|  |
| --- |
| 武汉大学logo  **研究生学位论文** |
| 题 目： 基于图数据库的研究团体搜索系统设计与实现  学 院： 计算机学院  专 业： 计算机科学与技术  学 号： 2019282110194  姓 名： 陈小龙  导师姓名： 祝园园  导师职称： 副教授 | |

目 录

[论文原创性声明 3](#_Toc64897924)

[摘要 3](#_Toc64897925)

[Abstract 3](#_Toc64897926)

[1绪论 4](#_Toc64897927)

[1.1 研究背景 4](#_Toc64897928)

[1.2 国内外研究现状 4](#_Toc64897929)

[1.3 本文主要工作 4](#_Toc64897930)

[1.4 论文组织结构 4](#_Toc64897931)

[2 相关技术与概念 4](#_Toc64897932)

[2.1 MVC 4](#_Toc64897933)

[2.2 Restful 4](#_Toc64897934)

[2.3 社区搜索 4](#_Toc64897935)

[2.4 NoSql 4](#_Toc64897936)

[2.5 Vue 5](#_Toc64897937)

[3 研究团体搜索系统分析与设计 5](#_Toc64897938)

[3.1 系统需求分析 5](#_Toc64897939)

[3.2 系统整体框架设计 5](#_Toc64897940)

[3.3 社区搜索 5](#_Toc64897941)

[3.4 API模块 5](#_Toc64897942)

[3.5 接口设计 5](#_Toc64897943)

[3.6 Unmanaged 5](#_Toc64897944)

[3.5 Produce 5](#_Toc64897945)

[4 Neo4j算法扩展 5](#_Toc64897946)

[4.1 Equitruss 5](#_Toc64897947)

[4.2 Srimrank 5](#_Toc64897948)

[5 系统实现与结果分析 5](#_Toc64897949)

[5.1 数据集 5](#_Toc64897950)

[5.2 实验环境搭建 5](#_Toc64897951)

[5.3 实验与结果分析 5](#_Toc64897952)

[5.3.1 功能试验与结果分析 5](#_Toc64897953)

[6 结果与展望 6](#_Toc64897954)

[6.1 工作总结 6](#_Toc64897955)

[6.2 未来工作展望 6](#_Toc64897956)

[参考文献 6](#_Toc64897957)

[致谢 7](#_Toc64897958)

# 论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下独立进行研究工作所取得的研究成果。除文中已经标明引用的内容外，本论文不包含任何其他人或集体已发表或撰写的研究成果。对本章的研究做出贡献的个人和集体，均已在文中以明确的方式表明。本声明的法律结果由本人承担。

# 摘要

互联网的高速发展使得“地球村”成为现实，世界各地高校、研究机构以及不同学者间的沟通交流与合作越来越频繁与紧密。他们之间交流合作的成果往往以学术论文的形式发布，并且根据交流合作的密切程度，不同的学者之间会形成一个个的研究团体。

同时，学术文章与文章的创作者之间的关系可以用图（网络）来描述：文章和作者作为图中的顶点，不同作者的合作关系以及文章和作者的从属关系作为图中顶点的边。那么如何从这个网络中方便的检索出作者信息、文章信息、作者所属研究团体信息以及作者之间的相似度等成为了现实的问题。然而，当前主流的搜索引擎，无法很好的解决上述问题。在此背景下，本文主要阐述基于DBLP数据，使用图数据库Neo4j[1]作为数据存储工具；利用“社区搜索”相关算法以及Web开发技术，设计并构建一个信息检索系统，用于解决上述研究团体检索及其关联信息查询等需求。

本文还针对系统不同功能进行测试，结果表明，研究团体检索系统不仅在搜索结果展示有简洁流畅的用户体验；同时，对于文章、作者等信息检索、使用不同算法实现研究团体搜索等功能有不错的表现。该系统为探寻学术研究的前沿方向、学术研究团体提供了一个有效的平台。

关键词：DBLP；社区搜索；Neo4j

# **Abstract**

The rapid development of the Internet has made the "global village" a reality, and the communication and cooperation between universities, research institutions and different scholars around the world have become more frequent and close. The results of exchanges and cooperation between them are often published in the form of academic papers, and according to the close degree of exchanges and cooperation, different scholars will form a research group.

