面向电子病历中文医学信息的可视组织方法①

徐天明[1,2], 樊银亭[3], 马翠霞[1], 滕东兴[1]

[1](中国科学院软件研究所 人机交互技术与智能信息处理实验室, 北京 100190)
[2](中国科学院大学, 北京 100190)
[3](中原工学院 计算机学院, 郑州 450007)

摘 要: 针对当前大量电子病历信息无法充分利用的问题, 研究了面向电子病历中文医学信息的主题建模及可视组织方法. 首先基于电子病历数据和医疗问答数据, 进行预处理并转换为纯文本语料, 然后采用基于 Mallet 的 LDA 主题模型训练算法进行主题建模, 并结合主题模型分析的需求进行可视组织与呈现, 最后构建了面向中文医学信息的可视分析系统. 实例验证表明该系统可以有效的辅助用户进行主题模型的构建与分析, 并有利于进一步的诊断.

关键词: 电子病历; 可视分析; 主题模型; 信息组织; 人机交互

Visual Organization Method for Chinese Medical Information

XU Tian-Ming[1,2], FAN Yin-Ting[3], MA Cui-Xia[1], TENG Dong-Xing[1]

[1](Intelligence Engineering Lab, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100190, China)
[3](School of Computer Science, Zhongyuan University of Technology, Zhengzhou 450007, China)

Abstract: To make the best of the Chinese medical information in electronic medical records, a visual organization method is proposed. Firstly, a medical information dataset based on electronic medical records and medical community web pages is constructed, which is preprocessed into text corpus. Secondly, a topic model using Mallet is trained and visualized the output of topic model. Finally, a visual analysis system for Chinese medical information is also built. Experiments showed that the system could effectively help the analyzers train topic models and diagnose.

Key words: electronic medical records; visual analysis; topic model; information organization; human-computer interaction

随着医疗信息化进程的加速和各类医疗信息系统的不断完善, 中文医学信息正以几何数字不断激增. 而医学信息资源的极大丰富, 正在对人们快速准确获取所需信息的需求造成挑战. 一方面由于缺乏方便获取信息的途径, 大量的医学信息资源被闲置; 另一方面, 医护人员也无法在大量的信息中, 快速准确地获取自己所需的信息. 目前海量的中文医学信息存在于各类医学数据库、医学信息组织系统、医学文献和互联网中, 而电子病历作为病人在医院诊断治疗全过程的原始记录, 是病人健康医疗信息的最主要载体, 包含大量有价值的中文医学信息资源. 相比于其他形式的中文医学信息, 病历医学信息的表示具有更多的语义特性, 主要体现在医学信息的专业表示和实际医疗过程的忠实表示[1].

目前, 电子病历系统的发展正处于初级阶段, 系统功能有限, 只是传统纸张病历的简单电子化, 此类典型的电子病历系统提供 MS Word 或者其他类似的文本编辑器由医生输入内容, 医生使用系统水平的参差不齐使得电子病历的质量不能得到保证[2]; 现有电子病历结构化研究仅仅是已有信息的文本形式的简单的存储和组织, 缺乏有序化、优质化的组织方式, 不能体现病历中的语义关系, 相对应的结构化电子病历系统中完全结构化的病历不但给输入造成了不便还有可能造成临床数据的曲解[3]; 病历结构化录入很不完善, 病

① 基金项目:国家高技术研究发展计划(863)(2012AA02A608);国家自然科学基金(61173057,61173058,61232013,U1304611)
收稿时间:2015-03-05;收到修改稿时间:2015-04-26

历的主体部分多以非结构化的文本形式存储, 很多研究只能基于有限的结构化数据进行[4,5], 无法满足临床信息获取和分析以及海量病历数据检索和数据挖掘的需求.

为此, 本文研究一种面向电子病历中文医学信息的可视组织方法, 在第 2 节中以病历中的非结构化文本信息为输入, 进行预处理和主题建模, 并在第 3 节中以可视化形态加以呈现. 第 4、5 节实现了基于语义的中文病历信息可视分析系统, 实现了自动化的病历文本信息预处理、主题模型训练, 并提供交互操作辅助分析人员进行模型训练、结果分析, 第 6 节验证系统有利于医务工作者对新的病例进行诊断.

1 相关工作

1.1 电子病历医学信息组织

目前的医学信息组织主要依赖各种医学信息组织系统, 例如: 临床医学知识库(CMKB)、中文医学主题词表(CMeSH)和中文一体化医学语言系统(CUMLS)等. 这些信息组织系统致力于对医学术语、知识、概念的标准化, 方便临床科研工作人员、医生迅速快捷地获取疾病诊断、治疗、用药等方面的系统的、权威的临床医学知识. 这些标准的术语概念, 可以成为病历检索和组织的依据, 但是仅仅依赖文中有没有出现给定的术语关键字进行检索并不能取得良好的效果, 缺乏上下文环境的利用.

