Web Archive Search Evaluation Metrics

Technical report

Abstract

Web Archive, Search satisfaction, Evaluation metrics, Online and Offline evaluation

Pedro Gomes, Foundation for Science and Technology: Arquivo.pt
pedro.gomes@fccn.pt
Daniel Gomes, Foundation for Science and Technology: Arquivo.pt
daniel.gomes@fccn.pt

In information retrieval (IR), evaluation metrics are crucial to measuring the performance of search engines. Defining metrics is the most powerful tool used in organizations to set long and short-term goals to decide which new products and features should be released to the users. Metrics decide the direction of an organization, and defining the best metrics is one of the most important and difficult problems an organization needs to solve. They try to define concepts such as success and engagement which are abstract and difficult to capture. This work will help us investigate the relationship between the data-driven approach from a commercial search engine (like Google) and a web archive search engine, which have as the most important parameter the time.

Index

Introduction

Defining metrics is the most powerful tool used in organizations to set long and short term goals to decide which new products and features should be released to the users. Metrics decide the direction of an organization, and defining the best metrics is one of the most important and difficult problems an organization needs to solve.

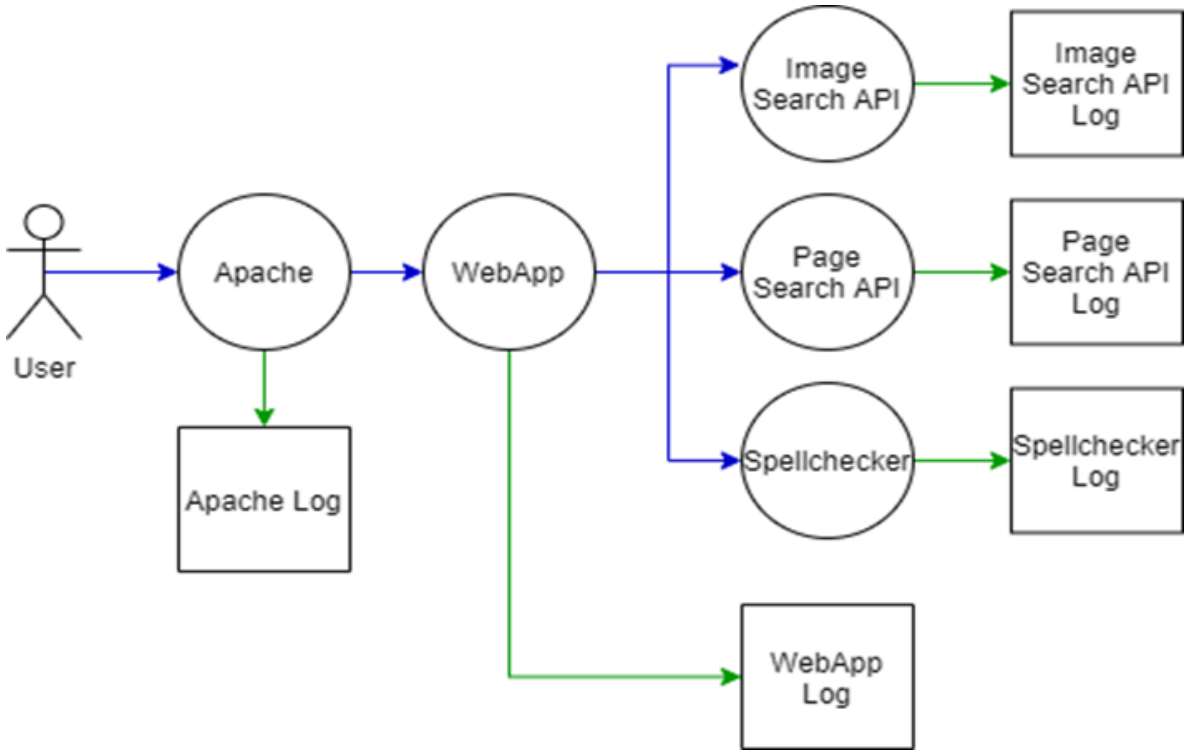
Work phases

1. Define the Log Format for each component;
2. Define general OKRs and KPIs for the organization;
3. Focus on simple and easy metrics related with Page Search;
4. Organize all metrics that are currently implemented in each platform;
5. Develop a set of metrics related with other components (e.g, Image Search);
6. Organize all metrics in one unique platform (e.g., Kibana or Grafana);

Log Formats

Geral

In general, the flow of interactions between the user and the different services are as follows:



Apache

Log Format

The log format follows the Common Log Format from Apache:

"%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-agent}i\""

Example

1.1.1.1 - - [29/Jan/2021:09:39:13 +0000] "GET /textsearch?q=Amadora HTTP/1.1" 200 15615 "http://arquivo.pt/" "Mozilla/5.0" 74441

Image Search API

Image Search API logs, only lines whose Log Format is similar to the one discussed below will be analyzed. The rest of the lines are only used for debugging.

Log Format

IP_Address\t**User_Agent**\t**Request**\t **Duration**\t**Search_Parameters**\t**Search_Results**

- **IP_Address** is a unique address that identifies a device on the internet or a local network.
- **User_Agent** is any software that retrieves and presents Web content for end-users.
- **Request** is the request made to Image Search API.
- **Duration** is the response time of the request.
- **Search_Parameters** are the parameters used in the request.
- **Search_Results** are the results that were returned to the user.

Example

1.1.1.1 Mozilla/5.0 http://arquivo.pt/imagesearch?q=example 160ms search_parameters:{"q":["example"]]} search_results:["1784b0bc37551b4f52fbf3f738e11e8c8903742192f379706f1a3b4f9761d7e3","699cf552fcf76b3d5021018a3dae46bdd4fc4fed7a143d57d4a3e58acd5c842d", ...]

