等［李树青和孙颖[11]、李树青[12]］利用加权关键词贡献的方法，发现学者的个性化时序研究路径，还提出了一种三词共现分析方法，通过获取映入词和紧密环的方式识别学者主要的研究兴趣特征，并在此基础上，推出个性化外文推荐服务。在动态多源性知识的处理方面，赵捧未等研究了对等网环境下知识地图的构建问题，提出了一种根据用户需求查询动态知识地图的构建方法，实现了语义层次上的知识地图，解决了静态生成的知识地图的死链接以及更新系统代价大等问题，具有较好的灵活性[13]。

综上所述，已有的学术研究对用户兴趣需求的挖掘多停留在数据层面，其重点集中于利用数据分析方法，以及本体、图论、概率统计、主题模型等进行分析，挖掘有效高频数据，为用户提供辅助决策，实现大数据价值外显的过程[14]。虽然部分已引入了时间因子进行趋势性分析，但在关联角度和挖掘范围的方法论上依旧相对单一。分析这类方法与研究路径可发现：①基于关键词或主题的兴趣模型不能很好地反映用户的知识兴趣；②基于词频数据统计和概率分析的模型对用户兴趣关联及发展过程缺乏指向性作用；③传统的数据关联模型仅仅从词共现等表层表达特征兴趣的关联特性，其数据模型构建极少涉入知识关联价值的探讨。而现有的信息资源具有大数据的特征基础，决定了资源性服务从数据到信息到知识的路径转化。传统的基于用户单纯兴趣需求的数据资源供给已无法满足用户的现实知识需求，须从知识角度探索用户兴趣提取，并结合多维度的知识关联发现，实现目标用户群潜在需求的对应性知识服务。多角度知识关联的引入能够更注重用户自身的知识结构，有针对性地进行知识需求分析，确定用户的时序兴趣发展，挖掘用户的潜在知识需求，形成与目标用户相匹配的个性化知识服务产品。

3 用户知识兴趣模型宏观架构

本文提出的用户知识兴趣模型的分析架构基本思路是通过用户检索过程中表达的个人兴趣或知

识需求的信息，进行知识提取和关联分析，构建基础逻辑表达；将知识术语的用户兴趣基本要素，建立语义关联；通过关联分析探索用户兴趣发展的知识空间定义，构建用户知识模型。其具体过程如图1所示。

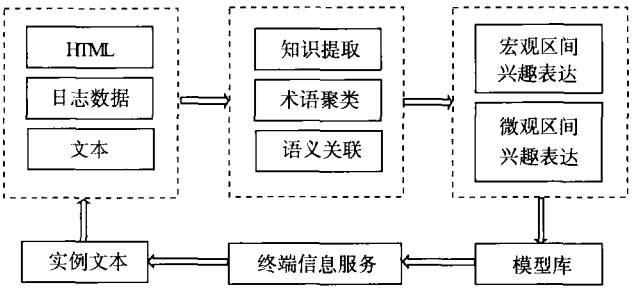

图1 用户动态知识需求模型宏观架构

用户的知识需求采集可来源于多种不同的数据源。在知识兴趣提取的过程中，用户知识点的采集是知识组织、知识关联的基础数据源。在现有的大数据信息资源环境中，知识点信息可以从多种渠道获取，如文献数据、事实数据、专业领域数据、日志数据以及用户知识行为数据等，因此用户知识点提取是对数据进行有效的清洗和聚类，并使之序化，形成统一的知识客体。分析分散知识点之间的相互关联规律，针对实际问题，探索用户知识需求的发展历程。

知识客体的形成则可实现术语语义的价值最大化表达。通常所说，术语是专门用途语言中专业知识的语言表达[15]。术语集合与术语之间的关系能够反映领域内的知识与知识结构，有助于实现人机之间的语义信息交互，通过语义关联模型建立术语间的语义关系，同时采用多种策略融合的语义模型形成多层次、多粒度的语义关联关系。该关联关系包含多种不同的形式，如时间关联、相似关联、共现关联等，进而形成面向特定领域的术语簇、知识链。

