

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

INTERNET SEARCH ENGINES – FLUCTUATIONS IN DOCUMENT ACCESSIBILITY

WOUTER METTROP
wouter.mettrop@cw.nl

*Centrum voor Wiskunde en Informatica (CWI), Kruislaan 413
1098 SJ Amsterdam, The Netherlands*

PAUL NIEUWENHUYSEN
paul.nieuwenhuysen@vub.ac.be

*Vrije Universiteit Brussel, Pleinlaan 2, B-1050, Brussel
Universitaire Instelling Antwerpen, Belgium*

An empirical investigation of the consistency of retrieval through Internet search engines is reported. Thirteen engines are evaluated: AltaVista, EuroFerret, Excite, HotBot, InfoSeek, Lycos, MSN, NorthernLight, Snap, WebCrawler and three national Dutch engines: Ilse, Search.nl and Vindex. The focus is on a characteristic related to size: the degree of consistency to which an engine retrieves documents. Does an engine always present the same relevant documents that are, or were, available in its database? We observed and identified three types of fluctuations in the result sets of several kinds of searches, many of them significant. These should be taken into account by users who apply an Internet search engine, for instance to retrieve as many relevant documents as possible, or to retrieve a document that was already found in a previous search, or to perform scientometric/bibliometric measurements. The fluctuations should also be considered as a complication of other research on the behaviour and performance of Internet search engines. In conclusion: in view of the increasing importance of the Internet as a publication/communication medium, the fluctuations in the result sets of Internet search engines can no longer be neglected.

1. INTRODUCTION

1.1 The WWW and search engines

The amount of information that is potentially available straight from the Internet keeps on growing. Estimates lead to the conclusion that in the year 2000, about one billion, that is 1,000 million, unique URLs or 'pages' were accessible in the total Internet (see for instance <http://searchenginesshowdown.com/> (Lawrence & Giles, 1999) and Inktomi Corporation & the NEC Research Institute (2000)).

Journal of Documentation, vol. 57, no. 5, September 2001, pp. 623–651

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

Those documents correspond to about ten terabytes (about 10,000 gigabytes) of text data. Clearly, some good retrieval systems are required to enhance the value of this unordered collection of information resources. Many systems are already available. Simplifying reality somewhat, we can distinguish several types:

- directories of selected sources categorised by subject, made by humans, mainly for browsing;
- search engines, based on databases with machine made indexes, mainly for word-based searching; and
- 'meta-search' or 'multi-threaded' search systems that rely mainly on the databases made available by the previous systems, for word-based searching.

During the last few years, an investigation of the performance of several well-known international (and a few smaller local) word-based Internet search engines has been undertaken by a group of information professionals from various institutes and companies in the Netherlands (and Belgium), as named below in the acknowledgements. This paper reports on a part of this work.

Many aspects/criteria can be considered in the evaluation of an Internet search engine, including:

- the coverage of documents present on the WWW (and most of the well-known investigations have already focused on this aspect, including Lawrence & Giles (1999));
- the elements of a document that are indexed to make them usable for retrieval; and
- the absence of dead links in the set of links suggested by the search system.

We started by studying the depth of indexing of some search engines and in this way we were confronted with fluctuations that exist in the performance of most systems. This phenomenon is relatively unknown, obscure and not well investigated, but it may hinder and influence serious quantitative investigations of other aspects. Therefore, we have made what we think is the first quantitative study in this area. This work has pointed out that most engines suffer from this 'incorrect variable behaviour', in the sense that unexpected and annoying fluctuations exist in the result sets of documents. Thus documents are not retrieved reliably.

1.2 Fluctuations in the performance of search engines

The result set of documents, shown by an Internet search engine as response to a query, changes over time. Broadly speaking, an alteration in this set is correct if it is a reflection of an alteration in the WWW reality, as documents are added to, removed from or changed in the WWW. If this is not the case, the change can be seen as incorrect. Such incorrect changes are not only to do with the incorrect removal of documents from the set of indexed documents, or incorrect (late) additions to this set. They can also occur when an engine has indexed a document, but does not always succeed in retrieving it afterwards.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

September 2001

SEARCH ENGINES

It is possible to demonstrate the existence of fluctuations by comparing the result set of an observation (i.e. one query or a set of queries) with an expected result set. What is expected depends on other information, obtained from other observations and from knowledge about the reality of the WWW. For example, when an engine retrieves a certain document on the basis of a certain query, and when it is known that this document still exists, then it is expected that this document will be found again when the query is resubmitted.

A fluctuation can be an increase in the number of documents in the result set, or a decrease. In this investigation we have only considered decreases. We have investigated the disappearance of documents from result sets, not the addition of documents to result sets, because adding documents to a search engine's database is a normal process for an Internet search system. So: a fluctuation appears, when the result set of an observation misses documents with respect to a certain expected result set. Here, one 'observation' is one query or a set of queries and the expected result set is determined by a certain frame of reference: observations and/or knowledge about web reality.

1.3 The search engines that have been investigated

Thirteen Internet search engines have been investigated: AltaVista, EuroFerret, Excite, HotBot, InfoSeek, Lycos, MSN, NorthernLight, Snap, WebCrawler and three Dutch engines: Ilse, Search.nl and Vindex. The popular Yahoo! has not been investigated, first because it is mainly a directory and second because the search engine, used as an addition to the directory, is external. Some of the investigated engines have several search modes: simple, advanced, super, fuzzy etc. We always used the simplest mode and did not change the mode during the tests.

2. MEASURING DIFFERENT TYPES OF FLUCTUATIONS

2.1 Introduction

Sections 2, 3 and 4 of this article are more theoretical than the conclusion. In Section 2 we define several types of fluctuations on the basis of the possibilities that we have for measuring them. In Section 3 we describe what we have measured. In Section 4 we present the results of our measurements. (Anybody who is interested in our interpretation can skip these sections and continue directly with Section 5.) In these three sections we consider search engines as black boxes that receive queries and send out answers; we will not give an interpretation of what might happen inside the search engines.

As mentioned in Section 1 one can track down fluctuations by comparing the result set of an observation with a frame of reference that is made up of other observations combined with knowledge about changes in the part of the WWW which is observed. If we know that this part of the web remains unchanged, we only have to compare result sets to detect fluctuations.

Our investigation has led us to distinguish three different dichotomies in fluctuations, based on differences in the way they are detected by comparing result sets of observations:

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

1. The compared observations are performed at the same time, or repeated regularly with a constant interval – which leads respectively to instantaneous fluctuations and to fluctuations detected through time.
2. The compared observations consist of just one query, or of a set of several queries – which leads respectively to fluctuations from singular observations and to fluctuations from plural observations.
3. The compared observations are the same (i.e. they consist of the same query or set of queries) or they are different – which leads respectively to fluctuations from the same observation and to fluctuations from different observations.

So we have eight classes of fluctuations, detected as follows:

1. instantaneously from the same singular observation;
2. instantaneously from the same plural observation;
3. instantaneously from different singular observations;
4. instantaneously from different plural observations;
5. through time from the same singular observation;
6. through time from the same plural observation;
7. through time from different singular observations;
8. through time from different plural observations.

We leave aside the question of whether the plural observations in 4 and 8 are disjunct. The classes here defined are not disjunct: fluctuations in one class can cause fluctuations in other classes.

We have investigated fluctuations of types 3, 5 and 6. The other types are more theoretical, although it might be interesting to investigate fluctuations of type 1. Because we have considered only one type of instantaneous fluctuations and only two types of fluctuations detected through time, we are able to name them for short respectively: ‘instantaneous fluctuations’ (type 3); ‘singular fluctuations’ (type 5) and ‘plural observations’ (type 6). We shall describe these more clearly in the following paragraphs.

