基于主题模型的信息系统研究热点分析*

杨婷，张瑾
（中国人民大学 商学院，北京 100872）

摘　要　为考察信息系统研究热点及其近年来的演变规律，本文以 MISQ（*Management Information Systems Quarterly*）、ISR（*Information Systems Research*）、JMIS（*Journal of Management Information Systems*）三本信息系统顶级期刊 2008~2017 年的文献为研究对象，采用 LDA（latent Dirichlet allocation，潜在狄利克雷分配）主题模型提取文献中潜在的十年内持续度较高的研究热点主题，并对其演变情况进行分析。此外，在当前大数据技术和信息系统研究深度融合的背景下，本文结合大数据驱动研究的全景式 PAGE（paradigm，analytics，governance，enabling）框架，进一步探究了信息系统研究与大数据核心研究方向间的映射关系。

关键词　信息系统，LDA 主题模型，文献计量，研究热点

中图分类号　C931.6

1　引言

伴随着经济全球化及信息全球化的进程，各行业数据获取壁垒逐渐降低，数据维度日益增长，信息技术快速发展，应用场景不断推陈出新。在此背景下，信息技术和信息系统应用逐渐成为促进社会经济发展的重要动力，受到的关注日益增加。作为近几十年来蓬勃发展的新兴研究领域，其中的研究方向和研究主题也随着时间的推移表现出不同的发展变化趋势，而研究的角度也呈现出百花齐放的姿态，不同的学者分别从个人、团体、组织和市场的角度探讨了信息系统和技术之间的相互影响。多样化的研究热点使人们得以对复杂的问题和事物获得更加深入的了解，进而促进了信息系统学科的发展。

追溯过去几十年，人们对信息系统领域的认识和兴趣有着不断的突破和创新，从最初仅限于系统和技术的基本概念和范畴，拓展至管理的层面和经济的视角，研究信息化过程中涉及的管理、系统和技术相关的现象，并通过不断引入归并新兴的技术，对现有系统更新换代，已成功建立了一套相对完善的理论体系[1~5]。

考察信息系统领域的研究热点及演变趋势，能够帮助相关学者和研究人员更好地把握本领域的发展趋势。尤其是在 2008~2017 年，信息系统的研究方法深度交叉，研究问题日益多样，有效地梳理已有研究热点能够帮助相应学者识别新的研究机会或研究视角，甚至对热点问题进行预测，为后继研究的开展提供必要借鉴。鉴于此，本文选取 MISQ、ISR、JMIS 三本信息系统领域的主流权威期刊 2008~2017 年十年时间内发表的文章进行主题建模分析，分析信息系统研究热点的发展情况与变化趋势，梳理热门主题与新兴主题的研究前沿，从而为相关学者提供本领域内研究机遇与挑战的借鉴。

* 基金项目：国家自然科学基金（71772177，72072177）。

通信作者：张瑾，中国人民大学商学院管理科学与工程系。E-mail：zhangjin@rmbs.ruc.edu.cn。

2 文献综述

2.1 信息系统学科的研究热点梳理

信息系统的概念面世并开始流行至今，其内涵和外延经历了多次迭代。在 20 世纪 70 年代，Walter T. Kennevan 定义管理信息系统为：“以书面或口头的形式，在合适的时间向经理、职员及外界人员提供过去的、现在的、预测未来的有关企业内部及其环境的信息，以帮助他们进行决策。”[6]随后，明尼苏达大学卡尔森管理学院教授 Gordon B. Davis 给出了更加完善的定义：“管理信息系统是一个利用计算机硬件和软件，手工作业，分析、计划、控制和决策模型，以及数据库的用户——机器系统，它能提供信息，支持企业或组织的运行、管理和决策功能。”[7]这个定义说明了信息系统的目标、功能和组成，且反映出信息系统当时已达到的水平。90 年代后，人类进入信息时代，信息资源对企业经营效益的巨大作用凸显出来，管理信息系统扩展出更广泛的发展空间，进入完善和创新阶段[8]。随着科学技术的推陈出新与社会经济的不断发展，管理信息系统的应用范畴愈加广阔，现在人们已普遍使用信息系统来指代这一意义上的系统。进入 21 世纪后，随着互联网的崛起，信息系统的研究和引用范围进一步延伸，客户关系管理、供应链管理等新方向出现[9]。随着信息化进程的持续前行，信息系统将面对更深入、更丰富的需求和更广阔的发展空间。

与此同时，信息系统领域的学者们也在不断对学科的研究热点进行分析和梳理，以帮助人们更好地把握信息系统的未来研究和应用方向。信息系统研究的热点内容随时间推移不断变化，新的研究方向层出不穷。Alavi 和 Carlson 系统地分析了 1968~1988 年 8 本信息系统核心期刊上发表的文章，得出信息系统领域流行的主要课题有信息系统管理、信息系统类型和特点，以及系统开发运行等[10]。Benbasat 和 Zmud 针对当时信息系统领域内的研究问题与基于信息技术的现象及问题愈加遥远，使得信息系统学科的中心身份逐渐模糊的问题，通稿讨论信息系统学者的研究，提出定义信息系统领域的一组核心属性，即概念和现象，来描述这一学科的中心身份[11]。Banker 和 Kauffman 回顾过往 50 年管理科学中信息系统及信息技术文献的发展情况，揭示这一领域的起点和进展，并通过定义信息系统领域的五个研究流派——决策支持与设计科学、人机交互、信息价值、信息系统组织及战略、信息技术经济学，总结五个研究流派的分析水平、相关理论、方法和相关学科，从而考察整个学科的进步情况[12]。Sidorova 等基于潜在语义分析方法对 1985~2006 年发表在信息系统 3 本顶级期刊的文章进行梳理归纳，确定了五个核心的研究方向：信息技术与组织，信息系统发展，信息技术与个体，信息技术与市场，信息技术与群组[13]。在该研究覆盖的时间段内，这些核心主题保持稳定，且每个方向内的具体研究主题都有明显的发展，反映相关研究更多侧重于信息技术与设计、使用的社会背景，而非技术发展本身。徐青等分析了 MISQ 期刊 2000~2009 年的文章信息，通过关键词分析得出结论：MISQ 的研究内容从信息系统的技术与管理的基本概念和范畴方面演化为更具体、前沿的话题，如电子商务、外包、用户接受等[14]。

2.2 主题模型在学科分析中的应用

主题模型是对文字隐含主题进行建模的方法。它克服了传统信息检索中文档相似度计算方法的缺点，能够在文本数据中自动寻找出文字间的语义主题[15]。应用主题模型对文献主题进行分析的国内外成果主要有：Liu 等使用创新的文本和图形挖掘算法及全文引文分析和主题建模，提高经典文献计量分析和发表排名[16]。Yau 等分析了基于 LDA 主题模型的科学文献分类方法，在 LDA 主题模型的基础上扩展研究，并选取不同领域的学术论文，验证其提出的分类方法的有效性[17]。Jiang 等聚焦于水电开

