ISSN (Online) : 2348 - 2001

International Refereed Journal of Reviews and Research

Volume 2 Issue 1 January-February 2014

International Manuscript ID : 23482001V2I101022014-30

HITS ALGORITHM - AN EFFECTIVE LINK INVESTIGATION APPROACH

Dr. Miji C. Kan
Beijing University of Technology
China

ABSTRACT

With the rapid increase in internet technology, users get easily confused in large hyper text structure. Providing the relevant information to user is primary goal of the website owner. In order to achieve this goal, they use the concept of web mining. Web mining is used to categorize users and pages by analysing the users behaviour, the content of the pages, and the order of the URLs that tend to be accessed in order. Web structure mining plays very important role in this approach. Its defined as the process of analysing the structure of hyperlink using graph theory. There are many proposed algorithms for web structure mining such as PageRank (PR), Weighted PageRank (WPR), and Hyperlink-Induced Topic Search (HITS) etc. Hyperlink-Induced Topic Search or simply HITS is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg. It was a precursor to PageRank. The idea behind Hubs and Authorities stemmed from a particular insight into the creation of web pages when the Internet was originally forming; that is, certain web pages, known as hubs, served as large directories that were not actually authoritative in the information that it held, but were used as compilations of a broad catalog of information that led users directly to other authoritative pages. In other words, a good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs. The scheme therefore assigns two scores for each page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages. In this manuscript, various aspects of the HITS algorithm is analyzed.

Keywords – HITS Algorithm, Web Ranking, Hubs, Authorities

ALGORITHMIC APPROACH

In the HITS algorithm, the first step is to retrieve the most relevant pages to the search query. This set is called the root set and can be obtained by taking the top n pages returned by a text-based search algorithm. A base set is generated by augmenting the root set with all the web pages that are linked from it and some of the pages that link to it. The web pages in the base set and all hyperlinks among those pages form a focused subgraph. The HITS computation is performed only on this focused subgraph. According to Kleinberg the reason for constructing a base set is to ensure that most (or many) of the strongest authorities are included.

Authority and hub values are defined in terms of one another in a mutual recursion. An authority value is computed as the sum of the scaled hub values that point to that page. A hub value is the sum of the scaled authority values of the pages it points to. Some implementations also consider the relevance of the linked pages.

The algorithm performs a series of iterations, each consisting of two basic steps:

- **Authority Update**: Update each node's Authority score to be equal to the sum of the Hub Scores of each node that points to it. That is, a node is given a high authority score by being linked to pages that are recognized as Hubs for information.
- **Hub Update**: Update each node's Hub Score to be equal to the sum of the Authority Scores of each node that it points to. That is, a node is given a high hub score by linking to nodes that are considered to be authorities on the subject.

The Hub score and Authority score for a node is calculated with the following algorithm:

- Start with each node having a hub score and authority score of 1.
- Run the Authority Update Rule
- Run the Hub Update Rule
- Normalize the values by dividing each Hub score by square root of the sum of the squares of all Hub scores, and dividing each Authority score by square root of the sum of the squares of all Authority scores.
- Repeat from the second step as necessary.

HITS, like Page and Brin's PageRank, is an iterative algorithm based on the linkage of the documents on the web. However it does have some major differences:

- It is query dependent, that is, the (Hubs and Authority) scores resulting from the link analysis are influenced by the search terms;
- As a corollary, it is executed at query time, not at indexing time, with the associated hit on performance that accompanies query-time processing.
- It is not commonly used by search engines. (Though a similar algorithm was said to be used by Teoma, which was acquired by Ask Jeeves/Ask.com.)
- It computes two scores per document, hub and authority, as opposed to a single score;
- It is processed on a small subset of ‘relevant’ documents (a 'focused subgraph' or base set), not all documents as was the case with PageRank.

To begin the ranking, $\forall p$, $\mathrm{auth}(p) = 1$and $\mathrm{hub}(p) = 1$. We consider two types of updates: Authority Update Rule and Hub Update Rule. In order to calculate the hub/authority scores of each node, repeated iterations of the Authority

Update Rule and the Hub Update Rule are applied. A k-step application of the Hub-Authority algorithm entails applying for k times first the Authority Update Rule and then the Hub Update Rule.

Authority Update Rule

$\forall p$, we update $\mathrm{auth}(p)$ to be the summation:

$$\mathrm{auth}(p) = \sum_{i=1}^{n} \mathrm{hub}(i)$$

where n is the total number of pages connected to p and i is a page connected to p. That is, the Authority score of a page is the sum of all the Hub scores of pages that point to it.

