CS 领域论文的可视化分析系统

25 组: NN Explorer (王为 2311605、苏威远 2311456、李泽椹 2311474)

目录

一、 分析目标

作为计算机专业的学生，在如今的信息时代，我们小组希望探究计算机科学（CS）各领域的发展态势。通过分析近三十年来（1990-2024 年）CS 领域的论文数据，我们的具体目标如下：

(1) 探究各个领域的发展态势；
(2) 研究当下的热门技术方向；
(3) 探究各个具体领域之间的关系；
(4) 分析 CS 领域发展的潜在规律；
(5) 探究某领域研究的方式和对象。

二、 我们的优点

本项目旨在构建一个交互性强、功能丰富且易于扩展的 CS 领域论文可视化分析系统。我们的程序设计以简洁易懂为核心，注重模块化和高效性，具体设计思路如下：

- **模块化设计：**系统采用模块化设计，将不同功能划分为多个子模块，便于理解和维护。主要模块包括数据存储、数据预处理、服务器响应、前端页面绘制以及高复杂度任务处理模块。我们使用 d3.dispatch 进行事件调度和模块间通信，在保证封装性的同时完成模块间交互。
- **前后端分离：**前端使用 HTML、CSS 和 JavaScript 实现页面设计与交互，后端通过 ‘server.js‘ 文件响应浏览器请求。前端与后端通过 HTTP 请求通信，确保页面动态更新和用户交互的响应速度。
- **高复杂度任务处理：**为提升复杂任务的效率，系统设计了专门的 Python 脚本，处理如大规模数据的分析和计算等任务。前端通过 JavaScript 向服务器发送请求，由 Python 脚本完成计算后返回结果。
- **交互性与可扩展性：**系统设计了丰富的交互功能，用户可通过总览页面、词云页面以及论文分析页面探索感兴趣的主题。系统具备良好的扩展性，用户只需重新运行数据预处理脚本，即可更新数据以适应新论文的发表。
- **数据权威性与可靠性：**系统使用 arXiv 官方 API 获取数据，确保数据的权威性和可靠性。同时，针对超过七万篇论文的分析由 Python 程序完成，耗时约 40 小时，确保数据处理的深度与准确性。
- **封装性与可维护性：**系统对复杂页面进行了模块化封装，分文件绘制，便于用户理解和修改。同时，优秀的封装性使得系统维护更加高效。

三、 设计流程

3.1 数据处理

- **数据源**：来自 arXiv 官网中 Computer Science 领域的论文数据，涵盖 1990 年至 2024 年的约 7 万篇论文。
- **数据处理流程**：原始数据经过清洗、格式化、分类处理后，提取了年份、领域、关键词等核心信息，形成适合可视化分析的数据集。

3.1.1 数据抓取与存储

为了从 arXiv 平台抓取计算机科学（CS）领域的论文数据，我们使用了 `requests` 库向 arXiv API 发送 HTTP 请求，并通过查询条件（如子类别、提交时间范围等）限制抓取范围。使用 `xml.etree.ElementTree` 库解析 API 返回的 XML 数据，提取论文的标题、作者、摘要、发表时间、链接和分类标签等信息。抓取的数据按年份和子类别组织，并通过 `json` 库存储为本地文件。我们解决了以下困难：

- 网络请求超时与限流：通过设置超时时间和加入延时机制（`time.sleep`），避免请求失败和服务器限流。
- 数据去重：通过论文链接去重，确保存储的数据唯一性。

3.1.2 按年份和子类别统计论文数量与总量

为了统计每年每个子类别的论文数量以及 CS 领域整体的论文总量，我们使用了 `json` 库读取数据，并通过 `defaultdict` 高效统计。对于每篇论文，我们解析其发表年份并按子类别进行分组，最终生成了按年份和子类别组织的论文数量数据，达到了展示子类别发展趋势及整体发展态势的效果。

3.1.3 子类别共现次数统计

为了量化各子类别之间的关联性，我们使用了 `itertools.combinations` 生成子类别对，并统计它们在同一篇论文中的共现次数。通过高效的数据结构存储共现关系，达到了展示子类别关系强度的效果。

3.1.4 停用词列表的定义与扩展

为了提高文本分析的准确性，我们使用了 `nltk` 库下载基础停用词列表，并结合 CS 领域特点扩展了领域特定停用词。通过去除无意义词汇，显著增强了关键词提取的效果。

图 1　程序运行截图

3.1.5 论文 PDF 文件的解析与信息提取

为了提取论文的关键信息（如页数、图片数量、文本内容等），我们使用了 `requests` 库批量下载 PDF 文件，`fitz`（`PyMuPDF`）库解析 PDF 内容，并通过 `tqdm` 显示处理进度，最终达到了高效批量处理 PDF 文件的效果。由于论文数量巨大，该过程持续约 40h.