本文对最终用户知识需求模型的表达则通过宏观及微观两种表现形式，分为宏观区间兴趣表达及微观区间兴趣表达两种形式，宏观区间的兴趣表达指的是依照传统统计理论跟随用户整体知识兴趣发展时序路径表达用户整体兴趣节点，而微观区间的兴趣表达则是在宏观兴趣节点的基础上通过不同的关联挖掘方法形成用户兴趣的微观术语簇扩展，在微观范围内关注用户的细节动态变化，最终形成相对完整的用户兴趣模型库。

4 面向用户兴趣的知识需求模型构建

4.1 模型假设

假设两个术语在一定时间动态微观范围内出现相邻关系，则该术语对之间存在相应的关联。

微观时间范围内的相邻共现指的是对于同一目标对象或目标对象群组，在相当小的连续时间区间内，产生了知识获取行为。

用户知识获取行为源于用户知识需求的驱动，也就是说在用户存在知识需求或兴趣需求的前提下，才会产生连锁的知识行为。因此，知识行为的提取是用户知识兴趣研究的基础。通常意义下，用户在时间上的连续知识获取行为以术语形式呈现，因而用户表达的术语间存在关联关系。多次连续行为中涵盖的术语或术语集合，能清晰地表达用户的知识需求；即使从形式上看术语之间没有任何知识关联，但连续呈现本身也体现了一定的关联性。

4.2 模型中的术语关联

在模型构建之前，我们将对用户兴趣数据源文本的属性关联进行必要的分析与梳理。其基本思路为将文本中的术语与术语（词与词）之间存在的某种联系，在形式化表达客观知识体系的基础上，从横向和纵向两方面进一步扩展其关联结构。本文将术语关联分为以下两类。

（1）术语相似度语义关联。

我们将某一领域概念下密切相关的术语统称为广义的同义词，这类词往往具有一定的共性。由于汉语中自然语言对概念表达灵活、自由，相关学者对这一课题有深入的研究。其在语言学、情报学中应用广泛。利用这一突出特点，我们提出语素相似度达到一定阈值的相似术语之间存在一定的关联性。

（2）相邻共现术语关联。

目前学术界的词共现分析方法均建立在此基本假设之上，即如果两个词语共现在同一文档中的同一个单元内，则可以认为这两个词语在意义上相互关联，或者说存在语义关联的两个词语应具有更大的可能经常共现于同一文档中的同一个单元内[16]。不同词语之间存在的两两共现关系将这些词语关联构成一个词语共现网络，这种词语共现网络可用于探索用户的知识兴趣。

