进展汇报

Codeforces 数据分析

廖嘉琦、曹健、张亦晴

April 26, 2024

Part I

数据获取及预处理

数据来源

- Codeforces API：主要来源，获取结构化的 JSON 数据
- Codeforces爬虫：辅助手段，获取 API 未提供，但对分析有用的数据

数据说明

数据将涵盖以下几个主要方面：

- **竞赛数据**：利用 contest.list、contest.standings 和 contest.status 接口（以及爬虫），获取竞赛的基本信息、排名和提交记录
- **用户数据**：通过 user.info、user.status 和 user.rating 接口，收集选手的基本信息、提交历史和等级变化
- **社区互动数据**：通过 blogEntry.view、blogEntry.comments 和 user.blogEntries 接口，分析社区的讨论热度和互动模式

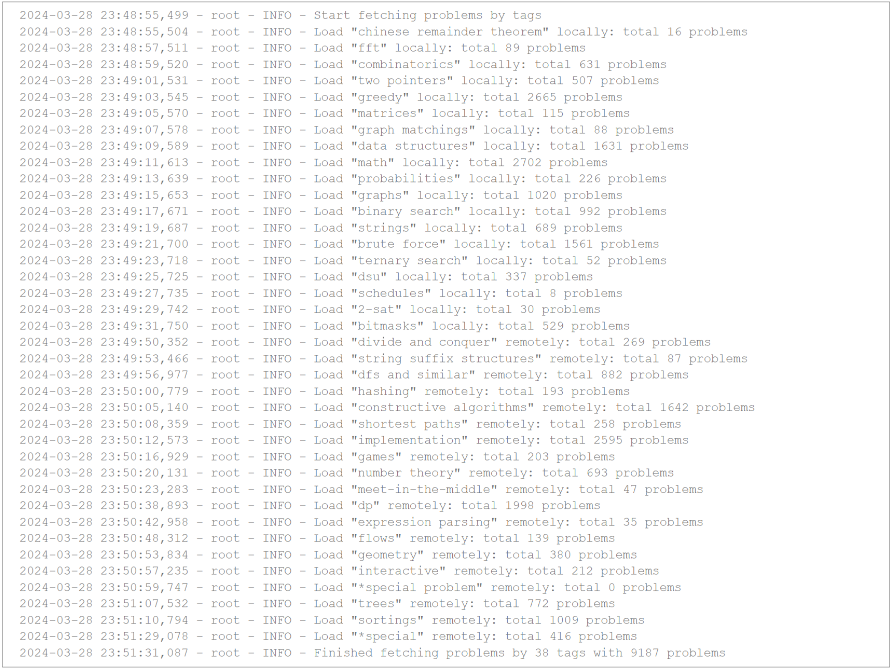
竞赛数据：1853 rows × 8 columns

	durationSeconds	name	type	phase	startTime	frozen	id	relativeTimeSeconds
0	10800	Codeforces Round (Div. 1 + Div. 2)	CF	BEFORE	2024-04-06 14:35:00	False	1951	-777732
1	7200	April Fools Day Contest 2024	ICPC	BEFORE	2024-04-01 14:35:00	False	1952	-345732
2	10800	CodeTON Round 8 (Div. 1 + Div. 2, Rated, Prizes!)	CF	BEFORE	2024-03-30 14:35:00	False	1942	-172932
3	8100	Codeforces Round 937 (Div. 4)	ICPC	BEFORE	2024-03-28 14:45:00	False	1950	-734
4	18000	European Championship 2024 - Online Mirror (Un...	ICPC	FINISHED	2024-03-24 10:00:00	False	1949	361968
...	...	...	...	...	...	...	...	...
1848	7200	Codeforces Beta Round 5	ICPC	FINISHED	2010-03-20 16:00:00	False	5	442535569
1849	7200	Codeforces Beta Round 4 (Div. 2 Only)	ICPC	FINISHED	2010-03-12 12:00:00	False	4	443241169
1850	7200	Codeforces Beta Round 3	ICPC	FINISHED	2010-03-07 12:00:00	False	3	443673169
1851	7200	Codeforces Beta Round 2	ICPC	FINISHED	2010-02-25 17:00:00	False	2	444519169
1852	7200	Codeforces Beta Round 1	ICPC	FINISHED	2010-02-19 12:00:00	False	1	445055569

```
2024-03-28 22:32:47,060 - root - INFO - Start fetching contests
2024-03-28 22:33:09,592 - root - INFO - Fetched 1853 contests hosted by Codeforces
2024-03-28 22:33:09,629 - root - INFO - Fetched 1887 contests in gym
```

