大数据时代图书馆信息系统的系统分析与设计

曹祺 著

System Analysis and Design of Library Information System in the Era of Big Data

WUHAN UNIVERSITY PRESS
武汉大学出版社

大数据时代图书馆信息系统的系统分析与设计

曹祺 著

System Analysis and Design of Library Information System in the Era of Big Data

WUHAN UNIVERSITY PRESS
武汉大学出版社

图书在版编目(CIP)数据

大数据时代图书馆信息系统的系统分析与设计/曹祺著. —武汉：武汉大学出版社，2020.5

ISBN 978-7-307-21475-0

Ⅰ.大… Ⅱ.曹… Ⅲ.图书馆—信息系统—研究 Ⅳ.G250.7

中国版本图书馆 CIP 数据核字(2020)第 073334 号

责任编辑：宋丽娜　　责任校对：李孟潇　　整体设计：马　佳

出版发行：**武汉大学出版社**　(430072　武昌　珞珈山)

(电子邮箱：cbs22@whu.edu.cn　网址：www.wdp.com.cn)

印刷：北京虎彩文化传播有限公司

开本：787×1092　1/16　印张：10　字数：243 千字　插页：1

版次：2020 年 5 月第 1 版　2020 年 5 月第 1 次印刷

ISBN 978-7-307-21475-0　定价：35.00 元

目　录

1 研究背景

学术论文是科研工作者的原料和成果,科研活动的目的是创新。科研活动是科研工作者反复阅读相关学者的科研成果(即图书馆信息系统管理的"科技大数据"),复盘前人的科学实验,不断深入研究,发现新的现象和问题,并最终以论文的形式向同行分享。

图书馆是服务科研活动的重要机构之一,一方面要保存和馆藏论文,另一方面要运营一个高效率的图书馆信息系统(Library Information System,LIS),管理各类科技大数据,来服务科研工作者。

随着大数据时代的来临,传统的图书馆信息系统有很多不足。笔者结合自己长期以来在图书情报与档案管理学科的科研经历和在软件工程领域的工程经验,开发了一款面向大数据时代的图书馆信息系统,主要面向科研工作者,分享相关经验和成果。

1.1 大数据的发展现状

过去30年,大数据改变了人类的生活方式和工作方式,数据变得越来越多,每天有2.5EB数据产生①。过去30年,大数据领域的发展主要分为以下3个阶段。

第一阶段为结构化数据领域的发展。1974年,Donald D. Chamberlin 和 Raymond F. Boyce② 提出了结构化查询语言(Structure Query Language, SQL)标准③,标志着人类可以处理结构化数据,其理论主要为ACID理论,相关研究公司为美国微软公司、甲骨文公司。美国微软公司的系统为SQL Server,甲骨文公司的系统为Oracle和MySQL,以及在移动互联网领域广泛应用的系统SQLite④。但是这些数据库底层存储原理主要为文件系统,其查询原理主要为B树。

第二阶段为半结构化数据领域的发展。2004年,谷歌公司提出了BigTable架构⑤,标

---

① Ergüzen Atilla, Ünver Mahmut.Developing a File System Structure to Solve Healthy Big Data Storage and Archiving Problems Using a Distributed File System[J].Applied Sciences,2018,8(6).

② Donald D Chamberlin, Raymond F Boyce. SEQUEL: A Structured English Query Language [J]. Proceedings of 1974 ACM-SIGFIDET Workshop on Data Description. Access and Control 1974,2.

③ ISO/IEC 9075-2: 2011 Information Technology-Database Languages[S]. 2011.

④ Kang Woon-Hak, et al.X-FTL: transactional FTL for SQLite databases[J].Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. ACM, 2013.

⑤ Chang Fay, et al.Bigtable: A distributed storage system for structured data[J].ACM Transactions on Computer Systems(TOCS),2008,26(2).

志着 NoSQL(Not Only SQL,NoSQL)标准的提出,人类可以处理结构化数据,其理论主要为 CAP 理论,相关研究公司如美国谷歌公司。NoSQL 与 SQL 不同的是,其数据存储范式相对比较松散。其中行数据库如 MongoDB①,列数据库如 Apache Hbase②、Google BigTable,图数据库如 Neo4J③,哈希数据库如 Leveldb、Redis④,XML 数据库如 XMLDB⑤。其存储原理主要基于分布式文件系统,如 Apache Hadoop。但是查询原理有不同的实现方式,具体而言,树状查询如 MongoDB;全文检索如 Apache Lucene;实时检索如基于 Apache Lucene 的 Elastic Search⑥,其原理为在内存中保留一棵 LSM 树,动态实时建立 Lucene 索引,并定时替换索引;哈希查询如内存 NoSQL 数据库 Redis;Xpath 查询如 XMLDB。

