

The lifespan, accessibility and archiving of dynamic documents.

Katarzyna Wegrzyn-Wolska
ESIGETEL, Ecole Supérieure d'Ingénieurs en Informatique et Génie des Télécommunications,
77215 Avon-Fontainebleau, France
katarzyna.wegrzyn@esigetel.fr

Abstract

Today most Web documents are created dynamically. These documents don't exist in reality; they are created automatically and they disappear after consultation. This paper surveys the problems related to the lifespan, accessibility and archiving of these pages. It introduces definitions of the different categories of dynamics documents. It describes also the results of our statistical experiments performed to evaluate their lifespan.

1 Introduction

Is a dynamic document a real document, or is it only the temporary presentation of data? Is it any document created automatically or is it the document created as a response to the user's action? The term "dynamic" can be used for the different signification; for the HTML Web page with some dynamic parts like a layers, scripts, etc., but this term is more often used for the pages created on-line by the Web server. This paper deal with the problems related to the documents created on-line.

2 The Lifespan and Age of Dynamic Documents

How can the lifespan of dynamic documents be evaluated? These documents disappear immediately from the computers' memory after their consultation. In this paper we define the lifespan of dynamic documents as the period where the given demand results in the same given response. This period is the time observed by the user as a documents lifespan. User when surfing the Web in his browser doesn't know how the document was created so he doesn't distinguish the difference between the static and the dynamic document.

How to determine age of the dynamic documents? Can we consider the value of the http header Modified and Expired or the value fixed in the HTML file with the META tag

Expires to indicate exactly when the document was changed or when it can be considered to have expired?

3 Dynamic Documents Categories

We distinguish two kinds of dynamic documents: documents created and modified automatically (news sites, chat sites, weblogs, etc.) and documents created as an answer to the user requests (the results pages given by the Search Engines, the responses obtained by filling in the data form, etc.).

We will analyse these documents separately in two categories. The first category is represented by the response-pages obtained from the Search Engines. The second category contains the pages from the different news sites and the Weblogs sites.

3.1 News Published on the Web

There are numerous web sites which publish the news. The news sites publish different kinds of information in different presentation forms [3]. News is a very dynamic kind of information, constantly updated. The news sites have to be interrogated frequently so as not to miss any of the news information. On the other hand, it is often possible to reach the old articles from the archival files available on their sites. The archival life is varied on the deferments sites. The updating frequency and the archival life for some news sites is presented in Table 1. This information, which we evaluated was confirmed by the sites administrators.

3.2 The Weblog Sites

A weblog, web log or simply a blog, is a web application, which contains periodic posts on a common webpage. It is a kind of online journal or diary frequently updated [1, 2].

Table 1. Updating news frequency and archival life.

Service news	Update	Archiving
French Google	about 20 min	30 days
Google	about 20 min	30 days
Voila actuality	every day	1 week
Voila news info	instantaneously	1 week
Yahoo!News	instantaneously	1 week
TF1 news	instantaneously	
News now	5 min	
CategoryNet	every day	never ending
CNN	instantaneously	
Company news	about 40 per day	2003, 2004, 2005 archived

Table 2. Index-database updating frequency.

Search Engine	Updating frequency
Google	4 weeks some pages are updating quasi daily
Yahoo!	3 weeks
All the Web	vary frequently publishes la date of the robots visit since 2004 index-data base together with Yahoo!
AltaVista	since 2004 index-data base together with Yahoo!

3.3 Search Engines

The Search Engines' response pages are the dynamic pages created on-line. The lifespan of the same response page (period when the Search Engine answer doesn't change) depends on the data retrieved from the Search Engine' index-database. It is evident that this time is correlated with the updating frequency of index-database.

Table 2 shows the examples of values of the Search Engines' index-database updating frequency.