3.1.6 模块出现频率统计与平均值计算

为了统计各子类别中各标准模块（如“引言”、“方法”等）的出现频率，以及论文的平均属性（如平均词数、图片数、页数等），我们使用了 `json` 库读取数据，通过 `defaultdict` 统计分布与均值，最终达到了展示模块分布规律和论文质量特征的效果。

四、 设计介绍

4.1 主视图系统设计总览

我们设计了一套简洁易懂、交互性强的可视化系统，旨在帮助用户分析 CS 领域的发展态势。图2和3是主界面的视图结构总览：

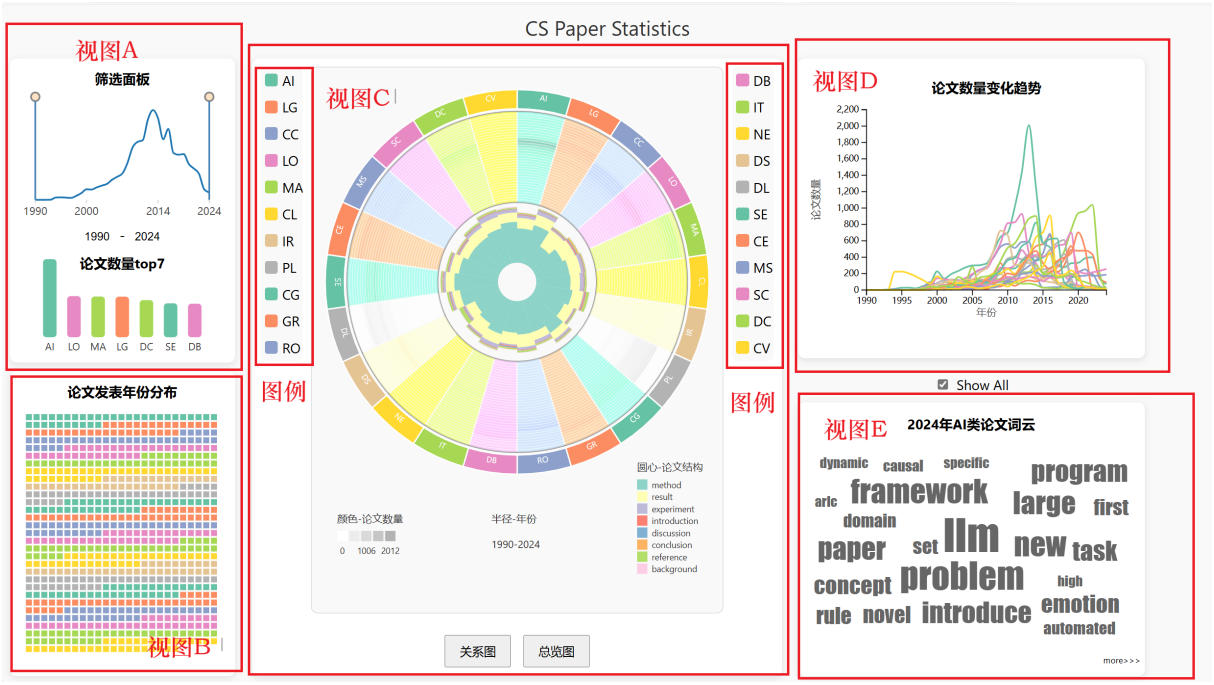

图 2　视图总览 1

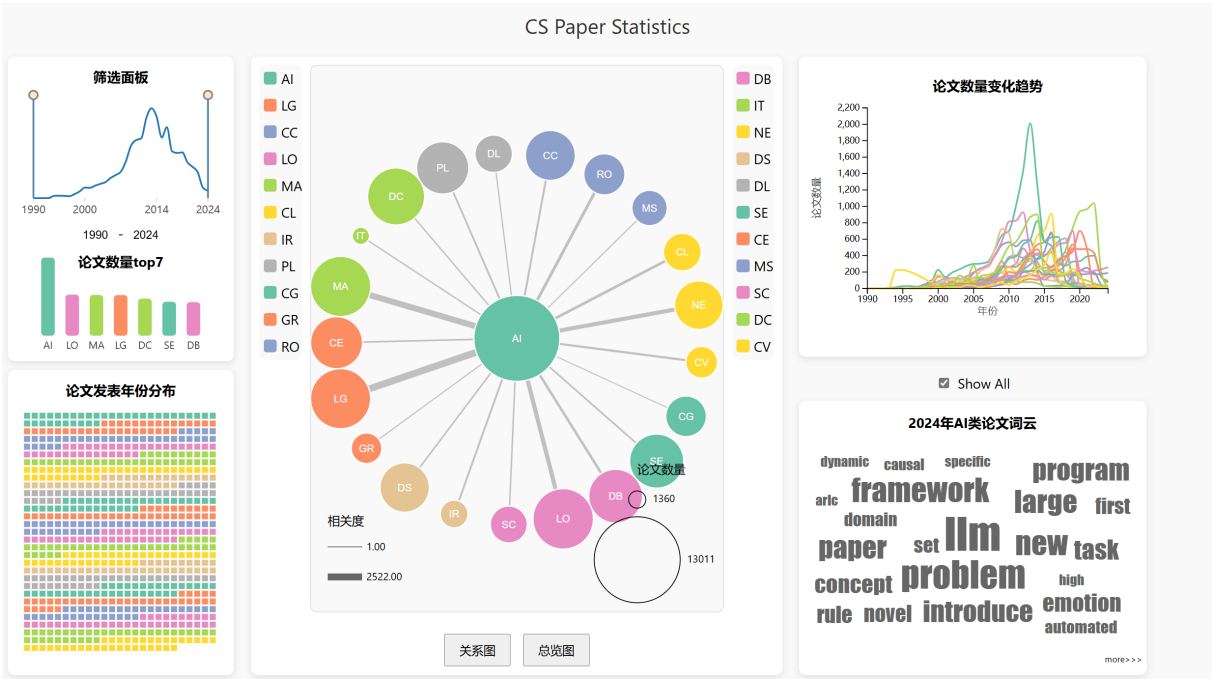

图 3　视图总览 2

(1) **视图 A**：年份范围选择与领域论文数量排行；
(2) **视图 B**：论文分布情况图；
(3) **视图 C**：领域关系图和领域总览图；
(4) **视图 D**：领域论文数量趋势图；
(5) **视图 E**：关键词词云。