第三阶段为非结构化数据领域的发展。最近 10 年,得益于 Geoffrey Hinton 在深度学习领域的研究⑦和 NVDIA GPU 硬件性能的提升,人类可以处理结构化数据并应用于产业,如 Google TensorFlow 引擎⑧。

近 5 年来,多个领域已经能分别处理结构化数据、半结构化数据和非结构化数据。但是产业的应用落后于科研的发展。其关键问题在于,尽管数据很多,但是缺乏个性化定制服务,而应用需要结合需求去定制。未来 30 年,大数据驱动下的个性化定制服务将越来越成熟,随着第五代移动通讯技术(5th-Generation,5G)通信技术,甚至第六代移动通讯技术(6th-Generation,6G)通信技术的发展,数据传输不会构成性能瓶颈。但是人类需要的不仅仅是大数据的简单查询,更多的是利用大数据提供服务。

在众多大数据服务中,最引人注意的是保存在各个科研机构图书馆的科技资源服务数据,国家科学技术部对此也发布了相关的指南并给予支持。

1.2 本书解决的问题

根据国家科学技术部为落实《国家中长期科学和技术发展规划纲要(2006—2020 年)》《国家创新驱动发展战略纲要》《国务院关于积极推进“互联网+”行动的指导意见》《国务院关于加快科技服务业发展的若干意见》《国家文化科技创新工程纲要》等提出的任务,国家

① Veronika Abramova,Jorge Bernardino.NoSQL databases: MongoDB vs Cassandra[J].Proceedings of the International C * Conference on Computer Science and Software Engineering. ACM, 2013.

② Vora Mehul Nalin. Hadoop-HBase for Large-Scale Data[J].2011 International Conference on Computer Science and Network Technology(ICCSNT), 2011,1.

③ Holzschuher Florian, René Peinl. Performance of Graph Query Languages: Comparison of Cypher, Gremlin and Native Access in Neo4j[J].Proceedings of the Joint EDBT/ICDT 2013 Workshops. ACM, 2013.

④ Carlson Josiah L. Redis in Action[J]. Manning Publications Co., 2013.

⑤ Al-Khalifa Shurug Cong Yu, H V Jagadish. Querying Structured Text in an XML Database[J]. Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data. ACM, 2003.

⑥ Gormley Clinton,Zachary Tong. Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine[J].O'Reilly Media, Inc., 2015.

⑦ LeCun Yann Yoshua Bengio, Geoffrey Hinton.Deep Learning[J]. Nature 2015(521): 436.

⑧ Abadi Martín et al. Tensorflow: A System for Large-Scale Machine Learning[J].OSDI,2016,16.

重点研发计划启动实施“现代服务业共性关键技术研发及应用示范”重点专项，并于2018年9月14日发布了《“现代服务业共性关键技术研发及应用示范”重点专项2018年度项目申报指南》（以下简称《指南》）。

《指南》“服务关键核心技术”中的“科技服务协同技术及平台研发（共性关键技术类）”为本研究指明了方向。《指南》指出：“研究跨平台科技资源与服务协同技术包括典型行业服务及资源模型与标准、跨平台服务业务流程及优化技术、跨平台服务描述/服务发现/服务选择/服务组合技术、大数据驱动的个性化定制服务以及服务价值链协同技术。”

本书所要解决的问题为《指南》“科技服务协同技术及平台研发”中的“大数据驱动的个性化定制服务”，主要针对图书馆管理的海量科技大数据，比如这些科技数据具备的行业新特性。

①海量异构数据源。在科技服务中，科技工作者申请一份专利，需要查询全球的专利数据库，如美国专利局、欧洲专利局、日本专利局和中国知识产权局。这些数据来自不同国家，数据语言、数据存储格式不同。海量的异构数据不适合进行结构化数据管理，比如MySQL单表存储容量上限仅仅为4GB。

②异构数据源的协同服务。发表专利不仅要查询海量异构数据，还要和其他数据进行协同，如专利查重需要和公开发表的论文进行对比，而论文专利的格式完全不同。常见的异构数据源协同一般采用无模式（Schemaless）的可扩展标记语言（Extensible Markup Language，XML）或者对象简谱（Java Script Object Notation，JSON）作为中间数据交换，而科技数据本身协同一般有行业文献管理格式，如RIS格式①，是具备科技数据行业特点的格式。科技数据经常需要和历史数据打交道，但是历史数据的管理系统和最新的系统不兼容，缺乏有效的集成开发环境（Integrated Development Environment，IDE）管理。目前在其他领域有微服务（Microservice）环境的成熟解决方案，如Docker②，但是专门针对科技数据服务行业的成熟解决方案却很少。

