Instability in Search Engine Results

Lessons learnt in the context of horizon scanning applications

Marco A. Palomino and Tim Taylor
ECEHH
University of Exeter Medical School
Truro, UK
[m.palomino, tjt205]@exeter.ac.uk

Geoff McBride
STFC Futures Programme
Rutherford Appleton Laboratory
Didcot, UK
geoff.mcbride@stfc.ac.uk

Richard Owen
University of Exeter
Business School
Truro, UK
r.j.owen@exeter.ac.uk

***Abstract*—Horizon scanning, the systematic search for information to identify potential threats, risks, emerging issues and opportunities, has become an increasingly important part of strategic decision making. Although horizon scanning has its roots in the pre-electronic information era, it has blossomed with the availability of Web-based information. Dedicated analysts responsible for scanning the horizon make frequent use of search engines to retrieve information. Regrettably, the results yielded by popular search engines are often inconsistent and redundant. Thus, post processing heuristics have to be employed to select the most relevant data. This paper focusses on the first steps of this process, and analyses the result counts provided by different search engine interfaces in response to a set of queries meant to gather information about new and emerging trends.**

Search engines; search engine instability; Web mining; trend discovery and tracking; horizon scanning

I. INTRODUCTION

The use of text mining techniques for futures research has increased recently [1]. The use of the World Wide Web for the extraction of information, and its eventual use in foresight activities, has increased too [2-4]. However, organisations keep making incorrect decisions and realising late that they ignored important *warning signals* [5]. Warning signals frequently appear in the form of disconnected data that at first resemble background noise, but which can then be recognised as part of a larger pattern when viewed through a different frame or by connecting it with other information [6]. Recognising warning signals involves knowing where to look for clues, how to interpret them and when to discard or act on faint and ambiguous stirrings. *Horizon scanning*, whose means and outcomes are the focus of our research, has proved to enhance the ability of an organisation to identify warning signals.

Horizon scanning has been defined as *"a systematic examination of information to identify potential threats, risks, emerging issues and opportunities, allowing for better preparedness and the incorporation of mitigation and exploitation into the policy making process"* [7]. The Web has been recommended as a source of information for horizon scanning [8], because it can corroborate information, augment the accuracy of forecasts, and enrich the acquisition of data [9].

An example of the usefulness of the Web in this context is *Google Flu Trends* [10]: by analysing data based on search terms that *Google* [11] has identified as indicators of influenza activity, Google Flu Trends has been able to forecast by up to several weeks where influenza outbreaks are most likely to occur on a geographical basis. This information has been productively used by the *U.S. Centres for Disease Control and Prevention*.

Horizon scanners—i.e., dedicated analysts responsible for *scanning the horizon*—make frequent use of search engines to retrieve information [12, 13]. Even though other Web resources, like blogs, RSS feeds and Twitter streams [14] are also employed, search engines play a significant part in the reports that horizon scanners communicate to decision makers. Regrettably, the results yielded by popular search engines are often inconsistent, obsolete and redundant. Thus, a number of post processing heuristics have to be employed to select the most relevant pieces of information from the extensive array of material available online. In this paper, we concentrate on the first steps of this process, which entail the retrieval of information via search engines and the use of search engine result counts to estimate the amount of information available.

The remainder of this paper is organised as follows: Section II reviews related work on the use of search engine result counts. Section III explains which search engine interfaces we evaluated in the context of horizon scanning applications. Section IV details the conditions of the experiment that we undertook to study the instability of the search engine interfaces. Section V reports on the evaluation of our experiments, and, finally, Section VI states our conclusions and highlights opportunities for future work.

II. RELATED WORK

Search engine result counts have been analysed largely in the natural language processing domain—especially with regard to the use of result counts as proxies for frequency estimates. Such studies have served in machine translation, spelling correction and adjective ordering. Keller and Lapata [15] and Kilgarriff and Grefenstette [16] are some examples. This paper, however, does not intend to examine language structures or derive meaning from natural language. It focusses on employing search engine results to discover trends and ultimately help horizon scanners to do their work.