4.1.1 视图 A 部分

- **上半部分**：一个年份范围选择的筛选面板，可以通过拖拽两个圆形按钮或者在输入框中输入起始和终止年份来选择年份范围，但是我们将其做成了每年度总论文数量的趋势图，这样可以让用户对整个计算机领域的发展有一个认知，并且可以据此趋势选择年份范围以探究论文数量发表高峰期的相关结论
- **下半部分**：根据所选年份各个具体领域的论文数量排行（top7），以探究在此时间段内哪些领域比较热门，其中颜色映射的是具体的领域，高度映射的是其数量。

4.1.2 视图 B 部分

论文发表年份分布情况图，每个颜色映射不同的领域，相同颜色按照顺序分别是从 1990 到 2024 年的领域的表示。有透明度高低之分（具体在交互设计中介绍）。

4.1.3 视图 C 部分

视图 C 中包含两个部分，用户可以通过点击“关系图”或“总览图”按钮在相应页面之间切换。

关系图部分：

- 每个节点的颜色表示一个领域；
- 节点的面积与所选年份范围内的论文数量成正比；
- 中间节点与其他领域之间连线的粗细表示其关联程度；
- 右下角展示两个圆形，分别代表论文数量最少和最多的领域节点大小，帮助用户理解节点大小与论文数量的关系。

总览图部分：

总览图部分提供了对各领域的整体概览，帮助用户直观地了解不同领域在各年份的论文数量分布和结构特点。视图的具体特征如下：

- 展示所有计算机科学子领域的论文数量，通过颜色和大小直观呈现各领域的相对规模。
- 右下角的图例展示了颜色与论文数量的关系，帮助用户理解颜色深浅与论文数量的对应关系。
- 图中圆心部分的颜色块表示论文的结构模块，如“方法”、“结果”等，帮助分析论文的组成特点。
- 半径的变化代表不同年份，图例中标明了年份范围，方便用户理解时间轴的分布。

4.1.4 图例部分

这是每个领域所映射的颜色的图例，全局代表各个领域的颜色都以此为映射标准，为了防止用户不了解各个领域的缩写，将鼠标移动至具体的图例上时会显示对应的中文名称。

4.1.5 视图 D 部分

根据所选年份范围，绘制所选领域论文数量的时间变化趋势图，便于用户分析领域的长期发展态势。

4.1.6 视图 E 部分

- 根据视图 B 中选定的领域和年份，生成对应的关键词词云，关键词颜色为灰色。
- 右下角 “More” 按钮可进入词云主界面：1) 可选择年份和领域，显示对应关键词；2) 点击关键词后，右侧显示关键词在近三十年的出现频率趋势以及占比。

