

硕士学位论文

**基于主题模型的病历文本特征提取的研究与应用**

学位申请人：崔喆森

指导教师：饶元 副教授

类别（领域）：工程硕士（软件工程）

2017年4月

**Research and Application of Medical Records Feature Extraction Based on Topic Model**

A thesis submitted to

Xi’an Jiaotong University

in partial fulfillment of the requirements

for the degree of

Master of Engineering

By

Zhesen Cui

Supervisor: Assoc.prof Yuan Rao

(Software Engineering)

April 2017

**论文题目：基于主题模型的病历文本特征提取的研究与应用**

**类别（领域）：工程硕士（软件工程）**

**学位申请人：崔喆森**

**指导教师：饶元副教授**

摘 要

医疗领域的信息化发展给医疗文本的分析带来了很多挑战，尤其是蕴含丰富医疗信息的病历文本。在文本挖掘领域中主题模型得到了广泛的应用和研究，它能够挖掘出文本在主题上的概率分布，并且在复杂的句法结构下仍能够达到不错的效果。所以，主题模型很适合处理病历文本，本文的工作概括如下：

首先，分析了层次狄利克雷过程主题模型存在的问题，并提出了一种改进的主题模型，即基于特征加权机制的层次狄利克雷过程。随后在病历文本数据集上进行模型训练，验证了改进后的模型在主题独特性和模型混乱度两个指标上有明显提高。在得到改进主题模型之后，本文分析了文本主题标签提取问题，原有的主题标签提取算法多用到语义相似度进行提取，这样会导致标签区分度不好，所以本文提出了综合考虑语义相似度和主题区分度的标签提取算法，利用其对病历文本的主题生成标签，生成的标签能够较好的代表当前主题。研究了病历文本挖掘中的内容检索问题，也就是相关项提取问题，本文通过计算文档词语关联度，利用主题模型生成与病历文本相关的检查项与治疗项，并且对比了模型改进前后在F值上的性能；最后，利用得到的研究成果设计并且实现了一套病历服务系统，其中包括文本训练模块，离线计算模块，实时计算模块。并对系统进行功能测试，结果表明系统能够满足需求。

综上所述，本文以病历文本数据集为语料库进行了模型训练，验证了提出的改进模型在性能上的提高，并且对相关的特征提取算法进行了改进，取得了一定的成果，具有一定借鉴意义和价值。

**关 键 词**：病历文本；主题模型；特征提取；语义关联度；主题区分度

**论文类型**：应用研究

**Title:** **Research and Application of Medical Records Feature Extraction Based on Topic Model**

**Professional Fields: Software Engineering**

**Applicant: Zhesen Cui**

**Supervisor: Assoc.prof Yuan Rao**

ABSTRACT

The development of information technology in the medical field has brought many challenges to the analysis of medical texts and medical records which have loose grammatical structure , diverse medical term and rich information.The problem brings many difficulties for the treatment of medical texts. Topic modeling is widely researched and applied in the field of text mining. It can dig out probability distribution of text on topic, can also dig out probability distribution of topic on terms, and still be able to achieve good effect in the complex syntactic structure. Therefore, Topic modeling is very suitable to deal with medical text, to help medical staff and related researchers. The major works are summarized as follows.

Firstly,In this thesis we analyze the characteristics of medical texts and propose an improved topic model,which is named Hierarchical Dirichlet Process model based on BM25F weighted.We complete the training and verify the effectiveness of this new model in text sets of medical records.Then the thesis analyzes the problems in tag generation and proposes a tag extraction model which gives a summary for the topic of medical records. We also study the Information Retrieval problem in text mining of medical records. this thesis explores the related items generation problem,according to some medical symptoms and physical condition in the text, the topic model generates related diagnosis items and treatment items based on semantic association as evaluation index, and we test performance in F-measure.Finally we build a medical records service system based on the research results talked above.This system provides the keyword extraction service,medical records summary extraction service,medical records matching service and related items extraction service.The system is tested and the results show that the system can meet the requirements.

In summary,according to the model training based on the medical records sets, we verify the improvement in the performance of the proposed model.The characteristics of medical records are extracted.this thesis has obtained certain achievement, has certain significance and value.

**KEY WORDS**: Medical Records; Topic Model; Feature Extraction; Semantic Association; Topic Discrimination

**TYPE OF THESIS**: Application Research

# 

目 录

[1 绪论 1](#_Toc482867631)

[1.1 研究的背景与意义 1](#_Toc482867632)

[1.2 国内外研究状况分析 2](#_Toc482867633)

[1.3 论文的主要研究内容 3](#_Toc482867634)

[1.4 论文的组织结构 4](#_Toc482867635)

[2 相关理论与技术 5](#_Toc482867636)

[2.1 潜在语义分析 5](#_Toc482867637)

[2.1.1 模型概述 5](#_Toc482867638)

[2.1.2 理论过程 5](#_Toc482867639)

[2.2 概率潜在语义分析 8](#_Toc482867640)

[2.2.1 模型描述和参数估计 8](#_Toc482867641)

[2.2.2 模型特性 9](#_Toc482867642)

[2.3 潜在狄利克雷分配 10](#_Toc482867643)

[2.3.1 共轭先验和狄利克雷分布 10](#_Toc482867644)

[2.3.2 文本建模 10](#_Toc482867645)

[2.3.3 模型特性 11](#_Toc482867646)

[2.4 层次狄利克雷过程 11](#_Toc482867647)

[2.4.1 模型概述 11](#_Toc482867648)

[2.4.2 狄利克雷过程 12](#_Toc482867649)

[2.4.3 层次狄利克雷过程的概率模型 13](#_Toc482867650)

[2.5 本章小结 14](#_Toc482867651)

[3 层次狄利克雷过程主题模型改进 15](#_Toc482867652)

[3.1 问题提出 15](#_Toc482867653)

[3.2 主题倾向性优化的HDP模型 16](#_Toc482867654)

[3.2.1 HDP模型引入BM25F和高斯函数 17](#_Toc482867655)

[3.2.2 改进模型的Gibbs采样 19](#_Toc482867656)

[3.3 实验内容与结果分析 20](#_Toc482867657)

[3.3.1 实验环境与评价指标 21](#_Toc482867658)

[3.3.2 文本数据集描述 22](#_Toc482867659)

[3.3.3 文本数据集预处理 23](#_Toc482867660)

[3.3.4 LDA、HDP和BM25F-HDP实验对比 24](#_Toc482867661)

[3.4 本章小结 28](#_Toc482867662)

[4 基于层次狄利克雷过程主题模型的病历文本特征提取 29](#_Toc482867663)

[4.1 病历文本特征提取问题分析 29](#_Toc482867664)

[4.2 病历文本的主题标签提取算法 31](#_Toc482867665)

[4.2.1 改进的主题标签提取算法 31](#_Toc482867666)

[4.2.2 主题标签提取算法实验设计及结果分析 33](#_Toc482867667)

[4.3 病历文本的相关项提取算法 35](#_Toc482867668)

[4.3.1 基于BM25F-HDP的相关项提取算法 35](#_Toc482867669)

[4.3.2 相关项测试方法 36](#_Toc482867670)

[4.3.3 相关项测试实验 37](#_Toc482867671)

[4.4本章小节 40](#_Toc482867672)

[5 病历服务系统的实现与测试 41](#_Toc482867673)

[5.1 需求背景 41](#_Toc482867674)

[5.2 系统需求分析 41](#_Toc482867675)

[5.3 系统设计 45](#_Toc482867676)

[5.3.1 系统架构设计 45](#_Toc482867677)

[5.3.2 系统功能模块设计 46](#_Toc482867678)

[5.3.3 数据库设计 48](#_Toc482867679)

[5.4系统各模块的实现 49](#_Toc482867680)

[5.4.1 文本训练模块 49](#_Toc482867681)

[5.4.2 离线计算模块 51](#_Toc482867682)

[5.4.3 实时计算模块 53](#_Toc482867683)

[5.5系统测试 56](#_Toc482867684)

[5.5.1 系统测试环境 56](#_Toc482867685)

[5.5.2 系统功能测试 56](#_Toc482867686)

[5.6本章小结 57](#_Toc482867687)

[6 结论与展望 58](#_Toc482867688)

[6.1 结论 58](#_Toc482867689)

[6.2 展望 58](#_Toc482867690)

[致 谢 60](#_Toc482867691)

[参考文献 61](#_Toc482867692)

声明

CONTENTS

[1 Preface 1](#_Toc478417820)

[1.1 Background and Significance of the Research 1](#_Toc478417821)

[1.2 Research Status 2](#_Toc478417822)

[1.3 Main Contents 3](#_Toc478417823)

[1.4 Structure of this Thesis 4](#_Toc478417824)

[2 Related Theory and Technology 5](#_Toc478417825)

[2.1 Latent Semantic Analysis 5](#_Toc478417826)

[2.1.1 Overview of Model 5](#_Toc478417827)

[2.1.2 Process of Theory 5](#_Toc478417828)

[2.2 Probabilistic Latent Semantic Analysis 8](#_Toc478417829)

[2.2.1 Overview of Model and Parameter Estimation 8](#_Toc478417830)

[2.2.2 The characteristics of Model 9](#_Toc478417831)

[2.3 Latent Dirichlet Allocation 10](#_Toc478417832)

[2.3.1 Conjugate Prior and Dirichlet distribution 10](#_Toc478417833)

[2.3.2 Text Modeling 10](#_Toc478417834)

[2.3.3 The Characteristics of Model 11](#_Toc478417835)

[2.4 Hierarchical Dirichlet Process 11](#_Toc478417836)

[2.4.1 Overview of Model 11](#_Toc478417837)

[2.4.2 Dirichlet Process 12](#_Toc478417838)

[2.4.3 Probabilistic Model of Hierarchical Dirichlet Process 13](#_Toc478417839)

[2.5 Chapter Summary 14](#_Toc478417840)

[3 The Improvement of Hierarchical Dirichlet Process 15](#_Toc478417841)

[3.1 Problems in Discussion 15](#_Toc478417842)

[3.2 The Improvement of Topic Orientation in HDP 16](#_Toc478417845)

[3.2.1 BM25F and Gaussian Function in HDP 17](#_Toc478417847)

[3.2.2 Gibbs Sampling of BM25F-HDP 19](#_Toc478417848)

[3.3 Experiment of Improved HDP Model 20](#_Toc478417849)

[3.3.1 Environment and Evaluation Index 21](#_Toc478417850)

[3.3.2 The Description of Text data set 22](#_Toc478417851)

[3.3.3 The Pretreatment of Text data set 23](#_Toc478417851)

[3.3.4 Comparison of LDA, HDP and BM25F-HDP 24](#_Toc478417852)

[3.4 Chapter Summary 28](#_Toc478417853)

[4 Research on Feature Extraction of Medical Records Based on Topic Model 29](#_Toc478417854)

[4.1 Analysis on Feature Extraction of Medical Records 29](#_Toc478417855)

[4.2 Tag Extraction Algorithm of Medical Records 31](#_Toc478417856)

[4.2.1 The Improvement of Tag Extraction Algorithm 31](#_Toc478417857)

[4.2.2 Results and Analysis of Tag Extraction Algorithm 33](#_Toc478417857)

[4.3 Correlation Extraction Algorithm of Medical Records 35](#_Toc478417867)

[4.3.1 Correlation Extraction Algorithm Based on BM25F-HDP 35](#_Toc478417868)

[4.3.2 Correlation Test Method 36](#_Toc478417869)

[4.3.3 Correlation Test Experiment 37](#_Toc478417870)

[4.4 Chapter Summary 40](#_Toc478417871)

[5 Medical Record Service System 41](#_Toc478417872)

[5.1 Requirements Background 41](#_Toc478417873)

[5.2 Requirements Analysis of System 41](#_Toc478417874)

[5.3 Design of System 45](#_Toc478417875)

[5.3.1 System Architecture Design 45](#_Toc478417876)

[5.3.2 System Module Design 46](#_Toc478417877)

[5.3.3 Database Design 48](#_Toc478417878)

[5.4 Implementation of Main Moduals 49](#_Toc478417879)

[5.4.1 Text Training Module 49](#_Toc478417880)

[5.4.2 Offline Computing Module 51](#_Toc478417881)

[5.4.3 Real-time Computing Module 53](#_Toc478417882)

[5.5 The Testing of System 56](#_Toc478417883)

[5.5.1 Testing Environment of the System 56](#_Toc478417884)

[5.5.2 Function Testing of the System 56](#_Toc478417885)

[5.6 Chapter Summary 57](#_Toc478417887)

[6 Summary and Outlook 58](#_Toc478417888)

[6.1 Summary 58](#_Toc478417889)

[6.2 Outlook 58](#_Toc478417890)

[Acknowledgements 60](#_Toc478417891)

[References 61](#_Toc478417892)

Declaration

# 绪论

## 研究的背景与意义

随着现在医学的研究和实验的快速发展，产生了大量的和医学有关的文本，面对如此多的医学文本数据，如何能够从中准确而又高效的获取关键信息成为了一个需要解决的问题[1]。而且不同医疗人员对于相同医学概念的描述可能是不一样的，这就会造成医学文本中会存在一些歧义、矛盾和冗余，从这些文本中获取关键信息变的更加困难。如果人工从这些医学文本中提取相关特征会非常耗时，而且人工成本高，所以不太现实。利用计算机自动化实现这些工作具有重要的意义。

对于病历文本来说，也面临着上述提到的问题。病历文本是医疗人员对患者疾病的发生、发展、转归，进行检查、诊断、治疗等医疗活动过程的记录[2]。也是对病历文本加以归纳、总结、分析，按照规范的要求和格式书写的患者医疗健康档案。病历是临床试验的归纳总结，对医疗、科研等有着重要的作用。由于病历文本是主治医师对患者的主观描述，语法结构会不尽相同，通常缩略语较多，或者短语拼凑在一起，对相同的医学概念描述也可能不一样。所以，从病历文本中提取某项医学概念的数值信息比较困难，传统的数据挖掘方法也不能取得很好的效果[3]。可是病历文本的特征提取对于医疗人员来说有着很重要的意义，比如医师需要对比相似病历来发现一些医学规律；从大量的病历文本中能够快速的获得病历的关键信息；从病历库中提取出和目前症状最接近的检查项和治疗项，提高医师的诊断效率等[4]。如果采用计算机自动化处理上述问题，利用主题模型为医疗人员提供良好的解释性信息，将会大大提高工作效率并且保证了准确度。

主题模型能够发现文本中的潜在主题，在各种文本挖掘领域中广泛应用。如果让医疗人员对病历文本进行编辑分类，可能会带来一些问题，比如主观的判断，很难给一篇病历文本以多个分类属性，分类维度无法控制，也无法衡量文本在某一个分类中的权重，如果使用主题模型的话就会解决上述的问题，这是因为主题模型从数据出发，是基于统计学的聚类。

主题模型不需要定量的测量指标，能够直接从病症描述和医学名词出发，找到病症描述和病症描述、病症描述和医学名词以及医学名词和医学名词之间的关联关系，这就使得病历文本能够在主题层面定量分析，也就是让病历文本在主题层面结构化。所以，以主题模型为基础进行研究，使得病历文本的特征提取研究与应用变的更加准确和容易。

本文针对病历文本的特点，提出改进的层次狄利克雷过程主题模型，将病历文本在主题上进行结构化，使得后续的特征提取工作能够顺利展开，最终建立一个面向医疗人员的病历服务系统，这对于主题模型在医学文本领域中的应用具有重要的意义。

## 国内外研究状况分析

提到主题模型，要从潜在语义分析技术谈起，其研究是从潜在语义分析(Latent Semantic Analysis，LSA)[5]开始的，LSA是S.T. Dumais等人在1988年提出的一种信息检索代数模型，其通过对文档-词项矩阵的奇异值分解(Singular Value Decomposition，SVD)[6]，将高维度的空间向量模型(Vector Space Model，VSM)[7]中的文档映射到低维度的潜在语义空间，获得潜在语义信息，其优点是降维，去除了部分噪声，充分利用了冗余数据，无监督并且与语言无关。由于LSA不能够解决一词多义的问题，没有刻画出概率模型，SVD的计算复杂度高，新文档的到来还需要重新训练模型。所以Hofmann于1999年在LSI的基础上，提出了概率潜在语义分析(probabilistic Latent Semantic Analysis，pLSA)[8]，其思想类似于LSA，但引入了潜在类别，用概率模型的方式描述LSA，是一种更先进的模型。其定义了概率模型[9]，并且概率分布有明确的物理解释，其隐含的多项分布更符合文本特性，优化目标是相对熵[10]最小，可以利用混乱度[11]评价指标确定主题个数。pLSA解决了一词多义的问题，使用强化的期望最大化算法(Expectation-maximization，EM)[12]训练潜在类，相比于LSA,有了更加坚实的统计学基础。

可是pLSA缺点也明显，在文档层面没有提供合适的概率模型，使得pLSA不是完备的生成式模型，文档的增加也会导致模型的线性增加，变的越来越大，EM算法也是其缺点，需要反复迭代，计算量很大。于是[David Blei](http://www.cs.princeton.edu/~blei/)等人于2003年提出了潜在狄利克雷分配(Latent Dirichlet Allocation ，LDA)[13]，LDA克服了pLSA的缺点，解决了文档生成概率的问题，是完备的生成模型[14]，还可以对不在训练样本中的文本进行语义识别，利用吉布斯采样(Gibbs Sampling)[15]训练潜在类。但也有其缺点，由于LDA参数贝叶斯模型的特性，主题个数需要提前给出，而主题个数的选择对模型训练的结果有很大的影响，如何确定主题个数是LDA需要解决的问题，LDA投射空间的维数也受到样本数量限制。所以[Yee Whye Teh](http://www.stats.ox.ac.uk/~teh/)等于2005年提出了层次狄利克雷过程(Hierarchical Dirichlet Process，HDP)[16],其在狄利克雷过程 (Dirichlet Process，DP)[17]的基础上，使用中国餐馆过程[18]构造狄利克雷过程，其可以自动学习主题个数，使用一个接近于无限的概率测度[19]取代LDA中的主题混合，根据不同的混合比例为每个文本建立新的狄利克雷过程，所有的狄利克雷过程分享提取的主题，但不同的狄利克雷过程拥有的混合比例不同，可以从上层狄利克雷过程中抽取。pLSA的使用目前一定很少，目前国内外对主题模型的研究主要集中在LDA和HDP这一部分，这些模型将文本看作是单词的组合。

