A Study of Link Analysis Using Components, Issues, Algorithms, Applications and Tools

[a]Arun K.C,[b] Dr.V.M.Ghodki, [c]Dr.S.B.Kishor and [d]B.K.Madhavi

[a]*Research Scholar , Gondwana University, Gadchiroli, Maharashtra, India*
[b]Associate Professor, Dept. of Computer Science, J.B.Science College, Wardha, Maharashtra, India
[c]Head, Dept. of Computer Science, Sardar Patel Mahavidyalaya, Chandrapur, Maharashtra, India
[d]Research Scholar, Gondwana University, Gadchiroli,

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: *Link Analysis is a data-analysis technique in Network theory used for evaluating relationships between nodes. The relationships between various types of nodes can be identified as people, property and transactions. This paper focuses on the study of Link Analysis components like Link generation, Explicit and Aggregate links and Infrared relationships. Also, this paper discusses Link Analysis Algorithm, like PageRank, Weighted PageRank and HITS. In addition to that, tools and applications used in Link Analysis research is also presented in this paper*.

Keywords: PageRank, HITS, Explicit, Aggregate, Weighted PageRank, Authority, Hub.

1.Introduction

Network science is a field which analyzes network data from cognitive, semantic, telecommunication, biological and computer networks by considering elements represented in the form of vertices and link between the elements are represented in the form of edges. This field is becoming popular because it draws on theories from mathematics, physics, computer science, statistical modelling and social structure from sociology. One obvious example of this field is the social network Facebook, in which nodes are represented in the form of persons and the links are represented in the form of relationships between those persons. Recent days, this field has grown tremendously where experts from various domains computer science, physics, maths, chemistry, and economists all started analyzing data. In computer science, the field is called as link analysis, and in physics, it is also known as network science. But both network science and link analysis is for analyzing and extracting information from entities like web pages, people, etc. That is, from a complex relational data. This paper focuses on link analysis, components, algorithms, research, applications and tools.

2 LINK ANALYSIS

In network science, link analysis is defined as a data-analysis method used for estimating the connections between nodes. The connections are recognized between different types of objects (nodes), that contains people, property and transactions. L. Getoor and C. Diehl [1], presented a survey on link mining. In which they have deliberated that various datasets of interest today are best described as a linked collection of interconnected objects. These could signify homogeneous networks, in which there is a single-object type and link type, or richer, heterogeneous networks, in which there could be multiple objects and link types (and probably other semantic information).

Example Heterogeneous networks are healthcare domain that includes patients, treatments and diseases, or bibliographic domains that includes publications and authors. Link mining refers data mining methods to consider these links for building descriptive or predictive models of the linked data.

Link analysis is a knowledge discovery that can be used to envisage information for enhanced analysis with respect to the background of links, whether it is related to Web or relationship links between various people and entities. It is mostly preferred in the optimization of search engine, security analysis and intelligence, medical and market research activities.

In addition to that, Link analysis is about examining the link among physical, digital or relational objects. This examination requires diligent data gathering. For example, in the website, all the links and its corresponding backlinks are to be studied and a tool has to be selected which examines codes and scripts in the page and discovers all the links to find out whether the existing links are active or dead? This process is quite essential for the optimization of search engine, as it allows the expert to check the search engine discovers and index the website or not.

In the field of networking, link analysis determines the integrity of the connection with respect to every network node by examining the data that is permitted over the virtual or physical links. Also, with the help of the data, experts discover problems and probable fault areas and are able to cover up rapidly or even provide to optimize the network.

Link analysis has three primary purposes and it is discussed in the following sections

2.1 Components of Link Analysis

There are two categories of Link analysis such as utilization of the resulting linkage graph and link generation.

A. Link Generation

Link generation is defined as a procedure that computes links, its attributes and node attributes. There are quite a few ways to define links but the key aspect in defining link is deciding which representation to use.

B. Explicit Links

An explicit link is referred to as a link that is created between the nodes in a transaction which corresponds to each pair of entities.