发领域，采用基于主题建模的文献计量分析方法，对全球水电科学文献进行定量评价[18]。王金龙等针对目前科研文献主题演化概率分布问题，阐述了主题与事件的关联关系，提出了一种新型的基于模块化的主题方法[19]。王萍在验证 LDA 主题模型对于文献知识挖掘可行性的基础上，提出了一种新的概率主题模型——Topic-Author 模型，对文献的文本信息和作者信息进行联合建模，研究趋势分析和主题关系挖掘[20]。任柯等提出一个基于主题模型的协作文献推荐，结合传统协同过滤、概率主题模型和知识协作网络模型，利用语义相似度的计算工具，提出基于概率的跨学科的检索推荐[21]。叶春蕾和冷伏海综合科研文献的关键词和引用文献，构建了一种引文-主题概率模型，经过分析，该模型可获取关键词及引文的分布情况，并能实现上述内容的主题识别[22]。王平针对科技文献主题多样、动态性强等特点，分析科技文献主题发现及演化具体方法，使用层次概率主题模型 hLDA（hierarchical latent dirichlet allocation），采用 Gibbs 抽样来进行模型参数估计，并通过计算互信息对主题词进行筛选，提取高质量的主题词[23]。

2.3 文献总结

综上，可以看出目前已有的信息系统研究热点分析的年代较为久远，尤其是缺少对 2008~2017 年研究热点的全面分析。信息系统研究的发展迭代速度不断加快，因此近期研究热点的全面分析对未来研究更具指导意义。此外，现有的分析信息系统文献的工作除一些借助主观文献梳理和综述之外，较多使用词频分析、潜在语义分析、关联分析等研究方法，关键词为主要的研究对象。虽然关键词能在一定程度上代表文献的研究内容，但每篇文献的关键词数量有限，且相同含义的关键词可能有多种表达方式，会导致词频分析精度的降低。因而，本文借助 LDA 主题模型，从话题语义的角度更全面地分析信息系统的研究热点。

另外，近年来借助主题模型进行学科研究热点分析的研究比较多见，如在水电开发、材料科学、生物科学、教育科学等领域，这些研究主要使用 LDA 主题模型对相关科学领域的文献进行文本分析，包括主题识别、主题演变和研究趋势分析等，也有一些研究在不同的科学文献分析领域开发 LDA 的改进模型。虽然 LDA 主题模型被应用在很多学科领域进行研究热点分析，但专门针对信息系统领域，尤其是 2008~2017 年信息系统领域的研究热点分析相对较少。因此，为更好地帮助信息系统领域的学者分析学科最近的研究热点和发展动向，本文采用 LDA 主题模型对本领域 MISQ、ISR、JMIS 三本顶级期刊 2008~2017 年文献进行主题分析，并结合大数据驱动研究的全景式 PAGE 框架，进一步探究信息系统研究与大数据核心研究方向间的映射关系。

3 信息系统研究文献的主题建模

3.1 数据来源

本文选择信息系统领域三本主要权威期刊 MISQ、ISR、JMIS 2008~2017 年的文章作为研究对象。对十年间三本期刊所有文章，剔除部分缺少关键词或摘要的短文章或特殊类型的文章，如 Editor's Notes、Executive Overview、Special Issue Introduction 等之后，得到可用于分析的 MISQ 论文 499 篇，ISR 论文 480 篇，JMIS 论文 404 篇。利用 Pdfparser 工具析取每篇文章的标题、摘要、关键词、引言，去除 Pdf parser 处理有误的文章后，得到 1 358 篇文章的代表文本。2008~2017 年每年的文章代表文本数量的统计结果如表 1 所示。

表 1　MISQ、ISR、JMIS 期刊 2008~2017 年文章的获取情况

年份	文章数量	年份	文章数量
2008	100	2013	178
2009	104	2014	139
2010	126	2015	132
2011	134	2016	138
2012	172	2017	135

3.2　数据处理

本文使用 Python 自然语言处理工具包（natural language toolkit，NLTK），对摘要文档进行数据清洗，将文档转化为单词列表。NLTK 创建于 2001 年，最初是宾夕法尼亚大学计算机与信息科学系计算语言学课程的一部分。在数十名贡献者的帮助下，NLTK 不断发展壮大，如今已被几十所大学的课程采纳，并作为许多研究项目的基础。本文分别使用 NLTK 中 corpus、tokenize、tag、stem 这几个模块进行分词、词性标注、词形还原等数据预处理工作。

由于文本语料中存在大量如同 the、a、and、of 等语义无价值的词，我们引入 NLTK 的停用词集合，将该类词汇筛除。对停用词集合之外、信息系统研究领域内的高频词汇，如 use、user、base、theory、study 等，考虑到其经常与其他词汇构成短语出现，如 user generated content（UGC）、theoretical basis（理论基础）、game theory（博弈论）、empirical study（实证研究），如仅根据出现频率将这些词汇筛除，可能造成文本语义的损失和主题识别的不完整，因此我们在语料清洗环节保留这些研究领域内的常用、通用词汇。但是由于以上列举的通用词汇不一定以短语形式出现，也会较常单独出现，将其保留在语料中输入 LDA 主题模型，对 LDA 主题模型提取主题的精度会造成一定的影响，这也是本文的局限。整体清洗流程如图 1 所示。

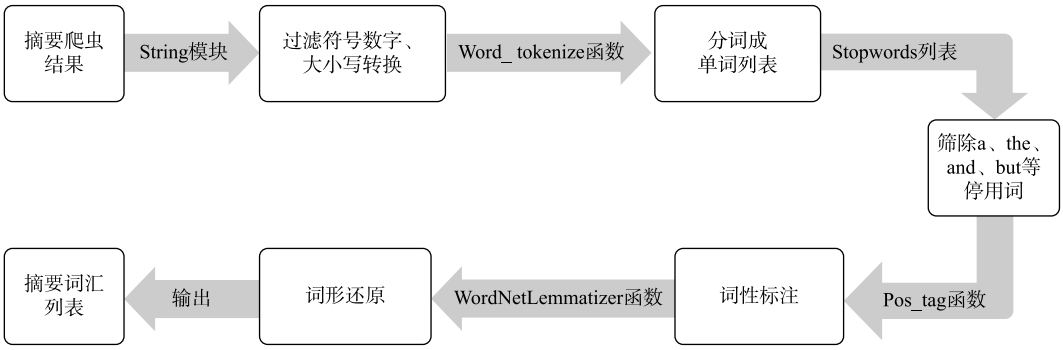


图 1　数据清洗流程图

经过以上步骤，代表文本中以字符串形式存储的标题、摘要、关键词、引言文本已转化成不含数字、符号、停用词、单词变形的词汇列表。将词汇列表作为后续分析的语料基础，完成代表文本的数据清洗工作。

3.3　模型训练

主题模型通过对文本隐含主题进行建模，能够在文本数据中自动寻找出文字间的语义主题。LDA 主题模型是一种基于概率的主题发现模型，能够提取文本隐含主题的非监督学习模型，是主题模型中最主要的经典模型，2003 年由 David M. Blei 等学者提出，现今 LDA 主题模型已经在机器学习的诸

多领域[24~27]及信息检索领域[28, 29]得到广泛应用。在文献主题发现领域，已有学者通过研究发现，在学术文献数据集中，LDA 主题模型能有效地发现有价值的隐含主题[30]。因此，本文选择基于 LDA 主题模型挖掘信息系统领域顶级期刊学术文章的主题信息。

LDA 主题模型是一个三层贝叶斯概率模型，由词、主题、文档三层结构组成。LDA 主题模型认为，每个主题是固定词表上的一个多项式分布，每个文档由多个主题混合，而文档到主题服从 Dirichlet 分布，主题到词服从多项式分布。因此，LDA 主题模型将每一篇文档视为一个词频向量，将文本信息转化为易于建模的数字信息，学习得到隐含在文本信息中的隐含主题，以及文档、词汇与主题之间的对应关系。

