INVESTIGATION OF DOCUMENTS PERSISTENCY ON THE WEB

Luigi Lancieri
France Telecom R&D
42rue des coutures14000 Caen (France)
luigi.lancieri@francetelecom.com

Matthieu Lecouvey
ENSICAEN (University of Caen)
6, boulevard du Marechal Juin 14000 CaenFrance)
ma2thieu@wanadoo.fr

Veronique Glasset
ENSICAEN (University of Caen)
6, boulevard du Marechal Juin 14000 Caen (France)
veroniqueg@free.fr

ABSTRACT

This article presents a study dealing with the appearance and disappearance of Web resources. Through two complementary approaches we presented a quantitative overview of these phenomena according to various criteria. For example, we study the reasons of the documents unavailability or the distribution of the more or less availability according to the topics evoked by the documents. We measure for example that documents related to some topics are more persistent than other or that the increase of the web is not the same in the different topics.

KEYWORDS

Internet, information availability, search engine, web measure, reusability

1. INTRODUCTION

One of the characteristics of Internet network is its capacity to store various kind of contents related to various topics and to make them available from almost all places of the planet. Each day, new contents appear whereas others disappear. Even if we have the feeling that the overall quantity of web information increases, it is difficult to know and to understand the reasons of this more or less availability. The objective of our study is to have a more precise vision of this informational dynamic and to analyze this phenomenon according to various sets of themes. In a first approach, the dynamics of information loss is evaluated starting from a fixed set of Web pages by checking their availability regularly. We can thus check if the loss of availability moves in the same way for the various topics. In a second approach, assorted sets of themes are associated with key words. With regular interval, these key words are provided to a search engine that gives URLs. Then, we compute for each topic the rate of new URLs provided by the search engine and the rate of URLs that disappeared. These results are compared with that of the previous period. A clearer vision of these phenomena can be interesting in several connections. From a theoretical point of view it makes possible a better take into account of the impact of the human behaviour on information systems. Indeed, it seems natural to think that the mechanisms of destruction and creation of resources do not intervene randomly, and that they seem to be directly related to the cognitive and social human behaviours. From a more practical point of view, this kind of evaluation can be interesting in an economic context. Indeed the follow-up of a temporal evolution of the rate of destruction or resources creation for each topic can give a helpful indication of topics on fashion or emerging interest. This article is segmented as follows. After having detailed our method, we present the various results obtained and before concluding, we present a state of the art with some studies or tools close to our work.

2. DESTRUCTION OF RESOURCES

For the first part of our study, we developed a program allowing testing a set of URLs that belongs to rather general topics representative of the Web content. Approximately 200 URLs were selected in each of the 16 topics of the directory of the Google search engine. Thus, we tested daily the availability of almost 3000 links over 97 days. In the general case, each time a user downloads a web page, the server gives in its HTTP response, a status-code that describes the success or the failure of the request. By testing this code, it is thus possible to know if a page has disappeared or if the web server has not been able to find it in a limited time. Table 1 gives the significance of some of these codes. (See also [1] for more details on HTTP status codes). From the various attempts of connections to these links, the servers gave different status codes.

We can observe that, in most cases (93.63 %), we correctly obtain the document requested. However, in 6.37 % of the cases (addition of the errors of the preceding table), the document is non-existent or inaccessible. These cases can be the consequences of various problems. Most of the errors are "404 Not Found". It means that the document corresponding to the URL is non-existent. Thus the majority of the inaccessibility appears to be related with the design of Internet sites and are, for example, the result of a lake in updating the Web site or in referencing it. These results can be different depending on the topics of the documents. These differences are visible on the rate of errors but also on the type of causes. One can compare, for example, the shopping and computers topics that have enough opposed results (see table 1). The first notable difference resides on the percentage of the average error. For a same amount of URLs, the “computers” topic has almost three times more inaccessible documents than the “shopping” one. The quantity of error 403 (prohibited access) is very higher in the case of the “shopping”. This can be explained by the use of secured procedures for the online payments that are largely used on the shopping web site. Thus, the accesses are more often restricted. Another great difference relates to the errors 404 which are more frequent in the computers category. This dissimilarity is logical since computers related content has often a short life and thus the documents are more frequently changed. It is in consequence more probable to find unreachable links than in others topics. Moreover, the computers-oriented web sites are rather often created by individuals whereas the shopping sites are generally managed by professionals.

