International Journal of Computer Applications Technology and Research
Volume 6–Issue 7, 316-325, 2017, ISSN:-2319–8656

# WHITS Algorithm for Detecting Web Communities: Using Link Structure Analysis by double weighting of links

Hemangini S. Patel
Bhagwan Mahavir College of Computer Application (BCA)
Surat, India

Apurva A. Desai
Veer Narmad South Gujarat University
Surat, India

**Abstract**: Recently two famous web page ranking algorithms are HITs and Page Rank. But Page Rank computed and refreshed off-line and not relevant to query term so not suitable for concept searching and finding topic-related communities instead HITs and SALSA are outperforms. In this paper, we discuss about to mine topic-related communities of web pages by HITs (Hyperlink-Induced Topic Search) and improve version of HITs, WHITs (weighted HITs) algorithm, which is based on hyperlink structure of web with double weighting of links matched with query term. The HITs and the WHITs algorithms are eigenvector based techniques for discovering "authoritative" web pages. Information Retrieval (IR) utilizes term based weighting method to discover relevant documents for a given query. Web IR utilities such as search engines tend to additionally process these relevant documents through link structure analysis and find rank score for each document within result set and present users for improving rank score rate of top ranked results. Existing link analysis algorithms are using principal eigenvector of resultant rank matrix for ranking. The multi topic or polymorphic query, the dominant topic discovers the major fraction within top ranked results and the sub-dominant topics are demoted. The improved version of HITs known as WHITs approach for link analysis serves for both ranking and grouping of pertinent links effectively.

**Keywords**: Eigenvector, Information Retrieval, Link Analysis, Web community, HITs

## 1. INTRODUCTION

Recently PageRank and HITs algorithms are most excellent and well-identified for ranking the web pages. In cooperation these algorithms hold a set of web pages and forms communal network. All web pages within the social network is connecting by hyperlinks structure to the other pages. By utilizing the Web page linkage arrangement in the social network; the importance of individual Web pages are evaluated. PageRank is not query dependent as well as it is not suitable for topic-related community detection instead HITs is query dependent and too slow at query time although outperforms for topic-related community detection and discovering authoritative web pages [1].

Too many non-relevant pages were found with relevant ones by search engine, and their rankings not often matched with users' requirements. Generally user sending queries to search engine are tended to short; generally contains 2-3 words, the troubles related to synonymy and polysemy make it mostly complicated to estimate which pages will be of concerned to user. By using this retrieved result, Users cannot recognize, which pages are pertinent or a highly pertinent to their query topics. It is always complicated and not the desirable case for users to search pertinent pages and obtains the vital information via browsing each and every individually. The user is further likely to be concerned in a page if it authoritative and it is pertinent to the user query.

To deal with this difficulty, one has to rearrange or else categorize the acquired pages in dissimilar clusters that are pertinent to the certain query focus to some area. These clusters outline a web page community with common interest.

Link analysis (frequently combined with content analysis) is then applied to develop the search accuracy through focusing the search in the graph neighbourhoods of these pages.

A community can be represented as a cluster of entities (People, Organization, and WebPages) that shares the common interest or an activity or an event. Its role in Web page ranking [2, 3], HITs algorithm revealed that there must be present numerous Web communities along with relevant Web pages when the query term has multiple meanings. This algorithm considers communities as association among 'authorities' as well as 'hubs'. Creators of web pages have a tendency to build association to various pages on interrelated topics. Via utilizing of these links, we can mine along with cluster of pages significant to the topics. In this paper, we describe these clusters of pages as "Web communities" [4, 5]. A Web community considered like a cluster of web pages with the purpose of further directly related to the peers inside the identical cluster than those exterior of the cluster. Generally in favour of the, query term "Jaguar", various key Web communities are around, correspondingly related to the Atari video game, automobile and the American Foot-ball players team. Through modelling the community network among a weighted graph, Web communities are able to expose based on the graph topology; the resultant Web communities are generally clusters or groups of Web pages among the identical topics.

Ahead of the hyper-linked Web situation, we consider that the thought of community as well be present in blogs as well as Web pages. Blogs are defined as Web pages that made up of journal entries among construction time stamps.

In graph theory, eigenvalue is found from the adjacency matrix of the graph. The eigenvector of the matrix is computed using the eigenvalues. The first principal eigenvector of the graph is referred as the principal eigenvector. This eigenvector plays an important role in computing a ranking in a social network. The most familiar example is web, the web pages are considered as individuals, the network structure is formed by providing link between the web pages. The ranking of the individual is computed by computing the marginal ranking by using the principal eigenvector [6]. The principal eigenvector is used in many ranking algorithm like PageRank, HITs, SALSA, Heigen[7].

This paper is structured as follows. In section 2, a few backgrounds regarding the HITs algorithm and some related work is discussed. Its enhancements are specified for improved understanding of current work. In section 3, apply the singular value decomposition (SVD) of a matrix on HITs and WHITs algorithm [8]. Hence, a few backgrounds regarding SVD also given within this section. In section 4, statistical experimental outcomes and their study are given to demonstrate the use along with achievability of WHITs. In the end, conclusion and advance research directions are in section 5.

## 2. RELATED WORK

Gibson et al. and Kleinberg [4, 5] first worked on community finding in social networks by the HITS algorithm. Within the HITs algorithm, a repetitive process was projected to calculate an authority weight and a hub weight for every page inside a set of associated Web pages. Once the computation meets up, the Web pages among peak authority ranks are authority pages, and those with peak hub ranks are hub pages. The HITs algorithm performs representation of each Web page in set by a directed graph, along with by finding the link weights through a matrix A. The entry (i, j) of A signifies the link power from page i to page j. $AA^T$ may have multi-set of eigenvalues. Well-splited eigenvalues often indicate the continued existence of several Web communities. In favour of the Web pages discovered for query "Jaguar", there are three most important Web communities exist.

Li et al. [9] deliberated the trouble of taking out communities within web pages as well as blogs, via utilizing the named component co-incident, to mapped Web pages into a named component graph.

