


The Journal[Cybermetrics News](#)[Editorial Board](#)[Guide for Authors](#)[Issues Contents](#)**The Seminars****The Source**[Scientometrics](#)[Tools](#)[R&D Policy & Resources](#)[World Situation Report](#)**VOLUME 2/3 (1998-9): ISSUE 1. PAPER 2**
**Daily time series of common single word
searches in AltaVista and NorthernLight**

Ronald RousseauKHBO, Faculty of Industrial Sciences and Technology
Zeedijk 101, 8400 Oostende, BelgiumRonald.Rousseau@kh.khbo.be**Abstract**

Over the past year the AltaVista search engine provided irregular numbers of hits. This is shown by collecting a time series of daily searches. The new AltaVista (on the Web since October 25, 1999) is much more stable, but not all problems have been eliminated. The same searches performed on NorthernLight gave stable results. It is suggested to use filtering techniques, such as median filtering, when dealing with irregular time series. An attempt is made to estimate the growth of the Web (as covered by a search engine), based on neutral single word searches. It turned out that this method systematically underestimates the growth of the Web. Collecting time series should be an essential part of Internet research.

Keywords

time series, median filtering, AltaVista, NorthernLight, estimating growth, nova-effect

1. Introduction

Nowadays the Internet is still growing exponentially (**Internet Software Consortium, 1999**). So, although it is known that search engines cover only parts (at most 16% according to Lawrence and Giles (**1998, 1999**)) of the whole Internet (**Brake, 1997**), one at least expects that the same search performed through the same search engine at consecutive moments in time would result in an increasing number of hits. Experience has shown, however, that this is not true (**Rousseau, 1999**). AltaVista in particular seems to yield irregular results. Hence, by taking daily samples in AltaVista (www.altavista.com) and in NorthernLight (www.northernlight.com) - as a control - we tried to determine the extent of these irregularities. NorthernLight was used for this as it was reported to have the broadest coverage at that time (**Lawrence & Giles, 1999**; see also Notess (**1999**)).

2. Method

In order to have an idea of the extent of the phenomenon we performed daily searches, over a 12 week period, for three common words (saxophone*,

trumpet*, pope) in AltaVista, Advanced Search with 'count documents matching the Boolean expression' on, and in NorthernLight, Simple Search. Note that the asterisk is the symbol for truncation. Searches were done at about the same time of the day. We were only interested in the number of hits, not in the exact pages that were found. A difference in retrieved pages is another aspect of the dynamics of the Internet as manifested by search engines ([Bar-Ilan, 1999b](#); [Bar-Ilan and Peritz, 1999](#)), but that was not the aim of this investigation. Because our first aim was to have an idea of the irregularities in the number of hits as reported by AltaVista we used single word queries (in order not to introduce additional complexity). These three simple queries were, moreover, 'neutral' words, i.e. words for which one does not expect a sudden increase or decrease in numbers. If, e.g., the pope had died in the period of investigation this would have been different.

Between the day of submission of the original manuscript and receiving the referees' reports AltaVista changed the contents and look of its database. It received one major update on October 25 (day 298 of the year), but from then on, it stopped the daily updates it had before the event. Moreover, the link 'count documents matching the Boolean expression' was eliminated. For this reason we will present [data](#) collected over a period of 21 weeks.

3. Results

Figures 1 and 2 show graphs of the time series of the number of hits over the period July 27, 1999 (day 208 of the year) – December 20, 1999 (day 354 of the year), or 21 weeks. It is clear that the NorthernLight results are precisely as expected. They show a general increase over the whole period of investigation. The AltaVista results on the other hand are highly irregular. In fact, there is only one conclusion possible: the number of hits as reported by (the old) AltaVista was often wrong. Dips of 20% lasting just one day clearly indicate wrong counting results. We also point out that these dips do not occur simultaneously for all searches, suggesting that they cannot be explained by heavy traffic on the Internet. This confirms earlier reports of a bug in AltaVista's counting algorithm ([Bar-Ilan, 1999a](#)).