现有对电子病历信息组织的研究, 大多集中在面向电子病历的检索引擎的研究. 王晓和胡恒文分别基于 Lucene 和 CLucene 研究了电子病历的全文检索引擎, 并且分别在查准率和查全率上取得了非常好的效果, 但是 Lucene 依然是根据关键词进行索引和检索, 分词的效果以及术语在文中的出现与否对检索的效果有重要的影响[6,7]. 为了解决检索关键字的语义问题, 赵洋提出了基于本体的电子病历检索系统, 利用本体库对用户查询进行语义扩展从而优化检索的查全率, 但是这种方法极大依赖于本体库的构建[8].

总之, 现有的电子病历医学信息组织方式, 包括各类医学信息组织系统和病历检索引擎, 并不能很好地兼顾语义和上下文信息, 限制了信息获取的效果, 不利于临床科研工作人员和医生迅速便捷地获取所需的信息, 进行相关分析和决策.

1.2 基于病例数据的数据挖掘

随着临床医学术语的丰富和数据挖掘技术的发展, 越来越多的研究人员将数据挖掘的方法应用于临床研究, 以挖掘出有价值的知识和规则. 刘立刚提出了基于兴趣度的 Apriori 算法在电子病历数据中提取有诊断价值的关联规则并提高医生的诊断效率[9]. 王欣萍将 BP 神经网络算法应用于确定子宫肌瘤的危险因素以及出生缺陷率的预估, 表现出良好的准确性[10].

但是这些研究中, 大都只利用了病历中的少数结构化数据, 并没有充分利用电子病历中占主体的非结构化的文本数据, 因而取得的效果极其有限.

1.3 LDA 主题模型

LDA[11]是由 David Blei 提出的三层贝叶斯模型, 其基本思想是: 文档由隐含的主题生成, 主题的概率分布符合 Dirichlet 分布. 主题用关键词的分布表示, 文档集由同一组主题生成, 而一个文档可以由不同的主题同时生成. LDA 能够表示丰富的语义信息, 但是 LDA 模型需要人为指定主题的个数.

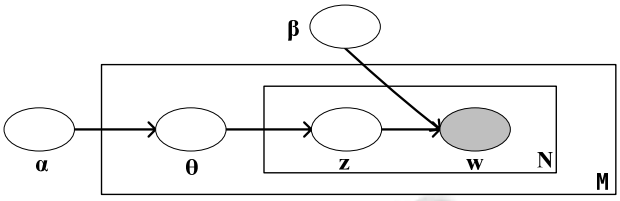


图 1 LDA 图概率模型图

在上述图概率模型图中, α 和 β 是语料级别的参数, 对一份语料中的所有文档都一样. θ 是文档级别的参数, 每个文档对应一个 θ, z 和 w 都是单词级别的参数. 其中 α 是 Dirichlet 分布的参数, 决定了 θ 的分布, θ_d 是文档 d 下的主题概率分布, β 为 Dirichlet 分布的参数. Z_n 为第 n 个词项文档产生的主题, W_n 为主题产生的词项, 由 β 和 z 共同决定. 生成 W_n 的过程产生了各个文档, 以上生成过程, 可以对应到下列公式:

$$p(\theta,z,w\mid\alpha,\beta)=p(\theta\mid\alpha)\prod_{n=1}^{N}p(z_n\mid\theta)p(w_n\mid z_n,\beta)$$

其中, N 为文档中词的个数, n 为词编号. Z_n 为第 n 个词项文档产生的主题, W_n 为主题产生的词项, 由 β 和 z 共同决定. 在迭代求解模型时, w 是观察所得, θ 和 z 是隐藏变量, 使用 EM 算法求解, E 步输入 α 和 β 计算似然函数, M 步最大化这个似然函数算出 α 和 β, 不断迭代直至收敛.

2 基于语义的中文医学信息主题建模

电子病历中信息的主要形式是自然语言, 包含着临床医生书写的住院志、病程记录、会诊记录、手术记录以及各种科室发出的放射、超声、内镜、心电等病历检查报告[12]. 这些非结构化的文本, 组成了绝大多数的病历内容. 由于电子病历结构化等基础技术的限制, 蕴含在这些非结构化文本中的信息并没有得到充分、有效的利用. 本节利用 LDA 基于语义对电子病历中的中文医学信息进行主题建模, 进而进行术语相关性分析和病历相似度分析, 实现了概念术语的语义关联和电子病历的语义相似度度量, 从而可对病历中的非结构化文本加以挖掘分析, 提取有价值的信息.