Tableau 1: Distribution of the HTTP status code

Status code	Meaning	General	Shopping	Computers
Error 403	Forbidden access	8.2 %	19.62 %	1.26 %
Error 404	Page not found	54.3 %	46.46 %	74.90 %
Error 500	Web server error	1 %	1.30 %	0.22 %
Error 503	Web server unavailable	22.5 %	27.85 %	12.79 %
Error 504	Time out	13.5 %	4.76 %	10.84 %
	Other Errors	0.5 %	0 %	0 %

The figure below represents over the tested period the percentage of inaccessible documents according to the selected topics. The segmentation of the set of themes can be debatable and it would be interesting further to consider this study with various forms of segmentation (set of themes or other). In this first version of our work, we begin to use the standard segmentation of the Google directory.

We also computed the evolution of the error rate over the 97 days of the experiment in order to qualify its trends. The trend seems to be linear (from 0 to 6.7 % over 97 days) with a correlation coefficient of 0.84. One interesting experiment not did yet could be to analyse this trend according to the different topics.

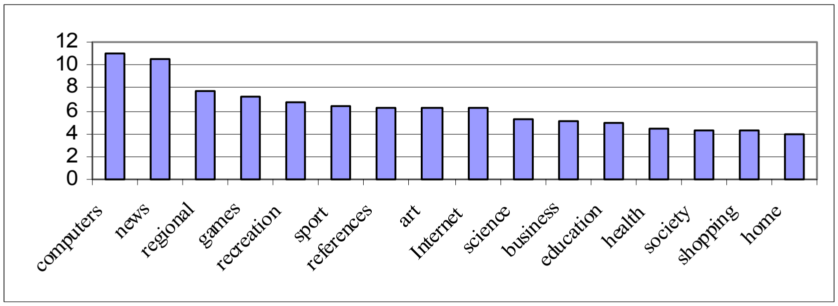

Figure 1: Percentage of error per topic

3. CREATION AND DESTRUCTION OF RESOURCES

Unlike the preceding study where we evaluated the real availability of documents on the origin Web server, we were rather interested here in the evaluation of the Web as it is presented by the search engines. When a surfer launches a request on a search engine, a certain amount of answers are provided but only some of them are presented to the user. These ignored URLs correspond to pages highly similar to others already presented or to pages where the level of relevance is considered as too weak. It is interesting to see that the level of ignored pages is near 50 % but can be very different from one topic to another (29 % education, health 37%, society 38 %, news 43 %, sport 48 %, business 50%, computers 60 %, and games 75 %).

Each day, our program gathers the URLs answering the various requests (fixed key words for each topic). The comparison with the results of the day before makes it possible to find the URLs which appeared or which disappeared daily. It appears that the levels of appearances and of disappearances are relatively equivalent in time. Indeed, we can see that about 4 % (maximum of 8 % and minimum of 0 %) of the contents moves daily (2 % of new URLs and as many disappearances). In order to have a more precise overview, we studied the evolution from the initial situation and by group of topics having similar behaviours regarding to the rate of appearance or disappearance. The graph below represents, for each day, the number of pages appeared (or disappeared) compared to the first day of test. . The lower curve represents the topics (education, computers, recreation, regional, sciences), the intermediate curve (shopping, Internet, games, home, references, health) and the upper curve (news, art, business, society, sport). Comparing the trends from the figure 3 and the figure 1 reveals that topics that growth faster are not always the same that those disappearing less rapidly. For example, the "news" category has a high level of unavailability but is also in the group that grows the faster. This is probably due to a higher dynamic of changes of these resources.

