



搜索引擎技术在移动互联网音乐领域的研究与实践

杨翊平, 张玉忠, 杨豪杰, 宋平波

(中国电信股份有限公司广州研究院 广州 510630)

摘要: 为了优化移动互联网用户对音乐搜索的体验,对搜索引擎的性能、搜索查准率及查全率、更新及时性等方面提出了新的要求。针对这些需求,对搜索引擎技术在音乐搜索领域的应用进行了深入研究,研发了一套可行的面向移动互联网音乐搜索的搜索引擎解决方案。方案采用层级式搜索策略提高了搜索引擎规则组合的灵活性,同时,内存索引和文件索引相结合的索引机制保证了索引更新的及时性。

关键词: 音乐搜索;搜索引擎;移动互联网

doi: 10.11959/j.issn.1000-0801.2015036

Study and Applying of Search Engine Technique in Mobile Internet Music

Yang Yiping, Zhang Yuzhong, Yang Haojie, Song Pingbo

(Guangzhou Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China)

Abstract: The music business of China Telecom have high performance, more reasonable precision ratio and recall ratio, timeliness of indexing and other needs in music search to meet the requirements for mobile internet users. The search engine technology for applications in the field of music search was explored and a solution of music search for the mobile internet users was put forward. The hierarchy searching strategies improve the flexibility of the search engine when combining it's searching rules. The indexing mechanism which implements by integrating the memory index and the file index ensures to present music in time.

Key words: music search, search engine, mobile internet

1 引言

对于大多数人来说,音乐是生活中不可或缺的一部分。数据显示,2013年底中国无线音乐市场用户规模达9.12亿户,同比增长9.6%^[1]。

天翼爱音乐文化科技有限公司作为一家音乐内容运营商,有超过70%的用户试听来自搜索,这就对搜索引擎的搜索性能、搜索查准率及查全率、资源更新及时性等提出了新的要求。

2 音乐搜索的关键问题

2.1 搜索查准率与查全率的权衡(搜索规则的权衡)

转化率是衡量搜索引擎质量优劣的最直观的指标,而查全率和查准率则是影响转化率的两个重要参数^[2]。高查全率可让更多的搜索结果呈现给用户,但某些结果与用户的查询条件的相关度不高;高查准率则为用户展现与查询条件的相关度高的结果,但是会影响搜索结果的数量。在实际实施过程中,两者受搜索相关度^[3]控制,不可兼得。在

音乐搜索中,高查全率要求搜索引擎尽可能多地返回歌曲,结果的搜索相关度较低。高查准率要求返回的歌曲中与查询条件相关的歌曲占比尽可能高,结果的搜索相关度较高。因此,在设计搜索引擎时,需要定义好搜索相关度,对查全率和查准率进行折中,并根据用户的行为反馈适时调整算法。

与通用搜索引擎不同,音乐搜索属于垂直领域,用户的诉求是查找歌曲,需求相对集中。而歌曲中相同歌名甚至演唱者同名的情况时有发生,在搜索结果中经常出现不同资源的搜索相关度相同的情况。因此,在进行搜索结果排序时,除了考虑相关度,应需要考虑其他因素,如歌曲的热度。两者综合考虑方能得到一个较佳的搜索结果。

绝大部分搜索引擎都会根据其自身的业务需求,将搜索结果按照特定的业务规则做二次排序。音乐搜索中,音乐提供商一般会在一段时期内根据业务运营的需要主打一批资源,这批资源需要在用户搜索时适时地展现给用户,但又不能太过影响用户的体验。因此,在设计搜索结果排序规则时,需要结合搜索相关度计算以及营销的业务规则综合考虑,并随着业务变化进行适时的调整策略。

2.2 多样化的搜索条件

当前,音乐已深入其他娱乐、文化相关领域,各种影视类节目甚至游戏、动漫等都充满着各种音乐元素,产生出许多脍炙人口的音乐内容。各种影视剧歌曲、娱乐节目相关歌曲等都开始进入各类音乐网站的歌曲热榜,成为脍炙人口的流行歌曲。百度音乐 20140806 期的热播榜前 10 名中,超过半数都是影视剧、节目类歌曲。用户在搜索此类歌曲时,往往会使用影视剧名称、节目名称或名称缩写等作为关键词。