Page Search API

Page Search API logs, only lines whose Log Format is similar to the one discussed below will be analyzed. The rest of the lines are only used for debugging.

Log Format

IP_Address\t**User_Agent**\t**Request**\t **Duration**\t**Search_Parameters**\t**Search_Results**

- **IP_Address** is a unique address that identifies a device on the internet or a local network.
- **User_Agent** is any software that retrieves and presents Web content for end-users.
- **Request** is the request made to Page Search API.
- **Duration** is the response time of the request.
- **Search_Parameters** are the parameters used in the request.
- **Search_Results** are the results that were returned to the user.

Example

1.1.1.1 Mozilla/5.0 http://arquivo.pt/pagesearch/textsearch?q=example 40 ms search_parameters: {"q":"example"} search_results: ["19961013180652/http://www.example.pt/1","19961013203212/http://www.example.pt/2", ..]

Arquivo Webapp

In the Arquivo-webapp Logs, only lines that contain the strings “ImageViewTracking” and “PageViewTracking” will be analyzed. The rest of the lines are only used for debugging.

Log Format

IP_Address\t**User_Agent**\t**Request**\t**Tracking_ID**\t**Session_ID**\t**Timestamp**\t**URL**

- **IP_Address** is a unique address that identifies a device on the internet or a local network.
- **User_Agent** is any software that retrieves and presents Web content for end-users.
- **Request** is the request made to Page Search API.
- **Tracking_ID** is a identifier that tracks the interaction between the user and the systems (e.g., User → Webapp → PageSearchAPI → Webapp → User) being constituted by <user_uid>_<search_id>_<position>.
    - **User_uid** is a hash generated on the client-side that identifies each user on subsequent searches, with a 1-day expiration date.
    - **Search_id** is a hash of the search sent by the client to the API to identify that unique search.
    - **Position** is the n position on the search result (e.g., 1, 2 or 3).
- **Session_ID** is the session ID from the request header.
- **Timestamp** is the timestamp of the page clicked.
- **URL** is the URL of the page clicked.

Example

'1.1.1.1' "Mozilla/5.0" 'http://arquivo.pt/page/view/XX_XX_1/19961013180652/https://example/' 'XX_XX_1' 'SESSION_ID' '19961013180652' 'https://example/'

Metrics

Introduction\Motivation from Papers

Measuring Metrics

“Metrics are a powerful tool used in organizations to set goals, decide which new products and features should be released to customers, which new tests and experiments should be conducted, and how resources should be allocated. To a large extent, metrics drive the direction of an organization, and getting metrics “right” is one of the most important and difficult problems an organization needs to solve. However, creating good metrics that capture long-term company goals is difficult. Metrics often try to capture abstract and subjective concepts such as *success*… These concepts represent real organizational goals for serving their customers, but there’s no standard way to formally define them.”

Meta-evaluation of Online and Offline Web Search Evaluation Metrics

The goal of IR researchers is to build search engine systems which can satisfy users' information needs. Offline metrics are usually based on relevance judgments of query-document pairs from assessors while online metrics exploit the user behavior data, such as clicks, collected from search engines to compare search algorithms. We find that both types of evaluation metrics signific
antly correlate with user satisfaction while they reflect satisfaction from different perspectives for different search tasks. Online metrics better align with user satisfaction in homogeneous search (i.e. ten blue links) whereas online metrics outperform when vertical results are federated.

However, although the offline metrics (e.g., NGCG) may provide easily interpretable outcomes, offline search evaluation has encountered two major problems:

- The editorial judgments are often less credible when measuring actual user experience. Recent studies show that assessors' judgments may significantly differ from users' assessments [31].
- The evaluation results based on offline metrics can be biased because they are usually generated with a small and incomplete dataset [13].

In addition, it is often cheap and fast to collect such data in modern search engines, making it particularly easy to scale up online evaluation and it can suffer from various biases present in typical search logs. Online behavior of users can be affected by many factors, with position bias being the most widely recognized effect, which requires de-biasing when inferring search success. In addition, online metrics may not be as reusable as offline metrics [42].

We found that while online and offline metrics measure users' search experience from different perspectives, they generally both significantly correlate with actual user satisfaction. The top-weighted offline metrics correlate extremely well with user satisfaction in navigational search while online metrics perform comparatively better in informational and transactional search tasks. We demonstrate that offline metrics work better in homogeneous search (i.e., search engine result pages (SERPs)) while online metrics outperform in heterogeneous search environments (i.e., videos or images) since offline metrics mainly rely on relevance assessments while the interaction-based online metrics may be more sensitive to the effect of vertical results. The results show that online metrics can better estimate user satisfaction when mouse hover information is incorporated.

OKRs and KPIs (Font)

OKRs (Objectives and Key Results) is a system used by Google designed to understand the business goals and need to be defined before choosing the specific key performance indicators for the website. After set OKRs, KPIs are used in organizations to measure the state of a system and need to have the following characteristics:

- KPIs are normally defined as a ratio, percentage, or average, allowing data to be presented in context.

- A KPI needs to be key to an organization's success.

Normally, for a small organization, it would be ideal to define 10 KPIs aligning with 10 OKRs or less, having only one KPI report will not cover the requirements of your entire organization. All KPIs need to have values defined.

In all KPIs, It is not correct to use the site-wide average (i.e., too general), thus it is more useful to compare the different sources. For example, compare average time on site and pages for new users versus returning users or to see the difference between users from google vs Facebook. A higher value can be a good thing or not. On the one hand, spending more time on your site and viewing more pages could mean visitors are highly engaged and interested in your content; on the other, they could be confused and lost in your navigation.

