面向非相关文献的知识关联发现系统设计与实现*

刘爱琴　安　婷

摘　要　面向非相关文献的知识关联能够促进新知识的产生,为科学研究寻找新的线索提供了一种便捷、有效的辅助手段。本文首先设计了面向非相关文献的知识关联发现系统,该系统以《中国分类主题词表》为主题词受控词表,通过对文献摘要的中文分词处理,提取主题词并标引,提取文档特征矩阵并利用计量分析技术和聚类技术分析文献间特征的相似、相异水平,从词的粒度层面对非相关文献之间的关系进行挖掘,揭示非相关文献的知识关联。其次,基于该系统为用户精确匹配相关的知识库,采用 TOP-K 算法反馈与用户相关的文献集,为用户提供满意度更高的知识发现及相关扩展服务。图 5。参考文献 29。
关键词　非相关文献　知识关联　中国分类主题词表　计量分析技术　知识发现

Design and Implementation of Knowledge Linkage Retrieval System for Non-Interactive Literature

Liu Aiqin　An Ting

Abstract: Knowledge linkage for non-interactive literature can promote the generation of new knowledge and provide a convenient and effective supplementary means for new clues of scientific research. This paper first designs knowledge linkage retrieval system for non-interactive literature, which uses the *Chinese Classification Thesaurus* as the subject word controlled vocabulary, with the Chinese word segmentation of the document abstract and the extraction of the subject words for indexing, extracting the document feature matrix and using the econometric analysis technique and clustering technique to analyze the similarity and difference level between documents, the relationship between non-interactive documents being mined from a more granular level, revealing knowledge linkage of non-interactive literature. Then, based on relevant knowledge base used to accurately match for users, this paper adopts TOP-K algorithm to feed back users-related document collections, so as to provide users with higher satisfaction knowledge retrieval and related extension services. 5 figs. 29 refs.
Keywords: Non-interactive Literature; Knowledge Linkage; China Classification Thesaurus; Metrological Analysis Technique; Knowledge Retrieval

1　面向非相关文献的知识关联研究现状

1986 年,Swanson 教授最先提出了基于非相关文献的知识发现,并提出将两个逻辑上有关联但相互独立的知识模块放一起研究,可能会有新的知识被发现,他将这部分文献称为“非交互文献”[1],后续国内相关研究将其称作“非相关文献”[2]。面向非相关文献的知识发现是一种利用计算机技术及相关算法来辅助学者从非相关文献中发现隐含关联的系统[3]。张云秋和冷伏海指出非相关文献知识发现有三个理论基础:检索理论、文献计量学理论和逻辑学理论[4]。黄水清等先后采用闭合式(2007)[5]、开放式(2008)[6]非相关知识发现方法,验证了 Swanson 的知识发现理论和方法在中文文献中同样适用。李勇、冷伏海等人在此基础上提出了基于非相关文献的三阶知识发现方法,通过对三个文献集的分析发

* 本文系山西大学人文社会科学科研基金项目“基于跨界思维的信息咨询新业态研究”(项目编号:115546003)的研究成果之一。

现潜藏的关联,挖掘不同主题之间的关系[7]。李勇基于闭合式非相关文献三阶知识发现方法的原理,发现三阶知识发现方法可以应用到研究活动中[8]。程趁娜和兰小筠整理现有的知识发现系统并进行研究,发现大部分是依据开放式的知识发现方法设计的,整个过程由开始词 A 出发,系统利用关联挖掘技术找出与其有直接关联的中间词 B,然后采用相同的方法关联出与每个中间词相关的目标词 C,最后对目标词进行排序,以供学者筛选[9]。郭卫宁与司莉以近几年国内发表的相关文献 236 篇作为分析数据,采用定量分析与定性分析相结合的方法,从时间、载文期刊、著者、机构、主题等 5 个角度分析了我国非相关文献知识发现的现状与特点[10]。闵波等利用复杂网络对文献集进行文本挖掘,探讨了关联知识的图结构组织对文献知识发现的作用,揭示了非相关文献中的隐含知识[11]。田瑞强、姚长青等人对关联文献知识发现的发展历程、技术变迁及发展方向进行了深入分析,并将其发展分为两个阶段:关联文献知识发现分析以标题和摘要字段为主,采用词频过滤技术;以叙词表和医学主题词表为主,分析单元从单词短语扩展到概念、语义,同时筛选采用语义过滤方法[12]。