Nomura et al. [10] proposed two sort of link analysis related alteration: *the projection technique* and *the base-set downsizing technique* to address *topic drift problem* that occurs due to authorities come together into closely linked unrelated pages, it is disreputable in the region of Information Retrieval.

Tianbao et al. [11] proposed unified model to combine link structure and content for community detection and introduce two models *conditional model and discriminative* model which obtains significant improvement over the state-of-the-art approaches for community detection.

Balaguru et al. [12] Surveyed about comparative study between PageRank, HITs, SALSA and Heigen for community discovery by computing principal eigenvalue and eigenvector.

Although Bharat et al. [13] enhanced HITs algorithm, they simply consider how to decrease the effect of un-wanted pages within the community building, not by removing these un-wanted pages. In the paper of Hou and Zhang [14], they proposed an un-wanted page removal algorithm (NPEA) to remove un-wanted pages within the base set of Web pages along with to get better the base set, which creates it feasible to build a high-quality Web page community. A worth web community is created via eliminating dissimilar web links.

Benzi et al. [15] have proposed technique which doesn't contain the outcome of noisy links by extending the notion of sub-graph centrality by eigenvector centrality for ranking authorities and hubs in web groups of associated community.

Eustace et al. [16] proposed algorithm, by utilizing a subspace of the entire links associated to query sent to search engine as well as their consequent pages returned, to discover a web community of associated hyperlinks within a query.

## 3. PRELIMINARIES TO DISCOVER COMMUNITIES

### 3.1 Methodology

Here we introduce the approach with link analysis to discover densely linked various web pages to identify numerous web communities from web graph. For this SVD is an important matrix decomposition method which is generally utilized in numerous areas, like computer vision, information retrieval, data noise reduction etc. SVD in linear algebra can disclose the inner association between matrix basics. PCA is closely related to the mathematical technique of singular value decomposition (SVD) [17].

The SVD could be utilized efficiently to mine positive key assets describing the organization of a matrix, such as the amount of autonomous columns or else rows, eigenvalues, estimate matrix. Here we are using SVD analysis with HITs and WHITs algorithms.

### 3.2 Detection of a community from query retrieved web-pages

In Link analysis algorithm the base set is the neighborhood graph N where each page (link) represents a node and a hyperlink from one page linking to another page is represented as directed edge. This neighborhood graph represents linked structure of web pages in base set. This neighborhood graph is in form of adjacency matrix as an input to link analysis algorithm. Considering L as the adjacency matrix on neighborhood graph N, the authority matrix, AUTMAT is derived from adjacency matrix L as $AUTMAT = L^{T*}L$.

Singular value Decomposition (SVD) is applied on HITs and WHITs algorithm and make use of eigenvectors of adjacency matrix to identify the principal components i.e. web communities. So, next step is to calculate the set of eigenvalues and corresponding eigenvectors for authority matrix AUTMAT. By using the SVD of the connectivity matrix, our WHITs algorithm allows the topic-based pages to get the key association information.

It is identified that, within a community of query discovered web pages, query based pages are well associated as contrast to query dissimilar pages. Some pages are not related to query topic although they are present in Web community.

## 4. EXPERIMENTATION AND RESULTS

To estimate the concert of WHITs algorithm for mining pertinent links and find out web communities, we experimented with the WHITs algorithm on numerous real data sets and compare it with HITs algorithm.

### 4.1 Data sets

We utilized the some queries from data set of Yue et al. [18] and [19] which was created according to methods specified by [19]. We utilized 13 queries; Java, Jaguar, Harvard, Search Engine, Kyoto University, Toyota, Honda, Olympic, Abortion, Alcohol, Artificial intelligence, Basketball and Architecture. Some queries topic discovers broad topics whereas some discovers specific topics.

We utilized an extended data set that was created through allowing for the entire out-links and anchors of the entire out-links as of the all root set web page and in-links as well as titles of in-links of root set. General data of every query term in dataset is presented in Table 1.

**Table 1. Experimental Data for Various Queries**

| Query (Q) | Root Set (R) | Out-links | In-links | Base Set (B) | normalized Base Set (B) |
|---|---|---|---|---|---|
| Java | 102 | 11546 | 1912 | 13560 | 10806 |
| Jaguar | 102 | 16527 | 744 | 17373 | 12711 |
| Harvard | 95 | 27243 | 4271 | 31609 | 13192 |
| Search engine | 100 | 8264 | 2273 | 10637 | 9152 |
| Kyoto University | 94 | 6393 | 700 | 7187 | 6070 |
| Toyota | 107 | 9116 | 497 | 9720 | 7802 |
| Honda | 109 | 3711 | 595 | 4415 | 3693 |
| Olympic | 105 | 5449 | 320 | 5874 | 4637 |
| Abortion | 98 | 7334 | 60 | 7492 | 6620 |
| Alcohol | 97 | 7960 | 84 | 8141 | 6702 |
| Artificial intelligence | 101 | 8614 | 113 | 8828 | 7296 |
| Basketball | 102 | 6013 | 412 | 6527 | 4833 |
| Architecture | 107 | 11813 | 331 | 12251 | 9731 |

We are finding communities which contribute within query related communities like authoritative link of pages. The authoritative links are associated to "authoritative" web pages as described within Kleinberg's HITs algorithm to consider the excellence of contributing nodes inside the community construction; we have allowed the excellence of the top 10 ordered web pages.

## 4.2 Results and evaluation

This section describes the outcome along with the excellence of community constructions produced results by our algorithm WHITs. HITS algorithm favours TKC construction as when calculating hub and authority scores via the HITS algorithm, SVD picks maximal S values which match with the tightly associated mechanism.