③数据的高频查询特点。科技数据服务，如查询论文，需要保持高频率的查询，以防止创新点失效。例如，通过SQL进行查询，SQL存储采用的是B树，而索引主要是B+树（如MyISAM索引）和B*树（如InnoDB索引），树的查询速度是O(log(N))。SQL查询中，关联表太多会导致查询速度极慢。另一方面，NoSQL中部分类别数据库采用哈希查询，查询速度是O(1)，但是此种模式不支持范围查询和排序，同时，大多数NoSQL不支持定式查询，如不支持CRON表达式③；

① RIS(file format).Wikipedia [DB/OL].(2018-08-03)[2019-02-02].

② Boettiger Carl.An Introduction to Docker for Reproducible Research[J].ACM SIGOPS Operating Systems Review,2015,49(1): 71-79.Bernstein David.Containers and Cloud: From lxc to Docker to Kubernetes[J].IEEE Cloud Computing,2014(3): 81-84.Anderson Charles.Docker [Software Engineering][J].IEEE Software,2015,32(3):102-103.

③ Lenz Moritz.Silent-Cron, a Cron Wrapper[J]. Perl 6 Fundamentals. Apress, Berkeley, CA, 2017:43-59.

④数据的深度分析。科技工作者使用科技数据往往要进行全文反复阅读,不会只停留在元数据的分析上,深度分析,如分析科技专利,需要反复地人机交互和询问专家①,并且还会根据过去几十年的论文数据去判断未来科研的方向②,或者根据当下的专利数据分析当下科技成果的转移情况③。但采用半结构化数据管理,如 NoSQL,因为数据缺乏统一的模式(Schema),往往会带来分析的不便。同时,NoSQL 分析数据需要编写代码,对非计算机专业的科技工作者来说,会带来使用难度。

⑤数据权限管理。科技数据往往来自不同实验室,因为不同实验室之间需要交换数据,但是各个实验室之间缺乏有效的互联互通,有效的数据权限管理机制缺乏导致很多实验室进行重复研究。

⑥个性化定制服务。科技数据的使用者每天使用大量的科技数据,缺乏个性化定制很难有效地管理数据。目前处理结构化数据的 SQL 和处理非结构化数据的 NoSQL,往往不是针对科技数据行业的,比如 SQL 通常采用正整数标识符作为主键,NoSQL 通常采用通用唯一识别码(Universally Unique Identifier,UUID)标识符作为主键。但在科技行业中,发表论文的唯一标识符的通用标准为 DOI④ 标准或者 Handle⑤ 标准。科技数据可以采用数字对象唯一标识符(Digital Object Unique Identifer,DOI)作为主键。因此,如果进行个性化定制,相关服务供应商需要编写大量程序代码。

基于以上国内外研究的不足,本书研究“大数据时代图书馆信息系统的系统分析与设计”,基于编译原理技术⑥,提出一种新的数据查询语言,该查询语言基于新的扩充巴科斯-瑙尔范式(ABNF)⑦,专门针对科技服务大数据。同时,为了提高项目的稳定性和扩展性,基于微服务(Microservice)的系统架构,开发了扩展程序。

根据中国科学技术协会的相关统计,中国有 1 亿名科研工作者,这些科研工作者日常需要和图书馆中的科研大数据打交道,本书通过优化图书馆信息系统,提高科研工作者使用科研数据的便利性和使用效率,服务中国的科研工作者。

---

① BUBELA T, GOLD E R, GRAFF G D, et al. Patent Landscaping for Life Sciences Innovation: Toward Consistent and Transparent Practices[J]. Nature Biotechnology, 2013, 31(3): 202-206.

② MUKHERJEE S, ROMERO D M, JONES B, et al. The Nearly Universal Link between the Age of Past Knowledge and Tomorrow's Breakthroughs in Science and Technology: the Hotspot[J]. Science Advances, 2017, 3(4): e1601315.

③ PADMANABHAN S, AMIN T, SAMPAT B, et al. Intellectual Property, Technology Transfer and Manufacture of Low-cost HPV Vaccines in India[J]. Nature Biotechnology, 2010, 28(7): 671-678.

④ Paskin Norman. Digital Object Identifier(DOI ®) System[J]. Encyclopedia of Library and Information Sciences 2010(3): 1586-1592.

⑤ Sun Sam Larry Lannom, Brian Boesch. Handle System Overview[Z]. No. RFC 3650, 2003.

⑥ Parr T. The Definitive ANTLR 4 Reference, Oreilly and Associate Series, Pragmatic Programmers, LLC, 2013.

