主题模型在临床文本挖掘中的应用现状*

李 燊 夏晨曦 马敬东

（华中科技大学同济医学院医药卫生管理学院 武汉 430030）

〔摘要〕 采用文献调研分析法对近10 年国内外运用主题模型方法挖掘临床文本的研究进行归纳分析，总结研究现状和常用的主题模型方法，阐述主题模型在文本挖掘领域存在的局限性，以期为相关领域的研究提供借鉴。

〔关键词〕 主题建模；临床文本；文本挖掘

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2018.05.012

Current Application Status of Topic Modeling in Clinical Text Mining *LI Shen, XIA Chen－xi, MA Jing－dong, School of Medicine and Health Management, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China*

〔**Abstract**〕 By investigating and analyzing literature, the paper carries out inductive analysis of domestic and overseas studies that mined clinical text through topic modeling over the past 10 years, summarizes status quo of the studies and topic modeling methods that are often adopted, and expatiates on the existing limitations of topic modeling in the text mining field to provide reference for studies in relevant fields.

〔**Keywords**〕 Topic modeling; Clinical text; Text mining

1 引言

信息技术的迅猛发展和“互联网＋”时代的到来大力推动医院信息化建设，由此产生大量的临床数据信息[1]，这些结构化（如性别）与非结构化（如护理记录）的临床数据中包含大量有价值的信息，对临床文本进行挖掘，以发现其中有价值的信息意义重大。但临床文本的数据挖掘面临诸多困难，自由文本不能结构化存储，使临床文本资源难以直接有效地利用。而且同种疾病的记录间存在联系，形成庞大、复杂的关联网络，使用常规的数据挖掘分析方法对临床文本进行挖掘分析并不适用。

在文本挖掘领域常用的方法有关联规则分析、聚类、分类等[2]，但对于处理大量无规则的文本数据而言，这些方法的效率和准确度都有所欠缺。主题模型是一种用于在文本中发现出抽象主题的方法，相对于传统的文本挖掘方法，它能高效地完成一些基本工作，如挖掘出文本的潜在关系、判断关联性、分类等[3]。因此对于临床文本的挖掘来说，应用主题模型是一个很好的方法。从 Web of Science 中关于"topic model" 主题的文献量粗略地了解到对主题建模研究情况，可知该方法是近 10 年来在文本挖掘领域逐渐兴起的一种挖掘方法。但主题建模目前在临床文本挖掘中还未得到广泛应用，本研究

〔修回日期〕 2018－01－16

〔作者简介〕 李燊，硕士研究生；通讯作者：马敬东，副教授。

〔基金项目〕 中央高校基本科研业务费资助项目“区域医疗机构知识网络形成机制研究”（项目编号：2015AE017）。

对国内外运用主题模型方法来挖掘临床文本的研究情况进行细致调研，对其现状和存在的问题进行讨论分析。

2 资料与方法

本研究以 Web of Science 和 PubMed 两个数据库中的文献作为主要数据来源，因相关文献数量不多，检索时扩大范围，故检索式分别为 TS = (("topic model") AND (EHR OR EMR OR electronic health records OR electronic medical records OR clinical data))、"topic model" [All Fields]，时间限定为 2008 – 2017 年，文献类型限定为期刊文章。其中 Web of Science 检出 34 篇，PubMed 112 篇（检索时间为 2017 年 10 月 25 日），汇总去重后得到 121 篇文献。同时还使用中国知网、万方数据库对国内临床文本主题挖掘研究的期刊文献进行检索，中国知网检出 4 篇，万方数据库 3 篇，去重后共计 5 篇。去除非临床文本的数据源以及只用结构化的临床数值数据以及其他方法进行文本挖掘的文献，选择主要运用主题建模、其他与主题建模相结合的方法对临床文本进行分析的文献，最终筛选出最契合的文献共 26 篇。本研究从多方面对这些文献信息进行归纳，临床文本主体建模总结，见表 1（省略同一作者不同年份发表的应用相似文献）。其中采用英文语料的文献 14 篇，中文 11 篇，其他语言 1 篇。