Table I reviews the most representative studies on search engine results' consistency, the specific engines that were investigated in such studies and their key findings. The study performed by Rayson et al. [17] bears a close resemblance to ours; yet, Rayson et al. were interested in using the Web as a corpus for language analysis, while we want to employ the results of search engines to discover trends and scan the horizon. The differences between the results provided by the interfaces of various engines have been studied by Bar-Yossef and Gurevich [18] and McCown and Nelson [19], but mostly for frequently used queries, whereas we focus on queries typically associated with horizon scanning purposes.

TABLE I. RELATED WORK

Study	**Search engines**	**Key findings**
McCown, F. and M.L. Nelson, 2007 [19]	Google, MSN Search and Yahoo!	It can take over a year for half of the top 10 results to a popular query to be replaced in Google and Yahoo; for MSN it may take only 2-3 months.
Bar-Yossef, Z. and M. Gurevich, 2008 [18].	Google, MSN Search and Yahoo!	At the time (2006), Yahoo! had the largest index, Google the second largest index and MSN the smallest one.
Rayson *et al.*, 2012 [17].	Bing, Google and Yahoo!	The agendas of commercial search engines and language researchers are widely divergent—search engines are interested in returning "useful" results fast. Further investigation is required to study the order in which search results are returned, which makes a significant difference to the collection of documents gathered from search engine results.

Horizon scanning has proved useful to identify new and emerging health technologies [20, 21]. To establish how exactly the Web can be used in health technology assessments, Douw et al. [20] circulated a questionnaire among organisations known to use the Web for horizon scanning purposes. The questionnaire concentrated on the type of websites scanned and the frequency of the scanning. Nevertheless, our work is directed towards the automation of horizon scanning. Hence, rather than surveying organisations, we have focussed on methods to carry out an automated, Web-based scan of the horizon.

III. SEARCH ENGINE INTERFACES

For the vast majority of users, Web browsers are the primary mode of interaction with a search engine. We will refer to the user interface that allows a person to interact with a search engine via a Web browser as a *Web user interface*, or simply *WUI*.

Based on our previous work on horizon scanning [12, 13], we can confirm that horizon scanners make frequent use of WUIs. They subscribe to *RSS feeds* and news websites too, and employ *feed readers* and *aggregators* to browse the information that they receive automatically, but when they require a search engine, they normally employ WUIs.

WUIs, however, are not the only way to interact with a search engine: it is also possible to do it via an *application programming interface (API)*. This alternative way of interaction is very convenient when developing software. Formerly, applications that required automated access to a search engine's results had to download the results first—formatted in HTML—and "scrape" the data off to find what they were looking for—this is a technique known as *Web scraping*. Obviously, whenever the engine changed the format of its results, the Web-scraping code became ineffective. Supplying regulated access to a search engine's repository via an API removes the need for Web-scraping. Unfortunately, the inner workings of the APIs, as well as most of the technology and algorithms that support the operation of a search engine, are considered proprietary and they are not publicly available for scrutiny. Hence, studies like this become necessary.

In 2002, Google became the first search engine to release a freely-available API for accessing its index [19]. *Yahoo!* and *Microsoft* released their APIs in early and late 2005, respectively. To exceed Google's offer, both Yahoo! and Microsoft reduced the usage restrictions and allowed programmers to create applications that did not require license keys for third parties [19]. During the past few years, Google, Yahoo! and Microsoft have expanded their APIs to include features such as maps, news and multimedia material. Google's API has evolved from its original implementation to the current *Google's Custom Search API* [22]. Similar improvements have been carried out by Yahoo! to create its *BOSS API* [23] and by Microsoft to produce *Bing's Search API* [24].

IV. EXPERIMENTS

To determine the extent of the instability of Google's interfaces, we released a specific set of queries daily for 100 days, starting on 3 August 2012. We released the queries at approximately the same time of the day for the entire length of the experiment—between 9:00 GMT and 10:00 GMT. The specific set of queries is listed in Table II. Note that the queries were meant to discover new and emerging issues in *eHealth* and *telemedicine*. The queries combine a series of keywords originally suggested by the *Defence Science and Technology Laboratory (Dstl)* as descriptors to look for new developments [25]—e.g., `revolutionary`; `ground-breaking`; `paves the way`; etcetera.

