The Evolution of Web Search Engines: Past, Present and Future

Aung Kwaw Htet, Michael Rothgang, Paige Snow and Hang Yuan
Jacobs University Bremen

We give an account of the evolution of web search engines through the ages. They were created when after the conception of the World Wide Web the amount of available information continued to grow rapidly and became too large to the processed only manually. We look at the technical aspects of how web search engines originated and explain how and why they evolved to their current state. We analyze the psychological implications of these changes (such as the Google effect, search biases and stereotype amplification) and effects on society. We glimpse through the ethical aspects of search engines, mostly concerning privacy rights of end users and violations thereof. We close with an outlook on search engines' possible future evolution.

Keywords: World Wide Web, search engines, PageRank, Google, Big Data

Contents

Introduction

When the internet was first created, it was far from being easy to use; merely establishing a connection to it required expert knowledge. This didn't change until 1990, when Tim Berners-Lee came up with the *World Wide Web*, an information space of internet resources linked together to form what he called *hypertext*. Berners-Lee wrote the first versions of HTML (Hypertext Markup Language, a language in which to specify hypertext documents), URI and URL specifications (a scheme for uniquely identifying resources on the web) and HTTP (a protocol for transferring data over the web). These open standards allowed the creation of programmes for viewing and editing the web, called *web browsers*. The first browser for the WWW was written by Berners-Lee himself and published in early 1992.

Other browsers were created soon after; Mosaic was the first popular one in 1993. Being easy-to-use and incorporating a graphical interface as well as support for video and sound, it greatly helped to increase knowledge about and usage of the web. As a consequence, the number of web pages grew exponentially and finding one's way around this growing forest became difficult.

The appearance of web search engines simplified this task. Search engines traverse the web and index the visited pages; upon input of a few keywords by the user, search engines iterate through the indexes for matching pages. Initially, engines only indexed a site's URL and title and listed all pages matching the query in the order they found them. With increasing growth of the internet, this quickly became impractical, hence engines started to rank their results by perceived user importance. Also, further metadata was taken into account, up to web search engines searching a page's full text.

With online services embracing the internet, search engines themselves became a hot topic also with an increasingly fierce competition. In 1996, Sergey Brin and Larry Page came up with an algorithm called *PageRank* to rank websites using information on the internet itself. To commercialise this idea, they launched the Google search engine in late 1997. Google quickly gained significant market share owing to its minimalist user interface and much better results compared to its competitors.

With Google becoming this influential, the underlying PageRank algorithm changed how search engines are used, as users found relevant search results very quickly and hence didn't bother to click further. Nowadays, users only glance at the very few first search results (Joachims, Granka, Pan, Hembrooke, & Gay, 2005). When search engines became one of the primary sources for information, this turned into

truly big business not only for search engine developers, but also for site owners: Search engines (such as Google in 2000) had begun placing advertisements (whose ranking was auctioned) next to their search results and thus profited from any usage. Site owners could make their site better ranked by search engines by manipulating the engine's algorithm, yielding more visitors on the site. A new business of *search engine optimisation* (SEO) was born.

This led to an arms race between search engine optimisers trying to optimise their pages' ranking to maximise profits for commercial sites and search engines adapting their algorithms in order to still deliver the results most relevant to their users. With many companies competing for the very few top spots on Google's result page with the associated potential billion dollars revenue stream, PageRank proved insufficient to counteract the optimisers' techniques while still delivering relevant results. Web search engines had to evolve again.

In this paper, we will describe how web search engines worked in the past, why they had to evolve and how they managed to overcome search engine optimisers again. However, this process changed their inner workings drastically with large practical consequences. We will discuss economic implications, influences on society and individual psychology as well as ethical aspects. Finally, we look ahead at how web search engines will evolve in the future.

Technical background and details

In this section, we will go into more technical detail. We will explain how web search engines' algorithms work (using PageRank as an example), illustrate how SEO can game their algorithms and explain how search engines function now, having found a way to counteract bad search engine optimisers.

PageRank as example for web search engine algorithms

Search engines' algorithms are the core of their workings; hence precise details are usually kept secret. One exception to this is the original PageRank algorithm which is even described in an academic paper (Page, Brin, Motwani, & Winograd, 1998), allowing it to be discussed in detail. We also chose to describe how PageRank works since it was revolutionary for the Web and its usage in general; additionally PageRank is important since it allowed Google, in particular to gain wide market share, providing the basis for Google's web search monopoly (and corresponding influence on the market) today.

As the Web's pages are hypertext documents, they contain links to other sites and are linked to from pages as well. The PageRank algorithm uses this link structure on the web to compute an approximate notion of a page's relevance. The reasoning is that if a page is highly relevant, many other sites will link to it. Merely counting this number of *backlinks* to a page is not sufficiently telling, though as the individual backlinks can have different importance. Intuitively speaking, PageRank rates a page as highly relevant when the sum of its backlinked pages' ranks is high—so being linked to from either a single highly relevant page or several moderately relevant pages will yield a high relevance ranking.