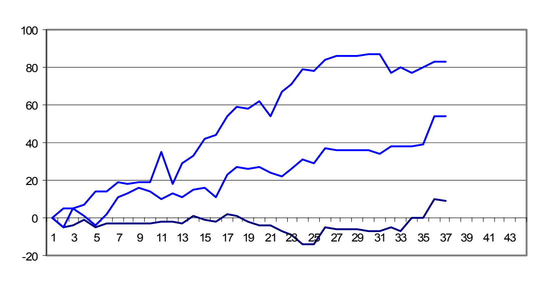

Figure 3: Compared evolution of 3 groups of topics

If we postulate that the search engine database gives a reasonable overview of the web content, we can say that the Web continue to expand itself over the time in almost all the fields but with different speeds depending on the topic.

4. STATE OF THE ART

The accessibility of online documents was the subject of several studies or tools. Many of these tools use the capacities of the response HTTP headers [1]. For example, the link-checkers (e.g. [2]) are tools that make possible to check systematically if links are broken or not. They can check the availability directly from a set of links or by extracting all the imbedded links from a HTML page. There are also many tools or sites which offer statistics of access to Web sites (ISP, xiti, INABIS 2000 ([3]). In particular some of these tools allow a count of the errors when reaching the resources. With these results, we can make precise statistics on the accessibility of the documents depending on one or few servers. But these analyses cannot be generalized at a national or international level. From a more general point of view, a lot of initiative or study aims at measuring the web in order to build resources or connection maps [7], [8],[9]. This works can be helpful in order to have a best understanding of the growth of the Web.

Some other studies were undertaken to know the evolution of the various search engines databases. The following site [4] provides some statistics regarding this aspect. In particular, we can find a report on the dead links given by some popular search engines. The results show the tendency to the reduction of the number of dead links. Indeed, on average, in the major search engines (Northern light, google, hotbot, fast, excite, altavista), the percentage of dead link was 16.6 (sept 99), 9.3 (nov 99) and 5,8 (nov 2000). Even if we cannot take into account the invisible part of the Web (not referred, dynamic...), the search engines make possible having an overview of the web evolution. On average, the indexed documents increase from 0.1 billion in 97 to 0.5 in 2000, 1.5 in 02, and 3.5 in 03. (Google, AllTheWeb, Inktomi, Teoma, AltaVista) [5].

5. DISCUSSION

This study enabled us to have a first overview of the dynamics of information availability on the Web. We also could stress tracks of future works or open questions. As an example, even if we saw that the topic is a key parameter that explains the more or less availability of Web resources, we highlighted the importance of the thematic segmentation. Another aspect to be taken into account is to well dissociate the availability of the information given by the search engines with the reality of the availability of the Web server. Regarding the reasons of the variations of the resources life duration, we highlighted several possible causes. First of all, there are technical reasons dealing with update or servers failures that influence the level of broken links. Except these problems, it is clear that some topics are more disposed to change than others. Since the life duration of some resources is variable, it can be interesting to replicate it. Some technologies as Peer to peer networks (Kazaa, E-mule, Napster, etc) have shown that the availability of data is directly linked to its level of replication. Other replications based techniques as proxy-caches can have a huge impact on the information availability. In the active mirror, [11] the most popular Web content is recycled, based on phenomenon of implicit cooperation among users. In this case resources' life duration can be extended as long as they remain popular. In such strategy of resources reuse, it can be useful to estimate how long native resources can be accessible.

REFERENCES

[1] Hypertext Transfer Protocol -- HTTP/1.1, Request for Comments 2616; available on http://www.ietf.org

[2] Link checker web site; http://www.poisontooth.com/linkcheck/

[3] Ianabis web site http://www.uclm.es/inabis2000/webstat.htm

[4] Search engine show down Web site www.searchengineshowdown.com

[5] Search engine watch web site: www.searchenginewatch.com

[7]CAIDA, the Cooperative Association for Internet Data Analysis; http://www.caida.org/

[8]The National Laboratory for Applied Network Research (NLANR); http://www.nlanr.net/

[9]National Internet Measurement Infrastructure; http://www.ncne.org/nimi/

[10]A.Broder, R.Kumar, F.Maghoul, P.Raghavan, S.Rajagopalan, R.Stata, A.Tomkins, J.Wiener ; Graph structure in the Web ; 9 th WWW conference 2000 http://www.almaden.ibm.com/cs/k53/www9.final/

[11]L.Lancieri, 2004, Reusing implicit cooperation, a novel approach in knowledge management, TripleC Journal