⑦ Crocker Dave, Paul Overell. Augmented BNF for Syntax Specifications: ABNF[Z]. No. RFC 5234, 2008.

1.3 研究内容

本书针对科技服务大数据驱动的个性化定制服务，其研究内容主要包括以下几方面。

①数据存储技术，具体包含单机存储技术和分布式存储技术。单机存储技术主要研究存储过程的读写速度。分布式存储技术主要研究网络环境下多服务器之间的数据寻址，并且在数据存储技术研究过程中要紧密结合科技数据的特点。

②数据维护语言解释器技术，具体包含词法研究、语法研究、语义研究。词法研究主要针对个性化定制服务中的数据操作研究词法规则的设计。语法研究主要根据个性化定制服务中的不同模块和不同数据，从可扩展性、可维护性角度研究语法规则的设计。语义研究指根据科技数据个性化过程中不同的模块和接口，从执行速度、扩展性、操作体验研究语义规则的设计。

③数据维护语言代码执行技术，具体包含数据关键词查询技术、数据过滤查询技术、数据范围查询技术、数据排序技术和数据更新技术。数据关键词查询技术主要结合科技服务数据的特点，研究如何快速查询，同时研究性能瓶颈。数据过滤查询技术主要结合个性化定制服务的特点，研究过滤技术的算法。数据范围查询技术和数据排序技术主要结合科技数据的特点，从性能和可用性角度，设计查询算法和排序算法。数据更新技术主要考虑在单机存储和分布式存储等不同存储环境下，如何进行适配和性能优化。

④数据权限管理技术，具体包含单机权限管理技术和远程权限管理技术。单机权限管理技术主要研究本机和远程服务器的安全授权机制。远程权限管理技术主要研究不同数据服务商之间如何协作，并在为用户提供个性化定制服务过程中防止数据失控。

本书拟解决的问题主要包括以下关键科学问题。

①大数据时代图书馆信息系统的系统架构分析。针对科技服务海量数据的特点，研究新的数据存储系统和架构，研究数据的性能瓶颈，同时和旧系统兼容。通过研究数据存储原理、数据查询过程的机制，在不降低数据查询性能的前提下，基于微服务系统的架构，以期提高个性化定制的扩展效率。

②大数据时代图书馆信息系统的查询语言设计。针对个性化定制服务的需求，研究新的数据维护语言算法。针对科研工作者个性化定制服务的特点，研究数据维护语言的扩展机制。通过研究科技服务大数据的数据特点和服务机制、数据使用过程中的生命周期，基于研究数据查询语言的编译原理，设计新的领域特定语言(Domain Specific Language，DSL)，以期提高个性化定制的使用体验和查询效率。

③大数据时代图书馆信息系统的共享安全问题。安全问题既包含信息系统的程序安全，又包含共享过程中的知识产权安全。通过研究共享过程中科技大数据(主要是学术论文的文献数据)的访问机制，基于区块链技术和浏览器扩展技术设计相关程序，以期达到不同角色的使用方便和共享安全。

④大数据时代图书馆信息系统的可视化分析。利用图书馆信息系统导出的数据进行元分析，来进行学科预测等服务。图书馆信息系统并不是孤立的系统，和第三方工具进行整合才能更好地服务科研工作者。对图书馆信息系统设计导出数据接口，通过接口，使导出的数

据能够被常用的情报分析与计量工具直接使用，这些计量工具包括 CiteSpace①、VOSViewer② 等。同时，还基于长尾理论，通过数据库跨库对比，研究图书馆数据的长尾关键词分布。

另外，本书还研究了最近三年（2018—2020 年）国外流行的 FOLIO 系统，研究了 FOLIO 系统的架构，发现采用 FOLIO 系统后能更好地管理图书馆的信息。