目前主题模型主要有三种用途：

1) 从文本中发现主题[20]，可以直接使用文本在主题上的分布和主题在词项上的分布进行进一步的文本处理。

2) 将文本表示为低维度的主题分布形式[21]，利用文本的主题分布将文本映射到主题空间，支持文本在主题上的相似度匹配[22]。

3) 可以用到信息检索领域中[23]的扩展查询。

在医学领域中，主题模型的应用研究还是一个相对较新的话题，Rada等人于1989年提出了一个基于主题网络的的主题距离度量方法，将很多医学相关词汇作为主题网络的组成部分，并且以层次结构进行整理。Lord等人[24]于2003年将WordNet[25]的相似度度量方法应用在基因功能注解中，利用主题层级结构建立了基因和细胞产品的联系。这些在医学领域的主题模型研究都是基于人工标注的主题网络之上，主题模型能够在已有的语料中找到潜在主题分布关联。Chute等人曾经利用LSI对病历进行分类和检索，他们使用了UMLS(Unified Medical Language System)[26]，在之后的研究中也都是关于LSI的一些应用。

国内，医学领域主题模型的研究与应用并不多，张小平[27]在2011年发表的博士论文中，对中医临床诊疗中潜在狄利克雷分布的应用进行了详细的研究。使用特征加权的思想消除高频词对描述病症词汇的影响，取得了很好的效果。李博[28]在2012年提出一种基于主题模型的主题树的医学文本自动标签方法，目的是提取文本中主题结构信息，然后利用主题结构对文本作自动标注。可是病历文本有其独特性，语法结构松散，短语居多，很难提取语法结构信息来支持主题分析。

## 论文的主要研究内容

本文分析了病历文本的特点，利用主题模型将病历文本在主题上结构化，从而使得之后的文本特征提取能够以主题模型为基础，从病历文本中提取出有用的医疗信息。最终，本文建立了一套面向医疗人员的病历服务系统。这个系统的建立以病历文本的固有特点为基础，提供了相关服务。本文的主要工作内容有以下几个方面：

1）本文分析了层次狄利克雷过程主题模型存在的一些问题，提出了改进的主题模型BM25F-HDP。原始模型在Gibbs采样中，若要得到对应主题，会受到词项在主题中的概率和词项对应的文本所拥有的主题概率两个因素影响，又由于高频词在文本和主题中所占的比例较大，而且每个词项都是按照相同的权重计算条件概率的，导致给词项分配新主题时会向高频词对应的主题倾斜。为了降低高频词对模型的影响，本文借鉴了特征加权的概念，应用到HDP模型中，得到改进的主题模型，并且通过实验得出，本文改进之后的模型在性能上有了较大的提升。

2）提出一种综合考虑语义相似度和主题区分度的主题标签提取算法，对病历文本主题进行标签提取。原有的主题标签提取算法多用到语义相似度进行提取，这样会导致标签区分度不好，所以在这里引入主题区分度的概念，使得主题的标签具有独特性，和别的主题语义相似度较低。针对病历文本的特点，还研究了其信息检索，具体就是基于语义相关度的病历相关项的提取。本文提出在主题模型的基础上，利用文档词语关联度为相关项和病历文本的相关性进行打分，以此为依据给出病历文本最相关的相关项，相关项包括检查项和治疗项，并且对比了HDP模型和BM25F-HDP模型在此应用上的性能。证明了改进后模型在平衡F值上要优于HDP模型。

3）在改进的主题模型和相关特征提取算法的基础上，设计了一套病历服务系统，该系统实现了病历文本主题标签提取，病历总结提取，相似病历匹配，相关项提取等功能。并对系统进行了测试，测试结果表明系统已经达到了功能要求。

## 论文的组织结构

1. 绪论

本章简单针对本文的应用背景，选题意义和应用价值进行了介绍。 描述了主题模型的研究与应用在国内外的发展现状，最后说明了本文的主要研究内容，并给出了本文的结构框架。

1. 相关理论与技术

本章主要对主题模型的各项技术和背景进行了介绍，分别介绍了潜在语义分析、概率现在语义分析、潜在狄利克雷分配和层次狄利克雷过程，分析了这四种模型的使用情况和优缺点。

1. 层次狄利克雷过程主题模型改进

本章阐述了病历文本的特点，并且分析了已有主题模型HDP中存在的不足，引入加权特征的思想对其进行改进，提出了具有针对性的BM25F-HDP模型，并进行了对比实验。

1. 基于层次狄利克雷过程主题模型的病历文本特征提取

本章在主题模型的基础上，提出了综合考虑语义相似度和主题区分度的病历文本主题标签提取算法，还进行了相关项提取实验，并且测试了相关项生成算法，对比了HDP模型和BM25F-HDP模型在相关项生成上的性能。

1. 病历服务系统的实现与测试

本章详细的阐述了病历服务系统的需求分析、设计与实现，并且对系统进行了功能测试。

1. 结论与展望

对论文所做工作进行了总结，并对今后的研究工作提出了展望。

# 相关理论与技术

主题模型(Topic Models)是一个在大量文档中尝试发现潜在主题的机器学习模型,通过分析文本中的词项来发现文本中主题之间的关联和发展。利用主题模型我们可以挖掘文本中的信息。主题模型有潜在语义分析、概率潜在语义分析、潜在狄利克雷分布、层次狄利克雷过程以及很多衍生模型，本章将具体介绍这些模型。

## 潜在语义分析

### 2.1.1 模型概述

潜在语义分析(Latent Semantic Analysis，LSA)是在自然语言处理中应用很多，其通过矢量语义空间来分析文档与词之间的关系。LSA设定，如果两个词项多次出现在同一文档中，则这两个词项在语义上具有相似性。LSA使用大量的文本构造矩阵，其中行代表词项，列代表文档，矩阵中的值代表词项在文档中出现的次数，再将矩阵进行奇异值分解（SVD），在保留列信息的情况下将矩阵行数减少，再利用余弦相似度计算每两个词项之间的相似性。相似度越接近于1说明两个词语越相似，反之则越不相似。

LSA 使用词项-文档矩阵(Occurences Matrix)描述词项是否在一篇文档中。其一般是稀疏矩阵，矩阵中的元素可以是词频，也可以是词语的权重值，如tf-idf [29]。词项-文档矩阵和传统的主题模型相比没有本质区别，只是传统的语义模型不使用矩阵进行这样的描述。

### 2.1.2 理论过程

1）降维

在构建好词项-文档矩阵之后，LSA将对其进行降维，找到一个低阶近似。降维的原因有如下几点：

（1）一般构造的词项-文档矩阵太大会导致计算机处理溢出，将为后的计算将更为实际和快捷。

（2） 一般构造的词项-文档矩阵中有噪音，降维后的矩阵就是一个去噪矩阵。

（3）一般构造的词项-文档矩阵过于稀疏。

（4）一般构造的词项-文档矩阵能够反映词项属于某一个矩阵，而关键点往往在于和一篇文档相关的所有词项是什么。

降维的结果就是语义相关的词汇会合并。而且可以解决一部分同义词和二义性问题。

因此我们需要发掘一个词的各种同义词的情况。降维的结果是不同的词或因为其语义的相关性导致合并，降维可以解决一部分同义词的问题，也能解决一部分二义性问题。

2）推导

假设X是词项-文档矩阵，其元素xi,j表示词项i在文档j中的词频，则X矩阵如下(2-1)所示：

 (2-1)

一行代表一个词项向量，式(2-2)描述了此词项在所有文档上的映射。

 (2-2)

一列代表一个文档向量，式(2-3)描述了此文档在所有词项上的映射。

 (2-3)

词向量可以表示这两个词项在文档集合上的相似性。矩阵 中的元素表示是词向量的点乘结果，按对角线对称的元素具有相同的值，表示的是两个词项的相似度。类似的，矩阵中的元素表示文档向量的点乘结果，也就是每两篇文档之间的相似度。

现在将矩阵X进行奇异值分解(SVD)，分解成正交矩阵U、V和对角矩阵\Sigma的乘积。如下(2-4)所示：

 (2-4)

词项和文档的相关性矩阵如下(2-5)和(2-6)所示：

 (2-5)

 (2-6)

由于和是对角矩阵，所以U是由的特征向量组成的矩阵，V是由特征向量组成的矩阵。这些特征向量对应的特征值就是中的元素。综上所述，

奇异值分解公式详细如下(2-7)所示：



(2-7)  
 在上式的矩阵分解中，  和分别叫做左奇异向量和右奇异向量,原

始矩阵中的 只和U的第i行有关，这里定义第i行为 。原始矩阵中的只和中的第j列有关，这里定义这一列为。和都不是特征值，但是由矩阵的特征值所决定。

选择m个最大的奇异值，和Ｕ与Ｖ中对应的向量相乘，就能得到一个X矩阵的m阶近似，此时该矩阵和X矩阵比较有着最小的Frobenius范数[30]。而且能够将文档向量和词向量映射到主题空间。

向量和含有k个奇异值的矩阵相乘，实现了矩阵从高维空间到低维空间转换，与此同时向量 也能够实现从高维空间到低维空间的转换，这种转换的公式如下(2-8)所示：

 (2-8)

在实现维度转换后，就可以完成以下工作：

（1）判断两篇文档在低维空间的相似度。比较向量和的余弦相似度就可以求解出。

（2）通过比较与可以比较两个词项的相似度。

（3）可以对文本进行聚类。

（4）可以计算给定字符串和文档在语义上的相似度。

如果比较给定字符串和文档的相似度，需要将字符串和文档都映射在语义空间上，文档的映射公式如下(2-9)所示：

 (2-9)

给定字符串的映射公式如下(2-10)所示：

 (2-10)

可以通过计算中非零元素的倒数得到其逆矩阵。进而计算字符串和文档的相似度。

3）应用

低维语义空间的应用有以下几点:

（1）在低维语义空间上，文档相似度计算将更加便捷，并且还可以对文档进行聚类。

（2）在翻译好的文档上进行模型训练，能够发现跨语言的相似文本文档，实现跨语言检索。

（3）发现词项与词项之间的关联，可用于同义词检测、歧义词检测等。.

（4）利用在语义空间上的相似性进行信息检索。

（5）从语义主题的角度发现词项和词项，文档和词项之间的关联，可用于选择题回答模型。

## 概率潜在语义分析

### 2.2.1 模型描述和参数估计

概率潜在语义分析(probabilistic Latent Semantic Analysis，pLSA)，又被叫做概率潜在语义索引，是一种基于概率统计的主题模型。概率潜在语义分析并不来源于线性代数，而来源于基于潜在语义模型的混合分解。

pLSA作为一种主题模型，认为一篇文档是在多个主题上的概率分布，而每个主题都是在词项上的概率分布。确定一篇文档d，以一定的概率选择文档d对应的主题z，再以一定概率选择主题z中的词项w，可以得到文档在主题上的分布和主题在词项上的分布。概率模型图如图2-1所示：

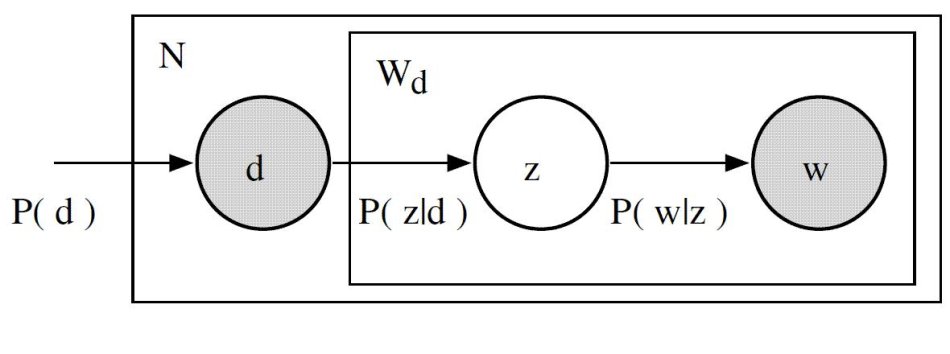


图2-1 PLSA概率模型图

上图中d表示文档，z表示主题，w表示文本中的词项，表示文档di中出现词项的概率，表示主题zk中的词项在文档di中出现的概率，表示主题zk中出现单词wj的概率。每个主题在所有词项上服从多项分布，每篇文档在所有主题上服从多项分布。整个文档的生成过程如下所示：

1）以的概率选中文档di；

2）以的概率选中主题zk；

3）以的概率产生一个词项。

其中可看到的数据就是和，而zk是隐含变量。所以的联合分布如下式(2-11)和(2-12)所示：

 (2-11)

 (2-12)

第m篇文档中词项的生成概率如下式(2-13)所示：

 (2-13)

整篇文档的生成概率如下式(2-14)所示：

 (2-14)

根据上述公式得到的生成概率，使用极大似然估计法求解参数，得到似然函数如下式(2-15)所示：

 (2-15)

对于这种含有缺失概率和隐含变量的概率模型参数估计问题，采用EM算法进行解决。

EM算法的步骤如下：

1）E步骤：求隐含变量当前估计的参数条件下的后验概率。

2）M步骤：将数似然函数的期望最大化，再使用E步骤中隐含变量的后验概率，计算出新的参数。

一直迭代直到函数收敛。

为了解决概率潜在语义分词中参数估计问题，在E步骤中，需要使用贝叶斯[31]公式计算隐含变量在当前参数值下的后验概率，公式如下(2-16)所示：

 (2-16)

在上述公式中，假设和两个概率值是已知的，由于开始时赋值是随机的，所以在后面迭代的过程中取前一次迭代M步骤中得到的参数值。

之后，在M步骤中，根据后验概率重新计算参数，公式如下(2-17)，(2-18)所示：

 (2-17)

 (2-18)

最后使用新的参数值，再次进入E步骤，计算隐含变量Zk 在当前估计的参数条件下的后验概率。一直进行迭代直到达到终止条件。

### 2.2.2 模型特性

基于统计的概率潜在语义分析相比于潜在语义分析具有严谨的数理统计基础，而且模型训练花费时间更少。其主要用途如下：

1）根据文档在主题上的概率分布进行文本相似度计算，而且规避了同义词，多义词带来的干扰。

2）文本聚类。

3）去噪音，清除干扰，保留重要的主题，更好的进行文本应用。

概率潜在语义分析的优势还是比较明显的，其可以解决了一词多义的问题，利用了强化的期望最大化算法对潜在类进行训练。而且相比于潜在语义分析拥有更加坚实的统计学基础。

当然pLSA也有缺点，首先其不是完整的生成式模型，在文档层面没有提供合适的概率模型；随着文档数目和词项数目的增加，概率潜在语义分析模型也会线性增加，变的越来越大，训练参数值会随着文档数和词项数线性增加；EM算法的计算量很大，需要反复迭代。

## 潜在狄利克雷分配

### 2.3.1 共轭先验和狄利克雷分布

潜在狄利克雷分配(Latent Dirichlet Allocation，LDA)是一种完备的生成模型，是在pLSA的基础上添加了一层贝叶斯框架。也就是说，LDA是pLSA的贝叶斯版本。LDA中的文档和词项生成模型中的多项分布的参数是变化的，并且拥有先验参数，也就是超参数(hyperparameter)[32]，这些参数的分布服从狄利克雷分布(Dirichlet distribution)。

利用共轭先验[33]的方式可以方便贝叶斯模型的计算，贝叶斯公式的计算公式如下式(2-19)所示：

 (2-19)

其中表示条件概率，也就是似然概率，是先验概率，是后验概率，如果后验概率和后验概率有相同的函数形式，则这两个概率都称之为共轭分布，其中先验概率是条件概率的共轭分布，在潜在狄利克雷分配中狄利克雷分布就是多项分布的共轭先验。

多项分布的概率密度函数如下式(2-20)所示：

 (2-20)

其中，p是多项分布的参数，表示时间发生的概率，n表示进行了n次重复多项实验。

维度为k的狄利克雷分布在参数α上有一个概率密度函数如下式(2-21)所示：

 (2-21)

其中可以理解为多项Beta函数，如下式(2-22)所示：

 (2-22)

### 2.3.2 文本建模

在LDA中，“文档-词项”生成模型的生成方式如下所示：

1）按照先验概率选择一篇文档di;

2）根据狄利克雷分布超参数α产生文档di的主题分布;

3）从主题分布中取样生成文档di的第j个词项的主题zi,j;

4）根据狄利克雷分布超参数β产生主题zi,j的词项分布;

5）从词项的多项式分布中取样生成词项wi,j;

以上可以看出，LDA相比于pLSA模型，为主题分布和词项分布添加了两个狄利克雷先验。

在LDA中，主题分布和词项分布不是唯一确定的，但也服从一定的分布，由狄利克雷先验随机决定，属于贝叶斯派思想。

潜在狄利克雷分布的概率模型图如下所示：

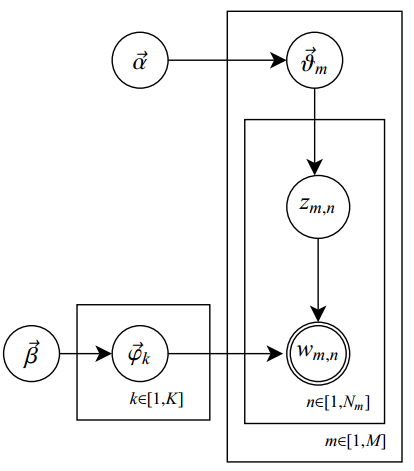


图2-2 潜在狄利克雷分配概率模型图

上图中的M是文档个数，K是文档生成的主题个数，Nm是第m篇文档的词个数，是每篇文档下主题的多项分布的狄利克雷先验参数，是每个主题下词项的多项分布的狄利克雷先验参数，wm,n是第m篇文档中第n个词，zm,n是第m篇文档中第n个词项所属的主题，隐含变量表示第m篇文档下的主题分布，表示第k个主题下词项的分布。

