Comparative Study of Ranking Algorithms for Network Analysis

Sarthak Kulkarni*, Pritam Bagad*, Hrishikesh Joshi*, Himanshu Randad*

*B.E., Information Technology Department, AISSMS Institute of Information technology, Pune, Maharashtra, India

Abstract - Key Opinion Leaders - Key opinion leaders, who are subject experts also sometimes considered as influencers are trusted and well respected in the community because of their knowledge and expertise in the research field. As these experts can have enormous impact while decision making for important subjects, it is of utmost importance to identify the right key leaders. While building on this idea, our project goal is to create and provide a platform that will be helpful to find these key leaders. Our platform is integration of different modules which are described in detail in this paper. Additionally, the paper consists of I) literature survey II) Drawbacks of different approaches which are already in practice and importance of nodes according to its position and neighborhood III) Different network analysis Algorithms with their implementations.

Introduction

While developing a new drug it goes through various phases including research, drug discovery, preclinical research, clinical research and finally drug approval. All these phases require to be led by a panel of experts. These panels can be hard to form and can be biased using the traditional methods such as interviewing, self identification, and observational methods. Key Opinion Leaders are the people who can govern, influence the decisions of these healthcare companies. In this paper we propose a method to identify the key opinion leaders using a centrality algorithm to analyse the network of experts and also comparison of these methods with other methods like page rank algorithm, K-shell decomposition etc.
The platform comprises of various modules which are as follows -

Module 1 - In this module we take the input of the keyword from the user about the area in which KOL needs to be found. Once the keyword is extracted then the keyword is forwarded to the next module.

Module 2 - It involves scraping of data for the keyword on the central database for medical research of Europe PubMed central. Keywords are provided as a filter for papers and then the required information is extracted and stored in the KOL database.

Module 3 - Centrality algorithms (Degree centrality, Closeness Centrality, Betweenness centrality, Eigenvector Centrality) are applied to the database formed from extraction of paper detail from European PubMed. The result is then filtered using the frequency of occurrence of author in each centrality algorithm.

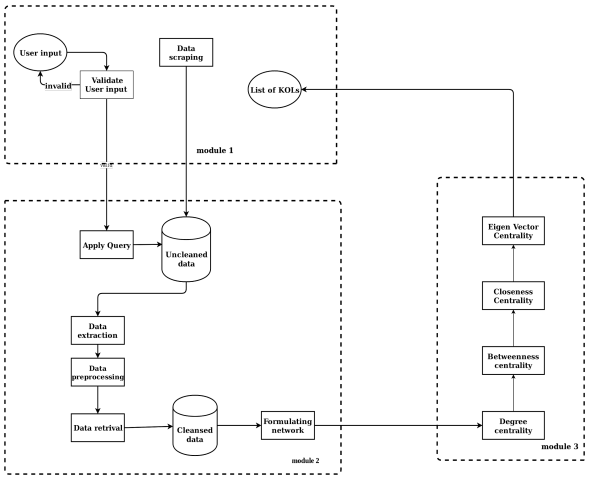

Fig 1. Block Diagram of system

Analysing different Ranking algorithms with respect to author network

a. Page Rank-

Page Rank is a modification to Katz's Centrality measure. There are various factors taken into consideration for determining the Page rank of a node.

1. The number of links it receives i.e the in-degree of a node
2. The link inclination of the linkers
3. The centrality of the linkers

The first point seems more obvious. The more a node is endorsed by other nodes the more prior it becomes. The value of the endorsement depreciates proportionally to the number of links given out by the endorsing node: links coming from parsimonious nodes are worthier than those emanated by spendthrift ones [1]. The links coming from valuable vertices are more important than others. So, to summarize Page rank -

"*A node is said to be important if it is connected to other important nodes or if the node is highly linked*"

The problem with page rank is if a node is connected to a highly ranked node the page rank of that node will be more. This characteristic does not favor while identifying key opinion leaders. Consider the diagram given below

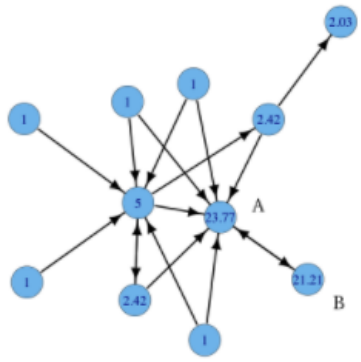

Fig 2. Page rank of various nodes

Notice how node B got a high page rank only because it was connected with A. This result is not appreciated in the case of the network of authors.

