基于共现分析的语义信息检索研究*

邱均平　楼　雯

摘　要　提高信息检索系统的用户体验度可以从查询优化的算法和增强可视化展示的研究等方面入手。本文利用文献调研、共现分析和构建本体等方法,设计基于共现分析的语义信息检索和流程,并利用武汉大学图书馆的书目检索系统中"世界考古"类目的数据进行实验分析。经过文献调研发现,目前语义信息检索主要集中在基于本体的查询技术、语义标注问题和语义关系检索等方面的研究,语义信息检索目前只能做到表层相关的检索。构建的基于共现分析的语义信息检索模型包括规范器、分析器和本体构建器三个部分,其中分析器是本模型的核心。经过实验分析发现共现分析可以应用于语义信息检索,并比较得出基于共现分析的语义检索比传统检索更具人性化、引导性。图7。表1。参考文献19。

关键词　语义信息检索　本体　共现分析　作者关键词耦合

分类号　G250

Semantic Information Retrieval Research Based on Co-occurrence Analysis

Qiu Junping & Lou Wen

ABSTRACT　In order to improve user experience of information retrieval system, researchers always explore into query optimization algorithms, visualization and other aspects. This paper used literature review, co-occurrence analysis, ontology building and other methods to design a model and process of semantic information retrieval based on co-occurrence analysis. And the paper selected the world's archaeological category data from Wuhan University Library's bibliographic retrieval systems to experimental analysis. Literature review summarized that semantic information retrieval research mainly concentrates in the ontology-based query techniques, semantic annotation and semantic relation retrieval. And most recent system can only realize obvious relations retrieval. This paper constructed a model to realize potential relations retrieval. The model is divided into three parts: Normalizer, Analyzer and Ontology builder. Analyzer is the model core. Experimental analysis proved the feasibility of co-occurrence analysis used in semantic information retrieval. Compared with traditional retrieval, semantic information retrieval based on co-occurrence analysis is more user-friendly and instructive. 7 figs. 1 tab. 19 refs.

KEY WORDS　Semantic information retrieval.　Ontology.　Co-occurence analysis.　Author keyword coupling analysis.

知识管理和知识经济的兴起,引发了知识社会化和社会知识化的趋势。计算机和网络技术的广泛应用,克服了时间、地域、机构之间的隔离,使科研人员、信息资源、科学仪器设备和计算工具等紧密联系在一起,营造了 e-science 的协同科研环境和 e-learning 的协同学习环境。在新的环境下,用户的信息需求更加多样化、知识化,要求也更高了。面对浩如烟海的各类信

*　本文系国家社科基金重大项目"基于语义的馆藏资源深度聚合与可视化展示研究"(批准号:11&ZD152)的研究成果之一。

通讯作者:邱均平,Email:jpqiu@whu.edu.cn

息,用户并不满足于图书馆只提供一个文献线索这种简单的服务,或者提供大量相关性不强的信息列表。当前,由于缺乏对信息资源深入的知识组织和规范控制,虽然身处"信息海洋",却面临"信息泛滥、知识匮乏"的困境,图书馆等传统信息机构的信息服务工作面临着巨大的挑战。

正是这种难题与挑战,才使得快速存取、有序组织、深度挖掘和有效利用数字化、网络化信息资源,成为图书馆适应知识经济时代的必然要求,基于语义的文本挖掘、信息组织和信息检索等新课题应运而生,越来越多的研究使得该领域在近几年成为图书情报领域的研究热点并得到长足发展。但大多研究都停留在模型、框架和体系的设计等层面上,很少在技术层面或微观层面解决图书馆个性化服务或书目检索技术等问题。本文在分析了基于语义的信息检索研究现状的基础上,将共现分析引入语义信息检索,提出基于共现分析的语义信息检索的模型,并以实验验证其提升馆藏资源的检索效率,提高用户使用数字图书馆的满意度。

1 语义信息检索研究

早在上世纪80年代,对语义检索的讨论就出现在SIGIR会议论文中,但语义检索研究始终受制于语义信息处理发展水平的局限。随着自然语言处理、人工智能的发展,尤其是语义网技术的兴起与发展,语义检索研究自上世纪末以来得以迅速发展[1]。语义信息检索就是要让用户在输入自然语言作为检索词的时候,能出现与该检索词相关的更多词,而不是机械地将与该检索词匹配到的所有信息一一列入检索结果。目前国内外语义信息检索研究主要集中在以下三个方面。