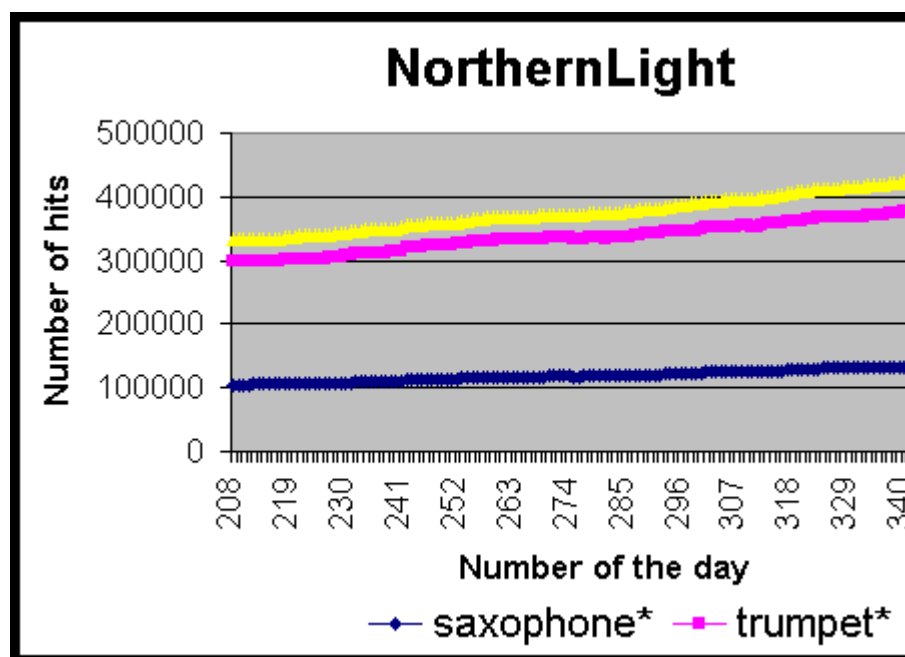


Fig. 1 The NorthernLight time series

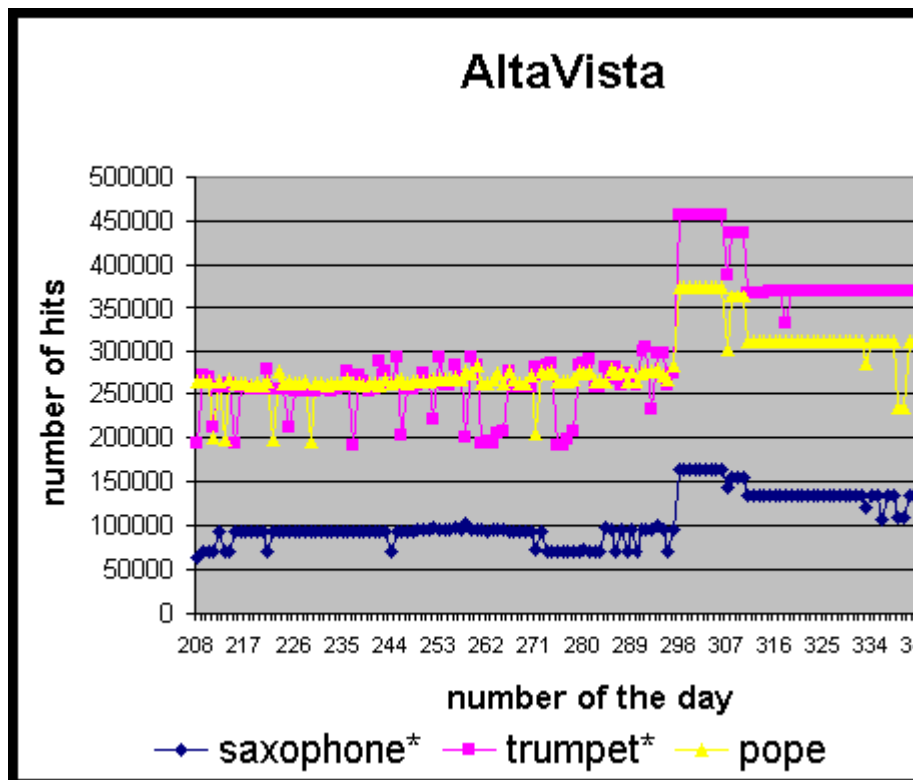


Fig. 2 The AltaVista time series

4. A discussion of the NorthernLight results

Contrary to the case of AltaVista, the NorthernLight hit results for the three queries changed on the same day (with very few exceptions, see data). This indicates that on that day the database itself changed. (We use the term 'database' in the sense of 'the Internet as covered by this search engine'.) When there was a change in the number of hits, this number usually, but not always, increased. We tried to see if those changes occurred on the same days of the week (see Table 1) but the hypothesis of a uniform distribution (meaning that every day is as likely as any other to show a change) cannot be rejected (the observed χ^2 value is 6.2). Yet, most changes occurred on a Monday and the least ones on a Sunday. Note that a change that we observe on a Sunday can easily have occurred on a Saturday (because of the time difference between Europe and America).

