数据科学导论 -HW 3 报告

孙育泉 10234900421
2025.10.02

总览

I 实验要求

这次实验，我们需要从 ESI 网站 中获取本校的各个学科数据，进而对于本校的学科情况进行数据分析.

II 具体实现

II.1 数据收集

首先，我们要利用爬虫来收集我们数据. 对于动态网页，可以考虑直接找到 post/get 请求的接口来获取数据，也可以使用 selenium 来模拟浏览器操作.

但是，尝试了 selenium 之后，我发现，ESI 网站的连接不是很稳定即使使用了科学上网，而且使用 product login 会报错 system error，所以使用了 requests 发送请求来获取数据. （一共有 22 个文件，这里只有 14 个，因为一次性下载有概率掉连接，所以分了三次进行下载）

In [94]:

```python
import requests
import os
from bs4 import BeautifulSoup

URL = "https://esi.clarivate.com/IndicatorsAction.action"
s = requests.Session()

res = s.get(URL, headers=headers)
html = res.text
soup = BeautifulSoup(html, "lxml")
text = soup.find_all("div", class_="checkbox-2columns filter-values")

sum = 0

def download_pdf(url, file_name, download_path):
    global sum
    response = requests.get(url, headers=headers)
```

```
    if response.status_code == 200:
        sum += 1
        file_name = file_name + ".csv"
        file_name = file_name.replace("%26", "&")
        file_name = file_name.replace("/", " ")
        with open(download_path + file_name, "wb") as file:
            file.write(response.content)
        print("完成对" + file_name + "的下载")
    else:
        print("未能将" + file_name + "成功下载")


categories_name = ""
download_path = "./download/"

os.makedirs(download_path, exist_ok=True)
for next_text in text:
    lines = next_text.text.splitlines()
    lines.pop(0)
    print(lines)
    for name in lines[8:]:
        categories_name = name.replace("&", "%26")
        categories_name = categories_name.upper()
        s1 = "&show=Top&sort=%5B%7B%22property%22:%22cites%22,%22direction%22:
        %22DESC%22%7D%5D&colFilterVal=&exportType=indicators&colNames=RowSeq,,Institutio
        Regions,Web%20of%20Science%20Documents,Cites,Cites/
        Paper,Top%20Papers&fileType=CSV&f=IndicatorsExport.csv"
        file_url = (
            "https://esi.clarivate.com/IndicatorsExport.action?exportFile&_dc=
            1368621151464&groupBy=Institutions&start=0&limit=1381&filterBy=
            ResearchFields&filterValue="
            + categories_name
            + s1
        )
        download_pdf(file_url, categories_name, download_path)
print("成功下载了" + str(sum) + "个文件")
```

txt

```
['Agricultural Sciences', 'Biology & Biochemistry', 'Chemistry', 'Clinical Medicine',
'Computer Science', 'Economics & Business', 'Engineering', 'Environment/Ecology',
'Geosciences', 'Immunology', 'Materials Science', 'Mathematics', 'Microbiology',
'Molecular Biology & Genetics', 'Multidisciplinary', 'Neuroscience & Behavior',
'Pharmacology & Toxicology', 'Physics', 'Plant & Animal Science', 'Psychiatry/
Psychology', 'Social Sciences, General', 'Space Science']
完成对GEOSCIENCES.csv的下载
完成对IMMUNOLOGY.csv的下载
完成对MATERIALS SCIENCE.csv的下载
完成对MATHEMATICS.csv的下载
完成对MICROBIOLOGY.csv的下载
完成对MOLECULAR BIOLOGY & GENETICS.csv的下载
完成对MULTIDISCIPLINARY.csv的下载
完成对NEUROSCIENCE & BEHAVIOR.csv的下载
完成对PHARMACOLOGY & TOXICOLOGY.csv的下载
完成对PHYSICS.csv的下载
完成对PLANT & ANIMAL SCIENCE.csv的下载
完成对PSYCHIATRY PSYCHOLOGY.csv的下载
完成对SOCIAL SCIENCES, GENERAL.csv的下载
完成对SPACE SCIENCE.csv的下载
```

```
16 成功下载了14个文件
17
```

II.2 准备工作

我们引入一些进行数据处理、数据分析、数据可视化的常用库，同时进行一些简单是初始化操作.

