Google Web Searches and Wikipedia Results: a Measurement Study*

Vittoria Cozza[1], Van Tien Hoang[2], and Marinella Petrocchi[3]

[1] Electrical & DEI, Polytechnic University of Bari, Italy
vittoria.cozza@poliba.it
[2] IMT Institute for Advanced Studies, Lucca, Italy
vantien.hoang@imtlucca.it
[3] Institute for Informatics and Telematics (IIT-CNR), Pisa, Italy
m.petrocchi@iit.cnr.it

Abstract. How are users exposed to Wikipedia results, in return to their web searches? Where are such results positioned on the screen? In this study, we experimentally measure the ranking of Wikipedia pages on Google Italia.

1 Introduction

As observed by a recent article of Nature News [5], "Wikipedia is among the most frequently visited websites in the world and one of the most popular places to tap into the world's scientific and medical information". One of the seventh most visited websites[4], the online encyclopaedia is a dominant source of Internet knowledge. Remarkably, a 2012 study assessed that Wikipedia pages appeared in 96% of the results all the searches through Google UK [7]. In his popular work on Filter Bubbles [6], Pariser was one among the first ones to theorize the phenomenon according to which users are unknowingly trapped into protective bubbles, created by search engines and social platforms to automatically filter contents. Given that users usually focus on the first few results of their web searches [1, 4], the exclusive privilege of Wikipedia at the very first positions could bias the informative content reachable on the Internet.

In this study, we measure the ranking of Wikipedia pages on Google Italia. To the best of our knowledge, it is the first study of this kind for the Italian language. The procedure is as follows. For our web searches, we concentrate on Italian keywords (and set of keywords). To collect the most popular search keywords, they have been chosen from Google Trend,from the Google Display Planner[5], and from the Italian trending words on Twitter. Google Trend gives the most searched terms in a year, as well as trending searches in the past 24 hours[6] and trending searches in the recent past[7]. Google Display Planner is a

* Work partially funded by the Registro.it project MIB (My Information Bubble)

[4] http://www.alexa.com/topsites (7th March, 2016)
[5] https://support.google.com/adwords/answer/3056115?hl=en
[6] https://www.google.com/trends/home/all/IT
[7] https://www.google.com/trends/hottrends#pn=p27

tool providing a series of websites linked to specific categories. It has also been used, to look for suggested keywords tied to particular categories, like, e.g., Sport and Vehicles. We have performed all the searches on Google Italia, from Italy. To avoid personalised results, we have used newly created browser instances and we have simulated users not logged into Google. The default settings for searches were the Google default settings. Among the obtained results, we considered only organic results and not sponsored one (i.e., Advertisements, Google News, and so on). In the following, we present the experiments and the results.

2 Experiments and results

Experiment settings. Our reference date is April 7th, 2016. We have collected the top search terms on Google Italia from 2011 to 2015. Also, we have extracted the trending news for the reference date and the hot trends stories of the ten days prior to that date. As a whole, we obtained 1169 unique search terms. For a wider view, we have further gathered the top Italian trending keywords from Twitter (updated three times and within three hours, on April 7th, 2016), leading to 40 unique terms from Twitter.

For the experiments, we have extended AdFisher[2], an automated tool for information flow experiments, freely available at GitHub[8]. In our work, AdFisher runs browser-based experiments that emulate search queries and store the results. AdFisher interacts with Selenium, a web browser automation tool. Selenium allows to run a unique instance of Firefox creating a fresh profile, with new associated cookies.

For each keyword (or keywords set), a new browser instance has been launched and we have searched such keyword(s). The browser instance has been destroyed after saving the query results. New browser instances have been used to avoid the so called carry-over effects [3], which would lead to results for the current search being influenced by the previous searches.

Examples of keywords that we have searched for are "Elezioni presidenziali negli Stati Uniti d'America del 2016" "Credito Valtellinese", "Una lama di luce". The complete list of keywords (and keywords set) is at `https://goo.gl/9KasJc`.

Results. Experiments were performed with more than 1,200 keywords, spanning 33 categories, with an average of 21 keywords per category. In order to evaluate the search results, we have considered only those keywords for which the corresponding Wikipedia link appeared in the first page of the result list - this corresponds to 708 keywords. Searching for those keywords, we found that Wikipedia pages in the result lists are ranked *1st* 41.1% and *2nd* 19.9%. Overall, they appear in the first five positions 78.8 times over 100. Figure 1 shows the occurrence of Wikipedia pages according to their position.

Searching keywords belonging to "Topic Emergenti" and "Mostre d'Arte" always yield Wikipedia links as the first result. Searching keywords belonging to "Assicurazioni", "Offerte di Lavoro" and "Economia e Finanza" yield Wikipedia links as first result for less than 10% of such keywords, see Figure 2.

