**To:** Instructor Nikhil Bhardwaj

**From:** Yeison Montoya - ID: 300375916

Link to my [Jupyter Notebook](https://github.com/yeisonmontoya1815/Special-Topics-in-Data-Analytics-CSIS-4260-002/blob/main/Seminar%202%20-%20HITS%20Algorithm.ipynb)

# Web Mining – HANDLING WITH UNSTRUCTURED DATA

## Executive Summary

Web mining is a process that involves three essential techniques to uncover the complexities of the World Wide Web. The first technique is web content mining, which involves extracting information from textual, visual, and multimedia elements on web pages. This process is fundamental for decoding the wealth of data available on the web (Han et al., 2011). The second technique is web structure mining, which focuses on hyperlink relationships and HTML tags to unveil the structural organization of web pages (Jiawei et al., 2012). The third technique is web usage mining, which systematically analyzes user interactions and behaviours on the web, including the examination of clickstreams and access logs to reveal patterns in user engagement (Wu et al., 2014). By using these techniques, it is possible to gain a comprehensive understanding of content structures, inter-page relationships, and user engagement patterns, which contribute to applications such as information retrieval and recommendation systems.

This report emphasizes more on web structure mining (HITS Algorithm) in particular. It extracts patterns and relationships from the structure of the World Wide Web and focuses on semi-structured data like hyperlinks and HTML tags, providing insights into web content organization, and analyzing hyperlink structures that help researchers identify patterns of importance and influence between web pages. HITS aids in information retrieval, link prediction, and identifying influential nodes.

## Analysing HITS (Hyperlink-Induced Topic Search)

HITS (Hyperlink-Induced Topic Search) is a web structure mining algorithm that scores web pages based on their role as hubs or authorities. It identifies these roles within a hyperlink structure, making it useful in information retrieval, social network analysis, and recommendation systems. Wang et al. (2021) demonstrated its effectiveness in identifying influential nodes in social networks, highlighting its relevance in contemporary web structure mining research.

The concept of "Hub" and "Authority" is central to the HITS algorithm. In the context of the web, a "Hub" is a web page that extensively links to other pages, acting as a central point that provides links to various sources of information. On the other hand, an "Authority" is a web page that is linked to many "Hubs". Essentially, an Authority is seen as valuable and trustworthy because it is referenced by many other pages considered important (Kleinberg, 1999). The HITS algorithm assigns Hub and Authority scores to each page on the web. During each iteration of the algorithm, Hub scores are updated based on the Authorities the page points to, and Authority scores are updated based on the Hubs pointing to the page. These scores are normalized in each iteration to prevent score inflation.