2.1 电子病历中文医学语言数据集的构建和预处理

2.1.1 电子病历中文医学语言数据集采集

由于电子病历数据涉及隐私, 科研用的脱敏数据集有限, 所以本文在进行实验时, 只使用了少量真实电子病历数据. 其他的数据来自于各大医疗问答论坛, 在医疗论坛问答中, 包含了病人对自己症状的描述以及医生的初步诊断结果、用药建议以及注意事项, 与病历中的文本有很大的相似性. 实验所用的数据集合计有一万左右的医疗文本数据.

在数据采集的过程中, 使用基于 Python 的 BeautifulSoup 编写爬虫程序, 获取医疗问答数据. 按照文本长度去除过短的低质量内容之后, 存储成纯文本格式. 结合已有的电子病历文档, 组成了训练主题模型所需的语料.

2.1.2 数据集分词去除标点和停用词

对收集到的中文医学语言数据集进行进一步预处理, 利用开源搜索引擎框架 Lucene 提供的 IKAnalyzer 进行分词和去除停用词的处理.

经过分词器的处理, 病历文本中无意义的停用词(如“的”, “了”)以及标点符号被去除, 文本成为空格分隔的关键字的集合, 以此作为 LDA 训练器的输入.

2.1.3 LDA 主题模型的训练

经过预处理的语料, 作为 LDA 训练算法的输入, 可以进行 LDA 模型的训练. LDA 模型训练使用了基于 Java 的开源工具包 Mallet. 主题模型的训练需要人为指定主题的个数, 主题的个数的选择需要基于分析人员对于训练语料的先验知识.

基于 Mallet 的 LDA 主题模型训练, 会生成两个模型文件. 一个包含了所有的主题的关键词表, 另一个包含了所有文档的主题概率分布. 更具体来讲, 主题的关键词表, 一方面代表了该主题语义上的含义, 另一方面是对关键词基于潜语义空间的聚类, 可以认为同一个主题下的所有的词, 都具有语义关联性; 文档的主题概率分布, 则是在潜语义空间下, 文档的一种表示方式, 可以作为文档的词向量空间表示方式的语义替代.

2.2 医学概念术语的语义相关性分析

利用构建完成的中文医学语料集基于 Mallet 完成 LDA 的训练, 并获得了各个主题的关键词表, 如图 2 所示.

```
0   1   胎儿 正常 超 周 描述 孕 怀孕 健康 发育 定期 检查 囊
1   1   咳嗽 痰 止咳 液 孩子 健康 排 一些 喝 感冒 里 呼吸道
2   1   运动 体育锻炼 平时 营养 身体 加强 每天 增加 锻炼 健康
3   1   心电图 描述 检查 甲亢 医院 健康 说 问题 请问 请 心脏
4   1   食物 油腻 饮食 辛辣 避免 清淡 刺激性 要吃 多吃 要注意
5   1   皮肤 斑 方法 激光 胎记 美 脱毛 后 疤痕 描述 色素 脸上
6   1   因素 发 生长 植 健康 头发 咨询 遗传 脱发 现在 关系 病
7   1   检查 原因 建议 明确 诊断 可能是 一下 考虑 是否 确诊
8   1   怀孕 月经 避孕药 早孕 排卵期 月 指导 试纸 意见 同房
```

图 2 各主题关键词表

主题的关键词表代表了本主题的含义, 另一方面是关键词在潜语义空间上的聚类. 同一个主题下的关键词具有语义上的关联, 是广义上的同义词或者相似词. 由此, 可以解决给定关键词的语义问题, 通过主题关键词表将关键词扩展成为了一组语义相同或者相关的词. 假设在 LDA 主题模型的训练结果中有主题 $T_1=\{w_i, w_{i+1}...W...w_n\}$和主题 $T_2=\{w_j, w_{j+1}...W...w_m\}$, 其中均包含词 W 则可以将其扩展为由 w_p 组成的词袋, 其中 $p=i...n,j...m$, 从而扩展词 W 的语义, 达到利用词 W 的上下文语义对 W 的含义进行描述的目的. 进一步, 利用 T_1 和 T_2 的共现的关键词个数, 可以度量两个主题之间的相似度.

根据主题关键词表对关键词进行语义扩展, 可以更充分描述关键词的含义, 达到了“升维”的作用.

2.3 电子病历文档的语义相似度分析

基于 LDA 主题模型训练过程中产生的文档在各个主题上的概率分布, 如图 3 所示, 可以作为潜语义空间中文档的表示形式, 用来替代传统的词向量空间. 而基于文档在各个主题上概率分布产生的文档向量的相似度, 可以代表文档的潜语义相似度.