另外,随着智能手机的普及,越来越多的用户通过手机浏览器、手机应用来查找、欣赏音乐,但是手机触屏的操作方式也导致了用户输入错字的概率提升。搜索引擎需要解决因用户输入错词(如同音错字、形似错字)引起的搜无问题。

2.3 海量信息及时更新

曲库每天都有新歌上线、歌曲版权到期下线、歌曲信息变更等,需要及时将相关信息更新到搜索引擎,保证搜索结果的准确性。另外,版权音乐存在集中到期下线、批量续期、批量导入等情况,这些都涉及索引的大批量更新操作,大批量索引更新往往引起搜索性能显著下降,影响用户体验。以上列举的情况需要一种合适的索引更新策

略,既要确保将一些实时性要求高的操作及时更新到索引,又要避免在大批量更新索引时,影响前端搜索的响应速度。

3 音乐搜索引擎系统

本文基于 Apache 组织的开源项目 Lucene^[4],研发了一套可灵活组合各种业务规则,可满足移动互联网时代用户对音乐搜索需求的音乐搜索引擎。

Lucene 是一款优秀的开源全文检索工具,被广泛应用于全文检索领域。

本文研发的音乐搜索引擎具有以下特点。

灵活的业务规则组合方式:支持将每一种业务规则以插件的方式加载入搜索引擎,各规则的组合顺序可动态调整。

搜索结果按相关度和歌曲热度综合排序:可根据需要,设定歌曲不同信息字段的权重、设定各种相关度计算因子等;根据用户历史行为计算歌曲热度。

支持歌曲延伸信息搜索:将影视剧名称等延伸信息与歌曲关联,支持通过影视剧名称、节目名称等方式搜索相关歌曲。

支持快速索引和批量索引:歌曲更新可快速同步到搜索引擎并被用户搜索到。批量的歌曲上下线操作的同时保证用户搜索体验不受影响。

图 1 所示是搜索引擎的系统架构。

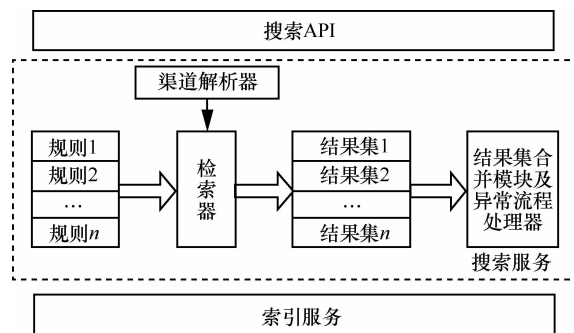


图 1 搜索引擎系统架构

3.1 层级式搜索策略

数据显示,在互联网上搜索音乐时,90%以上的中国网民选择输入歌曲名称进行音乐搜索,64.7%的网民通过输入歌手姓名进行音乐搜索^[5]。另外,音乐提供商不时会有营销活动,需要在搜索结果中向用户推荐营销资源。

根据上述情况,可将音乐搜索引擎系统的业务规则设计为层级式,定义搜索的规则为:

$$S=(s_1,s_2,\cdots,s_n) \quad (1)$$



其中 s_i 是某一层级的子规则, 定义 r_i 是用户搜索关键词通过 s_i 产生的一个结果集, 则将搜索规则 S 的结果集 R 定义为:

$$R=r_s \cup r_1 \cup r_2 \cup \dots \cup r_n \quad (2)$$

R 是有序集, 其元素根据在 r_i 中出现的顺序排序, 即 $sr_1 > sr_2$, $sr_1 \in r_1$, $sr_2 \in r_2$ 且 $sr_1 \neq sr_2$ 。 r_s 是营销集合, 当有营销活动, 将营销歌曲加入 r_s , 则当集合中的歌曲满足用户搜索条件时, 会优先展现。 r_1 、 r_2 等可根据系统中用户搜索行为习惯动态组合调整。

表 1 是根据上述分析定义的一种具体的音乐搜索业务规则。 r_n 对应了图 1 中的规则 n , 构成层级式搜索规则, 每一个搜索规则解析搜索关键词, 得到一个 Lucene 可识别的查询语句, 调用检索器得到搜索结果集。最后由结果合并模块将结果集整合输出一次搜索的结果。

表 1 音乐搜索业务规则

规则序号	规则描述
r_s	营销集合歌曲名、歌手名全文检索
r_1	歌曲名匹配
r_2	歌手名匹配
r_3	专辑名匹配
r_4	歌曲名、歌手名、专辑名等全文检索
r_5	歌词模糊匹配