2.2 *Plural fluctuations*

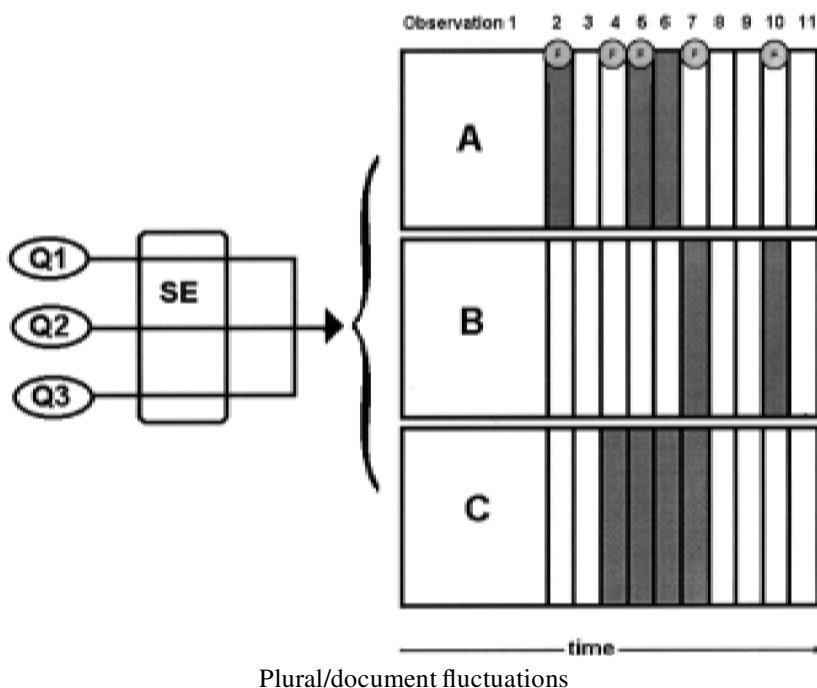
Plural fluctuations show the phenomenon that a set of queries, submitted regularly, stops retrieving one or more documents that still exist in reality. Figure 1 visualises the appearance of plural fluctuations: these occur every time that one of the result sets does not contain a document that was contained in the preceding set and that did not disappear from the web in reality. One observation is a set of queries and the frame of reference (i.e. what is expected) for each observation is made up by the preceding observation combined with knowledge about the WWW.

Later in this paper we interpret this kind of fluctuation as the apparent disappearance, for whatever reason, of documents out of the indexed database of an engine, and call them accordingly ‘document fluctuations’. We also consider some less trivial possible causes for the phenomenon of plural fluctuations.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

September 2001

SEARCH ENGINES



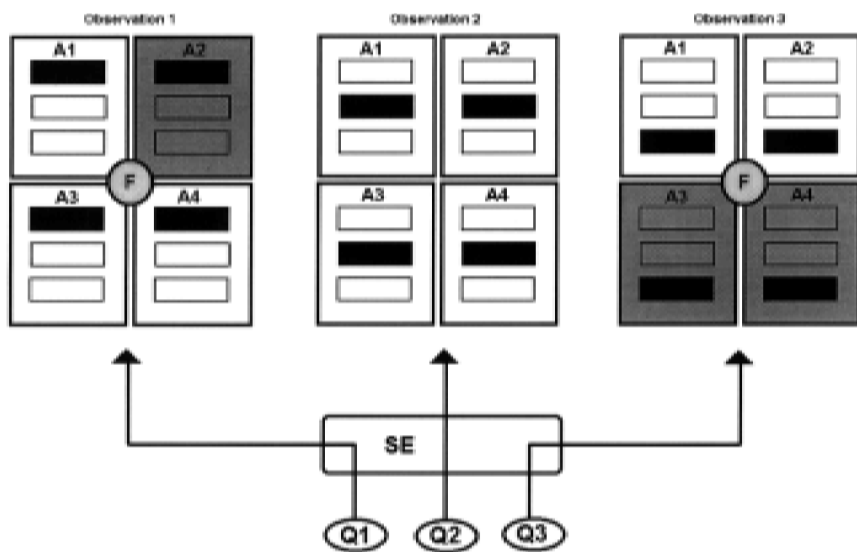
Plural/document fluctuations

Figure 1. An imaginary example of plural fluctuations (that we later call 'document fluctuations'): the queries *Q1*, *Q2* and *Q3* are submitted at the same time to the search engine *SE*; this is one observation. This is repeated 11 times. Each time their result sets are joined. Documents that are retrieved are white, and documents that are not retrieved are grey. Time is on the horizontal axis, from left to right. The three queries start retrieving 3 documents (*A*, *B* and *C*), then only 2 documents – the topmost (*A*) is dark, so only *B* and *C* are retrieved, etc. A plural fluctuation occurs every time that one of the result sets does not contain a document that was contained in the preceding set and that did not disappear from the web in reality. In this example plural fluctuations occur in observations 2, 4, 5, 7 and 10 (at the letters *F*)

2.3 Instantaneous fluctuations

Instantaneous fluctuations show inconsistencies in result sets. They occur when several queries, submitted at the same time, retrieve a set of identical documents and when one of these queries does not retrieve all of these identical documents. This definition is based on the existence of a set of identical documents and on a situation where a number of queries is submitted at the same time. This is an artificial environment used only for demonstrating these fluctuations. The phenomenon occurs in all practical situations. Figure 2 shows an imaginary example of instantaneous fluctuations. One observation is one query in a set of queries, and the frame of reference for each observation is made up of the other observations. Later in this paper we interpret this kind of fluctuation as the unexpected absence of documents in the result sets of queries, and call them 'element fluctuations'.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.



instantaneous/element fluctuations

Figure 2. An imaginary example of instantaneous fluctuations (that we later call 'element fluctuations'). There are 4 identical documents, placed on different sites in the web: A1, A2, A3 and A4. Three queries (Q1, Q2 and Q3) are submitted to the search engine SE at the same time. Documents that are retrieved are white and documents that are not retrieved are grey. Each query is related to one of the document elements, symbolised by the three bars in the documents. (A document element is a set of terms in a document with common properties, depending on the document. See Section 3.) The result set of all queries together contains all 4 identical documents. So now it is expected that each query will retrieve all 4 documents, or none of them. But only Q2 comes up to this expectation; Q1 and Q3 miss documents. Their result sets show instantaneous fluctuations (as indicated by the letters F)

2.4 Singular fluctuations

Next we compare the result sets of one single query that is repeatedly submitted through time. The definition is analogue with the plural fluctuation: a singular fluctuation occurs when an engine stops retrieving a document in response to a query that is submitted repeatedly. One observation is one query and the frame of reference for one observation is made up of the preceding observation combined with knowledge about the WWW. Figure 3 shows an imaginary example of singular fluctuations.

The difference between singular fluctuations and plural fluctuations is not obvious, but reveals itself in the interpretation given later in this paper, where we call singular fluctuations 'indexing fluctuations', because they seem to represent changes in the indexing policy of engines. An additional condition for that is that they are not due to plural or instantaneous fluctuations.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

September 2001

SEARCH ENGINES

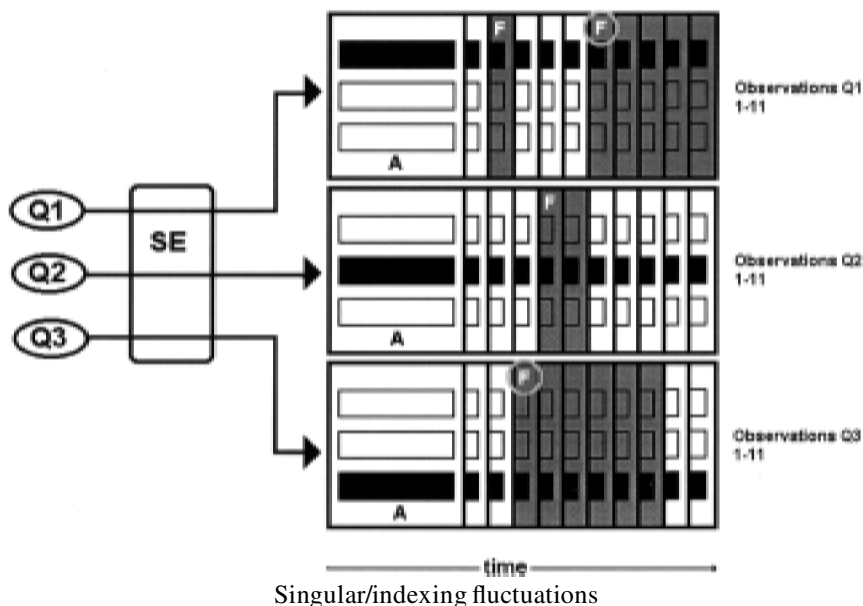


Figure 3. An imaginary example of singular fluctuations. *Q1*, *Q2* and *Q3* are submitted at the same time. This is repeated 11 times. The 11 result sets of *Q1* are compared through time; so are the 11 result sets of *Q2* and the 11 result sets of *Q3*. Documents that are retrieved are white and documents that are not retrieved are grey. They all start with retrieving document A, each because of a match with one specific document element, here symbolised by the three bars in the document. (A document element is a set of terms in a document with common properties, depending on the document. See Section 3.) The second time all queries find this document again, the third time only *Q2* and *Q3* retrieve the document, etc. When an engine stops retrieving a document in response to a certain query, and when this document did not disappear in reality we say that a singular fluctuation occurs. In this example, these fluctuations occur at the letters F, in observations 3, 4, 5 and 7.