(1)基于本体的查询技术。查询技术首先涉及查询语言,由W3C推出的RDF、SPARQL等系列查询语言已经可以实现对语义数据的查询并且应用广泛,如余传明博士[2]阐述并比较了三种基于查询语言的检索机制。国外的研究一般集中在利用语言本体(如WordNet)中的同位词、上下位词以及上下文检索技术对所要查询的内容进行语义消歧并进行查询扩展[3-4],文献[5]利用WordNet实现了地理信息的语义检索,另外Yongxiang Dou, Xiaoxian Bei[6]等人在WiCOM会议上对P2P网络信息内容进行语义查询和实现。不论是查询技术还是针对不同的查询对象,都涉及一项关键的技术,即语义相似度(或称相关度算法),语义相似度是指两个概念间的相似程度,目前多以研究路径长度方法、信息论方法和基于概念特征的方法三个方面为主[7-8]。

(2)语义标注问题。语义标注可以标注元数据、概念、网页、文档,这些被标注的内容便是语义构建系统的内容,所以它是实现资源语义化的基础,董慧[9]等人设计了层叠隐马尔可夫中文分词模型,对历史文献进行了语义信息提取和语义标注。文献[10]引入了向量空间模型,设计了加权算法和排序算法,对大规模文献的关键词进行了语义标注,再进行关键词的检索时用实验证明了语义标注模型和算法的优势;文献[11]也利用了向量空间模型和聚类模型设计了自动二进制分类算法,挑选了特定的网页进行分析;荆涛[12]、熊荣东[13]等在学位论文中分别研究了自然语言处理技术和基于信息容量的相似度算法,进而提出了针对不同领域的语义标注方法,刘海学[14]提出了一种针对元数据的语义标注方法,利用数据集中已有的语义标注信息自动构建生成元数据。

(3)语义关系检索。2009年Payam Barnaghi[15]等人撰文回顾了语义关系检索的历史,最早研究语义关系检索的是Aleman-Meza B, Halaschek C等人,他们发表了一系列文章讲述概念、文档、网页之间的语义关系也可以作为检索的重要内容[16-17];另外还有基于语义的多媒体信息检索和本体构建技术两个方面的研究。

目前语义信息检索的基本思路和技术路线是(见图1),先后利用概念提取、格式转换、信息整合等技术将资源库存储成本体库,再使用查询优化和语义关联等技术把本体库中的信息显现在用户眼前。这里提到的概念提取、格式转

换和信息整合是通过语义标注实现的，查询优化和语义关联则是依靠基于本体的查询技术和语义关系检索实现的。传统信息检索一般利用词匹配的结果和排序算法将检索集展现给用户，语义信息检索的结果则包含了与检索结果相关的其他信息，目前数字图书馆的检索系统一般实现了表层相关的语义检索。比如在某图书馆书目检索系统中搜索与《红楼梦》有关的书籍，传统信息检索系统只会搜索得到有关《红楼梦》的书籍列表，语义检索系统会搜索到《红楼梦》的作者曹雪芹的有关书籍，但却搜索不到与《红楼梦》相似的古典小说，或与曹雪芹关联性高的其他作者的书籍。若要实现此类检索结果，关键技术是在查询优化和语义关联的研究上，探索新的检索模型和检索流程。

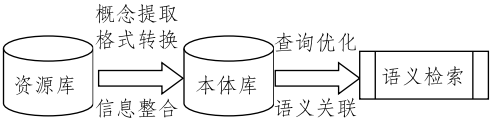


图 1　语义信息检索基本模型

2　基于共现分析的语义信息检索模型

本文在借鉴国内外已有研究的基础上，设计了一种基于共现分析的语义信息检索模型（见图 2）。该模型共分为三个模块，由规范器、分析器和本体构建器组成，每个模块都由不同的参与人员、设备和方法构成。模型基于以下目标设计：其一，概念提取便捷科学。模型的处理对象是馆藏资源，馆藏资源的外部特征是较为方便提取的数据，同时外部特征是经过主题标引的，也是科学的数据。其二，语义相似度定量化。早在 2006 年，王曰芬教授撰文描述了共现分析在知识服务中的应用，说明了共现分析可用于构建概念空间从而支持语义信息检索[18]。上文说语义相似度有三种方法，共现分析的结果则是基于概念特征的相关度，它能定量化描述语义相似度。其三，可检索语义关系，概念间的关系强弱可为检索内容。目前研究的语义关系往往是用模糊词汇，比如 is-a，belongs-to，have-a-subclass 等，本文将语义关系定量化，能够提高检准率和检全率。