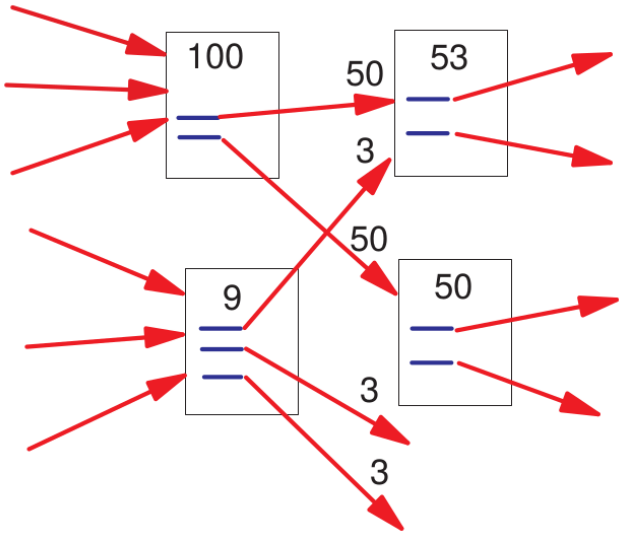

Figure 1. Graphical depiction of the simplified PageRank calculation. Source: Page, Brin, Motwani, and Winograd, 1998, page 4

More mathematically, let u be any web page, B_u be the set of its backlinks, F_u be the set of pages it links to (u's *forward links*) and $N_u = |F_u|$ be the number of u's forward links. Then a slight simplification of the page rank of u is

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

where c is a normalisation constant, ensuring that the total rank of all pages stays constant. We see that a page's rank is divided equally among all its forward links to contribute to their rank.

This is a recursive equation—given the ranks of a page's backlinks, we can compute its rank. In practice, pages' ranks are computed iteratively: One will start with some estimated initial ranking of the page corpus. Then, one will run PageRank on these weights, generating some new ranks. To these ranks, the algorithm is applied again, again and again until the ranks converge, i.e. they don't change much during one iteration. This happens quite fast in practice (Page et al., 1998).

Examples of search engine optimisation

If engines just used this simplified version of PageRank, their algorithms were easy to manipulate simply by creating a

cycle of webpages, each page linking only to the subsequent one, and then placing a link on your homepage pointing to one of the pages in the cycle. During the iterative computation of the web pages' rank, the pages in this cycle would accumulate more and more page rank, without contributing to other pages' ranks as there are no other outgoing links. As the web's total page rank stays constant by construction, the other pages' ranks would decrease and the overall importance ranking would be distorted.

This was already observed in the original PageRank paper. To address this, Page et al. introduced a vector E over all pages (i.e. each page u is mapped to some weight $E(u)$—in practice they often used the same weight for all pages) acting as a source of additional rank. Then the page rank of u is defined as to satisfy

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v} + cE(u)$$

where B_u, N_v are defined as above and R is chosen as to allow the maximum possible constant c.

As its working principle is public, gaming PageRank is possible. Just looking at the algorithm, it is clear how to get a high rank for your page (and thus possibly a better ranking by Google and other search engines relying on it)—simply increase the number and quality of backlinks to your page. Improving the site's content can have this effect, structuring it nicely can also help. These measures will probably also make the site more useful to users, are thus beneficial and are called white hat search engine optimisation for this reason.

A different way of SEO is to simply buy backlinks—pay a (ideally high-quality) site for linking to your page, so search engines will pick that up and adjust their rankings. In the incarnation of buying links from sites with totally different topics, this is clearly acting against user interest; this is an example of what is called black hat SEO. Search engines have worked on detecting such cases and discarding such bought links or possibly even penalising the concerned site (Cutts, 2007 and Cutts, 2005).

Another example of (black hat) SEO is to mostly copy content of high-quality sites; this works since modern search engines also look at the page's textual content and structure. This is at best redundant and distracting and thus penalised as well; Google issued its "Panda" update in 2011 specifically on this issue (Official Google blog, 2011).

As search engine optimisation is a big and vibrant business nowadays, even attempting to briefly mention all of it is beyond the scope of this paper, hence we leave it at that.

New technologies to defeat search optimisers

To defeat black hat SEO, web search engines needed new techniques beyond analysing the web's link structure. Hence, search algorithms have been expanded to include many other criteria (e.g. Google or Bing look at over 200 different signals nowadays, Sullivan, 2010 and Bing Search Blog, 2011) into consideration. The key innovation, however, has been a simple one: use user data to see which pages are relevant.

Nowadays, most web search engines record which query results users click on, and then use this data to adjust their ranking using machine learning techniques. About 90 percent of the time, engines do actually search past searches for matches—only in the remaining cases they actually search the Web (Offical Google Blog, 2008, Clark, 2015, Taft, 2014, DuckDuckGo Team, n.d.).

Web search engines employing this technology then have to store huge amounts of user data on their search preferences and be able to access them fast (latency is important for search engines)—we have a true Big Data initiative at hand.

Customised search results and user tracking

The users' data of search engines is aggregate data at first: it says which proportion of users clicked on a specific page, but does not identify the individual user. This has moved a step further with what engines call "personalised search". For instance, it is possible to sign in to an account with the search provider; then one's history of searches and preferred pages will be recorded and used to customise the search results. Today, many search engines go even further by also tracking users when not logging in to any account: By default, their trail over the web is followed, creating a big database of user behaviour information. This is used to adjust search results as well as advertisements shown (Bing Search Blog, 2011, "Google Privacy Policy," n.d., DuckDuckGo Team, n.d.). By using that search engine, the user consents to this policy—and to prevent this collection one can either change the search engine (e.g. Ixquick or DuckDuckGo do not individually track users, cf. DuckDuckGo Team, n.d. and "Ixquick Protects Your Privacy!" N.d.) or use anonymising software to make provider's life much harder (Eckersley, Schoen, Bankston, and Slater, 2006).