In Section 5 we define an indexing fluctuation as a singular fluctuation that is not due to document fluctuations or to element fluctuations. The first condition means that the document should be known by the engine all of the time. And as can be seen, this condition is fulfilled in this example. Every time the observation was performed, i.e. every time the 3 queries were submitted, at least 1 of the queries found the document. The second condition means that we have to expand the definition of an indexing fluctuation: the disappearance of a document out of the result set of a query should last several successive observations. In our test procedure we say that the disappearance should last at least 4 successive observations; indexing fluctuations then occur at the circled letters F in observation 4 and 7

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

3. TEST PROCEDURE: INVESTIGATION METHODS

Sixteen identical test documents were placed on different sites in the Netherlands and in Belgium; seven documents were submitted to the engines and linked to a document known by the engines. (All sixteen URLs are mentioned in the appendices.) These sixteen documents remained unchanged during the test period.

Thirty-two queries were formulated so that each one was related to a specific 'document element' of the test document. A document element is a set of terms in a document with common properties, depending on the document, such as the appearance of the term in HTML tagged fields, or the number of appearances of the term in the document, or in certain HTML tagged fields, or the place of the term in the document (e.g. only terms near the beginning are indexed) etc. The properties can also be document-independent, such as the number of appearances of the term in the indexed database of the engine, or the number of times that the term was submitted within a search query, etc. These document-independent criteria cannot be readily investigated and so we restricted our investigation to document elements.

In our study we considered HTML tagged fields as document elements, and we worked out to what extent each engine indexed the full text of the test document, i.e. we considered as document elements terms with a certain distance from the beginning of the document: e.g. terms between lines 600 and 700 of the document. Furthermore, we investigated whether the URL of a document was indexed (document element: all terms in the URL). See the appendices. We made the assumption that a term in a HTML tagged field is always indexed or never, as if the appearance in a HTML tagged field is the only criterion for an engine to index a term. Maybe there are other document elements involved (a term can belong to more than one document element). Maybe there are document-independent properties, such as the number of appearances of the term in the indexed database of the engine, or the number of times that the term was used in other search queries or other more complicated criteria that might determine whether the engine indexes the term or not, but that we cannot readily investigate. Similar assumptions have been made for the other investigated document elements. See the appendices.

These queries were repeatedly and regularly submitted: 1 per 29 minutes. One complete observation of $32 \times 13 = 416$ queries (all queries to all engines) took nine days. Forty-three rounds were included in the experiment reported here, during fourteen months, from October 1998 up to December 1999. Vast differences among search engines were observed in the number of test documents retrieved (Table 1) and in the number of document elements indexed (Table 2). According to our definitions of the fluctuations, all thirty-two queries should be submitted to one engine at the same time. For technical reasons we did not submit all thirty-two queries at one time, but one after the other in 31×29 minutes = 15 hours. In the appendices we explain why we think that this does not affect our conclusions.

The following were observed/counted/measured:

- which of the identical documents were retrieved by each engine;
- which document elements were indexed by each engine;
- the number of plural, singular and instantaneous fluctuations per engine;

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

September 2001

SEARCH ENGINES

Table 1. *Total number of our test documents retrieved at least once by each search engine during the test period*

<i>Search engine</i>	<i>Documents</i>
AltaVista	5
EuroFerret	6
Excite	6
HotBot	9
Ilse	5
InfoSeek	2
Lycos	3
MSN	4
NorthernLight	3
Search.nl	5
Snap	9
Vindex	8
WebCrawler	2

Table 2. *Total number of studied document elements indexed at least once by each search engine during the test period*

<i>Search engine</i>	<i>Document elements</i>
AltaVista	20
EuroFerret	4
Excite	16
HotBot	18
Ilse	16
InfoSeek	13
Lycos	17
MSN	27
NorthernLight	16
Search.nl	18
Snap	18
Vindex	28
WebCrawler	19

and

- the size of these fluctuations, i.e. the number of documents missing, expressed as a percentage of the number of documents expected:
 1. plural fluctuations: $0 < \text{size} \leq 100$
 2. instantaneous fluctuations: $0 < \text{size} \leq 100$
 3. singular fluctuations: $\text{size} = 100$ because all expected documents are missing, every time that such a fluctuation occurs.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

Table 3. *Test procedure*

	<i>Observation</i> <i>01</i>	<i>Observation</i> <i>02</i>	...	<i>Observation</i> <i>n</i>	...	<i>Observation</i> <i>43</i>
Q01	r(01.01)	r(01.02)		r(01. <i>n</i>)		r(01.43)
Q02	r(02.01)	r(02.02)		r(02. <i>n</i>)		r(02.43)
Q03	r(03.01)	r(03.02)				
...						
<i>Qm</i>	r(<i>m</i> .01)	r(<i>m</i> .02)		r(<i>m</i> . <i>n</i>)		
...						
Q31	r(31.01)	r(31.02)				
Q32	r(32.01)	r(32.02)		r(32. <i>n</i>)		r(32.43)
	R(01)	R(02)	...	R(<i>n</i>)	...	R(43)

Table 3 shows the test procedure. One observation is the submission of all thirty-two queries (Q1 ... Q32) at one moment. Time goes from left to right. Per engine forty-three observations were performed. Each query results in a result set that contains only the test documents that were retrieved: query *m* in observation *n* retrieves result set $r(m.n)$. Each observation *n* results in a result set $R(n)$, which is the joining of all the result sets $r(i.n)$.

Plural fluctuations are measured by comparing the sets $R(n)$. The frame of reference for an observation was the preceding observation: if $R(n)$ contains a document that $R(i+1)$ does not contain, we say a plural fluctuation appears in observation *i*+1. See the grey row in Table 3. Because each observation is made from thirty-two single queries, we can exclude any influence from singular fluctuations and from instantaneous fluctuations.

Singular fluctuations are measured by comparing the sets $r(m.n)$ for some *m* – see the row *Qm*. A singular fluctuation occurs in $r(m,i)$ if $r(m,i)$ contains no test documents while $r(m, i-1)$ contains one or more test documents.

Sometimes it is not clear if an empty result set is due to a singular fluctuation or to the disappearance of one or more documents because of a plural fluctuation or an instantaneous fluctuation; especially if the search engine knows only one test document. To avoid double counting of fluctuations, we decided to count the disappearance of all known documents in the result set of one query as a singular fluctuation, when the loss of the documents cannot be due to a plural fluctuation (in that case only the plural observation is counted) and when no documents are found in at least four successive observations (otherwise we say one or more instantaneous fluctuations occur).