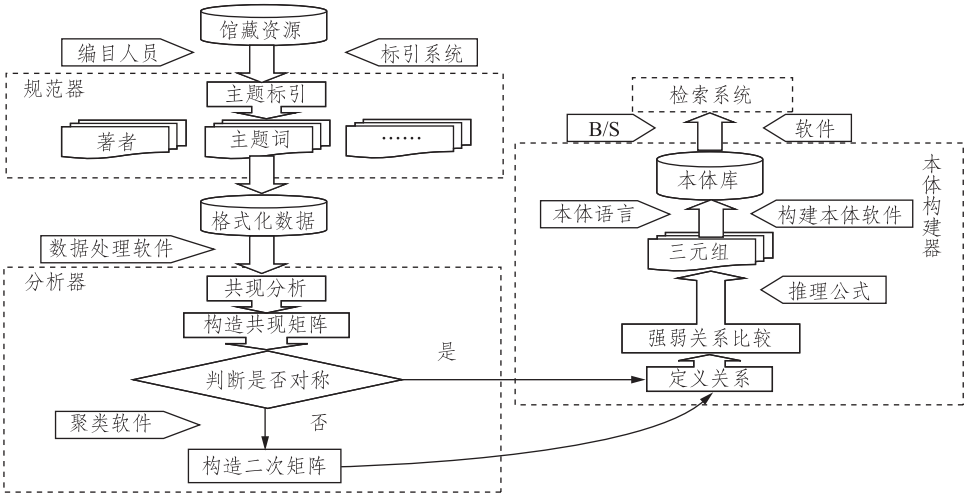


图 2　基于共现分析的语义信息检索模型

2.1　规范器

规范器主要由图书馆编目人员执行，编目人员首先对实体图书馆馆藏书目的内容进行主题分析，并根据分析结果，到主题词表中查找相

应的主题词,将馆藏资源的自然语言转换成规范化的主题语言形态。自上世纪90年代后期出现了计算机辅助标引后,主题标引工作更标准化、规范化,计算机辅助标引使得主题标引得到的数据形成MARC数据,本模型能且仅能利用到的数据便是MARC数据,针对中文馆藏资源,规范器最后得到的数据是CNMARC格式的数据[19]。

规范器是该模型的数据来源,它的稳定和可靠是整个模型正确实施的保证,而主题标引是规范器的核心,因此主题标引的科学与规范决定了规范器的可靠,做好主题标引才能保证整个模型的运行。

2.2 分析器

分析器是该模型的核心模块,它主要进行共现分析。首先用数据处理软件将规范器得到的格式化的数据处理成不同类型的数据格式,再按照共现分析的一般流程进行。第一步是数据抽取,在规范器中得到的数据包括馆藏资源的多方面数据,比如题名、著者、主题词、分类号等,可以根据需要选择特定字段,从而进行下一步骤——构造共现矩阵或数据向量,又因为选择字段的不同,构造共现矩阵或数据向量时会产生多种形式:若只选择著者或主题词字段,那么所构造的共现矩阵或数据向量则为对称矩阵或向量;若同时选择著者和主题词字段,那么所构造的共现矩阵或数据向量则为非对称矩阵或向量。这一判断尤为关键,若为对称矩阵或向量,分析器则直接进入下一模块的进程;若为非对称矩阵或向量,即为著者主题词耦合矩阵,则需要多一项程序,将主题词进行聚类分析,继而构造一个著者主题领域耦合矩阵,才能进入下一模块的进程。这是由于著者主题词耦合形成的矩阵或关系网络一般情况下较为复杂,同一著者对应多个主题词,加之主题词表中语义相近的主题词有多个,便导致了同一著者对应多个语义相近的主题词,造成冗余的数据和关系,因此要将表示相同语义的主题词按照聚类分析的方法凝聚在一起。

2.3 本体构建器

本体构建器是该模型的最终环节,由分析器得到的数据类型有多种,比如表示著者和著者关系的数据、表示主题词之间关系的数据、表示著者和主题领域关系的数据,那么本体构建器的第一道程序就是要明确定义这些关系。而要进行的下一步骤也至关重要,因为在共现分析得到的关系网络中,不管是著者之间、主题词之间还是著者主题之间的关系,都是多对多的关系,那么必然存在节点之间的强弱关系,因此根据一定的推理公式(在下述实验中讨论得出)得出所有关系的定义后,便得到包含概念、属性和实例的三元组。所有的三元组经过本体语言或构建本体的软件处理后形成本体库,本体库中的内容以备检索系统的设计和使用。

3 实验与讨论

为了评估基于共现分析的语义信息检索的可行性,笔者建立了以作者关键词耦合为方法的实验环境。该实验主要考查的指标有两个:一是共现分析和本体论的融合性;二是实验结果用于用户检索的可行性。实验步骤如下:

3.1 数据来源和处理

为实现馆藏资源的聚合和可视化,选择了武汉大学图书馆作为数据来源进行研究。选择武汉大学图书馆基于以下三点原因:

(1)熟悉性。笔者在武汉大学学习工作多年,熟知武汉大学图书馆的发展状况和馆藏资源的特色等,选择武汉大学图书馆作为研究对象,更容易实现本研究的最终目标。

(2)数据可获得性。限于人力和精力,研究数据不可能在实体图书馆的海量图书中获得,而要选择电子数据。一方面,武汉大学图书馆的馆藏书目可以在检索系统中查询得到,同时又有基于中图法和科图法的分类浏览馆藏目录的服务,这样便可获得馆藏图书的数据;另一方面,在武汉大学图书馆检索到的数据可以保存到本地,便于对数据的处理。