表 1 临床文本主题建模总结

序号	第 1 作者	时间（年）	应用
1	Corey W Arnold	2017	评估主题模型描述医疗报告的能力
2	Liangying Yin	2017	发掘治疗模式并分析医疗行为随时间推移的变化
3	Mark Hoogendoorn	2017	直肠癌建模预测
4	William Speier	2017	改进医疗临床报告质量
5	Aaron Zalewski	2017	评估患者健康状态
6	刘玉文	2017	识别疾病特征及分布
7	Jonathan H Chen	2016	比较自动生成医嘱和传统医嘱的效能
8	Yen – Fu Luo	2016	ICU 后死亡率预测及主题特征发掘
9	A. Rumshisky	2016	早期精神病患者再入院预测
10	Zhengxing Huang	2015	临床危险评估
11	Zhengxing Huang	2015	挖掘疾病潜在的治疗模式
12	Karla Caballero	2015	动态预测 ICU 再入率
13	A. Fong	2015	评估安全事件报告程度类别
14	Liang Yao	2015	发掘中医临床治疗模式
15	徐天明	2015	构建中文医学信息可视化分析系统
16	Yijun Shao	2015	识别精神错乱记录的文本
17	Li – wei Lehman	2014	预测出院后死亡率
18	Marzyeh Ghassemi	2014	ICU 死亡率建模
19	Raphael Cohen	2014	对患者记录的冗余建模
20	许珠香	2013	预测给定症状患者的诊断
21	Li – wei Lehman	2012	ICU 患者风险分层
22	Filip Gintera	2009	护理记录的主题分割与标注

3 结果与讨论

3.1 概述

基于文献调研情况，归纳出进行主题建模分析的数据来源主要有入院、诊断、检查、治疗、病史、出院记录和安全事件报告，以下将对目前利用主题建模对临床文本进行挖掘的具体应用和前景进行描述，对常用的主题建模预测临床文本主题建模方法进行讨论分析。

3.2 研究现状

3.2.1 国外 对于临床数据的挖掘研究较多，近年来开始运用主题建模方法对临床文本进行主题挖掘，从而更好地发现临床文本中特定疾病的临床治疗模式，进行疾病预测，为医生诊断提供决策支持。随着主题建模在临床文本挖掘中的发展，所运用的建模方法也呈现多样化，运用概率主题模型拓展到根据研究目的对模型进行适应性调整以达到更具针对性的挖掘效果，多种方法相结合来进行挖掘分析。

3.2.2 国内 目前对于临床数据的挖掘大多是从

结构化数据中提取，运用常用的统计学方法进行整合分析，运用主题建模方法进行临床文本挖掘的研究不多，近 10 年中文临床文本挖掘运用主题建模方法的研究期刊文献不超过 20 篇，且应用较为局限。由于临床文本数据获取困难，用主题建模对临床文本进行挖掘大都是对临床文本相关文献建模分析，而用在医院临床文本挖掘的研究不多，仅黄正行等[4-5]运用主题建模对临床文本挖掘以发现疾病特征、临床治疗路径模式和进行临床危险评估；刘玉文[6]等用潜在犹利克雷分布（Latent Dirichet Allocation，LDA）来识别疾病特征以辅助诊断；许珠香[7]等用 LDA 对中医临床文本挖掘以发现中医临床文本的主题结构等。由此可见，在我国运用主题建模进行临床文本挖掘仍有很大的发展空间。

根据表 1 结果可发现，当前运用主题建模方法来分析临床文本数据的应用主要有以下几个方面：（1）预测。根据现有的临床文本，分析主题建模结果来对患病风险、死亡率进行预测。如 A. Rumshisky[8]对早期精神患者再入院概率进行预测，Li - wei Lehman[9]对出院后患者死亡率进行预测等。但就目前利用临床文本主题建模预测的情况来说，整体预测准确率不高，结果只能提供一定参考。（2）发掘主题。在文本挖掘领域主题建模方法能够发掘潜在的主题信息，特别适用于大量临床文本数据的主题挖掘，为提高医疗诊断和服务提供支持。如 Zhengxing Huang[5]、殷良英[10]等运用主题挖掘方法发掘临床路径治疗模式，优化临床治疗的路径，提高医疗服务水平。（3）为临床决策提供支持。利用临床文本主题挖掘可识别出主要疾病，根据特征概率对就诊患者生成诊断建议，为医生诊断提供有价值的参考。如刘玉文[6]使用对临床文本中病症特征进行主题识别，辅助医生对疾病进行诊断。（4）评估。对临床文本主题建模分析后，根据主题的生成结果对患者健康状态、安全事件等级等进行评估，为今后的治疗服务提供改进建议。如 Aaron Zalewski[11]用主题建模评估患者健康状态，为患者提供更好的治疗。

3.3 常用主题建模方法

3.3.1 概述 文献调研中发现运用主题建模方法对临床文本进行分析挖掘的方法主要为 LDA 及其衍生模型相关文献有 22 篇，占 85%；3 篇使用层次犹利克雷过程（Hierarchical Dirichlet Process，HDP），占 11%；采用其他方法分析的很少，仅有 1 篇，占 1.4%。