The choice of the number four seems rather arbitrary, but is motivated by the interpretation given later in this paper where we call singular fluctuations indexing fluctuations because they seem to represent changes in the indexing policy of search engines. Therefore, they should last for several successive observations. How many observations are necessary to determine if an engine has changed its indexing policy? We think that considering only two or three successive observations is not enough to distinguish instantaneous fluctuations from indexing fluctuations. We chose four. Our results show that the choice of three or five should not change the results significantly.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

September 2001

SEARCH ENGINES

Singular fluctuations are detected by comparing the result sets $r(m,i)$ for any query Qm : a singular fluctuation occurs in $r(m,i)$ if $r(m,i-1)$ does contain one or more test documents, and when $r(m,i)$, $r(m,i+1)$, $r(m,i+2)$ and $r(m,i+3)$ are empty while $R(i)$, $R(i+1)$, $R(i+2)$, $R(i+3)$ are not empty. Because we examined here a period of four observations, we have had to take into account the situation that an engine did not know the document at all in one or more of these four observations and the situation that an engine was unreachable in one or more of these four observations. So we have expanded the definition a little bit so that a singular fluctuation also can be detected in these situations: a singular fluctuation occurs in $r(m,i)$ if $r(m,i-1)$ does contain one or more test documents, and when $r(m,i)$, $r(m,i+1)$, $r(m,i+2)$ and $r(m,i+3)$ are empty while:

1. $R(i)$, $R(i+1)$, $R(i+2)$, $R(i+3)$ are not empty; or
2. $R(i)$ is not empty and two successive results sets out of the set $[R(i+1)$, $R(i+2)$, $R(i+3)]$ are not empty, while the third one is empty because no documents were found at all in this observation; or
3. $R(i)$ and $R(i+1)$ are not empty and one of the result sets out of the set $[R(i+2)$ and $R(i+3)]$ is empty because no documents were found in this observation, while the other one is empty because of the fact that the engine was unreachable when query m was submitted.

Singular fluctuations detected with this extended definition occurred only with AltaVista (twice).

Instantaneous fluctuations are measured by comparing the results sets $r(m,n)$ for some n – see the column ‘Observation n ’ in Table 3. They appear when, in observation n , at least two results exist that are not equal; i.e. when there is at least one m so that the set $r(m,n)$ misses test documents compared to $R(n)$. This set $r(m,n)$ may be empty, if it is part of a sequence of at the most three successive empty result sets for query m , so that there is no influence from singular fluctuations. Plural fluctuations cannot influence this because they are measured by comparing whole observations and not single queries.

Not all investigated engines were observed for the whole test period. The investigation of EuroFerret started, for technical reasons, in observation 13, Excite in observation 3, Ilse in observation 13, InfoSeek in observation 12, MSN in observation 18, NorthernLight in observation 9, Search.nl in observation 18 and Snap in observation 13.

4. EXPERIMENTAL RESULTS

All quantitative data about the result sets per query per observation per search engine are available at <http://www.cwi.nl/cwi/projects/IRT/index.html>

4.1 Plural fluctuations observed

Figure 4 visualises the appearance of plural fluctuations from our test results for one engine, as an example. Visualisations for all engines are included in the appendices. Here it can also be seen how some documents disappear and then come back again. This happened with several engines.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

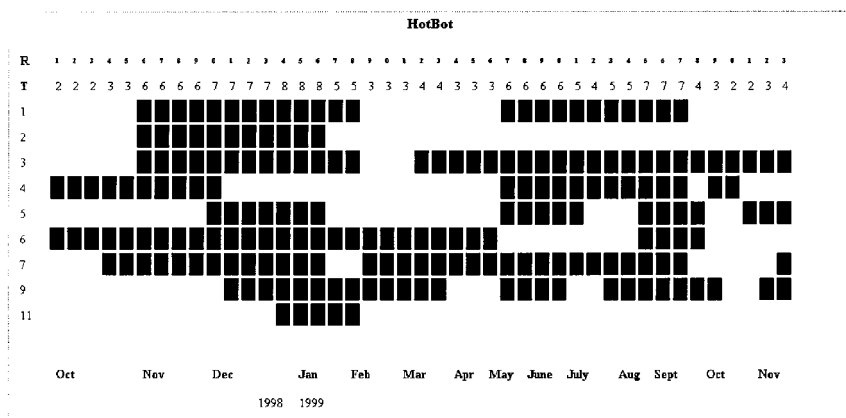


Figure 4. *HotBot* retrieved during the test period 9 different test documents: 1, 2, 3, 4, 5, 6, 7, 9, 11. Appendix 2 contains the corresponding URLs. Here it is shown when *HotBot* found them. Similarly to Figure 1, the result of each observation is represented in a column; time goes from left to right. The upper line (R) shows the observation numbers (1 to 43; expressed in only the second numeral); the second upper line (T) shows the total number of test documents that were found in each observation. The following (numbered) rows show which of the test documents were retrieved in each observation: a filled cell (at the row with number *n* and the column belonging to observation *m*) shows that test document *n* was retrieved with observation *m* (i.e.: at least one of the 32 test queries belonging to observation *m* resulted in this test document); an empty cell means that this test document was not retrieved with this observation. A plural fluctuation is recognised when one of the horizontal rows with successive filled cells ends. Plural fluctuations (with one or more missing documents) are measured in observations 11, 17, 19, 24, 27, 31, 32, 38, 39, 40, 41

Table 4 shows the experimental results of our investigation on plural fluctuations for all engines. The number of fluctuations and the size of the fluctuations are measured.

4.2 Singular fluctuations observed

We have measured the number of singular fluctuations and the size of the fluctuations. Every time that such a fluctuation occurs, all expected documents are missing, so that the average percentage of missed documents per result set is equal to the percentage of result sets with missing documents. Table 5 shows the experimental results of our investigation on singular fluctuations for all engines.

4.3 Instantaneous fluctuations observed

We have measured the number and the size of the instantaneous fluctuations. Figure 5 shows an example of the appearance of instantaneous fluctuations in our test. Table 6 shows the experimental results.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

September 2001

SEARCH ENGINES

Table 4. *Plural fluctuations: experimental results showing average percentage of missing documents per observation and percentage of observations with one or more missing documents*

<i>Search engine</i>	<i>% of observations with one or more missing documents</i>	<i>Average % of missing documents per observation</i>
AltaVista	7.1	4.8
EuroFerret	18.5	7.2
Excite	35.0	33.8
HotBot	26.2	9.1
Ilse	0.0	0.0
InfoSeek	3.2	3.2
Lycos	6.9	5.2
MSN	16.0	12.3
NorthernLight	5.9	5.9
Search.nl	0.0	0.0
Snap	43.3	14.3
Vindex	21.4	3.4
WebCrawler	20.0	16.7

Table 5. *Singular fluctuations: experimental results*

<i>Search engine</i>	<i>Average % of missing documents per result set = % of result sets with missing documents</i>
AltaVista	0.4
EuroFerret	0.0
Excite	0.0
HotBot	0.7
Ilse	0.0
InfoSeek	1.5
Lycos	0.0
MSN	1.7
NorthernLight	0.0
Search.nl	0.0
Snap	0.6
Vindex	1.2
WebCrawler	0.0

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

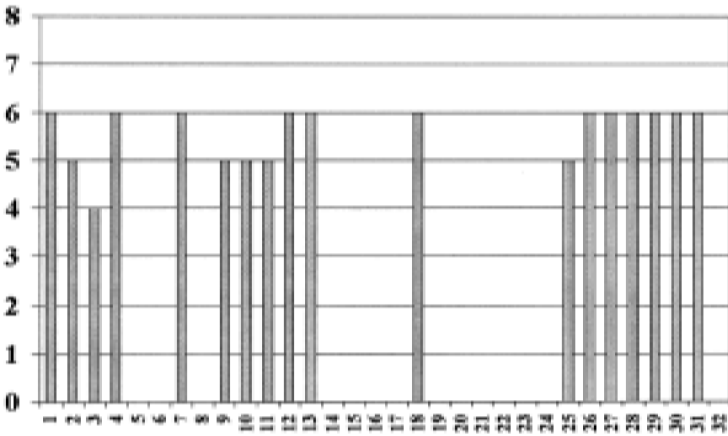


Figure 5. We see, as an example, the results of observation 16 of one of the investigated engines (HotBot). The horizontal axis shows the various queries, while the vertical axis shows the number of retrieved (identical) documents. In a simple world, one would expect that every query retrieves all known documents or none, but we see here result sets that contain sometimes 6 documents and sometimes 5 or 4. So there are instantaneous fluctuations. The empty result sets could indicate theoretically also instantaneous fluctuations, but analysing the data from the preceding and successive observations shows that this is not true.