(3)实用性。武汉大学图书馆作为全国高

校图书馆中的一份子，它的各类服务及先进技术的改变，都代表着高校图书馆的形象。本研究是基于武汉大学图书馆的馆藏资源的研究，研究结果便可应用于其他各大高校图书馆，甚至扩展应用于公共图书馆和企业图书馆等。

笔者在基于中图法的分类浏览检索系统中，选择了“K86 世界文物考古”类目，将其全选，依著者、题名和主题词三个字段按照卡片格式下载存盘，用 EXCEL 软件简单处理后，共得到馆藏 106 本图书，分别提取主题词和著者字段，得到不同主题词共 91 个，不同著者共 82 个。要说明的是，由于图书独著较多合编较少，我们只选取第一责任者，且排除译者参编的情况。

3.2　著者主题词耦合关系构建

在上述模型中，分析器的主要分析对象就是著者主题词的关系，若要实现用户检索时的著者主题词关联，首先要实现的就是挖掘出馆藏资源中著者和主题词之间的关系。分析器按下列步骤进行：

（1）构建著者主题词共现网络

每一条馆藏资源条目都有著者和主题词项，每个著录项都是由编目人员经主题标引得来，主题词是经过受控词表主题标引的，具有规范、科学的特点，一般三到五个不等，每一个著者都可对应多个主题词，每一个主题词也可对应多个著者。若著者和主题词共同出现一次，则把著者和主题词用一条线连接在一起，这样便可形成多个著者和多个主题词用多条线连接在一起的一张网络图。

首先，利用自编 VBA 程序，将 91 个主题词和 82 位著者一对一共同出现的次数形成一个 91 * 82 的共现矩阵，将共现矩阵导入 UCINET 软件，用 NETDRAW 形成的共现网络如图 3 所示。图中方形代表著者，圆点代表主题词，有箭头的线条代表著者主题词之间有关系。

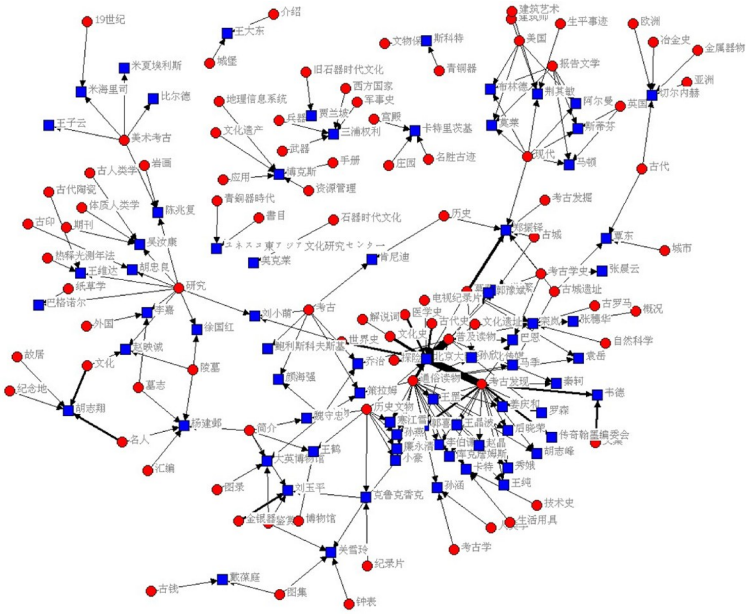


图 3　著者主题词共现网络图

（2）构造共词矩阵及聚类

图 3 中 91 * 82 的矩阵用社会网络分析软件呈现出的网络图已经让人眼花缭乱了，若是更大的学科得到的更大的矩阵，呈现的网络图则会更加复杂。因此需要将主题词进行聚合，这里不进行语义层面的深度聚合，只示例性地用

主题词共现的次数构造共词矩阵,最后用 SPSS 软件进行聚类分析。利用聚类树图,可以将所有 91 个主题词划分成 11 个类团,归纳得到类团大致代表一个研究领域。

(3)构建著者主题领域共现网络

按照构建著者主题词共现网络的原理,同样可以构建著者主题领域共现网络,将 91 个主题词分别归并入各自所属的类团,用 VBA 自编程序得出类团和 82 位著者分别共现的次数,形成 11 * 82 的共现矩阵,最后得到著者主题领域共现网络(见图 4)。

