国内推荐引擎学术研究知识图谱分析

奉国和,梁晓婷

(华南师范大学 经济管理学院,广东 广州,510006)

摘　要:利用SPSS软件,对CNKI数据库中2005~2010年间的国内推荐引擎领域论文进行共词分析,并借助聚类分析和因子分析,研究关键词之间的关系,绘制该领域的战略坐标图,探讨了国内推荐引擎领域的研究现状与热点。

关键词:推荐引擎;推荐系统;个性化;知识图谱
中图分类号:G350　　**文献标识码:**A　　**文章编号:**1007-7634(2012)01-144-05

Analysis on Knowledge Mapping of Academic Research on Recommender Engine in China

FENG Guo-he,LIANG Xiao-ting

(School of Economics & Management,South China Normal University,Guangzhou 510006,*China)*

Abstract: By the use of SPSS software,this paper gives a co-word analysis of the articles on recommender engine in China issued in the periodicals in recent 6 years and collected by CNKI database.The relationships among keywords were researched with the help of the factor analysis method and the hierarchical clustering method.Then the strategic coordinates figure was draw.Finally,it explored the status and hot spot of the researches on recommender engine in China.

Keywords:recommender engine;recommender system;personalization;konwledge mapping

互联网的迅猛发展给网民带来丰富资源的同时也带来了信息超载问题。搜索引擎的本质是帮助用户过滤信息,满足大众的信息需求,但没有个性化服务。相对于搜索引擎而言,推荐系统不仅能满足个性化需求而且能解决信息过载问题。

在学术界,推荐系统在电子商务、网络经济学和人类社会学等领域一直保持很高的研究热度并逐渐成为一门独立的学科。近几年来,国际学术界针对计算机网络信息整合的推荐相关的研究大量出现。推荐系统结合社会网络和语义网络的研究,面向互联网发展中出现的新问题和新技术需求,具有广泛的研究和应用前景[1]。

1 材料来源与研究方法

1.1 材料来源

笔者分别用关键字“推荐引擎”和“推荐系统”在CNKI“中国学术期刊网络出版总库”中的进行篇名、主题、关键词检索,检索日期限定为2005年到2010年,共得到1667篇,为了提高研究准确度经查重取掉无关键词和相关度不大的文献后得到有效文献703篇。

收稿日期:2011-09-05

作者简介:奉国和(1971-),男,湖南永州人,副教授,博士,主要从事数字图书馆、数据挖掘研究.

1.2 研究工具

目前构建知识图谱应用较多的是一些社会网络分析软件[2]，如Pajek、UCINET、Netdraw、Citespace、HistCite等。而共词分析一般都采用自编的软件进行统计，如CnkiRef[3]，Bicomb等。邱均平[4]等利用CnkiRef对我国图书馆学研究做了实证分析。刘向阳[5]等利用Bicomb对我国搜索引擎领域做了研究。本文采用的是中国医科大学崔雷老师开发的Bicomb共词分析软件和统计学分析软件Spss18.0。

1.3 词频分析与共词分析

词频分析法是利用能够揭示文献核心内容的关键词或主题词在某一研究领域中出现的频次高低来确定该领域研究热点和发展动向的文献计量方法[6]。如果某一关键词或主题词在其所在领域的文献中反复出现，则可反映出该词所表征的研究主题是该领域的研究热点。

共词分析法是内容分析法的一种，它通过对一组词两两统计它们在同一文献中出现的次数，并以此为基础对这些词进行聚类分析，从而反映出这些词之间的亲疏关系，进而分析这些词所代表的学科研究热点、主题的结构变化和转移趋势。共词分析法是对当前发表文献的直接统计，所寻找的是当前论文所集中关注的主题，适合寻找前沿领域。因为前沿领域的研究往往人数众多而不集中，作品比较分散，被引用情况不稳定，而关键词却很好地体现了该学科的研究热点、发展方向[2]，反映了关键词之间的联系强度及研究的深度。

