

Southwest University of Science and Technology

本科毕业设计（论文）



基于NoSQL的网络文学监测分析系统优化

|  |  |
| --- | --- |
| 学院名称 | 计算机科学与技术学院 |
| 专业名称 | 软件工程 |
| 学生姓名 | 曾丹 |
| 学号 | 20134484 |
| 指导教师 | 杨春明 副教授 |

二〇一七年六月

基于NoSQL的网络文学监测分析系统优化

摘要：网络文学监测分析系统是一款具有统计分析网络文学作者作品地域分布、性别分布、类型分布、排行分布以及评论词云等功能的系统，能够通过该系统清晰明了的洞察网络文学的发展趋势。目前每日可采集到相关作者作品信息上千条，随着时间的推移，数据的累积，现已采集作者数达30万，作品数近30万，评论数165万。由于该系统采用的传统关系型数据库Mysql，关系型数据库需提前设计表结构以及表之间关系，导致部分查询需要至少3个表的联表查询，导致部分统计分析接口响应时间超过5分钟，十分影响用户体验。针对上述问题，结合当下已成熟的应对海量数据的解决方案，决定更换数据库并修改逻辑，采用NoSQL数据库存储数据。NoSQL数据库即非关系型数据库的总称，分别有Key-Value键值对形式、面向列、面向文档的存储形式存储数据，其中主流NoSQL数据库有Hbase、Redis、MongoDB等数据库。Hbase功能强大，但相较于其他数据库略微笨拙；Redis是一款内存数据库，若内存增长过快需定期清理，更适合于数据变化快的系统；MongoDB为最类似于关系型数据库的NoSQL数据库，十分适合注重海量数据访问效率提升的系统。经过比对，基于MongoDB数据库的系统优化，页面平均响应时间小于3秒，比对优化前后的页面响应速度，性能有了明显的提升。

首页1min以上，近期更新图标30+s，top作者20s+

作者统计页面，统计25s+；作者列表，25s+；

作品详细页面，

关键词：网络文学监测分析系统；NoSQL数据库；MongoDB数据库

**Base on NOSQL database on the monitoring and analysis of network literature system optimization**

**ABSTRACT**

The network literature monitoring and analysis system is a system which has the functions of statistical analysis of the geographical distribution, gender distribution, type distribution, ranking distribution and commentary clouds of the author's work, which can clarify the development trend of network literature through the system. At present, the daily collection of relevant works of thousands of works, with the passage of time, the accumulation of data, has been collected the number of 300,000, the number of works nearly 300,000, the number of comments 1.65 million. As the system uses the traditional relational database Mysql, the relational database needs to design the table structure in advance and the relationship between the table, resulting in some inquiries need at least three tables of the table query, resulting in part of the statistical analysis of the interface response time of more than 5 minutes, Affect the user experience. In response to the above problems, combined with the current solution to deal with massive data, decided to replace the database and modify the logic, using NoSQL database to store data. NoSQL database that is non-relational database of the general term, respectively, Key-Value key pairs of form, oriented, document-oriented storage form of storage data, which mainstream NoSQL database Hbase, Redis, MongoDB and other databases. Redis is a memory database, if the memory growth too fast to be cleaned regularly, more suitable for data fast changes in the system; MongoDB is the most similar to the relational database NoSQL database, but the database is not the same as the database, Very suitable for focusing on massive data access to improve the efficiency of the system. By comparison, based on the MongoDB database system optimization, the page average response time is less than 3 seconds, compared to the optimization of the page before and after the response speed, performance has been significantly improved.

**Key words:** The network literature monitoring and analysis system, NoSQL database, MongoDB database

1. 绪论

1.1课题的背景

伴随信息时代的高速发展，数据以爆炸式方式增长。面对海量数据的存储与处理，传统关系型数据库已无法满足用户需求，甚至严重影响系统性能以致用户体验呈直线下降。随着Internet网的普及，人们的生活方式也发生了巨大变化，由现金支付转为网上支付，由纸质书转为电子书等等。