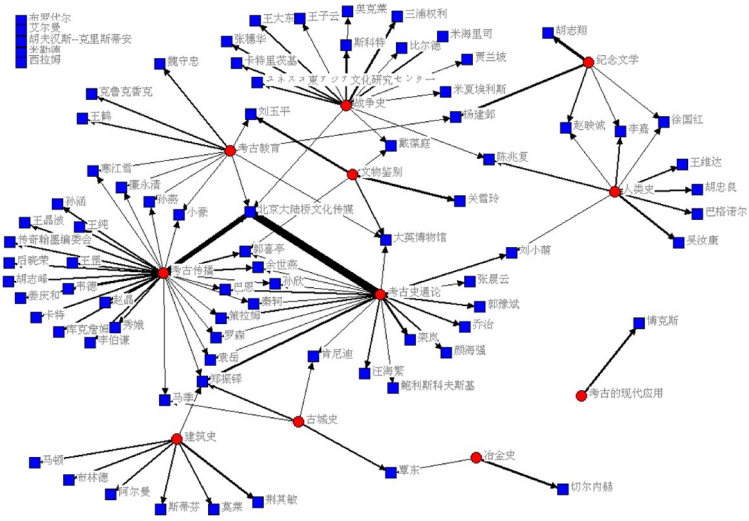


图 4　著者主题领域共现网络图

3.3　著者主题词耦合关系语义化

事实上,上述著者主题领域网络图已经可以表示著者和主题领域间的关系,但他们之间只用带有箭头的粗细不同的线条连接,具体是何种关系,无法呈现给用户,使得用户在进行检索时,只要有连线的节点,都可以被检索出来,形成信息冗余。因此,将著者和主题领域之间的关系形成多个三元组,即语义化,以使检索有章可循。

本例中,著者和主题领域已经呈现出来,也就是说,本体中的主语和宾语都已存在,仅缺的是谓语——著者和主题领域间的关系。经过前人多年的研究和探索,作者和一个研究领域可以存在多种关系,包括作者是本领域的专家,作者是本领域的学者,作者是本领域的潜在力量,作者是本领域的过客,作者是本领域的新手等。本文仅以不同数量的耦合强度来表示作者和领域间的关系,即由强到弱为耦合强度 5—耦合强度 4—耦合强度 3—耦合强度 2—耦合强度 1—耦合强度 0(耦合强度 0 指作者仅唯一属于一个领域),指的是当一个作者隶属于不同的主题领域时,一定会有强弱关系的区别,因此可用不同数量的耦合强度表示。那么这种强弱关系又依何而定呢?它由图中连线的粗细而定,而连线的粗细则由著者和主题领域中所包含的关键词共现次数的总和表示,共现次数越多,线条越粗,代表该著者越善于本领域的研究,比如“北京大陆桥文化传媒”这一节点,最粗的线连接到“考古史通论”,则定义简易三元组 <“北京大陆桥文化传媒”,耦合强度 5,“考古史通论” >,而“北京大陆桥文化传媒”另一条比较粗的线条是连接“考古传播”的,因此可以定义三元组 <“北京大陆桥文化传媒”,耦合强度 4,“考古传播” >,以此类推,可以完整定义有关“北京大陆桥文化传媒”的所有关系。以上推理可凝集成如下数学公式:

$$L_{ij} = c \cdot \sum F(A_i, S_{jk_m})$$

$$R_{ij} = g \cdot L_{ij}$$

其中，K_m 代表第 m 个主题词，S_{jk_m} 代表第 m 个主题词所属的第 j 个领域，A_i 代表第 i 个著者，函数 $F(x,y)$ 代表主题词和著者共现频次，L_{ij} 代表著者主题领域网络图中线条的粗细程度，R_{ij} 代表著者和主题领域的关系，c，g 为参数。

从公式中可以看到线条的粗细和共现频次成正比，而著者和主题领域关系和粗细成正比，所以当出现同一著者属于不同领域时，则可用 R_{ij} 比较得出粗细程度，进而得知其关系强弱。

按照上述逻辑，可以将所有的著者和主题词定义成三元组，即语义化（见表 1）。

表 1　著者主题耦合关系语义化简表

序号	著者	主　题	关　系	简易三元组
1	阿尔曼	建筑史	耦合强度 0	<阿尔曼，耦合强度 0，建筑史>
2	巴恩	考古史通论	耦合强度 4	<巴恩，耦合强度 4，考古史通论>
3	巴恩	考古传播	耦合强度 2	<巴恩，耦合强度 2，考古传播>
			……	
106	郑振铎	考古传播	耦合强度 2	<郑振铎，耦合强度 2，考古传播>
107	郑振铎	考古史通论	耦合强度 5	<郑振铎，耦合强度 5，考古史通论>
108	郑振铎	古城史	耦合强度 4	<郑振铎，耦合强度 4，古城史>