[8] `https://github.com/tadatitam/info-flow-experiments`

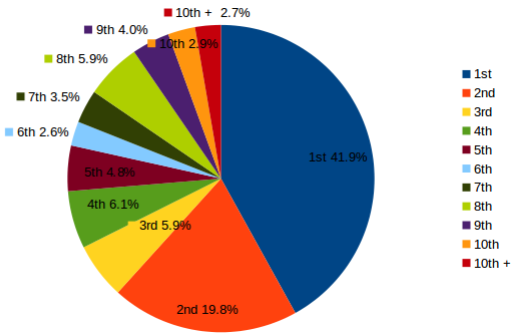

Fig. 1. How frequent Wikipedia pages are ranked *i-th*, in return to *google.it* searches

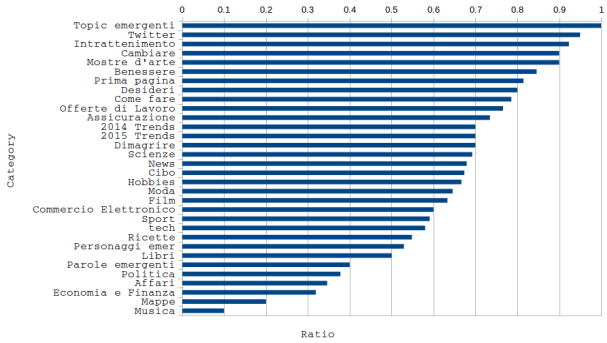

Fig. 2. Result pages with a Wikipedia link as 1st result, per category

Amongst all, "Mostre d'Arte" is a *top* category: all the keywords belonging to that category always lead to results with a Wikipedia link located within the first three Google Italia results. Examples of keywords are "Picasso - Milano", "Renoir - Pavia" and "Leonardo - Venaria". "Scienze" and "Benessere" are top categories too. More than 50% of searches of related keywords lead to Wikipedia results in the top three positions, see Figure 3. Instead, only 2% of keywords in the "News" category lead to results in the top three positions, while results for keywords related to "Politica", "Economia e Finanza", and "Hobbies" are ranked in the top three positions less than 10% of times (Figure 3). Figure 2 and Figure 3 show only the categories with the largest number of keywords.

Work in [7] performed a similar analysis on Google UK, focusing on encyclopaedic subjects, like scientific and natural sciences. Wikipedia scored extremely well, being its links in the top two result positions. The keywords belonging to the "Scienze" category (the Italian word for "Science") obtained more than 60% of Wikipedia links in the top three result positions.

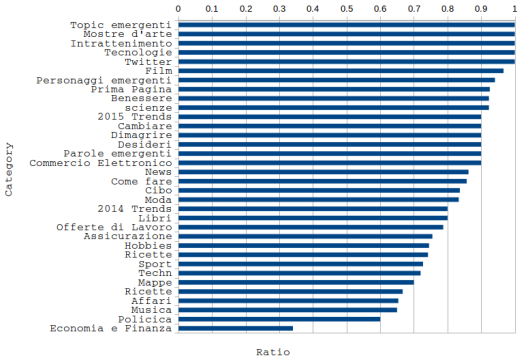

Fig. 3. Result pages with a Wikipedia link in the top three results, per category

3 Conclusions

This preliminary study measures the position of Wikipedia links, resulting from searching set of keywords over Google Italia. The outcome provides evidence that Wikipedia is dominant on Google Italia search results ranking: more than 78% of the times, there is one Wikipedia page within the first five search results. A closer look shows that keywords related to "Mostre d'Arte" category always have an associated Wikipedia page in the top three results, while those related to "News" are less than 10%. We have focused on quantifying the phenomenon, without investigating the semantics motivation behind the difference rankings for categories. We leave this study as a future work.

References

1. Edward Cutrell and Zhiwei Guan. What are you looking for?: An eye-tracking study of information usage in web search. In *Human Factors in Computing Systems*, CHI '07, pages 407–416. ACM, 2007.
2. Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *CoRR*, abs/1408.6491, 2014.
3. Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. Measuring personalization of web search. In *22nd World Wide Web*, WWW '13, pages 527–538. ACM, 2013.
4. Nadine Höchstötter and Dirk Lewandowski. What users see - structures in search engine results pages. *Inf. Sci.*, 179(12):1796–1812, 2009.
5. Richard Hodson. Wikipedians reach out to academics. *Nature News*, Sept. 2015.
6. Eli Pariser. *The Filter Bubble: What the Internet Is Hiding from You.* Penguin Group , The, 2011.
7. Sam Silverwood-Cope. Wikipedia: Page one of Google UK for 99% of searches. *pi-datametrics.com Blog*, 2012.