In order to quantify the instability of the result counts provided by Google's WUI and API, we employed the *shift width* measure [15], according to which the calculated shift width between two result counts obtained from a particular search engine interface for the same query in consecutive days is defined as follows:

$$s_i = \frac{r_i - r_{i-1}}{r_{i-1}},$$

where i is the number of days since the start of the experiment, r_i is the result count recorded on the i-th day of the experiment, and r_{i-1} is the result count recorded on the previous day.

We averaged the values of the calculated shift widths for the 26 queries in Table II for each day of the experiment. The number obtained provided the average shift width for the whole set of queries for a particular day—note that the shift width may be a negative value, when the result count decreases between one day and the previous one. Then, we repeated the same procedure with the result counts yielded by Google's API. The results are discussed below.

TABLE II. QUERY SET

`ehealth; breakthrough`
`telemed; breakthrough`
`ehealth; "closer to reality"`
`telemed; "closer to reality"`
`ehealth; "first time"`
`telemed; "first time"`
`ehealth; groundbreaking`
`telemed; groundbreaking`
`ehealth; "new development"`
`telemed; "new development"`
`ehealth; "new threat"`
`telemed; "new threat"`
`ehealth; novel`
`telemed; novel`
`ehealth; "paves the way"`
`telemed; "paves the way"`
`ehealth; "previously impossible"`
`telemed; "previously impossible"`
`ehealth; "previously unknown"`
`telemed; "previously unknown"`
`ehealth; revolutionary`
`telemed; revolutionary`
`ehealth; unprecedented`
`telemed; unprecedented`
`ehealth; "world's first"`
`telemed; "world's first"`

V. RESULTS

A query that showed a rather irregular behaviour when using Google's WUI was `ehealth; breakthrough`. Figure 1 plots the time series for the result counts for this query. As it can be seen, the result counts displayed large instability, shifting up and down continuously from a maximum of 429,000 results on 7 August 2012 to a minimum of 10,800 on 17 September 2012. Over the length of the experiment, the result counts tended to decrease, but not smoothly. The work of a horizon scanner cannot be based, uncritically, on these results.

Figure 1 also plots the time series for the result counts produced by Google's API for the same query—`ehealth; breakthrough`. The counts provided by Google's API also varied daily, but fluctuating over a much smaller interval: a maximum of 33,500—on 07 August 2012—and a minimum of 18,600—on 22 August 2012. No significant drops were recorded from one day to another.

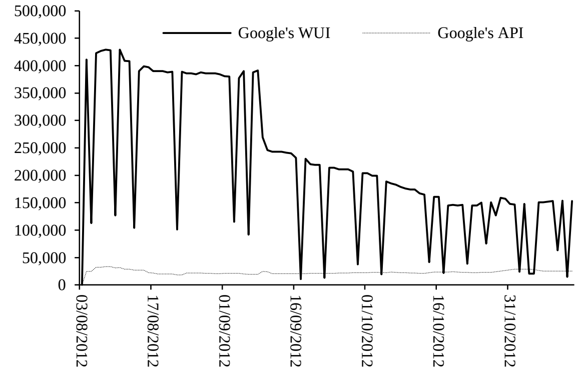

Figure 1. Time series for the result counts yielded by Google's WUI for the query `eHealth; breakthrough`

As explained in Section IV, we computed the average shift width for the results yielded by Google's WUI for the 26 queries involved in the experiment for each particular day. These values are plotted in Figure 2. The instability exhibited by Google's WUI did not occur, exclusively, over a specific period of time, which could have indicated temporary difficulties in Google's operation. The permanent instability emphasises the unpredictable nature of Google's WUI. The highest variation in shift widths was recorded on 17 September 2012, when the shift width went up by 7.41—i.e., a 74.1% increase in the average result counts—followed two days later (19 September 2012) by a drop by -0.02—i.e., a 2% decrease in the average result counts.