3.2 搜索结果排序规则

根据式(2), r_i 之间是有序关系, 因此通过子规则搜索得到的结果实际上已根据子规则的层级关系做了一次排序。因存在同名歌曲、同名歌手的情况, 在对子规则的搜索结果集进行排序时, 需综合考虑相关度和歌曲热度。

Lucene 将搜索结果中每一个结果的相关度计算定义为:

$$score(q, d) = coord(q, d) \cdot queryNorm(q) \cdot \sum_{t \in q} (tf(t \text{ in } d) \cdot idf(t)^2 \cdot t.boost \cdot norm(t, d)) \quad (3)$$

该算法比较适合传统的全文检索。在音乐搜索中, 因为歌曲名称、歌手名称等关键信息的文本长度都比较短, 且同一个词语多次匹配以提高相关度的意义不大。因此, 式(3)可修改为:

$$score(q, d) = coord(q, d) \cdot queryNorm(q) \cdot \sum_{t \in q} (idf(t)^2 \cdot t.boost \cdot norm(t, d)) \quad (4)$$

各项定义描述见表 2。

歌曲热度有多种定义, 可直接使用试听量或下载量。本文结合歌曲在各种业务中的订购量以及从互联网收集到的数据, 对歌曲进行综合打分来获得歌曲热度。

表 2 歌曲相关度计算式说明

式项	描述
$coord$	查询词在歌曲信息中出现的个数
$queryNorm$	搜索条件因子
idf	逆向文档率 ^[6] , 词语在所有歌曲信息中的区分度
$t.boost$	搜索词分词后, 每个词语的权重因子
$norm$	每首歌曲的索引因子, 和可用于检索每个字段的权重相关

定义一首歌曲的热度 H 如式(5)所示, 其中 u 是某一用户对歌曲的综合分数, 定义如式(6)所示。

$$H = 1 / (\sum u + \sum m'f) \max \quad (5)$$

$$u = \sum_j (m^2 \cdot f_j) / M \quad (6)$$

式(5)通过归一化为每首歌曲在统一的数值区间内计算一个热度。其中 $m'f$ 是某一首歌曲从某一互联网渠道收集到的歌曲热度值。 $\max$ 是系统中设定的最大热度值。

式(6)中 m 是用户在一个计算周期内通过业务 j 访问了歌曲 m 的次数加权重, f_j 是业务 j 权重因子, M 是用户在一个计算周期内通过业务 j 访问的所有歌曲的次数加权重。用户在所有业务中访问歌曲 m 所得分数即为用户对歌曲 m 的综合打分。

在进行搜索排序计算时, 每个子规则得到的结果集先根据搜索相关度进行排序, 然后再根据热度排序。考虑到热度值区分粒度太细导致相关度不起作用, 可对热度进行分区间及平滑处理。具体为, 首先设定热度阈值 d , 热度值在 $[d, \max]$ 之间方能起效。然后将热度区间划分为最热区间及普通区间。最热区间中的歌曲按先热度后相关度排序。普通区间内的歌曲划分若干等级, 每个等级内的歌曲视为具有相同热度, 按相关度排序。

3.3 歌曲延伸信息整合及搜索

歌曲延伸信息整合及搜索可分为 3 步: 搜集、整合关联、搜索。

采用网络爬虫技术从各音乐网站的影视金曲榜搜集歌曲延伸信息, 如百度影视金曲榜、酷狗电影歌曲等。可从其歌名、专辑名等信息中提取到相关歌曲延伸信息。另外从各百科网站可抓取影视剧介绍信息中的音乐相关信息, 从中提取相关歌曲名称, 与曲库歌曲匹配。

