

Analysing the Importance of different weight factors on Weighted PageRank Algorithm

Isha Potnis
Computer Science
and Electrical Engineering,
University of Maryland,
Baltimore County, Maryland 21227
Email: ipotnis1@umbc.edu

Akanksha Bhosale
Computer Science
and Electrical Engineering,
University of Maryland,
Baltimore County, Maryland 21227
Email: akanksh1@umbc.edu

Shantanu Sengupta
Computer Science
and Electrical Engineering,
University of Maryland,
Baltimore County, Maryland 21227
Email: ssen1@umbc.edu

Abstract—**Lorem ipsum dolor sit amet, consectetur adipiscing elit.** Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdier mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Keywords—*PageRank, Weighted PageRank, Web Structure Mining, HITS, Web Mining, SEO*

I. INTRODUCTION

PageRank Algorithm was introduced by Sergey Brin and Larry Page as a novel method of ordering pages in the World Wide Web by assigning ranks to them, which indicated their importance. The web was assumed to be a graph, where the pages were analogous to nodes and the hyperlinks were analogous to weighted edges. The algorithm, in a nutshell, worked by recursively calculating the rank of the page on the basis of its previous rank and the rank metrics of all the hyperlinks that connected to this page. Though this system was one of the first algorithms to rank web pages, it was also an easy system to manipulate because inbound links from falsely influenced PageRank misrepresented the rank of a webpage. In order to obtain better and relevant search results, Weighted Page rank was introduced as an extension to simple Pagerank, which considered the inbound and outbound links of a webpage and distributed the rankings based on the popularity of the web page.

Currently, Google Hummingbird focusses on more than 200 factors for ranking webpages and each factor has a different amount of influence on the rankings. It also introduced conversational search that worked on determining the context of the searches being made. With that in mind, our intent for this project is to identify an ideal distribution of different factors that produce the most relevant results for a page. We also intend to explore different scoring rubric for rating the results obtained.

II. BACKGROUND

Background is what is needed to understand your report. It is not new research.

Need to include formulas and basic Page Rank and Weighted Page Rank Algorithm

III. PREVIOUS WORK

Gupta et al.[1] suggests that the Weighted PageRank Algorithm constitutes an average of the weights of four different attributes such as content, user preference, history and popularity of the page. Whereas, we would like to analyse the effect of assigning different weights to different attributes.

The findings for this paper were obtained as a result of a user study where the users were graduate students with an objective to visit at most a collection of 50 web pages in the web and search for a total of 12 queries. On the basis of this study, a discrete precision value was assigned to each page that indicated its relevance. We would like to undertake a wider scope of users for the user study on fewer but diverse topics and have the users rank on multiple attributes rather than just relevance.

IV. METHODS

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdier mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

V. RESULTS

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdier mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue,

a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

VI. DISCUSSION

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

VII. OPEN PROBLEMS

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

VIII. CONCLUSION

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

ACKNOWLEDGMENT

The authors would like to thank...

REFERENCES

- [1] H. Kopka and P. W. Daly, *A Guide to L^AT_EX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

APPENDIX A ADDITIONAL NOTES

A. Progress

We have implemented a working code for the weighted Pagerank algorithm in Java based on the paper Weighted PageRank Algorithm[2] by Wenpu Xing and Ali Ghorbani. We have created classes for Pages and Links. We have used jsoup, a Java open source library that parses HTML Files to find inbound and outbound links. We have tested this code to accurately rank a series of webpages.

For the remaining half of the project, we intend to create a distribution of weights for different factors and produce a ranking system. Using the rankings from these different distributions and the original Weighted Pagerank, we intend to perform a user study to understand if the ranking system so produced, allowed for returning a more relevant set of webpages.

B. Difficulties

One of the major challenges for this project is how to create different distributions for the different factors and figuring out if there is an ideal solution to this problem. The second major challenge is to figure out how to assign weights to the different factors and on what basis will one factor be weighted more than other and by what scale.

C. Plan to Overcome them

For the scope of this project, we intend to deal with both the problems by researching the preferences that are assigned to different factors by Google Hummingbird and creating a sufficiently good rating system for the modified weighted Pagerank algorithm.

D. Significant changes to the proposal

E. Revised Schedule

1) Phase 1 - 10/01/17- 10/08/17: Formulating Interesting Research question and determining its impact on society

2) Phase 2 - 10/09/17-10/18/17: Required modifications in project proposal and initial development setup

3) Phase 3 - 10/19/17-10/25/17: Development and Experimentation in determining dominant weight factor

4) Phase 4 - 10/26/17-10/27/17: Mid evaluation and Deliver progress report

5) Phase 5 - 10/28/17-11/22/17: Final Testing, Inference and Code cleanup

6) Phase 6 - 11/23/17-12/4/17: Documentation and Final Presentation submission