But the worst has yet to come: the number of known documents in this observation is not 6 but 8. So, all result sets show instantaneous fluctuations and are incomplete, except for the queries that found nothing

Table 6. Instantaneous fluctuations: experimental results

Search engine	% of result-sets that were incomplete	Average % of missing documents per result-set
AltaVista	7.2	5.2
EuroFerret	0.0	0.0
Excite	6.6	3.6
HotBot	56.2	22.5
Ilse	12.9	9.1
InfoSeek	5.0	2.8
Lycos	0.0	0.0
MSN	57.3	29.6
NorthernLight	0.0	0.0
Search.nl	0.5	0.5
Snap	9.7	5.3
Vindex	93.2	22.9
WebCrawler	7.1	3.7

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

September 2001

SEARCH ENGINES

5. DISCUSSION AND CONCLUSIONS

5.1 *Interpretation*

In this section we present our interpretation of what happened when fluctuations occurred in our test.

5.1.1 Plural fluctuations: document fluctuations Plural fluctuations (fluctuations detected through time from the same plural observation) occurred every time that one of the result sets did not contain a document that was contained in the preceding set and that did not disappear from the web in reality. Because we used thirty-two different queries for one observation, we have interpreted an occurrence of this type of fluctuations as the disappearance, for whatever reason, of a document out of the indexed database of an engine. This document could simply not be retrieved. Therefore we call them document fluctuations. In Section 5.8 we also consider some less trivial possible causes for the phenomenon of plural fluctuations.

5.1.2 Instantaneous fluctuations: element fluctuations Instantaneous fluctuations (fluctuations detected at one moment from different singular observations) occurred every time that an engine did not find what should be found according to the result sets of other queries. What does the definition mean? Result sets can be inconsistent and they can be incomplete. In this type it is, oddly enough, not important if a document exists in reality, but if it exists in the result set of one of the queries. We call a fluctuation of this type an 'element fluctuation', because it corresponds with a certain query, which corresponds with one or more document elements (sets of terms in a document with common properties, depending on the document – see Section 3.) An element fluctuation can be caused by an error in the indexing process, an error in the retrieving process (e.g. incomplete use of the database(s) of the search engine) or by an error in both processes.

In order to learn more about what really happens, one should analyse element fluctuations more deeply. This can be done in many ways. Are there many successive element fluctuations for the same query? Are there many element fluctuations within one or more successive observations? Are there many element fluctuations for specific queries or clusters of queries? Are there many element fluctuations at a specific time of the day, or day of the week, etc.? These analyses can lead to information about what causes element fluctuations: errors or limited activities in the indexing process, in the retrieving process or in both processes. Broadly speaking, an engine makes errors of the first kind if it shows element fluctuations for the same queries in successive observations; it makes errors of the second kind if it shows many element fluctuations in one observation. We found the following:

- HotBot, Ilse, InfoSeek, MSN, Snap and Vindex showed element fluctuations in specific queries in successive observations;

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

- AltaVisa, Excite, HotBot, MSN and Vindex showed element fluctuations in specific observations;
- most engines showed element fluctuations in the result sets of all queries, or in none of them. Only three engines had difficulties with specific queries:
 - Ilse only showed element fluctuations in queries 26–32 (which means that Ilse showed element fluctuations in 38% of all queries that retrieved test documents). This could indicate that structural errors were made in the indexing process;
 - Infoseek showed difficulties only with queries 3–18 (46%) which also indicates indexing errors;
 - Vindex, MSN and HotBot showed element fluctuations in all queries, but also in clusters of observations. Both indexing and retrieving errors occurred. Snap showed element fluctuations in queries 4–31 (78%).

5.1.3 Singular fluctuations: indexing fluctuations Singular fluctuations (fluctuations detected through time from the same singular observation) occurred every time that an engine stopped retrieving a document with a certain query for a long time (at least four successive observations), so we think that the engine stopped indexing the search term(s) that defined the query. This is why we call them indexing fluctuations. What are the properties of a term that determine whether the term is indexed or not? This, of course, is one of the secrets of a search engine.

As we have described in Section 3, the properties can be dependent on the document and they can be independent of the document. Document elements are classes of search terms with one or more common properties that depend on the document. Examples of document-dependent properties are: the appearance of the term in HTML tagged fields, or the number of appearances of the term in the document, or in certain HTML tagged fields, or the place of the term in the document (e.g. only terms near the beginning are indexed) etc. In this case of document dependency, the attitude of the engine towards these properties apparently has changed: it has stopped indexing terms with this particular property or set of properties. Examples of properties that are document-independent are: the number of appearances of the term in the indexed database of the engine, or the number of times that the term was submitted within a search query, etc. In this case the attitude of the search engine towards the property did not change, but the environment of the property changed. The term became less important. These document-independent criteria cannot readily be investigated and so we restricted our investigation to document elements.

We think that what we can now observe is the indexing policy of an engine with respect to the investigated document elements. By analysing all the result sets of one query through time, it is possible to establish if an engine indexes the document element corresponding with the query, and if this indexing policy changes: the engine can start or stop indexing this document element. When it stops indexing, we say an indexing fluctuation occurs.

As reported in Section 3, we made the assumption that a term in a HTML tagged field is always indexed or never, as if the appearance in a HTML tagged

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

September 2001

SEARCH ENGINES

field is the only criterion for an engine to index a term. Maybe there are other document elements involved (a term can belong to more than one document element). Maybe there are document-independent properties, such as the number of appearances of the term in the indexed database of the engine, or the number of times that the term was used in other search queries, or other more complicated criteria that determine if the engine indexes the term or not, that we cannot readily investigate. Similar assumptions have been made for the other investigated document elements.

5.2 *Conclusion: search engines make mistakes*

The Internet retrieval systems investigated:

- forget documents completely (document fluctuations);
- miss documents in their result sets (element fluctuations); and
- are subject to changes in indexing policy (indexing fluctuations).

5.3 *Published research on fluctuations of search engines*

The first quantitative results of this research on fluctuations in retrieval by Internet search engines have already been published (Mettrop & Nieuwenhuysen, 1999). A case study by Bar-Ilan (1999) of a few international search engines has also revealed document fluctuations, in agreement with our results. A case study of two search engines at the end of 1999 revealed considerable fluctuations in retrieval by AltaVista but not in that by NorthernLight (Rousseau, 1999). Our research has been presented at two international conferences; the slides with data and graphs used for the presentation can be found in the printed published proceedings (Mettrop et al., 2000), as well as a first text version (Mettrop & Nieuwenhuysen, 2000). This present paper is the full report of the investigation. After submitting the first version of this paper for publication, a comparison was published of two search tools; this study revealed significant daily fluctuations in the results presented by HotBot and a high stability in the hits displayed by Snap (Bar-Ilan, 2000), even though both systems base their results on the Inktomi database.

5.4 *Fluctuations compared*

The influence of the three types of fluctuations is compared in Table 7, by considering the percentages of documents lost by each fluctuation. It is seen that the influence of indexing fluctuations can be neglected.

5.5 *What is correct and what is incorrect?*

The fluctuations have one thing in common: users may miss documents, including documents that they expect to retrieve. The impact of each type of fluctuation is different (which is measured in this investigation) and so is the degree to which search engines should be blamed for it. In Section 5.2 mistakes are mentioned: is every fluctuation we define a mistake? One can say that fluctuations are correct if they are a reflection of alterations in the WWW reality (documents are added or

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

Table 7. *Percentage of documents missed due to fluctuations*

<i>Search engine</i>	<i>Lost by indexing fluctuations</i>	<i>Lost by document fluctuations</i>	<i>Lost by element fluctuations</i>
AltaVista	0.4	4.8	5.2
EuroFerret	0.0	7.2	0.0
Excite	0.0	33.7	3.6
HotBot	0.7	9.1	22.5
Ilse	0.0	0.0	9.1
InfoSeek	1.5	3.2	2.8
Lycos	0.0	5.2	0.0
MSN	1.7	12.3	29.6
NorthernLight	0.0	5.9	0.0
Search.nl	0.0	0.0	0.5
Snap	0.6	14.3	5.3
Vindex	1.2	3.4	22.9
WebCrawler	0.0	16.7	3.7

removed). If not, they are incorrect. According to this definition, all of the three types of fluctuations that we consider are incorrect. Another definition could say that fluctuations are correct if they are a reflection of alterations in the indexed database of an engine. From that perspective, document fluctuations and indexing fluctuations are correct, but element fluctuations are still incorrect. In practice most users will not care about what is 'correct' or 'incorrect'; they simply cannot find a document that still exists in reality. If it is not their own mistake, it is the engine's mistake.

