An Evaluation of Reverse Image Search Performance of Google

Yıltan Bitirim
Computer Engineering Department
Eastern Mediterranean University
Famagusta, North Cyprus, via Mersin 10, Turkey
yiltan.bitirim@emu.edu.tr

Selin Bitirim
Computer Engineering Department
Eastern Mediterranean University
Famagusta, North Cyprus, via Mersin 10, Turkey
selin.bitirim@emu.edu.tr

Duygu Çelik Ertuğrul
Computer Engineering Department
Eastern Mediterranean University
Famagusta, North Cyprus, via Mersin 10, Turkey
duygu.celik@emu.edu.tr

Önsen Toygar
Computer Engineering Department
Eastern Mediterranean University
Famagusta, North Cyprus, via Mersin 10, Turkey
onsen.toygar@emu.edu.tr

***Abstract*—This study investigates reverse image search performance of Google, in terms of Average Precisions (APs) at various cut-off points, on finding out similar images by using fresh Image Queries (IQs) from the five categories "Fashion", "Computer", "Home", "Sports", and "Toys", in order to have an insight about reverse image search performance of Google and then, motivate the researchers and inform the users. Five fresh IQs with different main concepts were created for each of the five categories. These 25 IQs were run on the search engine and for each, the first 100 images retrieved were evaluated with binary relevance judgment. APs at the cut-off points 20, 40, 60, 80, and 100 were calculated for each category and for all 25 IQs. The performance range is from ~42% for Toys category at the cut-off point 100 to 71% for Home category at the cut-off point 20. When the categories are ignored, Google's performance range is from ~52% at the cut-off point 100 to ~57% at the cut-off point 20. It seems that reverse image search performance of Google needs to be improved.**

Keywords—reverse image search, Google, search engine, performance evaluation

I. INTRODUCTION

There are huge amount of available images on the Internet and they are steadily increasing [1]. Therefore, image search has been a necessity for the users. In order to search images on the Internet, one of the alternatives is to use reverse image search. Reverse image search is a query method of content-based image retrieval that is used to provide the content-based image retrieval system that searches based on the sample image given [2].

Google is the most popular search engine in the world [3]. It is one of the search engines that uses reverse image search method. Reverse image search can be used for various reasons such as finding out similar images, authenticating images, and tracking images. For finding out similar images, Image Query (IQ) that is either based on an image on the Internet or fresh can be used.

In this study, we investigated reverse image search performance of Google, in terms of Average Precisions (APs) at various cut-off points, on finding out similar images by using fresh IQs from the five categories "Fashion", "Computer", "Home", "Sports", and "Toys". It is aimed to have an insight about reverse image search performance of Google and then, motivate the researchers and inform the users.

Some of the related studies encountered in the literature are [4-10].

This paper is organized as follow: Section 2 describes the methodology employed in details; the next section is experiment results and discussion; and section 4 gives the conclusion and the future work.

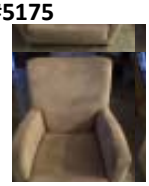
II. METHODOLOGY

First of all, five categories were determined for IQs. These categories are "Fashion", "Computer", "Home", "Sports", and "Toys". Then, five main concepts were determined for each category. These main concepts are as follows: For Fashion, "earring", "watch", "bag", "shirt", and "dress"; for Computer, "mouse", "keyboard", "computer case", "speaker", and "printer"; for Home, "candle", "curtain", "armchair", "vase", and "stock pot"; for Sports, "dumbbell", "wristband", "bicycle", "ball", and "mat"; and for toys, "play vehicle", "springer roller", "building block", "kung fu panda", and "doll". Later, a photo of a related item was taken for each main concept. (The same device that is "Apple iPhone XS" was used for photo shooting of every item.) Afterwards, every photo has been cropped in order to bring main concept of the corresponding photo to the fore more. Finally, the images were obtained to be used as fresh IQs. List of the IQs is given in Table I in detail. Note that the original IQs are available at "https://doi.org/10.6084/m9.figshare.12336275.v1".

