Formulating a Method for Using Search Query Trends as a Measure of Mass-User Interest

Shaown Sarker
N. C. State University
ssarker@ncsu.edu

Andrew McNamara
N. C. State University
ajmcnama@ncsu.edu

Jessica Staddon
N. C. State University
jessica.staddon@ncsu.edu

1 Introduction

It is considerably difficult to estimate the mass-user interest in a specific topic over time often due to the sensitivity of the self-reported data to bias such as privacy related surveys, which is a part of the privacy paradox. In 2015, Pew Research Center published a report discussing how the on-line activities and habits of American adults have changed since Edward Snowden began leaking documents about government intelligence programs [6]. Their survey identified how people changed their on-line behavior as they became aware of surveillance. These findings indicate that revelations of privacy-invasive programs make people change their on-line behaviors to better protect their sensitive data. Privacy-related decision-making like this has been actively studied for some time with much of the research focusing on the "privacy paradox" in which behavioral decisions seem at odds with reported preferences (see, for example, [2, 3]). In our work, we propose a novel methodology to measure the population's interest over time by creating a robust index based on on-line searches and leveraging of the search trends of the queries in this index. We verify our index-based process using the Pew Research Center report on the Snowden revelations.

2 Methodology

Our process is divided into two parts - constructing the index and creating a combined search trend for the index. For the index, we start by adding search queries that broadly represent the topic being explored, naming these as 'seed queries', to the index. We then proceed to add other search queries similar to the seed queries to the index. Our measure of query similarity uses the widely known web-kernel similarity function which determines similarity between two short snippets of text based on how similar are their on-line search results.

Researchers have been using Google trend data for forecasting [4] and estimating public interest over time [5] for sets of single search keyword. But, in our work we scale and create a representative trend over time for an entire search index. To achieve this, we retrieve the chronological search trend data from Google Trends [1] for each search query in the index. Then we standardize each of the trend data to a common search query and combine them. This process is repeated until we end up with an overall trend for our entire index. This overall trend represents the interest over time for our index topic.

3 Verification

We used the Pew Research Center report on the change in on-line privacy-related behaviors post Edward Snowden revelations [6] for verification. From the Pew report, of the 87% of respondents who were familiar with the surveillance program, 34% had taken at least one step to hide or shield their information from the government, while 25% had claimed to change their technological patterns of use. We identified 8 main themes of behavioral changes. As an initial investigatory step, we conducted a survey on Amazon Mechanical Turk (AMT) to determine how individuals might learn more about these topics and how to achieve specified tasks (available options were "ask a friend", "visit a specific website", "read a book about it", "use a search engine", "use something else", and "none of these"). The most common method of finding out more information for each of the topics was to use a search engine (56%) with Google being the most common one; the second most common method was to ask a friend (28%).

We started with 51 seed queries over these 8 identified categories and created an index containing 5203 search queries. We then created our overall trend from 3173 Google trend chronological data over nearly 6 years spanning on before and after the Snowden revelations. Using this overall trend, we are able to answer our research questions regarding the validity of the index and the association between the Pew Research report results and the overall trend.

References

[1] `https://www.google.com/trends/`.

[2] A. Acquisti and J. Grossklags. Privacy and rationality in individual decision making. *IEEE Security & Privacy*, 2(2005):24–30, 2005.

[3] S. B. Barnes. A privacy paradox: Social networking in the united states. *First Monday*, 11(9), 2006.

[4] A. F. Dugas, M. Jalalpour, Y. Gel, S. Levin, F. Torcaso, T. Igusa, and R. E. Rothman. Influenza forecasting with google flu trends. *PloS one*, 8(2):e56176, 2013.

[5] S. Funk and D. Rusowsky. The importance of cultural knowledge and scale for analysing internet search data as a proxy for public interest toward the environment. *Biodiversity & Conservation*, 23(12):3101–3112, Nov. 2014.

[6] L. Rainie and M. Madden. Americans' Privacy Strategies Post-Snowden. Technical report, Pew Internet & American Life Project, March 16 2015. Accessed: December 12 2016.