In [95]:

```python
import pandas as pd
import os
import matplotlib.pyplot as plt
import seaborn as sns

sns.set_theme(style="whitegrid", context="paper")
plt.rcParams['font.sans-serif'] = ['Source Han Sans SC']
plt.rcParams['axes.unicode_minus'] = False # 解决负号显示问题

```

II.3 数据加载与分析

这一步，我们将对 csv 文件进行处理

- 去掉名称中的空格和特殊字符
- 将一些数值型的列转换为数值类型

In [101]:

```python
# --- 1. 数据加载与清洗 ---
print("--- 步骤 1: 正在加载与清洗数据... ---")

data_folder = './download/'
all_data_frames = []
for filename in os.listdir(data_folder):
    if filename.endswith(".csv"):
        subject_name = os.path.splitext(filename)[0].upper()
        file_path = os.path.join(data_folder, filename)
        try:
            df = pd.read_csv(file_path, skiprows=1, encoding='latin-1')
            df['Subject'] = subject_name
            all_data_frames.append(df)
        except Exception as e:
            print(f"处理文件 {filename} 时出错: {e}")

master_df = pd.concat(all_data_frames, ignore_index=True)
master_df.rename(columns={'Unnamed: 0': 'Rank', 'Web of Science Documents': 'Documents', 'Cites/Paper': 'Cites_Per_Paper', 'Countries/Regions': 'Country'}, inplace=True)
numeric_cols = ['Rank', 'Documents', 'Cites', 'Cites_Per_Paper', 'Top Papers']
for col in numeric_cols:
    master_df[col] = pd.to_numeric(master_df[col], errors='coerce')
master_df.dropna(inplace=True)
int_cols = ['Rank', 'Documents', 'Cites', 'Top Papers']
master_df[int_cols] = master_df[int_cols].astype(int)
print("数据清洗完成！总共加载有效数据行数:", len(master_df))
```

```txt
```

```
--- 步骤 1: 正在加载与清洗数据... ---
数据清洗完成！总共加载有效数据行数: 30960