C. Aggregate Links

Aggregate links are referred to as a link that is created from multiple transactions and various links constitute together in a single aggregate link.

D. Inferred Relationships

Inferred relationships are referred as a link created between pairs of nodes with respect to inferred strengths of relationships between them. These are sometimes called as soft, association or co-occurrence links. Also, there are classes of algorithms for these computations which consists of context vectors, association rules and Bayesian belief networks. For instance, a link is made among any node pairs whose context vectors will lie within a certain radius of one another. Normally, a link's attribute shows the relationship strength it represents.

Inferred relationships time plays an important role that exposes linkages that could be lost due to many data analysis methods. For example, assume that temporal analysis is performed on wire transfer records state that a transfer from account X to person A at one bank is quite temporally proximate to the account transfer from an account Y to person B at another more bank. This gives you an inferred link between accounts X and Y. Sometimes, it is also identified for carrying out further inspections in money laundering activities.

In other words, inferred relationships are used to classify two nodes that resembles an identical physical entity, like an account or a person. Then with the help of Link analysis it is collaborated to a single node and this is done by creating rules or selecting parameters to make these nodes united to form a single node [2] [3] [4].

E. Utilization

Once the linkage graph is well defined by including the node and link attributes, it is utilized for creating variables which can be used as an input to a decision system.

2.2 Link Analysis Algorithms

The algorithms of Link Analysis provides information related to hyperlinks through which various webpages are connected. The World Wide Web (WWW) is seen as a direct labelled graph whose nodes are pages and edges are hyperlinks between those pages. This direct labelled graph structure is called as a web graph. There are number of algorithms based on link analysis and the three most prominent algorithms are below. That is, PageRank, Weighted PageRank and HITS.

2.2.1 PageRank

World Wide The PageRank algorithm was proposed by founders of Google, Larry Page and Sergey Brain at Stanford University in the year 1996[5]. This algorithm is utilized by Google search engine. The algorithm determines the page by calculating different pages are connected to it. This is called as backlinks. The rank score is revealed by page inlinks and the rank of contiguous pages [6] are selected by outinks. Also, the page's rank score is distributed among all outlinks, and the page possess a high rank if aggregate's backlink is high. So, the probability of a page visited by a web user with respect to random surfer model is called as PageRank (PR).

The PageRank's equation is specified by:

$$PR(x) = a \sum_{y \in M(y)} \frac{PR(y)}{N_y} \qquad (3.1)$$

Where,

d = damping factor between 0 and 1, frequently set to 0.85.

The theory of PageRank explains that any surfer who performs random click on links will stop clicking at one stage. So, the probability that a person will remain in the process is referred as a damping factor (d).

A probability distribution curve over the Web pages are formed by the PageRank and it can be computed mathematically by normalized eigenvector equations using iterative process.

2.2.2 Weighted PageRank

Weighted PageRank (WPR) algorithm [7] was developed by Wenpu Xing and Ali Ghorbani whose functions are quite similar to PageRank algorithm, but the rank score will depend on the web page's importance. Also the larger rank score is assigned to the important pages rather than it is divided equally among its linked pages [8] that are going outward. Here, each and every outgoing link will get a score that is proportional to its importance and the importance is attributed to weight values of a web page with respect to inlinks and outlinks. It is also denoted as W^{in} (x, y) and W^{out}(x, y) respectively.

