Proc. of Int. Conf. on Intelligence Computing & Information Technology , ICIT

Efficient and Effective PageRank for Custom Search Engines using In-Page Semantic Ranking

Leena Giri G[1], Manjula S. H[2] and Venugopal K. R[3]
[1-3]Department of Computer Science and Engineering, University Visvesvaraya College of Engineering,
Bangalore University, Bangalore – 560 00, India.
Email: leenagiri.g@dr-ait.org

***Abstract*—Custom search engines are search engines that websites offer to search through the content of their own website. The link structure that we observe on the web is not similar to the link structure of a single website. In an individual website, the many links between pages aid easy navigation and increase accessibility while not indicating the quality or the importance of the pages they point to. Using In-Page Semantic Ranking to filter the links results in a much smaller network graph with less edges to process than the original graph and is extremely efficient in getting rid of dubious links which are not useful or calculating the quality of pages since In-Page Semantic Ranking allows tuning based on the semantic structure of a website.**

***Index Terms*— Semantic search, ranking, In-Page Semantic Ranking.**

I. INTRODUCTION

Snippets are textual information present on web pages and are the basic foundation for the index building process of most search engines. Previous works in web page ranking use techniques for ranking web pages while considering snippets to be plain text, without exploiting the rich semantic information available from the HTML content. An HTML document not only contains snippets, but also describes semantics of that snippet and the structure of the entire document. A snippet ranking algorithm which uses semantic HTML tags, that embeds the snippet, and the underlying structure of the web page for assigning ranks is developed. This allows a more precise ranking compared to considering snippets as plain text.

A. Motivation

PageRank is a link analysis algorithm which assigns a numerical weight to each document on the web based on the graph created from all web pages considering web pages as nodes and hyperlinks as edges. While PageRank is effective in ranking the importance of web pages on the world wide web, there are limitations. PageRank is ineffective when used with the pages of an individual website. This accounts from the fact that there are links within a website that is used to aid navigation and accessibility which spams the graph of the website with dubious edges (corresponding to links) which do not contribute to the importance of a webpage. This makes PageRank less effective on individual websites and hence the custom search engines that depend on PageRank are less efficient in ranking and retrieving the relevant web pages.

Grenze ID: 02.ICIT.2018.2.508

B. Contribution

In-Page Semantic Ranking assigns ranks to the text snippets of a web page based on the semantic structure and hierarchy of a web page. A modified version of In-Page Semantic Ranking which ranks only hyperlinks on a webpage instead of text snippets to rank hyperlinks is considered. We propose a dynamic threshold method to filter hyperlinks that should be used for building the web graph to be used with PageRank. It would allow individual websites to get rid of links that are of very low quality and help in improving the precision of Page Rank. It is shown that using Page Rank without filtering hyperlinks, with modified In-Page Semantic Ranking, is less efficient when compared to filtering hyperlinks with modified In-Page Semantic Ranking.

II. METHOD

In-Page Semantic Ranking algorithm [1] provides a quantitative approach to ranking snippets based on the semantic tags they are embedded in and the position of the snippet in the parse tree hierarchy. The algorithm builds a foundation for snippet ranking which can be further used by clustering or page ranking algorithms to increase the precision of their results. In-page semantic ranking algorithm is a recursive ranking algorithm which traverses the HTML parse tree until a text node is reached, wherein the snippet is extracted along with its calculated rank. The algorithm is invoked starting from <body> tag as the root node. level( ) is a routine which matches the given node with the rule sets considered and returns the matching level. Since the rule sets are of the same syntax as that of CSS Selectors, level( ) is simplified implementation of a CSS Selector matching procedure.

In-Page Semantic Ranking algorithm has the following steps:

1. If Doctype is <!DOCTYPE html> which indicates the page used HTML 5, use the rule sets in Table 1 without any adjustments. If the page has used any other Doctype, then the rule sets matching HTML 5 sectioning elements will be removed before invoking the algorithm.

2. Extract the snippet from the Title tag < title > and give it the highest preference. The content in the title tag is the most important snippet for the page and it is given the highest rank over all other snippets on the page.

3. Invoke the In-Page Semantic Ranking algorithm rank beginning from the <body> node of the page. Snippet set with the snippets and their associated ranks are obtained on the completion of the execution of the algorithm.

4. The final snippet set is obtained after the algorithm terminates. A snippet set is a collection of pairs - a snippet and its rank as given in Table I.