本文采用 Python 的 scikit-learn 包进行 LDA 主题模型训练。scikit-learn 提供的 LDA 主题模型以变分推断（variational inference）和 EM（expectation maximization，最大期望）算法来得到模型的文档主题分布和主题词分布。首先，采用 CountVectorizer 函数对清洗后得到的 1 358 篇代表文本统计词频并保存为词频向量；其次，采用 LDA 进行主题模型训练，由于模型样本容量为 1 358，可以不将样本分批训练，因而求解算法参数选择“batch”，最大迭代次数设置为 1 000，doc_topic_prior、topic_word_prior 参数使用默认值 1/K。

LDA 主题模型的性能很大程度上受到主题数目 n_components 的影响，常见的确定主题数目的方法有以下三种：使用模型困惑度（perplexity）的方法[15]、非参数化主题模型层次狄利克雷过程（hierarchical Dirichlet processes，HDP）的方法[31]、应用贝叶斯模型确定最优主题数的方法[32]等。在此基础上，也有学者提出基于密度的自适应最优主题数、困惑度与主题相似度结合的最优主题数方法[33, 34]。其中，HDP 模型能够从文档集中自动训练最合适的主题数 K，但需要为同一集合分别建立一个 HDP 模型和一个 LDA 主题模型，且算法时间复杂性较高，存在效率不高的问题。以贝叶斯模型确定最优主题数目的方法则只能确定主题数目，缺乏对模型泛化能力的刻画[31]。引入主题相似度的方法多是以科技文献作为研究对象，而非研究特定领域的文献主题[34]。综合考虑之下，由于本文仅选取信息系统领域三本顶级期刊十年间的文献进行研究，样本量适中，因此本文选择困惑度作为度量指标，选择最优主题数 K。

LDA 的困惑度是概率语言模型中用来评估语言模型优劣的指标之一，其基本含义为训练得到的主题模型对一篇文章属于某个主题的不确定性，较小的困惑度意味着模型对新文本有较好的预测作用，因而困惑度一般随着潜在主题数量的增加呈现递减的规律。因此，在其他条件固定的情况下，能使困惑度最低的主题数目，即该模型的最佳主题数[35]。

通过测试，在其他参数固定的条件下，本文所建 LDA 主题模型的困惑度与主题数间的对应关系如图 2 所示。

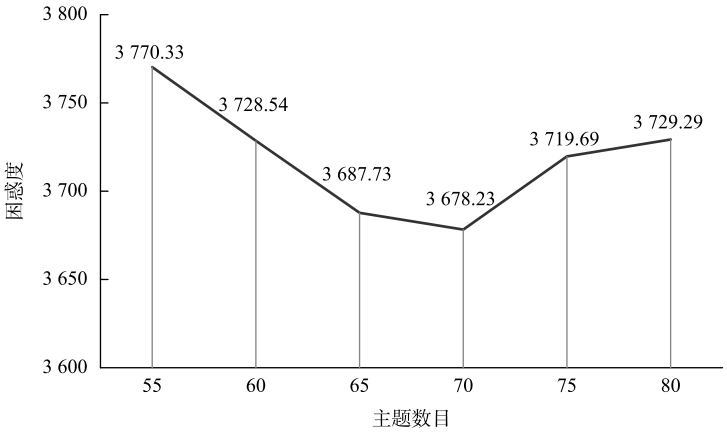


图 2 困惑度−主题数目关系图

由图 2 可知，LDA 主题模型的困惑度曲线在主题数取 70 时出现明显的拐点，此时得到最低的困惑度为 3 678.23。故选择模型参数 n_components=70 建立 LDA 主题模型。

4 信息系统研究热点分析与结果讨论

4.1 主题-代表词汇结果

根据上文建立的 LDA 主题模型，经过实验得到 70 个主题及其代表词汇列表。由于篇幅限制，文中仅列出每个主题的前 10 个代表词汇。然而每个主题在文档中的分配并不均匀，对此，Griffiths 借助文档的时间属性统计抽取出主题在时间维度上的分布情况，以此考察固定时间窗口内每个主题的强度[32]。本文参考 Griffiths 和 Steyvers 对主题强度的定义方法，通过计算所有文档在 70 个主题内的概率分布累加和，考察 2008~2017 年十年时间内每个主题的强度。对任一主题 z，其主题强度计算公式如式（1）所示：

$$I_z = \frac{1}{N_d}\sum \theta_{zd} \tag{1}$$

其中，N_d 表示主题模型中文档 d 的数量；θ_{zd} 表示文档 d 中主题 z 的概率。因此，主题强度表达了主题模型中所有文档在该主题的概率均值。对 70 个主题的强度进行计算，发现强度分布呈现长尾分布，强度最高的 7 个主题占据全部主题强度总和的 49.86%，前 10 个主题占据全部主题强度总和的 57.41%。因此，本文首先选择主题强度最高的 10 个主题，即 70 个主题中最具代表性的 10 个核心主题进行展示，其代表词汇如表 2 所示（主题标识括号内的数字为该主题在模型生成的 70 个主题中出现的顺序）。

表 2 主题提取结果（前 10 个主题）

主题标识	代表词汇
主题 1（6）	use research information system study model base theory process provide
主题 2（34）	firm investment risk information value decision business study industry performance
主题 3（66）	service business innovation technology digital customer value provider strategy resource
主题 4（9）	product consumer online information purchase sale customer effect study quality
主题 5（40）	user behavior use social individual technology influence perceive employee intention
主题 6（18）	market price product cost consumer model platform pricing effect offer
主题 7（11）	network social medium user share diffusion effect peer model content
主题 8（60）	team group work collaboration task technology virtual performance member coordination
主题 9（44）	software development source virtual developer open vendor world patch design
主题 10（7）	community online member participation knowledge social innovation participant open contribution

从 LDA 主题模型结果可以看出，以上最具代表性的 10 个主题的内容相对独立，覆盖技术、系统、管理、创新多个层面，涉及当下的新兴信息技术、大数据概念，并且都在信息系统的研究领域内，体现 LDA 主题模型对信息系统期刊文献进行主题提取的方法是有效的。

强度最高的主题 1 描述的是“信息系统”方面的内容，也是本文研究的核心母题。在文本数据清

洗环节，我们为避免语义损失保留了信息系统研究领域内的通用词汇，如 study、model、base、theory。在 LDA 主题模型提取出的主题结果中，可以获知通用高频词汇基本出现在强度最高的主题 1 的代表词汇中，而在其他提取出的主题中出现频率很低，或是以 user behavior（用户行为）的短语形式与相关词汇共同出现在某个主题中。相比于主题 1 的普遍性，其余主题的代表词汇皆可指向信息系统领域内的细分方向，如商业服务、产品消费等，由此可知本文 LDA 主题模型提取信息系统研究主题的有效性。