Economic implications of the new technology

Consequently, web search engines crucially depend on having comprehensive and representative data sets—searching past searches only yields relevant results when there is sufficient and representative data to search. This means that there is a natural monopoly for search engines: If a search engine has a large market share, it can collect lots of user data and continue to deliver very relevant results to users, hence they will continue to use it. Even further, other search engines without such data sets will be unable to offer as precise hits and thus tend to lose market share towards the leading competitor.

This is one reason why Google has such a high market share among search engines: having gained an advantage to-

ward their competitors via a better algorithm, this advantage is perpetuated through the reliance on user data for search result ranking. Google's business model works by earning money from every search, allowing it to fund further research to sustain its lead.

For other search engines to compete with a monopolist, it is crucial that they obtain a similarly good data set on user preferences. One way is to cooperate with internet service providers to gain user data, but Google starting to encrypt its searches in 2011 making this impossible (as a side effect or maybe also intended—the Official Google Blog, 2011a cites "protecting the personalised search experience we deliver" as one of the reasons for this measure).

An alternative is to cooperate with popular websites to track from where its visitors are coming, called referrer header: when a user clicks on a link, the target site is told where the user came from (from where the user was "referred to" the site). If the user came from a search results page, its address will contain the search terms so that one can reconstruct the user's action. This data is still very useful because the vast majority of web traffic goes over a few top sites. That option also became moot when Google encrypted its referrer header as well.

The last way to still obtain significant collections of user data is a browser's search box, recording what terms browser users search for and which sites they prefer directly in the web browser. In 2011, Google accused Microsoft of doing exactly that, capturing Internet Explorer users' searches with Google and their preferred results and then using this data to generate better results for Bing searches (Sullivan, 2011b, Official Google Blog, 2011b). This generated quite a bit of public coverage (e.g. CNN Wire Staff, 2011, Bing Search Blog, 2011, Sullivan, 2011a])—and even though it seems that Bing is only using this data as one of many signals, this illustrates well how competitive the web search market is now, and the power of monopolist Google's position.

Psychological implications

Web search engines have not directly influenced society, but the way people have used web search engines has created a change in society. The time we live in now is called the *digital revolution* (Helbing, 2015) where a web search engine is simply a couple clicks away. It can be accessed at any moment of the day. Web search engines have led to wonderful advancements in technology; however, this advancement has come at a price. One of these being the support that web search engines indirectly give to stereotypes, biases, and closed-mindedness. Since there are numerous definitions for each of these words, we will provide clarity on the definitions referred to here.

A *stereotype* is a generalization or simplification of a group of people that can be either positive or negative. It has been proven through psychological testing that, although we try the best we can to avoid stereotypes, our brains naturally organize information in a categorical manner; therefore, stereotypes are necessary for information processing. When we refer to stereotypes here, we will be referring to the negative, untrue stereotypes. *Bias* is a specific opinion, inclination, or trend that does not necessarily have any reasoning behind it. We will be referring to the negative connotation of bias as well. *Closed-mindedness* is being "obstinately resistant to argument or to unfamiliar or unwelcome ideas" (Merriam-Webster, n.d.).

These perceptions are amplified in a user through the user preference tool within web search engines. User preference was created to make web search engines more efficient, but this has led to search engines offering results that fit the user's beliefs instead of challenging them. This can be detrimental in several issues, especially when the user preference function is not widely known. Many people trust web search engines to give them accurate results, when in reality, web search engines give results that are accurate to the user. The small frame of reference that is provided by web search engines only contributes the facts that the user would want to see; therefore, supporting stereotypes and biases that they might have. Since opposing opinions are not presented in the search results, closed-mindedness continues. This defeats the purpose that people use web search engines. Many people use web search engines to find more information about a topic, where the information is presented in a neutral way. Many of the above-mentioned dispositions are combated through the education of that group or the topic that the person would hold a stereotype, bias, or be closed-minded about. This is not possible with the user preference system on web search engines. The impact of this drawback can be seen very clearly in an example that will be mentioned later on.

The technology of web search engines is not only supporting our perceptions, but also changing our psychology. A very interesting psychological effect that comes out of the usage of web search engines is called the *Google Effect*. Although Google is in the name of the effect, this effect holds true for all web search engines. This is when people use Google to look up what they want to know instead of remembering the fact. The part of the brain that is changed by this is the short-term memory. Our short-term memory relies more so on outsourcing rather than committing facts to memory. We use the Internet for memory much like we would use a friend or family member. For example, if you couldn't remember your grandmother's favorite type of flower, you would ask your parents when you needed to know this fact, rather than committing it to memory.

A large-scale example of how the Google Effect can impact society can be seen in a series of five psychological experiments across two countries performed by Robert Epstein and Ronald Robertson. In these experiments, they showed

that biased web search engine results led undecided voters to shift their votes by 20% or more. They state that these biased rankings can be masked so that the user is unaware of the manipulation (Epstein & Robertson, 2015). This type of manipulation done by web search engines, whether intended or unintended, is prevalent across many spheres of society, not just in politics.