**Table 2(A).**
**Comparisons among extracted principal eigenvector for query 'Java': The HITS algorithm and WHITs algorithm.**

| 'Java' | |
|---|---|
| ***Weights*** | ***HITs(principal eigenvector)*** |
| 0.0322 | https://plus.google.com |
| 0.0232 | http://www.oracle.com/technetwork/java/index.html |
| 0.0191 | http://www.youtube.com |
| 0.0174 | http://www.oracle.com |
| 0.0157 | http://java.com |
| 0.0153 | http://www.facebook.com |
| 0.0148 | http://www.oracle.com/technetwork/java/javase/downloads/index.html |
| 0.0147 | https://twitter.com |
| 0.0145 | http://twitter.com |
| 0.0141 | https://www.oracle.com |

**Table 2(B).**
**Comparisons among extracted principal eigenvector for query 'Java': The WHITs algorithm.**

| 'Java' | |
|---|---|
| ***Weights*** | ***WHITs(principal eigenvector)*** |
| **0.0397** | http://www.oracle.com/technetwork/java/index.html |
| **0.0369** | http://www.oracle.com |
| **0.0328** | http://java.com |
| **0.0319** | http://www.oracle.com/technetwork/java/javase/downloads/index.html |
| **0.0279** | https://www.oracle.com |
| **0.0264** | http://www.java.net |
| **0.0264** | https://cloud.oracle.com |
| **0.0230** | https://community.oracle.com |
| **0.0227** | http://education.oracle.com |
| 0.0216 | https://blogs.oracle.com |

Generally queries sent to search engine by users are short, unclear and ambiguous. For example a short term query 'Java' can mean by; the 'Java Programming Language' or the 'Java Islands in Indonesia' or 'java coffee'. By sending this query probably primary information can obtain easily but it is difficult to recognize the exact context of the searcher.

If the query is sent by computer programmers then he probably tend to interested in java programming language. However for traveller or geographically will be interested in pages related to Java Islands in Indonesia. Here almost all results returned by search engine are related to java programming language via HITs and WHITs algorithm, although WHITs returns more authoritative results as shown in Table 2(B).

**Table 3(A).**
**Comparison among extracted principal and non principal eigenvector for query 'jaguar': the hits algorithm.**

| 'Jaguar' | |
|---|---|
| ***Weights*** | ***HITs(principal eigenvector)*** |
| 0.0211 | http://www.jaguarusa.com/index.html |
| 0.0153 | http://www.jaguar.co.uk/index.html |
| 0.0146 | http://www.jaguar.com/index.html |
| 0.0139 | http://www.jaguar.com.au/index.html |
| 0.0134 | http://www.jaguar.in/index.html |
| 0.0134 | http://www.jaguar.ie/index.html |
| 0.0132 | http://www.jaguar.co.za/index.html |
| 0.0114 | http://www.jaguar.com |
| 0.0105 | http://jaguar.pl |
| 0.0104 | http://www.jaguarlaos.com |
| ***Weights*** | ***HITs(4th non-principal eigenvector) Community of Foot ball team*** |
| 0.0503 | http://www.jaguars.com/ |
| 0.0255 | https://twitter.com/jaguars |
| 0.0248 | http://twitter.com |

| | |
|---|---|
| 0.0220 | http://www.nfl.com |
| 0.0220 | http://www.news4jax.com |
| 0.0215 | http://prod.preview.jaguars.clubs.nfl.com |
| 0.0212 | http://www.jaguarsarcade.com/ |
| 0.0211 | http://www.giants.com |
| 0.0211 | http://www.atlantafalcons.com |
| 0.0211 | http://www.ticketexchangebyticketmaster.com |

**Table 3(B).**

**Comparison among extracted principal and non principal eigenvector for query 'jaguar': the whits algorithm.**

| 'Jaguar' | |
|---|---|
| ***Weights*** | ***WHITs(principal eigenvector)*** |
| **0.1509** | http://www.jaguarusa.com/index.html |
| 0.1331 | http://www.jaguarusa.com/ |
| **0.0313** | http://www.jaguar.com/index.html |
| **0.0294** | http://www.jaguar.co.uk/index.html |
| **0.0248** | http://www.jaguar.co.za/index.html |
| **0.0247** | http://www.jaguar.com.au/index.html |
| **0.0241** | http://www.jaguar.in/index.html |
| **0.0221** | http://www.jaguar.ie/index.html |
| 0.0124 | https://twitter.com |
| 0.0079 | https://www.youtube.com |
| ***Weights*** | ***WHITs (5th non-principal eigenvector) Community of Foot ball team*** |
| **0.1187** | http://www.jaguars.com/ |
| **0.0400** | http://jaguarsblack.com/ |
| 0.0400 | http://www.jaguars.com |
| 0.0364 | https://twitter.com/jaguars |
| 0.0240 | http://twitter.com |
| 0.0212 | http://www.nfl.com |
| 0.0212 | http://www.news4jax.com |
| 0.0210 | http://prod.preview.jaguars.clubs.nfl.com |
| 0.0203 | http://www.jaguarsarcade.com/ |
| 0.0200 | http://www.giants.com |

According to HITs algorithm based on the hyperlink information Kleinberg argued [19] that it is helpful in extracting various tightly linked collections of hubs and authorities on multiple eigenvectors. As shown in above Table 3 by our experiment we get two web communities, for example, with respect to the topic 'jaguar,' not only the community of automobile (principal eigenvector), but also the community of Jacksonville jaguar NFL football community (on the 4th non-principal eigenvector) were extracted by HITs. But by our experiment on WHITs it is clearly extracts jaguar automobile community at principal eigenvector with increased weight as compared to HITs, Similarly, it extracts the community of Jacksonville jaguar NFL football community (on the 5th non-principal eigenvector) clearly and also some links have weight increased.