4.3 用户兴趣模型构建

在用户知识行为文本中的语义关联的基础上，本文重点讨论知识兴趣链模型，见图 2。

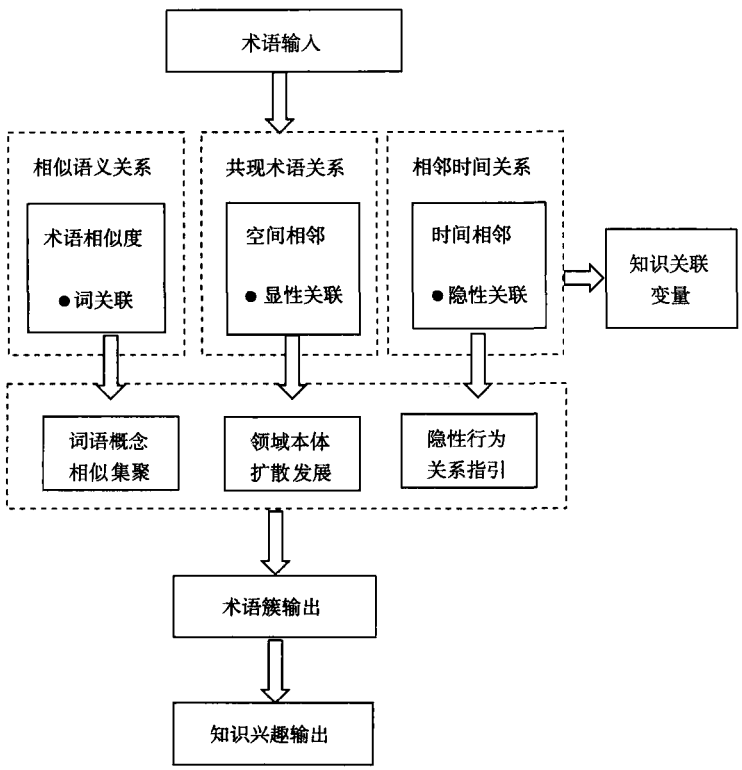

图 2 知识需求模型

该模型遵循上文所述的提取过程宏观架构，核心部分为用户兴趣的知识关联，实现方法为通过术语相似度算法进行概念集合、采用统计理论对时间上具备相邻条件的隐性关联术语进行统计整合以及通过双词共现分析方法对具备空间相邻条件的术语进行本体关联扩散定义。下文将对知识关联变量的具体逻辑表达方式进行详细探讨。

5 知识关联的逻辑表达

5.1 术语语义相似度关联

本文利用传统计算机识别汉语语义相似度的算法——基于单汉字的字面相似度算法进行同义词识别和字面相似度匹配，反映知识术语间的相似程度。利用术语相似度进行知识点“关联”分析，其目的在于探索目标用户在同一范畴内使用不同行为方式表达需求的问题，如用户先后使用“甲醛催化氧化”和“甲醛氧化”进行知识探索的行为。

在对两个术语进行术语相似度分析时，需综合考虑两个因素，分别为两个词中共同包含的语素数量以及相同语素在两个术语中的位置因素。具体方法及权重设定有效性采用朱毅华等的研究成果，在此不做多余赘述[17]。设定上述两个因素的权重：两个词含有相同语素的个数的影响占60%；相同语素在各个词中的位置关系的影响占40 %。由此拟定相似度匹配公式：

$$\mathrm{SIM}_{\mathrm{word1,word2}}=60\%\cdot\frac{\left[\frac{\mathrm{xsword}}{\mathrm{ctrlword}}+\frac{\mathrm{xsword}}{\mathrm{keyword}}\right]}{2}+40\%\cdot\mathrm{dp}\cdot\frac{\left[\sum\frac{c\_\mathrm{xsword}(i)}{\sum\mathrm{ctrlword}(i)}+\sum\frac{k\_\mathrm{xsword}(i)}{\sum\mathrm{keyword}(i)}\right]}{2}$$

其中，xsword表示两词含有相同语素，即匹配字的个数；$\sum\frac{k_\mathrm{xsword}(i)}{\sum\mathrm{keyword}(i)}$表示匹配字在被匹配词中所处位置的权数之和；$\sum\frac{c_\mathrm{xsword}(i)}{\sum\mathrm{ctrlword}(i)}$表示匹配字在待匹配词中所处位置的权数之和；dp表示位置系数，其值为被匹配词与待匹配词语素总和之比。

5.2 时间相邻共现关联

以4.1的模型假设为理论基础，通常文本所涵盖的时间信息中包含着不同术语之间的时序关联，按照这个思路，我们在前期知识关联的研究基础上，探索了直接利用数据文本中获取的时间信息，完成时间连续簇的构建。

在传统的语义信息表达方法中，每一个知识单元必然包含其出现的相应时间信息，从而构成时间元，表达方式如下：

$$\mathrm{TimeGram}(x_n)=\langle\mathrm{Unit},\mathrm{Time}\rangle$$

其中，知识点之间的时间差表示为

$$\Delta\mathrm{Time}(x_{ij})=\mathrm{Time}_i-\mathrm{Time}_j$$

根据比例相对指标，时间相邻共现研究的重点在于阈值的设定，当时间差小于设定阈值时，我们默认该术语对构成时间相邻共现关系。阈值设定采用抽样调查方法，采用研究术语集合中抽取的部分样品$x_1,x_2,\cdots,x_n$作为样本，样本容量为n，人工判断n个相互独立且与总体有相同分布的随机变量，以φ定义术语间的语义相关度：