由上述描述可知,当前学者主要在理论和技术方面探索非相关知识发现的方法,没有构建完整的非相关知识发现体系,本文将关联文献知识发现的技术应用于非相关文献领域,揭示非相关文献中的隐含知识,构建合理的非相关知识发现系统的完整体系,以帮助学者理解学科的发展与外延,进行有针对性的创新研究。

2 面向非相关文献的知识关联的研究脉络

首先,将中国分类主题词表作为受控词表,利用主题词标引与用户个性化推荐中的标签系统相类似的特征,对文献进行标引。其次,文献用主题词作为标签,依据主题词的层次结构以及主题词间的语义关联程度或相似程度,构建文献的隶属、相关结构,此层次结构具有文献、文献主题词列表、文献主题词的重要程度标注[13]。最后,借助知识相关算法,挖掘文献之间知识相关联的程度,并实现知识关联的有效揭示。

2.1 理论依据

《中国分类主题词表》是人类知识的体系结构和语义网络,有效且形象化地展示了知识载体间丰富多样的关系[14]。它集合了分类法和主题法的优点,更加规范、科学,是一种层次较为丰富、等级结构严谨的检索语言,也是一种检索工具[15]。其创建的知识体系图谱,既描述了知识的等级体系、主题的语义关联,还描述了这两者间的对应联系,在一定程度上创造了知识。将《中国分类主题词表》中的词语作为检索词,重新编码、组织,可以形成一定结构的主题词数据库[16]。通过输入的主题词,结合电子化的词表,浏览分类的词等级树,不仅可以检索到与其匹配的文献信息,而且能够查询到特定类别和子类的文献[17]。

2.2 研究方法

邱均平、余凡指出计量分析的馆藏资源语义化方法在语义关系深度、计算机自动处理程度、可重复性和实用性方面比基于元数据和领域本体的馆藏资源语义化方法更好[18]。计量分析的主要方法有共词分析、耦合分析和共引分析等,其中,共被引矩阵和耦合矩阵是基于引用关系的共现矩阵[19]。这两种分析方法是基于文献间的引用情况,利用期刊引文网络探讨学科研究的交叉关系,或者是利用关联规则挖掘、文本挖掘等现代数据挖掘技术探讨学科间的相关性和交叉知识,以达到分析文献之间关系的目的,但无法实现对文献内部隐含关联的有效揭示[20]。本文首先采用共词分析方法计算两个主题词共同出现的次数,以此为依据来确定两者之间的关系强度。随后,基于从文献中提取出来的主题词,获取高频主题词的共词矩阵,从更细的粒度层面利用 Kmeans 聚类的方法对文献之间的关联关系和关联强度进行深入挖掘。

刘爱琴等提出通过对论文的题目或者关键词进行关联发现的语义化知识发现方法不足以清晰、准确地反映文章的主旨,而从摘要中提取主

题词则更具有代表性[21]。因此,本文利用《中国分类主题词表》作为主题词受控词表,对文献摘要进行中文分词处理并提取主题词进行标引,提取出文档特征矩阵并利用计量分析技术和聚类技术分析文献间特征的相似、相异水平,对非相关文献之间的关系从更细的粒度层面进行挖掘,揭示非相关文献之间的知识关联。基于此,设计了面向非相关文献的知识关联发现系统,并基于该系统为用户精确匹配相关的知识库,采用 TOP-K 算法反馈与用户相关的文献集,为用户提供满意度更高的知识发现及相关扩展服务。

3 面向非相关文献的知识关联发现系统的设计与实现

3.1 系统设计

依据非相关文献的知识关联特征,本文设计的知识关联系统从非相关文献的关联、用户检索词与主题词的关联、知识的匹配与反馈三方面展开。非相关文献的关联是对提取的主题词表进行相关的关联,以及在知识库中匹配与用户输入的检索词相关联的资源,并通过 TOP-K 算法反馈给用户相关的文献集,详见图 1。