W^{in} (x, y) represents the weight of link (x, y) that is based on the computation of the number of inlinks of page y and the number of orientations pages inlinks of page x.

$$\text{Win}(x,y) = \frac{I_y}{\sum_{m \in A(x)} I_m} \quad (3.3)$$

Where,
I_m, I_y = page m and y's inlinks
A(x) = page x's allusion page list
W^{out} (x, y) = Weight of link (x, y) that is computed based on the number of page y's outlinks and the number of all reference pages outlinks of page x.

$$W^{out}(x,y) = \frac{O_y}{\sum_{m \in B(x)} O_m} \quad (3.4)$$

Where,
O_m, Oy = page m and y's outlinks.
B(x) = page x's allusion page list
The WPR's proposed equation is given as:

$$WPR(y) = (1-d) + d \sum_{x \in A(y)} WPR(x) W^{in}(x,y) W^{out}(x,y) \quad (3.5)$$

Where,
d = damping factor between 0 and 1 and it is frequently set to 0.85.
(1-d) = page rank distribution from pages which are not linked directly in order to avoid some page ranks loss. The recursive equation is proposed. In other words, it is computed using an initial value along with iterating the computation until all the values are converged.

2.2.3 HITS

HITS stands for Hypertext Induced Topics Search, an algorithm developed by Jon Kleinberg [9]. This algorithm is used for rating webpages and it is divided into pages like Hubs and Authorities.
Hubs pages will act as a resource list, consists of good source of links and authorities are pages that contains good source of content. It is a fact that good hub page is referred to a page that points authoritative pages with the same content with the same content and a good authoritative page is a page that points to many good hubs. So, a webpage can be of both types that shows mutual relationship.
The HITS algorithm considers WWW (World Wide Web) in the form of directed graph in which V represents a set of vertices and E is a set of edges with respect to hyperlinks which are used for linking different webpages [5]. The HITS algorithm assigns two scores in the each page of its authority that estimates the value of the page contents whereas its hub in turn estimates the value of links which are attached to other pages.

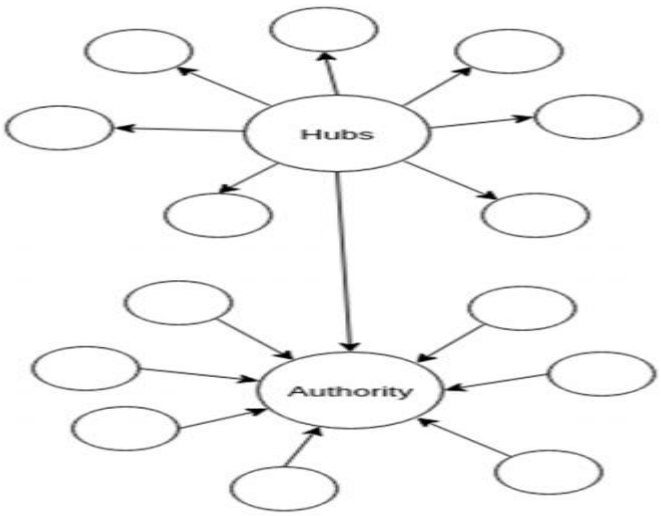

Figure 3.1. Hubs and Authority

2.2.3.1 Authorities and Hubs Rules

In order to begin the ranking, hub(s) is taken as 1 and auth(s) is also taken as 1, where s is a page.
Updates are of two types, Hub and Authority Update Rule. In order to calculate hub and authority rule, update rules are applied to the repeated iterations. For instance, an n-step application of the HITS algorithm applies n times the Authority Update Rule first and then the Hub Update Rule.

Authority Update Rule:
∀s, we update auth(s) to be:

$$\sum_{j=1}^{m} hub(j) \qquad (3.6)$$

Where
m = total number of pages linked to page s and page j which is linked to s. Therefore, the score page's Authority is the summation of Hub pages score at that particular point [5].

Hub Update Rule:
∀ s, we update hub(s) to be:

$$\sum_{j=1}^{m} auth(j) \qquad (3.7)$$

Where
m = is the total number of pages linked to s, and j is a page which s is inked to. Thus it is concluded that a page's Hub score is the sum of the Authority scores of all its linking pages [5].
Next, normalization step is applied and after repeated iteration algorithms, the final authority-hub scores of nodes are identified. This repeated iterations tend to diverge the values of the Authority and Hub Update Rules. Therefore, it is a must for matrix normalization after each and every iteration. So, whatever values obtained from this method will eventually converge.