网络文学监测分析系统是一款统计分析网络文学作者作品相关信息的网站，以图表形式清晰明了的展现给用户，能够快速了解与掌握网络文学作者作品的现状与发展趋势。

本文基于传统关系型数据库Mysql为网络文学监测分析系统主要的制约因素研究在海量数据环境下产生的NoSQL数据库在该系统的应用，比对各NoSQL数据库优缺点以及适用场景，通过更换数据库及系统逻辑实现，旨在优化网络文学监测分析系统性能，提升系统的响应速度，使得用户体验有质的飞跃。

1.2课题的研究意义

由于网络科技的高速发展与移动设备的普及，基本人手一台手机，达到足不出户便能知晓天下事。小说阅读是当今人们生活中重要的娱乐方式之一，李克强曾指出：读书不仅事关个人修为，国民的整体阅读水准，也会持久影响到整个社会的道德水平。随着2015年“两会”的圆满结束，“全民阅读”的发展进程进一步加快。在移动互联网时代的大背景下，移动设备已逐渐取代了传统的实体书，成为了人们主要的阅读工具。据统计，截止2015年，网络文学用户规模已达3.5亿人，而纸质图书作为最“正统”的阅读工具，用户仅占15%，因此阅读移动化已经势不可挡，也正是由于网络文学作品的大量出产，所以掌握网络文学作品发展趋势以方便对其管控和了解用户需求是尤为重要也是极其必要的。而网络文学监测分析系统正是以简单有效的图表形式展现给用户，可更方便了解网络文学用户的需求以及网络文学作品发展态势。

1.2.1 必要性

网络文学监测分析系统的性能随着爬取的网络文学作者作品信息数量递增，出现明显的下降，经过几次维护与修改，仍然未能从根源彻底解决问题，导致系统使用者往往想要查看数据分析的图表需要花费大量的时间等待，甚至有时会超过用户等待时间的最大限度，导致系统用户渐渐减少，如果这样的情况不能得到及时有效的解决，此系统将会成为一款无人使用的系统，就如同一栋烂尾楼，无人问津。为避免网络文学监测分析系统成为一死系统，所以对系统的优化不能停滞不前，要从根源解决问题才是根本，经过问题的排查和之前的维护与修改经验，最终确定是使用传统关系型数据库Mysql存储数据的原因，Mysql数据库需要提前定义好表字段以及表字段间关系，由于要满足三范式，所以获取一条数据往往需要关联多张表数据，在大数据量的情况下，查询效率差，因此更换数据库，基于NoSQL数据库完成对系统的优化是极其必要的。

1.2.2 重要性

基于非关系型数据库的网络文学监测分析系统与基于传统关系型数据库的网络文学监测分析系统相比较，具有以下6个优点：

1. 将关系型数据库Mysql替换为非关系型数据库MongoDB能够从根源解决系统性能瓶颈问题；
2. MongoDB采用的是面向文档存储方式，不需要提前定义好表字段及表之间关系，内部使用的是BSON语法，能够更直观易理解；
3. MongoDB海量数据下，查询效率高效快速，进而解决系统性能低下问题；
4. MongoDB使用简单，可以支持JSON、mapreduce、正则表达式以及Javascript函数等查询方式；
5. MongoDB相较于Mysql，节省系统资源，对CPU等资源的耗费较小；
6. MongoDB具有高可扩展性，支持分布式集群，通过分片数据缩放处理理论可实现更高的吞吐量。

1.3 课题研究现状

1.3.1 NoSQL数据库研究现状

NoSQL数据库最早可追溯到1991年Berkeley DB的第一版Berkeley DB是一个键值存储(Key-value Store)的Hash数据库，适用于数据类型相对简单但需要极高的插入和读取速度的嵌入式场合。NoSQL从2007年开始蓬勃发展。Google和Amazon的工程师们分别发表了有关BigTable和Dynamo数据库的论文，描述他们己经在使用的新型数据库的设计思想。BigTable提出了列存储模型，证明了数据持久存储可以扩展到数以千计的节点, Dynam。则提出了最终一致性的思想以达到更高的可用性和扩展性。分布式缓存系统Memcached证明了内存分布式数据存储可以达到极高的性能。事实上，Google的BigTable, Amazon的Dynamo和分布式缓存系统Memcached这三者可以说是所有NoSQL数据库的启蒙者。很多公司和组织受到他们思想观点的启发，开发了自己的开源 NoSQL数据库产品，例如Hypertable就是BigTable思想的开源实现。

