Google is the new Plato's cave

Paul Planchon

Politecnico Di Milano & CyTech Paris

The purpose of this paper was to show that Google is a modern century new Plato's cave. The Plato's cave is a metaphore for the knowledge, in which prisoners of a cave never see true knowledge (the sun), only the shadows of a dimed light. This paper will show that Google is the new Plato's cave through multiple angles. The first one is the way Google is design, multiple part of the design are not perfect, because making them perfect is impossible. Google web spider could be in a big unrelated graph, that way they will never sees other part of the knowledge graph. Also information is loose when Google crawlers tries to summarize and classify website. This led to the first cave: the one of pure knowledge, the one Google will never find or present to the user, by design. This paper shows that an other cave exists, the one created by user's data content. Lighting an other Plato's cave on Google : the big filter bubble it creates by influencing search results using user's data. These two arguments show that Google is a big Plato's cave.

Keywords: Google, free thinking, Plato's cave, filter bubble

CONTENTS

1. INTRODUCTION

Google takes a big place in the information based world we live in. Google is the portal to the internet, not only with Google Search : Google offers multiple services such as Google Chrome, Youtube and much more. This paper will focus on Google Search Engine, the service that allows the user to make queries using words, images or even sounds to search the deep ocean of information that is the Internet. Plato's cave is a philosophical thought experiment introduced by Plato in his book Republic to compare "the effect of education and the lack of it on our nature".

In this thought experiment, Plato describes a group of people who have lived all their lives chained to the wall of a cave, facing a blank wall (1). The people only watch the shadows projected on the wall from objects passing in front of a candle. They can only hear the echoes of the real world. They never have felt the real world, only a simplified projection. In Republic, Socrates, a student of Plato, suggests that the shadows are reality for the prisoners because they have never seen anything else. So they do not realize that what they see is only a projection of reality.

My thesis is that Google is our cave.

For the last 15 years, Google has been the most used Search Engine in the world (2), making it our cave. The white google.com page is our blank wall and the result of Google algorithm, the search results are the shadow of Google's reality.

To make the link between Google and Plato's cave I will explain to you how Google Search Engine works, what are the filter bubbles and the different dangers that personalized search results introduce. Then, I will have enough material to make the link between Google and Plato's cave.

2. WE ONLY SEE THE WORLD THROUGH GOOGLE

2.1. Internet is the new World

The world has seen three revolutions for the moment : the coal revolution of Britain in the late XVIII century, the Oil revolution in Germany and the USA at the end of the XIX century and the numeric revolution at the end of the XX century.
While the last two revolutions were not fully controlled by one entity, the numeric one that started in 1980 is much controlled by one big company.

The information age is gazed by the web, the HTTP protocol and web browser : Internet in one word. More than all the different services the numeric revolution offers : better services, faster communication, stronger science research. The information age has created a frictionless tool that can give access to the answer of the hardest question, providing trillions of results in less than a second : the Google Search Engine.

This tool can search the very deep ocean of information that is the Internet, providing you results ranging from videos with Youtube to scientifics papers with Google Scholar.

The fact that this tool has such big power and is not regulated raises a big concern : Google Search has the monopoly of the information. Google Search Engine is the only thing we ask nowadays when we have a question, and we never have questioned the integrity of Google data, or if Google has always the right answer.

2.2. Google's Bubble of information

Nobody is searching in books anymore. Why spend many hours in a library when a few words on GSE (Google Search Engine) can give you many answers to your questions in less than a second ?

That raised the question of the integrity of the information on Google. For the past 15 years, at least, Google has become the portal of knowledge. The unique portal of knowledge to be true : more than 91.2% (2) of web searches are processed through Google. It is the most prefered way to get access to information. If not the only way : to yahoo has never become a verb compared to Google(3). Having only one source of information is a problem : it creates a bubble.

If you do research on a specific topic, and all the information you get comes from Google, the research you do will be polarized by the GSE prism. Even more, when the research you did will arrive on Google, you will have completed the Google Search Engine information loop.

Google creates the information, or at least power the research to do it and distributes it to you. At first sight that appears to be a good thing because Google has become a very powerful tool to do research : it can search all the books and papers ever published in a fraction of a second. This is the problem : if we only trust Google, we will never be sure that other information exists. Trapping us in our cave, and reducing information within the reach of humanity.

It's Google algorithm and Google process of the web that could create this bubble. The next section will focus on the different algorithms and software that makes Google and why these tools can create the bubble I spoke about.

3. A SEARCH ENGINE MUST BE BIASED

A search engine is just a big algorithm whose job is to find the data you want, using clever sorting algorithms. At its core a search sorts the data it has access to. This process must decrease the entropy, quantity of information, of the dataset presented to the user.

