International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 12 Dec 2021, pp: 1083-1085 www.ijaem.net ISSN: 2395-5252

Search Engine: Unexplored Google Hilltop Algorithm

Miss Gauri Banore, Prof. A.A. Chinchamalatpure
ME Final year CSE, Dr. Sau. K.G.I.E.T. Amravati, Maharashtra
Prof, Dr. Sau. K.G.I.E.T. Amravati, Maharashtra

Submitted: 05-12-2021 Revised: 17-12-2021 Accepted: 20-12-2021

ABSTRACT — This paper, propose a novel ranking scheme for broad queries that places the most authoritative pages on the query topic at the top of the ranking. This algorithm operates on a special index of "expert documents." These are a subset of the pages on the WWW identified as directories of links to non-affiliated sources on specific topics. Results are ranked based on the match between the query and relevant descriptive text for hyperlinks on expert pages pointing to a given result page. We present a prototype search engine that implements this ranking scheme and discuss its performance. In response to a query a search engine returns a ranked list of documents. If the query is broad (i.e., it matches many documents) then the returned list is usually too long to view fully. Studies show that users usually look at only the top 10 to 20 results.

I. INTRODUCTION

When searching the WWW broad queries tend to produce a large result set. This set is hard to rank based on content alone, since the quality and "authoritativeness" of a page (namely, a measure of how authoritative the page is on the subject) cannot be assessed solely by analyzing its content. In traditional information retrieval we make the assumption that the articles in the corpus originate from a reputable source and all words found in an article were intended for the reader. These assumptions do not hold on the WWW since content is authored by sources of varying quality and words are often added indiscriminately to boost the page's ranking. For example, some pages are created to purposefully mislead search engines, and are known popularly as "spam" pages. The most virulent of spam techniques involves deliberately returning someone else's popular page to search engine robots instead of the actual page, to steal their traffic. Even when there is no intention to mislead search engines, the WWW tends to be crowded with information on topics popular with users. Consequently, for broad queries keyword matching seems inadequate.

II. RELATED WORK

Three approaches to improve the authoritativeness of ranked results have been taken in the past:

1.Ranking Based on Human Classification**:** Human editors have been used by companies such as Yahoo! and Mining Company to manually associate a set of categories and keywords with a subset of documents on the web. These are then matched against the user's query to return valid matches. The trouble with this approach is that: (a) it is slow and can only be applied to a small number of pages, and (b) often the keywords and classifications assigned by the human judges are inadequate or incomplete. Given the rate at which the WWW is growing and the wide variation in queries this is not a comprehensive solution.

2.Ranking Based on Usage Information**:** Some services such as Direct-Hit collect information on: (a) the queries individual users submit to search services and (b) the pages they look at subsequently and the time spent on each page. This information is used to return pages that most users visit after deploying the given query. For this technique to succeed a large amount of data needs to be collected for each query. Thus, the potential set of queries on which this technique applies is small. Also, this technique is open to spamming.

3.Ranking Based on Connectivity**:** This approach involves analyzing the hyperlinks between pages on the web on the assumption that: (a) pages on the topic link to each other, and (b) authoritative pages tend to point to other authoritative pages.[1][5]

III. PAGE RANK ALGOTITHM

Page rank algorithm are used before hilltop algorithm. Page rank is an algorithm to rank based on assumption of usage information. It computes a query independent authority score for every page on the web and uses this score to rank the result set. Page rank is query independent it cannot by itself distinguish between pages that are authoritative in general and pages that are authoritative on the query topic in

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 12 Dec 2021, pp: 1083-1085 www.ijaem.net ISSN: 2395-5252

particular a website i.e., authoritative in general may contain a page that matches a certain query but, in this algorithm not an authority on the topic of the query.

Drawback of Page Rank-: A problem is that the topic distillation is that computing the subgraph of www. Which is the query topic being hard to do in real time. This approach can fail because it is dependent on the comparativeness of the selected set of success.[3]

IV. NEED OF HILLTOP ALGORITHM

User usually expect the perfect result and result is shown firstly at the top list. This algorithm provides top results at firstly. In this algorithm first computes a list of the most relevant expect on the query topic and then identify relevant links within the selected set of experts and follow them to identify target web pages. Most important is that when such a pool of experts is not available, then hilltop provides no results. Thus, Hilltop is tuned for result accuracy and not coverage.