4.2 词云视图设计

在视图上方可以选择年份和领域，会显示对应的关键词云，点击关键词会在右侧上面显示这个词在近三十年出现的频率的趋势，下面显示关键词的占比。

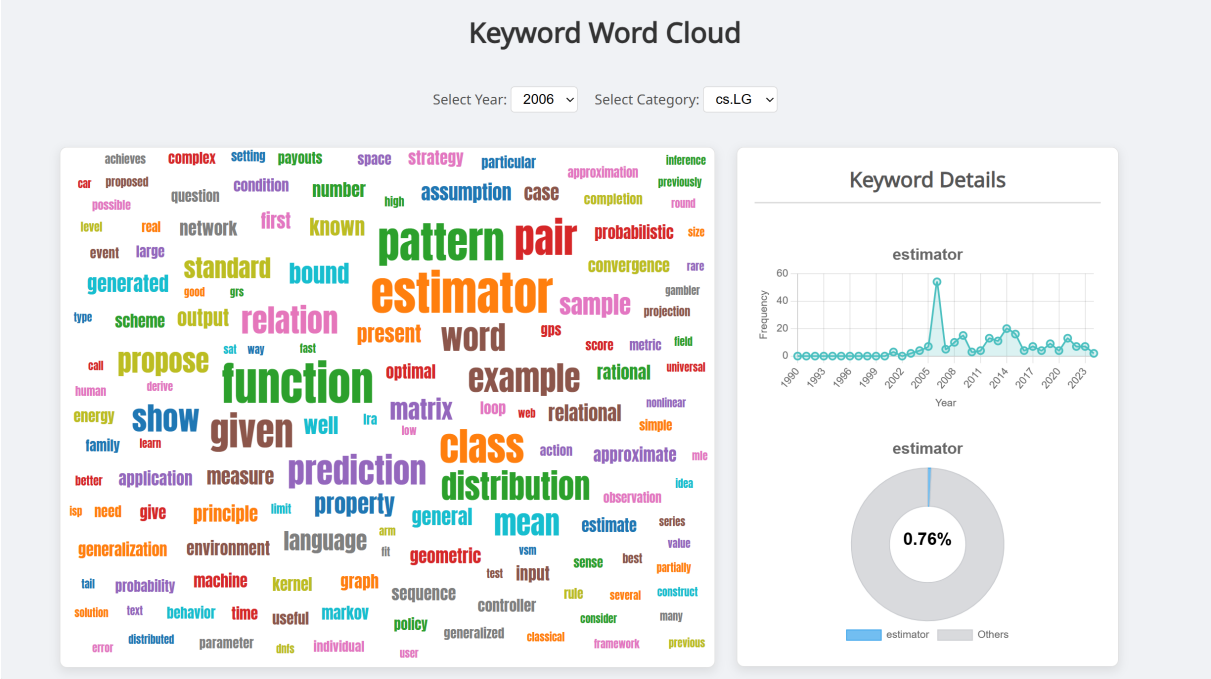

图 4　词云视图总览

4.3 主要交互设计

部分交互设计在上文中已经提及，在此不作赘述，接下来介绍，主要交互设计的效果。

(1) **年限选择交互**：

- 如图5所示趋势图的横轴范围根据所选年份范围动态变化；
- 如图6所示，论文分布图中，所选年份范围内的方格高亮显示。

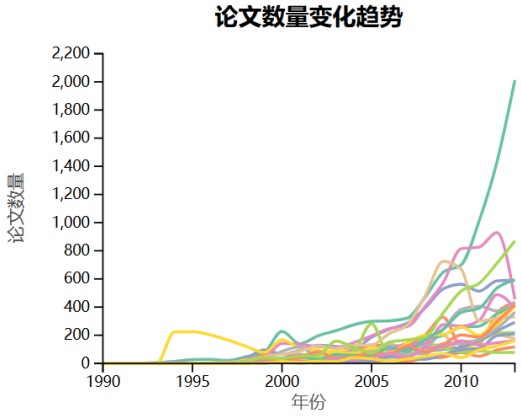

图 5

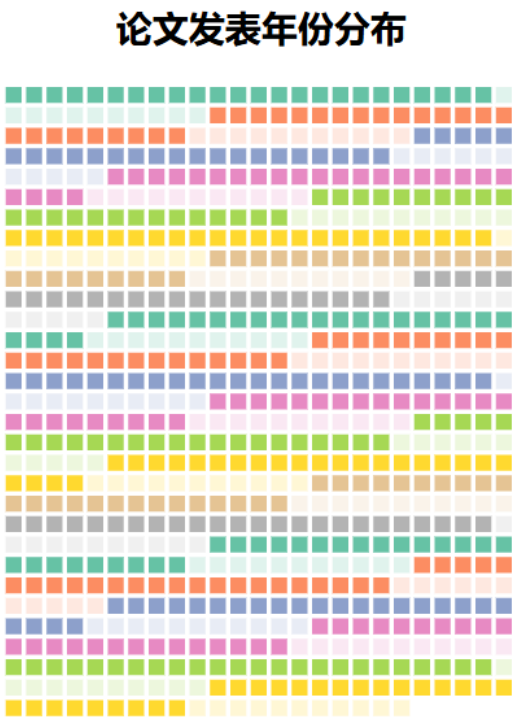

图 6

(2) **领域选择交互**:

- 如图7和图8所示，图例中选择领域后，其他视图同步更新，仅显示所选领域的数据；
- 视图 D 提供“Show All”按钮，可一键选中或取消选中所有领域。

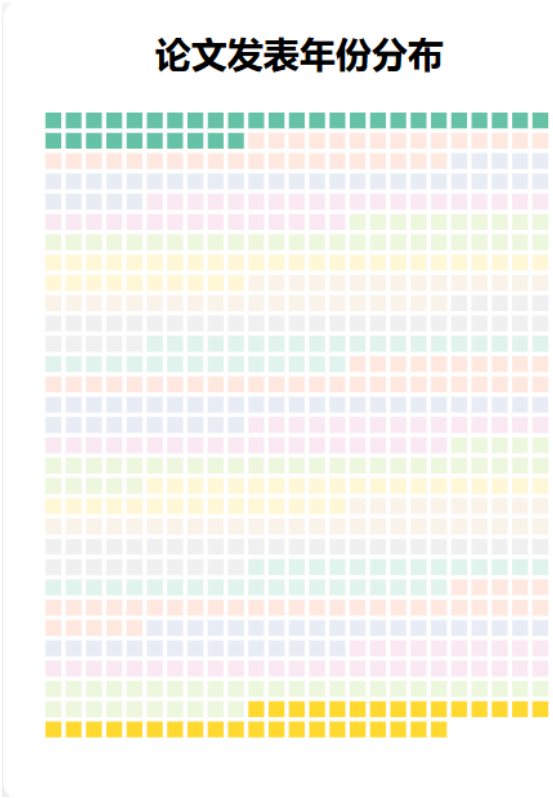

图 7

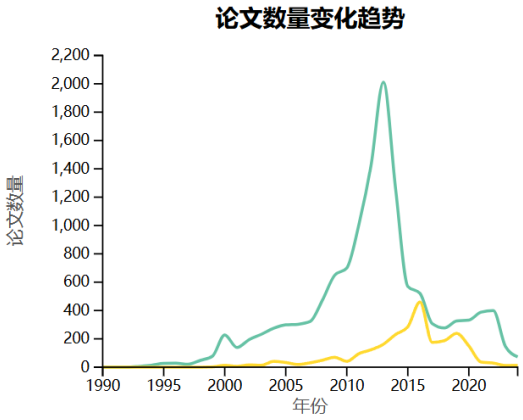

图 8

(3) **关系图交互**:

- 如图9所示点击某领域节点时，以该节点为中心放大显示与其他领域的关系；
- 若连续点击多个领域，可能导致多个中心节点，如图10所示，看似有种几何美，其实对可视化分析毫无用处，建议操作前重置分散状态。

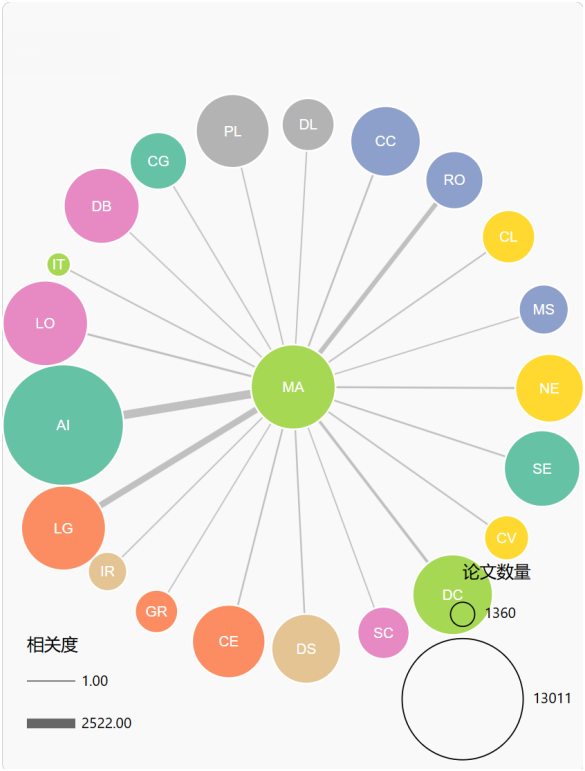

图 9

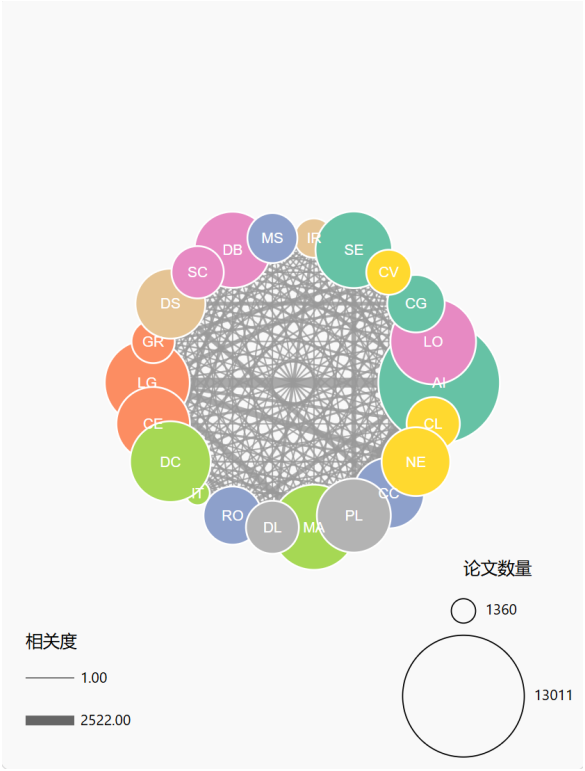

图 10

五、 结论

通过对计算机科学（CS）领域近三十年的论文数据分析，我们得出了以下结论：

5.1 对目前新兴领域的探究

人工智能（AI）是当前最热门的话题之一。我们通过查看 2024 年 AI 领域的论文关键词，探究其研究热点方向。

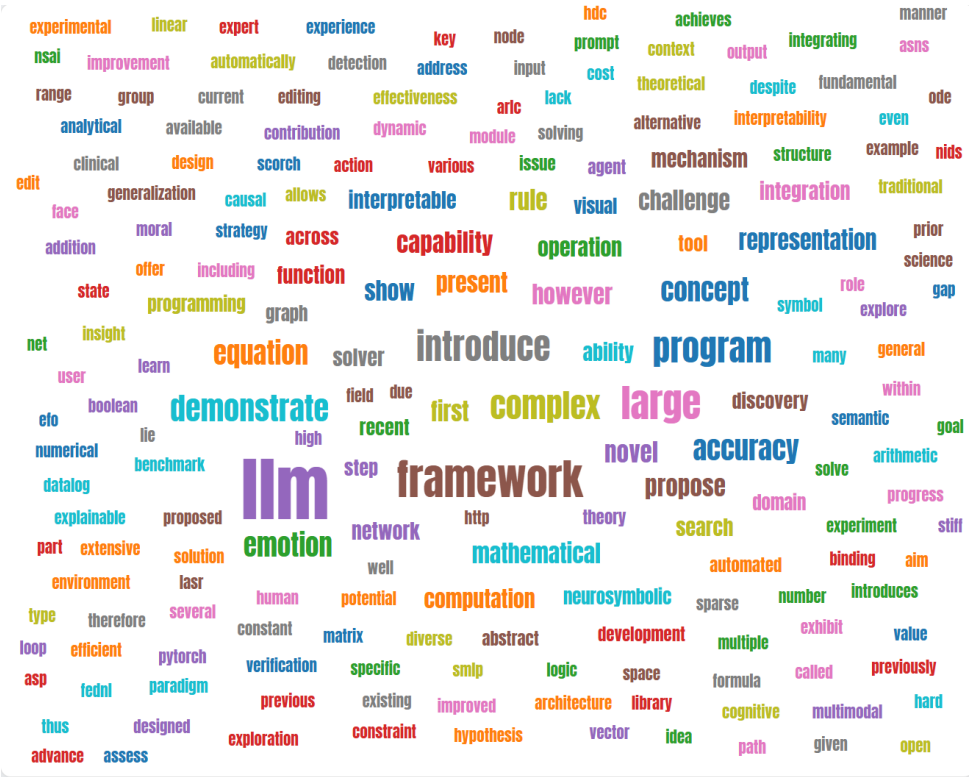

图 11　2024 年 AI 领域的关键词词云

如图 11 所示，2024 年 AI 领域的关键词中，“llm”（大语言模型）是一个显著热点。点击“llm”可查看其频率随时间的变化趋势。