This section will explain how a Search Engine like Google works, and what the trade off Google's engineers do to make the GSE as efficient as possible. And the impact of these trade offs on Google Data.

3.1. Internet spiders

The Internet is way too big for people to index it, or even, to just try to have a good picture of its size. The job of a web spider is to find and visit as many websites as possible. Helping Google have the best and most honest representation of the web.

A spider works by looking for all the links in a page, and visiting all of them (4). The internet spiders are like Sisyphe on his mountain. The spiders will never finish visiting all the internet like Sisyphe stone will always roll down to the bottom of the mountain. The job is too big. But internet spiders are not designed to cover all the internet : they are designed to cover as much as possible.

The design of a spider is flawed. All the web spiders can only search the convexe graph they started in. If the internet knowledge graph is not connexe and an island of knowledge exists, no spiders could find it. If such an island of knowledge exists, we may never know it. Because this is by definition, some knowledge we don't know about.

Also, in order to optimize the spider jobs, and to not flood the software that operates after this phase, the frequency of which a website is visited must be regulated (5). This means that Google will not analyse every website every time a spider comes around one.
This means that if a new piece of information is created, Google doesn't have it immediately. In our very fast world, that is a big problem. Information is too time sensitive nowadays to allow a delay. Fake news spreads very fast, and if for example the counter-argument of a fake news is on one of the non-revisited websites, the information may come too late for the user.

Nethertheless, the problem of the "connexivity" of the spider graph is not a real problem, the probability of having such and "island" of information is very low : the web is too much related for that.
Once the spider has captured a new link, the link is processed by the crawler, the software that is responsible to create the index and to summarise the webpage. This process must lead to the biggest loss of entropy in Google data aggregation.

3.2. Google's summarizing tools : the crawlers

The spiders are responsible for finding the website to analyse and add the website to the Google Search Engine database.

Storing all the websites is too big, even for Google. So they have to create an index (5), a simplified view of the website they visit. They transform the website in keywords, classify it, tty to put it in different groups of related websites. Even with summarisation made by humans (which is completely impossible), summarizing is a very complex subject, that can't be perfect. Summarizing cannot be classified as perfect. It's too subjective for that.

It's obvious that the way Google Engineers chose to operate the crawler introduces bias in the GSE process. No summarisation of anything can be as good as the original piece. This summarisation of the Internet led to another loss of entropy in Google Dataset of Internet and world knowledge.

The crawlers have the biggest job of Google's aggregation pipeline. Their work is primary, but the task is too difficult to do it perfectly.

Once this stage is finished, Google has to present the data to you. But reducing a multi-dimensional dataset down to a single dimension sets loose entropy.

3.3. Indexing and sorting a very complex dataset

Once the data is in Google servers, it just waits for a user to trigger a query keyword that would make this data appear on the search result webpage.

The Google HomePage is using a linear way of sorting the result (the most convenient for humans). That means that Google has to sort and filter their dataset in order to accomodate for that criteria. Searching results on Google is asking Google to minimize the distance between your query and result data. Data in Google server are multi-dimensional : results are ranked using the pagerank, but also keywords, cluster... This reduction of the dimension leads to a very big loss of information in the result dataset.

But Google's result page is only made of web results. It also contains ads, the first link in a Google results are ads, most of the time.
The probability of a user clicking on a link on a Google result page is very quickly decreasing : 25% of probability for the first link, then 25% for the second, then 11%... (6)

Ads may not be the most relevant piece of information Google has to offer, but it still takes the first place on Google result page. And these ads are not only query words triggered. They are user-data-powered. Also ads on Google result page are presented as normal Google results, just disguised with a little "ad" icon. Tricking the user into showing him what ad companies are willing to pay is a very bad thing. It's manipulation. This increases the size of Google's bubble.

Officially this is the only utilisation of personal data on the Google Search Engine.

But many reports show that Google is using personal data to influence the general search results. This use of personal data is a very big issue, this phenomenon has a name : Filter Bubbles.

4. FILTER BUBBLE

Filter bubble is a term invented by Eli Pariser and refers to the intellectual isolation that social medias like Facebook and Twitter do on their users. This phenomenon has been accepted for social media, even if the subject is very important. (7)

This section will try to understand if the phenomenon has arrived at Google, and what are the implications of that.