题目数据：9187 rows × 8 columns

	name	type	rating	tags	contestId	points	solvedCount	index
0	Amanda the Amoeba	PROGRAMMING	NaN	['graphs', 'implementation', 'trees', 'two poi...	1949	NaN	193	J
1	Clique Partition	PROGRAMMING	2100.0	['brute force', 'constructive algorithms', 'gr...	1948	NaN	2164	E
2	Array Fix	PROGRAMMING	1100.0	['brute force', 'dp', 'greedy', 'implementation']	1948	NaN	17820	B
3	Birthday Gift	PROGRAMMING	1900.0	['bitmasks', 'brute force', 'constructive algo...	1946	1750.0	3031	D
4	Tree Cutting	PROGRAMMING	1600.0	['binary search', 'dp', 'greedy', 'implementat...	1946	1500.0	7101	C
...	...	...	...	...	...	...	...	...
9182	Circular RMQ	PROGRAMMING	2200.0	['data structures']	52	1500.0	7983	C
9183	Dancing Lessons	PROGRAMMING	1900.0	['data structures']	45	NaN	842	C
9184	Queue	PROGRAMMING	2300.0	['data structures']	38	NaN	670	G
9185	Points	PROGRAMMING	2800.0	['data structures']	19	NaN	2032	D
9186	Bindian Signalizing	PROGRAMMING	2400.0	['data structures']	5	NaN	1987	E



```
2024-03-28 23:48:55,499 - root - INFO - Start fetching problems by tags
2024-03-28 23:48:55,504 - root - INFO - Load "chinese remainder theorem" locally: total 16 problems
2024-03-28 23:48:57,511 - root - INFO - Load "fft" locally: total 89 problems
2024-03-28 23:48:59,520 - root - INFO - Load "combinatorics" locally: total 631 problems
2024-03-28 23:49:01,531 - root - INFO - Load "two pointers" locally: total 507 problems
2024-03-28 23:49:03,545 - root - INFO - Load "greedy" locally: total 2665 problems
2024-03-28 23:49:05,570 - root - INFO - Load "matrices" locally: total 115 problems
2024-03-28 23:49:07,578 - root - INFO - Load "graph matchings" locally: total 88 problems
2024-03-28 23:49:09,589 - root - INFO - Load "data structures" locally: total 1631 problems
2024-03-28 23:49:11,613 - root - INFO - Load "math" locally: total 2702 problems
2024-03-28 23:49:13,639 - root - INFO - Load "probabilities" locally: total 226 problems
2024-03-28 23:49:15,653 - root - INFO - Load "graphs" locally: total 1020 problems
2024-03-28 23:49:17,671 - root - INFO - Load "binary search" locally: total 992 problems
2024-03-28 23:49:19,687 - root - INFO - Load "strings" locally: total 689 problems
2024-03-28 23:49:21,700 - root - INFO - Load "brute force" locally: total 1561 problems
2024-03-28 23:49:23,718 - root - INFO - Load "ternary search" locally: total 52 problems
2024-03-28 23:49:25,725 - root - INFO - Load "dsu" locally: total 337 problems
2024-03-28 23:49:27,735 - root - INFO - Load "schedules" locally: total 8 problems
2024-03-28 23:49:29,742 - root - INFO - Load "2-sat" locally: total 30 problems
2024-03-28 23:49:31,750 - root - INFO - Load "bitmasks" locally: total 529 problems
2024-03-28 23:49:50,352 - root - INFO - Load "divide and conquer" remotely: total 269 problems
2024-03-28 23:49:53,466 - root - INFO - Load "string suffix structures" remotely: total 87 problems
2024-03-28 23:49:56,977 - root - INFO - Load "dfs and similar" remotely: total 882 problems
2024-03-28 23:50:00,779 - root - INFO - Load "hashing" remotely: total 193 problems
2024-03-28 23:50:05,140 - root - INFO - Load "constructive algorithms" remotely: total 1642 problems
2024-03-28 23:50:08,359 - root - INFO - Load "shortest paths" remotely: total 258 problems
2024-03-28 23:50:12,573 - root - INFO - Load "implementation" remotely: total 2595 problems
2024-03-28 23:50:16,929 - root - INFO - Load "games" remotely: total 203 problems
2024-03-28 23:50:20,131 - root - INFO - Load "number theory" remotely: total 693 problems
2024-03-28 23:50:23,283 - root - INFO - Load "meet-in-the-middle" remotely: total 47 problems
2024-03-28 23:50:38,893 - root - INFO - Load "dp" remotely: total 1998 problems
2024-03-28 23:50:42,958 - root - INFO - Load "expression parsing" remotely: total 35 problems
2024-03-28 23:50:48,312 - root - INFO - Load "flows" remotely: total 139 problems
2024-03-28 23:50:53,834 - root - INFO - Load "geometry" remotely: total 380 problems
2024-03-28 23:50:57,235 - root - INFO - Load "interactive" remotely: total 212 problems
2024-03-28 23:50:59,747 - root - INFO - Load "*special problem" remotely: total 0 problems
2024-03-28 23:51:07,532 - root - INFO - Load "trees" remotely: total 772 problems
2024-03-28 23:51:10,794 - root - INFO - Load "sortings" remotely: total 1009 problems
2024-03-28 23:51:29,078 - root - INFO - Load "*special" remotely: total 416 problems
2024-03-28 23:51:31,087 - root - INFO - Finished fetching problems by 38 tags with 9187 problems
```

