Mechanisms to Discover the Real News on the Internet

Yuta Nemoto
Graduate Department of Computer and Information Systems
University of Aizu
Aizu-Wakamatsu, Japan
e-mail: m5231153@u-aizu.ac.jp

Vitaly Klyuev
Software Engineering Lab
University of Aizu
Aizu-Wakamatsu, Japan
e-mail: vkluev@u-aizu.ac.jp

Abstract— **Over the last decade, news media on the Internet have been developing at a high pace. The latest advances in information technology have helped to influence the growth of the number of people searching for information from their mobile devices. On the other hand, the risk of incorrect or false information spreading has also become higher. This problem is serious on social media and users need to distinguish between what is true and what is false. Meaningful tools to support them are strongly demanded in today's society. In this paper, we discuss the implementation of a tool that helps users find real news on the Internet. The key feature is to encourage users to look at the information from an objective perspective. To achieve this, an approach based on the idea of the "metasearch engine" can be applied. Although the popularity of this instrument has declined since the rise of Google, the mechanism itself is effective in preserving the neutrality of search results.**

Keywords - Internet; fake news; metasearch; Web; development.

I. INTRODUCTION

Nowadays, the impact of the information provided by online news media on users is growing with a rapid increase in the number of people owning a mobile device. The largest portion of this growth is in the younger generation (Generation Z) because they are growing up in a highly sophisticated technological environment and becoming familiar with computers, specifically smartphones [1]. Most tend to satisfy their information needs through Internet media on their smartphones. In other words, they are living in the environment where they can always access the latest information sources. On the other hand, the risk of false information spreading on the Internet is growing because of the specific characteristics of online media: uncertainty and lack of responsibility for information sources, and the high speed of information delivery to recipients.

This situation has attracted scientists' attention since the US election in 2016 [2]. According to [3], more than 27 percent of American people who have the right to vote visited at least one website with fake news during that election campaign. The report also mentioned that social media, especially Facebook, played an integral role in exposing people to fake news. This means the issue of fake news is now a serious social problem that affects both politics and economics on a large scale. Tools to help users recognize and reject false information are strongly demanded.

In a modern highly mobile society, people of any age tend to use a smartphone for lots of different actions, such as making phone calls, taking pictures, communicating with peers via messenger app, searching for information in a browser, purchasing goods online, and writing blogs. The main tendency in this communication is to use one application. For many users, Facebook or Google are used. These information instruments analyze a user's query and provide relevant ads. Utilizing the same approach, they may hide key documents that are important for the user. This altering of the results may influence the view of the users on news stories, politicians, etc. Search engines may influence users in the decision-making process during election campaigns. Traditional search engines are not independent judges of document quality and hearings in the US Congress in December 2018 [17] illustrate this conclusion.

Increasing plurality in the search results and reducing biased searches is the main goal in our development. To achieve this, the proposed tool works in the manner of the metasearch engine [4] to propagate the user query to several general search engines. The merging mechanism should place representatives retrieved by these search engines on the first output page within the search results.

We discuss a practical method to implement this. We expect this tool to be helpful for users who want to find real news related to their information needs. The real vision of events, facts, etc. can be created by the user from the different pieces of information like a mosaic. These pieces of information will be created by the proposed tool.

The remaining part of the paper is organized as follows. In Section II, we look over other publications in this area. In Section III, we discuss how to develop the tool using the mechanism of the metasearch engine. In Section IV, we specify the requirements for the development, and report the current progress in Section V. In Section VI, we illustrate the usage scenario that we expect by using an example. We explain the difficulties in the development in Section VII. The concluding remarks are in Section VIII.

II. RELATED WORK

To see the bigger picture of the mass media news area, Yap et al. [5] presents and roughly classify solutions for the problem of fake news into two categories: proactive and reactive variants. Their final goal is the same: to minimize the effect of fake news. According to the proactive solutions, the study above explains that Internet users should educate themselves about the existence of fake news, and it is

ISBN: 978-1-61208-760-3

effective if they validate news by finding at least two or more sources to check the credibility of the researched information.

Klyuev [6] proposes an approach using the mechanism of the metasearch engine to simply provide the needed information to users. It can be an efficient way to allow mobile device users, who are the main players in the searching process, to have the information on their mobile device.