3.4　构建本体

著者和主题关系已用三元组表示，而一个本体库就是由众多三元组构成的，现在可以利用 Protégé 软件实现本体库的构建。在此建立两个类，即作者和领域类，并创建所有实例，接着创建所有关系，最后用关系将各个实例联系起来，得到本体库。Protégé 软件自带的 OntoGraf 绘图工具可以描绘出构建的本体库的面貌，图 5 展示的是世界文物考古的资源本体库的全貌，其中显示了作者类下属的 82 位著者实例和领域类下属的 11 个主题，以及部分著者和主题间的关系。图中紫色实线表示下位类，如 Thing 的下位类有领域类和作者类；蓝色实线表示类包含实例，如领域类包括冶金史、战争史实例等；不同颜色的虚线表示实例间的不同关系，鼠标静止在虚线上会显示特定关系，如覃东等。

3.5　检索流程及检索效果

遗憾的是，笔者所用版本的 Protégé 软件尚不能把 ObjectProperies 中的词语作为检索词，用于查找不同关系的著者和主题词。但事实上，由 Protégé 软件形成的. owl 文件中的代码可以用于检索，基于此，本文设计了基于共现分析的语义信息检索流程（见图 6），并对传统的检索效果和本文设计的检索效果进行对比，可看出共现分析应用于语义检索的优势。

期望达到的检索效果应该是：当用户查询一条记录时，记录中有作者一项，系统除返回该记录外，也返回与该记录的作者有关的其他信息，比如作者所属于或善于或不善于的研究领域，点击该领域可返回本领域中其他作者的信息。根据基于共现分析的语义信息检索流程，首先是用户提出检索请求，这时检索系统会根据用户选择的检索途径将检索词输入搜索引擎，然后利用 SPARQL 语言和匹配算法与基于共现分析的语义信息检索模型中本体库的各种三元组内容进行匹配查询，得到一个初始检索结果集，其中包括文献最初标引时的三元组集，再从中提取作者、题名等字段，即二次结果集，此处针对预期效果只提取作者字段，再利用本体