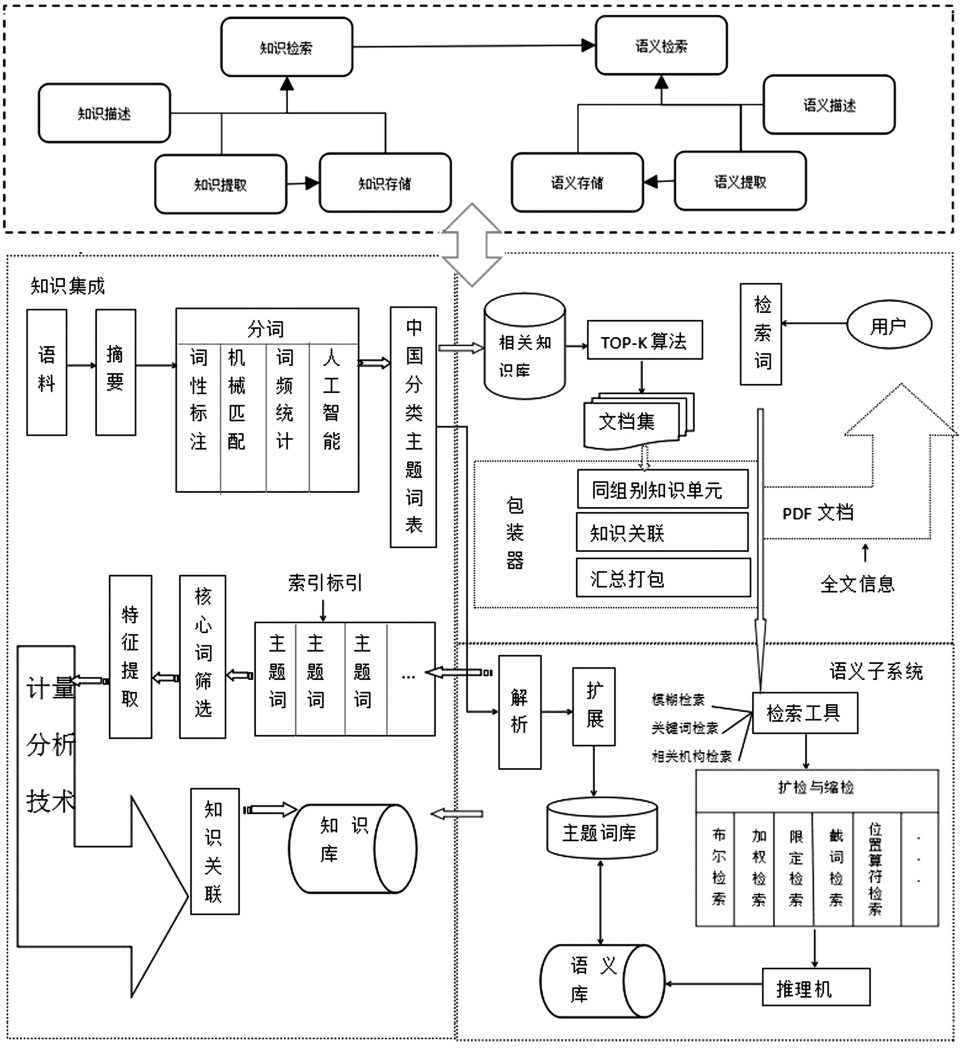

图 1 面向非相关文献的知识关联发现系统的设计结构

首先,对资源进行知识描述并提取其特征,建立特征标引并将其存储于知识库中以待检索访问。其次,利用语义子系统对用户的检索词进行语义分析与提取。最后,知识库与语义子系统提取的主题词进行关联匹配并返回结果。该体系包括三个层次:

(一)知识集成层

先将语料按照计算机可识别的结构整理存储入知识库:系统接收到文献语料后,首先提取文献的摘要进行中文分词处理,将文献的摘要切分成逐个的词语,并将切分好的词输入到语义子系统中。然后通过语义系统的解析并使用经过人工规范化处理的《中国分类主题词表》作为主题词受控词表,将解析后经过规范的主题词表返回系统,并且将这些主题词扩展到语义子系统的主题词库。最后收集整理子系统获得规范化的主题词表,提取核心主题词,并提取文献特征,利用计量分析技术以及聚类技术分析文献间特征的相似、相异水平,揭示出文献之间的潜在关联并传输到存储子系统。存储子系统接收到数据后进行索引组织和存储操作,将数据整理、汇总存储到知识库中,完成文档、语义以及知识的索引工作。