```

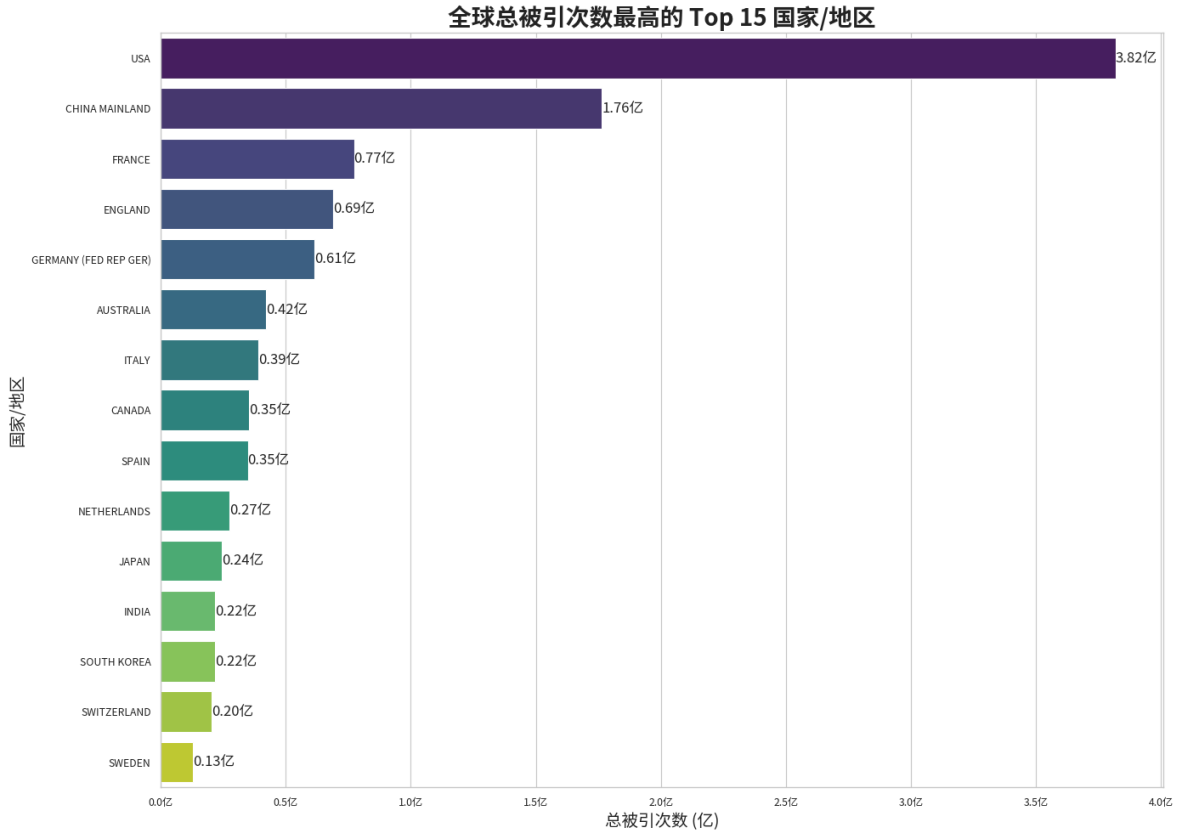
II.4 国家/地区分析

这一步，我们来找出论文被引用次数最多的一些国家/地区．结果如下图，次数最多的三个国家分别是：美国、中国、法国.

In [97]:

```python
# --- 2. 宏观分析与精美可视化 ---
print("\n--- 步骤 2: 正在进行宏观分析... ---")

# 2.1 按国家/地区分析
country_analysis = master_df.groupby('Country')['Cites'].sum().sort_values(ascending=False).head(15)

# 使用面向对象接口创建图表
fig, ax = plt.subplots(figsize=(14, 10))
barplot = sns.barplot(x=country_analysis.values, y=country_analysis.index, palette='viridis', ax=ax, hue=country_analysis.index, legend=False)

ax.set_title('全球总被引次数最高的 Top 15 国家/地区', fontsize=20, weight='bold')
ax.set_xlabel('总被引次数 (亿)', fontsize=14)
ax.set_ylabel('国家/地区', fontsize=14)
# 格式化X轴标签，将数字转换为以“亿”为单位
ax.xaxis.set_major_formatter(lambda x, pos: f'{x/1e8:.1f}亿')

# 添加数据标签
for p in ax.patches:
    width = p.get_width()
    ax.text(width, p.get_y() + p.get_height() / 2,
            f'{width/1e8:.2f}亿',
            va='center', ha='left', fontsize=12)

plt.tight_layout()
plt.show()