将搜集到的延伸信息进行规范化处理, 将处理后的规范信息填入曲库中歌曲的延伸信息字段 $extInfo$ 。图 2 所示是规范化处理过程。

在搜索阶段, 根据业务规则的设计, 将 $extInfo$ 加入表 2 中的子规则 r_1 和 r_4 , 当搜索关键词通过 r_1 能匹配到歌

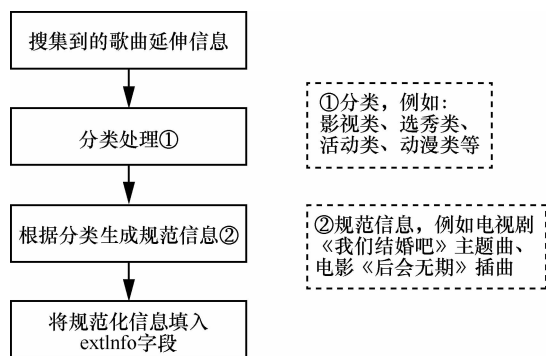


图 2 歌曲延伸信息规范化处理

曲时,可保证匹配的歌曲排名靠前;否则,由 r_4 确保相关歌曲被搜索到。

3.4 搜无异常处理

若经过上述层级式的搜索过程仍无法搜到相关结果,系统将进行搜无异常处理。搜无异常处理包括同义词处理和搜索词纠错。

搜索引擎维护一个同义词表,例如“王菲”和“王靖雯”、“好声音学员”和“李琦”、“姚贝娜”是一组同义词,若搜索关键词能够在表中找到同义词,且同义词搜索结果数量达到设定的阈值,搜索引擎将返回同义词搜索的结果。若找不到同义词或同义词搜索的结果数量达不到阈值,进入纠错功能。

搜索引擎的输入纠错功能^[7]是指用户在输入关键词进行搜索的时候,如果搜索引擎在返回结果中计算出与此关键词相似的另一形式(如词组中出现同音不同字或者某一错别字现象)得到大量的搜索结果,会尝试将用户输入的关键词“纠正”为系统推测的关键词项。

对于英文纠错,因为英文中每个词之间有空格,不用考虑分词,只需要对每个单词进行拼写检查即能满足大部分纠错需求。中文则需要进行分词之后再行错误检查和纠正,一般处理中文纠错的方法有基于上下文语言统计模型纠错、基于词典纠错等。本文研发的搜索引擎采用 Lucene 社区的开源组件 spellchecker 实现关键词纠错。该组件是基于 N-gram 模型^[7],配合编辑距离算法^[8],借助歌名、歌手名等关键信息构建的语料库进行关键词纠错。对于一个错误关键词,经纠错模块得到一个关键词,若该关键词搜到的结果数量达到阈值,搜索引擎将返回这个纠错后的关键词及其搜索结果。

3.5 索引更新

索引更新是指存储在曲库中的歌曲信息发生变更时,

通过一种机制将变更同步到搜索引擎的索引,确保用户搜索到最新的歌曲信息。

在本文研发的系统中,以下情况会引起歌曲数据变更:

- 基本信息变更,包括歌曲名变更、专辑变更、歌词变更、歌曲标签变更;
- 歌曲延伸信息变更;
- 歌曲到期自动下线;
- 新歌上线。

Lucene 的索引由若干段组成,每次提交更新会生成一个新的段,称为索引碎片。若将索引完全存储在磁盘文件系统上,随着索引段数增加,索引碎片过多会导致搜索性能下降。Lucene 提供了合并索引碎片的方法,但合并索引过程系统资源开销大,过多的触发索引合并操作会严重影响搜索引擎性能。

本文研发的搜索引擎将索引分为内存索引和文件索引两部分,图 3 是系统索引的更新机制。其中内存索引是更新索引,存放一天内有变更的歌曲。文件索引是主索引,存放曲库所有歌曲的数据。系统维持一个待索引记录队列,队列中的记录按优先级排序,索引器在收到有歌曲信息变更通知后,从队列读取数据更新索引。对于新增和修改的操作,若记录存在于文件索引中,将记录标记为删除后在内存索引中新增一个记录;对于删除操作,找到对应的记录后标记删除。完成一次提交操作后,索引器向检索器发送一个刷新通知,让检索器加载有变更的索引。

具体的,待索引队列中的记录记为 (S_x, P_x) , S_x 是记录内容, P_x 是权重因子,将索引区分为紧急索引、普通索引、批量索引。对于紧急索引和普通索引,索引器对索引更新