假设指定主题的个数为 N, LDA 主题模型训练完成之后会产生每个文档在 N 个主题上的概率分布, 由此可以利用 N 维的向量来表示每个文档, 每一维代表

一个主题, 各个维度的数值为此文档属于该主题的概率, 作为向量的权重. 这样可以将文档映射到 N 维的潜语义空间, 完成降维, 进而可以利用降维之后的结果进行文档的相似度分析, 具体可以采用各种向量的距离度量方法, 比如余弦相似度等. 经过降维之后文档的主题向量, 虽然损失了相当一部分的信息, 但是仍然可以保持潜语义层面上的相关性, 也就是说, 潜语义空间上相关的文档往往具有语义上潜在的关联关系, 例如“油腻”和“辛辣”是词义上毫无关系的两个词, 但是在潜语义层面上, 两词均常用来形容饮食习惯, 多见于病历文本的饮食注意事项中, 在语义和上下文关系上存在着比较高的相似性. 降维之后的主题信息也可以作为特征, 结合传统的词向量空间的特征, 对文档相似性度量的效果有显著提升.

```
#doc name topic proportion ...
0   segmented_data/57370000
    45  0.2242424242424242
    47  0.060606060606606061 39  0.060606060606606061
    19  0.048484848484848485 18 0.048484848484848485
    44  0.03636363636363636 16  0.03636363636363636
    9   0.03636363636363636 43  0.024242424242424242
    40  0.024242424242424242    31  0.024242424242424242
    38  0.018181818181818 36  0.0181818181818181818
    41  0.012121212121212121    37  0.012121212121212121
    34  0.012121212121212121    33  0.012121212121212121
    49  0.006060606060606061    48  0.006060606060606061
    46  0.006060606060606061    42  0.006060606060606061
```

图 3 文档在各个主题上的概率分布

3 面向中文医学信息主题模型的可视形态

LDA 主题模型的训练以及基于主题模型的术语关联关系分析和病历语义相似度分析是一个循环往复的渐进的探索过程. 分析人员需要根据训练的效果, 不断调整主题的个数、重新清理数据以达到最优的训练效果, 获得最有效的数据分析结果. 因此, 为了辅助分析人员进行主题模型训练并进行术语关联关系分析和病历语义相似度分析的过程, 本文针对 2.1 节、2.2 节和 2.3 节的 LDA 主题模型训练和分析过程构建了自然的、可以帮助用户快速对训练分析效果做出判断的、降低分析人员交互负担的可视化形态, 用于辅助分析人员 LDA 主题模型的训练和分析过程, 进而有助于医务工作者进行病例诊断.

3.1 面向中文医学信息主题模型的可视形态构建方法

Card 认为, 信息可视化是从数据到可视化形式再到人的感知系统的可调节的映射过程[13]. 在 Card 提出的可视模型中, 首先利用数据变化把原始数据预处理为结构化数据, 然后利用可视化映射将结构化数据表现为可视化形态, 从而将原始数据传递给人的感知系统. 本文结合 Card 的可视模型特点, 针对中文医学信息主题模型训练结果构建了术语语义关联关系和病历语义相似度的可视形态模型.

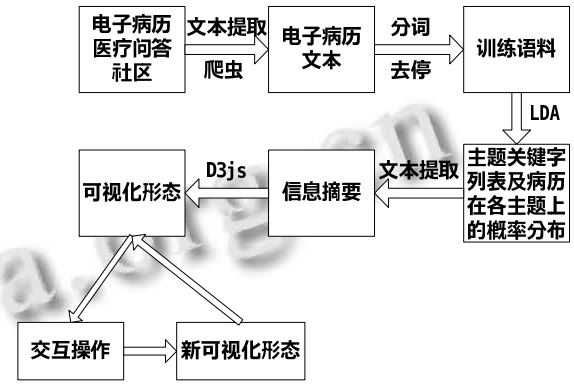


图 4 主题模型可视形态交互与数据迁移图

如图 4 所示, 先利用相应的数据转换算法, 将主题模型训练结果转换为可视形态可以映射的可视结构, 映射到用户的可视视图上; 用户再根据自己的领域知识和先验知识对可视视图进行交互, 如拖拽、圈选、钻取等; 可视视图再根据用户的交互动作, 将对数据的修改映射到相应的可视结构上, 可以是结果的调整修改、筛选、排序, 甚至是修改模型训练的参数重新训练主题模型.

3.2 面向中文医学信息主题模型的可视分析形态集

可视分析系统基于 LDA 主题模型训练结果, 提供可视形态呈现数据并提供交互操作帮助用户调整分析数据. 本文结合基于语义的医学术语关联关系以及病历的语义相关性分析需求, 针对在数据分析的各个阶段中遇到的问题构建了一系列可视形态集合, 如图 5 所示.

3.2.1 医学术语语义关系可视形态

医学术语语义关系可视形态用于对语料中包含的医学术语的关联关系进行可视化, 可以直观地将术语关键字之间以及术语集合之间语义关联关系表示出来. 如图 6(a)所示, 同一个主题的所有术语关键字形成簇, 代表术语集合, 被选中的术语位于簇的中心; 同时包含被选中术语的术语集合簇被链接在一起. 术语集合之间的距离远近, 代表了两个集合术语关键字重叠的个数, 表现出两个术语集合的相似度.