At the same time, the relationship between an academic article and the author of an article can be described by a graph (network): the article and the author are the vertices in the graph, and the cooperation relationship between different authors and the affiliation between the article and the author are the edges of the vertices in the graph. So how to easily retrieve author information, article information, author's research group information, and the similarity between authors from this network has become a real problem. However, the current mainstream search engines cannot solve the above problems well. In this context, this article mainly elaborates on DBLP data, using graph database Neo4j as a data storage tool; using "community search" related algorithms and Web development technology to design and build an information retrieval system to solve the above research group retrieval and its Related information query and other requirements.

This article also tests different functions of the system. The results show that the research group retrieval system not only has a concise and smooth user experience in the display of search results; at the same time, it is good for searching articles, authors and other information, using different algorithms to achieve research group search and other functions. which performed. This system provides an effective platform for exploring the frontier direction of academic research and academic research groups.

**Key Words:** DBLP;Community Search;Neo4j

# 1绪论

## 研究背景及意义

随着物联网、社交网络、云计算等技术的高速发展，互联网技术不断的改变着每一个人的生活。同时，计算能力、存储空间、网络带宽的发展更是日新月异。这些变化大大加快了各行各业积累数据的速度，数据量更是以PB计。与此同时，在新增的海量的数据中，传统的结构化数据只占一小部分，绝大部分是非结构化的数据。典型的非关系型数据包括文本、图片、音频、视频、XML、HTML、图数据等，而传统关系型数据库的处理数据的模式难以满足非结构化数据的处理要求。因此，如何从海量非结构化数据中提取需要的信息已经成为极具价值的热点问题。

就非结构化数据的存储而言，主流的工具是各种NoSql数据库。NoSql数据库可以按数据模型分成4类：键值存储库（如：Redis[2]）、列族数据库（如：Hbase[3]）、文档数据库（如：MongoDB[4]）、图形数据库（如：Neo4j）。其中图形数据库专门用来处理高度关联的数据，适用于推荐系统、社交网络、模式识别等领域。图形数据库具有高度的灵活性，并且支持复杂的图形算法，方便构建复杂的关系图。

计算机科学文献库DBLP(DataBase systems and Logic Programming)，提供计算机科学领域的文献检索服务。它收录了国际期刊和会议等公开发表的论文，并且文献质量较高，更新的速度快，能够反应国外学术研究的前沿方向。DBLP使用典型的非结构化XML数据结构存储标题、作者、发表日期等元数据。DBLP作为文献集成的数据库系统，只能满足以作者为中心的信息检索需求（如：搜索作者、搜索作者所有论文、获取论文详细信息等），无法满足诸如研究团体搜索、作者相似度查询等需求。

本系统以DBLP的XML数据集为基础进行数据建模，将原数据构造为包含作者和文献两类顶点的图，并使用图数据库Neo4j作为存储数据库。另外，本系统通过集成不同的图的算法（比如：中心性算法、社区检测算法），来提供更加高效和传统信息检索无法实现的信息搜索服务。该系统实现的功能对提取文献库更高层次的信息和充分利用现有文献库数据具有重要的意义。

## 国内外研究现状

* + 1. 图论的发展

## 本文主要工作

## 1.4 论文组织结构

# 2 相关技术与概念

## 2.1 MVC

## 2.2 Restful

## 2.3 社区搜索

## 2.4 NoSql

## 2.5 Vue

# 3 研究团体搜索系统分析与设计

# 

## 3.1 系统需求分析

## 3.2 系统整体框架设计

## 3.3 社区搜索

## 3.4 API模块

## 3.5 接口设计

## 3.6 Unmanaged

## 3.5 Produce

# 4 Neo4j算法扩展

# 

## 4.1 Equitruss

## 4.2 Srimrank

# 5 系统实现与结果分析

# 

## 5.1 数据集

## 5.2 实验环境搭建

## 5.3 实验与结果分析

## 5.3.1 功能试验与结果分析

5.3.2 性能试验与结果分析

# 6 结果与展望

# 

## 6.1 工作总结

## 6.2 未来工作展望

# 

# 参考文献

[1]Neo4j.Neo4j官网[EB/OL].https://neo4j.com/,2020.

[2]Redis Labs.Redis官网[EB/OL].https://redis.io/,2015-06.

[3]Apache Hbase.The Apache Software Foundation[EB/OL].https://hbase.apache.org/,2007.

[4]MongoDB.MongoDB简介[EB/OL].https://www.mongodb.com/cn,2021.

# 致谢

感谢CCTV