### 2.3.3 模型特性

LDA模型可以用来分类和降维，在文本主题挖掘领域是得力的工具，其克服了pLSA的缺点，是完整的生成模型，还可以对不在训练文本中的文本进行语义识别，当然也有缺点，作为参数贝叶斯模型，主题个数需要提前给出，而k值的不同对模型的性能有很大的影响。

## 层次狄利克雷过程

### 2.4.1 模型概述

层次狄利克雷过程(Hierarchical Dirichlet Process，HDP)是一种非参数贝叶斯模型[34]，其中狄利克雷过程(Dirichlet Process)是一个随机过程，在非参数贝叶斯模型中有广泛运用。最常见的是作为混合模型[35]的先验。它是一个分布的分布，狄利克雷过程中每次样本的抽取都是一个分布。它被称为狄利克雷过程过程是因为它在有限维上的边缘分布是狄利克雷分布。像高斯分布[36]一样，它的有限维边缘分布是高斯分布。从狄利克雷过程中抽取的分布是离散的，但无法使用有限个参数描述，因此它被归为非参数模型。

概率模型[37]被用在机器学习中为数据建立分布模型。当模型复杂度和有效数据量之间不匹配的时候，传统的参数模型使用固定的和有限维度的参数容易产生过拟合或者不拟合的情况。因此，在参数化模型建模时，在适当的复杂度下进行模型选择是非常重要的问题。然而，不管我们是使用交叉验证，还是使用边缘概率，模型选择都面临很多问题。而非参数贝叶斯模型不需要限制模型的复杂性，因此会缓解不拟合情况，同时通过计算参数的全部的后验概率可以减缓过拟合的情况，其中过拟合是因为层次模型的结构。

通常，我们都假设我们希望从数据中推断一个潜在的未知的分布。假设xi~F是来自同一个分布且相互独立的样本，其中F是未知的分布。贝叶斯方法解决这个问题时，会在在计算数据条件下F后验条件概率给F一个先验。传统环境下，这个先验会限制在一个范围内，指定参数的范围。而非参数贝叶斯会将这个先验放在一个分布空间中，以获取更好的结果。通常，这个分布空间是所有分布的集合。因此，如此大的选择范围会使后验概率的计算比较麻烦。目前，狄利克雷过程是最流行的非参数贝叶斯模型。

### 2.4.2 狄利克雷过程

要介绍狄利克雷过程，首先要从狄利克雷分布说起，设置Θ={θ​1,...,θ​m​​}，狄利克雷分布的定义如下式(2-23)和(2-24)所示：

Θ~Dir(α​1,...,α​m​​) (2-23)

 (2-24)

狄利克雷分布是多项式分布的共轭分布，贝塔分布是狄利克雷分布的特殊情况，是二元形式。狄利克雷分布是一组连续的多元概率分布，其参数是一个正直向量α。狄利克雷分布通常在贝叶斯中用作先验分布。而狄利克雷分布的无限维扩展就是狄利克雷过程。

将狄利克雷分布形式推广到分布上，也就是将狄利克雷分布中的参数换成某个分布，就可以得到狄利克雷过程如下式(2-25)所示：

G~DP(α,H) (2-25)

其中α是正的尺度参数，H是基分布（base distribution）。对比狄利克雷分布和狄利克雷过程我们可以发现，前者是对参数空间的分布，后者是对基分布的分布。前者扩展到无限维度即可得到后者。α是一个大于0的值，它决定了G有多离散。当α=0的时候，G就是一个离散的值，当α→∞时，G和H就是一样的了。从狄利克雷过程中抽取的样本是一个完整的分布，它可以理解成是一个测度空间的某种划分结果，也就是说混合模型中划分成几个部分，每个部分是什么样的一个结果。

狄利克雷过程是一个随机过程，其样本轨道是概率测度。随机过程是函数空间的分布，其样本轨道是从分布中抽取的随机函数。在狄利克雷过程中，就是概率测度的分布，概率测度有一些特殊的属性可以用来当作分布进行推断。因此，从狄利克雷过程中抽取样本可以理解成抽取随机分布。

### 2.4.3 层次狄利克雷过程的概率模型

层次狄利克雷过程[39]是一组随机测度的分布，也就是宽分布，狄利克雷过程的抽样**中，产生一个离散分布，就是G0，离散样本中的原子会按照基础分部H抽样，公式为，如果用离散分布G0代替公式中的基础分布H，狄利克雷过程中的随机测度Gj之间会共享样本中的原子，也就是将共享G0的原子，其中基础分布的离散特性使得多个随机测度共享离散原子，这样的结构就是层次狄利克雷过程，其中第一层狄利克雷过程建立了基础分布G0，第二层狄利克雷过程建立了随机测度Gj。

设置一组数据源xi，则数据源的观测数据为xi,j，如果数据源xi和xj的分布是由和两个参数决定的，参数是由随机测度产生，而随机测度是由构造，如果两个数据源相关，那对应的两个参数也会共享原子，记为，而观测数据会增强这个值的统计力度，这种模型就是层次狄利克雷混合模型。公式表示如下式(2-26)所示：

 (2-26)

上式表示第j组中第i个观测数据，表示观测数据的分布参数，表示产生的分布。下图2-3所示是层次狄利克雷过程的概率模型图：

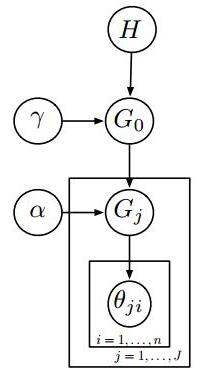


图2-3 层次狄利克雷过程的概率模型图

## 本章小结

本章介绍了主题模型的四种技术：LSA、pLSA、LDA和HDP，其中后三种都是基于概率模型的方式进行主题挖掘的，所以模型解释性更强，但是pLSA并不是完备概率生成模型，文档的增加会使得模型复杂度越来越高，而LDA弥补了pLSA的不足，是完备的概率生成模型。LDA也属于参数贝叶斯模型，主题个数的设置对模型性能影响很大，而非参数贝叶斯模型解决了这个问题，会自动生成主题个数的参数。

# 层次狄利克雷过程主题模型改进

本章分析了层次狄利克雷过程主题模型存在的问题，并且提出了改进方法，最终建立了一种基于特征加权机制的BM25F-HDP模型，为之后的病历文本特征提取提供算法支持。最后对改进后的HDP模型和改进之前的HDP模型，LDA模型进行了性能实验对比。

## 问题提出

层次狄利克雷过程(Hierarchical Dirichlet Process，HDP)是一种非参数贝叶斯模型，在主题建模的过程中，每篇文本都可以视为词项的集合，词项按照主题聚类，最终确定文档在主题上的分布，还有主题在词项上的分布，各个文档的主题服从相同的狄利克雷过程，这样的就使得不同的文档之间共享相同的主题。相比于参数贝叶斯模型潜在狄利克雷过程分配(Latent Dirichlet Allocation，LDA)，在HDP模型上进行文本训练之后的主题个数并不需要提前确定，而是在模型分布参数推导过程中自行确定。HDP模型的求解步骤如下所示：

1. 采用CRF方法来对狄利克雷过程进行构造，对文本中词项的主题进行指派。
2. 利用模型参数后验分布方法Gibbs Sampling对HDP模型进行采样
3. 马尔科夫链收敛到平稳分布，获取文本集的主题分布和主题的此项分布

Gibbs Sampling对HDP模型进行采样的算法如下：

**算法1 HDP模型的Gibbs采样算法**

输入：病历文本向量

输出：主题分配向量，主题个数K，多项分布参数，

1初始化：将全局统计量njk,mjk为0，隐含变量zji设为-1

2最大迭代次数t为100；

3 For t=1,…,T do

4 For j=1,…,J do

5 For i=1,…,I do

6 移除xji对应的隐含变量主题zji，将各个对应的统计量减1，并且给zji重新采样kji。

7 If zji==K+1then

8 得到新的主题K, 将对应的统计量加1，并且给xji重新采样隐含变量主题。

9 Else

10 采样得到的是已有的主题，更新对应的统计量，并且给xji重新采样隐含变量主题。

11 If njk==0 then

12 xji对应的主题下没有词项，则将此主题和相关参数一并删除，并且将相对应的主题个数减1；重新调整xji的主题索引，确保每个索引下的词项数目不为空；

13 采样mjk和β；

14 计算似然概率；

15 If 收敛||t=T then

16 Return和

可以看出，在HDP模型的Gibbs采样中，就是根据已经采样到的文本在主题上的分布和主题在词项上的分布，估计当前采样的词项对应的主题，并且反复迭代得到所有隐变量的过程。采样算法如上算法1所示。若要得到对应主题，会受到词项在主题中的概率和词项对应的文本所拥有的主题概率两个因素影响，又由于高频词在文本和主题中所占的比例较大，而且每个词项都是按照相同的权重计算条件概率的，导致给词项分配新主题时会向高频词对应的主题倾斜。也就是说，词项的频度高低会直接影响其在主题中的权重高低。而且很多词项本身并不能代表某种语义，这些词项的存在会增加模型混乱度。

在分析了HDP模型中存在的一些问题后，再结合具体的训练文本，也就是病历文本集进行分析。HDP模型中词频的分布会导致很多中频词项和低频词项的主题分布受到高频词项的主题分布的影响。那是否可以通过建立停用词库来降低高频词对主题模型的影响呢，在病历文本中，涉及很多专业术语，像“谵语”，“紫绀”这些具有很高的专业性，这些词可能词频不高，但是对于语义却有着重要的意义。还有很多词项词频很高，对于文档语义有着重要的影响，而且包含着重要的病症信息，不能去除。可是如果保留这些停用词，还是会导致这些词项在很多主题中的条件概率很高，但这样的词项并不能很好的说明这个主题的特征，在主题模型的不断迭代过程中，这些词项会频繁的出现在各个主题中，使得主题中的词项分布受高频词项的影响很大。所以只是单纯使用停用词库的方法并不能解决上述的问题。

所以，本节中就将针对HDP模型中，当利用Gibbs采样方法来估计模型参数，根据所有词的主题分配估计当前词分配各个主题的概率时，主题分配向高频词主题倾斜的问题提出解决方案，并且验证解决方案的的有效性。

## 主题倾向性优化的HDP模型

针对上述问题，本文提出了改进的HDP模型BM25F-HDP主题模型，该模型在采用中国餐馆连锁CRF(Chinese Restaurant Franchise)方法进行构造的过程中，将当前词项在主题K中出现的次数减少或者增加词项的特征权重，而不是仅仅的进行统计量减1和加1的操作。这样就起到了降低病历文本中高频词的作用。至于特征权重的选择对于主题模型的优化有着至关重要的作用，在文本特征加权领域中常用的方法有很多，比如特征的互信息值、几率比、期望交叉熵、信息增益以及信息检索领域的加权方法等。前面的四个方法多被应用在有监督学习的相关特征提取过程中，而信息检索领域的加权方法多用在无监督的学习过程中。在这里选择BM25F这个加权方法，在之后会给出详细的分析。同时，也引入了高斯函数对词项的主题倾向性进行优化，在综合考虑这两种因素的情况下对模型进行优化。

### 3.2.1 HDP模型引入BM25F和高斯函数

HDP模型将文本表示成词袋形式，也就是表示成词项的集合。并且将所有的词项一视同仁，也就是说所有的词项并不考虑其权重的大小，也就是说在重新选择词项的主题时，会更倾向于高频词的主题。针对文档中的高频词汇，可以运用高斯函数对词项定义权重， 也可以运用信息检索领域的加权方法降低其权重[41]，使得模型主题的倾向性不再偏向高频词的主题。本文中定义在同一篇文本中的词项如果一样，那它的权重值也是一样的。

1. 高斯函数对主题倾向性的影响

高斯函数在统计领域应用很多，在这里利用高斯函数进行词项频率的重新定义，能够降低高频词项的权重，可以通过调整fmid的定义来提高部分词项的权重，在这里定力fmid为中频词项权重，则增加了中频词项的权重，使得主题的倾向性不再偏向于高频词的主题，高斯函数中对文本中词项加权公式如下所示：

 (3-1)

式中:

——词项i在文档d中的高斯函数值；

fi——词项i的词频值；

V——文本中的词项集合；

——词项中词频数居中的词频值；

2）BM25F对主题倾向性的影响

信息检索领域中的加权方法很多，其中最基本加权方法的就是将逆向文件频率作为词项的权重，其意义为总的文档数目和包含此词语的文档数目的比值再取对数，不过使用较少，这是因为其有特征信息太少的缺陷，所以在此基础上又有了BM25模型，这个模型有了词频和文档长度这两个特性。可是BM25模型还是有其缺陷，就是一篇文档作为一个整体统计词频，但是忽略了文档中不同部分的重要性，所以BM25F就是为了解决这个问题在BM25上进行改进。

BM25F模型的计算公式如下所示：

 (3-2)

其中第二部分中的变量公式如下：

 (3-3)

式中：

Bv——不同文本段的长度情况；

v——文档有v个文本段；

——第i个单词在不同文本段中的加权和；

wk——不同的文本段对应的权重；

fvi­——在此文本段中词项i的词频；

其中Bv作为不同文本段的长度，具有其相关计算公式，公式的计算方法如下式所示：

 (3-4)

式中：

bv——各个文本段的调节因子；

——某片文本段的长度；

avvlv——某片文本段的平均长度；

病历的不同用途决定了不同文本段的权重值的大小，在本节中，主要讨论Gibbs采样中词项权重的改进对于模型的优化，所以在此将问题进行简化，将病历文本作为一个文本段进行处理，验证权重改进对于模型的优化。

所以在上式中，令v = 1，令权重值w=1，此时权重公式如下：

 (3-5)

式中：

——第d篇病历文本中第i个词项的特征权重；

N——全部文档数；

n(q)——包含词项q的文档数；

fi——在病历文本中词项i的词频；

k1——实验参数；

b——调节文档长度的因子；

dl——当前病历文档的长度；

avdl——全部病历文档的平均长度；

3）综合考虑BM25F和高斯函数的HDP主题模型

在模型的吉布斯采样过程中，会迭代到每篇病历文本的每个词项，当迭代词项t的时候，传统的非参数贝叶斯模型HDP会将词项t在主题k中出现的次数njk进行采样前的减1操作和采样后的加1操作，将属于主题k的餐桌数mjk进行迭代前的减1操作和迭代后的加1操作.而改进的非参数贝叶斯模型BM25F-HDP不会对njk和mjk进行简单的减1和加1操作，而是通过累加权重值的方式，权重值的计算方法已经在上一节中提到，而如果只是单纯的引入权重带入算法计算，会导致加权后的病历文本集的总词数和加权之前的病历文本集的总词数发生变化。会导致非参数贝叶斯模型选择的主题数k值越来越大而且很难进入下一次迭代。所以引入如下公式，保证在吉布斯采样每一次迭代完成之后保持词数不变。

 (3-6)

式中：

——词项t在文档m中的综合权重值

words——数据集中的总词数；

,——调节因子；

在上式中，综合考虑了高斯函数和BM25F两种方式对词项的主题倾向性进行优化，进而提升模型质量。其中有两个调节因子,，在这里设置两个调节因子为1，为之后的实验做准备。

在对词项的主题进行重新采样的前后进行此权重的加减操作，就不会出现主题数k越来越大而不会进入下一次迭代的问题了。而且使得所有的词项统计量不同，加权后的Gibbs采样算法描述如下：

for mM{

for tm{

njk  njk, mjk  1 mjk

重新采样主题

njk +  njk, mjk + 1 mjk

}

}

上式中mM表示一次迭代中，迭代到某一篇病历文本m，tm表示在此文本下迭代到某个词项。这只是一次迭代过程中对主题的重新采样，如果需要多次迭代直到马尔科夫链收敛，需要进行多次迭代。此算法描述只是为了突出模型改进的重点，在之后会给出改进HDP模型的详细采样过程。

### 3.2.2 改进模型的Gibbs采样

在模型中，使用吉布斯采样方法对其进行采样，此方法基于中国餐馆连锁。而中国餐馆连锁根据可观测变量，对z，β，进行直接采样，而变量π经过积分可以消除。改进模型的吉布斯采样过程如下所示，对中国餐馆连锁中的统计量餐馆中属于主题k的词项数njk进行主题倾向性优化，不仅仅是简简单单的加1和减1操作，而是综合考虑高斯函数和BM25F的模型改进。具体的改进点在第6、7和10步骤中进行了体现，此算法和本节开头提到的算法1形成了对比。