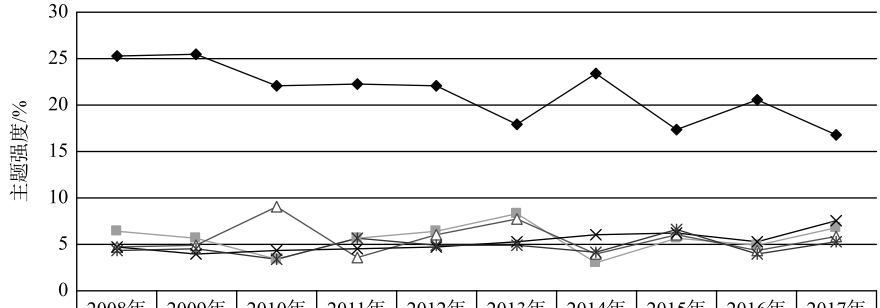
主题 2 的描述内容更聚焦于公司组织，主要研究公司组织的投资战略、决策制定、风险分析等；主题 3 主要描述数字技术创新与服务方面的内容；主题 4 与主题 6 的描述内容为产品与消费者，并且主题 6 的定价跨界到与信息系统多有关联的市场营销领域；主题 5 的描述内容为用户个体行为；主题 7~主题 10 分别描述社交网络、团队协作、软件开发、网络社区等信息系统领域的经典问题。综合表 2 中各主题的代表性词汇，将主题强度最高的 10 个主题分别概括为信息系统、企业投资、商业服务、产品消费、用户行为、营销定价、社交网络、团队协作、软件开发、网络社区。接下来，将从时间维度进一步对这 10 个主题进行细粒度分析。

4.2 主题强度随时间演变趋势

4.1 节在不考虑时间因素的情况下，利用 LDA 主题模型提取出文献主题，并通过计算主题强度分析信息系统的热点话题。本节将进一步分析抽取出的主题随年份变化的演变趋势。根据上文对主题强度的定义，以年为单位重新划分时间窗口，计算每个主题在 2008~2017 年内年度主题强度，计算公式见式（2）：

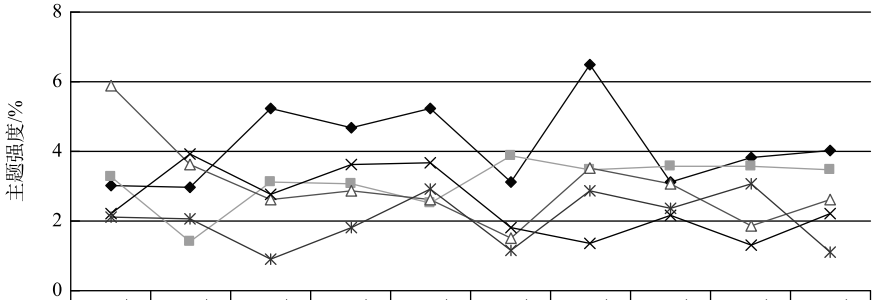
$$I_{yz}=\frac{1}{N_{yd}}\sum_{d\in y}\theta_{zd} \tag{2}$$

其中，y 表示以年为单位的时间窗口；I_{yz} 表示某一年内主题 z 的强度；N_{yd} 表示某一年内文档的数量；θ_{zd} 依然表示文档 d 中主题 z 的概率。4.1 节展示的 10 个代表主题在 2008~2017 年十年间的演化趋势如图 3 和图 4 所示。



	2008年	2009年	2010年	2011年	2012年	2013年	2014年	2015年	2016年	2017年
信息系统	25.20	25.54	22.06	22.20	21.99	18.00	23.48	17.39	20.59	16.85
企业投资	6.37	5.61	3.35	5.59	6.49	8.21	3.10	5.70	4.89	6.82
商业服务	4.68	4.82	9.13	3.50	6.08	7.65	3.87	6.06	4.35	5.80
产品消费	4.74	3.91	4.26	4.60	4.81	5.21	6.03	6.23	5.31	7.50
用户行为	4.36	4.48	3.38	5.66	4.91	4.83	4.15	6.56	3.92	5.30

图 3 主题演化趋势（一）



	2008年	2009年	2010年	2011年	2012年	2013年	2014年	2015年	2016年	2017年
营销定价	3.03	2.98	5.22	4.67	5.23	3.11	6.51	3.11	3.84	4.01
社交网络	3.27	1.39	3.11	3.08	2.54	3.87	3.49	3.58	3.55	3.47
团队协作	5.89	3.61	2.60	2.88	2.63	1.53	3.51	3.05	1.87	2.60
软件开发	2.20	3.94	2.75	3.60	3.68	1.81	1.38	2.16	1.31	2.23
网络社区	2.13	2.07	0.91	1.81	2.90	1.14	2.86	2.36	3.08	1.09

图 4　主题演化趋势（二）

2008~2017 年，信息系统对本领域的关注与研究基本呈递减趋势，“信息系统”主题的强度始终远高于其他相关主题，但从 2009 年的 25.54%逐渐降至 2017 年的 16.85%，表明这近十年间随着信息系统领域的发展，学者们研究的热点逐渐分散至信息系统领域更细致的问题，“信息系统”作为关键词或主题出现的频率降低。“企业投资”“商业服务”“用户行为”主题强度波动幅度较大，但十年内始终保持较高的研究热度；“产品消费”主题的强度平稳增长，自 2008 年的 4.74%稳步上升至 2017 年的 7.50%。

“营销定价”主题 2014 年的强度达到高峰，而后则恢复至平均热度并保持平稳；在大数据发展迅猛的背景下，“社交网络”主题的强度在 2013~2017 年明显提升；“网络社区”主题在 2008~2017 年十年时间内的强度波动较大，但整体趋势较为平稳；“团队协作”主题强度呈现整体下降的趋势，“软件开发”主题强度呈现先升后降趋势，两个主题在信息系统领域皆保持一定的热度。

从以上结果可以看出，信息系统领域内存在十年间持续受到关注和研究的热门主题。但这些热门的方向仍然包含更多细粒度的研究问题，因而本文以“用户行为”这一方向为例，使用 LDA 主题模型进一步探索该方向内更细致的研究问题。

4.3　细分领域发展趋势——以“用户行为”为例

从 4.2 节 LDA 主题模型的结果中，我们提取出在“用户行为”这一主题中获得高主题强度的 117 篇文章，并以此为新的训练语料训练新 LDA 主题模型，当提取主题数 n=10 时，得到最低困惑度 p=2 914.80，10 个主题按主题强度降序排列，其代表词汇如表 3 所示（主题标识括号内的数字为该主题在模型生成的 10 个主题中出现的顺序）。

表 3　用户行为领域细粒度主题提取结果

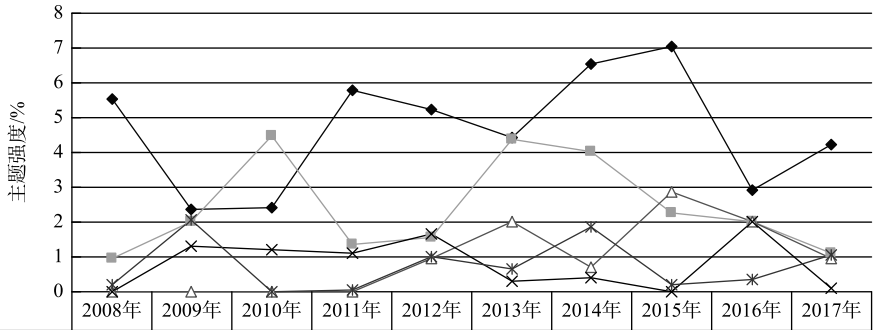
主题标识	代表词汇（topic_top10_words）
主题 1（10）	use social user behavior technology research study model network individual
主题 2（6）	service online social network community user product information model knowledge
主题 3（1）	learn study use research community practice game training outcome user
主题 4（5）	project control innovation research organization soa organizational alignment business community
主题 5（4）	information consumer product firm online share user decision informational individual

续表

主题标识	代表词汇（topic_top10_words）
主题 6（7）	privacy information behavior individual threat concern team online exchange user
主题 7（2）	medium security social technology cyberbullying research discourse digital information study
主题 8（8）	quality commentator moderation reputation comment compete high click user position
主题 9（3）	market pricing phishing price memory attack card effect network spot
主题 10（9）	security user end identity norm job perceive nms attitude performance