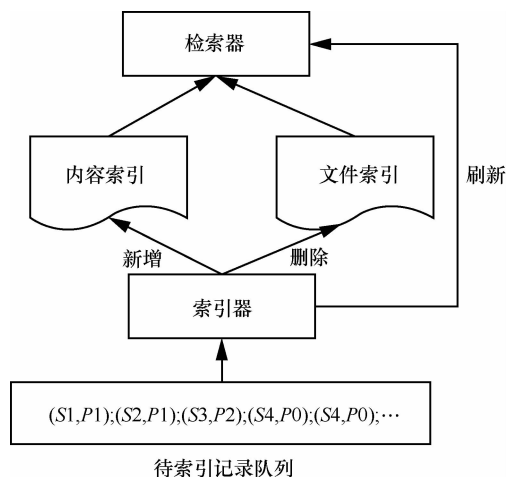


图 3 索引更新机制



之后即提交并通知检索器刷新。当遇到批量索引时,索引器会对大批量更新进行分批处理,同时确保在一次索引更新过程中处理尽可能多的记录,以减少因频繁提交导致索引碎片过多。

索引器维护一定时器,在每天闲时将内存索引合并到文件索引,清空内存索引,通知检索器重新加载索引后,索引合并维护进入一个新的周期。每周中一天定时全量重建文件索引。

4 音乐搜索引擎的应用

本文研发的音乐搜索引擎已应用于天翼爱音乐文化科技有限公司各门户,为“爱音乐”Web、WAP、客户端用户提供一致的搜索体验,取得了良好的应用效果。

音乐搜索引擎灵活的层级式搜索组合策略良好地支持了爱音乐各种业务需求,以近两年热门的选秀节目“中国好声音”为例,运营人员将节目相关的歌曲资源放入营销推荐集合(表2的 r_s 规则),当用户搜索“中国好声音”、“好声音”、“趁早”等该节目相关的关键词时,会优先返回营销推荐集合中的歌曲。而节目中更受欢迎的歌曲,其热度值也更高,通过综合相关度及热度的排序方式,在返回的结果中也能够排名靠前,减少了用户的搜索、翻页次数,显著提升用户体验。当节目档期结束后,关闭营销推荐的规则,则搜索引擎恢复按正常规则搜索歌曲。

为检验引入层级式搜索组合策略的方式是否影响搜索引擎性能,本文使用实际生产系统的用户搜索数据进行实验,结果见表3。实验表明,在同等数据量及访问量的情况下,引入层级式搜索组合策略在满足了音乐搜索引擎的

业务规则灵活组合的同时,系统的请求响应速度与未引入该策略时相当。

“爱音乐”的音乐资源均为正版授权,在授权到期后需要自动下线。授权一般采用批量的方式,因此会出现在某一天有大批量歌曲下线需要批量更新索引的情况。搜索引擎系统在遇到大批量索引更新时,自动分批处理更新索引。Lucene的索引特性决定了频繁进行增量索引的操作会产生大量索引碎片,导致搜索性能下降。本文对索引结合“爱音乐”实际的音源资源更新情况,试验了多种处理情况,表4是试验结果。实践证明,将索引碎片控制在80个以下,搜索引擎仍可以保持较好的搜索性能。笔者统计了半年来的歌曲更新记录,每天最大的更新量约为100 000首。因此,将每批增量索引的记录控制在2 000个,在索引处理时间及索引碎片的增量上可以取得较好的折中。

从用户访问统计看,访问高峰期的平均响应时间低于300 ms,一天内用户搜索的点击转化率接近90%,由此可见搜索引擎为用户提供良好的响应体验的同时,满足了绝大部分用户的搜索需求。

5 结束语

移动互联网正在飞速发展,用户查找音乐的方式也日趋多样化、个性化,搜索引擎仍将作为基础设施为用户服务。本文针对大部分用户搜索音乐的需求,对音乐搜索技术进行了研究和实践,将来还将结合自然语言识别技术,继续优化完善,满足移动互联网时代追求更加个性化需求的音乐搜索引擎。

表3 音乐搜索引擎搜索策略性能对比

场景	索引记录数/个	搜索关键词数/个	并发连接数/个	测试时间/s	搜索平均响应时间/ms	每秒事务数/个
无层级式搜索组合策略	268 416	130 648	50	300	204	220
	518 966	130 648	50	300	235	217
层级式搜索组合策略	268 416	130 648	50	300	198	226
	518 966	130 648	50	300	243	209

表4 增量索引策略及其对搜索性能影响的对比

索引记录数/个	搜索关键词/个	更新量/个	每批增量索引/个	索引碎片/个	并发连接数/个	搜索平均响应时间/ms
268 416	130 648	50 000	1	≥ 200	50	980
268 416	130 648	50 000	500	80~200	50	612
268 416	130 648	50 000	2 000	< 80	50	301