2 研究及其结果分析

共词分析法应该通过四个步骤[7]：第一，确定推荐引擎领域文献的高频关键词，确定高频词的方法有两种：一种是结合研究者的经验在选词个数和词频高度上予以平衡；另一种是结合齐普夫第二定律低频词分布规律判定高频词的界限。第二，建立共词矩阵，以便下一步使用。第三，基于共词矩阵选取因子分析、聚类分析、战略坐标图等绘制推荐引擎领域的知识图谱。第四，对得到的数据进行分析。在实际操作中，这些步骤并非固定的，而是可以根据研究对象和目标有选择地予以省略或重复。

2.1 词频分析

用Bicomb共词分析软件的词频统计功能对文献的关键词字段进行统计，在共词分析中，重点就是对高频关键词进行统计，因为高频词最能代表文献所在的主题领域。本文选取累计频次不小于5的关键词为高频词，去掉web、图书馆、推荐等过于抽象的词，合并一些相似的词组，如平均绝对偏差和平均绝对误差、相似性和相似度。最后得到高频词44个（见表1），这些关键词能从较大程度上代表推荐引擎领域的研究热点。

2.2 建立共词矩阵并标准化

利用Bicomb共词分析软件中共词统计功能对确定的44个高频关键词统计出它们在703篇论文中两两出现的频次，生成一个4444的共词矩阵。为消除原始共词矩阵绝对值差异的影响，真正揭示关键词之间的共现关系，我们利用Ochiia[8]系数把原始矩阵转换成相似矩阵（见表2）和相异矩阵。用1减去相似矩阵中的所有数值便得到相异矩阵。相似矩阵中的数值越接近1，表明两个词之间的关系越密切，表中对角线上的数据表示某个词与自身的相关程度，上式中均为1。

2.3 多元统计分析

多元统计方法是共词分析法的核心内容，研究

表1　2005~2010年我国推荐引擎领域高频关键词

序号	关键词	频次	序号	关键词	频次	序号	关键词	频次	序号	关键词	频次
1	推荐系统	293	12	个性化推荐系统	23	23	电子商务推荐系统	8	34	用户聚类	5
2	协同过滤	257	13	数字图书馆	22	24	项目相似性	8	35	图书荐购系统	5
3	个性化	154	14	平均绝对偏差	22	25	信息推荐	8	36	Web日志挖掘	5
4	电子商务	135	15	推荐技术	15	26	云模型	7	37	知识管理	5
5	数据挖掘	57	16	智能代理	15	27	个性化信息服务	7	38	自适应	5
6	Web挖掘	49	17	用户兴趣模型	14	28	信息检索	7	39	邻居用户	5
7	关联规则	46	18	模糊聚类	13	29	Web数据挖掘	7	40	网页推荐	5
8	推荐算法	45	19	本体	11	30	远程教育	7	41	决策支持系统	5
9	聚类	42	20	推荐引擎	10	31	隐私保护	6	42	信息过滤	5
10	相似性	40	21	稀疏性	9	32	搜索引擎	6	43	神经网络	5
11	个性化服务	28	22	用户模型	9	33	兴趣度	6	44	文本分类	5

表2 Ochiia系数相似矩阵(部分)

	推荐系统	电子商务	个性化	数据挖掘	关联规则	个性化服务	web挖掘	推荐系统
推荐系统	1.000	0.392	0.245	0.132	0.138	0.114	0.155	0.163
电子商务	0.392	1.000	0.139	0.239	0.165	0.061	0.016	0.241
个性化	0.245	0.139	1.000	0.085	0.059	0.028	0.000	0.189
数据挖掘	0.132	0.239	0.085	1.000	0.254	0.000	0.050	0.844
关联规则	0.138	0.165	0.059	0.254	1.000	0.000	0.000	0.231
个性化服务	0.114	0.061	0.028	0.000	0.000	1.000	0.000	0.000
web挖掘	0.155	0.016	0.000	0.050	0.000	0.000	1.000	0.106
推荐系统	0.163	0.241	0.189	0.844	0.231	0.000	0.106	1.000