(二)用户检索层

用户登录用户界面上的检索栏,借助关键词检索、模糊检索、相关机构检索等检索方式输入目标信息,进行查询检索。获取用户需求概念后借助语义子系统的推理机依次进行特征描述、特征提取、概念扩展等处理,进而调用语义库解析出的用户需求所对应的相关语义主题词。

(三)知识发现层

语义子系统可以高效地理解和处理数据间的语义关系,使知识发现的过程在语义查询的能力支持下进行,通过已有的语义关系发现相关的资源[22]。再利用 TOP-K 算法,将关联度最大的前 K 条记录组成一个文档集返回给用户,最终借助文件打包器对聚合的文档进行汇总打包,实现知识获取的智能可视化,将目标资源的全文信息以 PDF 文档格式条理化、可视化地呈献给用户。用户可继续进行浏览、检索或其他操作。并且,根据“二八原则”,检索系统要对热点检索结果进行缓存操作,为之后的检索行为提高效率,提高系统的检索性能和可用性。利用可视化的知识描述来取代先前简单的检索目录,用户即可获得与其查询内容相匹配的文献数字资源的全文信息 PDF 汇编文档,以期给用户带来更加智能化、便利化、柔性化的文献资源发现服务。

3.2 系统实现

3.2.1 数据收集

“数据挖掘”属于计算机科学学科下的重要概念,“管理”属于管理学学科下面的重要概念。在此我们提出这两个不同的学科之间存在相互融合的观点,并通过下面的实现过程进行验证。从 Web of Science 的“中国科学引文数据库 SM”中导出关键词为“数据挖掘”和“管理”的文献数据,“数据挖掘”数据 5000 条和“管理”数据 5000 条。

3.2.2 中文分词

中文文本挖掘技术,都是建立在“词”的基础上,但汉语文字中没有类似英文文本中独特的切分标志——空格。在汉语中,每个词的长短不一,人在理解语义过程中,会下意识地将词从句子中突显出来分别进行理解,词的组合构成句子,最终理解一句话的语义。对于计算机来说,理解语义的能力还不够完善,需要有一种其他的方式对中文进行分词处理。本文将机械匹配、词频统计、人工智能的中文分词算法进行糅合,精确地将连续的字切分成符合人理解的词语,在对词进行切分的同时还对其打上词性标签。

3.2.2.1 机械匹配的中文分词算法

机械匹配分词技术采用人工创建词典,该词典中包含尽量多的词语以实现最大匹配。本文对一系列字符串使用预先设计好的规则——最大匹配原则匹配字典中的词,机械地将字符串分割成子串,获得最终的词语序列。最大匹配原则使用简单,切词速度快,但由于词语二义性问题,切词的效果相对较差。

3.2.2.2 词频统计的中文分词算法

机器学习方法允许海量电子资源在没有词典的情况下也可以进行分词[23]。词频统计的中文分词算法以词为基础,词是较稳定的文字组合。如果汉字之间的共现频率较高,频繁地组合在一起使用,便可将其整体作为一个语义单元进行处理。本文将词频统计结果作为是否最终将这些汉字划分到一个词语(即作为分词)中的依据。

词频统计的中文分词算法基于某一个词的出现概率只由当前词语所在的上下文语义或者其他特征决定,而与其他因素都无关的原理。N-gram 是一种统计模型的分词手段,以词共现的概率为依据,计算出整个句子合理出现的分数值,确定一个满意分数值作为分界值,进行最终分词[24]。由于如此进行的分词是基于统计方式,分词统计模型需要大量文本才可以获得较高的准确率,这就要求更大的语料库和更长的切词计算时间。

3.2.2.3 神经网络分词算法

神经网络分词算法[25]是扩展到文本分析领域的特殊应用。本文使用神经网络分词算法将汉字组成复杂的神经网络,由句子的输入构建动态的链接表,经过计算获得最终有效的分词结果输出。该方法要求大数据集和较高的计算密集度。由于神经网络分词算法单机处理时间较长,一般作为分布式服务进行实施,且技术要求非常高。

3.2.3 自动赋词标引方法

自动赋词标引方法,可以自动地对切分好的文本标注主题词及词性,本文采用比较常见的自动赋词标引方法 tf-idf[26]。如果某个词在全体文档语料集中出现的概率比较低而在当前文档中呈现的概率比较高,说明此词在当前个别文档中比较重要,即可作为当前文档的主题词。使用该方法,可以很简便地获得一篇文章或一段话的主题词。