2.4 Comparison of Link Analysis Algorithm
This section shows (Table 3.1) the comparative analysis of Page Ranking Algorithms using various parameters such as the main technique used, methods employed, complexity, parameters used, etc.

Algorithm	HITS	Weighted PageRank	Page Rank
Mining technique	Content Mining and Web structure	Web structure	Web structure
Employed methods	Hub and authority score for each web page is computed	Page rank score at index time is computed.	Page rank score at index time is computed as well.
input parameters used	Back & forward links and content	Back & forward links	Back links
Complexity of the Algorithm	Less than O(log N)	Less than O(log N)	O(log N)
Result Relevancy	High	Low	Low
Quality of Result	It is low	It is High	It is Medium
Outcomes importance	Hub & authority values are utilized moderately	Pages are sorted High according to the	Importance is High due to use of back links

		importance	
Limitations	Efficiency problem	Query-independent is its limitation	Query-independent
Search Engine	Clever	Google	Google

Table 3.1. Comparison of Page Ranking Algorithms

2.5 Research, Applications and Tools

In the late 1990s, several workshops were organized to bring AI and link analysis communities together. There was a workshop titled AI Approaches to Fraud Detection and Risk Management [10], [11] organized in the year 1997.In this workshop lot many papers were presented the idea of link analysis. This workshop was followed by a symposium titled Artificial Intelligence and Link Analysis [12], which brings a direct focus on using AI techniques to linked data.

In the year 1998, there was a workshop hosted by Carnegie Mellon University on the topics Data Mining, Machine Learning and Knowledge Discovery [13]. The workshop recommendations were grouped into few areas like active, incremental, cumulative and learning by using prior knowledge. All these areas have been studied in the context relational data as well as proportional data. In other words, in the context of Link Analysis.

In the year from 1995 to 2005, there were many workshops on the topics like statistical models from relational data, statistical relational learning, link analysis for detecting complex behavior, link analysis for group detection, link discovery and on multi-relational data mining. This clearly shows that research work has been carried out a lot in the field Network theory.

Also, when it comes to Link Analysis applications, rapid increase in the research activities of link analysis in late 1990s paved a way for the appearance of many applications from data mining to link analysis. The application FinCEN AI System (FAIS) attempted to combine automated detection with link analysis techniques. [14] In FAI's system the patterns that represent potential money laundering were not derived from automated data mining, instead it was derived from expert consultation. Also, the system's computational load was focused toward the report consolidation of transactions by enterprise, people and an account. Analysts were able to generate potential leads by applying patterns of suspiciousness to these enterprise, people and accounts

Other applications like combating cellular telephone fraud [15], [16] and improper activity detection in the Nasdaq Stock Market [17]. In Research and Development, link analysis techniques are used in probabilistic relational models [18], and graph-based data mining [19].

When it comes to Link analysis tools, there are quite a number of tools used for searching nodes, links, and groups of nodes and links based on connectivity structure. Also, the tool supports searching nodes and links based on graph theory and social network metrics. Some advanced tools support graphical displays, temporal evolution views and larger datasets. A recent link analysis tool also detects multi-link paths based on common data values in a large volume of databases and data streams.

3 CONCLUSION

In this paper, the introduction to Link Analysis and the components of link analysis such as link generation, explicit links, aggregate links, and inferred relationships utilization are discussed. The Link Analysis issues are presented with the help of real-life example. Link analysis algorithms like PageRank algorithm, Weighted PageRank, HITS are also discussed and comparison is also made between them in terms of Mining technique, Complexity, employed methods, result relevancy, quality of result and limitations. In addition to that, Link Analysis techniques in research activities, its applications and tools are also discussed in this study.