**Table 4(A).**

**Comparison among extracted principal and non-principal eigenvector for query 'harvard': the hits algorithm.**

| 'Harvard' | |
|---|---|
| ***Weights*** | ***HITs(principal eigenvector)*** |
| 0.0275 | http://twitter.com |
| 0.0243 | https://twitter.com |
| 0.0224 | http://www.harvard.edu |
| 0.0220 | https://www.facebook.com |
| 0.0187 | http://www.harvard.edu/ |
| 0.0177 | https://plus.google.com |
| 0.0174 | http://www.facebook.com |
| 0.0154 | http://www.youtube.com |
| 0.0151 | http://www.linkedin.com |
| 0.0137 | http://news.harvard.edu |

**Table 4(B).**

**Comparison among extracted principal and non-principal eigenvector for query 'Harvard': The WHITS algorithm.**

| 'Harvard' | |
|---|---|
| ***Weights*** | ***WHITs(principal eigenvector)*** |
| 0.0275 | http://twitter.com |
| 0.0243 | https://twitter.com |
| 0.0224 | http://www.harvard.edu |
| 0.0220 | https://www.facebook.com |
| 0.0187 | http://www.harvard.edu/ |
| 0.0177 | https://plus.google.com |
| 0.0174 | http://www.facebook.com |
| 0.0154 | http://www.youtube.com |
| 0.0151 | http://www.linkedin.com |
| 0.0137 | http://news.harvard.edu |

As shown in Table 4(A) for query 'Harvard' HITs algorithm returns only one or two links home page of Harvard University in principal eigenvector. While as Table 4(B) WHITs returns almost all pages highly relevant to Harvard University in principal eigenvector. Thus the 1st community at principal eigenvector for the topic "Harvard" consist of a fusion of pages for schools at Harvard, pages on business school, medical school, school of public health, graduate school of design, Harvard alumni, Harvard athletics, library at Harvard, the home page of Harvard University etc.

**TABLE 5(A).**

**Comparison among extracted principal and non-principal eigenvector for query 'Search Engine': The HITS algorithm.**

| 'Search Engine' | |
|---|---|
| ***Weights*** | ***HITs(principal eigenvector)*** |
| 0.0052 | http://www.google.com |
| 0.0051 | http://www.bing.com |
| 0.0049 | http://www.ask.com |
| 0.0047 | http://www.yahoo.com |
| 0.0044 | http://www.lycos.com |
| 0.0042 | http://www.facebook.com |
| 0.0042 | http://www.ixquick.com |
| 0.0042 | http://www.webcrawler.com |
| 0.0040 | http://www.excite.com |
| 0.0040 | http://www.galaxy.com |
| ***Weights*** | ***HITs(5th non-principal eigenvector) Community of multimedia content*** |
| 0.0109 | http://www.clipblast.com |
| 0.0108 | http://www.scribd.com |
| 0.0107 | http://www.metatube.net |
| 0.0107 | http://issuu.com |

| | |
|---|---|
| 0.0107 | http://www.dorble.com |
| 0.0107 | http://megadownload.net |
| 0.0107 | http://lazylibrary.com |
| 0.0107 | http://deals.hongkiat.com |
| 0.0107 | http://www.gig-listing.co.uk |
| 0.0107 | http://www.megarapidsearch.com |

**Table 5(B).**
**Comparison among extracted principal and non-principal eigenvector for query 'Search Engine': The WHITS algorithm.**

| **'Search Engine'** | |
|---|---|
| ***Weights*** | ***WHITs(principal eigenvector)*** |
| 0.0052 | http://www.google.com |
| 0.0051 | http://www.bing.com |
| 0.0049 | http://www.ask.com |
| 0.0047 | http://www.yahoo.com |
| 0.0043 | http://www.lycos.com |
| 0.0042 | http://www.facebook.com |
| 0.0042 | http://www.ixquick.com |
| 0.0041 | http://www.webcrawler.com |
| 0.0040 | http://www.excite.com |
| 0.0040 | http://www.galaxy.com |
| ***Weights*** | ***WHITs (2nd non-principal eigenvector) Community of meta search engine*** |
| 0.1165 | http://www.searchenginecolossus.com/ |
| 0.1152 | http://searchenginecolossus.com/ |
| 0.0812 | https://www.ixquick.com/ |
| 0.0812 | http://ixquick.com/ |
| 0.0701 | http://search.aol.com/ |
| 0.0691 | http://www.dogpile.com/ |
| 0.0545 | http://searchengineshowdown.com/ |
| 0.0545 | http://www.hotbot.com/ |
| 0.043 | http://www.searchengineguide.com/ |
| 0.0388 | http://searchenginewatch.com/ |

In above table 5(A) it will display all available well-known search engines in principal eigenvector via HITs algorithm and WHITs Algorithm with similar weights. And 5th non-principal eigenvector community discovered by HITs is multimedia content like find music videos, tv shows, Movies and funniest videos, digital documents library, magazines, catalogs and publications, search and download shared files from different file hosting sites, newest software, gadgets & web services, Search File, EBook etc. Similarly, as shown in Table 5(B) WHITs discovered community of search engines which are combining results from two or more search engines like meta-search engines at 2nd non-principal eigenvector.

**Table 6(A).**
**Comparison among extracted principal and non-principal eigenvector for query 'Kyoto University': The HITS algorithm.**

| **'Kyoto University'** | |
|---|---|
| ***Weights*** | ***HITs(principal eigenvector)*** |
| 0.1237 | http://www.kyoto-u.ac.jp/en |
| 0.1077 | http://www.kyoto-u.ac.jp/en/ |
| 0.0201 | http://www.kyoto-u.ac.jp |
| 0.0128 | http://www.opir.kyoto-u.ac.jp |
| 0.0091 | http://www.kyoto-u.ac.jp/en/faculties-and-graduate/ |
| 0.0091 | http://www.oc.kyoto-u.ac.jp/en/ |
| 0.0089 | http://twitter.com |
| 0.0087 | http://www.asafas.kyoto-u.ac.jp/en/ |
| 0.0086 | https://www.facebook.com |
| 0.0076 | http://www.opir.kyoto-u.ac.jp/kuprofile/ |

**Table 6 (B).**
**Comparison among extracted principal and non-principal eigenvector for query 'Kyoto University': The WHITS algorithm.**

| **'Kyoto University'** | |
|---|---|
| ***Weights*** | ***WHITs(principal eigenvector)*** |
| **0.1326** | http://www.kyoto-u.ac.jp/en |
| 0.1067 | http://www.kyoto-u.ac.jp/en/ |
| **0.0402** | http://www.kyoto-u.ac.jp |
| **0.0173** | http://www.opir.kyoto-u.ac.jp |
| 0.0096 | http://www.opir.kyoto-u.ac.jp/kuprofile/ |
| 0.0087 | http://www.t.kyoto-u.ac.jp/en |
| 0.0081 | http://sph.med.kyoto-u.ac.jp |
| | http://www.kyoto-u.ac.jp/en/faculties-and-graduate/ |
| 0.0080 | http://www.oc.kyoto-u.ac.jp/en/ |
| 0.0078 | http://www.t.kyoto-u.ac.jp |

As shown in Table 6(A) principal eigenvectors for HITs algorithm returns pages related to 'Kyoto University' except two results. While results shown in Table 6(B) with WHITs returns almost all pages related to query 'Kyoto University' in Japan.