In-Page Semantic Ranking aids in removing links which are insignificant and which could result in dampening the effect of PageRank algorithm. Filtering the links using In-Page Semantic Ranking results in much smaller network graph with less edges to process than the original graph and is highly efficient in getting rid of dubious links which are not useful for calculating the quality of pages since In-Page Semantic Ranking allows tuning based on the semantic structure of a website. This gets rid of links that are of very low quality and help in improving the precision of PageRank on a single website. This is because the links and backlinks which are significant in calculating PageRank behave differently for the web as a whole in contrast with the links and backlinks of a single website.

The ranking algorithm provides a quantitative approach to rank snippets based on the semantic tags which embed them and the position of the snippet in the parse tree hierarchy. The algorithm builds a foundation for snippet ranking which can be further used by clustering or page ranking algorithms to increase the precision of their results. A modified In-Page Semantic Ranking was used which was used to rank snippets which are exclusively wrapped by anchor tags only. Once the semantic score for all the hyperlinks are obtained, the hyperlinks are ranked based on the following dynamic threshold:

$$T(r) = li/count(li) = topCount,$$

which is the topCount unique ranks of all hyperlinks. Then,

$$thresholdRank = min(Tr).$$

TABLE I. SNIPPET AND IT'S RANK

Snippet	Rank
WebPage Title	**Highest Prominence** (title)
Page Heading	1.05
Section Heading	0.90
Navigation Heading	0.65
Article Heading 1	0.61
Article Heading 2	0.61
Paragraph Snippet 1	0.47
Paragraph Snippet 2	0.47
Aside Heading	0.33
Navigation Link 1	0.28
Navigation Link 2	0.28
Footer Paragraph	0.25
Link Text 1	0.13
Link Text 2	0.13
Link Text 3	0.13

III. EXPERIMENTAL RESULTS

The algorithm was tested on a set of New York Times webpages. In-Page Semantic Ranking permits flexible movement of RuleSets among the three proposed ranking levels based on the semantic structure of the website. A hyperlink is used to construct the webpage for usage with PageRank only if *ri* for a link *li* is greater than threshold Rank.

TABLE II. SNIPPET AND IT'S RANK

	Without Filtering	With Filtering
No. of Nodes	14167	**324595**
No. of Edges	14167	589886

The filtered hyperlinks are exported as an edge list. A web graph of pages and hyperlinks was created from the edge list and were considered to run pagerank method of networkX python library. The reduction in the number of edges is by 50% as shown in Table 2. Once the PageRank was obtained, it was used as a rankfile for solr to run the search.

A few Accepted Hyperlinks

u'https://www.nytimes.com/store/art.html' (4387667856)
u'https://www.nytimes.com/store/art/most-popular-art.html' (4387750128)
u'https://www.nytimes.com/store/art/paintings-prints.html' (4387751952)

A few Discarded Hyperlinks

u'https://www.nytimes.com/store/store-policies/' (4388557952)
u'https://www.nytimes.com/store/sports/autographed/football.html' (4388009760)

u'https://www.nytimes.com/store/image-licensing/' (4388559104)

IV. Conclusions

It is observed that the filtering is able to filter out pages for the e-commerce cart page, the faq page and the policies and licensing page which are clearly not of importance to a search engine where users mainly search for content. But the filtering also discards the page related to sports related content and can be tested.
have been defined in the abstract. Do not use abbreviations in the title unless they are unavoidable.

Acknowledgment

The authors wish to thank Praveen Gowda for his valuable inputs.

References

[1] Leena Giri G., Praveen Gowda I. V., Manjula S. H., Venugopal K. R., " In-Page Semantic Ranking of Snippets for WebPages," In *Proceedings of the Sixth International Conference on Advances in Computing, Control and Telecommunication Technologies(ACT-2015)*, De Gruyter, pages 279-283, Trivandrum, India, August 2015.

[2] http://journals.sagepub.com/doi/abs/10.1177/016555150102700605

[3] http://infolab.stanford.edu/pub/papers/google.pdf

[4] https://en.wikipedia.org/wiki/PageRank

[5] https://networkx.github.io

[6] Apostolos Kritikopoulos, Martha Sideri, Iraklis Varlamis, "Wordrank: A Method for Ranking Web Pages Based on Content Similarity," *24th British National Conference on Databases (BNCOD'07) IEEE Computer Society*, 0-7695-2912-7/07.

[7] Po-Hsiang Wang, Jung-Ying Wang, Hahn-Ming Lee, "QueryFind: Search Ranking Based on Users' Feedback and Expert's Agreement," Proceedings *of the 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'04)*, pp 299-304.

[8] Lara Srour, Ayman Kayssi Ali, Chehab, "Personalized Web Page Ranking Using Trust and Similarity," *IEEE/ACS International Conference on Computer Systems and Applications, 2007(AICCSA '07)* pp 454-457.