During the searches of the IQs, Google account was kept as signed out and settings in Google were used as follows: (*) "Search settings" -> "Search results" -> "Safe search filters" -> off; (*) "Search settings" -> "Search results" -> "Results per page" -> "100"; (*) "Search settings" -> "Search results" -> "Spoken answers" -> "Speak answers for voice search"; (*) "Search settings" -> "Search results" -> "Where results open" -> "Open each selected result is in a new browser window" is unchecked; (*) "Search settings" -> "Search results" -> "Search

978-1-7281-7303-0/20/$31.00 ©2020 IEEE
DOI 10.1109/COMPSAC48688.2020.00-65

TABLE I. LIST OF THE IQS

IQ #5121  Cat.: Fashion M.C.: Earring S.: 1503*1105	**IQ #5122**  Cat.: Fashion M.C.: Watch S.: 3333*729	**IQ #5123**  Cat.: Fashion M.C.: Bag S.: 2827*2006	**IQ #5125**  Cat.: Fashion M.C.: Shirt S.: 2596*3406	**IQ #5128**  Cat.: Fashion M.C.: Dress S.: 1884*3689
IQ #5145  Cat.: Computer M.C.: Mouse S.: 2016*1186	**IQ #5473**  Cat.: Computer M.C.: Keyboard S.: 3911*1428	**IQ #5476**  Cat.: Computer M.C.: Computer case S.: 2336*3913	**IQ #5477**  Cat.: Computer M.C.: Speaker S.: 3071*2509	**IQ #5480**  Cat.: Computer M.C.: Printer S.: 2440*2235
IQ #5043  Cat.: Home M.C.: Candle S.: 1074*2282	**IQ #5174**  Cat.: Home M.C.: Curtain S.: 2197*3908	**IQ #5175**  Cat.: Home M.C.: Armchair S.: 2336*3913	**IQ #5180**  Cat.: Home M.C.: Vase S.: 1727*3436	**IQ #5717**  Cat.: Home M.C.: Stock pot S.: 2837*1848
IQ #5272  Cat.: Sports M.C.: Dumbbell S.: 3071*1970	**IQ #5283**  Cat.: Sports M.C.: Wristband S.: 2451*1594	**IQ #5301**  Cat.: Sports M.C.: Bicycle S.: 2916*2796	**IQ #5305**  Cat.: Sports M.C.: Mat S.: 2156*3895	**IQ #5330**  Cat.: Sports M.C.: Ball S.: 2522*2454
IQ #5222  Cat.: Toys M.C.: Play vehicle S.: 2774*2119	**IQ #5231**  Cat.: Toys M.C.: Spring roller S.: 2065*1780	**IQ #5235**  Cat.: Toys M.C.: Building block S.: 2619*2253	**IQ #5241**  Cat.: Toys M.C.: Kung fu panda S.: 2222*2035	**IQ #5257**  Cat.: Toys M.C.: Doll S.: 1715*4003
Note: "Cat." stands for "Category"; "M.C." stands for "Main Concept"; and "S." stands for "Size".				

customisation" -> "Signed out search activity" -> off; (*) "Search settings" -> "Search results" -> "Region settings" -> "Turkey"; (*) "Search settings" -> "Languages" -> "Which language should Google products use?" -> "English"; (*) "Search settings" -> "Languages" -> "Currently showing search results in" -> "English" & "Turkish"; (*) "Your data in search" -> "Ad personalization" -> "Ads Personalization Across the Web" was kept as default; and (*) "Your data in search" -> "Ad personalization" -> "Ads Personalization on Google Search" was kept as default.

