

## 面向临床数据中心的信息检索研究与应用

武学鸿<sup>1,2</sup>,朱建平<sup>2</sup>,李建华<sup>1,3</sup>

(1.中南大学计算机学院,湖南 长沙 410083;

2.湖南科医云健康科技有限公司,湖南 长沙 410012;

3.湖南科创信息技术股份有限公司,湖南 长沙 410012)

**摘要:**临床数据中心往往具有数据量大、关联关系复杂、数据增长快、主题多样性等特征,从中准确、高效的检索出相关信息是一项具有挑战的工作。本文提出了一种基于分布式搜索引擎的临床数据信息检索方法,以病案首页信息为核心,结合其关联的检验、文书、医嘱、费用、手术、诊断、检查信息构建了面向主题的父子关联索引模型,并提出了针对索引的优化方法,该方法能够提供有效的临床数据信息检索服务,可快速为临床医生、科研人员等提供准确的临床信息。

**关键词:**临床数据中心;大数据;搜索引擎;索引模型

**中图分类号:**R197

**文献标识码:**B

**DOI:**10.3969/j.issn.1006-1959.2022.02.003

**文章编号:**1006-1959(2022)02-0010-06

## Research and Application of Information Retrieval for Clinical Data Center

WU Xue-hong<sup>1,2</sup>,ZHU Jian-ping<sup>2</sup>,LI Jian-hua<sup>1,3</sup>

(1.School of Computer Science and Engineering,Central South University,Changsha 410083,Hunan,China;

2.Hunan Keyiyun Health Technology Co.,Ltd,Changsha 410012,Hunan,China;

3.Hunan Creator Information Technology Co.,Ltd,Changsha 410012,Hunan,China)

**Abstract:** Clinical data centers often have the characteristics of large amount of data, complex correlation, rapid data growth and subject diversity. It is a challenging work to retrieve relevant information accurately and efficiently. This paper proposes a clinical data information retrieval method based on distributed search engine, taking the home page information of medical records as the core, combined with its associated examinations, documents, advices, expenses, operation, diagnosis and inspection information constructs a topic-oriented parent-child index model, and puts forward an optimization method for the index. The results of clinical data analysis show that this method can provide effective clinical data information retrieval services, and can provide clinicians and researchers with a fast and accurate way to obtain information and tap value from it.

**Key words:** Clinical data center;Big data;Search engine;Index model

临床数据中心(clinical data centers,CDR)随着电子病历应用的不断丰富而持续发展<sup>[1-3]</sup>,其包含了患者所有重要的临床数据,可集成院内各科室级临床信息系统(医嘱、病历、检验、手术、心电、超声、病理等),实现所有临床诊疗数据的整合与集中展现,并为医疗诊断决策提供支持信息。临床数据中心具有数据量大、增长快、关联关系复杂、价值高等特点<sup>[4-6]</sup>。面对如此庞大的数据规模,传统的数据库在存储能力、检索效率,尤其是多表关联检索等方面,往往无法有效满足临床医生、科研人员等信息获取的需求<sup>[7-9]</sup>。本文提出了应用 Elasticsearch 分布式搜索引擎技术实现面向临床数据中心的信息

检索方法<sup>[10-13]</sup>,结合数据本身及搜索引擎技术特性<sup>[14]</sup>,制定相应的优化策略,并通过实际检索场景验证本方法的效果,现总结如下。

## 1 数据模型梳理及索引构建

1.1 数据模型梳理 临床数据主要是以患者为中心,本次围绕患者住院信息选择了具有代表性的八类数据来进行相关分析,八类数据信息分别是:病案首页、检验信息、病历文书、医嘱信息、费用信息、手术信息、诊断信息、检查信息,其描述见表1。将表1中八类数据以面向主题的方式进行整合,以病案首页为核心,其他数据与之形成关联,见图1。

表1 临床数据中心中八类核心信息

表名	表标识	表描述
病案首页	page	患者住院登记信息包括患者基本信息、主治医生、病房等信息
检验信息	lab	患者住院期间的各类检验信息
病历文书	docc	病历中描述的病历信息
医嘱信息	advice	医生在病历中给患者记录的医嘱信息
费用信息	coststat	患者住院期间的费用信息
手术信息	operation	患者住院期间的手术记录
诊断信息	diag	患者住院期间的诊断记录
检查信息	exam	患者住院期间的检查记录