本书分为 4 个子任务进行研究，主要的研究方法和技术路线如下。

子任务①：大数据时代图书馆信息系统的系统架构。

步骤 1：根据科技服务数据的特点，设计单机存储系统，通过不断的数据分层技术，实现单机的图书馆信息系统的架构。

步骤 2：利用微服务架构进行多数据节点的管理，设计微服务的通信技术，保证图书馆信息系统维护的可扩展性、信息服务的容错性。

子任务②：大数据时代图书馆信息系统的查询语言。

步骤 1：根据编译原理设计词法规则，借鉴结构化查询语言的词法规则进行扩展，设计新的数据领域语言。

步骤 2：根据编译原理设计语法规则，考虑半结构化数据的处理机制及非结构化数据处理模型的整合，同时考虑个性化需求的扩展性，在语法层进行扩展机制设计。

步骤 3：根据编译原理设计语义规则，考虑语义执行的性能，面对不断增长和变化的个性化需求，考虑语义执行时的版本兼容性问题。

子任务③：大数据时代图书馆信息系统的共享机制。

步骤 1：从知识产权保护和技术实现角度研究国外同行的共享机制。

步骤 2：参考行业研究成果，基于区块链技术开发相关程序，进行技术实践。

步骤 3：参考行业研究成果，基于浏览器技术开发相关程序，进行技术实践。

子任务④：大数据时代图书馆信息系统的数可视化分析实践。

步骤 1：利用本系统数据接口导出的数据，通过整合第三方科学计量工具，即 VOSViewer，进行情报学领域的学科分析实践。

步骤 2：利用本系统数据接口导出的数据，基于 t-SNE 算法，对“双一流”大学自然科学基金的数据进行可视化分析建模实践。

步骤 3：利用南京大学 CSSCI 数据库和知网 CNKI 数据库进行跨库对比研究，基于长尾理论，研究图书馆关键词的分布。

① http://cluster.ischool.drexel.edu/~cchen/citespace/.

② https://www.vosviewer.com/.

2 大数据时代图书馆信息系统的系统架构

2.1 图书馆信息系统的需求用例分析

2.1.1 领域热点分析需求用例

领域热点分析需求模块主要包含以下用例(Use Case),如图 2-1 所示。

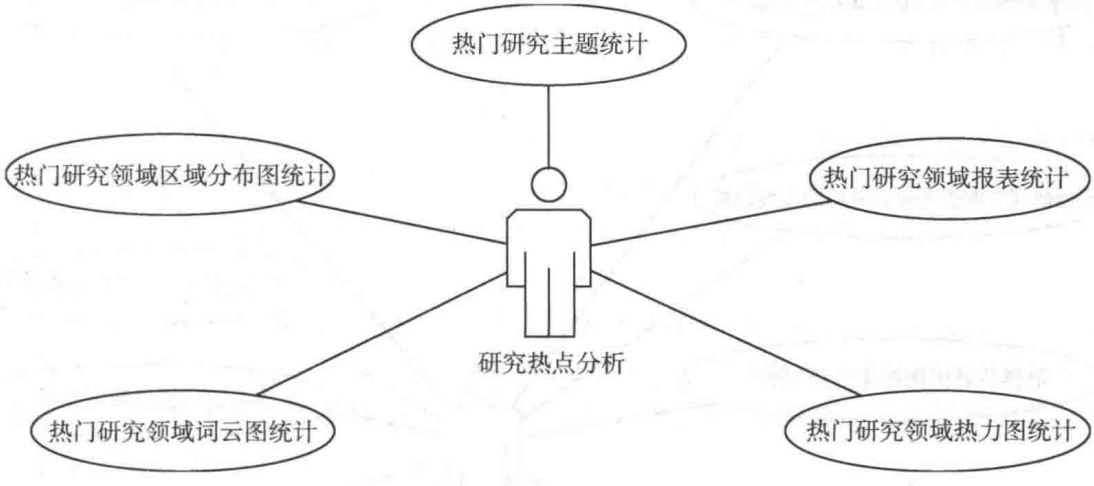

图 2-1 领域热点分析需求用例图

上述用例图主要包含以下用例。

①热门研究主题统计。该模块的主要功能是为用户提供热门研究领域关键词统计分析。目前,该模块根据用户选择的学科领域,由查询系统提前计算好该学科领域研究热点的关键词列表,更好地满足用户跟踪该学科的研究热点方向。

②热门研究领域报表统计。该模块的主要功能是为用户提供研究热点发现与分析。用户在使用本数据库系统的时候,通过系统的导航功能,根据系统的提示信息,输入相关的数据,查询分析数据库,得到研究热点的分析汇总结果。科研热点的分析汇总结果可以以多种形式向用户进行展示。

③热门研究领域热力图统计。该模块的主要功能是以热力图的形式,为用户提供热门研究领域关键词统计分析。用户在上述系统的输入部分选择学科或者关键词进行输入后,根据

输入的相关数据,查询分析数据库,以热力图的形式将统计结果输出呈现给用户。热力图呈现的是研究主题与研究热门程度的映射关系。

④热门研究领域区域分布图统计。该模块的主要功能是为用户提供热门研究领域关键词区域统计分析。目前,该模块根据用户在上述系统的输入部分选择学科或者关键词进行输入后,根据用户输入的相关数据,查询分析数据库,以地域分布图的形式将统计结果输出呈现给用户。地域分布图呈现的是研究主题与研究地区分布的映射关系。