3.3.2 LDA 及衍生模型 LDA 主题模型是 2003 年由 Blei 等人提出的一种文档主题生成模型，是一种非监督机器学习技术，不需要手工标注就能够根据大规模语料信息进行主题生成，对于识别其中潜藏的主题信息具有很好的效果[12]。该方法在文本挖掘领域有广泛应用，包括文本主题识别、分类、相似度计算等方面。近年来在临床文本挖掘方面 LDA 方法开始引起关注，目前对大量的临床文本挖掘时大都使用该方法。由于临床文本中的特殊名词、过程关联、结果影响等有别于其他类型文本，所以在使用 LDA 进行临床文本挖掘时有时还需要根据临床文本的特点或挖掘目的对模型进行调整，使其更适合对临床文本进行挖掘，如 Zhengxing Huang[4]所采用的概率风险分层模型（Probabilistic Risk Stratification Model，PRSM）在 LDA 建模基础上还生成患者子特征风险层级，Raphael Cohen[13]采用的 Red - LDA 则可考虑到在建模过程中临床文本的患者记录固有的冗余等。由此可知 LDA 运用广泛，具有很高的拓展性，可根据实际需要对模型进行调整，以获取更好的挖掘结果。

3.3.3 HDP 该模型是由 Teh 于 2005 年提出的一种主题生成模型，在狄利克雷过程（Dirichlet Process，DP）的基础上，使用 Stick - breaking、Polya Urn 或 Chinese Restaurant Process 构造狄利克雷过程[14]。Li - wei Lehman[9]利用 HDP 对重症监护室患者风险进行分层和预测出院后 1 年内的死亡率，从而有针对性地为患者预后提供建议；Aaron Zalewski[11]用 HDP 将临床护理记录与时间值结合，从而更好地根据时间推移来评估患者健康状态等，可发现运用 HDP 对含有时间节点的临床文本挖掘效果较好，因此 HDP 也是临床文本挖掘中较为常见的方法。

3.3.4 其他 在临床文本主题挖掘中，LDA 和 HDP 是最常见的两种方法，此外临床文本挖掘还可

以综合运用不同的方法来实现其挖掘目标，如 Karla Caballero[15] 采用 LDA 与广义动态线性模型（Generalized Dynamic Linear Models，GLDM）相结合，克服时间变化带来的影响，从而使模型能够根据时间推移而进行调整；Filip Gintera[16] 将潜在语义分析（Latent Semantic Analysis，LSA）与隐马尔可夫模型结合，可不需标注地自由定义感兴趣的话题效果。除主要利用概率主题模型方法对文本进行挖掘外，还可以通过统计学的聚类、K－means 等方法，但这些方法对文本和潜在主题挖掘效果不佳。目前在文本挖掘领域，还存在着奇异值分解（Singular Value Decomposition，SVD），基于词向量等主题建模方法，但在临床文本挖掘中还未有所使用。

3.4 现有研究存在的局限性

3.4.1 概述 综合文献分析结果，目前用主题建模方法进行临床文本分析仍处于起步阶段，还存在一些问题，以下将对现有文献中的局限性进行归纳分析，如结果需要人工标注使可解释性受主观因素影响，分词工具不成熟使预处理结果影响最终结果，数据来源有限，未考虑数据间、过程间的关联性，见表2。

表2 现有部分文献中的局限性

作者	方法	局限性							
		可解释性	分词工具	数据源	未考虑关联性	普适性	动态性	参数影响	准确性
Corey W Arnold	LDA 及衍生模型	√	–	√	–	√	√	–	–
Liangying Yin	LDA 及衍生模型	–	√	√	√	√	–	–	–
Mark Hoogendoorn	LDA 及衍生模型	√	–	√	–	–	–	–	–
Jonathan H Chen	LDA 及衍生模型	√	–	–	–	–	√	√	√
Yen－Fu Luo	LDA 及衍生模型	√	–	–	–	–	–	–	–
A Rumshisky	LDA 及衍生模型	–	√	–	√	√	–	–	√
Zhengxing Huang	LDA 及衍生模型	–	–	–	√	–	√	√	–
Zhengxing Huang	LDA 及衍生模型	–	–	√	–	√	–	√	–
Karla Caballero	LDA 及衍生模型	–	–	√	–	–	–	–	–
A Fong	LDA 及衍生模型	–	√	√	–	–	–	–	√
Liang Yao	LDA 及衍生模型	–	√	–	–	–	–	–	√
Li－wei Lehman	HDP	–	–	√	√	–	–	–	√