基金项目:国际科技创新合作基地项目(编号:2019CB1007)

作者简介:武学鸿(1988.1-),男,江苏南京人,硕士,工程师,主要从事临床医学大数据、医学知识图谱及 AI 辅助决策应用研究

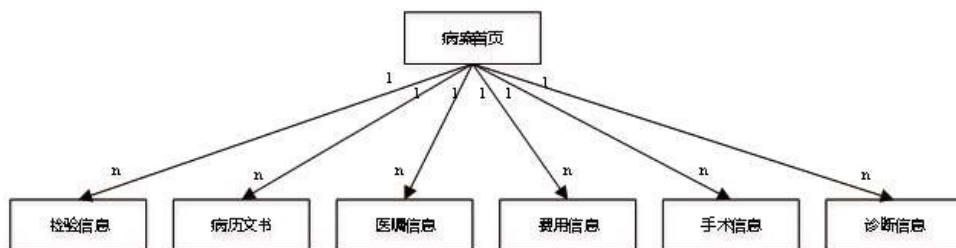


图 1 以病案首页为核心的关联关系模型

病案首页中包含了患者的基本信息,以病案首页信息为中心,其他数据表信息与其构成了父子关联模型,即病案首页信息为父表,检验信息、病历文书、医嘱信息等都为子表。通过该模型,在检索的业务需求中可以根据 1 个或者多个子表中的 1 个或者多个字段来查询病案首页信息或者根据病案首页信息来查询任意 1 个子表的信息。上述模型结构在面向极端场景时,即根据 7 个子表同时关联来查询病案首页信息,尤其是表的数据规模达到亿级别以上

时,传统关系型数据库往往难以支撑。Elasticsearch 不仅支持分布式索引数据存储还还原生的支持父子关联索引模型,同时在父子关联模型查询接口上提供了很好的支撑,可实现由父查子以及由子查父的关联检索场景<sup>[15-17]</sup>。

1.2 索引模型的构建 结合数据模型间的复杂关联关系<sup>[18,19]</sup>,基于 Elasticsearch 创建索引并配置各索引类型之间的关联关系映射,形成父子索引模型,索引映射文件配置部分信息见图 2。

```
{
  "es_pc_all": {
    "mappings": {
      "coststat": { "_parent": { "type": "page"}, "_routing": { "required": true}, "_source": { "enal
      "docc": {
        "_parent": {
          "type": "page"
        },
        "_routing": { "required": true},
        "_source": { "enabled": false},
        "properties": { "creator_id": { "type": "string", "index": "not_analyzed", "null_value": ""},
      },
      "page": {
        "properties": { "abo_blood_name": { "type": "string", "analyzer": "xy_ngram_analyzer"},
      },
      "diag": { "_parent": { "type": "page"}, "_routing": { "required": true}, "properties": { "diag_c
      "exam": { "_parent": { "type": "page"}, "_routing": { "required": true}, "_source": { "enable
      "advice": { "_parent": { "type": "page"}, "_routing": { "required": true}, "_source": { "enabl
      "operation": { "_parent": { "type": "page"}, "_routing": { "required": true}, "_source": { "en
      "lab": { "_parent": { "type": "page"}, "_routing": { "required": true}, "_source": { "enabled":
    }
  }
}
```

图 2 父子索引模型的配置信息

索引模型定义了 8 种类型分别对应的 8 个信息表,各类型包含了一系列属性对象的定义。其中子表与父表之间的关联关系是通过表中 `_parent` 属性定义实现,如:病历文书(docc)表指定其 `_parent` 属性值为病案首页(page),即父表是病案首页(page),子表是病历文书(docc);其他数据表,如检验信息、费用信息、诊断信息、医嘱信息等与病历文书的定义方式一致。定义完索引模型后,通过 Elasticsearch 所提供的索引创建 API 实现索引的建立。

1.3 数据索引及分词 Elasticsearch 提供了 bulk 接口支持数据索引,即将数据从关系型数据库或者其他数据源导入到索引库中,数据索引也有多种工具可选择,如 Elasticsearch river 插件、Logstash 工具等。这些工具都可以解决数据从关系型数据库索引