In this technical report, we will split into two large OKRs of Arquivo.pt.

Audience engagement

The first OKR is **audience engagement**.

- How much time do people spend on our website?
- How much do people do queries and clicks?
- Will users regress to the site? or are they new?

For instance, if the users read a single page and move on to another site or close the tab, leaving the site is a signal of dissatisfaction. To increase your engagement, we want users to spend more time interacting with Arquivo.pt (e.g., read more pages), in which they will be exposed to more functionalities and content, increasing the likelihood that they will click more times and return to Arquivo.pt next time. We define the following key topics to measure our audience engagement:

- **Bounce rate:**
    - A bounce is when a user arrives on your website, views the page, and then bounces off to another site or closes their browser without doing anything. For instance, the Save Page Now service needs to have a low bounce rate.
    - **Definition:**
        - Is the number of single-page visits with zero interaction divided by the total of website visits.
    - **High**: 50% + , **Medium**: 25 – 50%, **Good**: < 25%
    - High bounce rate pages can be due to:
        - Out-of-date content.
        - Errors on the page.
        - Content or features are not relevant.
- **Average time on site and clicks per query:**
    - The average time on site is the length of time visitors spend interacting with Arquivo.pt, being useful to help understand whether users are engaging.
    - **Definition:**

        - The difference in time between the last and first pageview, since we can not know when the user leaves Arquivo.pt.
- **Percent new users vs returning users and user recency:**
    - Recency is defined as the amount of time that passes between sequential visits. The range of values will depend on each website. Arquivo.pt wise to use:
        - High = within one week
        - Medium = between 8 and 30 days
        - Low = more than 30 days

System Performance

On the other hand, Webmasters have different needs since they have the responsibility for keeping the website running. The OKR is **System Performance**.

- Are the servers overloaded?
- Is the response time good?
- The ranking function is returning quality results?
- What are the features/functionalities more used?

We define the following key topics to measure our system performance:

- **Number of users, queries, and clicks:**
    - Average number of users per time-frame;
    - Average number of unique users per time-frame;
    - Average number of queries per time-frame;
    - Average number of clicks per time-frame;
- **Percentage of users from different countries and languages:**
    - The more insight about the user's demographics the better (e.g., language settings) → https://www.internetworldstats.com/stats7.htm.
- **Percentage of users using different devices:**
    - Web browsers and operating systems render web pages differently. The browser usually has the greatest impact in user experience.
        - https://netmarketshare.com/browser-market-share.aspx
- **Average time to response or to load:**
    - Average response time of the APIs/SOLR/QueryServer and average loading time of the SERPs results.
    - Studies affirm that **two seconds** is the average online website expectation for a web page to load and **79 percent of users who were not satisfied are less likely to return**. The average time to load can also be influenced by connection Speed, old browser versions, and old PCs.
- **Percentage of users receiving an error page:**
    - A target for this metric could be to maintain this level at less than 0.1 percent of our total queries or clicks.
- **Internal search query performance (Page and Image):**
    - To capture the search query performance we can define the following metrics:
        - Percentage of users that use special parameters (e.g., site search);
        - Average number of search results viewed per search;

        - Percentage of queries with at least one click;
        - Percentage of people conducting multiple searches during their visit (excluding multiple searches with the same keywords);
        - Average time in Arquivo.pt for each user (with search);
- **Internal search quality:**
    - To capture the site search's result quality without asking users is difficult, although we can define the following metrics:
        - Percentage of number of zero-result search page;
        - Average position clicked;

Correlation between OKRs and KPIs

In the following table we detail the previous OKRs and KPIs.

OKR	**KPIs**	**Value**	**Results**
To see more traffic	Percentage of user in Arquivo.pt and Sobre: ● General. ● Year by Year. ● Month by Month. ● Day by Day. ● Hour.	> 5% over Year;	Analyze user behavior from different sources and regions: ● Users from Google are more likely to do a query? ● Users from Facebook only see pages? ● Most of our user are from Portugal? ● Increase the budget for pay-per-click campaigns works? Cost per acquisition? Looking for seasonality in Arquivo.pt activity.
	Percentage of visits from external sources (e.g., google or facebook);	> 50%	
	Percentage of visits to Arquivo.pt and Sobre from different geography;	> 10% outside Portugal	
To see visitors engaging with our website more	Percentage of visits that do a query following by a click;	> 80%	Analyze user behavior when interacting with the system. Analyze quality SERPs results. Does our website work in all major browsers?
	Percentage of users complete the task (query and click) in each browser. ● Internet explorer vs Edge vs Chrome vs Firefox	> 80%	

	Percentage of visits that do a query and then leaves;	< 2%	Do users spend a lot of time on Arquivo.pt? ● It's good? ● Does this look like a news site? ● Do we want the user to spend as much time as possible researching?
	Percentage of visits that do a query and then reformulate the query (add or remove terms);	< 5%	
	Average time on site per visit;	< 10 min	
	Average search depth per visit;	< 2 clicks	
	Percentage of visits that see a page and then clicks in one or more versions;	> 50%	
Improve the customer experience and usability	Percentage of visits that bounce (single-page visits);	< 1%	Analyze user behavior and satisfaction: ● Is loading time too high when the user leaves? Analyze the quality of ranking function based on: ● Position clicked; ● Page of the position clicked; ● Task query + click; ● Quick and accurate Is your site easy to navigate?
	Percentage of searches that produce zero results;	< 5%	
	Average click position;	< 5	
	Percentage of click in the first page;	> 50%	
	Average loading time for SERPs and replay page;	< 10 sec	
Improve the UI from the different features in Arquivo.pt	Percentage of clicks in each features, for instance: ● Search Page ● Search Image ● Export Results ● Technical details	> 40 % Search Page	
Coverage of Arquivo.pt	Percentage of URL search not in Arquivo.pt.	< 5%	Understand if the crawls are covering the website that the users are looking for. Test the use of Save Page Now within the context of Arquivo.pt.
	Percentage of users that visits the page “URL Not found” and	> 90%	

	then clicks on the button “Save Page Now”.		
Mobile	Compare all the metric above between Desktop vs Mobile:	-	Analyze user behavior between Desktop and Mobile.
	Percentage of users that switch between desktop and mobile.	< 5%	

