

Examples of search algorithms used in information retrieval*

Jaroslav Ertl

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií
`xertlj@stuba.sk`

1. Novemeber 2023

Abstract

Many people use search engines on a daily basis for information retrieval, whether it's for school, work or personal use, but few understand the complex algorithms that power these systems. In the realm of Information Retrieval (IR), search algorithms play a crucial role in delivering relevant results from vast datasets. There are many types of search algorithms used in IR, each of them having its own unique purpose.

You can have a graph-based algorithm, keyword-based algorithm and many more. In this article, we will dive into the world of different search algorithms and their application in IR. In each section, we will discuss different search algorithm, how they work and their purpose.

You can find the article that guided me towards the topic of search algorithms in references.[2]

1 Introduction

In the digital age, the internet has become a crucial part of our lives and search engines have become our number one trusted companions. Whether we're seeking for answers, writing research for academic purposes or simply satisfying our unending curiosity. But have you ever stopped and thought how did that keyword entry had turned into relevant results you were seeking for? While the act of typing a keyword into a search bar may seem straightforward, beneath the surface lies a world of complexity, driven by search algorithms.

In this article, we embark on a voyage through the realm of search algorithms. Our first destination on this journey will be the island of graph-based algorithms, where we will delve into the details of Hyperlink-Induced Topic Search (HITS) and the revolutionary PageRank algorithm, developed by Google's founders, Larry Page And Sergey Brin. These algorithms have put down the foundation of web search and have transformed the way we discover information online.

From there, we will set sail for the world of keyword-based algorithms. At this stage of our adventure, we will uncover the mysteries of the Vector Space

*Semestrálny projekt v predmete Metódy inžinierskej práce, ak. rok 2023/24, vedenie: Ing. Richard Marko, PhD.

Model (VSM), a mathematical structure that underpins the way documents and queries are represented in the digital space. The VSM allows us to understand how documents are transformed into mathematical vectors and how these vectors are used to determine the relevance of documents to a search query. We will also navigate through the logic of Boolean Search, which includes logical operators to retrieve documents that correspond to our search criteria. And let's not forget Term Frequency-Inverse Document Frequency (TF-IDF), a cornerstone of keyword-based algorithms. This algorithm has paved the way for innovative approaches in information retrieval. In particular, it laid the foundation for the BM25 search algorithm, which has become one of the most common and influential algorithms used in many search engines today. We will explore the core principles of TF-IDF and how BM25 builds upon this concept to deliver relevant search results to users.

References

- [1] T. Chou. Hits algorithm: Link analysis explanation and python implementation. *Towards Data Science*, 2021.
- [2] W. Zhang, R. Shang, Z. Li, R. Sun, and J. Du. Personalized web page ranking based graph convolutional network for community detection in attribute networks. *IEEE Access*, 11:84270–84282, 2023.
- [1]