Mining Insights from the Past: Analyzing Query Logs in Web Archive

Adam Jatowt (adam.jatowt@uibk.ac.at UIBK) and Ricardo Campos (ricardo.campos@ubi.pt, Universidade da Beira Interior)

Objectives

In recent years, the analysis of user interaction logs has become a critical area in information retrieval research [1, 2, 3]. These logs capture invaluable insights into user behavior and preferences, which can be leveraged to improve search engines and enhance user experience [4]. Query logs over web archives are a unique resource of data that allow researchers to delve into the evolution of online information-seeking patterns and track the changing behavior of user demands over time. This proposal outlines a master thesis work aimed at conducting a comprehensive analysis on top of a large query log dataset provided by the Arquivo.pt, the Portuguese web archive. These logs (see Fig. 1) were accumulated over several years as a result of user interactions with the Arquivo.pt search engine, a "Google-like" service that enables searching pages and images collected from the web since the 1990s.

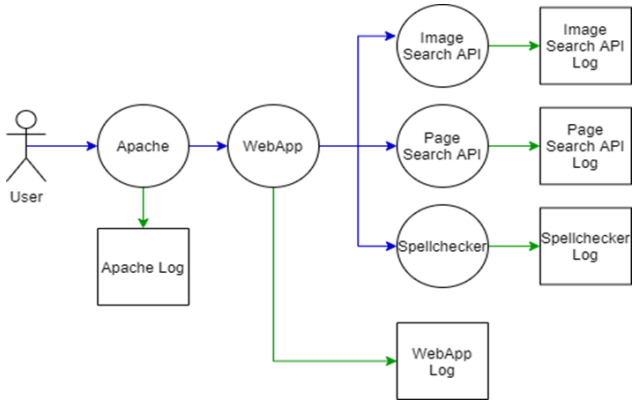

Fig. 1: Arquivo.pt query logs

The primary objective of this research is to preprocess and explore the dataset to understand its structure, identify patterns, and potential challenges. To this regard, the student should extend a recent work by Gallois, F. (2023) [5] who has performed an initial analysis over a 3-month query log provided by the Arquivo.pt. Based on the knowledge obtained and on the analysis conducted, the student is then expected to identify areas for potential improvement of the search engine and to propose and develop some innovative technique, potentially

leveraging machine learning approaches for query log anomaly detection, query understanding or search recommendation to name a few.

Approximate Workplan

T1: Literature Review and State-of-the-Art Analysis (2 months)

- Conduct a comprehensive review of the existing literature in the fields of web search behavior analysis, and information retrieval.
- Identify the most recent and relevant research papers, publications, and cutting-edge techniques in the analysis of user interaction logs.
- Summarize and synthesize the state-of-the-art approaches and findings in these areas.
- Identify potential areas where your research can contribute novel insights or improvements.
- Modify the research plan, as necessary, to incorporate state-of-the-art insights.

T2: Project Setup and Data Preparation (2 months)

- Set up a robust platform, potentially leveraging big data technologies, for efficient analysis of the query log dataset.
- Acquire the query logs from Arquivo.pt.
- Familiarize with the dataset.
- Data Cleaning and pre-processing to ensure data quality and completeness.

T3: User Behavior Analysis (3 months)

- Explore the dataset to understand its structure and potential challenges.
- Analyze user behavior patterns to gain insights into how users interact with the Arquivo.pt search engine. Suggested dimensions include temporal trends, click-through rates, query patterns, session behavior, and user demographics.
- Identify key insights and patterns in user behavior that can result in improvements to the search engine.

T4: Platform for Data Analysis (1 month)

- Make sure results are reproducible by creating the scripts for data analysis.
- Share code and findings on appropriate platforms.

T5: Proposal, Development and Evaluation of Novel Technique (3 months)

- Propose and develop innovative techniques, potentially leveraging machine learning approaches, for possible improvement of the search engine. Options go from developing a system for classifying user queries into different intent categories (e.g., informational, navigational, transactional) to better understand user needs and tailor search results accordingly; content or query recommendation based on user profiles and historical interactions; query log anomaly detection to identify unusual patterns or outliers in the query logs.
- Assess the impact of proposed improvements.

T6: Final Research Documentation and Dissemination (2 months)

- Write the thesis.
- Prepare the presentation.
- Write a scientific paper.

Table 1 presents the distribution of the tasks for the 9-month schedule.

Table 2: Chronology of tasks (T) per month (M).

	M1	M2	M3	M4	M5	M6	M7	M8	M9
T1									
T2									
T3									
T4									
T5									
T6									

Technical and Academic Prerequisites

- **Proficiency in Python**: Strong programming skills in Python for data analysis (e.g., pandas) and potentially machine learning. Familiarity with statistical analysis and data visualization tools (e.g., matplotlib).
- **Information Retrieval**: Solid understanding of Information Retrieval concepts.
- **GitHub and Open-Source Collaboration**: Experience with version control using GitHub and collaboration within open-source communities.
- **Big Data**: Experience (or willingness to learn) big data technologies such as Spark.

Complexity

This proposal outlines an exciting research opportunity at the intersection of information retrieval and data analysis, offering the potential to make substantial contributions to the field and enhance the user experience of Arquivo.pt's search engine. The project involves some level of complexity, as it requires handling a large and complex dataset and developing innovative analysis techniques. Students interested in this project should possess a strong technical background and a passion for information retrieval and data analysis.

Expected Results

- Comprehensive analysis of the Arquivo.pt user interaction logs.
- Identification of actionable insights to improve the search engine.
- Development of innovative analysis techniques.
- Project code and findings shared on Github.
- A written thesis.
- A research paper to be submitted to one of IR conferences: SIGIR; ECIR or CIKM conferences.

Bibiography

[1] Costa, M. (2014). Information Search in Web Archives. PhD. Universidade de Lisboa.

[2] Costa, M. and Silva, M. (2011). Characterizing Search Behavior in Web Archives. In Proceedings of the TempWeb Workshop @ The Web Conference 2011.

[3] Reimer, J., Schmidt, S., Fröbe, M., Gienapp, L., Scells, H., Stein, B., Hagen, M. and Potthast, M. (2023). The Archive Query Log: Mining Millions of Search Result Pages of Hundreds of Search Engines from 25 Years of Web Archives. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'23)

[4] Silverstein, C., Henzinger, M., Marais, H. and Moricz, M. (1999). Analysis of a very large web search engine query log. In ACM SIGIR Forum, vol. 33(1).

[5] Gallois, F. (2023). Analyzing User Search Behaviour in Temporal Web Repositories through Search Query Log Analysis. University of Innsbruck. Bsc of Computer Science.