5.6 *Comparison of search engines, concerning fluctuations*

Document fluctuations have been found in the result sets of all the investigated search engines except for two of the search engines restricted to the Netherlands: Ilse and Search.nl. Most document fluctuations were shown by Snap and Excite. Excite also missed most documents per observation due to document fluctuations. Element fluctuations have been found in the result sets of all search engines except EuroFerret, Lycos, NorthernLight and (almost) Search.nl. Most element fluctuations were shown and most documents per result set were missed by element fluctuations by HotBot, MSN and Vindex.

Some engines were only slightly subject to document or element fluctuations: AltaVista, InfoSeek, Lycos, NorthernLight and Search.nl. Their result sets were almost complete and did not change significantly. Moreover, Lycos, NorthernLight and Search.nl did not change their indexing policy during our research.

Many indexing, document and element fluctuations were observed with HotBot, MSN and Vindex. Moreover, they missed the highest percentage of documents by element fluctuations.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

September 2001

SEARCH ENGINES

5.7 *Relations concerning fluctuations*

- Search engines that showed many element fluctuations also showed many document fluctuations. Moreover, they index many document elements and showed many indexing fluctuations.
- Engines showing many document fluctuations did not always show many element fluctuations.
- Engines that index few document elements showed few element fluctuations.

5.8 *How can fluctuations occur?*

Why do documents disappear from the database of an engine? Why does an engine stop indexing a search term? How can inconsistent result sets be explained? The first two events can lead to document fluctuations and indexing fluctuations. They can be caused of course by the conscious decisions of (the builders of) an engine to remove a document (because it is pornographic, aggressive or ...) from its database, or to stop indexing a document element (because, maybe, nobody searches for it). For indexing fluctuations this might be a conclusive explanation, but we think that it is unlikely that all document fluctuations can be explained that simply. Moreover, all test documents are identical and so all of them should be removed or none of them. Element fluctuations cannot be explained on the basis of such a conscious decision, or it should be the decision to limit the burdening of the search engine's computer system, which can lead to poor work in the indexing process, the retrieving process or in both processes.

In addition, we can mention four less trivial possible causes:

- *Several databases.* Andrei Broder (AltaVista) said in an interview that AltaVista is aware of the fluctuations described here. The most important reason probably is the fact that, although it appears to the user that he is querying one database, he is actually searching a construction of several databases. This implies that the indexes of each database must be replicated regularly. The query is not always taken care of by the same database. So unfinished replication can cause fluctuations in the search result.
AltaVista tries to limit the influence to a minimum, but gives priority to producing the search results as quickly as possible. Avoiding all fluctuations should cause an unacceptable reduction of the search speed and cost too much. In order to know when to change the arrangements, they measure the size of the fluctuations every thirty minutes. See the AltaVista HELP http://doc.altavista.com/help/search/search_help.html (visited 2000).
- *Different indexes.* Another important explanation, caused by technical inadequacies, for fluctuations in the search results, mentioned by Broder, is the fact that for different parts of the Internet, there are different indexes. If one of these indexes is, for some reason, unreachable, this temporarily causes incomplete search results.
- *Engines use performance-dependent algorithms.* Also mentioned by Broder. The more queries that need to be handled, the less these algorithms are activated, the more chances for fluctuations. Element fluctuations can be explained by this policy.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

- *Duplicate removals.* Some search engines try to detect duplicate identical documents among all the documents that they harvest from the WWW, so that they can remove identical copies from their database and index, because their viewpoint is that one copy is sufficient while the other copies make the search engine less efficient. Then there is one representative for a set of identical documents. Changing the representative, for whatever reason can lead to some of the document fluctuations that we have detected.

There are three engines that never retrieved more than one document in our tests (see the appendices): Excite, NorthernLight and WebCrawler. This is most important for Excite, which showed many document fluctuations. NorthernLight only twice changed the representative test document, showing three different test documents in the test period and WebCrawler also changed twice showing only two test documents, while Excite retrieved six different documents and changed fourteen times!

5.9 *Are fluctuations in search results important?*

Fluctuations are important – we think – for both the users of search engines and the providers of information on the web and therefore deserve to be investigated.

Most users working with databases (and information systems in general) expect identical results for identical queries (or following identical actions performed with an information system). This is in particular true for users of more classical databases provided by well-known, classical online information retrieval systems like Dialog and certainly by users of databases on CD-ROM. So fluctuations and inconsistencies are unexpected for most users. Either they do not realise that they occur, or they observe them in a non-systematic, more informal way than reported here.

Users who are unaware of fluctuations can be confused when they cannot retrace a document found earlier, can make inferences about the performance, size, contents of the search engine that are not correct, can make quantitative measurements about citations (links) to web-based documents that are not correct, and so on ... So first it is important to know that fluctuations and inconsistent result sets do exist. Then it is important to observe them in a systematic way. When users remain ignorant about fluctuations, they can realise that ‘something strange is happening’, but still be puzzled and lose confidence in a search engine or even in the Internet as a whole as a medium for accessing and distributing information; also they have a problem in discussions when someone finds a document while another person does not, even at the same moment. And of course users who search for all information on a specific topic (e.g. in the case of the application for a patent) should be aware of this.

Now to the providers of information. They can see the absence of their documents in search results due to fluctuations as a problem. This is true for scientific information as well as for commercial information. Many companies exist only because of their appearance on the web. They do not like the fact that their pages are only retrieved most of the time and sometimes not at all.

Finally, in any case, any quantitative analysis of the web can be impaired by fluctuations and inconsistencies. This can be important, for instance, in the

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

September 2001

SEARCH ENGINES

domain of scientometrics where citation analysis is used to assess production, distribution and impact of research results; this is particularly important now that the role of printed documents is taken over more and more by web-based documents. This can also be important in the commercial world, where numbers of visitors and links determine the costs of advertising on a site.

ACKNOWLEDGEMENTS

Many thanks go to Hanneke Smulders. Without her valuable remarks this paper would never have been completed.

The following persons have been involved in some way and in some phase of the research project on Internet information retrieval tools (in alphabetical order):

- Louise Beijer (Hogeschool van Amsterdam, The Netherlands)
- Hans de Bruin (Unilever Research Laboratorium, Vlaardingen, The Netherlands)
- Hans de Man (JdM Documentaire Informatie, Vlaardingen, The Netherlands)
- Marten Hofstede (Universiteit Leiden, The Netherlands)
- Wouter Mettrop (CWI, Amsterdam, The Netherlands)
- Paul Nieuwenhuysen (Vrije Universiteit Brussel and Universitaire Instelling Antwerpen, Belgium)
- Eric Sieverts (Hogeschool van Amsterdam and Rijksuniversiteit Utrecht, The Netherlands)
- Hanneke Smulders (Infomare, Terneuzen, The Netherlands)
- Hans van der Laan (Computer and Internet consultant, Leiderdorp, The Netherlands)
- Ditmer Weertman (Nederlands Jazz Archief, The Netherlands)

REFERENCES

- Bar-Ilan, J. (1999). Search engine results over time – a case study on search engine stability. *CyberMetrics*, 2/3 (Issue 1, paper 1) <http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html> (visited April 2000).
- Bar-Ilan, J. (2000). Evaluating the stability of the search tools HotBot and Snap: a case study. *Online Information Review*, 24(6), 439–449.
- Inktomi Corporation & the NEC Research Institute, Inc. (2000). *Web surpasses one billion documents: Inktomi and NEC Research Institute complete first web study*. <http://www.inktomi.com/webmap/index.html> and <http://www.inktomi.com/new/press/billion.html> (visited February 2000).
- Lawrence, S. & Giles, C.L. (1999). Accessibility of information on the web. *Nature*, 400, 8 July, 107–109.
- Mettrop, W. & Nieuwenhuysen, P. (1999). Some empirical research on the performance of Internet search engines. In: Aparac, T., Saracevic, T., Ingwersen, P. & Vakkari, P., eds. *Digital Libraries: Interdisciplinary Concepts, Challenges and Opportunities. Proceedings of CoLIS 3, the Third*

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

JOURNAL OF DOCUMENTATION

vol. 57, no. 5

International Conference in Library and Information Science, Inter-University Centre Dubrovnik. Zagreb, Lokve, Croatia: Benja Publishing, 329–333.