Conclusion

- Click Position Average
    - The percentage of clicks in the first position describes the quality of the search. If the user clicks on the first result it is because he analyzed the spinnet, the title, and the URL, which motivated the user to click on the result. If the results are extremely bad, the users would not click on any result. This metric can be influenced by several aspects:
        - Different types of users (e.g., beginner or expert);
        - Interface changes (e.g., color scheme or the number of SERPs results);
        - Different ranking function;
        - Add new collections;
    - The results will help describe the quality of our ranking function.
- Which functions\features are more used in Arquivo.pt?
    - Understanding what are the most used functionalities/features in the Arquivo.pt affects the interface design. For instance, most of the websites in the web have every functionality in one interface since they can not decide which are the best features for each case. In our case, Arquivo.pt makes a progressive presentation of each feature, being a simple, easy, and cleaning interface. It will be also important to know what are the most used parameters in when the users use the advanced search, advanced parameters (e.g., the parameter collection), and the parameters used in the API.
    - The results will help decide with greater awareness what are the functionalities/features that can appear in each interface and improve the design.
- Which domains are more often viewed?
    - It will be important to know what is the most viewed content in Arquivo.pt, in which we could use this information to improve the ranking function.
    - For example, the most viewed content would likely be what the users want (e.g., Google gives higher priority to trusted sites like wikipedia). However, there is an important disadvantage, if the content is already the most viewed even with the current ranking function, if the ranking function was changed to prioritize the most viewed content it could cause an even greater bias (i.e., the most viewed content would have even more views).

    - The results will help to improve the focus of our collections or repair versions of the most viewed websites.
- The users are satisfied?
    - The big question is knowing how to classify user satisfaction, since it is a subjective concept. How can we rate user satisfaction?
        - Make a set of thresholds over a set of metrics?
        - Build a predictive model based on a dataset made by a limited set of people? Will we be able to represent the wide range of users of the Arquivo.pt?
        - User level satisfaction vs Query level satisfaction vs Session level satisfaction will have the metrics?
    - We can only extrapolate if users are satisfied, if we combine all the previous questions as well as more user characterization metrics.
    - The results will help to improve the focus on the behavior of the user on Arquivo.pt and will help to decide which new products/features should be released to the users, and new experiments that should be conducted.

All relevant metrics based on previous papers

Queries

- **Percentage of queries per user\session**.
    - Percentage of queries per user is a metric highly used by researchers and the industry to understand the quality of the ranking function as well as the behavior of users when using the system, being easy to measure.However, this metric can not be used alone as it may have different meanings.
    - For instance, if a user does a lot of queries it can be a good sign since it can mean that the user is using your system a lot. We also have the opposite assumption, if a user does a lot of queries and we do not click on any position it is because the ranking function has some problems since the user needs to do more queries to satisfy the information needs, which can lead to dissatisfaction.
    - Thus, for this metric to be as accurate as possible, it is necessary to relate the following information:
        - Clicks per user;
        - Bounce rate;
        - Type of search (i.e., Navigational, Informational, or Transactional);
        - Use of advanced parameters (e.g., time range);
    - **Goal (normally)**
        - Lower percentage is better.
- **Percentage of query reformations per user\session.**
    - The percentage of queries reformations per user is a more reasonable metric since more reformulations normally means more difficulties for the user to satisfy certain information. Although, depending on the type of query, more queries or

reformation not all is a bad thing (e.g., informational queries). To keep it simple, we are just going to account for addition and removal of terms.
    - **Goal (normally)**
        - Lower percentage is better.

Clicks

- **Percentage of clicks per query, user, and session**
    - The percentage of clicks per query, user, and session will give an overview of how the user interacts with the system and will be key to understanding the quality of our ranking function and our UI SERPs.
    - **Goal (normally)**
        - Lower percentage is better
- **Percentage of queries with a click (i.e., Click-through rate (CTR))**
    - Percentage of queries with a click is the portion of users who clicked on a result when they are in the SERPs.
    - **Goal (normally)**
        - Higher percentage is better since relevant results should be clicked, so we want a higher *CTR*
- **Percentage of queries with a click per user**
    - Percentage of queries with a click per user will help make user clustering.
    - **Goal (normally)**
        - Higher percentage is better
- **Position of the clicks**
    - Analyzing the position of the clicks is the most used metrics in log analysis.
    - For instance, we can measure the average position of the clicks rank of clicked results and our goal is to the most relevant results to be clicked first (i.e., low average rank), but it is not a very reliable metric, since we may have outliers (caused by bots) that cause the position average to be higher. A possible solution would be to use the median instead of the average, and compare between the different days of the week which could indicate different types of users.
    - It would be interesting to know what was the line where the result clicked by the user has been shown in the image search. Will it reveal how the user analyzes the results?
    - **Goal (normally)**
        - Lower the metric is better

Sessions

Session in this analysis will be defined as a sequence of activities followed by one individual to satisfy an information need, regardless of the elapsed time, number of interactions with the system, or the existence of interruptions on these interactions (font).

