Critique on “What’s New on the Web? The Evolution of the Web from a Search Engine Perspective”

Authors-Alexandros Ntoulas, Junghoo Cho and Christopher Olston

Critique by-Pragathi Reddy Mallu, Vivek Nuthalapati

PAPER SUMMARY

This paper deals with the evolution of web from search engine perspective and what is new on the web. The authors approach is based on evolution of link structure, rate of creation of new pages and new distinct content in web and the rate of change of existing pages. The authors has collected weekly snapshots of 150 websites to do the analyses. Their findings indicate that there is a rapid turnover rate in webpages. Pages that change by great amount in one week will continue to experience the great change in next weeks. Similarly pages that experience a little change in one week will experience little change in further weeks.

STRENGTHS

- The title of the article is clear. The abstract is specific. The purpose of the article is very clear in the introduction. The discussion in the paper is very relevant.
- The website selection and the sampling of data collection was done quite extensively covering a large number of features robustly.
- The comparative analysis of the TF. IDF Cosine Distance, word count and shingles was very exhaustive.
- The paper gave a good justification for correlation of every feature.
- The authors considered birth, death rate and also replacement rate which is a very good factor to analyze the new content in the web.
- The authors extracted all the links from every snapshot and measured how many of the links from the first snapshot existed in the subsequent snapshots to analyze the overall link structure change which gave more accuracy to results.
- The authors also tried to predict future changes by considering frequency of change and degree of change.

WEAKNESS

- The paper considers only the top order pages on the internet and generalizes the conclusions for all the pages.
- Coming to the newness of the content. Firstly a very vague definition of newness of content on the web has been used. Even the changes in the URL is referred as new content. This definition increases the percentage of new content on the web.
- The pages considered for this experiment were the pages that have the highest ratings in all categories. But, the OCLC analysis of the number of pages created and destroyed per year is for all types of pages which cannot be generalized.
- Categorical generalization is also not justifiable. Although Matrimony websites and even management websites fall in the same category of “Life Events”. Generalizing the rate of change for these 2 websites would not be right. As the changes in these websites are totally independent of each other and are dependent on a very few common features.

- The paper considers only top web sites for link structure evolution which might not hold true in case of lower ranked websites. The lower ranked websites usually put it more change and more efforts to increase their rankings. The lower ranked websites might be eventually having more changes in link structure than the higher ranked sites.
- The paper emphasizes more on the TF.IDF Cosine Distance than on the word count. But it has been mentioned in many papers that both the metric have their merits and demerits.
- The degree of change and frequency of change cannot be related as per the paper. But, a relation can be formulated at least on categorical basis, if not on all the categories combined.

CONCLUSION

The main important finding is that existing pages are being removed from the web and replaced by new pages at a very rapid rate, but new pages are having the content mostly from existing pages. Altogether, this can be considered as very good experiment conducted by authors to know what’s new on the web.