

Sic Transit Gloria Telae: Towards an Understanding of the Web's Decay

Reviewed By
Chris Wojciechowski

1. Introduction

Sic Transit Gloria Telae: Towards an Understanding of the Web's Decay was written with the intent of developing an algorithm to determine the decay, or obsolescence, of a web page. Decay, as described in the context of this paper, is a measure of the lack of maintenance that a web page receives over time to provide current and relevant data. This paper will measure a page's relative data by the integrity of its out links. Decaying web pages affect the validity and reliability of information on the Internet, and make information retrieval more difficult for users. In determining the decay of a link, the authors define what they call, dead links. A dead link is defined as a link which responds with either a 404 or soft-404 response. A soft-404 response is a response by the web-server which states that a page is dead but does not offer a direct 404 response. Determining which pages respond with a 404 or soft-404 allows the authors to create an algorithm which can score a web page with a decay value based off of a calculation of the pages it points to. The authors present the problem and provide a solution which is useful for determining patterns of decay in various types of websites ranging from commercial to scientific.

2. Method for Detecting a Soft-404 Page

The paper defines a soft-404 page as a page which states that the given URL has not been found, but does not provide a 404 response to the client. The problem with this practice is that it can be computationally more difficult to determine if a URL is dead, because a soft-404 page does not return a definite 404 response. To work around this special case, the authors have presented a method and algorithm to determine if a soft-404 page has been reached versus having reached actual content. Their method is well defined and makes sense as a broad application to solving this problem. The authors search for a web page, r , which has a low probability of existing on the web server.

Their method of selecting a web page consisting of 25 random letters is valid, as it would be very rare to find a page in the directory of that server with a similar name. This is compared under the same directory as the web page in question to acknowledge the possibility that the current directory could be handled by a different server. The behavior of the random URL is compared with the behavior of the test URL. The random URL can have two responses: a 404 response, or a valid web page (soft-404). In the case of the random URL producing a web page, the content and path taken is then compared to the original test URL. The authors have mapped out each possible outcome and have come to very reasonable assumptions to the varying output expected from the algorithm. They also acknowledge that the algorithm is not flawless, and admit that they make the assumption that a root of a website could never be a soft-404. This assumption can affect the validity of their results, however if used in a controlled environment on valid root pages, then this is of less concern and should not overshadow the benefits of the algorithm.

3. Computing Page Decay

The authors present a process for computing decay which is recursive in nature. It computes a score based on the traversal of out-links by a random surfer. A set of pages is traversed to determine dead pages. Of this set the pages are randomly traversed in a tree-like manner to determine a decay value for each page. As the surfer traverses the link tree, the effect that a dead link has on the decay value of the page decreases exponentially. Repeating the random surfer test 300 times to ensure a good distribution of random walks through the tree ensures. Their method of determining a decay value is logical and gives a good perspective on how a page is decaying by the level the pages it points to are decaying. This is a useful calculation as it can represent a page that is dead, is nearing death, or may not be maintained as often.

Results and Observations

Their results are consistent with their claims and methods for determining page decay. The authors tested their methods on three fairly distinct data sets. Each result set exhibited different qualities of link decay. In the case of the Yahoo! data set they pulled 30 nodes from the Yahoo! ontology and pulled the archived information from the previous 80 months. The results appear to be consistent with Yahoo!'s method of page organization and maintenance. They show close to zero dead pages in recent months, which points to good link analysis and management. The results however show a flat-lining of page decay towards recent months. This is possibly because of the way that Yahoo! performs link management, in that they do not search far within URLs that they link to, thus raising the possibility that the pages they point to may not have been maintained in a longer period of time.

The results obtained from the FAQ data set were also interesting because the trend towards page decay was larger in the middle of the time period. This is representative of the unmaintained state that the FAQ is kept in, as it is hand maintained. A peak was reached in 1998, and from there onward, the decay of pages slowly declines. What should be mentioned in the paper is why this curve vaguely resembles an upside-down bell curve. The FAQ data which is newer would obviously have fewer dead links and page decay, but it is interesting to note that the older data also has a lower amount of dead links and page decay. A possible reason for this could be that FAQ articles that are tested, reliable, and deal with a general or time-invariant issue, link to FAQ articles and pages that are similar. The pages in the middle fall into a more volatile state as they experience page decay and are either forgotten or maintained to point to more reliable sources.

The WWW data set shows a decline in page decay on more recent years, as papers and their references are more current in the recent years. The authors made an observation that the papers from earlier conferences are more likely to be dead as well as point to pages that are dead.[1] These results are consistent with trends regarding these papers.

Recommendations and Conclusion

The authors have presented the idea that page decay could be determined from more extensive link analysis, and have provided an adequate algorithm to do so. The results show that page decay does exist in the con-

text of this paper's claims that unmaintained links are representative of page decay. The results show that there is no specific model for a general website, but that depending on the website being analyzed, there are varying approaches to understanding the data and make decisions regarding the relevance and validity of data on a time based scale. They have provided a method that can be applied to many existing algorithms and practices pertaining to ranking, maintaining taxonomies, crawling, and determining topic popularity. The algorithm could be improved to include other decay features such as dates and slang, however as a standalone algorithm the current implementation works well to track and analyze data decay.

References

- [1] Bar-Yosseff et. al, "Sic Transit Gloria Telae: Towards an Understanding of the Web's Decay" , 2004.