Keeping Web Indices up-to-date

Marcel Dasen,
ETH Zürich, Computer Engineering and Networks Laboratory

dasen@tik.ee.ethz.ch

Erik Wilde,
ETH Zürich, Computer Engineering and Networks Laboratory

net.dret@dret.net

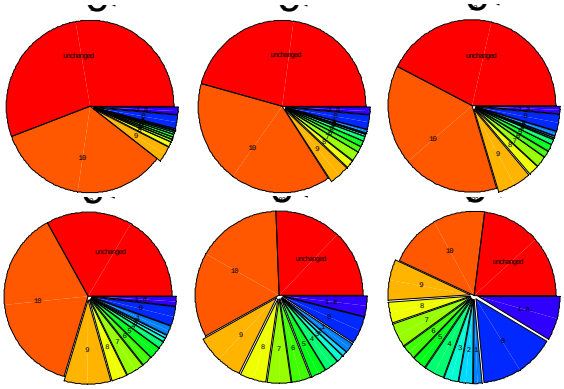

Figure 2: The change of documents on the Web. After 1 day, 7 days, 13 days, 1 month, 3 month and 6 month. Unchanged documents are bit-identical, 10 documents with no index relevant change, 9..1 are documents which have changed to a certain degree, the smaller the value the higher the change (N refers to the vector similarity is.N, e.g. 9 denotes a similarity of 0.9). "i.a." are documents which were inaccessible at the time.

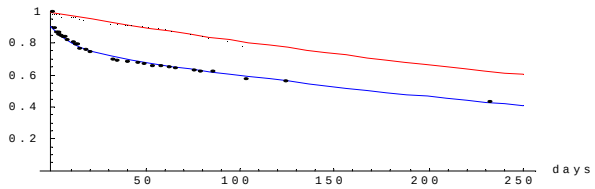

Figure 4: The fraction of documents without index relevant changes versus time. The blue graph shows the modelled change of a generic sample form the Web, while the red graph above shows change rate of documents out of the .edu top level domain.

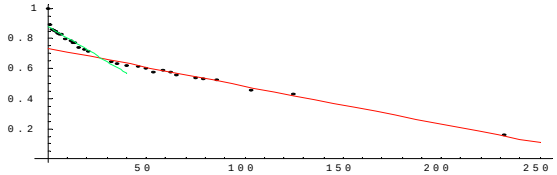

Figure 5: Dual Poisson modeling for the change rates of documents. In the first 20 days a different process must be assumed than thereafter.

Abstract

Search engines play a crucial role in the Web. Without search engines large parts of the Web becomes inaccessible for the majority of users. Search engines can make new and smaller sites accessible at low cost. Without them, other media, such as Television, would be needed to advertise the existence new site on the Web, only large commercial sites can follow this path. The Web would be endangered to become dominated by a few, well known sites. A crucial problem of search engines is to keep their index up-to-date. Especially if the index grows, the effort needed to update the index increases, since Web documents are dynamic and thus already stored data becomes obsolete. There have been various attempts to monitor the evolvement of the Web [1][2]. However, we believe, that change model used in prior work over-estimates the rate of change due to an inadequate change model. Our change model has been adapted from the information retrieval field to distinguish index relevant changes from irrelevant modifications in Web documents, e.g. simple spelling corrections or dynamic advertisement links. We have monitored multiple smaller collections of documents over a time period of six month to measure the documents change.

1. Index Relevant Changes

Not all changes in Web documents need to be index relevant. E.g. links in the documents might have been updated, some spelling has been improved or the document has been extended with more material of the same kind. Therefore, Web change estimations based on Bit identity, e.g. using checksum, typically over-estimate the change in documents [1][2].

We have applied a more refined change model, by using a well understood technique from information retrieval, the vector retrieval model [7][8]. In this model the frequency of occurrence of all words in a document form a vector describing this document. To account for the relative relevancy of words, the words are additionally weighted inversely to their appearance in documents (inverse document frequency). This model has been widely applied in digital library indexing and also in the context of the Web [3][4][5]. The change of two instances is calculated by forming the scalar product of their vectors.