```

```txt

--- 步骤 2: 正在进行宏观分析... ---

```



II.5 机构分析

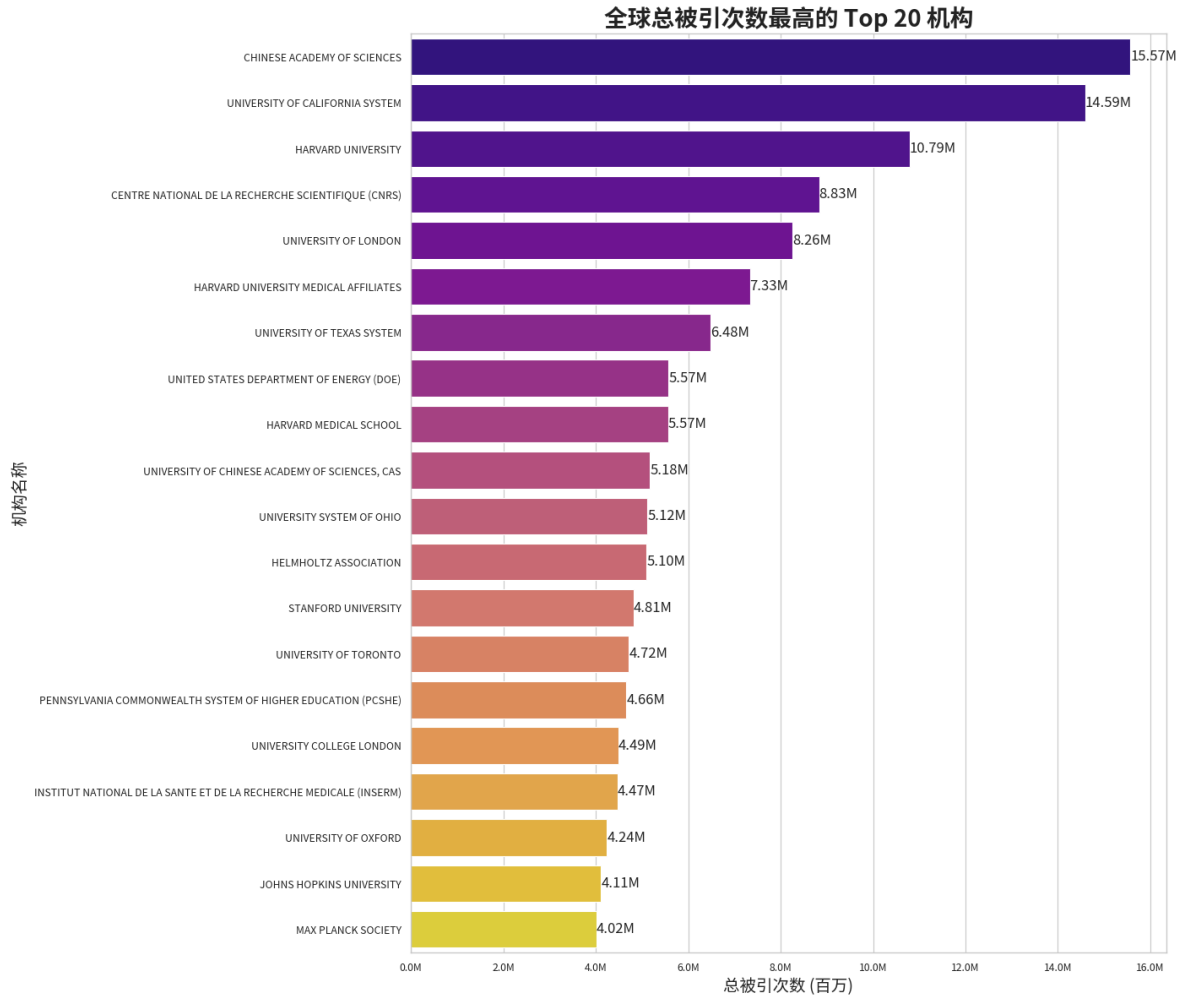
这一步，我们来找出论文被引用次数最多的一些机构．结果如下图，次数最多的三个机构分别是：中科院、美国加州大学、哈佛大学.

In [98]:

python

```python
# --- 2.2 按机构分析 ---
institution_analysis = master_df.groupby('Institutions')['Cites'].sum().sort_values(ascending=False).head(20)

# 使用面向对象接口创建图表
fig, ax = plt.subplots(figsize=(14, 12)) # 增加了图表高度以容纳20个机构名称
barplot = sns.barplot(x=institution_analysis.values, y=institution_analysis.index, palette='plasma', ax=ax, hue=institution_analysis.index, legend=False)


ax.set_title('全球总被引次数最高的 Top 20 机构', fontsize=20, weight='bold')
ax.set_xlabel('总被引次数 (百万)', fontsize=14)
ax.set_ylabel('机构名称', fontsize=14)
# 格式化X轴标签，将数字转换为以“百万”为单位
ax.xaxis.set_major_formatter(lambda x, pos: f'{x/1e6:.1f}M')

# 添加数据标签
for p in ax.patches:
    width = p.get_width()
    # 在柱子末端稍微靠右的位置添加文本
    ax.text(width, p.get_y() + p.get_height() / 2,
            f'{width/1e6:.2f}M', # M 代表百万 (Million)
```

```
            va='center', ha='left', fontsize=12)

plt.tight_layout()
plt.show()
```



II.6 ECNU 纵向分析

接下来，关注 ECNU 的情况，这里统计了 ECNU 在各个学科的论文被引用的次数的数据，可以发现，前三引用量最高的学科是：化学、材料科学、环境生态学.