从 LDA 主题模型的细粒度提取结果中可知，在用户行为这一领域内热度最高的话题仍为核心母题，这一结果与整体 LDA 主题模型的结果一致。紧随其后的热点话题分别为在线服务、社区研究、项目控制、信息消费、隐私信息、媒体安全、质量评价、营销定价、用户安全，皆是包含在用户行为范畴内或与之高度相关的研究主题。

纳入时间因素，对以上 10 个主题随年份的演变趋势进行分析，得到其 2008~2017 年十年间的变化情况，如图 5 和图 6 所示。



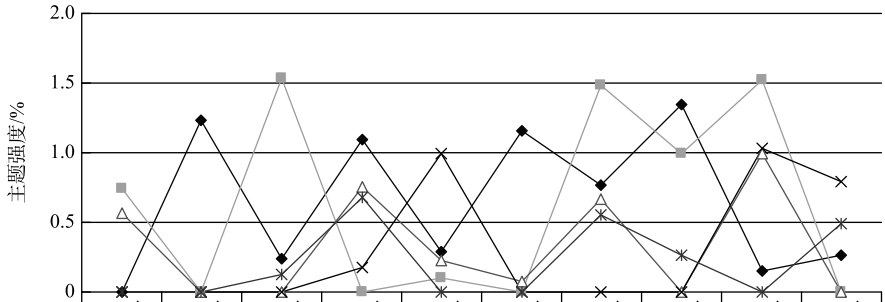
	2008年	2009年	2010年	2011年	2012年	2013年	2014年	2015年	2016年	2017年
用户行为	5.55	2.38	2.41	5.77	5.21	4.41	6.52	7.06	2.93	4.24
在线服务	0.94	2.00	4.47	1.36	1.55	4.38	4.03	2.26	2.00	1.10
社区研究	0.00	0.00	0.00	0.00	0.96	2.00	0.68	2.86	2.00	0.96
项目控制	0.00	1.33	1.21	1.10	1.65	0.31	0.41	0.00	2.00	0.09
信息消费	0.19	2.06	0.00	0.06	1.00	0.65	1.88	0.20	0.36	1.08

图 5 用户行为领域内主题演化趋势（一）

与整体 LDA 主题模型的结果不同，在此我们展示的是在用户行为领域内提取得到的所有热点主题而非强度最高的主题，因此各个主题的强度随时间的分布情况更加随机。在整体 LDA 主题模型中，2015 年“用户行为”主题的强度最高，为 6.56%，2010 年的主题强度最低，为 3.38%。这一分布特征与 LDA 主题模型在“用户行为”这一细粒度领域中的“社区研究”主题呈现出的结果相类似。

4.4 长尾主题分析

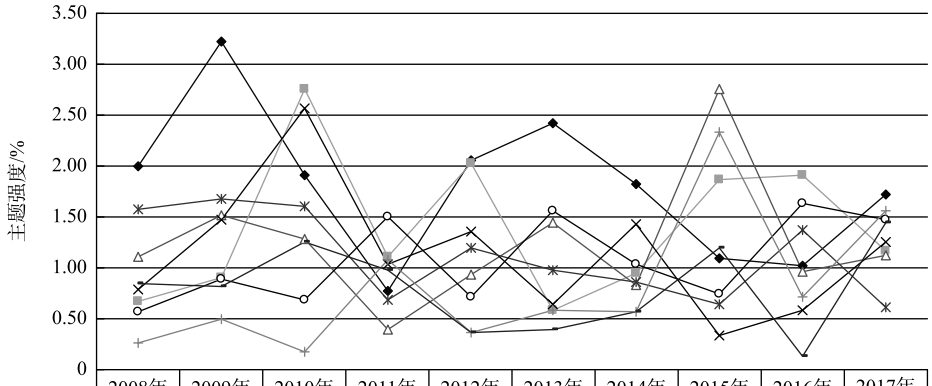
在提取出的其他 60 个主题中，本文发现一些主题在特定时间内出现研究高峰，但由于其他年份主题强度较低，所以综合十年的主题强度不高。而在对长尾分布中的“尾巴”部分主题分析时，本文以 2015~2017 年主题强度与其余年份主题强度差值作为排序依据，差值越大，说明该主题越有可能成为未来受到广泛关注的主题。将 60 个主题进行排序并选取排名靠前的 8 个主题作为“潜在新兴”或“再次兴起”的候选主题，其在 2008~2017 年十年间的主题强度演化趋势如图 7 所示。为确定候选主题的



	2008年	2009年	2010年	2011年	2012年	2013年	2014年	2015年	2016年	2017年
隐私信息	0.00	1.23	0.24	1.10	0.29	1.16	0.77	1.34	0.15	0.26
媒体安全	0.74	0.00	1.54	0.00	0.10	0.00	1.49	1.00	1.52	0.00
质量评价	0.56	0.00	0.00	0.76	0.23	0.07	0.67	0.00	1.00	0.00
营销定价	0.00	0.00	0.00	0.17	1.00	0.00	0.00	0.00	1.03	0.79
用户安全	0.00	0.00	0.13	0.68	0.00	0.00	0.55	0.26	0.00	0.49

图 6　用户行为领域内主题演化趋势（二）

研究前景，本文引入 2018 年及 2019 年上半年三本期刊的文献，统计 8 个主题在这些文献中的出现频次，从而判断信息系统领域的当下趋势及未来的研究机会。



	2008年	2009年	2010年	2011年	2012年	2013年	2014年	2015年	2016年	2017年
项目管理	2.00	3.23	1.91	0.77	2.05	2.42	1.83	1.10	1.02	1.72
搜索广告	0.67	0.90	2.76	1.11	2.03	0.58	0.95	1.86	1.91	1.16
信息安全	1.11	1.52	1.28	0.39	0.93	1.44	0.83	2.76	0.96	1.13
竞争能力	0.79	1.47	2.57	1.03	1.35	0.64	1.43	0.33	0.59	1.25
外包合同	1.58	1.68	1.60	0.68	1.19	0.98	0.86	0.64	1.37	0.61
隐私	0.57	0.89	0.69	1.50	0.71	1.56	1.04	0.74	1.64	1.48
医疗	0.26	0.49	0.18	1.08	0.37	0.58	0.57	2.34	0.72	1.56
供应链	0.85	0.82	1.26	0.97	0.36	0.40	0.57	1.20	0.13	1.44

图 7　候选主题演化趋势

从表 4 中可以看出：以上 8 个从长尾主题中筛选出来的主题中，医疗、信息安全的后续研究数量相对最高，三本期刊中分别共有 12 篇相关文献，表现出成为受到广泛关注研究主题的潜力，这两个主题也确实是近年来实践领域和学术领域共同关注的新兴热点话题；与竞争能力、隐私、项目管理主题相关的研究均在 6 篇左右，表现出一定的研究潜力；而搜索广告、外包合同、供应链主题的研究文献数量较少，呈现出较弱的研究潜力。因此，通过对“尾巴”部分的主题进行分析，并以主题在 2015~2017

年强度与其余年份强度的差值作为依据，可以有效地识别出那些潜在的“新兴主题”，当然这些主题中也存在很大的比例最终并没有受到广泛的关注。信息系统领域的学者可参考这一趋势，进一步关注和挖掘医疗、信息安全等方向的研究内容。