到 Elasticsearch 集群的过程,并支持增量索引。Elasticsearch 默认支持英文单词的分词方式,通过安装配置分词插件可实现中文分词,本文中采用的是 ik 分词器<sup>[13,20]</sup>,该分词器目前应用较广泛,无论是原生的分词效果还是其扩展性都能够满足业务检索的需求。

## 2 索引的优化方法

2.1 数据索引优化 数据索引即把数据导入到 Elasticsearch 的过程,如果数据体量较大,那么在不做优化方案的情况下往往会导致数据索引过慢,而且数据的索引过程并不是一次性的工作,当索引字段变更,索引映射文件变更的时候就需要将所有数据进行重新索引,每次索引过程都比较耗时。为此,本研究对索引过程进行了优化:①在数据索引阶段禁

用数据的副本:数据副本能够有效保障数据安全性,但是在数据索引过程中启用副本会消耗一定时间在数据的复制过程中,通过禁用副本可以提升数据索引的效率,当所有数据索引完成后即可打开副本;②设置数据提交刷新时间为手动刷新:数据索引过程会利用数据缓冲策略,数据缓冲默认实时刷新缓存到持久化层,通过禁用自动刷新,可以有效利用缓存策略,提升数据索引的吞吐量和效率;③设置增大数据索引提交批量:与上一点同原理,通过提升数据索引的提交量,可提升数据索引的吞吐量;④按需调整增加分片数:Elasticsearch 原生支持分布式能力,索引分片是其基础的分布式单元,通过增加分片数,可以提升其并发处理能力,从而提升数据索引的吞吐量;⑤增大 Elasticsearch 服务节点内存:与第③点同原理,启用缓存,增大提交量就会占用更多服务节点内存,通过增大内存保障吞吐量;⑥原始数据不存储:Elasticsearch 主要是实现倒排索引的构建与存储,其本身默认存储原始数据,但是原始数据过多会导致无论是内存、IO,还是在磁盘空间占用方面都会对索引数据形成一定的影响,因此通过禁用原始数据存储可以有效释放资源,保障数据索引效率;⑦提升服务器硬件配置:从硬件层面来提升数据的处理性能,从而保障数据索引效率。

**2.2 检索效率优化** 采用 Elasticsearch 分布式搜索引擎的默认配置信息即可提供有效的数据检索性能,然而在实际应用过程中随着数据量的剧增,默认的配置信息往往无法满足业务检索的需求,在硬件配置环境一定情况下可以通过以下方式进行优化配置:①采用分索引方案:按时间维度对索引进行分区划分,如对临床数据的检索一般都会有检索时间段条件,当数据体量较大时可以将索引数据按时间维度划分索引,2002–2015 年共 14 个索引,其中索引命名方式为 index\_2002,index\_2003....,当业务查询需要查询 2014–2015 年的病案首页信息时,后端执行 API 只需要查询 index\_2014,index\_2015 这两个索引即可,这样有效缩小了数据的检索范围,提升了数据的检索效率;②合理设置索引分片:分片数越多会带来越高的并发度,但并不是分片数越多越有效,分片数越多也会带来检索过程中数据的合并与 IO 的消耗,因此需按实际应用情况合理调整分片的数量;③根据实际情况可考虑去除 \_all 字段:\_all 字段是默认启用,主要用于全文检索,如果实际场景中只需要实现精确检索功能,可以去除 \_all 字段带来的索引负载;④采用 Elasticsearch warmer 实现数据热加载:基于缓存技术提升检索效率。

**2.3 检索准确率优化** 数据检索的准确率主要是体

现在分词的准确率上,而分词的准确率需要有业务相关的专有名词库支撑。如“门脉高压”一词,在 ik 分词器默认分词配置下,ik 分词器无法识别“门脉高压”一词会将其进一步切分,而如果将“门脉高压”作为专有名词库配置到 ik 分词中,其就能准确识别出“门脉高压”,在检索时可以准确检索出该词所对应的信息。另一类场景是同义词库的应用,在实际检索过程中,检索的信息不仅要精确出现,与检索信息意思相同或相近的结果也需要能够检索出来,如同样检索“门脉高压”一词,需要能够把包含“肝硬化”记录信息也能够检索出来,而通过配置同义词库可以实现该效果。