$$\varphi=\varphi(\Delta t_{11},\Delta t_{12},\cdots,\Delta t_{1n})$$

其中，判断相关时$\varphi=1$，非相关时$\varphi=0$。最终以样本函数φ所对应时间差的样本k阶中心矩作为最终阈值，计算公式如下：

$$M_k=\frac{1}{n}\sum_{i=1}^{n}(\varphi_i-\bar{\varphi})^k,k=2,3,\cdots,n$$

5.3 空间相邻共现关联

空间相邻是指整个用户知识兴趣网络中部分语义关联关系会通过同一知识点进行发散。基于上述分析，为了更为准确地描述相邻共现关联现象，给出如下定义。

视整体用户知识兴趣集合为：$G=(C,R)$，其中，C 为用户兴趣术语集，即顶点集合，$C=\{c_1,c_2,\cdots,c_n\}$；R 为 C 上的关联关系集合，$R=\{r_1,r_2,\cdots,r_m\}$。若任取的两个术语 C_i、C_j 之间存在关联关系 r_k，可表示为 $\forall c_i,c_j\in C$，$\exists r_k\in R$，$\langle c_i,c_j\rangle\in r_k$。依次类推，如果 $\forall c_i,c_j,c_k\in C$，$\exists r_p,r_q\in R$，$(c_i,c_j)\in r_p$，$(c_j,c_k)\in r_q$，则关系 r_p 与关系 r_q 空间相邻[18]。

依据共现关系方向性的不同，本文将空间相邻语义关系共现分为四类。图 3(a)表示为：$\langle c_i,c_j\rangle\in r_p$，$\langle c_j,c_k\rangle\in r_p$，即概念 c_i 与 c_j 间通过关系 r_p 相关联，概念 c_j 与 c_k 间通过关系 r_q 相关联，如图 3（a）所示，$c_i\xrightarrow{r_p}c_j\xrightarrow{r_q}c_k$。

图 3（b）、图 3（c）、图 3（d）现象同理。通过这四类情况的综合相邻关系发现共现隐含的规律。按照上述共现关系分类，结合在实际操作情况中出现概率的观察，本文将有代表性的相邻共现关系列举如下。

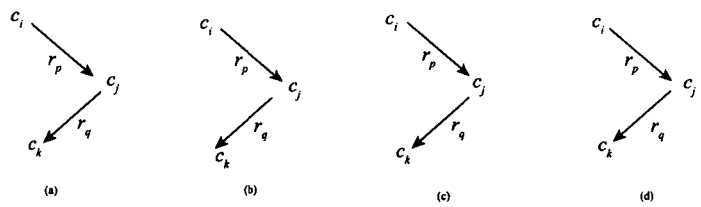

图 3 四类空间相邻共现关系

（1）闭合环式共现——强并行关联关系。

术语共现分析方法能够在词与词本身建立连接，但有很大概率会产生环式共现。针对实际案例研究，我们不难发现一些特定规律，即随着时间的延续性发展，在 c_i、c_j、c_k 等术语之间逐一发生共现关系后，时间链后期出现的某个术语与 c_i 产生回归共现的情况。我们认为，若共现形式以闭合环的样式出现，则表示为共现知识点之间的强并行关联性。这里定义的“并行”，指的是闭合环式共现中的所有知识单元节点具有统一的权重和概念表达程度（图 4）。

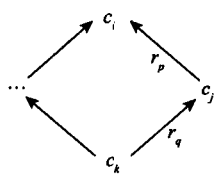

图 4 环式空间相邻共现关系

（2）发散式共现——增补关系。

发散式共现可以认为是共现关系中较为常见的共现形式。这类共现形式中通常具有一到两个核心词（参见图 5 中的 c_j），术语 c_j 与其他某几个术语 c_i、c_k、c_l 均发生共现。若共现以这种一对多的发散式出现，共现知识点间（除主知识点外）不存在重叠性，我们认为其属于外围知识点对应主知识点的

增补关系。

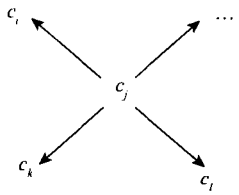

图5 发散式空间相邻共现关系

当然在实际的知识组织过程中，共现往往比较复杂，可能并不是以上述几种单一模式出现，而是以多种模式的混合出现，所以在计算过程中需综合考虑。

在用户需求对应的知识网络空间中进行需求判定，我们需要通过全部数据概率网络集合中所有关联信息。通过统计和汇总，梳理出贴合用户兴趣发展的最高价值推荐知识集合，最终形成用户需求知识元关联表达方式为语义关联及共现关联并集。