- **Duration of the sessions**:
    - The duration of the sessions will be calculated with the difference between the first interaction of the user with the system and the last click related to the search

or the next search if it does not belong to the same information need and are closing in time.
    - Example:
        - Same user:
            - (1) Home page
            - (1) Query → “Lisbon”
            - (1) Click
            - (1) After 5 min
            - (2) Query → “History”
            - (2) Click
            - After 5 days
            - (3) Home page
            - (3) Query → “Lisbon”
            - (3) Click
    - This metric is related with:
        - The time spent on the site
        - The engagement
        - The ranking function quality and speed
        - The features used (e.g., the time range implies shorter sessions?)
        - Type of search (informational queries imples longer sessions?)
    - **Goal (normally)**
        - Higher the metric is better
- **Number of sessions per user**.
    - The number of the sessions per user shows the behavior of the user, since the more sessions the user does, the more satisfied users are.
    - **Goal (normally)**
        - Lower the metric is better
- **Percentage of queries per session**.
    - The percentage of queries per session shows how the user interacts with the system since more queries per session the user had difficulties satisfying his information needs.
    - **Goal (normally)**
        - Lower the metric is better
- **Percentage of query reformations per session.**
    - The percentage of queries reformations per session shows the quality of the ranking function since more reformations mean that the ranking function does not return the correct results in the previous query.
    - **Goal (normally)**
        - Lower the metric is better
- **Absence Time (**Here)
    - Absence time of a user is defined by the time between two consecutive sessions of the user.
    - This metric reflects how often and how soon a user is coming back to use the system again.

    - Lower absence time indicates higher user engagement and is an evidence of better satisfaction, helping identify different users (regular vs sporadic).
    - **Goal (normally)**
        - Lower the metric is better
- **Session satisfaction (based Baidu Search Company)**:
    - Model with:
        - Number of click in a session (search outcome)
        - Average query length in a session (search cost)
        - Number of queries without clicks in a session (user effort)
        - The difference of the last and first query in the sum of dwell time (outcome and effort change)
    - **Goal (normally)**
        - Lower the metric is better

Time

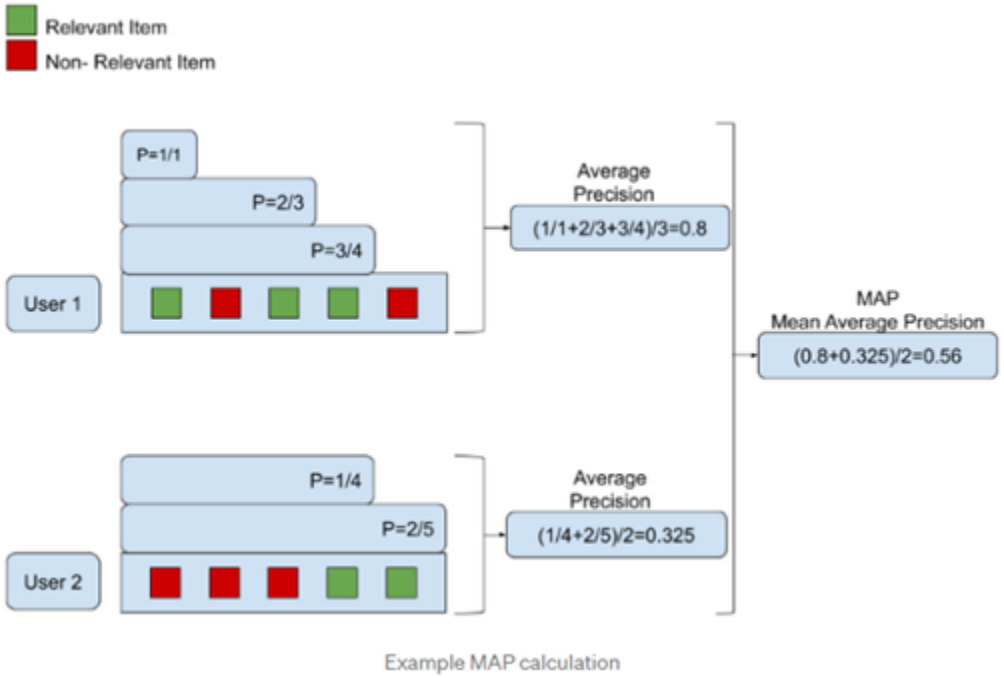
- **Dwell Time.**
    - Dwell time is the amount of time that goes by from the moment a user clicks on a search result to the moment they return to the search engine results pages (SERPs). Pages abandoned too quickly may not present much satisfactory information.
    - **Time to Long Click**, where we define “long” as the user not returning to the search engine for at least 30 seconds after the click. Intuitively, a long click should be a better indicator of the user actually finding what they wanted.
    - **Goal (normally)**
        - Longer the metric is better
- **Time to Click**.
    - The better the results are and the more clear the page is, the sooner the user will be able to decide where to click.
    - **Goal (normally)**
        - Lower the metric is better

Offline metrics

The offline metrics are usually used to do relevance judgments based on a dataset to draw conclusions about ranking functions. We can divide in the following metrics:

- **Precision and Recall**
    - **Precision** → “is the fraction of relevant instances among the retrieved instances.”
    - **Recall** → “ is the fraction of relevant instances that were retrieved.”
    - Technically we can not calculate Recall since the correct results may never appear.
    - **Precision** is to find the most relevant results while recall is to find all the relevant documents.
- **Mean Average Precision (MAP)**

- Measures the relevance of each item in the results list to the user's query with a specific cutoff N. The top N number is usually chosen arbitrarily or based on the number of paginated results. MAP is calculated by averaging the AP scores for each query in our dataset. The result is a measure that penalizes returning irrelevant documents before relevant ones. Normally, It can be very deep (i.e., it can take into account relevant results in position 500). However, in our case 500 results is not very important (35/50 for pages and 100/150 for images will be right).