One author node should not be ranked high only because it's connected to another highly ranked author. Individual contribution to the network should be considered.

Another problem with page rank is link farms. A link farm is a website or node created for the sole purpose of boosting the page rank of the main page. In the following two scenarios the page rank of the main page is very good even though the average page rank is very low.

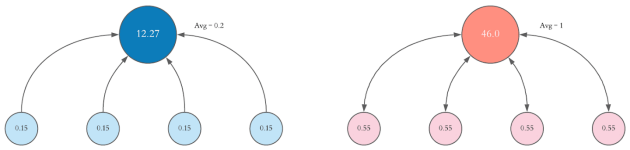

Fig 3. Links Farm

b. HITS Algorithm

Hyperlink-Induced Topic Search also known as Hubs and authorities is primarily a link analysis algorithm. In this the link network is divided into two sets, one is called the root set and other is the base set. Root contains all the nodes or pages that can be potential authority and Base contains the pages or nodes that are linked to root in some way and are potential hubs. Unlike the page rank algorithm which looks at network as whole and then analyses the network to find the highest ranked page, HITS algorithm look at a subset of the network looking first at subset of the network i.e the roots and then the base this process is iterated to and each node is provided with two scores hubs and authority and at every iteration the score is normalised.

This method might seem relevant in case of finding KOL's but it suffers through some caveats.

- One major caveat, namely topic drift, the neighbour graph could contain nodes having high authority scores having a topic unrelated to the original query. Also, the web graph constructed from the pages in the base set, will not have the most relevant nodes and as a result the algorithm will not be able to find the highest ranked authorities and hubs for a given query.

- Since HITS is a query-based algorithm, time taken for query evaluation is expensive.

- A situation might occur where a page containing links to a large number of different topics may receive a high hub rank and is not relevant to the provided query. Even though this page is not the most relevant source for any information, it has a very high hub rank if it points to highly ranked authorities.

New approach

The above given algorithm takes under consideration a single attribute for example - indegree of a node, number of links going out of the node to decide the popularity of the node. Even though this attribute should be taken under consideration, this is not enough for ranking the key opinion network . As this network can have various characteristics i.e number of citations for a given author/paper, number of co-authors in a particular paper etc. Therefore there is a need to consider a method for accessing multiple attributes of the network.

a. Using Centrality Algorithm

We have applied standard centrality algorithms using Python's NetworkX functions. While applying the standard centrality algorithms on our weighted network we have performed some small set of modifications in the basic centrality algorithms. We have applied the distance (number of edges) between two nodes as a sum of reciprocal of the weight of each edge present between those two nodes.

1. Degree centrality algorithm

In case of the degree centrality algorithm, we can calculate the degree of each node in the network. Degree is nothing but with how many nodes that particular node is directly connected to. Its mathematical formula can be given as follows:

$$C_D(i) = k(i) = \Sigma_j A_{ij} = \Sigma_j A_{ij} \quad (1)$$

Where,
i, j are nothing but nodes i. e. authors A is an edge from i to j.

We can normalize this formula in order to see proper comparison between all the nodes i.e. all authors that which author have more degree centrality. We can normalize these values between 0 and 1 as follows:

$$C_D(i) = \left(\frac{1}{n-1}\right) C_D \quad (2)$$

Where,
n is the total number of authors in a network.

2. Closeness Centrality Algorithm

In case of closeness centrality algorithm, we can calculate minimum distance of each node from every other remaining n-1 nodes in the network and do the sum of all. Means distance of particular node from remaining all other nodes can be calculated here. Its mathematical formula can be given as follows:

$$C_C(i) = \frac{1}{\Sigma_j d(i,j)} \quad (3)$$

Where,
i is node whose closeness centrality we are calculating. d (i , j) is minimum distance between node i and j.
It can be normalized in the range 0 to 1 as follows:

$$C_C(i) = (n-1)\, C_C \quad (4)$$

Where,
n is the total number of nodes (authors).

If clusters are present in network and network is disconnected then distance between two disconnected nodes will become infinite in this case. In order to overcome this problem we can use following mathematical formula to calculate closeness centrality of node. It is known as Harmonic Centrality algorithm.

$$\text{Harmonic Centrality} = \Sigma_j \frac{1}{d(i,j)} \quad (5)$$

3. Betweenness Centrality Algorithm

In case of betweenness centrality we calculate the sum of the minimum distance between two nodes if the path between these two nodes is passing through the node for which we are calculating betweenness centrality and divide it by the total number of shortest paths present between earlier two nodes. Its mathematical formula can be given as follows:

$$C_B(i) = \frac{\Sigma_{s \neq t \neq i} \sigma_{st}(i)}{\sigma_{st}} \quad (6)$$

Where,

i is a node.
s, t are another two nodes.