In [123]:

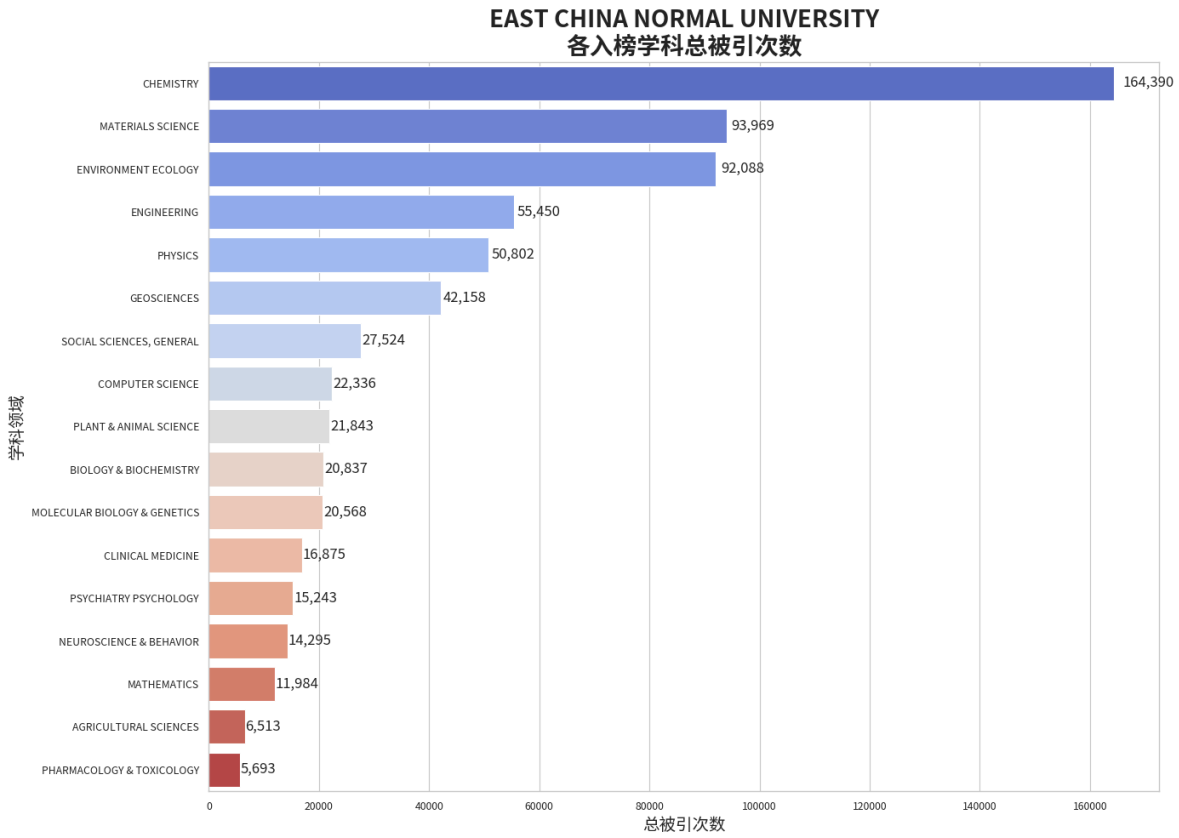
python

```
# --- 3. ECNU 深度分析与精美可视化 ---
print("\n--- 步骤 3: 正在分析 East China Normal University... ---")
university_name = "EAST CHINA NORMAL UNIVERSITY"
# university_name = "BEIJING INSTITUTE OF TECHNOLOGY"
# university_name = "PEKING UNIVERSITY"
# university_name = "TSINGHUA UNIVERSITY"
# university_name = "SHANGHAI UNIVERSITY"
# university_name = "UNIVERSITY OF TOKYO"

ecnu_data = master_df[master_df['Institutions'] == university_name]

if not ecnu_data.empty:
```

```
    ecnu_sorted_by_cites = ecnu_data.sort_values(by='Cites', ascending=False)

    fig, ax = plt.subplots(figsize=(14, 10))
    sns.barplot(x='Cites', y='Subject', data=ecnu_sorted_by_cites, palette='coolwarm',
    ax=ax, hue='Subject', legend=False)
    ax.set_title(f'{university_name}\n各入榜学科总被引次数', fontsize=20, weight='bold')
    ax.set_xlabel('总被引次数', fontsize=14)
    ax.set_ylabel('学科领域', fontsize=14)

    # 添加数据标签
    for p in ax.patches:
        width = p.get_width()
        ax.text(width * 1.01, p.get_y() + p.get_height() / 2,
                f'{int(width):,}', # 格式化为千位分隔符
                va='center', fontsize=12)

    plt.tight_layout()
    plt.show()
else:
    print(f"未在数据中找到 {university_name}. ")