选用的统计方法有因子、聚类分析,同时结合战略坐标图分析。

2.3.1 因子分析

因子分析的目标是用尽可能少的因子去描述众多的指标或因素之间的联系,其基本思想是根据关键词间的相关性大小把研究对象的变量分组,使得同组内的变量之间相关性较高,而不同组的变量相关性较低。每组变量代表一个基本结构,这个基本结构称为公共因子,这样较少的几个公共因子就可以反映原资料的大部分信息。本文利用Spss18.0将共词矩阵转化为斯皮尔曼相关矩阵借此消除由共词频次差异所带来的影响,然后在该相关矩阵的基础上,利用主成分法进行因子分析,取9个因子时,累积方差贡献率为91.366%,而取4个因子时,累积方差贡献率超过60%,因子分析碎石图如图1。

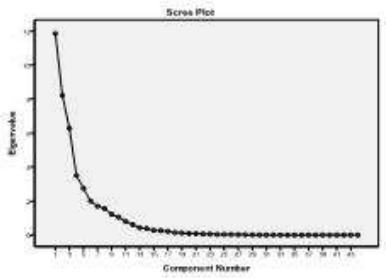

图1 因子个数碎石图

2.3.2 聚类分析

聚类分析的基本思想是利用变量间不同程度的相似性,对事物进行分类,分为一类的个体间相似性较高。本文采用聚类分析法中最常用的系统聚类法对这44个关键词分类(见图2)。

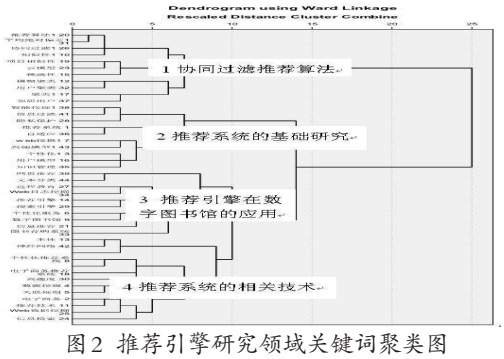

图2 推荐引擎研究领域关键词聚类图

2.3.3 战略坐标图分析

聚类和因子分析给出的只是研究结构的静态特征,而战略坐标图能更好地展现这些研究热点和研究结构的发展动态。它是以向心度和密度为参数绘制成的二维坐标图,一般X轴为向心度,Y轴为密度,原点为二者的均值。向心度是用来量度一个类团和其他类团相互影响的程度。密度是该类的内部强度,它表示该类维持自己和发展自己的能力。这种方法用来描述不同研究领域或研究结构的内部联系和相互影响情况[9]。论文结合因子和聚类分析绘制该领域的战略坐标图,图3中的1,2,3,4的含义与聚类图(图2)中的含义相同。具体分析见2.4这一节。

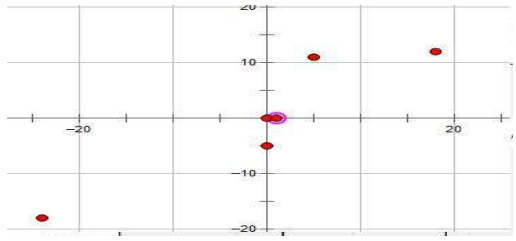

图3 战略坐标图

2.4 结果分析

图1中取4个因子时就已经可以该领域60%以上的信息,图2中分为4大类时比较合理,所以笔者认为该领域包括四个知识群。并结合图3研究知识群的研究现状和热点,在分析的过程中得出四大知识群共包括九大研究热点。

2.4.1 协同过滤推荐算法研究知识群(知识群1)

该群包括推荐算法、平均绝对偏差、协同过滤、相似性、项目相似性、云模型、稀疏性、模糊聚类、用户聚类、聚类、邻居用户等高频关键词。表明该知识群主要围绕协同过滤推荐算法研究。该知识群位于第一象限,但向心度和密度都排第二,其中向心度与2号知识群相差13,说明该知识群虽然与2号知识群同处于推荐引擎领域的核心领域,但与其他知识群的联系紧密程度并不如2号知识群,关注度没有2号群高。但两知识群的密度相差仅为1,说明此二者

内部结构联系紧密,研究较为成熟。

在该知识群中,可分为三个研究热点。

(1)协同过滤。从表1中可以看出,协同过滤这个关键词的频次为257,仅次于推荐系统位居第二位。目前关于协同过滤的研究主要是基于传统协同过滤算法的基础上提出各种改进的算法以提高推荐效果。

