中文医学摘要主题建模方法评估*

王 凡 夏晨曦

（华中科技大学同济医学院医药卫生管理学院 武汉 430030）

〔摘要〕 以中文医学论文摘要为语料，采用 3 种特征筛选方法以及两种主题模型方法的组合进行主题建模，结果表明建模方法中 LDA 的预测能力和拟合度优于 CTM，而特征筛选方式中 IDF 拟合度、TF 预测能力更好。

〔关键词〕 主题模型；评估；中文；医学摘要

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2018.02.014

Evaluation of Topic Modeling Methods in Chinese Medical Abstracts *WANG Fan, XIA Chen－xi, School of Medical and Health Management of Tongji Medical College, HUST, Wuhan 430030, China*

〔**Abstract**〕 Taking abstracts of Chinese medical papers as corpus, the paper adopts the combination of 3 feature selection methods and 2 topic modeling methods to carry out topic modeling. The result shows that in the modeling methods, prediction ability and fitting degree of LDA are better than that of CTM, but in the feature selection methods, fitting degree of IDF and prediction ability of TF are better.

〔**Keywords**〕 Topic model; Evaluation; Chinese; Medical abstract

1 引言

网络信息产生大量文本数据，主题模型（Topic Model）作为挖掘文档隐含主题的概率统计模型已广泛应用于文本分析之中。主题模型是一种用于发现文档隐含主题的概率模型。主题以一系列相关词汇表述，词汇频率越高与主题越相关，即主题是词汇的条件概率分布。文档中各主题出现的概率不同，即文档是主题的概率分布。主题模型自动分析每个文档，以文档词汇的概率分布状况判断主题以及各主题的概率分布。主题模型是一种高效的、无监督的[1]，用于寻找语义主题[2]的文本处理方法，它能够处理词汇多义性[3]的问题，适用于各种语言。

主题模型这一构想最早在 1988 年提出，即隐性语义检索（Latent Semantic Indexing，LSI）[4]，通过奇异值分解（SVD）得到文档和主题、词和词义以及语义和主题的相关度。1990 年概率潜在语义检索（Probabilistic Latent Semantic Indexing，PLSI）[5]建立相似度规则和数据产生式模型，解决 LSI 缺乏统计学基础的问题。2003 年隐含狄利克雷分配（Latent Dirichlet Allocation，LDA）[6]融入贝叶斯层次模型，以超参数来控制参数。目前大部分主题模型由 LDA 发展而来，如相关主题模型（Correlated Topic Model，CTM）[7]，关系主题模型（Relational

〔修回日期〕 2017－10－16

〔作者简介〕 王凡，硕士研究生；通讯作者：夏晨曦，博士。

〔基金项目〕 中央高校基本科研业务费华中科技大学青年创新基金项目“面向社交网络的情感分析与观点挖掘方法研究”（项目编号：2015AC028）。

Topic Models, RTMs)[8]，多模态事件主题模型(Multi - modal Event Topic Model, mmETM)[9]等。主题模型向提高训练算法效率、跨语言处理等方面发展[10]。在中文医学领域，相关文献[11]基于临床数据进行主题建模来帮助临床诊断，也有研究通过对医学文献的主题建模来发现生物医学领域的主题热点[12]及疾病与主题间的关系[13]。

当前中文医学领域大多以修改主题模型的加权方式来提高主题建模性能，对医学领域的应用还缺乏全面、深入地认识和评估，对文本预处理方法，特别是特征筛选方法对主题建模性能的影响研究更为缺乏。中文医学文本作为专业文本，疾病名称、诊疗方式、药物名称等都有特定术语表达。语料作为影响主题模型效果的因素之一，不仅影响主题挖掘，还关乎计算性能。因此本研究将以中文医学文献摘要集为测试语料，对该领域的主题建模方法和特征筛选方法进行深入评估，以推进主题建模方法在中文医学领域的应用，更有效地利用中文医学领域的文本资源。