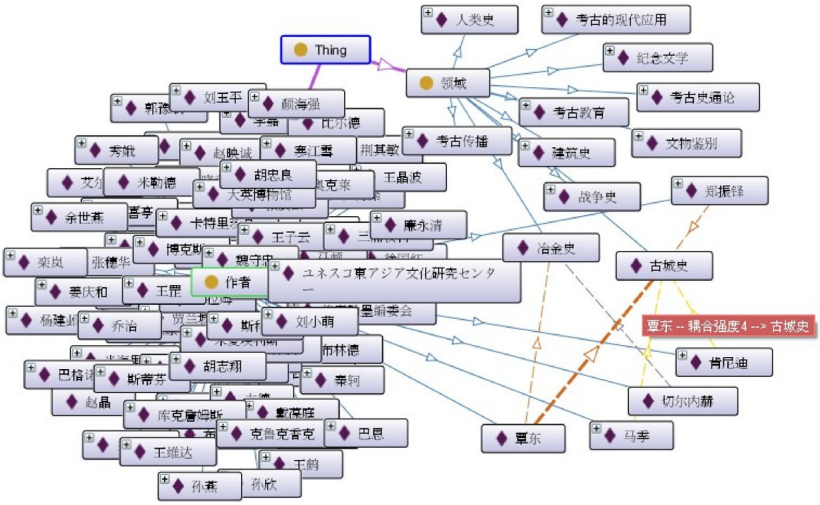


图5　世界文物考古资源本体库整体图

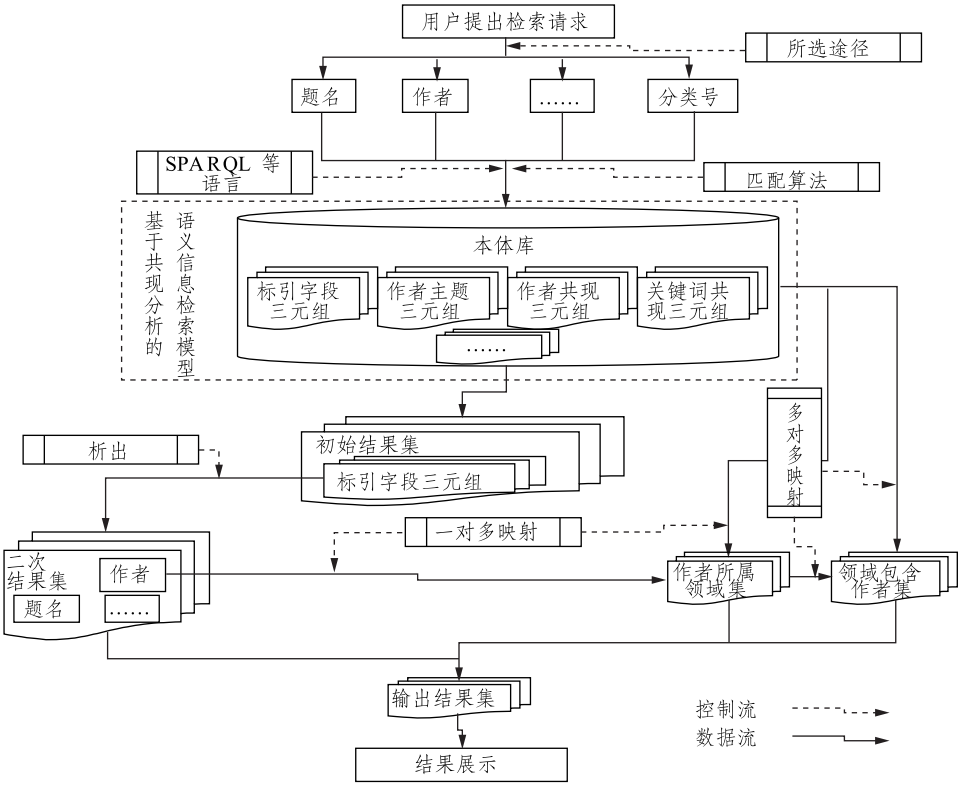
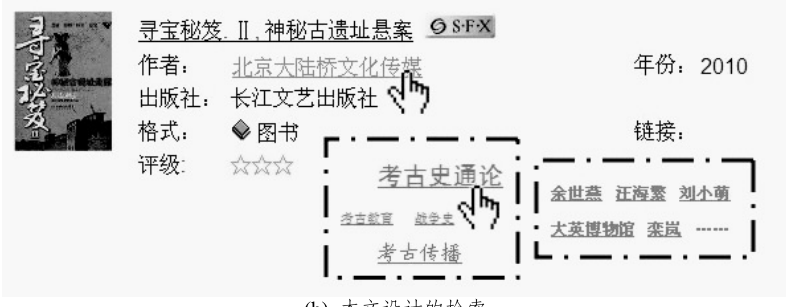


图6　基于共现分析的语义信息检索流程示例图

库中的作者主题三元组,与该作者进行一对多映射,找到该作者所属的领域结果集,而领域结果集中包含的多个领域又可以与本体库中的作者主题三元组进行多对多映射,进而找到某一

领域所包含的作者集。最后把上述所得的二次结果集、作者所属领域集和领域包含作者集合并为输出结果集，为结果可视化展示而用。若二次结果集中提取的是题名、主题词和分类号等字段，此流程也同样适用。

(a) 传统检索



(b) 本文设计的检索

图 7　传统检索效果与本文设计的检索效果对比图

图 7(a/b)展示了传统检索效果与本文设计的检索效果的对比图，假设这样一个场景，用户想要找一本有关世界考古的某一专家的书，但却记不清该专家的姓名，因此该用户仅会凭借自己记忆中的词汇，找到世界考古的所有书籍(见图 7a)，再一页一页地查找，这样又费时又费力，用户体验满意度一定很低。图 7(b)是改进后的原第四条记录的结果，每一条记录的标引

字段都会有更多的信息与其对应,比如当用户鼠标划过第三条记录的作者时,会出现该作者所属于或善于的著作领域,且用云标签显示出来,更易让用户从感观上发现该作者最善于的领域,而继续点击某一领域,会出现本领域内所有作者。这时用户便有可能会发现最初的检索目标,若没有发现,这些记录还可以继续点击下去,直到查到最初检索目标。用户既可快速查找信息,又可轻松了解自己所检索的或偏好的领域或作者的更多信息,图书馆也更好地实现了个性化推送服务。

4 结语

通过总结语义信息检索的研究现状,提出了基于共现分析的语义信息检索模型,最后用实验佐证共现分析融入信息检索的可行性。因此可以得出以下几点结论:

(1)基于共现分析的语义信息检索具有引导性的特点。随着共现分析的引入,本体中概念与概念间的关系更能够被定量化表示出来,定量的数据和信息更有利于发现用户需求,从而更贴心地为用户服务。模型配套的基于共现分析的语义信息检索流程细化了语义检索的方式,对于检索结果的可视化展示提供了结构化数据。

(2)本体应用于语义信息检索的研究还有更宽更广的领域可以探索。本文仅以共现分析为方法探索语义信息检索,也仅以作者关键词为例进行实验和展示,而本体的应用领域之广泛,是无法使用单一的技术和方法来概述的,因此还需要倾注更多的学者和科研人员的才华和智慧来发掘探索更宽广的领域。

本文将共现分析应用于语义信息检索,还有诸多不尽人意之处,比如系统要随时更新这一领域的知识网络,这是对技术人员和系统软硬件的重大挑战,因此还有待更多的研究和探讨。

参考文献:

[1] 黄敏,赖茂生. 语义检索研究综述[J]. 图书情报工作,2008(6). (Huang Min, Lai Maosheng. Survey of semantic search[J]. Library and Information Service, 2008(6).)