The main page of Google ("https://www.google.com") was opened on the browser "Google Chrome" and "Images" was clicked. "Google images" page came. "Search by image" was clicked and IQ was uploaded. Afterwards, by clicking "Visually similar images" on the coming page, the full list of images retrieved were displayed. For each IQ in Table I, search was done on Google with these steps and the first 100 images retrieved were evaluated one by one with binary relevance judgment (relevant or non-relevant). If the corresponding IQ's main concept itself was seen fully or partially on an image retrieved, that image was accepted as relevant; otherwise, that image was accepted as non-relevant.

Main concepts were used in the crop operations of the photos. Furthermore, each one was also used while examining the corresponding images retrieved. Main concepts have broad meanings; for example, "bag" is the main concept for IQ #5123. Instead of "main" concepts, "specific" concepts could be used; for example, it could be "leather bag" or "black leather bag" for IQ #5123. However, using main concepts was decided since it was assumed that the user's intention is to look for broad meaning.

Precision is the fraction of retrieved items that are relevant as shown in (1). Precisions at various cut-off points, i.e. 20, 40, 60, 80, and 100, were calculated for every IQ. Afterwards, APs at these cut-off points were calculated for each category and for all 25 IQs.

$$Precision = \frac{number\ of\ relevant\ items\ retrieved}{number\ of\ retrieved\ items} \quad (1)$$

III. Experiment Results and Discussion

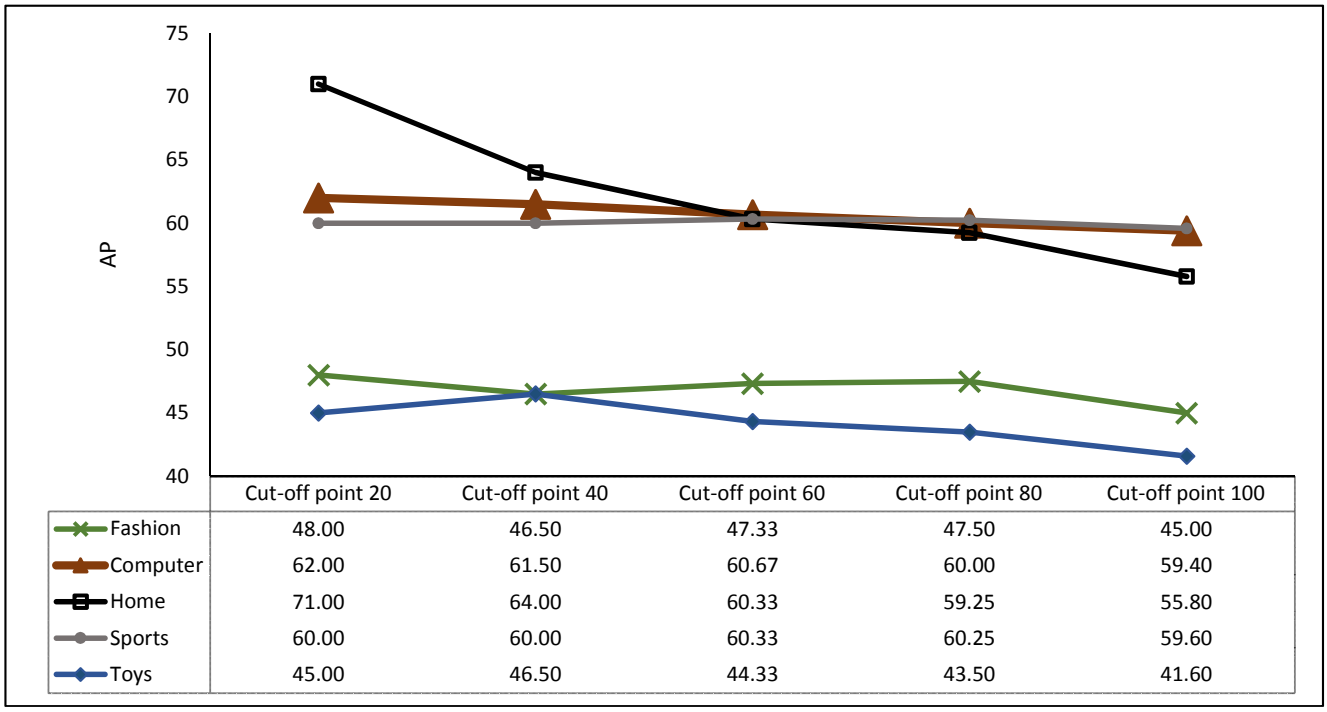
Searches were done between December 09, 2019 and January 03, 2020. The number of relevant images retrieved and the number of non-relevant images retrieved for 25 IQs from five categories are given in Table II. The first 100 images retrieved were evaluated for each IQ, except IQ #5473. For IQ #5473, total 31 images were retrieved; therefore, only 31 images were evaluated. Total 2431 images were evaluated for all IQs, 1238 of them were relevant and 1193 of them were non-relevant. No non-relevant image was retrieved for four IQs, i.e. #5145, #5473, #5301, and #5330, and no relevant image was retrieved for three IQs, i.e. #5477, #5283, and #5235. Two IQs with no non-relevant image and one IQ with no relevant image are from Computer category. The same also happened for Sports category. The last one from the IQs with no relevant image is in Toys category.