The Google Effect is an example of the consequence of web search engines, but whether this consequence is helpful or hurtful to society is debatable. Knowing *where* some information is rather than *what* the information is can be either a good thing or a bad thing. It is a good thing because we have more "room" within our minds to store information that we perceive as being less trivial. Web search engine use is making us more intuitive in accessing the answers that we want to find. It is simpler to look something up on the Internet rather than memorize all of these facts.

On the other side of this argument, the use of web search engines for outsourcing memory causes us to have fewer connections with information that we have stored in our brain because it is simply not there. There is a classic saying in Neuroscience: neurons that wire together, fire together. Just to clarify, neurons are the particular cells within the brain. They communicate with each other through the use of electrical signals. These signals are caused by a change in the charge of the membrane of the cell. This change in membrane potential is initiated by the summing up of electrical signals from surrounding neurons. This saying also is connected to a term called neuroplasticity, which states that the changes in the neural pathways are caused by changes in behavior, environment, emotions, and thinking.

With this in mind, the use of web search engines as information storage is literally changing the connections that we make between our neurons. If only the location of this information is present within the brain, there is not as rich of a connection of this particular fact with another fact that we have stored. This changes the very psyche of a person because our psyche has a foundation in making connections. Without this information being present, we lack the fullness of life that we previously had when facts were memorized instead of searched. However, one could argue that the facts that are chosen to be searched rather than memorized are not facts that would make that much of an impact on us if we remembered them.

When looking at the advantages of search engines, many believe that it is not the change in memory storage that is concerning people, but rather, the ominous future of web search engines in our lives. The Google Effect was identified recently, and the full implications of this effect are not yet understood. It is argued that it is the huge transition that society is going through that makes people wary of web search engines, not the direct use of web search engines.

Another consequence of web search engines usage that is not widely discussed is how the usage of web search engines changes our perception of time. Our sense of time can be explained by three perspective biases that were proposed by Professor Phil Zimbardo of Stanford University. These are known as Future-Oriented, Past-Oriented, and Present-Hedonistic. Professor Zimbardo suggests that we all come into this world as Present-Hedonists, which means that we live in the now, pursuing present pleasure. It is not that Present-Hedonists don't know the consequence that can come from their pleasure-seeking attitude; it is just that they don't plan for it. The job of school systems is to make Present-Hedonists into Future-Oriented people. Future-Oriented people are focused more on the upcoming consequences of alternative courses of action, and they plan according to these perceived consequences. Past-Oriented people develop depending upon the cultural setting or group within the society that the person grows up in. This type of person will look more at previous consequences of actions, and use this information as a base for present decisions (Zimbardo & Boyd, 2008).

Web search engines change this perspective of time by stopping the development of Present-Hedonists to Future-Oriented people. They do this by supporting the immediate satisfaction that is a trademark in Present-Hedonists. It is not only web search engines that contribute to this phenomenon, but technology in general as well. We live in a time where patience is truly a virtue. With one-day shipping, instant streaming, and having information at the click of a button, there is no need to wait for our desires to be met. The increasing amount of technology use has been proposed as a block in the development of a Future-Oriented person. The modern world is dependent upon the existence of Future-Oriented people, but instant gratification hinders this.

In conclusion, web search engine technology enhances stereotypes, biases, and closed-mindedness by not displaying opposing views. Web search engines change the way our brains store memory by acting as an outsourcing device. This change can either be good or bad: good that we have more space to remember important information, bad that we can't make deep connections within our brain because the information doesn't exist there. Finally, web search engines change our sense of time by causing us to be Present-Hedonists rather than Future-Oriented people.

On search engines and homo ethicus

It is worthwhile looking into the ethical aspects of the technology society of man has produced because ethics is the very foundation of social institutions (Berger & Luckmann, 1966). Because the human by its nature maximizes both individual and social utility (O'Boyle, 1994), completely ignoring the latter for the former would lead to its own demise.

Search engines act as the gatekeeper to the ultimate goal of the Enlightenment project; the consolidation of humanity

and rationality. The enlightenment ideals, especially by Mills and to a certain degree, Hegel, the advancement of humanity as the whole is to be achieved by the rational contestation of ideas through freedom of speech and shared experience (from diversity). The truth, historically and contemporarily, is that human society is highly fractionalized with the rational self-interest of each community. In the case of search engines, the rationally interested groups can generally be identified as: 1. the shareholders and the executives of the search engines, 2. the associated businesses such as data mining companies, advertisers and producers, 3. the web content authors, 4. the societies affected, and lastly, 5. the general users like you and me.

Because of phenomena such as the Google Effect the public opinion has further become multi-polarized, contrary to the enlightenment ideals. Due to how most search engines work, unpopular views tend to be suppressed as they will be accessed less frequently and receive fewer links from other pages. This bias against new pages is called entrenchment effect: "Heavy reliance on a search engine that ranks results according to popularity can delay widespread awareness of a high-quality page by a factor of over 60, compared with a simulated world without a search engine in which pages are accessed through browsing alone" (Davis, n.d.). According to Davis, "All results of this kind are necessarily based on models that are necessarily highly idealized, since you can't create an alternate world in which there are no search engines" (Davis, n.d.). If true, this problem would have been latent dysfunction of the technology over objectivity, the very basis for shared experience.