REFERENCES

1. Rozyyev, A., Hasbullah, H., & Subhan, F. (2011). Indoor child tracking in wireless sensor network using fuzzy logic. Research Journal of Information Technology, 3(2), 81-92.
2. Singh, G., & Kapoor, I.V. (2017). Performance evaluation of Zigbee routing protocols using NETSIM simulator. International Journal of Advanced Research in Computer Science. 8(3): 852-855.
3. Biswas, P. K., & Phoha, S. (2006). Self-organizing sensor networks for integrated target surveillance. IEEE Transactions on Computers, 55(8), 1033- 1047Lee, L. T., & Chen, C. W. (2008). Synchronizing sensor networks with pulse coupled and cluster based approaches. Information Technology Journal, 7(5), 737-745.

4. Sabri, N., Aljunid, S. A., Ahmad, B., Yahya, A., Kamaruddin, R., & Salim, M.S. (2011). Wireless sensor actor network based on fuzzy inference system for greenhouse climate control. Journal of Applied Sciences, 11(17), 3104-3116.
5. Yick, J., Mukherjee, B., & Ghosal, D. (2008). Wireless sensor network survey. Computer networks, 52(12), 2292-2330.
6. Arampatzis, T., Lygeros, J., & Manesis, S. (2005, June). A survey of applications of wireless sensors and wireless sensor networks. In Intelligent Control, 2005. Proceedings of the 2005 IEEE International Symposium on, Mediterrean Conference on Control and Automation (pp. 719-724). IEEE. Tseng, Y. C., Pan, M. S., & Tsai, Y. Y. (2006). Wireless sensor networks for emergency navigation. Computer, 39(7), 55-62.
7. Gama, J., Rodrigues, P. P., & Lopes, L. (2011). Clustering distributed sensor data streams using local processing and reduced communication. Intelligent Data Analysis, 15(1), 3-28.
8. Aghbari, Z. A., Kamel, I., & Awad, T. (2012). On clustering large number of data streams. Intelligent Data Analysis, 16(1), 69-91.
9. Chi, Y., Wang, H., Yu, P. S., & Muntz, R. R. (2004, November). Moment: Maintaining closed frequent itemsets over a stream sliding window. In Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on (pp. 59-66). IEEE.
10. Gama, J., Ganguly, A., Omitaomu, O., Vatsavai, R., & Gaber, M. (2009). Knowledge discovery from data streams. Intelligent Data Analysis, 13(3), 403- 404.
11. George, B., Kang, J. M., & Shekhar, S. (2009). Spatio-temporal sensor graphs (stsg): A data model for the discovery of spatio-temporal patterns. Intelligent Data Analysis, 13(3), 457-475.
12. Mahmood, A., Shi, K., & Khatoon, S. (2012). Mining data generated by sensor networks: a survey. Information Technology Journal, 11(11), 1534-1543.
13. Rabatel, J., Bringay, S., & Poncelet, P. (2009, July). SO_MAD: SensOr mining for anomaly detection in railway data. In Industrial Conference on Data Mining (pp. 191-205). Springer, Berlin, Heidelberg.
14. Guralnik, V., & Haigh, K. Z. (2002, July). Learning models of human behaviour with sequential patterns. In Proceedings of the AAAI-02 workshop "Automation as Caregiver (pp. 24-30).
15. Huang, S., & Dong, Y. (2007). An active learning system for mining time- changing data streams. Intelligent Data Analysis, 11(4), 401-419.
16. Beringer, J., & Hüllermeier, E. (2007). Efficient instance-based learning on dat streams. Intelligent Data Analysis, 11(6), 627-650.
17. Spinosa, E. J., de Leon, F., Ponce, A., & Gama, J. (2009). Novelty detection with application to data streams. Intelligent Data Analysis, 13(3), 405-422.
18. Puccinelli, D., & Haenggi, M. (2005). Wireless sensor networks: applications and challenges of ubiquitous sensing. IEEE Circuits and systems magazine, 5(3), 19-31.
19. Mainetti, L., Patrono, L., & Vilei, A. (2011, September). Evolution of wireless sensor networks towards the internet of things: A survey. In Software, Telecommunications and Computer Networks (SoftCOM), 2011 19th International Conference on (pp. 1-6). IEEE.