**Table 7(A).**
**Comparison among extracted principal eigenvector for query 'toyota': the hits algorithm.**

| **'Toyota'** | |
|---|---|
| ***Weights*** | ***HITs(principal eigenvector)*** |
| 0.0514 | https://www.facebook.com |
| 0.0504 | https://plus.google.com |
| 0.0494 | https://twitter.com |
| 0.0478 | http://www.toyota.com |
| 0.0444 | https://www.youtube.com |
| 0.0344 | http://www.toyota.com/ |
| 0.0165 | http://www.dealer.com |
| 0.0127 | https://www.google.com |
| 0.0119 | https://instagram.com |
| 0.0114 | http://instagram.com |

**Table 7(B).**
**Comparison among extracted principal eigenvector for query 'Toyota': WHITs algorithm.**

| **'Toyota'** | |
|---|---|
| ***Weights*** | ***WHITs(principal eigenvector)*** |
| 0.0944 | http://www.toyota.com/ |
| 0.0615 | http://www.toyota.com |
| 0.0417 | https://plus.google.com |
| 0.0325 | https://www.facebook.com |
| 0.0316 | https://www.youtube.com |
| 0.0296 | https://twitter.com |
| 0.0189 | http://www.toyota-global.com/ |

| | |
|---|---|
| 0.0181 | http://www.dealer.com |
| 0.0149 | http://instagram.com |
| 0.0147 | http://www.toyotaracing.com/ |

In above table 7(A) results returned by HITs algorithm contains home page of Toyota Company with lower weights while results as shown in Table 7(B) returned by WHITs contains higher weights for Toyota home page and returns more related results.

**Table 8(A).**
**Comparison among extracted principal eigenvector for query 'Honda': The HITS algorithm.**

| 'Honda' | |
|---|---|
| ***Weights*** | ***HITs(principal eigenvector)*** |
| 0.1961 | http://www.honda.com/ |
| 0.1728 | http://powersports.honda.com/ |
| 0.1728 | http://powersports.honda.com/index.aspx |
| 0.0571 | http://automobiles.honda.com/ |
| 0.0443 | http://world.honda.com/ |
| 0.0402 | http://marine.honda.com/ |
| 0.0314 | http://powerequipment.honda.com/ |
| 0.0190 | http://www.hondafinancialservices.com/ |
| 0.0183 | https://plus.google.com/%2BHonda |
| 0.0173 | http://www.hondacenter.com/ |

**Table 8(B).**
**Comparison among extracted principal eigenvector for query 'Honda': The WHITS algorithm.**

| *'Honda'* | |
|---|---|
| ***Weights*** | ***WHITs(principal eigenvector)*** |
| 0.1714 | http://powersports.honda.com/ |
| 0.1714 | http://powersports.honda.com/index.aspx |
| 0.1610 | http://www.honda.com/ |
| **0.0909** | http://automobiles.honda.com/ |
| **0.0631** | http://marine.honda.com/ |
| **0.0502** | http://world.honda.com/ |
| **0.0419** | http://powerequipment.honda.com/ |
| **0.0349** | http://www.hondafinancialservices.com/ |
| **0.0296** | https://plus.google.com/%2BHonda |
| **0.0228** | http://automobiles.honda.com/civic-sedan/ |

As shown in table 8 WHITs principal eigenvectors returns similar results to HITs results but principal eigenvectors with higher weights.

**Table 9(A).**
**Comparison among extracted principal eigenvector for query 'Olympic': The HITS algorithm.**

| 'Olympic' | |
|---|---|
| ***Weights*** | ***WHITs(principal eigenvector)*** |
| 0.0790 | http://www.olympic.org/ |
| 0.0306 | https://twitter.com |
| 0.0303 | http://www.nbcolympics.com/ |
| 0.0281 | https://plus.google.com |
| 0.0238 | https://www.facebook.com |
| 0.0198 | http://www.rio2016.com |
| 0.0194 | http://www.youtube.com |
| 0.0182 | https://www.instagram.com |
| 0.0182 | https://www.rio2016.com/en |
| 0.0179 | https://www.linkedin.com |

**Table 9(B).**
**Comparison among extracted principal eigenvector for query 'Olympic': The WHITS algorithm.**

| 'Olympic' | |
|---|---|
| ***Weights*** | ***WHITs(principal eigenvector)*** |
| **0.2625** | http://www.olympic.org/ |
| **0.2258** | http://www.nbcolympics.com/ |
| 0.0666 | http://www.specialolympics.org/ |
| 0.0558 | http://en.beijing2008.cn/ |
| 0.0364 | http://sports.yahoo.com/olympics/ |
| 0.0247 | http://history1900s.about.com/od/greateventsofthecentury/a/olympicfacts.htm |
| 0.0180 | http://www.teamusa.org/ |
| 0.0135 | http://www.itftennis.com/olympics/ |
| 0.0124 | http://www.olympicholidays.com/ |
| 0.0099 | https://www.rio2016.com/en |

As shown in Table 9(B) WHITs principal eigenvectors returns highly related results for query Olympic as compared to HITs results.