Example MAP calculation

- **ERR (Expected reciprocal rank)**:
  - Expected reciprocal rank is based on the cascade model of search. The cascade model assumes a user scans through ranked search results in order, and for each document, evaluates whether the document satisfies the query.
  - One problem remaining for this metric (and others) is correlated documents. Lots of queries are ambiguous. For instance, consider the query "john langford", which is highly ambiguous.
  - https://github.com/skondo/evaluation_measures
  - https://github.com/scikit-learn/scikit-learn/issues/16813
- **Rank-Biased Precision (Font)**
  - RBP deals with unjudged documents by offering uncertainty in the evaluation, providing a range covering the evaluation scores that would have been obtained if the unjudged documents were relevant or irrelevant. The evaluation is based on a user persistence model that requires the choice of a user persistence value before the evaluation can take place.
- **NDCG for recommendation systems (Font)**
  - Widely used to measure the quality of ranking algorithms (i.e., recommendation systems), since good ranking results correlates with user satisfaction. By

definition, NDCG measures the similarity between a list of results with a perfect order.
    - The biggest advantage of the NDCG compared to other metrics is that the NDCG further tunes the recommended lists evaluation since it is able to use the fact that some documents are more relevant than others.Highly relevant items should come before medium relevant items, which should come before non-relevant items.
    - NDCG is the ratio (range [0,1]) between Discounted Cumulative Gain(DCG) of recommended order shown to the user by the system and iDCG, which is the ideal order (scikit-learn.org).

All the metrics described above are defined in this python package.

Conclusion Offline metrics:

The best metric is NDCG since it takes into account the graded relevance values using more than the binary relevant/non-relevant annotations used by MAP. However, as it is possible to make several levels of partial relevance, it can become subjective depending on the user.
It is possible that the iDCG is equal to zero (there are no relevant results), in this case NDCG is zero.

Benefits offline metrics:

- Quality control of annotations;
- Uses a set of relevance metrics based on a test collection to give an overview of the quality of the IR system.

Disadvantages offline metrics:

- Need to build a test collection.
- User satisfaction does not depend only on the clicked position.

Top quality system metrics

1. **Time to loading:**
    a. **Results (SERPs);**
    b. **Replay (Wayback);**
2. **Response time (APIs)**
    a. Information available on imagesearch.log and pagesearchwebapp.log
3. **Percentage of click on the query suggestion:**
    a. /spellchecker/checker?query=lisboa&l=pt HTTP/1.1" 200 227 "https://preprod.arquivo.pt/page/search?hitsPerPage=10&query=lisboa&l=pt&ion-dt-0=1996-01-01&ion-dt-1=2021-07-13&dateStart=01%2F01%2F1996&dateEnd=13%2F07%2F2021&**spellchecked=true**"
4. **Number of requests 200 vs !200 status code (wayback).**

List of functions\features in Arquivo.pt

1. **Advanced Search** can be identified by the following keywords in bold:
    a. “**/page/advanced/search**” or “**/image/advanced/search**”
    b. /page/search?l=pt&adv_and=fccn&adv_phr=&**adv_not=**&ion-dt-0=1996-01-01&ion-dt-1=2021-07-12&dateStart=01%2F01%2F1996&dateEnd=12%2F07%2F2021&format=all&site=&hitsPerPage=10&btnSubmitBottom=Pesquisar+no+Arquivo HTTP/1.1" 200 9077 "https://preprod.arquivo.pt/**page/advanced/search**?l=pt"
2. **Table/List Versions:**
    a. Missing distinction (dev.arquivo.pt solve the problem)
3. **Change Language (English):**
    a. Changing language is always a new request (new textsearch, new spellchecker..)
    b. https://preprod.arquivo.pt/page/search?hitsPerPage=10&query=card&**l=pt**&ion-dt-1=1996-01-01&ion-dt-0=2021-07-12&dateStart=01%2F01%2F1996&dateEnd=12%2F07%2F2021"
4. **Copy Link:**
    a. Missing
5. **Sobre:**
    a. I can only know if the user clicked the “Sobre” button if they are going to analyze the wordpress logs.
6. **Replay Options**:
    1. **Technical details** can be identified by the following keywords in bold:
        i. /textsearch?**metadata=**http%3A%2F%2Fwww.publico.clix.pt%3A80%2F%2F20060104061100
    2. **Save** can be identified by the following keywords in bold:
        i. **/screenshot/**?url=https%3A%2F%2Fpreprod.arquivo.pt%2FnoFrame%2Freplay%2F20190319211818%2Fhttp%3A%2F%2Fdef.pt%2F&**width=1920&height=1080** HTTP/1.1" 200 20432 "https://preprod.arquivo.pt/wayback/20190319211818/http://def.pt/"
    3. **Print** can be identified by the following keywords in bold:
        i. **/screensho**t?url=https://preprod.arquivo.pt/noFrame/replay/20060104061100/http%3A%2F%2Fwww.publico.clix.pt%3A80%2F&**download=false** HTTP/1.1" 200 935191 "https://preprod.arquivo.pt/wayback/20060104061100/http://www.publico.clix.pt:80/"
    4. **Complete Page** can be identified by the following keywords in bold:
        i. /noFrame/patching/record/
    5. **Full screen** can be identified by the following keywords in bold::
        i. /noFrame/replay
    6. **Old Browser** can be identified by the following keywords in bold:
        i. /wayback/static/img/old-browser-icon-blue.svg
7. **Export Results:**
    a. The only way is to check the loading of the image and we can not see which one you clicked:
        i. /img/export-results-hover.svg