6 实证研究

笔者以检索日志为例，对某机构在某一中文学术文献数据库的检索日志进行了用户需求行为数据获取，时间跨度为2010年1月至2014年12月，共5年，总共获得8 714 727条有效记录。

依据上述用户知识需求动态分析的宏观架构，首先对用户检索行为数据进行数据清洗与提取，过滤掉部分定义文献类型、检索要素的标识符等不相关字符串。其次通过分词方法进行粒度分类等操作，排除用户误操作产生的数据，最终形成统一的术语集合。

6.1 术语相似语义关联实验结果

我们抽取了某一特定用户其在五年时间跨度范围内的所有信息获取行为记录，共计10 284条，数据样例如表1所示。

表1 实验中所抽取用户名的需求行为记录信息

检索关键词	时间跨度	请求	记录	页码	开始时间	身份
切割液	244	0	3 999	3	2011-08-07 11：13：29	0
切割液	218	0	3 999	4	2011-08-07 11：13：41	0
切割液	229	0	3 999	5	2011-08-07 11：14：39	0
太阳能硅片 切割	269	0	25	1	2011-08-07 11：15：04	0
太阳能硅片 切割	193	0	25	2	2011-08-07 11：16：41	0
太阳能硅片 切割	194	0	25	3	2011-08-07 11：17：10	0
电子电镀 银	74	11	165	1	2011-08-22 11：43：25	0
电子电镀 银	20	11	77	1	2011-08-22 11：43：25	0
电子电镀 银	176	11	21	1	2011-08-22 11：43：25	0
电子电镀 银	230	11	88	2	2011-08-22 11：43：41	0
电子电镀 银	202	11	21	2	2011-08-22 11：43：41	0
光转换薄膜	328	0	956	6	2011-07-13 09：45：15	0
电子电镀	250	11	1 121	8	2011-08-22 11：42：18	0
电子电镀	224	11	363	8	2011-08-22 11：42：18	0

我们对全部数据集合中的所有术语进行了相似度计算，得到的术语相似度计算结果样例如表 2 所示。

表 2 特定时间范围内术语相似度排名前 15 的术语对

k_1	k_2	SIM
臭氧过氧化氢	过氧化氢	0.885 317
超声波细胞粉碎机	超声波细胞粉碎	0.873 611
甲醛催化氧化	甲醛氧化	0.872 619
过氧化物酶的测定	植物过氧化物酶的测定	0.857 091
催化甲苯	光催化甲苯	0.849 333
低矿物质水	矿物质水	0.849 333
光催化氧化	光催化	0.848
超声波浓度计	超声波浓度	0.835 714
蛋白溶解	蛋白盐溶解	0.828
氢氧化镍复合电极材料	氢氧化镍复合材料	0.816 364
比色法测定维生素 C	钼酸铵比色法测定维生素 C	0.813 462
甲苯降解产物	催化甲苯降解产物	0.812 5
浸提 pH	提 pH	0.81
油稳定性	稳定性	0.81
超级电容	超级电容器	0.806 667

通过术语相似度的计算我们不难发现，中文术语相似度的计算解决了术语单元中由于前置、后置等产生的术语多样性问题，同时为原术语单元的知识定义提供了补充扩展的空间，其一方面有利于我们对用户知识兴趣的聚类，另一方面也对用户的需求给予了更准确的指向定义（如蛋白溶解指向为蛋白盐溶解）。

6.2 时间相邻关联实验结果

由于用户需求行为的多样性，某些术语并未与其他术语之间构成语义共现关系，同时从术语相似度的角度也无法定义其相关的知识集合。例如，表 3 中“椰子”，利用术语相似度计算和共现计算都无法识别其与“植物油酸”“酸化油”等术语存在的相关性。而实际研究中，“椰子”富含丰富的椰子油，经硫酸化后可生成椰子酸化油，是良好的工业净洗剂、润滑剂、泡沫稳定剂，在制备聚酯多元醇中发挥重要作用。因此“椰子”与“植物油酸”“聚酯多元醇”等知识元之间具备明显的隐性知识关联。采用时间连续簇的计算能有效弥补现有的知识关联分析方法的不足。