**Table 10(A).**
**Comparison among extracted principal and non-principal eigenvector for query 'Abortion': The HITS algorithm.**

| 'Abortion' | |
|---|---|
| ***Weights*** | ***HITs(principal eigenvector)*** |
| 0.0087 | http://www.guttmacher.org |
| 0.0086 | http://www.youtube.com |
| 0.0085 | http://www.ncbi.nlm.nih.gov |
| 0.0084 | http://www.rcog.org.uk |
| 0.0084 | http://www.ama-assn.org |
| 0.0084 | http://www.cdc.gov |
| 0.0083 | http://www.nytimes.com |
| 0.0083 | http://www.plannedparenthood.org |
| 0.0082 | http://facebook.com |
| 0.0082 | http://query.nytimes.com |
| ***Weights*** | ***HITs($7^{th}$ non-principal eigenvector) community of health and medical association*** |
| 0.0829 | http://www.guttmacher.org |
| 0.0810 | http://ama-assn.org |
| 0.0810 | http://www.utahmed.org |
| 0.0810 | http://www.msmaonline.com |
| 0.0810 | http://www.rcsed.ac.uk |
| 0.0810 | http://deutsch.medscape.com |
| 0.0810 | http://www.asrm.org |
| 0.0810 | http://espanol.medscape.com |
| 0.0810 | http://www.endocrine.org |
| 0.0810 | http://francais.medscape.com |

**Table 10(B).**
**Comparison among extracted principal and non-principal eigenvector for query 'Abortion': The WHITS algorithm.**

| 'Abortion' | |
|---|---|
| ***Weights*** | ***WHITs(principal eigenvector)*** |
| 0.0173 | http://www.guttmacher.org |

| | |
|---|---|
| 0.0153 | http://www.nrlc.org |
| 0.0149 | http://www.cdc.gov |
| 0.0143 | http://www.justfacts.com/abortion.asp |
| 0.0143 | http://www.pollingreport.com/abortion.htm |
| 0.0143 | http://www.afterabortion.org |
| 0.0142 | http://www.gallup.com |
| 0.0142 | http://www.justfacts.com |
| 0.0142 | http://blogs.abcnews.com |
| 0.0142 | http://medical.merriam-webster.com |
| ***Weights*** | ***WHITs (6th non-principal eigenvector) Community of health and national abortion federation*** |
| 0.2502 | http://www.guttmacher.org |
| 0.1832 | http://www.prochoice.org |
| 0.1832 | http://prochoice.org |
| 0.0916 | http://ama-assn.org |
| 0.0916 | http://www.utahmed.org |
| 0.0916 | http://www.msmaonline.com |
| 0.0916 | http://www.rcsed.ac.uk |
| 0.0916 | http://deutsch.medscape.com |
| 0.0916 | http://www.asrm.org |
| 0.0916 | http://espanol.medscape.com |

As shown in Table 10(B) for the query topic 'abortion,' WHITs returns the health related, the pro-life community, such as NRLC (National Right to Life), CDC (Centres for Disease Control and Prevention), Search *medical* terms and abbreviations with the most up-to-date and comprehensive *medical* dictionary from the reference experts and so on come to the front as compared to HITs as shown in Table 10(A).
Second 6th non-principal eigenvector come in front by WHITs is on health, national abortion federation, AMA (American Medical Association) , Utah medical association, MSMA (Mississippi State Medical Association), the royal college of surgeons of Edinburgh, The *ASRM* is an organization devoted to advancing knowledge and expertise in reproductive medicine and biology, with a particular focus on infertility etc. Similarly, at 7th non-principal eigenvector HITs detects community of health and medical association.

**Table 11(A).**
**Comparison among extracted principal and non-principal eigenvector for query 'Alcohol': The HITs algorithm.**

| **'Alcohol'** | |
|---|---|
| ***Weights*** | ***HITs(principal eigenvector)*** |
| 0.0398 | https://twitter.com |
| 0.0384 | https://www.facebook.com |
| 0.0274 | http://twitter.com |
| 0.0266 | https://plus.google.com |
| 0.0196 | http://www.youtube.com |
| 0.0196 | https://www.youtube.com |
| 0.0185 | http://www.facebook.com |
| 0.0143 | http://www.cdc.gov |
| 0.0130 | http://www.ncbi.nlm.nih.gov |
| 0.0124 | http://www.who.int |
| ***Weights*** | ***HITs(2nd non-principal eigenvector) Community of NHS*** |
| 0.2588 | http://twitter.com |
| 0.2531 | http://www.facebook.com |
| 0.2475 | http://www.youtube.com |
| 0.2044 | http://www.nhs.uk/livewell/alcohol |
| 0.1935 | http://www.nhscarerecords.nhs.uk |
| 0.1935 | http://www.show.scot.nhs.uk |
| 0.1935 | http://www.nhs.uk |
| 0.1935 | http://www.hscni.net |
| 0.1935 | http://www.nhsdirect.wales.nhs.uk |
| 0.1935 | http://www.jobs.nhs.uk |

**Table 11(B).**
**Comparison among extracted principal and non-principal eigenvector for query 'Alcohol': The HITs algorithm.**

| **'Alcohol'** | |
|---|---|
| ***Weights*** | ***WHITs(principal eigenvector)*** |
| 0.0356 | https://twitter.com |
| 0.0324 | https://www.facebook.com |
| **0.0245** | http://www.cdc.gov |
| 0.0214 | http://www.niaaa.nih.gov |
| 0.0210 | https://plus.google.com |
| **0.0202** | http://www.ncbi.nlm.nih.gov |
| 0.0196 | https://www.youtube.com |
| 0.0178 | http://twitter.com |
| **0.0160** | http://www.who.int |
| 0.0155 | http://pubs.niaaa.nih.gov |
| ***Weights*** | ***WHITs (2nd non-principal eigenvector) Community of NHS*** |
| 0.1524 | http://www.nhs.uk/livewell/alcohol |
| 0.1405 | http://www.nhs.uk |
| 0.1006 | http://twitter.com |
| 0.0948 | http://www.youtube.com |
| 0.0917 | http://www.facebook.com |
| 0.0703 | http://www.nhscarerecords.nhs.uk |
| 0.0703 | http://www.show.scot.nhs.uk |
| 0.0703 | http://www.hscni.net |
| 0.0703 | http://www.nhsdirect.wales.nhs.uk |
| 0.0703 | http://www.jobs.nhs.uk |

As shown in above table 11(B), WHITs returns principal eigenvectors as CDC (centres for daises control and prevention), NIH (national institute on alcohol abuse and alcoholism), WHO (World Health Organization) etc. WHITs retunes more related community with higher weights as compared to HITs algorithm shown in Table 11(A). In 2nd non-principal eigenvector community of NHS (National Health Service) is come in front successfully by WHITs algorithm as compared to HITs algorithm.