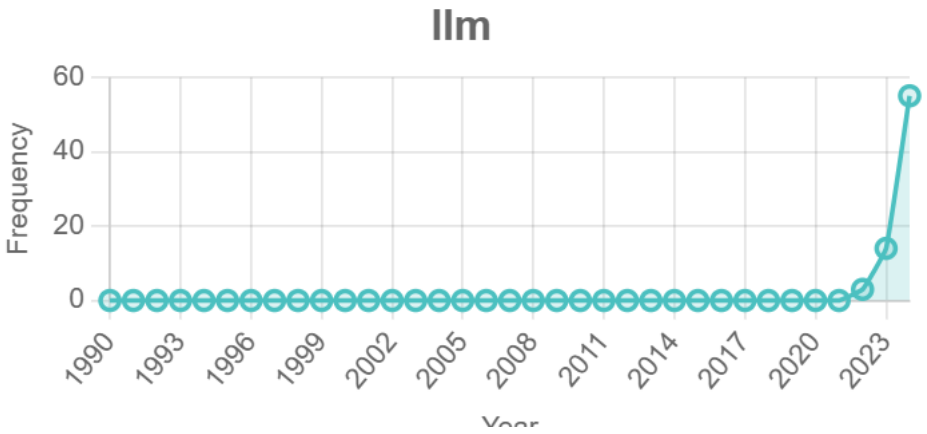

图 12　“llm”关键词的频率变化趋势

图 12 显示，“llm”在两年前开始兴起，此前并未被提及，表明它是一个新兴方向。短时间内快速增长，说明大语言模型正处蓬勃发展之势。

5.2 对 AI 与其他领域关系的探究

分析 AI 与其他领域的关系发现，AI 与以下领域关联最为紧密（参见图 13）：

- **MA（多智能体系统）**：发展与 AI 息息相关；
- **LG（机器学习）**：是 AI 领域的核心组成部分。

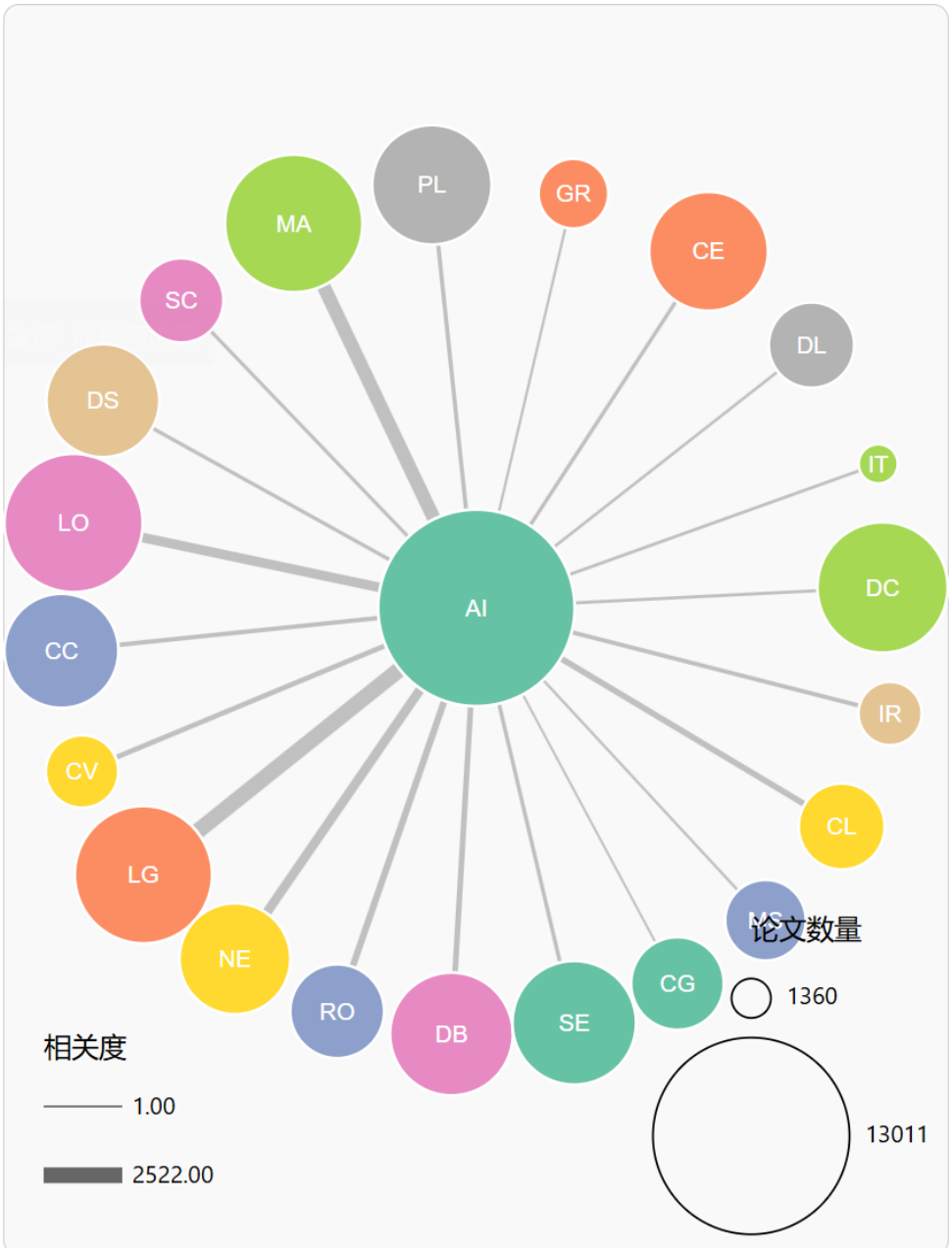

图 13　AI 领域与其他领域的关系图

如图 13 所示，与 AI 连线显著粗的领域有 MA（多智能体系统），LG（机器学习）说明了机器学习是 AI 领域非常关键的部分，MA 的发展与 AI 的发展息息相关