### 3 实际检索场景验证

**3.1 实验数据与环境** 实验数据采用某医院 2009 年的临床电子病历数据信息,所有数据已经过脱敏处理,数据总记录数为 12 696 458 条,其中各个表记录数见表 2。

表 2 临床数据中各表记录数

表名	表标识	记录数
病案首页	page	111 084
检验信息	lab	5 238 581
病历文书	docc	1 422 574
医嘱信息	advice	4 884 239
费用信息	coststat	110 947
手术信息	operation	61 174
诊断信息	diag	637 313
检查信息	exam	230 546

测试所用的 Elasticsearch 集群服务包含 3 个节点,其中每个节点服务器配置信息为:Centos7 64 位操作系统、64 GB 内存,CPU 双路 24 核 [Intel(R) Xeon (R) CPU E5-2620 v3 @ 2.40GHz], 磁盘空间 600 GB。

**3.2 数据索引性能分析** 以记录数最多的检验信息(lab)数据为例,在保证数据导入源端一致的情况下,通过优化 Elasticsearch 集群及索引的配置,分析优化操作对数据导入性能的影响,形成对比结果见图 3。图中 A 是采用 Elasticsearch 集群默认的服务节点及索引配置,其中默认服务节点为 1 G 内存,默认索引配置为启用 1 个副本,5 个分片,自动索引刷新;B 在 A 的基础上去除分片副本;C 在 A 的基础上取消索引自动刷新并增大索引数据提交批量;D 在 A 的基础上增大分片数,10 个分片;E 在 A 基础上增大各个 Elasticsearch 节点内存;F 为所有优化集成。可以看出,通过对服务节点及索引的配置优化,B~F 条件下的数据导入性能相较于 A 都有明显提升,F 配置下的数据索引性能最优。

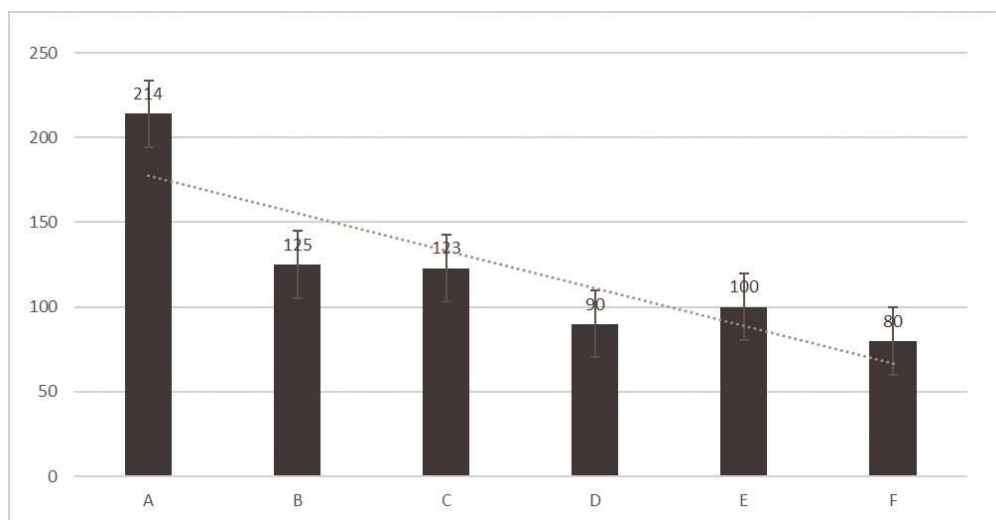


图 3 数据索引性能分析

3.3 关联检索结果分析 关联检索主要是针对临床数据中心部分复杂的关联查询需求,通过父查子/子查父两类检索场景验证 Elasticsearch 检索服务的有效性。其中“子查父”指根据检验信息、病历文书、医嘱信息、费用信息、手术信息等子表信息,以及病案首页本身查询条件来检索病案首页信息;“父查子”指根据病案首页信息,以及任意子表本身的查询条件来检索相应的子表信息。实验结果见图 4。

图 4A 是 1 个子查父的关联检索场景,其具体检索需求为:患者诊断为“慢性肾炎”且手术中采用了“全身麻醉”的所有病案首页信息,即根据诊断信息以及手术信息来关联检索病案首页信息。图 4B 是 1 个父查子的关联检索场景,其具体检索需求为:患者性别为男性,且入院日期为 2009 年 3 月 18 日至 2009 年 4 月 19 日之间的所有诊断记录信息,即根据病案首页信息来检索诊断信息。两种检索场