**Table 12(A).**
**Comparison among extracted principal and non-principal eigenvector for query 'Artificial Intelligence': The HITS algorithm.**

| ***'Artificial Intelligence'*** | |
|---|---|
| ***Weights*** | ***HITs(principal eigenvector)*** |
| 0.0211 | https://twitter.com |
| 0.0193 | http://www.facebook.com |
| 0.0154 | https://plus.google.com |

| | |
|---|---|
| 0.0125 | http://twitter.com |
| 0.0105 | https://www.youtube.com |
| 0.0090 | https://www.facebook.com |
| 0.0079 | https://www.linkedin.com |
| 0.0060 | http://www.youtube.com |
| 0.0046 | http://blogs.barrons.com |
| 0.0046 | https://itunes.apple.com |
| ***Weights*** | ***HITs(7$^{th}$ non-principal eigenvector) Community of AI*** |
| 0.1220 | http://www.aaai.org/ |
| 0.0975 | http://www.wired.com |
| 0.0350 | https://www.youtube.com |
| 0.0347 | http://www.aaai.org |
| 0.0336 | https://secure.customersvc.com |
| 0.0336 | https://vip.wordpress.com |
| 0.0336 | https://subscription.timeinc.com |
| 0.0336 | http://subscription-assets.timeinc.com |
| 0.0293 | https://www.pinterest.com |
| 0.0279 | http://en.wikipedia.org |

**Table 12(B).**
**Comparison among extracted principal and non-principal eigenvector for query 'Artificial Intelligence': The WHITS algorithm.**

| **'Artificial Intelligence'** | |
|---|---|
| ***Weights*** | ***WHITs(principal eigenvector)*** |
| 1.0000 | http://www.imdb.com/title/tt0212720/ |
| 0.0000 | http://www.gizmodo.in |
| 0.0000 | http://www.gale.cengage.com |
| 0.0000 | http://www.zoomtv.in |
| 0.0000 | http://www.ceser.in/ceserp/index.php/ijai |
| 0.0000 | http://homes.cs.washington.edu/~lazowska/cra/ai.html |
| 0.0000 | https://piratebay.host |
| 0.0000 | http://news.blogs.nytimes.com |
| 0.0000 | http://opennero.googlecode.com |
| 0.0000 | http://www.cnetnews.com.cn |
| ***Weights*** | ***WHITs (5$^{th}$ non-principal eigenvector) Community of AI*** |
| 6.8614 | http://www.aaai.org/ |
| 5.7750 | http://www.airccse.org/journal/ijaia/ijaia |
| 3.5836 | http://aima.cs.berkeley.edu/ |
| 3.5392 | http://www.cs.washington.edu |
| 3.2918 | http://www.neci.nj.nec.com |
| 3.2918 | http://zelda.thomson.com |
| 3.2918 | http://www.research.ibm.com |
| 3.2918 | http://www.mcs.anl.gov |
| 3.2918 | http://sls-www.lcs.mit.edu |
| 3.2918 | http://207.68.137.59 |

As shown in Table 12(A) results returned by HITs are not related to AI but WHITs returns more related results as shown in Table 12(B) like, the International Journal of Artificial Intelligence (*IJAI*) is a peer-reviewed online journal, A magazine of Artificial Intelligence CRA (Computing Research Association), Game platform for Artificial Intelligence research and education.

At 7$^{th}$ non-principal eigenvector HITs detects community of AI, similarly at 5$^{th}$ non-principal eigenvector community of AI related topics is extracted.

**Table 13(A).**
**Comparison among extracted principal eigenvector for query 'Basketball': The HITS algorithm.**

| **'Basketball'** | |
|---|---|
| ***Weights*** | ***HITs(principal eigenvector)*** |
| 0.1910 | http://sports.yahoo.com/nba/ |
| 0.1751 | http://espn.go.com/nba/ |
| 0.1601 | http://www.nba.com/ |
| 0.1416 | http://www.basketball-reference.com/ |
| 0.0995 | http://www.fiba.com/ |
| 0.0485 | http://www.euroleague.net/ |
| 0.0461 | http://espn.go.com/mens-college-basketball/ |
| 0.0163 | http://www.usab.com/ |
| 0.0158 | http://www.basketball.net.au/ |
| 0.0128 | http://www.onlinegames.com/basketball/ |

**Table 13(B).**
**Comparison among extracted principal eigenvector for query 'Basketball': The WHITS algorithm.**

| **'Basketball'** | |
|---|---|
| ***Weights*** | ***WHITs(principal eigenvector)*** |
| 0.1591 | http://espn.go.com/nba/ |
| 0.1578 | http://www.nba.com/ |
| 0.1568 | http://sports.yahoo.com/nba/ |
| 0.1183 | http://www.basketball-reference.com/ |
| 0.1045 | http://www.fiba.com/ |
| 0.0816 | http://espn.go.com/mens-college-basketball/ |
| 0.0388 | http://www.euroleague.net/ |
| 0.0313 | http://www.onlinegames.com/basketball/ |
| 0.0257 | http://www.usab.com/ |
| 0.0222 | http://www.flashscore.com/basketball/ |

As shown in Table 13(B) WHITs and shown in Table 13(A) HITs extracts the similar community related to basketball like NBA (National Basketball Association), FIBA (*International* Basketball Federation), Euroleague Basketball, online basketball etc.