为了实现知识检索“高召回率、高检全率”的目标,本文要求语料切词粒度尽可能地细化,实现词语的全部切分;为了更贴近检索意图,本文将多种方法融合并借助词典,获得折中的切词效果。

3.2.4 特征提取与知识关联提取

本文对经过中文分词后切分到的词语,使用语义系统进行标准化操作,筛选出系统可识别并具有代表性的词语作为最终特征,即进行自动赋词标引。而其他一些在中文分词表中没有出现的词也同样可以作为主题词,这就涉及到自动增词标引,在标引词语的同时通过扩展重要词来扩展字典,赋予系统相对成长性。由于词语间都会有语义的相似性,一些词可以用另一些词替换来表示相同的含义。而主题词一般比较规范,本文将一部分语义相同的词语转化成主题词,既缩减了文档向量空间的大小,降低了计算复杂性,同时也使得知识关联系统更加精炼,提高知识利用的效率。

本文使用语义模型,提取出文档特征矩阵,使用特征矩阵进行知识关联隶属度从属关系挖掘,获得知识关联模型。首先,本文将上文所述的一

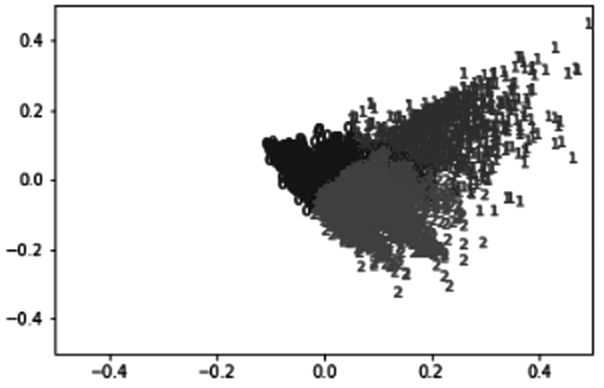

图 2 KMeans 分类结果

万条数据进行 KMeans 聚类处理,将数据分成三类,可以将语义相似度较高的文档聚类到一起。从感性上理解各文档之间的区分度和关联程度,分类较多则会导致图形较为繁杂,展示不够直观,本文简单地将其分为三类作为直观的反映图。其次,为了实现多维度数据的可视化,将高维度数据经过 PCA 算法进行降维处理,获得二维数据,并进行可视化。

聚类处理的结果如图 2 所示,所划分的三类数据并未相互分离,而是紧紧贴合,界限模糊,验证了之前的学科之间存在相互融合的观点[27]。通过 KMeans 聚类,将数据分成三类,同时在使用的时候可以划分为更多的类,用以缩小所寻找知识的范围。使用 LDA 主题分析,并且可视化,得到图 3 所示结果。

