Machine Learning in Search Engines: Improving Information Selection for Users*

Marián Kurčina
Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií
xkurcinam@stuba.sk

30. október 2023

Abstract

In an era of the internet, search engines have become the gateway to knowledge. They play a prime role in connecting users with the information they seek, but achieving optimal information selection remains a difficult challenge. From personalized search results that align with individual preferences to context-sensitive responses that adapt to specific needs, machine learning is driving improvements in relevance and user satisfaction. This article delves into the impact of machine learning in search engines and how it is changing the way information is selected and presented to users. We will explore the evolution of search engines, the integration of neural networks, and the incorporation of user behaviour data in shaping search outcomes.

1 Introduction

In today's digital age, the internet has revolutionized how we access information. With the projected surge of data worldwide from 33 zettabytes in 2018 to 175 zettabytes by 2025 [2], it can be a daunting task for the average online user to go through and locate the desired information. From the beginning stages of the internet, search engines have been essential in providing users with the ability to search and filter the information available online. However, with the exponential growth in the amount of data available, search engines have had to adapt to meet the ever-changing demands of users by providing the most relevant and accurate results. The solution to this problem was found through the implementation of machine learning. The integration of machine learning in search engines has revolutionized the search process, providing users with personalized results based on their preferences. In this article, we will discuss how machine learning shaped search engines, but first, we have to look into what search engines are and how they work.

2 Search Engines

Search engines [7] are tools that enable users to locate information on the internet. These search engines provide results that can consist of various types of content, such as web pages, images,

*Semestrálny projekt v predmete Metódy inžinierskej práce, ak. rok 2023/24, vedenie: Ing. Mohammad Yusuf Momand, MSc.

videos, files, and other types of information. The thousands of web pages are processed and presented in the form of a list called "Search results". The search results can contain thousands of pages with the most suitable being at the top. Working of a search engine consists of 3 main processes:

- web crawling
- indexing
- searching

Web crawling [4] is a process in which a web crawler, program/software or programmed script, systematically browses the Internet. The Internet has a graphical structure, which means the page may be connected to other pages through URL links. A web crawler utilizes a graphical structure to navigate the vast expanse of the Internet. During its journey, it may discover and visit new web pages, which are subsequently added to a local repository. While traversing the World Wide Web, a web crawler systematically retrieves information about various web pages and subsequently archives it in its designated database (indexing). Indexing is the most important part of the whole process with size of the indexes of the biggest search engines in billions.1 The last part of the searching process is searching itself. The browser uses a search interface, which connects the user and the database, it helps with searching through the database and provides the user with a search result page.

Search engine	Reported size (billion)
Ask Jeeves	2.5
Google	8.1
MSN Search	5.0
Yahoo! (estimate)	4.2

Figure 1: Reported size of each Web search engine index [9]

2.1 Types of search engines

There are 4 types of search engines. [5]They are:

- **Crawler based search engines**

  The listings are generated automatically by crawling through the internet. Changes made to web pages can affect the listings and the rank they will be presented at in search results. The examples of crawler based search engines are Google and Bing.

- **Human powered directories**

  Listings depend on human submissions made by the page owner. The submissions contain an address, title and a description of their site. These submissions are then approved or denied by editors. The advantage of this type of search engine is a better quality of web pages, but on the downside, after committing any changes to the page it has to go through a process of submitting again.

- **Hybrid search engines**

  Hybrid search engines use features from crawler-based search engines as well as from human-powered directories. This makes the search process more effective.

- **Meta search engines**

  Metasearch engines use search results from other search engines. The results are combined and listed according to their relevance across all used search results. MetaCrawler is one example of a Meta search engines.

3 History of search engines

Historical connections. The first tool [11] for searching through content online, named Archie, was created in 1990 by Alan Emtage, Bill Heelan and J. Peter Deutsch. Archie did not index the contents of the sites, with the amount of data that wasn't needed, it was possible to search manually. Then in 1993, Matthew Gray produced the first web robot, World Wide Web Wanderer, which he used to generate an index called "Wandex". His purpose was to measure the size of the World Wide Web. The same year the Aliweb search engine was created. This search engine as Archie wasn't using a web robot, instead, it depended on notifications from website administrators. In late 1993 the Jump Station was created, a search engine which used a web robot to find web pages and to build its index. The next advancement came in the year 1994 when the WebCrawler was created. Unlike its predecessors it allowed its users to search for any word on a webpage. It was also the first one to be widely known by the public. During the early 2000s, Google emerged as the leading player in the market. The company achieved better search results with an innovation called PageRank [8].