The change of two documents $\cos(\phi)$ is calculated by the following equality in the vector retrieval model.

$$\dot{d} \bullet d' = |d||d'|\cos(\phi)$$

Multiple samples of 10K documents from the Web have been regularly revisited and the change to the original documents has been assessed using the vector retrieval model (Figure 2). The sample has been taken such, that statistics of the sample largely corresponds to the statistics of a 100 times larger sample of the Web. Each monitored samples included more than 500 domains predominantly from the .com top level domain.

2. A model for the change

From the data, the rate of change has been modelled. We have used dual Poisson processes (equation) for this purpose, one for the first few days and another for the time thereafter. Using two processes was suggested by the fact the rate of change is significantly faster in the first days of the life span of a document. We suspect that there is a class of documents which is constantly being worked on, e.g. a developing news story. Poisson process is used to model random events, which occur independent at a fixed rate over time. We believe that Poisson is a good model for evolvement of a large set of pages. It can safely be assumed that pages change independent of each other [2].

It can be seen that different domains evolve differently. An exterm case are the sample collection of documents from .edu have a much higher persistence, than e.g. documents in .com domain.

The model below has been fitted to the data of the experiment (Figure 4: blue line).

$$F(t) = \upsilon_1 e^{\lambda_1 \cdot t} + \upsilon_2 e^{\lambda_2 \cdot t}$$

We have found the half life period for the first process to be 65 days and 276 days for the second process. The respective half life period for .edu domain are 286 and 334 days. This implies that a single process adequately models .edu domain.

Furthermore our investigation shows a tendency of smaller documents to change faster (Figure 3). This fact could be exploited to improve scheduling of index updates of search-engines.

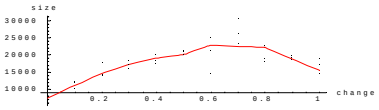

Figure 3: Document change versus size

3. Conclusion

We have found that, although, a large portion of the monitored documents did evolve rapidly if monitored for bit identity, many of those changes are not relevant to the index of a search engine. Users of the index would thus still find the document and experience high precision of their query results even though the source document had changed. We have found that a single Poisson processes does not fit well the more dynamic part of the Web such as the .com domain. However, the half life time of documents in the longer turn has been consistently found to be around 300 days with not too much variation between the .edu (334 days) and .com (276 days) top level domains.

4. References

[1] B. E. Brewington, G. Cybenko, Keeping Up with the Changing Web, IEEE Computer, p. 52 - 58, May 2000

[2] J. Cho, H. Garcia-Molina, The Evolution of the Web and Implications for an Incremental Crawler, Proceedings of 26th International Conference on Very Large Databases (VLDB), September 2000

[3] Jeffrey Dean, Monika Henzinger, Finding related web pages in the World Wide Web, Proceedings of th 8th International World wide web Conference, Toronto Canada, Elsevier Science B.V., p. 389-401, 1999.

[4] Yanhong Li, Towards a qualitative Search Engine, IEEE Internet Computing, Vol. 2, No. 4, p. 24-33, 1998

[5] J. Cho, H. Garcia-Molina, L. Page, Efficient Crawling Through URL Ordering, Proceedings of the 7th International WWW Conference, Brisbane, Australia, p.161-172, 1998

[6] Kihong Park, Walter Willinger, Self-Similar Network Traffic and Performance Evaluation, John Wiley & Sons, New York, ISBN 0-471-31974-0, 2000

[7] C. J. van Rijsbergen, Information Retrieval, Butterworths, London, Second Edition, 1979.

[8] G. Salton, C. Buckley. Improving Retrieval Performance By Relevance. Journal of the American Society for Information Science, Vol. 41, No. 4, p. 228-297, 1990.