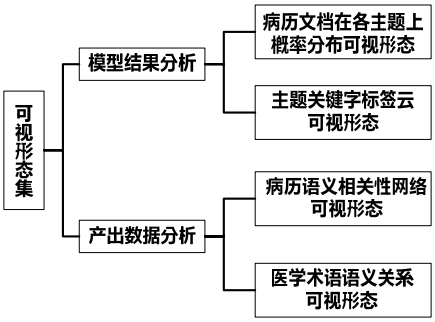
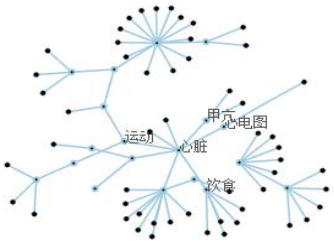


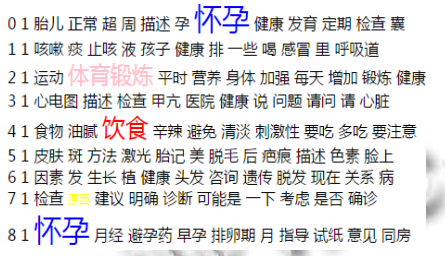
图 5　可视形态分类图

3.2.2 主题关键字标签云可视形态

主题关键字标签云可视形态利用标签云的方式将每个主题中的所有关键词显示出来. 如图 6(b)所示, 每一行代表一个主题, 同样的关键字用同样的颜色和字体标识, 可以快速探查某个特定关键字出现在哪些主题中. 对于选中的关键字用字体的大小代表关键字的权重, 字体越大, 对主题的区分度越强.

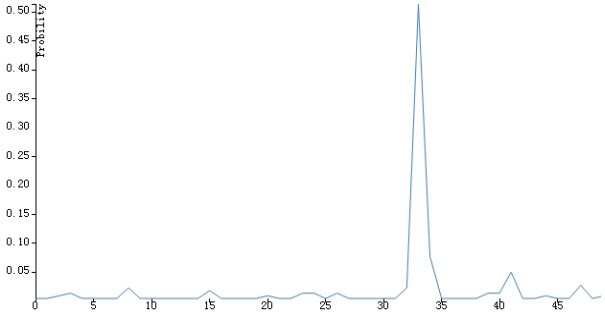


(a) 医学术语语义关系可视形态



(b) 主题关键字标签云可视形态

(c) 病历语义相关性网络可视形态



(d) 病历文档在各主题上概率分布可视形态

图 6　面向中文医学信息主题模型的可视分析形态集

3.2.3 病历语义相关性网络可视形态

病历语义相关性网络可视形态描述病历文档之间的语义相似度. 如图 6(c)所示, 每一个结点代表一个病历文档, 结点之间的距离表示文档之间的相似度. 距离越近相似度越高, 距离越远相似度越低, 相似的病历可以为病患的诊断提供参考. 为方便探查, 相似度网络会以指定病历为中心进行布局.

3.2.4 病历文档在各主题上概率分布可视形态

病历文档在各个主题上的概率分布可视形态, 用折线图的方式展示出病历在各个主题上的概率分布, 如图 6(d)所示. 可以直观地掌握每个病历由哪几个主题生成, 根据大部分病历文档的主题组成可以直观的判断主题模型训练的效果. 若大部分病历在概率分布图上均呈现出“单峰”或“双峰”则模型训练的结果较好; 若大部分病历在各个主题上均没有很明显的倾向性, 则主题模型训练得出的信息量非常少, 不足以从语义上区分各个病历文档.

4　面向中文医学信息的可视分析架构

本文构建了面向中文医学信息的可视分析系统, 系统体系架构如图 7 所示. 系统架构分为 6 层.

① 多源异构数据源层: 主要用于描述在各个电子病历系统以及各种电子病历数据源的多源异构的数据源. 由于缺乏对现有电子病历标准的贯彻实施, 大部分数据源均有各自的存储方式和格式, 如数据库、Word 文档、Web 页面、纯文本文档等, 因此首先要为各种数据源编写对应的接口, 将所需的中文医学信息文本统一转化为纯文本文档.

② 数据预处理层: 主要用于对纯文本文档进行分词、去停用词等预处理操作, 经过预处理模块, 纯文

本文档被处理成为可以作为数据挖掘算法输入的语料. 本文集成了基于Lucene的IKAnalyzer作为分词器并用医学语言词典作为分词字典, 使用了搜狗在线语料库中的停用词表. 此外, 由于很多病历语料内容过短, 不能有效提供信息, 还编写了低质量数据清洗程序, 用以过滤低质数据.