(2)聚类。表1中与聚类相关的关键词有用户聚类、聚类、模糊聚类。目前聚类在推荐系统中的应用主要是从项目和用户角度出发进行聚类,以期缩小查询空间而提高系统的实时性。聚类通常采用离线方式建立模型以保证其实时性,但由于时间滞后性可能导致推荐与用户兴趣不符,因此支持用户动态更新的增量机制将是改进其推荐质量的一个新思路。当聚类涉及事物之间的模糊界限时,需运用模糊聚类分析方法。

(3)基础性研究。如稀疏性问题、平均绝对偏差、相似度。面对日益增多的用户和项目,用户–项目矩阵无限增大,协同过滤算法的扩展性问题成为制约推荐系统实施的重要因素。目前学者针对稀疏性问题提出各种解决方案,如维数简化、引入用户基本信息等方法。平均绝对偏差是衡量推荐系统推荐效果的重要指标,也是最常使用的指标。相似性的获取是求解最近邻的前提,用户之间的相似性计算主要包括四种方法:欧几里得距离,余弦相似性,相关相似性,修正的余弦相似性。

2.4.2 推荐系统的基础研究知识群(知识群2)

该群包括推荐系统、自适应、web挖掘、兴趣模型、个性化、用户模型、知识管理等关键词。该知识群主要围绕推荐系统的基础研究。2号知识群向心度和密度都位于第一位,说明该知识群不仅是我国推荐引擎领域的核心,而且研究最为成熟。

该知识群可分为二个研究热点。

(1)推荐系统的应用。推荐系统是为解决信息过载问题而产生的,最早应用于电子商务推荐系统,目前已广泛应用于各领域。从表1中我们看出推荐系统这个关键词的频次为293,位于第一位,说明推荐系统是该领域的研究热点而且研究也逐步趋于成熟。

(2)关于兴趣模型的研究。用户兴趣模型是推荐引擎的基础和核心,兴趣模型的质量直接关系到推荐结果的质量。只有当用户的兴趣、偏好和访问模式等用户信息可以很好地被系统“理解”的时候,才可能实现理想的个性化服务[10]。我国学术界主要包括用户兴趣模型的信息来源与获取方法,用户模型的表示、更新以及评价方法,主要的建模技术等[11]。用户模型在信息检索、电子商务推荐系统、数字图书馆等领域应用广泛。

2.4.3 推荐引擎在数字图书馆的应用知识群(知识群3)

该群包含了网页推荐、文本分类、远程教育、web日志挖掘、搜索引擎、个性化服务、数字图书馆、信息推荐、图书荐购系统等关键词,说明该知识群主要研究推荐引擎在数字图书馆的应用。其位于第三象限,其向心度和密度相比较其他知识群都是最低的,说明其内部结构松散,研究尚不成熟,处于整个研究网络的边缘。事实上关于搜索引擎和数字图书馆的研究已经很成熟,但是本文研究的是数字图书馆、搜索引擎在推荐引擎领域的应用程度,所以处于第三象限并不足为其。

该知识群仅包括一个研究热点:即数字图书馆的研究。与其相关的关键词有数字图书馆、图书荐购系统等。国外比较常见的数字图书馆推荐系统有Tapestry、斯坦福大学的Fab系统、Citeseer系统、基于图表的数字图书馆推荐系统、俄勒冈的SERF系统、加州大学Melvyl推荐系统、Amazon网络书店的推荐系统[12]。冯克鹏[13]、黄晓斌[14]等提出构建基于协同过滤的数字图书馆推荐系统,将协同过滤技术应用数字图书馆中。王燕[15]针对数字图书馆的推荐系统的模型,结合用户的特征,采用Apriori算法挖掘出反映用户需求的强关联规则,实现同用户群内的个性化的推荐。从高频关键词表中可以看出该知识群的关键词位于表1的最后,且在相似矩阵中相似系数基本都为零,所以从严格意义上来说不是一个知识群,也说明这些关键词所代表的研究方向还不够成熟。