```

txt

```

--- 步骤 3: 正在分析 East China Normal University... ---

```



II.7 ECNU 横向对比

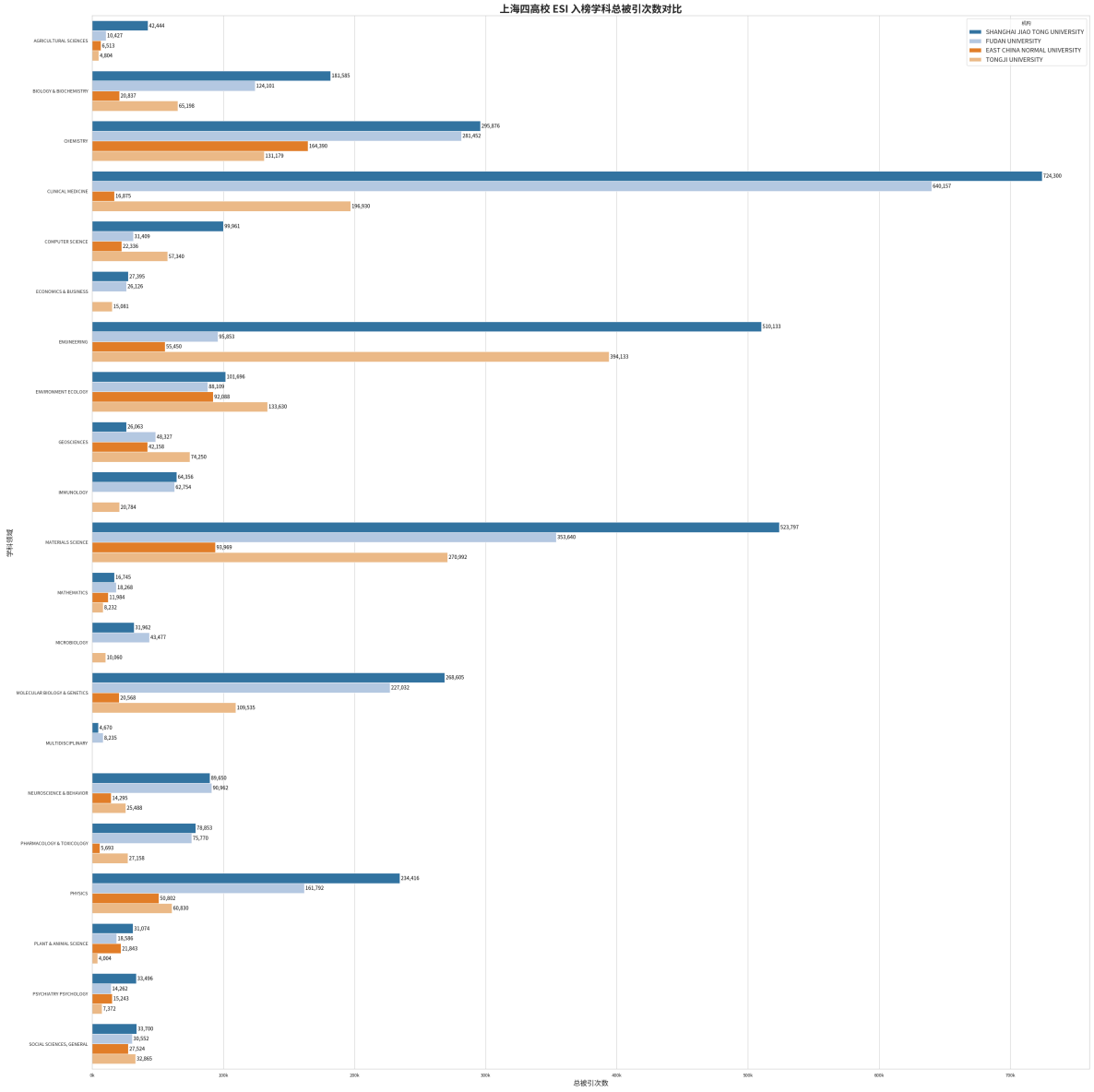
我们对比一下上海四个高校复交同华的学科情况. 不难发现，我们华师大和统计的水准不相上下 😆，特别是在化学以及植物动物科学上碾压同济!