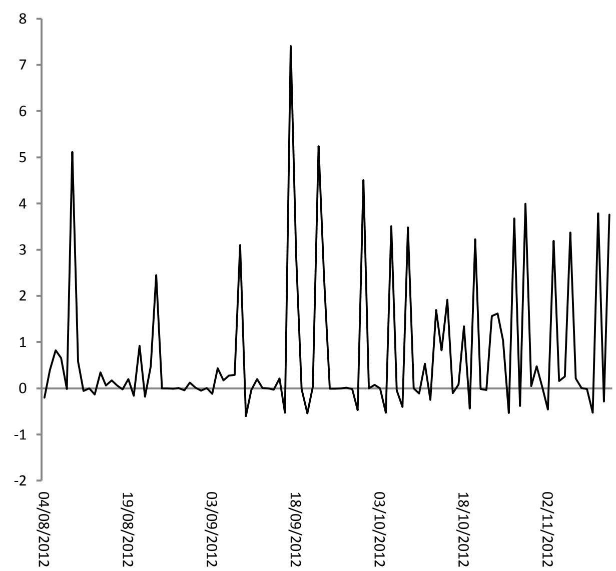

Figure 2. Time series for the average shift width yielded by Google's WUI

Note the continuous appearance of negative shift widths shown in Figure 2. Negative shift widths indicate a decrease in the result counts, which may signify a reduction in the number of documents indexed by Google, perhaps as a consequence of documents disappearing from the Web. Nonetheless, the continuous fluctuation of the average shift width represents a significant limitation for horizon scanners who utilise Google's WUI.

There may be a correlation between the different periods of instability, especially in the last month of the experiment, when the shift width drops and rises intermittently. In any case, the high instability demonstrates that, regardless of the moment when we issue a query, we may not obtain a "correct" estimate of the number of results actually available online. The longest periods of stability took place between 25-28 August 2012, and then between 24-28 September 2012, when the average result counts remained almost the same, causing the average shift width to assume a value close to zero over those two periods. Yet, the same pattern was not repeated in the next two months or at any other time.

To quantify the instability of the result counts provided by Google's API, we repeated the same analysis explained above with the result counts supplied by Google's API. Figure 3 shows the time series for this analysis. Note that the result counts derived from Google's API are far more stable than those yielded by Google's WUI. For the most part of the experiment, the average shift width calculated for the result counts yielded by Google's API remained in the range between ±0.05—i.e., between a 5% increase and a 5% decrease. There are only three peaks and one valley that escape the ±0.05 interval in Figure 3: 13 August 2012 (8% increase), 9 September (14% increase), 11 September 2012 (6% decrease) and 31 October 2012 (20% increase, and the major peak of instability recorded in the experiment).

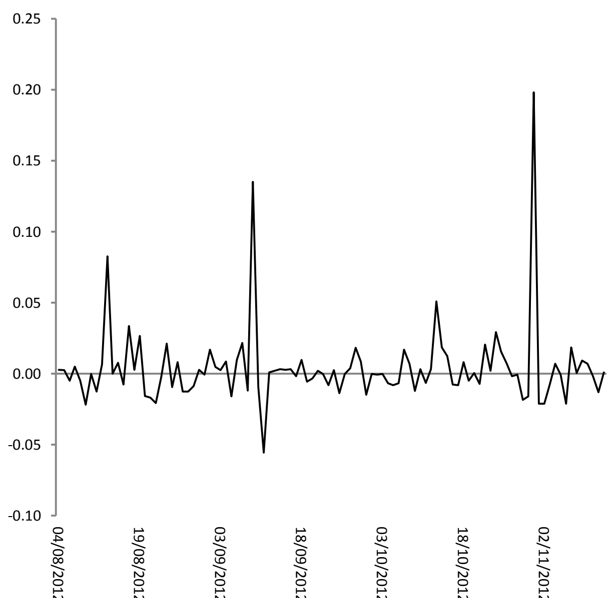

Figure 3. Time series for the average shift width produced by Google's API

Google's API appears to be a far better choice for futures research than Google's WUI—simply because its result counts are more consistent over time. We hypothesise that the estimated search count on the WUIs is high because it has become a marketing tool [17]—users may think that greater means better. Since Web browsers are the primary mode of interaction with search engines, the results yielded by the WUIs are likely to impact strongly on the users' opinions of the comprehensiveness of engines and hence their use.

Another reason why search engines may not show stable figures for the number of results is speed. WUIs aim to provide results in the shortest possible time. Thus, retrieving the number of results for each particular query must involve an algorithm to "approximate" the actual value, rather than precisely computing it.