景都准确的检索出了相关结果,并且在千万级数据规模场景下检索效率分别为 42 ms 以及 104 ms。

3.4 数据检索效率分析 影响检索效率的因素有很多,包括硬件设施配置、集群中节点内存配置等<sup>[21]</sup>。本次主要从索引角度出发,在保证硬件配置一致,集群环境配置一致的前提下,通过调整索引分片数量来分析检索效率。实验过程中,基于同一份测试数据创建了 10 个索引,其中每个索引依次是 1 个分片到 10 个分片。通过模拟客户端发起 100 次 ES 检索请求,求取所有请求的平均值,分析不同分片索引下的检索效率,统计结果见图 5。可以看出,从 1 分片到 4 分片,分片数越多检索效率越高,而从 4 分片到 10 分片,分片数越多检索效率反而有所下降,主要原因是分片增多,数据并发处理能力虽然提升了,但是数据 IO 以及数据的合并所消耗的时间增加了。



注:A:子查父;B:父查子

图 4 关联检索分析



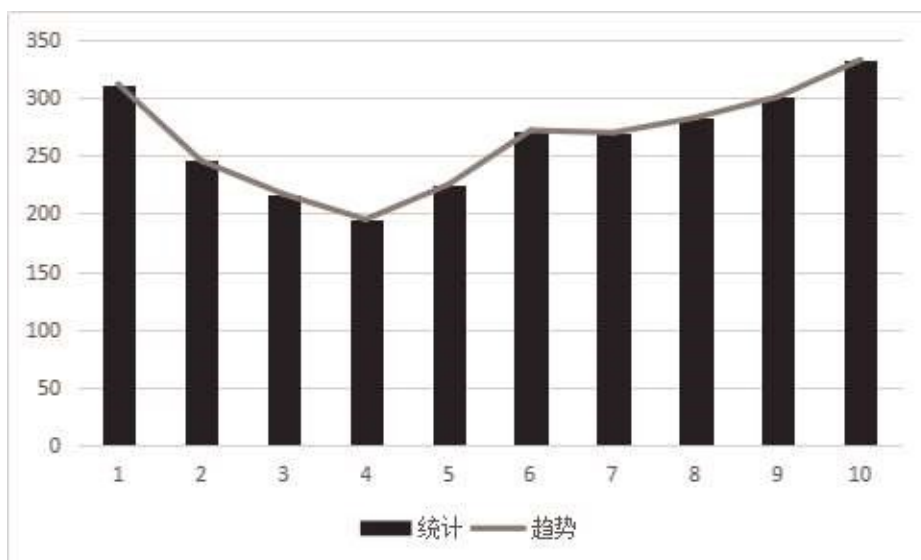


图5 不同分片下检索效率分析

3.5 检索准确率分析 IK 分词器支持自定义词典库的配置,在临床检索需求中,专有名词的准确识别对于检索准确率十分重要。使用 IK 默认词典库,无法达到准确分词的效果。为了进一步验证查询的准确率,本次建立了3个对比测试索引:ik\_all, ik\_none\_all 和 ik\_zymc\_all,其中3个索引的词库配置信息见表3。

表3 索引配置对比

索引名	专有名词库	同义词库
ik_all	√	√
ik_zymc_all	√	×
ik_none_all	×	×

按照是否配置专有名词、同义词库组合划分,本应还有1个索引,即只配置了同义词库,而没有配置专有名词库的,然而当没有专有名词库时,专有名词无法划分,那么仅配置同义词库将失去意义,因此用于实验的索引仅有3个。表中3个索引库中索引的数据一致,分别从3个索引库中检索“肝硬化”或“门脉高压”一词,查询统计结果见表4。

表4 检索结果对比

检索词	ik_all	ik_none_all	ik_zymc_all
肝硬化	6583 条	0 条	5331 条
门脉高压	6583 条	0 条	1549 条

可以看出,配置了专有名词库和同义词库,无论是搜索“肝硬化”,还是“门脉高压”都可以准确搜索出所有的记录;而仅使用专有名词的情况下,检索结果的查全率不足;若没有配置专有名词库、以及同义词库,那么在检索时将无法检索出任何信息。因此,专有名词库提升了检索的准确率,而配合上同义词