Even though that would have been bad enough, the traditional PageRank algorithm still represents a "technical and objective meaning" of popularity (Hinman, 2005, p. 22). The post-PageRank era of Web 2.0 is much darker. Take a look at the following real life case. The Bush administration in 2005 has ordered Google, Yahoo, AOL and MSN to turn over one million random web addresses and records of all Google searches for a one week period to revive child protection law which would assist the government to identify the number of porn links that show up in a child's searches (Sullivan, 2006). Google refused initially (but the other three were assumed to have complied). In January 2006, a court ruled that Google did not have to comply with the initial demand by the Bush administration but had to turn over 50,000 URLs to the government (Nissenbaum, 2010, pp. 29–30). Only from this case, many learned that search engine giants such as Google had been keeping a log of all search queries for long times. While technically justified, de facto there was no informed consent from the users; this indicates how search providers and the justice system have failed to protect our right to privacy from the government.

Search engine enterprise's powerlessness against the sovereign states can further be confirmed in the censorship by the authoritarian regimes. This is antipodal to the core enlightenment virtues of freedom and rationality. In 2006 when entering the Chinese market, Google's agreed to comply to China's censorship laws that censors entries such as "Democracy," "Tiananmen Square," "Free Tibet," and "Dalai Lama." By doing so, Google has violated its own core principle of "don't be evil" by succumbing to "facilitate and support" China's censorship regime (Spinello, 2011).

Since 2010, Google has changed its policy in mainland China by directing users to a site in Hong Kong where censorship is less strict. Yet, we could still question that Google's broken promise and four years of cooperation with anti-democratization for profit's sake remain unjustified. Even when granting the benefit of doubt that Google would resist Big Brothers' power as much as it possibly could, the reality is search engine enterprise is a ready-made piece of panopticon for powerful state espionage institutions like NSA. Leaked documents by Snowden in 2013 revealed that the National Security Agency has

"secretly broken into Google and Yahoo data centers outside the U.S. worsen the crisis the U.S. government faces over its cyber-espionage activities and undermine confidence in Google and Yahoo as global communication businesses" (Fidler, 2013).

Such lack of accountability of both the US government and search engine enterprise on mass violation of privacy of digital communications indicates that, on the ground of virtue ethics, search engines are just another pair of lens for the Big Brother and are not to be trusted.

Contrary to the traditional PageRank algorithm, the personalized nature of modern web search engines is to give search results that are adjusted to the user's past behavioral patterns. This has further pulled the average users away from the enlightenment norm of objectivity of information. Diaz (2008) warned that that major search engine companies such as Google direct "hundreds of millions of users towards some content and not others, towards some sources and not others." Thus systematically, though arguably unintentional, "independent voices and diverse viewpoints" that are core to democracy are the filtered out by search engines.

Because the very being of search engine in-itself is very tempting to the rational self-interest of the businesses and also because shareholders and executives of the search engines are financially dependent on the paid business, Kantian deontology of "Never use humanity as a mean" is grossly violated. In a 2011 book review, Halpern states that there are approximately 500 companies that can track all our activities from the web and sell to data mining companies (Halpern, 2011). Google in particular integrates information gathered from

> "Web cookies, detailed server logs, and user accounts [...] from Google applications such as Gmail, Google +, and Google Chrome [...]

> which provides a powerful infrastructure of dataveillance to monitor, record, and aggregate users' online activities" (Zimmer, 2008, p. 77).

According to Pariser (2011), these information includes your logged in location, browser that you use and your sequential order of queries. Based on this information, Google and other major search engines use "prediction engines" to categorize its users as commodities to be "auctioned off within microseconds to the highest bidding consumer." The data mining company Acxiom, according to Pariser at the time of his writing had "accumulated an average of 1500 pieces of data on every person in its database—personal data that ranges from credit scores to medications used."

Enterprises by nature have a natural monopoly of symbolic dominance. Google as the search engine giant is no exception. According to Hedström and Bearman (2009) "status is the stock of accumulated acts of deference and therefore a signal of an underlying quality," meaning status arises from exchange relationships and over time becomes a producer of qualities ("Matthew Effect"). Status increases when qualities are uncertain, as people rely on status as the signifier of quality.

Podolny and Lynn's theory explains the rise of Google as monopolist in the search engine enterprise in terms of status: Superior service delivered by Google by in the early days has crowned Google as the embodiment of excellence among the jungle of inferior alternatives. Over time, people spend less time and effort to assess the true quality of their used brands (the increasingly complex technology involved and the market specialization with increasing competition didn't make it easy to follow by normal customers), used status as a heuristic instead and chose to stick with Google for efficiency sake. The natural monopoly within the enterprise makes the end users even more vulnerable. Not only are they barred from objectivity, and their privacy, autonomy, and liberty compromised, but also are they facing a lack of true choice (as the monopolist's competitors deliver inferior results).

To sum things up from sociological perspectives, as the informed consent from an end user that his or her information will be stored is hard to achieve, the existence of the web search engine inherently manifests violation of the right to privacy of the end user. Also, because web search engine companies are legally subject to the sovereigns of the markets in which they operate, they become just another big data source for mass surveillance to the states, and thus instrumentally assist in the violation of the ethic of individual autonomy in some cases, even censorship is accepted. In addition to sovereignty threat to individual freedom, search engines as an enterprise are a complimentary business model to data mining companies and advertisers. The natural monopoly of the enterprise renders the end users more vulnerable to all the threats above.