2 数据来源与研究方法

2.1 数据来源

相关文献[14]表明摘要作为语料在主题模型中表现良好。本研究将万方医学数据库以“病”为关键词采集到 27 779 篇文献摘要作为语料库，最短的摘要为 15 字（包括符号），最长的摘要为 2 840 字，平均摘要长度约 272 字，标准差为 160.892 6。主题模型[15]文档采用 342 篇文献摘要，经过预处理和特征词筛选后的特征数量为 1 461；也有文献[16]采用 PNAS 的 20 551 篇文摘，词汇总量为3 026 970。因此本研究采集 27 779 篇文献摘要，词汇总量为3 500 525，足以作为中文医学摘要语料进行研究。

2.2 研究方法

由于预处理和建模方法对主题模型的效果有重要影响，本研究对 3 种特征词过滤方法与两种主题模型的组合进行主题模型性能评估。3 种特征词过滤方法为常用的计算特征权重的方法[17]，即词频(TF)、逆文档频（IDF)、TFIDF。两种模型即 LDA 和 CTM，LDA 是最经典的主题模型，服从 Dirichlet 分布，CTM 则是服从 Logistic 正态分布。本研究采用困惑度（Perplexity)、拟合度（Log Likelihood)评价，公式如下：

$$\text{perplexity}(w) = exp\left\{-\frac{\log(p(w))}{\sum_{d=1}^{D}\sum_{j=1}^{V} n^{(jd)}}\right\} \quad \text{（公式 1）}$$

$$log(p < w \mid z >) = k\log\left(\frac{\Gamma(V\delta)}{\Gamma(\delta)^{V}}\right) + \sum_{K=1}^{k}\{[\sum_{j=1}^{V}\log(\Gamma(n_K^{(j)} + \delta))] - \log(\Gamma(n_K^{(j)} + V\delta))\} \quad \text{（公式 2）}$$

困惑度通过预测词的不确定性程度来表示模型生成词的性能，困惑度越小预测能力最强[18]。拟合度说明拟合程度，拟合度越大模型的拟合能力越强[19]。

3 实验流程与结果

3.1 实验流程

本实验流程，见图 1。（1）将文献摘要按照每行 1 篇的要求储存在 1 个文本文档中。（2）进行数据清洗及预处理，包括 4 个部分：①将 ICD - 10 疾病编码及诊断等医学词库导入用户词典，②使用 Rwordseg 分词同时实现中文分词和词性标记。③去停用词。④去除连词、副词、介词等虚词。文献[20]表明以名词、动词作为特征词更利于文本分类，因此本研究筛选词性为“n”（名词)、“vn”（动名词)、“v”（动词)、“en”（英文词汇）以及用户词典的词汇。（3）将筛选的词汇形成文本词矩阵。(4）进行特征词筛选，分别计算词汇的 TF、IDF 和 TFIDF，将低于权重 20% 的词汇剔除。（5）将 3 种筛选方法分别与 LDA 和 CTM 两种主题模型组合建模，共 6 种主题模型组合。（6）采用困惑度和拟合度评价模型的性能。

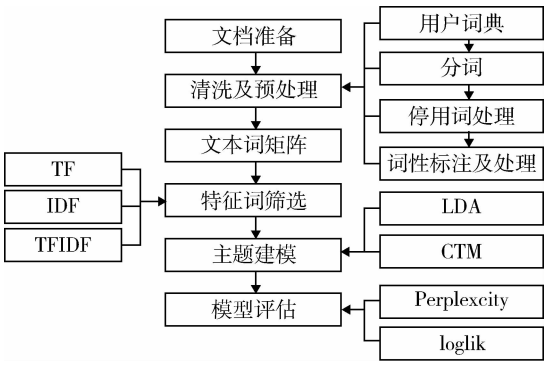

图 1 实验流程

3.2 结果

3.2.1 数据预处理 本研究共读取 22 738 条中文医学文献摘要，预处理结果，见表 1。去除停用词后，词汇量减少约 22%，词性筛选后词汇量减少约 45%，剩余 1 913 767 词。

表 1 数据预处理操作及结果

操作	结果
读取中文医学文献摘要	27 779 篇
中文分词及词性标注	27 779 篇；3 500 525 词
停用词处理	27 779 篇；2 740 004 词
词性筛选	27 779 篇；1 913 767 词