During retrieval output evaluations, some anomalies were noticed. The anomalies and the solutions to them are as follows. (*) For IQ #5305 -> Almost the same images which had very tiny differences because of editing were seen in the retrieval output. Since these images had different Web addresses, they were considered as different images. Furthermore, this situation happened more than one time. (*) For IQ #5257 -> The same item's two images with different poses were seen in the retrieval output. Since they had different Web addresses, they were considered as different images. (*) For IQ #5717 -> The same item's two images with different poses were seen in the retrieval output. In addition to this, another same item's three images with different poses were seen in the retrieval output. In both situations, the corresponding images were considered as different images since they had different Web addresses.

TABLE II. Number of Relevant/Non-Relevant Images Retrieved

Category	IQ #	# of relevants	# of non-relevants	Total
Fashion	5121	99	1	**100**
	5122	3	97	**100**
	5123	19	81	**100**
	5125	50	50	**100**
	5128	54	46	**100**
Computer	5145	100	0	**100**
	5473	31	0	**31**
	5476	93	7	**100**
	5477	0	100	**100**
	5480	4	96	**100**
Home	5043	5	95	**100**
	5174	86	14	**100**
	5175	41	59	**100**
	5180	92	8	**100**
	5717	55	45	**100**
Sports	5272	97	3	**100**
	5283	0	100	**100**
	5301	100	0	**100**
	5305	1	99	**100**
	5330	100	0	**100**
Toys	5222	62	38	**100**
	5231	1	99	**100**
	5235	0	100	**100**
	5241	90	10	**100**
	5257	55	45	**100**
	Total	**1238**	**1193**	**2431**

APs at the five cut-off points (i.e. 20, 40, 60, 80, and 100) for each of the five categories (i.e. Fashion, Computer, Home, Sports, and Toys) are shown in Fig. 1. Although Home has the highest APs at the cut-off points 20 and 40, its AP decreased when the cut-off point increased. Computer has the highest AP at the cut-off point 60, while Sports category is the best at the cut-off points 80 as well as 100. Computer's APs decreased slightly, when the cut-off point increased. Sports category has 60% or very close to 60% APs at all the cut-off points and it is the most stable category based on all the cut-off points. Although sometimes Computer category has higher AP than Sports category and sometimes Sports category has higher AP than Computer category, the gap between Computer category's AP and Sports category's AP decreased when the cut-off point increased. Home, Computer, and Sports have the same or almost the same APs at the cut-off point 60 and they have almost the same APs at the cut-off point 80. Computer and Sports have almost the same APs at the cut-off point 100. Fashion and Toys have exactly the same AP at the cut-off point 40. They are the worst at this cut-off point; however, Toys category is the worst alone at the other cut-off points. The best AP, 71%, is obtained at the cut-off point 20 for Home category and the worst AP, 41.60%, is obtained at the cut-off point 100 for Toys category. When the average of APs at all the cut-off points is considered for every category, the performance rank of the categories is as