All in all, as much as search engines are a machine for predictability for businesses, Big Brothers and philosophers on the ivory tower, their abuse by few makes them a de facto weapon of mass deception and control of the many; they are just another self-defeating product of the enlightenment project.

Outlook on web search engines

In the fifth annual IDC Digital Universe study, the world's data is predicted to grow fiftyfold in the next decade. How to get the information we need from the web becomes an even more important question. In web search 1.0, algorithms like PageRank gave us good enough results; in web search 2.0, web search engines became capable of natural language processing to answer the questions that people have; in web search 3.0, context search provided the users with information that were optimized using heuristics like intent, location and conversation. (Monago, 2015) Now our challenge has become how to do better.

In this section, we will discuss some impending technological advancements and current works in progress that facilitate producing better search results in web search 3.0. This section is by no means intended to be comprehensive, but just to give a taste of how web search engines and the use of web search engines are going to evolve.

Web search 3.0

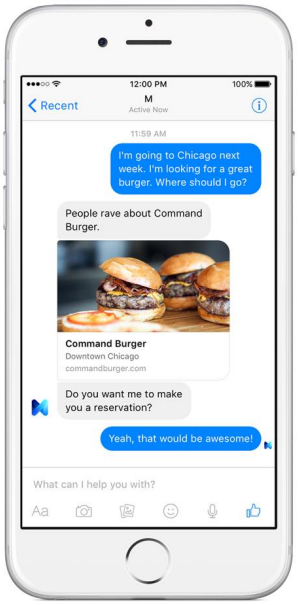

Figure 2. Source: Facebook.com

Conversational search. Today's leading search engines companies are providing means to make search easier than ever. Conversational search is one of the new areas that search engines companies are developing to let machines better understand the context of a search.

Facebook's M, Apple's Siri, Google's Now and Microsoft's Cortana, are intelligent "personal assistants" that are trying to understand what a user really means by looking at the queries from a much more holistic approach. The picture on the left shows how M is suggesting a burger place when one user is going on a trip to Chicago. Just like having a real conversation with a friend, M can respond to the user with the information about the recommended restaurant

using natural languages. Another example will be Now from Google. Let's say if a user is listening to a song using his phone, that user can ask Now some questions like, "Who is the singer?" or "When was this song published?" Now will respond to the questions based on the song that is being played.

"Personal assistants" are powered with the context of a search query, so that the search engines have more information to figure out what exactly the user wants. However, our privacy is further exploited by the trend of conversational search as the search engines need to watch what we do with our phones or computers to glean the necessary cues to generate context-sensitive results.

Understanding User Goals. Let us imagine we already have a perfect search engine, albeit, is having such a search engine good enough to get the information we want? The answer is no, because sometimes we don't even know what we want and thus, even if a search engine can give us an excellent answer when provided with a question, we still cannot get the information we need from the search engine. The real problem is we cannot even ask the right question. Going one step further, if we have a perfect search engine and know exactly what we are looking for, this information might still not be enough for the search engine to figure out what the right answer is because we might not be able to specify the needed information. In a study of Orienteering Behavior in Directed Search (Teevan, Alvarado, Ackerman, & Karger, 2004) when one participant is looking for a file in a large system, she said something like this:

"I don't know how I could have the directory in mind without knowing its name, but I felt sure which it was." This participant eventually found the document after a series local context searches.

Similar directed pattern is also found in another group of studies by Suchman in 1987; we call the behavior that does not search for the information directly but performs a series of suited actions *orienteering* (Teevan et al., 2004). Teevan et al.'s study concludes, "often keyword-based search engines were not used when searching, and when they were used, it was usually part of an orienteering strategy." By adopting this interdisciplinary approach between psychology and computer science, incorporating users' orienteering strategy nonetheless can help web search engines deliver better results.

Use of nonverbal behaviors. Another use of a psychological approach is to collect user's status via nonverbal behavior. Given the widespread of wearable devices and phones that contain HD front cameras, lots of physical information about a user can be easily collected—such as heartbeat, blood pressure, gestures, and facial expressions. Since we don't always mean what we say, nonverbal behaviors are used to facilitate communication via means of contradiction, substitution or regulation (Knapp, Hall, & Horgan, 2013).

Our understanding of nonverbal behavior is good enough for us to even have a nonverbal behavior generator for embodied conversational agents, (Lee & Marsella, 2006) however the use of nonverbal behavior has not been reasonably incorporated into neither natural language processing nor search engine heuristic gathering due to privacy and ethical reasons. Having your camera constantly monitoring your facial expressions might not sound pleasant, even if you might eventually benefit from it.

The future of web search engines is largely based on the collection of more useful information about the users, whereas the fact that more information needs to be collected might hinder the progress of those new advancements. The questions such as how much information we are willing to make available to those electronic devices and what is the bottom line of such "Big Brother", or if it is even a bottom line at all, need to be answered.

The development of technology gallops and the public accepted standards also change as time goes on. Perhaps one day we won't be intimidated by the massive information collection anymore, as we will get used to it sooner or later.