In [122]:

```python
# --- 4. 四校对比分析 (ECNU, Fudan, Tongji, SJTU) ---
print("\n--- 步骤 4: 正在进行四校对比分析... ---")

# 定义要比较的四所高校的名称（必须与CSV中的完全一致）
comparison_list = [
    # "PEKING UNIVERSITY",
    # "TSINGHUA UNIVERSITY",
    "EAST CHINA NORMAL UNIVERSITY",
    "FUDAN UNIVERSITY",
    "TONGJI UNIVERSITY",
    "SHANGHAI JIAO TONG UNIVERSITY"
    # "BEIJING INSTITUTE OF TECHNOLOGY",
    # "SHANGHAI UNIVERSITY",
    # "UNIVERSITY OF TOKYO"
]

# 筛选出这四所高校的所有数据
comparison_df = master_df[master_df['Institutions'].isin(comparison_list)]

# 按学科和总引用数排序，以获得更好的可视化效果
comparison_df_sorted = comparison_df.sort_values(by=['Subject', 'Cites'],
ascending=[True, False])

if not comparison_df.empty:
    # 增加图表高度以容纳更多学科
    fig, ax = plt.subplots(figsize=(30, 30))

    sns.barplot(x='Cites', y='Subject', hue='Institutions', data=comparison_df_sorted,
    palette='tab20', ax=ax)

    ax.set_title('上海四高校 ESI 入榜学科总被引次数对比', fontsize=20, weight='bold')
    ax.set_xlabel('总被引次数', fontsize=14)
    ax.set_ylabel('学科领域', fontsize=14)
    ax.legend(title='机构', fontsize=12)
    ax.xaxis.set_major_formatter(lambda x, pos: f'{int(x/1000)}k') # 以 "k" (千) 为单位

    # 为分组条形图添加数据标签（可选，因为可能很密集）
    for p in ax.patches:
        width = p.get_width()
        if width > 0: # 只为有数据的条添加标签
            ax.text(width, p.get_y() + p.get_height() / 2,
                    f' {int(width):,}', # 格式化为千位分隔符
                    va='center', ha='left', fontsize=10, color='black')

    plt.tight_layout()
    plt.show()
else:
    print("在数据中未能找到指定的四所高校. ")
```

```txt

--- 步骤 4: 正在进行四校对比分析... ---
```

```
3
```



II.8 一些建议

通过分析上海四所顶尖高校（上海交通大学、复旦大学、华东师范大学、同济大学）在 ESI（基本科学指标数据库）入榜学科的**总被引次数**对比情况.