**算法2 改进模型BM25F-HDP的Gibbs采样算法**

输入：病历文本向量

输出：主题分配向量，主题个数K，多项分布参数，

1初始化：将全局统计量njk,mjk为0，隐含变量zji设为-1

2最大迭代次数T为100；

3 For t=1,…,T do

4 For j=1,…,J do

5 For i=1,…,I do

6 移除xji对应的隐含变量主题zji，将餐馆中当前词项在主题K中出现的次数减去词项的权重，将属于主题k的桌子数减1，并且给zji重新采样kji。

7 If zji==K+1then

8 得到新的主题K, 将餐馆中当前词项在主题K中出现的次数增加词项的权重，将属于主题k的桌子数加1，并且给xji重新采样隐含变量主题。

9 Else

10 采样得到的是已有的主题，将餐馆中当前词项在已有主题中出现的次数增加词项的权重，将属于已有主题的桌子数加1，并且给xji重新采样隐含变量主题。

11 If njk==0 then

12 xji对应的主题下没有词项，则将此主题和相关参数一并删除，并且将相对应的主题个数减1；重新调整xji的主题索引，确保每个索引下的词项数目不为空；

13 采样mjk和β；

14 计算似然概率；

15 If 收敛||t=T then

16 Return和

## 实验内容与结果分析

本节针对LDA潜在语义模型，HDP非参数贝叶斯模型，BM25F-HDP非参数贝叶

斯模型进行实验对比，并且进行分析。首先会介绍混乱度的概念，混乱度作为衡量模型性能的指标被广泛应用。然后会对实验中涉及到的超参数进行介绍和设置。之后会对三种模型进行实验，在主题独特性，模型混乱度，以及迭代运行时间三个方面进行性能对比。

### 3.3.1 实验环境与评价指标

模型运算的实验软硬件环境如下表所示：

表3-1 实验软硬件环境

|  |  |
| --- | --- |
| 计算机 | 联想G470(lenovo) |
| 操作系统 | Windows 7 Ultimate |
| 处理器 | Intel(R) Core(TM) i5-2430M |
| 数据库软件 | MySQL |
| Java | jdk1.7.0\_25 |

模型测试数据集为1185条病历文本，本次试验选取950条病历文本作为训练集，235条病历文本作为测试集，然后在训练数据及上分别运行潜在语义模型LDA，非参数贝叶斯模型HDP，改进非参数贝叶斯模型BM25F-HDP。

在主题模型的评价指标中，选择主题独特性，运行时间，混乱度(perplexity)三个特性作为评价标准。所谓混乱度，就是在一片病历文本d中，模型对于文本d属于哪个主题的不确定性，这个不确定性的程度大小就是混乱度。混乱度越小，说明模型性能越好。在实验中，首先使用训练集进行模型训练，再在训练出来的模型基础上运行测试集，然后计算测试集文本数据上所有token的似然值几何平均数的倒数。其计算公式如下：

 (3-7)

式中：

Wtest——测试病历文本集；

T——训练好的模型参数；

M——病历文本测试集个数

Nm——病历文本m的长度；

——病历文本m的第i个词项；

——词项w在主题k下的概率；

——主题k在病历文本m下的概率；

在LDA模型中，涉及到的超参数主要包括主题个数K,狄利克雷分布超参数α、β，经过实验的参数调优，设置K为20，α为0.6，β为0.01，在HDP模型及其改进模型BM25F-HDP中，设置基分布H的超参数η为0.5，concentration参数γ，α 0由gamma先验分布决定。总体来说，超参数起到平滑数据的作用。当α变小的时候，会尽可能的让一篇文档只有一个主题，当β变小的时候，会让一个词项尽可能的属于同一个主题。也就是α和β过小的话，就会让主题模型的多项式分布参数θ和φ变的稀疏。γ表示的是在中国餐馆连锁中，顾客想要开一张新桌的意愿，γ越大，表示开新桌子的意愿越强。

### 3.3.2 文本数据集描述

病历文本就是医疗人员对于患者的身体症状进行检查，诊断，治疗等一系列医疗活动的记录，对采集到的病历信息进行归纳总结，综合分析，按照一定的规范和要求编写的病患医疗档案。病历是对医疗工作实践的总结，也是寻找疾病医学规律的重要依据，其对科研，医疗，疾病预防等方面都有重要的作用。本文采用的病历文本集的结构包括四大部分：入院情况、查体、诊断和治疗原则。如图3-1、图3-2、图3-3、图3-4所示为病历文本样例图：

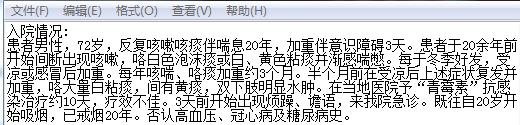


图3-1 入院情况文本段示意图

图3-1所示的是入院情况文本段，此文本段是指病患入院后，由医师通过问诊，辅助检查等方法获得相关资料，并对这些资料进行归纳总结并且整理成记录。此记录中，病患的情况一般含有姓名，性别，年龄等基本信息，包含病患就诊时的体征症状和持续时间，病患诊疗疾病的发展、演化、诊疗等方面的基本情况，主要症状的特点、发展变化情况、伴随症状、发病后诊疗过程和结果、饮食和休息睡眠等情况的变化。包含病患以前的健康和疾病情况，具体就是既往的健康状况、疾病史、传染病史、预防接种史、手术外伤史、输血史和食物药物过敏史等，有无烟酒或者药物等喜好，有无家族疾病遗传倾向的疾病。

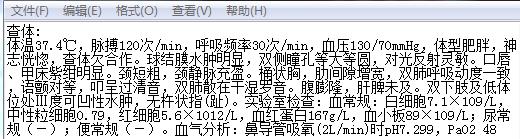


图3-2 查体文本段示意图

图3-2所示的是查体文本段，包含生命特征的检查，内容具体包括体温、脉搏、呼吸频率、血压、心跳等，器官的功能性检查，包括血液检查、尿液检查、心电图、听力、视力、肺功能等各项在器官功能检验方面的检查，器官的结构性检查，包括胸腹X光，上肠胃道摄影，超音波等，还有其他的一些体格检查，包括皮肤、粘膜、淋巴结、头部、颈部、胸部（心脏、血管、肺部、胸廓）、腹部（肝、脾）、脊柱、四肢、外生殖器、神经系统等。

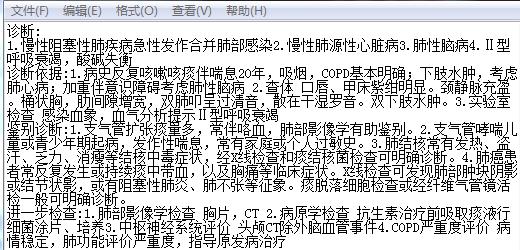


图3-3 诊断文本段示意图

图3-3所示的是诊断文本段，描述的是医师对病患病史，查体情况进行全面分析，根据病患特点，提出的初步诊断和诊断依据，对诊断不明的需要另写出鉴别诊断并且进行分析，对接下来的诊疗、治疗措施进行分析。

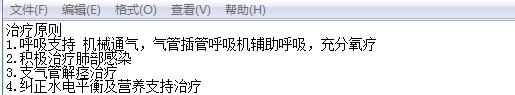


图3-4 治疗原则文本段示意图

图3-4所示的是治疗原则文本段，描述的是医师根据诊断依据，提出的具体的检查治疗措施安排。

在病历文本中，是由不同的描述模块组成，入院情况模块主要是对于病患的基本情况和症状进行描述，能够使得医疗人员有个大致的判断。而查体主要是针对身体检查项，来判断哪些器官有病症。诊断部分针对病患特征和查体状况判断疾病，并且包含鉴别诊断内容。如果需要的话还包括进一步检查、鉴别诊断。治疗原则主要是针对疾病进行相关的治疗，包含要药物治疗、仪器治疗、注意事项等。

病历文本描述具有其专业性，包含对于各种疾病现象的描述，而不同的医生可能对于一些医学主题词有不同的描述方式，但其在语义方面是一样的。

### 3.3.3 文本数据集预处理

在为病历文本建立主题模型之前，需要针对文本进行预处理，预处理的主要工作包含两部分：

1）在分词系统中添加专业医学名词词库

2）使用停用词库去停用词

本文中在处理病历文本时建立了自己的专业医学名词词库，词项的来源如下：

1）搜狗细胞词库医疗部分

2）病历文本中提取的部分医疗词汇

本文采用的分词工具是中科院汉语分词系统。此分词器支持拓展词库，停用词库，并且采用了层叠马尔科夫模型，将词法分析放在了一个完整的理论框架中，最终获得了最好的总体效果，拓展词库添加的就是专业医学名词词库，添加的目的就是为了使得分析器能够正确的对医学用词进行分词。

### 3.3.4 LDA、HDP和BM25F-HDP实验对比

本节从三个方面来对比潜在语义模型LDA、非参数贝叶斯模型HDP、改进非参数贝叶斯模型BM25F-HDP的性能，分别是病历文本训练集训练完成后的主题独特性[42]对比，混乱度对比和运行时间对比。

1）主题独特性

主题质量的指标包括多样性，有效性，独特性等，在这里选择主题独特性这个指标对主题质量进行评价。通过实验得出三个模型中吉布斯采样迭代次数和主题数目变化的关系，如图3-5所示：

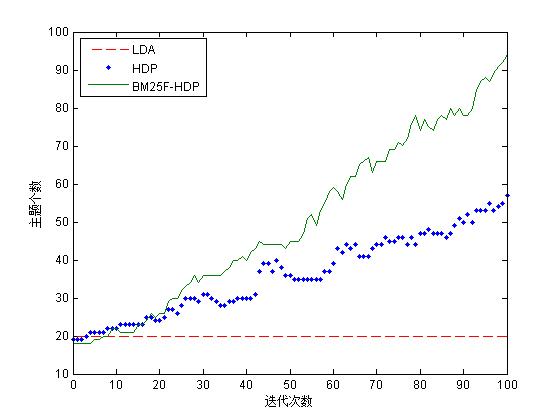


图3-5 模型主题变化对比图

图3-5中显示的是LDA模型，HDP模型，BM25F-HDP模型的主题数对比图.由于LDA是参数贝叶斯模型，需要先设置主题个数，在这里设置为20。在非参数贝叶斯模型中，主题的个数不需要具体给出，只是在模型迭代运行的过程中自动选择主题的个数，这就会导致每次迭代时的主题数目很有可能不一样，就如上图所示。上图中HDP模型中在迭代前的主题数量确定为19，在迭代过程中主题个数大体上呈现增长趋势。但也偶尔会有下降。BM25F-HDP模型中初始主题数量确定为18，在迭代过程中主题个数大体上也呈现增长趋势。部分吉布斯采样迭代次数下，三种模型的主题个数如下表3-2所示：

表3-2 模型主题个数对比

|  |  |  |  |
| --- | --- | --- | --- |
| 迭代次数 | LDA模型主题个数 | HDP模型主题个数 | BM25F-HDP主题个数 |
| 1 | 20 | 19 | 18 |
| 10 | 20 | 22 | 22 |
| 20 | 20 | 24 | 26 |
| 30 | 20 | 31 | 36 |
| 40 | 20 | 30 | 40 |
| 50 | 20 | 36 | 45 |
| 60 | 20 | 39 | 59 |
| 70 | 20 | 44 | 66 |
| 80 | 20 | 47 | 74 |
| 90 | 20 | 50 | 78 |
| 100 | 20 | 57 | 94 |

可以得出结论，以吉布斯采样迭代次数为自变量时，改进后的模型BM25F-HDP的主题增加趋势比HDP模型明显，相同的迭代次数下，BM25F-HDP模型的主题个数要比HDP模型多，而且迭代次数越多，主题数量相差越大。这可以理解为，BM25F-HDP模型挖掘出了比HDP模型更多的主题，也就是说BM25F-HDP模型能够区分出相似度很高的主题，挖掘出了更多的潜在主题，主题的独特性有所增强，对于主题的质量也有所提升。

2）模型混乱度

通过实验得出三个主题模型中吉布斯采样迭代次数和混乱度之间的关系，下面展示的就是参数贝叶斯模型LDA、非参数贝叶斯模型HDP和BM25F-HDP的混乱度对比，如图3-6所示。

图3-6中显示的是主题个数k为20的LDA模型、HDP模型和BM25F-HDP模型的混乱度对比图。LDA模型混乱度值在前20次迭代中下降速度较快，混乱度降低了230左右，而后下降速度逐渐放缓，在迭代次数达到90和100时，混乱度值基本稳定在了706，说明马尔可夫过程基本收敛到了均衡。HDP模型混乱度值在前40次迭代中下降速度较快，混乱度降低了160左右，后来下降速度逐渐放缓，之后的混乱度基本稳定在了670左右。BM25F-HDP模型混乱度值在前50次迭代中下降速度较快，混乱度降低了180左右，之后还有缓慢的上升趋势，但是上升幅度很微小。而在迭代次数进一步的增加之后又降低了，随后混乱度基本稳定在630。可以看出不论迭代次数的多少，BM25F-HDP模型的混乱度都是最低的。预计在之后随着迭代次数增加，马尔科夫链收敛之后BM25F-HDP模型的混乱度仍然是最低的，HDP模型次之，LDA模型混乱度最高。

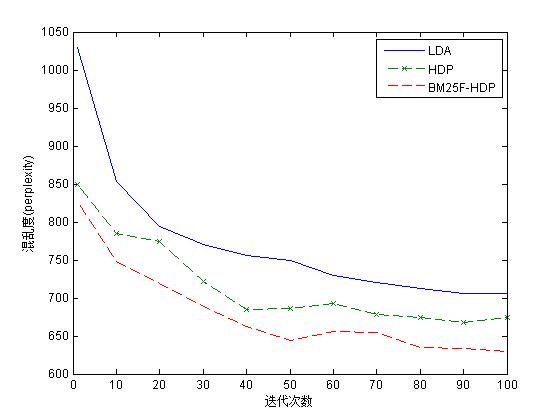


图3-6 LDA、HDP和BM25F-HDP混乱度对比图

在不同的吉布斯采样迭代次数下，三种模型的混乱度数值如下表3-3所示，其中混乱度舍去小数点后数字。由于不可能将所有迭代次数下的混乱度对比值列出来，所以迭代次数只选取了从1到100的整数。

表3-3 模型混乱度对比

|  |  |  |  |
| --- | --- | --- | --- |
| 迭代次数 | LDA模型混乱度 | HDP模型混乱度 | BM25F-HDP模型混乱度 |
| 1 | 1029 | 849 | 827 |
| 10 | 853 | 785 | 748 |
| 20 | 794 | 775 | 719 |
| 30 | 770 | 722 | 689 |
| 40 | 756 | 685 | 662 |
| 50 | 749 | 686 | 644 |
| 60 | 729 | 693 | 656 |
| 70 | 720 | 678 | 654 |
| 80 | 713 | 674 | 635 |
| 90 | 706 | 668 | 634 |
| 100 | 706 | 675 | 630 |

可以得出结论，以吉布斯采样迭代次数为自变量时，改进后的非参数贝叶斯模型BM25F-HDP的混乱度比非参数贝叶斯模型HDP和参数贝叶斯模型LDA都要低，而HDP模型由于主题个数在迭代训练的过程中自动确定，这说明了自动确定主题个数是提升主题模型性能的关键要素。在混乱度这个评价指标上，BM25F-HDP模型表现最优，HDP模型表现次之，LDA模型表现最差。所以，对于非参数贝叶斯模型来说，性能相比于参数贝叶斯模型要好。

3）迭代时间

实验得出三个主题模型中吉布斯采样迭代次数和运行时间的关系，如图3-7所示，图3-7中显示的是参数贝叶斯模型LDA、非参数贝叶斯模型HDP、非参数贝叶斯模型BM25F-HDP的运行对比图。可以看出主题模型迭代次数的增加对于LDA运行时间的影响很小，增加很缓慢，几乎没有变化。而HDP模型和BM25F-HDP模型随着迭代次数的增加，运行时间增加明显，并且改进的非参数贝叶斯模型要比改进前的运行时间长。说明BM25F-HDP模型由于迭代过程中还需要对主题个数进行不断的调整，增加了运行时间。

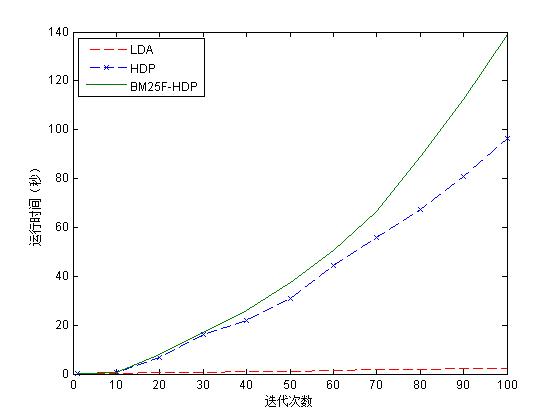


图3-7 LDA、HDP和BM25F-HDP迭代运行时间对比图

下表中显示的是在不同吉布斯采样迭代次数下，三种模型的运行时间对比，取值单位是秒，精确到小数点后第3位，迭代次数还是选择1到100中的整数，将模型迭代时间进行展示。由于不可能将全部迭代次数下的运行时间进行展示，在这里只展示了迭代次数为10的倍数的情况下三种模型的运行时间。可以看到LDA的运行时间保持在个位数秒之内，而另外两种模型运行时间多很多，不过这两种模型相差并不是很多，只是改进后的模型和改进前模型时间差越来越大。

表3-4 模型迭代时间对比

|  |  |  |  |
| --- | --- | --- | --- |
| 迭代次数 | LDA模型迭代时间（秒） | HDP模型迭代时间（秒） | BM25F-HDP模型迭代时间（秒） |
| 1 | 0.074 | 0.079 | 0.073 |
| 10 | 0.253 | 0.739 | 0.709 |
| 20 | 0.493 | 6.751 | 7.780 |
| 30 | 0.738 | 16.007 | 17.159 |
| 40 | 0.989 | 21.789 | 25.846 |
| 50 | 1.225 | 30.876 | 37.319 |
| 60 | 1.480 | 44.425 | 50.419 |
| 70 | 1.716 | 56.055 | 66.547 |
| 80 | 1.951 | 67.313 | 88.990 |
| 90 | 2.198 | 80.879 | 112.458 |
| 100 | 2.440 | 96.502 | 138.814 |

可以得出结论，虽然LDA模型在三个模型中性能最差，但是运行迭代时间最快。BM25F-HDP模型性能最好，运行时间最慢。相比而言BM25F-HDP模型和HDP模型的迭代时间比较相近。