Conclusion

In our review, we not only touched on the underlying mechanism behind the existing web search engines, but also extended our discussion into the psychological and social implications of web search engines. The development of web search engines is growing faster than ever, and surely the concerns brought by these technological advancements do not seem to cease—despite all the latent dysfunctions of search engine technology, the warnings in this paper would not possible without the use of the very technology. Indeed, we cannot overlook the good intention that many creators of technologies bear in mind when they first start to build these amazing things-to make people's life better.

Undoubtedly, the tendency of Big Data is not going to change, and in fact the plateau of Big Data won't come until five or ten years later (Gartner Inc., 2014). The real question is how are we going to prepare ourselves for tomorrow's massive amount of information. As humans, we can only rely on our brains, which have a limited capacity, whereas technology can advance fast enough to help us tackle more onerous situations, and in this process, how are we going to deal with the ever challenged ethical bottom line, and how are we going to remain calm instead of being lost in the midst of technological hyperbolas that require imminent answers. Otherwise the seemly "big boon" might turn into the next "Frankenstein," and we as humans would have to go on a path of redemption—but at that time, things may have been irrevocable.

Acknowledgements

We would like to thank seven anonymous referees for their helpful comments. We would also like to thank Andreas Gal for inspiring us on this topic by pointing out the Big Data aspects of today's search engines.

References

Berger, P. L. & Luckmann, T. (1966). *The social construction of reality*. Anchor Books.

Berners-Lee, T. (1989, March). Information management: a proposal. Retrieved November 1, 2015, from http://www.w3.org/History/1989/proposal.html

Bing Search Blog. (2011, February 1). Thoughts on search quality. Retrieved November 1, 2015, from https://blogs.bing.com/search/2011/02/01/thoughts-on-search-quality/

Clark, J. (2015, October 26). Google turning its lucrative web search over to ai machines. Retrieved November 1, 2015, from http://www.bloomberg.com/news/articles/2015-10-26/google-turning-its-lucrative-web-search-over-to-ai-machines

CNN Wire Staff. (2011, February 11). Google: sting proves bing copied search results. Retrieved November 2, 2015, from http://edition.cnn.com/2011/TECH/web/02/02/google.bing.sting/index.html

Cutts, M. (2005, September 15). Text links and pagerank. Retrieved November 1, 2015, from https://www.mattcutts.com/blog/text-links-and-pagerank/

Cutts, M. (2007, April 14). How to report paid links. Retrieved November 1, 2015, from https://www.mattcutts.com/blog/how-to-report-paid-links/

Davis, E. (n.d.). The politics and ethics of search engines. Retrieved from http://cs.nyu.edu/courses/spring11/G22.2580-001/lec14.html

Decadeology. (n.d.). An accurate in-depth history of the internet. Retrieved October 31, 2015, from http://www.decadeology.com/an-accurate-history-of-the-internet.html

Diaz, A. (2008). *Through the google goggles: sociopolitical bias in search engine design*. Springer.

DuckDuckGo Team. (n.d.). Duckduckgo privacy. Retrieved November 1, 2015, from https://duckduckgo.com/privacy

Eckersley, P., Schoen, S., Bankston, K., & Slater, D. (2006, September 14). Six tips to protect your search privacy. Retrieved November 2, 2015, from https://www.eff.org/wp/six-tips-protect-your-search-privacy

Epstein, R. & Robertson, R. E. (2015). The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, *112*(33), E4512–E4521.

Fidler, D. (2013, October 30). Iu expert: nsa infiltration of google and yahoo networks threatens u.s. tech leadership. Retrieved from http://news.indiana.edu/releases/iu/university-wide/2013/10/nsa-muscular-expert-comment.shtml

Gal, A. (2015a, March 30). Data is at the heart of search. but who has access to it? Retrieved November 1, 2015, from http://andreasgal.com/2015/03/30/data-is-at-the-heart-of-search-but-who-has-access-to-it/

Gal, A. (2015b, February 23). Search works differently than you may think. Retrieved November 1, 2015, from http://andreasgal.com/2015/02/23/search-works-differently-than-you-may-think/

Gartner Inc. (2014, August 11). Gartner's 2014 hype cycle for emerging technologies maps the journey to digital business. Retrieved from http://www.gartner.com/newsroom/id/2819918

Google Inc. (n.d.). Google through the years. Retrieved October 31, 2015, from https://www.google.de/intl/en/about/company/timeline/

Google Privacy Policy. (n.d.). Retrieved November 2, 2015, from https://www.google.com/policies/privacy/

Google Support. (n.d.). Webmaster guidelines. Retrieved November 1, 2015, from https://support.google.com/webmasters/answer/35769

Halpern, S. (2011, June 23). Mind control & the internet. *New York Review of Books*, *58*(11). Retrieved November 1, 2015, from http://www.nybooks.com/articles/archives/2011/jun/23/mind-control-and-internet/

Hedström, P. & Bearman, P. (2009). *The oxford handbook of analytical sociology*. Oxford University Press.

Helbing, D. (2015, May 20). The world after big data: what the digital revolution means for us. Retrieved November 21, 2015, from http://futurict.blogspot.de/2014/05/the-world-after-big-data-what-digital.html

Hinman, L. M. (2005). Esse est indicato in google: ethical and political issues in search engines. *International Review of Information Ethics*, *3*(6), 19–25.