For the reactive solutions, as a major approach to curbing the expansion of misinformation especially on social networks, the most needed task is to identify the articles that require fact checking. Tschiatschek et al. [7] illustrate the possible approaches of detection using crowd signals: users vote for the candidates of online articles needing to be checked. Kim et al. [8] discuss the crowd-powered solution and presents an algorithm to assist in the decision-making processes. Also, the study by Reis et al. [9] discusses the system that successfully predicts the articles to be validated at an acceptable accuracy by clarifying the features of detection. However, it also points out that the final judgment depends on an expert. If the article is about sensitive issues such as politics, the decision is an even tougher task. This means that automating the process will be difficult, and the efficiency as a solution is limited in practice.

In this work, we focus on the approach to give an unbiased perspective to the users and propose a concrete schema of the tool that implements the mechanism of the metasearch engine. Its main idea is presented in [6].

III. Approach Overview

To design the system, we extend the proposition presented in [6]. The approach presents the following filtering schema for the search results.

- Only textual data are used as result items;
- The result items are classified into encyclopedias, famous news agencies, online newspapers, portals, and blogs, by matching the URL to the prepared list of news agencies and newspapers or by analyzing the content of the documents;
- The maximum number of result documents has the following limitations: no more than 9; the result should include an encyclopedia and 2 items for each category mentioned above;
- The latest documents (in terms of time) should be selected if there is more than one document from the same source;
- Previous steps form the pool of documents. The final list for presenting is created by picking up one or two documents of each category from the pool;
- The ranks of documents given by the search engines should also be reflected in the presented search results. They should be ordered randomly when multiple documents have the same rank.

This work defines the filtering schema for mobile devices and computers with a powerful CPU, by considering the difference of computing speed and response time. If the device is the computer equipped with high computational ability, the following step is added to the schema.

- Selection of highest-ranked and lowest-ranked documents is carried out if two or more documents are from the same source or the same category.

All search result items are shown on the screen after this selection process. They are ordered according to the ranks assigned to each document.

A different application will be used for mobile devices and PCs, with high processing ability, and the application will work as a stand-alone system. In other words, the applications can work without any server running the service. The program on the devices sends the query, collects and then orders the search results on the device. This explains the necessity to use different schemas for mobile devices and PCs.

IV. Implementation

The data process in this application consists of three layers: metasearch, selection of the result items to be presented, and presentation of the chosen information.

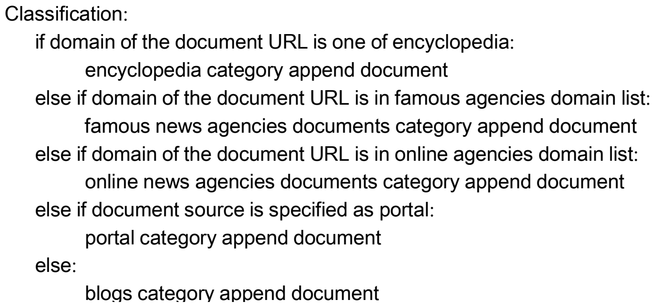
A. Metasearch Layer

According to [10], the metasearch engine's systems can be classified as follows.

- Real: They are similar to traditional search engines and work on the server.
- Semi-pseudo: They propagate queries to multiple search engines and present the results grouped by engines in a scrollable easy to read list.
- Pseudo: They open multiple search engine pages simultaneously in multiple browser windows/ frames.
- Client-side: Their components reside on a user's machine.

Although it needs frequent updating and client software installation, here we adopt the client-side metasearch approach because it is assumed that the form of online native applications for mobile devices is required by users ([11] reports that mobile Internet has grown more than 500% in daily media consumption since 2011).

To implement this part, we utilize the method of scraping in programming. Beautifulsoup [12] is one useful library to realize the scraping from multiple search systems. The created program sends requests to the predefined search engines, which include Google, Yahoo! and Bing. To send the query to these search engines, we need to prepare the methods that correspond with each engine to retrieve the search result pages. For example, in the case of Yahoo!, the search result page can be obtained by requesting the URL, "https://search.yahoo.com/search?p=[keyword]". When users send multiple keywords, the program needs to combine the words with '+'. The following example illustrates this: "hongkong+protest". After receiving the search result page, the HTML parser retrieves the needed elements for each result item on the page, such as the document (Web page) title, URL, page description, and rank in the search result.



```
Classification:
    if domain of the document URL is one of encyclopedia:
          encyclopedia category append document
    else if domain of the document URL is in famous agencies domain list:
          famous news agencies documents category append document
    else if domain of the document URL is in online agencies domain list:
          online news agencies documents category append document
    else if document source is specified as portal:
          portal category append document
    else:
          blogs category append document
```