③ 数据挖掘算法层: 主要用于将预处理完成的语料作为输入, 进行模型训练, 输出模型文件. 本文集成了开源的基于 Java 的 Mallet 工具作为 LDA 模型的训练算法, 为了方便进一步的文本分析挖掘工作, 还编写了若干工具算法模块, 例如 TFIDF 计算算法、各种相似度计算算法等. 此外还设计了系统接口用于接入更多的数据挖掘算法, 以便进一步丰富系统的功能.

④ 可视化结构数据提取层: 结合可视化需求, 以模型文件为依据进行进一步的计算, 生成可以直接映射为可视化形态的可视化结构数据. 本文针对面向中文医学信息主题模型中的常见问题, 为各种可视分析需求编写了相应的算法库, 提取相应的可视化结构数据, 用于支持中文医学信息的可视分析过程.

⑤ 可视形态层: 主要由可视形态集合交互任务集两个模块组成. 本文从易于用户理解和认知的角度构建了一套面向中文医学信息主题模型的可视形态集, 并提供了易于交互使用的交互任务辅助用户的数据质量管理与分析.

⑥ 支撑数据接口层: 将可视分析产生的结果数据, 按照指定的格式进行导出. 用于进一步的数据分析、系统构建等.

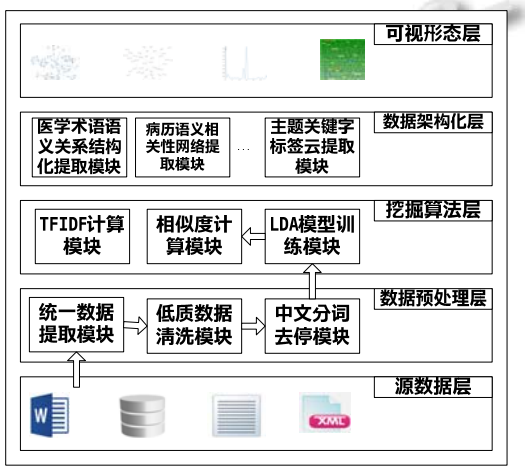


图 7 面向中文医学信息的可视分析架构

5 可视化结构数据提取算法

5.1 主题关键字标签云布局算法

当分析人员针对 LDA 主题模型评估各个主题聚类的效果时, 往往会先通过每个主题包含的主题关键字推测该主题的大致含义. 为了帮助分析人员快速把握每个主题的含义, 本文利用标签云的方式展现每个主题的关键字. 为了保持自上而下的阅读习惯, 主题的关键字依然按行列出, 为了区分同一个关键字在不同主题中的不同语义, 同一个关键字使用了相同颜色和大小. 此外, 同一个主题不同的关键字对于主题含义的区分度是不同的, 而现有的主题模型中, 并没有对这些关键字的重要性作区分, 本文拟用词频率逆文档频率作为衡量关键字重要性的权重, 关键字在标签云中的显示大小与权值大小成正比. 其计算公式如下:

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \quad idf_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|}$$

$$tfidf_{i,j}=tf_{i,j}\times idf_i$$

其中, $tf_{i,j}$的计算公式中, $n_{i,j}$为该词在文档d_j中出现的次数总和, 而分母为文档 d_j 中所有词出现的次数总和. 在 idf_i的计算公式中, $|D|$为语料库中文档总数, 分子为包含词 t_i的文档总数.

5.2 病历文档语义相关性布局算法

病历文档基于语义的相关性度量, 是面向中文医学信息主题模型的重要输出数据. 文档相关性是文档信息检索的基本问题, 因此基于语义的文档相似度度量算法对于基于语义的病历文档检索有重要意义.

对于面向中文医学信息的主题模型, 假设指定的主题个数为 N, 则在主题模型输出的文档关于各个主题的概率分布中, 每个文档可以表示为 $R=\{S_1,S_2,S_3,...S_i\}$, 其中 $i\in\{1,2...N\}$, Si 为文档属于主题 i 的概率. 为了计算文档的相似性, 可以用向量之间的夹角余弦值来表示:

$$Similar(R_i,R_j)=\cos(R_i,R_j)=\frac{R_i\bullet R_j}{|R_i|\times|R_j|}$$

为了方便用户对相关性网络进行可交互的探查, 采用了有限制条件的 Verlet 算法[14]进行节点布局. Verlet 算法是一种用于求解牛顿运动方程的数值方法, 被广泛应用于分子动力学模拟以及视频游戏中. 在交互操作中, 当用户改变某个节点的位置, 需要对每个节点的位置用以下公式进行重新计算, 从而完成整个

网络的重新布局:

$$r_p = \frac{w_q(d-|\mathbf{p}-\mathbf{q}|\mathbf{pq})}{(w_p+w_q)|\mathbf{p}-\mathbf{q}|}$$

其中, r_p 即为 p 点的位移; q 点为固定点的位置, w 为 p、q 两点的权重. 由于 q 为固定点, $w_q \gg w_p$.