表 4 2018 年和 2019 年上半年文献中候选主题出现频次

主题	出现频次
医疗	12
信息安全	12
竞争能力	7
隐私	6
项目管理	5
搜索广告	4
外包合同	4
供应链	2

进一步深入探索“医疗”主题中的具体研究，可以发现：Yan 和 Tan 调查在线医疗社区中患者与其他患者的信息交流对治疗效果的影响，发现交流过程中达成共识对患者的治疗效果有积极影响，而这种积极影响受到共享信息的数量、患者预先承诺和社会联系的负向调节作用[36]。该研究的结果可以帮助医疗从业人员制定有效的干预措施，帮助患者更好地治疗。在未来研究中，可以扩展至治疗本身之外的评价信息，如对医生、服务提供者的评价信息共享是否能影响患者治疗效果。Lin 等的研究结合健康信息技术、大数据和预测分析方法，通过搭建贝叶斯多任务学习（Bayesian multitask learning，BMTL）的综合预测模型，对慢性疾病患者的健康风险进行预测，证明 BMTL 方法可靠的预测性能，在临床实践方面的应用可减少预防性干预的失败和延误[37]。未来的研究可以在更多不同的数据集上试验 BMTL 方法，探索 BMTL 方法的适用场景与边界。

5 讨论

本文分析了 2008~2017 年十年间信息系统领域持续关注的研究热点，以及逐渐趋热的研究主题。需要特别注意的是，大数据已成为当前信息系统研究的一项重要的时代特征，在回顾梳理这些热点对应的研究时，在 MISQ、ISR、JMIS 十年的文献中发现存在许多与大数据概念和应用相结合的研究，而由于大数据在其中较多体现在数据获取、数据处理和数据分析方法层面，而较少直接体现在研究主题当中，因而大数据概念在本文提取的研究主题结果中并不突出。结合陈国青等针对管理决策情境下大数据驱动的研究应用提出的融合大数据特征与重要研究方向的全景式 PAGE 框架[38]（图 8），本文进一步探究信息系统领域与大数据研究中理论范式、分析技术、资源治理、使能创新四个核心研究方向的呼应关系。

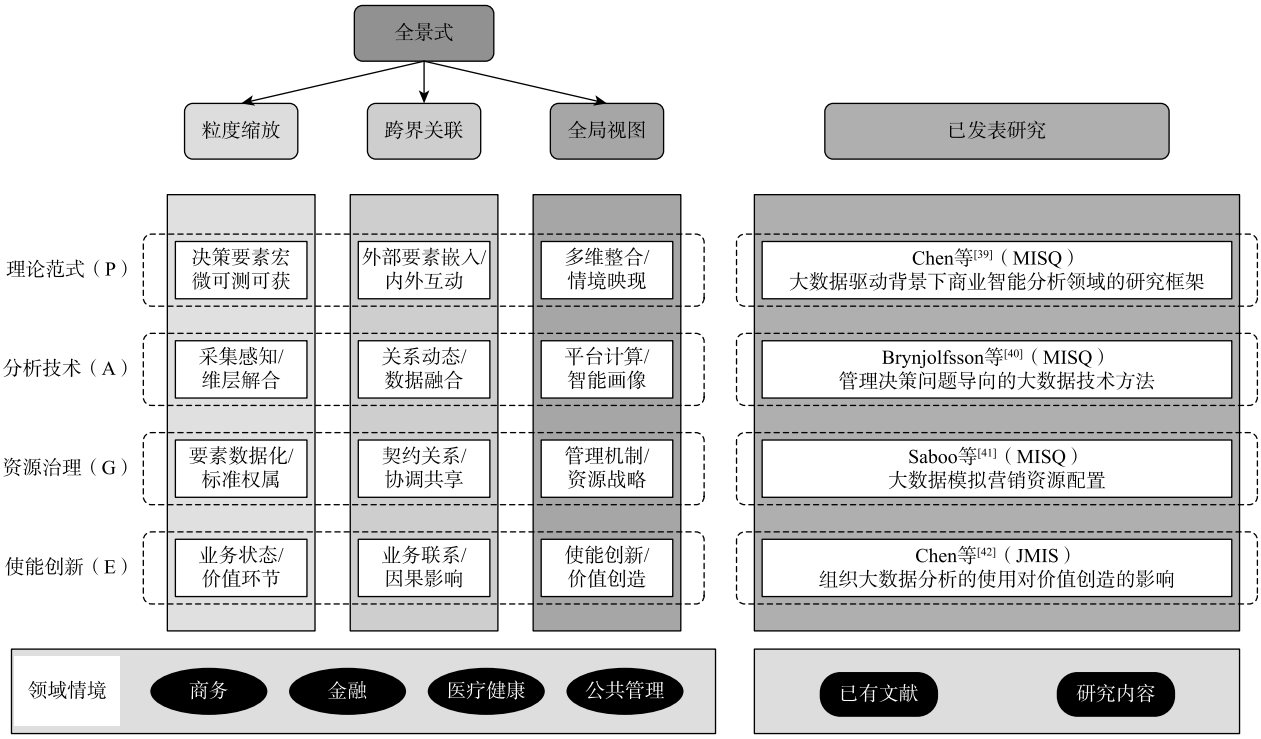


图 8　全景式 PAGE 框架及对应文献

选取本文 LDA 主题模型提取得到的相对具体的 9 个主题对应文献，邀请两名信息系统方向的研究学者对这些文献进行判断，先判断其是否属于大数据研究范畴；对于属于大数据研究范畴的文献，进而判断其与 PAGE 框架的四个研究方向之间的对应关系并进行分析，分别将其映射至 PAGE 模型的理论范式、分析技术、资源治理、使能创新四个研究方向。对于映射结果中的冲突内容，我们又邀请第三名信息系统方向的学者进行判断，以求给出准确的映射关系。统计主题与方向交叉的文献数量，得到潜在主题与 PAGE 模型映射汇总情况，如表 5 所示。

表 5　热点主题与 PAGE 模型映射汇总

主题	企业投资	商业服务	产品消费	用户行为	营销定价	社交网络	团队协作	软件开发	网络社区
P	3	3	0	0	0	2	1	4	0
A	1	1	3	5	1	2	0	0	2
G	2	1	0	0	0	0	0	2	0
E	1	0	1	0	1	0	0	0	0

基于表 5 的结果，我们可以总结出，首先，在时间维度上，由于 PAGE 模型中的四个研究方向皆有鲜明的特征主题，同时也因为 PAGE 框架提出时间相对较新，所以我们映射后取交集得到的文献数量并不多，从时间维度我们尝试进一步拆分后并未观测到显著的演化趋势。其次，目前在信息系统顶级期刊的研究成果中，理论范式和分析技术两个研究方向相对较多，而资源治理和使能创新相对欠缺，这也是未来信息系统研究可以重点拓展的两个方向。具体地，在理论范式方向，较多研究软件开发、企业投资、商业服务主题的文章提及理论框架的构建；在分析技术方向，研究用户行为、产品消费的文献数量居多；在资源治理方向，出现较多对企业投资、软件开发主题的研究；在使能创新方向，企