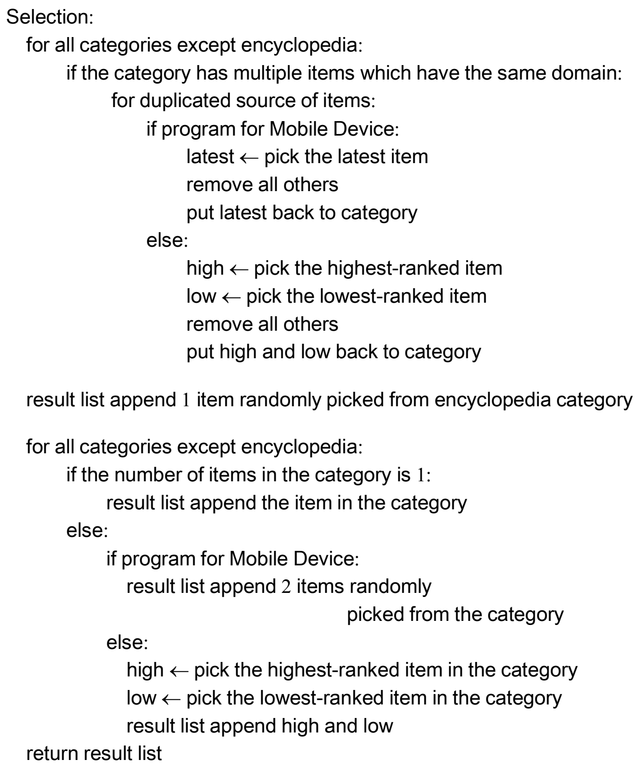
Figure 1. Pseudo-code for the classification

The program packs a set of this documented information as the result item object and goes to the analyzing step of the layer. In the analyzing step, it classifies each document into 5 categories (as mentioned in the previous section) and carries it to the next layer. The pseudo-code for this process is shown in Figure 1.

To specify whether the specific document is classified as "Portal" or "Blog", the program checks the content of the item. If the document is installed by the owner, it is classified as the item from a portal. In other cases, it is considered a blog item. The lists of famous news agencies, online newspaper agencies, and encyclopedia for source specifications are created in advance. They can be edited by the user. The domain part of the source URL and the page descriptions are the important factors in this classification.

B. Selection of the Result Items to be Presented to the User

This layer is the key part of the application. It works for choosing the information to be presented to the user. The following concrete method is necessary to get a maximum of 9 items: an encyclopedia page, 2 items from famous news agencies, 2 from online newspapers, 2 from portals, and 2 from blogs. Also, as mentioned in the previous section, the method chooses 2 items in the case of a PC if multiple documents are presented from the same source or belonging to the same category. Fixing the numbers mentioned above is carried out by analyzing ordinary users' behavior: most users look only at the first page. If they are not satisfied with the results of the search, they change the query. The number of documents presented to the user by general purpose search engines is in the range of 10 to 15. To increase the polarity of views on the topic of user interests, the method selects two documents from each category: they have the highest and lowest ranks in the retrieved set. This process is shown in Figure 2.



```
Selection:
   for all categories except encyclopedia:
        if the category has multiple items which have the same domain:
             for duplicated source of items:
                  if program for Mobile Device:
                       latest ← pick the latest item
                       remove all others
                       put latest back to category
                  else:
                       high ← pick the highest-ranked item
                       low ← pick the lowest-ranked item
                       remove all others
                       put high and low back to category

   result list append 1 item randomly picked from encyclopedia category

   for all categories except encyclopedia:
        if the number of items in the category is 1:
             result list append the item in the category
        else:
             if program for Mobile Device:
                result list append 2 items randomly
                                      picked from the category
             else:
                high ← pick the highest-ranked item in the category
                low ← pick the lowest-ranked item in the category
                result list append high and low
   return result list
```

Figure 2. Pseudo-code for the selection

In the case of mobile devices, selection from the same source is completed by choosing the latest document. Yet, in the case of a computer with high processing ability, the process is more complicated: it selects 2 items with the highest rank and the lowest rank. For example, in the case of 3 Web documents published by the BBC: news A with rank 1, news B with rank 1, and news C with rank 3 are classified in the "items from famous news agencies" category. It picks up A or B randomly and C is added as the second item from this class.