3.4.2 方法准确度不高 主题建模方法适用于大量临床文本的挖掘，目前常用非监督机器学习的概率主题模型方法，其中 LDA 需要先验提供主题数目，对模型结果的准确率、质量影响很大，而现有求取主题数目大多采用困惑度的方法，但尚未有公认的最佳求取主题数目方法[17]。而 HDP 模型虽然能够自主选择最佳主题数目并能适应时间变化，但原理复杂，所需计算量大，对于非专业研究人员来说使用存在困难[18]。概率主题模型在主题分类后，需要人工对分类的主题进行标注，具有主观性，致使结果准确率不高。相对于其他挖掘方法，主题建模具有处理文本量大、无监督、准确率高等优点，但其召回率仍不够高，目前分析的文献数据中 AUC 值在 0.7～0.9 之间，结果只能为使用者提供参考，不能作为标准的衡量依据。其余运用主题建模方法来挖掘临床文本的不多，目前尚未有比 LDA、HDP 更优的主题建模挖掘临床文本的方法。

3.4.3 分词工具不成熟 由于临床文本主题建模前需要对文本进行切分、去停用词等预处理，临床文本的语言对此过程影响较大。现有的文本分词工具对英文处理较为成熟，一定程度上对英文临床文本处理的发展有所帮助。而其他语言因使用范围有限，语言本身存在的特点使分词工具开发不够成熟，如中文由汉字组成，而非字母，词与词之间没

有空格分开，词也没有形态上的变化[19]，对中文的文本挖掘处理就造成一定困难。目前较为常用的中文分词工具有中国科学院的 ICTCLAS、哈工大的 LTP、jieba 分词等[20]，对于像临床文本挖掘这类专业领域的文本处理效果还有待提升。

3.4.4 数据使用具有局限性 由于分析所使用的临床文本都来自于医院，包含大量患者的隐私信息。目前有很多研究者开始关注去除临床文本隐私，但大部分医院还没有开展该工作。出于法律对公民个人隐私的保护，临床文本信息不能向公众公开和共享使用[21]。目前的临床文本挖掘研究大多是高校与医院达成一致后，再抽取所需数据信息进行分析研究。而且医疗机构间的数据具有关联性，患者可能不仅在一所医疗机构中就医，这对于产生的结果准确性有一定影响，而同时获取不同医疗机构间的数据是非常困难的。由于临床文本的难获取性，不能广泛、联合运用临床文本进行信息挖掘，对发现更好的主题建模方法也存在一定阻碍。

4 讨论

4.1 主题建模在临床文本挖掘领域潜力巨大

根据文献分析结果可知目前主题建模方法在临床文本挖掘中还处于起步阶段，但其发展潜力巨大。通过主题建模处理大量无规则临床文本信息，减少因人工方式造成的不良后果。因为其属于非监督的机器学习技术，能够自动生成结果，减少人工消耗。在未来临床文本挖掘的主题建模中，模型会更具有普适性，不仅局限于一种疾病或数据形式，能动态考虑分析数据及其关联性，对于不同数据来源都能有稳定的输入结果，能结合不同领域的专业知识给出准确性更高的结果。主题模型在医疗领域的应用不仅局限于临床文本，还能处理医疗机构中的各类医疗卫生数据。

4.2 主题建模的不同方法各有其优势和不足

LDA 方法能够无监督地处理大量文本数据，拓展性高，但在构建时需要预先指定主题的数量，而模型质量直接依赖于主题数量的选取，具有较大的主观性和随机性[22]。相比 LDA，HDP 能够根据文本数据集自动确定主题数目，对于未见过的新文本能够产生新的主题，故其也能适合于对不断变化的文本集进行主题建模，但其原理和建模过程较 LDA 更为复杂，计算耗时长，所以 HDP 在临床文本挖掘中应用并没有 LDA 广泛。在临床文本的主题挖掘中，运用 LDA 及其衍生模型的方法较为普遍，HDP 使用率并不高，而其他主题建模方法在临床文本挖掘中尚未有深入应用。由此可见临床文本挖掘的主题建模方法仍有很大的发展空间。对于准确度不高的问题，可以对数据进行动态分析，考虑关联性，引入知识库，改进模型使其更具有普适性；对于分词工具不成熟的情况，除提升分词工具的性能外，还可结合专业知识来提升分词结果准确性；针对临床文本数据来源的局限性，可依法将隐私信息去除后供研究者申请使用，医疗机构间也需要加强信息的互联互通，实现数据共享。

5 结语

本文总结分析近 10 年主题建模方法在临床文本挖掘中的研究现状、存在的挑战，发现主题建模的数据挖掘方法在临床文本中的应用还处于起步阶段，模型结果准确度还有待提高，方法有待完善和发展。相信随着信息的技术不断发展，在医疗卫生信息化建设的浪潮下，能够克服目前存在的困难，更准确、便捷、高效地处理临床文本，为临床决策支持提供服务，推动医疗卫生事业信息化的发展。

参考文献

1 都丽婷，罗维，李磊，等. 临床文本自动去识别方法比较［J］. 医学信息学杂志，2017，38（4）：44－49.