3.2.2 特征词筛选 3 种特征筛选方式的实验结果，见表 2，剔除 5% 特征权值较低的词汇，约减少 2 000 个特征词。TF 筛选下文档变化较小，说明 TF 筛选方式剔除的特征词分布比较分散，而 IDF 筛选方式减少的文档数量较多，说明 IDF 筛选方式下，有较多的文档特征词 IDF 值过低，导致文档被剔除。

表 2 特征词筛选结果

特征筛选	筛选范围	经过特征筛选的文本词矩阵结果
未筛选	–	27 779 篇；39 610 词
TFIDF	0.099 7	27 769 篇；37 640 词
TF	0.007 4	27 779 篇；37 601 词
IDF	9.149 5	27 432 篇；37 623 词

3.2.3 主题模型评估 主题模型根据困惑度和拟合度来进行评估。为将主题数量控制在合理范围内，实验以主题数量为 50、100、150 进行，见图 2。TFIDF 特征筛选的主题模型，困惑度随主题数量的增加而降低，拟合度随主题数量的增加而增加。LDA 的困惑度始终低于 CTM，且拟合度始终高于 CTM。因此主题数量为 50 ~ 150 时，采用 TFIDF 进行特征筛选的 LDA 是相比 CTM 而言预测能力和拟合程度更好的主题模型。TF 特征筛选的主题模型，CTM 的困惑度随主题数量的增加而增加，拟合度随主题数量的增加而降低，表明随着主题数量的增加，TF 特征筛选下的 CTM 的性能逐渐降低；TF 特征筛选下的 LDA 模型的性能随主题数量的增加而提高，LDA 的困惑度始终低于 CTM，拟合度始终高于 CTM。因此主题数量为 50 ~ 150 时，采用 TF 进行特征筛选的 LDA 是相比 CTM 而言预测能力和拟合程度更好的主题模型，主题数量为 150 时模型达到最优。IDF 特征筛选的主题模型，CTM 的困惑度低于 LDA，拟合度高于 LDA，说明在主题数为 50 ~ 100 时，CTM 的预测能力和模型拟合程度优于 LDA。但随着主题数量增加到 150 时，LDA 的困惑度低于 CTM，拟合度高于 CTM，此时模型达到最佳状态。

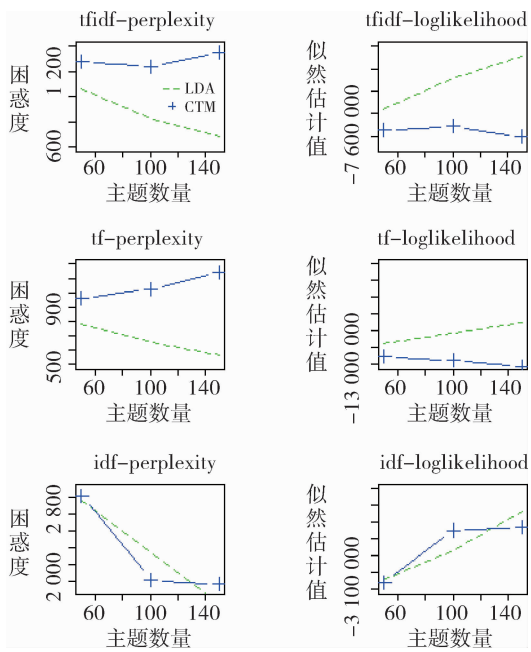

图 2 主题模型评估

4 讨论

4.1 不同预处理方法的影响

4.1.1 用户词典 有助于提高分词的准确性。由于中文医学文献专业词汇较多，而分词工具通常用于新闻语料，因此缺少用户词典时会降低中文医学文献分词的准确性。缺少用户词典导致专业词汇切分细碎、语义模糊，而使用用户词典时专业词汇划分较准确，见表 3。