C. Presentation of Chosen Information

Finally, the aforementioned items are presented to the user on the screen. In particular, the interfaces are different on mobile devices to the display on a PC. It is more desirable to prepare optimal presentation format (character size, size of the description text, back/ forward functions, etc.) appropriate to the assumed display size. The responsive design technology should be applied for this purpose.

V. CURRENT PROGRESS

By the time of submission of this paper, we have realized the stage of obtaining the search results, classification and function to select the result items.

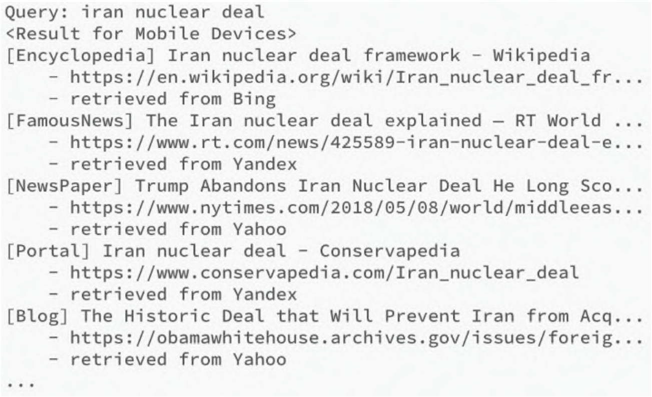
To retrieve the search results from several search engines, we use a Python library, Beautifulsoup [12], and APIs from third parties.

The classification of blogs and portals is specified by their URLs. For example, if the document is provided by predefined domains, such as "news.yahoo.com", the item is classified as a portal document. The function to check its content is under development because the formats of the documents on the search results depend on the publishers.

For the selection method, the app preserves the ranks in the search result given by each engine. It also checks the published date of the document. However, this part is not yet complete. We need to consider how to manage items without any date information. The presentation layer is planned to be ready after solving all of the aforementioned issues.

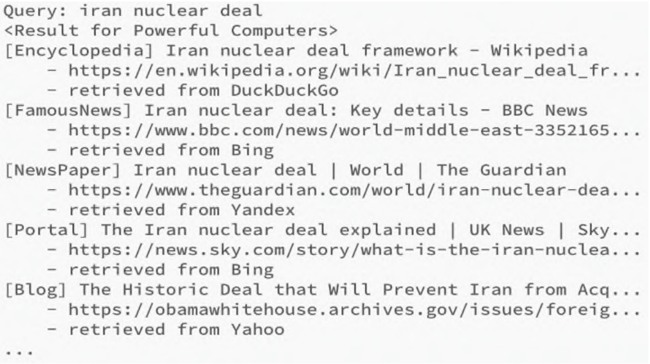
VI. RESULTS OF THE EXPERIMENTS

To demonstrate the implemented features, we illustrate the outcomes of the system using an example scenario:



```
Query: iran nuclear deal
<Result for Mobile Devices>
[Encyclopedia] Iran nuclear deal framework - Wikipedia
    - https://en.wikipedia.org/wiki/Iran_nuclear_deal_fr...
    - retrieved from Bing
[FamousNews] The Iran nuclear deal explained — RT World ...
    - https://www.rt.com/news/425589-iran-nuclear-deal-e...
    - retrieved from Yandex
[NewsPaper] Trump Abandons Iran Nuclear Deal He Long Sco...
    - https://www.nytimes.com/2018/05/08/world/middleeas...
    - retrieved from Yahoo
[Portal] Iran nuclear deal - Conservapedia
    - https://www.conservapedia.com/Iran_nuclear_deal
    - retrieved from Yandex
[Blog] The Historic Deal that Will Prevent Iran from Acq...
    - https://obamawhitehouse.archives.gov/issues/foreig...
    - retrieved from Yahoo
...
```

Figure 3. Example output for mobile devices with the query "Iran nuclear deal"



```
Query: iran nuclear deal
<Result for Powerful Computers>
[Encyclopedia] Iran nuclear deal framework - Wikipedia
    - https://en.wikipedia.org/wiki/Iran_nuclear_deal_fr...
    - retrieved from DuckDuckGo
[FamousNews] Iran nuclear deal: Key details - BBC News
    - https://www.bbc.com/news/world-middle-east-3352165...
    - retrieved from Bing
[NewsPaper] Iran nuclear deal | World | The Guardian
    - https://www.theguardian.com/world/iran-nuclear-dea...
    - retrieved from Yandex
[Portal] The Iran nuclear deal explained | UK News | Sky...
    - https://news.sky.com/story/what-is-the-iran-nuclea...
    - retrieved from Bing
[Blog] The Historic Deal that Will Prevent Iran from Acq...
    - https://obamawhitehouse.archives.gov/issues/foreig...
    - retrieved from Yahoo
...
```