⑤热门研究领域词云图统计。该模块的主要功能是为用户提供热门研究领域关键词统计分析。目前,该模块根据用户在上述系统的输入部分选择学科或者关键词进行输入后,根据用户输入的相关数据,查询分析数据库,以词云图的形式将统计结果输出呈现给用户。词云图呈现的是研究主题关键词和热门高频词汇的一种混合动态,给用户直观的视觉感受。

2.1.2 科研竞争力分析需求用例

科研竞争力分析需求模块主要包含以下用例,如图 2-2 所示。

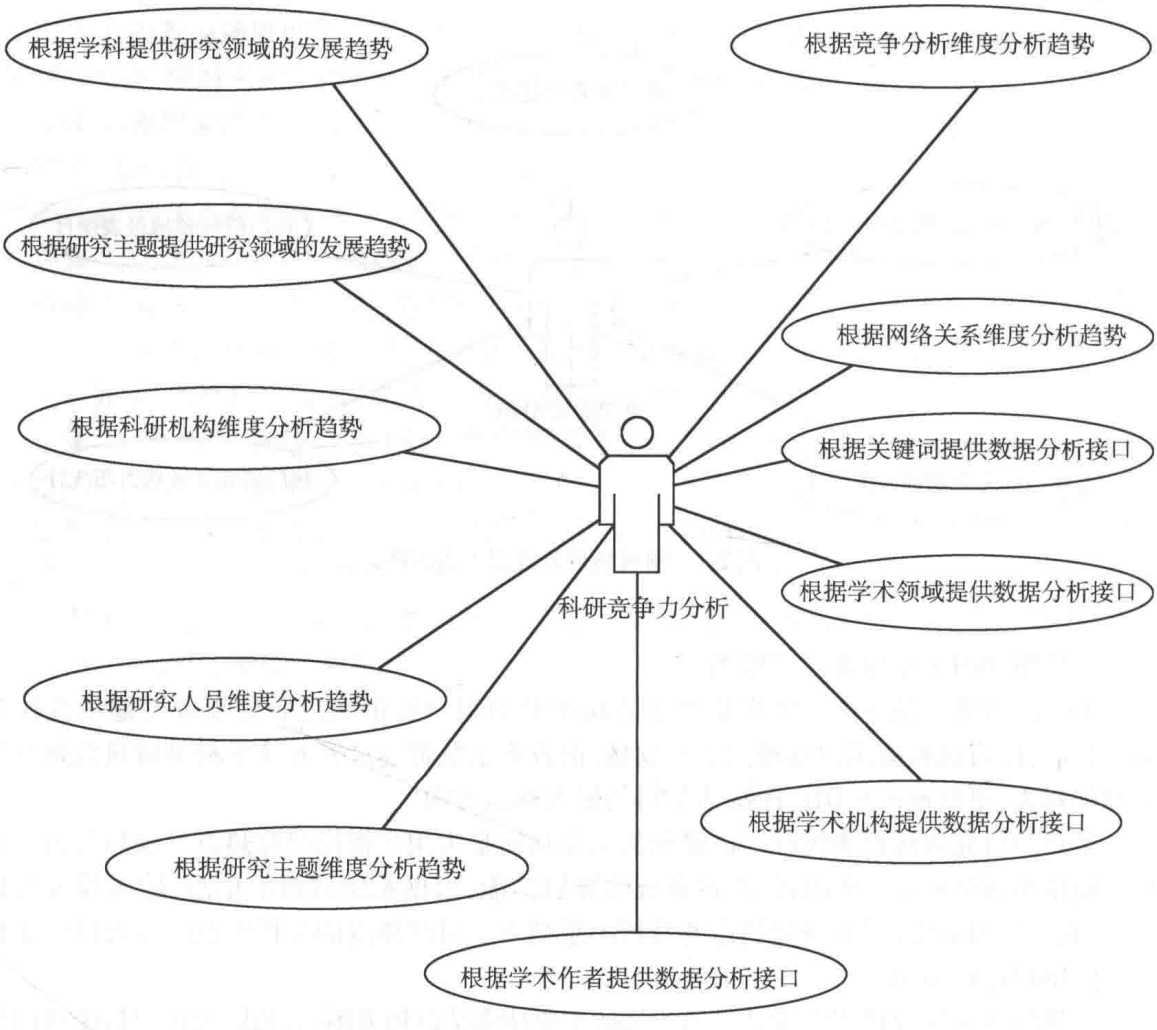

图 2-2 科研竞争力分析需求用例图

上述用例图主要包含以下用例。

①根据学科提供研究领域的发展趋势。该模块的主要功能是为用户提供研究趋势的统计分析。目前,该模块根据用户选择的学科领域,输入提前计算好的该学科字典作为关键词,系统在后台使用 SQL 分析与查询数据库,查询的论文结果数据集以数据报表的形式返给用户。