4 Archiving

Dynamic documents can be printed, saved by the user or put to special caching and archiving systems. There are many Web applications, which store the current web image the Web (example Wayback Machine developed by The Internet Archive¹). These applications try to retrieve and save all of the visible Web [4]. It is evident that this task is very difficult. The WWW is enormous and it changes and

¹<http://www.archive.org/index.php>

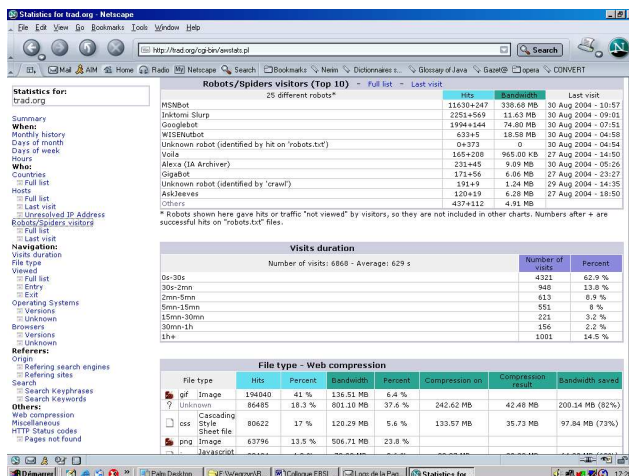


Figure 1. Visit Frequency of indexing robots.

grows very fast. An unfortunate side effect of this continual growth and dynamical modification is that it is impossible to save the totality of Web images. We have compared the data from the GoogleNews and BBC archives presented by the Wayback Machin with our statistical data (Table2, Table1). This comparison shows clearly that this archive is incomplete.

5 Statistical Evaluation

We have carried out the following statistical evaluation: index-databases updating frequency (for Search Engines and Meta Search Engines) and different statistical tests of the News sites and the Weblogs.

5.1 Search Engines

To estimate the updating frequency of index-databases we have analysed the differents logs files and we have calculated frequency of access to Search Engines carried out by different indexing robots [5, 6, 7]. Figure 1 shows the example of logs' data concerning the robots visits.

5.2 News sites and Web logs

We have carried out some statistical tests to evaluate the updating frequency (lifespan) [7] of News. The results showed the different behavior of interrogated sites (Figure 6a, Figure 6b, Tableau 3). We have analyzed four categories of sites:

- Sportstrategies the sport news service,
- News on the site of French television TF1,
- News from BBC site,
- Weblog site (Slashdot.org).

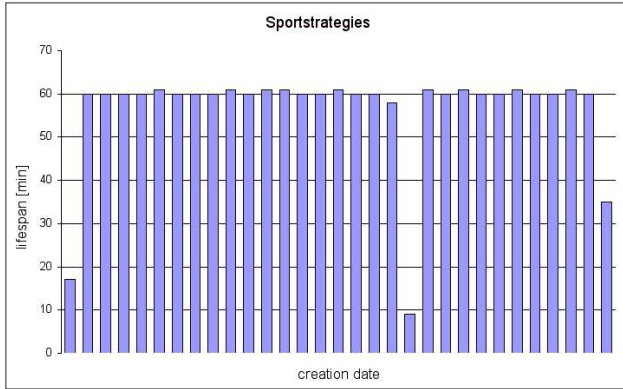
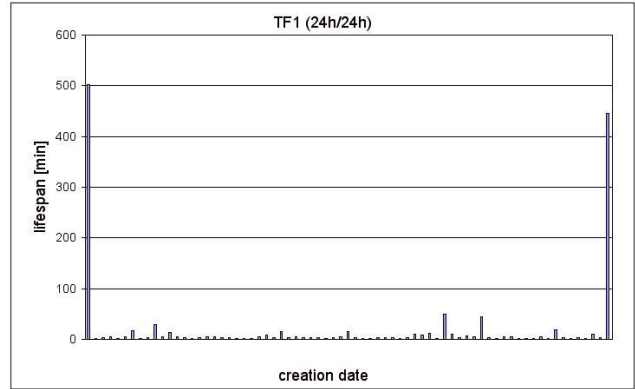