Figure 4. Example output for powerful computers with the query "Iran nuclear deal"

results for the query "Iran nuclear deal". The output for the query on mobile devices and for powerful computers is shown in Figure 3 and Figure 4. There are 41 items collected including duplicate items: 6 from Bing, 15 from Yandex, 10 from Yahoo! and 10 from DuckDuckGo. Table I shows the classification of documents obtained (uncounted items are non-textual sources, such as YouTube links). The "Blogs" category is retrieved with the predefined set of URLs. Overall, the result pages look balanced compared to every source search system. The outcome may be even better if the sources of search include diverse types of engines.

TABLE I. RETRIEVED RESULTS FOR THE EXAMPLE QUERY

Docs class	***Encyclopedia***	***Famous News Agencies***	***Online News Papers***	***Portal Websites***	***Blogs***
No. of items	4	18	7	8	2

The implicit limitation is that this selection criteria does not always reflect the importance of documents presented by search engines, especially for results on mobile devices. There is a search result item entitled "Iran nuclear deal: Key details – BBC News". It is retrieved by all 4 systems and classified as the "Document from famous news media". However, the result for mobile devices does not include the item in this trial because the final selection of the items to be presented to the user is random after arranging the documents by publication date. The accuracy of the calculation level and the output level of the application for mobile devices should be adjusted after tests on real devices.

VII. DISCUSSION

Scraping is implemented for Yahoo!, Bing, Yandex and DuckDuckGo. We have difficulty with this phase for Google and Baidu: Google officially prohibits the computational scraping of search results. Although there are several APIs to scrape the search results from Google published on GitHub, most of them do not work correctly because Google also takes measures against scraping actions.

Baidu is another search engine of this kind. All of the links written on the results page are URLs to the Baidu server. All of the original URLs on the results page are stored in the Baidu database and users need to request the real ones by accessing the URL on the Baidu server. To incorporate Baidu in our metasearch subsystem safely, we need to find another way to retrieve these links.

Another problem is the weight of the search result items obtained from search engines. Right now, we consider the 6 main search engines: Google, Bing, Yahoo!, Baidu, Yandex, and DuckDuckGo. They have the most search engine market share in the world. However, not all of them provide the actual search engine system which crawls the Web. For example, Yahoo! uses Bing as the source to present its search results [13][14] and Yahoo! Japan switched its search technology to use Google [15]. If several search engines present the result items from the same source and our system evaluates them equally, the results from one engine may skew the search results. To develop the tool to work correctly, we need to remove these duplicated items from the search engine list or reduce the weighting for such items.

The way to evaluate the quality of the search outcomes is one of the foreseen difficulties. The metasearch system gives users less-biased search results. Hence, the advantage of this application for mobile and PC devices is in the relatively better neutrality of search results. Normally, the measure TREC-Style Average Precision (TSAP) [16] is used to evaluate the score by analyzing the relevance of the top N result items as a traditional way to assess the performance of the search engine. The points to quantify are not only the relevance of the search results to the intention of the given query, but also the fairness of the given results. In other words, we require a method that evaluates the bias affecting the search results or different ways to penetrate opinions in the result items. This is especially true when the searched issue is about politics: the presented results should include as many views as possible. The computational indicator should thus evaluate the obtained results objectively. Still, in many studies this process depends on experts' judgments in practice.

VIII. CONCLUSION

To develop a practical approach to discover real news on the Internet, the implementation of information retrieval from multiple search engines, classification, and the selection layer are reported as a work in progress. There are general and technical problems with compiling search results. Still, the discussion on how to deal with the search platforms having the same search system is also insufficient.

The next step in the development of the classification layer is to classify the items in the portals and blogs categories. We consider the general factors of the documents or domains, such as the amount of text, grammar, functional contents on the page, and so on.

Finally, the assessment method for the whole work including the evaluation of neutrality of the search result content is needed. We look forward to identifying a measure to determine the score of the approach comprehensively.