[2] 余传明. 基于本体的语义信息检索系统研究[D]. 武汉:武汉大学,2005:91 - 97. (Yu Chuanming. Research on ontology-based semantic information system[D]. Wuhan: Wuhan University, 2005:91 - 97.)

[3] Albanese M, Capasso P, Picariello A, et al. Information retrieval from the web: An interactive paradigm[C]// Candan K S, Elentano A. MIS 2005: LNCS 3665. Berlin, Heidelberg: Springer Verlag, 2005:17 - 32.

[4] Moldovan D I, Mihalcea R. Using WordNet and lexical operators to improve Internet searches[J]. IEEE Internet Computing, 2000(2).

[5] Buscaldi D, Rosso P, Arnal E S. Using the WordNet ontology in the GeoCLEF geographical information retrieval task[J]. Lecture Notes in Computer Science, 2006, 939 - 946.

[6] Dou Yongxiang, Bei Xiaoxian. Ontology-based semantic information retrieval systems in unstructured P2P networks [C]//Proceedings of Wireless Communications Networking and Mobile Computing 2008 WiCOM 08 4th International Conference, 2008:1 - 4.

[7] 刘宏哲,须德. 基于本体的语义相似度和相关度计算研究综述[J]. 计算机科学,2012(2). (Liu Hongzhe, Xu De. Ontology based semantic similarity and relatedness measures review[J]. Computer Science, 2012(2).)

[8] 俞扬信. 基于语义相似度的信息检索研究[J]. 情报杂志,2009(9). (Yu Yangxin. Study of information retrieval based on semantic similarity[J]. Journal of Intelligence, 2009(9).)

[9] 董慧,余传明,姜赢,等. 基于本体的数字图书馆检索模型研究(Ⅱ)——语义信息的提取[J]. 情报学报,2006(8). (Dong Hui, Yu Chuanming, Jiang Ying, et al. Research on the ontology-based retrieval model of digital library(II)—Semantic information acquisition[J]. Journal of the China Society for Scientific and Technical Information, 2006(8).)

[10] Castells P, FernáNdez M, Vallet D. An adaptation of the vector-space model for ontology-based information retrieval[J]. IEEE Transactions on Knowledge and Data Engineering, 2007:161 - 272.

[11] Kwong L W, Ng Y K. Performing binary-categorization on multiple-record web documents using information retrieval models and application[J]. Ontologies World Wide Web Archive, 2003(6).
[12] 荆涛. 面向领域网页的语义标注若干问题研究[D]. 长春:吉林大学, 2011:5-6. (Jing Tao. Research on semantic annotation for domain-specific web pages[D]. Changchun: Jilin University, 2011:5-6.)
[13] 熊荣东. 结合 WordNet 的领域语义标注研究[D]. 重庆:重庆大学, 2011:3-4. (Xiong Rongdong. Research on the semantic annotation of domain with WordNet[D]. Chongqing: Chongqing University, 2011:3-4.)
[14] 刘海学. 基于语义标注的元数据自动构建及其相关技术研究[D]. 上海:华东师范大学, 2010:6-9. (Liu Haixue. Research on the issues of semantic annotation based automatic metadata construction[D]. Shanghai: East China Normal University, 2010:6-9.)
[15] Barnaghi P, Wei Wang, Kurian J. Semantic association analysis in ontology-based information retrieval[J]. Handbook of Research on Digital Libraries Design Development and Impact, 2009:131-141.
[16] Aleman-Meza B, Halaschek C, Arpinar I B. Context-aware semantic association ranking[C]// Cruz I F, Kashyap V, Decker S, et al. Proceeding of the 1st international Workshop on Semantic Web and Databases. Berlin, Germany: Humboldt-Universität, 2003:33-50.
[17] Halaschek C, Aleman-Meza B, Arpinar I B. Discovering and ranking semantic associations over a large RDF metabase[C]// Michael C, Serge A, Lockemann P C. Proc. of the 30th VLDB Conference. Toronto, Canada: Elsevier Science & Technology Books, 2004:1317-1320.
[18] 王曰芬, 宋爽, 苗露. 共现分析在知识服务中的应用研究[J]. 现代图书情报技术, 2006(4). (Wang Yuefen, Song Shuang, Miao Lu. Application study of co-occurrence analysis in knowledge service[J]. New Technology of Library and Information Service, 2006(4).)
[19] 朱芊. 全国中文机读书目主题标引格式问题分析[J]. 中图图书馆学报, 2002(1). (Zhu Qian. An analysis of format problems in subject indexing in Chinese MARC[J]. Journal of Library Science in China, 2002(1).)

邱均平　武汉大学信息管理学院教授、博导。
通讯地址:湖北武汉大学信息管理学院。邮编:430072。
楼　雯　武汉大学信息管理学院博士生,通讯地址同上。

(收稿日期:2012-05-26)