HUB UPDATE RULE

$\forall p$, we update $\mathrm{hub}(p)$ to be the summation:

$$\mathrm{hub}(p) = \sum_{i=1}^{n} \mathrm{auth}(i)$$

where n is the total number of pages p connects to and i is a page which p connects to. Thus a page's Hub score is the sum of the Authority scores of all its linking pages

NORMALIZATION

The final hub-authority scores of nodes are determined after infinite repetitions of the algorithm. As directly and iteratively applying the Hub Update Rule and Authority Update Rule leads to diverging values, it is necessary to normalize[disambiguation needed] the matrix after every iteration. Thus the values obtained from this process will eventually converge.[4]

```
G := set of pages
for each page p in G do
    p.auth = 1 // p.auth is the authority score of the page p
    p.hub = 1 // p.hub is the hub score of the page p
function HubsAndAuthorities(G)
    for step from 1 to k do // run the algorithm for k steps
        norm = 0
        for each page p in G do  // update all authority values first
            p.auth = 0
            for each page q in p.incomingNeighbors do // p.incomingNeighbors is the set of pages that link to p
                p.auth += q.hub
            norm += square(p.auth) // calculate the sum of the squared auth values to normalise
        norm = sqrt(norm)
        for each page p in G do  // update the auth scores
            p.auth = p.auth / norm  // normalise the auth values
        norm = 0
        for each page p in G do  // then update all hub values
            p.hub = 0
            for each page r in p.outgoingNeighbors do // p.outgoingNeighbors is the set of pages that p links to
                p.hub += r.auth
        norm += square(p.hub) // calculate the sum of the squared hub values to normalise
    norm = sqrt(norm)
    for each page p in G do  // then update all hub values
        p.hub = p.hub / norm   // normalise the hub values
```

The hub and authority values converge in the pseudocode above.

The code below does not converge, because it is necessary to limit the number of steps that the algorithm runs for. One way to get around this, however, would be to normalize the hub and authority values after each "step" by dividing each authority value by the square root of the sum of the squares of all authority values, and dividing each hub value by the square root of the sum of the squares of all hub values. This is what the pseudocode above does.

NON-CONVERGING PSEUDOCODE

```
G := set of pages
for each page p in G do
    p.auth = 1 // p.auth is the authority score of the page p
    p.hub = 1 // p.hub is the hub score of the page p
function HubsAndAuthorities(G)
    for step from 1 to k do // run the algorithm for k steps
        for each page p in G do  // update all authority values first
            p.auth = 0
            for each page q in p.incomingNeighbors do // p.incomingNeighbors is the set of pages that link to p
                p.auth += q.hub
        for each page p in G do  // then update all hub values
            p.hub = 0
            for each page r in p.outgoingNeighbors do // p.outgoingNeighbors is the set of pages that p links to
                p.hub += r.auth
```

LIMITATIONS WITH HITS ALGORITHM

Problems of HITS Algorithm are Although HITS provides good search results for a wide range of queries, HITS did not work well in all cases due to the following three reasons:

- Mutually reinforced relationships between hosts.
- Sometimes a set of documents on one host point to a single document on a second host, or sometimes a single document on one host point to a set of document on a second host.
- External Links not identified
- Internal Investigation not supported
- Score calculations for the links on the same domain.
- Automatically generated links. Web document generated by tools often have links that were inserted by the tool.
- Non-relevant nodes. Sometimes pages point to other pages with no relevance to the query topic.

REFERENCES

[1] Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze (2008). "Introduction to Information Retrieval". Cambridge University Press. Retrieved 2008-11-09.

[2] Kleinberg, Jon (December 1999). "Hubs, Authorities, and Communities". Cornell University. Retrieved 2008-11-09.

[3] Pankaj Gupta, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Bosagh Zadeh WTF: The who-to-follow system at Twitter, Proceedings of the 22nd international conference on World Wide Web

[4] von Ahn, Luis (2008-10-19). "Hubs and Authorities" (PDF). 15-396: Science of the Web Course Notes. Carnegie Mellon University. Retrieved 2008-11-09.

[5] Kleinberg, Jon (1999). "Authoritative sources in a hyperlinked environment" (PDF). Journal of the ACM 46 (5): 604–632. doi:10.1145/324133.324140.

[6] Li, L.; Shang, Y.; Zhang, W. (2002). "Improvement of HITS-based Algorithms on Web Documents". Proceedings of the 11th International World Wide Web Conference (WWW 2002). Honolulu, HI. ISBN 1-880672-20-0.