2 Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assef, et al. A Brief Survey of Text Mining: classification, clustering and extraction techniques［EB/OL］.［2017－07－28］. https://arxiv.org/abs/1707.02919.

3 唐晓波，向坤. 基于 LDA 模型和微博热度的热点挖掘［J］. 图书情报工作，2014，58（5）：58－63.

4 Zhengxing Huang, Wei Dong, Huilong Duan. A Probabilistic Topic Model for Clinical Risk Stratification from Electron-

ic Health Records [J]. Journal of Biomedical Informatics, 2017, (58): 28 - 36.

5 Zhengxing Huang, Wei Dong, Lei Ji, et al. Discovery of Clinical Pathway Patterns from Event Logs Using Probabilistic Topic Models [J]. Journal of Biomedical Informatics, 2014, (47): 39 - 57.

6 刘玉文, 张钰, 杨枢. 基于 LDA 模型和电子病历的疾病辅助诊断方法 [J]. 宿州学院学报, 2017, 32 (2): 114 - 116, 124.

7 许珠香, 江弋. 基于潜在狄利克雷分配模型的医疗数据研究 [J]. 厦门大学学报 (自然科学版), 2013, 52 (3): 356 - 359.

8 A Rumshisky, M Ghassemi, P Szolovits, et al. Predicting Early Psychiatric Readmission with Natural Language Processing of Narrative Discharge Summaries [J]. Translational Psychiatry, 2016, (6): 1 - 5.

9 Li - wei Lehman, William Long, Mohammed Saeed, et al. Latent Topic Discovery of Clinical Concepts from Hospital Discharge Summaries of a Heterogeneous Patient Cohort [C]. Chicago: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2014.

10 殷良英, 董蔚, 黄正行, 等. 面向临床路径的医疗行为变化趋势检测与分析 [J]. 中国生物医学工程学报, 2015, 34 (3): 272 - 280.

11 Aaron Zalewski, William Long, Alistair E W Johnson, et al. Estimating Patient's Health State Using Latent Structure Inferred from Clinical Time Series and Text [C]. Florida: IEEE - EMBS International Conference on Biomedical and Health Informatics, 2017.

12 David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3 (4 - 5): 993 - 1022.

13 Cohen R, Aviram I, Elhadad M, et al. Redundancy - Aware Topic Modeling for Patient Record Notes [J]. PLoS ONE, 2014, 9 (2): e87555.

14 颜端武, 陶志恒, 李兰彬. 一种基于 HDP 模型的主题文献自动推荐方法及应用研究 [J]. 情报理论与实践, 2016, 39 (1): 128 - 132.

15 Caballero K, Akella R. Dynamic Estimation of the Probability of Patient Readmission to the ICU using Electronic Medical Records [C]. Chicago: AMIA Annual Symposium Proceedings, 2015.

16 Filip Gintera, Hanna Suominena, Sampo Pyysalob, et al. Combining Hidden Markov models and Latent Semantic Analysis for Topic Segmentation and Labeling: Method and clinical application [J]. International Journal of Medical Informatics, 2009, (78): e1 - e6.

17 Jonathan H Chen, Mary K Goldstein, Steven M Asch, et al. Predicting Inpatient Clinical Order Patterns with Probabilistic Topic Models vs Conventional Order Sets [J]. Journal of the American Medical Informatics Association, 2017, 24 (3): 472 - 480.

18 魏强, 金芝, 许焱. 基于概率主题模型的物联网服务发现 [J]. 软件学报, 2014, 25 (8): 1640 - 1658.

19 周小甲. 中文病历文本的时间信息提取研究 [D]. 杭州: 浙江大学, 2011.

20 魏光泽. 中文分词技术在搜索引擎中的研究与应用 [D]. 青岛: 青岛科技大学, 2016.

21 中国青年政治学院互联网法治研究中心. 封面智库. 中国个人信息安全和隐私保护报告 [R]. 北京: 中国政治青年学院, 2016.

22 廖列法, 勒孚刚, 朱亚兰. LDA 模型在专利文本分类中的应用 [J]. 现代情报, 2017, 37 (3): 35 - 39.