8. **Not Found:**
    a. /url/search/20141123221153/http://publico.pt
9. **Search Other Archives:**
    a. Missing

All Online Metrics

Geral

Geral métrics:

- Top most frequent queries (diversification of queries\topics)
- Word cloud of users' queries for three salient social events;
- Topics of user's queries;
- Number of users per hour, per day, per month, and per year (Average and Median);
- Number of queries per hour, per day, per month, and per year (Average and Median);
- Number of pages clicked per hour, per day, per month, and per year (Average and Median);
- Number of sessions per hour, per day, per month, and per year (Average and Median);
- Percentage of queries without clicks;
- Percentage of queries from Page search to Image Search (vice versa)

User

User métrics:

- Number of users (i.e., per hour, per day, per month, and per year), and general métrics (i.e., Average and Median);
- Queries per user (i.e., per hour, per day, per month, and per year), and general métrics (i.e., Average and Median);
- Clicks per user (i.e., per hour, per day, per month, and per year), and general métrics (i.e., Average and Median);
- Sessions per user (i.e., per hour, per day, per month, and per year), and general métrics (i.e., Average and Median);
- Duration of the visit:
    - http://logs.arquivo.pt/awstats/awstats.pl?config=arquivo.pt&framename=mainright#sessions

Query

- Query length distribution;
- Terms per query distribution;
- Query frequency distribution;
- Distribution of query count over per hours (24h), per day, per month, and per year;

- Geographical distribution of the number of queries:
    - http://logs.arquivo.pt/awstats/awstats.pl?config=arquivo.pt&framename=mainright#countries
- Weekdays query count distributions;
- Number of terms per query;
- How often distinct queries are asked;
- Distribution of the number of clicks per query;
- Percentage of unique queries;
- Percentage of unique terms;
- Queries never repeated;
- Terms never repeated;
- Percentage of reformulated queries using Bing Query Suggestion;
- Percentage of types of queries (navigational, transactional, and informational) and related with previous metrics;

Devices

- Distribution of query count over 24-hours for both mobile and Web users;
- Geographical distribution of the number of queries posted by both mobile and Web users.
- Weekdays query count distributions for both mobile and Web users;
- Percentage of mobile sessions vs browser sessions;

Session

- The average of session duration;
- The average number of queries per session.
- The average number of clicks per session.
- % Modified queries per Session (Modified, Identical, Terms Swapped, New);

Advanced Queries

- Advanced queries vs normal queries;
- Most used parameters;

Example reference

1. Request from google with query already in place:
    a. 172.16.10.90 - - [05/Apr/2021:11:44:55 +0100] "GET /images.jsp?query=bola&dateStart=01/01/1996&dateEnd=31/12/2018&pag=prev&start=2230&l=pt HTTP/1.1" 302 - "https://www.google.com/" "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/89.0.4389.114 Safari/537.36" 664
2. Request from google without query already in place (“Google is lost”):

a. 172.16.10.90 - - [05/Apr/2021:12:07:11 +0100] **"GET / HTTP/1.1"** 200 9463 **"https://www.google.com/"** "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/89.0.4389.114 Safari/537.36" 3663
b. 172.16.10.90 - - [05/Apr/2021:12:07:17 +0100] **"GET /page/search?hitsPerPage=10&query=sim&l=pt&ion-dt-1=1996-01-01&ion-dt-0=2021-04-05&dateStart=01%2F01%2F1996&dateEnd=05%2F04%2F2021 HTTP/1.1"** 200 9070 **"https://arquivo.pt/"** "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/89.0.4389.114 Safari/537.36" 5405

Questions?

- Session Satisfaction is given by the average of the values from Query Satisfaction?
  - No, it depends. For instance, queries reformation.
  - Why the average? Why not the satisfaction of the last query?
- User Search Intent Classifier (Navigational, Informational, and Transactional):
  - Navigational: “fccn” ou “fccn.pt”
  - Informational: “covid”
  - Transactional: “hotel afonso III Albufeira”
- Log mouse tracker (scroll, not scroll);
- Bounce Rate (Google Analytics)
- https://en.wikipedia.org/wiki/Web_analytics
- Evaluation 14: query logs and click deviation → https://www.youtube.com/watch?v=-hVq7OCWqDE
- Is it smart to drop user-agent entries? without IP?
- Is it interesting to identify questions in queries?
- Is it interesting to identify query language, arquivo.pt request language, language of the browser?
- Attributes of a Page? Readability? Accessibility? Number of elements present (Percentage of links 200 / (links 200 + links 404)?
- Number of times you went back? (Click Undo)
- How to know that a user has left Arquivo.pt?
- https://github.com/arquivo/pwa-technologies/issues/1146

Conclusions from papers

- “We see that de-dupped versions (Remove Duplicate Queries) of metrics perform better on most evaluation criteria, for most metrics.” Here
- “The metrics based on click behaviors in general correlates more weakly with user satisfaction, compared with dwelltime-based metrics. This may be because a clicked result does not always necessarily mean a high quality document hence the click-based metrics may fail.In contrast, some metrics based on scroll (MaxScroll) and dwelltime information (SumClickDwell, QueryDwellTime and TimeToLastClick) have stronger

(moderate) negative correlation with user satisfaction, which means scrolls and dwelltime information are quite important behavior signals to infer user satisfaction." Here

- "To evaluate a search system, satisfaction can be considered as regarding not only to the whole search experience but also to some specific aspects [46], such as the precision or completeness of search results, response time and so on." Here

In this framework, both the benefit factors (document relevance) and cost factors (the effort users spend on examining search engine result pages (SERPs) and landing pages) are used to estimate satisfaction.