②根据研究主题提供研究领域的发展趋势。该模块的主要功能是为用户提供不同研究主题研究领域的发展趋势。目前,该模块根据用户选择的研究主题,输入提前计算好的主题列表作为关键词,系统在后台使用 SQL 分析与查询数据库,查询的论文结果数据集以数据报表的形式返给用户。

③根据科研机构维度分析趋势。该模块的主要功能是为用户提供基于研究机构数据的统计分析。目前,该模块根据用户选择的科研机构,输入提前计算好的科研机构作为关键词,系统在后台使用 SQL 分析与查询数据库,查询的论文结果数据集以数据报表的形式返给用户。

④根据研究人员维度分析趋势。该模块的主要功能是为用户提供基于研究人员数据的统计分析。目前,该模块根据用户选择的研究人员,输入提前计算好的研究人员作为关键词,系统在后台使用 SQL 分析与查询数据库,查询的论文结果数据集以数据报表的形式返给用户。

⑤根据研究主题维度分析趋势。该模块的主要功能是为用户提供基于研究主题数据的统计分析。目前,该模块根据用户选择的科研主题,输入提前计算好的科研主题作为关键词,系统在后台使用 SQL 分析与查询数据库,查询的论文结果数据集以数据报表的形式返给用户。

⑥根据竞争分析维度分析趋势。该模块的主要功能是为用户提供基于竞争分析数据的统计分析。目前,该模块根据用户选择的需要对比分析的研究人员,输入提前计算好的研究人员作为关键词,系统在后台使用 SQL 分析与查询数据库,查询的论文结果数据集以数据报表的形式返给用户。

⑦根据网络关系维度分析趋势。该模块的主要功能为用户提供基于网络关系分析数据的统计分析。目前,该模块根据用户选择的需要对比分析关系网络的研究人员,输入提前计算好的研究人员作为关键词,系统在后台使用 SQL 分析与查询数据库,查询的论文结果数据集以数据报表的形式返给用户。

⑧根据关键词提供数据分析接口。该模块的主要功能是为用户提供基于关键词统计分析的接口。目前,该模块根据用户选择查询的关键词,系统在后台使用 SQL 查询数据库中的关联论文数据,查询的论文结果数据集以 CSV 等格式文件返回结果集,提供给用户下载。

⑨根据学术领域提供数据分析接口。该模块的主要功能为用户提供基于学科领域统计分析的接口。目前,该模块根据用户选择查询的领域关键词,系统在后台使用 SQL 查询数据库中的关联论文数据,查询的论文结果数据集以 CSV 等格式文件返回结果集,提供给用户下载。

⑩根据学术机构提供数据分析接口。该模块的主要功能是为用户提供基于学术机构关键词统计分析的接口。目前,该模块根据用户选择查询的学术机构关键词,系统在后台使用SQL查询数据库中的关联论文数据,查询的论文结果数据集以CSV等格式文件返回结果集,提供给用户下载。

⑪根据学术作者提供数据分析接口。该模块的主要功能是为用户提供基于学术作者关键词统计分析的接口。目前,该模块根据用户选择查询的作者关键词,系统在后台使用SQL查询数据库中的关联论文数据,查询的论文结果数据集以CSV等格式文件返回结果集,提供给用户下载。

2.1.3 学术产出分析需求用例

学术产出分析需求模块主要包含以下用例,如图2-3所示。

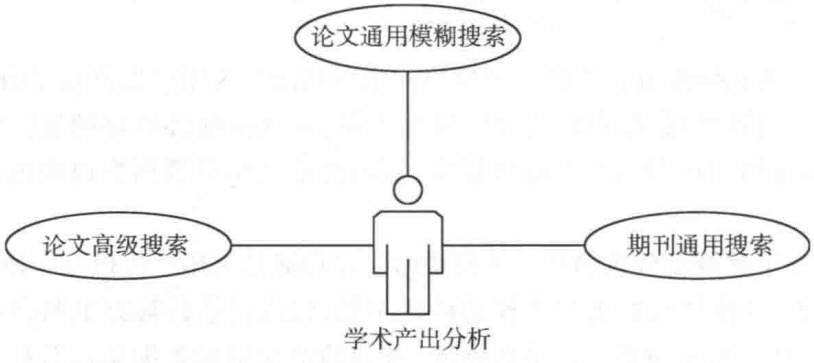

图2-3 学术产出分析需求用例图

上述用例图主要包含以下用例。

①论文通用模糊搜索功能。该模块中,用户在使用学术产出数据库时,首先在系统中对应的论文搜索功能模块处,输入查询的搜索关键词,确认提交后,系统会在后台数据库中检索海量的与所选择关键词相关的论文数据,通过计算以及汇总,分析得到与所选相关的论文数据集结果。搜索结果中将高亮显示查询得到的论文标题中包含的关键词。后台搜索将使用Sphinx搜索引擎作为底层服务。