References

[1] M. Dimock, *Defining generations: Where Millennials end and Generation Z begins*. [Online]. Available from: http://tony-silva.com/eslefl/miscstudent/downloadpagearticles/defgenerations-pew.pdf, [retrieved: 1, 2020].

[2] D. M. J. Lazer et al., “The science of fake news,” *Science*, vol. 359, issue 6380, pp. 1094-1096, Mar. 2018, doi: 10.1126/science.aao2998

[3] A. Guess, B. Nyhan, and J. Reifler, *Selective Exposure to Misinformation: Evidence from the consumption of fake news during the2016 U.S. presidential campaign*. [Online]. Available from: http://www.ask-force.org/web/Fundamentalists/Guess-Selective-Exposure-to-Misinformation-Evidence-Presidential-Campaign-2018.pdf, [retrieved: 1, 2020].

[4] S. R. Lawrence and C. L. Giles, “Meta Search Engine” United States Patent 6,999,959 B1, Feb. 14, 2006

[5] A. Yap, L. G. Snyder, and S. Drye, “The Information War in the Digital Society: A Conceptual Framework for a Comprehensive Solution to Fake News,” *Academy of Social Science Journal*, vol. 03, issue 07, pp. 1214-1221, Jul. 2018, Available from: http://www.innovativejournal.net/index.php/assj/article/view/2202, [retrieved: 1, 2020].

[6] V. Klyuev, “Finding the Real News in News Streams,” 2019 Amity International Conference on Artificial Intelligence (AICAI), Dubai, United Arab Emirates, 2019, pp. 21-24. doi: 10.1109/AICAI.2019.8701334

[7] S. Tschiatschek, A. Singla, M. G. Rodriguez, A. Merchant, and A. Krause, “Fake News Detection in Social Networks via Crowd Signals,” Companion Proceedings of The Web Conference 2018 (WWW '18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 517-524, March 2, 2018. doi: 10.1145/3184558.3188722

[8] J. Kim, B. Tabibian, A. Oh, B. Schölkopf, and M. Gomez-Rodriguez, “Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation,” Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18). ACM, New York, NY, USA, pp. 324-332, November 27, 2017. doi: 10.1145/3159652.3159734

[9] J. C. S. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, “Supervised Learning for Fake News Detection,” *IEEE Intelligent Systems*, vol. 34, no. 2, pp. 76-81, March-April 2019. doi: 10.1109/MIS.2019.2899143

[10] M. Manoj and E. Jacob, “Information retrieval on Internet using meta-search engines: A review,” *Journal of Scientific & Industrial Research,* vol. 67, pp.739-746, October 2008, Available from: http://ir.niist.res.in:8080/jspui/handle/123456789/1953, [retrieved: 1, 2020].

[11] A. Rodriguez, *The internet is finally going to be bigger than TV worldwide*. [Online]. Available from: https://qz.com/1303375/internet-usage-will-finally-surpass-tv-in-2019-zenith-predicts/ 2018.06.13, [retrieved: 1, 2020].

[12] Crummy, *Beautiful Soup*. [Online]. Available from: https://www.crummy.com/software/BeautifulSoup/ 2019.11.09, [retrieved: 12, 2019].

[13] BBC News, *Microsoft and Yahoo seal web deal*, July 29, 2009. [Online]. Available: http://news.bbc.co.uk/2/hi/business/8174763.stm 2009.07.29, [retrieved: 1, 2020]

[14] L. Dirk, “Living in a world of biased search engines,” *Online Information Review*, vol. 39, issue. 3, pp. 278-280, June 8, 2015. doi: 10.1108/OIR-03-2015-0089.

[15] H. Tabuchi, *Yahoo Japan Teams With Google on Search*. [Online]. Available from: https://www.nytimes.com/2010/07/28/technology/28yahoo.html 2010.07.27, [retrieved: 1, 2020].

[16] D. Hawking, N. Craswell, P. Bailey, and K. Griffiths, “Measuring Search Engine Quality,” *Information Retrieval (2001),* vol. 4, issue. 1, pp. 33-59, November 28, 2000. doi: 10.1023/A:1011468107287

[17] D. Wakabayashi and C. Kang, *Google’s Pichai Faces Privacy and Bias Questions in Congress*. [Online]. Available from: https://www.nytimes.com/2018/12/11/technology/google-pichai-house-committee-hearing.html 2018.12.11, [retrieved: 1, 2020]