Table 1: Days and number of times the number of hits changed in NorthernLight (for the query saxophone*).

Day	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Number of times changed	16	8	11	8	10	11	6

Table 2 shows the distribution in the number of days the database stayed unchanged.

Table 2 Number of days the database stayed unchanged (for the query saxophone*).

Length of period without change	1	2	3	4	5	6	7	8

Number of days	25	27	10	3	2	1	0	1
----------------	----	----	----	---	---	---	---	---

The mean number of days without change was 2.10, indicating a regular update of the NorthernLight database.

5. A discussion of the AltaVista results

The raw (old) AltaVista data are clearly useless. There is too much noise and there are too many irregularities. Techniques to reduce this noise must first be applied. We eliminated occasional dips by using a 5-point median filter. This means that the number of hits assigned to one particular day is the median of the period beginning two days earlier and ending two days later. Median filters are often used in signal processing. They are useful for eliminating noise in one-dimensional as well as in two-dimensional signals ([Castleman, 1996](#); [Bylsma, 1999](#)). In pictures they are often applied to remove speckles (Lim, 1990). Figure 3 gives the AltaVista time series after 5-point median filtering has been applied. There still are ups and downs, especially for the term 'trumpet*', but most of the irregularities have been removed. The technique is (almost) completely successful in eliminating irregularities in the search results of the new AltaVista.

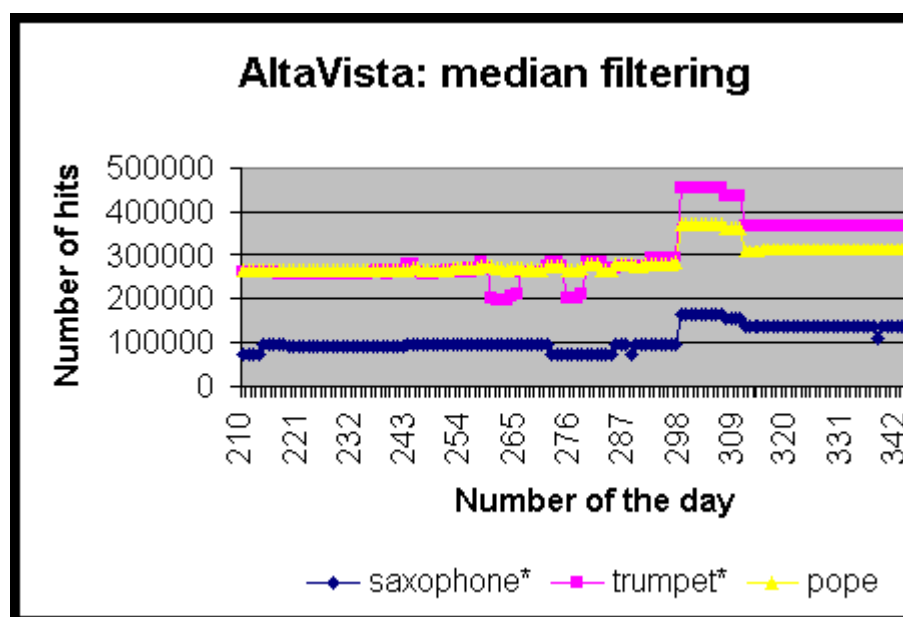


Fig. 3 The AltaVista signals after application of median filtering

6. The nova-effect

As I knew the exact date of the launching of New AltaVista just one day in advance, I felt somewhat like an astronomer observing a nova. In one day (day 298 of the year) the number of hits, as counted by AltaVista, increased by about 35 to 70% (depending on the query), see Fig. 4. Note that the number of hits decreased considerably after 13 days. We can only guess at the reason for this decline, but removal of dead links is probably a good suggestion.