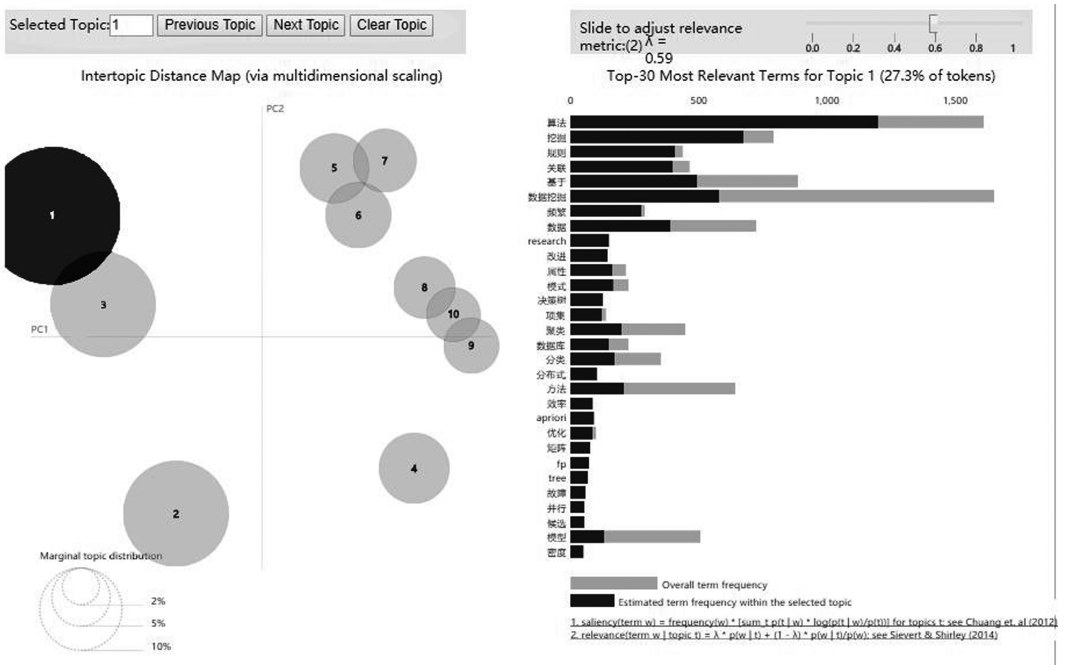

图 3　LDA 主题分析结果

图 3 左侧圆表示各主题,圆的大小表示主题重要程度,突出显示了重要程度最大的主题(最左侧圆),右侧统计了出现频率前 30 的词语以及其词频。本文的十个主题中主题 1 比其他主题的重要程度大,其他各主题出现的词频为前 30 的词语较少且频率较低,所以本文采用不同颜色用于区分。左图中最重要的主题为主题 1,右图展示了主题 1 中出现的词语以及词频,主题 1 中出现的各个词语的频率代表了其占有的比重。利用该结果,可以将同一主题的相关知识进行整合,并通过主题词共现方式提取知识关联。

曹祺等采用深度学习的 Doc2Vec 模型,随机挑选专利用来研究专利相似度检测问题,并和传统的相似度检测模型进行对比研究,结果表明该方法对分析人员的专利领域知识要求较低,不需要对专利数据进行基于专利领域知识的数据清洗,提升了研究效率[28]。唐明、朱磊等人将 Word2Vec 模型与聚类算法结合形成 Doc2Vec 模型,能有效地表达文档信息,并指出该模型用于中文文档分类的可行性[29]。基于上述研究,本文采用 Doc2Vec 模型,将语义蕴含在特征创建过程中,挖掘、获得特征矩阵,并利用不同方法验证各文档、各主题之间的关系。通过语义挖掘获得各主题知识之间的关联,从而形象化地描绘出各主题

间的关联，如图4展示的即是各文档之间的关联，将文档之间的关联通过线条连接予以表示。

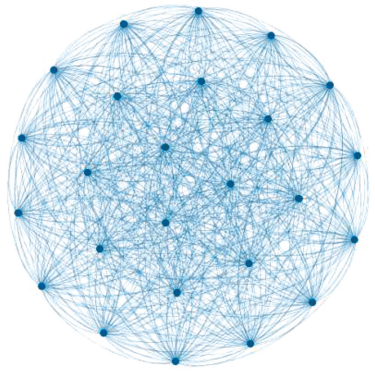

图4　不同主题知识间的关联

借助语义挖掘模型，进行各文档之间相似程度的计算，可以获得相似矩阵。通过文档语义相似矩阵，构建简单的语义挖掘系统。将随机抽取的主题词“算法”“医疗”“规则”“数据挖掘”“效率”“评价”“模型”“知识”“仿真”“异常”“安全”“预测”“组合”“评估”“治疗”“基因”“融合”“推荐”“空间”导入模型，输出一个term-document矩阵A，如图5中上半部分是经过降维处理的结果。矩阵中的每一行A_i代表着词典里的一个词，矩阵的每一列A_j代表着语料里的一篇文献，A_{ij}表示词w_i在文献D_j中的重要程度（类似于因子分析），可以提取行向量作该词的语义向量，列向量为文档向量。本文所统计的w_i词频是在收集的语料文献D_j的摘要中统计的。