Mettrop, W. & Nieuwenhuysen, P. (2000). The reliability of Internet search engines: fluctuations in document accessibility. In: Williams, H.E., ed. *Proceedings of National Online Meeting 2000, May 16–18, 2000, New York*. Medford, NJ: Information Today Inc., 271–281.

Mettrop, W., Nieuwenhuysen, P. & Smulders, H. (2000). Internet search engines: fluctuations in document accessibility. In: Nixon, C. & Burmood, J., eds. *Proceedings of Internet Librarian International 2000, March 20–22, 2000, London, UK*. Medford, NJ: Information Today Inc., 153–163.

Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight. *Cybermetrics*, 2/3. <http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html> (visited January 2001).

(Revised version received 5 April 2001)

APPENDIX 1 – INDEXED DOCUMENT ELEMENTS

The thirty-two queries that we used for one observation correspond to thirty-two different document elements. A document element is a set of terms in a document with common properties, depending on the document (see Section 3). We worked out which HTML tagged fields were indexed, to what extent each engine indexes the full text of the test documents and if the URL of a document is indexed. The test document has been made especially for this purpose.

We considered the following document elements:

HTML tagged fields:

- title tag
- meta tags: keywords, description and author
- comment tag
- alt tag
- text/URL of a link to a document
- H3 tag
- table header
- text of: an internal link, a reference anchor, a link to a sound file
- name of a sound file (au/wav/aiff/ra)
- text of a link to an image
- name of an image file (gif or jpg; inline or linked to)
- name of a Java applet (with or without extension class)

Full text investigation:

- terms after the first 100 lines in a document (200/.../700)

URL of document:

- the URL of a document

All queries are mentioned on our web site at <http://www.cwi.nl/cwi/projects/IRT/index.html>.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

September 2001

SEARCH ENGINES

Table 8. *Tags indexed by investigated search engines*

<i>Tag</i>	<i>Indexed by</i>
title tag	all 13 engines
meta tag keywords	AltaVista, HotBot, Ilse, InfoSeek, MSN, Snap and Vindex
meta tag description	AltaVista, HotBot, Ilse, InfoSeek, MSN, Search.nl and Snap
meta tag author	HotBot, MSN and Snap
comment tag	Vindex
alt tag	AltaVista, EuroFerret, Lycos and Search.nl

The number of studied document elements that were indexed by each search engine at least once during the test period is shown in Table 2. We have also investigated which meta tags were indexed at least once during the test period, by each engine: the results are shown in Table 8.

APPENDIX 2 – TEST DOCUMENTS

In our test we used sixteen test documents, divided into two classes. The first class contains the test documents that we tried to make known by the search engines. We submitted them to the search engines and we made a link from the WWW home page of our research project (<http://www.cwi.nl/cwi/projects/IRT/index.html>) to them. This class contains seven test documents. The second class contains the nine other test documents that were not submitted or linked. Test documents in class 2 were hardly retrieved by the engines. The URLs used were:

Submitted and linked:

- 1 <http://home.worldonline.nl/~edirks/pagina.html>
- 2 <http://www.cwi.nl/~wouter/irttest/pagina.html>
- 3 http://www.vub.ac.be/BIBLIO/IRT_linked/pagina.html
- 4 <http://www.algerak.nl/irttest/pagina.html>
- 5 <http://www.xs4all.nl/~weertman/irttest/pagina.html>
- 6 <http://www.dsv.nl/~asse-nl/irttest/pagina.html>
- 7 <http://www.globalxs.nl/home/jjdmocinf/irttest/pagina.html>

Not submitted and not linked:

- 8 http://www.algerak.nl/irttest_niet_gelinkt/pagina.html
- 9 <http://home.worldonline.nl/~edirks/no-link/pagina.html>
- 10 <http://www.cwi.nl/~wouter/irttest-nolink/irttest/pagina.html>
- 11 <http://www.cwi.nl/~wouter/search-tools/IRT/themadag/pagina.html>
- 12 <http://www.cwi.nl/~wouter/irttest-link/irttest/pagina.html>
- 13 <http://www.cwi.nl/~wouter/irttest-htaccess/irttest/pagina.html>
- 14 <http://www.xs4all.nl/~weertman/geenlink/pagina.html>
- 15 <http://www.dsv.nl/~asse-nl/irtnolink/pagina.html>
- 16 <http://www.globalxs.nl/home/jjdmocinf/irtnolink/pagina.html>

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

JOURNAL OF DOCUMENTATION

vol. 57, no. 5

APPENDIX 3 – DOCUMENT FLUCTUATIONS

Figure 6 shows per search engine which of the test documents is found per observation. Similar to Figure 1, the result of each observation is represented in a column; time goes from left to right. The first column contains the numbers 1 – 8; they refer to our test documents. The numbers 1 – 7 correspond with the seven test documents that we submitted to the search engines and that we linked from our IRT homepage. Because the test documents that were not submitted and not linked were hardly retrieved by the search engines, we let number 8 represent all nine of them (see Appendix 2).

For each column there is shown in the upper line (R) the observation number and in the second upper line (T) the total number of test documents found in that specific observation. The following (numbered) rows show which of the test documents were retrieved in each observation: a filled cell (at the row with number n and the column belonging to observation m) shows that test document n was retrieved with observation m (i.e.: at least one of the thirty-two test queries belonging to observation m resulted in this test document); an empty cell means that this test document was not retrieved with this observation. A filled cell in row 8 means that test documents were found that were not submitted and linked. The total number of these is mentioned within the cell. A document fluctuation is recognised when one of the horizontal rows with successive filled cells ends. (The document fluctuations for the documents in row 8 are also measured, although they are not all visualised here.) One observation can be affected by several document fluctuations. An observation with the letter X in all rows means that the engine was not investigated in that observation (for some technical reasons).

In Section 5 we stated that three search engines only retrieved one document at a time in the tests: Excite, NorthernLight and WebCrawler. Here it can be seen that Excite twice retrieved two documents in the same result set and WebCrawler once. This is because of the fact that all the observations were not performed at the same time but took fifteen hours. Each single query in the observations of Excite, NorthernLight and WebCrawler retrieved only one document. See Appendix 5.

APPENDIX 4 – UNREACHABILITY

Sometimes an engine was not reachable, or returned an error message. General data for all engines:

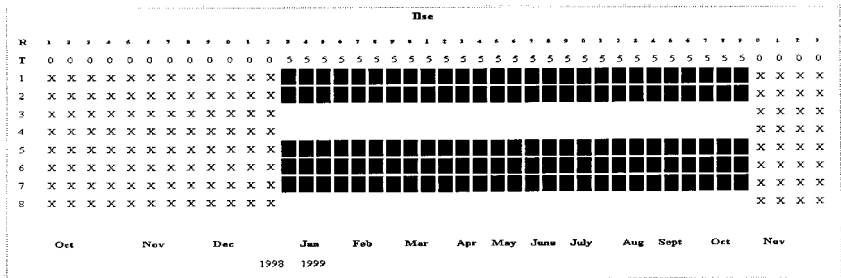
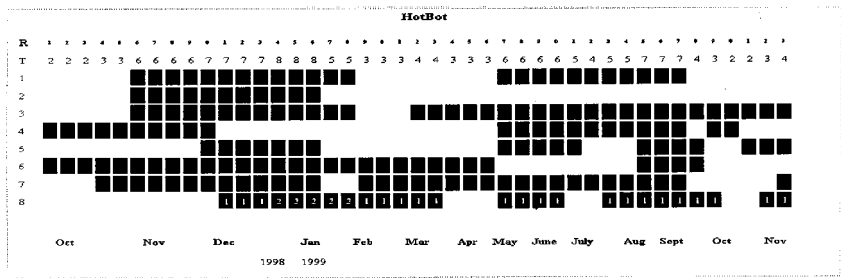
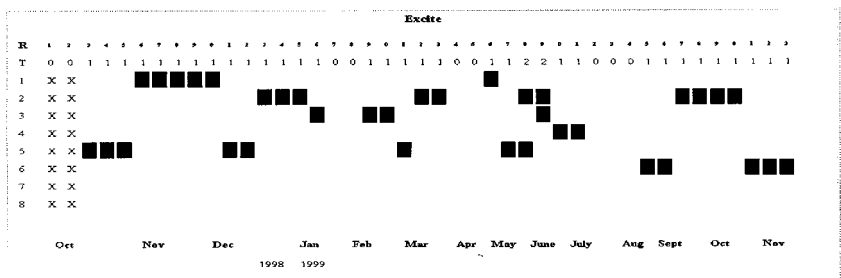
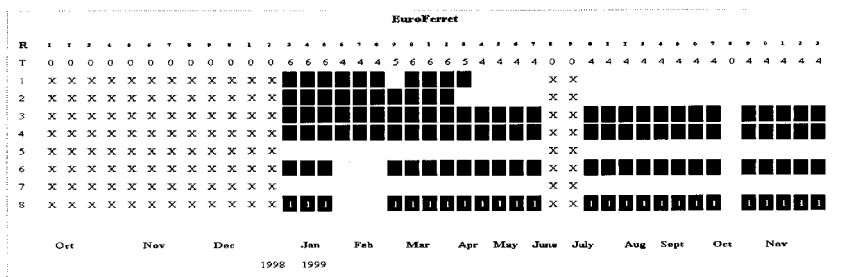
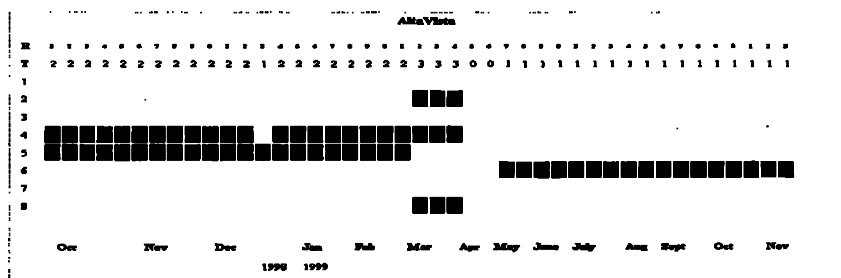
- 14,528 queries were sent to 13 search engines;
- search engines were unreachable 721 times;
- the percentage of unreachability for each engine varied from nearly 0% to nearly 15%;
- the studied search engines were reachable for 95% of the queries.

The percentages of failures per engine are shown in Table 9. Some engines produced failures for specific queries. The queries with most failures per engine are shown in Table 10.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

September 2001

SEARCH ENGINES



All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

September 2001

SEARCH ENGINES

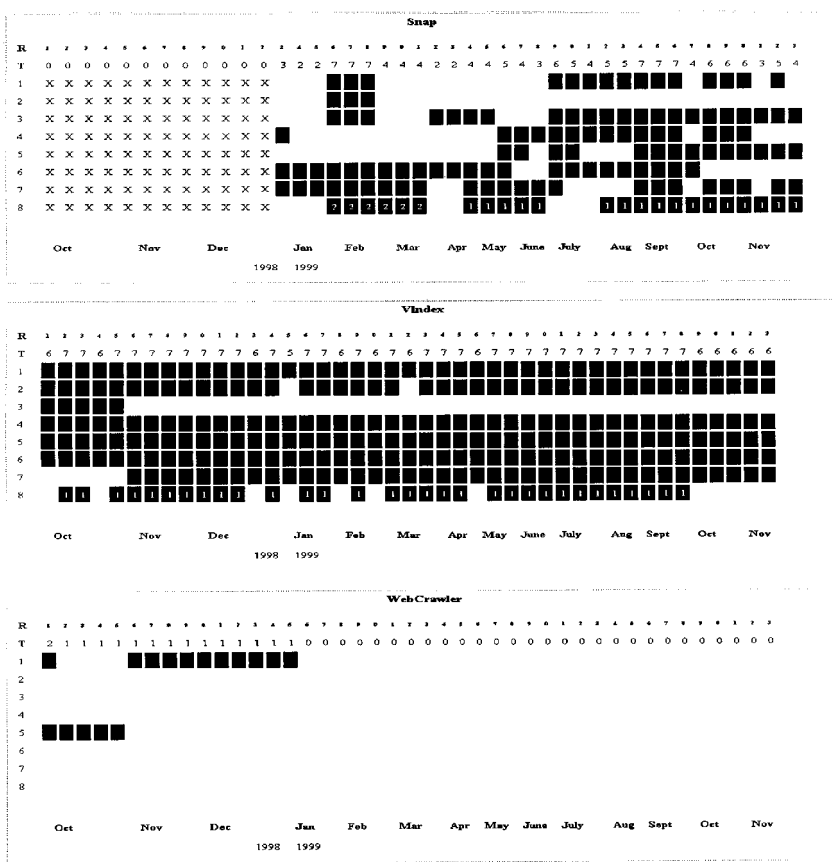


Figure 6. Document fluctuations per search engine

Table 9. Percentage of submissions that failed due to unreachability of the engine

Search engine	% of submissions that failed due to unreachability of the engines
AltaVista	1.2
EuroFerret	14.9
Excite	6.3
HotBot	3.1
Ilse	3.6
InfoSeek	9.0
Lycos	4.5
MSN	0.4
NorthernLight	5.4
Search.nl	2.5
Snap	2.7
Vindex	10.0
WebCrawler	1.2

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

Table 10. *Queries with most failures due to unreachability per engine*

<i>Search engine</i>	<i>Query number</i>	<i>Number of failures</i>
AltaVista	22	3
EuroFerret	7 and 25	17
Excite	19	9
HotBot	20	5
Ilse	2/5/8/13-16/23/27	2
InfoSeek	19	18
Lycos	6	33 (successive times)
MSN	9, 19 and 29	1
NorthernLight	4	15
Search.nl	13-16	2
Snap	8 and 12	6
Vindex	4 and 13	16
WebCrawler	6	3

APPENDIX 5 – COMMENTS ON THE TEST PROCEDURE

We want to make here two comments on our test procedure. Both deal with the fact that we do not know exactly when an engine changes the part of the indexed database that contains the test documents. The first concerns the time between two successive observations and the second concerns the time that one observation took.

1. Time between two successive observations

The time between two observations is a variable in our test procedure. We used an interval of nine days between two successive observations. The length of this interval was chosen for technical reasons only. What would happen if we used another interval? Of course, this should not directly affect our conclusions on element fluctuations and indexing fluctuations, but it could affect our conclusions on document fluctuations.

Document fluctuations occur after an alteration in the part of the indexed database that contains the test documents. We call such an alteration an update. Let us assume that these updates occur regularly with a constant interval between each two successive updates. If we call this interval $i(u)$, which of course is different for each engine, and the interval between two successive observations $i(o)$, we can conclude that investigating document fluctuations will be most accurate when $i(o) = i(u)$. If $i(o) < i(u)$, then the percentage of observations with document fluctuations will be too small; if $i(o) > i(u)$, then the investigation is less accurate, because the number of observations is smaller.

Conclusion: a shorter or longer interval could change the number of detected document fluctuations. Our conclusions are for all engines based on a nine day interval.

All rights reserved. Except as otherwise permitted under the Copyright, Designs and Patents Act 1988, no part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise without the prior written permission of the publisher.

September 2001

SEARCH ENGINES

2. *One observation takes some time*

The definitions of fluctuations that we used are based on a test procedure, where all queries in one observation are submitted at the same time. In our test however – as was mentioned before – the queries were submitted one after the other and not at the same time. In fact, one complete observation lasted about fifteen hours.

Does this affect our conclusions? When an update occurs within one observation, we submit queries before and after the update. So it is possible that, although the set of test documents in the indexed database may have changed, we did not measure a document fluctuation, because document fluctuations are measured by comparing whole observations. Moreover, we may have measured wrongly element fluctuations and maybe indexing fluctuations. Every update that occurs within an observation can generate element fluctuations and can contribute to the detection of an indexing fluctuation. In fact the conclusions that we made are an upper limit for the number of element fluctuations and indexing fluctuations, and a lower limit for the number of document fluctuations.

We think the chance of the part of the indexed database that we study being updated within the fifteen hours of one observation is small. So we think that the chance of this happening so often in all of the observations we made (forty-three observations for most engines) that it could affect our conclusions significantly is negligible.

We have worked this out in more detail in: <http://www.cwi.nl/cwi/projects/IRT/doc/appendix5.txt>