5.3 医学术语关键字集合相似性度量算法

医学主题由医学术语关键字集合组成, 医学术语关键字集合的相似性表现了各个医学主题之间的相似性. 在主题模型的训练中, 主题个数设置过大往往导致产生相似性比较高的主题, 对主题相似性进行呈现有助于分析人员优化模型训练参数, 提升模型训练效果. 医学术语关键字集合相似性计算公式如下:

$$Similar(T_i,T_j) = \frac{2\times N(w)}{N(w_i)+N(w_j)}$$

其中, T_i 和 T_j 代表两个术语关键字集合, $N(w)$代表两个术语集合中相同的关键字个数, 而 $N(w_i)$和 $N(w_j)$分别代表两个术语关键字集合中的关键字个数.

6 应用实例

基于以上研究, 我们开发了面向中文医学信息的可视分析系统, 如图 8, 针对已构建的中文医学语言数据集进行了 LDA 主题模型的训练和可视分析, 进行了效果验证.

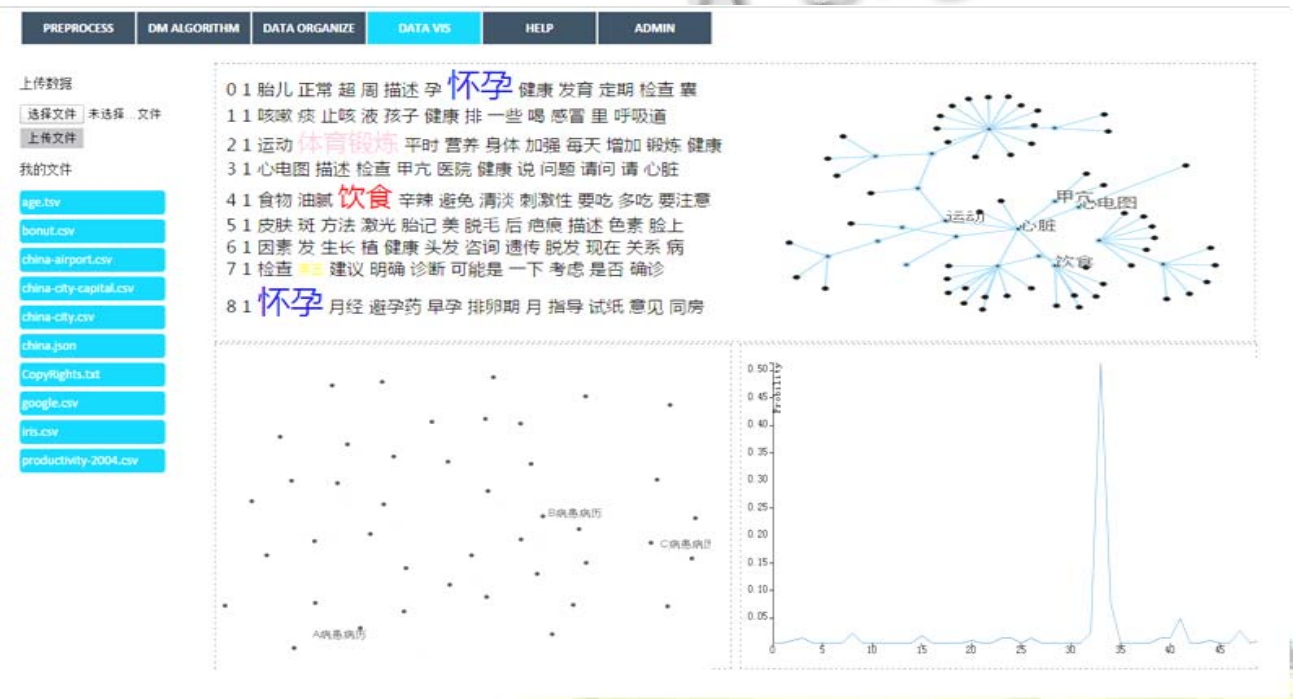


图 8　面向中文医学信息的可视分析系统界面图

可视分析系统根据指定的源数据集所在的目录, 根据数据类型调用对应的数据源处理接口, 将数据处理为纯文本格式, 再通过预处理模块进行分词、去停用词等预处理, 形成模型训练的语料. 语料经过 LDA 模型训练之后, 输出模型文件, 通过可视化结构数据构建算法计算出可视化结构数据, 映射到图 6 所示的各个可视形态中.

