Keywords for Data Discovery

Lightning Talk

Brigitte Mathiak, Fidan Limani, Yousef Younes

14.9.2021

Introduction

- Data discovery is a complex process [1]
  - 75% of researchers often rely on literature review
  - 59% of researchers rely on search engines
  - 41% use domain data repositories
- We are looking at keywords
  - Keywords are important for both finding and providing data
  - Keyword position affects (webpage) dataset ranking
- Methodology
  - Look at search queries made in web search
  - Cluster them to gain insight

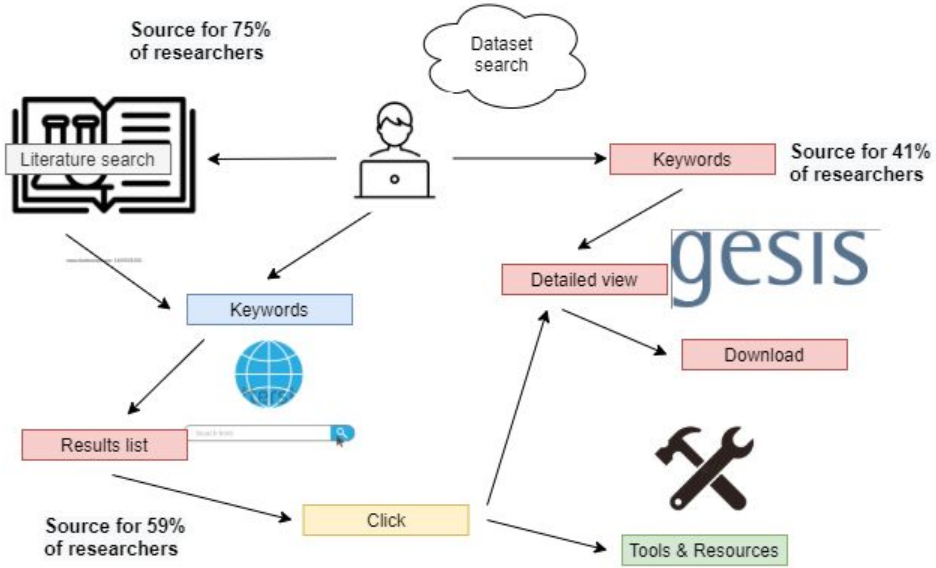


Figure 1. Data search paths

Data

Data Source	Data type	Format	Collection size
GESIS https://www.gesis.org/	Search queries	CSV	1.000 queries
DBK https://dbk.gesis.org/	Search queries	CSV	1.000 queries
DZWH https://www.dzhw.eu/	Search queries	CSV	1.000 queries
SSOAR https://www.gesis.org/ssoar/home	Dataset mentions from papers	JOSN	63426 items

Table 1. Data sources for the analysis

Clustering Based on CCTR

Preprocessing

- Tokenization
- Remove stop words
- Remove punctuations and numbers

Calculating CCTR

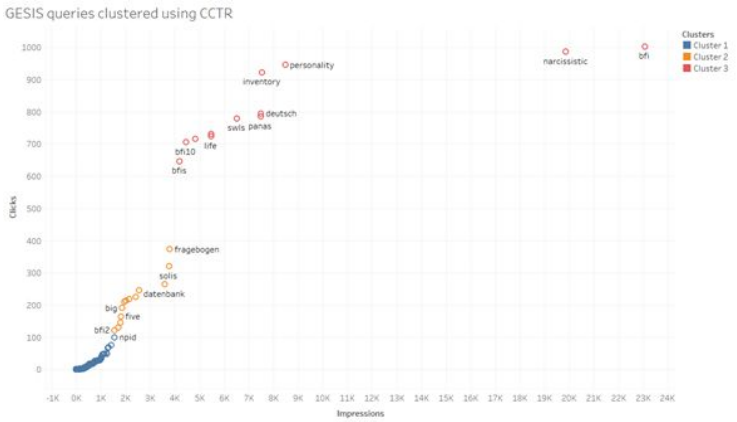
For each keyword, compute normalize the number of clicks into the range [0,1] and the number of impressions into the range [1,2].

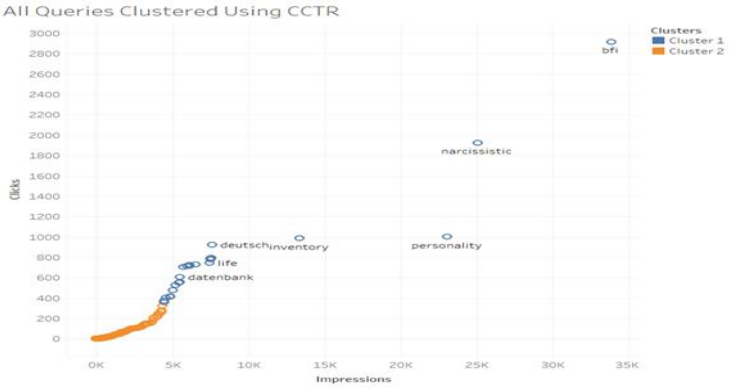
Compute the CCTR= NC/NI; where NC stands for normalized clicks and NI= normalized impressions.

Clustering

Cluster the keywords using K-means [2] with Calinski-Harabasz [3] depending on their CCTR.

Analyzing Queries Using CCTR





Result

- Names of datasets and abbreviations/acronyms as well as name components are very common
- Specifier terms such as “Question”, “Questionaire”, etc. are quite common as well
- Descriptive subject terms are fairly rare

Take Home Message

- Names of datasets and abbreviations/acronyms are very important for web findability
- Do use specifier terms, such as “dataset”, to allow disambiguation of search terms
- For German Social Science these terms are:

  Frage, Fragebogen, Daten, Skala, ...

References

[1] Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. (2020, April 30). Lost or Found? Discovering Data Needed for Research. Harvard Data Science Review, 2. doi:10.1162/99608f92.e38165eb

[2] Ralambondrainy, H., 1995. A conceptual version of the k-means algorithm. Pattern Recognition Letters, 16(11), pp.1147-1157.

[3] Halkidi, M., Batistakis, Y. and Vazirgiannis, M., 2001. On clustering validation techniques. Journal of intelligent information systems, 17(2), pp.107-145.

Questions / Comments

Thank you!

brigitte.mathiak@gesis.org

Notes

- Dates
  - The event's programme is published: Lightning talks are scheduled on Sep 21 and 22;
- What is the focus of OSFair 2021?
  - "OS Fair 2021 aims to bring together and empower open science communities and services; to identify common practices related to open science; to see what are the best synergies to deliver and operate services that work for many; and to bring experiences from all around the world and learn from each other"
  - #toDo Try to highlihght practices or findings from our research that aligns with Open Science;

- Lightning talk format
  - "For lightning talks we will apply the 24/7 rule: 7 minute presentations comprising no more than 24 slides. Successful presentations are fast paced and have a clear focus on one idea."

Notes (2)

Today's SEO focuses on #optional (https://www.searchenginejournal.com/seo-101/why-is-search-important/#close)

- Understanding personas
- Data-driven insights
- Content strategy
- Technical problem solving.

Our research relates to these SEO foci to a different extent;