## 本章小结

本章在原有的非参数贝叶斯模型HDP的基础上，引入了加权机制，改进了计算条件概率时的权重值，提出了一种新的BM25F-HDP模型。并在病历文本数据集的基础上，对LDA、HDP、BM25F-HDP三种模型进行了对比实验。得出BM25F-HDP模型在主题独特性、模型混乱度两项指标上有最优表现。

# 基于层次狄利克雷过程主题模型的病历文本特征提取

本章在改进的层次狄利克雷过程主题模型的基础上，研究了在病历文本的特征提取[43]。为后一章的病历服务系统提供数据服务。首先从病历文本训练集中挖掘出主题信息，运用综合考虑了语义相似度和主题区分度的主题标签提取算法，为每一种主题生成最优的主题标签。利用相关项提取算法生成文本的检查项和治疗项，并且从信息检索的角度对比了HDP模型和改进后的BM25F-HDP模型在平衡F值这个指标上的性能。

## 病历文本特征提取问题分析

在文本信息领域，文本的检索，摘要，过滤等都是技术的重点，这些技术的主要问题就是文本特征的提取。中科院使用词性类别的互信息量进行的特征提取。甲骨文公司利用百科全书进行概念扩充，根据概念的使用频度进行特征提取。还有相关研究者在向量空间模型的基础上使用词频和词性对结果进行标注，并且进行特征提取。还有通过语言分析技术提取文本摘要句子。可以看出这些特征提取方法只是利用了词法，句法等信息，并没有考虑病历文本中大量存在的同义和多义现象，也没有考虑到词性褒贬的因素，而这些因素对于特征提取会产生非常大的影响。所以就需要从语义的方面对特征提取进行研究。

本章就是在语义的基础上设计了特征提取相关算法，这些算法使用的语义模型是一种病历文本的形式化结构，从文本主题的角度给出文本的特征。最初的主题模型LDA是一种参数贝叶斯模型，需要提前给出主题个数，而主题个数的多少对于文本语义的分布会有很大的影响，所以在这里引入了改进的非参数贝叶斯模型BM25F-HDP，它能够在采样过程中不断调整主题个数，能够保证主题个数是一个比较合理的值，在上一章中已经实验得出在模型混乱度和主题独特性两个指标上BM25F-HDP模型优于LDA模型和HDP模型，本章将在BM25F-HDP模型的基础上进行病历文本的特征提取。

本章中的特征提取包括病历文本主题标签提取和相关项的提取，如图4-1所示。分别采用综合考虑语义相似度和主题区分度的主题标签提取算法和基于主题模型的相关项提取算法来实现，本章将分别对两种特征提取算法进行问题描述、实验和结果分析。



图4-1 特征提取图

1）病历文本主题标签提取算法主要问题分析

主题标签的提取指的就是，对病历文本数据集进行了改进HDP模型的训练后，定义表示在文档di中全部主题z的一个多项分布，则可以表示成一个向量，表示在主题zk中所有词项v的一个多项分布，也可以表示成一个向量。对于改进的模型HDP可以获取的信息包括文档di的主题分布和主题zk的词项分布，在主题标签候选集中，需要选择最合适的主题标签和zk进行匹配，对主题进行标注。进而挖掘出病历文本中的医疗主题信息，这也是特征提取的一部分，利用病历的语义信息进行病历文本的特征提取。针对这个问题，展开分析。

在特征提取领域中，利用两个对象之间的相似度作为评价指标已经得到了广泛应用，也是文本挖掘中的核心问题。这是因为自然语言的词语之间存在着同义、反义等复杂关系，需要用一种简单的标准对这些关系进行度量。语义相似度在文本分类、语义消歧等领域应用广泛。语义相似度评价指标因为在统计模型的基础上考虑到文本的语义特征，所以在自然语言处理问题中也逐渐成为热点。所谓语义相似度，是指两个对象在语义分布上的相似程度。一般运用相对熵(relative entropy)的概念来计算两个对象主题的语义相似度，相对熵又名KL散度（Kullback – Leibler divergence）,简称KLD。相对熵是两个概率分布M和N差别的非对称性的度量，用来度量使用基于M的编码来编码来自N的样本平均所需的额外的比特个数。在典型情况之中，用M表示数据的理论分布，用N表示数据的真实分布。在本节中用来度量标签和主题在词项上的多项分布的相似性。衡量两个分布的距离公式如下所示，在下图中，KL-divergence始终大于等于0，当两个多项分布完全相同时，KL散度的值为0。

 (4-1)

病历文本经过模型训练之后一般会产生多个主题类别，有的标签可能和多个主题的相似度都较高，如果只是使用语义相似度作为评价指标对主题标签进行提取，这样的标签并不能很好的代表这个主题，如果标签只和当前主题具有较高的语义相似度，而和其他主题的具有较低的语义相似度，我们就会说这样的标签是一个优秀的标签，所以针对这个问题，如何能够给主题选择最合适的标签，使其具有较好的区分性，成为亟待解决的问题。

2）相关项提取算法的主要问题分析

在病历文本中相关项包含检查项和治疗项，所谓检查项指的就是病患需要检查的身体器官或者指标，而治疗项指的是针对患者的一些治疗手段。本节研究的内容就是基于语义信息提取病历信息的相关项。

在医学学术界对于疾病还有药品的关联挖掘做了许多工作,也实现了一些突破，其中很多都是从关联规则挖掘的角度实现的，可是此方法只能寻找到项目的频繁模式，无论如何进行频繁项集挖掘，都只是在浅层含义上的文本挖掘工作，并没有深入到语义层面进行挖掘。所以需要在语义层面进行相关项的挖掘，而本文中的主题模型正好能够从语义方面对病历文本进行挖掘。

所以本节提出一种基于主题模型的相关项生成方法，使用BM25F-HDP模型生成文本矩阵，用这个生成模型得到语义上的相关项，也就是检查项和治疗项，供医疗人员查看。

## 病历文本的主题标签提取算法

### 4.2.1 改进的主题标签提取算法

针对主题标签提取中出现的问题，如何能够选择合适的标签，不仅能够和当前主题具有较高的语义相似度，还要和别的主题语义相似度较低，进而能够为主题选择最合适的标签。所以在这里引入主题区分度的概念，利用主题区分度正好能够起到去除和不止一个主题具有较高语义相似度的标签，增强了标签的独特性和区分度。在下面会对主题区分度进行介绍。

1）主题标签候选集的生成

首先我们要解决主题标签的生成问题，主题标签可以是句子，短语，还有词项，但是句子比较长，容易产生冗余，而且用句子作为标签并不能简明其意，直接说明标签内容，所以这里不会是用句子作为语义标签。病历文本还有其特点，就是存在大量医学实体，能够起到很好的语义表达的效果。因此，短语和词项相比而言是病历文本标签的较优选择。

针对模型中的主题，首先需要构造主题标签候选集。对于病历文本数据集，本节使用两种方法来构造主题标签候选集。

（1）搜狗细胞词库中的医学词汇大全

（2）根据病历文本语料库的分词。得到的病历文本主题标签候选集部分如图4-2所示：

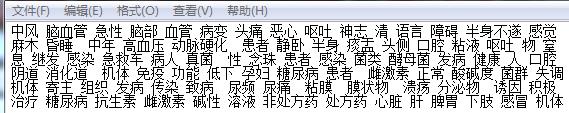


图4-2 部分分词结果图

2）引入主题区分度的标签提取

一个好的主题标签应该有的特性就是具有好的主题相似性，当然，有的词项和很多主题都有好的相似性，所以还需要考虑词项对不同主题的区分度，如何衡量标签和主题的相似性和标签对主题的区分度，可以通过其在词项表上的多项分布来实现，这样就可以根据语义相似度和主题区分度两个评价指标综合考虑，选择最合适的标签来代表主题。

首先定义一个标签由n个单词组成，表示为g=w1，w2，…wn，这些单词相互独立，如果这些单词中包含主题φk中概率值较高的词项，那么这个标签就和这个主题相关，相关度的公式如下所示：

 (4-2)

公式表示的是标签被这个主题k生成的似然率。但是这个公式有一个缺陷，并没有用到标签的语义信息。所以在这里选择一种比较合理的方法，利用标签的语义信息，即其在词项上的概率分布。

可以通过相对熵来计算标签和主题k的语义相似度，不过由于并没有使用改进的BM25F-HDP模型对给定候选标签集进行语义训练，我们需要对公式进行推导，使其不需要求取标签在词项上的多项分布。计算公式的推导过程如下：

 (4-3)

在上式中，所求就为标签和主题的语义相似度，值越小，说明标签和主题语义相似度越高。由于后一项对于所有的标签值是一样的，所以我们只需要考虑前一项值，上式可以简化为

 (4-4)

上式中将公式前面的符号去掉，所以结果变成值越大，标签和主题语义相似度越大。只考虑语义相似度会导致一个问题，就是如果标签和每个主题的语义相似度都很高，那这个标签并没有很好地区分度，不能够很好地说明各个主题之间的区别，这样的标签并不是一个好标签。如果一个标签和当前的一个主题语义相似度很高，而和别的主题语义相似度较低时，就可以认为这个标签是一个优秀的标签，而且和越多的主题语义相似度越低，说明这个标签的区分度越好，所以在这里引入标签和主题的区分度公式如下所示，其中α是调节参数，n表示主题个数减1，这是因为如果不进行相除的话相关度相加值会比较大，结果会出现负值。可以通过调整调节参数控制标签和主题的区分度。包含语义相似度的主题区分度的综合公式如下所示，在之后的实验中会用到以下公式。

 (4-5)

上式的标签提取公式就是在考虑语义相似度的基础上考虑了语义区分度，将两个因素结合起来进行考量，最终得到的主题标签提取公式，通过调节参数可以获得最优的主题标签。接下来的实验就是要对标签和主题进行语义相似度和主题区分度的计算和展示。

### 4.2.2 主题标签提取算法实验设计及结果分析

本次实验使用BM25F-HDP模型对病历文本集进行训练，迭代次数为100，生成主题个数为94，在这里选择第1个主题进行测试。在主题标签候选集中，取条件概率排名在前244个的关键词，进行标签提取公式运算实验。

下图所示是条件概率排名前15的关键词的柱状图，图中可以看到这15个词包含治疗、诊断等和具体病症没有直接联系的词，也包含心绞痛、心律、甲状腺等和疾病关系比较紧密的词项。之所以要列出该主题下条件概率排名靠前的词项，就是为了说明这个主题的主要语义是什么，描述的主要内容是什么，为之后的主题标签提取结果的分析提供依据。

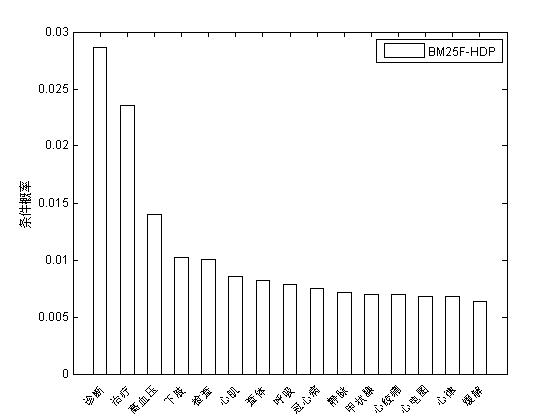


图4-3 条件概率排名前15的关键字

下图所示是主题标签候选集中的标签和第1个主题相似度排名最高的8个标签。图中看出，排名前8的标签分别是治疗、诊断、高血压、检查、下肢、查体、呼吸和心肌，而根据上述的条件概率排名较高的关键字可以看出，排名前7个的主题标签并没有体现出该主题中所包含的和心相关的语义信息，而且可以观察出治疗，诊断这些词汇就是语义区分度不高的词汇，因为这些词项在所有的病历文本中几乎都会存在，而不能仅仅代表某一个主题，丧失了主题的区分度。

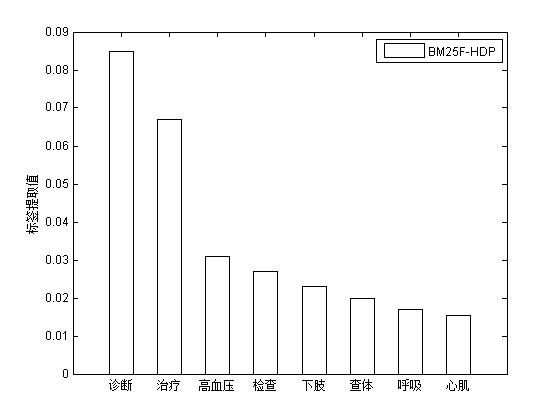
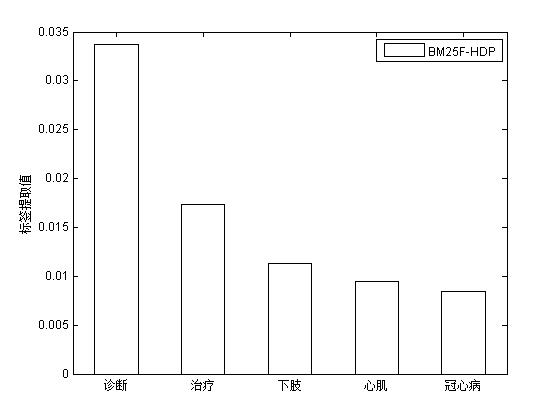
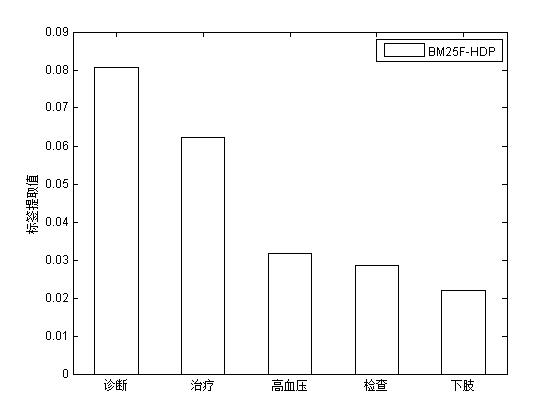


图4-4 仅考虑语义相似度的标签提取值排名前8的标签

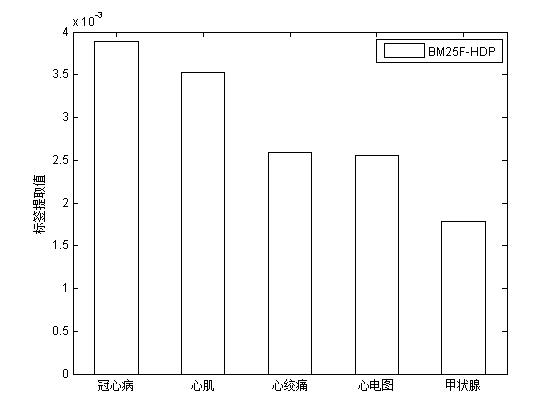
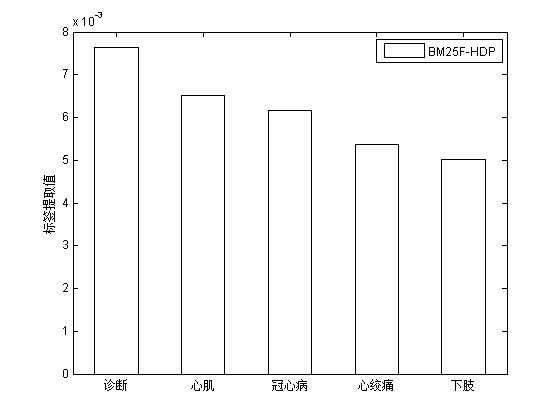
针对只利用语义相似度对主题标签进行提取的问题，下面的实验在语义相似度的基础上考虑了主题区分度，利用最终的标签提取公式进行实验，而标签提取公式中有调节参数α，调节参数的设置在一定程度上会影响标签提取的结果，只有在不断的调试过程中，才能够选择最优的参数值。下图所示就是不同的调节参数值下提取的排名最高的5个主题标签。下图所示是α为1和10时考虑主题区分度的标签提取值排名前5的标签。



(a) 在α=1时标签提取值排名前5的标签 (b) 在α=10时标签提取值排名前5的标签

图4-5 在α为1和10时考虑主题区分度的标签提取值排名前5的标签

下图所示是α为15和20时考虑主题区分度的标签提取值排名前5的标签。



(a) 在α=15时标签提取值排名前5的标签 (b) 在α=20时标签提取值排名前5的标签

图4-6 在α为15和20时考虑主题区分度的标签提取值排名前5的标签

可以看出，在α取1时，标签提取结果和只使用语义相似度进行标签提取的结果没有变化，当α取10时，标签提取结果中开始出现和心相关的词项心肌和冠心病，当然，治疗和诊断的标签提取排名还是前两名。而在α取到15时，出现了和心相关的词项三个，治疗这个词项已经跌出排名前5，诊断这个词项还是排名第1，在α取到20的时候，能够看到5个标签全部是和直接病症相关的，其中四个和心相关。诊断和治疗这两个主题区分度不高的词项已经不再标签提取的结果当中。

最终可以得出，第一个主题中条件概率排名前15个词项是治疗 ，检查、诊断、静脉，呼吸、高血压、甲状腺、心绞痛、心肌、冠心病、缓解、心电图、心律、查体、下肢和缓解，而经过主题标签提取算法提取出的主题标签是冠心病，心肌，心绞痛，心电图和甲状腺，可以利用已知的条件概率排名靠前的词项对这些提取出的标签能否代表当前语义进行评价。