表 3 用户词典分词效果

例句	观察联合应用 - 氨基酮戊酸光动力疗法（ALA - PDT）与放射疗法（RT）治疗老年 Bowen 病（BD）的疗效
无用户词典分词结果	“观察”、“联合”、“应用”、“氨基”、“酮”、“戊”、“酸”、“光”、“动力”、“疗法”、“ALA”、“PDT”、“与”、“放射”、“疗法”、“RT”、“治疗”、“老年”、“Bowen”、“病”、“BD”、“的”、“疗效”
有用户词典分词结果	“观察”、“联合”、“应用”、“氨基酮戊酸”、“光动力疗法”、“ALA”、“PDT”、“与”、“放射疗法”、“RT”、“治疗”、“老年”、“Bowen”、“病”、“BD”、“的”、“疗效”

4.1.2 词性筛选 词性不仅有助于降低语料数据，还能提高主题模型的性能[21]。Rwordseg 所标注的汉语词性共有 40 种，如助词、代词、形容词、连词等[18]意义较小，通过词性筛选减少 45% 的数据量，不仅降低噪音词汇对主题建模的影响，还减小了语料库的冗余和计算量。

4.1.3 特征筛选 预处理后文本以文本词条矩阵表示并进行词频统计。矩阵中存在一些没有区分度的词汇，因此对矩阵降维能提高模型准确性和减少计算复杂度。研究采用 3 种特征降维方法，使矩阵的词汇减少约 2 000 个，见图 3。以 TF 进行筛选时困惑度最低，但拟合度也最低，说明 TF 特征筛选是预测能力最好但拟合度最差的方式；而以 IDF 进行筛选时困惑度最高、拟合度最好。因此需中文医学摘要主题模型有较好的预测能力时，可以选择 TF 特征筛选方式；如需主题模型有较高的拟合度时，可采用 IDF 特征筛选办法。

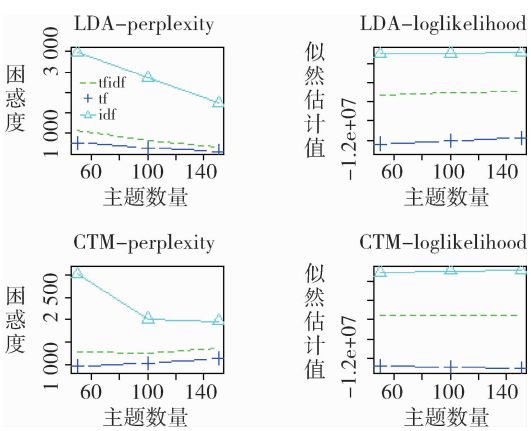

图 3 不同特征筛选的建模效果评价

4.2 不同主题建模方法的影响

本研究采用两种主题建模方法，LDA 是较 CTM 而言预测能力及模型拟合程度更好的主题模型。虽然 CTM 能够将维度较大的文本数据转化为低纬度的数据[22]，但是对已进行过降维处理的文本数据，CTM 在中文医学摘要的主题建模能力还是低于 LDA。CTM 考虑了主题的相关性，增加了模型的复杂度。主题数量是影响主题建模的重要因素。由于 LDA 和 CTM 的主题数量没有确定的标准，本研究仅说明主题数量在 50 ~ 150 时主题模型的性能，不能排除在主题数量更大时 CTM 性能将会高于 LDA 的可能性。

4.3 与现有研究结果或结论的异同

相关文献[23]以新闻为语料，对 LDA、OnlineLDA、CTM 进行测试，通过人工判断的方式来评价 3 种模型的性能，认为 LDA 主题模型得出的主题词汇易于理解和区分。本研究是针对中文医学领域的研究，以中文医学文献摘要作为语料，对 LDA 和 CTM 进行比较得出 LDA 的预测能力和拟合程度优于 CTM。Jian Tang 等[24]通过设置参数来研究主题模型的影响因素，研究表明 LDA 的性能受文档数量、长度、超参数及主题数量的影响。本研究从模型的预测能力和拟合程度来判断主题模型的性能，主题数量的大小在不同的主题模型中的影响不同，不是数量越大性能越低，但是主题数量越大模型的