现如今，国外的NoSQL数据库品种己经超过100个。除了Google的BigTable,Amazon的Dynamo，主流的数据库还有FaceBook的Cassandra, lOgen的MongoDB, Amazon的SimpleDB, Apache的HBase, Redis, CouchDB, Tokyo Cabinet/Tyrant, MemcacheDB等等。这些数据库在互联网中应用广泛，表现优异。Cassandra的用户包括Twitter, Cisco, Urban等多个公司，已知最大的Cassandra集群有超过400个节点，存储数据量超越300TB最受欢迎的MongoDB数据库在电子商务、政府、数据存储、实时统计、社交网络、游戏等众多领域都应用广泛，其中著名的有github, source forge, Disney等，国内的淘宝网、大众点评和视觉中国也在使用MongoDB国内的 N0SQL数据库领域在2009年也开始活跃起来。人人网下属UGC团队参考Dynamo设计文档，开发了自己的NoSQL数据库Nuclear。豆瓣推出了开源数据存储系统BeansDB。盛大创新院也基于Tokyo Cabinet开发了TCDatabase开源数据库。相比国外的NoSQL数据库，国内的这几款数据库并不成熟，也没有在生产中有更广泛的应用。国内大型的互联网公司多数还是选择国外支持度高的NoSQL数据库，除了淘宝、大众点评、视觉中国选用MongoDB外，新浪微博选用Redis存储数据，盛大也推出了基于MongoDB的“盛大云”MongoIC的数据库云服务。

在NoSQL蓬勃发展途中，有着成功案例的同时，NoSQL数据库也有一些失败的经历。2010年7月Twitter提出暂停使用Cassandra替代MySQL存储feed的计划，原因是Cassandra仍缺少大并发海量数据访问的案例及经验，新产品需要一定稳定期。2010年10月，使用MongoDB的foursquare网站由于数据碎片化和监控不力的原因，经历长达11小时的两次宕机。随着不断的完善，NoSQL也愈加成熟，得到广泛应用。

1.3.2 MongoDB数据库研究现状

MongoDB是一个性能卓越的开源的NoSQL数据库。它的存储模式自由，能够存储文档型数据等结构数据，灵活的存储方式在存储数据方面对传统的关系型数据库是一种强大的冲击，它逐渐的替代了Key/V alue存储方式。MongoDB的这些优点省略了开发过程中很多不必要的麻烦和繁琐的操作。同时，MongoDB能够很好的应对大多数互联网应用对大数据量、高并发的需求。MongoDB的水平扩展能力，使应用具备了处理上千万甚至上亿级别的数据处理能力。这些性能极大的满足了Web2.0和移动互联网系统对大数据存储的需求，同时，它的即插即用的特点也从某种程度上大大降低了相关行业网站的运行和维护的成本。两年来，跟随着Web2.0的发展势头，国内对NoSQL数据库出现了热忱的局面，MongoDB作为NoSQL中性能比较优越的一个，自然的成为了最有影响的一个。很多的互联网公司都将MongoDB作为数据库存储的最佳选择，在大数据的存储上也给专家、学者们在科研上提供了很好的存储服务，在行业内创造了很多具有不可估量的价值的互联网、移动应用。

1.4课题目标和研究范围

1.4.1 目标

在当前Web2.0时代的互联网应用和大数据时代的海量数据管理的背景下，NoSQL数据库相较于传统的关系型数据库更适合处理高并发读写、海量数据处理、可用性和高扩展性的应用需求。本文目的在于针对NoSQL中表现优异的MongoDB在网络文学监测分析系统中展开研究与应用，使系统功能响应时间小于5秒。