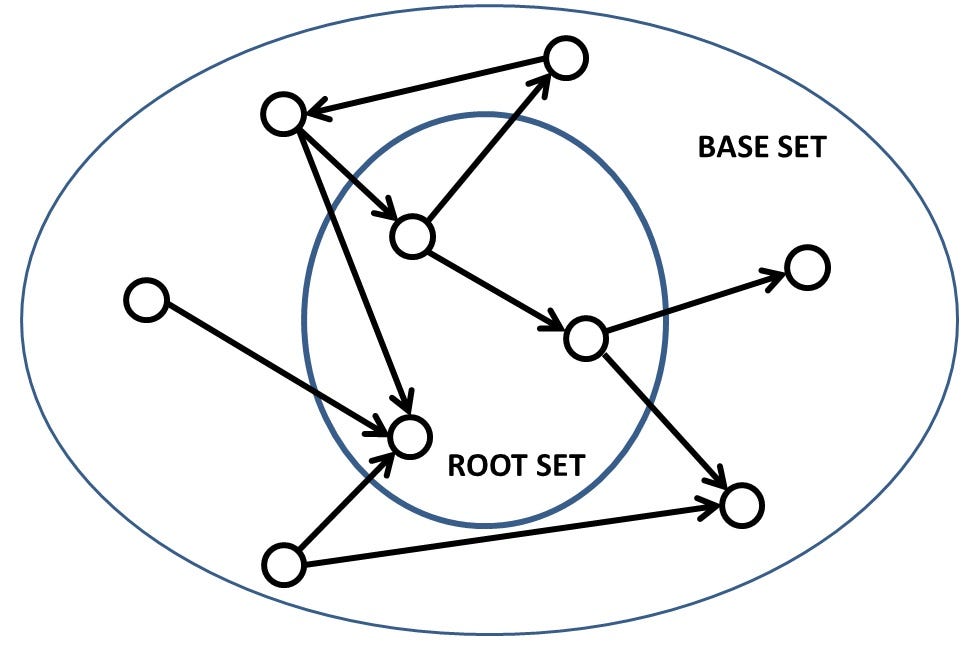
## Mathematical Model of HITS

HITS, commonly known as the Hubs and Authorities algorithm, provides a formal representation of web page importance based on hyperlink structures. Let G = (V, E) denote a directed graph where V represents the set of nodes (web pages) and E represents the set of directed edges (hyperlinks). The algorithm iteratively assigns to each node: Authority (ai) and Hub (ℎi). the general scope of the algorithm is as follows:

* Set the initial values of hub and authority for each node to 1.
* In each iteration, modify the hub and authority values for every node in the graph.
* The updated authority score is calculated as the sum of the hub scores of its parent nodes.
* The updated hub score is determined by summing the authority scores of its child nodes.
* Normalize the newly calculated authority and hub scores.

To begin with, we need to identify a group of relevant web pages. These may be web pages that contain the query string in the text, or pages that the search engine considers important based on some other reason. These potential pages are known as authorities, as they are important to the user's query. This group is referred to as the root set. In this example, the potential authorities we have identified make up the root set.

#### Figure 1. Roots and Base sets for the HITS initial analysis



Retrieved by. HITS Algorithm: Link Analysis Explanation and Python Implementation from Scratch ([Link](https://towardsdatascience.com/hits-algorithm-link-analysis-explanation-and-python-implementation-61f0762fd7cf))

1. Initialize Authority and Hub scores for each node to 1:
2. At iteration , update the Authority scores:

Where is the set of nodes pointing to (in-neighbors).

1. Also, at iteration , update the Hub scores:

Where is the set of nodes pointed to by (out-neighbors).

1. Normalize Authority and Hub scores:

where ||•|| denotes the Euclidean norm.

1. Repeat steps 2-4 for a predefined number of iterations or until the scores converge. The HITS algorithm iteratively updates Hub and Authority scores until convergence.

## Practical Example

Consider a simplified web graph with three pages (A,B,C) and hyperlinks (A→ B,A → C,B → A,C → A,C → B). Initialize Authority and Hub scores to 1 for each page. Iteratively update the scores according to the rules outlined above.

* **Iteration 1:**

2

1

* **Iteration 2:**

Continue iterations until convergence. The final scores represent the relative authority and hub values for each page in the web graph.

## HITS algorithm in Python environment

This Python script employs the NetworkX library to implement the Hyperlink-Induced Topic Search (HITS) algorithm on a directed graph representing a hyperlink structure. The directed edges between nodes (web pages) illustrate the hyperlink relationships. The script initializes a DiGraph object in NetworkX, adding edges based on the given relationships. Following the graph creation, the HITS algorithm is applied to calculate hub and authority scores for each node in the graph. The nx.hits function from NetworkX is utilized for this purpose, with a maximum of 100 iterations specified. To get a more detailed understanding, please refer to my [Jupyter Notebook](https://github.com/yeisonmontoya1815/Special-Topics-in-Data-Analytics-CSIS-4260-002/blob/main/Seminar%202%20-%20HITS%20Algorithm.ipynb).

* **NetworkX (Python):** NetworkX is a Python library for the creation, manipulation, and study of complex networks. It provides an implementation of the HITS algorithm.
* **Gephi:** this is an open-source visualization and exploration platform for network analysis. It can be used to visualize and analyze the results of the HITS algorithm.

### import networkx as nx

### # Create a directed graph representing the hyperlink structure

### G = nx.DiGraph()

### G.add\_edges\_from([(1, 2), (1, 3), (2, 1), (3, 1), (3, 2)])

### # Apply HITS algorithm

### hub, authority = nx.hits(G, max\_iter=100)

### # Display Hub and Authority scores for each page

### for node in G.nodes():

### print(f"Page {node}: Hub Score = {hub[node]}, Authority Score = {authority[node]}")

## Results

The resulting hub and authority scores are stored in the hub and authority dictionaries, respectively. The script then proceeds to display these scores for each page in the graph, iterating through the nodes and printing the corresponding hub and authority scores.