将主题词和文献都转换成Embedding的词向量与文档向量，通过计算余弦相似度来匹配与主题词相似度较高的文档，如图5下半部分所示，最后5行数字由两部分组成，“:”右半部分表示分档编号，左半部分表示相似度。每一行表示所输入的主题词与某一编号的文档所具有的相似度，根据相似度进行检索匹配保证了检索结果的质量。

```
[-1.19080395e-01 -3.06584179e-01  1.19743630e-01 -3.33216667e-01
  2.90593177e-01 -2.16276184e-01 -9.98405740e-02  1.29800647e-01
 -8.17458518e-03  8.31889063e-02 -1.66609481e-01 -1.78287640e-01
 -5.62797070e-01 -2.30598520e-03 -2.45640010e-01 -9.66278277e-03
  1.91498876e-01 -2.45888829e-01 -2.11732492e-01 -4.10041124e-01
  3.65106076e-01  1.60516530e-01  1.36672720e-01 -1.74975455e-01
  2.96586245e-01  3.06278199e-01 -6.50808513e-02 -1.49308100e-01
  7.30900839e-02  7.86485970e-02 -2.11133674e-01 -4.22205627e-01
 -5.35301119e-02  1.04570091e-01 -1.87790900e-01 -3.20476472e-01
  1.04467385e-01  9.75991264e-02 -2.28692248e-01  1.23806857e-02
  1.40828416e-01 -2.56550670e-01  2.41673775e-02  6.03541434e-02
  2.06319243e-01 -8.15978721e-02 -3.21221501e-01  3.22411925e-01
  5.19802049e-02 -1.41537651e-01  3.33032429e-01 -7.38634542e-02
 -1.94506824e-01  1.82982877e-01 -1.27192046e-02  2.42773909e-02
  9.64538939e-03 -8.25571537e-01  1.05036728e-01 -2.10355878e-01
 -4.50166911e-01 -3.66456479e-01 -1.54161558e-01 -3.44834387e-01
  2.07846418e-01  1.87671036e-01  1.50786862e-01  3.64745036e-02
  1.99071929e-01 -2.23290578e-01 -3.29352431e-02 -5.72083779e-02
 -1.88749924e-01  4.39724997e-02 -3.85808051e-01 -3.85625869e-01
 -9.95771736e-02 -6.92145303e-02  9.53248069e-02  3.68963093e-01
 -1.02174856e-01 -1.97986886e-01 -1.68264374e-01 -5.81303835e-01
 -1.96693242e-01 -2.64474392e-01  2.62887299e-01 -4.06423301e-01
 -9.23980027e-02  2.67497420e-01 -3.58613878e-01  2.57072132e-02
 -8.07533190e-02 -5.75208454e-04  3.65719944e-02 -1.85193010e-02
  2.66937524e-01  4.09068987e-02  1.00731544e-01 -2.00584203e-01]
0.494681:8938
0.341949:8406
0.316444:1978
0.307366:7002
0.303170:3851
```

图5　不同主题词知识关联挖掘结果

3.2.5　知识发现和推送

在知识检索的过程中，一般的用户会直接输入未经过规范的检索词，但这些不规范的检索词在检索系统中常常无法准确地匹配到用户所需

的信息资源。为了提高检索质量并为用户提供灵活的检索服务,本文在检索后端通过语义子系统调用语义库识别用户输入的检索词,并将其转换为知识库可以识别、匹配的主题词,然后将转化来的主题词输入系统进行知识匹配,并且把经过知识关联的资源共同展示给用户。

在知识关联的过程中,使用计量分析、聚类分析等方法,将显性知识以及难以发现的隐性知识一并挖掘出来后进行提炼,并通过有效的组织方式存储起来。在用户提出知识需求时,本文从预先建立好的知识库中提取方案,再借助语义子系统的分析,获得切实可行的解决方案,并在方案实施过程中及时跟踪评价实施效果,最终高效解决用户的问题。

在知识检索结果推送的过程中,我们会得到数量极多的符合条件的记录,一般的知识检索系统会返馈给用户关联度最强的前 n 条记录,例如中国知网,每次检索可以返回 10、20、50 条记录等。为了精准高效地完成此任务,本文采用 Top-K 算法,根据关联度快速排序并获得前 K 条记录,收集用户浏览结果,记录浏览次数,从中挖掘出更深度的知识关联,用于提高用户知识检索的准确度和效率。

4 结论

本文使用《中国分类主题词表》作为主题词受控词表,首先对文献摘要进行分词处理,选取文献的主题词并进行标引,提取文献核心主题词及文献特征,利用语义挖掘模型的分析技术对文献进行语义分析,并利用计量分析技术和聚类技术分析文献间特征的相似、相异水平,从更细的粒度层面对非相关文献之间的关系进行挖掘,揭示非相关文献之间的知识关联;其次,将语义关联的文献进行组织并存储;最后,集成知识库为知识关联系统提供知识支持,为系统便捷高效地进行知识服务提供保障,通过语义子系统对用户检索词进行分析与规范,经过规范化的检索词与知识库的知识特征进行匹配,采用 TOP-K 算法为用户反馈精确、关联度高的前 K 条记录,实现知识服务的高质量化。