表 3 时间连续簇

检索关键词	开始时间
植物油酸	2012-04-18 12：22：11
植物油聚氨酯	2012-04-18 12：22：19
聚酯多元醇	2012-04-18 12：25：45
酸化油	2012-04-18 13：01：00
椰子	2012-04-18 13：23：54

6.3 空间相邻共现关联实验结果

对术语共现关联的研究，我们将全部数据集合中的所有共现术语进行了实验分析。其中，对产生闭合环式共现和发散共现的术语进行了集中处理。图 6 中，“聚氨酯”“有机硅”“太阳能电池”“减反射材料”“减反射涂层”这五个术语，在逐一发生共现关系后，又出现了其中某个词回归共现的情况；图 7 中，“量子点”成为核心术语，分别与其他术语——“光电转换”“光伏电池”“发光材料”“太阳能”“制作方法”等发生共现。这类知识元均在宏观区间的需求表达网络中展示（表 4 和表 5）。

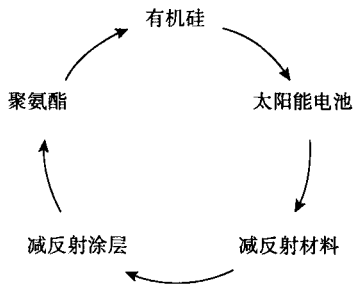

图 6 环式共现案例

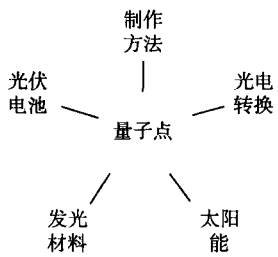

图 7 发散共现案例

表 4 闭合环式共现

k_1	k_2	k_3
有机硅	太阳能	
有机硅	减反射	涂层
太阳能电池	减反射材料	
减反射涂层	太阳能	
有机硅	聚氨酯	
聚氨酯	有机硅材料	
减反射涂层	减反射材料	

表 5 发散式共现

k_1	k_2
量子点	光电转换
量子点	太阳能
量子点	发光材料
量子点	光伏电池
量子点	制作方法

6.4 用户知识结构网络表达

最终为了整体检验知识关联挖掘模型的可行性，在排除人工参与的情况下，对整体数据文本基于上述三个层面的相似度计算自动建立用户知识结构网络体系。将上述实验结果利用 gephi 工具进行可视化分析，生成了该用户的知识结构网络图谱，进一步基于用户兴趣模型对用户检索行为文本进行语

义关联和共现关联的整合，生成用户实际需求知识结构网络，通过查看微观层面的图谱则可以发现用户的知识结构分支状态（图 8）。

图 8　用户知识结构网络

7　结论

目前知识关联挖掘的研究已渗透于信息处理分析理论及应用研究过程的各个环节，在知识分类、知识组织、知识检索、知识发现、知识网络、知识地图和知识图谱等应用过程中均发挥着至关重要的作用。本文的研究目标及创新性在于从用户兴趣角度出发构建多维度细化知识关联挖掘模型，其目的在于尽可能依循用户自身知识点—知识簇—知识结构（网络）的有效路径，对用户兴趣知识内容进行系统性挖掘分析。

文中通过面向用户需求的知识关联挖掘宏观架构搭建模型系统。该模型从词语概念相似集聚、领域本体扩散发展以及隐性行为关系指引三个方面入手建立知识关联变量，具体实现方法分别为通过术语相似度算法进行概念集合、基于统计理论的时序上相邻条件的隐性关联术语整合以及采用双词共现分析方法对具备空间相邻条件的术语进行本体关联扩散，实现知识关联挖掘。除此之外，本文的知识关联挖掘模型在关联判断、关联强度以及关联方式上都做了深入详细的探讨和新的尝试，如列举了发散式共现、闭合环式共现等多种关联方式，并对其进行了关系定义，指出其所对应的增补关系、强并行关联关系等，能够动态地表达用户的知识需求，以在结构网络属性基础上定义其相关知识兴趣的发展规律。从目前的实验结果来看，初步实现了预期的研究目标。知识提取是一个复杂、持续且要求精准的过程。本文在研究方法上仍有一些缺陷，如对知识时序变化的细微趋势识别以及从微观层面上定义用户的小范围需求漂移情况等，仍然需要更精细的研究。

参 考 文 献

[1] 徐绪堪，蒋勋，苏新宁. 面向知识服务的知识组织框架体系构建[J]. 情报学报，2013，32（12）：1278-1287.