主题标签的评价是一个非常主观的过程，并不能用具体的数字来说明主题标签生成的好坏，所以在这里用到了人工的评价。当然，在最终的结果中结果中可以看到主题标签和当前主题具有良好的相似度，提取出的词项基本涵盖了该主题中条件概率排名较高的关键词，像心肌和心绞痛这些和心相关的词语还是提取出来了，而关键词中多次提到了和心脏相关的词语。也具有良好的区分度，像“治疗”，“诊断”这些词语比较宏观，在很多的主题中都会出现，并不能很好的代表某个语义，所以随着调节参数α的增大，特征提取算法对于主题标签的区分度影响越大，导致区分度不高的词项在特征提取排名中不断后移。覆盖度还有待提高，比如“高血压”这些词并没有覆盖到其语义。

## 病历文本的相关项提取算法

### 4.3.1 基于BM25F-HDP的相关项提取算法

1）相关项候选集的生成

本节中的相关项候选集分为两部分，相关检查项候选集和相关治疗项候选集。训练文本数据集使用全部四个文本段，相关检查项的测试文本数据集只使用入院情况文本段，相关治疗项的测试文本数据集使用入院情况、查体、诊断三个文本段。

本文中的候选集分为检查项和治疗项，检查项是对查体文本段分词并且处理后得到的词项，而治疗项是对治疗原则文本段进行分词并且处理后得到的词项，如下所示：

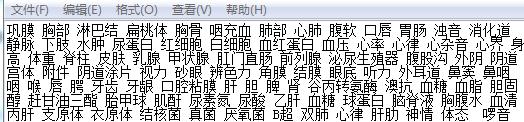


图4-7 相关检查项候选集图

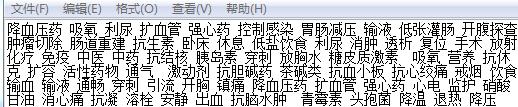


图4-8 相关治疗项候选集图

2）语义关联度计算

对于一个检查项或者治疗项，可以通过建立的BM25F-HDP对其进行打分，在模型训练结果中知道，每一个主题都对应其在词项上的多项概率分布，换一种思路，也可以认为是一个词项在各个主题上的多项分布，这个时候我们可以通过利用相对熵来求解词项和文档之间的相关度，相对熵本来所求是已知两个变量的多项概率分布，求其相似度，在这里这两个变量分别是词项和文档，并且在模型中知道其在主题上的多项概率分布是什么，这样就可以求得其相似值是多少，在这里定义主题k在词项w上的概率值是rw(k),也可以理解为词项w在主题k上的概率值是rw(k)。词项和文档相关度公式如下所示：

 (4-6)

其中得到各个候选词项的相关度值后，进行排序，并且将排名前n项输出为所得的相关项。

### 4.3.2 相关项测试方法

相关性的实验测试大量用于信息检索领域，测试的是检索项和搜索出来的相关项的关系，而在本实验中，测试的是病历文本文档和检查项或者病历文本文档和治疗项的相关性。

在本节中的测试集由以下部分组成：

1）一篇病历文档；

2）一组检查项和一组治疗项；

3）对于选取的检查项或者治疗项，需要赋予其相关或者不相关的二值判断结果。

在信息检索领域中，具有两个重要的度量值，准确率(Precise)和召回率(Recall)，是评价检索结果的重要指标，在本节中，用以衡量检查项和治疗项的相似度结果评价指标。其中召回率是检索结果中相关项数目和所有相关项的比值，准确率是检索出的相关项和检索出的所有项的比值。准确率和召回率的取值都在0和1之间，数值越接近于1，则相关准确率和相关召回率就越高。精确率的计算公式如下所示:

 (4-7)

上式中的精确率反映了分类器对整个诊疗项和治疗项的判定能力，但由于相关项存在着不均衡性，并不能很好的适应本节中相关检查项和治疗项的生成性能，而召回率和准确率只是针对返回的检查项和治疗项中真正的相关项进行评估。而在通常意义来说，好的模型能够同时保证较高的召回率和准确率，所以，需要找到一种综合考虑两个方面的方法。

由于召回率和准确率会出现矛盾的情况，所以为了更好地运用召回率和准确率，综合考虑他们，本文使用综合评价指标F值(F-Measure)作为衡量指标来测量性能。如下所示：

 (4-8)

其中α的取值一般取为1，这是因为召回率和准确率权重值相等，P表示准确率，R表示召回率，所以，平衡值F记成F1，如下所示：

 (4-9)

在以下试验中，会以此公式进行相关计算和实验对比，并且对实验结果进行分析和模型优劣的评价。

### 4.3.3 相关项测试实验

本次实验中利用病历文本数据集来测试基于主题模型的相关项生成算法，其中主题模型分别使用BM25F-HDP和HDP来对比相关项生成的性能优劣。

其中选择190篇病历文本作为训练集，47篇病历文本作为测试集。分别在不包含查体文本段的病历文本数据集和不包含治疗原则文本段的病历文本数据集上进行模型训练，这样就使得入院情况和诊断文本段分别和查体文本段、治疗原则文本段建立语义上的关联。

改进的HDP模型中设置不包含治疗原则的每一篇病历文本返回的关键词数目N为50，不包含查体的每一篇病历文本返回的关键词数目N也为50。评价指标选择平衡F值F1。

下表所示就是相关检查项和相关治疗项的实验统计表。表中包含相关项返回数目，相关项未返回数目，非相关项返回数目等。

表4‑1 相关项表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 检查项 | | 治疗项 | |
| HDP | BM25F-HDP | HDP | BM25F-HDP |
| 相关项返回数目 | 46 | 71 | 11 | 20 |
| 相关项未返回数目 | 528 | 503 | 104 | 95 |
| 非相关项返回数目 | 2304 | 2279 | 2339 | 2330 |

下表所示是相关检查项和相关治疗项的准确率、召回率和平衡F值的计算结果。

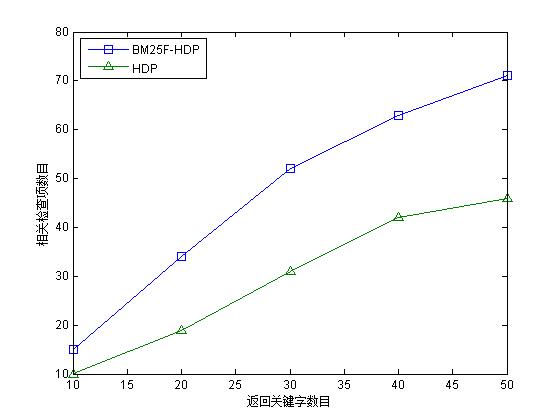
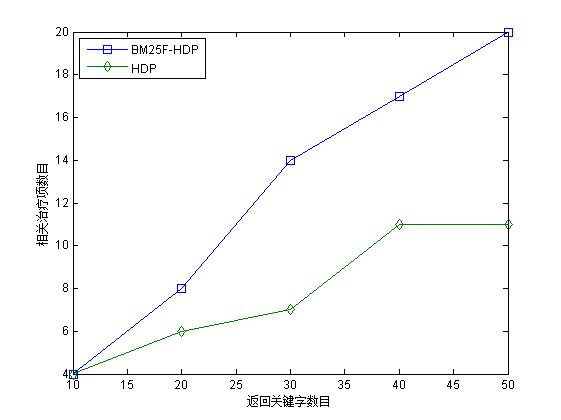
表4‑2 相关项计算表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 检查项 | | 治疗项 | |
| HDP | BM25F-HDP | HDP | BM25F-HDP |
| 召回率 | 8.01% | 12.37% | 9.57% | 17.39% |
| 准确率 | 1.96% | 3.02% | 0.47% | 0.85% |
| 平衡F值 | 3.15% | 4.85% | 0.90% | 1.62% |

由上述实验结果可以看出，不论是返回相关检查项还是相关治疗项，在召回率，准确率还有平衡F值F1三个指标上BM25F-HDP模型都是优于HDP模型的。

在上述试验中，返回关键词的个数N设置为50，其实N值的增大会导致返回关键词中相关项的增加，所以为了更加全面的测试利用主题模型进行相关项生成的性能，对不同的返回关键词个数N进行实验。

如下图所示就是在不同的返回关键字数目下返回的相关检查项数目和相关治疗项数目，可以看出随着返回关键字数目的增加，返回相关检查项的数目也增加。并且对比了HDP模型和BM25F-HDP模型在返回相关检查项上的表现，可以看到，当返回相同的关键字数目时，BM25F-HDP的表现要比HDP模型要好，说明了BM25F-HDP的相关项生成性能优于HDP模型。



(a) 相关检查项变化趋势 (b) 相关治疗项变化趋势

图4-9 相关项变化图

上图可以看出当返回相同的关键字数目时，BM25F-HDP模型在相关治疗项上的表现要比HDP模型要好。

下表4-2所示是不同的返回关键字数目下，在不同的模型下检查项和治疗项的返回数目表。

表4‑3 相关项数目表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 返回关键字数目 | 检查项 | | 治疗项 | |
| HDP | BM25F-HDP | HDP | BM25F-HDP |
| 10 | 10 | 15 | 4 | 4 |
| 20 | 19 | 34 | 6 | 8 |
| 30 | 31 | 52 | 7 | 14 |
| 40 | 42 | 63 | 11 | 17 |
| 50 | 46 | 71 | 11 | 20 |

下图4-10所示为BM25F-HDP和HDP在不同N下平衡F值F1的对比图，

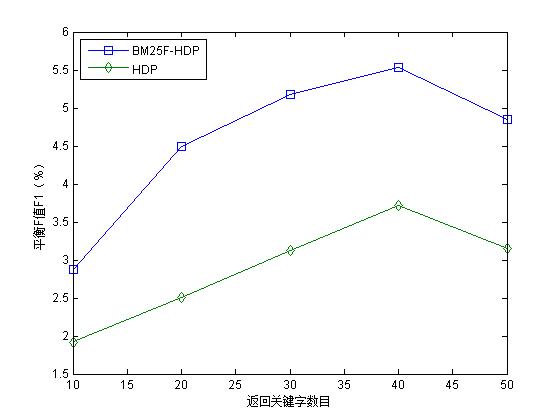


图4-10 相关项生成对比图

可以看到。N值不同时，BM25F-HDP的性能都要优于HDP，在N取值从10到40的过程中，可以看到平衡F值F1在增加，这说明召回率的增加和准确率的降低对于相关项的生成还是有所提高的。不过在N=40之后，两个模型都出现了平衡值的下降，召回率的增加带来的正面影响已经不能够抵挡准确率的降低带来的负面影响。所以可以得出，N的选取对于相关项生成的性能影响并不是线性的，而是有一个最合适的值，而且这个值需要在实验中获得。

下表4-3是不同关键字返回数目和平衡F值的具体数值，可以看到在平衡F值的提升都在1到2个百分点之间。

表4‑4 平衡F值对比表

|  |  |  |
| --- | --- | --- |
| 关键字返回数目 | HDP平衡F值 BM25F-HDP平衡F值 | |
| 10 | 1.92% | 2.87% |
| 20 | 2.51% | 4.49% |
| 30 | 3.12% | 5.18% |
| 40 | 3.72% | 5.53% |
| 50 | 3.15% | 4.85% |

## 4.4本章小节

本章通过在病历文本数据集上建立BM25F-HDP模型，改进了只利用语义相似度对主题标签进行提取的算法，考虑了主题区分度。利用主题模型进行了相关项的生成实验，对比了本文提出的BM25F-HDP模型和普通HDP模型性能，并证明了在此病历文本数据集上，BM25F-HDP的性能是优于普通HDP模型的。

# 病历服务系统的实现与测试

前面章节针对病历文本数据集的各种应用进行了测试和分析。本章将会在改进的主题模型BM25F-HDP的基础上搭建一个病历服务系统，将病历的相关应用在系统中是实现。

## 需求背景

互联网的发展给很多人的生活和工作带来了很多便利，其中医疗领域的茁壮成长正在造福很多的医疗人员和病患。然而针对而大量的病历管理是一个棘手的问题。所以通过建立一个系统对病历进行特征提取，为医疗人员提供辅助信息支持，就能够极大的减小医疗人员的工作量。所以，本章设计并且实现了一个病历服务系统，将改进的BM25F-HDP模型，主题标签提取算法，基于似然率最大化的病历文本总结提取方法，通过文本在主题分布上的余弦相似度计算进行相似病历匹配，还有基于主题模型的相关项提取算法应用到系统当中，医疗人员就可以快速直观的对病历文本的特征提取结果进行应用和分析，为医疗人员提供可靠的决策支持。

## 系统需求分析

病历文本系统主要以主题模型为基础，数据预处理后的病历文本为数据集，以实现病历主题标签的提取，病历总结的提取，相似病历的匹配，病历相关检查项和治疗项的生成等功能，目的是让医疗人员从可视化的界面中直观的获取病历关键信息以辅助医疗人员分析。同时也可以进一步说明本文提出的基于主模型的病历特征提取的实用性和有效性。

本系统中主要包括三类用户，第一类是医疗人员，他们是系统的使用者，可以进行登录、注册等基本的操作，也可以查看主题标签，查看病历总结，查询相似病历，查询病历相关检查项和相关治疗项等操作。在查看主题标签时需要选择具体主题；在查看病历总结时需要选择具体病历；在查询相似病历和相关检查项时需要输入拥有入院情况和查体的病历；在查询相关治疗项时需要输入拥有入院情况、查体和诊断的病历。第二类用户是业务管理员，他们主要负责病历文本的预处理，模型超参数的配置和训练，提取主题标签，病历总结，相关项候选集并且存储到数据库。第三类用户是系统管理员，主要进行系统管理和文本整理工作。其中系统管理又包括用户管理，角色权限管理，数据字典管理，索引管理等。文本整理包括病历文本的采集，病历文本的更新。

根据上述的需求分析，使用UML建模方法可得到的病历服务系统的用例图如图5-1所示：



图5-1 用例图

病历服务系统的功能性需求主要包括以下几部分：

* + - 1. 病历文本预处理，包括文本分词，停用词过滤，专有名词过滤等；
      2. 模型训练，并将相关结果存入数据库；
      3. 病历文本的主题标签提取；
      4. 病历文本的总结提取；
      5. 病历文本的相似度计算；
      6. 病历文本相关项的提取；

其中病历文本的预处理是模型训练的前提，为了保证分词效果，使用停用词对常用的无意义词语进行过滤，并且为了保证分词能够不将专业名词进行拆分，在这里使用搜狗细胞词库中的专有医学名词库。而模型的训练会生成病历文本-主题的二维矩阵和主题-词项的二维矩阵，为之后的病历文本特征提取做准备。其中特征提取分为四个方面，主题标签的提取、病历总结提取、相似病历匹配和病历相关检查项和相关治疗项的提取。标签提取采用综合考虑语义相似度和主题区分度的主题标签提取算法，这样提取的标签不仅仅能够很好地代表当前的主题，还能够较好的和别的主题进行区分。这样能够辅助医疗人员知道病历文本集中会有哪几类的病历。病历总结的提取是在主题模型的基础上，运用似然率最大化的方法对病历的关键词进行提取，这些关键词组成了病历的总结，辅助医疗人员能够更快的了解病历主要描述的内容是什么，然后决定是否去详细的查看病历。相似病历的匹配是医疗人员希望查看是否有和输入病历相类似的病历。采用病历在语义分布上的余弦相似度进行匹配。而相关项的提取包括相关检查项和相关治疗项的提取。相关检查项是指根据病历摘要文本系统给出患者可能需要进行检查的项目有哪些。而治疗项是指根据摘要，查体和诊断文本给出患者可能需要接受的治疗手段或者治疗药物。相关项的提取是根据以往的诊断经验，给医疗人员提供辅助的医疗支持。

如下表所示是相似病历匹配的用例描述：

表5-1 相似病历用例描述

|  |
| --- |
| 用例：相似病历匹配 |
| 简单描述：将用户输入的病历进行相似病历匹配  主参与者：医疗人员  副参与者：数据库  前置条件：本系统已被运行并且用户成功登陆  主要流程：   1. 用户打开网站并且登录成功 2. 输入需要进行相似匹配的病历 3. 点击匹配按钮 4. 系统为用户输入的病历进行相似匹配 5. 得到匹配结果，用户可以看到相似病历内容   后置条件：病历匹配结束，用户得到相似病历 |

如下图所示是相关检查项提取的用例描述：

表5-2 相关检查项提取用例描述

|  |
| --- |
| 用例：相关检查项提取 |
| 简单描述：为用户输入的病历入院情况文本段生成检查项  主参与者：医疗人员  副参与者：数据库  前置条件：本系统已被运行并且用户成功登陆  主要流程：   1. 用户打开网站并且登录成功 2. 输入需要提取检查项的病历入院情况文本段 3. 点击查询按钮 4. 系统为用户输入的病历入院情况文本段进行检查项提取 5. 得到提取结果，用户可以看到提取的检查项的柱状图   后置条件：检查项提取结束，用户得到相关检查项 |

在病历服务系统中，首先要在主题模型上对预处理后的病历文本分词进行训练，所以需要一个文本训练模块，文本训练模块是系统实现其他模块的基础。在文本训练完成后，要根据文本训练结果计算标签候选集中的词项和模型主题的相对熵，选择排名靠前的词项作为标签；计算关键词在一篇病历中的似然率，选择较大的关键字作为病历总结；实现这些功能和医疗人员无关，只需要业务人员在模型训练完成之后进行操作即可。所以需要一个这样的模块实现这些功能，在这里称之为离线计算模块。如下图5-2所示是文本训练模块和离线计算模块的活动图：



图5-2 文本训练和离线计算活动图

在系统中医疗人员可以进行相似病历匹配查询、相关项的查询，也需要一个模块来完成这些功能，在这里称之为在线计算模块。下图5-3所示是在线计算模块中相似病历匹配的活动图：



图5-3 相似度计算活动图

## 系统设计

### 5.3.1 系统架构设计

本系统使用B/S架构，前段使用JSP技术，客户端界面运行在浏览器中，服务器端使用JAVA语言实现所有功能，服务器采用开源的Apache Tomcat 7.0，数据库管理使用Mysql数据库。