②论文高级搜索功能。该模块中,用户在使用学术产出数据库时,首先在系统中对应的论文搜索功能模块处,输入查询的高级搜索表达式,确认提交后,系统会在后台数据库中检索海量的所选择表达式,查询对应的论文数据,通过计算以及汇总,分析得到与所选相关的论文数据集结果。搜索结果中将高亮显示查询得到的论文元数据中包含的关键词。后台搜索将使用Sphinx搜索引擎作为底层服务。

③期刊通用搜索功能。该模块中,用户在使用学术产出数据库时,首先在系统中对应的期刊搜索功能模块处,输入查询的搜索关键词,确认提交后,系统会在后台数据库中检索海量的与所选择关键词相关的期刊数据,通过计算以及汇总,分析得到与所选相关的期刊数据集结果。搜索结果中将高亮显示查询得到的期刊名称中包含的关键词。后台搜索将使用

Sphinx 搜索引擎作为底层服务。

2.1.4 图书信息数据集成需求用例

图书信息数据集成需求模块主要包含以下用例，如图 2-4 所示。

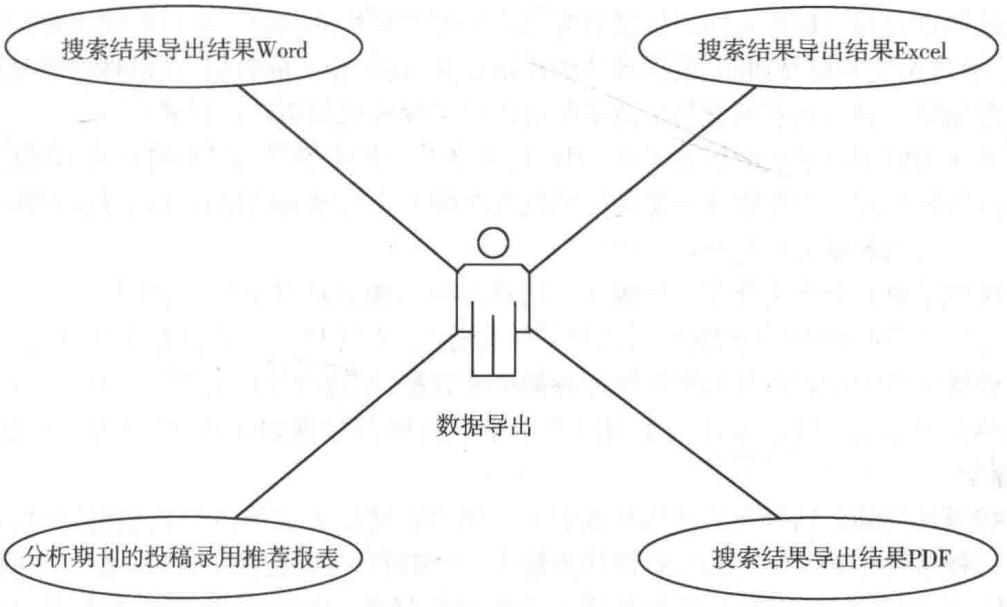

图 2-4 图书信息数据集成的需求用例图

上述用例图主要包含以下用例。

①搜索结果导出结果 Excel。该模块的主要功能是为用户提供搜索论文结果数据集导出为 Excel 格式。根据用户选择查询的关键词，该模块系统在后台使用 SQL 查询数据库中的关联论文数据，查询的论文结果数据集以 Excel 格式返回结果集，提供给用户下载。

②搜索结果导出结果 PDF。该模块的主要功能是为用户提供搜索论文结果数据集导出为 PDF 格式。根据用户选择查询的关键词，该模块系统在后台使用 SQL 查询数据库中的关联论文数据，查询的论文结果数据集以 PDF 格式返回结果集，提供给用户下载。

③搜索结果导出结果 Word。该模块的主要功能是为用户提供搜索论文结果数据集导出为 Word 格式。目前，根据用户选择查询的关键词，该模块系统在后台使用 SQL 查询数据库中的关联论文数据，查询的论文结果数据集以 Word 格式返回结果集，提供给用户下载。

④分析期刊的投稿录用推荐报表。该模块的主要功能是为用户提供搜索论文和期刊的结果数据集导出影响因子报表。目前，根据用户选择查询的关键词，该模块系统在后台使用 SQL 查询数据库中的关联论文和期刊数据，查询的论文和期刊以及影响因子结果数据集以报表的形式返回结果集，提供给用户查看使用。