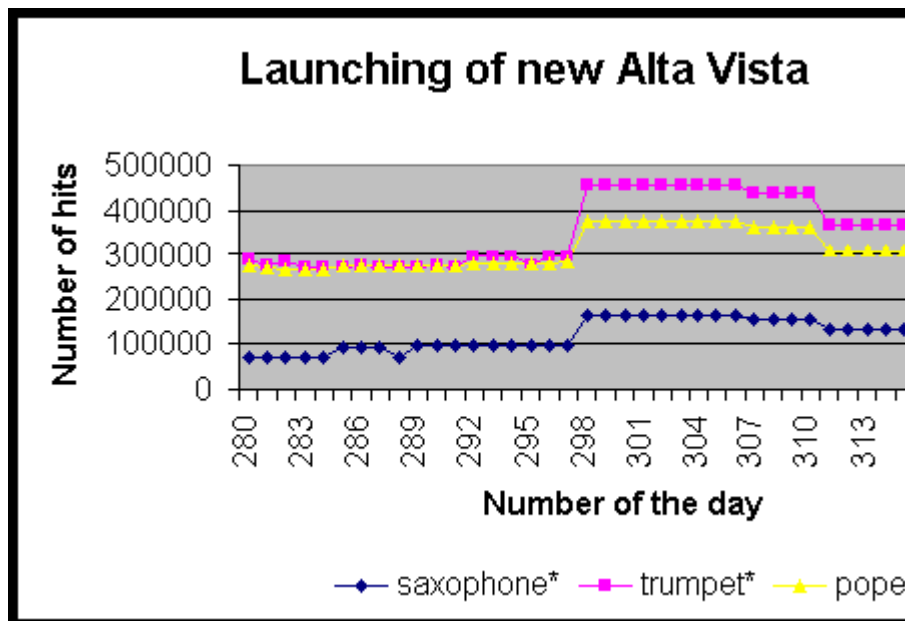


Fig. 4. Launching of new AltaVista: a nova effect

7. Can we predict the growth of the NorthernLight database based on simple single word queries?

In order to answer this question we fitted a linear and an exponential function to the number of web pages in NorthernLight as provided by Greg Notess (1999), Table 3.

Table 3. Notess' data on NorthernLight's web coverage during the year 1999

Date	Number of the day in the year (d)	Number of web pages covered by NorthernLight (W)
5 January 1999	5	115 455 526
5 March 1999	64	128 540 264
5 May 1999	125	140 609 561
4 August 1999	216	153 586 380
9 September 1999	252	169 222 122
29 November 1999	333	200 352 984

Best fitting curves are:

$$W = (114347.8 \times e^{0.0015886 d}) \times 1000, \text{ with } R = 0.99 \quad (1)$$

for the exponential case, and

$$W = (111082 + 242.485 d) \times 1000, \text{ with } R = 0.979 \quad (2)$$

for the linear case. Note that the fitting for the exponential case is slightly better, confirming the exponential growth of the Internet. Fig.5 shows the data of Table 3 and a best-fitting exponential curve.

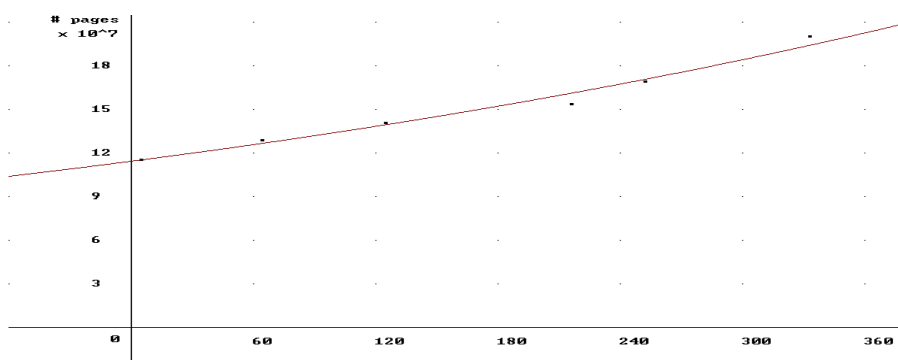


Fig. 5. Growth of the Internet as covered by NorthernLight and a best-fitting exponential curve

Next, we did the same for a number of data provided by NorthernLight (collected by us), see Table 4. Note that Notess (1999) has checked that the numbers provided by NorthernLight are correct.

Table 4. Coverage of the Web by NorthernLight (period October-December 1999)

Date	Number of the day in the year (d)	Number of web pages covered by NorthernLight (W)
22 October 1999	295	181 489 807
2 November 1999	306	185 000 698
15 November 1999	319	192 208 478
22 November 1999	326	197 970 728
7 December 1999	341	202 363 520
19 December 1999	353	205 506 440

Best fitting curves of the data of Table 4 are:

$$W = (93171.8 \times e^{0.00226644 d}) \times 1000, \text{ with } R = 0.985^{(3)}$$

for the exponential case and

$$W = (52309.5 + 438.496 d) \times 1000, \text{ with } R =$$

$$0.986^{(4)}$$

for the linear case. Here the linear and exponential function fit equally well due to the slight S-shape of the data.