系统的采用springMVC框架，分为Controller层、Service层和Dao层。其中Controller层负责系统的功能模块流程的控制，并且和界面结合比较紧密。Service层主要负责功能模块的逻辑应用设计，Dao层主要做数据持久层的工作。

### 5.3.2 系统功能模块设计

根据系统的需求分析，主要的功能模块包括系统管理模块、文本训练模块、离线计算模块和实时计算模块。系统的功能结构图如图5-4所示：



图5-4 系统功能模块图

病历服务系统中各个模块的具体介绍如下所示：

1）文本训练模块

文本训练模块的主要作用就是完成对病历文本的分词，过滤停用词，添加专业名词等预处理操作。使用停用词，专用名词就是为了避免高频无意义词对模型训练的影响，而添加专业名词就是为了防止分词器无法识别专有名词而错误分词。在预处理完成后，让主题模型读取分词进行模型训练，产生病历文本到主题，主题到词项的两个二维矩阵，还有分词字符串，并将它们存储到数据库中，以便后续的其他模块进行进一步的计算。

2）离线计算模块

离线处理模块包括主题标签的提取、病历总结的提取，这些提取运算和医疗人员没有关系，只需要拥有主题标签候选集，模型训练好之后就可以由业务管理员操作完成。主题标签提取中需要计算标签候选集中词项和病历文本的提取值，选取排名靠前的词项代表这个主题，并存储到数据库中；总结提取中需要计算关键字和病历的似然率，选择似然率最大的几个关键词作为这一篇病历的总结，并存储到数据库中。当医疗人员登录系统后，可以直接查看主题标签和病历总结这些内容。

以上两个模块的时序图如下图5-5所示：



图5-5 文本训练模块和离线计算模块时序图

3）实时计算模块

实时计算模块包括相似病历匹配，相关项的生成。这两种都是需要医疗人员输入相关的病历文本，后台会在主题模型上进行相关计算。此模块的时序图如图5-6所示：



图5-6 实时计算模块时序图

### 5.3.3 数据库设计

数据库选用Mysql数据库存储系统相关数据。其中文本训练模块，对病历进行预处理后的词项需要进行存储，模型训练之后的文本-主题矩阵、主题-词项矩阵以字符串的形式进行存储。主题标签，病历总结也需要存储以便医疗人员后续查看。检查项候选集和治疗项候选集需要存储在表中以便进行相关度计算。相似病历病历匹配和相关项生成是实时性质的查看，结果不需要存储，所以没有设计数据库表。具体的表结构信息如图5-7所示：



图5-7 数据库概念模型图

从图中可以看出，用户表、词项表、病历表、主题表和标签表等几个表都是相互关联的，存在一对多的关系，接下来选择展示数据库中关键核心表的属性字段，如下所示。

下表是主题表，包含病历文本主题相关字段，如下表5-3所示：

表5-3 topic

|  |  |  |  |
| --- | --- | --- | --- |
| 列名 | 数据类型 | 完整性约束 | 说明 |
| topic\_id | CHAR(32) | 主键 | 代码生成编号 |
| topic\_number | INT |  | 主题编号 |
| words\_distribution | VARCHAR(10000) |  | 词分布字符串 |
| topic\_label | VARCHAR(50) |  | 标签 |
| similarity\_words\_distribution | VARCHAR(10000) |  | 摘要和查体词分布 |

下表是病历表，表示病历相关特征字段，如下表5-4所示：

表5-4 medical\_record

|  |  |  |  |
| --- | --- | --- | --- |
| 列名 | 数据类型 | 完整性约束 | 说明 |
| medical\_record\_id | CHAR(32) | 主键 | 代码生成编号 |
| medical\_record\_number | INT |  | 病历编号 |
| content | VARCHAR(1000) |  | 病历内容 |
| medical\_record\_words | VARCHAR(1000) |  | 分词字符串 |
| topic\_distribution | VARCHAR(500) |  | 主题分布字符串 |
| summary | VARCHAR(50) |  | 总结 |
| similarity\_words | VARCHAR(1000) |  | 摘要和查体分词 |
| similarity\_topic\_distribution | VARCHAR(500) |  | 摘要和查体主题分布 |
| similarity\_content | VARCHAR(1000) |  | 摘要和查体内容 |

## 5.4系统各模块的实现

### 5.4.1 文本训练模块

本模块的实现就是在病历文本数据集上进行BM25F-HDP模型训练，首先从数据库中读取全部病历，进行分词，分词器使用中科院NLPIR汉语分词系统，实现Clibrary接口，将病历进行分词，分词器中要使用停用词库，并且添加专业词库，避免分词器将专业词语进行拆分。生成的分词结果存储到数据库中。然后根据分词结果进行词频统计，形成每篇病历文本的词频生成结果。结果的格式如下，每一行表示一篇病历文本，其中最前面的数字代表这篇文档中分词的个数，后面每隔一个空格表示一个词项，冒号前的数字代表这个词项的编号，这个编号是针对所以病历中词项的编号，冒号之后的数字表示这个词项在这篇文档中出现的次数。而模型训练函数需要读取这样一个结果。

层次狄利克雷过程训练的过程使用吉布斯采样进行模型训练，在迭代过程中需要进行相关操作。吉布斯采样就是不停地进行迭代，对所有词项重新估计主题分布，直到主题分布实现收敛，不再变化，这样就生成了最终每篇文档的主题分布和每篇主题下的词项的分布，所以在迭代到当前词项时，需要排除当前词项的主题分配，并且排除该主题的文档分配，根据其他词项的主题分配来计算当前词项分配各个主题的概率，得到当前词项属于所有主题的概率分布后，为该词项采样一个新的主题，然后利用同样的方法不断更新下一个词项的主题，直到发现每个文档下的主题分布和每个主题下的词项分布收敛。

1）在迭代到文档m中的第n个词项时，将主题k下所有词项的个数减1，词t在主题k中出现的次数减1，主题k在文档m下出现的次数减去词项的特征权重值。代码如下所示：

//文档m的第n个词语的的主题编号

k = z[m][n];

//主题k在文档m下出现的次数

nmk[m].set(k, nmk[m].get(k) - weightVocabulary[m][n]);

//词t在主题k中出现的次数

nkt.get(k)[t]=nkt.get(k)[t]-1;

//主题k下所有词项的个数

nk.set(k, nk.get(k) - 1);

2）为词项采样一个新的主题，并将产生的新主题k赋值给z[m][n]，代码如下所示：

double psum = 0;

for (int kk = 0; kk < K; kk++) {

k = kactive.get(kk);

pp[kk] = (nmk[m].get(k) + alpha \* tau.get(k)) \*

(nkt.get(k)[t] + beta) / (nk.get(k) + V \* beta);

psum += pp[kk];

}

//新组件的似然率

if (!fixedK) {

pp[K] = alpha \* tau.get(K) / V;

psum += pp[K];

}

double u = rand.nextDouble();

u \*= psum;

psum = 0;

int kk = 0;

for (; kk < K + 1; kk++) {

psum += pp[kk];

if (u <= psum) {

break;

}

}

if (kk < K) {

k = kactive.get(kk);

z[m][n] = k;

}

3）在给词项选择新的主题后，需要将将主题k下所有词项的个数加1，词t在主题k中出现的次数加1，主题k在文档m下出现的次数加上词项的特征权重值，和1)中的代码相对应，代码如下所示：

nmk[m].set(k, nmk[m].get(k) +weightVocabulary[m][n]);

nkt.get(k)[t]=nkt.get(k)[t]+1;

nk.set(k, nk.get(k) + 1);

4）当主题k下的词数为0时，将这个主题去掉，并且将主题数减1，也就是层次狄利克雷过程中的主题个数自动减少，代码如下所示：

if (inited && nk.get(kold) == 0) {

kactive.remove((Integer) kold);

kgaps.add(kold);

assert (Vectors.sum(nkt.get(kold)) == 0

&& nk.get(kold) == 0 && nmk[m].get(kold) == 0);

K--;

//更新CRP中的参数Tau

updateTau();

}

5）此方法将主题个数加1，并将增加的主题编号作为词项t的新主题，也就是层次狄利克雷过程中主题个数自动增加。

private int spawnTopic(int m, int t) {

int k;

if (kgaps.size() > 0) {

k = kgaps.first();

kgaps.remove(k);

kactive.add(k);

nmk[m].set(k, 1.0);

nkt.get(k)[t] = 1;

nk.set(k, 1);

} else {

k = K;

for (int i = 0; i < M; i++) {

nmk[i].add(0.0);

}

kactive.add(K);

nmk[m].set(K, 1.0);

nkt.add(new int[V]);

nkt.get(K)[t] = 1;

nk.add(1);

tau.add(0.);

}

K++;

if (pp.length <= K) {

pp = new double[K + ppstep];

}

return k;

}

### 5.4.2 离线计算模块

本模块是在文本训练模块的基础上进行离线计算，分别为标签集相对熵计算和关键字似然率计算。主要实现主题标签的提取和病历总结的提取。

1）以下方法是病历总结提取中关键词似然率计算方法，并对似然率进行排序，下面代码计算的是第i个篇病历文本的关键字似然率。的其中m\_k[][]表示文档-主题矩阵，number[k]表示第k个词语编号，并对求出的各个词项的似然率进行排序。

for (int k = 0; k < number.length; k++) {

for (int j = 0; j <m\_k[i].length; j++) {

//本篇文档第k个词语的似然概率 关键词提取

keyword[k] =keyword[k] + m\_k[i][j]\*k\_t[j][number[k]];

}

}

double temp;

for (int i = 0; i < keyword.length; i++) {

int k = i;

//词项似然率从小到大排序

for (int j = keyword.length - 1; j >i; j--) {

if (keyword[j] > keyword[k]) k = j;

}

temp = keyword[i];

keyword[i] = keyword[k];

keyword[k] = temp;

}

2）下面方法计算标签候选集的特征提取值，由于排序和上面似然率排序一样，所以没有贴出。

double[][] kld = new double[k\_t.length][nu.length];

double[] su = new double[kld[1].length];

double[] dis = new double[kld[1].length];

for (int i = 0; i < k\_t.length; i++) {

for (int k = 0; k < nu.length; k++) {

//kl散度结果

kld[i][k] = k\_t[i][nu[k]]\*Math.log10(vocabularyFrequency[nu[k]]);

}

}

//存储某个词项和所有主题的kl散度和主题个数的商

for (int i = 0; i < kld[1].length; i++) {

for (int j = 0; j < kld.length; j++) {

su[i] = su[i] + kld[j][i];

}

su[i] = 20\*su[i]/k\_t.length;

}

//计算区分度

for (int i = 0; i < kld[1].length; i++) {//i表示第i个词项

dis[i] = kld[1][i]-su[i];

}

首先进入病历服务系统登陆后的界面如下图5-8所示，下图中会在首页展示数据库中的病历，其中病历编号倒序展示，病历内容截取前244个字符串展示，如果医疗人员想要了解此病历的全部内容是什么，可以点击进入查看，在下面还有查看更多病历的按钮，医疗人员可以查看更多病历。



图5-8 系统首页图

下图5-9所示是主题标签提取的界面，其中医疗人员可以选择自己想要了解的主题，然后会在下面显示相关内容，最前面试主题编号，之后是主题描述，由于主题有在词项上的多项分布，所以在主题描述中选择分布概率最高的15个词项作为主题的描述，而在标签中选择特征提取值排名前5的词项作为主题标签，可以通过主题描述来观察主题标签是否能够较好的覆盖主题语义，可以看出标签主要说明的是和心脏相关的描述。则第一个主题主要描述心脏相关的内容。



图5-9 主题提取图

下图5-10所示是病历总结提取的界面，首先选择第51号病历，会出现病历编号，病历的详细内容，还有病历总结，其中病历总结选取似然率排名前13个关键词，帮助医疗人员了解这篇病历的大概内容，进而决定是否详细查看这篇病历。



图5-10 病历总结图

### 5.4.3 实时计算模块

本模块在文本训练结果的基础上进行实时计算，即相似病历匹配的语义相似度计算，检查项和治疗项的相关度计算。

1）下面代码对病历文本在主题分布上的相似度进行了计算，其中t表示新输入的第t篇病历，它会和每一篇病历进行相似度计算，相似度值存储在cos[]数组中。

double[] cos = new double[m\_k.length];

double a = 0,b = 0,sqrta = 0,sqrtb = 0;

for (int j = 0; j <m\_k[t].length; j++) {

sqrta =sqrta + m\_k[t][j]\*m\_k[t][j];

}

sqrta = Math.sqrt(sqrta);

for (int i = 0; i < m\_k.length; i++) {

for (int j = 0; j <m\_k[i].length; j++) {

a =a + m\_k[i][j]\*m\_k[t][j];

sqrtb =sqrtb + m\_k[i][j]\*m\_k[i][j];

}

sqrtb = Math.sqrt(sqrtb);

b=sqrta\*sqrtb;

cos[i] = a/b;

a=0;

sqrtb = 0;

}

2）下面代码对输入病历和相关项的相关度进行了计算，首先要计算输入病历和所有词项的相关度，而这些词项中不仅仅有相关项。

//相关度存储数组

double[] sim = new double[k\_t[0].length];

int[] index = new int[k\_t[0].length];

int[] item = new item[5];

for (int k = 0; k < k\_t[k].length; k++) {

for (int j = 0; j < m\_k[i].length; j++) {

// sim[k]表示输入编号为i的文档和第k个词项的sim相似度，主题编号为j

if (k\_t[j][k] > 0.00001 && m\_k[i][j] > 0.00001) {

sim[k] = sim[k] + m\_k[i][j]

\* Math.log10(m\_k[i][j] / k\_t[j][k]) + k\_t[j][k]

\* Math.log10(k\_t[j][k] / m\_k[i][j]);

}

}

}// 内两层循环得出的是第i篇文档和所有词项的sim相似度 存储在sim[]中

3）要对所以词项的相关度进行排序，在这里选择冒泡排序。

double middle;

int middleIndex;

for (int j = 0; j < sim.length; j++) {

for (int j2 = 0; j2 < sim.length - j - 1; j2++) {

if (sim[j2] < sim[j2 + 1]) {

middle = sim[j2];

sim[j2] = sim[j2 + 1];

sim[j2 + 1] = middle;

middleIndex = index[j2];

index[j2] = index[j2 + 1];

index[j2 + 1] = middleIndex;

}

}

}

4）输出新病历关键词排名前30的词项,然后判断其中哪些属于相关项，并将相关项的下标存储在数组中。

int x = 0;

for (int j = 0; j < 30; j++) {

// 数组内存储的值为1时，说明该词项是检查项，词项下标是index[j]

if (zhenliao[index[j]] == 1) {// 确定检查项的下标,并将检查项下标赋值给item数组

item[x] = index[j];

x++;

}

}

下图5-11所示是相似病历匹配界面，首先需要输入病历的入院情况和查体文本段，点击匹配按钮，后台会计算出输入文本段的主题分布，然后计算和数据库中病历在主题分布上的的相似度，输出相似度最高的病历。此模块是基于只有入院情况和查体文本段的模型训练。



图5-11 相似病历匹配图

下图5-12所示是病历相关检查项的运行界面，需要输入入院情况文本段，点击查询，系统会训练出输入文本段在主题上的概率分布，然后计算检查项和输入文本段的相关度，选择排名最高的3个检查项并且输出结果。此模块的实现是基于病历全部文本段的模型训练。针对入院情况中的描述，系统给出的诊疗的项目有虹膜，双肺，心音等。



图5-12 检查项图

下图5-13所示是病历相关治疗项的运行界面，需要输入入院情况、查体和诊断3个文本段，点击查询，系统会训练出输入文本段在主题上的概率分布，然后计算治疗和输入文本段的相关度，选择排名最高的3个治疗项输出。此模块的实现是基于病历全部文本段的模型训练。给出的治疗手段有降压，抗感染，透析等。



图5-13 治疗项图

## 5.5系统测试

### 5.5.1 系统测试环境

系统的测试环境如下表所示：

表5-5 系统测试环境

|  |  |
| --- | --- |
| 项目 | 数据类型 |
| 操作系统 | Windows 7 Ultimate |
| CPU | Intel(R) Core(TM) i5-2430M |
| 内存 | 4GB |
| 开发工具 | Eclipse |
| 数据库管理系统 | Mysql 5.7 |

### 5.5.2 系统功能测试

本章的前几节将系统的分析到设计进行了详细的介绍，接下来会检测系统能否满足功能需求。在测试中，功能测试又称为数据驱动测试或者黑盒测试，测试了产品的可操作行为和特性，而且相较于白盒测试，不需要了解系统内部的代码实现，也就是说，不需要了解系统内部的详细细节，只需要测试软件中提供的各个功能能否满足预期需求，一般从系统软件的架构和界面出发，按照需求编写测试用例，输入用例来测评实际结果和期望结果，进而提出使产品更加满足用户使用的要求。其中功能测试用例表如下所示：

表5-6功能测试用例表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 名称 | 功能点 | 预想的成功结果 | 实际结果 | 测试结论 |
| 文本训练模块 | 文本分词 | 生成分词结果 | 生成分词结果 | 成功 |
|  | 模型训练 | 生成两个二维矩阵 | 生成两个二维矩阵 | 成功 |
| 离线计算模块 | 标签相对熵计算 | 返回主题标签 | 返回主题标签 | 成功 |
|  | 关键字似然率计算 | 返回病历关键字总结 | 返回病历关键字总结 | 成功 |
| 实时计算模块 | 相似病历匹配 | 返回相似病历 | 返回相似病历 | 成功 |
|  | 检查项相关度计算 | 返回相关检查项 | 返回相关检查项 | 成功 |
|  | 治疗项相关度计算 | 返回相关治疗项 | 返回相关治疗项 | 成功 |