	Cut-off point 20	Cut-off point 40	Cut-off point 60	Cut-off point 80	Cut-off point 100
Fashion	48.00	46.50	47.33	47.50	45.00
Computer	62.00	61.50	60.67	60.00	59.40
Home	71.00	64.00	60.33	59.25	55.80
Sports	60.00	60.00	60.33	60.25	59.60
Toys	45.00	46.50	44.33	43.50	41.60

Fig. 1. APs at the cut-off points for each of the five categories.

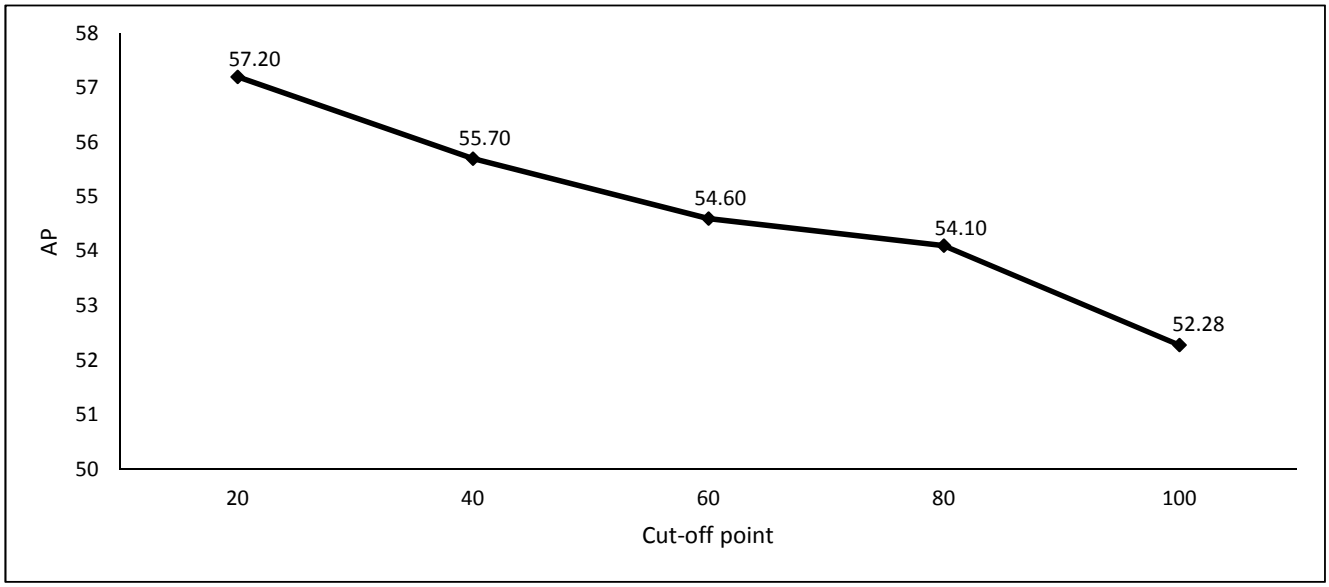


Fig. 2. APs at the cut-off points for all IQs.

follow: Home with 62.08%, Computer with 60.71%, Sports with 60.04%, Fashion with 46.87%, and Toys with 44.19%. Generally speaking; mostly, about half of the images on the retrieval output were not related to what the user was looking for.