Figure 2. Sportstrategies: News lifespan.



a) News 24hours/24

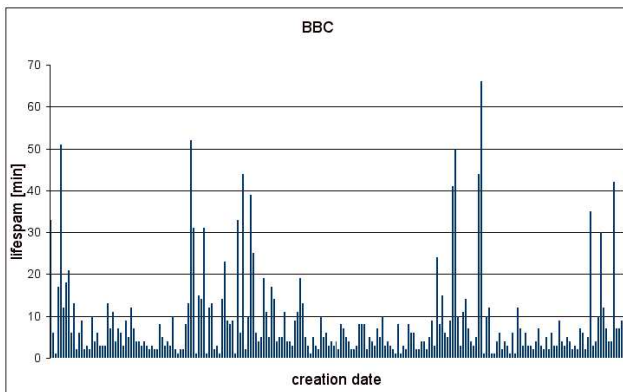
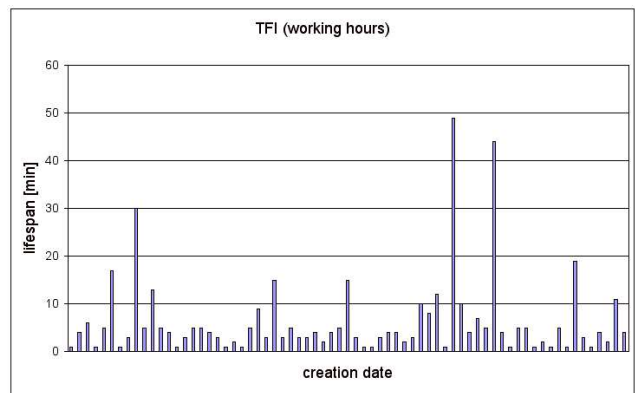


Figure 3. BBC: lifespan of the news.



b) News at working hours

Sportstrategies is an example of the very regular News site, with a constant update time (every hour : Figure 2).

BBC News is diffused online, the lifespan is very irregular because the information is updated instantaneously when present (Figure 3).

TFI News is updated frequently during the day. On the other hand there are no modifications by night. The lifespan of the News pages is very different in these two cases. We have presented it in two separated graphs. (Figure 4a et Figure 4b). Two high peaks in the extremes of the graph in Figure 4a correspond to the long period without any changes during the night.

Slashdot.org Weblog site, represents the last category of site. This collective weblog is one of the more popular blogs oriented on the Open Source. The data changes here very quickly, the new articles are diffused very often and the actual discussions continue without any break. The lifespan of these dynamically changed pages is extremely short (Figure

Figure 4. TF1 News: lifespan of the news.

5); the mean lifespan is equal to 77 sec. (Tableau 3).

Updating frequency values. In the next graphs (Figure 6) and Table3 comparatives of updating frequency values for some tested sites are presented. We have found the maximal and minimal values of the updating frequency and calculated the mean. The results confirm that the content of the news sites changes very often.

6 Conclusion

Dynamic documents don't exist in reality, they disappear from the computer memory directly after consultation. Their real lifespan is very short. The sites can be classified into different categories depending on the news-updating period; very regular -with a constant update time, irregular - information updated when present. The sites can be also classified into two categories depending on the refresh time; slow -with a refresh time greater then 10 minutes, fast