## 5.6本章小结

本章为了验证基于BM25F-HDP主题模型的病历文本特征提取的有效性，实现了一套病历服务系统，其中具体实现了文本训练模块、离线计算模块和实时计算模块。利用病历文本对系统进行了实验，验证了系统的可行性和实用性。最后对系统进行测试和界面展示，同时发现了系统还有需改进之处。

# 结论与展望

## 结论

随着这些年医疗信息的飞速发展，产生了大量的病历文本，并且含有丰富的医疗信息。这些病历文本语法结构松散，医学用词不一致给计算机的处理带来了很多障碍。本文从统计学的角度出发，提出一种思路，用主题模型将病历文本在主题上结构化，利用量化的主题信息做后续的病历特征提取工作。

总的来说，本文的研究内容分为对主题模型的研究和在主题模型的基础上对病历文本进行特征提取的应用研究，总结如下：

1）提出改进的BM25F-HDP模型。本文分析了HDP模型存在的不足，运用特征加权的思想，降低了高频词对模型的影响，并在病历文本数据集上验证了该模型的性能优于原始HDP模型和LDA模型。

2）研究了主题标签提取问题，原始的利用语义相似度进行标签提取会导致标签无法仅仅代表某一个主题，这是因为这个标签可能和很多主题有较高的语义相似度，所以在这里又考虑了主题区分度，使得提取的标签尽量只和某一个主题保持较高的语义相似度，这样的标签能够更好的代表当前主题。还研究了病历文本的检索功能，具体来说就是相关检查项和相关治疗项的提取。本文设计了基于主题模型的词语文档相关度方法，在病历文本上取得了很好的性能。基于本文提出的BM25F-HDP模型的相关项生成算法，在检查项和治疗项上的平衡F值相对于原始的HDP模型都有所提升。

3）利用主题模型的研究成果，建立了一个初步应用的病历服务系统，该系统实现了病历文本主题标签提取，病历总结提取，相似病历匹配，相关检查项和治疗项的提取等功能。

## 展望

本文提出的基于主题模型的病历文本特征提取，应用在了病历服务系统中，取得了较好的效果，但是仍然存在一些不完善的地方，还需要进行进一步的研究，主要包含以下几个方面：

1）除了使用特征加权的思想对HDP模型改进外，是否还有别的思路对模型进行改进。

2）在语义标签的提取中，只是运用了相对熵的概念进行语义相似度计算和主题区分度计算，是否还有别的方法能够计算主题和标签的覆盖度，已能够达到最准确的主题标签提取。

3）本文并没有使用Spark这一类能够进行迭代任务计算的分布式计算框架，没有很好地实现线性扩展。

4）本文提出的病历服务系统没有自学习的反馈模块，如果引入这个模块，就能够使医疗人员以最小的代价对本系统的结果进行评价，这能保证系统的性能能够得到不断的提高。

# 致 谢

时光荏苒、岁月如梭，不知不觉三年的研究生生活就快落下帷幕了。回顾三年的求学生涯，我收获了很多专业知识和技能，学到了踏实做事，虚心求学的人生态度，各方面都有了长足的进步。借此论文完成之际，特别向那些帮助过我的人表示我最真诚的感谢。

首先，我要感谢我的研究生导师饶元老师，还记得研究生刚开学时，饶老师让我们这一届参加实验室的项目验收报告会，让我第一次感受到了实验室这个大家庭的温暖。之后饶老师给我们指定了详细的学习计划，让我的学习生活丰富而又充实，学到了很多知识。饶老师的一万小时定律一直激励着我要付出持续不断的努力。本文从选题到最后的论文审阅都离不开饶老师的指导和帮助。饶老师不仅在学习上给予了我很大的指导和帮助，在人生态度上也给予了我积极的影响。能够师从饶老师，我感到非常的荣幸，在此谨向饶老师表达最诚挚的敬意和感谢！

其次，我要感谢软件学院的所有老师，你们为了我们的学业倾注了大量的心血，传授了很多专业知识。

还要感谢社会智能与复杂数据处理实验室的同门师兄师姐、同届的同学们，他们给予我很多帮助和指导，让我不仅收获了知识，也收获了友情，我很感激。

特别感谢我敬爱的父母，感谢父母二十多年的养育之恩，感谢父母一路走来对我的教导和支持，我将为家庭的美好未来而奋斗。

最后，由衷的感谢所有在百忙之中审阅论文和参加答辩的各位专家、教授！感谢你们给本论文提出的宝贵意见！

# 参考文献

1. Kamna Solanki, Sudhir,Parul Berwal, Sandeep Dalal. Analysis of Application of Data Mining Techniques in Healthcare [J]. International Journal of Computer Application,2016,148(2):16-21.
2. 周小甲，周庆利．中文病历文本中时间信息自动标注[C]．浙江生物医学工程学会年会， 2011:434-439.
3. Arnold C W, El-Saden S M, Bui A A, et al. Clinical Case-based Retrieval Using Latent Topic Analysis[J]. Amia Annu Symp Proc, 2010, 2010:26-30.
4. Wang X, Chused A, Elhadad N, et al. Automated knowledge acquisition from clinical narrative reports.[J]. AMIA. Annual Symposium proceedings / AMIA Symposium. AMIA Symposium, 2008, 6:783-787.
5. 盖杰，王怡，武港山．潜在语义分析理论及其应用[J]．计算机应用研究，2004，21(3):9-12.
6. Bullinaria J A, Levy J P. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD.[J]. 2012, 44(3):890-907.
7. Guo Q L, Yan-Mei L I, Tang Q. Similarity computing of documents based on VSM[J]. Application Research of Computers, 2008, 5186(11):585-586.
8. 张玉芳, 朱俊, 熊忠阳．改进的概率潜在语义分析下的文本聚类算法[J]．计算机应用，2011，31(3):674-676.
9. 刘霄．一种短文本主题建模方法[D]．浙江大学，2012.
10. Bai S, Huang C L, Ma B, et al. Semi-supervised learning of language model using unsupervised topic model[J]. 2010, 6493(10):5386-5389.
11. [陈涛](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e5%bc%a0%e7%8e%89%e8%8a%b3%22)，林杰 ，罗贤春．基于混乱度的自动初建语义词典评价方法研究[J]．情报学报，2010，29(2):240-245.
12. O’Hagan A, Murphy T B, Gormley I C. Computational aspects of fitting mixture models via the expectation-maximization algorithm[J]. Computational Statistics & Data Analysis, 2012, 56(12):3843-3864.
13. [张晨](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e5%bc%a0%e7%8e%89%e8%8a%b3%22)逸，[孙建伶](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e5%ad%99%e5%bb%ba%e4%bc%b6%22)，[丁轶群](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e4%b8%81%e8%bd%b6%e7%be%a4%22)．基于MB-LDA模型的微博主题挖掘[J]．计算机研究与发展，[2011,48(10)](http://c.wanfangdata.com.cn/periodical/jsjyjyfz/2011-10.aspx):1795-1802.
14. [王桢文](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e7%8e%8b%e6%a1%a2%e6%96%87%22)，[肖卫东](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e8%82%96%e5%8d%ab%e4%b8%9c%22)，[谭文堂](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e8%b0%ad%e6%96%87%e5%a0%82%22)．基于概率生成模型的网络数据分类方法[J]．计算机研究与发展，[2013， 50(12)](http://c.wanfangdata.com.cn/periodical/jsjyjyfz/2011-10.aspx):2642-2650.
15. Venugopal D, Gogate V. On lifting the Gibbs sampling algorithm[J]. Advances in Neural Information Processing Systems, 2012, 3:1655-1663.
16. [严宇宇](http://xueshu.baidu.com/s?wd=authoruri%3A%28b13061977dabaaf0%29%20author%3A%28%E4%B8%A5%E5%AE%87%E5%AE%87%29%20CADCG%E5%9B%BD%E5%AE%B6%E9%87%8D%E7%82%B9%E5%AE%9E%E9%AA%8C%E5%AE%A4&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson&sort=sc_cited)，[陶煜波](http://xueshu.baidu.com/s?wd=authoruri%3A%284fc926047bde3f62%29%20author%3A%28%E9%99%B6%E7%85%9C%E6%B3%A2%29%20%E6%B5%99%E6%B1%9F%E5%A4%A7%E5%AD%A6CADCG%E5%9B%BD%E5%AE%B6%E9%87%8D%E7%82%B9%E5%AE%9E%E9%AA%8C%E5%AE%A4&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson&sort=sc_cited)，[林海](http://xueshu.baidu.com/s?wd=authoruri%3A%2856137a7de035b649%29%20author%3A%28%E6%9E%97%E6%B5%B7%29%20%E6%B5%99%E6%B1%9F%E5%A4%A7%E5%AD%A6CADCG%E5%9B%BD%E5%AE%B6%E9%87%8D%E7%82%B9%E5%AE%9E%E9%AA%8C%E5%AE%A4&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight%3Dperson&sort=sc_cited) ．[基于层次狄利克雷过程的交互式主题建模](http://xueshu.baidu.com/s?wd=paperuri%3A%289b6f1d0a40542884674a1e19daa89f0b%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fd.wanfangdata.com.cn%2FPeriodical%2Frjxb201605006&ie=utf-8&sc_us=506195517969536604)[J]．软件学报，[2016，27(5)](http://c.wanfangdata.com.cn/periodical/jsjyjyfz/2011-10.aspx):1114-1126.
17. Du L, Buntine W, Jin H. A segmented topic model based on the two-parameter Poisson-Dirichlet process[J]. Machine Learning, 2010, 81(1):5-19.
18. Blei D M, Griffiths T L, Jordan M I. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies[J]. Journal of the Acm, 2007, 57(2):7-8.
19. 何岩．统计稀疏学习中的贝叶斯非参数建模方法及应用研究[D]．浙江大学，2012.
20. [高玮军](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e9%ab%98%e7%8e%ae%e5%86%9b%22)，[马栋林](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e9%a9%ac%e6%a0%8b%e6%9e%97%22)，[张其文](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e5%bc%a0%e5%85%b6%e6%96%87%22)．一种基于本体的文本主题提取方法研究[J]．计算机应用与软件，[2012，29(2)](http://c.wanfangdata.com.cn/periodical/jsjyjyfz/2011-10.aspx):54-56.
21. [张小平](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e5%bc%a0%e5%b0%8f%e5%b9%b3%22)，[周雪忠](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e5%91%a8%e9%9b%aa%e5%bf%a0%22)， [黄厚宽](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e9%bb%84%e5%8e%9a%e5%ae%bd%22)．基于词相似性与CRP的主题模型[J]．模式识别与人工智能，[2010，23(1)](http://c.wanfangdata.com.cn/periodical/jsjyjyfz/2011-10.aspx):72-76.
22. [王振振](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e7%8e%8b%e6%8c%af%e6%8c%af%22)，[何明](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e4%bd%95%e6%98%8e%22)，[杜永萍](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e6%9d%9c%e6%b0%b8%e8%90%8d%22)．基于LDA主题模型的文本相似度计算[J]．计算机科学，[2013，40(12)](http://c.wanfangdata.com.cn/periodical/jsjyjyfz/2011-10.aspx):229-232.
23. Crnic J. Introduction to Modern Information Retrieval[M]. McGraw-Hill, 2011.
24. Lord P W, Stevens R D, Brass A, et al. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.[M]. Oil shales of the world Pergamon Press, 2003:1275-83.
25. Liu H, Bao H, Xu D. Concept vector for semantic similarity and relatedness based on WordNet structure[J]. Journal of Systems & Software, 2012, 85(2):370-381.
26. Demner-Fushman D, Mork J G, Shooshan S E, et al. UMLS content views appropriate for NLP processing of the biomedical literature vs. clinical text[J]. Journal of Biomedical Informatics, 2010, 43(4):587-594.
27. 张小平，主题模型及其在中医临床诊疗中的应用研究[D]．北京交通大学，2011．
28. 李博，基于LDA和LSA的医学文本和影像分析模型及应用研究[D]．吉林大学，2012．
29. [黄承慧](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e9%bb%84%e6%89%bf%e6%85%a7%22)，[印鉴](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e5%8d%b0%e9%89%b4%22)，[侯昉](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e4%be%af%e6%98%89%22)．一种结合词项语义信息和TF-IDF方法的文本相似度量方法[J]．计算机学报，[2011，34(5)](http://c.wanfangdata.com.cn/periodical/jsjyjyfz/2011-10.aspx):856-864.
30. [孟纯军](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e5%ad%9f%e7%ba%af%e5%86%9b%22)，[李桃珍](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e6%9d%8e%e6%a1%83%e7%8f%8d%22)．矩阵方程AXB=C的最小二乘解的定秩研究[J]．湖南大学学报，2013，40(7):92-94.
31. [朱军](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e6%9c%b1%e5%86%9b%22)， [胡文波](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e8%83%a1%e6%96%87%e6%b3%a2%22)．贝叶斯机器学习前沿进展综述[J]．计算机研究与发展，2015，52(1):16-25.
32. Hoffman M D, Blei D M, Bach F R. Online Learning for Latent Dirichlet Allocation[C]// Advances in Neural Information Processing Systems 23:, Conference on Neural Information Processing Systems 2010. Proceedings of A Meeting Held 6-9 December 2010, Vancouver, British Columbia, Canada. DBLP, 2010:856-864.
33. [杨颖涛](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e6%9d%a8%e9%a2%96%e6%b6%9b%22)，[王跃钢](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e7%8e%8b%e8%b7%83%e9%92%a2%22)，[邓卫强](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e9%82%93%e5%8d%ab%e5%bc%ba%22)，[徐洪涛](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e5%be%90%e6%b4%aa%e6%b6%9b%22)．基于共轭先验分布的贝叶斯网络分类模型[J]．控制与决策，2012，27(9):1393-1396.
34. [徐谦](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e5%be%90%e8%b0%a6%22)，[周俊生](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e5%91%a8%e4%bf%8a%e7%94%9f%22)，[陈家骏](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e9%99%88%e5%ae%b6%e9%aa%8f%22)．Dirichlet过程及其在自然语言处理中的应用[J]．中文信息学报，[2009，23(5)](http://c.wanfangdata.com.cn/periodical/zwxxxb/2009-5.aspx):25-32.
35. [罗林](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e7%bd%97%e6%9e%97%22)，[苏宏业](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e8%8b%8f%e5%ae%8f%e4%b8%9a%22)，[班岚](http://s.wanfangdata.com.cn/Paper.aspx?q=%e4%bd%9c%e8%80%85%3a%22%e7%8f%ad%e5%b2%9a%22) ．Dirichlet过程混合模型在非线性过程监控中的应用[J]．浙江大学学报，[2015， 49(11)](http://c.wanfangdata.com.cn/periodical/zjdxxb-gx/2015-11.aspx): 2230-2236.
36. Hankin R K S. The Complex Multivariate Gaussian Distribution[J]. R Journal, 2014, 7(1):73-80.
37. 张文进．文本信息检索中的概率模型[J]．情报杂志，[2005，24(3)](http://c.wanfangdata.com.cn/periodical/zjdxxb-gx/2015-11.aspx):107-110.
38. 汪海波．[扩展DPMM模型在短文本主题识别中的应用](http://xueshu.baidu.com/s?wd=paperuri%3A%28473b06372564296d724845501207a4b5%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fd.wanfangdata.com.cn%2FPeriodical%2Fjsjyyyrj201408048&ie=utf-8&sc_us=12665347136901521119)[J]．计算机应用与软件，2014，31(8):191-195.
39. 严宇宇，陶煜波，林海．基于层次狄利克雷过程的交互式主题建模[J]．计算机应用与软件，2016， 27(5):1114-1125.
40. 李国垒，陈先来，夏冬．中文病历文本分词方法研究[J]．中国生物医学工程学报，2016, 35(4):477-481.
41. 张小平，周雪忠，黄厚宽．一种改进的LDA主题模型[J]．北京交通大学学报，2010，34(2):111-114.
42. 刘少鹏，印鉴，欧阳佳．基于MB- HDP模型的微博主题挖掘[J]．计算机学报，2015，38(7):1408-1419.
43. 周小甲，中文病历文本的时间信息提取研究[D]．浙江大学，2011．

学位论文独创性声明（1）

本人声明：所呈交的学位论文系在导师指导下本人独立完成的研究成果。文中依法引用他人的成果，均已做出明确标注或得到许可。论文内容未包含法律意义上已属于他人的任何形式的研究成果，也不包含本人已用于其他学位申请的论文或成果。

本人如违反上述声明，愿意承担以下责任和后果：

1．交回学校授予的学位证书；

2．学校可在相关媒体上对作者本人的行为进行通报；

3．本人按照学校规定的方式，对因不当取得学位给学校造成的名誉损害，进行公开道歉。

4．本人负责因论文成果不实产生的法律纠纷。

论文作者（签名）： 日期： 年 月 日

学位论文独创性声明（2）

本人声明：研究生 所提交的本篇学位论文已经本人审阅，确系在本人指导下由该生独立完成的研究成果。

本人如违反上述声明，愿意承担以下责任和后果：

1．学校可在相关媒体上对本人的失察行为进行通报；

2．本人按照学校规定的方式，对因失察给学校造成的名誉损害，进行公开道歉。

3．本人接受学校按照有关规定做出的任何处理。

指导教师（签名）： 日期： 年 月 日

学位论文知识产权权属声明

我们声明，我们提交的学位论文及相关的职务作品，知识产权归属学校。学校享有以任何方式发表、复制、公开阅览、借阅以及申请专利等权利。学位论文作者离校后，或学位论文导师因故离校后，发表或使用学位论文或与该论文直接相关的学术论文或成果时，署名单位仍然为西安交通大学。

论文作者（签名）： 日期： 年 月 日

指导教师（签名）： 日期： 年 月 日

(本声明的版权归西安交通大学所有，未经许可，任何单位及任何个人不得擅自使用)