[2] 殷蜀梅. 判断新兴研究趋势的技术方法分析[J]. 情报科学，2008，26（4）：536-540.
[3] Borth D，Ulges A，Breuel T. Dynamic vocabularies for web-based concept detection by trend discovery[C]. Proceedings of the 20th ACM international conference on Multimedia，Nara：IEEE Press，2012：977-980.
[4] 苏雪阳，左万利，王俊华. 基于本体与模式的网络用户兴趣挖掘[J]. 电子学报，2014，42（8）：1556-1563.
[5] 任沁，刘伟. 本体技术在用户兴趣建模中的应用研究[J]. 信息系统工程，2012，（5）：108-137.
[6] 周红卫，周宏印. 基于向量空间用户兴趣模型的态势情报信息分发机制[J]. 指挥信息系统与技术，2015，6（6）：90-95.
[7] 陈冬玲，王大玲，于戈，等. 基于PLSA方法的用户兴趣聚类[J]. 东北大学学报，2008，29（1）：53-56.
[8] 龚卫华，杨良怀，金蓉，等. 基于主题的用户兴趣域算法[J]. 通信学报，2011，32（1）：72-78.
[9] 王金龙，徐从富，耿雪玉. 基于概率图模型的科研文献主题演化研究[J]. 情报学报，2009，28（3）：347-355.
[10] 王冰怡，刘杨，聂长新，等. 基于用户兴趣三维建模的个性化推荐算法[J]. 计算机工程，2015，41（1）：65-70.
[11] 李树青，孙颖. 基于加权关键词共现时间元的个性化学术研究时序路径发现及其可视化呈现方法[J]. 情报学报，2014，33（1）：55-67.
[12] 李树青. 基于三词共现分析的学者主要研究兴趣识别及个性化外文推荐服务的实现[J]. 情报学报，2013，32（6）：629-639.
[13] 赵捧未，王亚楠，窦永香. 对等网环境下动态知识地图的构建研究[J]. 信息系统学报，2010，（1）：340-344.
[14] 张引，陈敏，廖小飞. 大数据应用的现状与展望[J]. 计算机研究与展望，2013，50（S2）：216-230.
[15] 揭春雨，冯志伟. 基于知识本体的术语定义（下）[J]. 术语标准化与信息技术，2009，（3）：14-23.
[16] Peat H J，Willett P. The limitations of term co-occurrence data for query expansion in document retrieval systems[J]. Journal of the American Society for Information Science，1991，42（5）：378-383.
[17] 朱毅华，侯汉清，沙印亭. 计算机识别汉语同义词的两种算法比较和测评[J]. 中国图书馆学报，2002，28（4）：82-85.
[18] 裘江南，张彬，王慧丽，等. 客观知识体系中语义关系相邻共现规律研究[J]. 情报学报，2012，31（2）：126-135.

A Model of Knowledge Association Mining Based on User Interest

YING Xuan，SUN Jiqing
（East China University of Science and Technology，School of Business，Shanghai 200237，China）

Abstract The mining analysis of user interest is an important cornerstone of high-efficient precise knowledge services. On the basis of summarization of current data mining analysis methods，through semantic relevance and co-occurrence association，this article aims to build a comprehensive user knowledge interest correlation model from the perspective of knowledge. Based on spatial structure of user knowledge，to find hot spots of user knowledge needs efficiently and precisely，to provide strong support for knowledge service of information resource service institute.

Key words user interest model，semantic relevance，co-occurrence association，temporal association，knowledge structure

作者简介

应璇（1986—），女，华东理工大学商学院博士研究生，研究方向为数据分析、知识管理、大数据挖掘等。E-mail：nhwhzyh2006@sina.com。

孙济庆（1952—），男，博士生导师，研究馆员，研究方向为技术管理与信息系统、信息检索、知识管理、现代情报学等。E-mail：jqsun@ecust.edu.cn。