5.3 CS 领域的发展态势

通过观察 1990 年至今 CS 领域的论文数量变化趋势，我们可以总结出 CS 整体及各领域的发展状态。

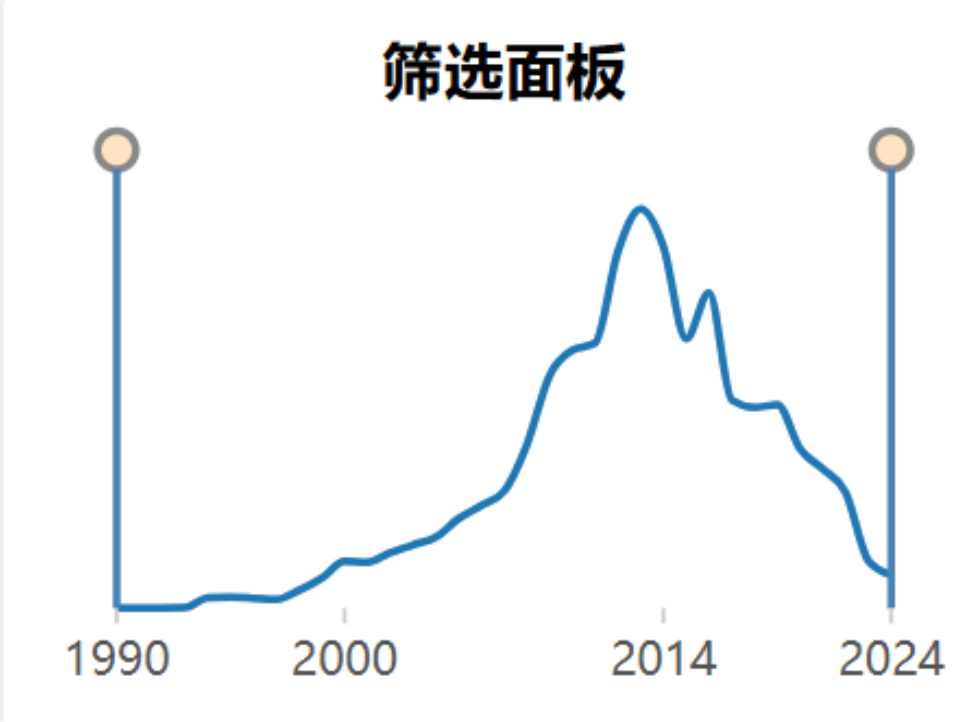

图 14　1990 年至今 CS 领域及各领域的论文数量变化趋势

如图 14 所示：

- CS 领域的论文发表数量在十几年前达到顶峰后，整体逐年下降，表明许多领域已趋于成熟；
- 图中紫色曲线代表的 MA（多智能体系统）在近些年达到顶峰。

进一步探究 MA 领域，在词云主界面选择其高峰年份，生成如下关键词词云（见图 15）。

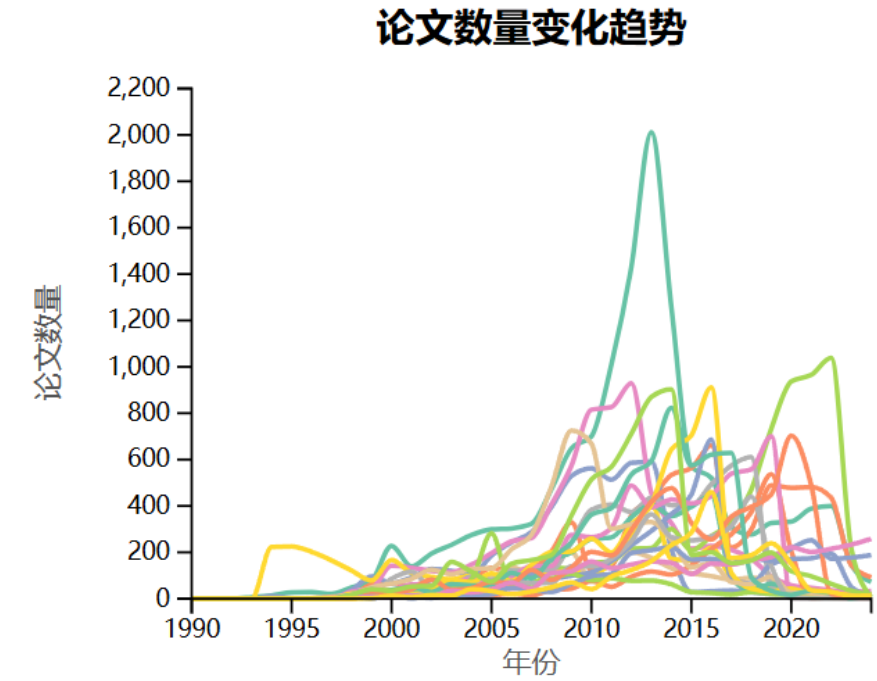

图 15　MA 领域高峰年份的关键词词云