业投资、产品消费、营销定价主题相关的文献更多关注价值的创造。

由于在时间维度上未观测到 PAGE 框架四个研究方向上的显著演化趋势，为给信息系统领域学者提供更好的借鉴，本文选取了四个方向的代表性文章进行分析和讨论，如图 8 所示。具体地，在十年的文献当中，Chen 等提出了在大数据驱动背景下，商业智能分析（business intelligence & analytics，BI&A）这一研究领域的框架，确定 BI&A 的发展脉络、应用场景和新兴方向，与 PAGE 框架中理论范式形成呼应[39]。Brynjolfsson 等发现人群产生的大数据对人们意图、偏好和意见的预测潜力，选择“搜索趋势数据”作为数据选取的方法，进一步挖掘对用户行为的预测潜力，并通过实证研究证实其方法的可靠性[40]。其研究与关注管理决策问题导向的大数据技术方法的分析技术研究方向吻合。Saboo 等利用大数据模拟研究营销资源配置的时变效应，使用时变效应模型（time-varying effect model，TVEM）为大数据的时间变化建模，发现营销邮件、交易特征和人口统计因素对销售的影响会随时间而变化[41]。其研究结果可以指导公司纠正忽略时变导致的资源错配，重新配置企业资源，提升销售收入。此研究基于大数据提出公司资源配置的有效建议，呼应了关注大数据资源治理、组织资源战略制定的资源治理研究方向。Chen 等利用动态能力理论，对组织层面的大数据分析的使用前提与对价值创造的影响进行分析，提供了对组织使用大数据分析的前因后果的理解，同时也提供了关于管理者使用这一新兴技术的指导[42]。大数据分析对供应链管理中价值创造的研究与关注数据价值创造的使能创新研究方向对应。在 PAGE 框架所指出的各个方向中均已有研究发表在 MISQ、ISR、JMIS 等信息系统的顶级期刊中。

6 总结

本文运用主题模型中的 LDA 主题模型对 2008~2017 年信息系统领域的研究热点进行主题提取和分析，并梳理了热点主题十年间的演变情况。本文研究表明：2008~2017 年，信息系统研究热点有企业投资、商业服务、产品消费、用户行为、营销定价、社交网络等。在 2008 年前后，对企业投资、团队协作主题的研究较多；在 2010 年，对商业服务主题进行的研究远多于其他领域内的主题；2012~2013 年，对企业投资、商业服务主题的关注程度较高；2014 年则是与营销领域相关的主题更受关注，如营销定价、产品消费；2015 年围绕用户行为进行的研究热度达到十年间的最高值；2016~2017 年对企业投资、产品消费主题进行的研究较多。此外，纵观十年间的整体演变趋势，产品消费与社交网络的主题强度呈现上升趋势，团队协作主题强度呈现下降趋势，软件开发主题强度呈现先升后降趋势。

本文的研究也存在一定的局限性。首先，在文献的选择上，本文选取信息系统领域三本顶级期刊的文献作为研究对象，文献对于信息系统领域各方面研究的覆盖面略有不足。其次，本文使用 LDA 主题模型提取得到 70 个主题，但大部分主题的出现频率较低，长尾效应明显，本文主要针对热点高频主题和长尾主题中部分新兴主题进行了分析和讨论，如何进一步挖掘其他低频主题，并从中提取出对信息系统学科发展有益的结论，也是未来需要进一步探讨的问题。

参 考 文 献

[1] Cooper R B. Review of management information systems research：a management support emphasis[J]. Information Processing & Management，1988，24（1）：73-102.

[2] Avgerou C. The significance of context in information systems and organizational change[J]. Information Systems Journal，2001，11（1）：43-63.

[3] Wade M，Hulland J. Review：the resource-based view and information systems research：review，extension，and suggestions for future research[J]. MIS Quarterly，2004，28（1）：107-142.
[4] Avgerou C. Information systems in developing countries：a critical research review[J]. Journal of Information Technology，2008，23（3）：133-146.
[5] 章以佥，宗乾进，袁勤俭. 国际管理信息系统研究热点及趋势[J]. 情报杂志，2013，（4）：80-84，90.
[6] 廖燕，曹建安. 企业内部绩效评价信息系统特征与结构研究[J]. 情报杂志，2006，25（9）：43-44，47.
[7] 佘维，罗爱静. 企业 MIS 理论与发展对策初探[J]. 科技情报开发与经济，2007，17（29）：147-149.
[8] 黄梯云，李军一. 管理信息系统导论[M]. 第 3 版. 北京：机械工业出版社，2004.
[9] 孙卫军. 管理科学与工程类专业人才培养模式研究[D]. 天津大学硕士学位论文，2005.
[10] Alavi M，Carlson P. A review of MIS research and disciplinary development[J]. Journal of Management Information Systems，1992，8（4）：45-62.
[11] Benbasat I，Zmud R W. The identity crisis within the IS discipline：defining and communicating the discipline's core properties[J]. MIS Quarterly，2003，27（2）：183-194.
[12] Banker R D，Kauffman R J. 50th anniversary article：the evolution of research on information systems：a fiftieth-year survey of the literature in Management Science[J]. Management Science，2004，50（3）：281-298.
[13] Sidorova A，Evangelopoulos N，Valacich J S，et al. Uncovering the intellectual core of the information systems discipline[J]. MIS Quarterly，2008，32（3）：467-482.
[14] 徐青，李天智，张腾跃，等. 基于 CiteSpace 的 MIS 前沿研究：对 2000-2009 年 MISQ 发表论文的分析[C]. 中国管理现代化研究会. 第五届（2010）中国管理学年会——管理科学与工程分会场论文集，2010.
[15] Blei D M，Ng A Y，Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research，2003，3：993-1022.
[16] Liu X Z，Zhang J S，Guo C. Full-text citation analysis：enhancing bibliometric and scientific publication ranking[C]. ACM International Conference Proceeding Series，2010：1975-1979.
[17] Yau C K，Porter A，Newman N，et al. Clustering scientific documents with topic modeling[J]. Scientometrics，2014，100：767-786.
[18] Jiang H C，Qiang M S，Lin P. A topic modeling based bibliometric exploration of hydropower research[J]. Renewable and Sustainable Energy Reviews，2016，57：226-237.
[19] 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.
[20] 王萍. 基于概率主题模型的文献知识挖掘[J]. 情报学报，2011，30（6）：583-590.
[21] 任柯，黄智兴，邱玉辉. 基于主题模型的跨学科协作文献推荐[J]. 计算机科学，2012，39（9）：235-239，261.
[22] 叶春蕾，冷伏海. 基于引文—主题概率模型的科技文献主题识别方法研究[J]. 情报理论与实践，2013，36（9）：100-103.
[23] 王平. 基于层次概率主题模型的科技文献主题发现及演化[J]. 图书情报工作，2014，58（22）：70-77.
[24] Dhillon I S，Modha D S. Concept decompositions for large sparse text data using clustering[J]. Machine Learning，2001，42（1/2）：143-175.
[25] Kabán A，Girolami M A. A dynamic probabilistic model to visualise topic evolution in text streams[J]. Journal of Intelligent Information Systems，2002，18（2/3）：107-125.
[26] Chua F C T，Lauw H W，Lim E P. Generative models for item adoptions using social correlation[C]. IEEE Transactions on Knowledge and Data Engineering，2013，25（9）：2036-2048.
[27] Panahi N，Shayesteh M G，Mihandoost S，et al. Recognition of different datasets using PCA，LDA，and various classifiers[C]. 2011 5th International Conference on Application of Information and Communication Technologies（AICT），IEEE，2011：1-5.
[28] Hassan S U，Haddawy P. Analyzing knowledge flows of scientific literature through semantic links：a case study in the field of energy[J]. Scientometrics，2015，103（1）：33-46.
[29] Wu Q Q，Zhang C D，Hong Q Q，et al. Topic evolution based on LDA and HMM and its application in stem cell research[J]. Journal of Information Science，2014，40（5）：611-620.
[30] 范云满，马建霞. 基于 LDA 与新兴主题特征分析的新兴主题探测研究[J]. 情报学报，2014，33（7）：698-711.
[31] Teh Y，Jordan M，Beal M，et al. Hierarchical dirichlet processes[J]. Journal of the American Statistical Association，2006，101（476）：1566-1581.
[32] Griffiths T L，Steyvers M. Finding scientific topics[J]. Proceedings of the National Academy of Sciences of the United States of America，2004，101（S1）：5228-5235.
[33] Cao J，Zhang Y，Li J，et al. A method of adaptively selecting best LDA model based on density[J]. Chinese Journal of Computers，2009，31（10）：1780-1787.
[34] 关鹏，王曰芬. 科技情报分析中 LDA 主题模型最优主题数确定方法研究[J]. 现代图书情报技术，2016，（9）：42-50.