1.4.2 研究范围

1. NoSQL技术综述：NoSQL非关系型数据库技术使用松祸合的数据模式，支持水平伸缩，拥有在磁盘和(或)内存中的数据持久化能力，支持多种“Non-SQL”接口来进行数据访问。NoSQL的数据模型包括Key-Value键值对、Key一结构化数据、Key一文档、列簇式存储和图结构存储。NoSQL支持复杂的查询、弱事务机制，支持冗余备份保证单机可靠性、多种数据同步方式实现多机可靠性，支持哈希分区和范围分区来进行分布式扩展，提供强一致性和最终一致。
2. MongoDB服务机制分析和应用优化：MongoDB是当前应用广泛的NoSQL数据库。本文全面研究了MongoDB应用及优化方法与原则：
3. 针对MongoDB面向文档的数据模型，提出反范式模式设计方法；
4. 根据MongoDB的自动分片机制，提出基于分区粒度、写扩展和查询隔离的分片键选择机制、分片集群部署策略和磁盘碎片化问题的解决方案；
5. 针对MongoDB基于简单模式分类、多计划并行的查询策略及其存在问题，提出了加入语义关联、细粒度模式分类的改进查询算法。
6. MongoDB应用：网络文学监测分析系统抓取存储多个大型小说网站的作者和作品信息，存储量大且增长快。本文根据网络文学监测分析系统的存储、查询、扩展以及复制冗余等需求，给出了MongoDB的解决方案，并与之前应用的Mysql数据库时的查询效率以及系统功能响应时间进行了对比。

1.5课题实施方案

本课题将采用先对系统优化总体规划，然后系统功能优化分模块实施，需求调研与应用开发紧密结合的技术实施方案。

1. 在海量数据情景下，对NoSQL数据库的应用进行充分的调研，分析并抽取完成系统优化必要流程和步骤，对系统优化进行可行性分析。
2. 学习和掌握实现网络文学监测分析系统优化所需的必要技术支持，指定系统优化实现的技术方案；
3. 找出实现系统优化所要解决的技术关键点，给出解决该技术难点的可行解决方案；
4. 完成系统优化总体设计，包括实现数据库的更换，数据库的设计与数据的迁移，提取系统功能等；
5. 完成系统优化详细设计，包括功能模块优化具体实施方案的选择；
6. 根据前期的总体设计和详细设计，逐步分模块实现系统功能优化；
7. 完善系统所有功能优化，并加强系统在功能、安全和网络适应性等方面的性能。

1.6 本章小结

本章对课题所要研究的基于NoSQL的网络文学监测分析系统优化进行了简要介绍，与使用传统关系型数据库相比较，给出了本课题的研究意义，并且介绍了非关系型数据库的研究现状，明确了本课题所要实现的项目目标和研究范围，给出系统优化的实施方案，为后期系统的优化明确了方向和思路。

1. 系统开发关键技术介绍

网络文学监测分析系统原本采用的是SSM框架开发，即Spring+SpringMVC+Mybatis的集成开发框架，也是目前主流的开发框架，相较于SSH更轻量级更易于搭建与开发。由于系统数据与日俱增，Mysql数据库性能直线下降，所以本文主要介绍MongoDB的shell命令、Spring与MongoDB的整合以及Java中如何操作MongoDB。

2.1 MongoDB的Shell命令

MongoDB简单操作（增删查改，正则表达式，索引）

MongoDB聚合操作（aggregate，mapreduce）

MongoDB之JavaScript函数（function（element））

2.2 Spring与MongoDB的整合

Xml方式

2.3 Java操作MongoDB

2.4 本章小结

1. 系统需求分析

3.1 系统介绍

3.2 系统优化可行性分析

3.3 系统面向的用户群体

3.4 系统中术语与缩写解释

3.5 系统优化功能性需求分析

3.6 系统非功能性需求分析

3.7 本章小结

1. 系统概要设计

4.1 系统运行环境

4.2 系统优化策略

4.3 系统结构设计

4.3.1 总体结构

4.3.2 功能模块

4.4 数据库设计

4.5 接口规范设计

4.6 本章小结

1. 系统优化详细设计与实现

5.1 技术难点解决

5.1.1 Spring与MongoDB的整合

5.2 系统模块设计与优化

5.2.1 文学作者统计分析模块

5.2.2 文学作品统计分析模块

5.2.3 文学网站统计模块

5.2.4 评论词云分析模块

5.3 本章小结

1. 系统测试

6.1 测试原则

6.2 测试环境

6.3 测试结果

6.3.1 单元测试

6.3.2 功能性测试

6.3.3 非功能性测试

6.4 本章小结

1. 参考文献

[1]沈姝. NoSQL数据库技术及其应用研究[D].南京信息工程大学,2012.

[2]刘卓. 基于NoSQL的空间数据云存储的研究[D].河南大学,2014.