Finally, we fitted the exponential function to the (Tuesday) query data for the observed period (21 data points). This gave the following equations, with Y denoting the number of hits and d again the number of the day in the year:

$$Y = 71718.2 \times e^{0.00179635 d} \quad \text{for 'saxophone*'} \quad (5)$$

$$Y = 209567.3 \times e^{0.00171656 d} \quad \text{for 'trumpet*'} \quad (6)$$

$$Y = 227840 \times e^{0.00181044 d} \quad \text{for 'pope'} \quad (7)$$

The most interesting observation we can make here is that the growth rate as shown by the exponent of the exponential function is the largest for the whole NorthernLight database as measured over the 21-week period. Then we have the three rates of the three queries and finally, we have the rate for the whole year. This means that 'predictions' for the last known date (December 19, 1999) based on the smaller growth rates all underestimate the real number of web pages, see Table 5. Predictions based on single word queries moreover use the ratio of the fitted curves on the 208th day of the year (Tuesday, 27 July 1999), assuming the same ratio on December 19. This means that the web, as covered by NorthernLight, increases faster than a fixed exponential (the growth rate increases).

Table 5. Predictions of the number of NorthernLight web pages on December 19, 1999 based on seven fitted equations

Exact number of Web pages	205506 440 \approx 205.5 10^6
Prediction by equation (1)	207.4 10^6
Prediction by equation (2)	207.1 10^6
Prediction by equation (3)	200.3 10^6
Prediction by equation (4)	196.7 10^6
Prediction by equation (5)	193.7 10^6
Prediction by equation (6)	191.5 10^6
Prediction by equation (7)	194.1 10^6

8. Recommendations for further research

It would be very interesting to collect time series using more and different search engines. This would help an Internet researcher (cybermetrician) to find the best search engine for his/her purpose. If one is interested in quotients of the number of hits, as when calculating web impact factors, (**Ingwersen, 1998**) results obtained through (the old) AltaVista can lead to highly doubtful results (both the denominator and the numerator can be wrong by as much as 20%). So we would recommend not using AltaVista for informetric research on the Web, unless one needs a unique feature of this particular search engine. In that case filtering, e.g. by a median filter, is certainly necessary. Finally, we would like to draw the attention of the research community to the need of collecting more time series to study different aspects of the Internet.

Acknowledgements. I like to thank the anonymous referees for making me think somewhat deeper about the obtained results. I also thank the editor Isidro Aguillo, for a job well done.

References

Bar-Ilan, J. (1999a). Personal communication (July 1999).

Bar-Ilan, J. (1999b). *Search engine results over time – A case study on search engine stability*. **Cybermetrics**, 2/3(1):1.
<www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html>

Bar-Ilan, J. and Peritz, B.C. (1999). *The availability and life span of a specific topic on the Web; the case of "informetrics": a quantitative and content analysis*. In: **Proceedings of the seventh conference of the international society for scientometrics and informetrics (C.A. Macias-Chapula, ed.)**. **Universidad de Colima**, 11-19.

Brake, D. (1997). *Lost in cyberspace*. **New Scientist**, 154 (2088), 12-13.

Bylsma, Wesley (1999). *Median filtering*. **Dr. Dobb's Journal**, 24(10), 119-121.

Castleman, K. (1996). *Digital image processing*. Englewood Cliffs (NJ): Prentice Hall.

Ingwersen, P. (1998). *The calculation of web-impact factors*. **Journal of Documentation**, 54, 236-243.

Internet Software Consortium (1999). **Internet Software Consortium – Domain Survey** – Internet Domain Survey. Available at:
<<http://www.isc.org/ds>>

Lawrence, S. and Giles, C.L. (1998). *Searching the World Wide Web*. **Science** 280 (5360), 98-100.

Lawrence, S. and Giles, C.L. (1999). *Accessibility of information on the web*. **Nature**, 400 (6740), 107-109.

Lim, J.S. (1990). *Two-dimensional signal and image processing*. Englewood Cliffs (NJ): Prentice Hall.

Notess, Greg R. (1999). *Search engine statistics: database total size estimates*. Available at:
<<http://www.notess.com/search/stats/sizeest.shtml>>. Visited December 19,

1999.

Rousseau, R. (1999). *Time evolution of the number of hits in keyword searches on the Internet*. Talk presented at the Colima Cybermetrics'99 conference. **Summary** available at
<www.cindoc.csic.es/cybermetrics/cybermetrics99.htm>

Received 22/October/1999
Accepted 21/December/1999



[Copyright information](#) | [Editor](#) | [Webmaster](#) | Updated: 11/25/2003

[TOP](#)