Ixquick Protects Your Privacy! (n.d.). Retrieved November 2, 2015, from https://www.ixquick.com/eng/protect-privacy.html

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. *Proceedings of the Conference on Research and Development in Information Retrieval (SIGIR)*. Retrieved November 2, 2015, from http://www.cs.cornell.edu/People/tj/publications/joachims_etal_05a.pdf

Knapp, M., Hall, J., & Horgan, T. (2013). *Nonverbal communication in human interaction*. Cengage Learning.

Lee, J. & Marsella, S. (2006). Nonverbal behavior generator for embodied conversational agents. In *Intelligent virtual agents* (pp. 243–255). Springer.

Merriam-Webster. (n.d.). Merriam-webster dictionary: closed-mindedness. Retrieved November 21, 2015, from http://www.merriam-webster.com/dictionary/closed-minded

Monago, A. (2015, September 23). Brave new world: the future of search. Retrieved from https://www.thoughtworks.com/insights/blog/brave-new-world-future-search

Nielsen, J. (2005, September 26). The power of defaults. Retrieved November 1, 2015, from http://www.nngroup.com/articles/the-power-of-defaults/

Nissenbaum, H. (2010). *Privacy in context: technology, policy, and the integrity of social life*. Palo Alto, CA: Stanford University Press. Retrieved from http://www.sup.org/books/title/?id=8862

O'Boyle, E. J. (1994). Homo socio-economicus: foundational to social economics and the social economy fr1:0034-6764/94/0601-286/ 1.50/0 ef. *Review of Social Economy*, *52*(3), 286–313.

Offical Google Blog. (2008, September 15). Search evaluation at google. Retrieved January 1, 2015, from https://googleblog.blogspot.de/2008/09/search-evaluation-at-google.html

Official Google Blog. (2011a, October 18). Making search more secure. Retrieved November 1, 2015, from https://googleblog.blogspot.de/2011/10/making-search-more-secure.html

Official Google Blog. (2011b, February 1). Microsoft's bing uses google search results—and denies it. Retrieved November 1, 2015, from https://googleblog.blogspot.de/2011/02/microsofts-bing-uses-google-search.html

Official Google blog. (2011, February 24). Finding more high-quality sites in search. Retrieved November 1, 2015, from https://googleblog.blogspot.de/2011/02/finding-more-high-quality-sites-in.html

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The pagerank citation ranking: bringing order to the web. Retrieved November 1, 2015, from http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

Pariser, E. (2011). *The filter bubble: what the internet is hiding from you*. Penguin UK.

Spinello, R. A. (2011). Google in china: corporate responsibility on. *Investigating Cyber Law and Cyber Ethics: Issues, Impacts and Practices: Issues, Impacts and Practices*, 239–253.

Stewart, B. (n.d.-a). Mosaic – the first global web browser. Retrieved October 31, 2015, from http://www.livinginternet.com/w/wi_mosaic.htm

Stewart, B. (n.d.-b). Online service rush to the web. Retrieved November 2, 2015, from http://www.livinginternet.com/w/wi_online.htm

Sullivan, D. (2006, January 19). Bush administration demands search data; google says no; aol, msn & yahoo said yes. Retrieved from http://searchenginewatch.com/sew/news/2059843/bush-administration-demands-search-data-google-says-no-aol-msn-yahoo-said-yes

Sullivan, D. (2010, September 17). Schmidt: listing google's 200 ranking factors would reveal business secrets. Retrieved November 1, 2015, from http://searchengineland.com/schmidt-listing-googles-200-ranking-factors-would-reveal-business-secrets-51065

Sullivan, D. (2011a, February 4). Bing: why google's wrong with its accusations. Retrieved November 1, 2015, from http://searchengineland.com/bing-why-googles-wrong-in-its-accusations-63279

Sullivan, D. (2011b, February 1). Google: bing is cheating, copying our search results. Retrieved November 1, 2015, from http://searchengineland.com/google-bing-is-cheating-copying-our-search-results-62914

Taft, D. K. (2014, March 23). Microsoft using machine learning, ai across product lines. Retrieved January 1, 2015, from http://www.eweek.com/database/microsoft-using-machine-learning-ai-across-product-lines.html

Teevan, J., Alvarado, C., Ackerman, M. S., & Karger, D. R. (2004). The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 415–422). ACM.

Wall, A. (n.d.). History of search engines: from 1945 to google today. Retrieved October 31, 2015, from http://www.searchenginehistory.com/

WordStream. (n.d.). History of search engines - chronological list of internet search engines. Retrieved October 31, 2015, from http://www.wordstream.com/articles/internet-search-engines-history

World Wide Web Consortium. (n.d.-a). Facts about w3c - history. Retrieved November 1, 2015, from http://www.w3.org/Consortium/facts#history

World Wide Web Consortium. (n.d.-b). What is the difference between the web and the internet? Retrieved October 31, 2015, from http://www.w3.org/Help/#webinternet

Zimbardo, P. & Boyd, J. (2008). *The time paradox: the new psychology of time that will change your life*. Simon and Schuster.

Zimmer, M. (2008). The gaze of the perfect search engine: google as an infrastructure of dataveillance. In A. Spink & M. Zimmer (Eds.), *Web search* (Vol. 14, pp. 77–99). Information Science and Knowledge Management. Springer Berlin Heidelberg. doi:10.1007/978-3-540-75829-7_6