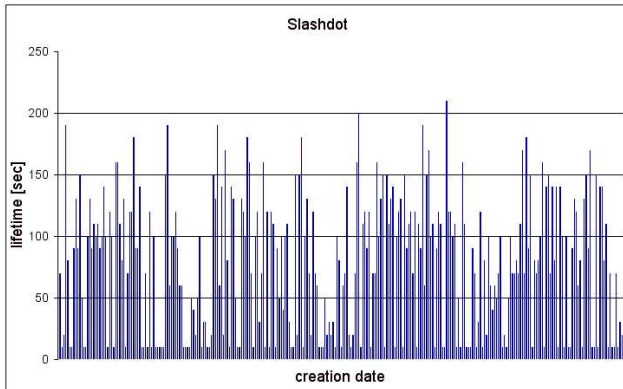
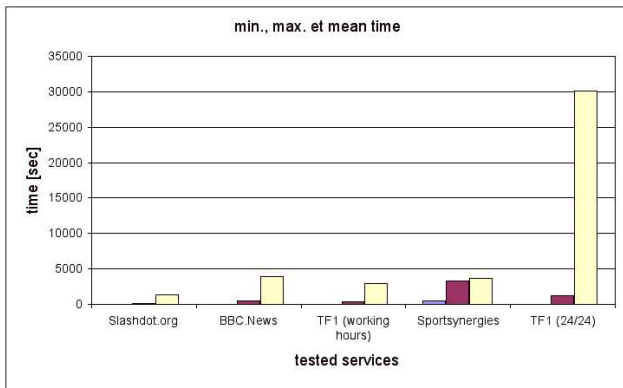
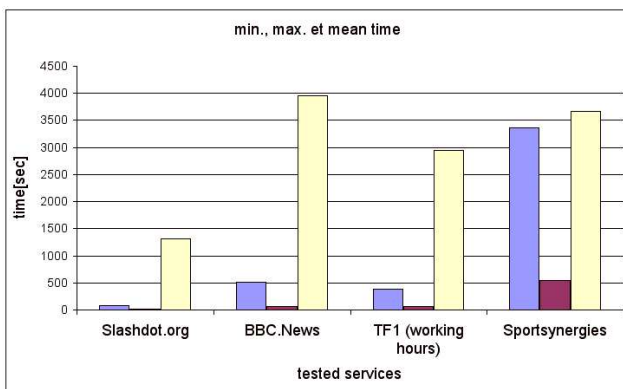


Figure 5. Slashdot lifespan of articles.



a) TF1: 24/24



b) TF1: working hours

Figure 6. Updating frequency.

Table 3. Updating frequency

tested site	lifespan		
	mean	min.	max.
Slashdot.org	77 sec	10 sec	22 min
BBC News	8,5 min	1 min	66 min
TF1 news (24/24)	19,5 min	1 min	502 min
TF1 News (working hours)	6,3 min	1 min	49 min
Sportsynergies	56 min	9 min	61 min

-information refresh even about 10 seconds. Some news sites present periodic activity: ex. the news site of the French television channel TF1 is updated only during working hours.

On the other hand, dynamic documents can be stored by special archiving systems and in fact, users can access them for a long time. Management of the archived dynamic document's lifespan is identical to that of static documents, because the dynamic documents are stored in the same way as static ones.

References

- [1] R. Blood. *The Weblog Handbook: Practical Advice on Creating and Maintaining your Blog*. 2002.
- [2] S. Booth. *C'est quoi un Weblog*. 2002.
- [3] A. Christophe. Chercher dans l'actualite recente ou les archives d'actualites francaises et internationale, on-line <http://c.asselin.free.fr>, 2004.
- [4] S. Lawrence. Online or invisible ? *Nature*, 411(687):521, Jan 2001.
- [5] K. Wegrzyn-Wolska. *Etude et realisation d'un meta-indexeur pour la recherche sur le Web de documents produits par l'administration francaise*. PhD thesis, Ecoles Superieures de Mines de Paris, DEC 2001.
- [6] K. Wegrzyn-Wolska. Fim-metaindexer: a meta-search engine purpose-bilt for the french civil service and the statistical classification and evaluation of the interrogated search engines using fim-metaindexer. In G. J.T.Yao, V.V.Raghvan, editor, *The Second International Workshop on Web-based Support Systems, In Conjunction with IEEE WIC ACM WIAT'04*, pages 163–170. Sainr Mary's University, Halifax, Canada, 2004.
- [7] K. Wegrzyn-Wolska. Le document numerique: une etoile filante dans l'espace documentaire. *Colloque EBSI-ENSSIB; Montreal 2004*, 2004.