4 Machine Learning

The solution to better and more relevant information selection is machine learning. Machine learning is a branch of computer science that makes computers learn without being directly programmed. It learns by improving its performance by working with training data and adapting according to its results. (As shown in the Figure 2) After learning it is supposed to be able to work with any set of data provided to it. There are 3 main types of learning [6]:

- **Supervised learning**

  With this type of learning the value of outcome is known for each observation.

- **Unsupervised learning**

  Opposite to supervised learning, the outcome values are not known. An algorithm has to identify common characteristics within data without any reference.

- **Reinforcement learning**

  The system learns through trial and error over time and gets either positive or negative feedback according to their performance. through time it self adapts and gets better at a task.

The learning of a system itself is done through learning algorithms. The most famous algorithm used for machine learning is an artificial neural network which will be briefly introduced to you below.

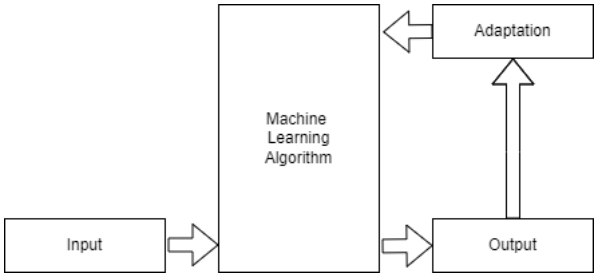

Figure 2: Machine learning process

4.1 Neural Networks

Artificial neural networks [12] are inspired by the functioning of the brain and its neurons (therefore its name). The network is made of 3 types of neuron layers: input layer, hidden layer and output layer. Neurons are connected and based on a connection strength they give output to another layer of neurons. The learning process consists of letting the system change connection strength between neurons according to the output. Learning can be supervised, unsupervised or also reinforced.(An example of such a network can be seen at figure3)

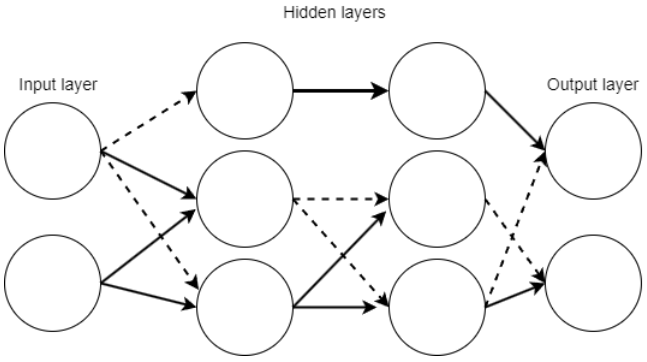

Figure 3: Artificial neural network
(dashed lines represent weaker connection)

5 Applications of Machine Learning in Search Engines

Machine Learning can be applied in multiple ways to improve searching experience [10]:

- **Better search ranking results**
- **Query understanding**
- **Weider search selection**
- **More relevant ads**
- **Pattern recognition**

- **Detecting unwanted messages**
- **Search related features**
- **Better crawlers**
- **More effective work**

5.1 Better search ranking results

[10]Machine learning can substantially improve search results, by making them more personalized, by detecting duplicates and stolen content. Without machine learning, the law would have to be reinforced manually, which would be practically impossible. Identifying stolen content is this way automatic. Personalization plays a dramatic role in better search results. Providing a user with personalized content searching for desired content takes less time and makes searching more pleasant.

5.2 Query understanding

[10]While searching and writing expressions in a search engine, the search engine needs to understand what is the user trying to ask. If it can't understand the question the search results won't be helpful. This is where machine learning helps. Machine learning can identify and correct spelling mistakes and it can identify and work with synonyms.

5.3 Weider search selection

[10]Thanks to machine learning, text searching isn't the only search option. Machine learning can understand multiple other inputs, like photo search, music search or voice search. Through long learning, search engines now provide more ways to search. By taking a photo, the search engine finds similar photos and identifies an object. It makes searching also accessible to people who find it difficult to write, by having a build in feature that converts a speech to text. Machine learning takes a part in that too.