2.4.4 推荐引擎与其他领域技术的结合知识群(知识群4)

该群包含了本体、神经网络、个性化推荐系统、电子商务推荐系统、兴趣度、数据挖掘、关联规则、web数据挖掘、信息检索等关键词,说明该知识群的中心是推荐引擎与其他领域技术的结合。其位于第三象限与第四象限的交界处,但其密度和向心度相比较3号知识群要高很多,从这点上来说该知识群具有一定程度的向心度,与其他群体有一定的联系,而且该领域研究尚不成熟,这也说明此象限的研究主题有进一步发展的空间。

该知识群主要包括三个研究热点。

(1)推荐引擎与其他领域技术的结合,其中关联规则、神经网络、本体等关键词体现了这一点。基于关联规则的推荐算法根据生成的关联规则模型和用户当前的购买行为向用户产生推荐[16]。通过离线生成关联规则虽然解决了实时性问题,但在一定程度不能及时反映用户的兴趣变化,而关联规则的频繁生成会增加成本。所以定期增量更新既可以适应用户变化,又可以节约成本。与神经网络的结合主要是利用[17]BP神经网络能够有效地处理非完整信息的特点进行预评分减少候选最近邻数据集的稀疏性及并行性[18]提高了推荐速度。目前本体已经广泛应用于人工智能、知识工程及其相关领域。本体在推荐系统中的应用主要是基于领域知识的研究。佘名高[19]等利用用户兴趣与领域本体中概念的映射关系,构建用户兴趣本体,发掘用户兴趣模式,通过融合评分项目相似度和用户相似度的计算,使用户在评分的共同项目很少或为零的情况下也能找到最近邻进行协同推荐。李珊[20]引入领域本体,对特征项语义进行扩展,构建电影领域本体。根据信息论思想,改进相似度的计算方法,构建基于本体的用户四元组多兴趣细粒度表示模型和相应更新机制。郑晓娟[21]等把领域本体构建的知识库引用到电子商务研究中,提出了一种基于知识库的新型电子商务结构框架。

(2)数据挖掘在推荐系统中的应用。按照数据挖掘应用的行业,可分为旅游、食品、图书馆、服装等。李晓城[22]等提出基于web数据挖掘的健康餐饮分析推荐系统的设计,通过跟踪用户的饮食习惯,推荐有利于用户健康的食品。丁雪[23]等对图书馆借阅记录进行关联规则挖掘,构建基于数据挖掘的图书智能推荐系统。齐杨[24]等从构建的服装电子商务网站上获取的有关消费者消费行为的数据,经过决策树算法挖掘出的规则知识,进而确定特定消费群体,有针对性地提供推荐服务。

(3)电子商务推荐系统。电子商务推荐系统是推荐引擎应用最成功和最广泛的领域。研究电子商务推荐系统的文献很多,主要围绕电子商务推荐系统的设计、关键技术、算法进行研究。

3 结 语

研究结果表明,推荐引擎关键技术是学术界最为关注的主题,是支撑推荐引擎学术研究发展的主要领域。本文应用共词分析法,借助于聚类分析和因子分析对我国推荐引擎学术研究的知识图谱分析。分析结果受关键词规范化、样本数目的选择、高频阈值的确定、聚类分析方法的选择等因素影响。另外,聚类分析本身就是一个无监督的方法,多数情况下要靠经验积累才能找到较好的聚类途经和聚类结果的解释。本文只是从宏观上绘制国内推荐引擎的知识图谱,为研究者提供一个指导方向,具体需要参考推荐系统方面的文章。