计算复杂度越高，降低主题模型收敛的速度。

5 结语

以中文医学文献摘要为测试语料，对中文医学领域的主题建模方法和特征筛选方法进行了深入评估。从结果来看，LDA 是较 CTM 更适用的主题模型，其预测能力和拟合程度都优于 CTM；采用 IDF 特征筛选拟合度更好，TF 筛选预测能力更好，若同时考虑预测能力和拟合度，则采用 TFIDF 性能更佳。中文医学领域的主题建模方法评估在未来还有很大的空间，在此只针对最常用 3 种特征筛选方式和两种主题模型、使用文献摘要做为测试语料进行研究。还有更多的、改进的特征筛选方法和主题模型可以纳入研究范围，其他类型的中文医学语料也可能会带来不同的结果。

参考文献

1 李晨曦，谢罗迪. 基于 LDA 模型的文本分类与观点挖掘［J］. 电子技术与软件工程，2017，(4)：209 - 210.

2 袁芳. 基于语义分析的文本检索模型技术研究［D］. 武汉：华中师范大学，2016.

3 姜在兴. 面向中医临床处方分析的主题模型研究［D］. 北京：北京交通大学，2013.

4 Deerwester S, Dumais S T, Furnas G W, et al. Indexing by Latent Semantic Analysis［J］. Journal of the American Society for Information Science, 1990, 41 (6)：391 - 407.

5 Hoffman T. Unsupervised Learning by Probabilistic Latent Semantic Indexing［J］. Sigir Audit Reports, 1999, 40 (22)：28 - 31.

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

7 David M Blei, John D Lafferty. A Correlated Topic Model of Science［J］. Annals of Applied Statistics, 2007, 1 (1)：17 - 35.

8 N Chen, J Zhu, F Xia, et al. Discriminative Relational Topic Models［J］. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37 (5)：973 - 986.

9 S Qian, T Zhang, C Xu, et al. Multi - Modal Event Topic Model for Social Event Analysis［J］. IEEE Transactions on Multimedia, 2016, 18 (2)：233 - 246.

10 徐戈，王厚峰. 自然语言处理中主题模型的发展［J］. 计算机学报，2011，34 (8)：1423 - 1436.

11 王瑶. 朱建贵教授调气化痰治疗失眠症经验总结及温胆汤临床疗效观察［D］. 北京：中国中医科学院，2015.

12 佘玉轩，熊赟. Medas：一个基于 Medline 的生物医学文献分析系统［J］. 计算机研究与发展，2015，52 (S1)：102 - 106.

13 崔明亮. 基于主题模型的生物医学文献知识发现［D］. 长春：吉林大学，2017.

14 关鹏，王曰芬，傅柱. 不同语料下基于 LDA 主题模型的科学文献主题抽取效果分析［J］. 图书情报工作，2016，(2)：112 - 121.

15 Grun B, Hornik K. Topicmodels：an R package for fitting topic models［J］. Journal of Statistical Software, 2011, 40 (13)：2011.

16 Griffiths TL, Steyvers M. Finding Scientific Topics［J］. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101 (1)：5228 - 5235.

17 孙挺，耿国华，周明全. 一种有效的特征权重计算方法［J］. 郑州大学学报（理学版），2008，40 (4)：48 - 51.

18 彭时名. 中文文本分类中特征提取算法研究［D］. 重庆：重庆大学，2006.

19 Brown P F, Pietra V J D, Mercer R L, et al. An Estimate of an Upper Bound for the Entropy of English［J］. Comoutational Linguistics, 1992, 18 (1)：31 - 40.

20 张勇. 基于词性与 LDA 主题模型的文本分类技术研究［D］. 合肥：安徽大学，2016.

21 程彬彬. 词性在汉语科技文献检索中的作用与影响［D］. 南京：南京农业大学，2008.

22 王燕霞. 基于相关主题模型的文本分类方法研究［D］. 苏州：苏州大学，2010.

23 赵云. 主题模型的评价方法研究［D］. 大连：大连海事大学，2014.

24 Tang J, Meng Z, Nguyen X, et al. Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis［C］. Proceedings of the 31st International Conference on Machine Learning, 2014, 32 (1)：190 - 198.