4.1. Personalized search in Google

Duck Duck Go's google filter study (8) shows that Google is ranking search results based on personal data. The study was simple : ask participants to do a very simple Google search, in private browsing mode (so, with no cookies, trackers...).
If the results were not personalized, we expect to see the same page every time. But DuckDuckGo's study found that Google results were indeed personalized, even in private browsing mode. Showing multiple things :

(i) GSE is always tracking the user, even when no cookies are available
(ii) GSE is create a personalized search result for every user

The DuckDuckGo study uses very sensitive keywords such as Gun Control. In a state where carrying a gun is allowed, the first link would be to promote guns. In other states, links are more related to gun violence and regulation. This study shows that Google is personalising our search result.

This led to a total bubble between states, between citizens. Google only shows results we want to read, results a community is agreeing with. That is the definition of a filter bubble.

4.2. Is Google Filter Bubble worse than Facebook's one ?

Facebook's filter bubble is a very big problem. But we never trust Facebook. We know that Facebook is creating a feed to keep you on their website, to make you interact with your friends, create information for them to sell you better ads. On the other hand, it's mostly accepted, even if we should not, that Google results are not influenced.
So we tend to trust Google more than Facebook, or other user generated content of other social media. That gives a lot of power to Google's filter bubble, because we tend to acknowledge it less, we tend to forget it.

Maybe it's the blind trust that humanity has in Google that has blind us.

5. A NEW CAVE FOR MANKIND

5.1. Google is our cave

Plato's cave is a metaphor for knowledge. In Plato's book Republic, he describes a caves. In this cave are a group of people who have lived chained to the wall of this cave all their lives. The group faces a blank wall. The people can only see the shadows projected on the wall from objects passing in front of the fire behind them. Because they have never seen anything else, these shadows are their reality. But, and this is where the metaphor of knowledge comes from, they never see the real world, light by the sun, with colors, shapes...

The Google Search Engine is Google flagship's product, it's capable of finding virtually anything on the Internet, at least everything that is in Google Database's bubble. As I explained before, this bubble must be created by multiple factors : the method of collecting data (using spiders that are prawn to island of knowledge), the way they analyse and index it the data they collected and finally the way they present it to the user (reducing the dimensionality of n-Dimensional dataset to 1D).

This is the first step in the cave of Google Search Engine. This is the biggest cave Google can create, the cave of "knowledge" : the knowledge Google can propose and the knowledge Google cannot propose on their service, just because they don't have it, by design.

The Engine will never be perfect, nethertheless, the problem is that we step even more in the cave with personnalisation of the search results.
The DuckDuckGo study shows that every query on Google is personalized, even the private browsing one. GSE must be proposing different results depending on the region, the hardware... Region, Language or Hardware put you in a different cluster of people in Google Search Engine results. This led to the creation of multiple bubbles, for every cluster of people in the world. Sorting every mind in a different way of thinking.

Everybody is chained with multiple handcuffs in Google's cave. Like the prisoners in Republic, we are trapped in at least 2 big caves : the first one is a cave of available knowledge created by the GSE process. The second one is personal data one, the most important one. The cave that traps you in your own thoughts.

5.2. Will we ever see the sunlight ?

In Republic, the book in which Plato's book explains what the cave is, one of the prisoners succeeded to escape. When he freed himself, the prisoner was confronted with the Sun, the power of the light coming from it temporarily blind him, giving pain to him. But after a while the prisoner is used to the power of the sun and accepts reality.

Freeing us of Google means that we understand and prove that Google is a cave in the first time.

6. CONCLUSION & OUVERTURE

Google is our new Plato's cave. The over-abused of user's data in order to "make the search more relevant" have trapped us in at least one cave, the filter bubble. The knowledge bubble may not exists, at least this paper is here to describe it.

A good ouverture to this subject would be the design of a Turing Test for filter bubble and cave like Google has trap us in.

REFERENCES

Wikipedia's Allegory of the cave, https://en.wikipedia.org/wiki/Allegory_of_the_cave, 2021.

Internet and Search Engine Usage By Country http://ptgmedia.pearsoncmg.com/images/9780789747884/supplements/9780789747884_appC.pdf, 2011

Oxford Dictionnary, https://www.oxfordlearnersdictionaries.com/definition/english/google, 2021.

The Anatomy of a Large-Scale HypertextualWeb Search Engine, http://infolab.stanford.edu/pub/papers/google.pdf, 1998.

How the search works ? https://developers.google.com/search/docs/beginner/how-search-works, 2021.

Google results click rate, https://www.searchenginejournal.com/google-first-page-clicks/374516/, 2020.

Beware online "filter bubble", Eli Pariser, https://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles, 2011.

Measuring the "Filter Bubble": How Google is influencing what you click, DuckDuckGo, https://spreadprivacy.com/google-filter-bubble-study/, 2018