参考文献

1 许海铃,吴 潇,李晓东,阎保平.互联网推荐系统比较研究[J].软件学报,2009,(2):351-359.
2 秦长江,侯汉清.知识图谱---信息管理与知识管理新领域[J].大学图书馆学报,2009,27(1):30-37.
3 周春雷.CNKI输出文件在文献计量中的应用[J].图书情报工作,2007,51(7):124-126.
4 邱均平,丁敬达,周春雷.1999-2008年我国图书馆学研究的实证分析(上)[J].中国图书馆学报,2009,(5):72-79.
5 刘 阳,宋余庆.搜索引擎学术研究知识图谱[J].图书情报知识,2010,(6):105-110.
6 马费成,张 勤.国内外知识管理研究重点---基于词频的统计分析[J].情报学报,2006,25(2):163-165.
7 张 勤,徐绪松.定性定量结合的分析方法---共词分析法[J].技术经济,2010,29(6):20-24.
8 宫玉雯,王鲁燕.2005~2008年我国图书情报学web2.0研究热点分析[J].农业图书情报学刊,2010,20(4):57-59.
9 邱均平,丁敬达.1999-2008年我国图书馆学研究的实证分析(下)[J].中国图书馆学报,2009,(5):79-87.
10 关庆珍.基于本体的个性化信息搜索的用户模型研究[D].重庆:西南大学,2008.
11 陈玉娥.个性户服务中用户模型的研究与设计[D].山东:山东科技大学,2007.
12 黄晓斌,张海娟.国外数字图书馆推荐系统评述[J].情报理论与实践,2010,33(8):125-128.
13 黄晓斌.基于协同过滤的数字图书馆推荐系统研究[J].大学图书馆学报,2006,(1):53-57.
14 冯克鹏.基于协同过滤的数字图书馆推荐系统研究[J].软件导刊,2010,9(5):17-18.
15 王 燕.基于关联规则的推荐系统在数字图书馆中的应用[J].情报科学,2007,25(6):878-880.
16 哈进兵,郑 锐,甘利人.一种基于加权关联规则的协同推荐算法[J].情报学报,2010,29(4):718-722.
17 张 锋,常会友.使用BP神经网络缓解协同过滤推荐算法的稀疏性问题[J].计算机研究与发展,2006,43(4):667-672.
18 张 磊,陈俊亮,孟祥武,沈筱彦,段 锟.基于BP神经网络的协作过滤推荐算法[J].北京邮电大学学报,2009,32(6):43-46.

(下转第160页)

所以生产阅读器如同生产电脑和手机一样。在数字出版的行业标准统一之后，各类产品植入或下载应该像手机接收信号一样无差别，差异只在显示质量、操作、功能、色彩上。为此，由技术提供商与中盘商结成合作伙伴，依照各中盘商的技术要求，与电子产品生产商一起创造自己的品牌[7]。这可能才是中国电子书出版的取胜之道。

参考文献

1 王续文.关于数字出版产业发展的思考[N].中国新闻出版报，2010-05-24(9).
2 明 慧.电子书时代是否已经到来[N].中国改革报，2010-10-16(10).
3 韩 成，周中华.中美电子书市场分析与比较：商业模式制胜数字出版产业[N].中国新闻出版报，2010-08-26(7).
4 刘成勇.探索数字出版商业模式与发展路径[N].中国新闻出版报，2010-10-25(5).
5 钟健华.从全球视角看国内数字出版发展趋势[N].中国新闻出版报，2010-10-25(5).
6 于 帆.《关于电子书产业的意见》出台提高准入“门槛”避免资源浪费[N].中国文化报，2010-10-16(6).
7 黄国荣.突破电子书出版发展瓶颈[N].中国新闻出版报，2010-12-27(10).

（责任编辑：刘凤琴）

（上接第148页）

19 佘名高，张照亮，胡锦红.基于领域本体的个性化推荐在健康系统中的研究[J].电子设计工程，2010，18(11)：23-26.
20 李 珊.个性化服务中用户兴趣建模与更新研究[J].情报学报，2010，29(1)：67-71.
21 郑晓娟.基于领域本体的个性化推荐系统研究与应用[D].武汉：武汉理工大学，2009.
22 李晓城，张增杰，夏勇明，钱松荣.基于web数据挖掘的健康餐饮分析推荐系统的设计[J].微型电脑应用，2011，27(1)：44-46.
23 丁 雪.基于数据挖掘的图书智能推荐系统研究[J].情报理论与实践，2010，33(5)：107-110.
24 齐 扬，朱欣娟.基于数据挖掘的服装推荐系统研究[J].西安工程大学学报，2010，24(4)：439-441.

（责任编辑：刘凤琴）