库后进一步提升了检索结果的查全率。

#### 4 总结

临床数据中心的建设为临床分析、业务优化、决策支持等提供了良好的数据支撑。然而,数据持续增长及业务场景的复杂性都使得传统关系型数据库无法有效满足临床医生及科研人员对海量数据信息的检索与分析需求。本研究提出的基于 Elasticsearch 分布式搜索引擎的临床信息检索方法,可实现复杂业务关联信息的检索,同时结合一系列的优化策略进一步提升了临床信息的索引效率、检索效率,以及检索准确率,可快速为临床医生、科研人员等提供准确的临床信息。

#### 参考文献:

- [1]韩煜.医院临床数据中心构建的思路分析[J].医学信息,2020,33(17):18-19.
- [2]周瑜,李永林.医院信息集成平台与临床数据中心建设探讨[J].中国信息化,2020(5):82-83.
- [3]Hak F,Guimares T,António Abelha,et al.Trends and Innovations in Information Systems and Technologies [M].Berlin: Springer,2020.
- [4]Solar M,Araya-Lopez M,Cockbaine J,et al.An Interoperable Repository of Clinical Data [C]// 2020 Seventh International Conference on eDemocracy & eGovernment (ICEDEG).2020.
- [5]磨云鲜.医院临床数据中心的建设与应用探讨[J].数字技术与应用,2019,348(6):100,102.
- [6]管雅文.整合生物信息的临床数据中心建设方案[J].中国数字医学,2019,14(2):57-59.
- [7]Lin MU,Zhang YY,Han Y.Research and Realization on Big Data of Science and Technology Resources Search Engine Based on SOLR[J].DEStech Transactions on Social Science Education and Human Science,2020.

(下转第30页)

(上接第14页)

- [8]Madec J,Bouzille G,Riou C,et al.eHOP Clinical Data Warehouse: From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network [M]//OhnoMachado L,Seroussi B.Studies in Health Technology and Informatics.2019.
- [9]Yu YW,Weber GM.Balancing Accuracy and Privacy in Federated Queries of Clinical Data Repositories: Algorithm Development and Validation [J].Journal of Medical Internet Research,2020(22):e1873511.
- [10]Heart T,Ben-Assuli O,Shabtai I.A review of PHR, EMR and EHR integration: A more personalized healthcare and public health policy [J].Health Policy and Technology,2017(6):20-25.
- [11]王伟,魏乐,刘文清,等.基于 ElasticSearch 的分布式全文搜索系统[J].电子科技,2018,31(8):56-59.
- [12]Shenoi SJ.Developing a Search Engine for Precision Medicine. AMIA Joint Summits on Translational Science proceedings [J].AMIA Joint Summits on Translational Science,2020(2020): 579-588.
- [13]宋玉红,吴琼贵.基于 Elasticsearch 大数据搜索引擎全网流监测系统设计与实现 [C]//第十四届全国信号和智能信息处理与应用学术会议.2021.
- [14]许雪晶,陈捷,林辰玮.基于 Lucene 的医疗搜索引擎排序算法的研究[J].长春师范大学学报,2020,39(6):54-58.
- [15]Taylor R,Ali MH,Varley I.Automating the processing of data in research. A proof of concept using elasticsearch[J].International Journal of Surgery,2018(55):S41.
- [16]李敏波.基于 JSON 文档结构的工业大数据多维分析方法[J].中国机械工程,2020,31(14):1700-1707,1716.
- [17]程彪,张晓明,阮晨.基于 Elasticsearch 的知识库和病案检索服务平台的设计与实现[J].中国病案,2021,22(3):44-48.
- [18]Elasticsearch BV."Frozen Indices" in Patent Application Approval Process (USPTO 20200326986) [J].Computer Weekly News,2020(2020):46-47.
- [19]Jonassen S.Efficient query processing in distributed search engines[J].ACM SIGIR Forum,2012,47(1):111-135.
- [20]梁爽,赵宝军,张海霞.基于 Elasticsearch 的海量数据入库及快速检索方法研究[J].测绘与空间地理信息,2020,43(12):74-76.
- [21]范朗.Elasticsearch 海量数据存储查询优化[J].工业控制计算机,2020,33(10):85-87.

收稿日期:2021-09-06;修回日期:2021-09-18

编辑/成森