Initially, we were only interested in measuring the instability of Google's result counts on a day by day basis; yet, we noticed that even on the same day Google provides different result counts for the same queries, when they are submitted at different times. To prove this, we submited our queries to Google's WUI on 5 December 2012—starting at 9:45 AM—and then submitting them again on the same day, but 8 hours later—i.e., starting at 5:45 PM. The queries were also submitted to Google's API on the same day with a difference of 8 hours between submissions. Table III shows the results for this experiment. Some of the result counts provided by the WUI remained constant after the 8 hour period, but more than a third of the queries employed yielded different result counts. Moreover, the changes seem completely irregular: for example, the result count for the query `ehealth; unprecedented` increased from 12,100 to 54,200 after 8 hours; whereas the result count for `ehealth; paves the way` decreased from 9,350 to 8,830. Note that the WUI's result counts that changed during the experiment are highlighted in bold font in Table 5.

VI. CONCLUSIONS

By taking daily samples of the result counts provided by two Google's interfaces, we have explored the instability of a particular search engine, and the differences between Google's WUI—typically employed by human users browsing the Web—and Google's API—normally utilised by computer applications retrieving results automatically. Forthcoming research could focus on collecting statistics associated with other engines, in the same fashion in which we have done it with Google, as this would help horizon scanners to distinguish the best choice for their purposes.

When we began our investigation, we assumed that the result counts supplied by Google's interfaces would fluctuate over time, but we expected to see the fluctuation hovering around some central values—which would increase gradually, as Google discovers new content. The outcome of our analysis, however, shows, unmistakeably, that result counts vary at an extremely fast pace—sometimes in only a matter of hours—, and not necessarily in a growing fashion.

Indications are that horizon scanning needs to be supported by continuous searching to obtain a sensible snapshot of the information on the Web. This is consistent with the results of our previous study [13], which was carried out in collaboration with *Lloyd's of London*. Compared to Lloyd's current practice, we improved the number of relevant documents about emerging risks retrieved from the Web, by performing daily searches using Google's API.

On the basis of our research, result counts provided by WUIs seem to be overestimated. In our opinion this may be due to their use for marketing purposes; search engines are economically motivated tools, and the higher the number of the results reported, the larger the market they may approach.

TABLE III. RESULT COUNTS REPORTED BY GOOGLE'S WUI AND API ON 5/12/2012 AT TWO DIFFERENT TIMES

Query	Google's WUI 05/12/2012 09:45	Google's WUI 05/12/2012 17:45	Google's API 05/12/2012 09:40	Google's API 05/12/2012 17:40
`ehealth; breakthrough`	167,000	167,000	21,500	21,400
`telemed; breakthrough`	**377**	**197**	669	665
`ehealth; "closer to reality"`	**3,340**	**3,470**	470	469
`telemed; "closer to reality"`	12,300	12,300	457	457
`ehealth; "first time"`	222,000	222,000	52,300	52,600
`telemed; "first time"`	1,780	1,780	1,750	1,760
`ehealth; groundbreaking`	13,800	13,800	6,070	6,120
`telemed; groundbreaking`	**108**	**107**	283	280
`ehealth; "new development"`	**3,120**	**3,140**	1,980	1,990
`telemed; "new development"`	53	53	79	79
`ehealth; "new threat"`	123	123	244	243
`telemed; "new threat"`	424	424	181	176
`ehealth; novel`	227,000	227,000	40,700	40,800
`telemed; novel`	663	663	3,540	3,540
`ehealth; "paves the way"`	**9,350**	**8,830**	1,460	1,460
`telemed; "paves the way"`	22,800	22,800	1,140	1,170
`ehealth; "previously impossible"`	82	82	103	103
`telemed; "previously impossible"`	**61**	**603**	501	502
`ehealth; "previously unknown"`	4,130	4,130	1,020	1,020
`telemed; "previously unknown"`	1,750	1,750	902	907
`ehealth; revolutionary`	21,200	21,200	34,200	34,400
`telemed; revolutionary`	**364**	**594**	1,740	1,750
`ehealth; unprecedented`	**12,100**	**54,200**	9,650	9,650
`telemed; unprecedented`	**589**	**585**	621	623
`ehealth; "world's first"`	14,700	14,700	4,370	8,350
`telemed; "world's first"`	65,900	65,900	5,540	5,540