APs at the cut-off points 20, 40, 60, 80, and 100 for all 25 IQs are given in Fig. 2. It is obviously seen that AP decreased when the cut-off point increased. This means that the search engine's performance decreased when the cut-off point increased. The search engine got its best performance at the cut-off point 20 with 57.20% AP. The user evaluates images on the retrieval output easily and fast; so, the user can check more and more. From the figure, it seems that when more images were evaluated, the performance decreased. At the cut-off point 100,

the performance decreased to 52.28%. When the average of APs at all the cut-off points is considered, the performance is 54.78%. Generally speaking; almost half of the images on the retrieval output were not related to what the user was looking for.

IV. Conclusion and Future Work

Reverse image search performance of Google, in terms of APs at various cut-off points, on finding out similar images by using fresh IQs from the five categories "Fashion", "Computer", "Home", "Sports", and "Toys" is investigated in this study.

For each of Computer, Sports, and Toys categories, no relevant image was display on the retrieval output for one IQ. The performance range is from ~42% (Toys – cut-off point 100) to 71% (Home – cut-off point 20). The highest performances on the cut-off points 20, 40, 60, 80, and 100 are obtained for Home, Home, Computer, Sports, and Sports, respectively. In the same manner, the lowest performances are obtained for just Toys, except the cut-off point 40 that Fashion is with Toys.

When the categories are ignored, Google's performance range is from ~52% (cut-off point 100) to ~57% (cut-off point 20). When the cut-off point increased -more images were evaluated-, the performance decreased.

The search engine has troubles to recognize the same images.

It is believed that reverse image search performance of Google needs to be improved.

Searches would be made for both main concepts and specific concepts; then, it would be studied to find out how the performance of Google changes when the user makes searches for specific concepts instead of main concepts.

Data Availability Statement

The original IQs are available at "https://doi.org/10.6084/m9.figshare.12336275.v1".

References

[1] A. Karczmarczyk, J. Jankowski, and W. Sałabun, "Linguistic query based quality evaluation of selected image search engines," *Procedia Computer Science*, vol. 112, pp. 1809-1818, 2017, doi: 10.1016/j.procs.2017.08.215.

[2] Wikipedia, "Reverse image search." https://en.wikipedia.org/wiki/Reverse_image_search (accessed February, 2020).

[3] StatCounter, "Search engine market share worldwide." https://gs.statcounter.com/search-engine-market-share#monthly-201901-202001-bar (accessed February, 2020).

[4] P. Nieuwenhuysen, "Image search process in the Web using image copy," *Journal of Multimedia Processing and Technologies*, vol. 9, no. 4, pp. 124-133, December 2018, doi: 10.6025/jmpt/2018/9/4/124-133.

[5] K. Smelyakov, D. Sandrkin, I. Ruban, M. Vitalii and Y. Romanenkov, "Search by image. New search engine service model," *2018 International Scientific-Practical Conference - Problems of Infocommunications. Science and Technology (PIC S&T)*, Kharkiv, Ukraine, 2018, pp. 181-186.

[6] N. Hegde et al., "Similar image search for histopathology: SMILY," *npj Digital Medicine*, vol. 2, no. 56, pp. 1-9, 2019, doi: 10.1038/s41746-019-0131-z.

[7] E. Kelly, "Reverse image lookup of a small academic library digital collection," *Codex: the Journal of the Louisiana Chapter of the ACRL*, vol. 3, no. 2, pp. 80-92, 2015.

[8] F. H. D. Araujo et al., "Reverse image search for scientific data within and beyond the visible spectrum," *Expert Systems with Applications*, vol. 109, pp. 35-48, November 2018, doi: 10.1016/j.eswa.2018.05.015.

[9] P. M. Chutel and A. Sakhare, "Evaluation of compact composite descriptor based reverse image search," *2014 International Conference on Communication and Signal Processing*, Melmaruvathur, 2014, pp. 1430-1434.

[10] D. Guinness, E. Cutrell, and M. R. Morris, "Caption Crawler: Enabling Reusable Alternative Text Descriptions using Reverse Image Search," *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*, New York, USA, April 2018, paper no. 518, pp. 1–11, doi: 10.1145/3173574.3174092.