**Table 14(A).**
**Comparison among extracted principaland non-principal eigenvector for query 'Architecture': The HITS algorithm**.

| 'Architecture' | |
|---|---|
| ***Weights*** | ***HITs(principal eigenvector)*** |
| 0.0145 | http://www.archdaily.com/ |
| 0.0388 | https://twitter.com |
| 0.0190 | https://www.facebook.com |
| 0.0186 | http://boty.archdaily.com |
| 0.0182 | http://www.archdaily.com |
| 0.0182 | http://www.fastcodesign.com |
| 0.0177 | http://www.architectural-review.com |
| 0.0174 | http://www.archdaily.pe |
| 0.0174 | http://www.archdaily.cn |
| 0.0174 | http://www.plataformaarquitectura.cl |

**Table 14(B).**
**Comparison among extracted principaland non-principal eigenvector for query 'Architecture': The WHITS algorithm.**

| 'Architecture' | |
|---|---|
| ***Weights*** | ***WHITs(principal eigenvector)*** |
| 0.0949 | http://www.archdaily.com/ |
| 0.0412 | http://www.archdaily.com |
| 0.0351 | https://twitter.com |
| 0.0211 | http://www.architectural-review.com |
| 0.0210 | http://boty.archdaily.com |
| 0.0210 | http://www.fastcodesign.com |
| 0.0202 | http://www.archdaily.pe |
| 0.0202 | http://www.archdaily.cn |
| 0.0202 | http://www.plataformaarquitectura.cl |
| 0.0202 | http://www.archdaily.mx |

As shown in Table 14(B) WHITs detects the community of Broadcasting Architecture Worldwide: Architecture news, competitions and projects, global *architecture* magazine for the 21st century etc for the query 'Architecture'. As above WHITs returns more relevant results then HITs.

## 5. CONCLUSION AND FUTURE WORKS

In our model, community can be represented by densely linked web pages for the short term query topics such that a collection of topic- based web pages are retrieved having common interest.

In our work, introduced in this paper, use SVD to describe a community of query based web pages as of the returned results. In adding up, our WHITs algorithm which uses it to discover a community of query based web pages. We experimented our WHITs algorithm via existent data set along with the precision of community discovery, was estimated adjacent to trendy existing ranking algorithms. Experimental estimation specifies that, our method do well in the discovered web community for weighted graph. We look forward to examine the probable expansion of our model in removing unrelated and noisy links. As well as centred on utilizing less links for construct a dependable community of query-based web pages.

## 6. REFERENCES

[1] Richardson, M. and Domingos, P. 2004. Combining link and content information in Web search. In *Web Dynamics*. Springer Berlin Heidelberg. pp. 179-193.

[2] Efe, K., Raghavan, V. and Lakhotia, A. 2004. Content and link structure analysis for searching the web. *Series in Machine Perception and Artificial Intelligence* 58. pp. 431-452.

[3] Smyth, B., Balfe, E., Freyne, J., Briggs, P., Coyle, M. 2004. Exploiting query repetition and regularity in an adaptive community-based web search engine. *User Modeling and User-Adapted Interaction* 14, no. 5. pp. 383-423.

[4] Gibson, D., Kleinberg J. and Raghavan, P. 1998. Inferring web communities from link topology. In *Proceedings of the ninth ACM conference on Hypertext and hypermedia: links, objects, time and space---structure in hypermedia systems: links, objects, time and space---structure in hypermedia systems*. ACM. pp. 225-234.

[5] Kleinberg, J. M. 1999. Hubs, authorities, and communities. *ACM computing surveys (CSUR)* 31, no. 4es: 5. pp. 1-3.

[6] Ng, A. Y., Zheng , A. X. and Jordan, M. I. 2001. Link analysis, eigenvectors and stability. In *International Joint Conference on Artificial Intelligence*, vol. 17, no. 1, LAWRENCE ERLBAUM ASSOCIATES LTD. pp. 903-910.

[7] Ding, C., He, X., Zha, H. and Simon, H. 2002. PageRank, HITS and a unified framework for link analysis. In Proceedings of the 25th ACM SIGIR Conference, Tampere, Finland. pp. 353–354.

[8] Patel, H. S. and Desai, A.A. An Improvement of Link Analysis Algorithm to Mine Pertinent Links: Weighted HITS Algorithm based on additive fusion of graphs by Query Similarity. Unpublished. (Unpublished manuscript).

[9] Li, X., Liu, B. and Philip, S. Y. 2006. Mining Community Structure of Named Entities from Web Pages and Blogs. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 108-114.

[10] Nomura, S., Oyama, S., Hayamizu T. and Ishida, T. 2004. Analysis and improvement of hits algorithm for detecting web communities. *Systems and Computers in Japan* 35, no. 13, pp. 32-42.

[11] Jin, T. R., Chi, Y. and Zhu, S. 2009. Combining link and content for community detection: a discriminative approach. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. pp. 927-936.

[12] Balaguru, S., Nallathamby, R. and Rene Robin, C. R. 2015. Community discovery: comparitive survey between page rank, hits, salsa and heigen. *Elysium journal of engineering research and management*: 46, Volume - 2, Issue – 3, 1-4.

[13] Bharat, K. and Henzinger, M. R. 1998. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. pp. 104-111.

[14] Hou, J. and Zhang, Y. 2002. Constructing good quality web page communities. In *Australian Computer Science Communications*, vol. 24, no. 2, Australian Computer Society, Inc., pp. 65-74.

[15] Benzi, M., Estrada, E. and Klymko, C. 2013. Ranking hubs and authorities using matrix functions. *Linear Algebra and its Applications* 438, no. 5, pp. 2447-2474.

[16] Eustace, J., Wang, X. and Li, J. 2014. Approximating web communities using subspace decomposition. *Knowledge-Based Systems* 70, pp. 118-127.

[17] Shlens, J. 2014. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

[18] He, Y., Qiu, M. Jin, M. and Xiong, T. 2012. Improvement on HITS Algorithm. *Applied Mathematics & Information Sciences* 6, no. 3, 1075-1085.

[19] Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)* 46, no. 5, 604-632.