Part II

数据分析与可视化

Codeforces 官方比赛的年月分布

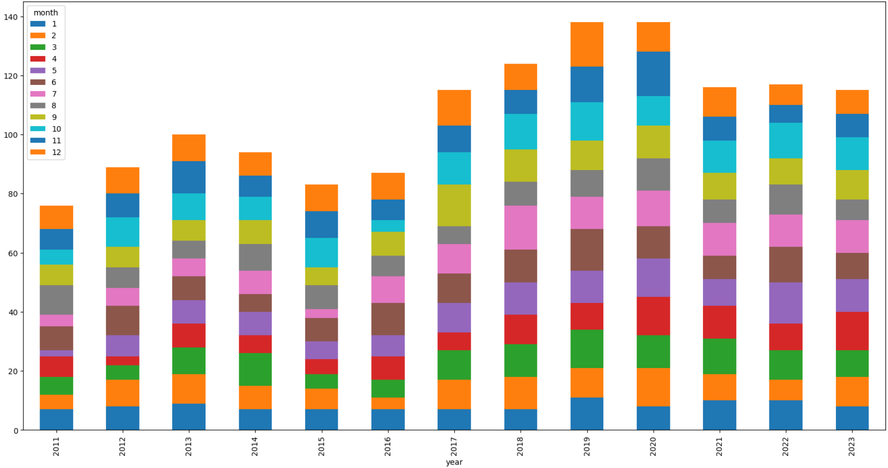
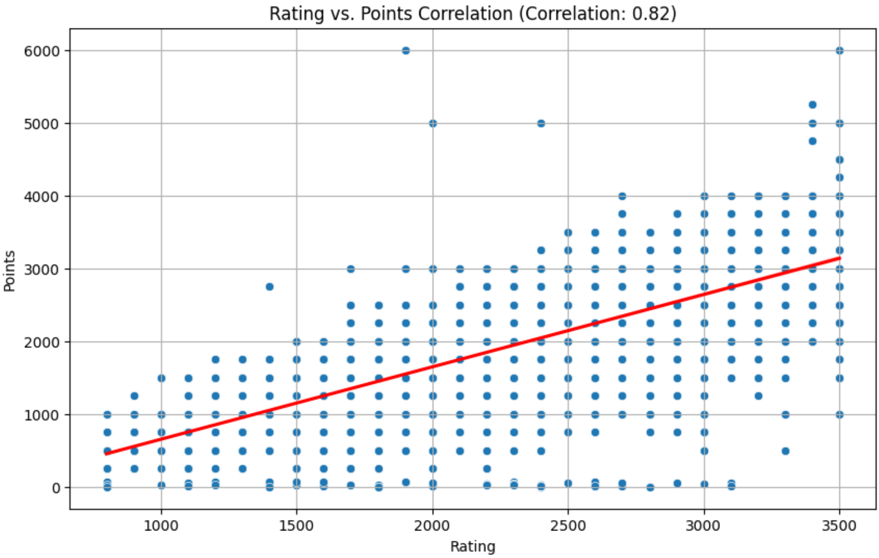
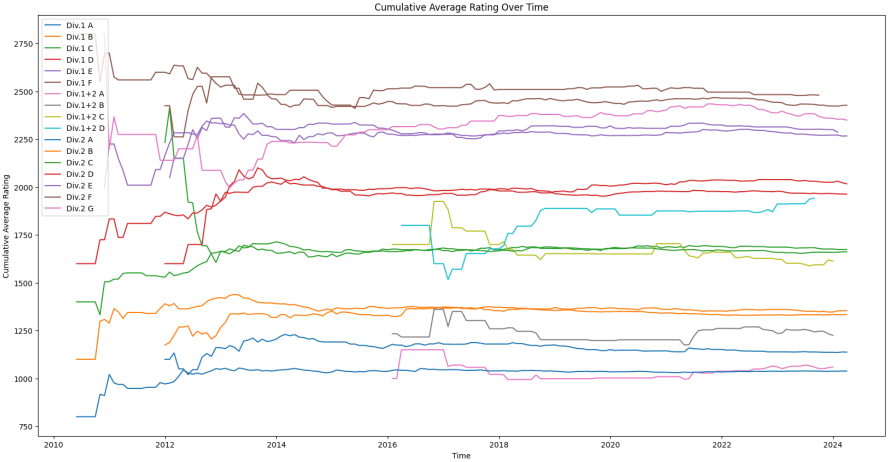


Figure: 2011-2024 年 CF 官方比赛分布

赛前（Points）-赛后（Rating）题目难度打分的相关性



比赛题目难度的变化



难度整体比较稳定

Part III

其他

模型选取

- **关联规则学习**：发现不同编程问题之间的关系，例如哪些类型的题目经常一起出现，或者哪些技能在解决某类问题时特别重要。这对于理解竞赛题目的结构和参赛者的解题模式非常有帮助。
- **关联规则学习**：发现选手表现与学习资源、讨论话题之间的关联，例如通过 Apriori 算法寻找常见的题目组合或讨论主题。
- **聚类算法**：如 K-means 或层次聚类，这可以用来发现具有相似特征的参赛者群体或题目类型，从而帮助理解数据中的模式和关系。

系统交互设计

- **数据浏览**：允许用户浏览和搜索 Codeforces 的历史比赛和题目数据
- **数据分析**：提供各种预设的数据分析选项，如趋势分析、参赛者表现评估等
- **报告生成**：用户可以生成和下载分析报告，包括图表和统计摘要
- **实时数据追踪**：跟踪实时比赛数据和用户表现

问题及下一步工作

1. 部分数据 API 的记录不完全，需要写爬虫重新获取
2. 部分数据未获得，需要获取
3. 进一步分析数据，挖掘数据的结果

任务分工与进度

- 廖嘉琦：数据获取、算法实现、文档撰写：进行中
- 张亦晴：系统设计、可视化、文档撰写：进行中
- 曹健：算法实现、可视化、文档撰写：进行中