REFERENCES

1. Chan, S.W.K. and J. Franklin, A text-based decision support system for financial sequence prediction. Decision Support Systems, 2011. 52(1): p. 189-198.
2. Linstone, H.A. and M. Turoff, Delphi: A brief look backward and forward. Technological Forecasting and Social Change, 2011. 78(9): p. 1712-1719.
3. Gheorghiu, R.C., Adrian; Paunica, Mihai; Holeab, Cosmin, Web 2.0 and the emergence of future oriented communities. Economic Computation & Economic Cybernetics Studies & Research, 2009. 43(2): p. p1.
4. Gnatzy, T., et al., Validating an innovative real-time Delphi approach - A methodological comparison between real-time and conventional Delphi studies. Technological Forecasting and Social Change, 2011.
5. Wissema, H., Driving through red lights: How warning signals are missed or ignored. Long Range Planning, 2002. 35(5): p. 521-539.
6. Ansoff, I., Managing Strategic surprise by response to weak signals. California Management Review, 1975. 18(2): p. 21-33.
7. Chairman of the Joint Intelligence Committee, Review of cross-government horizon scanning, 2012, Cabinet Office: London, UK.
8. Stevens, A.J., et al., Keeping pace with new technologies: Systems needed to identify and evaluate them. BMJ, 1999. 319.
9. Robert, G., A.J. Stevens, and J. Gabbay, 'Early warning systems' for identifying new healthcare technologies. Health Technology Assessment, 1999. 3(13): p. 1-108.
10. Google. Google Flu Trends. 2013; Available from: http://www.google.org/flutrends/.
11. Google. About Google. 2012; Available from: http://www.google.com/intl/en/about/.
12. Palomino, M.A., et al., Web-based horizon scanning: Concepts and practice. Foresight, 2012. 14(5): p. 355-373.
13. Palomino, M.A., A. Vincenti, and R. Owen, Optimising Web-based information retrieval methods for horizon scanning. foresight, 2013. 15(3).
14. Truvé, S., Big Data for the Future - Unlocking the Predictive Power of the Web, 2011, Recorded Future: Cambridge, MA.
15. Keller, F. and M. Lapata, Using the Web to obtain frequencies for unseen bigrams. Comput. Linguist., 2003. 29(3): p. 459-484.
16. Kilgarriff, A. and G. Grefenstette, Introduction to the special issue on the Web as corpus. Comput. Linguist., 2003. 29(3): p. 333-347.
17. Rayson, P., O. Charles, and I. Auty. Can Google count? Estimating search engine result consistency. in Proceedings of the ACL SIG on Web as Corpus (WAC7). 2012. Lyon, France.
18. Bar-Yossef, Z. and M. Gurevich, Random sampling from a search engine's index. Journal of the ACM (JACM), 2008. 55(5): p. 1-74.
19. McCown, F. and M.L. Nelson, Agreeing to disagree: Search engines and their public interfaces, in Proceedings of the 7th ACM/IEEE-CS Conference on Digital Libraries, 2007, ACM: Vancouver, Canada. p. 309-318.
20. Douw, K., et al., Use of the Internet in scanning the horizon for new and emerging health technologies: A survey of agencies involved in horizon scanning. Journal of Medical Internet Research, 2003. 5(1): p. e6.
21. Palomino, M.A., et al., Web-based Horizon Scanning: Recent Developments with Application to Health Technology Assessment. Business Informatics, 2012. 3(25): p. 139-159.
22. Google. Custom Search. 2012; Available from: https://developers.google.com/custom-search/.
23. Yahoo! Yahoo! Search BOSS API. 2013; Available from: http://developer.yahoo.com/search/boss/.
24. Microsoft. Bing API, Version 2. 2013; Available from: http://msdn.microsoft.com/en-us/library/dd251056.aspx.
25. Wilson, J.C. and D.J. Holland-Smith, White Paper: Dstl S&T horizon scanning, 2008, Defence Science and Technology Laboratory (Dstl).