从图 15 可见，MA 领域关键词多集中在无人机、机器人等实际应用方向，表明在这些方面有显著研究进展。结合关键词随年度的变化趋势并查询相关文献资料，可对该领域的发展形成更深认识。

5.4 对各领域论文结构的认知

通过图16不难看出，所有领域的论文的描述重点都在方法与结果上，体现了 CS 领域对技术创新、实践应用的高度重视，体现了其作为技术驱动学科的特点。

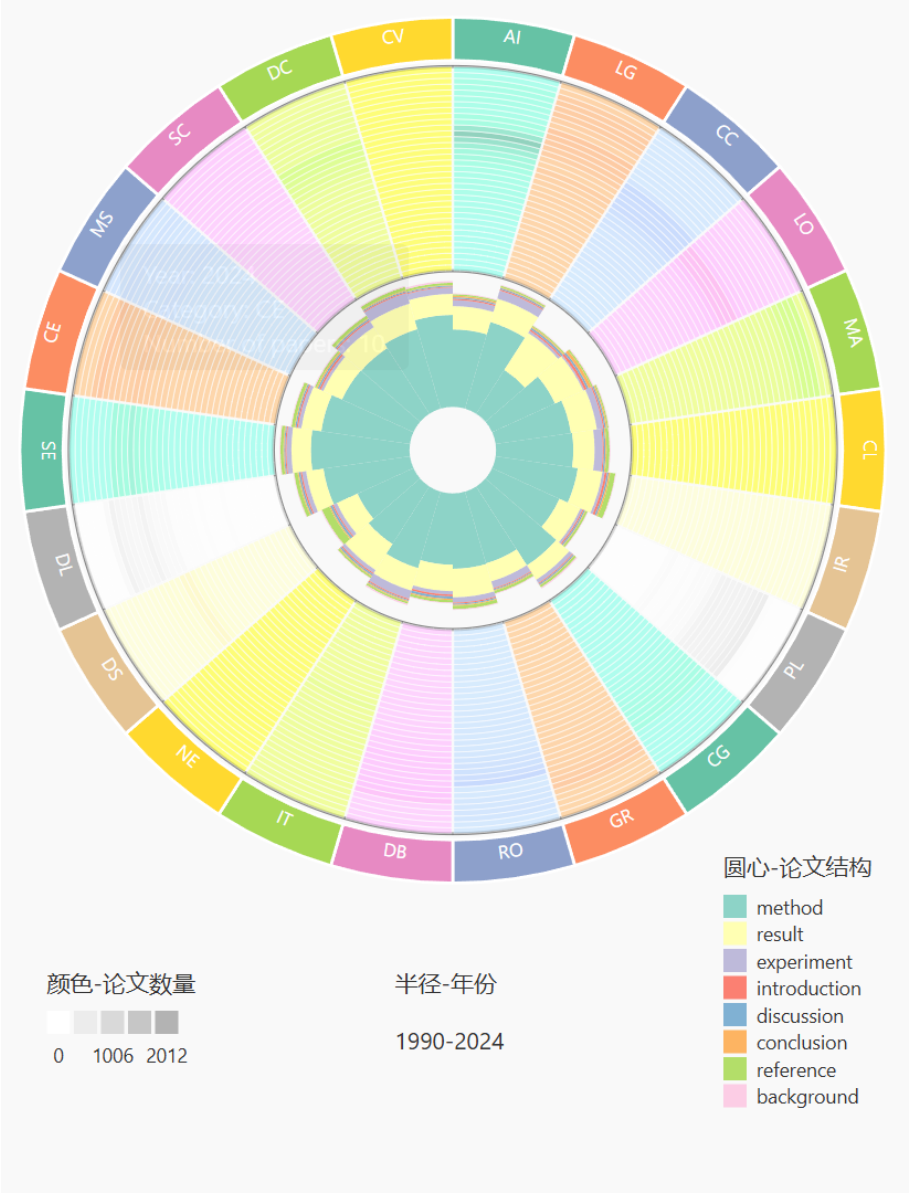
AI
LG
CC
LO
MA
CL
IR
PL
CG
GR
RO
DB
IT
NE
DS
DL
SE
CE
MS
SC
DC
CV
圆心-论文结构
method
result
experiment
introduction
discussion
conclusion
reference
background
颜色-论文数量
0
1006
2012
半径-年份
1990-2024

图 16　总览图