在下一步研究中,可以考虑通过收集、分析用户信息来提取出用户特征,然后将用户特征与文献资源特征进行相关匹配,实现知识的精准推送。

参考文献

1 Swanson DR. Undiscovered public knowledge [J].Library Quarterly, 1986, 56:103-118.

2 张云秋,冷伏海.非相关文献知识发现的理论基础研究[J].中国图书馆学报,2009(4):25-30.

3 程趁娜.图书情报学的专门研究方法——基于非相关文献的知识发现方法[J].内蒙古科技与经济,2012(14):90-91,93.

4 黄水清,等.闭合式非相关文献知识发现方法在中文文献中的验证[J].中国图书馆学报,2007(5):83-87.

5 黄水清,等.开放式非相关文献知识发现方法在中文文献中的验证[J].情报理论与实践,2008(2):246-250.

6 李勇.基于非相关文献的 3 阶知识发现——闭合式方法探讨[J].医学信息学杂志,2014(5):56-60.

7 李勇,等.基于非相关文献的三阶知识发现方法探讨[J].中国图书馆学报,2011(4):21-26,69.

8 程趁娜,兰小筠.基于非相关文献知识发现系统的算法[J].情报科学,2009(12):1824-1828.

9 郭卫宁,司莉.国内非相关文献知识发现研究综述[J].图书情报知识,2013(6):97-105.

10 闵波,等.构建基于文献信息网络的知识发现系统应用模型的设想[J].中华医学图书情报杂志,2016(5):30-34.

11 田瑞强,等.关联文献的知识发现与创新研究进展[J].情报理论与实践,2013(8):117-123.

12 赵萍萍.从传记文献的主题标引谈文献主题

标引规则[J].现代情报,2004(6):137-139.
13 曹玉强.《中国分类主题词表》(第2版)各科教材与教学参考书的标引[J].图书馆建设,2009(7):14-15,46.
14 吴雯娜.我国叙词表的编制历史与发展模式[J].情报理论与实践,2018(6):39-44.
15 陈树年,刘惠敏.从网络信息组织看《中国分类主题词表》[J].国家图书馆学刊,2006(2):21-27.
16 余丰民.国内主题词表研究脉络初探[J].情报科学,2014(5):12-17.
17 王曰芬,等.共现分析在文本知识挖掘中的应用研究[J].中国图书馆学报,2007(2):59-64.
18 邱均平,余凡.基于计量分析的馆藏资源语义化理论研究[J].中国图书馆学报,2012(4):71-78.
19 邱均平,楼雯.基于共现分析的语义信息检索研究[J].中国图书馆学报,2012(6):89-99.
20 李育嫦.文献检索中提高查全率与查准率的方法探讨[J].图书馆学研究,2002(11):92-93,95.
21 刘爱琴,等.基于爬虫技术的关键词关联推荐算法优化与实现[J].情报理论与实践,2018(4):134-138.
22 李楠,张学福.基于关联数据的知识发现模型研究[J].图书馆学研究,2013(1):73-77,67.
23 黄昌宁,赵海.中文分词十年回顾[J].中文信息学报,2007(3):8-19.
24 尹锋.基于神经网络的汉语自动分词系统的设计与分析[J].情报学报,1998(1):41-50.
25 陈白雪,宋培彦.基于用户自然标注的TF-IDF辅助标引算法及实证研究[J].图书情报工作,2018(1):132-139.
26 祝清松,冷伏海.非相关文献知识发现的融合评价方法构建研究[J].情报理论与实践,2013(7):106-109,105.
27 李纲,巴志超.共词分析过程中的若干问题研究[J].中国图书馆学报,2017(4):93-113.
28 曹祺,等.基于Doc2Vec的专利文件相似度检测方法的对比研究[J].图书情报工作,2018(13):74-81.
29 唐明,等.基于Word2Vec的一种文档向量表示[J].计算机科学,2016(6):214-217,269.

(刘爱琴　副教授　山西大学经济与管理学院,安　婷　山西大学经济与管理学院物流工程专业2017级硕士研究生)

收稿日期:2018-11-26