首先, 分析人员可以在主题关键字标签云中探查各个主题的含义以及每个关键字在不同主题中的语义, 如图 6(b)所示, 其中“怀孕”存在于两个主题中, 并且语义有所不同, 前者倾向于计划中的怀孕, 而后者倾向于避孕. 在大致了解每个主题的含义之后, 分析人员对本次主题模型的训练效果已经有了大致的认识, 并且可以对各个主题的含义进行标注. 然后分析人员可以利用图 6(d)中的可视形态, 探查各个文档在各个主题上的概率分布, 并利用先验知识验证合理性. 进行完如上的合理性验证之后, 基本可以确定主题模型的训练效果, 决定是否需要调整参数重新训练甚至重新清理数据集. 图 6(a)以及图 6(c)中的可视形态, 可以帮助分析人员直观判断可视分析系统的产出数据是否合理, 并有利于分析人员对新的病例进行诊断. 比如, 当分析人员拿到一份新的与“怀孕”有关的病历时, 可

以首先根据图 6(a)和图 6(b)中的主题关键词关系网络和标签云找到与该病历相关的主题关键词, 比如“早孕试纸”、“排卵期”等, 进而利用这些关键词对病历文档进行检索, 找到相关的历史病历及诊断方案进行参考. 进一步, 可以利用如图 6(c)的文档主题相似性对检索出的文档进行筛选, 得出高相关性的文档进行参考. 综上可知, 利用如上的可视形态, 分析人员可以在较短时间内, 判断主题模型的训练效果, 验证输出的合理性. 此外, 分析人员还可以利用交互任务集, 对输出结果进行微调, 修正明显的错误用例, 最终产出合理可以进一步使用的数据集.

7 结论与展望

本文构建了中文医学信息数据集, 基于 Mallet 进行了主题模型训练并针对在主题模型训练和分析中可能遇到的问题构建了可视化结构数据并映射为可视化形态, 同时, 为了辅助主题模型训练和分析, 减轻分析人员的认知负担, 设计了对应的交互任务集, 帮助分析人员对中文医学信息进行分析与管理. 最后, 开发了一个面向中文医学信息的可视分析系统, 并结合具体实例进行验证. 结果表明, 该系统能够有效地进行主题模型的训练与分析.

由于时间所限, 本系统还存在以下不足: 仅仅使用了一种主题模型, 从中文医学信息中获得的信息有限, 此外辅助主题模型训练和分析的可视形态还不够丰富, 在未来的工作中还有许多在电子病历中文医学信息的组织和分析中有实用价值的算法可以应用. 例如 K-Means、KNN 等基于距离的聚类算法可以基于给定的文档的向量表示方法, 对电子病历文档进行聚类, 帮助分析人员发现病历文档之间的关系和规律; Apriori 和 FP-Growth 算法可以用于发现电子病历文本中的频繁模式, 挖掘出具有实用价值的临床医学规律; 在给定电子病历医学分类的情况下, 还可以利用已标注的训练数据集基于各种文本分类算法(如 SVM 等)对病历进行分类, 方便电子病历文档的组织和管理.

参考文献

1 温有奎,焦玉英.基于语义三元组的电子病历潜在知识发现研究.情报学报,2011,30(7):675–681.

2 曹原,齐静.电子病历在医院实施 HIS 系统中的优点及不足.青海医药杂志,2006,35(10):33–33.

3 陈衡.结构化电子病历综述.湖南省图书情报学研究生论坛,2010.

4 丁卫平,施佺,管致锦.基于频繁概念格的电子病历关联规则挖掘研究.微电子学与计算机,2008,25(8):125–128.

5 曾勇.关联规则在脑科电子病历挖掘中的应用.医学信息学杂志,2014,35(10):55–58.

6 王晓,张健.基于 Lucene 检索引擎的电子病历全文检索系统.医疗卫生装备,2009,29(12):43–44.

7 胡恒文,高智勇,王辉.基于 Clucene 的电子病历全文检索系统研究与设计.计算机与数字工程,2014,42(3):521–525.

8 赵洋,李万龙,白杰英.基于本体的电子病历检索系统研究.计算机技术与发展,2010,20(3):211–213.

9 刘立刚,钟锐,杨娟.基于兴趣度的 Apriori 算法在电子病历数据分析中的应用.江西理工大学学报,2013,34(5):72–76.

10 王欣萍,孙昕,孙尧.基于 BP 人工神经网络模型构建电子病历系统的数据分析.中国组织工程研究,2011,15(35):6592–6595.

11 Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. the Journal of machine Learning research, 2003, 3: 993–1022.

12 李昊曼,段会龙,吕旭东.医学语言处理技术及应用.中国数字医学,2008,3(11):11–13.

13 Card SK, Mackinlay JD, Shneiderman B, eds. Readings in Information Visualization: Using Vision to Think. San Francisco: Morgan Kaufmann, 1999:15–20.

14 Dwyer T. Scalable, versatile and simple constrained graph layout. Computer Graphics Forum, 2009, 28(3): 991–998.