[35] Grossman D A. Information Retrieval：Algorithms and Heuristics[M]. Berlin：Springer Science & Business Media，2004.
[36] Yan L，Tan Y. The consensus effect in online health-care communities[J]. Journal of Management Information Systems，2017，34（1）：11-39.
[37] Lin Y K，Chen H，Brown R A，et al. Healthcare predictive analytics for risk profiling in chronic care：a bayesian multitask learning approach[J]. MIS Quarterly，2017，41（2）：473-495.
[38] 陈国青，吴刚，顾远东，等. 管理决策情境下大数据驱动的研究和应用挑战——范式转变与研究方向[J]. 管理科学学报，2018，21（7）：1-10.
[39] Chen H，Chiang R H L，Storey V C. Business intelligence and analytics：from big data to big impact[J]. MIS Quarterly，2012，36（4）：1165-1188.
[40] Brynjolfsson E，Geva T，Reichman S. Crowd-squared：amplifying the predictive power of search trend data[J]. MIS Quarterly，2016，40（4）：941-961.
[41] Saboo A R，Kumar V，Park I. Using big data to model time-varying effects for marketing resource（re）allocation[J]. MIS Quarterly，2016，40（4）：911-939.
[42] Chen D Q，Preston D S，Swink M. How the use of big data analytics affects value creation in supply chain management[J]. Journal of Management Information Systems，2015，32（4）：4-39.

附录 LDA主题模型所有主题代表词汇

主题编号	代表词汇
1	memory volunteer motivation sponsor transactive participant bank trade card perceive
2	research method neurois brain neuroscience mixed actor failure qualitative researcher
3	phishing deception detection credibility message assessment fraud theory detect decision
4	project control governance mechanism development performance formal offshore process prediction
5	page web user time status customer app visitor mobile link
6	use research information system study model base theory process provide
7	community online member participation knowledge social innovation participant open contribution
8	user idea generate ideation content comment quality time reputation status
9	product consumer online information purchase sale customer effect study quality
10	review rating bias average product rmation negative star positive helpfulness
11	network social medium user share diffusion effect peer model content
12	cloud service center compute open department content quality cost article
13	construct formative critical indicator realism latent structural reflective path bias
14	data informational cascade plan environmental consumer user isp content isps
15	exception demand cost share platform forecast resolution cpfr data record
16	document category demand public new technology website variety evolution concentration
17	organic brand substitution conflict path camp market middle marketplace equity
18	market price product cost consumer model platform pricing effect offer
19	coupon seller market couponing valuation investor board homophily participation quality
20	contract outsource client vendor cost relational contractual transaction incentive party
21	supply chain partner supplier coordination information integration stock use inter

续表

主题编号	代表词汇
22	capability competitive process industry relationship human information action capital performance
23	expectation model rmation technology discon cache use addiction utaut experience
24	model good depreciation pricing cost information consumer sell lease individual
25	knowledge editor university share edu senior school erent learn history
26	content technology social knowledge learn communication alliance study site medium
27	attacker end information policy target prim accountability effect jump game
28	security information threat attack breach behavior insider policy cop organization
29	alignment business unit organizational management level organization strategic social strategy
30	negotiation agent market price research customer design buyer spot offer
31	use user adoption research relationship attraction design svw base heuristic
32	mobile app apps user social change facebook cyberbullying size medium
33	learn game engagement outcome learning self trainee date technology strategy
34	firm investment risk information value decision business study industry performance
35	cio executive structure reporting leadership ugc ceo relationship career cfo
36	healthcare network adoption social information model process practice probability centrality
37	agility alignment organizational firm interdependency performance business change environmental ambidexterity
38	reputation seller open work buyer transparency source development openness strategic
39	query reuse content database anchor sql work adjustment neutrality search
40	user behavior use social individual technology influence perceive employee intention
41	privacy information user disclosure concern personal seal individual share change
42	diversity journal process error work social research eld automation discipline
43	isd distance temporal boundary work project member team stress developer
44	software development source virtual developer open vendor world patch design
45	government service diversity cultural development citizen technology country supplier customer
46	compliance security control policy reward punishment employee enforcement protection emotion
47	cyberloafing announcement formal cyberloaf expressive neutralization past peer logical learn
48	value fashion problem sample large decision advisor criterion regression failure
49	rule provider hit mip recommendation item moral quality past double
50	trial cation software sociomaterial free experience practice time erp fit
51	channel distribution travel new acquisition airline self electronic diversification market
52	auction bid piracy bidder price market consumer bidding share seller
53	infant rural social energy mortality broadband care qos neutrality content
54	ict digital divide country use communication develop access social research
55	data analytics big auction record intelligence combinatorial predictive ciency linear

续表

主题编号	代表词汇
56	search advertising quality price click model online advertiser strategy sponsor
57	trust web blog site culture commerce cultural online Internet satisfaction
58	bundle crm consumer good behavioral price argument effect purchase customer
59	social medium user platform content release generation forum generate employee
60	team group work collaboration task technology virtual performance member coordination
61	research student knowledge doctoral scale practice value professional community practitioner
62	price social piracy dispersion model market control public use sector
63	patient health care hospital medical telemedicine physician clinical visit disease
64	governance fashion Internet resource ios integration ownership control operational business
65	data content medium model program broadcast quality distribution sale revenue
66	service business innovation technology digital customer value provider strategy resource
67	data input spillover use capital social medium productivity production estimate
68	recommendation performance recommender portal customer base use web strategy effect
69	idea itc capability task heterogeneity resource durable leadership dependence representation
70	value social resource technology potential outcome absorptive website competency capacity

Research Hotspot Analysis of Information System based on Topic Modeling

YANG Ting，ZHANG Jin

（School of Business，Renmin University of China，Beijing 100872，China）

Abstract In order to investigate the hotspots of information system research and the evolution in recent years，we take literature from MISQ，ISR，and JMIS during the past ten years as research object，use LDA topic model to extract the potential hot topics with high sustainability and analyze their evolution in the past decade.In addition，in the context of the current deep integration of big data technology and information system research，this paper combines the panoramic PAGE framework of big data-driven research to further explore the relationship between information system research and the core research direction of big data.

Keywords Information systems，LDA topic modeling，Bibliometric analysis，Research hotspot

作者简介

杨婷（1995—），女，中国人民大学商学院2020级博士研究生，研究方向为商务分析、内容消费、文本挖掘等。E-mail：yangting_rmbs@ruc.edu.cn。

张瑾（1984—），男，中国人民大学商学院副教授、博士生导师，研究方向为大数据管理与分析、数字经济、电子商务、文本挖掘、商务智能等。E-mail：zhangjin@rmbs.ruc.edu.cn。