V. HOW HILLTOP ALGORITHM WORK

Hilltop algorithm computes the Expert Score and Target Score.

1)Expert Score: - Thus, we compute the score of an expert as a 3-tuple of the form (S_0, S_1, S_2). Let k be the number of terms in the input query, q. The component S_i of the score is computed by considering only key phrases that contain precisely k - i of the query terms. E.g., S_0 is the score computed from phrases containing all the query terms.[4]

S_i =SUM $_{\{\text{key phrases p with k - i query terms}\}}$ Level-Score(p) * Fullness-Factor (p, q)

- If m <= 2, Fullness-Factor (p, q) = 1
- If m > 2, Fullness-Factor (p, q) = 1 - (m - 2) / plen

Our goal is to prefer experts that match all of the query keywords over experts that match all but one of the keywords, and so on. Hence, we rank experts first by S_0. We break ties by S_1 and further ties by S_2. The score of each expert is converted to a scalar by the weighted summation of the three components:

Expert Score = $2^{32} * S_0 + 2^{16} * S_1 + S_2$.

2)Target Score: -The target score T is computed in three steps:

1. For every expert E that points to target T we draw a directed edge (E,T). Consider the following "qualification" relationship between key phrases and edges:
- The title phrase qualifies all edges coming out of the expert
- A heading qualifies all edges whose corresponding hyperlinks occur in the document after the given heading and before the next heading of equal or greater importance.
- A hyperlink's anchor text qualifies the edge corresponding to the hyperlink.

For each query keyword w, let occ(w, T) be the number of distinct key phrases in E that contain w and qualify the edge (E,T). We define an "edge score" for the edge (E,T) represented by Edge Score(E,T), which is computed thus:

- If occ(w, T) is 0 for any query keyword then the Edge Score(E,T) = 0.
- Otherwise, Edge-Score(E,T)=Expert-Score(E) * Sum$_{\{\text{query keywords w}\}}$ occ(w, T)

2. We next check for affiliations between expert pages that point to the same target. If two affiliated experts have edges to the same target T, we then discard one of the two edges. Specifically, we discard the edge which has the lower Edge Score of the two.

3. To compute the Target Score of a target we sum the Score of all edge's incident on it.

The list of targets is ranked by Target Score. Optionally, this list can be filtered by testing if the query keywords are present in the targets. Optionally, we can match the query keywords against each target to compute a Match Score using content analysis, and combine the Target Score with the Match Score before ranking the targets. [6][7]

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 12 Dec 2021, pp: 1083-1085 www.ijaem.net ISSN: 2395-5252

Figure 1. Hilltop Ranking for the Query: "jobs"

VI. BENEFITS OF HILLTOP ALGORITHM

1) Quick result is available.
2) Thus, the results shown on the top of the list.
3) Hilltop Algorithm first calculate the target score and expert score then provide the result.
4) In this experiment Hilltop Algorithm classified 2.5 million pages over the 140 million pages.

VII. CONCLUSION

Hilltop Algorithm generates a lot of target pages which are likely to be very authoritative pages on the topic of the query. In computing the usefulness of a target page from the hyperlinks pointing to it, we only consider links originating from pages that see to be experts. In blind evaluation we found that hilltop deliver a high level of relevance given broad queries and performs comparably engines tested.

REFERENCES

[1]. J. Kleinberg. Authoritative sources in a hyperlinked environment. To appear in the Journal of the ACM, 1999. **http://www.cs.cornell.edu/home/kleinber/auth.ps**

[2]. S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. **http://decweb.ethz.ch/WWW7/1898/com1898.htm**

[3]. S. Chakrabarti, M. van den Berg and B. Dom. Focused crawling: A new approach to topic-specific Web resource discovery.. **http://www.cs.berkeley.edu/~soumen/doc/www99focus/html/**

[4]. K. Bharat and M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. **ftp://ftp.digital.com/pub/DEC/SRC/publications/monika/sigir98.pdf**.

[5]. Oliver A. McBryan. GENVL and WWWW: Tools for Taming the Web. **http://www.cs.colorado.edu/home/mcbryan/mypapers/www94.ps**

[6]. Krishna Bharat, George A. Mihaila:HillTop Algorithm at a glance.www.cs.toronto.edu/~georgem/**hilltop/**

[7]. S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In WWW Conference, volume 7, 1998.