Fonts

Measuring Metrics
Meta-evaluation of Online and Offline Web Search Evaluation Metrics
On Correlation of Absence Time and Search Effectiveness
Building Smarter Search Products: 3 Steps for Evaluating Search Algorithms Shopify
Search Engine Pictures: Empirical Analysis of a Web Search Engine Query Log
A Search Log Analysis of a Portuguese Web Search Engine
How Questions are Posed to a Search Engine?

Interesting repository

https://github.com/varepsilon/clickmodels
https://github.com/thukg/query-intent-classification
https://github.com/reza-sohrabi/Intent-Classification
https://github.com/iai-group/ecir2018-intents
https://github.com/MasonCaiby/custy_sat
https://github.com/jiepujiang/ir_metrics
https://github.com/deepmipt/DeepPavlov

- https://demo.deeppavlov.ai/#/en/ner

https://github.com/open-guides/og-search-engineering
https://github.com/mitchellkrogza/nginx-ultimate-bad-bot-blocker/commit/629429aacc9f688216b72353fbf8015284ad1558
https://github.com/vyskoto4/SBot
https://www.kaggle.com/datasets?search=URL
https://www.kaggle.com/remosin/bot-detection/code
https://www.kaggle.com/surya0307/weblogdata
https://www.kaggle.com/ramyaa98/nasa-access-log
https://datasetsearch.research.google.com/search?query=Server%20logs&docid=IfKiWhSR5923vOSqAAAAAA%3D%3D
https://webscope.sandbox.yahoo.com/
http://www.thuir.cn/data-sigir18-UserStudy/

Related with Bots search logs:

https://dl.acm.org/doi/10.1145/1772690.1772742
https://dl.acm.org/doi/pdf/10.1145/1718487.1718540
→ https://github.com/vyskoto4/SBot

(LogStash, Elastic, Kibana) vs (Mysql, Grafana)

Both Kibana and Grafana are powerful visualization tools. However, at their core, they are both used for different data types and use cases. Grafana, together with a time-series database such as Graphite or InfluxDB is a combination used for metrics analysis; on the other hand. Kibana is part of the popular ELK Stack, used for exploring log data.

https://www.youtube.com/watch?v=xXmOmFyN3Hs
https://logz.io/blog/grafana-vs-kibana/

Grafana

Main Features:

- Metrics Analysis :
    - E.g., Number of request per day; Position Clicks;
- Monitoring:
    - E.g., CPU; Memory; Disk; I/O utilization;

Benefits:

- Ease of creation and implementation;
- TimeSeries Data;
- Separation into different dashboards;
- Different types of source;
- It can also motorize the system;
- Possibility of having alerts;

Disadvantages:

- Difficult to display queries;
- Make queries;

Kibana

Main Features:

- Explore the data:
    - E.g., Queries (text\sql) on the data (exp: statuscode:200); Word clouds;
- Troubleshooting, forensics, development, security;

Benefits:

- Queries display;
- Make queries;
- TimeSeries Data;

Disadvantages:

- Difficult system creation and implementation;
- Data can only come from elasticsearch;
- There are no alerts;

More detail

Installation:

- Both Kibana and Grafana are pretty easy to install and configure. Since Kibana is used on top of Elasticsearch, a connection with your Elasticsearch instance is required.
- Grafana is configured using an .ini file which is relatively easier to handle compared to Kibana's syntax-sensitive YAML configuration files.

Data Source:

- Grafana was designed to work as a UI for analyzing metrics. As such, it can work with multiple time-series data stores, including built-in integrations with Graphite, Prometheus, InfluxDB, MySQL, PostgreSQL, and Elasticsearch, and additional data sources using plugins.
- Kibana on the other hand, is designed to work only with Elasticsearch and thus does not support any other type of data source.

Query:

- Querying and searching logs is one of Kibana's more powerful features. Using either Lucene syntax, the Elasticsearch Query DSL or the experimental Kuery, the data stored in Elasticsearch indices can be searched with results displayed in the main log display area in chronological order. Lucene is quite a powerful querying language but is not intuitive and involves a certain learning curve.
- Grafana, users use what is called a Query Editor for querying. Each data source has a different Query Editor tailored for the specific data source, meaning that the syntax used varies according to the data source. Graphite querying will be different than Prometheus querying, for example.

Dashboards and visualizations:

- Both Kibana and Grafana boast powerful visualization capabilities.

- Kibana offers a rich variety of visualization types, allowing you to create pie charts, line charts, data tables, single metric visualizations, geo maps, time series and markdown visualizations, and combine all these into dashboards. Dashboards in Kibana are extremely dynamic and versatile — data can be filtered on the fly, and dashboards can easily be edited and opened in full-page format. Kibana ships with default dashboards for various data sets for easier setup time.
- Grafana dashboards are what made Grafana such a popular visualization tool. They are infamous for being completely versatile. Visualizations in the software are called panels, and users can create a dashboard containing panels for different data sources. Grafana supports graph, singlestat, table, heatmap and freetext panel types. The software's users can make use of a large ecosystem of ready-made dashboards for different data types and sources.
- Both Grafana and Kibana offer many customization options that allow users to slice and dice data in any way they want. Although, the grafana has a wider array of customization options and also makes changing the different settings easier with panel editors and collapsible rows.

Alerts:

- Grafana has shipped with a built-in alerting engine that allows users to attach conditional rules to dashboard panels that result in triggered alerts to a notification endpoint of your choice (e.g. email, Slack, PagerDuty, custom webhooks).

Problems?

- Is MYSQL the most suitable? due to being a relational database?