5.4 More relevant ads

[10]The companies owning search engines get revenue from advertisements that they display on their websites. By showing more relevant ads to the users the chance of them clicking on them is higher which is beneficial for the search engines, but also for the company providing the product or service, with their sales being greater. Machine learning can identify the correct group to target and then show them the advertisement.

5.5 Pattern recognition

[10]Search engines use machine learning to detect patterns, which can help in a multitude of ways. Search engines can recognize spam and unwanted messages, but also they're able to recognize handwritten text and other patterns which help with analysing the content on the web pages and thus can eliminate low-quality pages from its ranking page.

5.6 Detecting unwanted messages

[10]As discussed in the last paragraph, machine learning can detect unwanted content and eliminate it before it gets to the user. One of the biggest problems with children having access to the internet is the explicit content they might be exposed to. The search engines allow parents to turn on a kids mode which filters content and removes anything not suitable for young.

5.7 Search related features

[10]Machine learning is used to enhance the search experience, with search features like related searches, recommended content or user classification. With machine learning content is recommended to users according to the stuff they would enjoy. User classification helps with that. The search engine gets to know what kind of user you are, which is then used to recommend more precise content for them.

5.8 More effective work

People and work. I see a possibility of machine learning helping with work and planning. Machine learning could help with planning a sprint or even replacing a scrum master. It is capable of doing small tasks and it can help with information search which would make work more easy and more effective. All the said features can help with workflow.

6 Compications with implementation

Sustainability and ethics. [3]As with everything, the implementation of machine learning and AI in search engines has some complications. The learning process and the actual running of the programs are costly. If Google was to change their whole system to AI it would cost them tens of billions of dollars. Afterwards, the cost of electricity required for its running would be another problem. The estimated power consumption would be around 29.2 terawatt-hours per year. This consumption level is equivalent to the consumption of Ireland and would be double the company's total energy consumption of 15.4 terawatt hours in 2020 [1]. With the world trying to be more eco-friendly, this change could generate a lot of public rage.

7 Conclusion

In conclusion, I think the addition of Machine learning in search engines has many advantages even though the troubles of its implementation. Since its implementation, there have been many ways it has helped with searching and it is a perfect solution to increase of data available on the World Wide Web.

References

[1] Google environmental report 2021. Technical report, Google, December 2021.

[2] John Gantz David Reinsel and John Rydning. The digitization of the world from edge to core. Technical Report US44413318, International Data Corporation, November 2018.

[3] Alex de Vries. The growing energy footprint of artificial intelligence. *Joule*, 7:2191–2194, 2023.

[4] V. S. Dhaka Md. Abu Kausar and Sanjeev Kumar Singh. Web crawler: A review. *International Journal of Computer Applications*, 63(2):31–36, February 2013.

[5] Ms. D. Kavitha Mr.K. Tarakeswar. Search engines:a study. *Journal of Computer Applications (JCA)*, IV:29–33, 2011.

[6] Joshua Kaminsky Qifang Bi, Katherine E. Goodman and Justin Lessler. What is machine learning? a primer for the epidemiologist. *American Journal of Epidemiology*, 188(12):2222–2239, 2019.

[7] Vishwas Raval and Padam Kumar. Serelec (search engine result refinement and classification) - a meta search engine based on combinatorial search and search keyword based link classification. pages 627–631, 01 2012.

[8] Ian Rogers. The google pagerank algorithm and how it works. 2002.

[9] Amanda Spink, Bernard J. Jansen, Chris Blakely, and Sherry Koshman. A study of results overlap and uniqueness among major web search engines. *Information Processing and Management*, 42(5):1379–1391, 2006.

[10] Neenu Ann Sunny. Machine learning in search engines. *International Journal of Engineering Development and Research*, 8:155–161, 2020.

[11] Dean Frantsvog Tom Seymour and Satheesh Kuma. History of search engines. *International Journal of Management and Information Systems*, 15(4):47–58, 2011.

[12] Jinming Zou, Yi Han, and Sung-Sau So. Overview of artificial neural networks. *Artificial neural networks: methods and applications*, pages 14–22, 2009.