我们发现:

1. **整体格局：两强领跑，各具特色**
   - **上海交通大学** (SJTU) 和 **复旦大学** (Fudan) 在绝大多数高引用学科中，形成了明显的“第一梯队”，总被引次数远超另外两所高校.
   - SJTU (**深蓝色**) 的优势集中在**工科和理科**，尤其在**材料科学、工程学、化学** 这三个领域展现出强大的统治力.
   - **复旦大学** (**浅蓝色**) 的顶峰优势在**临床医学**，其被引次数遥遥领先，体现了其医科的卓越地位. 同时，在化学、材料科学等领域也具备极强的实力.
2. **华东师范大学 (ECNU) 的定位：优势突出，潜力巨大**

- **传统优势学科 (高影响力区)**：ECNU (橙色) 的总被引次数最高的学科主要集中在**化学、工程学、材料科学** 和 **环境与生态学**. 这些是 ECNU 科研体量和影响力的“基本盘”.
- **特色与领先学科 (相对优势区)**：在 **精神病学与心理学** 领域，ECNU 的总被引次数在四校中处于**领先位置**. 这表明 ECNU 在该领域具备了独特的竞争优势和学术声誉. 此外，在**地球科学** 和 **社会科学总论** 等领域也表现出强劲的竞争力，与同济大学相当或略有优势.
- **待发展学科 (追赶区)**：在临床医学、物理学等领域，ECNU 虽然也进入了 ESI 前 1%，但与交大、复旦相比，总引用量存在较大差距，属于需要持续投入和追赶的领域.

3. **同济大学 (Tongji) 的观察** 同济大学 (棕色) 的优势符合其传统认知，主要体现在**工程学**和**计算机科学**，在这两个领域其被引次数非常可观，是主要的贡献者.

基于以上数据分析，可以为 ECNU 的学科发展战略提出以下几点建议：

- 建议一：巩固传统优势，打造“高峰”学科
  - **化学、材料科学、工程学、环境与生态学**是 ECNU 学术影响力的基石.
  - **策略**：应继续加大对这些“高峰”学科的战略投入，吸引顶尖人才，构建大科研平台. 目标不仅是维持 ESI 前 1%，更应是缩小与上海交大、复旦在这些主流赛道上的差距，力争在某些细分方向上取得突破性领先.
- 建议二：强化特色学科，形成“不可替代”的标签
  - **精神病学与心理学、地球科学、社会科学**等是 ECNU 的特色和相对优势所在，也是与 ECNU“师范”底蕴和文理综合优势高度相关的领域.
  - **策略**：将这些“人无我有，人有我优”的学科作为学校的“名片”来打造. 资源配置上可以有所倾斜，鼓励其建立跨学科研究中心（例如，结合教育学的“教育心理学”，结合计算机科学的“计算社会科学”），形成其他高校难以复制的交叉学科优势，成为国内乃至国际上的领跑者.
- 建议三：推动学科交叉，寻找新的“增长点”
  - ECNU 在许多领域虽有入榜，但并非顶尖. 盲目追赶不如“弯道超车”.
  - **策略**：主动设计和推动“优势学科”与“待发展学科”的深度交叉融合. 例如，可以利用**计算机科学**（ECNU 在此领域有一定基础）和领先的**心理学/神经科学**优势，去赋能**临床医学**研究，聚焦于“计算精神病学”、“智能心理健康”等前沿方向. 又如，将**环境生态学**的优势与**社会科学**结合，深入研究“可持续发展政策”、“气候变化社会影响”等议题. 通过这种方式，可以在竞争激烈的领域中开辟出新的、有增长潜力的赛道，化“追赶”为“引领”.
- 建议四：实施数据驱动的精准资源配置
  - 本次分析本身就是一个很好的例子.
  - **策略**：建议学校层面常态化地利用 ESI、Scopus、InCites 等数据库进行数据分析，定期对本校及对标高校的学科表现进行“体检”. 基于客观数据，动态调整资源分配、人才引进方向和科研考核指标，使决策更加科学、精准.

总而言之，ECNU 的发展策略可以概括为：**稳固理科工科的“基本盘”，高举心理社科的“特色牌”，善用学科交叉的“催化剂”，最终实现从“多点开花”到“高峰凸显、高原广阔”的战略升级**.

> **Remark**
>
> - 完整代码见 `./hw3.ipynb`.
> - ESI 数据位于 `./download` 文件夹下.