σ_{st} are the total number of shortest paths between nodes s and t.

$\sigma_{st}(i)$ are the total number of shortest paths between nodes s and t passing through node i.

4. Eigenvector Centrality Algorithm:

In this algorithm we first formulate adjacency matrix A (as we have given in formula) n*n matrix A and then we consider vector v of having dimension n*1. Vector v contains values of each node in terms of their influence; greater value nodes will have greater influence in the network. And one dumping factor is also used in order to control increasing vector v values after each iteration. We keep on doing iterations till it converges to some value. Its mathematical formula can be given as follows:

$$v(i) = \frac{1}{\lambda}\left(\lambda_j A_{ij} V_j\right) \qquad (7)$$

Where,
A is adjacency matrix.
V(i) is vector.
λ is dumping whose value is generally equal to 0.85

b. Case Study for Keyword – Diabetes

We implemented our approach for the keyword 'Diabetes' to find the Key Opinion Leader in that network. The output provided was then cross validated against the author's publicly available data and scopus database.

Information of author was collected on basis of following parameters:

- H-Index (Scopus)
- Citation for the author
- Number of Co-Authors the author has worked with

Algorithm	Avg. h-index (Scoupus)
Degree	83.2
Betweenness	68.4
Closeness	63.4
EigenVector	99.6
Page Rank	61
HITS	74

Author	H-index Scopus	Citation
Murray CJ(Christopher J L Murray)	175	191526
Remuzzi G(Remuzzi, Giuseppe)	168	157481
Benjamin EJ(Benjamin, Emelia J)	156	162303
Salomon JA	88	87083
Vos T(Vos Theo)	80	26966
Lozano R(Lozano, Rafael)	77	82591
Shibuya K(Kenji Shibuya)	75	83721
Cushman WC	66	67185
Lackland DT(Lackland, Daniel T)	64	64817
Carnethon MR(Mercedes R Carnethon)	58	28560

c. Analysing the evaluation

On implementing centrality measures on the author data, we considered the top 10 authors from each of the centrality measures. On removing the redundant authors we found most of the authors had a very good h-index. According to the paperpile.com, an h-index of 60 and above is considered "remarkable". Each author has a lot of work experience and has a lot of achievements (referred from Impactstory.org). Many of the authors held leadership positions in important organizations related to the medical field also most of

the leaders provided in page rank and HITS were covered in our output.This validates our results and thus they can be considered as key opinion leaders.

CONCLUSION

In this approach, we have formed a network of authors using the publicly available data. We applied various centrality measures such as degree centrality, betweenness centrality, eigenvector centrality, closeness centrality for finding the KOL's. The results obtained are cross-verified by the scopus h-index value and thus they are accurately validated as KOL's.

REFERENCES

[1] https://www.sci.unich.it/~francesc/teaching/network/pagerank

[2] https://www.slideshare.net/shatakirti/pagerank-and-hits

[3] https://en.wikipedia.org/wiki/HITS_algorithm

[4] https://www.irjet.net/archives/V4/i12/IRJET-V4I1251.pdf

[5] Rekha Jain, Dr G. N. Purohit, "Page Ranking Algorithms for Web Mining," International Journal of Computer application, Vol 13, Jan 2011.

[6] S. Brin, and L. Page, "The Anatomy of a Large Scale Hypertextual Web Search Engine", Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998.

[7] Prof. Anuja Phapale, Sarthak Kulkarni, Pritam Bagad, Hrishikesh Joshi, Himanshu Randad, "Identification of Key Opinion Leaders in Pharmaceutics Using Network Analysis", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 8 Issue 3, pp. 01-05, May-June 2021.

[8] "Finding Key Opinion Leaders Using Social Network Analysis - A New Paradigm in Key Opinion Leader Identification," LNX Research second white paper in a series 2007.

[9] "Identifying Key opinion leaders using social network analysis," Cognizant System Private Limited's White Paper on June 2015.

[10] Pooja Devi, Ashlesha Gupta and Ashutosh Dixit, "Comparative Study of HITS and PageRank Link based Ranking Algorithms," International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 2, February 2014.