### Page 1: Hub Score = 0.35689586789220956, Authority Score = 0.35689586789220934

### Page 2: Hub Score = 0.1980622641951617, Authority Score = 0.44504186791262895

### Page 3: Hub Score = 0.44504186791262884, Authority Score = 0.19806226419516174

Here's an explanation of the scores for each page:

**Page 1:** The Hub score for Page 1 is **0.3569**, meaning it is considered a **Good hub**, linking to other important pages. The Authority score of **0.3569** suggests that Page 1 is also linked to other authoritative pages.

**Page 2:** The Hub score for Page 2 is **0.1981**, indicating that it is **not as significant** as a hub compared to Page 1. However, it has a **high Authority** score of **0.4450**, meaning it is linked to other important pages.

**Page 3:** it has a high Hub score of **0.4450**, suggesting a **significant hub**, linking to other important pages. On the other hand, its Authority score is **0.1981**, **less authoritative** pages compared to P1.

After reordering some lines of code, more detail in my ([Jupyter Notebook](https://github.com/yeisonmontoya1815/Special-Topics-in-Data-Analytics-CSIS-4260-002/blob/main/Seminar%202%20-%20HITS%20Algorithm.ipynb)) here's the results plot:

Figure 2. Relationship between nodes and their HUB & Authority scores

A diagram of a triangle

Description automatically generated

Retrieved by. Montoya Y. ([Link](https://github.com/yeisonmontoya1815/Special-Topics-in-Data-Analytics-CSIS-4260-002/blob/main/Seminar%202%20-%20HITS%20Algorithm.ipynb))

More examples can be analyzed from the University of Michigan (2022) a complete video about the HITS Algorithm by following this [link](:%20https:/www.youtube.com/watch?v=-kiKUYM9Qq8) and a more complete example with 14 nodes is available in my ([Jupyter Notebook](https://github.com/yeisonmontoya1815/Special-Topics-in-Data-Analytics-CSIS-4260-002/blob/main/Seminar%202%20-%20HITS%20Algorithm.ipynb)).

## When to use the HITS Algorithm

* **Directed Graph Structures:** HITS excels in scenarios where the data is structured as a directed graph, making it particularly valuable for applications involving hyperlink relationships.
* **Web Structure Mining:** The algorithm is highly effective for web structure mining, offering insights into the importance and influence of specific web pages based on hyperlink relationships.
* **Identification of Authorities and Hubs:** HITS is well-suited for situations where identifying authoritative sources and hubs within a network is crucial, providing a nuanced understanding of the significance of specific web pages.
* **Hierarchical Data Structures:** In domains where the data exhibits a clear hierarchical structure, HITS proves to be a powerful tool for uncovering authoritative sources within a linked environment.

## Where NOT to use it

* **Lack of Graph Structure:** HITS may not be optimal for datasets lacking a discernible graph structure or meaningful hyperlink relationships.
* **Loosely Connected Entities:** In cases where the dataset comprises unconnected or loosely connected entities without clear directional links, the effectiveness of HITS diminishes.
* **Undefined Entity Importance:** If the relationships between entities do not inherently indicate authority or hub status, HITS may not provide meaningful insights.
* **Non-Hyperlink Data Types:** HITS might not be the best choice for datasets where the data types do not involve hyperlink structures; alternative algorithms may offer more accurate results.

## Summary

The HITS algorithm identifies the importance of web pages in terms of their role as hubs (pages that link to other important pages) and authorities (pages that are linked to other important hubs). Pages with high Hub scores are good at linking to other pages, and pages with high Authority scores are linked from other significant hubs. The interpretation of these scores depends on the specific structure of the hyperlink network in the web data mining tested.

# References

Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Morgan Kaufmann.

Jiawei, H., Jian, P., & Micheline, K. (2012). Data Mining: Concepts and Techniques. Elsevier.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5), 604-632. <https://doi.org/10.1145/324133.324140>

University of Michigan. (2022, January 15) HITS Algorithm and HUBS and AUTHORITIES Explained. YouTube. <https://www.youtube.com/watch?v=-kiKUYM9Qq8>

Wang, C., Li, Y., Wang, Y., & Zhang, X. (2021). Identifying Influential Nodes in Social Networks Based on Hyperlink-Induced Topic Search. IEEE Access, 9, 51118-51128. <https://doi.org/10.1109/ACCESS.2021.3069919>

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. IEEE Transactions on Knowledge and Data Engineering, 26(1), 97-107. <https://doi.org/10.1109/TKDE.2013.109>