参考文献

- 1 2013-2014 年中国无线音乐市场年度报告. <http://www.techweb.com.cn/news/2014-03-13/2015698.shtml>, 2014
2013-2014 China wireless music market annual report. <http://www.techweb.com.cn/news/2014-03-13/2015698.shtml>, 2014
- 2 毛振鹏, 胡滨, 代海岩. 搜索引擎质量评价体系研究. 晋图学刊, 2005(5)
Mao Z P, Hu B, Dai H Y. Search engine quality appraisal system research. Shanxi Library Journal, 2005(5)
- 3 贺林. 搜索引擎中的相关度. <http://abc.wm23.com/lin1144896348/112185.html>, 2011
He L. The relevance of search engine. <http://abc.wm23.com/lin1144896348/112185.html>, 2011
- 4 Apache Lucene. <http://lucene.apache.org/>, 2014
- 5 2009 年歌曲名称为中国网民最常用的音乐搜索关键词. http://www.iresearch.com.cn/html/consulting/search_engine/DetailNews_id_104891.html, 2009
2009 Song names used commonly as music search by keywords for Chinese Internet users. http://www.iresearch.com.cn/html/consulting/search_engine/DetailNews_id_104891.html, 2009
- 6 TF-IDF. <http://zh.wikipedia.org/wiki/TF-IDF>, 2014
- 7 陈智鹏, 吕玉琴, 刘华生等. 基于 N-gram 统计模型的搜索引擎中文纠错. 中国电子科学研究院学报, 2009(3)
Chen Z P, Lv Y Q, Liu H S, *et al.* Chinese spelling correction in search engines based on N-gram model, Journal of CAEIT, 2009(3)
- 8 吉胜军. 基于 Levenshtein distance 算法的句子相似度计算. 电脑知识与技术, 2009(9)
Ji S J. The similitude calculate of sentence based on Levenshtein distance algorithm. Computer Knowledge and Technology, 2009(9)

[作者简介]



杨翊平, 男, 中国电信股份有限公司广州研究院工程师, 主要研究方向为搜索引擎技术、云计算、互联网应用。

张玉忠, 男, 中国电信股份有限公司广州研究院工程师, 主要研究方向为数据库技术、数据挖掘技术。

杨豪杰, 男, 中国电信股份有限公司广州研究院工程师, 主要研究方向为推荐引擎技术、大数据技术及应用。

宋平波, 男, 中国电信股份有限公司广州研究院工程师, 主要研究方向为移动 App 研发、性能管理等。

2015036-6

(上接 2015039-5)

- 5 虞凯, 范平志, 郝莉. 基于大系统方法的上行 CDMA 蜂窝系统性能分析. 通信学报, 2011(1): 21~26
Yu K, Fan P Z, Hao L. Performance analysis of uplink CDMA cellular systems based on the large system method. Journal on Communications, 2011(1): 21~26
- 6 黄韬, 刘韵洁, 张智江等. LTE/SAE 移动通信网络技术. 北京: 人民邮电出版社, 2009
Huang T, Liu Y J, Zhang Z J, *et al.* LTE/SAE Mobile Communication Network Technology. Beijing: Posts & Telecom Press, 2009
- 7 钱毅. 板状天线作机械倾斜的垂直方向性函数的计算及分析. 电视工程, 1997(2): 10~15
Qian Y. Calculation and analysis of mechanical inclined vertical directivity function for the plate antenna. Television Engineering, 1997(2): 10~15
- 8 魏玉杰. LTE-A 系统载波聚合下基于用户分组的比例公平调度算法研究. 重庆邮电大学学报, 2013(6): 310~312
Wei Y J. PF scheduling algorithm research based on user grouping for LTE-advanced system with carrier aggregation. Journal of Chongqing University of Posts and Telecommunications, 2013(6): 310~312
- 9 李佳俊, 文博, 许国平. FDD LTE 系统容量研究. 邮电设计技术, 2013(3): 36~41
Li J J, Wen B, Xu G P. Research on the Capacity of FDD LTE System. Designing Techniques of Posts and Telecommunications, 2013(3): 36~41

[作者简介]



周戈, 男, 中国电信深圳无线网络运营中心高级工程师、网优一级工程师, 主要研究方向为移动通信网络系统模型和大数据分析。

王军文, 男, 中国电信深圳无线网络运营中心高级工程师、网优一级工程师, 主要研究方向为移动通信网络规划和优化。

2015039-6